Incorporating Uncertainty in Sensor Data into Bayesian Grape Bunch Growth Models

Marina Chen

Under the supervision of Prof. Elena Moltchanova and Dr. Daniel Gerhard

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistics

School of Mathematics and Statistics,
University of Canterbury,
2021

Abstract

Accurate and timely predictions of grape yield are required by the wine industry for logistics planning, crop management, and wine marketing strategies. The Grape Yield Analyser project is an interdisciplinary collaboration aiming to predict grape yield in a timely and efficient manner. A Bayesian growth model assuming a double sigmoidal curve has been developed by **ellis** to predict grape yield. The model requires measurements of grape bunch mass at different times during the growing season. Such measurements require substantial trained staff and are also time-consuming and destructive. Hence, there is increasing research into the use of sensors in the industry. Since most sensors do not directly measure the mass of grape bunches, it can be difficult to obtain precise measurements of the grape bunch weights.

In this thesis, we provide the modelling framework for incorporating sensor-based measurements into the existing growth model. We assume the sensors produce measurements of grape bunch mass with known uncertainties. We present models which can incorporate uncertainties in continuous response variables and produce accurate (unbiased) and precise (minimal variability) predictions. MCMC algorithms are provided to estimate the proposed models for the two situations: (i) when the uncertainty is reported in the form of a parametric distribution, and (ii) when the uncertainty is reported in the form of a sample of values representing a nonparametric distribution.

We use simulation studies to evaluate the resulting model. In the first situation, our Bayesian model which incorporates uncertainty assuming a normal error distribution can perform well when the uncertainty is smaller than 80% of the population variation. In the second situation, when we have a sample of values representing a nonparametric distribution instead of a precise measurement, a naïve analysis can still produce accurate and precise measurements, given that the sample mean is an unbiased estimator of the actual value and for a large enough sample size. The models in this thesis can be applied to regression

problems in other fields, such as chemistry, medicine and physics, when there are continuous response variables reported with uncertainties.

Acknowledgements

First and foremost, I would like to thank the Lord, my God and Saviour Jesus Christ. Without you I would not have been able to get to this point. Thank you for always providing for me and being with me in the joys and difficult moments of doing this thesis. You are my good Shepherd and truest friend. Words cannot express everything you have done for me.

Thank you to my supervisors Elena and Daniel. Thank you for teaching me how to do research, sharing your knowledge and experiences with me and training me to work independently. Thanks also for your understanding, encouragement, sense of humour, and for being a sounding board for my worries. It has been a privilege to learn from both of you.

To the staff in the School of Mathematics, thank you for your kindness, and the postgraduate students for the peer support and company, which really helped my mental well-being and made my load feel less heavy.

To Ian Platt, thank you for kindly and patiently explaining to me how the microwave sensor works and providing me with more information about the project. Thank you to Mike Trought and Linlin Yang, who collected the 2017/2018 grape bunch mass data used in this thesis.

Thank you to my parents for supporting me and sharing in my joys and pains the past year. Thank you to my brother Gordon for being a really good listener, proofreading my thesis and supporting me, particularly towards the end. Thank you to my dog Maddie-Mikayla. Thank you to my dear friends for their support, love and kindness. To my church family, thank you for the encouragement in the faith, care, prayers and understanding.

Finally, thank you to MBIE and New Zealand Winegrower for funding this research and to Lincoln Agritech Ltd for the opportunity to work on this project.

Contents

1 Introduction **1**

2 Background theory **5**

2.1 Bayesian inference 5

2.2 Measurement uncertainty 8

2.3 Simulation studies 9

2.4 Methods to deal with measurement error 12

3 Parametric error model **14**

3.1 Overview 14

3.2 Bayesian inference for a normal sample in the absence of measurement uncertainty 14

3.3 Bayesian model for a univariate normal sample of measurements with uncertainty 16

3.3.1 MCMC algorithm: Gibbs sampler 18

3.4 Simulation study: a sample of values with normal error 21

3.5 Simple linear regression 28

3.5.1 MCMC algorithm: Gibbs sampler 30

3.6 Simulation study: simple linear regression with normal error 32

3.7 Example: double sigmoidal growth model 34

3.8 Discussion 52

4 Nonparametric model **55**

4.1 Overview 55

4.2 A single data point ($n = 1$) 55

4.3 Multiple data points ($n > 1$) 58

4.3.1 MCMC algorithm: Gibbs sampler 60

4.4 Preliminary study 61

4.5 Simulation study: a sample of values with a nonparametric error distribution 63

4.6 Example: double sigmoidal growth model 69

4.7 Discussion 71

5 Conclusion and Discussion **77**

5.1 Summary 77

5.2 Discussion 78

5.2.1 Contributions 78

5.2.2 Implications 80

5.2.3 Limitations 82

5.2.4 Future directions 83

Bibliography **83**

Appendix A Gibbs sampler for a sample of values with a normal error in R 87

Appendix B Gibbs sampler for a simple linear regression with a normal error in R **90**

Appendix C Gibbs sampler for the Bayesian nonparametric model in R **94**

Appendix D Metropolis-Hastings algorithm for the double sigmoidal curve for measurements with uncertainty (normal error) in R **97**

List of Figures

3.1 Directed acyclic graph for the Bayesian model for a normal sample. 16

3.2 Directed acyclic graph of the assumed relationships for the Bayesian model for a univariate sample of values with measurement uncertainty. 17

3.3 Simulated data for the simple linear regression, where **y** are the actual values and **m** are the measurements with uncertainty. The uncertainty is 50% of the population variation. The red line is the true regression line. The line segments show the magnitude and direction of the error in the measurements. Data was generated following the model specified in Equation 3.17—3.22 with $n = 100,\ \alpha = 0,\ \beta = 1,\ \tau = 10^{-2}$ and $u_i = 5^{-2}$. 31

3.4 Simple linear regression comparison of the three models and the posterior predictive envelopes when the uncertainty is 50% of the population variation. The posterior predictive means are shown by the solid lines. The envelopes represent the 95% credible interval of the posterior predictive distributions. The settings used are $\alpha = 0,\ \beta = 2,\ \tau = 10^{-2}$ and $u = 5^{-2}$. 37

3.5 Simple linear regression estimate of the posterior predictive distribution at $x = 200$ for the three models and posterior predictive means and 95% credible intervals. The uncertainty is 50% of the population variance. The settings used are $\alpha = 0,\ \beta = 2,\ \tau = 10^{-2}$ and $u = 5^{-2}$. 38

3.6 Simple linear regression comparison of the three models and the posterior predictive envelopes when the uncertainty is the same size as the population variation. The posterior predictive means are shown by the solid lines. The envelopes represent the 95% credible interval of the posterior predictive distributions. The settings used are $\alpha = 0$, $\beta = 2$, $\tau = 10^{-2}$ and $u = 10^{-2}$. . 39

3.7 Simple linear regression estimate of the posterior predictive distribution at $x = 200$ for the three models and posterior predictive mean and 95% credible intervals. The uncertainty is the same size as the population variance. The settings used are $\alpha = 0$, $\beta = 2$, $\tau = 10^{-2}$ and $u = 10^{-2}$. 40

3.8 Scatterplot of the 2018 grape bunch mass data with actual values, $\mathbf{y}$ and simulated measurements with uncertainty $\mathbf{m}$ using a setting of $u = 0.1^{-2}$. Jitter has been applied to the points to reduce overplotting. The bunch masses are plotted on the log scale. 44

3.9 Scatterplot of the 2018 grape bunch mass data with actual values, $\mathbf{y}$, simulated measurements with uncertainty $\mathbf{m}$ using a setting of $u = 0.1^{-2}$ which is approximately 20% of the original values (based on 1.96*se, where se is 10%). The bunch masses have been back transformed to be on the original scale in grams. Jitter has been applied to the points to reduce overplotting. 45

3.10 Double sigmoidal curve comparison of the three models and the posterior predictive envelopes for simulated uncertainty of 10% of the original values. The solid lines are the means of the posterior predictive distributions. The envelopes show the 95% credible interval of the posterior predictive distributions. The grape bunch masses are plotted on the log scale. 46

3.11 Double sigmoidal curve comparison of the three models and the posterior predictive envelopes for simulated uncertainty of 10% of the original values. The solid lines are the means of the posterior predictive distributions. The envelopes show the 95% credible interval of the posterior predictive distributions. 47

3.12 Double sigmoidal curve comparison of the estimated posterior predictive distribution at day 120 for the three models. This was for simulated uncertainty of 20% of the original values. Below the densities plot are the posterior predictive means and 95% credible intervals. 48

3.13 Double sigmoidal curve comparison of the three models and the posterior predictive envelopes for simulated uncertainty of 50% of the actual values. This is approximately 2 times the actual values, **y** (based on 1.96*se, where se is 50%). The solid lines are the means of the posterior predictive distributions. The envelopes show the 95% credible interval of the posterior predictive distributions. The grape bunch masses are plotted on the log scale. 49

3.14 Double sigmoidal curve comparison of the three models and the posterior predictive envelopes for simulated uncertainty of 50% of the actual values. This is approximately 2 times the actual values, **y** (based on 1.96*se, where se is 50%). The solid lines are the means of the posterior predictive distributions. The envelopes show the 95% credible interval of the posterior predictive distributions. Bunch masses are on the original scale in grams (g). 50

3.15 Double sigmoidal curve comparison of the estimated posterior predictive distribution at day 120 for the three models. This was for simulated uncertainty of 100% of the original values. Below the densities plot are the posterior predictive means and 95% credible intervals. 51

4.1 Illustrating the simulated sample of values, **d** observed instead of a single precisely measured observation y as a (a) density plot (obtained using kernel density estimation in R) and (b) as a frequency histogram. The true value of y is represented by the vertical dashed line. Data is generated from $d_k \sim \mathrm{N}(y = 200, u = 5^{-2})$ for $k = 1, \ldots, K$ where $K = 100$. 56

4.2 Directed acyclic graph for the Bayesian nonparametric model. 59

4.3 Probability density function of the Gamma($\alpha = 1,600,\ \beta = 40,000$) prior. . 63

4.4 Illustrating the nonparametric distributions using boxplots and half violin plots. These are simulated observed samples. True values of $\mathbf{y}$ are shown by the red line segments. The sample means are shown by the green line segments. The true population mean is represented by the horizontal dashed line. The samples were generated following the data-generating mechanism given by Equation 4.11 and Equation 4.12. The settings used are: $\mu = 200$, $\tau = 5^{-2}$, $u_i = 2^{-2}$, $n = 200$ and $K = 100$. Only the first 10 samples were plotted because it would be difficult to display all $n = 200$ in a single plot. 64

4.5 Cumulative mean error of the posterior mean of the (a) mean, (b) precision and (c) variance using our model and a Gamma($\alpha = 0.01,\ \beta = 0.01$) prior for the precision. This was for the case of $n = 100$ and $K = 100$. 65

4.6 Cumulative mean error of the variance when a vague $\tau \sim \text{Gamma}(0.1, 0.1)$ prior is used. The true variance is 25. The cumulative mean error converges to approximately -4 all cases, including when $n = 200$ and $K = 10,000$. . . . 66

4.7 Cumulative mean error of the variance using an informative $\tau \sim \text{Gamma}(\alpha = 1,600,\ \beta = 40,000)$ prior with mean 4 and standard deviation 0.001. The cumulative mean error of the variance converges to approximately -0.08. . . . 67

4.8 Scatterplot of the 2018 grape bunch mass data with corresponding simulated nonparametric data (a sample of values for each data point) distribution displayed as boxplots. The uncertainty is set to $u_i = 0.1^{-2}$. The colours of the boxplots correspond to the original data points. The plot is only showing three out of 14 days of data because it is difficult to clearly display the simulated data for all days in the dataset. 72

4.9 Illustrating the distributions of the simulated data and displaying the 2018 grape bunch mass data for a single day (day 41). The simulated samples are displayed as half violin plots (kernel density estimation i.e. smoother applied). Uncertainty was set to $u_i = 0.1^{-2}$. The colours of the distributions representing each sample correspond to the colours of the original data points. Data is only displayed for a single day as it is difficult to clearly display the simulated data for all days in the dataset. 73

4.10 Double sigmoidal curve comparing the naive analysis and precise measurements. The 2018 bunch mass data are represented by the black points and the simulated data are represented by the blue points. The setting used for the uncertainty is $u_i = 0.5^{-2}$. The grape bunch masses are plotted on the log scale. 74

4.11 Double sigmoidal curve comparing the naive analysis and precise measurements fitted with 2018 bunch mass data represented by the black points and simulated samples representing nonparametric distributions are represented by the blue points. The uncertainty used here is $u_i = 0.5^{-2}$. 75

4.12 Double sigmoidal curve with simulated nonparametric data using the setting $u_i = 0.5^{-2}$ comparing naive analysis and precise measurements and their estimated posterior predictive distribution at day 120. Below the plot of the densities are the posterior predicted means and 95% credible intervals. 76

5.1 An illustration of the two different data-generating processes for the sample of values situation in Chapter 3 (left) and Chapter 4 (right). Both (a) and (b) have the same true values shown in red, however the measured data that is observed shown in blue differs between the two plots. In (a) the filled red circles are the true values and the blue circles are the measured (observed) values. In (b) the red lines are the true values, the densities show the distributions of the samples observed, the green lines represent the means of each sample. In both (a) and (b) the horizontal dashed line represents the true population mean. 79

List of Tables

3.1 Data structure for sample of measurements with stated uncertainty 17

3.2 Estimates of performance measures for the mean and variance with Monte Carlo standard errors are reported in parentheses and obtained using the R package `rsimsum`. This is for the case of $\mu = 200,\ \tau = 5^{-2}$ and $u_i = 2^{-2}$. . . 25

3.3 Estimates of performance measures for the mean of the posterior predictive distribution for different data-generating mechanisms. Monte Carlo standard errors are reported in parentheses and obtained using the R package `rsimsum`. 25

3.4 Performance of the parameter estimates α, β, and σ^2 for the Bayesian linear model with a comparison against the Bayesian naive analysis and the Bayesian linear model with precise measurements. The true values used are: $\alpha = 0, \beta = 1, \sigma^2 = 10^{-2}$ and $u_i = 5^{-2}$. 35

3.5 Comparing the performance of the estimated posterior predictive distribution at $x = 100$ of the three different models and for different data-generating processes. 36

3.6 Prior distributions used for the double sigmoidal growth models 41

4.1 Estimates of performance measures for the mean and variance for the three different approaches. The data-generating process was: $n = 200$ and $K = 100$. For the nonparametric model we compared the use of an informative prior for the precision Gamma($\alpha = 1,600,\ \beta = 40,000$) and a vague prior for the precision Gamma($\alpha = 0.01,\ \beta = 0.01$). Monte Carlo standard errors are reported in parentheses and obtained using the R package `rsimsum`. 68

4.2 Simulation study results comparing performance estimates of the posterior predictive distribution for the four different models: (1) a Bayesian nonparametric model with a vague prior for τ, (2) a Bayesian nonparametric model with an informative prior for τ, (3) a Bayesian naive analysis and (4) a Bayesian analysis with precise measurements. 69

4.3 Assessing the performance of the naive analysis, where we let $y_i = \overline{\mathbf{d}}_i$, considering data-generating mechanisms, K= 3, 20, 30 and 50 for estimates of the mean and variance. Monte Carlo standard errors are reported in parentheses and obtained using the R package `rsimsum`. The setting of $n = 200$ is used for the data-generating process. A total of 300 simulations were performed. . 70

List of symbols

Some commonly used symbols in this thesis are as follows:

$\mathbf{x}$	a vector of values for a single predictor variable
$\mathbf{y}$	a vector of the true values of the response variable
$\mathbf{m}$	a vector of the measured values of the response
$\mathbf{u}$	a vector of the reported values of the uncertainty in the measured response
n	the number of observations
K	the sample size
$\mathbf{d}$	the observed sample of values instead of a single precise measurement of the response
$\mathcal{D}$	the list of n samples observed instead of n precise measurements

Chapter 1

Introduction

Grape yield prediction is an essential part of the wine industry. Accurate and timely predictions of grape yield are required by both grape growers and winemakers. Predictions of grape yield at harvest time assist grape growers in crop management practices and logistic planning. Planning includes the number of workers, capacity of trucks and tank space required (Tan et al., 2019; Ellis et al., 2020). Excessive yield can slow the development of fruit and result in losses in revenue for grape growers as they are required to pre-sell their crops to winemakers (Martin et al., 2003). For winemakers, accurate predictions are required for planning, including the amount of machinery required and ordering supplies such as bottles, labels and packaging (Tan et al., 2019). Accurate yield estimation can also support wine marketing strategies (Liu et al., 2017). Ellis et al. (2020) have developed a modelling framework for a Bayesian double sigmoidal curve to model grape bunch growth in order to predict yield. The double sigmoidal curve is a nonlinear regression that models how the grape bunch mass changes over time. It requires measurements of individual grape bunch mass at different points of time during the growing season.

Traditional yield estimation procedures involve manually cutting off bunches and are destructive and labour-intensive. The in-field measurements are time-consuming, have a high manual labour cost and can be prone to error (Tan et al., 2019). Research and

development is currently being conducted into sensors that indirectly measure the mass of the grape bunches. The Precision Grape Yield Analyser project funded by the New Zealand Ministry of Business, Innovation and Employment (MBIE) from the 2016 Endeavour fund and New Zealand Winegrowers is a five-year research programme focussing on developing sensor-based equipment and grape yield forecasting models to automate and improve the prediction of grape yield. The project is lead by Lincoln Agritech, in collaboration with the University of Canterbury, Plant and Food Research, Lincoln University and CSIRO. It is a multi-disciplinary project, including viticulture experts, grape phenology experts, a statistical modelling team, a team developing an optical sensor and another team developing a microwave sensor. The microwave sensor team and optical sensor team are developing a range of sensors both with the aim of measuring grape bunch mass and counting the number of grape bunches to ultimately produce yield predictions.

The optical sensor being developed takes photographic images. Image processing techniques are then applied to the images to identify grape bunches. However, one of the biggest issues faced when optical sensors are used is the leaves can obscure the grape bunches during the growing season, making it difficult to identify them (Eccleston et al., 2019; Parr et al., 2020). A microwave sensor, more specifically a synthetic-aperture-radar (SAR), is being developed to detect grape bunches on vines (Eccleston et al., 2018). The microwave sensor can penetrate leaves to detect grape bunches by adjusting the wavelength of the sensor (Eccleston et al., 2018). There are challenges to measuring grape bunches using microwave sensors, as the grape bunches are embedded in a complex microwave scattering environment containing leaves, stems and supporting wires (Eccleston et al., 2018). The microwave sensor estimates the bunch mass by measuring a proxy of biomass, and their performance is carefully calibrated and assessed. However, there is noise from the leaves and other sources.

Since the sensors do not directly measure the grape bunch mass and we did not have clarity on the final form of the output from the sensor as they were still under development, we assume that the sensors produce measurements of individual grape bunch masses with known uncertainty. We also assume that the size of the uncertainty cannot be ignored. The

current regression model described in Ellis et al. (2020) is valid for precisely measured grape bunches. However, additional statistical methodology is required to incorporate the data with known uncertainty into the double sigmoidal growth curve.

The aim of this thesis is to develop a modelling framework for incorporating sensor data with the reported uncertainty into the Bayesian double sigmoidal grape growth model developed by Ellis et al. (2020) to predict grape yield at harvest time.

Having models to incorporate sensor data with reported uncertainty will help enable the integration of the sensors with the grape yield prediction model and will be a step towards the development of a tool for automating yield predictions for the wine industry. If the sensor can be successfully coupled with the grape yield prediction model while making accurate and precise predictions, it could mean that sensors can be relied solely upon in the future to provide data for predictions of yield. Automation of yield prediction can reduce the manual labour effort required, reduce costs and enable data collection without the destructive sampling of grapes. Furthermore, if additional sensor data can be included into the current grape yield prediction model, it could improve the accuracy and precision of predictions by providing more data on grape growth.

The algorithms developed here can be applied to any regression problems where a continuous response variable is measured with uncertainty. Such situations are not limited to viticulture. Data with uncertainties frequently arise in physics, chemistry and medicine.

An overview of what lies ahead

In this thesis, we will start by explaining why it is important to account for reported (or known) uncertainty in measurements. We will then look at incorporating this uncertainty when it is reported in the form of a parametric distribution in Chapter 3, and in the form of non-parametric distribution in Chapter 4. For each situation, we propose a Bayesian framework to incorporate the specific form of uncertainty into a regression model. We

conduct simulation studies to evaluate our models, and to compare our Bayesian models for incorporating uncertainty with a naive analysis where the uncertainty is ignored, and with their precisely measured counterparts, respectively. We illustrate our models for the double sigmoidal growth model. Finally, we conclude by discussing the advantages and drawbacks of the suggested methods, and outlining the scope of future work.

As a guide to what follows, we would like to highlight the two different assumed data-generating processes of the sensor in this thesis. The assumption made in Chapter 3 is that we are given measured values and their reported uncertainty. We assume that the measured data arise from a normal distribution with a mean equal to the actual value and precision equal to the uncertainty. The assumption made in Chapter 4 is that we observe a sample of values instead of a precise measurement of the actual value. In this situation, assume that there is no bias in the sensor and we treat the sample nonparametrically (we do not apply any distributional assumptions).

Chapter 2

Background theory

2.1 Bayesian inference

The model we propose to extend in this thesis has been developed within a Bayesian framework. Therefore, we begin with a brief outline of the basic Bayesian model and MCMC algorithms mentioned in this research.

In Bayesian inference, parameters have a distribution instead of a fixed value in frequentist statistics. It requires a likelihood model which describes how the data arises. It also requires a prior distribution for the parameters, which summarises the information about the parameters before the data is observed. In Bayesian inference, the likelihood model and the prior distribution can be combined using Bayes' theorem to get the posterior distribution:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int p(y \mid \theta)p(\theta)d\theta} \tag{2.1}$$

$$\propto p(y \mid \theta)p(\theta) \tag{2.2}$$

where $p(y \mid \theta)$ is the likelihood and $p(\theta)$ is the prior distribution. The denominator in Equation 2.1 is a normalising constant that ensures that the posterior distribution is a proper

probability distribution. In many situations, the integral in Equation 2.1 is high dimensional and difficult to evaluate analytically. Because of this, Markov chain Monte Carlo (MCMC) algorithms which are numerical algorithms to sample from the posterior distribution are often used. (Gelman et al., 2014).

For predictive inferences which are inferences made about an unknown observable (Gelman et al., 2014), we can use the posterior predictive distribution for a new (future observation), which is defined as follows

$$p(\tilde{y} \mid \mathbf{y}) = \int p(\tilde{y} \mid \theta) p(\theta \mid \mathbf{y}) d\theta$$

where $\tilde{y}$ is a new value of y.

Two popular MCMC algorithms are the Metropolis-Hastings algorithm and its special case, the Gibbs sampler. We briefly describe both algorithms below.

Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm was introduced by Hastings (1970). It is a very useful tool in Bayesian inference as it allows us to estimate the posterior distribution when only the unnormalised posterior density in Equation 2.2 is known. The algorithm proceeds as follows:

Step 1. Start with an arbitrary initial value $\theta = \theta_0$.

Step 2. Sample a proposed value θ^* from a proposal distribution $J(\theta^* \mid \theta)$ conditional on the current value.

Step 3. Evaluate the Metropolis-Hastings acceptance ratio:

$$R = \frac{p(y \mid \theta^*) p(\theta^*) / J(\theta^* \mid \theta)}{p(y \mid \theta) p(\theta) / J(\theta \mid \theta^*)}$$

.

