Statistical analysis of agricultural soils climate data to aid food security under environmental change

Emily Grace Mitchell

Thesis submitted to The University of Nottingham

for the degree of Doctor of Philosophy

November 30, 2021

Acknowledgements

First of all, I would like to thank my supervisors Professor Andrew Wood, Dr Gilles Stupfler, Professor Neil Crout and Professor Paul Wilson for their continuous support, encouragement and patience throughout my PhD and allowing me the opportunity to work with them. I would also like to thank Dr Karthik Bharath for stepping in and supervising me over the critical final years of my PhD. The knowledge I have gained from them is immense and I am extremely grateful to have worked with them all. I would also like to express my gratitude to the Leverhulme Trust for providing me with the financial support without which this PhD would not be possible.

I would like to thank the Probability and Statistics group in the School of Mathematical Sciences and statisticians further afield for making my PhD such a pleasurable experience. I would like to thank the Modelling and Analytics for a Sustainable Society (MASS) community and Prof Markus Owen for allowing me to be part of a fantastic group of scholars, both mathematically and socially.

I would also like to thank Conal, to whom has had to put up with me writing a thesis during a global pandemic, and to my brother Jack for the inspiration to pursue a career in Statistics. Finally, I would like to thank my parents - Nichola and William, for their endless love and encouragement.

Abstract

Wheat is one of the most important food crops in the world for human consumption, livestock feed and biofuels. Demand for wheat has increased due to a rising population and crop growth concerns resulting from a changing climate. By exploring novel uses of data gathered on farming practices from the Farm Business Survey, this thesis aims to identify key farming practices which are most associated with high yields.

The first part of this thesis is concerned with modelling wheat yield based on a linear combination of data from the Farm Business Survey, such as annual crop protection costs, labour costs and organic status of the farms, and data from the UK Met Office, such as annual monthly rainfall. We compute coefficient estimates in the linear model using quantile regression, linear regression and principal component regression. We also take a two-step approach by fitting a linear regression model after selecting variables based on either forward stepwise regression, with and without orthogonalisation after every step, or Lasso regression. Variable selection methods consistently select organic status, crop protection and rainfall in June to be included in the model first. Comparing all models based on their mean squared prediction error for an average year, we find that a model created based on linear regression applied to a subset of variables selected with forward stepwise regression with orthogonalisation after every step achieves the smallest mean squared prediction error. This model included the majority of the variables corresponding to farming practices and a small number of weather conditions.

To account for the uncertainty at both the variable selection stage and the parameter estimation stage, focus is next shifted to Bayesian shrinkage priors as a means of simultaneous model selection and inference. If uncertainty is only accounted for after

variable selection, the confidence intervals of the coefficient estimates will be unrealistically narrow and lead us to be overconfident about our estimates. The Bayesian Lasso, which is the analogue of the frequentist Lasso, and the horseshoe prior provide credible intervals for the parameters of the linear model. In order to apply these shrinkage priors, we use the Gibbs sampler when the global shrinkage parameter is allowed to vary and Hamiltonian Monte Carlo when the global shrinkage parameter is fixed. We find that these methods also consistently select organic status, crop protection and rainfall in June to be important factors when modelling wheat yields. However, the horseshoe prior finds appropriate credible intervals capturing the combined uncertainty of the model selection and parameter estimation stages for these factors which the two-step frequentist approach aims to account for, but fails to do.

The second part of this thesis is specifically concerned with modelling the highest yields under current technologies and growing conditions. We address this by performing an extreme value analysis, which in our context translates to modelling the highest-yielding farms. We find that wheat yields have an upper finite bound estimated at 17.60 tonnes per hectare and therefore the scope to improve yields for high-yielding farms diminishes when yields per hectare approach this bound. Furthermore, we find there is no difference between the maximum attainable yields for macro-regions west England and Wales, north England and east England. Lastly, we show that the difference between the maximal yields of medium and high spenders on crop protection and fertilisers is not statistically significant.

Relevant publications

Much of the content in Chapter 4 of this thesis has previously appeared in the following publications:

- E.G. Mitchell, N.M.J. Crout, P. Wilson, A.T.A. Wood, and G. Stupfler. Operating at the extreme: estimating the upper yield boundary of winter wheat production in commercial practice. *Royal Society Open Science*, 7(4):1-12, 2020.

The work contained in Chapters 2 and 3 is currently being prepared for submission to a peer-reviewed journal.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Wheat production

Wheat is one of the most important food crops in the world. Current global annual production levels of wheat stand at 756.8 million tonnes (FAO (2018)), two-thirds of which is used for human consumption in food staples such as bread. As a result of sustained improvements to crop varieties and agricultural technology, there was a progressive and very large increase in wheat yields over the second half of the last century (Calderini and Slafer (1998)). Despite this, there are concerns for the future growth of crop yield, the main one arguably being climate change. Recent literature has focused on forecasting the behaviour of crops in a changing climate (Asseng et al. (2013), Challinor et al. (2009), Iizumi et al. (2017), Olesen and Bindi (2002)), and found that a global temperature increase may lead to a yield reduction in cereal crops in certain regions. At the same time, current projections point to major increases in demand for food and livestock feed, as well as rising demand for biofuels due to a progressive shift of major economic powers to generating energy via renewable sources (Spiertz and Ewert (2009)). Understanding the factors associated with crop yield is of crucial importance to successfully address the challenge of global food security.

We look to understand the factors associated with UK wheat yields with two different perspectives. The first part of this thesis investigates factors associated with an average yield. Although association does not necessarily imply causation, the identification of factors largely associated with yields can provide actionable information on which farming practices may improve wheat yields in England and Wales. The second part looks at

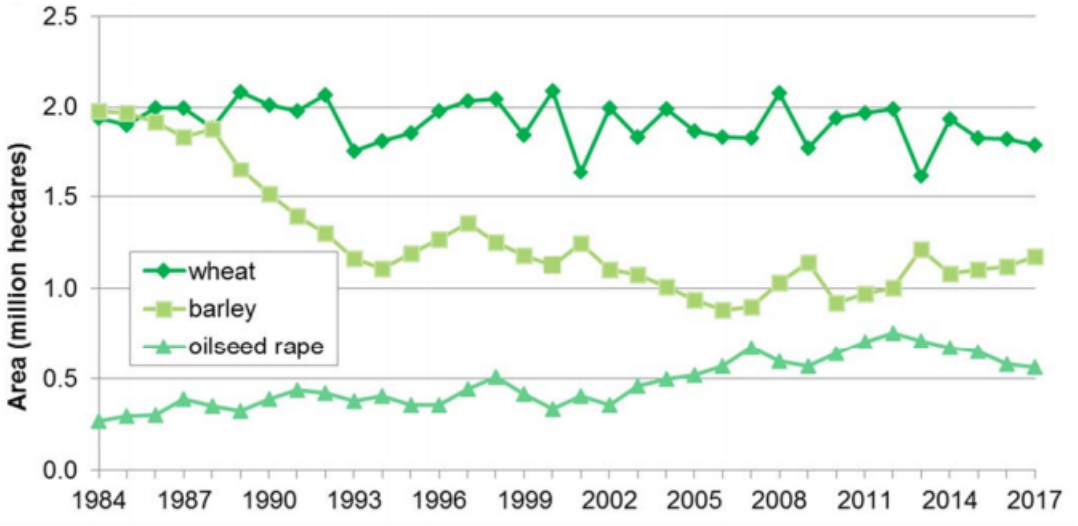


Figure 1.1: Crop areas in the UK between 1984 and 2017.

modelling the highest yields under current technologies and growing conditions. Figure 1.1 shows crop areas have remained approximately constant over the last two decades, therefore if wheat yields have stagnated, then a rising population may not be fully catered for. This thesis looks to address these with novel uses of the Farm Business Survey data, about which we give details in the next section.

1.2 Data

1.2.1 Farm Business Survey

The Farm Business Survey collects information about farm businesses in England and Wales, to give a yearly overall perspective of the agricultural and economic performance of farms. Each year, approximately 2300 farms take part in the survey. On average, 695 were involved with the production of winter wheat from 2006 to 2015, each reporting 76 observed variables, among which were yield per hectare, region, and fertiliser and crop protection costs. To be able to share our results, we must ensure anonymity of the farms in the survey. Therefore no variables can be used in our analysis such that the location of the farm can be determined. Table 1.1 describes the variables to be extracted from the Farm Business Survey, including those used when modelling yield per hectare.

Throughout the thesis, yield per hectare will be modelled instead of total yield. If total

Variable	Description
`Yield`	Yield (tonnes/hectare)
`Net margin`	Net margin (£/hectare)
`Seeds`	Seeds costs (£/hectare)
`Fert`	Fertiliser costs (£/hectare)
`Sprays`	Crop protection costs (£/hectare)
`OtherVC`	Other variable costs (£/hectare)
`Fuel`	Fuel costs (£/hectare)
`Labour`	Labour costs (£/hectare)
`Contract`	Contract work costs (£/hectare)
`Machinery`	Machinery costs (£/hectare)
`TOFC`	Total other fixed costs (£/hectare)
`LAND`	Land costs (£/hectare)
`UAA`	Utilised agricultural area
`Organic`	Organic status
`Education 1`	GCSE level or equivalent education
`Education 2`	A level or equivalent education
`Education 3`	College level or equivalent education
`Education 4`	Degree level or equivalent education
`Education 5`	Postgraduate level, Apprenticeship and other education

Table 1.1: Variables in the Farm Business Survey

yield was modelled to assess performance, this would be determined by the size of the farm rather than how productive a farm is regardless of its size.

To take inflation into account, the financial data are adjusted to their 2010 equivalent (DEFRA (2018)). Analysis on this dataset has been scarce. Stochastic frontier analysis has been conducted to model the variation of technical inefficiency based on managerial objectives of UK potato producers (Wilson et al. (1998)) and winter wheat producers in the east of England and Wilson et al. (2001) find there is little potential for eastern England farmers to improve their technical efficiency. Since this thesis is concerned with identifying key farming practices which are most associated with high yields rather than taking a business perspective, we shall not be concerned with this kind of analysis. Ritchie (2015) combines variables from the Farm Business Survey, including those in Table 1.1, and monthly UK Met Office data in a forward stepwise algorithm to identify important variables for modelling yield and compares the results to a mechanistic model, however their work does not consider other regression or variable selection techniques to construct the parameter estimates. Furthermore, they do not critique the variable selection algorithm and assess the model fit for yields departing from the average (i.e.

high and low yields).

In our analyses, we shall also use the variables from the Farm Business Survey in Table 1.1 alongside the following Met Office data.

1.2.2 UK Met Office

The Met Office provides daily reports on weather and climate conditions in the UK to inform the public such that precautions can be put in place to minimise risk. Monthly averages of mean daily temperature ($^{\circ}C$), daily rainfall (mm) and daily sunshine hours can be extracted from their high-dimensional database. If monthly weather conditions for each individual farm location are used in our analyses and identifies key weather conditions which may be associated with a larger yield, this will not preserve the anonymity of these farms. To ensure anonymity of these farms, the monthly averaged climate variables are also averaged over each of the Nomenclature of Territorial Units for Statistics 1 (NUTS1) administrative regions, such as east England, south west England and north east England, in Figure 1.2. If the number of farms located within each of these regions are too small to perform analyses based on, the data will first need to be cleaned to accommodate for this by either removing these regions (as in Chapters 2 and 3) or aggregating these regions (as in Chapter 4). The following section looks at any data cleaning required prior to performing the analyses in Chapter 2 and 3.

1.2.3 Data cleaning

Data cleaning involves preparing the data in advance of performing the analyses such that our modelling is not biased towards these anomalies. Data cleaning is required for our analyses Chapters 2 and 3, whilst no data cleaning is required for Chapter 4.

First, a very small proportion of farms taking part in the Farm Business Survey are in Wales. To avoid Welsh farms having a large influence on our linear model, these farms are removed from the analyses in Chapters 2 and 3. Farms in Wales are not an issue for Chapter 4 because we combine the NUTS1 regions into larger regions according to whether they are in west England and Wales, east England and north England to ensure

Figure 1.2: Administrative subdivision of the UK in NUTS1 regions (source: Met Office).

our sample sizes within each region are sufficiently large.

Furthermore, any farm spending considerably more on aspects of farming practices, out of line with similar farm businesses in the Farm Business Survey is also removed in Sections 2 and 3. Again, this is to avoid certain farms having a large influence on our modelling when looking at typical yields.

1.3 Aims of the thesis

This thesis aims to explore novel uses of the dataset from the Farm Business Survey on farming practices alongside the UK Met Office dataset on weather conditions to focus

on two main problems:

- Determine which farming practices or environmental conditions are most associated with an average yield. This will involve fitting linear models using different minimisation criteria and assessing their predictive performance. We shall assess the importance of variables in modelling yields according to their t-statistic for regression methods and their order of entry for variable selection techniques. This thesis will also aim to fully account for parameter uncertainty in our best fitting model. This will involve estimating the coefficients in the linear model instead using Bayesian shrinkage priors.
- Determine whether a finite upper bound for wheat yields exists and estimate this upper bound, to act as a target for wheat producers and indicate whether there is room to improve on their current yield. This will involve modelling the high yields between 2006 and 2015 and not as a curve of maximum yield through time. We also compare the finite upper bound estimated under various scenarios of location and input use.

1.4 Structure of the thesis

Chapter 2 concerns determining important variables for modelling typical wheat yields using methods familiar to non-specialists but yet to be used with the Farm Business Survey data. Quantile regression and linear regression are applied in Section 2.4.1, where their coefficients are compared to ensure we can use the normality assumption of the residuals. Section 2.4.2 applies principal component regression by modelling yields based on a linear combination of the covariates which captures the largest variation. Rather than applying standard regression techniques, Sections 2.6.1 and 2.6.2 apply variable selection techniques instead, namely forward stepwise and Lasso regression. Models at all steps of the variable selection method are compared and we end with a comment on the model which achieves the smallest mean squared prediction error.

Chapter 3 looks to use Bayesian shrinkage priors to find credible intervals for variables selected which will account for the uncertainty associated with parameter estimation and

the uncertainty associated with variable selection. If uncertainty is only accounted for after variable selection, the confidence intervals of the coefficient estimates will be unrealistically narrow and lead us to be overconfident about our estimates and overconfident about our predictions based on the parameter estimates. For the amount of shrinkage both varying and fixed, we apply the Bayesian Lasso (Sections 3.3.4 and 3.3.5) and the horseshoe prior (Sections 3.5.3 and 3.5.4) and comment on their behaviour when shrinking coefficients in Section 3.4. We comment on the posterior predictive distributions for yield for each Bayesian shrinkage prior in their respective applications.

Chapter 4 focuses on modelling the highest yields in an extreme value framework to find a target yield per hectare for wheat producers to indicate whether there is room to improve on their current yield. Section 4.3.2 finds the maximum attainable yield. Sections 4.3.3 and 4.3.4 find the maximum attainable yield conditional on geographical regions and agricultural inputs. Furthermore, we look at what happens for net financial margins in Section 4.3.5. Final remarks are made in Chapter 5.

2 | Exploratory data analysis, regression and variable selection

2.1 Introduction

UK crop yields are known to be influenced by the climate and the farming practices employed. Predictions for crop yields often involve mechanistic models based on complex interactions of biophysical processes e.g. Vanuytrecht et al. (2014) and Mirschel et al. (2014), however these do not give a sense of how each factor in the model marginally influences wheat yield. Complex simulation models have been used in the past to predict outputs (Absalom et al. (2001), Walter and Heimann (2000)), yet more recently these have been found to be over-parameterised (Crout et al. (2009), Cox et al. (2006), Crout et al. (2014), Tarsitano et al. (2011), Gibbons et al. (2010)). Although these do not attempt to predict wheat yields, these suggest a suitable initial model for wheat yields would be to start with a linear model and compare to models with increasing complexity. Non-linear regression models have also been studied for maize yields (see Hawkins et al. (2013)), but these too can lead to overparameterised systems.

Lobell and Burke (2010) combined farming practices of Sub-Saharan Africa maize farmers and simulated weather conditions in a regression model to simulate maize yields and assess the impact climate change will have on these yields in the future. Qian et al. (2009) used stepwise regression to select variables based on weather conditions and simulated soil properties, such as water content, at key growth stages of spring wheat in the Canadian prairie province. These studies recognise a model for yields must incorporate weather

conditions and farming practices, however both consider crops in a different climate to that experienced in the UK.

Landau et al. (1998) and Landau et al. (2000) both predicted UK wheat yields using a statistical regression framework on monthly climatic data in the UK, with the latter incorporating climate-yield mechanistic models from other studies, however neither include factors representing farming practices. More recently, Nkurunziza et al. (2020) used socio-ecological factors to determine crop performance of spring barley in Sweden using partial least squares however do not take account of the climate in which the barley is grown. Ritchie (2015) performed forward selection on farming practices from the Farm Business Survey and monthly weather conditions from the UK Met Office to assess their influence on wheat yields, however lacked in critique of the model fit for yields departing from the average, such as high or low yields.

We will use a variety of parameter estimation and variable selection techniques for the coefficients in the linear model using monthly weather conditions from the UK Met Office and the agronomic and socio-ecological factors from the Farm Business Survey to model per hectare wheat yields, critique the statistical methods and comment on their model fits.

This chapter consists of the following: Section 2.2 looks at the distribution and the marginal influence of variables on yield per hectare. Sections 2.3 and 2.4 discuss and apply parameter estimation methods for the linear model. We will model the conditional median using quantile regression (Section 2.3.1), the conditional mean using linear regression (Section 2.3.2) and apply principal component regression (Section 2.3.3) to predict wheat yields. In the second half of this chapter, we look at forward stepwise regression (Section 2.5.2) and Lasso regression (Section 2.5.3) to model yield per hectare based on a select few covariates.

2.2 Exploratory data analysis

We first analyse the variables to be included in our linear model for yield per hectare, perform data pre-processing, and comment on their marginal influence on yield and the

dependencies amongst the variables. This will ensure the distributional assumptions of the methods in Section 2.3 hold and give an insight of which variables the variable selection methods in Section 2.6 will most likely select for our model.

2.2.1 Wheat yields

The distribution of wheat yields per hectare recorded in the Farm Business Survey (Figure 2.1) has a slight negative skew with a median of 8.006 tonnes per hectare. Figure 2.2 shows the distribution of yields per hectare annually from 2006 to 2015. In terms of productivity, 2012 was a year of low wheat yields as a result of poor weather conditions, according to the UK Department for Environmental, Food and Rural Affairs (DEFRA) report (see DEFRA (2012)). In 2015, DEFRA (2015) reported wheat yields reached their highest level since 1990; that year, the crops benefited from optimal growing conditions during the spring and summer months.

Extreme cold temperatures were experienced during the winter of 2009 (see Prior and Kendon (2011)) when the winter wheat sprouts are underground, but there are no reports of further exceptional weather conditions during the rest of the growing year. Comparing the yields per hectare in other years to those attained in 2009, Figure 2.2 suggests yields per hectare in 2009 are representative of typical yields. When assessing the model fit, we shall look to predict yields for 2009 given this chapter is concerned with linearly modelling a typical yield.

Yields between farms are assumed to be independent and since farms are allowed to enter and leave the survey, it is also reasonable to assume the yields within each farm are also independent. The following section looks at agronomic and socio-ecological factors from the Farm Business Survey to be used in our model for a typical wheat yield.

Figure 2.1: Violin plot for per hectare yields in the Farm Business Survey.

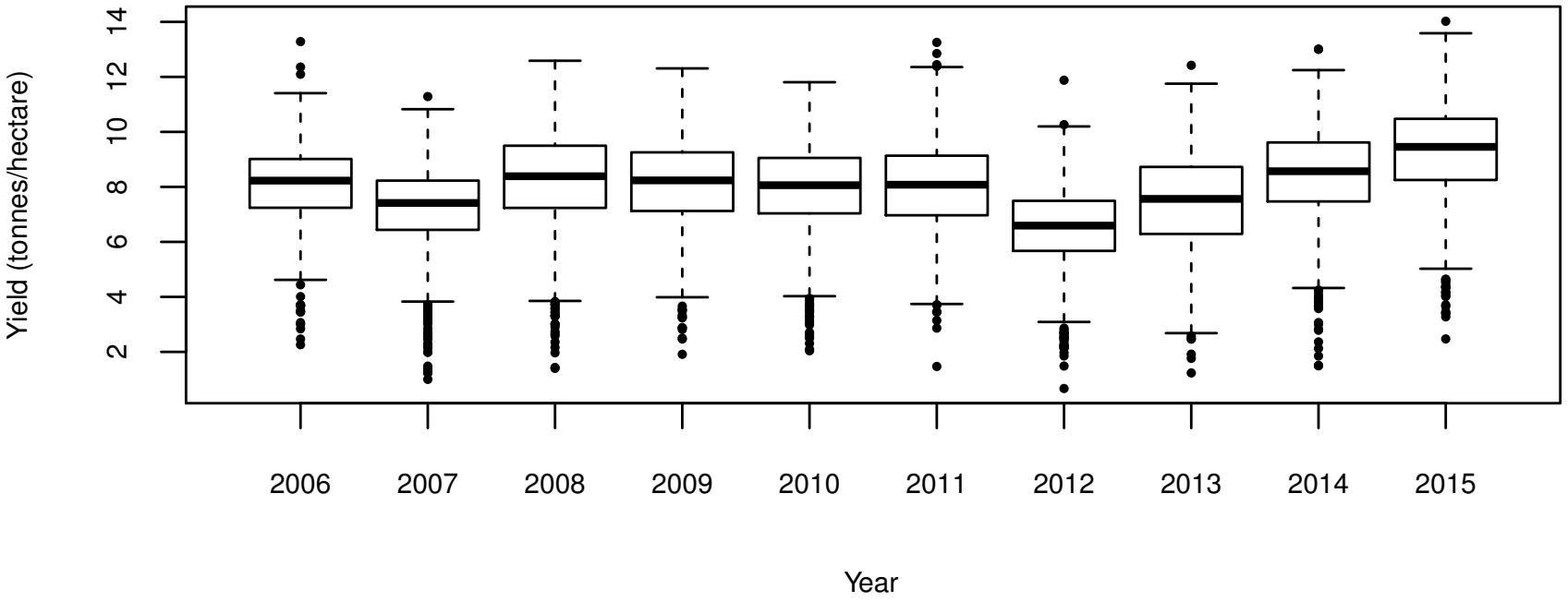


Figure 2.2: Annual yield boxplots using the Farm Business Survey data.

2.2.2 Farm Business Survey variables

Section 1 discussed the data available on farming practices from the Farm Business Survey to be incorporated in a linear model for wheat yield. This section takes a closer look at the variables the model will be based on. We also consider whether any variables would be more suited under a transformation to model yield, or whether an interaction term should be included to capture the relationship between two variables.

First we look at the densities of the variables as they appear in the Farm Business Survey and their marginal influence on yield. The marginal densities will give an insight into the spread of the data for each of the variables to be included in the model and the range of values they take. These variables will be referred to as the main effects henceforth.

The use of fertilisers and crop protection are correlated as a consequence of farmers either using intensive amounts of each in the hope of attaining a larger yield or adopting organic farming practices and using neither fertiliser or crop protection. Figure 2.3 shows the distributional similarities between fertiliser use and crop protection. Both density plots have a positive skew with peaks approximately at their respective medians and another at zero. Further to modelling with fertiliser and crop protection costs, an indicator variable for the farms' organic status (i.e. whether a farm uses crop protection) is created to account for the two peaks identified in the violin plots. Once stratified by organic status for yield per hectare, Figure 2.4 indicates wheat yields are generally much smaller for organic farms compared to conventional farms.

Figure 2.3 shows all other continuous variables taken from the Farm Business Survey also have a positive skew. Any farms spending considerably more on aspects of their farming practices, out of line with similar farm businesses in the survey, will be removed when fitting linear models in Sections 2.4.

One final variable proposed to use in the model for yield is the education status of the farmer. There are 6 possibilities for educational status: school only, GCSE or equivalent, A level or equivalent, college, degree or other. The other category captures those who have studied to postgraduate level, learnt through an apprenticeship, or through an unconventional pathway. The boxplots in Figure 2.5 suggest there is no noticeable

Seeds	-0.0927	LAND	0.0748
Fert	0.2621	UAA	0.1240
Sprays	0.3680	Organic	-0.4000
OtherVC	0.0834	Education 1	-0.0358
Fuel	0.0276	Education 2	-0.0079
Labour	0.1219	Education 3	0.0363
Contract	-0.0070	Education 4	0.0282
Machinery	0.0684	Education 5	-0.0065
TOFC	-0.0096		

Table 2.1: Correlations between yield per hectare and variables from the Farm Business Survey to be incorporated in the linear model as main effects.

difference between the 25%, 50% and 75% quantiles of yield for each education status.

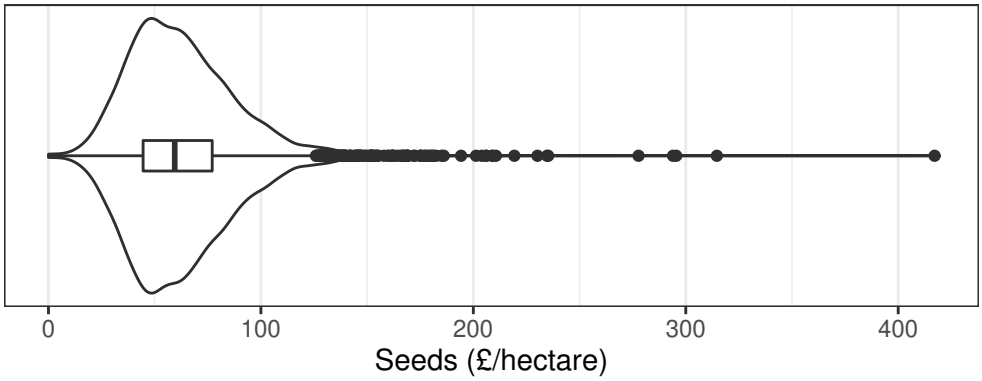
Calculating the correlation between each of the covariates discussed in this section and yield will give an indication of which variables marginally influence yield and are most likely to be included in our model. Table 2.1 contains the Pearson's correlation product moment coefficients

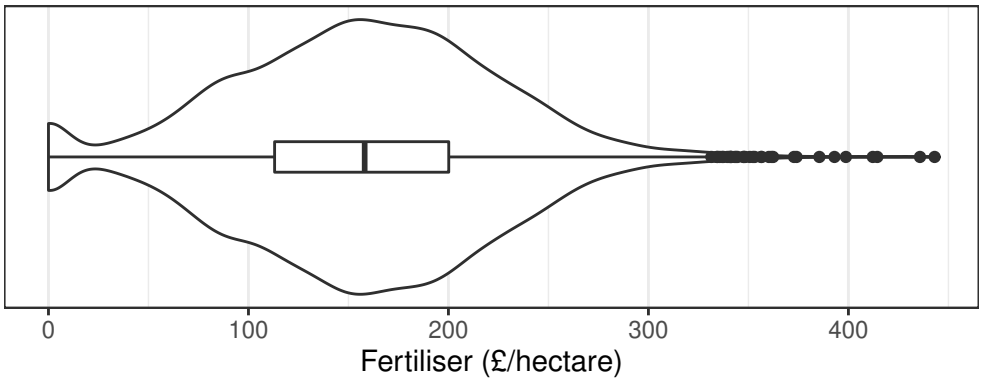
$$\rho_{x_j y} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

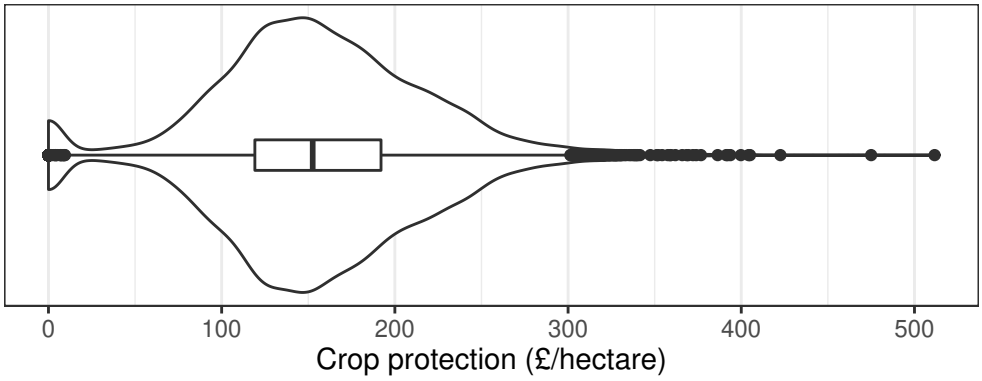
calculated for each variable $\boldsymbol{x_j}$ proposed from the Farm Business Survey and yield per hectare $\boldsymbol{y}$. The Point-Biseral correlation coefficient for the categorical variables is calculated in the same manner, provided they are first decomposed as variable indicators for each category. When modelling yield per hectare, it would be reasonable for the linear model to be largely based on the variables with a large correlation, in absolute value, with yield.

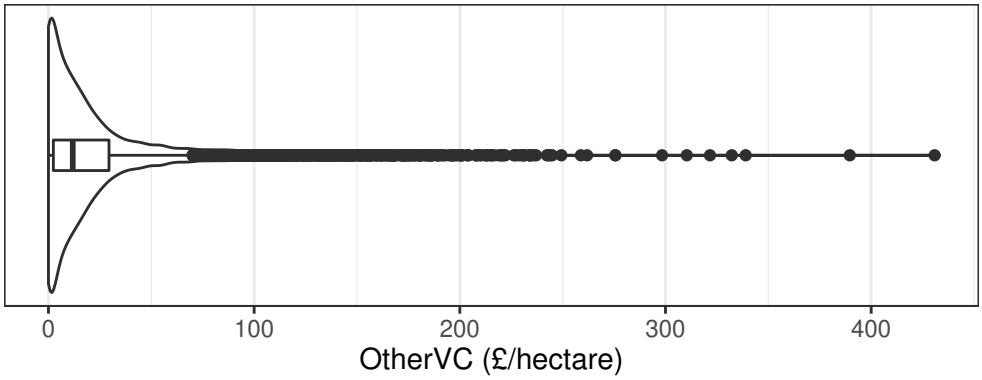
Most variables have approximately no correlation with yield per hectare, however fertilisers, crop protection, labour and utilised agricultural area have positive correlations with yield per hectare. Organic status is negatively correlated with yield as already seen in Figure 2.4. Although both fertiliser and crop protection are correlated with the yield, they are also correlated with one another. Partial correlation finds the correlation between the response and a variable, given another variable is already included to model the response. The partial correlation of a covariate $\boldsymbol{x}$ and the response $\boldsymbol{y}$, controlling for covariate $\boldsymbol{z}$, is found by calculating the correlation between the residuals from linearly regressing $\boldsymbol{z}$ onto $\boldsymbol{y}$ and linearly regressing $\boldsymbol{z}$ onto the variable of interest $\boldsymbol{x}$. Conditional

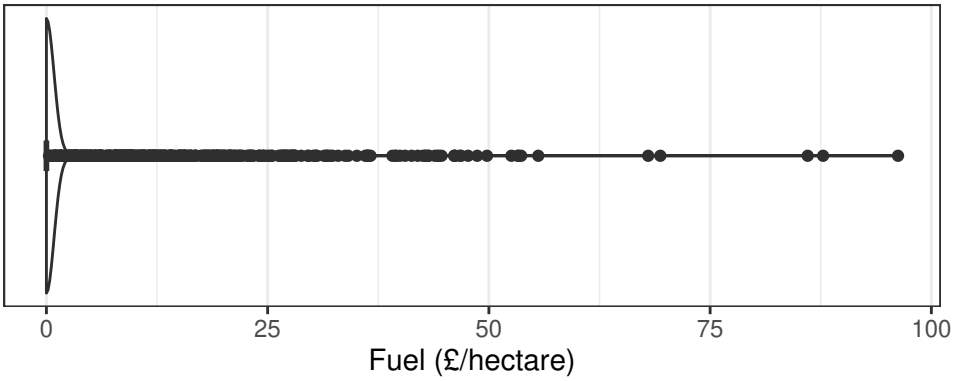
on crop protection already being in the model, the partial correlation between fertiliser and yield per hectare is 0.1194, which is less than the correlation with yield per hectare without considering the influence of crop protection. We shall see whether both variables appear individually in our model for yield when using variable selection techniques on the main effects in Section 2.6.

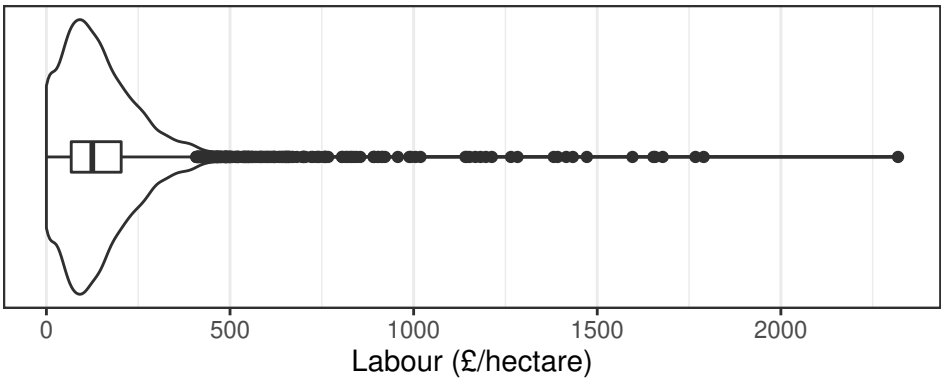

0
100
200
300
400
Seeds (£/hectare)

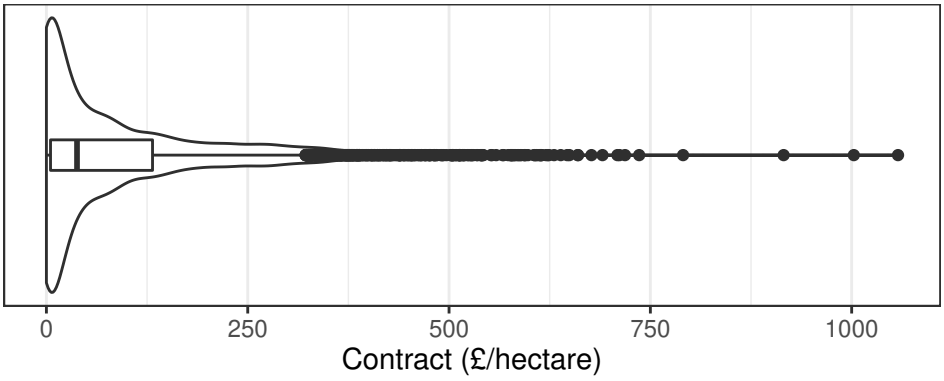

0
100
200
300
400
Fertiliser (£/hectare)

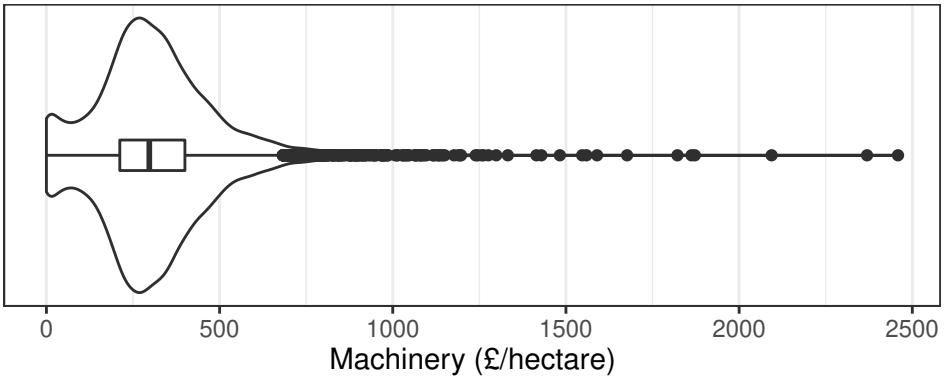

0
100
200
300
400
500
Crop protection (£/hectare)


0
100
200
300
400
OtherVC (£/hectare)


0
25
50
75
100
Fuel (£/hectare)


0
500
1000
1500
2000
Labour (£/hectare)


0
250
500
750
1000
Contract (£/hectare)


0
500
1000
1500
2000
2500
Machinery (£/hectare)

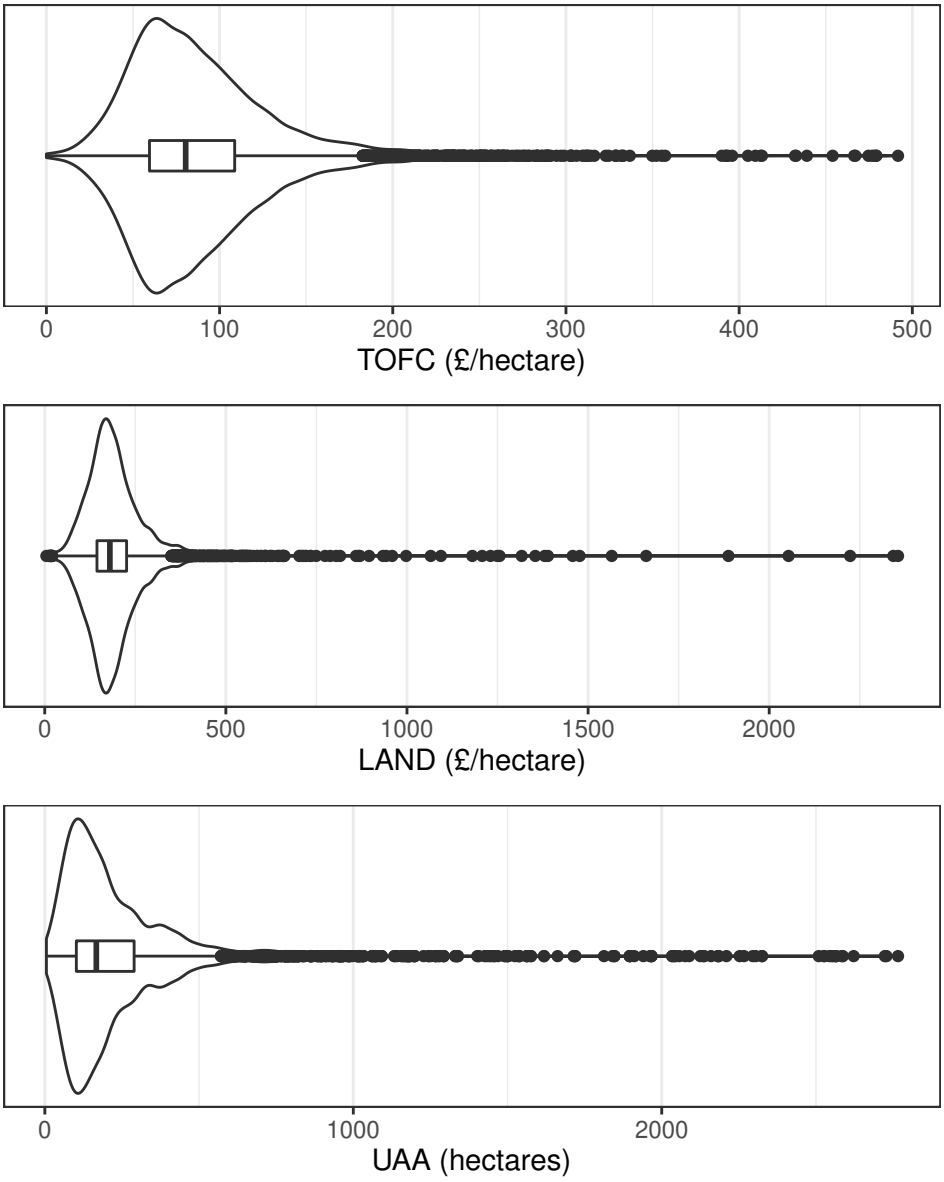


Figure 2.3: Violin plots of the continuous Farm Business Survey variables.

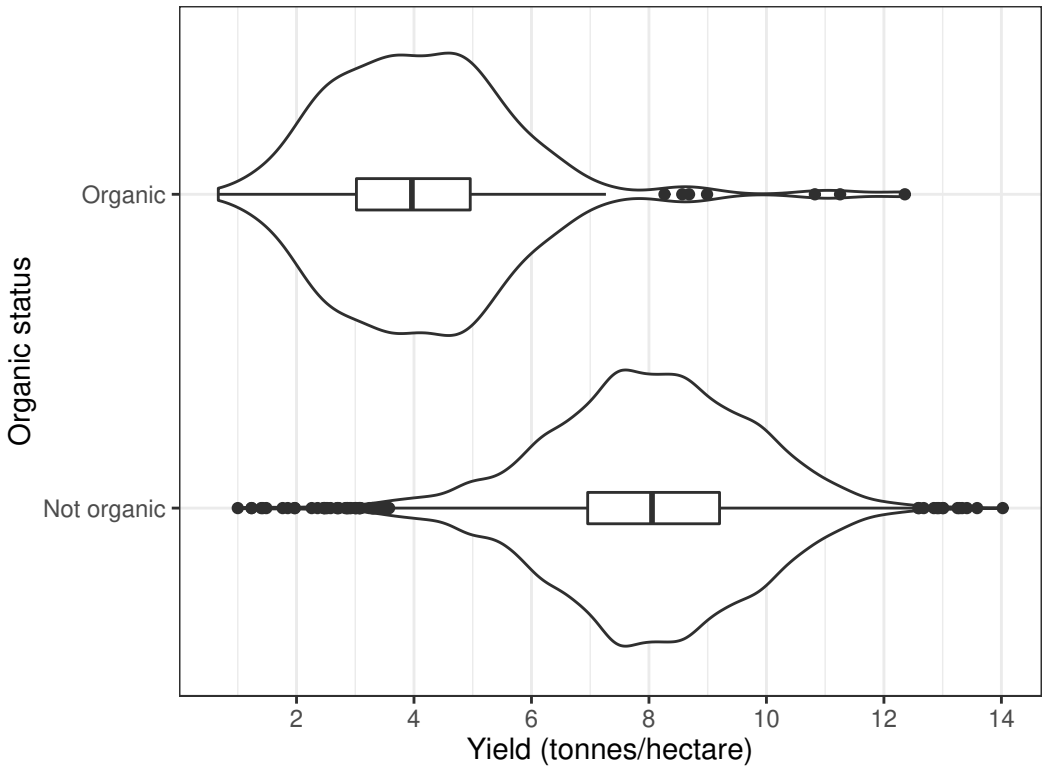


Figure 2.4: Violin plots of yield per hectare stratified according to organic status.