Step 4. Accept the proposed value with probability $\min(R, 1)$.

Repeat steps 2, 3 and 4 until the sampling process converges to the desired joint distribution (Gelman et al., 2014).

Gibbs sampler

The Gibbs sampler was invented by Geman and Geman (1984). It requires the full conditional distributions (also called the full conditional posterior distributions). The full conditional distributions, are the posterior distribution for a block of parameters with all other parameters held constant. The full conditional distributions can be derived from the joint posterior distribution. If the full conditional distributions can be analytically derived, and we are directly sample from them, then the Gibbs sampler algorithm proceeds as follows

Step 1. Assign each component of $\theta = (\theta_1, \ldots, \theta_J)$ an arbitrary initial value.

Step 2. Alternately sample from the conditional posterior distribution of each component given not only the data, but all the other components of θ using the most recently sample values of all the other components.

Repeat step 2 until the sampling process converges to the desired joint distribution (Gelman et al., 2014).

The Gibbs sampler is more efficient than the Metropolis-Hastings algorithm because every draw is accepted.

Gibbs sampler and Metropolis-Hastings algorithm as building blocks

Often the Gibbs sampler and the Metropolis-Hastings algorithm is used in various combinations to sample from complicated distributions (Gelman et al., 2014). The Gibbs

sampler is one of the simplest MCMC algorithms and is the preferred choice for conditionally conjugate models when we can directly sample from each of the full conditional distributions (Gelman et al., 2014). The Metropolis-Hastings algorithm can be used for models that are not conditionally conjugate, which can mean greater flexibility in model specification (Gelman et al., 2014).

Posterior summaries

Once we obtain a sample as an estimate of the posterior distribution, for example, using MCMC algorithms, we can summarise the estimated posterior using sample statistics such as means and quantiles (Gelman et al., 2014; Ellis et al., 2020). In this thesis, we use the mean of the posterior distribution as a point estimate. For interval estimation, in Bayesian inference, the notion of credible intervals is used instead of confidence intervals. There are two types of credible intervals: the central posterior interval and the highest posterior density (HPD) interval. The central posterior interval corresponds "to the range of values above and below which lies exactly $100(\alpha/2)\%$ of the posterior probability" (Gelman et al., 2014, p. 33). The highest posterior density interval is the shortest interval that contains 95% probability (Christensen et al., 2011).

2.2 Measurement uncertainty

In metrology, the scientific study of measurements, "measurement is the process of determining the value of a physical quantity" in an experimental manner with the help called measuring instruments (Rabinovich, 2005, p. 1). A measurable quantity, also called a measurand, can be defined as a property of phenomena, bodies, or substances that can be described qualitatively and expressed quantitatively (Rabinovich, 2005). When measurements are undertaken for scientific purposes, there is always uncertainty in the measurements. Measurement uncertainty is defined as the expression of the statistical dispersion of the values

attributed to a measured quantity (JCGM, 2008). In this thesis, we also refer to this measurement uncertainty as stated uncertainty. In contrast, measurement error is defined as the difference between a measured value of a quantity and its true value (Rabinovich, 2005).

The standard uncertainty is defined as the uncertainty characterised numerically by the standard deviation (Kirkup and Frenkel, 2006). Kirkup and Frenkel (2006) explain that from this standard deviation, it is common practice to obtain a $\pm$ numerical value referred as the "expanded" uncertainty in the GUM (Guide to the Expression of Uncertainty in Measurement). This value describes the range of values that is very likely to include the true value of the measurand (Kirkup and Frenkel, 2006). The number following the $\pm$ is normally about twice the standard deviation of the measurand. This has some parallels with the frequentist 95% confidence interval (Kirkup and Frenkel, 2006). In contrast, a standard uncertainty should be stated without the $\pm$ symbol and without any sign, for example the standard uncertainty for the measured mass of a grape bunch could be $u = 20$ grams.

The uncertainty about an estimation or prediction can be characterised through either Bayesian credible intervals or frequentist estimates of confidence interval limits.

2.3 Simulation studies

Simulation studies are "computer experiments that involve generating data by pseudo-random sampling" (Morris et al., 2019). They are a useful tool for statistical research, particularly for evaluating new methods and comparing the performance of different methods (Morris et al., 2019). Simulation studies are used to obtain empirical results about the performance of statistical methods in certain scenarios as opposed to more general analytic results which may cover a wide range of scenarios but may not always be possible to obtain (Morris et al., 2019). One of the first steps after determining the aim of the simulation study is to determine the data-generating mechanism. Morris et al. (2019) explains that the data-generating mechanism "denotes how random numbers are used to generate a dataset". The settings used for the data-generating mechanism usually refer to the values set for parameters

in the data-generating model. Simulation studies can also use unrealistic or extreme settings for data-generating mechanisms which can allow us to identify settings where the method may fail (Morris et al., 2019).

Performance measures

When simulation studies are used to evaluate methods, they are "typically motivated by frequentist theory and use frequentist properties of methods, even if the methods are Bayesian" (Morris et al., 2019). Desirable properties of estimators used in this thesis include unbiasedness and efficiency (Morris et al., 2019). Performance measures are numerical quantities used to assess the performance of a method (Morris et al., 2019). Common performance measures include, bias, mean squared error (MSE), coverage, and credible interval length or confidence interval length. Since simulation studies are empirical experiments, the performance measures are estimated and are thus subject to error. Therefore, Morris et al. (2019) have the view that estimates of uncertainty should be presented and they suggest to do this by reporting the Monte Carlo standard error.

The performance measures used in this thesis to evaluate our models are described below with formulae from Morris et al. (2019). The conceptual estimand and its true value is denoted by θ. An estimand could be, for example a parameter of the data-generating model. In addition, n_{sim} is the number of simulations and $i = 1, \ldots, n_{sim}$ indexes the repetitions of the simulations, and $\hat{\theta}_i$ is the estimator obtained in simulation i (Morris et al., 2019).

Bias is defined as $E[\hat{\theta}] - \theta$, where $E[\hat{\theta}]$ is the expected value of the estimator and θ is the true value of the parameter. In a simulation study, the bias can be estimated by the mean error (ME) given by:

$$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_i - \theta.$$

The Monte Carlo standard error of the mean error can be obtained using

$$\sqrt{\frac{1}{n_{sim}(n_{sim}-1)}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i-\bar{\theta})^2}.$$

The mean squared error (MSE) is defined as $E[(\hat{\theta}-\theta)^2]$. It is a combination of the bias and variance. The estimate of the MSE is

$$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i-\theta)^2$$

The Monte Carlo standard error of the MSE can be obtained using

$$\sqrt{\frac{\sum_{i=1}^{n_{sim}}\left[(\theta_i-\theta)^2-\widehat{M}\right]^2}{n_{sim}(n_{sim}-1)}}. \tag{2.3}$$

The coverage of the estimator is defined as

$$Pr(\hat{\theta}_{low}\leq\theta\leq\hat{\theta}_{high}),$$

and can be estimated using

$$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}I(\hat{\theta}_{low,i}\leq\theta\leq\hat{\theta}_{high,i}),$$

where $\hat{\theta}_{low,i}$ and $\hat{\theta}_{high,i}$ are the estimated posterior 2.5% quantile and 97.5 % quantile for the parameter of interest in simulation i. The Monte Carlo standard error of the estimated coverage can be obtained by

$$\sqrt{\frac{\widehat{Coverage}\times(1-\widehat{Coverage})}{n_{sim}}}.$$

We will also look at the efficiency of the estimators. A more efficient estimator has a smaller variance. To assess the efficiency, we will look at the coverage rates of the 95% credible interval and the average lengths of the 95% credible interval. We will use 95% confidence intervals when a frequentist approach is used.

2.4 Methods to deal with measurement error

It was challenging to do a literature review on methods for dealing with uncertainty in response variables since uncertainty is specific to each field. However, there is a large body of literature in statistics in the field called measurement error and exposure uncertainty. Within this field are methods for dealing with measurement error and theories pertaining to the consequences of ignoring measurement error. The problems it deals with are commonly known as measurement error modelling or errors-in-variables (Carroll et al., 2006).

The focus in the field of measurement error is primarily on measurement error in predictor variables, sometimes referred to as exposure variables in epidemiology. Carroll et al. (2006) writes extensively about how to deal with measurement error in predictors for linear and nonlinear models using frequentist methods with a chapter dedicated to Bayesian methods. They include methods for dealing with response error, however their focus is on binary response variables for logistic regression, rather than a continuous response variable which is of interest in this thesis. They include a section on Bayesian methods for dealing with measurement error, however they do not describe how to deal with measurement error in the response variable. Carroll et al. (1995) state that generally, in the field of measurement error, there has been more focus on methods for dealing with measurement error in predictors because they are known to cause biases in estimated regression coefficients. Furthermore, (Abrevaya and Hausman (2004) state that "classical measurement error (that is, additive error uncorrelated with the covariates) in the dependent variable is generally ignored in regression analysis because it simply gets absorbed into the error residual". Gustafson (2003) describes Bayesian methods for dealing with measurement error, however his focus is on predictors measured with error. He explains that his reason for focusing on measurement error in explanatory variables instead of response is because there is no systematic bias in the regression coefficients and it only affects inferential uncertainty, but he does not go into further detail (Gustafson, 2003). However, in this thesis, we are concerned with inferential uncertainty because we are interested in prediction. Predictive inferences depend

on the estimated coefficients. If the estimated coefficients or parameters of the model have greater variance (or uncertainty) then this will mean that the predictions will have greater uncertainty. We are concerned with the effect that models that adjust for uncertainty have on prediction, instead of coefficients of the model, since grape growers and winemakers require predictions of grape yield.

Buonaccorsi (2010) and McElreath (2015) both describe how to deal with measurement error in continuous response variables. Buonaccorsi (2010) describes how to deal with uncertainty in the response using frequentist methods-weighted regression. McElreath (2020) presents Bayesian approaches to dealing with measurement error in the response. He treats the true values of the response, $\mathbf{y}$ as parameters, the same way that missing values are dealt with in a Bayesian framework. He demonstrates his model on an example dataset in Stan (a software package for Bayesian modelling) (Stan Development Team, 2015). Carroll et al. (1995) describes this approach for predictor variables.

Throughout this thesis, we refer to a model or estimation procedure where we ignore uncertainty as a naive analysis or naive estimation. This term comes from Gustafson (2003) and Buonaccorsi (2010).

Chapter 3

Parametric error model

3.1 Overview

In this chapter, we will introduce a standard Bayesian model for a normal likelihood with conjugate priors. We will then extend it to account for data reported in the form of measurements with uncertainty, first a simple one-sample problem, then to a simple regression models and then to the double sigmoidal growth model. We derive MCMC algorithms for the parameter estimation and demonstrate their performance. Using simulation studies, we compare our proposed approach with the naive approach, which ignores uncertainty and also with the situation where we have precise measurements. We discuss how the normal distributed response error can be generalised further to other parametric error distributions.

3.2 Bayesian inference for a normal sample in the absence of measurement uncertainty

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a random sample of n independent and identically distributed observations from a normal population, $\mathrm{N}(\mu, \tau)$, with mean, μ and precision, τ. The precision

is defined as the inverse of the variance, σ^2, that is $\tau = 1/\sigma^2$. Our goal is to estimate the underlying population parameters, μ and τ and make predictive inferences regarding y. The likelihood model is

$$y_i \mid \mu, \tau \sim \mathrm{N}(\mu, \tau). \tag{3.1}$$

with probability density function

$$p(y_i \mid \mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\Big(-\frac{\tau}{2}(y_i - \mu)^2 \Big).$$

If we assume independent priors

$$p(\mu, \tau) = p(\mu)p(\tau)$$

and utilise conditional conjugacy then we can derive the full conditional posterior distributions and use a Gibbs sampler. Thus, we specify a normal prior for the mean

$$\mu \sim \mathrm{N}(\mu_0,\ \tau_0) \tag{3.2}$$

and a gamma prior for the precision, with shape parameter α and rate parameter β

$$\tau \sim \mathrm{Gamma}(\alpha,\ \beta). \tag{3.3}$$

Using Bayes' theorem, the joint posterior distribution of μ and τ given $\mathbf{y}$ is

$$p(\mu, \tau \mid \mathbf{y}) \propto \prod_{i=1}^{n} p(y_i \mid \mu, \tau) p(\mu) p(\tau). \tag{3.4}$$

For the Gibbs sampler, the full conditional posterior distributions for μ and τ with conditionally conjugate priors are standard results. See, for example, Christensen et al. (2011) for derivations and more details (pp. 120-121). The full conditional posterior distribution for the mean, $p(\mu \mid \tau, \mathbf{y})$ is a normal distribution

$$\mu \mid \tau, \mathbf{y} \sim \mathrm{N}\left(\frac{\tau n \bar{y} + \tau_0 \mu_0}{n\tau + \tau_0},\ n\tau + \tau_0 \right). \tag{3.5}$$

The full conditional posterior distribution for the precision, $p(\tau \mid \mathbf{y}, \mu)$ is a gamma distribution

$$\tau \mid \mu, \mathbf{y} \sim \mathrm{Gamma}\left(\alpha + \frac{n}{2},\ \beta + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right). \tag{3.6}$$

The directed acyclic graph (DAG) for the model is shown in Figure 3.1. It follows the usual convention where a square represents fixed or observed quantities, for example our observed data. A circle represents the unknowns or parameters to be inferred. The arrows show that the value of the parameter located at the end of an arrow is assumed to depend on the value of the parameter at its beginning (Scaccia and Green, 2003).

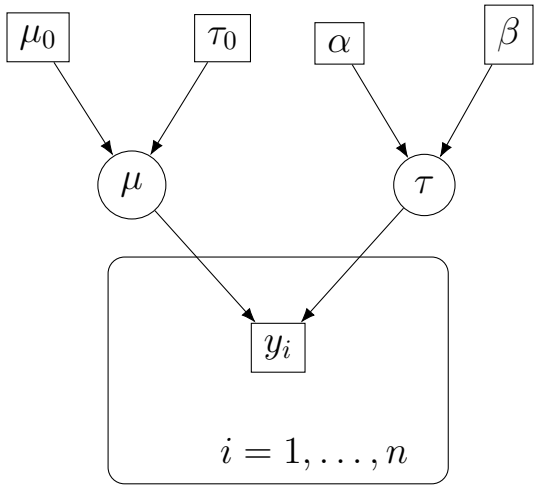


Figure 3.1. Directed acyclic graph for the Bayesian model for a normal sample.

3.3 Bayesian model for a univariate normal sample of measurements with uncertainty

Assume that instead of observing $\mathbf{y} = (y_1, \ldots, y_n)$ directly, we observe $\mathbf{m} = (m_1, \ldots, m_n)$, with respective stated uncertainty $\mathbf{u} = (u_1, \ldots, u_n)$. The generic set-up is described in Table 3.1. We assume that m_i is generated from a normal distribution with mean y_i and precision u_i as follows

$$m_i \mid y_i, u_i \sim \mathrm{N}(y_i,\ u_i) \quad \text{for all } i = 1, \ldots, n. \tag{3.7}$$

We can treat $\mathbf{y}$ as parameters, and $\mathbf{m}$ and $\mathbf{u}$ as data (see the DAG in Figure 3.2).

The probability density function of m_i given y_i and u_i is then

$$p(m_i \mid y_i, u_i) = \sqrt{\frac{u_i}{2\pi}} \exp\Big(-\frac{u_i}{2}(m_i - y_i)^2 \Big).$$

We refer to Equation 3.7 as a normal error distribution. The term error distribution is used by McElreath (2020). Equation 3.7 is also called a classical measurement error model where m_i is an unbiased measure of y_i (Carroll et al., 2006). This is a probabilistic model for how the measurements **m** are observed.

Table 3.1: Data structure for sample of measurements with stated uncertainty

Observation	True value (unobserved)	Measured value (observed)	Uncertainty
1	y_1	m_1	u_1
⋮	⋮	⋮	⋮
i	y_i	m_i	u_i
⋮	⋮	⋮	⋮
n	y_n	m_n	u_n

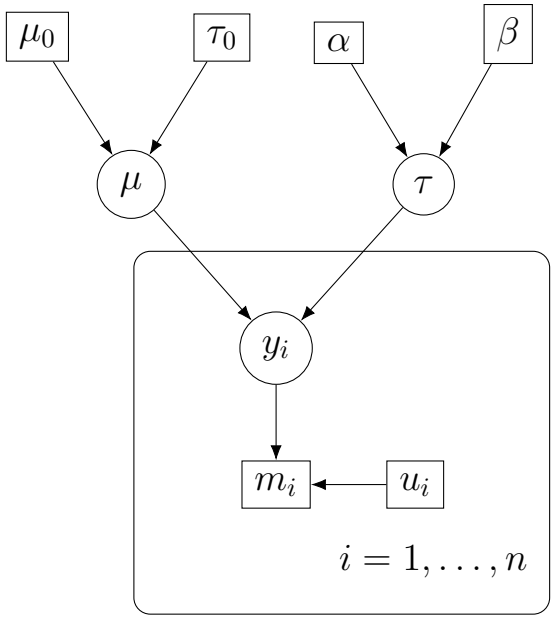


Figure 3.2. Directed acyclic graph of the assumed relationships for the Bayesian model for a univariate sample of values with measurement uncertainty.

The joint posterior distribution of the parameters μ, τ and $\mathbf{y}$ given $\mathbf{m}$ and $\mathbf{u}$ can be

derived using Bayes' theorem

$$p(\mu, \tau, \mathbf{y} \mid \mathbf{m}, \mathbf{u}) \propto p(\mathbf{m} \mid \mathbf{y}, \mathbf{u}) p(\mathbf{y} \mid \mu, \tau) p(\mu) p(\tau) \tag{3.8}$$

$$\propto \prod_{i=1}^{n} p(m_i \mid y_i, u_i) \prod_{i=1}^{n} p(y_i \mid \mu, \tau) p(\mu) p(\tau). \tag{3.9}$$

For our multiparameter model here, the parameters of interest are μ and τ. We are not as interested in the parameters $\mathbf{y}$ and treat them as "nuisance parameters" a term used by Gelman et al. (2014). The aim of our Bayesian model is to obtain marginal posterior distributions of μ and τ. Once we have the joint posterior distribution, we can obtain the marginal posterior distributions of μ and τ by integrating over parameters $\mathbf{y}$.

3.3.1 MCMC algorithm: Gibbs sampler

Because the marginal posterior distributions cannot be derived analytically we will use a numerical algorithm, the Gibbs sampler here, to approximate those distributions.

We already have the full conditional distributions for μ and τ from Section 2.1, therefore we only require the full conditional distribution for the parameters $\mathbf{y} = (y_1, \ldots, y_n)$. Since we assume that each y_i is independent of each other and the Gibbs sampler samples each parameter separately with all the other parameters held constant, we can sample each y_i one at a time. Thus, instead of deriving the full conditional distribution $p(\mathbf{y} \mid \mathbf{m}, \mathbf{u}, \mu, \tau)$, we can derive the full conditional distribution for each y_i of $p(y_i \mid m_i, u_i, \mu, \tau)$. The full conditional posterior distribution for y_i of $p(y_i \mid m_i, u_i, \mu, \tau)$ can be derived as follows

$$\begin{aligned}
p(y_i \mid m_i, u_i, \mu, \tau) &\propto p(m_i \mid y_i, u_i) p(y_i \mid \mu, \tau) p(\mu) p(\tau) \\
&\propto p(m_i \mid y_i, u_i) p(y_i \mid \mu, \tau) \\
&= \sqrt{\frac{u_i}{2\pi}} \exp\Big(-\frac{u_i}{2}(m_i - y_i)^2 \Big) \sqrt{\frac{\tau}{2\pi}} \exp\Big(-\frac{\tau}{2}(y_i - \mu)^2 \Big) \\
&\propto \exp\Big(-\frac{u_i}{2}(m_i - y_i)^2 \Big) \exp\Big(-\frac{\tau}{2}(y_i - \mu)^2 \Big) \\
&= \exp\Big(-\frac{u_i}{2}(m_i - y_i)^2 - \frac{\tau}{2}(y_i - \mu)^2 \Big)
\end{aligned}$$

$$
\begin{aligned}
&= \exp\left\{ -\frac{u_i}{2}\left[m_i^2 - 2m_i y_i + y_i^2\right] - \frac{\tau}{2}\left[y_i^2 - 2y_i\mu + \mu^2\right] \right\} \\
&= \exp\left\{ -\frac{u_i m_i^2}{2} + u_i m_i y_i - \frac{u_i y_i^2}{2} - \frac{\tau y_i^2}{2} + \tau y_i \mu - \frac{\tau \mu^2}{2} \right\} \\
&\propto \exp\left\{ u_i m_i y_i - \frac{u_i y_i^2}{2} - \frac{\tau y_i^2}{2} + \tau y_i \mu \right\} \\
&= \exp\left(-\left(\frac{u_i}{2} + \frac{\tau}{2}\right) y_i^2 + (u_i m_i + \tau\mu) y_i \right) \\
&= \exp\left\{ -\frac{u_i + \tau}{2}\left(y_i^2 - \frac{2(u_i m_i + \tau\mu)}{u_i + \tau} y_i \right) \right\} \\
&\propto \exp\left\{ -\frac{u_i + \tau}{2}\left(y_i - \frac{u_i m_i + \tau\mu}{u_i + \tau} \right)^2 \right\}.
\end{aligned}
$$

The result is a normal distribution

$$
y_i \mid m_i, u_i, \mu, \tau \sim \mathrm{N}\left(\frac{u_i m_i + \tau\mu}{u_i + \tau}, u_i + \tau \right) \quad \text{for all } i. \tag{3.10}
$$

Now we have the full conditional distributions we require for our Gibbs sampler.

Pseudo-code for the Gibbs sampler

Step 0. Set the arbitrary initial values for the parameters. E.g. $\mu^{(0)} = \overline{\mathbf{m}}$, $\tau^{(0)} = 1/s^2_{\mathbf{m}}$ and $y_i^{(0)} = \overline{\mathbf{m}}$ for $i = 1, \ldots, n$. Where $\overline{\mathbf{m}}$ is the sample mean and $s^2_{\mathbf{m}}$ is the sample variance of $\mathbf{m}$.

Step 1. Sample each y_i from a normal distribution with mean,

$$
\frac{u_i m_i + \tau\mu}{u_i + \tau}
$$

and precision, $u_i + \tau$, given the current values of μ and τ.

Step 2. Sample one value of μ from a normal distribution with mean,

$$
\frac{\tau n \bar{y} + \tau_0 \mu_0}{n\tau + \tau_0}
$$

and precision, $n\tau + \tau_0$, given the current values of $\mathbf{y}$ and τ.

Step 3. Sample one value of τ from a gamma distribution with shape parameter, $\alpha + \frac{n}{2}$ and rate parameter, $\beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2$, given the current values of μ and $\mathbf{y}$.

Repeat steps 1, 2 and 3 until convergence.

An implementation in R of this Gibbs sampler can be found in Appendix A.

A total of 5,000 iterations were used for the Gibbs sampler, with a burn-in of 500 iterations. Our final posterior sample had size 4,500. Convergence was visually assessed.

Naive analysis

For comparison, we will look at a naive analysis, which ignores the uncertainty in the measurements, using both a Bayesian and frequentist approach.

Model 2: Bayesian naive analysis

The Bayesian naive analysis ignores the uncertainty in measurements. Here we ignore the mechanism generating the uncertainty and treat $y_i = m_i$. A two step Gibbs sampler is used based on Equation 3.5 and Equation 3.6.