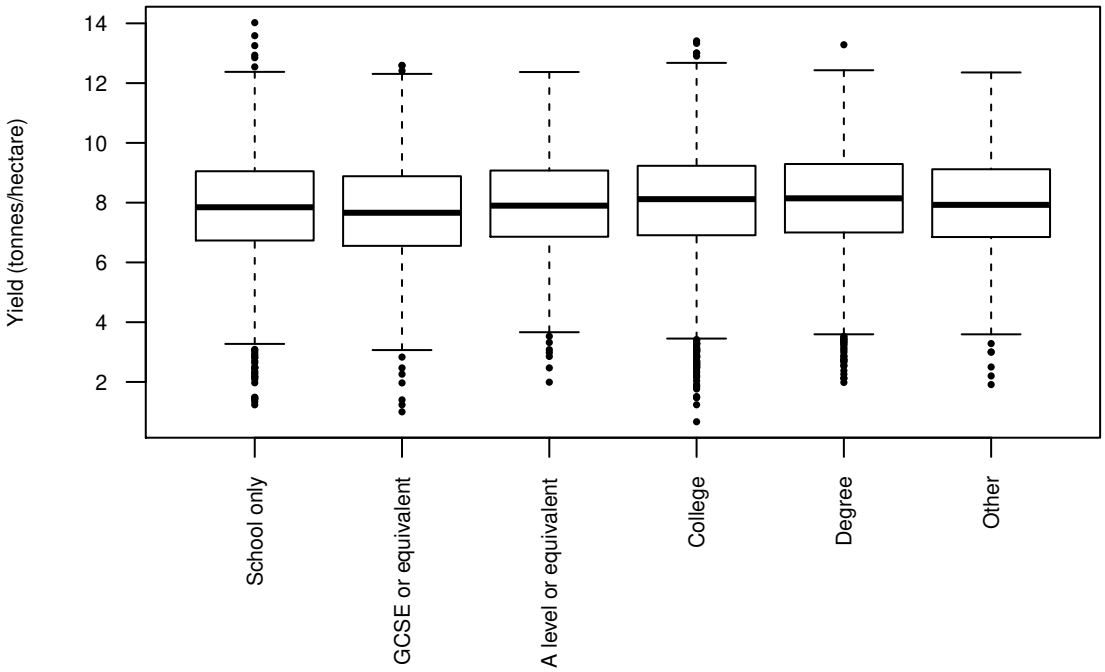


Figure 2.5: Boxplots for yield per hectare stratified about education status.

So far the relationship between the variables from the Farm Business Survey and yield per hectare has been assumed to be linear. Here, we look to see if there are any which could have a non-linear relationship with yield.

Figure 2.6 shows fertiliser costs, crop protection costs and utilised agricultural area have the strongest non-linear terms with yield per hectare. A sharp increase in yield per hectare is seen for small fertiliser costs and small utilised agricultural areas but the increase in yield per hectare becomes negligible as both fertiliser costs and utilised agricultural area increase beyond £100 per hectare and 500 hectares respectively. This may indicate fertiliser contributes to a larger yield per hectare overall however the increase in benefit diminishes as the amount of fertiliser increases. As for utilised agricultural area, Figure 2.6 shows the contrast of yield attained by small scale farms and conventional industrial farms, yet once the utilised agricultural area is already above 1000 hectares, little benefit will be gained from increasing the area of the farm beyond this.

Overall crop protection costs also increase yield per hectare, however when crop protection costs per hectare exceed approximately £300 per hectare, yield per hectare begins to decrease again. This could be a result of applying an excessive amount of crop protection since it is one of the only factors the farmer can control to try and improve yields yet there may be other factors that can not be controlled which will also influence yield.

Table 2.2 shows the Pearson's correlation product moment coefficients between yield and these non-linear effects, taken to be squared terms. The correlation between yield and the squared terms for fertiliser, crop protection and utilised agricultural area, 0.1892, 0.2634 and 0.0786 respectively, are smaller compared to the correlation between yield per hectare and their respective linear terms previously investigated.

To capture these non-linear relationships with yield per hectare, we can incorporate squared terms into another linear model for fertiliser costs, crop protection costs and utilised agricultural area with the main effects already discussed. In addition to non-linear terms, the linear model may also benefit from including interaction terms for those variables who have strong relationships amongst themselves.

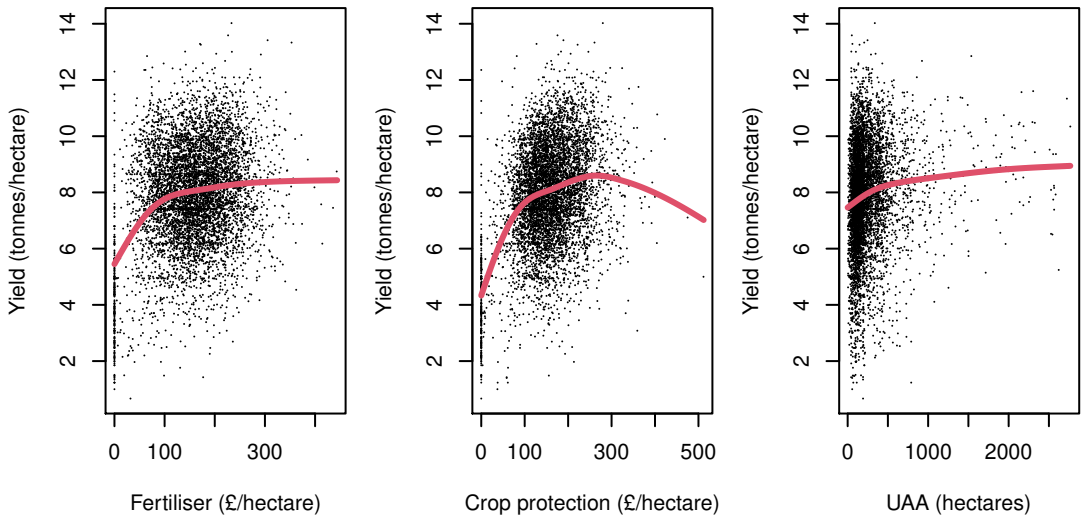


Figure 2.6: Plots of yield per hectare versus fertiliser costs per hectare, crop protection costs per hectare and utilised agricultural area with general trend superimposed.

Figure 2.4 showed there is a striking difference in yield per hectare between organic and non-organic farms. Here we look to see whether there are any differences in the Farm Business Survey variables between organic and non-organic farms that should be taken account of in our linear model. Figure 2.7 shows the difference in seed costs, fertiliser costs and utilised agricultural area between organic and non-organic farms. Organic farms may spend more on seeds by spreading more in an attempt to gain a larger crop due to not applying crop protection. If a farm is organic, then they will apply very little fertiliser compared to conventional farms and in general much smaller in size.

To find further interaction terms to include, the Pearson's correlation product moment coefficients in Equation 2.1 can again be calculated between each of the continuous variables with $y_i = x_{ik}$ for $k \neq j$. Here, we only look for the correlations between variables which are amongst the largest (above 0.2 in absolute value) to ensure the number of model parameters remains on a practical scale. The pairs of variables selected to include an interaction term between are crop protection and fertiliser costs, and each of the pairwise correlations between labour, contract costs and machinery costs. From the violin plots in Figure 2.3, we anticipated there would be a relationship between crop protection and fertiliser costs and that is confirmed here. The pairwise correlations between contract, labour and machinery costs could not be seen from the previous violin plots, however

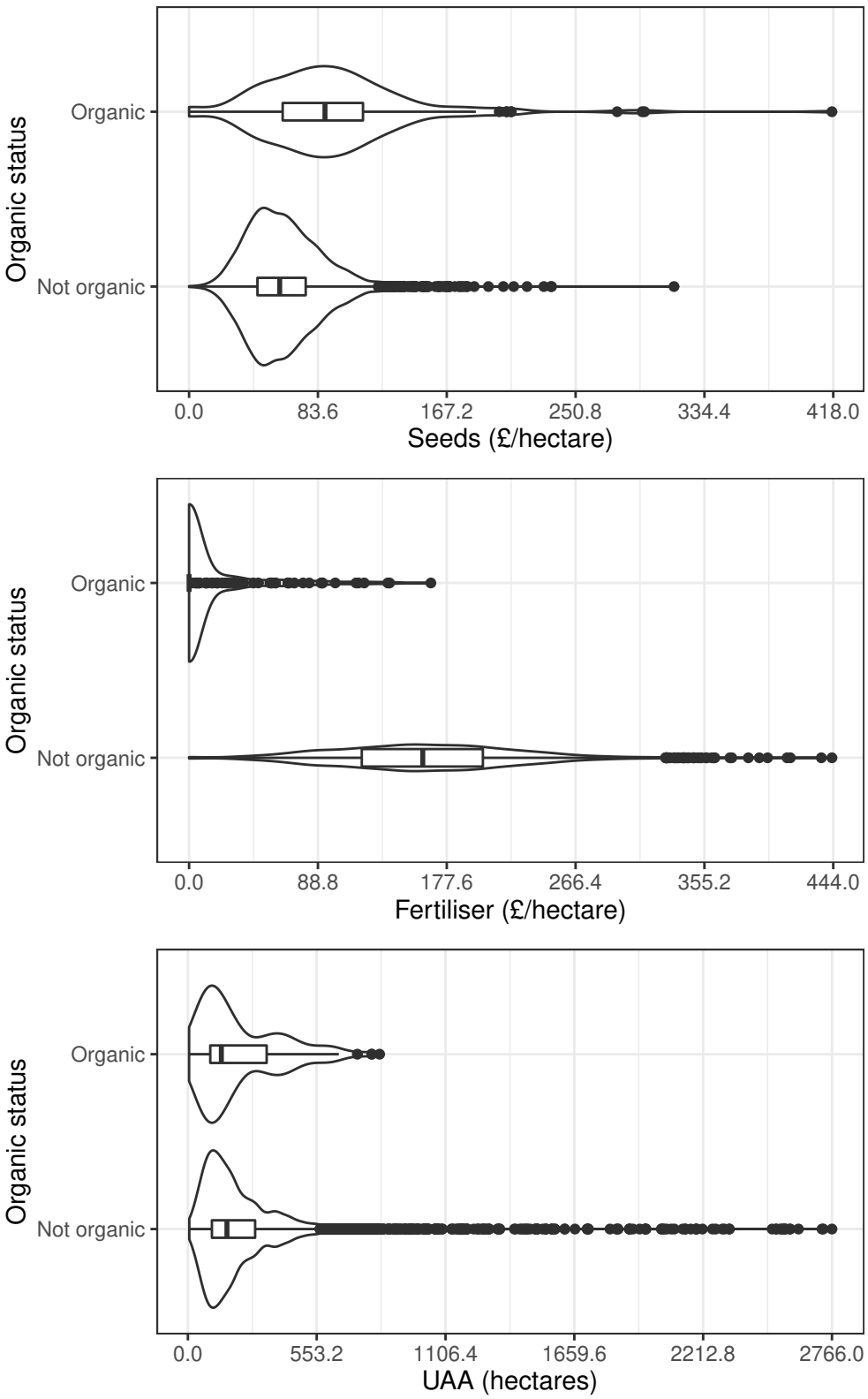


Figure 2.7: Violin plots of seed costs per hectare, fertiliser costs per hectare and utilised agricultural area stratified according to organic status.

the scatterplots in Figure 2.8 show the relationships between these. As a farm increases their labour and machinery costs, they are less likely to use contractors to undergo work on the farm since they already have the equipment and labour. Furthermore, as labour costs increase, machinery costs may also increase due to requiring sufficient equipment for the workforce.

Looking at the Pearson's correlation product moment coefficients between yield per hectare and these interaction terms in Table 2.2, the largest correlations with yield, in absolute value, are the interaction terms between seed costs and organic status and utilised agricultural area and organic status. Marginally, these are associated with yield per hectare, however we shall see in Section 2.3 whether these are still associated with yield once main effects are accounted for.

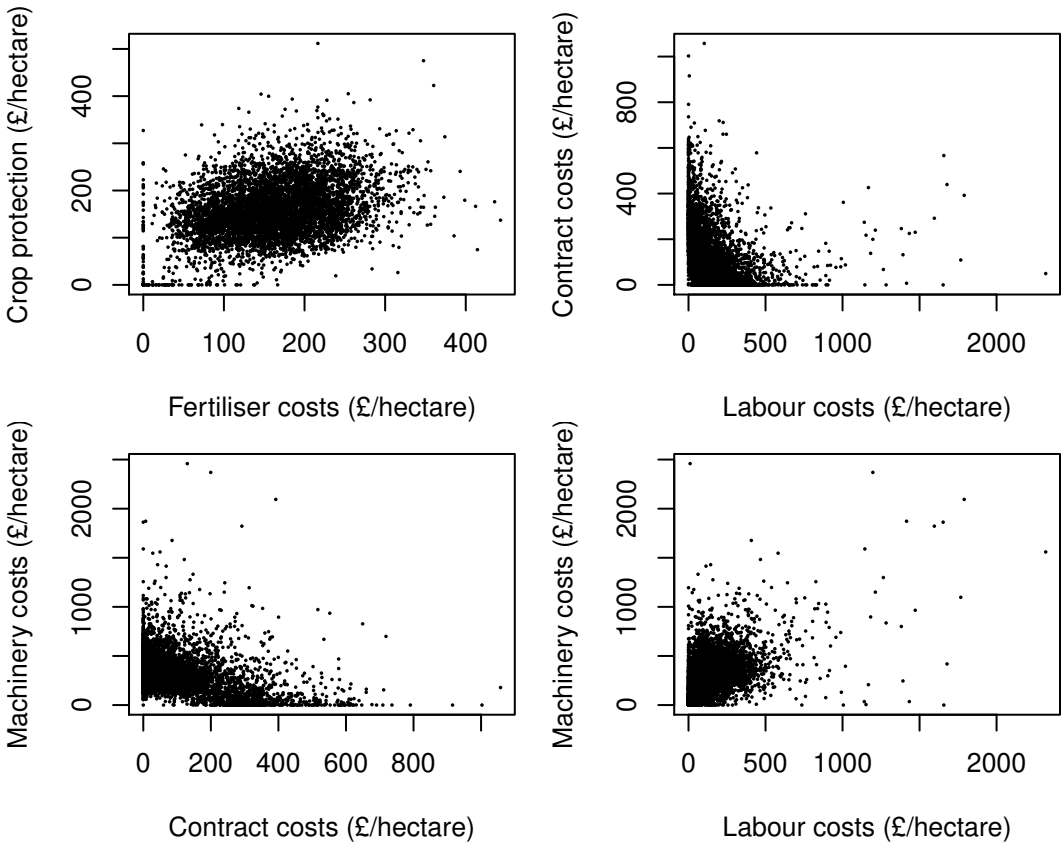


Figure 2.8: Plots of continuous variables with correlation coefficients larger than 0.2 in absolute value. Top left: fertiliser costs against crop protection costs, top right: labour costs against contract costs, bottom left: contract costs against machinery costs, bottom right: labour costs against machinery costs.

Seeds*Organic	-0.3532	Labour*Machinery	0.0611
Fert*Organic	-0.1368	Contract*Machinery	0.0076
UAA*Organic	0.3151	Fert*Fert	0.1892
Fert*Sprays	0.2820	Sprays*Sprays	0.2634
Labour*Contract	0.0466	UAA*UAA	0.0786

Table 2.2: Correlations between yield and non-linear and interaction terms for selected variables from the Farm Business Survey to be incorporated in another linear model with the main effects.

2.2.3 Climatic variables

Section 1 discussed the impact climate has on crop growth and the data available on climate conditions from the UK Met Office. We look to incorporate monthly daily rainfall, mean temperature and sunshine hours into our model for yields per hectare. These are averaged over each NUTS1 region to protect the farms' anonymity(see Section 1). Again, we look at the marginal densities of the climate data and correlations with yield to find which are marginally associated with yield.

Figure 2.9 shows the densities for rainfall in June and August. Both are examples of the densities found from the Met Office monthly climate conditions. The density of rainfall in June is representative of localised extreme weather conditions, whether that be extremely high in the case of rainfall, or extremely low in the case of sunshine, whereas the density of rainfall in August is representative of the case when there are no extreme weather conditions.

The correlation between the monthly climate variables and yield per hectare is calculated in Table 2.3. The most notable correlations with yield are in December, April, June and July and may be due to the cold climate in December, April showers and the warmer summer months of June and July. Figure 2.10 show scatterplots of yield per hectare against rainfall in December and June with the general trend superimposed. The vertical lines of points are a result of rounding in the Met Office data. The scatterplot between yield and rainfall in December is representative of weather variables with a smaller correlation with yield per hectare, whereas the scatterplot between yield per hectare and rainfall in June is representative of a larger correlation with yield per hectare. Given rainfall in June has a correlation with yield per hectare of -0.3050, the largest correlation in absolute value for the weather conditions in Table 2.3, the scatterplot indicates the

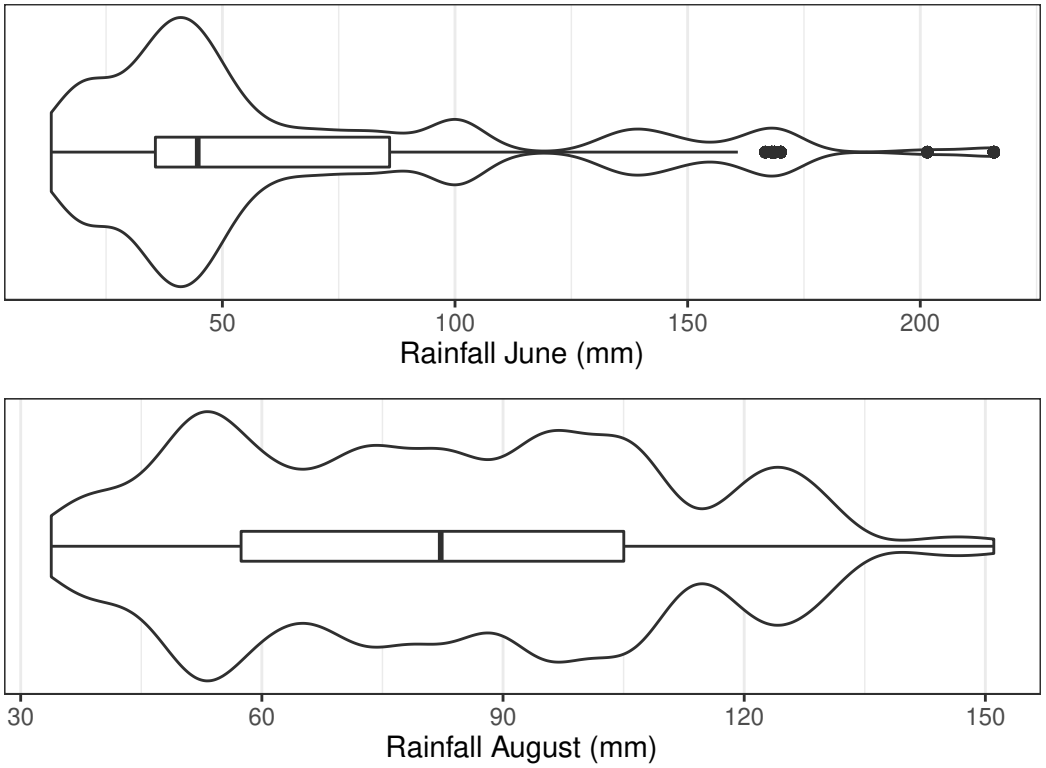


Figure 2.9: Violin plots for rainfall in June and August.

decrease in yield per hectare can be suitably captured by an increase rainfall in June, when looking at each variable marginally.

Month	Rainfall	Mean temp	Sunshine
October	0.0044	-0.0051	-0.0822
November	0.0003	-0.1374	-0.0531
December	-0.1899	-0.0874	0.2386
January	0.0099	-0.0927	0.0260
February	-0.0022	0.0132	0.1203
March	-0.0084	-0.0615	-0.1075
April	-0.2249	0.0748	0.1298
May	-0.0097	0.0097	0.0115
June	-0.3050	0.0724	0.2698
July	-0.2081	0.1078	0.0711
August	0.0491	-0.0206	-0.0670
September	-0.1212	0.0463	-0.0201

Table 2.3: Correlations between yield and rainfall, yield and mean temperature, and yield and sunshine hours for each month.

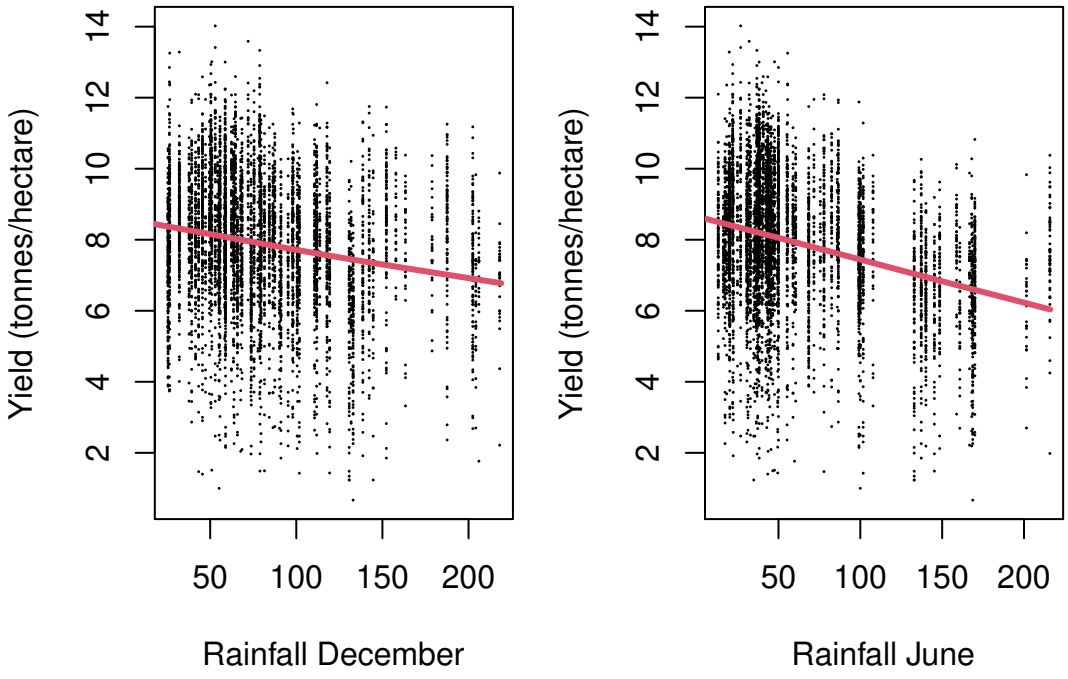


Figure 2.10: Plots of yield per hectare versus rainfall in December and June with general trend superimposed.

Briefly looking at the relationships between the weather variables within each month, Table 2.4 has, for each month, the pairwise correlations between rainfall, mean temperature and sunshine hours. The negative correlation between rainfall and sunshine hours for all months is plausible: it is less likely to rain when the sun is shining. Figure 2.11 shows the scatterplots between rainfall and sunshine hours in December and June with general trend superimposed. Again, due to the rounding of the Met Office data, points in the scatterplots may coincide. Since the summer months have longer daylight hours compared to winter months, the decrease in sunshine hours in June can be attributed to an increase in rainfall due to cloud cover, hence a larger correlation coefficient in absolute value. As for December, which will have a smaller number of daylight hours, the decrease in sunshine hours may not be attributed to rainfall, but instead down to the time of the year.

Month	`Rainfall` & `Mean temp`	`Rainfall` & `Sunshine`	`Mean temp` & `Sunshine`
`October`	0.1237	-0.5437	0.1065
`November`	0.0941	-0.1475	-0.2220
`December`	0.4320	-0.2128	0.0979
`January`	0.3525	-0.1513	0.1873
`February`	0.3865	-0.1898	0.1790
`March`	-0.4643	-0.4709	0.8964
`April`	-0.6146	-0.7168	0.6240
`May`	0.1048	-0.7243	-0.0098
`June`	-0.3393	-0.7675	0.5775
`July`	-0.6846	-0.5527	0.8076
`August`	-0.4010	-0.5557	0.1747
`September`	-0.3115	-0.2681	0.2251

Table 2.4: Correlations between each pairwise combination of rainfall, mean temperature and sunshine for each month.

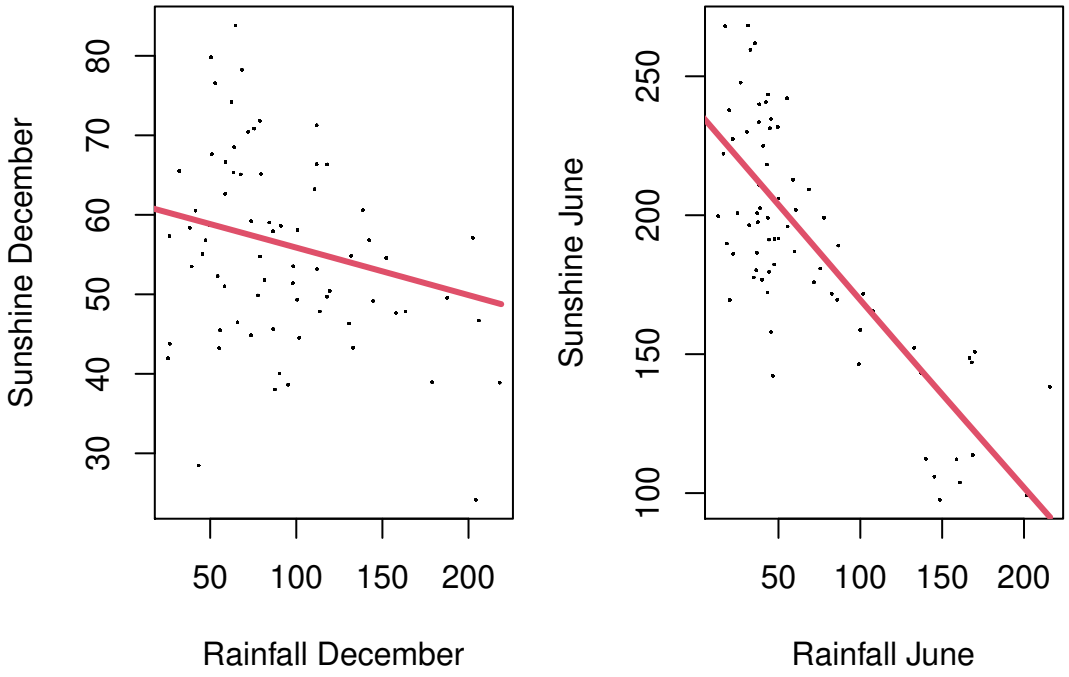


Figure 2.11: Plots of rainfall against sunshine hours for December and June.

We shall be incorporating weather conditions in our models as main effects only. Table 2.4 may suggest there should be interaction terms for the weather conditions, however since this thesis focuses on finding which farming practices are associated with yield, investigating these interactions to be studied as future work. Incorporating weather conditions as main effects only will provide a brief insight into whether yield per hectare still depends on climate conditions conditional on what we know from the Farm Business Survey variables.

2.3 Regression methods

So far we have looked at the marginal influence of each covariate from either the Farm Business Survey in Section 2.2.2 or the UK Met Office in Section 2.2.3 on yield per hectare. This section looks to combine these variables in one linear model with main effects only and one linear model with main effects, non-linear terms and interaction terms, using quantile regression, linear regression and principal component regression. Quantile regression is robust to high yielding farms. Linear regression may skew the

model to accommodate for outliers, however if these models approximately agree, linear regression is preferred since the ordinary least squares estimator has the smallest variance amongst all linear unbiased estimators. This is an advantage when fitting the model, however this could produce estimates far from the observed yields when testing the model on out-of-sample data. Principal component regression is based on capturing the largest variation between the explanatory variables rather than capturing the largest association with yield per hectare.

All of these methods are formulated from the following linear model. Suppose a set of fixed covariates $\boldsymbol{x}_1, ..., \boldsymbol{x}_p \in \mathbb{R}^n$ are to be used to describe a response $\boldsymbol{y} \in \mathbb{R}^n$, then a linear model takes the form $\boldsymbol{y} = \alpha + \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j + \boldsymbol{\epsilon} = \alpha \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of parameter coefficients, $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$, α the intercept parameter, $\mathbf{1}_n$ being the vector of length n composed of ones for the intercept term, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a zero-mean error term with constant covariance matrix $\sigma^2 \boldsymbol{I}_n$, with $\boldsymbol{I}_n$ being the identity matrix.

The following sections look at different methods to find estimates for the coefficients $\boldsymbol{\beta}$ using various minimisation criteria. We shall compare the predictions from each model with their mean squared prediction error (Section 2.3.4).

2.3.1 Quantile regression

Quantile regression models the conditional quantiles of yield, $Q_\tau(\boldsymbol{Y}|\boldsymbol{X}) = \alpha(\tau)\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}(\tau)$, see Koenker and Bassett (1978), where $\alpha(\tau)$ and $\boldsymbol{\beta}(\tau)$ are estimated by

$$(\hat{\alpha}(\tau), \hat{\boldsymbol{\beta}}(\tau)) = \underset{\alpha,\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i \in \left\{ i : y_i \geq \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} \right\}} \tau \left| y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta} \right| + \sum_{i \in \left\{ i : y_i < \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} \right\}} (1 - \tau) \left| y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta} \right| \right\},$$

and $|\cdot|$ denotes taking the absolute value. The minimisation problem allows for different weightings of the residuals through τ. For each i, if the residual $y_i - \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta}$ is positive, then the residual is weighted by τ, otherwise it is weighted by $1 - \tau$. As τ increases,

our estimates for α and β would be largely based on the upper quantiles of yield. As τ decreases, our estimates would be based on the lower quantiles of yield. Taking $\tau = 0.5$ gives the residuals equal weighting, corresponding to the conditional median, minimising the absolute loss

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ |\boldsymbol{Y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}|_1 \right\},$$

where $|\cdot|_1$ denotes the Euclidean L^1 norm

$$|\boldsymbol{r}|_1 = \sum_{i=1}^{n} |r_i|, \qquad \boldsymbol{r} \in \mathbb{R}^n. \tag{2.2}$$

This ensures the parameter estimates for α and $\boldsymbol{\beta}$ are robust to outliers.

No distributional assumptions about the residuals are required for quantile regression. If the error in our model is normally distributed, it would be preferable to use an estimator utilising this assumption to achieve smaller confidence intervals for our model parameters.

2.3.2 Linear regression

Linear regression models the conditional mean, $E(\boldsymbol{Y}|\boldsymbol{X}) = \alpha \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}$, where the asymptotic normality assumption of the residuals leads to desirable properties such as asymptotic normality, unbiasedness, and on most occasions a small variance for parameter estimates, whilst still being interpretable for non-specialists.

To fit a linear model, ordinary least squares finds $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ which will minimise the residual sum of squares

$$\text{RSS} := (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}}), \tag{2.3}$$

with $\hat{\boldsymbol{y}} = \hat{\alpha}\mathbf{1}_n + \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Letting $\boldsymbol{X}^* = (\mathbf{1}_n, \boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ and $\boldsymbol{\beta}^* = (\alpha, \beta_1, ..., \beta_p)$ be the merged coefficient vector, then $\hat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^{*T}\boldsymbol{X}^*\right)^{-1} \boldsymbol{X}^{*T}\boldsymbol{y}$, and so $\hat{\boldsymbol{y}} = \boldsymbol{X}^*\hat{\boldsymbol{\beta}}^*$, where $\boldsymbol{P}$ is the projection matrix $\boldsymbol{X}^* \left(\boldsymbol{X}^{*T}\boldsymbol{X}^*\right)^{-1} \boldsymbol{X}^{*T}$.

A normality assumption for $\boldsymbol{y}$ conditioned on $\boldsymbol{X}^*\boldsymbol{\beta}^*$ allows for inferences to be made about the coefficient estimates. To show $\boldsymbol{y}$ conditioned on $\boldsymbol{X}^*\boldsymbol{\beta}^*$ is approximately distributed according to a Normal distribution, the sample quantiles of the standardised

residuals should be approximately equal to the theoretical quantiles of a standard Normal distribution with zero mean and unit variance. A vector $\boldsymbol{r}$ is standardised by subtracting its mean $\bar{r}$ then dividing through by its standard deviation s_r, where

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \quad \text{and} \quad s_r = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2}. \tag{2.4}$$

A histogram of the standardised residuals should also appear as a standard Normal distribution. If the standardised residuals appear to be skewed, Box and Cox (1964) suggest first correcting for the skew by applying the Box-Cox transform

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \log(y_i), & \lambda = 0, \end{cases}$$

to the i^{th} observed yield y_i, where λ is estimated using profile likelihood.

Given the normality assumption for $\boldsymbol{y}$ conditioned on $\boldsymbol{X}^*\boldsymbol{\beta}^*$ holds, $\boldsymbol{X}^*\boldsymbol{\beta}^*$ is non-random, $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, the distribution of $\boldsymbol{Y}$ takes the same distributional form as the residuals, $\boldsymbol{Y} \sim N_n(\boldsymbol{X}^*\boldsymbol{\beta}^*, \sigma^2\boldsymbol{I}_n)$, with density function

$$\pi(\boldsymbol{y}|\boldsymbol{X}^*, \boldsymbol{\beta}^*, \sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\right)^T\left(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\right)\right). \tag{2.5}$$

The result for $\boldsymbol{\beta}$ can also be found using maximum likelihood estimation. Our distributional assumption for $\boldsymbol{Y}$ and the properties of the multivariate Normal distribution gives $\mathbb{E}(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta}^*$ and $\text{var}(\hat{\boldsymbol{\beta}}^*) = \sigma^2(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}$, therefore

$$\hat{\boldsymbol{\beta}}^* \sim N_p\left(\boldsymbol{\beta}^*, \sigma^2(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\right) \tag{2.6}$$

allows for confidence intervals to be constructed for the intercept α and the variable coefficients $\boldsymbol{\beta}$. The relevant hypothesis test statistic for testing the null hypothesis of a coefficient equalling zero, $H_0 : \beta_j^* = 0$, is $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^*/s\sqrt{d_{ii}}$, where s is the standard error of the residuals using Equation 2.4 and d_{ii} is the ii^{th} element of the matrix $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$. The test statistic is compared to a t-distribution with $n-p$ degrees of freedom. As the sample

size n increases, the t-distribution approximately resembles a Normal distribution. In this case, the $(1-\alpha)\%$ confidence intervals will be

$$\hat{\beta}_j^* \pm \Phi^{-1}(1-\alpha/2) \times s\sqrt{d_{ii}}, \tag{2.7}$$

where Φ^{-1} is the inverse of the cumulative distribution function for the Normal distribution. With $\alpha = 0.05$, this will correspond to the 95% confidence interval. These confidence intervals will also allow variable importance to be determined at the α significance level. If a variable is not important to model the yield per hectare then zero will be contained in the confidence interval and the null hypothesis would be accepted. Varying α will vary our tolerance to what is an important variable.

Although linear regression will achieve the smallest variance in the coefficient estimates, all of the data available is used to fit a linear model. This will capture the subtle differences in our dataset. For out-of-sample data which may not contain these subtle differences, predictions may be far from what was observed. Instead, the following section looks to fit a model to a projection of the data which captures the largest proportion of the variation in our training data.

2.3.3 Principal component regression

Assuming the columns of $\boldsymbol{X}$ have been standardised (see Equation 2.4), principal component regression first reduces the dimension of the standardised covariate matrix $\boldsymbol{X}$ by projecting the covariates onto the orthonormal columns of matrix $\boldsymbol{Q}$ which in turn maximise the correlation of the observations not already captured by the previous orthonormal vectors (see Figure 2.12). These are referred to as the principal component loadings, the projected covariates $\boldsymbol{XQ}$ are referred to as the principal component scores and the original covariates are referred to as the raw scores. Principal component regression involves linearly regressing upon a subset of principal component scores rather than the raw scores as done in the previous section.

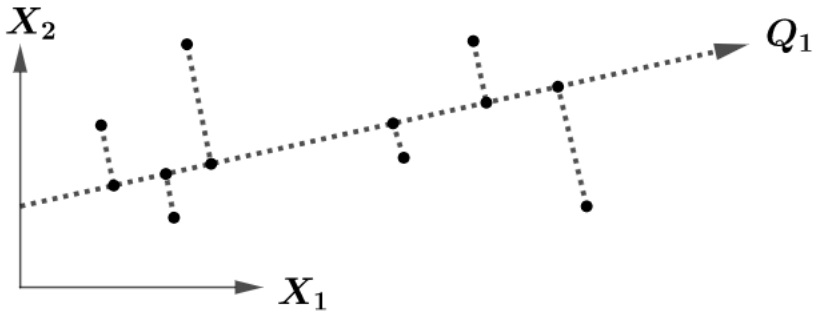


Figure 2.12: Projection of $\boldsymbol{X}$ onto the first principal component $\boldsymbol{Q}_1$.

Finding the principal component loadings and scores reduces down to finding the eigenvalues and eigenvectors of the correlation matrix $\boldsymbol{X}^T\boldsymbol{X}$. Given $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric, we can find the spectral decomposition

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T,$$

where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, ..., \lambda_p\}$ is the diagonal matrix of decreasing eigenvalues of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{Q}$ is the matrix of corresponding orthonormal eigenvectors.

The first principal component loading is the eigenvector of the covariance matrix $\boldsymbol{X}^T\boldsymbol{X}$ which corresponds to the largest eigenvalue. This will capture the largest proportion of correlation, the second principal component loading will capture the second largest proportion, and so on. Rather than regressing on all of the principal component scores, equivalent to using linear regression from Section 2.3.2, a subset of the principal component scores can be used to capture most of the correlation whilst reducing the dimension. Suppose the first $k < p$ principal component scores are to be used, that is $\boldsymbol{W}_k = \boldsymbol{X}\boldsymbol{Q}_k$, where $\boldsymbol{Q}_k$ is the $p \times k$ matrix of the first k principal component loadings, we linearly regress the principal component scores onto the centred response $\boldsymbol{y}$ through the linear model $\boldsymbol{y} = \boldsymbol{W}_k\boldsymbol{\gamma} + \boldsymbol{e}$, where $\boldsymbol{e} \in \mathbb{R}^n$ is a zero-mean error term analogous to $\boldsymbol{\epsilon}$ in the original linear model, to find an estimate for $\boldsymbol{\gamma}$. Furthermore, if we want to see how this translates back to the original formulation, $\hat{\boldsymbol{\beta}} = \boldsymbol{Q}_k\hat{\boldsymbol{\gamma}}$ where $\hat{\boldsymbol{\gamma}}$ is the vector of estimated coefficients of the principal component scores and $\hat{\boldsymbol{\beta}}$ is the estimated coefficients of the

standardised matrix $\boldsymbol{X}$.

Since the estimated parameters $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are estimated from the standardised vectors, the following transform is used to revert the coefficients back to their original scale:

$$\hat{\alpha}^* = \bar{y} - \sum_{j=1}^{p} \frac{\hat{\beta}_j \bar{x}_j}{s_{x_j}} \quad \hat{\beta}_j^* = \frac{\hat{\beta}_j}{s_{x_j}}, \tag{2.8}$$

where $\hat{\beta}_j$, $j = 1, ..., p$ are the estimated parameters from the standardised vectors and $\hat{\alpha}^*$, $\hat{\beta}_j^*$, $j = 1, ..., p$, are the transformed estimated parameters, $\bar{y}$ is the sample mean of $\boldsymbol{y}$ and $\bar{x}_j$ and s_{x_j} are the mean and standard deviation of the original $\boldsymbol{x}_j$ before standardising. The estimated coefficients $\hat{\beta}_j$ of the standardised vectors and the estimated coefficients $\hat{\beta}_j^*$ in the original model differ by a scale factor depending on the scale of their corresponding variable $\boldsymbol{x}_j$.

Although principal component regression reduces the dimension of the matrix $\boldsymbol{X}$ to capture the largest variation in $\boldsymbol{X}$, the model may still include variables which have little association with yield per hectare.

2.3.4 Prediction accuracy

All of the previous sections discussed different methods to find intercept and coefficient estimates $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ to use in the linear model $\boldsymbol{y} = \hat{\alpha}\mathbf{1}_n + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$. Linear regression and principal component regression find $\hat{\boldsymbol{\beta}}$ by minimising the residual sum of squares in Equation 2.3, whether that be with the full set of covariates in linear regression or with a projected set of covariates in principal component regression.

The following section compares out-of-sample predictions using the different methods discussed. To assess the accuracy of the out-of-sample predictions, we take the estimates for α and $\boldsymbol{\beta}$ and the data from 2009 to find $\hat{\boldsymbol{y}} = \hat{\alpha}\mathbf{1}_n + \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and the mean squared prediction error,

$$\text{MSPE} := \frac{1}{n} \left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right)^T \left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right). \tag{2.9}$$

This analogous to finding the mean of the residual sum of squares in Equation 2.3 with out-of-sample data.

2.4 Application

The methodology from Section 2.3 will be applied to the Farm Business Survey and UK Met Office data from Section 2.2 to study the relationship between the covariates and yield, conditional on other variables already being included in the models. Both the model with main effects only and the model with main effects, non-linear and interaction terms will be compared for each parameter estimation method: quantile regression and linear regression (Section 2.4.1) and principal component regression (Section 2.4.2).

2.4.1 Quantile and linear regression

To model the conditional median, estimates for the coefficients $\boldsymbol{\beta}$ are found by minimising the mean absolute difference (see Section 2.3.1). Table 2.5 shows the quantile regression estimates for each variable in the model with main effects only and the model with main effects, non-linear and interaction terms from Section 2.2. Comparing these to the estimates for the coefficients when minimising the residual sum of squares in Equation (2.3), we find these are approximately equal. Since the least squares estimator achieves the smaller variance between the two methods, we discuss the coefficients computed using linear regression and where appropriate, their associated confidence intervals.

Section 2.3.2 discussed checking the standardised residuals to ensure the linear model is appropriate to conduct inference based on the asymptotic normality property of ordinary least squares regression. Figure 2.13 indicates both models can be appropriately modelled linearly; the sample and theoretical quantiles are approximately equal and the standardised residuals are randomly scattered about zero, therefore there is no need to apply a Box-Cox transform to the yields per hectare.

Table 2.5 gives the coefficient estimates and their corresponding 95% confidence interval for each model proposed. If the 95% confidence interval does not contain zero, then the variable will be important at the $\alpha = 0.05$ level. This is equivalent to the t-statistic being larger than 1.96 in absolute value.

From the confidence intervals for the main effects model only, the Farm Business Survey

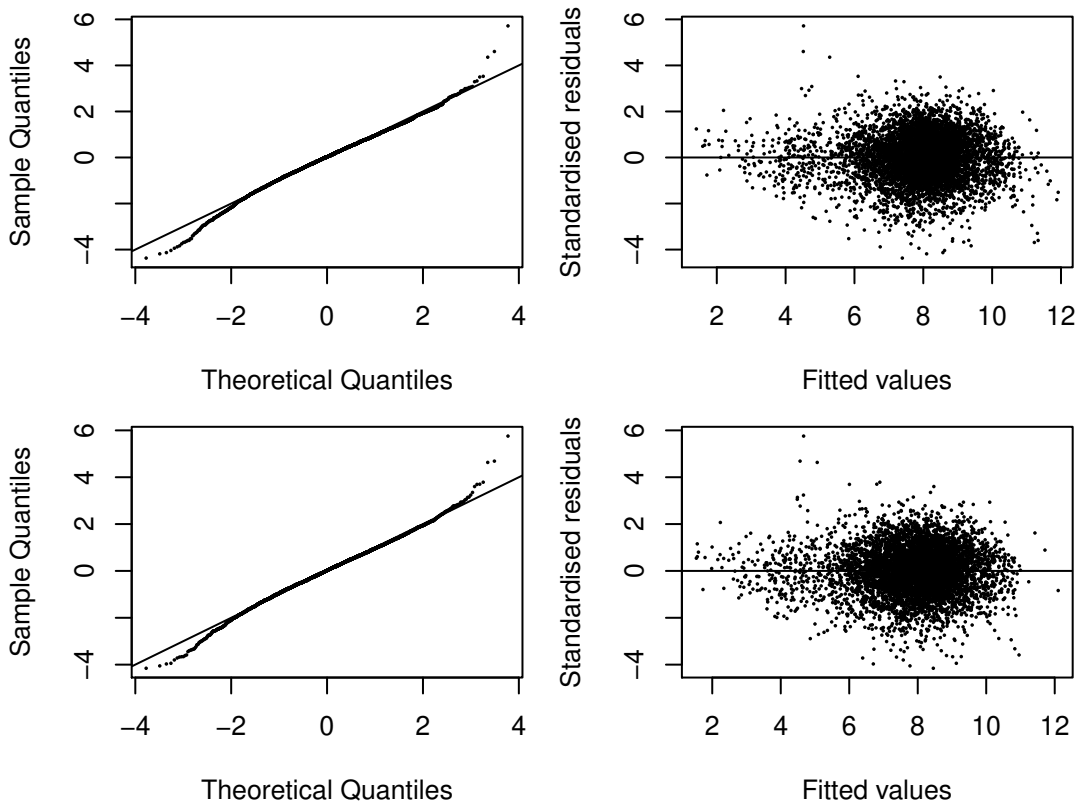


Figure 2.13: Left: plots of theoretical quantiles of the standard Normal distribution against sample quantiles of the standardised residuals from the linear model for, top: main effects only, bottom: main effects, non-linear and interaction terms. Right: plots of standardised residuals against fitted values for the linear model for, top: main effects only, bottom: main effects, non-linear and interaction terms.

variables indicated as not being important are all of the education levels except college level or equivalent, `fuel` and `TOFC`. For the model with main effects, non-linear and interaction terms, the main effects from the Farm Business Survey variables indicated as not being important remain the same. Even with the non-linear and interaction terms now being included, the main effects included in the non-linear and interaction terms are still important. Since the interaction terms between organic status and seed costs, fertiliser costs and utilised agricultural area have now been included in the model, the coefficient for organic status as a main effect has now decreased, since some of the marginal association between organic status and yield is now captured by the interaction terms.

The mean squared prediction error for the main effects model is 2.704 and for the model with main effects, non-linear and interactions terms is 3.155, hence we look at

the predictions for the main effects model. Figure 2.15 shows the predictions of yield per hectare using the linear regression coefficients for the main effects model. These predictions appear to capture the yields around the mean sufficiently well but struggle to predict yields departing from the mean. Although linear regression shows how a variable behaves in the presence of other variables in the model, the coefficients will capture subtle differences in the data the model is trained on. The following section looks to capture the largest variation in our data and neglects these nuances.

2.4.2 Principal component regression

Section 2.3.3 discusses the advantages to dimension reduction techniques as opposed to linear regression. Principal component regression assumes the columns of $\boldsymbol{X}$ have been standardised and $\boldsymbol{y}$ has been centred to compute the principal component loadings. Response vector $\boldsymbol{y}$ is centred by subtracting its mean. To standardise $\boldsymbol{X}$, each column is subtracted by their mean and divided by their standard deviation (see Equation 2.4). Basing our principal component analysis on the correlation matrix ensures the projections are not determined by the units of the covariate vectors.

We first look at principal components for the main effects model. Before regressing on the principal component scores, we look at the projections for each principal component. Figure 2.14 shows the contributions each covariate has in the first 2 principal components. For our data, the first principal component, accounting for 12.54% of the correlation in the observations, is composed of weather conditions. This may be down to weather conditions changing each year whereas farming practices often remain the same. The second principal component is also composed of the weather conditions. The first 2 principal components account for 23.87% of the total variation.

Comparing the mean squared prediction error (Equation 2.9) for each model of increasing number of principal components used in the modelling, the smallest mean squared prediction error is 1.944 when 32 principal components are used to model wheat yield, however there is little difference in the mean squared prediction error between using 10 and 32 principal components once the principal components are regressed upon. When we regress upon the first 10 principal components, capturing 66.02%, the mean squared

prediction error is 2.080. This is smaller than 2.704 which is the mean squared prediction error achieved using linear regression.

Looking at the principal component for the model with main effects, non-linear and interaction terms, the first 2 principal components are still composed of the weather conditions but now capture 20.14% of the total variation. The reduction in total variation is down to now including non-linear and interaction terms using data from the Farm Business Survey, yet these additional terms are not included in the first 2 principal components. The smallest mean squared prediction error is found using the first 42 principal components, however the mean squared prediction error is, again, not very different to using the first 10 principal components, now capturing 63.05%, achieving a MSPE of 2.124.

Predicting yields per hectare using the first 10 principal components for both models gives the coefficient estimates in Table 2.5 and the predictions for the main effects model are in Figure 2.15, since this achieves a smaller MSPE compared to the model with non-linear and interaction terms. The predictions appear to model the mean sufficiently well, but still overestimate lower yields.

Calculating coefficients based on a subset of principal component scores has reduced the mean squared prediction error, compared to those from linear regression, however variables in the principal component projections are those which capture the largest variation in the data rather than those which have the largest association with yield per hectare. The following section looks at methods to construct a linear model based on Farm Business Survey and Met Office variables which are most associated with yield per hectare.

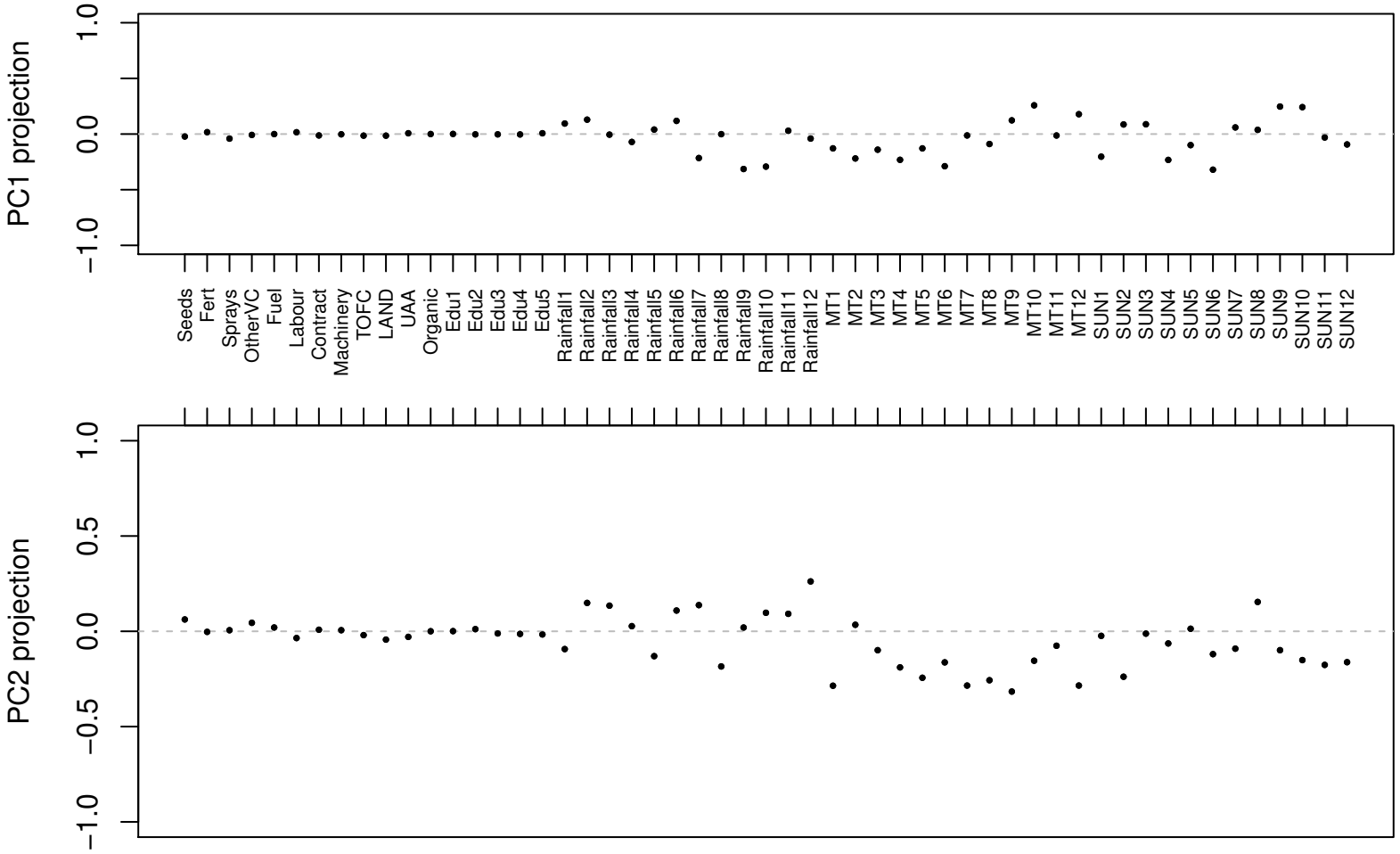


Figure 2.14: Principal component loadings for the 1st and 2nd principal components of the main effects model. Weather variable names with numbers 1 to 12 correspond to the months of the growing year, e.g. `Rainfall1` corresponds to rainfall in October.

	Main effects model				Model with interactions			
Variable	$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{QT})}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{LR})}$		$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{PC})}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{QT})}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{LR})}$		$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{PC})}$
Intercept α	4.1852	5.8600 (3.3432, 8.3762)	*	6.7156	3.3932	4.7164 (2.2447, 7.1881)	*	3.8229
Seeds	−0.0036	−0.0043 (−0.0058, −0.0028)	*	−0.0047	−0.0027	−0.0035 (−0.0051, −0.0019)	*	−0.0021
Fert	0.0033	0.0036 (0.0029, 0.0042)	*	0.0036	0.0076	0.0072 (0.0046, 0.0097)	*	0.0012
Sprays	0.0061	0.0065 (0.0057, 0.0072)	*	0.0056	0.0176	0.0200 (0.0174, 0.0226)	*	0.0023
OtherVC	0.0020	0.0026 (0.0017, 0.0035)	*	0.0004	0.0021	0.0027 (0.0018, 0.0035)	*	0.0003
Fuel	0.0039	0.0032 (−0.0033, 0.0097)		0.0098	0.0021	0.0009 (−0.0054, 0.0073)		0.0038
Labour	0.0005	0.0005 (0.0002, 0.0008)	*	0.0007	0.0021	0.0021 (0.0017, 0.0026)	*	0.0006
Contract	0.0013	0.0014 (0.0011, 0.0017)	*	−0.0009	0.0027	0.0026 (0.0022, 0.0030)	*	−0.0004
Machinery	0.0015	0.0014 (0.0012, 0.0016)	*	−0.0001	0.0024	0.0022 (0.0019, 0.0025)	*	0.0002
TOFC	−0.0005	−0.0004 (−0.0011, 0.0003)		−0.0010	−0.0006	−0.0005 (−0.0012, 0.0003)		0.0001
LAND	0.0013	0.0010 (0.0007, 0.0013)	*	−0.0004	0.0012	0.0011 (0.0008, 0.0014)	*	$2.825.10^{-5}$
UAA	0.0007	0.0006 (0.0005, 0.0007)	*	0.0001	0.0020	0.0019 (0.0017, 0.0022)	*	0.0002
Organic	−2.4896	−2.1358 (−2.3690, −1.9026)	*	−2.1239	−1.3770	−0.8366 (−1.3714, −0.3018)	*	−0.9710
Education 1	−0.1172	−0.0745 (−0.2035, 0.0545)		0.0719	−0.0821	−0.0458 (−0.1717, 0.0802)		0.0092
Education 2	0.0858	0.0389 (−0.1232, 0.2011)		0.0652	0.1013	0.0615 (−0.0970, 0.2200)		0.0499

Education 3	0.1526	0.1182 (0.0285, 0.2079)	*	0.0738	0.1456	0.1197 (0.0321, 0.2073)	*	−0.0413
Education 4	0.1049	0.1007 (−0.0115, 0.2129)		−0.1534	0.1641	0.1166 (−0.0068, 0.2263)		0.0469
Education 5	−0.0700	0.0378 (−0.1450, 0.2206)		−0.1435	0.0155	0.0286 (−0.1500, 0.2073)		0.0145
Seeds*Organic					0.0023	−0.0014 (−0.0024, 0.0051)		−0.0087
Fert*Organic					−0.0023	−0.0023 (−0.0090, 0.0043)		−0.0173
UAA*Organic		Interaction			−0.0006	−0.0009 (−0.0019, 0.0001)		−0.0030
Fert*Sprays		terms			$-3.6.10^{-6}$	$-1.1.10^{-6}$ $(-1.2.10^{-5}, 9.7.10^{-6})$		$6.0.10^{-6}$
Labour*Contract					$-3.6.10^{-6}$	$-2.5.10^{-6}$ $(-4.3.10^{-6}, -6.7.10^{-7})$	*	$1.1.10^{-6}$
Labour*Machinery					$-2.0.10^{-6}$	$-1.9.10^{-6}$ $(-2.4.10^{-6}, -1.4.10^{-6})$	*	$5.5.10^{-7}$
Contract*Machinery					$-2.6.10^{-6}$	$-1.9.10^{-6}$ $(-3.4.10^{-6}, -4.0.10^{-7})$	*	$-1.1.10^{-7}$
Fert*Fert					$-1.2.10^{-5}$	$-1.2.10^{-5}$ $(-1.9.10^{-5}, -4.9.10^{-6})$	*	$2.5.10^{-6}$
Sprays*Sprays		Non-linear terms			$-3.2.10^{-5}$	$-3.8.10^{-5}$ $(-4.5.10^{-5}, -3.0.10^{-5})$	*	$5.3.10^{-6}$
UAA*UAA					$-7.6.10^{-7}$	$-7.6.10^{-7}$ $(-9.2.10^{-7}, -6.0.10^{-7})$	*	$1.2.10^{-7}$
Rainfall October	0.0004	−0.0008 (−0.0046, 0.0030)		−0.0011	−0.0008	−0.0007 (−0.0044, 0.0031)		−0.0002
Rainfall November	−0.0085	−0.0086 (−0.0116, 0.0057)		0.0002	−0.0090	−0.0083 (−0.0112, −0.0054)	*	0.0007
Rainfall December	0.0048	0.0047 (0.0013, 0.0082)	*	−0.0028	0.0036	0.0040 (0.0006, 0.0074)	*	−0.0007
Rainfall January	0.0022	0.0005 (−0.0034, 0.0045)		0.0012	0.0025	0.0002 (−0.0037, 0.0041)		0.0009

Rainfall February	0.0027	0.0011 (−0.0023, 0.0045)		−0.0008	0.0040	0.0019 (−0.0013, 0.0052)	0.0008
Rainfall March	−0.0029	0.0006 (−0.0040, 0.0052)		0.0008	−0.0018	0.0002 (−0.0043, 0.0047)	−0.0001
Rainfall April	−0.0166	−0.0149 (−0.0208, −0.0090)	*	−0.0035	−0.0170	−0.0134 (−0.0191, −0.0076)	* −0.0010
Rainfall May	−0.0030	−0.0031 (−0.0069, 0.0007)		0.0007	−0.0030	−0.0028 (−0.0065, 0.0010)	−0.0016
Rainfall June	−0.0087	−0.0082 (−0.0111, −0.0053)	*	−0.0033	−0.0081	−0.0082 (−0.0110, −0.0054)	* −0.0019
Rainfall July	−0.0091	−0.0087 (−0.0120, −0.0054)	*	−0.0016	−0.0078	−0.0081 (−0.0113, −0.0048)	* −0.0021
Rainfall August	0.0061	0.0043 (−0.0004, 0.0089)		0.0011	0.0063	0.0033 (−0.0012, 0.0078)	0.0028
Rainfall September	0.0041	0.0046 (0.0004, 0.0087)	*	−0.0024	0.0031	0.0038 (−0.0003, 0.0078)	−0.0001
Mean temp October	0.3739	0.4735 (0.2792, 0.6678)	*	−0.0309	0.3571	0.5072 (0.3174, 0.6970)	* 0.0010
Mean temp November	−0.3053	−0.2948 (−0.5294, −0.0602)	*	−0.0384	−0.2856	−0.2618 (−0.4910, −0.0326)	* −0.0098
Mean temp December	0.0544	0.0438 (−0.1673, 0.2550)		−0.0114	0.0737	0.0377 (−0.1684, 0.2438)	−0.0182
Mean temp January	−0.3392	−0.3359 (−0.5287, 0.1431)		0.0103	−0.2875	−0.2949 (−0.4831, −0.1068)	* −0.0171
Mean temp February	−0.4473	−0.4199 (−0.5771, −0.2627)	*	0.0452	−0.5077	−0.4620 (−0.6154, −0.3085)	* 0.0190
Mean temp March	−0.0657	−0.0057 (−0.1559, 0.1444)		−0.0079	−0.0792	−0.0657 (−0.2125, 0.0812)	0.0089
Mean temp April	−0.0704	0.0845 (−0.1982, 0.3672)		0.0317	−0.0502	0.1180 (−0.1579, 0.3940)	0.0207
Mean temp May	−0.2070	−0.2847 (−0.5094, −0.0600)	*	0.0459	−0.1961	−0.2399 (−0.4593, −0.0205)	* 0.0444
Mean temp June	0.1838	0.0116 (−0.3053, 0.3285)		−0.0017	0.1786	−0.0631 (−0.3726, 0.2465)	0.0612

`Mean temp July`	-0.9848	-0.8384 $(-1.1239, -0.5528)$	*	-0.0040	-1.0257	-0.8175 $(-1.0965, -0.5386)$	*	0.0302
`Mean temp August`	0.4569	0.3047 $(0.1193, 0.4902)$	*	-0.0044	0.4306	0.2553 $(0.0735, 0.4371)$	*	0.0054
`Mean temp September`	0.8485	0.7179 $(0.4688, 0.9670)$	*	-0.0434	0.9174	0.7429 $(0.4997, 0.9862)$	*	0.0271
`Sunshine October`	-0.0034	0.0012 $(-0.0063, 0.0087)$		-0.0005	-0.0048	-0.0006 $(-0.0079, 0.0068)$		-0.0003
`Sunshine November`	-0.0070	-0.0069 $(-0.0136, -0.0001)$	*	-0.0033	-0.0146	-0.0086 $(-0.0152, -0.0020)$	*	-0.0014
`Sunshine December`	0.0055	0.0029 $(-0.0045, 0.0103)$		0.0126	0.0059	0.0020 $(-0.0053, 0.0092)$		0.0036
`Sunshine January`	0.0044	0.0008 $(-0.0101, 0.0118)$		0.0034	0.0081	0.0021 $(-0.0086, 0.0128)$		-0.0022
`Sunshine February`	0.0528	0.0526 $(0.0436, 0.0616)$	*	0.0072	0.0506	0.0520 $(0.0432, 0.0608)$	*	0.0012
`Sunshine March`	0.0119	0.0081 $(0.0022, 0.0140)$	*	-0.0015	0.0120	0.0082 $(0.0025, 0.0140)$	*	-0.0005
`Sunshine April`	0.0099	0.0081 $(0.0022, 0.0140)$	*	0.0029	0.0098	0.0083 $(0.0025, 0.0141)$	*	0.0010
`Sunshine May`	-0.0053	-0.0027 $(-0.0079, 0.0026)$		0.0001	-0.0059	-0.0021 $(-0.0072, 0.0030)$		0.0031
`Sunshine June`	-0.0069	-0.0049 $(-0.0087, -0.0012)$	*	0.0035	-0.0065	-0.0046 $(-0.0083, -0.0009)$	*	0.0027
`Sunshine July`	0.0005	0.0015 $(-0.0038, 0.0068)$		-0.0002	0.0026	0.0014 $(-0.0037, 0.0066)$		0.0002
`Sunshine August`	-0.0055	-0.0050 $(-0.0101, 0.0001)$		-0.0029	-0.0081	-0.0067 $(-0.0117, -0.0017)$	*	-0.0023
`Sunshine September`	-0.0127	-0.0117 $(-0.0173, -0.0061)$	*	-0.0022	-0.0156	-0.0140 $(-0.020, -0.0085)$	*	-0.0012
Variable	$\hat{\boldsymbol{\beta}}_j^{(QT)}$	$\hat{\boldsymbol{\beta}}_j^{(LR)}$		$\hat{\boldsymbol{\beta}}_j^{(PC)}$	$\hat{\boldsymbol{\beta}}_j^{(QT)}$	$\hat{\boldsymbol{\beta}}_j^{(LR)}$		$\hat{\boldsymbol{\beta}}_j^{(PC)}$
	Main effects model				Model with interactions			

Table 2.5: Linear model coefficient estimates with quantile regression $\hat{\boldsymbol{\beta}}^{(QT)}$, linear regression $\hat{\boldsymbol{\beta}}^{(LR)}$ with 95% confidence interval and principal component regression $\hat{\boldsymbol{\beta}}^{(PC)}$ based on 10 principal components. Stars represent significance at the $\alpha = 0.05$ according to their confidence intervals.

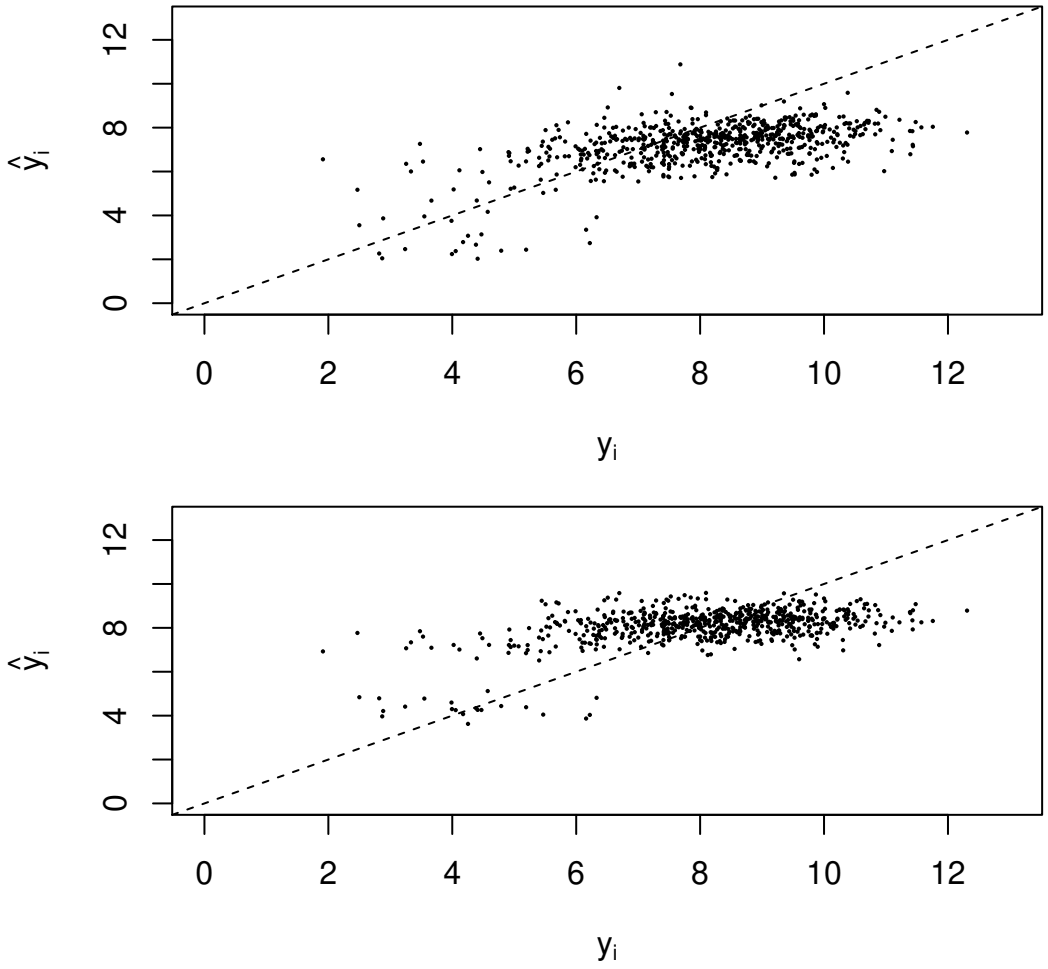


Figure 2.15: Observed yield per hectare versus predicted yield per hectare for the main effects model using, top: linear regression, bottom: principal component regression based on 10 principal components. The predictions using principal component regression with 10 principal components achieves a smaller mean squared prediction error compared to the predictions using linear regression, hence is better to model yield per hectare.

2.5 Variable selection methods

The previous sections looked at regression methods for linearly modelling yield per hectare using all of the covariates available. Here we use stepwise methods to reduce the set of covariates down to only include the variables which are most associated with yields.

Forward stepwise regression and Lasso regression in the following chapters require the covariates $\boldsymbol{X}$ to first be standardised and the response $\boldsymbol{y}$ to be centred by subtracting its mean (see Equation 2.4). This is to avoid shrinking and selecting variables based on the magnitude of each variables rather than their relationship with the response. Equation 2.8 discusses how to transform the coefficients from forward stepwise regression or Lasso regression back to their original scale.

2.5.1 Best subsets regression

Best subsets regression searches all 2^p combinations of inclusion and exclusion of the p variables in the linear model to estimate the conditional mean. For each $k \in \{1, ..., p\}$, best subset regression finds the subset of k explanatory variables which minimises the residual sum of squares amongst all subsets of size k. The advantage of this method is parsimonious models can be found which achieve a small residual sum of squares. There are $p!/((p-k)!k!)$ different combinations of explanatory variables to create subsets of size k. When p increases, the number of combinations to compute the residual sum of squares will increase exponentially, hence the calculations will quickly become infeasible for large p.

2.5.2 Forward stepwise regression

Forward stepwise regression looks to search amongst the 2^p possible combinations of covariates by sequentially adding variables which achieve the largest covariance with the residuals. Forward stepwise regression searches the combinations methodically, whereas best subset regression searched the combinations exhaustively. The covariates are stan-

dardised and the response vector is centred such that the selected variables do not depend on the centre of the response nor the magnitude of the observed variable.

At the first step, the variable $\boldsymbol{X}_j$ is selected if it minimises the residual sum of squares in Equation 2.3 with $\hat{\boldsymbol{y}} = \boldsymbol{X}_j\hat{\beta}_j$, over all j, where $\hat{\beta}_j$ is the coefficient from linearly regressing $\boldsymbol{X}_j$ onto centred $\boldsymbol{y}$ as in Section 2.3.2, without an intercept term. Consequently, the first variable j_1 to be selected satisfies

$$(j_1, s_1) = \underset{j=1,\ldots,p, s\in\{-1,1\}}{\operatorname{argmax}} \frac{s\boldsymbol{X}_j^T\boldsymbol{y}}{||\boldsymbol{X}_j||_2}, \tag{2.10}$$

where s_1 is the corresponding sign of the coefficient of $\boldsymbol{X}_{j_1}$. With $\boldsymbol{r}_{A_1} = \boldsymbol{y} - \boldsymbol{X}_{j_1}\hat{\beta}_{j_1}$ from the previous step, the next step looks to find the variable $\boldsymbol{X}_j$, $j \neq j_1$, which minimises the residual sum of squares in Equation 2.3 now with $\boldsymbol{y} = \boldsymbol{r}_{A_1}$ and $\hat{\boldsymbol{y}} = \boldsymbol{X}_j\hat{\beta}_j$, $j \neq j_1$. A_{k-1} refers to the set of variables already selected up to step k.

For general step k, the variable which enters the model at the k^{th} step satisfies

$$(j_k, s_k) = \underset{j\notin A_{k-1}, s\in\{-1,1\}}{\operatorname{argmax}} \frac{s\boldsymbol{X}_j^T\boldsymbol{r}_{A_{k-1}}}{||\boldsymbol{X}_j||_2} = \underset{j\notin A_{k-1}, s\in\{-1,1\}}{\operatorname{argmax}} \frac{s\boldsymbol{X}_j^T\boldsymbol{P}_{A_{k-1}}^{\perp}\boldsymbol{y}}{(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1/2}}, \tag{2.11}$$

where $\boldsymbol{P}_{A_{k-1}}^{\perp} = \boldsymbol{I} - \boldsymbol{P}_{A_{k-1}}$ is the projection on the space orthogonal to $\boldsymbol{X}_{A_{k-1}}$. This continues until all variables are included in the model, assuming $\boldsymbol{X}$ is of full rank.

Usually forward selection does not involve orthogonalising after each step. Tibshirani et al. (2016) introduced the idea of orthogonalising at each step such that fewer parameters are included to achieve predictions close to those from using linear regression. These methods are only guaranteed to produce the same path of active coefficients if all of the covariates are orthonormal. Furthermore, the residual sum of squares using the variables selected from orthogonalising after each step will always be smaller than or equal to the residual sum of squares from not orthogonalising. This alternative method starts in the same manner, where j_1 and s_1 satisfy Equation 2.10. When orthogonalising after each step, the general k^{th} term becomes

$$(j_k, s_k) = \underset{j \notin A_{k-1}, s \in \{-1,1\}}{\operatorname{argmax}} \frac{s\tilde{\boldsymbol{X}}_j^T \boldsymbol{r}_{A_{k-1}}}{||\tilde{\boldsymbol{X}}_j||_2} = \underset{j \notin A_{k-1}, s \in \{-1,1\}}{\operatorname{argmax}} \frac{s\boldsymbol{X}_j^T \boldsymbol{P}_{A_{k-1}}^{\perp} \boldsymbol{y}}{\left(\boldsymbol{X}_j^T \boldsymbol{P}_{A_{k-1}}^{\perp} \boldsymbol{X}_j\right)^{1/2}}, \tag{2.12}$$

where $\tilde{\boldsymbol{X}}_j$ is $\boldsymbol{X}_j$ orthogonalised with respect to the variables already included in the model, $\tilde{\boldsymbol{X}}_j = \boldsymbol{P}_{A_{k-1}}^{\perp} \boldsymbol{X}_j$. We will apply both these variations of forward stepwise regression on the Farm Business Survey and UK Met Office data and compare the sequences in which the variables are selected.

Although orthogonalising at each step in forward stepwise finds a suitable linear combination of variables which have a strong association with yield, forward stepwise is a greedy variable selection technique, going as far as possible in one direction before looking for another to include in the model. Progressing in a greedy manor means some of the variables which are highly correlated with yield per hectare are not selected.

2.5.3 Lasso regression

Rather than following as far as possible in the direction spanned by the variables included in the model before adding in another variable, Lasso regression (see, Tibshirani (1996) and Hastie et al. (2008)) selects variables according to their correlation with the residuals, gradually increasing the coefficient parameters until another variable is equicorrelated with the residual (Figure 2.16).

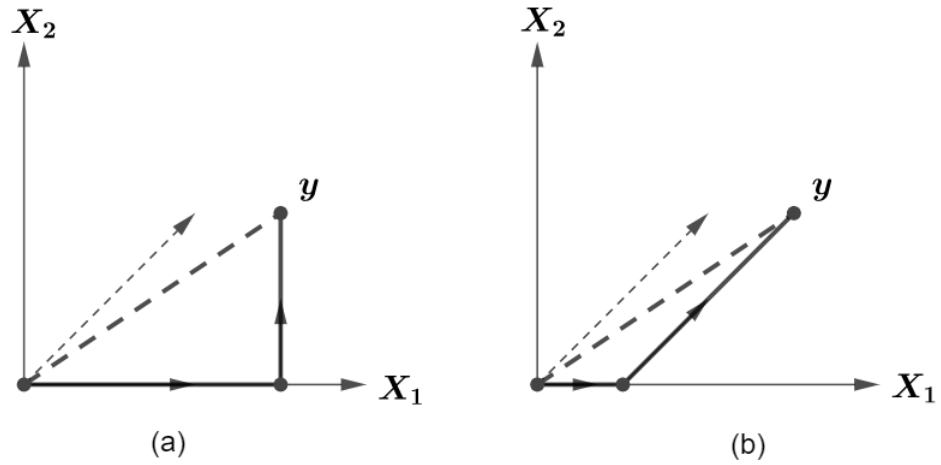


Figure 2.16: (a) Steps taken in forward stepwise regression. The coefficient increases as far as possible in the direction of the variable with the largest correlation with the response $\boldsymbol{y}$, $\boldsymbol{X}_1$ in this toy example. (b) Steps taken in Lasso regression. The coefficient increases until another variable is equicorrelated with the residual. Solid line represents the path taken, thick dashed line represents the shortest route between $(0, 0)$ and $\boldsymbol{y}$ and the thin dashed line represents the direction when the variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are equicorrelated.

Tibshirani (1996) showed this can be formulated as an optimisation problem, aiming to solve

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right)^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right) - \lambda |\boldsymbol{\beta}|_1 \right\} \tag{2.13}$$

where λ is the shrinkage parameter to restrict the length of the coefficient vector $\boldsymbol{\beta}$. Consequently, λ also controls the number of variables included in the model. Figure 2.17 shows how using an L_1 penalty on the size of the coefficient applies both shrinkage and selection. Figure 2.17(a) illustrates selecting a large enough λ such that only 1 variable is selected, whereas Figure 2.17(b) selects a smaller λ which includes both coefficients in our resulting model but shrinks the coefficients down. When $\lambda = 0$, the linear regression coefficients are retrieved. We shall see in Section 2.6.2 the order in which the variable are selected when decreasing λ.

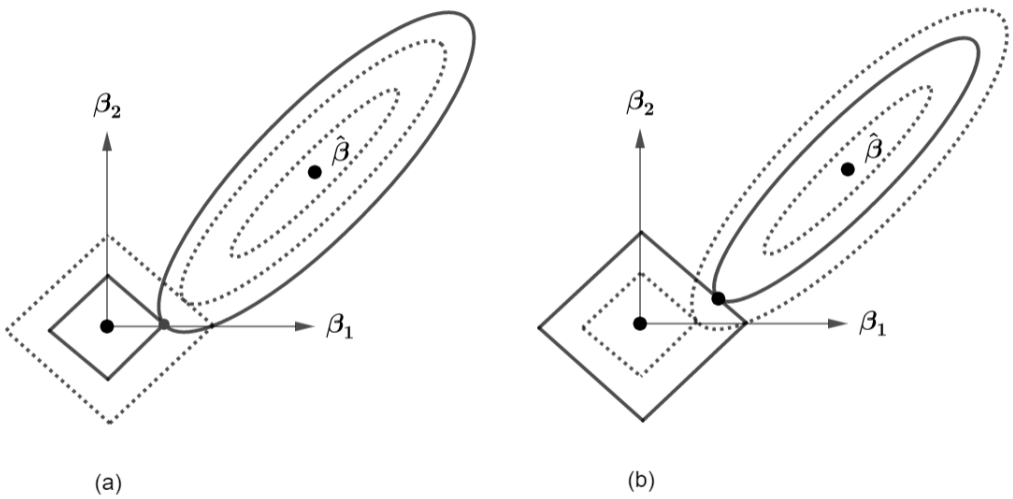


Figure 2.17: Constraint plots for Lasso regression. (a) large λ to only select 1 variable. (b) smaller λ to shrink the coefficients.

2.6 Application

We first note with $p = 53$, Section 2.5.1 suggests best subsets regression is not feasible due to the computing power required to calculate the coefficient estimates for 2^{53} different combinations of the 53 variables. Instead, forward stepwise regression would be feasible by sequentially searching through the models by adding variables in according to their covariance with the residuals from the previous step (see Section 2.5.2).

2.6.1 Forward stepwise regression

The order in which each variable enters the model shall determine the variable's importance in the model. Performing forward stepwise for our main effects model until all covariates are in the model, Tables 2.6 and 2.7 show the variables in order of entry for the non-orthogonalised forward stepwise approach and the orthogonalised approach attributed to Tibshirani et al. (2016). These methods select the same variables for the first 6 steps: `organic`, `rainfall June`, `sprays`, `rainfall December`, `rainfall January` and `machinery`. Organic status being selected first confirms what was seen in Section 2.2; organic status has the largest marginal influence on wheat yields and so enters the model first. Rainfall in June, December and January may be indicative of rain at key stages of

the crop's growing cycle. A large number of variables representing farming practices lie in the first half of the table for both forward stepwise applications. This could be a feature of the association of each of the weather conditions with yield per hectare potentially being encapsulated by another weather condition if they are highly correlated, hence pushing them further down the list. One final point to make is education of the farmer is not deemed important according to the order of the variables entering the model.

Order	Variable
1	Organic
2	Rainfall June
3	Sprays
4	Rainfall December
5	Machinery
6	Rainfall January
7	Sunshine January
8	Rainfall April
9	UAA
10	Sunshine February
11	Mean temp August
12	Contract
13	LAND
14	OtherVC
15	Rainfall November
16	Mean temp December
17	Labour
18	Fert
19	Rainfall September
20	Sunshine April
21	Rainfall February
22	Education 1
23	Seeds
24	Mean temp May
25	Mean temp January
26	Sunshine December
27	Fuel

Order	Variable
28	Education 3
29	Education 4
30	Sunshine October
31	Mean temp July
32	Mean temp October
33	Mean temp November
34	Rainfall May
35	Sunshine May
36	Sunshine November
37	Rainfall March
38	Rainfall August
39	Rainfall October
40	TOFC
41	Sunshine September
42	Mean temp September
43	Sunshine March
44	Mean temp June
45	Sunshine June
46	Sunshine August
47	Mean temp April
48	Education 2
49	Mean temp March
50	Education 5
51	Mean temp February
52	Rainfall July
53	Sunshine July

Table 2.6: Order of appearance for variables using forward stepwise regression without orthogonalisation after each step.

Order	Variable
1	Organic
2	Rainfall June
3	Sprays
4	Rainfall December
5	Rainfall January
6	Machinery
7	Rainfall April
8	Sunshine February
9	Fert
10	Mean temp January
11	UAA
12	Contract
13	LAND
14	Seeds
15	OtherVC
16	Mean temp August
17	Rainfall November
18	Sunshine April
19	Rainfall September
20	Sunshine January
21	Mean temp December
22	Mean temp October
23	Mean temp May
24	Rainfall May
25	Rainfall March
26	Sunshine June
27	Rainfall July

Order	Variable
28	Mean temp July
29	Mean temp September
30	Labour
31	Sunshine September
32	Mean temp February
33	Mean temp November
34	Sunshine August
35	Sunshine March
36	Education 3
37	Education 4
38	Mean temp June
39	Rainfall August
40	Education 1
41	Sunshine May
42	Sunshine November
43	TOFC
44	Sunshine December
45	Fuel
46	Rainfall February
47	Sunshine July
48	Mean temp April
49	Rainfall October
50	Education 2
51	Education 5
52	Sunshine October
53	Mean temp March

Table 2.7: Order of appearance for variables using forward stepwise regression with orthogonalisation after each step.

The minimum mean squared prediction error (see Section 2.3.4) found with main effect models constructed using forward stepwise regression without orthogonalising after every step is 2.045 achieved with the first 5 variables. With orthogonalisation after every step, the smallest mean squared prediction was 1.873, using the first 17 variables selected from Table 2.7 given in Table 2.8 .

Running both forward stepwise algorithms on the model with main effects, non-linear and interaction terms, the smallest mean squared prediction error 1.913 are achieved by selecting 15 variables from applying the forward stepwise procedure with orthogonalisation after every step. This is not smaller than the mean squared prediction error calculated using the first 17 selected from the model with the main effects only, hence the analyses based on the model of higher complexity are omitted.

Table 2.8 gives the coefficients estimated if we looked at the data once to perform variable selection and then a second time to linearly regress with the subset of covariates selected, with the predictions using these estimated coefficients in Figure 2.18. Confidence intervals here are computed based on the asymptotic normality assumption in Section 2.3.2 but for the selected variables only. We provide these confidence intervals with a caution that they will not capture the uncertainty associated with the variable selection algorithm (see Chapter 3).

We also calculate the increase in the R^2 value at each step when another variable is added. The R^2 value is calculated by subtracting the residual sum of squares (Equation 2.3), using the linear regression coefficients, from the total sum of squares (i.e. Equation 2.3 with intercept only $\hat{\alpha} = \bar{y}$) and dividing by the total sum of squares. This will measure the proportion of variation explained by the linear model, therefore finding the increase in R^2 will determine how much more variation is explained compared to the previous model. The largest noticeable increase in R^2 is for the `organic` variable and the increase in R^2 diminishes as less important variables are included in the model.

Order	Variable	$\hat{\beta}_{\text{LR}}$	R^2 increase
-	`Intercept` α	$8.4081\ (7.5363, 9.2799)$	-
1	`Organic`	$-2.1192\ (-2.3534, -1.8850)$	0.1600
2	`Rainfall June`	$-0.0054\ (-0.0066, -0.0042)$	0.0908
3	`Sprays`	$0.0064\ (0.0057, 0.0072)$	0.0352
4	`Rainfall December`	$-0.0088\ (-0.0102, -0.0075)$	0.0248
5	`Rainfall January`	$0.0032\ (0.0017, 0.0046)$	0.0164
6	`Machinery`	$0.0015\ (0.0013, 0.0017)$	0.0138
7	`Rainfall April`	$-0.0092\ (-0.0104, -0.0079)$	0.0091
8	`Sunshine February`	$0.0234\ (0.0208, 0.0260)$	0.0118
9	`Fert`	$0.0037\ (0.0031, 0.0044)$	0.0126
10	`Mean temp January`	$-0.1768\ (-0.2176, -0.1360)$	0.0086
11	`UAA`	$0.0006\ (0.0005, 0.0008)$	0.0079
12	`Contract`	$0.0012\ (0.0009, 0.0016)$	0.0066
13	`LAND`	$0.0011\ (0.0008, 0.0014)$	0.0048
14	`Seeds`	$-0.0048\ (-0.0062, -0.0034)$	0.0042
15	`OtherVC`	$0.0028\ (0.0019, 0.0037)$	0.0038
16	`Mean temp August`	$-0.1415\ (-0.1948, -0.0883)$	0.0020
17	`Rainfall November`	$-0.0031\ (-0.0044, -0.0018)$	0.0021

Table 2.8: Linear regression coefficients for the first 17 variables selected by forward stepwise regression with orthogonalisation. The 95% confidence intervals and increase in R^2 are computed based on performing linear regression on the selected variables.

Order	Variable
1	Organic
2	Sprays
3	Rainfall June
4	Sunshine June
5	Rainfall April
6	Rainfall December
7	Sunshine December
8	Sunshine February
9	Fert
10	UAA
11	Machinery
12	OtherVC
13	Mean temp August
14	LAND
15	Labour
16	Seeds
17	Mean temp December
18	Sunshine November
19	Contract
20	Education 1
21	Rainfall November
22	Rainfall July
23	Mean temp May
24	Education 3
25	Fuel
26	Sunshine April
27	Sunshine October

Order	Variable
28	Rainfall May
29	Mean temp October
30	Education 4
31	Rainfall September
32	Rainfall March
33	Sunshine January
34	TOFC
35	Mean temp November
36	Sunshine September
37	Sunshine July
38	Sunshine May
39	Mean temp February
40	Mean temp July
41	Rainfall August
42	Mean temp September
43	Mean temp January
44	Sunshine March
45	Education 2
46	Rainfall October
47	Education 5
48	Rainfall February
49	Sunshine August
50	Rainfall January
51	Mean temp April
52	Mean temp June
53	Mean temp March

Table 2.9: Order of appearance for variables using Lasso regression.