A total of 2,000 iterations were used for the Gibbs sampler, with a burn-in of 500 iterations. Our final posterior sample had size 1,500. Convergence was visually assessed.

Model 3: Frequentist naive analysis

For the frequentist naive analysis which ignores the uncertainty in measurements, the estimator for the population mean is the sample mean of $\mathbf{m}$:

$$\hat{\mu} = \overline{\mathbf{m}} = \frac{\sum_{i=1}^{n} m_i}{n} \tag{3.11}$$

with 95% confidence interval for the mean given by

$$\left(\hat{\mu} - 1.96\frac{s}{\sqrt{n}},\ \hat{\mu} + 1.96\frac{s}{\sqrt{n}}\right) \tag{3.12}$$

where s is the standard deviation of the sample, $\mathbf{m}$.

The estimate of the population variance is the sample variance of $\mathbf{m}$:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(m_i - \overline{m})^2}{n-1}. \tag{3.13}$$

Assuming that $\mathbf{m}$ comes from a normal population, the 95% confidence interval for the variance is given by

$$\left(\frac{(n-1)\hat{\sigma}^2}{\chi^2_{0.025,\,n-1}},\ \frac{(n-1)\hat{\sigma}^2}{\chi^2_{0.975,\,n-1}}\right) \tag{3.14}$$

where $\chi^2_{0.025,\,n-1}$ and $\chi^2_{0.975,\,n-1}$ are the 97.5% and 2.5% percentiles of the Chi-squared distribution, respectively with degrees of freedom, $n-1$.

The 95% prediction interval, that is for a new value of y, is given by

$$\left(\hat{\mu} - 1.96s\left(\frac{1}{n}+1\right),\ \hat{\mu} + 1.96s\left(\frac{1}{n}+1\right)\right). \tag{3.15}$$

3.4 Simulation study: a sample of values with normal error

To see how well the model works and assess the importance of accounting for measurement error, we will compare our three model: (1) a Bayesian model which incorporates uncertainty (described by the DAG in Figure 3.2), (2) the Bayesian naive analysis and (3) the frequentist naive analysis and we will also compare them against the situation where we have precise measurements. For this simulation study, we examined the special case all observation having the same size uncertainty, $u_i = u$, instead of having measurement specific uncertainty. For the Bayesian models, vague priors were chosen. This was to ensure that the data would provide the majority of the information, so that the results could be compared with frequentist results. The priors for the mean, μ and precision τ are specified as follows:

$$\mu \sim \mathrm{N}(\mu_0 = 0,\ \tau_0 = 10^{-4})$$

$$\tau \sim \text{Gamma}(\alpha = 0.001, \beta = 0.001).$$

Data are simulated on $n = 100$ observations, where each y_i is generated using

$$y_i \sim \text{N}(\mu\ \tau)$$

and each m_i is generated using

$$m_i \sim \text{N}(y_i,\ u_i).$$

We consider different values of μ, τ and u_i. For our cases, the precision τ and uncertainty, u_i have been written to emphasise the standard deviation. For example, $\tau = 10^{-2}$ means that the standard deviation is 10. We introduce some additional notation to describe our cases, where we let

$$\gamma = \frac{SD(\mathbf{m}|\mathbf{y})}{SD(\mathbf{y})} \tag{3.16}$$

following a similar convention to Gustafson (2003), where γ describes the magnitude of the uncertainty expressed as a fraction of the variability in the response variable $\mathbf{y}$ itself. For example $\gamma = 0.1$ can be interpreted as the uncertainty is 10% of the variability in $\mathbf{y}$. Or it can be viewed as yielding 10% imprecision in the measurement of $\mathbf{y}$ (Gustafson, 2003). We specifically considered six scenarios, starting with the case:

1. $\mu = 200,\ \tau = 5^{-2}$ and $u_i = 2^{-2}$ (In this case, $\gamma = 0.4$; the uncertainty is 40% of the variability in $\mathbf{y}$)

Then, we look at the effect of increasing the size of the uncertainty on prediction, so we chose a population standard deviation of 10 and uncertainty in terms of the standard deviation ranging from 2 to 15. We also sought to identify settings that would cause the method to fail.

2. Uncertainty is 20% of the variability in $\mathbf{y}$: $\mu = 200,\ \tau = 10^{-2}$ and $u_i = 2^{-2}$

3. Uncertainty is 50% of the variability in $\mathbf{y}$: $\mu = 200,\ \tau = 10^{-2}$ and $u_i = 5^{-2}$

4. Uncertainty is 80% of the variability in $\mathbf{y}$: $\mu = 200,\ \tau = 10^{-2}$ and $u_i = 8^{-2}$
5. Uncertainty is the same size as the variability in $\mathbf{y}$: $\mu = 200,\ \tau = 10^{-2}$ and $u_i = 10^{-2}$
6. Uncertainty is 1.5 times the variability in $\mathbf{y}$: $\mu = 200,\ \tau = 10^{-2}$ and $u_i = 15^{-2}$

We conduct 2,000 simulations. Performance is assessed by the bias (ME), MSE, coverage percentage of the 95% credible interval (or 95% confidence interval for the frequentist naive analysis) and average 95% credible interval length of the estimators. We include the Monte Carlo standard error to report the uncertainty in the performance measures. For the Bayesian models, to estimate the population mean, we use the mean of the marginal posterior distribution of the mean $p(\mu \mid \mathbf{m}, \mathbf{u})$ and we use the 95% credible interval to assess the coverage. Our estimate of the population variance is calculated as $1/\text{mean}(p(\tau \mid \mathbf{m}, \mathbf{u})$. To assess the coverage of the variance, we calculate the 95% credible interval as follows: the lower limit for the 95% credible is calculated as

$$1/\text{lower limit of the 95\% CI for } \tau$$

and the upper limit is calculated as

$$1/\text{upper limit of the 95\% CI for } \tau.$$

Data were simulated in R using the 32-bit Mersenne Twister for random number generation (R Core Team, 2020). For reproducibility, an input seed was used so that all models were applied to the same 2,000 datasets.

Table 3.2 compares the three models with the situation when precise measurements are made and their performance estimating the population mean, μ and variance, σ^2. All three models produce unbiased and efficient estimators of the population mean. All models exhibit bias close to 0, and coverage rates of the 95% credible interval or confidence interval close to 95%. Therefore, if we want to estimate the population mean, we can simply use a naive analysis and ignore the uncertainty in the data. When it comes to estimating the variance of the population, the Bayesian model for incorporating uncertainty gives unbiased

and the most efficient estimates of the population variance out of the three models. The bias is close to 0, MSE is low, and coverage rates of the 95% credible interval are close to 95%. In contrast, both the Bayesian and frequentist naive analysis, result in a positive bias and undercoverage in the estimate of the variance. From Morris et al. (2019) undercoverage is to be expected if the bias is not equal to 0.

Table 3.3 shows the performance of three models in terms of predictive inferences for different cases of μ, τ and u_i. All four models produce unbiased estimates in terms of prediction. The Bayesian and frequentist naive analyses have correct coverage when the uncertainty is 20% of the population variation. However, we observe overcoverage when the uncertainty is 40%, 50%, 80%, 100% and 1.5 times the population variation. The Bayesian model incorporating uncertainty performs well in terms of coverage for uncertainty that is up to 80% of the population variation. However, it starts to exhibit undercoverage when the uncertainty is the same as the population variation and also when it is 1.5 times the population variation.

The simulation study reveals that the Bayesian model which incorporates uncertainty produces unbiased and efficient predictions, up to a point where the uncertainty is 80% of the population variation. When the uncertainty is greater than 80% of the population variation, we start to observe undercoverage; this means that may convey a false sense of precision (the real precision is lower than what is declared).

The posterior predictive distribution depends on the posterior mean and posterior variance. A larger population variance will result in wider 95% credible intervals. If the estimated variance is larger, then the posterior prediction will have a greater variance and the credible interval of the posterior prediction distribution will be wider. For the grape yield analyser project, this means that if the estimated population variance is larger, it will mean that the predictions have greater uncertainty. We can see from Table 3.2 and Table 3.3 that the positive bias in the estimated variance from the Bayesian and frequentist naive analyses correspond to posterior predictive distributions with larger average 95% credible intervals, and larger average 95% prediction intervals, respectively.

Table 3.2: Estimates of performance measures for the mean and variance with Monte Carlo standard errors are reported in parentheses and obtained using the R package `rsimsum`. This is for the case of $\mu = 200,\ \tau = 5^{-2}$ and $u_i = 2^{-2}$.

Parameter	Performance measure	Bayesian model incorporating uncertainty	Bayesian naive analysis	Frequentist naive analysis	Precise measurements
μ	Bias	-0.015 (0.012)	-0.014 (0.012)	-0.009 (0.012)	-0.018 (0.011)
	MSE	0.284 (0.009)	0.284 (0.009)	0.284 (0.009)	0.251 (0.008)
	Coverage (%)	95.4 (0.5)	95.2 (0.5)	95.2 (0.5)	94.4 (0.5)
	Average 95% credible interval length	2.123	2.121	2.102	1.968
σ^2	Bias	-0.284 (0.092)	3.906 (0.092)	3.906 (0.092)	-0.142 (0.079)
	MSE	16.881 (0.534)	32.061 (1.005)	32.022 (1.002)	12.505 (0.406)
	Coverage (%)	95.8 (0.5)	80.2 (0.9)	80.6 (0.9)	94.9 (0.5)
	Average 95% credible interval length	16.657	16.668	16.725	14.334

Table 3.3: Estimates of performance measures for the mean of the posterior predictive distribution for different data-generating mechanisms. Monte Carlo standard errors are reported in parentheses and obtained using the R package `rsimsum`.

Performance measure	Bayesian model incorporating uncertainty	Bayesian naive analysis	Frequentist naive analysis	Precise measurements
Data-generating mechanism: $\mu = 200,\ \tau = 5^{-2}$ and $u_i = 2^{-2}$				
Bias	-0.0241 (0.013)	-0.169 (0.013)	-0.168 (0.012)	-0.074 (0.012)
MSE	24.851 (0.759)	25.333 (0.823)	24.523 (0.771)	25.048 (0.840)
Coverage (%)	95.8 (0.4)	96.2 (0.4)	96.3 (0.4)	95.0 (0.5)

Performance measure	Bayesian model incorporating uncertainty	Bayesian naive analysis	Frequentist naive analysis	Precise measurements
Average 95% credible interval length	19.761	21.291	21.233	19.744
Data-generating mechanism: $\mu = 200$, $\tau = 10^{-2}$ and $u_i = 2^{-2}$				
Bias	-0.062 (0.024)	-0.352 (0.024)	-0.341 (0.023)	-0.159 (0.024)
MSE	99.308 (0.017)	101.305 (3.287)	98.078 (3.085)	100.196 (3.360)
Coverage (%)	95.9 (0.4)	95.3 (0.5)	95.2 (0.5)	95.0 (0.5)
Average 95% credible interval length	39.484	40.292	40.184	39.488
Data-generating mechanism: $\mu = 200$, $\tau = 10^{-2}$ and $u_i = 5^{-2}$				
Bias	-0.060 (0.026)	-0.348 (0.026)	-0.334 (0.025)	-0.159 (0.024)
MSE	99.483 (3.040)	101.390 (3.297)	98.126 (3.083)	100.196 (3.360)
Coverage (%)	95.7 (0.5)	96.8 (0.4)	97.2 (0.4)	95.0 (0.5)
Average 95% credible interval length	39.521	44.217	44.098	39.488
Data-generating mechanism: $\mu = 200$, $\tau = 10^{-2}$ and $u_i = 8^{-2}$				
Bias	-0.061 (0.029)	-0.348 (0.030)	-0.326 (0.028)	-0.159 (0.024)
MSE	99.833 (3.048)	101.671 (3.312)	98.353 (3.087)	100.196 (3.360)
Coverage (%)	95.3 (0.5)	98.6 (0.3)	98.8 (0.3)	95.0 (0.5)
Average 95% credible interval length	39.446	50.688	50.551	39.488

Performance measure	Bayesian model incorporating uncertainty	Bayesian naive analysis	Frequentist naive analysis	Precise measurements
Data-generating mechanism: $\mu = 200$, $\tau = 10^{-2}$ and $u_i = 10^{-2}$				
Bias	-0.064 (0.032)	-0.350 (0.033)	-0.321 (0.031)	-0.159 (0.024)
MSE	100.149 (3.056)	101.970 (3.326)	98.604 (3.093)	100.196 (3.360)
Coverage (%)	94.4 (0.5)	99.2 (0.2)	99.4 (0.2)	95.0 (0.5)
Average 95% credible interval length	39.129	55.996	55.845	39.488
Data-generating mechanism: $\mu = 200$, $\tau = 10^{-2}$ and $u_i = 15^{-2}$				
Bias	-0.071 (0.042)	-0.362 (0.042)	-0.309 (0.040)	-0.159 (0.024)
MSE	101.569 (3.094)	103.107 (3.371)	99.578 (3.120)	100.196 (3.360)
Coverage (%)	85.4 (0.8)	100 (0.0)	99.9 (0.1)	95.0 (0.5)
Average 95% credible interval length	34.907	71.418	71.225	39.488

Note. For the frequentist naive analysis, instead of the coverage of the 95% credible interval for the posterior predictive distribution, we report the coverage of the 95% prediction interval. Similarly, instead of the average length of the 95% credible interval, report the average length of the 95% prediction interval.

3.5 Simple linear regression

We now extend our model for a sample of values measured with uncertainty to the situation of a simple linear regression model. The full Bayesian model specification is

$$y_i \sim \mathrm{N}(\mu_i,\ \tau) \tag{3.17}$$

$$\mu_i = \alpha + \beta x_i \tag{3.18}$$

$$m_i \mid y_i, u_i \sim \mathrm{N}(y_i, u_i) \tag{3.19}$$

$$\alpha \sim \mathrm{N}(\mu_\alpha, \tau_\alpha) \tag{3.20}$$

$$\beta \sim \mathrm{N}(\mu_\beta, \tau_\beta) \tag{3.21}$$

$$\tau \sim \mathrm{Gamma}(a,\ b) \tag{3.22}$$

where α is the intercept, β is the slope and τ is the precision (describing the error about the line). We assume a normal error distribution in Equation 3.19 for our measurements where each measurement m_i is generated from a normal distribution with mean y_i and uncertainty defined by the precision of u_i.

We choose conditionally conjugate priors for convenience so that we can use a Gibbs sampler and analytically derive the full conditional distributions of α, β and τ. We also assume independent priors as follows,

$$p(\alpha, \beta, \tau) = p(\alpha)p(\beta)p(\tau).$$

Therefore, the joint posterior distribution of α, β, and $\mathbf{y}$ can be written using Bayes' theorem as

$$p(\alpha, \beta, \tau, \mathbf{y} \mid \mathbf{m}, \mathbf{u}, \mathbf{x}) \propto p(\mathbf{m} \mid \mathbf{y}, \mathbf{u})p(\mathbf{y} \mid \mathbf{x}, \alpha, \beta, \tau)p(\alpha)p(\beta)p(\tau) \tag{3.23}$$

$$\propto \prod_{i=1}^{n} p(m_i \mid y_i, u_i) \prod_{i=1}^{n} p(y_i \mid x_i, \alpha, \beta, \tau)p(\alpha)p(\beta)p(\tau). \tag{3.24}$$

There is no analytical closed-form solution to the joint posterior distribution $p(\alpha, \beta, \tau, \mathbf{y} \mid \mathbf{m}, \mathbf{u}, \mathbf{x})$. Therefore, we will use a Gibbs sampler to obtain samples from the joint posterior. The full conditional distributions for α, β, τ and y_i are provided below:

Full conditional distribution for α:

$$\alpha \mid \beta, \tau, \mathbf{y}, \mathbf{x} \sim \mathrm{N}\Bigg(\frac{\tau \sum_{i=1}^{n} (y_i - \beta x_i) + \tau_\alpha \mu_\alpha}{n\tau + \tau_\alpha},\ n\tau + \tau_\alpha \Bigg) \tag{3.25}$$

Full conditional distribution for β:

$$\beta \mid \alpha, \tau, \mathbf{y}, \mathbf{x} \sim \mathrm{N}\Bigg(\frac{\tau \sum_{i=1}^{n} x_i(y_i - \alpha) + \tau_\beta \mu_\beta}{\tau \sum_{i=1}^{n} x_i^2 + \tau_\beta},\ \tau \sum_{i=1}^{n} x_i^2 + \tau_\beta \Bigg) \tag{3.26}$$

Full conditional distribution for τ:

$$\tau \mid \alpha, \beta, \mathbf{y}, \mathbf{x} \sim \mathrm{Gamma}\Bigg(a + \frac{n}{2},\ b + \frac{1}{2} \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2 \Bigg) \tag{3.27}$$

Full conditional distribution for y_i:

Starting with Equation 3.24, the derivation is as follows

$$\begin{aligned}
p(y_i \mid \alpha, \beta, \tau, m_i, u_i, x_i) &\propto p(m_i \mid y_i, u_i) p(y_i \mid x_i, \alpha, \beta, \tau) p(\alpha) p(\beta) p(\tau) \\
&\propto p(m_i \mid y_i, u_i) p(y_i \mid x_i, \alpha, \beta, \tau) \qquad (3.28) \\
&= \sqrt{\frac{u_i}{2\pi}} \exp\Big(-\frac{u_i}{2}(m_i - y_i)^2 \Big) \sqrt{\frac{\tau}{2\pi}} \exp\big(y_i - (\alpha + \beta x_i)\big)^2 \\
&\propto \exp\Big(-\frac{u_i}{2}(m_i - y_i)^2 \Big) \exp\big(y_i - (\alpha + \beta x_i)\big)^2 \\
&= \exp\Big(-\frac{u_i}{2}\big(m_i - 2m_i y_i + y_i^2\big) \Big) \exp\Big(-\frac{\tau}{2}\big(y_i^2 - 2y_i(\alpha + \beta x_i) + (\alpha + \beta x_i)^2\big) \Big) \\
&= \exp\Big\{ -\frac{u_i}{2} m_i^2 + u_i m_i y_i - \frac{u_i y_i^2}{2} - \frac{\tau y_i^2}{2} + \tau y_i(\alpha + \beta x_i) - \frac{\tau(\alpha + \beta x_i)^2}{2} \Big\} \\
&\propto \exp\Big\{ u_i m_i y_i - \frac{u_i y_i^2}{2} - \frac{\tau y_i^2}{2} + \tau y_i(\alpha + \beta x_i) \Big\} \\
&= \exp\Big(-\Big(\frac{u_i}{2} + \frac{\tau}{2}\Big) y_i^2 + \big(u_i m_i + \tau(\alpha + \beta x_i)\big) y_i \Big) \\
&= \exp\Big\{ -\Big(\frac{u_i + \tau}{2}\Big) \Big(y_i^2 - \frac{2\big(u_i m_i + \tau(\alpha + \beta x_i)\big)}{u_i + \tau} y_i \Big) \Big\} \\
&\propto \exp\Big\{ -\Big(\frac{u_i + \tau}{2}\Big) \Big(y_i - \frac{u_i m_i + \tau(\alpha + \beta x_i)}{u_i + \tau} \Big)^2 \Big\}
\end{aligned}$$

which is a normal distribution

$$y_i \mid \alpha, \beta, \tau, m_i, u_i, x_i \sim \mathrm{N}\Bigg(\frac{u_i m_i + \tau(\alpha + \beta x_i)}{u_i + \tau},\ u_i + \tau \Bigg) \quad \text{for all } i. \tag{3.29}$$

3.5.1 MCMC algorithm: Gibbs sampler

Pseudo-code for the Gibbs sampler

Step 0. Set the arbitrary initial values for the parameters, α, β, τ and $\mathbf{y}$.

Step 1. Sample each y_i from a normal distribution with mean,

$$\frac{u_i m_i + \tau(\alpha + \beta x_i)}{u_i + \tau}$$

and precision, $u_i + \tau$ given the current values of α, β and τ.

Step 2. Sample one value of α from a normal distribution with mean,

$$\frac{\tau \sum_{i=1}^{n}(y_i - \beta x_i) + \tau_\alpha \mu_\alpha}{n\tau + \tau_\alpha}$$

and precision, $n\tau + \tau_\alpha$ given the current values of β, τ and $\mathbf{y}$.

Step 3. Sample one value of β from a normal distribution with mean,

$$\frac{\tau \sum_{i=1}^{n} x_i(y_i - \alpha) + \tau_\beta \mu_\beta}{\tau \sum_{i=1}^{n} x_i^2 + \tau_\beta}$$

and precision, $\sum_{i=1}^{n} x_i^2 + \tau_\beta$, given the current values of α, τ and $\mathbf{y}$.

Step 4. Sample one value of τ from a gamma distribution with shape parameter, $a + \frac{n}{2}$ and rate parameter, $b + \frac{1}{2}\sum_{i=1}^{n}\big(y_i - (\alpha + \beta x_i)\big)^2$ given the current values of α, β and $\mathbf{y}$.

Repeat steps 1, 2, 3 and 4 until convergence.

An implementation in R of this Gibbs sampler can be found in Appendix B.

The number of iterations used for the Gibbs sampler was 5,000 with a burn-in period of 500 iterations. Convergence was visually assessed.

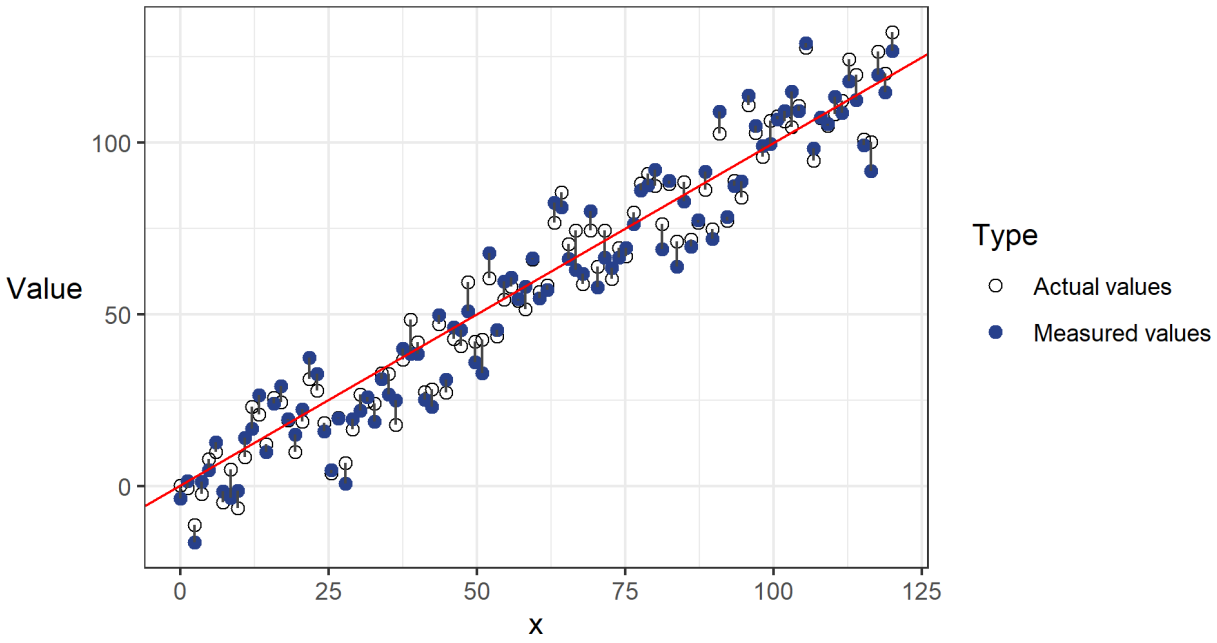


Figure 3.3. Simulated data for the simple linear regression, where $\mathbf{y}$ are the actual values and $\mathbf{m}$ are the measurements with uncertainty. The uncertainty is 50% of the population variation. The red line is the true regression line. The line segments show the magnitude and direction of the error in the measurements. Data was generated following the model specified in Equation 3.17—3.22 with $n = 100,\ \alpha = 0,\ \beta = 1,\ \tau = 10^{-2}$ and $u_i = 5^{-2}$.