2.6.2 Lasso regression

The order of appearance of the variables from applying Lasso regression (Section 2.5.3) is slightly different to the order found by applying forward stepwise regression to the model with main effects only. Table 2.9 shows the variables in order of appearance in the lasso regression path, with organic status, crop protection and rainfall in June again appearing as the first three. They are now followed by `sunshine June`, `rainfall April` and `rainfall December`.

The mean squared prediction error again can be calculated using Equation 2.9 for each model constructed along the Lasso path. The minimum mean squared prediction error, 1.903, is achieved with 18 variables, with those selected being listed in Table 2.10. Again, repeating the analysis for the model with main effects, non-linear and interaction terms, the same model is selected. Therefore even though non-linear and interaction terms were allowed to be included in the model, non-linear and interaction terms were not required to minimise the MSPE. The minimum mean squared prediction error achieved here is larger than that achieved using forward stepwise regression with orthogonalisation in the previous section due to selecting different important variables to be included in the modelling according to their respective variable selection algorithms. Section 2.5.3 suggested Lasso regression is preferable when creating a model based on the variables with the largest association with yield per hectare for the training data, however predictions made using the out-of-sample data may be further away from their observed yields per hectare.

Estimated coefficients in Table 2.10 are those from linearly regressing the selected 18 variables onto yield per hectare. Again, confidence intervals here are computed based on the asymptotic normality assumption in Section 2.3.2 for the selected variables only but again, these need to be examined with caution due to not accounting for uncertainty when selecting the model, which we are to investigate in the following chapter.

Figure 2.18 shows the predictions from the linear model consisting of the variables selected using Lasso regression with their coefficients estimated using linear regression. There is little difference between the predictions using Lasso regression and forward stepwise

with orthogonalisation from the previous section. Comparing these to the predictions using linear regression and principal component regression based on the first 10 principal components in Figure 2.15, there is also little difference however the models based on variable selection algorithms will be based on fewer variables, hence allows stakeholders to focus on key variables to improve wheat yields.

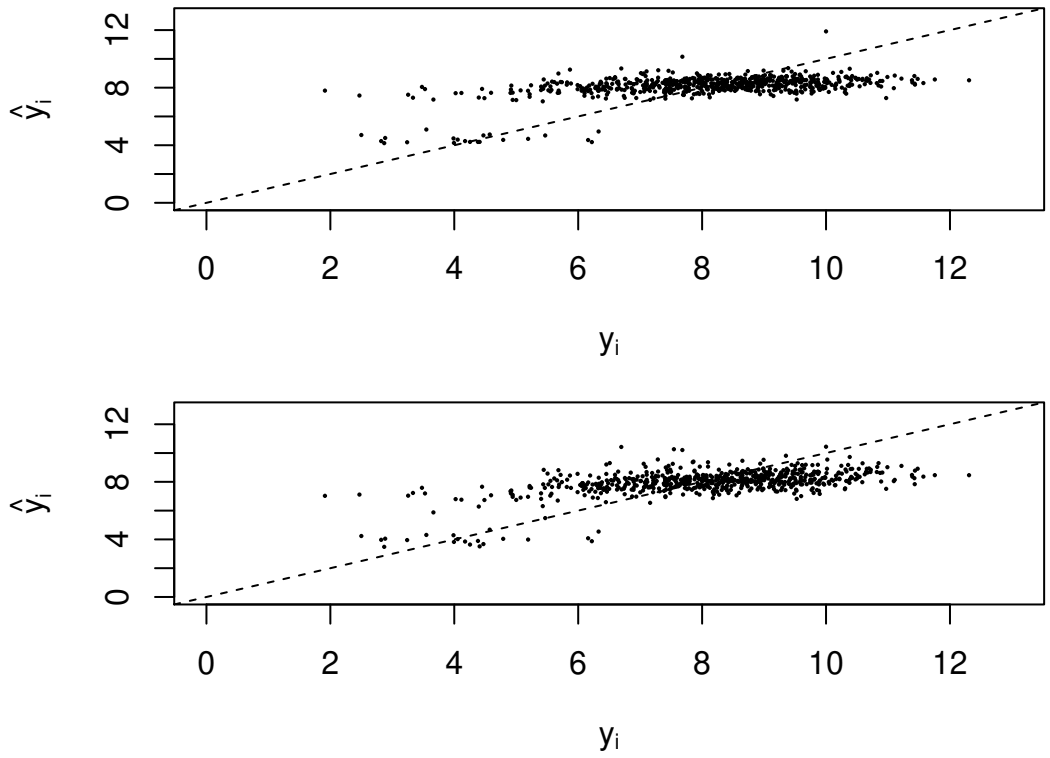


Figure 2.18: Observed yield versus predicted yield for the main effects model using, left: linear regression after selecting variables based on forward stepwise regression with orthogonalisation, right: linear regression after selecting variables based on Lasso regression.

Order	Variable	$\hat{\beta}_{\text{OLS}}$	R^2 increase
-	`Intercept` α	$8.8585\ (7.8971, 9.8198)$	-
1	`Organic`	$-2.1670\ (-2.4037, -1.9302)$	0.1600
2	`Sprays`	$0.0062\ (0.0054, 0.0069)$	0.0392
3	`Rainfall June`	$-0.0056\ (-0.0069, -0.0043)$	0.0867
4	`Sunshine June`	$0.0020\ (0.0005, 0.0035)$	0.0076
5	`Rainfall April`	$-0.0076\ (-0.0089, -0.0064)$	0.0144
6	`Rainfall December`	$-0.0070\ (-0.0082, -0.0059)$	0.0178
7	`Sunshine December`	$0.0058\ (0.0017, 0.0100)$	0.0020
8	`Sunshine February`	$0.0210\ (0.0186, 0.0233)$	0.0198
9	`Fert`	$0.0036\ (0.0029, 0.0043)$	0.0101
10	`UAA`	$0.0006\ (0.0005, 0.0007)$	0.0066
11	`Machinery`	$0.0010\ (0.0008, 0.0012)$	0.0147
12	`OtherVC`	$0.0029\ (0.0020, 0.0037)$	0.0058
13	`Mean temp August`	$-0.2049\ (-0.2478, -0.1619)$	0.0115
14	`LAND`	$0.0012\ (0.0009, 0.0015)$	0.0052
15	`Labour`	$0.0005\ (0.0002, 0.0008)$	0.0010
16	`Seeds`	$-0.0044\ (-0.0059, -0.0030)$	0.0022
17	`Mean temp December`	$-0.0623\ (-0.0954, -0.0291)$	0.0048
18	`Sunshine November`	$-0.0068\ (-0.0100, -0.0037)$	0.0017

Table 2.10: Linear regression coefficients for the first 18 variables selected by Lasso regression. The 95% confidence intervals are computed based on performing linear regression on the selected variables.

2.7 Conclusion

Our analysis from applying different regression and variable selection techniques to the data from the Farm Business Survey and UK Met Office found various interesting aspects of our data. Linear regression found how important each variable was conditional on all of the other variables being included in the model. This indicated which farming practices and weather conditions are most associated with yield per hectare by assessing

whether each variable should be included in the model, given all of the other variables are already included. Instead of using all of the data, we used principal component regression to first create vectors which will capture the largest correlation in the variables instead and then regress upon these. The first few principal components was based on weather conditions only. This achieved a smaller MSPE compared to linear regression, however principal component regression is not designed to include the variables most associated with yield. Forward stepwise and Lasso regression are designed to find the model which will capture the largest association with yield per hectare, both finding organic status, crop protection and rainfall in June are to be included in a model for yield per hectare.

The model which achieves the closest prediction to the observed yields, according to mean squared prediction error, is found using the main effects with forward stepwise regression with orthogonalisation, which includes most of the variables taken from the Farm Business Survey (`seeds`, `fert`, `sprays`, `otherVC`, `contract`, `machinery`, `LAND`, utilised agricultural area, organic status) and a small number of weather variables (monthly rainfall in November, December, January, June and April, monthly mean temperature in January and August, monthly sunshine hours in February). A model with interaction terms was also considered for each regression and variable selection method, however these did not achieve a smaller mean squared prediction error than the one described using main effects only. The coefficients for seed costs and organic status are found to be negatively associated with yield per hectare. This is a result of farmers applying excessive amounts of seeds in an attempt to achieve a better yield regardless of other factors which may influence wheat yield. Organic farms are found to do this more often due to not applying crop protection. All of the remaining farming practices from the model achieving the smallest mean squared prediction are positively associated with yield per hectare.

2.8 Discussion

To incorporate weather conditions into our model, we only considered the weather conditions as main effects and combined these with our variables from the Farm Business

Survey. This gave a brief insight into how weather may be associated with yield per hectare, however from previous studies, the interactions between weather conditions are often key to predicting wheat yield in a changing climate. Further work would involve looking at potential interaction terms within the weather conditions. Furthermore on reflection of our work, it is difficult to attribute an increase in annual yield to a monthly climate variable. Preferably, we would like to incorporate a measure of how much the yield has grown during each month and examine the association of growth with weather conditions, however this is difficult to observe whilst the crop grows, in particular when the crop is still underground. Future research would investigate potential simulation methods to incorporate this into our model.

Throughout this chapter, we looked to predict yields per hectare for 2009 since the yields per hectare attained during this year are representative of typical yields, hence we built a model based on the remaining 9 years of data. Due to the weather conditions being highly correlated, predicting a different year will most likely result in a different set of variables being selected in the variable selection algorithms. This is indicative of the ordering of our model selection stage being unstable to the choice of year. An fruitful avenue of future work would be to cross-validate in our regression and variable selection methods by predicting each year in turn or subsets of observations from each year.

Finally, here we only look at the confidence intervals based on the linear regression but emphasise this will only account for uncertainty in the parameter estimation and will not account for uncertainty in the variable selection stage. Appendix A discusses capturing both of these stages of uncertainty by conditioning on a polyhedral set (Tibshirani (1996)), however finds if the ordering of the variables selected is unstable, that is the ordering of the variables change when the training data changes, then the confidence intervals for the parameter estimates may be $\pm\infty$. Again, this motivates trying to fit a model based on cross-validating our data in the future to first confirm our ordering of the variables is robust to different subsets of data, and hence whether it is possible to conduct inference based on conditioning on a polyhedron.

3 | Bayesian inference for model selection coefficients

3.1 Introduction

Uncertainty in our predictions accumulates at two stages of our analysis: performing model selection and parameter estimation. Predictions of UK wheat yields should be accompanied by the associated uncertainty to assess a model's predictive performance and not solely rely on point estimates. In agricultural studies, uncertainty is often not fully accounted for by taking a two-step approach by fitting a regression model after selection (e.g. Landau et al. (1998), Landau et al. (2000)). Tibshirani et al. (2016) made progress on frequentist approaches to post-selection inference by visualising conditional polyhedrons, however Tibshirani et al. (2019) subtly mentioned this method computes unreasonable confidence intervals when the coefficient estimate using linear regression falls close to one of the edges of the conditional polyhedron; the tails of the truncated normal distribution are too short to find a confidence interval. Appendix A gives further details on this method. This manifests itself in the variable selection algorithm when two or more variables are equally important to model yields and enter the model in quick succession. This may occur more often for datasets with a large number of variables.

In the previous chapter, uncertainty was calculated for the selected variables based on the parameter estimation stage using the asymptotic normality assumption of linear regression coefficient estimates. This did not take account of the uncertainty in the model selection stage. In this chapter, we address this problem using Bayesian shrinkage priors

to induce sparsity and construct credible intervals accounting for both the uncertainty of model selection and the uncertainty of the parameters simultaneously. Bayesian analysis has been used in wheat genomics (e.g. Montesinos-López et al. (2018)) and field experiments (e.g. Besag and Higdon (1999)), however it is yet to be used for UK crop yields as a linear model of monthly weather conditions and agronomic and socio-ecological factors when producing wheat due to scarce datasets on UK farming practices. This is now possible using data from the Farm Business Survey.

We will apply two Bayesian shrinkage priors which have gained recent attention: the Bayesian Lasso (Park and Casella (2008)) and the horseshoe prior (Carvalho et al. (2009), Carvalho and Polson (2010)). The former is the Bayesian analogue of the frequentist Lasso and the latter is a modified version of the Bayesian Lasso prior hierarchy more suited to perform variable selection and account for uncertainty.

The structure of this chapter is as follows. Sections 3.1.1, 3.1.2, 3.1.3 and 3.1.4 discuss the preliminary methodology to perform Bayesian inference. Section 3.2 introduces the hierarchical prior structure for shrinkage. Section 3.3 details the sampling schemes for the Bayesian Lasso with applications in Sections 3.3.4 and 3.3.5. Section 3.4 exposes the overshrinkage feature of the Bayesian Lasso and how the horseshoe prior can rectify this. Finally, Section 3.5 discusses the horseshoe prior hierarchy and sampling schemes with applications in Sections 3.5.3 and 3.5.4.

3.1.1 Bayesian inference

Bayesian inference uses Bayes theorem to combine data and prior beliefs of the model parameters to find the posterior density of the set of parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ to be estimated in the likelihood $\pi(\boldsymbol{y}|\boldsymbol{\theta})$. The posterior density of $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \propto \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

is calculated from a combination of the likelihood density $\pi(\boldsymbol{y}|\boldsymbol{\theta})$, assumed from the distribution of the data and the prior density $\pi(\boldsymbol{\theta})$ of the model parameters. The prior distribution $\pi(\boldsymbol{\theta})$ incorporates knowledge about the parameter vector $\boldsymbol{\theta}$, before the anal-

ysis is undertaken. The posterior density in most practical cases can only be calculated up to proportionality due to the integral in the denominator often being difficult to evaluate.

Furthermore, given the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, the posterior predictive distribution

$$\pi(\tilde{y}|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \pi(\tilde{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}$$

in the present context finds the distribution of a new yield (per hectare) $\tilde{y}$, given what we have already seen from the sample of yields $\boldsymbol{y}$ through the posterior distribution. Credible intervals of the posterior predictive distribution can be compared to the observed values for $\tilde{y}$ to assess the performance of our model for new values.

3.1.2 Bayesian linear regression

In Section 2.3.2 for a set of fixed covariates $\boldsymbol{x}_1, ..., \boldsymbol{x}_p \in \mathbb{R}^n$, linear regression assumed the response $\boldsymbol{y} \in \mathbb{R}^n$ is distributed according to $\boldsymbol{y} \sim N_n(\alpha\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, where $\alpha \in \mathbb{R}$ is the unknown intercept parameter, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown coefficient vector, $\boldsymbol{X} = (\boldsymbol{x}_1, .., \boldsymbol{x}_p) \in \mathbb{R}^{n\times p}$, $\mathbf{1}_n$ is the vector of ones of length n, $\sigma^2 \in \mathbb{R}^+$ is the unknown constant error variance and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Bayesian linear regression looks to estimate the intercept α, coefficient vector $\boldsymbol{\beta}$ and error variance σ^2 whilst incorporating prior information on the model parameters $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \sigma^2)$ through the prior density $\pi(\alpha, \boldsymbol{\beta}, \sigma^2)$. Using Bayes theorem, the joint posterior distribution of these will be

$$\pi(\alpha, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) \propto \pi(\boldsymbol{y}|\boldsymbol{X}, \alpha, \boldsymbol{\beta}, \sigma^2)\pi(\alpha, \boldsymbol{\beta}, \sigma^2),$$

where $\pi(\alpha, \boldsymbol{\beta}, \sigma^2)$ is the joint prior distribution of the parameters of interest and

$\pi(\boldsymbol{y}|\boldsymbol{X}, \alpha, \boldsymbol{\beta}, \sigma^2)$ is the likelihood density assumed to be normally distributed with mean $\alpha\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}$ and variance $\sigma^2\boldsymbol{I}_n$. For Bayesian linear regression, little is known about the coefficients or the error variance σ^2 a priori except their respective supports. This leads to the parameters α, $\boldsymbol{\beta}$ and σ^2 assumed to have independent prior distributions $\pi(\alpha)$,

$\pi(\boldsymbol{\beta})$ and $\pi(\sigma^2)$, taken to be non-informative

$$\pi(\alpha) \sim N(0, \nu^2), \qquad \pi(\beta_j) \sim N(0, \lambda^2), \qquad \pi(\sigma^{-2}) \sim \text{Gamma}(b_1, b_2),$$

for $j \in \{1, ..., p\}$, ν^2 large, λ^2 large and b_1 and b_2 being close to zero, equivalently expressed as

$$\pi(\alpha) \propto 1, \qquad \pi(\beta_j) \propto 1, \qquad \pi(\sigma^2) \propto 1/\sigma^2,$$

to be used when finding the conditional posterior distributions. Section 3.2 will look at using a different prior distribution for $\boldsymbol{\beta}$ to induce sparsity, whilst keeping the same priors for α and σ^2.

In practice, the posterior distribution is difficult to sample from directly. The following section looks at how to approximately sample from the posterior distribution using Markov Chain Monte Carlo.

3.1.3 MCMC methods

Markov Chain Monte Carlo (MCMC) methods allow approximate samples to be generated from a target distribution, in our case the posterior distribution, by sampling from a proposal distribution. The proposal distribution only depends on the previous value drawn hence forming a Markov Chain. Since its inception, the Metropolis-Hastings algorithm (Metropolis et al. (1953), Hastings (1970)) underpins current extensively used MCMC methods including the Gibbs sampler and Hamiltonian Monte Carlo discussed here. Algorithm 1 describes the Metropolis-Hastings algorithm to sample from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ for the model parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$.

Algorithm 1: Metropolis-Hastings sampler

1. Initialise: $\boldsymbol{\theta^{(0)}} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$

2. Iterate: Propose a new vector $\boldsymbol{\theta}^*$ from proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})$.

Set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$ with probability

$$\alpha = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{y})q(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{t-1}|\boldsymbol{y})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})}\right)$$

else $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$.

3. Repeat: Repeat step 2 M times.

The resulting samples $\boldsymbol{\theta}^{(1)}, ...\boldsymbol{\theta}^{(M)}$ will approximately come from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. When the proposal distribution q is symmetric, $q\left(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}\right)$, this is referred to as the Metropolis algorithm (Metropolis et al. (1953)).

So far all samples generated from the Metropolis-Hastings algorithm $\boldsymbol{\theta}^{(1)}, .., \boldsymbol{\theta}^{(M)}$ are used to approximate the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. If no prior knowledge of the target distribution is used to initialise the sampler, the samples generated at the start of the Markov Chain may be far away from the target distribution. There can also be high dependency between sequential samples due to the transition kernel q. To ensure our samples are independent and approximately drawn from the target distribution, a burn-in period is used to correct for the initial samples and the resulting sample is thinned to remove dependency between the samples. We shall also initialise the sampler numerous times and combine the resulting samples to be sure our sampler has thoroughly searched the parameter space of the target distribution.

The Metropolis-Hastings algorithm also relies on a suitably chosen proposal distribution q to thoroughly search the parameter space. If not chosen appropriately, the sampled values of $\boldsymbol{\theta}$ will be far from the target distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. The following two approaches to MCMC avoid selecting a proposal distribution.

Gibbs Sampler

The Gibbs sampler (Geman and Geman (1984)) selects the proposal distribution q in the Metropolis-Hastings algorithm so the proposed new value $\boldsymbol{\theta}^*$ is always accepted. To

sample from the posterior distribution $\pi(\boldsymbol{\theta})$, the Gibbs sampler samples from each of the full conditional distributions in a deterministic order. Given the set of parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, the Gibbs sampling algorithm is described in Algorithm 2.

Algorithm 2: Gibbs sampler

1. Initialise: $\boldsymbol{\theta^{(0)}} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$

2. Iterate: Draw from the following conditional distributions

- $\theta_1^{(t)} \sim \pi(\theta_1 | \theta_2^{(t-1)}, ..., \theta_p^{(t-1)})$

  $\vdots$

- $\theta_j^{(t)} \sim \pi(\theta_j | \theta_1^{(t)}, ..., \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, ..., \theta_p^{(t-1)})$

  $\vdots$

- $\theta_p^{(t)} \sim \pi(\theta_p | \theta_1^{(t)}, ..., \theta_{p-1}^{(t)})$

3. Repeat: Repeat step 2 M times.

For most applications, the full conditionals are common distributions which are easy to sample from. If not, the Gibbs sampler can be computationally expensive. The following solution avoids finding full conditionals.

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (Neal (2011), Betancourt and Girolami (2013)) uses Hamiltonian dynamics to sample from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ by first introducing momentum variables $\phi_j \in \mathbb{R}$ for each of the positions of the parameters of interest $\theta_j \in \mathbb{R}$, $j = 1, ..., p$, and defines a Hamiltonian, $H(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$,

$$\begin{aligned} H(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) &= -\log(\pi(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})), \\ &= -\log(\pi(\boldsymbol{\phi}|\boldsymbol{\theta}, \boldsymbol{y})) - \log(\pi(\boldsymbol{\theta}|\boldsymbol{y})), \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p) \in \mathbb{R}^p$ and $\boldsymbol{\phi} = (\phi_1, ..., \phi_p) \in \mathbb{R}^p$. Assuming $\boldsymbol{\phi}$ is independent of $\boldsymbol{\theta}$ and the data $\boldsymbol{y}$, then the Hamiltonian becomes

$$\begin{aligned} H(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) &= -\log(\pi(\boldsymbol{\phi})) - \log(\pi(\boldsymbol{\theta}|\boldsymbol{y})) \\ &= K(\boldsymbol{\phi}) + V(\boldsymbol{\theta}|\boldsymbol{y}), \end{aligned}$$

where $K(\boldsymbol{\phi})$ is referred to as the kinetic energy of the system and $V(\boldsymbol{\theta}|\boldsymbol{y})$ is referred to as the potential energy of the system (Neal (2011)). With momentum vector $\boldsymbol{\phi}$ and mass matrix $\boldsymbol{T} \in \mathbb{R}^{p \times p}$, kinetic energy is commonly defined to be

$$K(\boldsymbol{\phi}) = \frac{1}{2}\boldsymbol{\phi}^T \boldsymbol{T}^{-1} \boldsymbol{\phi},$$

which conveniently corresponds to minus log probability density of the zero-mean Gaussian distribution with covariance matrix $\boldsymbol{T}$, hence

$$\boldsymbol{\phi} \sim N_p(\mathbf{0}, \boldsymbol{T}),$$

where $\boldsymbol{T}$ is typically a diagonal matrix (Neal (2011)) such that the components of $\boldsymbol{\phi}$ are also independent. The Hamiltonian system $H(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$ will then be governed by the Hamiltonian equations

$$\begin{aligned} \frac{d\theta_j}{dt} &= \frac{\partial H}{\partial \phi_j} = \frac{\partial K(\boldsymbol{\phi})}{\partial \phi_j} = \left[\boldsymbol{T}^{-1}\boldsymbol{\phi}\right]_j, \\ \frac{d\phi_j}{dt} &= -\frac{\partial H}{\partial \theta_j} = -\frac{\partial V(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_j} = \frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_j}. \end{aligned}$$

Hamiltonian Monte Carlo alternates between taking step size of $\epsilon \in \mathbb{R}$ from $\boldsymbol{\theta}$ along its gradient $d\boldsymbol{\theta}/dt$ and a step of size ϵ for $\boldsymbol{\phi}$ along its gradient $d\boldsymbol{\phi}/dt$, i.e.

$$\begin{aligned} \boldsymbol{\phi} + \frac{\epsilon}{2}\frac{d\boldsymbol{\phi}}{dt} &= \boldsymbol{\phi} + \frac{\epsilon}{2}\nabla \log \pi(\boldsymbol{\theta}|\boldsymbol{y}), \\ \boldsymbol{\theta} + \epsilon\frac{d\boldsymbol{\theta}}{dt} &= \boldsymbol{\theta} = \boldsymbol{\theta} + \epsilon \boldsymbol{T}^{-1}\boldsymbol{\phi}, \end{aligned}$$

where

$$\nabla \log \pi(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1}, ..., \frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_d} \right).$$

After iterating between these a fixed number of times to propose a new value $\boldsymbol{\theta}^*$, this process manifests itself as the Markov Chain transitions, previously seen through kernel q for the Metropolis-Hastings algorithm. Neal (2011) found these Markov Chain transitions to be reversible hence a Metropolis acceptance step is used to sample from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. With discretization step ϵ, the Hamiltonian Monte Carlo algorithm is given in its entirety in Algorithm 3.

Algorithm 3: Hamiltonian Monte Carlo

1. Initialise: Randomly generate $\boldsymbol{\theta^{(0)}} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$. Draw $\boldsymbol{\phi} \sim N_p(0, \boldsymbol{T})$.

2. Iterate

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{\epsilon}{2} \nabla \log \pi(\boldsymbol{\theta}|\boldsymbol{y}),$$
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon \boldsymbol{T}^{-1} \boldsymbol{\phi},$$
$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{\epsilon}{2} \nabla \log \pi(\boldsymbol{\theta}|\boldsymbol{y}),$$

3. Repeat: Repeat step 2 L times.

4. Metropolis acceptance step:

Propose a new vector $\boldsymbol{\theta}^* = \boldsymbol{\theta}$.

Set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$ with probability $\alpha = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{y})\pi(\boldsymbol{\phi}^*)}{\pi(\boldsymbol{\theta}^{t-1}|\boldsymbol{y})\pi(\boldsymbol{\phi}^{t-1})}\right)$

else $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$.

5. Repeat: Repeat steps 2-4 M times.

3.1.4 Posterior predictive distribution for Bayesian linear regression

Once one of the MCMC methods in Section 3.1.3 has been used to create a sample approximately coming from the posterior distribution, we can also sample approximately from the posterior predictive distribution for a new observation. Suppose

$$\boldsymbol{\theta}^{(t)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}), \qquad \boldsymbol{\theta}^{(t)} = (\alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \sigma^{2(t)})$$

is the t^{th} sample from the (approx) posterior distribution, for $t = 1, ..., M$, where $\boldsymbol{y} \in \mathbb{R}^n$

and $\boldsymbol{X}^{n\times p}$ are the samples used in the modelling. Given a new set of covariates $\tilde{\boldsymbol{x}}$ to predict a new response $\tilde{y}$, a sample approximately coming from the posterior predictive distribution is sampled from the likelihood $\pi(y|\boldsymbol{x} = \tilde{\boldsymbol{x}}, \boldsymbol{\theta}) \sim N(\alpha + \tilde{\boldsymbol{x}}^T\boldsymbol{\beta}, \sigma^2)$ with $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \sigma^2)$ replaced with their respective sample quantities $\boldsymbol{\theta}^{(j)} = \left(\alpha^{(j)}, \boldsymbol{\beta}^{(j)}, \sigma^{2^{(j)}}\right)$ randomly drawn from the posterior distribution with probability $1/M$. For draw i from the posterior predictive distribution, $i = 1, ..., N$,

$$\tilde{y}^{(i)} \sim N(\alpha^{(j)} + \tilde{\boldsymbol{x}}^T\boldsymbol{\beta}^{(j)}, \sigma^{2^{(j)}}),$$

where after each draw from the posterior predictive distribution, $\boldsymbol{\theta}^{(j)}$ is redrawn from the posterior distribution with probability $1/M$. The resulting sample $\tilde{y}^{(1)}, ..., \tilde{y}^{(N)}$ will approximately come from the posterior predictive distribution for the new response $\tilde{y}$, conditional on the associated set of covariates $\tilde{\boldsymbol{x}}$. In the present context, we shall find the posterior predictive distribution of a new yield (per hectare) conditional on the associated set of covariates from the Farm Business Survey and the UK Met Office.

3.2 Bayesian inference with shrinkage priors

Bayesian linear regression in Section 3.1.2 took the prior densities for the coefficients $\boldsymbol{\beta}$ to be distributed as

$$\pi(\beta_j|\lambda) \sim N(0, \lambda^2), \quad j = 1, ..., p, \tag{3.1}$$

with λ^2 large. Instead, suppose the shared error variance λ^2 is small such that the prior on β_j for all $j = 1, .., p$ has a peak at zero. This will shrink the coefficients at the same rate by placing a larger weight at zero. Solely using a fixed global shrinkage parameter will not only shrink the error signals to zero, but it will shrink all coefficients regardless of their importance to model the response. Furthermore, if each β_j, $j = 1, ..., p$, had their own variance parameter λ_j^2, then strict prior knowledge would be needed about each fixed λ_j^2, which would be difficult in most cases.

Ideally, with little prior knowledge, we would like to use a prior density for the coefficients which will shrink the error signals but not apply unnecessary shrinkage to the parameter

coefficients corresponding to the important variables. This will perform variable selection and estimate coefficients which will take account of the variable selection already performed. For this to be the case, a suitable prior will have a large sharp peak at zero and be flat for β_j away from zero. Examples with these properties are shown in Figure 3.1 with their associated prior densities in Table 3.1.

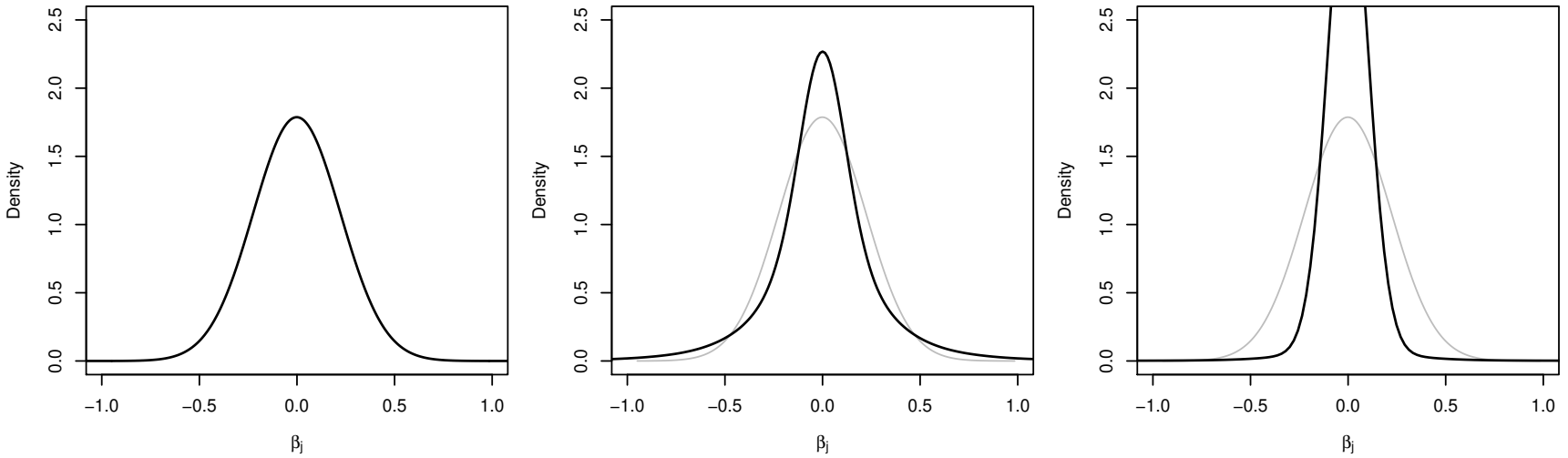


Figure 3.1: Left: normal prior density for $\boldsymbol{\beta}$, centre: Laplace prior density for $\boldsymbol{\beta}$, right: horseshoe prior density for $\boldsymbol{\beta}$. All prior densities are proportional. Grey density represents the normal prior density in all three plots.

Shrinkage prior	**Conditional prior density** $\pi(\beta_j\|...)$	**Conditional prior density** $\pi(\beta_j\|\tau_j...)$	**Hyperprior density** $\pi(\tau_j^2)$ or $\pi(\tau_j)$
Normal	$\pi(\beta_j\|\lambda) = \frac{1}{\sqrt{2\pi\lambda^2}}\exp\left(-\frac{\beta_j^2}{2\lambda^2}\right)$	NA	NA
Laplace	$\pi(\beta_j\|\lambda,\sigma) = \frac{\sqrt{2}}{\lambda\sigma}\exp\left(-\frac{\sqrt{2}\|\beta_j\|}{\lambda\sigma}\right)$	$\pi(\beta_j\|\tau_j,\lambda,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2\lambda^2\tau_j^2}}\exp\left(-\frac{\beta_j^2}{2\sigma^2\lambda^2\tau_j^2}\right)$	$\pi(\tau_j^2) = \exp\left(-\tau_j^2\right)$
Horseshoe	Not analytically tractable	$\pi(\beta_j\|\tau_j,\lambda,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2\lambda^2\tau_j^2}}\exp\left(-\frac{\beta_j^2}{2\sigma^2\lambda^2\tau_j^2}\right)$	$\pi(\tau_j) = 2\left(\pi\left(1+\tau_j^2\right)\right)^{-1}, \quad \tau_j \geq 0$

Table 3.1: Common shrinkage prior densities for β_j, $j = 1, ..., p$. Where required for MCMC sampling, the conditional prior density for β_j is decomposed into a conditional prior density $\pi(\beta_j|\tau_j, ...)$ and hyperprior density for τ_j^2 (Laplace) or τ_j (horseshoe).

Another suitable prior would be to use a scale mixture of two Normal distributions (Mitchell and Beauchamp (1988), George and McCulloch (1993)),

$$\pi(\beta_j|\nu_j) = \nu_j N(0, \lambda^2) + (1 - \nu_j)N(0, \epsilon^2),$$

referred to as a spike-and-slab prior (Ishwaran and Rao (2005)), where ν_j is a zero-one latent variable, $\lambda^2 > 0$ is fixed to be suitably large and $\epsilon^2 > 0$ is fixed to be suitably small, often taken to equal zero such that $N(0, \epsilon^2)$ becomes a delta spike at zero (Piironen and Vehtari (2017a)). Although the spike-and-slab prior is a popular choice for Bayesian variable selection, Piironen and Vehtari (2017a) found the horseshoe prior, a natural extension to the Laplace shrinkage prior, has comparable performance to the spike-and-slab prior. Since the Laplace shrinkage prior has analogous links to the frequentist Lasso in Chapter 2, the following section will look closer at the Laplace prior for β_j.

From here on in, λ will be referred to as the global shrinkage parameter. Furthermore, to avoid shrinking based on the magnitude of each variable rather than their relationship with the response, we will first find parameter estimates $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ using the standardised responses $\boldsymbol{y}$, with mean $\bar{y}$ and standard deviation s_y, and the standardised covariates $\boldsymbol{x}_j$, with mean $\bar{x}_j$ and standard deviation s_{x_j}. Standardised responses $\boldsymbol{y}$ and standardised covariates $\boldsymbol{x}_j$ are further divided by $\sqrt{n}$ to ensure the global shrinkage parameter λ is on a practical scale whilst not influencing our coefficient estimates. Estimating parameters from standardised vectors might not be a good idea since the coefficients will no longer reflect the magnitude of each variable. After finding parameter estimates $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ using the standardised vectors, the parameters are transformed back to reflect their original scales

$$\hat{\alpha}^* = \bar{y} + s_y \left(\hat{\alpha}\sqrt{n} - \sum_{j=1}^{p} \frac{\hat{\beta}_j \bar{x}_j}{s_{x_j}} \right) \quad \hat{\beta}_j^* = s_y \frac{\hat{\beta}_j}{s_{x_j}} \quad \hat{\sigma}^* = s_y \hat{\sigma}\sqrt{n},$$

where $\hat{\alpha}$, $\hat{\beta}_j$, $j = 1, ..., p$, and $\hat{\sigma}$ are the estimated parameters from the standardised vectors and $\hat{\alpha}^*$, $\hat{\beta}_j^*$, $j = 1, ..., p$, and $\hat{\sigma}^*$ are the transformed estimated parameters.

3.3 Bayesian Lasso

Section 2.5.3 discussed the frequentist Lasso to perform model selection by shrinking coefficients down to zero, such that variables of little importance become zero and only those important variables remain. The Bayesian analogue of the Lasso (see Park and Casella (2008)) induces sparsity by weighting the coefficients $\boldsymbol{\beta}$ towards zero using a Laplace prior on $\boldsymbol{\beta}$

$$\pi(\boldsymbol{\beta}|\omega) \propto \exp\left(-\frac{|\boldsymbol{\beta}|}{\omega}\right), \tag{3.2}$$

i.e. Laplace distribution with mean zero and scale parameter ω. Figure 3.1 compares the Normal prior for $\boldsymbol{\beta}$ to the Laplace prior. The Bayesian Lasso has flatter tails for the densities of $\boldsymbol{\beta}$, hence a smaller amount of shrinkage will be applied to important covariates.

Briefly looking at the conditional distribution for $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \alpha, \lambda, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \alpha\mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - \alpha\mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\right) - \frac{1}{\omega}|\boldsymbol{\beta}|\right),$$

it is not possible to sample from this distribution in a Gibbs sampling scheme; the conditional posterior distribution for β_j is not analytically tractable with the prior hierarchy in its current form. However a Gibbs sampling scheme does exist if the prior hierarchy is expanded using distributional properties of the Laplace distribution. Andrews and Mallows (1974) showed a centred Laplace distribution with scale $1/A$ can be decomposed into a centred normal distribution with variance a, and an exponential hyperprior on a with rate $A^2/2$, that is

$$\begin{aligned}
\pi(x|A) &= \int_{a=0}^{\infty} \pi(x|a)\pi(a|A)\mathrm{d}a \\
&= \int_{a=0}^{\infty} \frac{1}{\sqrt{2\pi a}}\exp\left\{-x^2/(2a)\right\}\frac{A^2}{2}\exp\left\{-A^2a/2\right\}\mathrm{d}a \\
&= \frac{A^2}{2}\exp(-A|x|)\cdot\frac{1}{|x|}\int_{a=0}^{\infty} a\frac{|x|}{\sqrt{2\pi a^3}}\exp\left(-\frac{A^2|x|^2\left(a-|x|/A\right)^2}{2|x|^2a}\right)\mathrm{d}a \\
&= \frac{A}{2}\exp\left(-A|x|\right).
\end{aligned}$$

Using this property, the Laplace prior in Equation (3.2) can be written as the prior hierarchy

$$\beta_j|\psi_j^2 \sim N(0, \psi_j^2), \qquad \psi_j^2|\omega \sim \text{Exponential}\left(\frac{1}{2\omega^2}\right). \tag{3.3}$$

To get rid of all dependencies in the hyperpriors for the variance parameters, after a change of variable $\psi_j^2 = 2\omega^2\tau_j^2$, Makalic and Schmidt (2016) showed the prior hierarchy can also be written as

$$\beta_j|\tau_j^2, \omega \sim N(0, 2\omega^2\tau_j^2), \qquad \tau_j^2 \sim \text{Exponential}\,(1)\,. \tag{3.4}$$

The Bayesian Lasso literature has selected ω in various forms: Park and Casella (2008) take $\omega = \sigma/\lambda$ to ensure a unimodal maximum and Makalic and Schmidt (2016) chose $\omega = \lambda\sigma/\sqrt{2}$ to make the conditional prior distribution of β_j comparable to other shrinkage priors. The posterior mode of β_j when $\omega = \sigma^2/\lambda$ and λ fixed, will approximately lead to the frequentist Lasso estimates from Hastie et al. (2008), however Gibbs sampling cannot be used since the full conditional for σ^2 will not be analytically tractable. Since we would like to compare what happens with different local shrinkage priors, we will use $\omega = \lambda\sigma/\sqrt{2}$ and continue to refer to λ as the global shrinkage parameter.

3.3.1 Gibbs sampling conditional posterior distributions for the Bayesian Lasso

To approximately sample from the posterior distribution using the Gibbs sampler, the closed forms of the full conditionals first need to be found. Makalic and Schmidt (2016) showed the prior hierarchy of the Bayesian Lasso is

$$\beta_j|\tau_j^2, \lambda, \sigma^2 \sim N(0, \sigma^2\lambda^2\tau_j^2), \qquad \tau_j^2 \sim \text{Exponential}\,(1)\,, \qquad \alpha \sim 1 \qquad \sigma^2 \sim \frac{1}{\sigma^2}, \tag{3.5}$$

where α, β_j for $j = 1, ..., p$ and σ^2 are the parameters in the normality assumption for the linear model

$$\boldsymbol{y} \sim N_n(\alpha\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n),$$

τ_j is the local shrinkage parameter for the j^{th} variable and λ is the global shrinkage parameter. α and σ^2 have the same priors here as in Section 3.1.2 such that shrinkage is not applied to these nor are the priors informative. From the likelihood and the scale mixture of Normals prior for $\boldsymbol{\beta}$,

$$\begin{aligned}\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\tau},\lambda,\sigma^2) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)-\frac{1}{2\sigma^2\lambda^2}\boldsymbol{\beta}^T\boldsymbol{\tau}^{-1}\boldsymbol{\beta}\right\}\\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^T\left(\boldsymbol{X}^T\boldsymbol{X}+\lambda^{-2}\boldsymbol{\tau}^{-1}\right)\boldsymbol{\beta}-2\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{y}-\alpha\mathbf{1}_n)\right)\right\}\end{aligned}$$

therefore $\boldsymbol{\beta}\sim N(\boldsymbol{\mu'},\boldsymbol{\Sigma'})$, where $\boldsymbol{\Sigma'}=(\boldsymbol{X}^T\boldsymbol{X}-\lambda^{-2}\boldsymbol{\tau}^{-1})^{-1}$ and $\boldsymbol{\mu'}=\boldsymbol{\Sigma'}\boldsymbol{X}^T(\boldsymbol{y}-\alpha\mathbf{1}_n)$. Similarly, for α

$$\begin{aligned}\pi(\alpha|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)\right\}\\ &\propto \exp\left\{-\frac{n}{2\sigma^2}\left(\alpha^2-\frac{2\alpha}{n}\sum_{i=1}^{n}y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right\}\end{aligned}$$

therefore $\alpha\sim N\left(\sigma^2/n,\frac{1}{n}\sum_{i=1}^{n}y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$, and for σ^2,

$$\begin{aligned}&\pi(\sigma^2|\boldsymbol{y},\boldsymbol{X},\alpha,\boldsymbol{\beta},\boldsymbol{\tau},\lambda)\\ &\quad\propto\left(\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)\right\}\\ &\qquad\times\left(\sigma^2\right)^{-p/2}\exp\left\{-\frac{1}{2\sigma^2\lambda^2}\boldsymbol{\beta}^T\boldsymbol{\tau}^{-1}\boldsymbol{\beta}\right\}\times\left(\sigma^2\right)^{-1}\\ &\quad\propto\left(\sigma^2\right)^{-(n+p)/2-1}\exp\left\{\frac{-1}{\sigma^2}\left(\frac{\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)}{2}+\frac{\boldsymbol{\beta}^T\boldsymbol{\tau}^{-1}\boldsymbol{\beta}}{2\lambda^2}\right)\right\},\end{aligned}$$

therefore σ^2 is inverse Gamma with shape parameter $(n+p)/2$ and scale parameter

$$\frac{1}{2}\left[\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)+\boldsymbol{\beta}^T\boldsymbol{\tau}^{-1}\boldsymbol{\beta}/\lambda^2\right].$$

The full conditionals for α, $\boldsymbol{\beta}$ and σ^2 hold for any shrinkage priors which can be written as a scale mixture of Normals.

From the priors for the coefficients and their respective local shrinkage parameters,

$$\begin{aligned}\pi(\tau_j^2|\boldsymbol{y},\boldsymbol{X},\beta_j^2,\lambda,\sigma^2) &\propto \left(\tau_j^2\right)^{-1/2}\exp\left\{-\frac{\beta_j^2}{2\sigma^2\lambda^2\tau_j^2}-\tau_j^2\right\}\\ &\propto \left(\tau_j^{-2}\right)^{1/2}\exp\left\{-\frac{\tau_j^2\beta_j^2}{2\lambda^2\sigma^2}\left[\frac{1}{\tau_j^4}+\frac{2\lambda^2\sigma^2}{\beta_j^2}\right]\right\},\end{aligned}$$

where after making a change of variables, $\mu' = \sqrt{\frac{2\lambda^2\sigma^2}{\beta_j^2}}$ and $\lambda' = 2$,

$$\pi(\left(\tau_j^2\right)^{-1}|\boldsymbol{y},\boldsymbol{X},\beta_j^2,\lambda,\sigma^2) \propto \frac{1}{\sqrt{2\pi\sigma^2}}\left(\tau_j^{-2}\right)^{-3/2}\exp\left\{-\frac{\lambda'}{2\left(\mu'\right)^2\tau_j^{-2}}\left(\tau_j^{-2}-\mu'\right)^2\right\}$$

shows $1/\tau_j^2$ has an inverse-Gaussian distribution with mean μ' and variance λ'. The full conditional for $1/\tau_j^2$ changes depending on the prior for the local shrinkage parameter.

3.3.2 Choice for global shrinkage parameter

So far the global shrinkage parameter λ has been assumed constant. The smaller λ is, the larger the level of shrinkage. However it is not clear what a suitable global shrinkage parameter should be to balance the trade-off between shrinking the error signals to zero but keeping the strong signals.

Fixed λ

In the first of our cases, we assume λ is fixed to control how much shrinkage to perform. However, Piironen and Vehtari (2017b) find there can be computational difficulties associated with using a point estimate for λ. For λ close to zero, the Gibbs sampler can not sample from the conditional posterior for $\boldsymbol{\beta}$ due to the collapse of the variance $(\boldsymbol{X}^T\boldsymbol{X}-\lambda^{-2}\boldsymbol{\tau}^{-1})^{-1}$ to zero. One possible solution is to change the MCMC method used such that we can approximately sample from the posterior distribution using an acceptance step, such as Hamiltonian Monte Carlo (see Section 3.3.3). Alternatively, a prior can be used for λ to continue using a Gibbs sampling scheme, which is now

considered.

Prior on λ

Gelman (2006) proposes the half-Cauchy distribution for the global shrinkage parameter due to its non-zero density for $\lambda \to 0$ (Figure 3.2) compared to the inverse-gamma conjugate prior for λ^2 (Polson and Scott (2011)). Given that the dependency on the error variance σ^2 has been moved further up the prior hierarchy (see Equation (3.4)), the prior for the global shrinkage parameter is

$$\lambda \sim C^+(0, 1)$$

where C^+ denotes the half-Cauchy distribution henceforth.

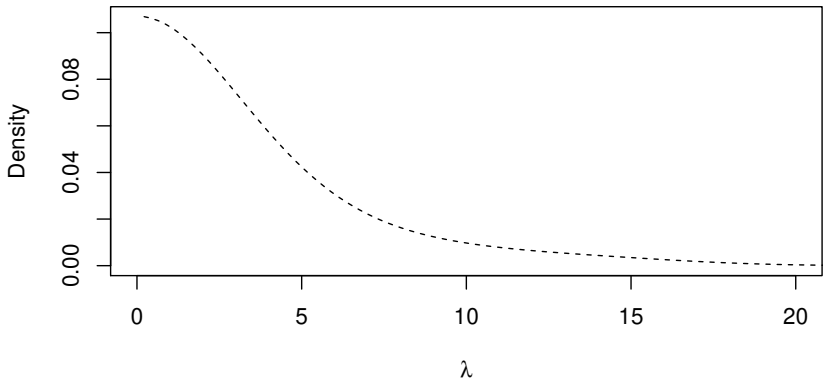


Figure 3.2: Plot of $C^+(0, 1)$ for $\lambda \in (0, 20)$.

To use the Gibbs sampling scheme with this prior for λ, Makalic and Schmidt (2015) makes use of the fact that the half-Cauchy distribution can be expressed as a scale mixture of inverse-gamma distributions (see Wand et al. (2011)). With

$$X^2|a \sim IG(1/2, 1/a) \qquad \text{and} \qquad a \sim IG(1/2, 1/A^2),$$

$$
\begin{aligned}
\pi(x|A) &= \left|\frac{\mathrm{d}(x^2)}{\mathrm{d}x}\right| \int_{a>0} \pi(x^2|a)\pi(a|A)\mathrm{d}a \\
&= 2x \int_{a>0} \frac{(\frac{1}{a})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}(x^2)^{-\frac{1}{2}-1}\exp\left\{-\frac{1}{ax^2}\right\} \frac{(\frac{1}{A^2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}a^{-\frac{1}{2}-1}\exp\left\{-\frac{1}{A^2 a}\right\}\mathrm{d}a \\
&\propto \frac{1}{Ax^2}\int_{a>0}\frac{1}{a^2}\exp\left\{-\frac{1}{a}\left(\frac{1}{x^2}+\frac{1}{A^2}\right)\right\}\mathrm{d}a \\
&= \frac{1}{A\left(1+\frac{x^2}{A^2}\right)},
\end{aligned}
$$

then $X \sim C^+(0, A)$. In our case, a new auxiliary hyperparameter γ is introduced into the prior hierarchy such that

$$
\lambda^2|\gamma \sim IG(1/2, 1/\gamma) \quad \text{and} \quad \gamma \sim IG(1/2, 1),
$$

and added to the Gibbs sampling scheme in Section 3.3.1 with the full conditionals

$$
\begin{aligned}
\pi(\lambda^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \gamma, \sigma^2) &\propto \left(\lambda^2\right)^{-p/2} \exp\left\{-\frac{1}{2\sigma^2\lambda^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2}\right\}\left(\lambda^2\right)^{-1/2-1}\exp\left\{-\frac{1}{\lambda^2\gamma}\right\} \\
&\propto \left(\lambda^2\right)^{-(p+1)/2-1}\exp\left\{-\frac{1}{\lambda^2}\left(\frac{1}{\gamma}+\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2}\right)\right\}
\end{aligned}
$$

i.e. $\lambda^2 \sim IG\left(\frac{p+1}{2}, \frac{1}{\gamma}+\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2}\right)$ and $\gamma \sim (1, \frac{1}{\lambda^2}+1)$. The Gibbs sampling scheme discussed can be implemented using the *bayesreg* package in R (see Schmidt and Makalic (2021) for documentation).

3.3.3 Gradient vector for Hamiltonian Monte Carlo using the Bayesian Lasso

Section 3.3.2 showed for λ fixed close to zero, the Gibbs sampler can not sample from the conditional posterior for $\boldsymbol{\beta}$ due to the variance term collapsing to zero. An alternative is to use Hamiltonian Monte Carlo with a small step size for efficient sampling (Betancourt and Girolami (2013)). Section 3.1.3 showed the posterior density is used in the Hamilton

Monte Carlo algorithm through the gradient of the log posterior densities for $\boldsymbol{\theta}$,

$$\nabla \log \pi(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1}, ..., \frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_p} \right),$$

in the governing Hamiltonian equations (Neal (2011)), where for the case of the linear model, $\boldsymbol{\theta} = (\alpha, \beta_1, ..., \beta_p, \tau_1^2, ..., \tau_p^2, \sigma^2)$. Given the log posterior distribution

$$\begin{aligned} \log \pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) \propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 - \frac{p}{2}\log(\sigma^2) - \frac{p}{2}\log(\lambda^2) \\ - \frac{1}{2}\sum_{j=1}^{p}\log(\tau_j^2) - \frac{1}{2\sigma^2\lambda^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2} - \sum_{j=1}^{p}\tau_j^2 - \sigma^2, \end{aligned}$$

the partial derivatives with respect to α, β_j for $j = 1, .., p$ and σ^2, namely

$$\frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \boldsymbol{x}_i^T\boldsymbol{\beta}),$$

$$\frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})}{\partial \beta_j} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\Big(x_{ij}^2\beta_j + x_{ij}(y_i - \alpha)\Big) - \frac{1}{\sigma^2\lambda^2}\frac{\beta_j}{\tau_j^2},$$

$$\frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \alpha - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 - \frac{p}{2\sigma^2} + \frac{1}{2\sigma^4\lambda^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2} - 1$$

remain the same regardless of the priors for the local shrinkage parameters τ_j, $j = 1, ..., p$. For the Bayesian Lasso,

$$\frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})}{\partial \tau_j^2} = -\frac{1}{2\tau_j^2} + \frac{1}{2\sigma^2\lambda^2}\frac{\beta_j^2}{\tau_j^4} - 1.$$

These partial derivatives are then plugged into the Hamiltonian Monte Carlo algorithm in Section 3.1.3. The computer language Stan through the R interface (see Guo et al. (2020)) can be used to generate samples using the HMC once the prior hierarchy has been specified.

3.3.4 Application: Bayesian Lasso with fixed λ

Section 2.6.2 used the frequentist Lasso to produce coefficient estimates by manually selecting a value of λ such that only a small number of variables are selected and using linear regression whilst ignoring the uncertainty associated with the variable selection stage. Often overlooked by agricultural studies, our goal is to first produce credible intervals for our parameters estimates to find the agronomic, socio-ecological and climatic variables from Chapter 1 which are most associated with wheat yields, regardless of the other variables included in the model. These will not only capture the uncertainty from the parameter estimation, but they will also capture the uncertainty from the model selection algorithm. Finding credible intervals which capture both stages of uncertainty will also enable predictions to also capture both stages of uncertainty, and therefore will not lead us to be overconfident in our predictions.

Reducing the model to include organic status, crop protection costs and rainfall in June, the first 3 covariates to be selected with the frequentist Lasso, their linear regression coefficients are -2.8298, 0.0063 and -0.0118 respectively. Here we use the Bayesian analogue to only determine 3 covariates as being important and construct credible intervals about their coefficients. Note the global shrinkage parameter λ in the frequentist Lasso is proportional to the reciprocal of the global shrinkage parameter taken here (see Section 3.3). To only select 3 variables as being important, the global shrinkage parameter here is set to $\lambda = 0.01$. Section 3.3.2 discussed the computational difficulties using a Gibbs sampler for small values of λ. Instead, a Hamiltonian Monte Carlo scheme is implemented with a small step size to avoid divergent sampling.

The HMC sampling scheme is repeated 11,000 times for the first chain, where the first 1,000 samples are used as a burn-in period, and of the remaining 10,000 samples, every 5^{th} sample is used to approximate the posterior distribution. This is repeated 4 times to ensure the chains are well mixed. Diagnostics of chain convergence can be seen in Appendix B.

Table 3.2 shows the effect the small global shrinkage parameter λ has on the credible intervals. The lower and upper bounds are the 2.5% and 97.5% sample quantiles of the

samples approximately coming from the posterior distribution. The credible intervals constructed from the Bayesian Lasso for the coefficients associated with organic status, crop protection costs and rainfall in June are closer to zero in comparison to the linear regression estimates. To give an indication of how far away from zero the credible intervals lie and the variable importance, the mean of the posterior samples for each β_j are divided by their respective standard deviations, which we refer to as the importance statistic for each variable. The stars represent whether the 90%, 95% and 99% credible interval contains zero with 1, 2 or 3 stars respectively. The stars will also indicate how far away from zero the credible interval lies and whether the coefficient for β_j should be non-zero for $j = 1, ..., p$.

Reassuringly, the 3 variables important at the 95% level with fixed $\lambda = 0.01$ using the Bayesian Lasso are organic status, crop protection and rainfall in June. The same justifications as Section 2.6.2 hold here as to why these variables are selected. Organic status and crop protection being selected highlights the reduction in yield when using more organic farming practices, whilst rainfall in June negatively impacting yield is indicative of too much rain and consequently a lack of sun as the crop approaches maturity.

The Bayesian Lasso not only determines which variables are important, but also allows for credible intervals to be constructed using the MCMC samples. The widest 95% credible interval of the coefficient parameters is for organic status. This fluctuation of the coefficient parameter may be indicative of the knowledge and resources a farmer needs in order to run a high-yielding organic farm compared to conventional farming methods.

To assess the performance of the model, we look at the posterior predictive distribution for each farm record in 2009 and find whether each credible interval for yield captures the true observed yield. Section 3.1.4 discusses how to sample from the posterior predictive distribution given the posterior samples for each of the parameters α, $\boldsymbol{\beta}$ and σ^2 in the linear model. Using 2000 samples from their respective posterior predictive distributions, Figure 3.5 shows the 95% credible intervals predicted for every yield in 2009 given their respective set of covariates on farming practices and environmental conditions.

There is little difference between the posterior predictive distributions for each new yield

in 2009 above 7 tonnes per hectare, with respect to the subset of significant variables found here. The credible intervals for the significant variables are $(-1.9668, -1.4533)$ for `organic`, $(0.0026, 0.0042)$ for `sprays` and $(-0.0055, -0.0030)$ for `rainfall June` and the approximate posterior predictive distributions for the yields per hectare in 2009 are in Figure 3.5. The approximate posterior prediction distributions show our model suitably captures larger yields but often over estimates smaller yields.

Since our approximate posterior predictive distributions are sampled from the likelihood

$$\pi(y|\boldsymbol{x} = \tilde{\boldsymbol{x}}) \sim N(\alpha + \tilde{\boldsymbol{x}}^T\boldsymbol{\beta}, \sigma^2),$$

shrinking and selecting the coefficients such that only 3 variables are deemed important means less variation in the response is captured by the term $\tilde{\boldsymbol{x}}^T\boldsymbol{\beta}$ and instead captured through the variance term σ^2. Therefore, our approximate confidence intervals for the new value of $\tilde{y}$ will be much wider than had we not applied as much shrinkage. Figure 3.3 shows the approximate posterior distribution of the error variance with approximate maximum a posteriori estimate at 2.88. We shall see how this compares when using a prior for λ in the next section.

Variable	Lower bound	Upper bound	Importance statistic	
Organic	-1.9668	-1.4533	12.9533	***
Sprays	0.0026	0.0042	8.3269	***
Rainfall June	-0.0055	-0.0030	6.5378	***
Rainfall April	-0.0016	$3.3463.10^{-5}$	1..5032	
Sunshine June	$-3.5390.10^{-5}$	0.0019	1.4191	
Rainfall December	-0.0013	$3.2504.10^{-5}$	1.3338	
Sunshine December	-0.0002	0.0028	1.0830	
Fert	$-3.4951.10^{-5}$	0.0003	0.9740	
Sunshine February	-0.0001	0.0006	0.8010	
UAA	$-9.2483.10^{-6}$	$4.3444.10^{-5}$	0.7611	
Sunshine April	-0.0001	0.0003	0.7598	
Rainfall July	-0.0003	0.0001	0.7550	
Mean temp November	-0.0081	0.0018	0.7485	
Labour	$-2.0324.10^{-5}$	0.0001	0.7288	
Machinery	$-1.3044.10^{-5}$	$4.8100.10^{-5}$	0.6703	
OtherVC	-0.0001	0.0002	0.6154	
LAND	$-2.7467.10^{-5}$	0.0001	0.5988	
Mean temp April	-0.0018	0.0049	0.5890	
Rainfall September	-0.0003	0.0001	0.5640	
Seeds	-0.0003	0.0001	0.5529	
Mean temp December	-0.0037	0.0016	0.5447	
Sunshine August	-0.0002	0.0001	0.4783	
Sunshine October	-0.0003	0.0002	0.4355	
Mean temp August	-0.0074	0.0031	0.4136	
Sunshine January	-0.0003	0.0006	0.3558	
Mean temp February	-0.0027	0.0043	0.3544	
Education 4	-0.0102	0.0150	0.3191	

Education 1	-0.0182	0.0121	0.3104
Sunshine March	-0.0002	0.0001	0.2924
Rainfall August	-0.0001	0.0002	0.2681
Mean temp January	-0.0034	0.0022	0.2670
Education 3	-0.0082	0.0109	0.2662
Fuel	-0.0007	0.0011	0.2633
Rainfall November	-0.0001	0.0001	0.2421
Sunshine November	-0.0003	0.0003	0.2413
Sunshine September	-0.0002	0.0002	0.1810
Mean temp June	-0.0044	0.0056	0.1745
Rainfall January	-0.0001	0.0001	0.1728
Sunshine May	-0.0002	0.0002	0.1371
Rainfall October	-0.0001	0.0001	0.1183
Mean temp May	-0.0038	0.0043	0.1155
Mean temp July	-0.0026	0.0032	0.1154
Contract	$-3.1623.10^{-5}$	$4.0269.10^{-5}$	0.0913
Mean temp March	-0.0030	0.0025	0.0875
Mean temp September	-0.0033	0.0034	0.0772
Education 2	-0.0200	0.0191	0.0588
Rainfall March	-0.0002	0.0002	0.0560
Mean temp October	-0.0031	0.0033	0.0401
Education 5	-0.0225	0.0209	0.0393
TOFC	-0.0001	0.0001	0.0372
Rainfall February	-0.0001	0.0001	0.0316
Rainfall May	-0.0001	0.0002	0.0187
Sunshine July	-0.0001	0.0001	0.0093
Variable	**Lower bound**	**Upper bound**	**Importance statistic**

Table 3.2: Bayesian Lasso with fixed $\lambda = 0.01$. Variables have been ranked according to their importance statistic, i.e. mean of the posterior samples divided by their respective standard deviation. Lower and upper bounds are the 95% credible intervals. Stars indicate whether the 90% (*), 95% (**) and 99% (***) credible intervals contain zero.

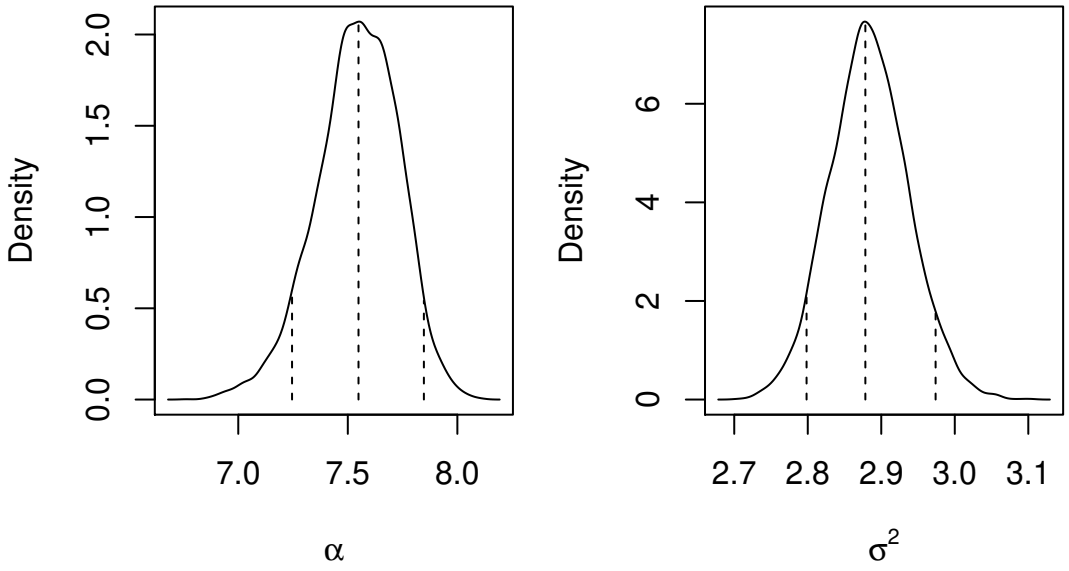


Figure 3.3: Approximations to the posterior densities for α and σ^2 using the Bayesian Lasso for fixed $\lambda = 0.01$. Vertical lines represent the 2.5% quantile, maximum a posteriori estimate and the 97.5% quantile.

3.3.5 Application: Bayesian Lasso with prior λ

Section 3.3.1 details the Gibbs sampling scheme in addition to Section 3.3.2 when using a prior for λ. Again, 4 chains are used, all with a burn-in period of 1,000 samples, thinning of every 5^{th} sample, resulting in a sample of size $8,000$ to approximate the posterior distribution.

Table 3.3 shows a larger number of variables are now considered important when varying the level of regularisation. The three variables from the fixed λ case, `organic`, `sprays` and `rainfall June` are still deemed important here.

With 95% credible intervals far away from zero, mean temperature in September and October positively influence yield whilst mean temperature in February and July negatively influence yield. Again, this may be an indicator of a high mean temperature being a good environment for growing wheat during the initial growth phase, but in the latter stages, a high mean temperature can bring the wheat to maturity quicker and dry the crop out before harvest.

For those parameters with 95% credible intervals close to but not including zero, `machinery`,

`contract` and `labour` may indicate better equipment and more experienced labourers lead to larger yields. Utilised agricultural area positively influencing yield may indicate larger farms often attain larger yields due to efficient farming practices, and fertiliser to provide crops with additional nutrients required for growth. There also appears to be a positive relationship between education to college level or equivalent, which may be indicative of the farmer's knowledge in growing crops.

Of the important Farm Business Survey variables, `seeds` is the only one with a 95% credible interval whose bounds are negative. This may be due to farmers applying excessive amounts of seeds to cover any crops lost during the year without due care for the other factors influencing wheat yields. All other weather conditions deemed important could again be indicative of desirable environmental conditions for the wheat production cycle at each stage.

Comparing the list of important variables found here in Table 3.3 to those found in Section 2.9, the frequentist Lasso is restricted to the variables selected according to the stepwise algorithm as the level of shrinkage decreases and suggests a variable is important to model yield if it captures the largest proportion with the response, conditional on the preceding variables already being included in the model. Instead, and more preferable, the Bayesian Lasso indicates variable importance regardless of the other variables included in the model by varying the amount of shrinkage applied to each individual coefficient in an MCMC sampling scheme and examines the credible intervals for each coefficient. Additional weather variables were not considered important in the frequentist Lasso yet are indicated as important when using the Bayesian Lasso. This could be due to other variables already included in the stepwise model already capturing these variables' relationships with the response, yet the Bayesian Lasso picks out these variables' relationships with the response by considering a model without these variables who have a stronger relationship with the response.

Using a prior for λ allows the Gibbs sampler to sample with a smaller level of shrinkage. Figure 3.4 shows the approximate posterior distribution for λ, with very few samples being close to zero. According to the 95% credible interval, indicated by two stars in Table 3.3, 27 are now deemed important, compared to 21 when considering 99%

credible intervals to indicate importance. Without manually controlling for λ ourselves, the number of significant variables may now be too many. Section 3.5 will look at an alternative shrinkage prior which allows for λ to be selected manually but will not overshrink the coefficients.

If we were to use model selection to only select those important parameters whose 99% credible intervals contain zero, the number of important variables would be smaller, however the credible intervals would be wider. Assuming the Gibbs sampler is run for long enough such that the posterior samples lying outside of the 99% credible intervals are sufficiently large, then using a prior for λ would avoid shrinking the active coefficients to zero, as done in the fixed λ case, but widens our credible intervals to show our uncertainty when reducing the number of important variables in the model.

The maximum a posteriori estimate for α still approximately equals 8 tonnes per hectare, however the width of the 95% credible interval for α has now increased significantly. This might be seen as a consequence of the variables for the farming practices and environmental conditions having a larger role in modelling the yields, and α responding accordingly rather than solely modelling the mean. The maximum a posteriori estimate for σ^2 is much smaller than the fixed λ case, yet the posterior variance of σ^2 is approximately the same. Again, this could be a consequence of capturing some of the strong signals that was previously taken as error, yet the spread of the variance remaining approximately the same since the variables do not play a role in the variance of the yields according to our linear model assumption in Section 3.1.2, $\boldsymbol{y} \sim N_n(\alpha \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$.

For yields in 2009, the posterior predictive distribution can be found with respect to their covariates, where Figure 3.5 gives the 95% credible intervals. The posterior predictive distributions for many large yields underestimates the yield observed. Possible reasons include the linear model assumption which is not suited for modelling high yields. The significant covariates may also not capture the increase in yield for these farms.

Variable	Lower bound	Upper bound	Importance statistic	
Organic	-2.3748	-1.9021	17.7468	***
Sprays	0.0057	0.0072	17.2368	***
Machinery	0.0012	0.0016	12.7231	***
Sunshine February	0.0356	0.0507	11.3061	***
Fert	0.0028	0.0042	10.0739	***
UAA	0.0005	0.0008	9.1795	***
Contract	0.0011	0.0017	8.4204	***
LAND	0.0007	0.0013	6.4998	***
Mean temp October	0.3552	0.6817	6.1460	***
Seeds	-0.0057	-0.0028	5.6851	***
Other VC	0.0016	0.0034	5.4502	***
Rainfall November	-0.0096	-0.0044	5.2291	***
Rainfall June	-0.0096	-0.0043	5.1177	***
Rainfall April	-0.0159	-0.0066	4.7209	***
Rainfall July	-0.0102	-0.0040	4.4955	***
Sunshine September	-0.0154	-0.0053	4.0594	***
Mean temp February	-0.4421	-0.1498	3.9948	***
Mean temp September	0.1789	0.6218	3.5792	***
Sunshine April	0.0037	0.0130	3.5304	***
Mean temp July	-0.6711	-0.1827	3.5283	***
Labour	0.0002	0.0008	3.4794	***
Mean temp May	-0.4208	-0.0609	2.6409	***
Mean temp November	-0.4582	-0.0674	2.5765	***
Education 3	0.0210	0.1961	2.4988	**
Rainfall May	-0.0071	-0.0003	2.1035	**
Sunshine November	-0.0128	-0.0003	2.0609	**
Sunshine March	0.0003	0.0106	2.0356	**

Education 4	-0.0152	0.1985	1.7371 *
Mean temp January	-0.2868	0.0177	1.6827
Rainfall September	-0.0009	0.0057	1.4086
Mean temp August	-0.0458	0.2778	1.4041
Mean temp December	-0.2590	0.0404	1.3509
Sunshine June	-0.0055	0.0008	1.3240
Rainfall August	-0.0010	0.0061	1.2591
Rainfall December	-0.0011	0.0048	1.2195
Education 1	-0.2020	0.0509	1.1335
Fuel	-0.0024	0.0102	1.0875
TOFC	-0.0011	0.0003	1.0604
Sunshine May	-0.0060	0.0020	0.9451
Rainfall March	-0.0021	0.0057	0.8621
Mean temp June	-0.3214	0.1126	0.8051
Rainfall February	-0.0015	0.0041	0.7853
Sunshine August	-0.0052	0.0027	0.6375
Sunshine July	-0.0055	0.0026	0.6117
Rainfall October	-0.0042	0.0023	0.5113
Education 2	-0.1196	0.1941	0.4560
Rainfall January	-0.0023	0.0035	0.3430
Education 5	-0.1437	0.2045	0.3286
Mean temp March	-0.1303	0.0915	0.3100
Sunshine October	-0.0064	0.0050	0.2329
Sunshine January	-0.0087	0.0099	0.1887
Sunshine December	-0.0058	0.0067	0.1157
Mean temp April	-0.1621	0.1697	0.0817
Variable	**Lower bound**	**Upper bound**	**Importance statistic**

Table 3.3: Bayesian Lasso with half-Cauchy prior on λ. Variables have been ranked according to their importance statistic. Lower and upper bounds are the 95% credible intervals. Stars indicate whether the 90% (*), 95% (**) and 99% (***) credible intervals contain zero.

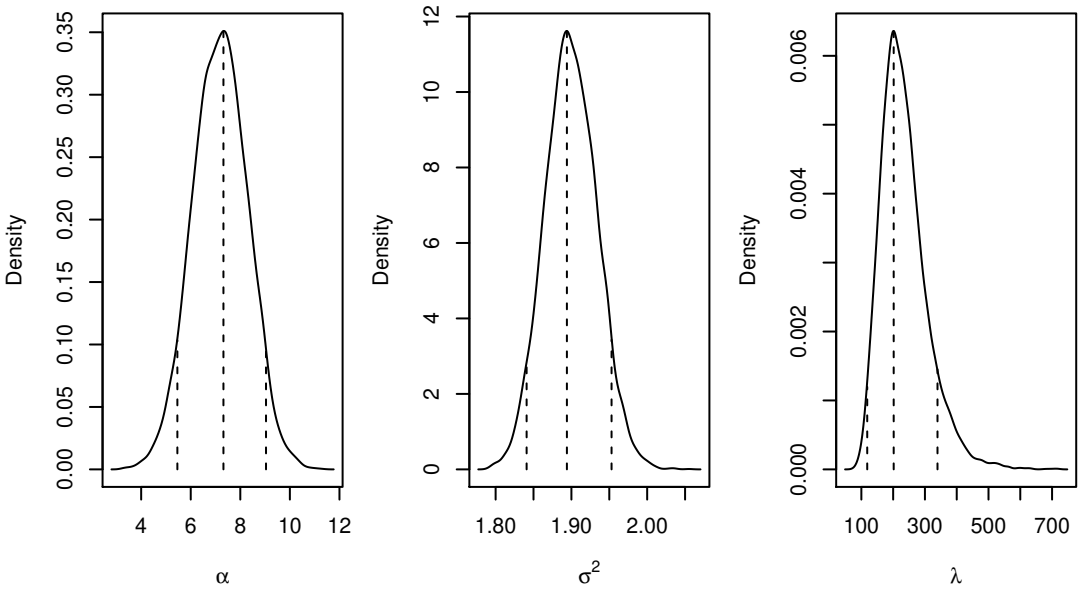


Figure 3.4: Approximations to the posterior densities for α, σ^2 and λ^2 using the Bayesian Lasso with a half-Cauchy prior for λ. Vertical lines represent the 2.5% quantile, maximum a posteriori estimate and the 97.5% quantile.

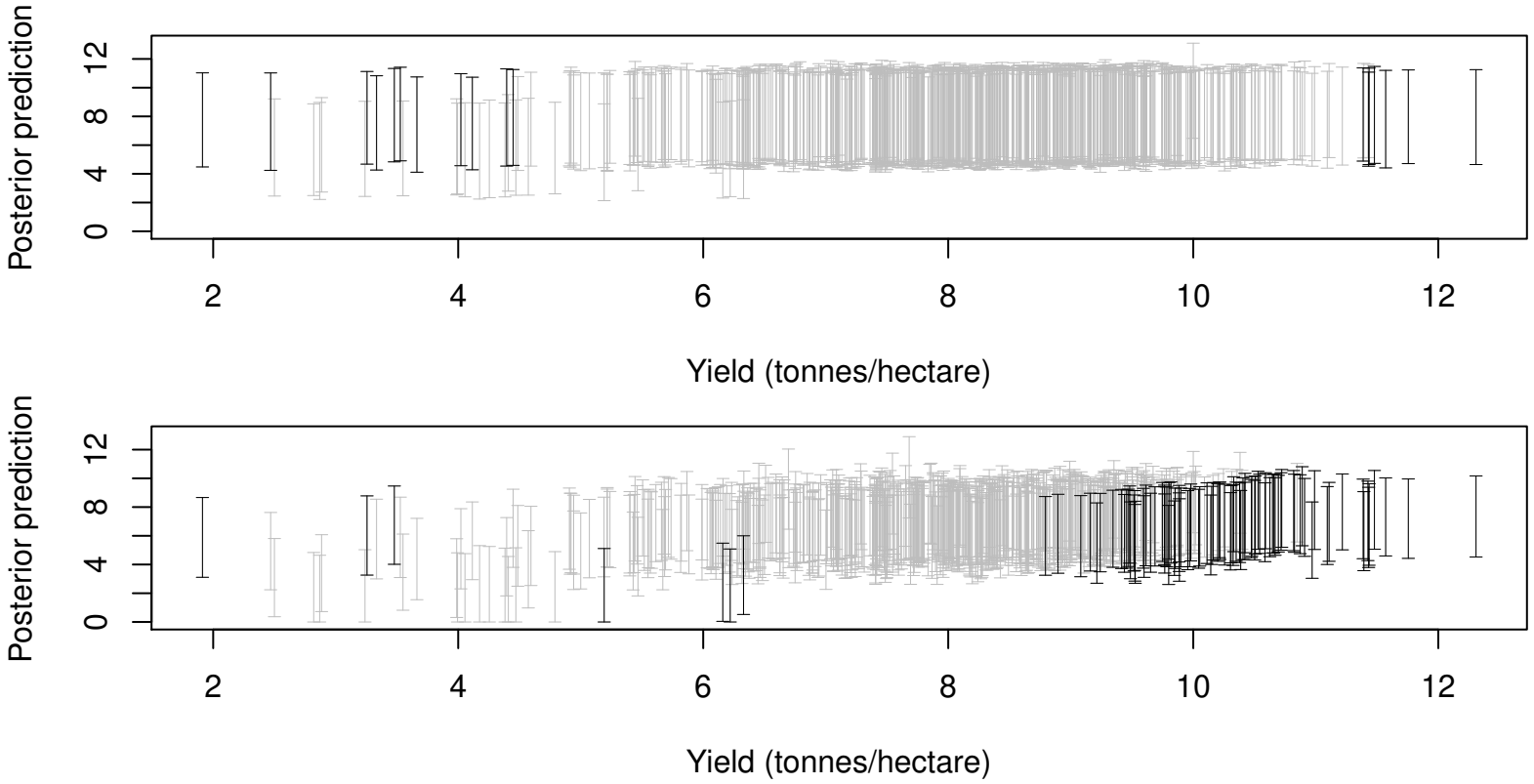


Figure 3.5: 95% credible intervals for each of the posterior predictive distributions of yields in 2009 with respect to their farming practices and environmental conditions, from the posterior samples of the Bayesian Lasso. Top: fixed $\lambda = 0.01$, bottom: prior λ. Credible intervals indicated in black do not contain their observed yields.

3.4 Implied shrinkage coefficient prior

Manually selecting λ in the Bayesian Lasso has shown the extent to which shrinkage can be influential. Provided $\lambda > 0$, the Bayesian Lasso will always apply shrinkage to all coefficients regardless of their associated variable's importance to model the response. Preferably, we would like apply a shrinkage prior that would shrink negligible coefficients to zero but would also allow for the possibility of coefficients being left alone. Shrinkage behaviour can be better understood by looking at the shrinkage profile of the implied shrinkage coefficient prior (see Carvalho et al. (2009), Carvalho and Polson (2010), Piironen and Vehtari (2017b)).

Still assuming $\boldsymbol{X}$ has been standardized and $\sigma^2 = 1$, then

$$\begin{aligned}\beta_j|\boldsymbol{y}, \alpha, \tau_j^2 &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \alpha - x_{ij}\beta_j)^2\right)\exp\left(-\frac{1}{2\lambda^2\tau_j^2}\beta_j^2\right)\\ &\propto \exp\left(-\frac{1+\lambda^2\tau_j^2}{2\lambda^2\tau_j^2}\left[\beta_j^2 - 2\beta_j\frac{\lambda^2\tau_j^2}{1+\lambda^2\tau_j^2}\sum_{i=1}^{n}(y_i-\alpha)x_{ij}\right]\right)\end{aligned}$$

Therefore, $E(\beta_j|\boldsymbol{y}, \alpha, \lambda, \tau_j^2) = (\lambda^2\tau_j^2/(1+\lambda^2\tau_j^2))\sum_{i=1}^{n}(y_i-\alpha)x_{ij}$, where $(\lambda^2\tau_j^2/(1+\lambda^2\tau_j^2))$ is the weight placed on $\boldsymbol{x}_j^T(\boldsymbol{y} - \alpha\mathbf{1}_n)$ and $1 - (\lambda^2\tau_j^2/(1+\lambda^2\tau_j^2)) = 1/(1+\lambda^2\tau_j^2)$ is the weight placed on 0. Letting $\kappa_j = 1/(1+\lambda^2\tau_j^2)$,

$$E(\beta_j|\boldsymbol{y}, \alpha, \tau_j^2) = (1-\kappa_j)\boldsymbol{x}_j^T(\boldsymbol{y} - \alpha\mathbf{1}_n) = (1-\kappa_j)u_j,$$

and since $\kappa_i \in [0, 1]$,

$$E(\beta_j|u_j) = \int_0^1 (1-\kappa_j)u_j\pi(\kappa_j|u_j)\mathrm{d}\kappa_j = [1 - E(\kappa_j|u_j)]\,u_j,$$

a linear function of u_j, and so a linear function of y_i (Carvalho et al. (2009), Bhadra et al. (2019)). This holds for all hierarchical priors where β_j is expressed in the same form as Equation (3.4), a scale mixture of Normals. With $\sigma^2 = 1$ and

$$\pi(\tau_j^2) \propto \exp(-\tau_j^2)$$

in the case of the Bayesian Lasso. With $\kappa_j = 1/(1+\lambda^2\tau_j^2)$, a change of variables yields

$$\pi(\kappa_j) \propto \frac{1}{\lambda^2\kappa_j^2}\exp\left(-\frac{1}{\lambda^2\kappa_j}\right), \quad \kappa_j \in (0,1).$$

Figure 3.6 shows the shrinkage profiles for $p(\kappa_j)$ for various levels of shrinkage. Although all three levels of shrinkage have non-zero probabilities to fully shrink a coefficient to zero, as $\kappa_j \to 0$, $\pi(\kappa_j) \to 0$ hence the Bayesian Lasso will always shrink coefficients regardless of their signal.

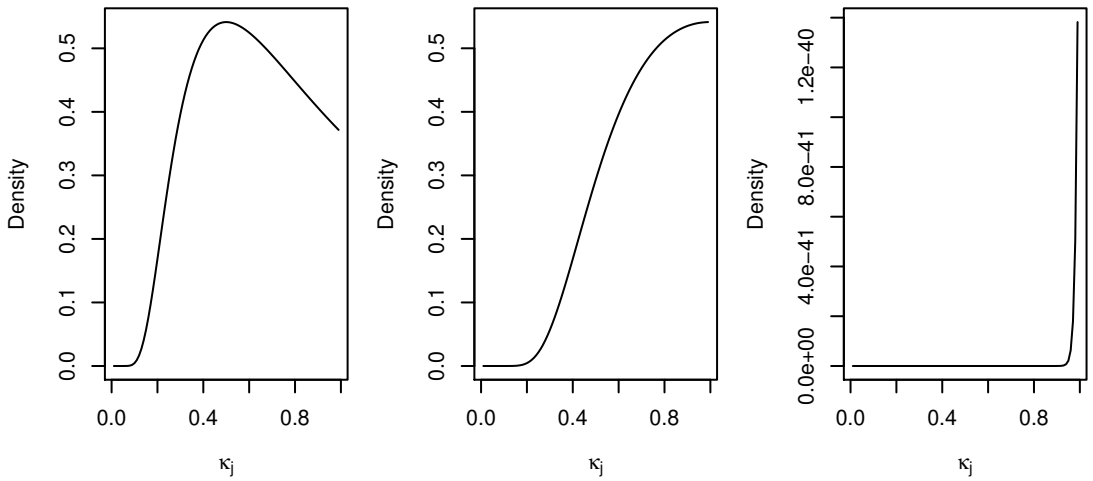


Figure 3.6: Shrinkage profiles when $\tau_j^2 \sim \text{Exp}(1)$ with $\lambda = 1$, $\lambda = 0.5$ and $\lambda = 0.01$.

Ideally, shrinkage profiles should have large probability densities for no shrinkage ($\kappa_j = 0$) and full shrinkage ($\kappa_j = 1$). This would leave the strong signals and shrink the error signals to zero.

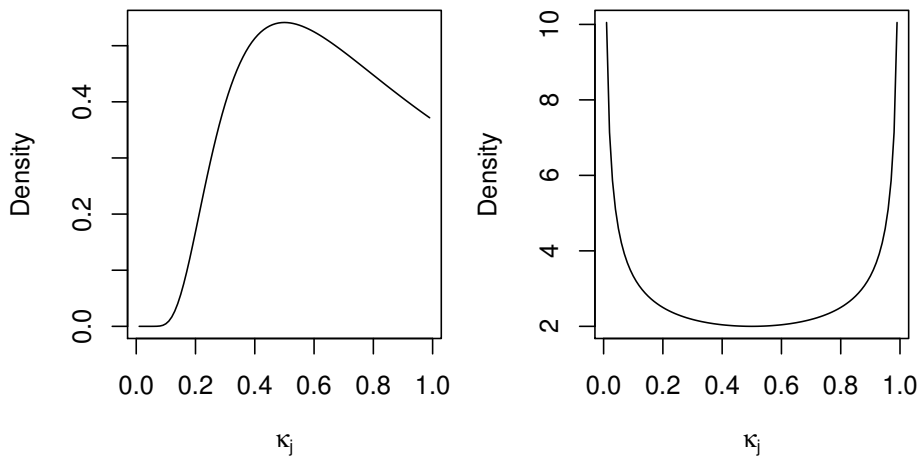


Figure 3.7: Proportional prior densities for the implied shrinkage coefficient κ when τ_j^2 is exponentially distributed with shape parameter 1 (left) and τ_j^2 distributed according to a half-Cauchy with 0 and 1 (right).

One such choice is the Beta$(1/2, 1/2)$ density $\pi(\kappa_j) \propto \kappa_j^{-1/2}(1-\kappa_j)^{-1/2}$ in Figure 3.7. With $\kappa_j = 1/(1+\tau_j^2)$ as before, when $\lambda = 1$,

$$\pi(\tau_j) = (1+\tau_j^2)^{1/2}\left(\frac{\tau_j^2}{1+\tau_j^2}\right)^{-1/2}\tau_j\left(\frac{1}{1+\tau_j^2}\right)^2 = \frac{1}{1+\tau_j^2},$$

i.e. $C^+(0,1)$. The prior hierarchy with $\tau_j \sim C^+(0,1)$ instead of an exponential with rate 1 is aptly named the horseshoe prior hierarchy after the shape of its shrinkage profile.

3.5 Horseshoe

To avoid overshrinking the coefficients $\boldsymbol{\beta}$ towards zero, the horseshoe estimator (Carvalho et al. (2009), Carvalho and Polson (2010)) replaces the exponential prior for the local shrinkage parameters τ_j with a half-Cauchy prior in the prior hierarchy

$$\beta_j|\tau_j^2 \sim N(0,\tau_j^2), \quad \tau_j|\lambda \sim C^+(0,\lambda) \quad \lambda|\sigma \sim C^+(0,\sigma) \quad \sigma^2 \propto \frac{1}{\sigma^2} \tag{3.6}$$

where λ is the global shrinkage parameter. Makalic and Schmidt (2015) showed this

prior hierarchy can be reformulated as

$$\beta_j|\tau_j^2 \sim N(0, \sigma^2\lambda^2\tau_j^2), \quad \tau_j \sim C^+(0,1) \quad \lambda \sim C^+(0,1) \quad \sigma^2 \propto \frac{1}{\sigma^2}, \tag{3.7}$$

and hence the difference between the Bayesian Lasso and the horseshoe prior can be seen through the prior distribution for the local shrinkage parameter τ_j. The prior for τ_j^2 was exponential whereas now the prior for τ_j is half-Cauchy. The half-Cauchy has a larger peak at zero and a heavier tail compared to the exponential, making it more desirable as a shrinkage prior.