3.6 Simulation study: simple linear regression with normal error

We conduct a simulation study to evaluate how well the model estimates the true population parameters, that is the intercept, slope and variance. Additionally we evaluate how well it performs in terms of predictive inferences. The data is generated according to the model following Equation 3.17—Equation 3.22. We chose uninformative priors for our model as follows

$$\alpha \sim \mathrm{N}(\mu_\alpha - 0, \tau_\alpha = 10^{-4}) \tag{3.30}$$

$$\beta \sim \mathrm{N}(\mu_\beta = 0, \tau_\beta = 10^{-4}) \tag{3.31}$$

$$\tau \sim \mathrm{Gamma}(a = 0.01,\ b = 0.01). \tag{3.32}$$

For our data-generating process, we used $n = 100$ observations and our $\mathbf{x}$ values were equally spaced from 0 to 100. Again we consider the same size known uncertainty for all observations i.e. $u_i = u$. We chose the following settings for our simulations:

1. $\alpha = 0$, $\beta = 1$, $\sigma^2 = 10^{-2}$ and $u_i = 5^{-2}$ (uncertainty is 50% of the error about the line)
2. $\alpha = 0$, $\beta = 1$, $\sigma^2 = 10^{-2}$ and $u_i = 8^{-2}$ (uncertainty is 80% of the error about the line)
3. $\alpha = 0$, $\beta = 1$, $\sigma^2 = 10^{-2}$ and $u_i = 10^{-2}$ (uncertainty is the same size as the error about the line)

A total of 2,000 simulations were performed. We report the bias (ME), MSE, coverage rates of the 95% credible intervals and the average 95% credible interval lengths. To estimate α and β, we use the posterior mean and 95% credible intervals of the marginal posterior distributions of α and β, respectively. To estimate the variance, we do this in the same way described in the simulation study for a sample of values measured with uncertainty in

Section 3.4. Data were simulated in R (R Core Team, 2020). For reproducibility, an input seed was used so that all models were applied to the same 2,000 datasets.

Table 3.4 shows that the Bayesian model which incorporates uncertainty estimates all population parameters well, under the setting where the uncertainty is 50% of the population variation. The estimates are unbiased for the intercept, slope and variance, with good coverage and are the most efficient estimators. The naive analysis gives unbiased estimates of the intercept, α and slope, β with correct coverage. However, when using the naive analysis there is a positive bias in the estimate of the variance and also undercoverage of the 95% credible interval.

Table 3.5 summarises the performance of the posterior predictive distribution at $x = 100$ for three models under different data-generating mechanisms. All three models produce unbiased predictions. The Bayesian model which incorporates uncertainty performs reasonably well when the uncertainty is 50% and 80% of the population variation. There may be slight undercoverage, but the average 95% credible length is very similar to the results from the Bayesian model with precise measurements. However, when the uncertainty is the same size as the population variation, the Bayesian model which incorporates uncertainty shows undercoverage, with a coverage percentage of 93.6% of the 95% credible interval. Figure 3.6 suggests there is undercoverage in the prediction, where the envelope for the Bayesian model incorporating uncertainty is more narrow than the Bayesian model using precise measurements. The naive analysis results in overcoverage of the 95% credible interval with a corresponding longer average 95% credible interval length under all settings in Table 3.5. This means that it produces a conservative prediction.

Figure 3.5 and Figure 3.7 show the estimated posterior predictive distribution at $x = 200$, and the posterior predicted mean and 95% credible interval for one simulation and for two different cases: when the uncertainty is 50% of the population variation and 100% of the population variation, respectively. In Figure 3.5 the Bayesian model which incorporates uncertainty has a shorter 95% credible interval for the posterior predictive distribution and is a similar length to the Bayesian model using precise measurements. The

naive analysis has a longer 95% credible interval for the prediction than the Bayesian model which incorporates uncertainty. In Figure 3.7, we observe the Bayesian model incorporating uncertainty has the shortest 95% credible interval for the prediction. The 95% credible interval for the Bayesian model which incorporates uncertainty is shorter than the 95% credible interval for the Bayesian model with precise measurements. This appears to be undercoverage which occurs for the Bayesian model which incorporates uncertainty from the results of our simulation studies in Table 3.5. The naive analysis results in a 95% credible interval for the prediction that is longer than the interval for the Bayesian model with precise measurements, which suggests that the 95% credible interval for the prediction exhibits overcoverage.

3.7 Example: double sigmoidal growth model

In this section, we extend our model for incorporating measurement uncertainty to the double sigmoidal growth curve. We use 2018 bunch mass data used in Ellis et al. (2020) collected by Mike Trought and Linlin Yang from Rowley Crescent in Marlborough, and simulate data with measurement uncertainty based on the real bunch masses in R (R Core Team, 2020). The data was collected over the period from December 2017 to March 2018. The response variable is the individual grape bunch mass and the predictor is the number of days since 1 December 2017. The mass is measured in grams (g). There were 14 days of data collected and a total of 448 observations. A scatterplot of the data is shown in Figure 3.9.

For reproducibility, we set the random number seed in R when generating the measurements with uncertainty for the double sigmoidal curve. We compare our model with a naive analysis and with the precisely measured (original data) in terms of prediction and analyse the results.

We extend previous work by Ellis et al. (2020), which assumes a double sigmoidal growth model for the grape bunch growth. The model contains a total of seven parameters;

Table 3.4: Performance of the parameter estimates α, β, and σ^2 for the Bayesian linear model with a comparison against the Bayesian naive analysis and the Bayesian linear model with precise measurements. The true values used are: $\alpha = 0, \beta = 1, \sigma^2 = 10^{-2}$ and $u_i = 5^{-2}$.

Parameter	Performance measure	Bayesian model incorporating uncertainty	Naive analysis	Precise measurements
α	Bias	0.031 (0.049)	0.028 (0.049)	0.015 (0.044)
	MSE	4.768 (0.149)	4.764 (0.148)	3.798 (0.120)
	Coverage (%)	95.0 (0.5)	95.4 (0.5)	95.1 (0.5)
	Average 95% credible interval length	8.723	8.754	7.818
β	Bias	-0.001 (0.001)	-0.001 (0.001)	-0.001 (0.001)
	MSE	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)
	Coverage (%)	94.9 (0.5)	94.9 (0.5)	95.1 (0.5)
	Average 95% credible interval length	0.126	0.126	0.113
σ^2	Bias	-1.631 (0.400)	24.659 (0.398)	-0.583 (0.318)
	MSE	321.237 (9.936)	924.999 (24.176)	203.015 (6.550)
	Coverage (%)	95.7 (0.5)	63.7 (1.1)	94.9 (0.5)
	Average 95% credible interval length	72.066	72.468	57.789

Table 3.5: Comparing the performance of the estimated posterior predictive distribution at $x = 100$ of the three different models and for different data-generating processes.

	Bayesian model incorporating uncertainty	Naive analysis	Precise measurements
Data-generating mechanism: $\alpha = 0$, $\beta = 1$, $\sigma^2 = 10^{-2}$ and $u_i = 5^{-2}$			
Bias	-0.008 (0.026)	-0.104 (0.051)	-0.116 (0.035)
MSE	104.331 (3.314)	105.249 (3.476)	97.335 (3.001)
Coverage (%)	94.4 (0.5)	97.3 (0.4)	95.3 (0.5)
Average 95% credible interval length	39.501	44.858	39.640
Data-generating mechanism: $\alpha = 0$, $\beta = 1$, $\sigma^2 = 10^{-2}$ and $u_i = 8^{-2}$			
Bias	0.000 (0.030)	-0.098 (0.058)	-0.116 (0.035)
MSE	104.749 (3.334)	106.799 (3.517)	97.335 (3.001)
Coverage (%)	94.3 (0.5)	98.7 (0.3)	95.3 (0.5)
Average 95% credible interval length	39.438	51.421	39.640
Data-generating mechanism: $\alpha = 0$, $\beta = 1$, $\sigma^2 = 10^{-2}$ and $u_i = 10^{-2}$			
Bias	0.005 (0.032)	-0.093 (0.064)	-0.116 (0.035)
MSE	105.088 (3.349)	108.239 (3.557)	97.335 (3.001)
Coverage (%)	93.6 (0.5)	99.4 (0.2)	95.3 (0.5)
Average 95% credible interval length	39.109	56.805	39.640

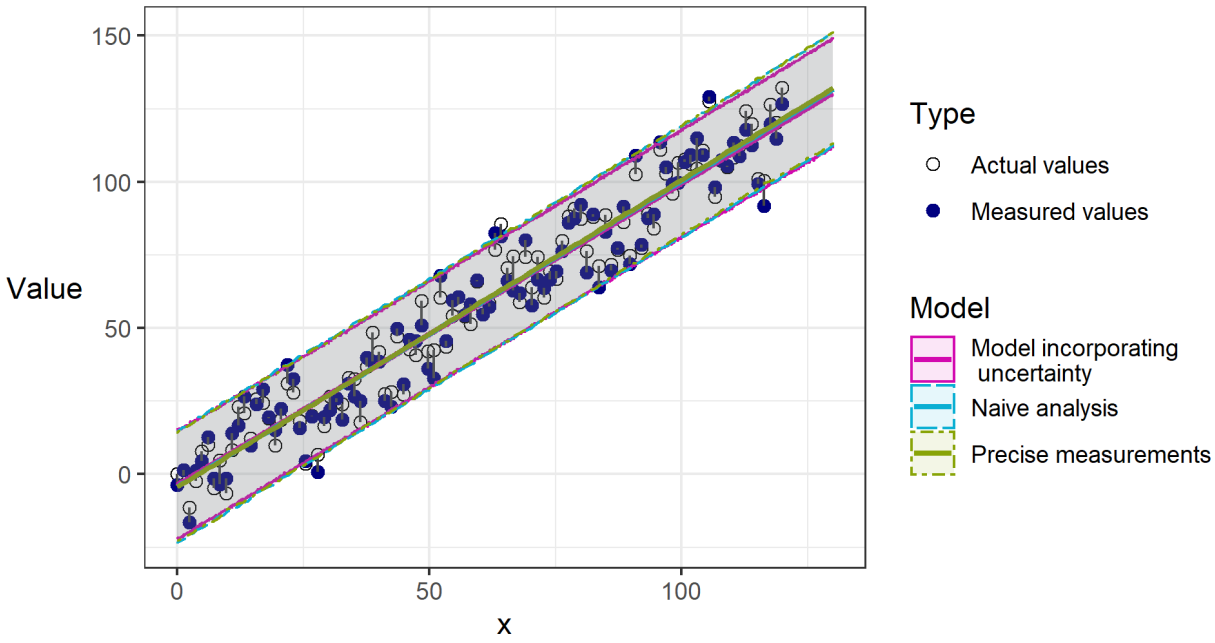


Figure 3.4. Simple linear regression comparison of the three models and the posterior predictive envelopes when the uncertainty is 50% of the population variation. The posterior predictive means are shown by the solid lines. The envelopes represent the 95% credible interval of the posterior predictive distributions. The settings used are $\alpha = 0$, $\beta = 2$, $\tau = 10^{-2}$ and $u = 5^{-2}$.

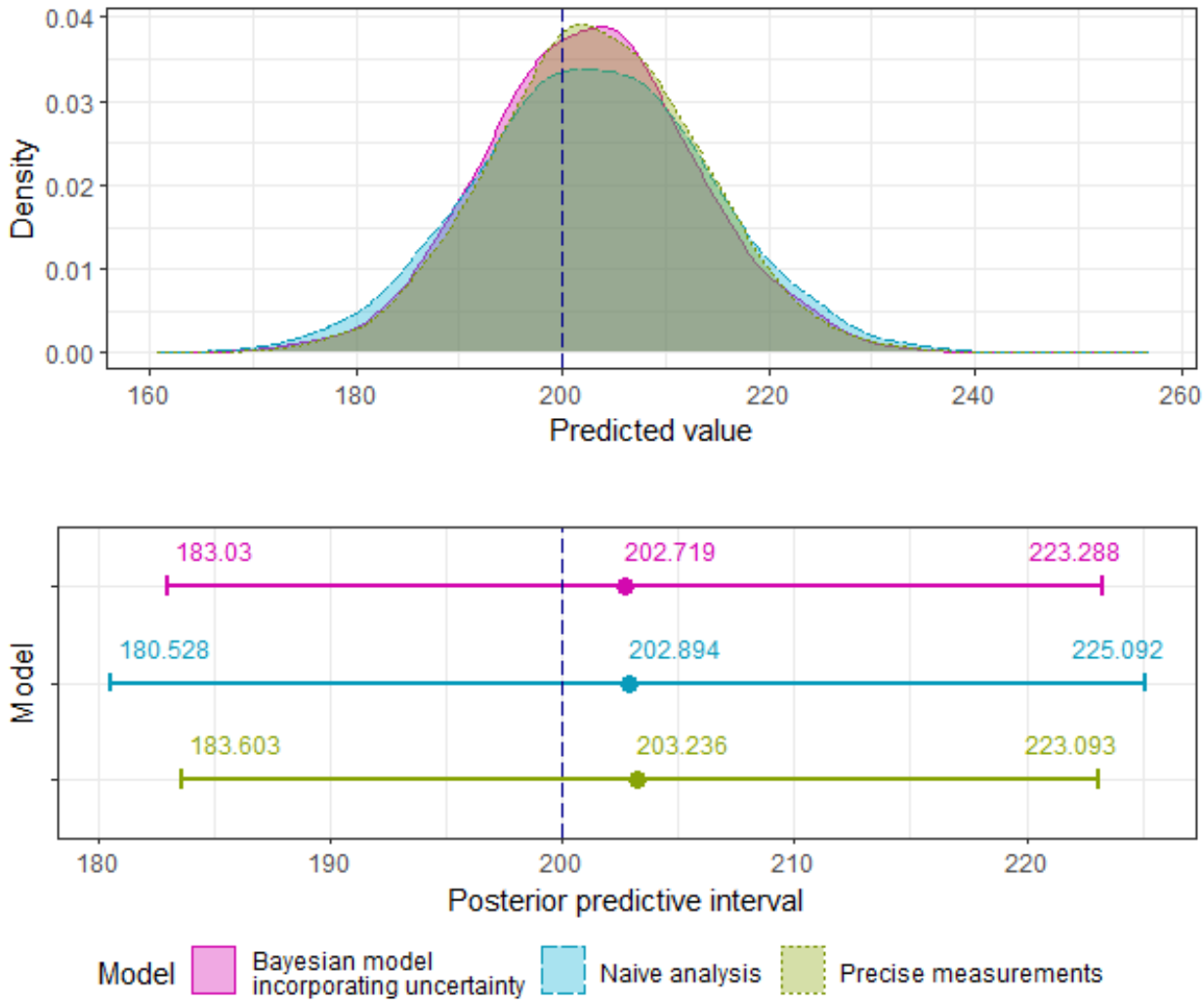


Figure 3.5. Simple linear regression estimate of the posterior predictive distribution at $x = 200$ for the three models and posterior predictive means and 95% credible intervals. The uncertainty is 50% of the population variance. The settings used are $\alpha = 0$, $\beta = 2$, $\tau = 10^{-2}$ and $u = 5^{-2}$.

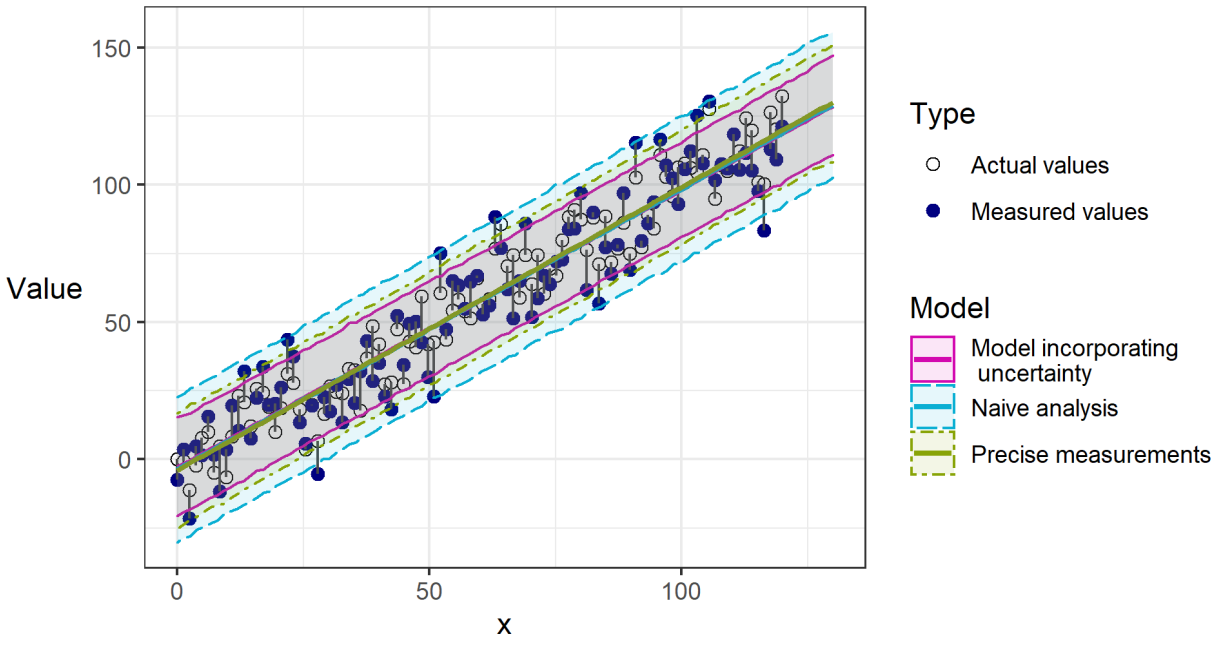


Figure 3.6. Simple linear regression comparison of the three models and the posterior predictive envelopes when the uncertainty is the same size as the population variation. The posterior predictive means are shown by the solid lines. The envelopes represent the 95% credible interval of the posterior predictive distributions. The settings used are $\alpha = 0$, $\beta = 2$, $\tau = 10^{-2}$ and $u = 10^{-2}$.

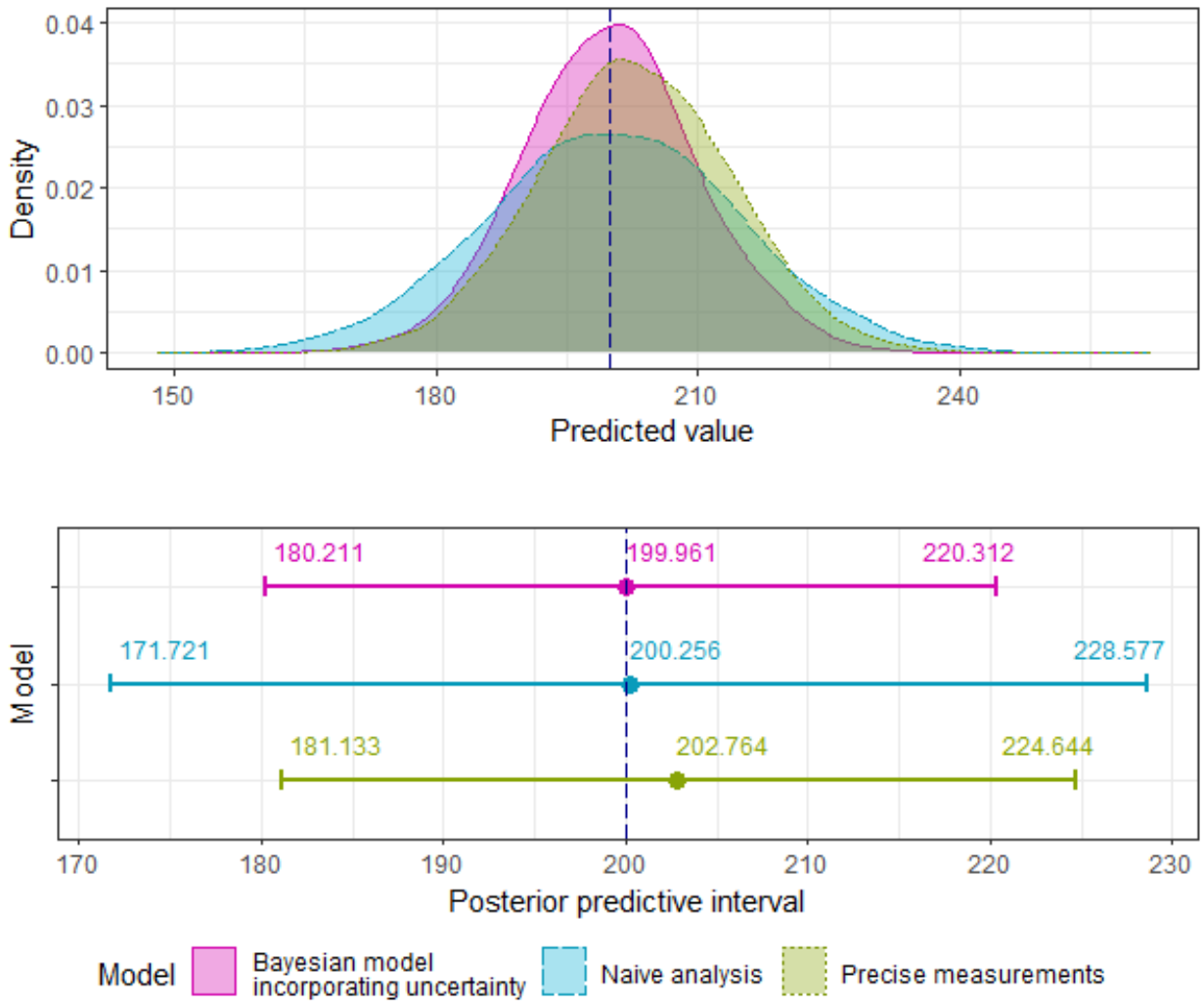


Figure 3.7. Simple linear regression estimate of the posterior predictive distribution at $x = 200$ for the three models and posterior predictive mean and 95% credible intervals. The uncertainty is the same size as the population variance. The settings used are $\alpha = 0$, $\beta = 2$, $\tau = 10^{-2}$ and $u = 10^{-2}$.

six coefficients for the double sigmoidal curve model and the precision parameter. For more details and explanations of all the parameters, see Ellis et al. (2020). The model is specified within a Bayesian framework. Here y_i denotes the individual grape bunch masses and t_i denotes the number of days since 1 December 2017. The logged individual grape bunch masses are assumed to come from a normal distribution

$$\log(y_i) \sim \mathrm{N}(\mu_i, \tau)$$

with mean

$$\mu_i = f(x_i,\ \alpha_0,\ \alpha_1,\ \beta_0, \beta_1,\ \gamma_0,\ \gamma_1) = \frac{\alpha_0}{1 + e^{-\gamma_0(t_i - \beta_0)}} + \frac{\alpha_1}{1 + e^{-\gamma_1(t_i - \beta_1)}} \tag{3.33}$$

and precision τ.