Section 3.4 showed the local shrinkage behaviour of the horseshoe prior with $\sigma^2 = \lambda = 1$. Here we now look at the influence of the global shrinkage parameter λ has on the shrinkage profile for the horseshoe. If we use a scale parameter in the prior distribution for the local parameter τ_j^2 then the implied prior distribution for κ_j will be

$$\pi(\kappa_j|\lambda) \propto \kappa_j^{-1/2}(1-\kappa_j)^{-1/2}\left(\frac{\lambda}{(\lambda^2-1)\kappa_j+1}\right).$$

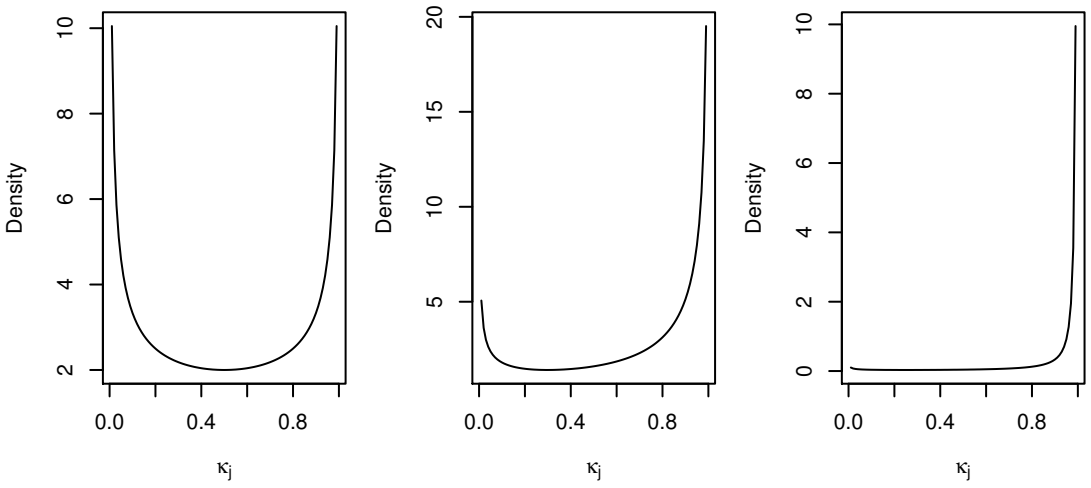


Figure 3.8: Shrinkage profiles when $\tau_j^2 \sim C^+(0,1)$ with $\lambda = 1$, $\lambda = 0.5$ and $\lambda = 0.01$.

Figure 3.8 shows the shrinkage profile for varying levels of λ. For $\lambda = 0.01$, the horseshoe prior always has a non-zero probability at $\kappa_j = 0$, hence always allows for the case of no shrinkage. Piironen and Vehtari (2017a) discuss how to manually select λ to reflect the

effective number of non-zero coefficients to be in the model. Again, an alternative is to use a half-Cauchy prior for λ as done for the Bayesian Lasso.

3.5.1 Conditional posterior distributions for Gibbs

The only difference between the Bayesian Lasso and the horseshoe is the prior on τ_j, therefore the only conditional posterior distribution which needs to change for the Gibbs sampling scheme is for τ_j (Makalic and Schmidt (2016)). Section 3.3.2 showed the centred half-Cauchy with scale 1 can be decomposed into a scale mixture of inverse-gamma distributions

$$\tau_j^2|\delta \sim IG(1/2, 1/\delta) \quad \text{and} \quad \delta \sim IG(1/2, 1),$$

with full conditionals in the Gibbs sampling scheme as

$$\begin{aligned}\tau_j^2|\beta_j, \tau_j, \lambda, \sigma, \delta &\propto \left(\tau_j^2\right)^{-1/2} \exp\left\{-\frac{\beta_j^2}{2\tau_j^2\sigma^2\lambda^2}\right\}\left(\tau_j^2\right)^{-1/2-1} \exp\left\{-\frac{1}{\delta\tau_j^2}\right\} \\ &\propto \left(\tau_j^2\right)^{-1-1} \exp\left\{-\frac{1}{\tau_j^2}\left(\frac{1}{\delta} + \frac{\beta_j^2}{2\sigma^2\lambda^2}\right)\right\},\end{aligned}$$

hence $\tau_j^2 \sim IG\left(1, \frac{1}{\delta} + \frac{\beta_j^2}{2\sigma^2\lambda^2}\right)$ and

$$\begin{aligned}\delta|\tau_j &\propto \delta^{-1/2}\exp\left\{-\frac{1}{\tau_j^2}\right\}(\delta)^{-1/2-1} \exp\left\{-\frac{1}{\delta}\right\} \\ &\propto \delta^{-2}\exp\left\{-\frac{1}{\delta}\left(\frac{1}{\tau_j^2} + 1\right)\right\},\end{aligned}$$

hence $\delta \sim IG\left(1, \frac{1}{\tau_j^2} + 1\right)$. These are added to the other full conditionals in Section 3.3.1 for the full horseshoe Gibbs sampling scheme.

3.5.2 Gradient vector for Hamiltonian Monte Carlo using the horseshoe prior

Since the only difference between the prior hierarchies for the Bayesian Lasso and the horseshoe prior is the prior for the local shrinkage parameters τ_j, the only difference in

the set of partial derivatives in Section 3.3.3 is the gradient for τ_j

$$\frac{\partial \log \pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})}{\partial \tau_j} = -\frac{1}{\tau_j} + \frac{1}{\sigma^2\lambda^2}\frac{\beta_j^2}{\tau_j^3} - \frac{2\tau_j}{1+\tau_j^2}.$$

3.5.3 Application: Horseshoe for fixed λ

Section 3.4 discussed the advantage of using the horseshoe prior as a shrinkage prior due to its desirable shrinkage properties. Sampling with Hamiltonian Monte Carlo with $\lambda = 0.01$, Table 3.4 gives the importance statistics for the agronomic and climatic variables.

Using a fixed λ here does not shrink parameters as much as the Bayesian Lasso due to the non-zero probability at zero for the shrinkage profile in Figure 3.8. This allows for the likelihood to dominate the shrinkage prior if the data is in favour of keeping the variable in the model. According to the 99% credible intervals, the variables said to be important here are the same as those in the Bayesian Lasso case with prior λ (see Section 3.3.5 for justifications).

However, the horseshoe prior allows for full shrinkage to be applied to coefficients of unimportant variables. In this case, only mean temperature in May remains when looking at the 95% credible intervals, whereas before, education to college level or equivalent, mean temperature in November and January, sunshine in November and rainfall in May were also important. Similarities between the Bayesian Lasso with prior for λ and this analysis can also be seen in the approximate posterior distributions for α and σ^2 in Figure 3.9 and the posterior predictive distributions in Figure 3.11.

Although selecting $\lambda = 0.01$ here proves more fruitful in keeping strong signals away from zero compared to the Bayesian Lasso, uncertainty in this arbitrary choice for λ is not taken account of. If there is no justifiable reason to fix λ, a prior should be placed on λ, as we shall do in our final analysis.

Variable	Lower bound bound	Upper bound	Importance statistic	
Organic	-2.3819	-1.9117	17.9447	***
Sprays	0.0058	0.0073	17.2917	***
Sunshine February	0.0374	0.0505	13.0689	***
Machinery	0.0012	0.0016	12.6657	***
Fert	0.0028	0.0042	9.9094	***
UAA	0.0005	0.0008	9.2732	***
Contract	0.0010	0.0017	8.2822	***
Mean temp October	0.3740	0.6743	6.9044	***
LAND	0.0007	0.0013	6.3243	***
Rainfall November	-0.0090	-0.0043	5.5677	***
Seeds	-0.0055	-0.0026	5.4602	***
Rainfall June	-0.0097	-0.0047	5.4494	***
Rainfall April	-0.0152	-0.0064	4.9919	***
Sunshine September	-0.0174	-0.0076	4.9636	***
Other VC	0.0013	0.0032	4.8802	***
Rainfall July	-0.0095	-0.0035	4.1527	***
Mean temp February	-0.4612	-0.1599	3.9792	***
Mean temp July	-0.7178	-0.2412	3.9074	***
Mean temp September	0.2121	0.6840	3.608	***
Sunshine April	0.0026	0.0125	3.0887	***
Labour	0.0001	0.0007	2.8561	***
Mean temp May	-0.4035	0.0046	2.1354	*
Education 3	-0.0046	0.1664	1.7256	
Mean temp November	-0.3663	0.0163	1.6586	
Rainfall May	-0.0059	0.0003	1.5837	
Sunshine November	-0.0111	0.0008	1.4205	
Mean temp January	-0.3057	0.0185	1.3266	

Mean temp December	-0.2815	0.0219	1.1501
Education 1	-0.1878	0.0316	1.0825
Sunshine March	-0.0009	0.0077	1.0330
Education 4	-0.0343	0.1584	1.0256
Rainfall August	-0.0008	0.0048	0.9885
Mean temp June	-0.3810	0.0384	0.9651
Rainfall September	-0.0009	0.0046	0.8787
Sunshine June	-0.0044	0.0008	0.8198
Fuel	-0.0027	0.0079	0.7869
Rainfall February	-0.0010	0.0033	0.6860
Mean temp August	-0.0453	0.2566	0.6582
TOFC	-0.0008	0.0003	0.6491
Rainfall December	-0.0009	0.0033	0.6310
Sunshine August	-0.0048	0.0016	0.6102
Rainfall January	-0.0011	0.0028	0.5471
Mean temp March	-0.0940	0.0448	0.3727
Sunshine December	-0.0057	0.0032	0.3264
Rainfall October	-0.0028	0.0017	0.3156
Sunshine July	-0.0041	0.0024	0.2726
Education 2	-0.0953	0.1392	0.2012
Sunshine January	-0.0071	0.0050	0.1956
Sunshine May	-0.0027	0.0022	0.1913
Sunshine October	-0.0031	0.0043	0.1432
Education 5	-0.1213	0.1417	0.1163
Mean temp April	-0.1015	0.0953	0.0198
Rainfall March	-0.0023	0.0027	0.0086
Variable	**Lower bound**	**Upper bound**	**Importance statistic**

Table 3.4: Horseshoe prior hierarchy with fixed $\lambda = 0.01$. Variables have been ranked according to their importance statistic. Lower and upper bounds are the 95% credible intervals. Stars indicate whether the 90% (*), 95% (**) and 99% (***) credible intervals contain zero.

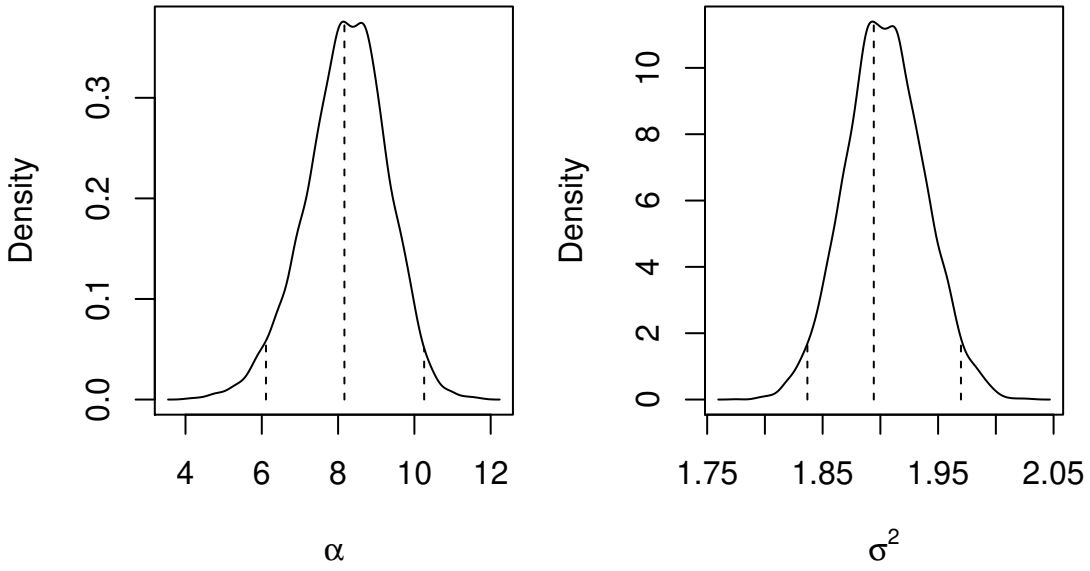


Figure 3.9: Approximations to the posterior densities for α and σ^2 using the horseshoe prior with fixed λ. Vertical lines represent the 2.5% quantile, maximum a posteriori estimate and the 97.5% quantile.

3.5.4 Application: Horseshoe for prior λ

Using a prior for λ instead, the same variables appear to be important when looking at the 99% credible intervals. However, if we do not know for sure what level of regularisation we would like to apply, using a prior for λ allows for other variables which may not indicated as important for the 99% credible intervals to be highlighted for narrower credible intervals. Compared to the fixed λ case in Section 3.5.3, mean temperature in May and November and education to college level or equivalent are also identified when using the 95% credible interval to determine importance. Again, there is little difference in the approximate posterior distributions for α and σ^2. Comparing the approximate posterior distribution for λ here to that in Section 3.3.5 under the Bayesian Lasso framework, the density is concentrated closer to zero leading to more samples with a stricter shrinkage enforced.

Although the horseshoe prior hierarchy is slightly more complex in its exposition compared to the Bayesian Lasso, the horseshoe prior hierarchy performs the inference agricultural studies attempt to quantify but now fully accounts for all stages of uncertainty. Furthermore, using a prior for λ is advisable when there is no prior knowledge for the

shrinkage parameter λ, hence the parameter estimates for λ will not be restricted and will respond to the data in our model.

Variable	Lower bound	Upper bound	Importance statistic	
Organic	-2.3729	-1.9165	18.1077	***
Sprays	0.0058	0.0072	17.1836	***
Machinery	0.0012	0.0016	12.7386	***
Sunshine February	0.0375	0.0523	11.8616	***
Fert	0.0028	0.0042	10.0168	***
UAA	0.0005	0.0008	9.2094	***
Contract	0.0010	0.0017	8.3407	***
Mean temp October	0.3681	0.6898	6.4051	***
LAND	0.0007	0.0013	6.3771	***
Seeds	-0.0056	-0.0027	5.6002	***
Rainfall November	-0.0095	-0.0045	5.4395	***
Rainfall June	-0.0099	-0.0045	5.1947	***
Other VC	0.0015	0.0033	5.1518	***
Rainfall April	-0.0165	-0.0071	4.8692	***
Sunshine September	-0.0169	-0.0068	4.5361	***
Rainfall July	-0.0100	-0.0038	4.4225	***
Mean temp February	-0.4833	-0.1802	4.2055	***
Mean temp July	-0.7765	-0.2797	4.0806	***
Mean temp September	0.2454	0.7190	3.9348	***
Sunshine April	0.0034	0.0128	3.4052	***
Labour	0.0002	0.0008	3.1898	***
Mean temp May	-0.4124	-0.0219	2.3495	**
Education 3	0.0061	0.1852	2.1290	**
Mean temp November	-0.4082	0.0031	2.0580	*
Mean temp January	-0.3295	0.0118	1.8475	
Rainfall May	-0.0061	0.0003	1.7195	
Sunshine November	-0.0118	0.0006	1.7104	
Sunshine March	-0.0006	0.0094	1.5850	

Education 4	-0.0184	0.1884	1.3480
Rainfall September	-0.0008	0.0053	1.2310
Sunshine June	-0.0054	0.0008	1.1681
Rainfall August	-0.0009	0.0054	1.1622
Mean temp August	-0.0462	0.2854	1.1439
Education 1	-0.1953	0.0424	1.1057
Rainfall December	-0.0008	0.0045	1.0251
Mean temp December	-0.2516	0.0430	0.9870
Mean temp June	-0.3482	0.0729	0.9428
Fuel	-0.0029	0.0090	0.8800
Sunshine August	-0.0056	0.0018	0.8560
TOFC	-0.0010	0.0003	0.8173
Rainfall February	-0.0011	0.0038	0.8168
Rainfall January	-0.0016	0.0032	0.5184
Sunshine May	-0.0044	0.0023	0.4874
Mean temp March	-0.1186	0.0645	0.4062
Rainfall October	-0.0037	0.0021	0.3986
Education 2	-0.1158	0.1608	0.3052
Sunshine July	-0.0047	0.0030	0.2526
Education 5	-0.1370	0.1731	0.2029
Rainfall March	-0.0032	0.0036	0.1689
Mean temp April	-0.1266	0.1565	0.0987
Sunshine December	-0.0058	0.0051	0.0952
Sunshine October	-0.0043	0.0051	0.0466
Sunshine January	-0.0079	0.0080	0.0311
Variable	**Lower bound**	**Upper bound**	**Importance statistic**

Table 3.5: Horseshoe prior hierarchy with half-Cauchy prior on λ. Variables have been ranked according to their importance statistic. Lower and upper bounds are the 95% credible intervals. Stars indicate whether the 90% (*), 95% (**) and 99% (***) credible intervals contain zero.

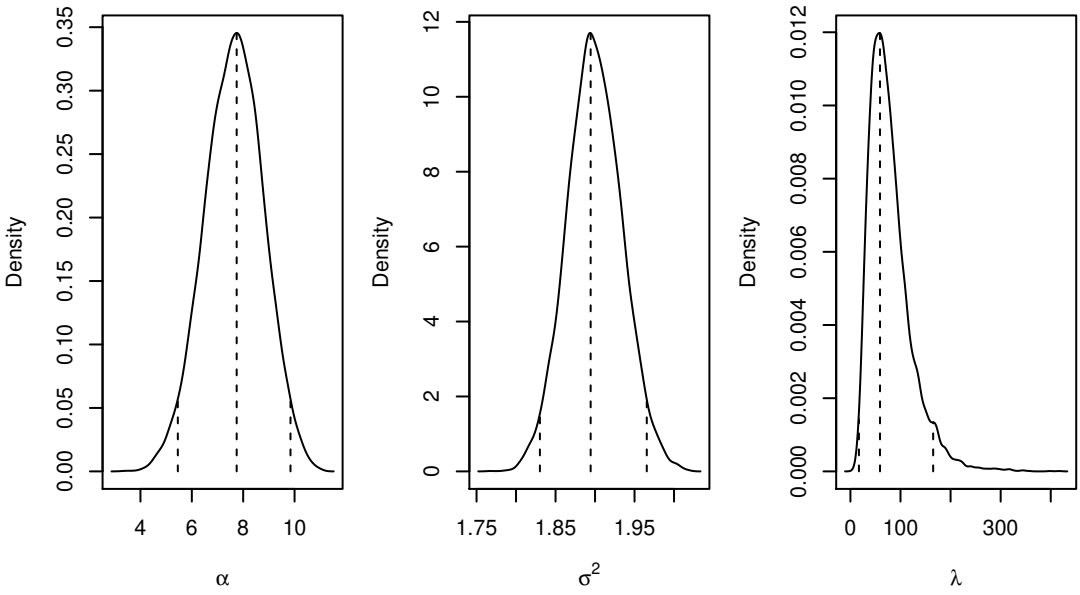


Figure 3.10: Approximations to the posterior densities for α, σ^2 and λ^2. Vertical lines represent the 2.5% quantile, maximum a posteriori estimate and the 97.5% quantile.

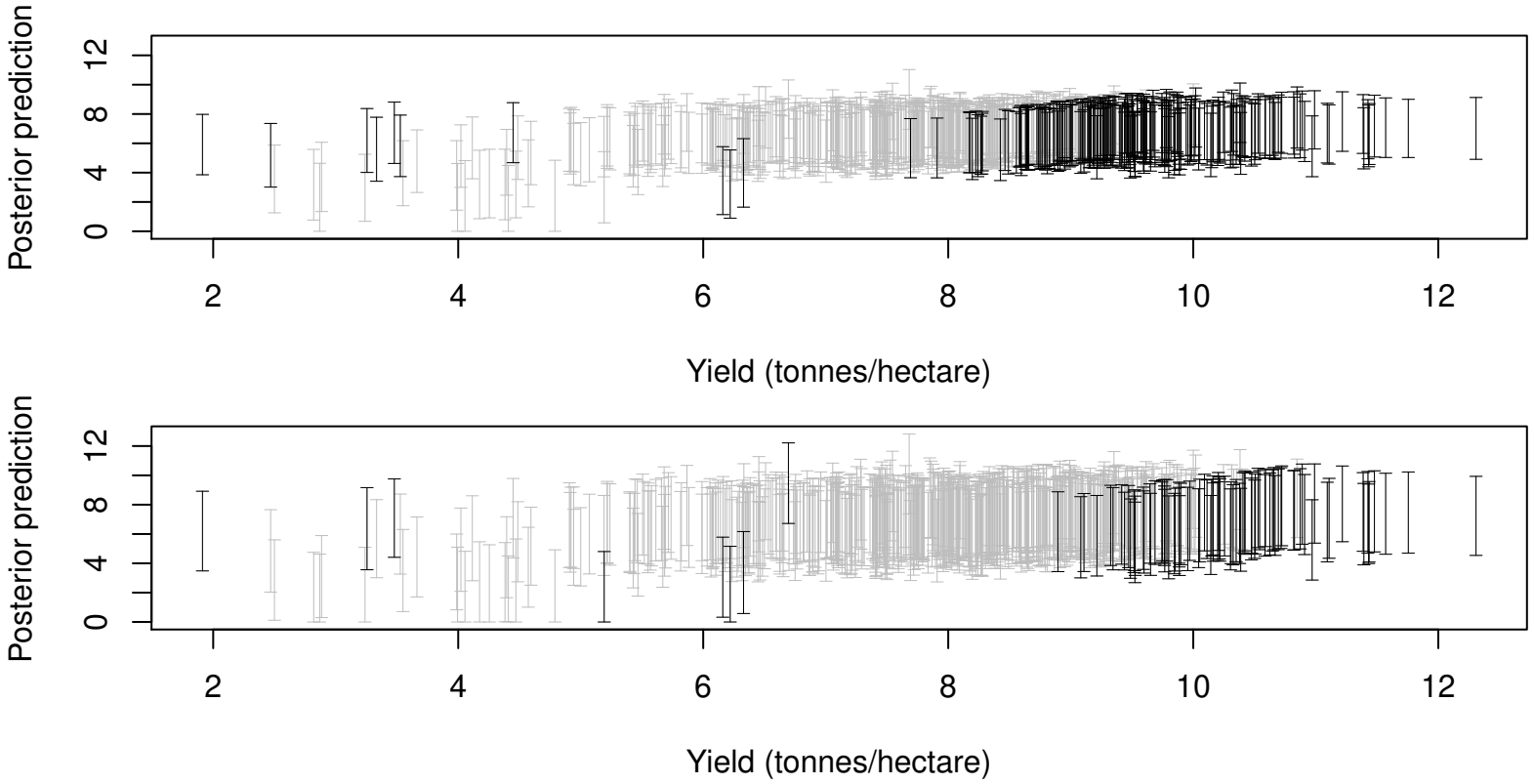


Figure 3.11: 95% credible intervals for each of the posterior predictive distributions of yields in 2009 with respect to their farming practices and environmental conditions, from the posterior samples of the horseshoe prior. Top: fixed $\lambda = 0.01$, bottom: prior λ. Credible intervals indicated in black do not contain their observed yields.

3.6 Conclusion

Our analysis using the Bayesian Lasso and the horseshoe prior hierarchy has found credible intervals for each variable in the Farm Business Survey data set and weather variables from the UK Met Office, assessed their importance once subject to regularisation and provided uncertainty intervals for all variables regardless of their importance. Using the horseshoe prior hierarchy due to its efficiency at removing the error signal and keeping coefficients of significant variables for modelling yields away from zero,

- organic status, crop protection and rainfall in June are indicated as significant variables when predicting wheat yields.
- the credible interval for organic status indicates a strong negative relationship with yields as expected from previous studies.
- other important variables which have credible intervals lying further away from zero are mean temperature in September and October, positively influencing, and February and July, negatively influencing yields as a result of growing towards maturity at a faster rate.

The horseshoe prior hierarchy captures the uncertainty for the parameter estimation stage and the uncertainty for the variable selection stage, which the two-step frequentist approach aims to account for, but fails to do. Accounting for all uncertainty means the credible intervals will be wider than the credible intervals constructed from the uncertainty in the parameter estimates, and not lead to overconfidence in the predictions for wheat yields. Furthermore, using a prior for λ is recommended when there is no prior knowledge on how much shrinkage to apply.

3.7 Discussion

All criticisms of modelling the conditional mean for yields from Section 2 still hold here. To use the Bayesian Lasso and horseshoe prior hierarchy, Section 3.1.2 made the

distributional assumption that the yields follow a normal distribution with the mean as a linear combination of the Farm Business Survey and Met Office covariates. This was suitable for modelling average yields, since the coefficients were estimated to minimise the mean squared error, however failed to model large yields, where the observed yields often fell outside of the credible interval. This can be a result of there being no deciding factor in our linear model distinguishing large yields from the average yields. A separate analysis can be done specifically for large yields but it needs to be emphasised this would no longer be suitable to model average yields.

The Bayesian Lasso and the horseshoe prior hierarchy were selected due to their efficient shrinkage properties and their natural extensions to Chapter 2. Section 3.2 suggested a popular alternative prior for the coefficients, with comparable performance to the horseshoe prior, is the spike-and-slab prior (Ishwaran and Rao (2005)) using a scale mixture of normal distributions. The spike-and-slab prior is a weighted sum of a point mass at zero (i.e. the spike) and a centred normal distribution with a large variance (i.e. the slab). Therefore, rather than shrinking the coefficients of the unimportant variables down to approximately zero, the spike-and-slab prior would weigh the coefficients in favour of the point mass at zero. An interesting question would be to find out whether the same parameters are considered important as those using a regularized version of the horseshoe prior, found to closely resemble the spike-and-slab prior (Piironen and Vehtari (2017a)). Other methods for Bayesian model selection using an additional prior for the model space (e.g. Forte et al. (2018)) are computationally infeasible when the number of model parameters is large. An efficient sampler would be required (e.g. Clyde et al. (2011), García-Donato and Martínez-Beneito (2013)), hence these methods are avoided in our study but could also be studied to follow up on to this work.

To assess the variable importance in our analyses, we calculate the posterior mean divided by the posterior standard deviation (Makalic and Schmidt (2016)) and in this work refer to the outcome as the importance statistic. Makalic and Schmidt (2016) suggested this is in fact an estimate for the t-statistic, however care needs to be taken as to whether these should be referred to as the t-statistic since this suggests a hypothesis test can be constructed by comparing each approximate t-statistic to the t-distribution. Bayesian

estimates in a frequentist framework. A final fruitful avenue of work would be to look at whether a t-test is in fact valid when the estimates for the coefficients are found using MCMC methods.

4 | Extreme value analysis

4.1 Introduction

UK wheat yields have risen from a little over 2 tonnes per hectare in the early 20th century (Brassley (2000)) to current averages of approximately 8 tonnes per hectare (DEFRA (2017)). Knight et al. (2012) suggested there has been a lack of progression in wheat yields which has led us to question whether wheat yields have stagnated at a maximal level under current technologies and growing conditions. We address this using extreme value analysis, a statistical framework used to model extreme events which occur with a very small probability.

In the first of our contexts, the extreme value analysis of yields translates to modelling the highest yields over all farms and over all years. We carry out this analysis on winter wheat yields collated by the Farm Business Survey between 2006 and 2015. Each yield used in our analysis is each farm's highest yield attained over this time period. Let us highlight here that our objective is not to estimate the notion of yield potential, which is equal to the yield of a crop under ideal conditions (no pest, disease, nutrient or water stresses), or the related notion of water-limited yield potential (see e.g. van Ittersum et al. (2013); van Wart et al. (2013)). The estimation of these quantities typically requires the use of sophisticated computer models to simulate crop growth in specified conditions, see for example Chen et al. (2017) and Gobbett et al. (2017); our goal in Section 4.3.2 is rather to estimate the distribution of the highest yields attained in a real-world setting and under observed farming practices in order to estimate a practical upper bound on yield given current technology and conditions.

Our analysis of the highest wheat yields can also be refined to take account of growing conditions. In the literature, forecasts for winter wheat yields have been calculated for geographical regions, such as the Nomenclature of Territorial Units for Statistics level 1 (NUTS1) regions in Germany and France (de Wit et al. (2005)) or administrative regions in the UK defined by the Met Office (Cho et al. (2012)). This makes it possible to assess the variation in yield levels depending on climate and practices. The effect of the use of agricultural inputs, mainly fertilisers and crop protection, on average yield levels is also of interest; it is important, in this respect, to assess the trade-off between an improvement in yield and potential damage to the environment that may result from excessive use of those inputs. It has thus been found in the literature that the use of crop protection and fertiliser does indeed generally improve yield, but that a moderate level of these inputs typically brings the same improvement as higher levels without incurring the same risks to the environment and human health (see e.g. Damalas (2009); Ecobichon (2001); Pimentel et al. (1993)). Sections 4.3.3 and 4.3.4 use further information contained in the Farm Business Survey database to carry out extreme value analyses of winter wheat yield depending on location and level of crop protection and fertiliser use. We then compare the conclusions of each of these analyses, and contrast them with the interpretation of the extreme value analysis of the full, non-stratified sample of yields.

So far we have proposed to look at modelling highest yields without any regard to the financial implications to attain the yield such as labour and seed costs. Net margin is the profit gained from selling wheat after removing costs contributing to production. Since a high yielding farm does not necessarily imply a high performing farm from a business perspective, our final application of extreme value theory is based on the net margins to find what we expect the top performing farms to be achieving as an annual income. Section 4.3.5 concludes by assessing whether UK winter wheat production is able to provide a sustainable income and whether average performing farms allocate their spending differently to top performing farms.

The structure of this chapter is as follows. Section 4.1.1 gives a brief account of fields extreme value theory has already been used in. Section 4.2 discusses the available methods

to perform an extreme value analysis. Sections 4.3.1, 4.3.2, 4.3.3 and 4.3.4 are based on the published paper (Mitchell et al. (2020)), which details the implementation of those techniques first on the full data set, then on the data stratified by location and spending on agricultural inputs for maximum yield levels. Finally, Section 4.3.5 instead analyses high net margins. The chapter concludes with a reflection on the analyses and discusses future directions in Sections 4.4 and 4.5.

4.1.1 Example applications of extreme value theory

Extreme value theory has found applications in numerous fields, the most prominent examples being environmental science (see e.g. Coles and Walshaw (1994); Eastoe and Tawn (2009); Katz (1999)) and insurance and finance (see e.g. Embrechts et al. (1997); Hao et al. (2005); Marimoutou et al. (2009)). Other applications include engineering (see e.g. Holmes and Moriarty (1999); Steinkohl et al. (2013)) and toxicology (see Tressou et al. (2004)). More recently, extreme value analysis has been used in epidemiology to estimate the probability of severe pneumonia and influenza epidemics (see Thomas et al. (2016)), and in the field of demography with a discussion of whether there is a finite upper bound on human lifespan (see e.g. Rootzén and Zholud (2017)). Applications of extreme value analysis in the agricultural sciences have concentrated on financial aspects, for instance commodity price fluctuations (see e.g. Fretheim and Kristiansen (2015); Gong et al. (2015)), rather than agronomic factors such as yield. Our published work is the first to use agricultural yields in an extreme value framework (see Mitchell et al. (2020)). Extreme value theory has also been widely used on financial data, such as net margin (see e.g. Embrechts et al. (1997), Resnick (2007)), however there are, to the best of our knowledge, currently no applications of extreme value theory on the incomes of agricultural producers once expenditure has been accounted for.

4.2 Methods

Extreme value analysis methods are required in order to model the highest yields within each scenario and estimate a practical upper bound for each case. This section reviews

the foundational and theoretical aspects of extreme value theory needed (see de Haan and Ferreira (2006) and Beirlant et al. (2004)) to perform an extreme value analysis; the univariate case will be considered throughout.

4.2.1 Limiting distribution of the sample maxima

Extreme value analysis concerns the behaviour of the maximum of a sample $X_1, ..., X_n$ as sample size $n \to \infty$. Supposing $X_1, ..., X_n$ is an independent and identically distributed sample from a distribution F with possibly infinite right endpoint $x^* = \sup\{x : F(x) \leq 1\}$, then

$$\max(X_1, X_2, ..., X_n) \xrightarrow{P} x^* \text{ as } n \to \infty; \tag{4.1}$$

a necessary and sufficient condition for the existence of a linearly normalised version of this maximum to converge to a nondegenerate distribution H_γ is

$$\lim_{t \to x^*} \frac{1 - F(t + xf(t))}{1 - F(t)} = (1 + \gamma x)^{-1/\gamma} = 1 - H_\gamma(x) \tag{4.2}$$

where f is a positive nondecreasing function, $\gamma \in \mathbb{R}$ (see de Haan and Ferreira (2006), Theorem 1.1.6). It follows that as $t \to x^*$, the exceedances $Y = X - t$ approximately follow a generalised Pareto distribution

$$\mathbb{P}(X - t \leq y | X > t) \approx H_{\gamma,\sigma(t)}(y) := 1 - \left(1 + \frac{\gamma y}{\sigma(t)}\right)^{-1/\gamma}, \tag{4.3}$$

for all $y > 0$ such that $1 + \gamma y/\sigma(t) > 0$. de Haan and Ferreira (2006) add that the shape parameter γ ultimately determines what type of tail our data exhibits, while $\sigma(t)$ is a positive scale parameter. If $\gamma < 0$, then the approximate distribution of the exceedances in Equation 4.3 will be short tailed and a finite maximum can be found. If $\gamma > 0$, then the approximate distribution of the exceedances will be heavy tailed and a finite maximum will not exist. For $\gamma = 0$, $H_{\gamma,\sigma(t)}(y)$ becomes $H_{\sigma(t)}(y) = 1 - \exp(-y/\sigma(t))$. The following section looks at methods to estimate γ to determine whether a finite maximum exists.

4.2.2 Parameter estimators for the generalised Pareto distribution

Consistent for all $\gamma \in \mathbb{R}$, the moment estimator (Dekkers et al. (1989)) takes a semi-parametric approach to estimate the shape parameter γ using log-moments

$$\hat{\gamma} := M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1}, \tag{4.4}$$
$$\text{with } M_n^{(j)} := \frac{1}{k}\sum_{i=0}^{k-1}\left(\log X_{n-i,n} - \log X_{n-k,n}\right)^j, j = 1, 2,$$

given the sequence of integers $k = k(n)$ satisfying $k(n) \to \infty$, $k(n)/n \to 0$. Here $X_{j,n}$ is the j^{th} largest observation in the sample of size n and k is the index corresponding to threshold t such that $X_{n-k,n} = t$. k can also be seen as the sample size above threshold t, hence k will be referred to as the effective sample size henceforth. Furthermore, it is known that (de Haan and Ferreira (2006)) as $n \to \infty$

$$\hat{\gamma}_k \approx N\left(\gamma, \frac{1}{k}V\right), V = \begin{cases} \gamma^2 + 1, & \gamma \geq 0, \\ \dfrac{(1-\gamma)^2(1-2\gamma)(1-\gamma+6\gamma^2)}{(1-3\gamma)(1-4\gamma)}, & \gamma < 0, \end{cases} \tag{4.5}$$

where the bias is assumed to be zero from carefully selecting the effective sample size k. To select a suitable effective sample size, the shape parameters are estimated for increasing effective sample sizes k and plotted against each other. We find the first stable region reached as k increases and find the corresponding threshold. Increasing k is equivalent to increasing the number of observations the parameter estimate is based on. Consequently, as the index increases, the width of the associated confidence intervals decreases; nevertheless, this also increases bias in the estimate from involving less-extreme observations. To fulfil this bias-variance tradeoff, the largest threshold such that the corresponding index k produces an estimate $\hat{\gamma}_k$ contained in the stability region is selected, and in which case, we can assume the bias to be approximately zero.

de Haan and Ferreira (2006) also provide the corresponding estimate for the scale parameter σ,

$$\hat{\sigma} = X_{n-k,n}M_n^{(1)}(1 - \hat{\gamma}_-), \qquad \hat{\gamma}_- = \hat{\gamma} - M_n^{(1)},$$

where $M_n^{(1)}$ is defined in Equation 4.4.

Given the estimates for γ and σ in the generalised Pareto distribution approximation of the exceedances (Equation 4.3), Coles (2001) suggests the generalised Pareto assumption can be checked by plotting the sample quantiles containing the data against the theoretical quantiles of the generalised Pareto distribution,

$$\left\{\left(t + H^{-1}_{\hat{\gamma},\hat{\sigma}}\left(\frac{k-i}{k+1}\right), x_{n-i,n}\right) : i = 0, ..., k-1\right\},$$

where $x_{n-i,n}$ is the i^{th} largest observation in the sample of size n and k is the number of large values ultimately taken in the analysis. If the generalised Pareto assumption is appropriate, the sample quantiles should approximately equal the theoretical quantiles. Section 4.3 will check the generalised Pareto assumption holds for each dataset introduced.

Another method to test the generalised Pareto assumption is to compare the estimates found using the moment estimator and the estimates found using an estimator which relies more on the generalised Pareto assumption, namely the maximum likelihood estimator. If these are approximately equal, then the generalised Pareto assumption holds.

Being a popular choice for parameter estimation, the maximum likelihood estimator, used in previous studies including Eastoe and Tawn (2009), Steinkohl et al. (2013) and Walshaw and Anderson (2000), achieves narrow confidence intervals as a consequence of using a parametric approach. Zhou (2009) found the maximum likelihood estimator exists and is consistent for $\gamma > -1$. The likelihood function of the scaled generalised Pareto distribution to approximate the exceedances is

$$L(\gamma, \sigma | x_{n-k,n}, ..., x_{n,n}) = \prod_{i=1}^{k} h_{\gamma,\sigma}(x_{n-i+1,n} - x_{n-k,n}), \tag{4.6}$$

where $h_{\gamma,\sigma}(x) = \partial H_{\gamma,\sigma}(x)/\partial x$. The maximum likelihood estimates of γ and σ cannot be found analytically.

Given $k = k(n) \to \infty$ and $k/n \to 0$, for $\gamma > -1/2$ Drees et al. (2004) proved the asymptotic normality for this estimator takes the form

$$\begin{pmatrix} \hat{\gamma}_k \\ \hat{\sigma}_k/\sigma_k \end{pmatrix} \approx N_2\left(\begin{pmatrix} \gamma \\ 1 \end{pmatrix}, \frac{1}{k}\mathbf{V}\right), \text{ with } \mathbf{V} = \begin{pmatrix} (1+\gamma)^2 & -(1+\gamma) \\ -(1+\gamma) & 1+(1+\gamma)^2 \end{pmatrix}. \quad (4.7)$$

The small variance associated with the ML estimator is a desirable property, however the maximum likelihood estimator is only consistent for $\gamma > -1$. For $\gamma < -1$, the likelihood function for the scaled generalised Pareto distribution in Equation 4.6 tends to ∞.

If both estimators discussed here approximately agree on a shape parameter, the maximum likelihood estimate will be used in the quantile estimator in Section 4.2.4 to achieve narrower confidence intervals.

4.2.3 Hypothesis testing for distributional differences

After fitting generalised Pareto distributions to two sets of observations, a test for differences in the tail behaviour between these will determine whether the samples above threshold t can in fact reasonably be thought to come from the same distribution, or whether two separate extreme value analyses would be best to describe their distinct behaviours. Based on classical likelihood ratio tests (Silvey (1970), Cox and Hinkley (1974)), Coles (2001) suggested for two different samples A and B:

(i) fit two generalised Pareto models separately to the two groups A and B using maximum likelihood from Section 4.2.2, to find $(\hat{\gamma}_A, \hat{\sigma}_A)$ and $(\hat{\gamma}_B, \hat{\sigma}_B)$. The two maximised likelihoods are calculated separately $L(\hat{\gamma}_A, \hat{\sigma}_A|\mathbf{x}_A)$ and $L(\hat{\gamma}_B, \hat{\sigma}_B|\mathbf{x}_B)$, and finally $L = L(\hat{\gamma}_A, \hat{\sigma}_A|\mathbf{x}_A) \times L(\hat{\gamma}_B, \hat{\sigma}_B|\mathbf{x}_B)$ is computed. This represents the maximum likelihood under the full model, with 4 parameters, describing A and B jointly.

(ii) fit a single generalised Pareto model to the combined data $\mathbf{x}_{A\cup B} = (\mathbf{x}_A, \mathbf{x}_B)$ using maximum likelihood to find $(\hat{\gamma}_{A\cup B}, \hat{\sigma}_{A\cup B})$. The corresponding maximum likelihood $L_0 = L(\hat{\gamma}_{A\cup B}, \hat{\sigma}_{A\cup B}|\mathbf{x}_{A\cup B})$ is computed, representing the maximum likelihood under the restricted model where A and B can be described by the same distribution.

Finally, the relevant likelihood ratio test statistic for testing the null hypothesis of equal models $H_0 : (\gamma_A, \sigma_A) = (\gamma_B, \sigma_B)$ is the deviance $D = -2\log(L_0/L)$, to be compared to the 95% quantile of the χ^2 distribution with 4 (from (i)) -2 (from (ii)) $= 2$ degrees of freedom, equal to 5.99.

4.2.4 Quantile estimators

Given an estimate of the shape parameter is made using one of the estimators in Section 4.2.2, plugging the estimates of the shape and scale parameters into Equation (4.3) will provide an approximate distribution to model the tail. Furthermore, the estimate of γ will also determine whether a finite maximum exists; de Haan and Ferreira (2006) show $\gamma < 0$ indicates a finite maximal value x^* exists, whereas $\gamma > 0$ indicates a finite maximal value does not exist; however, we can still construct $100(1-p_n)\%$ quantiles with $p_n \neq 0$ as the exceedance probability.

Recall Equation (4.3) where the exceedances $Y = X - t$ are assumed to approximately follow a generalised Pareto distribution $H_{\gamma,\sigma(t)}(y)$ for large n. To find the quantiles, Beirlant et al. (2004) suggest inverting

$$\mathbb{P}\left(X - t \geq y\right) = \mathbb{P}\left(X > t\right)\left[1 - \mathbb{P}\left(X - t \leq y | X > t\right)\right] \approx \frac{k}{n}\left(1 + \frac{\gamma y}{\sigma}\right)^{-1/\gamma},$$

where $p_n := \mathbb{P}\left(X - t \geq y\right)$ is the probability of exceedance, giving

$$\hat{x}_{p_n} := t + \frac{\hat{\sigma}_k}{\hat{\gamma}_k}\left[\left(\frac{k}{np_n}\right)^{\hat{\gamma}_k} - 1\right], \tag{4.8}$$

where t is the selected threshold such that $P(X > t) = k/n$, $\hat{\sigma}_k$ and $\hat{\gamma}_k$ are the estimates of the scale and shape parameters respectively based on the k largest observations and total sample size n. As $n \to \infty$, de Haan and Ferreira (2006) give the asymptotic distribution:

$$\sqrt{k}\frac{(\hat{x}_{p_n} - x_{p_n})}{\hat{\sigma} q_{\hat{\gamma}}(d_n)} \xrightarrow{d} \Gamma + (\gamma_-)^2 B - \gamma_- \Lambda - \lambda \frac{\gamma_-}{\gamma_- + \rho}, \tag{4.9}$$

where (Γ, Λ, B) are jointly Normal random variables corresponding to the shape, scale and location parameters of H respectively, $\gamma_- = \min(0, \gamma)$. For $d_n = k/(np_n) \to \infty$, de Haan and Ferreira (2006) also show

$$q_{\hat{\gamma}}(d_n) = \begin{cases} (d_n)^{\hat{\gamma}} \log(d_n)/\hat{\gamma}, & \hat{\gamma} > 0, \\ 1/\hat{\gamma}^2 & \hat{\gamma} < 0, \end{cases}$$

since $q_{\hat{\gamma}}(d_n)$ converges to $q_{\gamma}(d_n)$, which is used to ensure the limit distribution of Equation (4.9) is a linear combination of the limit distributions already found for the parameters. For $\hat{\gamma} < 0$, Equation (4.8) allows p_n to be replaced by zero without being undefined to produce the finite endpoint estimator

$$\hat{x}_k^* = t_k - \frac{\hat{\sigma}_k}{\hat{\gamma}_k}, \tag{4.10}$$

and $q_{\hat{\gamma}}(d_n) = 1/\hat{\gamma}^2$ in Equation (4.9) with asymptotic distribution

$$\hat{x}_k^* \approx x^* + \frac{1}{\sqrt{k}} \times \frac{\hat{\sigma}_k}{\hat{\gamma}_k^2} \times N(0, V_{\hat{\gamma}}) \tag{4.11}$$

where $V_{\hat{\gamma}} = \text{Var}\left(\Gamma + (\hat{\gamma}_-)^2 B - \hat{\gamma}_- \Lambda\right)$. When γ is estimated using the maximum likelihood estimator, the covariance matrix of the bivariate distribution (Γ, Λ) is known from Equation (4.7), and B, independent of Γ and Λ, is asymptotically standard normal. Independence between B and (Γ, Σ) is due to maximum likelihood estimator for γ and σ being built on exceedances above the threshold rather than the location parameter $X_{n-k,n}$ itself. For the maximum likelihood estimate of $\hat{x}^*$,

$$\hat{x}_k^* \approx x^* + \frac{1}{\sqrt{k}} \times \frac{\hat{\sigma}_k}{\hat{\gamma}_k^2} \times N(0, 1 + 4\hat{\gamma} + 5\hat{\gamma}^2 + 2\hat{\gamma}^3 + 2\hat{\gamma}^4). \tag{4.12}$$

For $\hat{\gamma} > 0$ and $(np_n/k)^{-\hat{\gamma}_k}$ to exist, p_n can not equal zero. Instead, quantiles in less extreme regions compared to the 100% quantile can be looked at. Sample quantile estimates for these cases would be unsuitable due to the average number of observations above each quantile being small. For $\hat{\gamma} > 0$, de Haan and Ferreira (2006), Theorem 1.2.5

states a simpler form of the quantile estimate in Equation (4.8), which uses the fact if the limit relation in Equation (4.2) holds for some $f > 0$, then it also holds with $f(t) = \gamma t$, where by rearrangement

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}. \tag{4.13}$$

Now

$$\mathbb{P}\left(\frac{X}{t} \geq y\right) = \mathbb{P}\left(\frac{X}{t} \geq y | X > t\right) \mathbb{P}\left(X > t\right) \approx \frac{k}{n}\left(y^{-1/\gamma}\right)$$

where $\mathbb{P}\left(\frac{X}{t} \geq y\right)$ is now the exceedance probability p_n. Inverting gives

$$\hat{x}_{p_n} = t_k d_n^{\hat{\gamma}_k}, \tag{4.14}$$

where $d_n = k/(np_n)$. Using the general asymptotic distribution in Equation (4.9) with $\gamma > 0$, the asymptotic distribution for $\hat{x}_{p_n}$ is

$$\hat{x}_{p_n} \approx x_{p_n}\left(1 + \frac{\log\left(d_n\right)}{\sqrt{k}} \times N(0, V_\gamma)\right), \tag{4.15}$$

where using the maximum likelihood estimator for γ becomes

$$\hat{x}_{p_n} \approx x_{p_n}\left(1 + \frac{\log\left(d_n\right)}{\sqrt{k}} \times N(0, (1+\gamma)^2)\right). \tag{4.16}$$

From Section 4.2.2, if the parameter estimates approximately agree on a shape parameter, then the maximum likelihood estimate is carried forward, however if we would prefer to not rely on the generalised Pareto assumption, the moment estimate can be used instead. For $\gamma > 0$, $V_\gamma = \gamma^2 + 1$ from Equation (4.5). For $\gamma < 0$,

$$\mathrm{Var}(\Gamma, \Lambda) = \frac{(1-\gamma)^2}{(1-3\gamma)(1-4\gamma)} \begin{pmatrix} (1-2\gamma)(1-\gamma+6\gamma^2) & -1+4\gamma-12\gamma^2 \\ -1+4\gamma-12\gamma^2 & \frac{2-16\gamma+51\gamma^2-69\gamma^3+50\gamma^4-24\gamma^5}{(1-\gamma)^2(1-2\gamma)} \end{pmatrix},$$

(see de Haan and Ferreira (2006), Corollary 4.2.2) B remains independent of Γ but $\mathrm{Cov}(\Lambda, B) = \gamma$ when γ is estimated using the moment estimator.

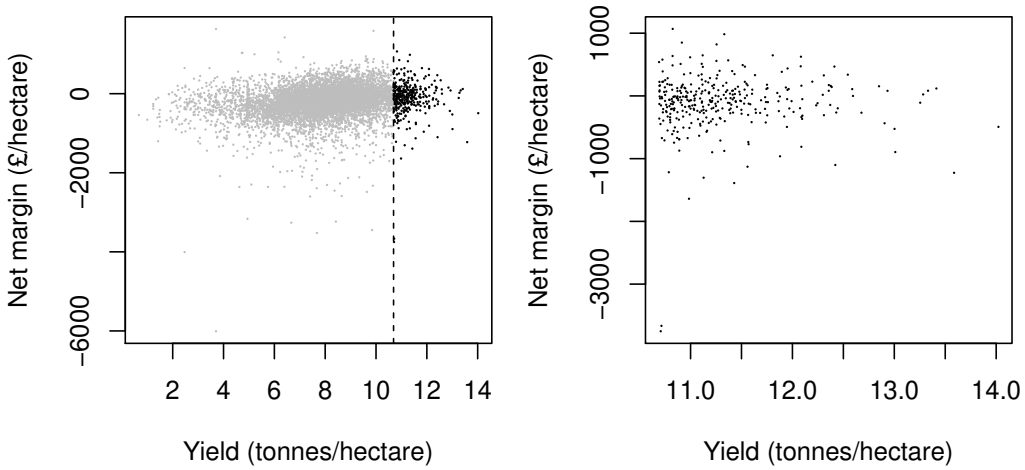


Figure 4.1: Left: plot of yield per hectare versus net margin per hectare for all 6951 observations. Vertical line is at 10.69 tonnes per hectare equivalent to the threshold taken in Section 4.3.2 for the extreme value analysis. Right: yield versus net margin for observations with yield above threshold 10.69 tonnes per hectare.

4.3 Application

Using data from the Farm Business Survey, we perform an extreme value analysis of wheat yields with techniques from Section 4.2 in order to estimate a practical upper bound on yield, if one exists. This will indicate whether wheat yields have stagnated under current technologies and growing conditions. The maximum yield under various scenarios is also estimated by performing an extreme value analysis separately for each case, where Section 4.3.3 concerns geographical location and Section 4.3.4 concerns input levels. Comparing the estimated maximums will indicate whether it is possible to achieve a larger yield by changing farming practices. Mitchell et al. (2020) is composed of Sections 4.3.2, 4.3.3 and 4.3.4. Figure 4.1 shows a high yielding farm does not necessarily go on to achieve a high net margin, therefore we also perform an extreme value analysis on net margins in Section 4.3.5. First, we explain how we select unique observations from the data set to use in the extreme value analyses.

4.3.1 Data selection

Section 1 gave details of the information collected which constitutes the Farm Business survey. To decide which out of the 6951 observations should be used to make up the

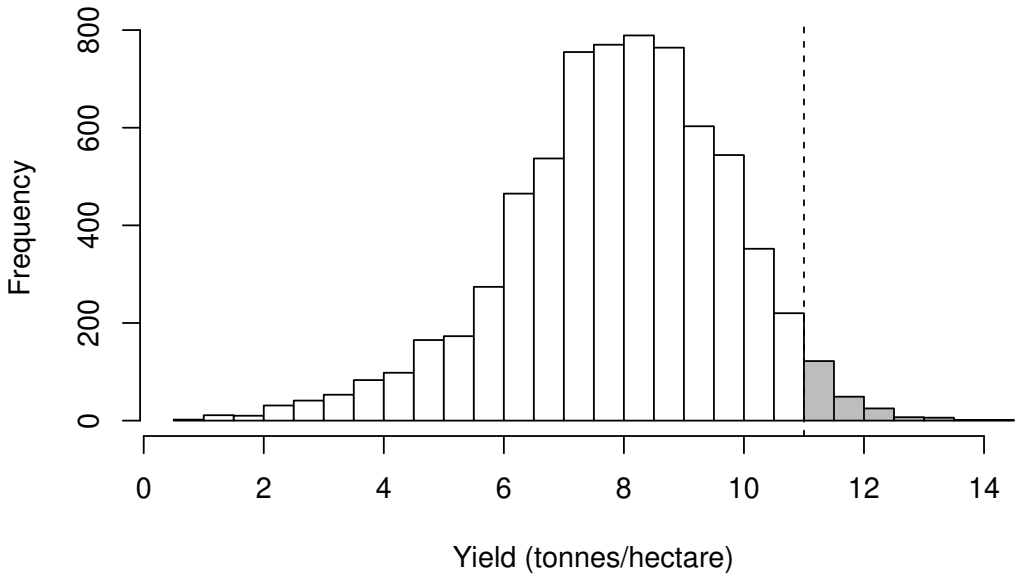


Figure 4.2: Histogram of yields for all 6951 observations with yields above 11 tonnes per hectare indicated in the tail of the distribution.

sample of high-yielding farms, intuitively all of the yields can initially be aggregated together without taking account of what years each yield is achieved, and extreme value techniques could be applied to the tail of the distribution of yield in Figure 4.2. However the selected observations exceeding the threshold may contain yields from the same high-yielding farms and therefore will not accurately represent all of the high-yielding farms across England and Wales.

Alternatively, each set of annual yields between 2006 and 2015 could be used in separate extreme value analyses to take account of temporal dependence. Even though this will ensure each farm attaining a high yield will only contribute one observation to the annual sample to carry out extreme value analysis, it needs to be emphasised that interest lies with finding the best yield possible over this time period by combining the observations and not the curve of maximum yield through time.

Instead, for each farm which took part in the Farm Business Survey, the maximum yield each farm attained between 2006 and 2015 is retained for the analyses in Sections 4.3.2, 4.3.3 and 4.3.4. This ensures the same high achieving farms are not duplicated in our analyses. If a farm consistently attains high yield, then our analysis would be biased towards this farm if yields are combined across years. Section 4.3.2 is based on yields

without stratification. Section 4.3.3 performs 3 separate analyses by splitting the highest yields into 3 groups according to location: west England and Wales, north England and east England. Section 4.3.4 performs 3 separate analyses by splitting the highest yields into 3 groups according to input use: low, medium and high. Since Section 4.3.5 concerns the high-income farms, the maximum net margin for each farm between 2006 and 2015 is retained instead.

4.3.2 Extreme value application on maximum wheat yields

To estimate the maximum value of yield, the threshold for our extreme value modelling of yield is chosen first, or equivalently the number k of high data points employed. This is done by representing the curve of ML estimates of the shape parameter γ as a function of k in Figure 4.4 and suggests the ML estimate $\widehat{\gamma}_k$ is stable for k between 100 and 250, implying that the largest 250 observed yields constitute a suitable sample of data on which to base our analysis of high yields. The choice $k = 250$ corresponds to taking the threshold $t = t_{250} = 10.69$ tonnes per hectare, and the ML estimate for the shape parameter γ is then $\widehat{\gamma}_{250} = -0.11$ with 95% confidence interval $(-0.22, 0.00)$ (throughout, all confidence intervals are calculated at the approximate 95% confidence level). The corresponding ML estimate for the scale parameter is $\widehat{\sigma}_{250} = 0.76$ $(0.65, 0.91)$. Since the moment estimator gave a shape parameter estimate close to the ML estimate, the generalised Pareto assumption holds and estimates using maximum likelihood are preferred due to their narrower confidence intervals. Figure 4.3 also confirms the generalised Pareto assumption holds since the sample quantiles of maximal yields above threshold $t = 10.69$ approximately equal the theoretical quantiles found using the ML estimates for γ and σ. With the negative shape parameter estimate, using Equation (4.10), a finite right endpoint estimate is $\hat{x}^*_{250} = t_{250} - \hat{\sigma}_{250}/\hat{\gamma}_{250} = 17.60\,(11.44, 23.75)$ tonnes per hectare. Confidence intervals have been calculated according to the asymptotic distribution of the endpoint estimator (Equation 4.12).

Since the lower bound of the Gaussian confidence interval is not constrained to be larger

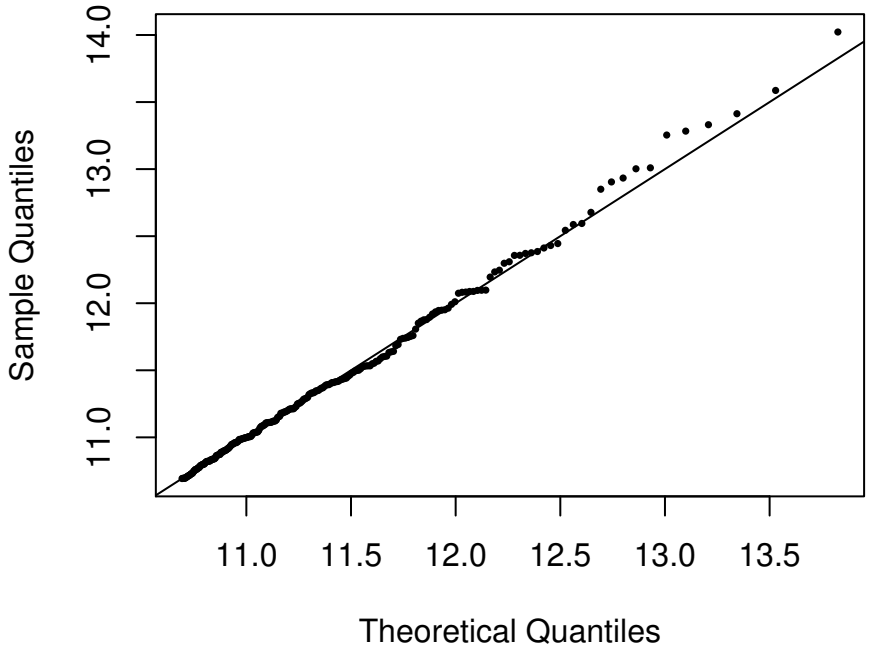


Figure 4.3: Plot of sample quantiles of the yields above the threshold $t = 10.69$ against the theoretical quantiles of the generalised Pareto distribution with the estimates $\hat{\gamma} = -0.11$ and $\hat{\sigma} = 0.76$.

than the maximum value in the sample, the interval

$$\left(\max\left[t_0, \hat{x}_k^* - \frac{1.96}{\sqrt{k}} \times \frac{\hat{\sigma}_k}{\hat{\gamma}_k^2} \times \sqrt{1 + 4\hat{\gamma}_k + 5\hat{\gamma}_k^2 + 2\hat{\gamma}_k^3 + 2\hat{\gamma}_k^4}\right], \right.$$
$$\left. \hat{x}_k^* + \frac{1.96}{\sqrt{k}} \times \frac{\hat{\sigma}_k}{\hat{\gamma}_k^2} \times \sqrt{1 + 4\hat{\gamma}_k + 5\hat{\gamma}_k^2 + 2\hat{\gamma}_k^3 + 2\hat{\gamma}_k^4}\right) \quad (4.17)$$

(Mitchell et al. (2020)) is used as an approximate 95% confidence interval for the maximum yield x^*, where t_0 denotes the maximum value in the sample. Truncating the interval at level t_0 does not affect its coverage probability in practice because, by definition, the true value x^* of the right endpoint must be larger than t_0 with probability 1. Using Equation (4.17) instead gives the confidence interval for the endpoint estimate as $(14.02, 23.75)$. These results, along with those of the subsequent analyses of maximum yields, are shown in Table 4.2. This estimate of a finite upper bound for winter wheat yield agrees with the physical intuition that yield per hectare should be bounded by a maximum yield which cannot be exceeded.

The current verified records for UK and worldwide wheat yields are 16.52 (observed in

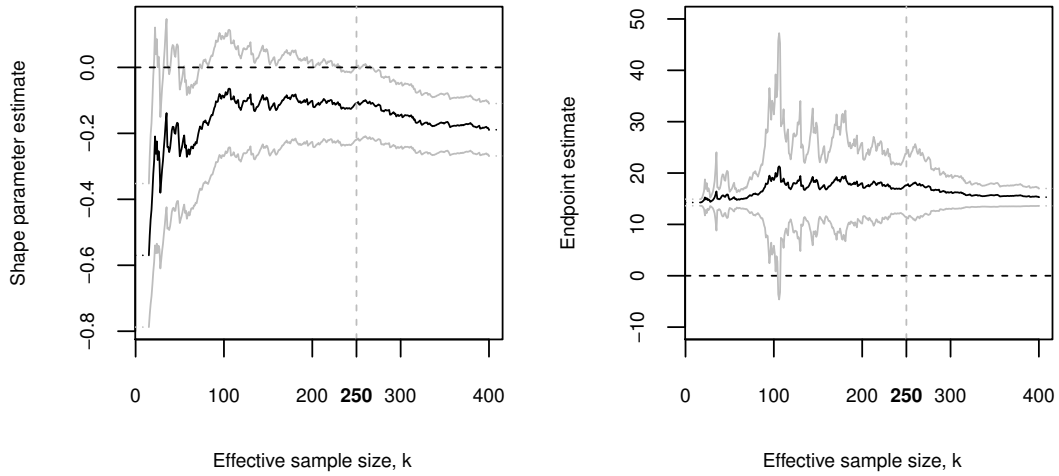


Figure 4.4: Left: plot of the ML estimate of the shape parameter γ, right: plot of the estimate of the endpoint x^*. Both plots give the estimates as a function of the effective sample size k taken for the estimation, with corresponding approximate 95% Gaussian confidence intervals. Estimates for sample sizes smaller than 15 and greater than 400 are omitted due to large variance and large bias respectively.

2015) and 17.40 (observed in New Zealand in 2020 and confirmed by Guinness World Records) tonnes per hectare, suggesting that our estimated value of 17.60 tonnes per hectare is a sensible estimate of this maximum possible yield. Although this extreme value analysis of winter wheat yield provides an estimate of the maximum attainable yield per hectare, this does not give any idea of the potential variation of wheat yields depending on geography or growing conditions. These two questions are the focus of the next two refined analyses.

4.3.3 Difference in geographical regions

Section 1 stressed the importance to carry out regional analyses of yield. We group farms using the macro-regions west England and Wales, north England and east England. This results in sample sizes of, respectively, 435, 331 and 770. These are adequately large sample sizes which ensure after each sample has been thresholded, there will still be a sufficiently large effective sample size to perform an extreme value analysis. We also note that, in addition to containing the highest number of farms, east England has a larger average yield per hectare figure compared to the other two regions.

Based on this geographical subdivision, we carry out an extreme value analysis similar to the global analysis of the previous section to model regional high yields. This is justified by likelihood ratio tests based on the generalised Pareto model from Section 4.2.3, which show that the model appropriate to the description of high yields depends indeed on the chosen region; see Table 4.1. The regional shape parameter estimates, as a function of effective sample size, are plotted in Figure 4.5.

As Table 4.2 shows, all three regions reassuringly give negative shape parameter estimates, albeit with wider confidence intervals; this was expected since stratifying decreases the available sample size and therefore increases uncertainty. Together with matching estimates of the regional scale parameter, Figure 4.7 indicates the generalised Pareto assumption is appropriate for all three cases. These shape and scale parameter estimates make it possible to produce estimates of regional upper bounds for yield using Formula (4.10). These estimates are $17.68\,(13.25, 29.11)$ tonnes per hectare for west England and Wales, $15.91\,(13.59, 21.20)$ for north England, and $17.81\,(14.02, 26.98)$ for east England.

The wide confidence intervals on these extreme value estimates make it impossible to suggest that, at the 95% level, there are regional differences between maximal yields across the three considered regions, although we do mention that the point estimate of maximal yield is noticeably lower for north England. We conclude this analysis by mentioning that although the point estimates of maximal yield in west England and Wales and east England are slightly higher than the point estimate across the whole data set, the increase is statistically insignificant and appears to be due to the fluctuations of the maximum yield estimate as a function of the effective sample size k. There is therefore no inconsistency between these stratified results and our earlier global analysis.

4.3.4 Difference in inputs

The effects of fertiliser and crop protection use for large-scale agricultural activities on public and the environment motivates our idea of assessing whether the effect of agricultural inputs on maximal wheat yield levels can be identified. We divide the sample

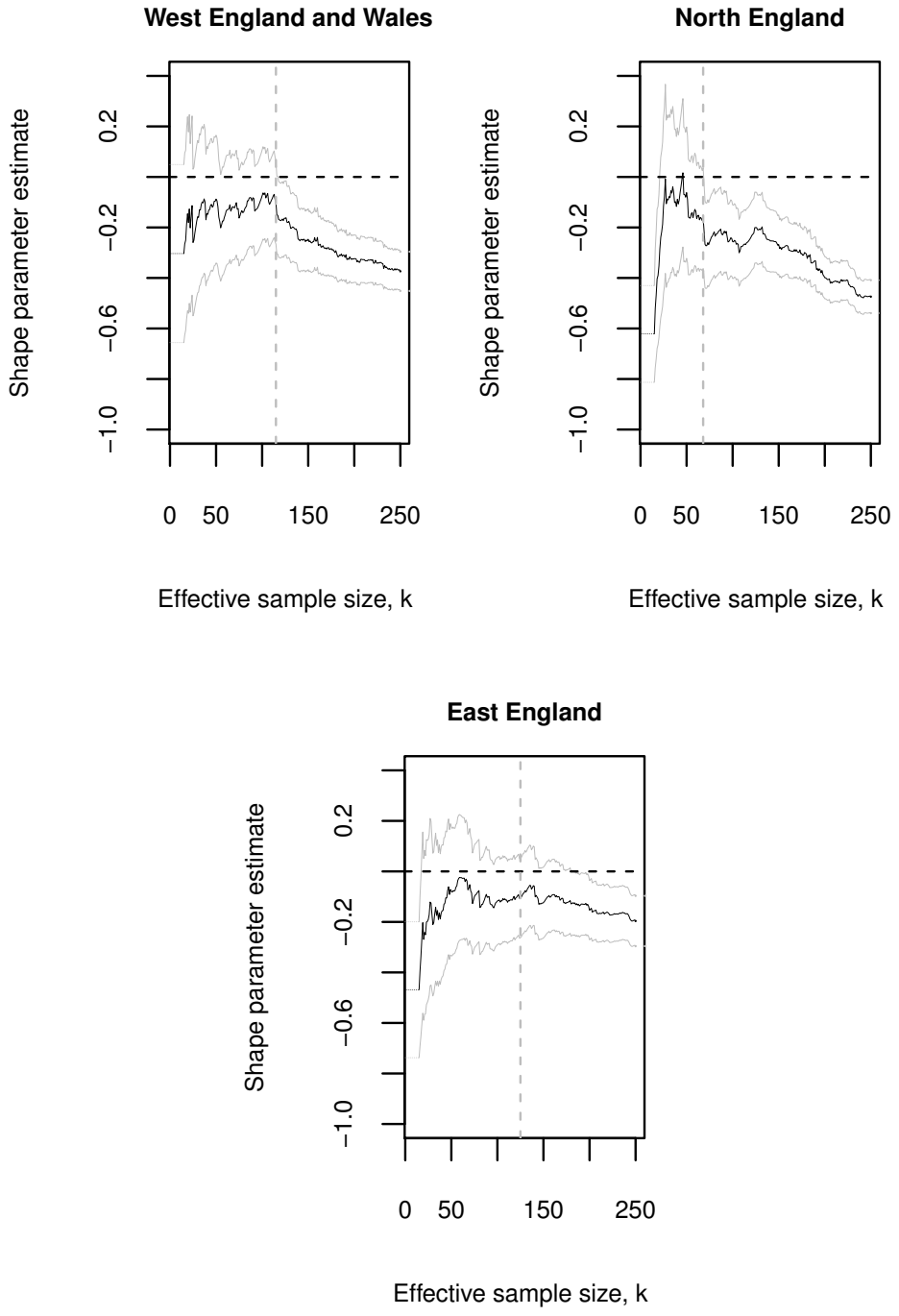


Figure 4.5: ML estimates of γ, for west England and Wales (top left), north England (top right) and east England (bottom).

of $n = 1536$ farms into three equally sized groups according to their expenditure on fertilisers and crop protection: low (less than £271.50 per hectare per year, corresponding to the bottom third in terms of expenditure), medium (between £271.50 and £370.10

per hectare per year, corresponding to the middle third), and high (greater than £370.10 per hectare per year, corresponding to the top third). An extreme value analysis based on organic farms only is not appropriate because the sample size is not large enough to threshold the data and still include enough data points to satisfy the bias-variance tradeoff in Section 4.2.2.

Based on this stratification by spending, and in view of the results of the likelihood ratio tests in Table 4.1 indicating that the appropriate model for high yields indeed depends on input level, we carry out an extreme value analysis for each scenario similar to the above regional analysis. Shape parameter estimates are represented in Figure 4.6. Figure 4.8 indicates the generalised Pareto assumption is valid, since the sample quantiles within each scenario are approximately equal to the quantiles of the generalised Pareto distribution with their respective parameter estimates from Table 4.2.

All three categories give negative shape parameter estimates, although the estimate for low input levels lies outside the confidence interval for the estimate of the shape parameter estimate of the full yield data, suggesting a significant difference in the behaviour of high yields for low spenders. The associated upper limit estimates for yield are 14.27 $(12.85, 16.52)$, 16.40 $(13.28, 24.99)$ and 19.18 $(14.02, 33.58)$ for low, medium and high use of inputs, respectively. The value and uncertainty on the maximal yield estimates for low spending on inputs do suggest that the use of fertiliser and crop protection improves the maximum attainable yield; however, and despite a point estimate of maximal yield being higher for the biggest consumers of these inputs than for average users, the uncertainty on our estimates does not provide significant evidence that spending a larger amount of capital on fertiliser and crop protection improves maximal yield levels.

Case	Deviance D	Conclusion
Tests for geographical regions		
A = North England, B = West England & Wales	107	Reject H_0
A = North England, B = East England	262	Reject H_0
A = West England & Wales, B = East England	240	Reject H_0
Tests for input use		
A = Low input, B = Medium input	113	Reject H_0
A = Low input, B = High input	173	Reject H_0
A = Medium input, B = High input	71.7	Reject H_0

Table 4.1: Likelihood ratio test statistics when testing if the samples come from the same distribution, $H_0 : (\gamma_A, \sigma_A) = (\gamma_B, \sigma_B)$ (see Section 4.2.3). The likelihood ratio statistics are compared to the 95% quantile of the χ^2 distribution with 2 degrees of freedom, equal to 5.99.

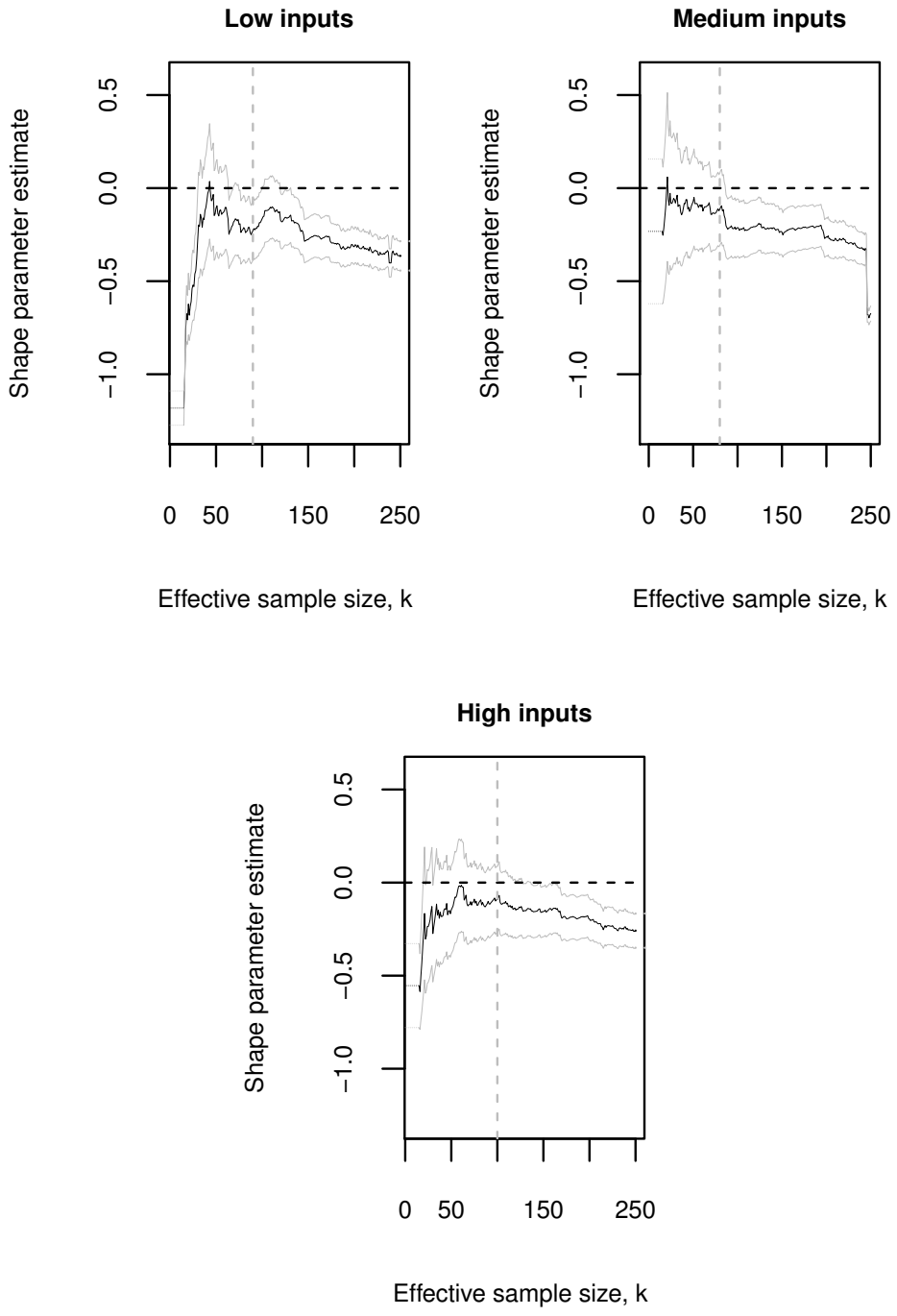


Figure 4.6: ML estimates of γ, for low input levels (top left), medium input levels (top right) and high input levels (bottom).

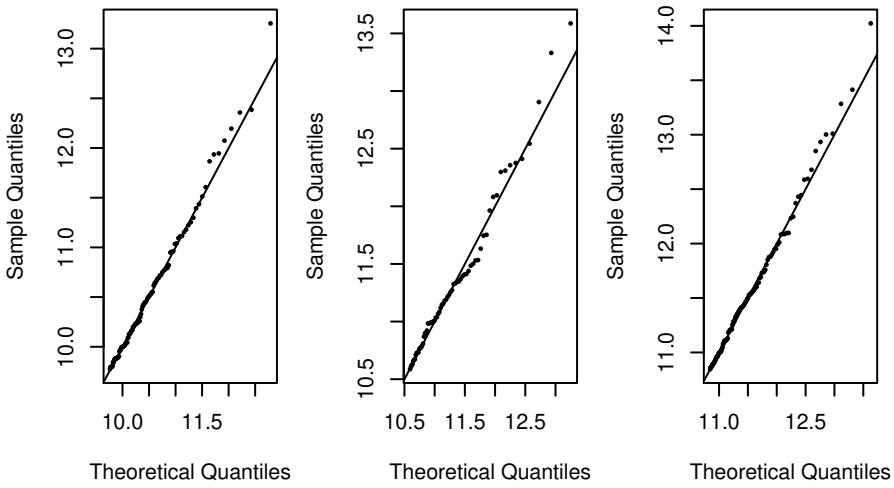


Figure 4.7: Plot of sample quantiles of the yields for low input levels (left), medium input levels (centre) and high input levels (right) above their respective thresholds against the theoretical quantiles of the generalised Pareto distribution with their respective parameter estimates for γ and σ in Table 4.2. For example, the sample quantiles of the yield for low input levels above threshold $t = 9.93$ is plotted against the theoretical quantiles of the generalised Pareto distribution with $\hat{\gamma} = -0.23$ and $\hat{\sigma} = 0.99$.

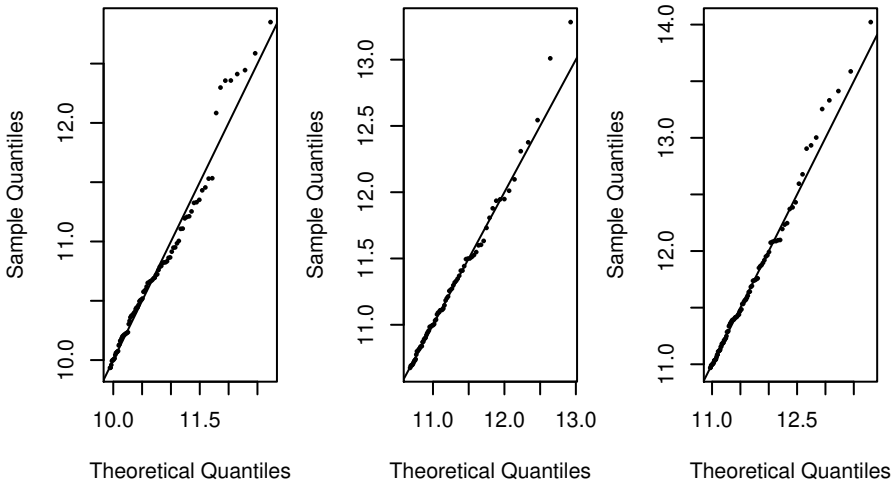


Figure 4.8: Plot of sample quantiles of the yields for west England and Wales (left), north England (centre) and east England (right) above their respective thresholds against the theoretical quantiles of the generalised Pareto distribution with their respective parameter estimates for γ and σ in Table 4.2. For example, the sample quantiles of the yield for west England and Wales above threshold $t = 9.76$ is plotted against the theoretical quantiles of the generalised Pareto distribution with $\hat{\gamma} = -0.10$ and $\hat{\sigma} = 0.80$.

Variable	n	k	t	Shape estimate $\hat{\gamma}$	Scale estimate $\hat{\sigma}$	$\hat{x}^* = t - \hat{\sigma}/\hat{\gamma}$
Yield	1536	250	10.69	-0.11 $(-0.22, 0.00)$	0.76 $(0.65, 0.91)$	17.60 $(14.02, 23.75)$
Location						
West England and Wales	435	115	9.76	-0.10 $(-0.27, 0.06)$	0.80 $(0.65, 1.07)$	17.68 $(13.25, 29.11)$
North England	331	68	10.58	-0.16 $(-0.36, 0.03)$	0.87 $(0.67, 1.27)$	15.91 $(13.59, 21.20)$
East England	770	125	10.84	-0.11 $(-0.26, 0.05)$	0.74 $(0.60, 0.96)$	17.81 $(14.02, 26.98)$
Inputs						
Low (Input < 271.5)	512	90	9.93	-0.23 $(-0.39, -0.07)$	0.99 $(0.79, 1.34)$	14.27 $(12.85, 16.52)$
Medium ($271.5 \leq$ Input < 370.1)	512	80	10.67	-0.11 $(-0.31, 0.08)$	0.65 $(0.50, 0.92)$	16.40 $(13.28, 24.99)$
High (Input > 370.1)	512	100	10.96	-0.09 $(-0.27, 0.09)$	0.75 $(0.60, 1.03)$	19.18 $(14.02, 33.58)$

Table 4.2: Maximum yield level estimates $\hat{x}^*$ for the full data set and the data stratified with respect to region or spending on agricultural inputs, along with a summary of sample sizes, threshold choices, shape estimates $\hat{\gamma}$ and scale estimates $\hat{\sigma}$. Numbers in brackets next to shape, scale and maximum yield estimates represent approximate 95% confidence intervals.

4.3.5 Extreme value application on net margin

Since yield does not take account of the financial implications of agricultural inputs, we instead look to estimate what we expect the top performing farms to achieve financially. Rather than using the maximum attained yields, we take the maximum attained net margin for each farm across the all of the years in which the farm took part in the survey and construct extreme quantile estimates of net margin for the top performing 1%, 0.5% and 0.1% farms.

To estimate the extreme quantiles for income, we initially follow a similar procedure to estimating maximum yields in Section 4.3.2 by selecting a threshold using a plot of the shape parameter estimates $\hat{\gamma}$ against the effective sample size k. The curve in Figure 4.9 is stable for k between 100 and 140 indicating our extreme value modelling may be based on the largest 140 net margins. The maximum likelihood estimate for γ with $k = 140$ is $\hat{\gamma} = 0.138$ $(-0.050, 0.327)$. In this case ($\gamma > 0$), a quantile estimate will be found since endpoint estimates do not exist for positive shape parameters. Given the shape parameter estimated using the moment estimator is approximately equal to the shape parameter estimated using the maximum likelihood estimator, we will continue to use the maximum likelihood estimate due to its narrow confidence intervals. For $p = 0.01$, 0.005 and 0.001 corresponding to the 99%, 99.5% and 99.9% quantiles for net margin, we estimate the top 1%, 0.5% and 0.1% of farms to achieve $£655.17(£462.45, £1,123.25)$, $£721.12(£466.01, £1,593.38)$ and $£901.00(£486.79, £6,043.95)$ per hectare where asymptotic 95% confidence intervals are constructed using the asymptotic distribution of quantile estimates in Equation (4.16).

Reflecting on our analysis, we have only been concerned with figures per hectare and not necessarily the financial result of the entire farm. Averaging over the number of hectares attributed to growing wheat for each farm in England and Wales, a typical farm is assumed to be approximately 160 hectares. Supposing a farm of this size achieves a net margin in the top 1% of all farms, estimated to be at least £655.17, then an estimated profit for the wheat enterprise alone is $£104,827.64$ per year, or equivalently $£8,735.55$ per month; four times the average wage in the UK. While comparing net margins with

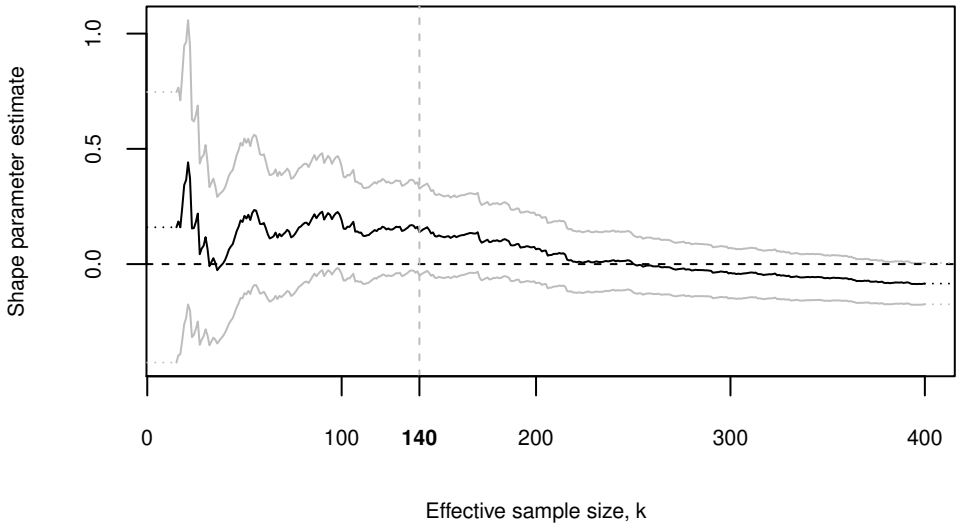


Figure 4.9: ML estimate of the shape parameter γ for net margins.

average wages does not provide identical like-for-like comparatives of return to a farmer, the net margin return in this context provides a measure of the return to the farmers managerial input to the running of the business. Analysing this figure in relation to society as a whole, this would place a top 1% performing farm amongst the top 3% in the UK in terms of annual salaries, according to HM Revenue and Customs (2017). Even though the income for a top 1% performing average-sized farm falls short of the top 1% annual salaries in the UK, we cannot imply the average income of average-sized farms falls short of the average UK salary from our extreme value application which models highest net margins.

We can however look at which expenditure top performing farms prioritise to achieve a high yield compared to average performing farms. By grouping seed, fertiliser and pesticide expenditure under the term "agricultural inputs", and labour, machinery and contracting costs under "labour inputs" from the Farm Business Survey data discussed in Section 1, Figure 4.10 indicates the top-performing farms may not attribute their costings any differently to other farms, and spend approximately the same per hectare. The cause of the difference in yields between top performing farms and average performing farms across England and Wales remains unknown from our analyses in the previous sections, yet here costing allocations appear to not play a role.