The prior distributions used are given in Table 3.6. For more details on the prior distributions see Ellis et al. (2020).

Table 3.6: Prior distributions used for the double sigmoidal growth models

Coefficient	Prior
α_0	$\mathrm{N}(4.09, 0.11)$
$\Delta\alpha$	$\mathrm{TN}(0.69, 1, 0)$
β_0	$\mathrm{N}(40, 0.02)$
$\Delta\beta$	$\mathrm{TN}(30, 0.11, 0)$
γ_0	$\mathrm{TN}(0.3, 44.44, 0)$
γ_1	$\mathrm{TN}(0.3, 44.44, 0)$
τ	$\mathrm{Gamma}(4, 1)$

Note. TN stands for the truncated normal distribution.

A Metropolis-Hastings sampler has been implemented in R by Ellis et al. (2020) to generate samples from the posterior distribution.

If we are to incorporate uncertainty, we can assume that each measurement m_i is generated from a normal distribution with mean $\log(y_i)$ and precision, u_i,

$$m_i \sim \mathrm{N}(\log(y_i), u_i) \tag{3.34}$$

Note that the generated measurements m_i will also in turn be on a log scale. We can assume any error distribution but we choose a normal error distribution here as it allows us to analytically derive the full conditional distributions. Then we can replace μ_i in Equation 3.18 with Equation 3.28. If we make the same assumption as we did for the linear regression, that each y_i is independent and identically distributed, then instead of Equation 3.29 our full conditional distribution for the parameter y_i would become

$$\log(y_i) \mid \alpha_0,\ \alpha_1,\ \beta_0, \beta_1,\ \gamma_0,\ \gamma_1,\ \tau, m_i, u_i \sim N\left(\frac{u_i m_i + \tau\mu_i}{u_i + \tau},\ u_i + \tau\right) \quad \text{for all } i. \tag{3.35}$$

where

$$\mu_i = \frac{\alpha_0}{1 + e^{-\gamma_0(t_i - \beta_0)}} + \frac{\alpha_1}{1 + e^{-\gamma_1(t_i - \beta_1)}}$$

is from Equation 3.33. Therefore, we can use a Gibbs step to directly sample from the full conditional distribution for $\mathbf{y} = (y_1, \ldots y_n)$, as an additional step in the Metropolis-Hastings sampler written by Ellis et al. (2020).

Additional Gibbs step for the Metropolis-Hastings sampler

Sample each y_i from a normal distribution with mean

$$\frac{u_i m_i + \tau\mu_i}{u_i + \tau}$$

where,

$$\mu_i = \frac{\alpha_0}{1 + e^{-\gamma_0(t_i - \beta_0)}} + \frac{\alpha_1}{1 + e^{-\gamma_1(t_i - \beta_1)}}$$

and precision $u_i + \tau$, using the current value of all the other parameters.

The Metropolis-Hastings algorithm implemented in R is included in Appendix D.

A total of 1,200,000 iterations were used for the Metropolis-Hastings sampler with a burn-in period of 20,000 iterations. The remaining sample was thinned to yield a posterior sample of size 5,000. Convergence was visually assessed.

We report the 95% highest posterior density interval using the `hdi()` function from the `HDInterval` package in R.

We investigated two cases of u_i: when $u_i = 0.1$ and $u_i = 0.5$. The case where the uncertainty is $u_i = 0.5$ can be interpreted as approximately $m_i \pm 100\%$. This could happen when two grape bunches are identified as one grape bunch, which can occur when the wavelength used for the microwave sensor is larger than the distance between two grape bunches. Additionally, there is noise from the leaves and from other unknown sources, which can add to the size of the uncertainty.

Figure 3.11 shows that the posterior predictive envelope is smaller for the Bayesian model incorporating uncertainty than the naive analysis. Figure 3.14 shows that the difference between the posterior predictive envelopes is much more apparent when the uncertainty is larger. The naive analysis exhibits the widest 95% HPD interval, and is much larger than the case of the Bayesian model using precise measurements of mass. Figure 3.12 shows that the posterior predicted mean for both the Bayesian model which incorporates uncertainty and the Bayesian naive analysis is relatively close to the mean predicted value when precise measurements are used. The Bayesian model which incorporates uncertainty has the smallest 95% HPD interval, which means that it is more precise. The Bayesian naive analysis produces the widest 95% HPD interval, which means there is more uncertainty in the prediction.

Figure 3.15 shows that the Bayesian naive analysis has the widest 95% HPD interval for the prediction. However the Bayesian model which incorporates uncertainty has a narrower 95% HPD interval for the prediction than the Bayesian model using precise measurements. This is possible undercoverage exhibited by the Bayesian model which incorporates uncertainty. In conclusion, by using the Bayesian model which incorporates uncertainty,

we can still obtain accurate and precise estimates, however caution may need to be taken when the uncertainty becomes large e.g. $u_i = 0.5$ as there may be undercoverage for the prediction.

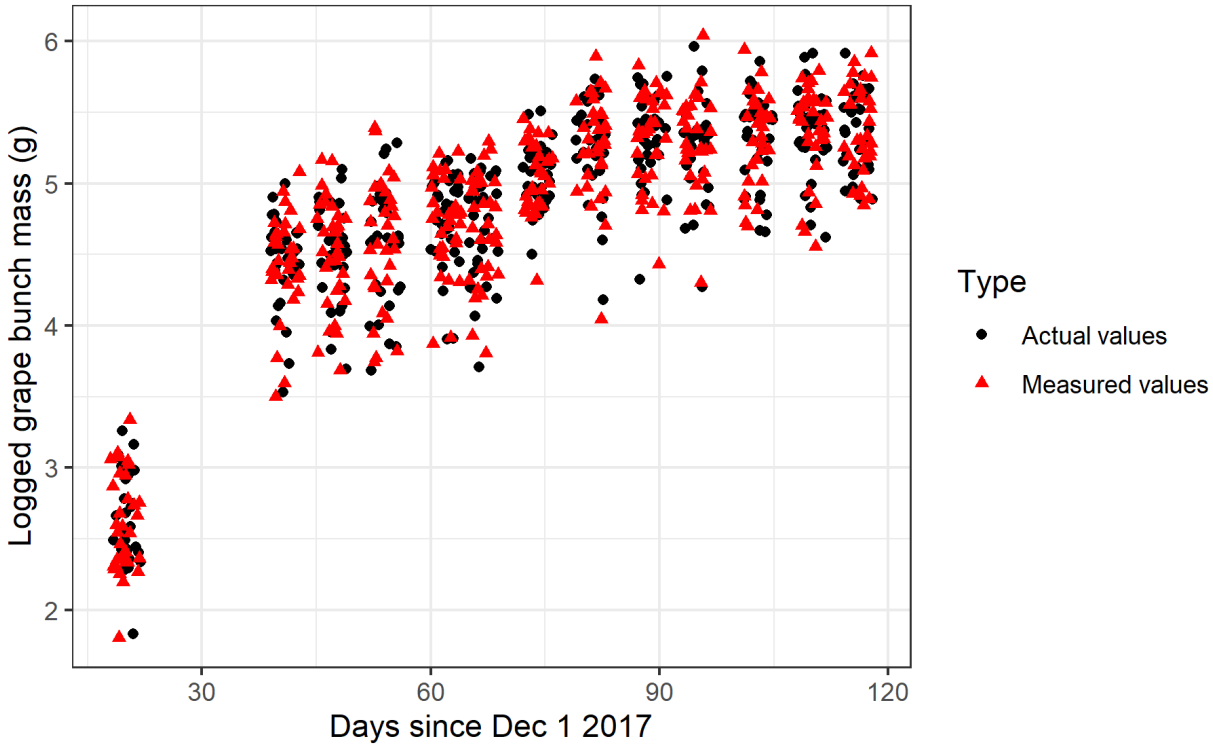


Figure 3.8. Scatterplot of the 2018 grape bunch mass data with actual values, $\mathbf{y}$ and simulated measurements with uncertainty $\mathbf{m}$ using a setting of $u = 0.1^{-2}$. Jitter has been applied to the points to reduce overplotting. The bunch masses are plotted on the log scale.

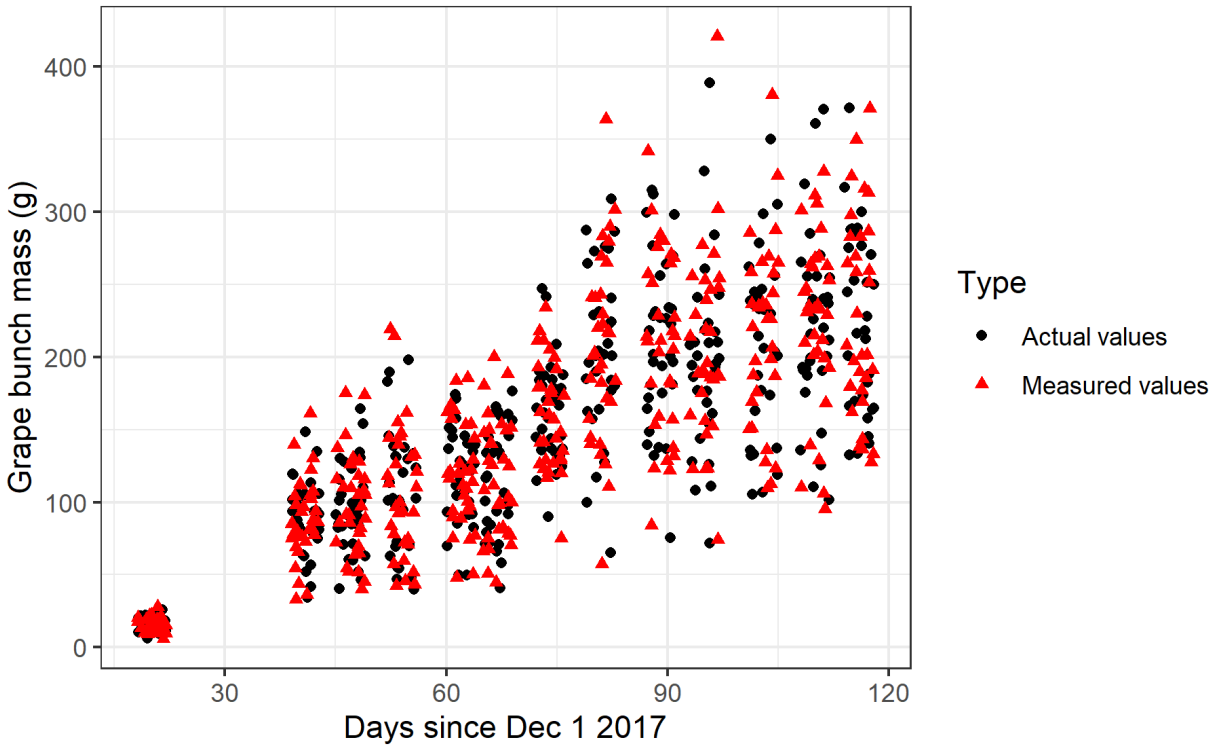


Figure 3.9. Scatterplot of the 2018 grape bunch mass data with actual values, $\mathbf{y}$, simulated measurements with uncertainty $\mathbf{m}$ using a setting of $u = 0.1^{-2}$ which is approximately 20% of the original values (based on 1.96*se, where se is 10%). The bunch masses have been back transformed to be on the original scale in grams. Jitter has been applied to the points to reduce overplotting.

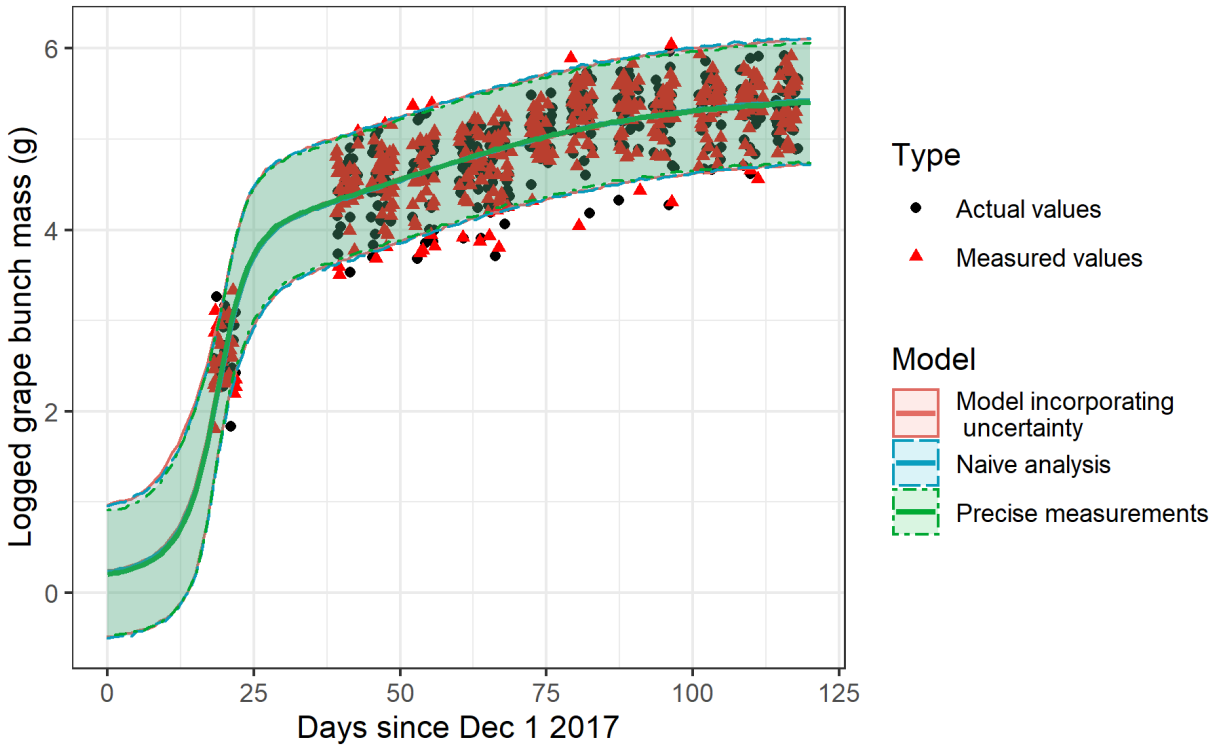


Figure 3.10. Double sigmoidal curve comparison of the three models and the posterior predictive envelopes for simulated uncertainty of 10% of the original values. The solid lines are the means of the posterior predictive distributions. The envelopes show the 95% credible interval of the posterior predictive distributions. The grape bunch masses are plotted on the log scale.

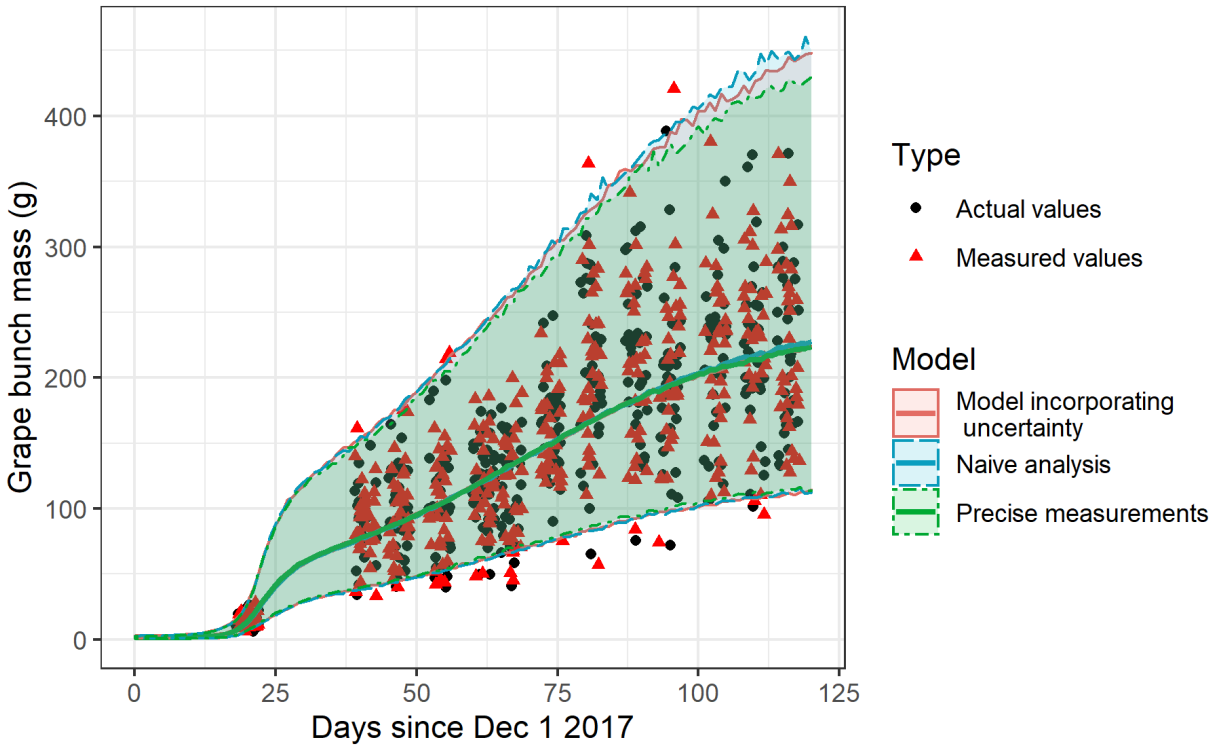


Figure 3.11. Double sigmoidal curve comparison of the three models and the posterior predictive envelopes for simulated uncertainty of 10% of the original values. The solid lines are the means of the posterior predictive distributions. The envelopes show the 95% credible interval of the posterior predictive distributions.

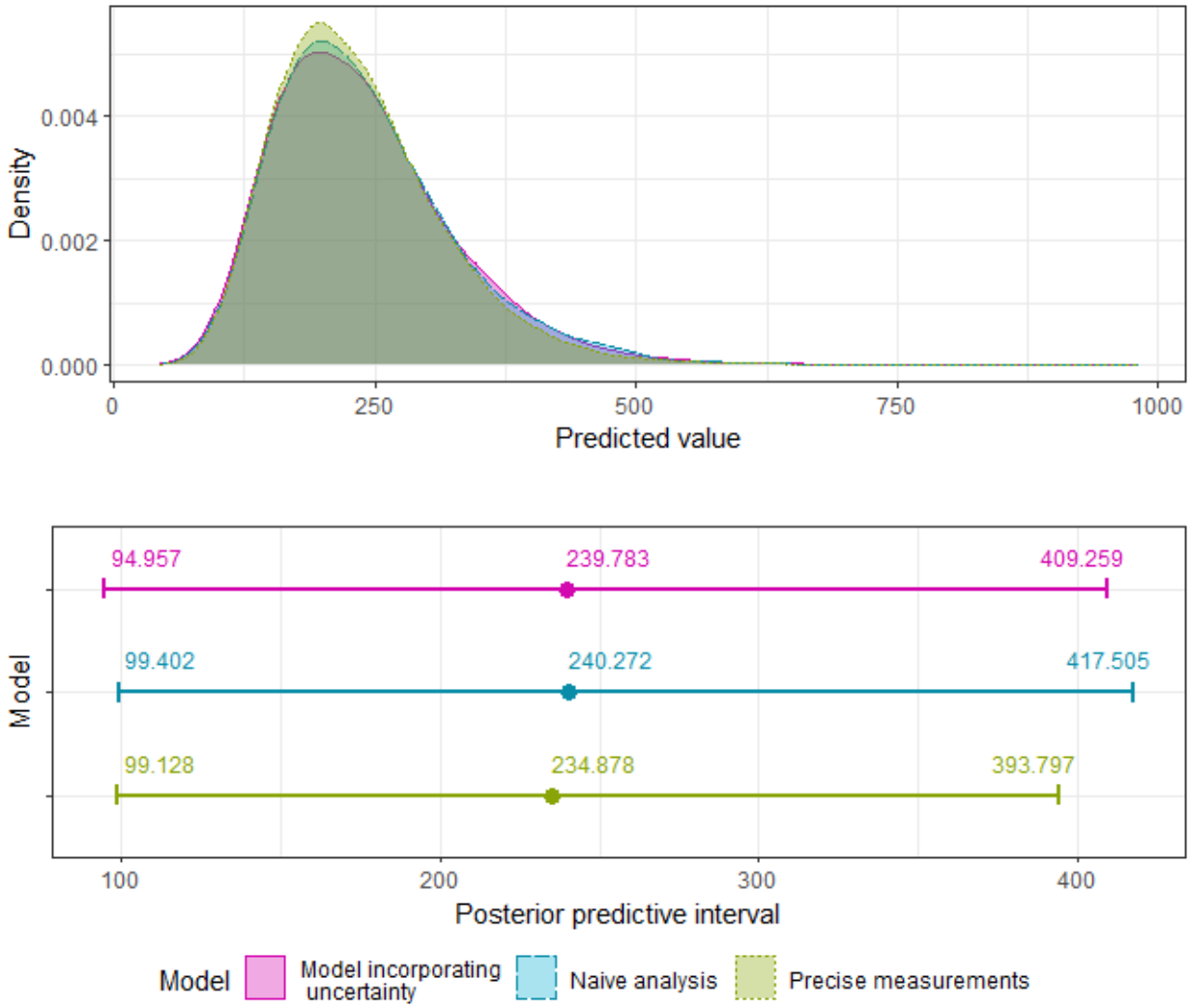


Figure 3.12. Double sigmoidal curve comparison of the estimated posterior predictive distribution at day 120 for the three models. This was for simulated uncertainty of 20% of the original values. Below the densities plot are the posterior predictive means and 95% credible intervals.

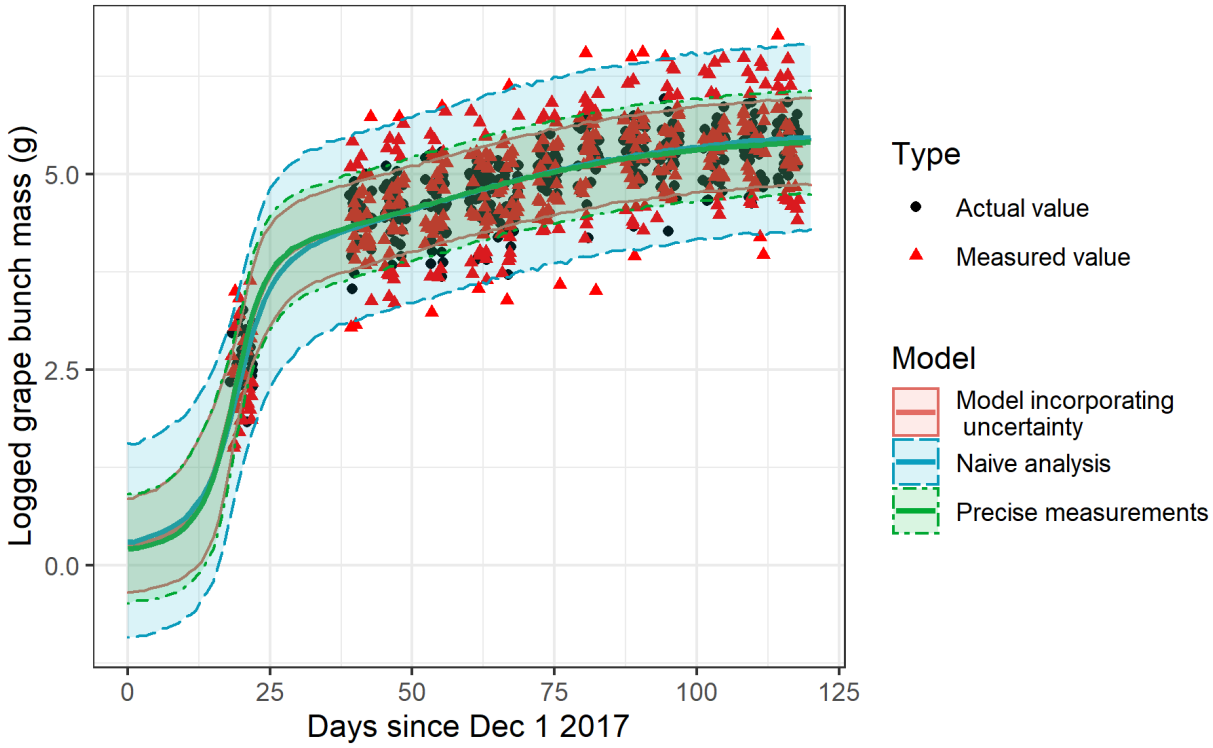


Figure 3.13. Double sigmoidal curve comparison of the three models and the posterior predictive envelopes for simulated uncertainty of 50% of the actual values. This is approximately 2 times the actual values, $\mathbf{y}$ (based on 1.96*se, where se is 50%). The solid lines are the means of the posterior predictive distributions. The envelopes show the 95% credible interval of the posterior predictive distributions. The grape bunch masses are plotted on the log scale.

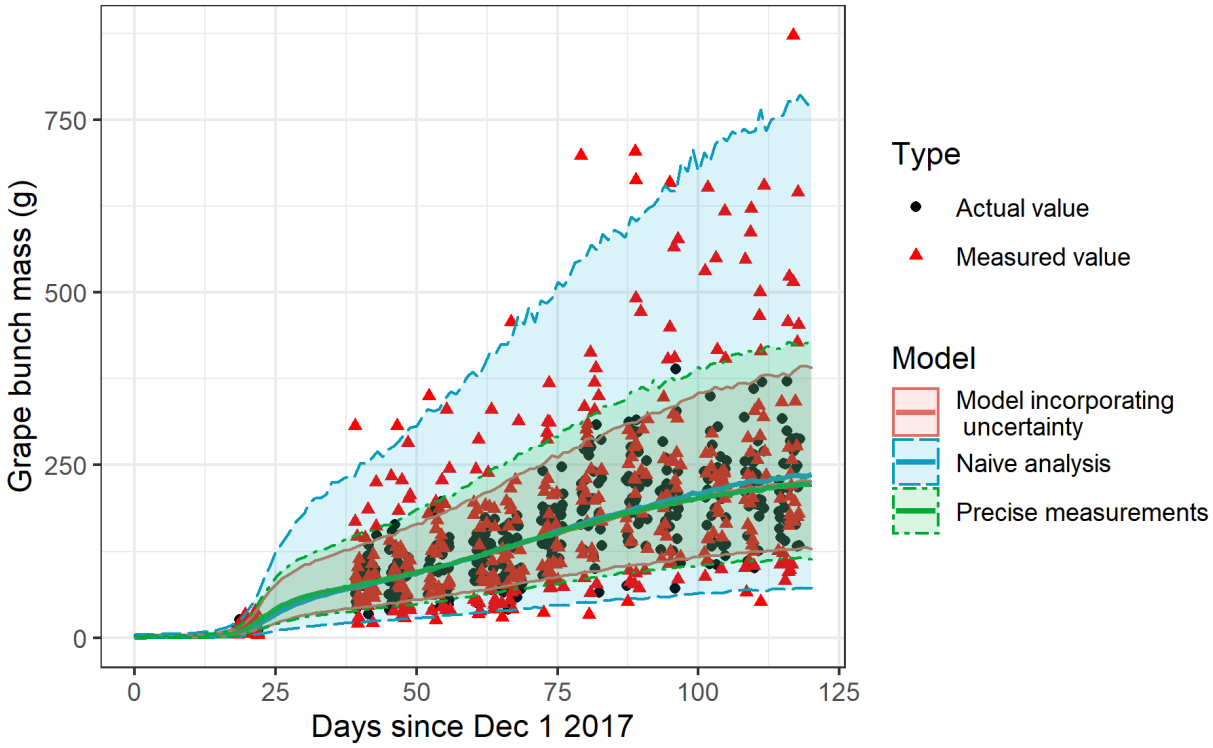


Figure 3.14. Double sigmoidal curve comparison of the three models and the posterior predictive envelopes for simulated uncertainty of 50% of the actual values. This is approximately 2 times the actual values, **y** (based on 1.96*se, where se is 50%). The solid lines are the means of the posterior predictive distributions. The envelopes show the 95% credible interval of the posterior predictive distributions. Bunch masses are on the original scale in grams (g).

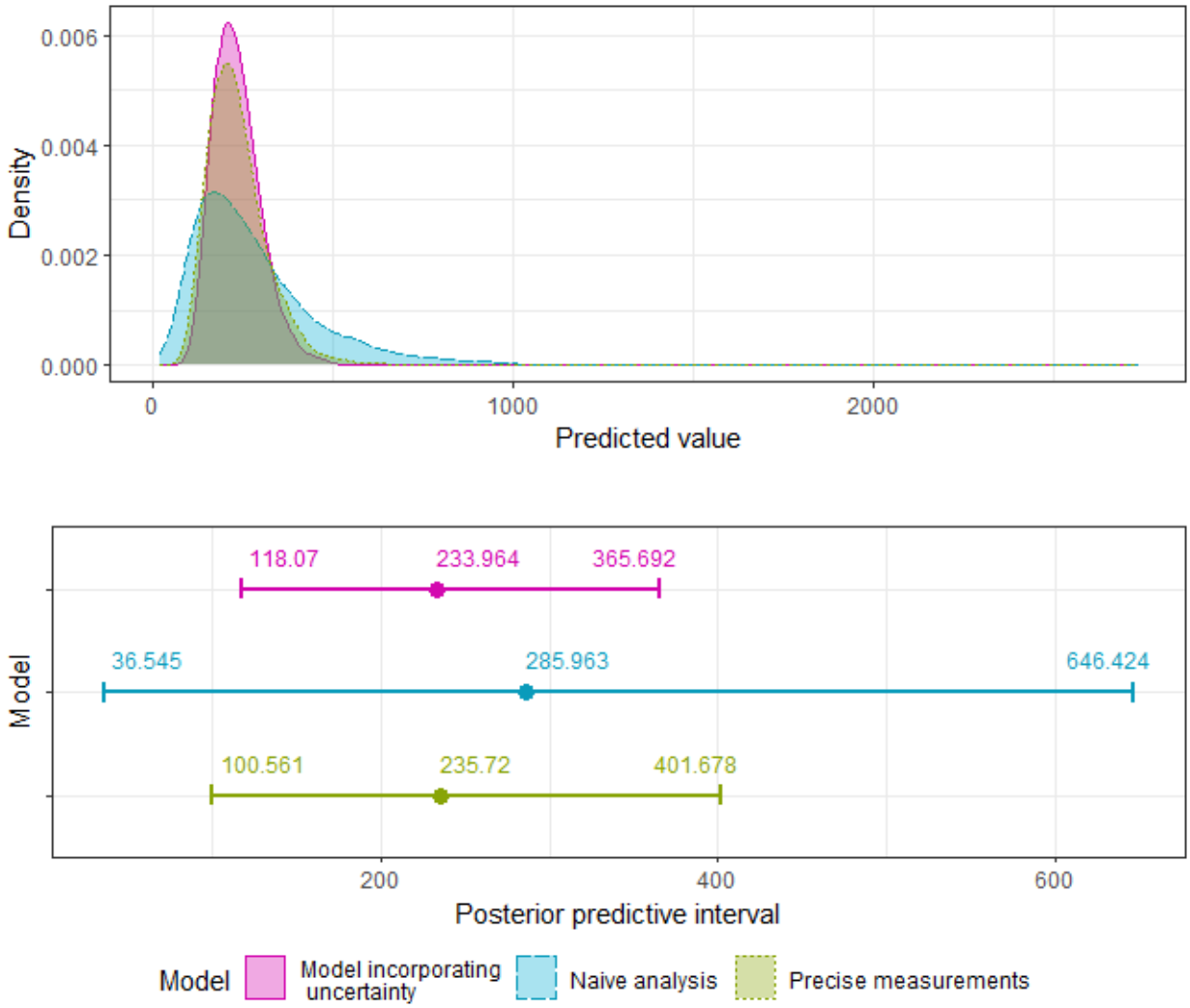


Figure 3.15. Double sigmoidal curve comparison of the estimated posterior predictive distribution at day 120 for the three models. This was for simulated uncertainty of 100% of the original values. Below the densities plot are the posterior predictive means and 95% credible intervals.

3.8 Discussion

In the examples above, we were able to use a Gibbs sampler due to the convenience of the assumed distribution and the resulting conjugacy. This may not always be possible, if the distribution $p(\mathbf{m} \mid \mathbf{y})$, follows another parametric distribution, for example a t-distribution or triangular distribution. In which case, a Metropolis-Hastings algorithm may be used instead of a Gibbs step. The Metropolis- Hastings acceptance ratio, R would be as follows

$$R = \frac{p(\mathbf{m} \mid \mathbf{y}^*, \mathbf{u})p(\mathbf{y}^* \mid \theta)/J_t(\mathbf{y}^* \mid \mathbf{y}^{t-1})}{p(\mathbf{m} \mid \mathbf{y}^{t-1}, \mathbf{u})p(\mathbf{y}^{t-1} \mid \theta)/J_t(\mathbf{y}^{t-1} \mid \mathbf{y}^*)} \tag{3.36}$$

where θ is the parameter vector, $J_t(\cdot)$ is the proposal (or jumping) distribution, y^* is the proposed value of y, and y^{t-1} is the value of y at the previous iteration. Unfortunately, unlike the Gibbs sampler which accepts every proposal, Metropolis-Hastings algorithm only accepts some proposal values. Thus, it generally requires longer runs and is more computationally intensive.

We have provided a framework for incorporating uncertainty given measured values $\mathbf{m}$ and stated uncertainty $\mathbf{u}$ for single predictor models. However, Ellis et al. (2020) discuss that other variables could affect grape growth such as temperature, the amount of solar radiation and characteristics of the land. If there are multiple predictors and a normal likelihood is assumed then the only part of the model that would change is μ_i in Equation 3.18. For example, if there are three predictors, then Equation 3.18 could be modified to become $\mu_i = \alpha + \beta_0 x_i + \beta_1 x_i + \beta_2 x_i$. McElreath (2020) provide an example of a model with two predictors where there is measurement error in the response and one predictor variable in Stan.

There are several other advantages to our Bayesian model which incorporates uncertainty. Even when a vague prior is used, we can still estimate population parameters and produce accurate and precise predictions for uncertainty that is smaller than 80% of the population variation. Our Bayesian model which incorporates uncertainty also allows for measurement specific uncertainty, i.e. if the size of the uncertainty is different for each

observation. Our Bayesian model which incorporates uncertainty can still produces accurate and precise (small variance) predictions. However, our simulation studies show that caution should be taken when the uncertainty becomes large, for instance when it is 80% of the population variation, then the Bayesian model which incorporates uncertainty can result in undercoverage in the prediction (posterior predictive distribution).

We have found that if a normal error model is assumed, then it is important to incorporate uncertainty. The Bayesian model which incorporates uncertainty always produces more precise predictions, then the naive analysis which ignores uncertainty. This is true for the sample of values, simple linear regression and double sigmoidal curve.

In a practical sense, we are more interested in the prediction and the uncertainty (variability) of the prediction from the model, rather than the estimated population mean. Which is why we focus our simulation studies on varying the settings of the uncertainty and population variance on the posterior predictive distribution, rather than the estimation of the population mean and variance. From frequentist statistics theory, the (nominal) prediction interval is always larger than the confidence interval for the mean. For a larger vineyard, it might be possible to use the estimated mean, as the 95% credible interval for the mean and for the posterior predictive distribution may be very similar. For smaller vineyards, we might be more interested in the predicted predictive distribution.

A desirable solution is one that can produce timely estimates of grape yield. In the case where $p(m_i \mid y_i, u_i)$ is assumed to have a normal distribution, we have shown that a Gibbs sampler can be used. The Gibbs sampler is computationally efficient, which means that the algorithm will not hinder the timeliness of the predictions.

The Bayesian model incorporating uncertainty always gives a more precise prediction (narrower credible interval) when the uncertainty is approximately less than 80% of the population variation. For practical purposes for the Grape Yield Analyser project, it is up to the grape growers and winemakers, whether it is worth using the Bayesian model which incorporates uncertainty. How "small" the difference in prediction intervals is between the

Bayesian model which incorporates uncertainty and naive analysis is and whether it is worth incorporating uncertainty, is to be determined by the grape growers and winemakers and how much additional uncertainty they are willing to tolerate.

Chapter 4

Nonparametric model

4.1 Overview

In this chapter we consider the situation where the sensor output is a sample of values representing a nonparametric distribution instead of a single precisely measured value. First, we develop a Bayesian model first for the simplest case of a single data point, then extending the model to a sample of values. We evaluate our Bayesian nonparametric model by first conducting a preliminary study to assess how well it estimates the true parameter values of the mean and variance in terms of bias. Then we conduct a full simulation study and compare the results with a Bayesian naive analysis and a Bayesian model with precisely measured values. Finally, we illustrate our recommended model of a Bayesian naive analysis, for a double sigmoidal growth model and discuss our results.

4.2 A single data point ($n = 1$)

We will start with the simplest case of a single data point. Consider a normal likelihood model, $p(y \mid \mu)$ with unknown mean μ and known precision τ. Let $p(\mu)$ be the prior

for the mean. Suppose that y is not observed directly but a sample of K equally possible values, $\mathbf{d} = \{d_1, \ldots, d_K\}$ is observed instead. For a single observation, we can define this as the random variable D with a discrete uniform distribution

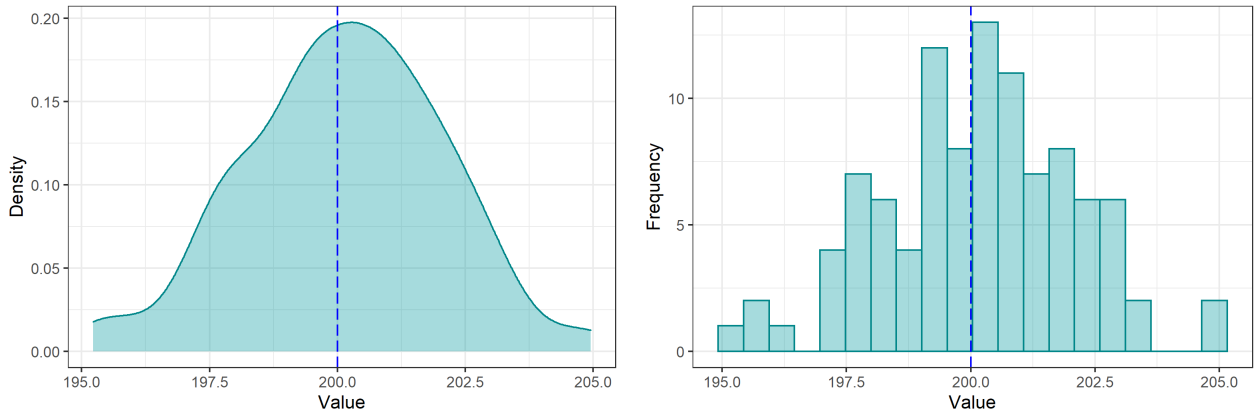
$$f(d \mid y) = P(D = d \mid y) = \begin{cases} \frac{1}{K} & \text{if } d \in \{d_1, \ldots, d_K\} \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

The likelihood of observing the sample of K values $\mathbf{d}$ is

$$f(\mathbf{d} \mid y) = P(d_1 \mid y) \times \ldots \times P(d_K \mid y) = \left(\frac{1}{K}\right)^K.$$

If we treat both μ and y as parameters, Bayes' theorem provides the joint posterior distribution as

$$p(\mu, y \mid \mathbf{d}) = \frac{f(\mathbf{d} \mid y)p(y \mid \mu)p(\mu)}{p(\mathbf{d})} = \frac{f(\mathbf{d} \mid y)p(y \mid \mu)p(\mu)}{\int\int f(\mathbf{d} \mid y)p(y \mid \mu)p(\mu)d\mu dy}$$
$$= \frac{\left(\frac{1}{K}\right)^K p(y \mid \mu)p(\mu)}{\left(\frac{1}{K}\right)^K \int\int p(y \mid \mu)p(\mu)d\mu dy}$$



(a) Density plot

(b) Histogram

Figure 4.1. Illustrating the simulated sample of values, $\mathbf{d}$ observed instead of a single precisely measured observation y as a (a) density plot (obtained using kernel density estimation in R) and (b) as a frequency histogram. The true value of y is represented by the vertical dashed line. Data is generated from $d_k \sim \mathrm{N}(y = 200, u = 5^{-2})$ for $k = 1, \ldots, K$ where $K = 100$.

$$= \frac{p(y \mid \mu)p(\mu)}{\int\int p(y \mid \mu)p(\mu)d\mu dy} \quad \text{for all } y \in \mathbf{d}.$$

Since we assume that y can only take values $d_1, \ldots, d_K$ (otherwise $p(y)$ is zero), the joint posterior distribution becomes

$$p(\mu, y \mid \mathbf{d}) = \frac{p(y \mid \mu)p(\mu)}{\sum_{k=1}^{K} \int p(d_k \mid \mu)p(\mu)d\mu} \quad \text{for all } y \in \{d_1, \ldots, d_K\}. \tag{4.2}$$

There is no closed-form analytical solution to the joint posterior and the marginal posterior distributions. However, it is possible to analytically derive the full conditional distributions for a Gibbs sampler. For the full conditional distribution $p(y \mid \mathbf{d}, \mu)$, using Bayes' theorem we have

$$\begin{aligned} p(y \mid \mathbf{d}, \mu) &= \frac{f(\mathbf{d} \mid y)p(y \mid \mu)}{p(\mathbf{d}, \mu)} \\ &= \frac{f(\mathbf{d} \mid y)p(y \mid \mu)}{\sum_{k=1}^{K} f(\mathbf{d} \mid y)p(d_k \mid \mu)} \\ &= \frac{\left(\frac{1}{K}\right)^K p(y \mid \mu)}{\left(\frac{1}{K}\right)^K \sum_{k=1}^{K} p(d_k \mid \mu)} \\ &= \frac{p(y \mid \mu)}{\sum_{k=1}^{K} p(d_k \mid \mu)} \quad \text{for all } y \in \{d_1, \ldots, d_K\} \text{ and 0 otherwise.} \end{aligned} \tag{4.3}$$

Given $\mathbf{d}$ and μ, we can sample y directly from a discrete uniform distribution on the set $\{d_1, \ldots, d_K\}$ with probability

$$p(y = d_i \mid \mu, \mathbf{d}) = \frac{p(d_i \mid \mu)}{\sum_{k=1}^{K} p(d_k \mid \mu)} \tag{4.4}$$

where $p(d_i \mid \mu)$ is a normal probability density function.

The full conditional distribution for μ for a single y is

$$\mu \mid y \sim \mathrm{N}\left(\frac{y + \mu_0\tau_0}{\tau + \tau_0}, \tau + \tau_0\right). \tag{4.5}$$

Now we have both full conditional distributions required for the Gibbs sampler to sample from the joint posterior $p(\mu, y \mid \mathbf{d})$. To obtain the marginal posterior of our parameter of interest μ, $p(\mu \mid \mathbf{d})$ we would simply take the posterior sample and look at the values of μ and ignore the values of y.

4.3 Multiple data points ($n > 1$)

Suppose that we now have a vector, $\mathbf{y} = (y_1, \ldots y_n)$ from a normal population with unknown mean, μ and precision, τ:

$$y_i \sim \mathrm{N}(\mu, \tau).$$

We observe a sample of K values, $\mathbf{d_i} = \{d_{i1}, \ldots, d_{iK}\}$ instead of each y_i. Let d_{ik} be k^{th} value in the sample for the i^{th} value of y, where $k = 1, \ldots, K$, and K is the size of the sample. In total we observe a list of n samples, $\mathcal{D} = \{\mathbf{d_1}, \ldots, \mathbf{d_n}\}$, instead of the vector $\mathbf{y}$.

For our Bayesian nonparametric model, let us assume that we have a sample of equally likely observations $\mathbf{d_i}$ for each y_i. The likelihood for a single observation d_{ik} given the true value of the observation, y_i is given by

$$f(d_i|y_i) = P(d_i \mid y_i) = \begin{cases} \frac{1}{K} & \text{if } d_i \in \{d_{i1}, \ldots, d_{iK}\} \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

Therefore, the joint likelihood for the sample of values $\mathbf{d_i}$ becomes

$$\begin{aligned} p(\mathbf{d_i}|y_i) &= p(d_{i1}, \ldots, d_{iK} \mid y_i) \\ &= P(d_{i1} \mid y_i) \times \ldots \times P(d_{iK} \mid y_i) = \left(\frac{1}{K}\right)^K. \end{aligned} \tag{4.7}$$

If we assume independent priors

$$p(\mu, \tau) = p(\mu)p(\tau)$$

then we can write the joint posterior distribution for μ, τ and $\mathbf{y}$ using Bayes' theorem as

$$p(\mu, \tau, \mathbf{y} \mid \mathcal{D}) \propto \prod_{i=1}^{n} p(\mathbf{d_i} \mid y_i) \prod_{i=1}^{n} p(y_i|\mu, \tau)p(\mu)p(\tau). \tag{4.8}$$

The DAG for the Bayesian nonparametric model is shown in Figure 4.2.

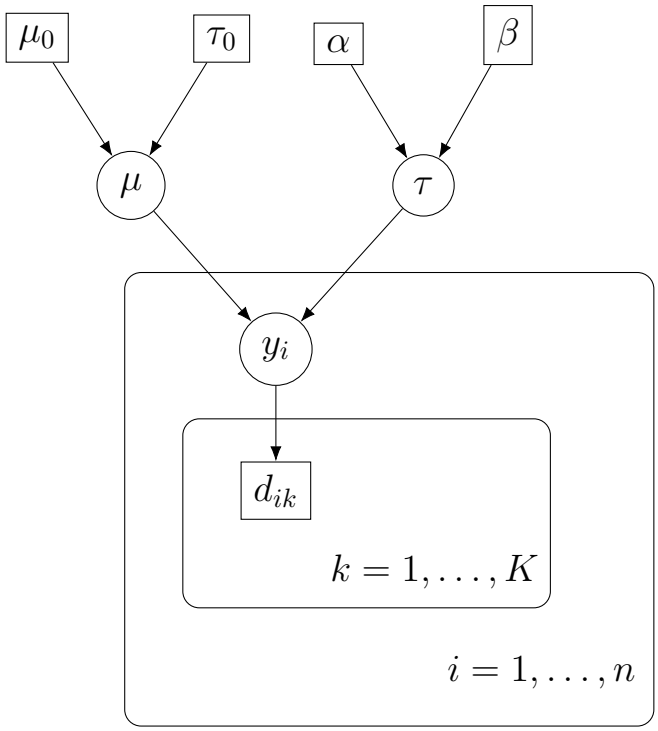


Figure 4.2. Directed acyclic graph for the Bayesian nonparametric model.