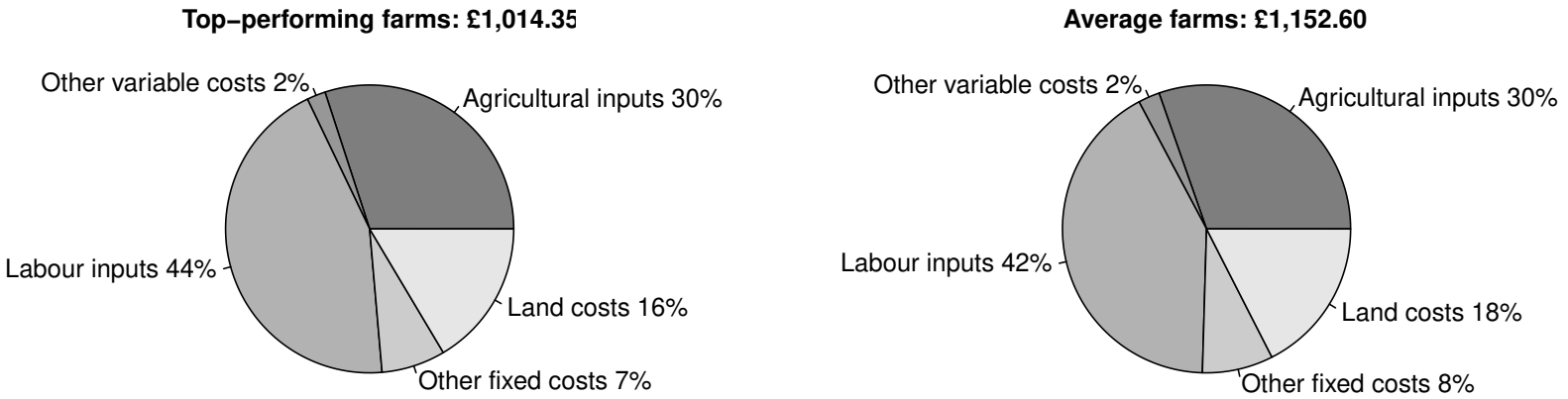


Figure 4.10: Average annual spending of top-performing farms and the remaining farms.

4.4 Conclusion

Our analysis of ten years of recent winter wheat production data, collected in England and Wales by the Farm Business Survey, indicates that annual winter wheat yields per hectare have a finite upper bound which we estimate to be 17.60 tonnes. Our model, based on the use of a generalised Pareto distribution suggested by the framework of extreme value analysis, was also adapted to the estimation of regional maximal yields and maximal yields as a function of spending on agricultural inputs. These estimates seem plausible, and show that:

- Although the maximum yield point estimate for north England is lower than the corresponding ones for west England and Wales and east England, there is insufficient statistical evidence to suggest that north England farms cannot reach the estimated maximum yield of 17.60 tonnes per hectare;
- There is an increase in maximum yield from low to high use of fertiliser and crop protection, although the difference between the maximal yields of medium and high spenders on these inputs is not statistically significant.

Furthermore, our analysis of annual net margins to assess performance financially indicates the top 1% of farms are expected to earn at least £655.17 per hectare when producing winter wheat, placing them amongst the top 3% earners in society as a whole, yet they do not allocate their spending any differently to average performing farms.

4.5 Discussion

To use our ML estimators of the shape and scale parameters, and then deduce an estimate of the right endpoint of yield, we had to make the distributional assumption that yields above a sufficiently high threshold approximately follow a generalised Pareto distribution. The quality of this approximation is a critical factor in the performance of the estimators, and may lead to poor estimates if the underlying distribution of high yields is too far from our model. The moment estimator is flexible in the sense that its validity is not rooted

in the generalised Pareto assumption, but the price to pay for this is a higher asymptotic variance compared to the ML estimator. In this work we checked the generalised Pareto assumption by examining quantile-quantile plots and comparing the maximum likelihood estimates to the moment estimates. Further tests on the validity of the generalised Pareto assumption have been proposed by de Haan and Ferreira (2006) and Hüsler and Li (2006) which can be used as alternatives to those used in this chapter.

The second part of our analysis was an effort to assess the dependence of the maximum yield on location of a farm. The point estimate of maximal yield in north England, which is 15.91 tonnes per hectare, is actually lower than the verified record for this region, which is also the UK record of 16.52 tonnes per hectare, attained in a Northumberland farm in 2015. This data point, which is not part of the data from the Farm Business Survey and hence not taken into account in our methodology, is well within the confidence interval calculated for the maximum yield in north England and thus not inconsistent with our results. Analysing the reasons behind this extremely high yield reveals that, while the north of England typically suffers from increased rainfall, lower temperatures and limited sunshine compared to the southern part of the UK, this was not the case in 2015 (DEFRA (2015)).

The third part of our extreme value analysis, stratified with respect to spending on agricultural inputs, suggested that there is not a statistically established increase in maximum yield arising from a large use of crop protection and fertilisers. Our findings, consistent with previous studies (Reader et al. (2018), Wilson et al. (2001)), indicate the potential for an upper-level marginal input use reduction while still obtaining high yields, providing high food production potential, increased farmer profit and reduced environmental footprint. Our statistical analyses demonstrate no significant difference in extreme yield between medium and high input use, and that additionally there was no significant difference in maximum yield across the three regions within the dataset, implying that soil type and weather variation are, on aggregate, not the main drivers of high yields within the data. It would also be informative to re-test the hypothesis of the difference in maximum attainable yields against different fertiliser and crop protection input use levels from a larger sample of data, for example, drawn from European wide

data or from the USA. This would reduce the width of the confidence intervals for the estimates of maximum yield stratified according to spending in agricultural inputs. The potentially large yield gains to be made, starting from average yield levels, imply that detailed farm level studies of agricultural practice with statistically relevant numbers of observations would be worthwhile.

The models based on wheat yields alone and various stratifications of input level and location were discussed in Mitchell et al. (2020). For this work, separate analyses were conducted for each scenario independently, where each case may have a different shape and scale parameter. Furthermore, the input levels were segmented into discrete classes rather than the original input value from the Farm Business Survey. An interesting question, which is beyond the scope of this work and to be addressed in future research, is to find a precise model for the description of high levels of yield as a function of agricultural input use and location. This could be done by, for instance, letting the shape or scale parameter (or both) of the generalised Pareto distribution vary smoothly as a function of input level or geographical coordinates, for example $\gamma = \gamma(x)$ or $\sigma = \sigma(x)$ where x is the continuous variable influencing yield, as described for instance in Chavez-Demoulin et al. (2016). This would allow the shape or scale parameter to vary with input use, hence the endpoint estimate would also vary with input use. Such an analysis would allow for the prediction of the high and maximum levels of yield attainable under various biological and physical circumstances, and would thus be important for agricultural policy and decision-making.

Another fruitful avenue of further work to build on Mitchell et al. (2020) would be to incorporate the agronomic, socio-ecological and climate variables from Chapters 2 and 3 to design a model which includes the influence of these variables in an extreme value analysis. Chapters 2 and 3 used regression and model selection techniques to model typical yields. Recent literature has looked to develop the Bayesian Lasso in the extreme value framework (de Carvalho et al. (2021)) to take account of the dependence of covariates on the variable of interest. In our case, we would look to model maximum yield depending on the covariates in Chapters 2 and 3. Furthermore, it would be interesting to compare the variables selected using Bayesian model selection to model a typical yield

and a maximal yield. Such a model would also be very useful when accounting for the effect of climate change on maximal yield levels.

The final part based on net margins suggested there is no difference between how the average performing farms and the top performing farms allocate their spending annually, however little is known about the timings of spending throughout the year. Attention to detail in agricultural production practice has been previously cited as a key profitability driver (Wilson (2014)), and exploring the managerial drivers of performance with an extreme value theory approach represents a potentially fruitful area of research work. The Farm Business Survey is undertaken annually; if we are to investigate the precision of farming practices, then data would need to be collected on the farmers' technical and business management decisions and practices each year.

5 | Summary

This thesis aimed to explore novel uses of the datasets from the Farm Business Survey and the UK Met Office to identify key farming practices which are most associated with high yields.

Sections 2 and 3 used data from the Farm Business Survey and UK Met Office to quantify the impact of important agronomic and climatic variables for modelling yield to inform stakeholders on how they can change their farming practices accordingly. Section 2 used statistical modelling techniques already frequently used in agronomical studies to provide insights into this dataset. The linear model with the smallest mean squared prediction error gave the 3 most important variables from our dataset to model yield as organic status, crop protection and rainfall in June. To allow for limitations in frequentist post-selection inference, Section 3 used a Bayesian framework for simultaneous model selection and inference to finding the 3 most important variables to be organic status, crop protection and machinery and reduced the importance of climatic variables. Both of the frequentist and Bayesian approaches show overall crop protection, and organic versus conventional status, to have a strong influence on winter wheat yields, and climatic variables indicative of environmental conditions during the wheat production cycle.

The models in sections 2 and 3 predict average yields accurately according to their squared prediction errors and posterior predictive distributions respectively. However, these models consistently underestimate yields achieved for top-performing farms. All models proposed in sections 2 and 3 suggest pesticide use increases yield. This confirms what has been found in past studies, however this analysis does not confirm whether a reduction in pesticide use will not significantly reduce crop yields.

The second of our contexts, Section 4 performed an extreme value analysis to model high yields, estimate a maximum attainable yield, and assess whether this maximum changes under different scenarios. One such scenario found maximum attainable yield is improved if spending on fertiliser and crop protection is increased for low spenders, but there is no significant evidence spending a larger amount on fertiliser and crop protection improves maximal yields otherwise. Using the macro-regions west England and Wales, north England and east England as a proxy for climate from the previous sections, there is also no difference in maximum attainable yield between these regions. Performing an extreme value analysis on net margins instead also suggests there is no difference in expenditure allocation between the top-performing farms and average farms.

From this analysis, we are still unable to distinguish between farms achieving large yields and those achieving average yields using the data from the Farm Business Survey. A potential area of future research would be to incorporate climate data in an extreme value framework, taking account of any extreme weather events occurring during the wheat production cycle and the timing of the events.

All of our analyses have been restricted to the variables available from the Farm Business Survey and the UK Met Office. Soil quality and components is also known to influence yield. Another avenue of future work would be to look at the interaction between soil quality, extreme weather events and farming practices to achieve larger yields. Future research would also benefit from making use of data on a smaller time scale, with more precise measurements. Currently, the Farm Business Survey collects limited annual data on farming practices. Specifically this does not capture precise farming practices during the wheat production cycle and only gives total expenditure each year. One such example being the amount of crop protection applied at specific stages of the wheat production process. An exception is the yields attained which will only be known at the end of the harvest year. The same criticism holds for the UK Met Office data taken to be monthly averages. This will fail to account for extreme fluctuations in weather conditions. A final criticism of the agronomic variables in the Farm Business Survey is they are recorded in financial metrics. Therefore, a larger expenditure may not necessarily translate to a larger application, but could reflect quality of the product instead, or differential product pricing

across businesses. Further studies would benefit from explicitly measuring quantities.

Finally, all 3 sections assumed each agronomic and climatic factors influenced yield linearly. This is not realistic due to the complex nature of climate models in the literature, however our analyses favoured simplicity over accuracy for ease of communication to stakeholders.

A | Post-selection inference using conditional polyhedron

Recent literature has conducted inference for the linear regression coefficients after variable selection by characterising the selection events as a polyhedral set (Tibshirani et al. (2016), Lee et al. (2016)) and visualising $\boldsymbol{y} \in \mathbb{R}$ falling into this conditional polyhedron. This allows for conditional hypothesis tests to be constructed based on this polyhedron and will capture both the uncertainty associated with parameter estimation and the uncertainty associated with variable selection.

Sections 2.3, 2.3.2 and 2.5.2 already discussed the linear model, linear regression and variable selection procedures respectively hence this work will be omitted here. Here we shall only discuss forward selection since the polyhedral sets follow on from the procedure discussed in the main body of the thesis.

Assuming $\boldsymbol{X}$ has been standardised and $\boldsymbol{y}$ is centred, Section 2.5.2 finds the first and k^{th} steps in the forward selection procedure satisfy

$$\frac{s_1 \boldsymbol{X}_{j_1}^T \boldsymbol{y}}{||\boldsymbol{X}_{j_1}||_2} \geq \pm \frac{\boldsymbol{X}_j^T \boldsymbol{y}}{||\boldsymbol{X}_j||_2} \qquad \text{and} \qquad \frac{s_k \tilde{\boldsymbol{X}}_{j_k}^T \boldsymbol{r}}{||\tilde{\boldsymbol{X}}_{j_k}||_2} \geq \pm \frac{\tilde{\boldsymbol{X}}_j^T \boldsymbol{r}}{||\tilde{\boldsymbol{X}}_j||_2}$$

respectively, where variable j_1 is selected as the first variable to enter the model as a result of achieving the smallest residual sum of squares when regressed upon to model the response, and j_k is selected at the k^{th} step due to achieving the smallest residual sum of squares when regressing upon to model the residuals $\boldsymbol{r}$ (see Section 2.5.2). Tibshirani et al. (2016) found these events can be expressed in the form $\Gamma y \geq 0$, where each

inequality is appended as rows in a matrix with the previous inequalities, and therefore describes a polyhedron $\{y : \Gamma y \geq 0\}$. Figure A.1 provides an illustration of the polyhedra conditioned on when selecting 2 variables out of the 3 from a variable selection algorithm.

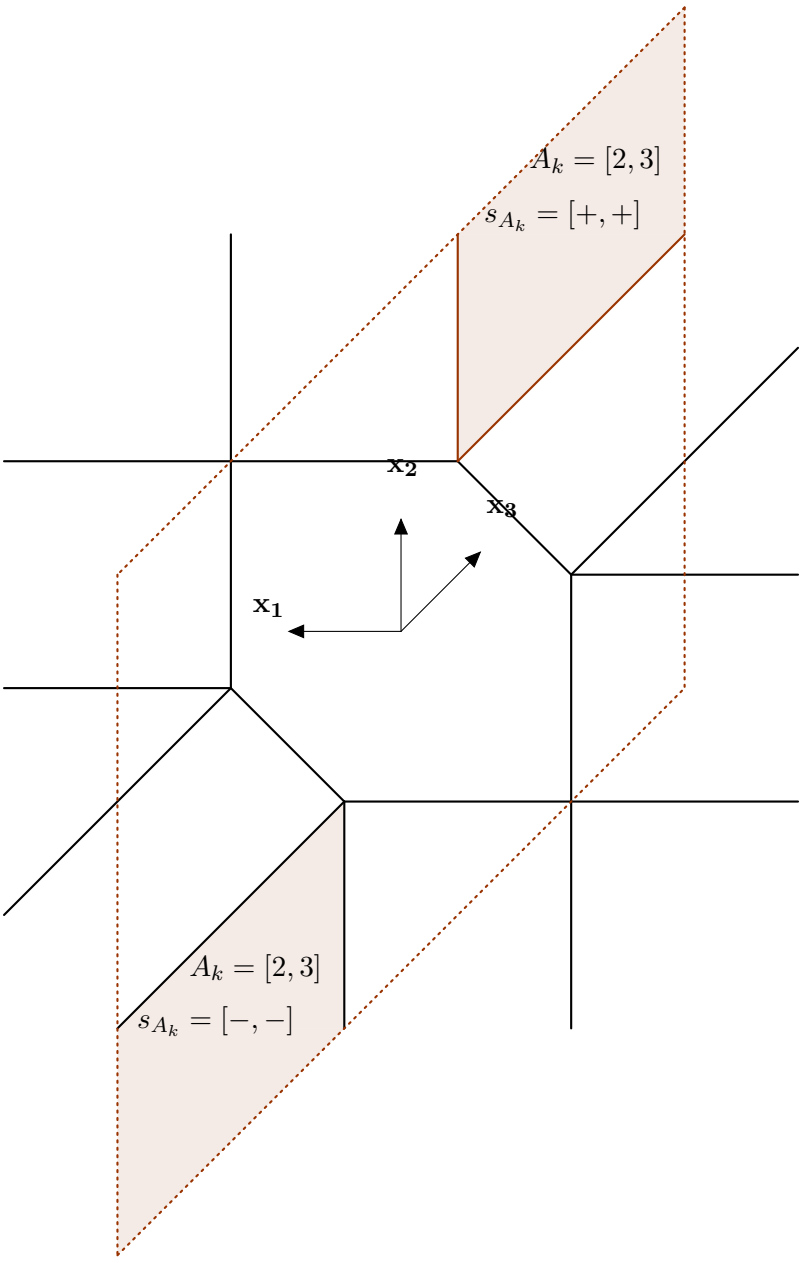


Figure A.1: Partitions of the space $\mathbb{R}^2$ into polyhedra when selecting 2 variables ($\boldsymbol{x}_2$,$\boldsymbol{x}_3$) out of the 3 ($\boldsymbol{x}_1$,$\boldsymbol{x}_2$,$\boldsymbol{x}_3$) from a variable selection algorithm. s_{A_k} denotes to signs of the coefficients selected (Lee et al. (2016)). The space in which the polyhedra are partitioned is visualised by the red dotted parallelogram.

Tibshirani et al. (2016) defines the boundaries of this conditional polyhedron in the direction of the k^{th} standard basis vector $\boldsymbol{e}_k$ to satisfy

$$\begin{aligned}\nu^{\text{lo}} &= \max_{j:(\Gamma\nu)_j>0} -(\Gamma y)_j \cdot ||\nu||_2^2/(\Gamma\nu)_j + \nu^T y, \\ \nu^{\text{up}} &= \max_{j:(\Gamma\nu)_j<0} -(\Gamma y)_j \cdot ||\nu||_2^2/(\Gamma\nu)_j + \nu^T y. \end{aligned} \tag{A.1}$$

where $\nu = \boldsymbol{X}_{A_k}(\boldsymbol{X}_{A_k}^T\boldsymbol{X}_{A_k})^{-1}\boldsymbol{e}_k$ is the k^{th} coefficient from regressing the selected variables A_k onto the response $\boldsymbol{y}$ and Γ in the conditioning polyhedron can easily found from the sets of inequalities above.

Now the bounds have been established to condition on, we look to the hypothesis test to conduct based on this polyhedron. Tibshirani et al. (2016) find, conditional on the polyhedral set, to perform the hypothesis tests

$$H_0 : \nu^T\theta = 0 \qquad \text{against} \qquad H_1 : \nu^T\theta \neq 0,$$

for each standard basis vector $\boldsymbol{e}_k$, and hence for each individual coefficient β_k^*, where $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}^*$ is the set of true coefficients. Therefore the α level confidence intervals will satisfy, based on the first k variables being selected

$$\mathbb{P}\left(\nu^T\theta \in [\delta_{\alpha/2}, \delta_{1-\alpha/2}] \middle| \hat{A}_k(y) = A_k, \hat{s}_{A_k} = s_{A_k}\right) = 1-\alpha,$$

where, under H_0, the confidence bounds for β_k, $\delta_{\alpha/2}$ and $\delta_{1-\alpha/2}$, are defined to satisfy

$$\begin{aligned} 1 - F^{[\nu^{\text{lo}},\nu^{\text{up}}]}_{\delta_{\alpha/2},\sigma^2||\nu||_2^2}(\nu^T y) &= \alpha/2 \\ 1 - F^{[\nu^{\text{lo}},\nu^{\text{up}}]}_{\delta_{1-\alpha/2},\sigma^2||\nu||_2^2}(\nu^T y) &= 1-\alpha/2 \end{aligned} \tag{A.2}$$

where the truncated Gaussian distribution

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} \tag{A.3}$$

is a pivotal statistic, with mean μ, variance σ^2 and truncation bounds a and b. This

approach to post-selection inference is suitable if the observed statistic ν^T falls well within the conditional polyhedral set.

Tibshirani et al. (2016) states in their code for post-selection inference based on polyhedral selection that at least one of the conditional confidence bounds $\delta_{\alpha/2}$ and $\delta_{1-\alpha/2}$ will be $\pm\infty$ if the observed statistic $\nu^T y$ is close to one of the truncation bounds ν^{lo} or ν^{up}, hence fails to provide confidence in our coefficient estimates. From Equation A.1, this happens if Γ_y is close to zero or $\Gamma\nu$ is close to ∞. We shall focus on the former case here for where this method collapses.

Supposing the forward stepwise algorithm is at step k, if $\Gamma\boldsymbol{y}$ is close to zero then

$$\frac{s_k \tilde{\boldsymbol{X}}_{j_k}^T \boldsymbol{r}}{||\tilde{\boldsymbol{X}}_{j_k}||_2} \approx \pm \frac{\tilde{\boldsymbol{X}}_j^T \boldsymbol{r}}{||\tilde{\boldsymbol{X}}_j||_2},$$

for one of the remaining j, where the residual sum of squares are approximately equal for two variables, one to be included in the model, j_k and one to be excluded. When considering the path of a variable selection algorithm, this will occur when two variables enter the model in quick succession and the model selection algorithm is unstable, i.e. the ordering of the variables can quickly switch depending on the data used. This becomes a particular issue as the number of variables in the model increase, since it becomes more likely two variables will achieve approximately the same residual sum of squares once the residual decreases as more variables are included. Cross-validating the data may be a solution, as discussed in Section 2.8, however with an increased interest in high-dimensional datasets, this may still become an issue when dealing with a model containing many more variables than we have included in our work.

B | MCMC convergence

To ensure the samples taken from the posterior distribution are well mixed, Figure B.1 shows the 2000 sampled coefficients from each of the 4 chains. By selecting a small step size in the HMC algorithm and thinning the sample, the chains in Figures B.1 and B.3 appear to be well mixed. For the Gibbs sampling schemes in Figures B.2 and B.4 also appear well mixed and converged.

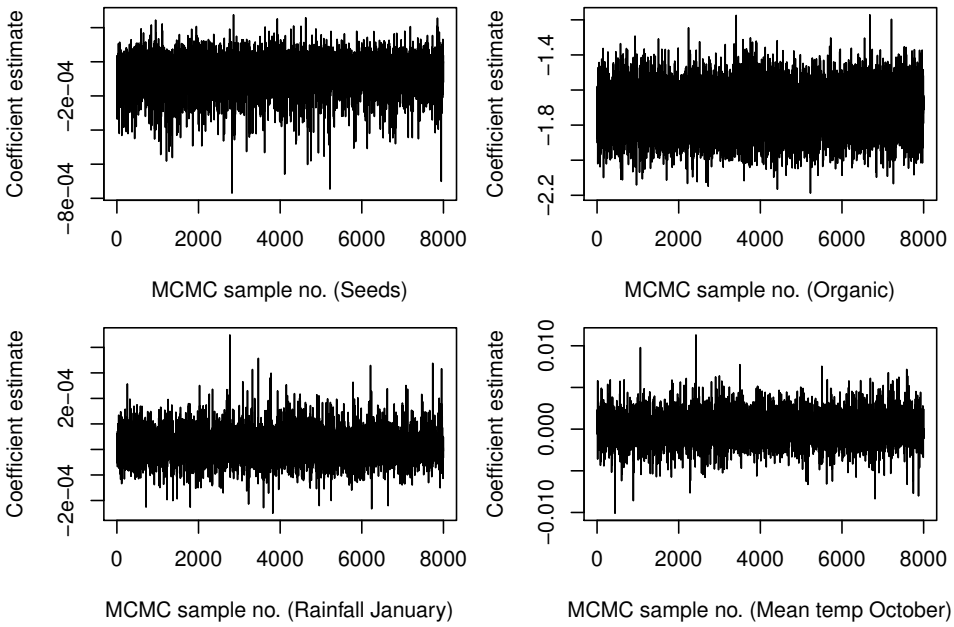


Figure B.1: Chains for the variables seeds, organic, rainfall in January and mean temperature in October for the Bayesian Lasso with fixed λ.

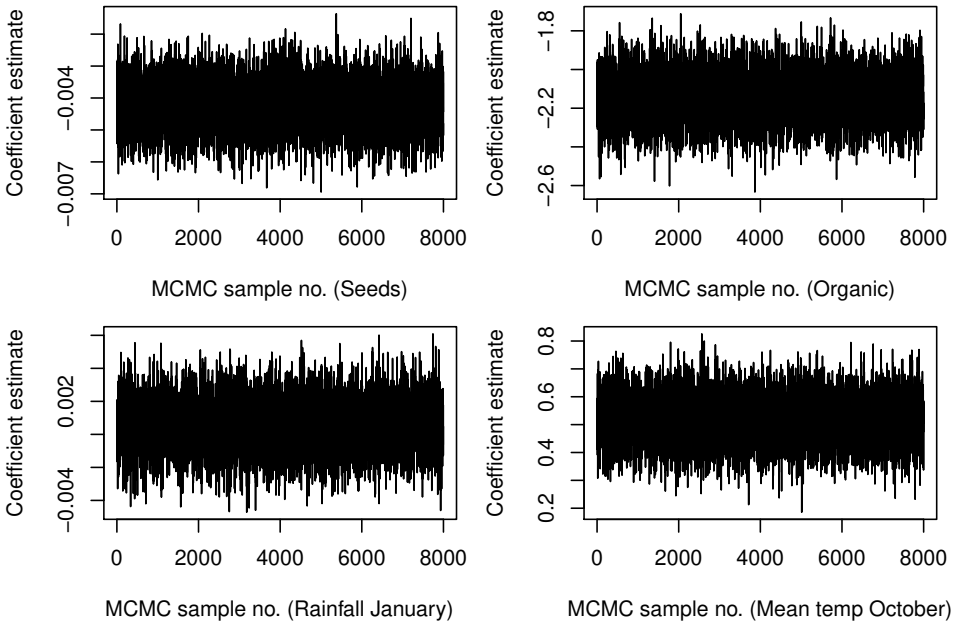


Figure B.2: Chains for the variables seeds, organic, rainfall in January and mean temperature in October for the Bayesian Lasso with prior λ.

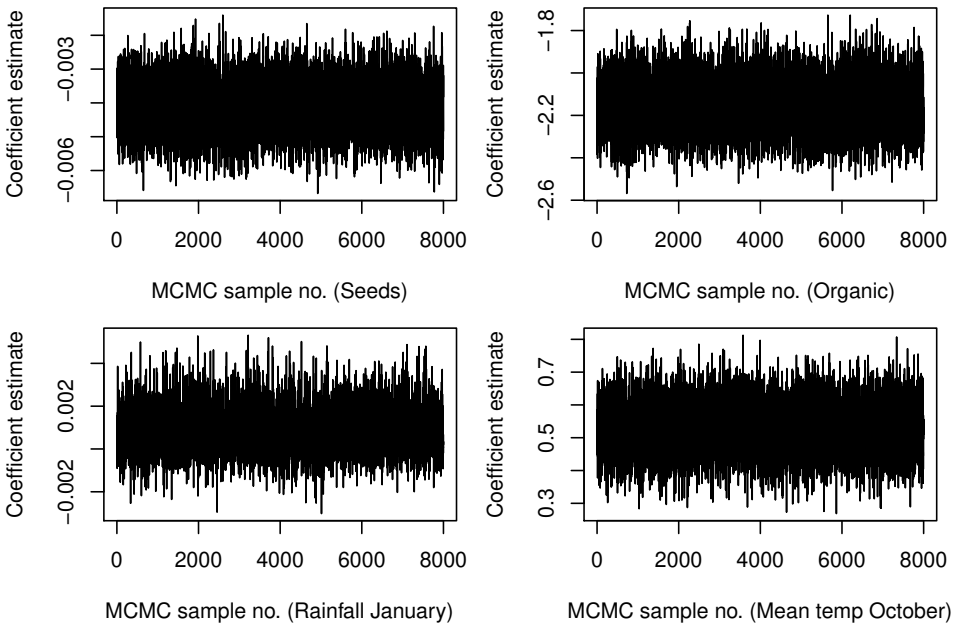


Figure B.3: Chains for the variables seeds, organic, rainfall in January and mean temperature in October for the horseshoe prior hierarchy with fixed λ.

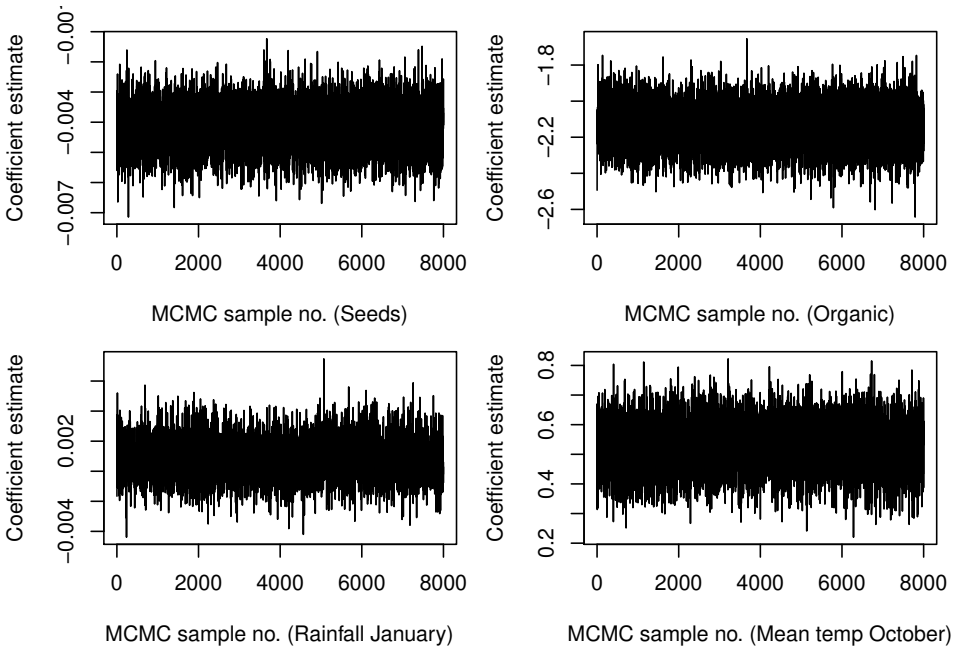


Figure B.4: Chains for the variables seeds, organic, rainfall in January and mean temperature in October for the horseshoe prior hierarchy with prior λ.

Bibliography

J.P. Absalom, S.D. Young, N.M.J. Crout, A. Sanchez, S.M. Wright, E. Smolders, A.F. Nisbet, and A.G. Gillett. Predicting the transfer of radiocaesium from organic soils to plants using soil characteristics. *Journal of Environmental Radioactivity*, 52:31–43, 2001.

D.F. Andrews and C.L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B*, 36(1):99–102, 1974.

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of Extremes*. Wiley, Chichester, England, 2004.

J. Besag and D. Higdon. Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B*, 61(4):691–746, 1999.

M.J. Betancourt and M. Girolami. Hamiltonian monte carlo for hierarchical models, 2013.

A. Bhadra, J. Datta, N.G. Polson, and B. Willard. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019.

G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26(2):211–252, 1964.

P. Brassley. Output and technical change in twentieth-century British agriculture. *The Agricultural History Review*, 48(1):60–84, 2000.

C.M. Carvalho and N.G. Polson. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

C.M. Carvalho, N.G. Polson, and J.G. Scott. Handling Sparsity via the Horseshoe. In 12^{th} *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

V. Chavez-Demoulin, P. Embrechts, and M. Hofert. An extreme value approach for modeling operational risk losses depending on covariates. *Journal of Risk and Insurance*, 83(3):735–776, 2016.

Y. Chen, Z. Zhang, F. Tao, P. Wang, and X. Wei. Spatio–temporal patterns of winter wheat yield potential and yield gap during the past three decades in North China. *Field Crops Research*, 206:11–20, 2017.

K. Cho, P. Falloon, J. Gornall, R. Betts, and R. Clark. Winter wheat yields in the UK: uncertainties in climate and management impacts. *Climate Research*, 54(1):49–68, 2012.

M.A. Clyde, J. Ghosh, and M.L. Littman. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20 (1):80–101, 2011.

S. G. Coles and D. Walshaw. Directional modelling of extreme wind speeds. *Journal of the Royal Statistical Society: Series C*, 43(1):139–157, 1994.

S.G. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, 2001.

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall, 1974.

G.M. Cox, J.M. Gibbons, A.T.A. Wood, Craigon J., S.J. Ramsden, and Crout N.M.J. Towards the systematic simplification of mechanistic models. *Ecological modelling*, 198 (1-2):240–246, 2006.

N.M.J. Crout, D. Tarsitano, and A.T.A. Wood. Is my model too complex? evaluating model formulation using model reduction. *Environmental Modelling & Software*, 24: 1–7, 2009.

N.M.J. Crout, J. Craigon, G.M. Cox, Y. Jao, D. Tarsitano, A.T.A. Wood, and M. Semenov. An objective approach to model reduction: Application to the sirius wheat model. *Agricultural and Forest Meteorology*, 189-190(100):211–219, 2014.

C. A. Damalas. Understanding benefits and risks of pesticide use. *Scientific Research and Essays*, 4(10):945–949, 2009.

M. de Carvalho, S. Pereira, P. Pereira, and P. de Zea Bermudez. An extreme value bayesian lasso for the conditional left and right tails, 2021.

L. de Haan and A. Ferreira. *Extreme Value Theory*. Springer, New York, 2006.

A. J. W. de Wit, H. L. Boogaard, and C. A. van Diepen. Spatial resolution of precipitation and radiation: The effect on regional crop yield forecasts. *Agricultural and Forest Meteorology*, 135(1-4):156–168, 2005.

DEFRA. Farming Statistics – Final crop areas, yields, livestock populations and agricultural workforce at 1 June 2012, United Kingdom. Technical report, National Statistics, 2012.

DEFRA. Farming Statistics – Final crop areas, yields, livestock populations and agricultural workforce at June 2015 - United Kingdom. Technical report, National Statistics, 2015.

DEFRA. Farming Statistics – Final crop areas, yields, livestock populations and agricultural workforce at June 2017 - United Kingdom. Technical report, National Statistics, 2017.

A. L. M. Dekkers, J. H. J. Einmahl, and L. de Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 17(4):1833–1855, 1989.

H. Drees, A. Ferreira, and L. de Haan. On maximum likelihood estimation of the extreme value index. *Annals of Applied Probability*, 14(3):1179–1201, 2004.

E. F. Eastoe and J. A. Tawn. Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C*, 58(1):25–45, 2009.

D. J. Ecobichon. Pesticide use in developing countries. *Toxicology*, 160(1-3):27–33, 2001.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events*. Springer, Berlin-Heidelberg, 1997.

A. Forte, G. García-Donato, and M. Steel. Methods and tools for bayesian variable selection and model averaging in normal linear regression. *International Statistical Review*, 86(2):237–258, 2018.

T. Fretheim and G. Kristiansen. Commodity market risk from 1995 to 2013: an extreme value theory approach. *Applied Economics*, 47(26):2768–2782, 2015.

G. García-Donato and M.A. Martínez-Beneito. On Sampling Strategies in Bayesian Variable Selection Problems With Large Model Spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2013.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

E.I. George and R.E. McCulloch. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

J.M. Gibbons, A.T.A. Wood, J. Craigon, S.J. Ramsden, and N.M.J. Crout. Semi-automatic reduction and upscaling of large models: a farm management example. *Ecological Modelling*, 221:590–598, 2010.

D. L. Gobbett, Z. Hochman, H. Horan, J. Navarro Garcia, P. Grassini, and K. G. Cassman. Yield gap analysis of rainfed wheat demonstrates local to global relevance. *Journal of Agricultural Science*, 155(2):282–299, 2017.

X. Gong, S. Sriboonchitta, S. Rahman, and S. Kuson. Modeling Value at Risk of agricultural crops using Extreme Value Theory. *Advanced Science Letters*, 21(5): 1339–1343, 2015.

J. Guo, J. Gabry, B. Goodrich, and S. Weber. *rstan*, 2020. R package version 2.21.2.

J. Hao, A. Bathke, and J. Skees. Modeling the tail distribution and ratemaking: An application of Extreme Value Theory. In *American Agricultural Economics Association Annual Meeting*, 2005.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2008.

W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

E. Hawkins, T.E. Fricker, A.J. Challinor, A.T. Ferro, C.K. Ho, and T.M. Osborne. Increasing influence of heat stress on french maize yields from the 1960s to the 2030s. *Global Change Biology*, 19(3):937–947, 2013.

HM Revenue and Customs. Percentile points from 1 to 99 for total income before and after tax. available at https://www.gov.uk/government/statistics/percentile-points-from-1-to-99-for-total-income-before-and-after-tax, 2017.

J. D. Holmes and W. W. Moriarty. Application of the generalized Pareto distribution to extreme value analysis in wind engineering. *Journal of Wind Engineering and Industrial Aerodynamics*, 83(1-3):1–10, 1999.

J. Hüsler and D. Li. On testing extreme value conditions. *Extremes*, 9:69–86, 2006.

H. Ishwaran and J.S. Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

R. W. Katz. Extreme value theory for precipitation: sensitivity analysis for climate change. *Advances in Water Resources*, 23(2):133–139, 1999.

S. Knight, S. Kightley, I. Bingham, S. Hoad, B. Lang, H. Philpott, R. Stobart, J. Thomas, A. Barnes, and B. Ball. Desk study to evaluate contributory causes of the current 'yield plateau' in wheat and oilseed rape. Technical report, Agriculture and Horticulture Development Board, 2012.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

S. Landau, R.A.C Mitchell, V. Barnett, J.J. Colls, J Craigon, K.L. Moore, and R.W. Payne. Testing winter wheat simulation models' predictions against observed uk grain yields. *Agricultural and Forest Meteorology*, 89(2):85–99, 1998.

S. Landau, R.A.C Mitchell, V. Barnett, J.J. Colls, J Craigon, and R.W. Payne. A parsimonious, multiple-regression model of wheat yield response to environment. *Agricultural and Forest Meteorology*, 101(2–3):151–166, 2000.

J.D. Lee, D.L Sun, Y. Sun, and J.E. Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016.

D.B. Lobell and M.B. Burke. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452, 2010.

E. Makalic and D.F. Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2015.

E. Makalic and D.F. Schmidt. High-dimensional bayesian regularised regression with the bayesreg package, 2016.

V. Marimoutou, B. Raggad, and A. Trabelsi. Extreme Value Theory and Value at Risk: Application to oil market. *Energy Economics*, 31(4):519–530, 2009.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6): 1087–1092, 1953.

W. Mirschel, R. Wieland, K. Wenkel, C. Nendel, and C. Guddat. Yieldstat - a spatial yield model for agricultural crops. *European Journal and Agronomy*, 52(A):33–46, 2014.

E.G. Mitchell, N.M.J. Crout, P. Wilson, A.T.A. Wood, and G. Stupfler. Operating at the extreme: estimating the upper yield boundary of winter wheat production in commercial practice. *Royal Society Open Science*, 7(4):1–12, 2020.

T.J. Mitchell and J.J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

A. Montesinos-López, O.A. Montesinos-López, G. de los Campos, J. Crossa, J. Burgueño, and F.J. Luna-Vazquez. Bayesian functional regression as an alternative statistical analysis of high-throughput phenotyping data of modern agriculture. *Plant Methods*, 14(46):1–17, 2018.

R.M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5. Chapman & Hall, 2011.

L. Nkurunziza, C.A. Watson, I. Öborn, H.G. Smith, G. Bergkvist, and J. Bengtsson. Socio-ecological factors determine crop performance in agricultural systems. *Scientific Reports*, 10(4232), 2020.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

J. Piironen and A. Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051, 2017a.

J. Piironen and A. Vehtari. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. In *Proceedings of the 20th International Conference on Artifical Intelligence and Statistics (AISTATS)*, 2017b.

D. Pimentel, L. McLaughlin, A. Zepp, B. Lakitan, T. Kraus, P. Kleinman, F. Vancini, W. J. Roach, E. Graap, W. S. Keeton, and G. Selig. Environmental and economic effects of reducing pesticide use in agriculture. *Agriculture, Ecosystems and Environment*, 46 (1-4):273–288, 1993.

N.G. Polson and J.G. Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9:501–538, 2011.

J. Prior and M. Kendon. The uk winter of 2009/2010 compared with severe winters of the last 100 years. *Weather*, 66(1):4–10, 2011.

B. Qian, R. de Jong, R. Warren, A. Chipanshi, and H. Hill. Statistical spring wheat yield forecasting for the Canadian prairie provinces. *Agricultural and Forest Meteorology*, 149(6-7):1022–1031, 2009.

M. A. Reader, C. Revoredo-Giha, R. J. Lawrence, I. D. Hodge, and B. Lang. Farmers' spending on variable inputs tends to maximise crop yields, not profit. *International Journal of Agricultural Management*, 7(1):1–11, 2018.

S. I. Resnick. *Heavy-Tail Phenomena*. Springer, 2007.

E.J. Ritchie. Analysis, development and application of wheat models of differing complexity at the farm and field scale, 2015.

H. Rootzén and D. Zholud. Human life is unlimited – but short. *Extremes*, 20(4):713–728, 2017.

D.F. Schmidt and E. Makalic. *bayesreg*, 2021. R package version 1.2.

S. D. Silvey. *Statistical Inference*. Chapman & Hall, 1970.

C. Steinkohl, R. A. Davis, and C. Klüppelberg. Extreme value analysis of multivariate high-frequency wind speed data. *Journal of Statistical Theory and Practice*, 7(1):73–94, 2013.

D. Tarsitano, S.D. Young, and N.M.J. Crout. Evaluating and reducing a model of radiocaesium soil-plant uptake. *Journal of Environmental Radioactivity*, 102:262–269, 2011.

M. Thomas, M. Lemaitre, M. L. Wilson, C. Viboud, Y. Yordanov, H. Wackernagel, and F. Carrat. Applications of extreme value theory in public health. *PLoS One*, 11(7): e0159312, 2016.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

R. Tibshirani, R. Tibshirani, J. Taylor, J. Loftus, S. Reid, and J. Markovic. *selectiveInference*, 2019. R package version 1.2.5.

R.J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

J. Tressou, A. Crepet, P. Bertail, M. H. Feinberg, and J. Ch. Leblanc. Probabilistic exposure assessment to food chemicals based on extreme value theory. Application to heavy metals from fish and sea products. *Food and Chemical Toxicology*, 42(8): 1349–1358, 2004.

M. K. van Ittersum, K. G. Cassman, P. Grassini, J. Wolf, P. Tittonell, and Z. Hochman. Yield gap analysis with local to global relevance – A review. *Field Crops Research*, 143:4–17, 2013.

J. van Wart, K. C. Kersebaum, S. Peng, M. Milner, and K. G. Cassman. Estimating crop yield potential at regional to national scales. *Field Crops Research*, 143:34–43, 2013.

E. Vanuytrecht, D. Raes, P. Steduto, T.C. Hsiao, E. Fereres, L.K. Heng, M.G. Vila, and P.M. Moreno. Aquacrop: Fao's crop water productivity and yield response model. *Environmental Modelling & Software*, 62:351–360, 2014.

D. Walshaw and C. W. Anderson. A model for extreme wind gusts. *Journal of the Royal Statistical Society: Series C*, 49(4):499–508, 2000.

B.P. Walter and M. Heimann. A process-based, climate-sensitive model to derive methane emissions from natural wetlands: application to five wetland sites, sensitivity to model parameteres, and climate. *Global Biogeochemical Cycles*, 14:745–765, 2000.

M.P. Wand, J.T. Ormerod, Padoan S.A., and R. Frühwirth. Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900, 2011.

P. Wilson. Farmer characteristics associated with improved and high farm business performance. *International Journal of Agricultural Management*, 3(4):191–199, 2014.

P. Wilson, D. Hadley, S. Ramsden, and I. Kaltsas. Measuring and Explaining Technical Efficiency in UK Potato Production. *Journal of Agricultural Economics*, 49(3):294–305, 1998.

P. Wilson, D. Hadley, and C. Asby. The influence of management characteristics on the technical efficiency of wheat farmers in eastern England. *Agricultural Economics*, 24 (3):329–338, 2001.

C. Zhou. Existence and consistency of the maximum likelihood estimator for the extreme value index. *Journal of Multivariate Analysis*, 100(4):794–815, 2009.