We specify conditionally conjugate priors; a normal prior for μ and a gamma prior for τ as follows

$$\mu \sim \mathrm{N}(\mu,\ \tau_0)$$
$$\tau \sim \mathrm{Gamma}(\alpha\ \beta).$$

so that the full conditional distributions for μ and τ can be analytically derived and so we can use a Gibbs sampler.

For the Gibbs sampler, we have the full conditional distribution for μ and τ in Chapter 3 (Equation 3.5 and 3.6, respectively). Since we assume that each y_i is independent of each other, we can derive the full conditional posterior distribution for each y_i separately as we treat the other parameters as constants. The full conditional distribution for each y_i can be derived as follows

$$p(y_i \mid \mathbf{d_i}, \mu, \tau) = \frac{p(\mathbf{d_i} \mid y_i)p(y_i \mid \mu, \tau)}{\sum_{k=1}^{K} p(\mathbf{d_i} \mid y_i)p(d_{ik} \mid \mu, \tau)}$$

$$= \frac{(\frac{1}{K})^K p(y_i \mid \mu, \tau)}{(\frac{1}{K})^K \sum_{k=1}^K p(d_{ik} \mid \mu, \tau)}$$
$$= \frac{p(y_i \mid \mu, \tau)}{\sum_{k=1}^K p(d_{ik} \mid \mu, \tau)} \quad \text{for all } y_i \in \{d_{i1}, \ldots, d_{iK}\}.$$

Since the likelihood of observing y_i given μ and τ, is a normal probability density given by

$$p(y_i \mid \mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\Big(-\frac{\tau}{2}(y_i - \mu)^2 \Big) \tag{4.9}$$

the full conditional distribution for y_i is

$$p(y_i \mid \mathbf{d_i}, \mu, \tau) = \frac{\sqrt{\frac{\tau}{2\pi}} \exp\big(-\frac{\tau}{2}(y_i - \mu)^2 \big)}{\sum_{k=1}^K \sqrt{\frac{\tau}{2\pi}} \exp\big(-\frac{\tau}{2}(d_{ik} - \mu)^2 \big)} \quad \text{for all } y_i \in \{d_{i1}, \ldots, d_{iK}\}. \tag{4.10}$$

This means that the full conditional distribution, $p(y_i \mid \mathbf{d_i}, \mu, \tau)$ is the likelihood and we can sample for each y_i in a Gibbs step.

4.3.1 MCMC algorithm: Gibbs sampler

Pseudo-code for the Gibbs sampler

Step 0. Set the arbitrary initial values for the parameters. E.g. $\mu^{(0)} = \overline{\mathcal{D}}$, $\tau^{(0)} = 1/s^2_{\mathcal{D}}$ and $y_i^{(0)} = \overline{\mathbf{d_i}}$ for $i = 1, \ldots, n$. Where $\overline{\mathcal{D}}$ is the mean of all the values in the list $\mathcal{D}$ and $s^2_{\mathcal{D}}$ is the variance of all the values in the list $\mathcal{D}$.

Step 1. Sample each y_i from

$$p(y_i \mid \mathbf{d_i}, \mu, \tau) = \frac{\sqrt{\frac{\tau}{2\pi}} \exp\big(-\frac{\tau}{2}(y_i - \mu)^2 \big)}{\sum_{k=1}^K \sqrt{\frac{\tau}{2\pi}} \exp\big(-\frac{\tau}{2}(d_{ik} - \mu)^2 \big)},$$

using the current value of the other parameters μ and τ.

Step 2. Sample one value of μ from a normal distribution with mean,

$$\frac{\tau n \bar{y} + \tau_0 \mu_0}{n\tau + \tau_0}$$

and precision, $n\tau + \tau_0$, using the current values of the other parameters $\mathbf{y}$ and τ.

Step 3. Sample one value of τ from a gamma distribution with shape parameter, $\alpha + \frac{n}{2}$ and rate parameter, $\beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2$, using the current values of the other parameters μ and $\mathbf{y}$.

Repeat steps 1, 2 and 3 until convergence.

An implementation of this Gibbs sampler in R can be found in Appendix C.

A total of 10,000 iterations were used for the Gibbs sampler with a burn-in period of 1,000 iterations. Convergence was visually assessed.

4.4 Preliminary study

Before doing a full simulation study, we wanted to determine how well the model would perform in terms of bias in the estimation of the mean and variance. This is because bias in the estimate of the mean and variance will mean that the model will produce less accurate or precise predictions. We examined the behaviour of the cumulative mean of the mean and variance. We varied n, the number of observations and K, the sample of values for each observation. Data was generated according to the following model:

$$y_i \sim \mathrm{N}(\mu, \tau) \quad \text{for } i = 1, \ldots, n \tag{4.11}$$

$$d_{ik} \sim \mathrm{N}(y_i, u_i) \quad \text{for } k = 1, \ldots, K. \tag{4.12}$$

The uncertainty was the same for all measurements i.e. $u_i = u$. The following settings were used: $\mu = 200$, $\tau = 5^{-2}$ and $u_i = 2^{-2}$. With these settings, we looked at the following combinations of n and K: $\{n = 100,\ K = 100\}$, $\{n = 200,\ K = 100\}$, $\{n = 100,\ K = 10,000\}$ and $\{n = 200,\ K = 10,000\}$. To estimate the mean, we used the mean of the estimated marginal posterior of the mean. The variance estimate was calculated as 1 over the posterior mean of the precision, $\hat{\sigma^2} = 1/\mathrm{mean}(\hat{\tau})$.

Figure 4.5 shows the cumulative mean error of the estimated mean converges to zero as the number of iterations becomes larger, which means that we get an unbiased estimate

of the population mean. However, the cumulative mean error of the precision converged to approximately 0.0099 after 300 iterations. The corresponding mean error of the variance converged to approximately -4.4 after 300 iterations. Thus, we conclude that when we use the setting $n = 100$, $K = 100$ for our Bayesian nonparametric model, we can obtain unbiased estimates of the population mean. However, we get a large positive bias in the estimate of the precision of approximately 0.01 when we use a noninformative Gamma($\alpha = 0.1$, $\beta = 0.1$) prior (with a corresponding mean of 1 and variance of 10) for the precision.

We wanted to see if the reason why the precision was overestimated (and therefore, the variance was underestimated) was because the sample size K of each observation used was small, and we were not sampling from the tails of the distribution for each y_i. However, even when we used the setting $\{n = 200,\ K = 10,000\}$ the cumulative mean error of the variance converged to approximately -4, illustrated in Figure 4.6. Increasing both n and K to $n = 200$ and $K = 10,000$ also did not lead to an unbiased estimate of the population variance. Therefore, we conclude that the Bayesian nonparametric model tends to underestimate the population variance.

To see if using a more informative prior for the precision could improve the estimate of the population precision, we tried a gamma prior with mean, $\alpha/\beta = 0.04$ and standard deviation, $\sqrt{\alpha/\beta^2} = 0.001$ to see if this would give us an unbiased estimate of the precision. Solving for α and β, we obtained the parameters of the prior distribution for τ which is a gamma distribution parameterised as Gamma($\alpha = 1,600,\ \beta = 40,000$).

Figure 4.7 shows that the cumulative mean error of the posterior mean of the precision converged to approximately -0.08 for a true variance of 25. Thus, we were able to obtain an acceptably small amount of bias in the estimate of the precision.

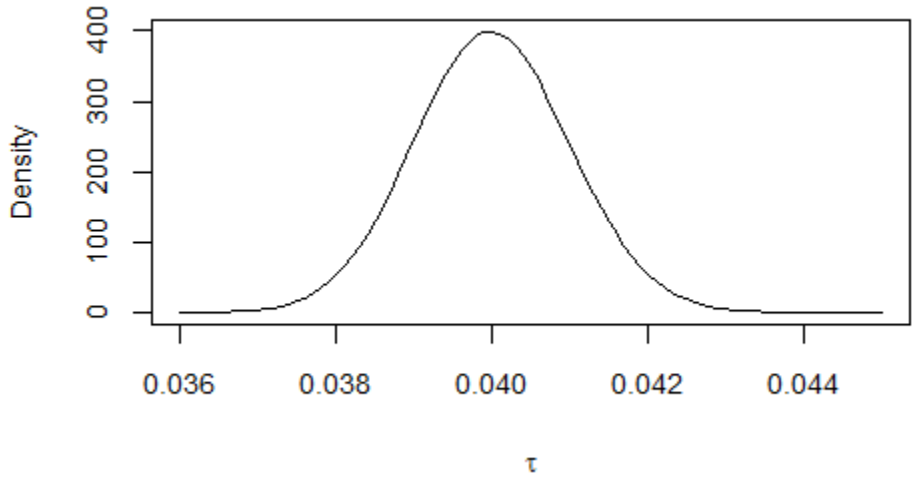


Figure 4.3. Probability density function of the Gamma($\alpha = 1,600,\ \beta = 40,000$) prior.

4.5 Simulation study: a sample of values with a nonparametric error distribution

The aim of this simulation study was to evaluate our model and compare the performance of our Bayesian nonparametric model with a Bayesian naive analysis and the Bayesian model where we have precise measurements. We evaluate the performance of the models in terms of estimation of the population parameters, μ and τ, and their performance in terms of prediction. For the Bayesian naive analysis, we take the mean of each sample. We use the Bayesian model with precise measurements as a comparison for our other two models. Data was generated from the following model:

$$y_i \sim \mathrm{N}(\mu, \tau) \quad \text{for } i = 1, \ldots, n$$
$$d_{ik} \sim \mathrm{N}(y_i, u_i) \quad \text{for } k = 1, \ldots, K.$$

The same sized uncertainty was used where $u_i = u$. The settings used were: $n = 200$, $K = 100$, $\mu = 200$, $\tau = 5^{-2}$ and $u_i = 2^{-2}$. For all models, a vague prior was specified for the mean as follows

$$\mu \sim \mathrm{N}(\mu_0 = 0, \tau_0 = 10^{-5}).$$

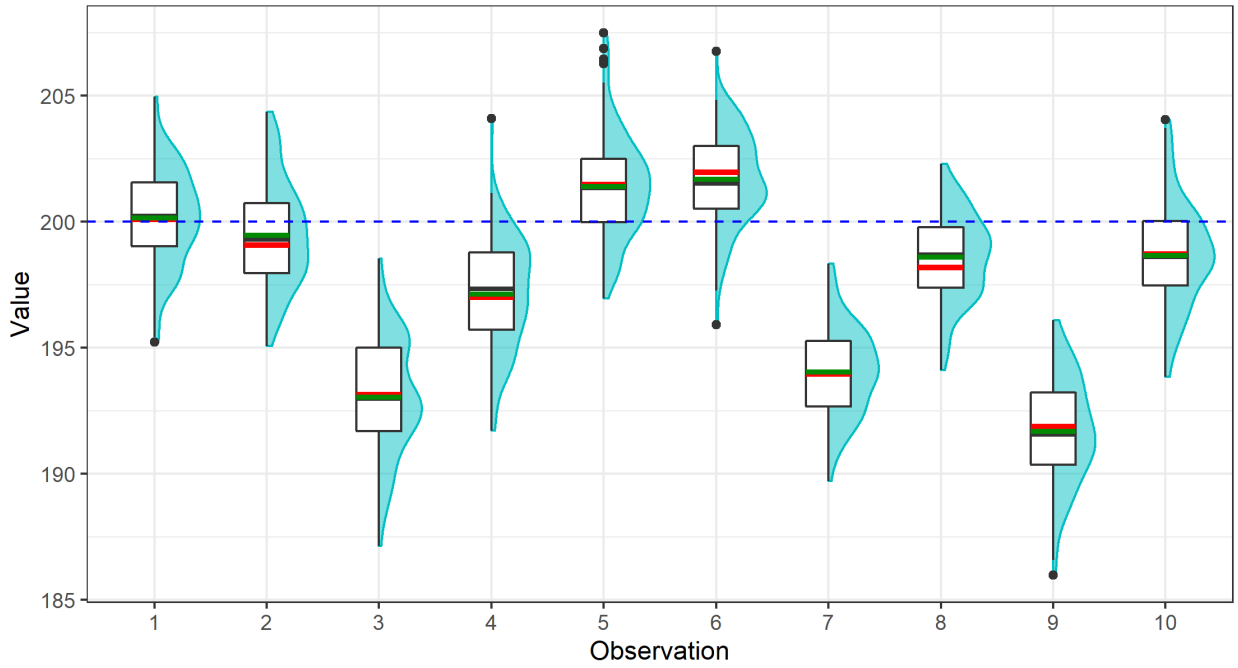


Figure 4.4. Illustrating the nonparametric distributions using boxplots and half violin plots. These are simulated observed samples. True values of $\mathbf{y}$ are shown by the red line segments. The sample means are shown by the green line segments. The true population mean is represented by the horizontal dashed line. The samples were generated following the data-generating mechanism given by Equation 4.11 and Equation 4.12. The settings used are: $\mu = 200$, $\tau = 5^{-2}$, $u_i = 2^{-2}$, $n = 200$ and $K = 100$. Only the first 10 samples were plotted because it would be difficult to display all $n = 200$ in a single plot.

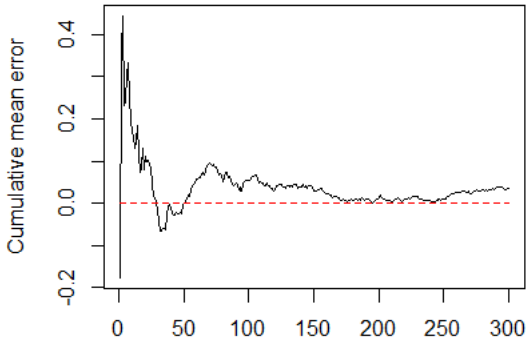
For the Bayesian nonparametric model, we compared the results from using a vague prior for the precision

$$\tau \sim \text{Gamma}(\alpha = 0.01,\ \beta = 0.01)$$

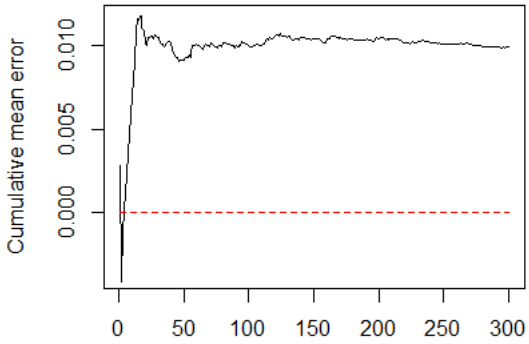
which has mean of 0 and variance 100. And an informative Gamma prior, from the preliminary study of:

$$\tau \sim \text{Gamma}(\alpha = 1,600,\ \beta = 40,000).$$

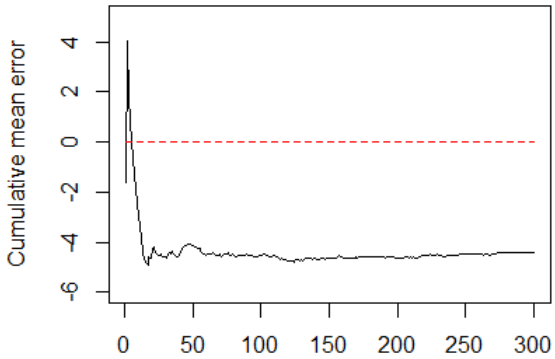
For the Bayesian naive analysis and Bayesian model with precise measurements, only the vague prior for the precision was used.



(a) Mean



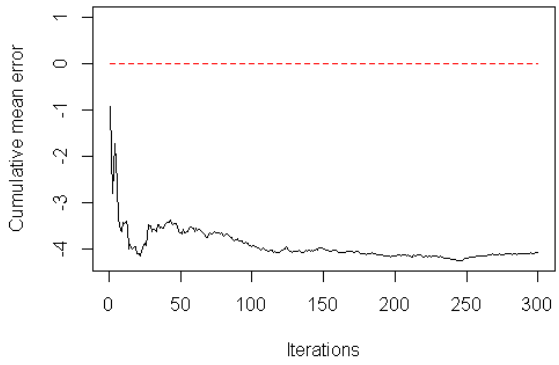
(b) Precision



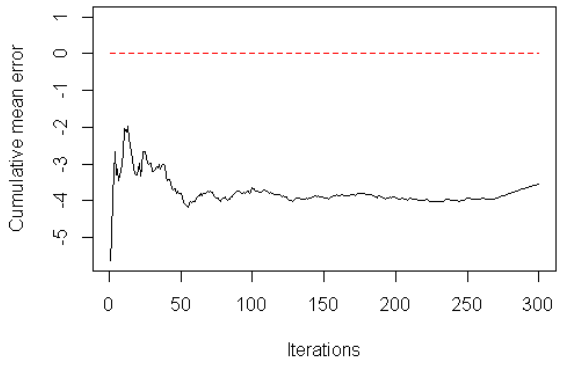
(c) Variance

Figure 4.5. Cumulative mean error of the posterior mean of the (a) mean, (b) precision and (c) variance using our model and a Gamma($\alpha = 0.01,\ \beta = 0.01$) prior for the precision. This was for the case of $n = 100$ and $K = 100$.

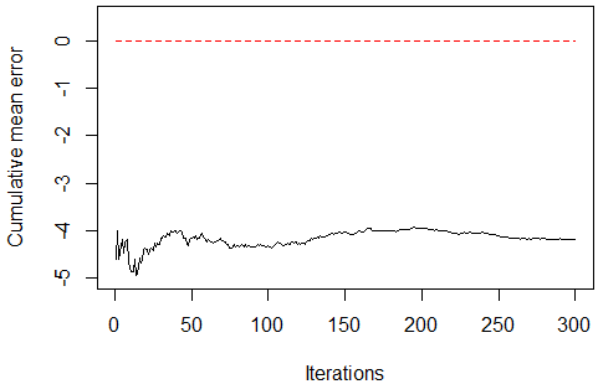
A total of 2,000 simulations were performed, unless otherwise stated in table captions. For the Bayesian nonparametric model we summarise the results when using a Gamma($\alpha = 1,600,\ \beta = 40,000$) and Gamma($\alpha = 0.1,\ \beta = 0.1$) prior for the precision. We report the ME (bias), MSE, coverage of the 95% credible interval and average credible interval width. Table 4.1 summarises the performance of the models in terms of their ability to recover the true population parameters, μ and σ^2. The nonparametric model leads to unbiased estimates of the population mean for both the informative and vague prior for the precision, for the values $n = 200$ and $K = 100$. The mean squared errors are small and coverage rates close to 95% for the 95% credible interval. However, when we assess the estimation of the variance,



(a) Variance for $n = 200, K = 100$



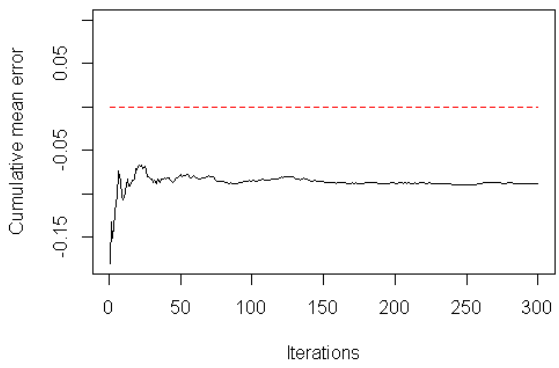
(b) Variance for $n = 100, K = 10,000$



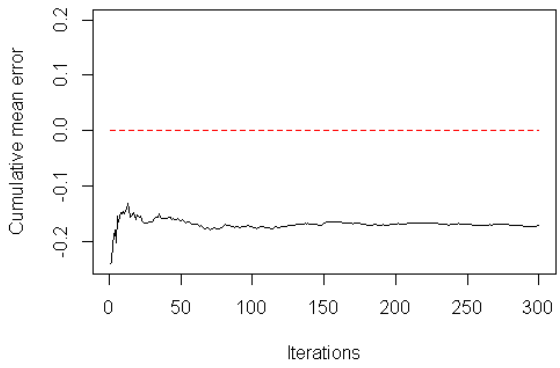
(c) Variance for $n = 200, K = 10,000$

Figure 4.6. Cumulative mean error of the variance when a vague $\tau \sim \text{Gamma}(0.1, 0.1)$ prior is used. The true variance is 25. The cumulative mean error converges to approximately -4 all cases, including when $n = 200$ and $K = 10,000$.

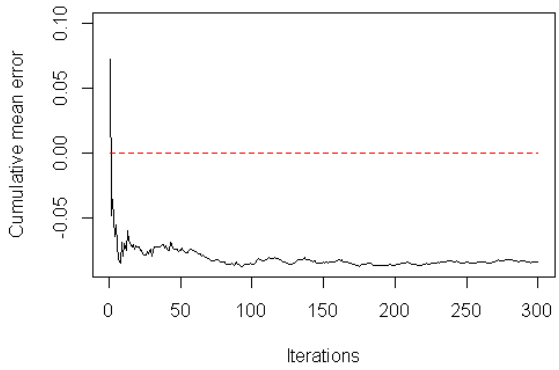
the Bayesian nonparametric model with an informative prior gives an unbiased estimate of the variance but there is overcoverage of the true population variance. The Bayesian nonparametric model with a vague prior for the precision resulted in a biased estimate of the variance of -4.043 and undercoverage of 70. The undercoverage is due to the bias in the estimate of the variance.



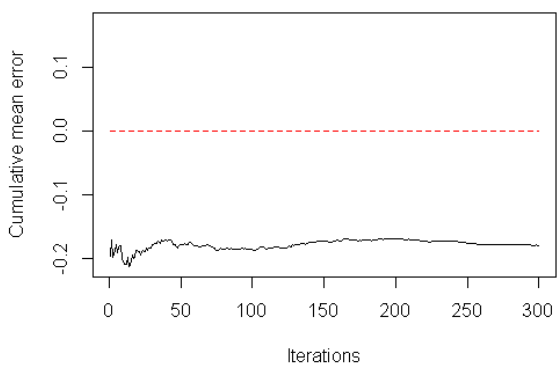
(a) Variance for $n = 100, K = 100$



(b) Variance for $n = 200, K = 100$



(c) Variance for $n = 100, K = 10,000$



(d) Variance for $n = 200, K = 10,000$

Figure 4.7. Cumulative mean error of the variance using an informative $\tau \sim \text{Gamma}(\alpha = 1,600,\ \beta = 40,000)$ prior with mean 4 and standard deviation 0.001. The cumulative mean error of the variance converges to approximately -0.08.

Table 4.2 summarises the performance of the models in terms of prediction. Both the Bayesian nonparametric model with an informative prior for the precision and the Bayesian naive analysis perform well in terms of bias and coverage. The average 95% credible interval width close to that of the precise measurements. However, the naive analysis has the advantage of not requiring knowledge of the population precision and still demonstrating satisfactory performance across all performance measures.

We also looked at the effect of varying the sample size K, to see if the naive approach would not perform as well for certain scenarios. Table 4.3 shows the naive analysis does not perform well when $K = 3$ as we observe a positive bias and undercoverage of the variance. However, the naive analysis appears to perform well for sample sizes $K = 20$, 30 and 50.

Table 4.1: Estimates of performance measures for the mean and variance for the three different approaches. The data-generating process was: $n = 200$ and $K = 100$. For the nonparametric model we compared the use of an informative prior for the precision Gamma($\alpha = 1,600$, $\beta = 40,000$) and a vague prior for the precision Gamma($\alpha = 0.01$, $\beta = 0.01$). Monte Carlo standard errors are reported in parentheses and obtained using the R package `rsimsum`.

Parameter	Performance measure	Bayesian nonparametric model		Naive analysis	Precise measurements
		Vague prior for τ	Informative prior for τ	Vague prior for τ	Vague prior for τ
μ	Bias	-0.001 (0.009)	-0.001 (0.008)	-0.001 (0.008)	-0.001 (0.008)
	MSE	0.121 (0.004)	0.121 (0.004)	0.121 (0.004)	0.121 (0.004)
	Coverage (%)	95.6 (0.5)	96.9 (0.4)	95.4 (0.5)	95.7 (0.5)
	Average 95% credible interval length	1.389	1.486	1.391	1.390
σ^2	Bias	-4.045 (0.056)	-0.001 (0.008)	-0.023 (0.056)	-0.062 (0.056)
	MSE	22.585 (0.465)	0.043 (0.001)	6.240 (0.195)	6.214 (0.193)
	Coverage (%)	70.6 (1.0)	100 (0.0)	94.8 (0.5)	95.1 (0.5)
	Average 95% credible interval length	9.963	2.377	9.995	9.980

Table 4.2: Simulation study results comparing performance estimates of the posterior predictive distribution for the four different models: (1) a Bayesian nonparametric model with a vague prior for τ, (2) a Bayesian nonparametric model with an informative prior for τ, (3) a Bayesian naive analysis and (4) a Bayesian analysis with precise measurements.

Performance measure	Bayesian nonparametric model with vague prior for τ	Bayesian nonparametric model with informative prior for τ	Naïve analysis with vague prior for τ	Precise measurements with vague prior for τ
Bias	-0.045 (0.008)	-0.005 (0.009)	0.077 (0.008)	0.003 (0.009)
MSE	25.654 (0.785)	25.902 (0.855)	24.799 (0.803)	23.989 (0.761)
Coverage	91.3 (0.6)	94.1 (0.5)	95.1 (0.5)	95.5 (0.5)
Average 95% credible interval length	18.027	19.497	19.655	19.619

4.6 Example: double sigmoidal growth model

Since we have shown that we can obtain accurate estimates using a naive analysis for the case of a sample of values, we will use this approach for the double sigmoidal growth model and analyse the results.

We use the 2018 grape bunch mass data described in Section 3.7 collected over the 2017/2018 growing period. However, this time we simulate data with uncertainty, where we observe a sample of values with variability within the sample, instead of precisely measured observation. We simulated data using the true grape bunch mass data. The data is generated as follows:

$$d_{ik} \sim \mathrm{N}(\log(y_i), u_i) \quad \text{for } k = 1, \ldots, K.$$

Table 4.3: Assessing the performance of the naive analysis, where we let $y_i = \overline{\mathbf{d}_i}$, considering data-generating mechanisms, K= 3, 20, 30 and 50 for estimates of the mean and variance. Monte Carlo standard errors are reported in parentheses and obtained using the R package `rsimsum`. The setting of $n = 200$ is used for the data-generating process. A total of 300 simulations were performed.

Parameter	Performance measure	K= 3	K = 20	K= 30	K = 50
μ	Bias	-0.0239	-0.0227	-0.0206	-0.0193
		(0.0203)	(0.0199)	(0.0199)	(0.0199)
	MSE	0.1238	0.1190	0.1188	0.1182
		(0.0108)	(0.0099)	(0.0098)	(0.0098)
	Coverage (%)	95.7 (1.2)	94.7 (1.3)	94.3 (1.3)	94.7 (1.3)
	Average 95% credible interval length	1.428	1.395	1.393	1.393
σ^2	Bias	1.2870	0.1770	0.0795	0.0113
		(0.1516)	(0.1440)	(0.1451)	(0.1443)
	MSE	8.5279	6.2278	6.3034	6.2252
		(0.7175)	(0.4759)	(0.4832)	(0.4684)
	Coverage (%)	91.0 (1.7)	96.0 (1.1)	95.3 (1.2)	96.0 (1.1)
	Average 95% credible interval length	10.515	10.069	10.036	10.008

First we take the natural logarithm of the actual grape bunch mass y_i to obtain, $\log(y_i)$. For each $\log(y_i)$, we generate K values from a normal distribution with mean $\log(y_i)$ and precision u_i. The uncertainty u_i was the same for all measurements. We analyse the results using the setting where the uncertainty is 50% of the true values: $K = 100$ and $u_i = 0.5$. This was because, we thought that if the model will work for the case where the uncertainty is 50% of the true values, it will work for the cases where the uncertainty is smaller than

50%.

The simulated data represented as box plots for three out the total 14 days of data collected is shown in Figure 4.8. Figure 4.9 represents the simulated data as half violin plots, using kernel density estimation (a smoother) to show the distributions of the data.

The results show that the naive analysis leads to predictions that are very close to their precisely measured counterpart, when there is the same prior information. When the uncertainty is 50% of the true values, the 95% posterior predictive envelopes in Figure 4.11 are very similar for both the Bayesian naive analysis and the Bayesian model with precisely measured grape bunch masses. The estimated posterior predictive distributions are very similar for both models, and we see that the 95% posterior predictive intervals (HDI) are very similar for both models; (97.47, 391.66) and (98.93, 397.40) respectively. This means that the Bayesian naive analysis can produce accurate and precise estimates in the face of considerably large uncertainty of 50% of the actual values.

4.7 Discussion

In our investigation, we found that in the situation that we had a sample of values representing a nonparametric distribution instead of a precise measurement, the Bayesian naïve analysis had the best performance. It produced the best estimates of the population parameters and gave the most accurate and precise predictions. If the mean of the sensor measurement, $\overline{\mathbf{d_i}}$ is an unbiased estimate of the true grape bunch mass y_i, then it does not matter how large the uncertainty is, you can still obtain accurate and precise (small variance) predictions. Figure 4.3 is from exploratory data analysis of the simulated data and illustrates that the sample means are quite close to the true values of y_i. An advantage of the naive analysis is that we can have different sample sizes, K, for each observation, as long as the sample mean is an unbiased and sufficient estimator.

Furthermore, the naive analysis is still expected to work even if the sample $\mathbf{d_i}$ does

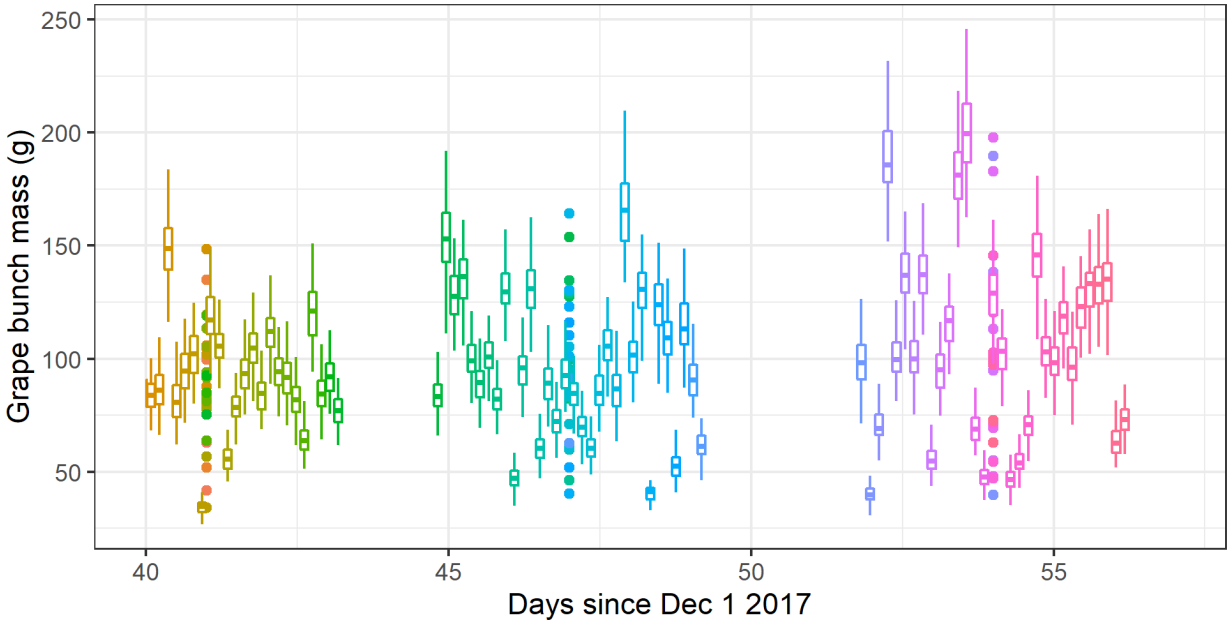


Figure 4.8. Scatterplot of the 2018 grape bunch mass data with corresponding simulated nonparametric data (a sample of values for each data point) distribution displayed as boxplots. The uncertainty is set to $u_i = 0.1^{-2}$. The colours of the boxplots correspond to the original data points. The plot is only showing three out of 14 days of data because it is difficult to clearly display the simulated data for all days in the dataset.

not come from a normal population; as long as the sample size is large enough that the central limit theorem can be applied. According to the central limit theorem, the sampling distribution of the mean is approximately a normal distribution for large enough K (where K here is our sample size) even if the original variables themselves do not come from a normal population. The naive analysis works under particular conditions. For the naive analysis to work, the sample mean needs to be an unbiased estimator of the actual value y_i . If the sample mean is a biased estimator of the actual value, then the model will result in inaccurate predictions. This could be checked, for example, as part of calibration of the sensor; checking that the mean of the observed sample is equal or very close to the actual grape bunch mass.

A possible explanation as to why the Bayesian nonparametric model does not work

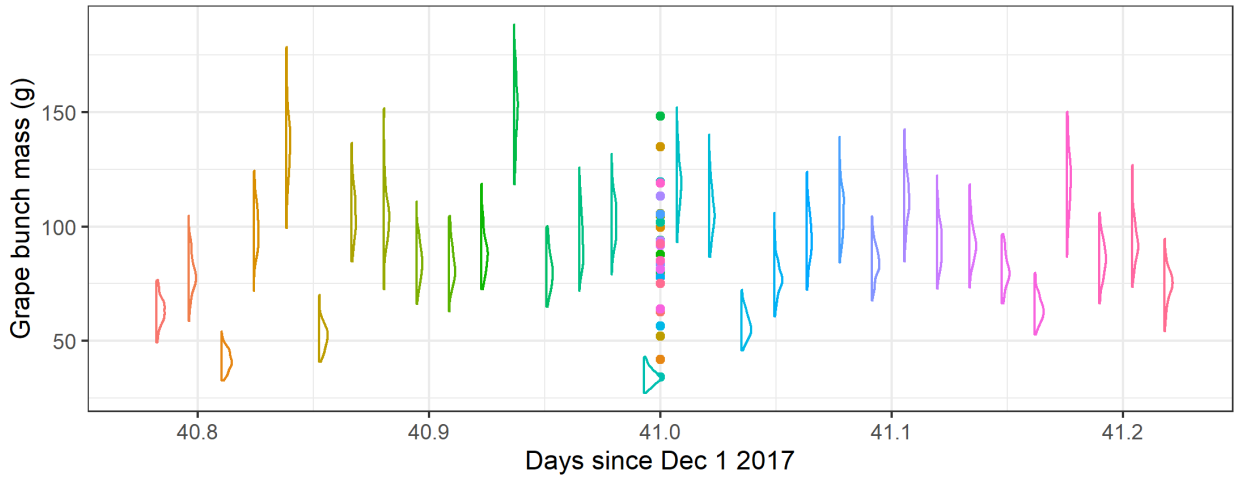


Figure 4.9. Illustrating the distributions of the simulated data and displaying the 2018 grape bunch mass data for a single day (day 41). The simulated samples are displayed as half violin plots (kernel density estimation i.e. smoother applied). Uncertainty was set to $u_i = 0.1^{-2}$. The colours of the distributions representing each sample correspond to the colours of the original data points. Data is only displayed for a single day as it is difficult to clearly display the simulated data for all days in the dataset.

so well, could be because of the problem of identifiability. That is, the model is unable to uniquely identify the parameters τ and u_i. Gustafson (2003) states that parameter identifiability is often an issue for models which account for mismeasurement in variables, "since a great deal must be known about the mismeasurement process in order to obtain a fully identified model" (p. 152). However, the advantage of Bayesian models is that we can often deal with the issue of non-identifiability by using an informative prior given we have knowledge of the parameter. This is what we have done here. However, a danger is that unless the guess is correct, the model will be somewhat misspecified (Gustafson, 2003).

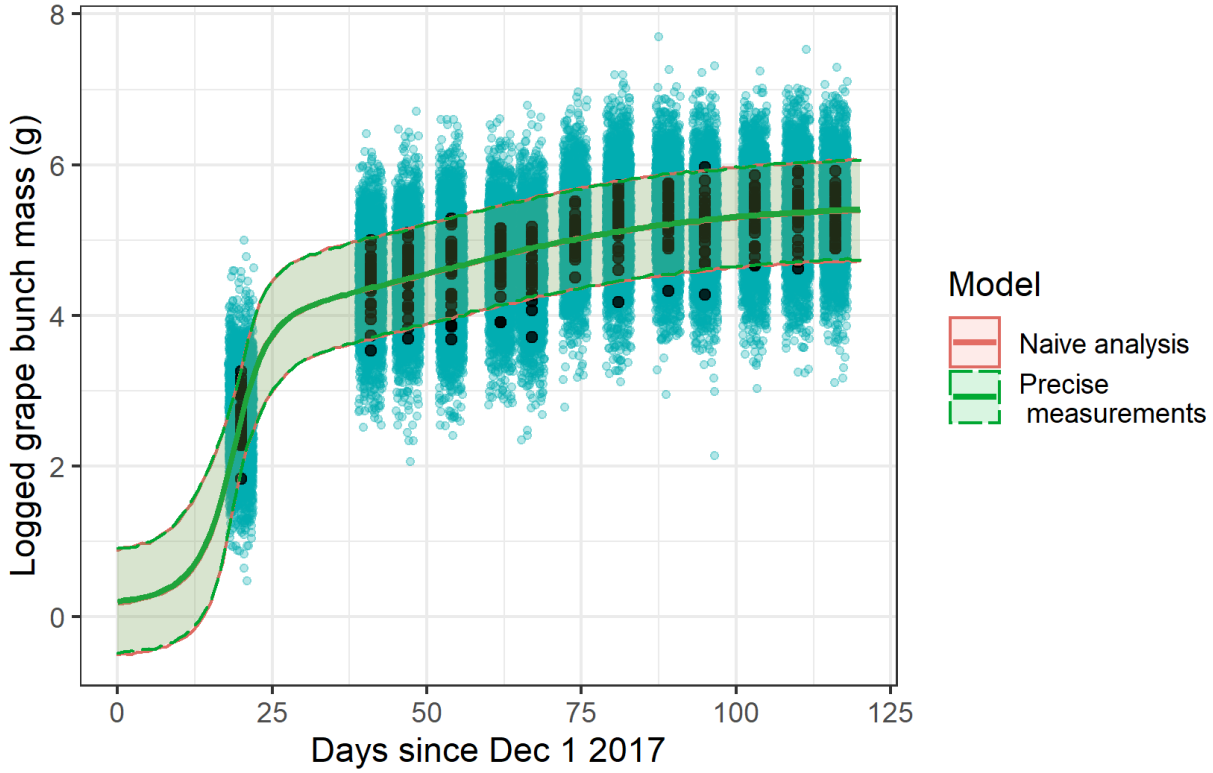


Figure 4.10. Double sigmoidal curve comparing the naive analysis and precise measurements. The 2018 bunch mass data are represented by the black points and the simulated data are represented by the blue points. The setting used for the uncertainty is $u_i = 0.5^{-2}$. The grape bunch masses are plotted on the log scale.

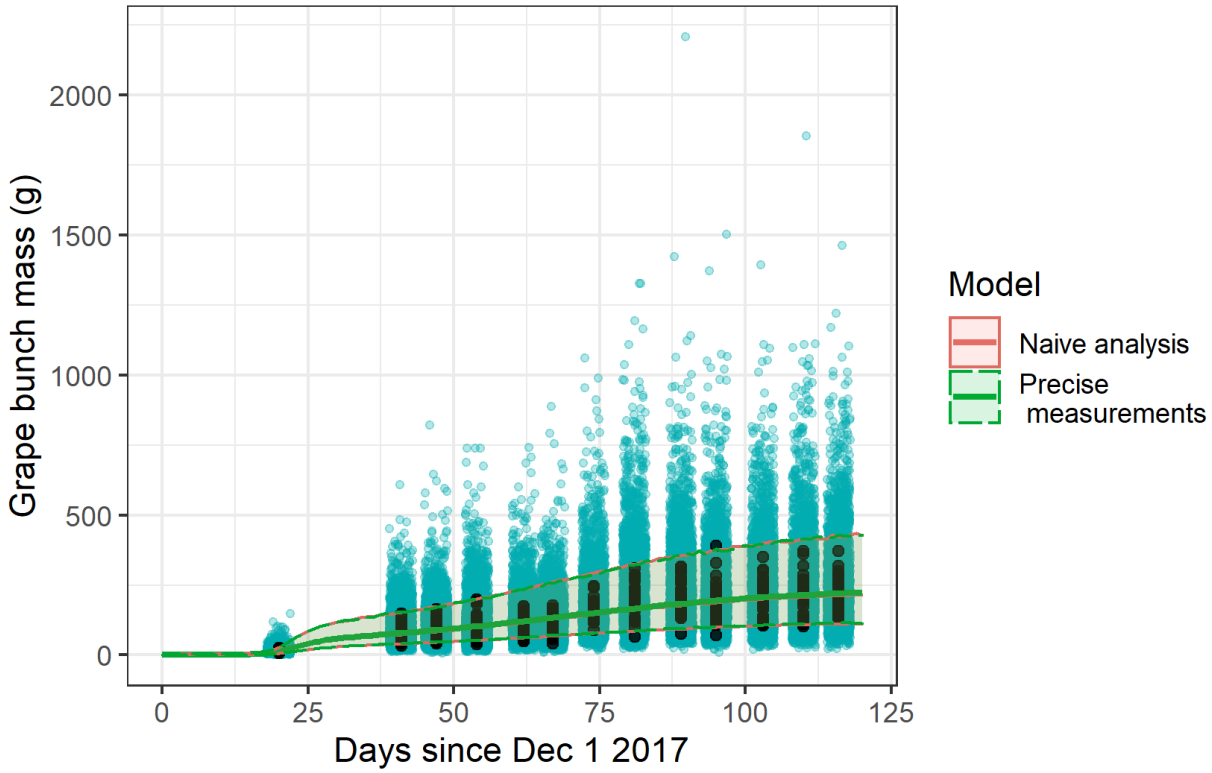


Figure 4.11. Double sigmoidal curve comparing the naive analysis and precise measurements fitted with 2018 bunch mass data represented by the black points and simulated samples representing nonparametric distributions are represented by the blue points. The uncertainty used here is $u_i = 0.5^{-2}$.

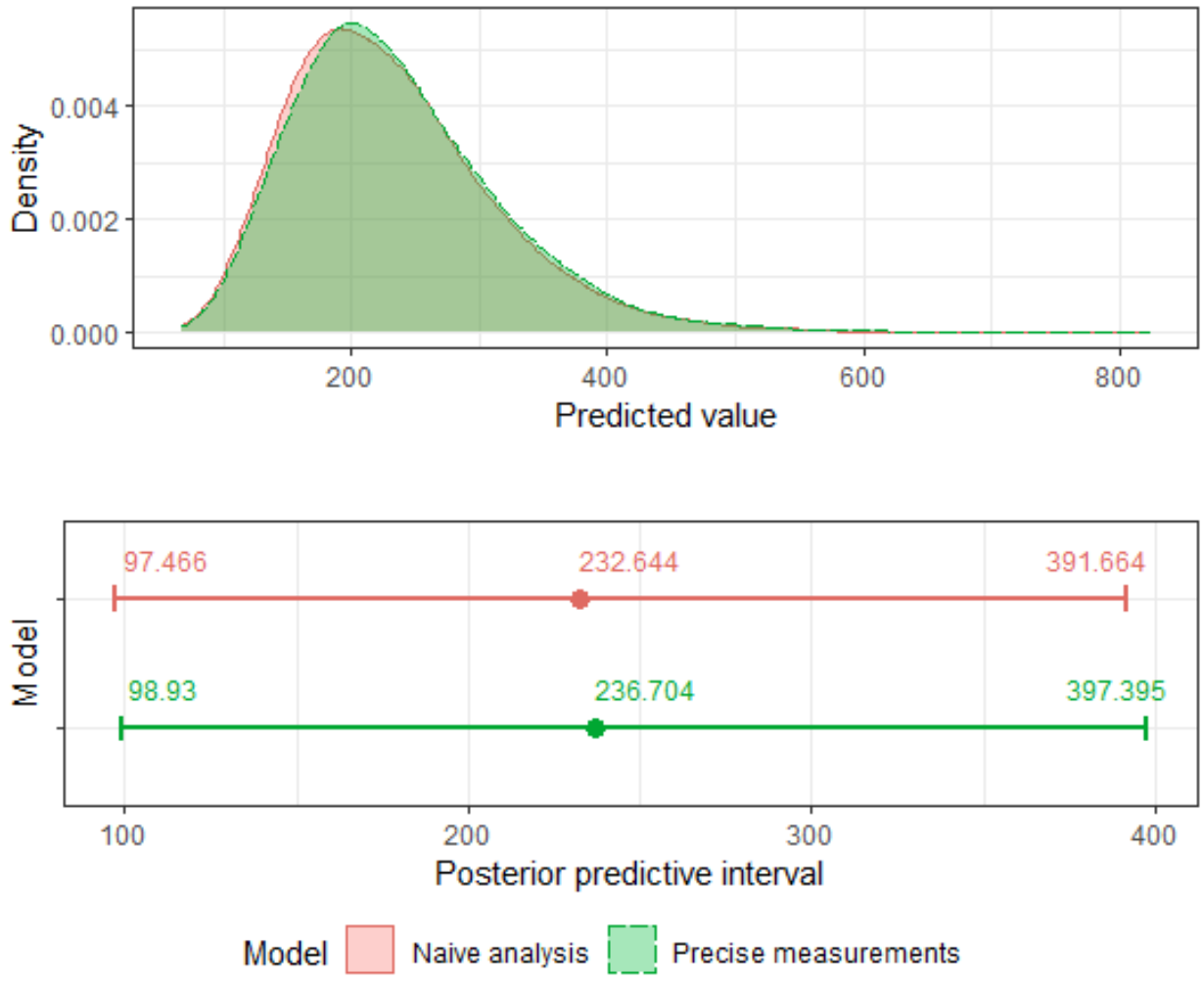


Figure 4.12. Double sigmoidal curve with simulated nonparametric data using the setting $u_i = 0.5^{-2}$ comparing naive analysis and precise measurements and their estimated posterior predictive distribution at day 120. Below the plot of the densities are the posterior predicted means and 95% credible intervals.

Chapter 5

Conclusion and Discussion

5.1 Summary

The aim of this study was to investigate methods to account for stated uncertainty in response variables with the ultimate goal of applying the methods to a Bayesian double sigmoidal growth model used to predict grape yield. We considered both the situation where the measurements are reported with uncertainty assuming a normal error distribution, and the situation of having a sample of values representing a nonparametric distribution, instead of a precisely measured observation. Additionally, we studied the impact of ignoring uncertainty in a continuous response variable on the inference of population parameters and for prediction. We constructed MCMC algorithms to incorporate continuous measurements with uncertainty in the response variable and implemented them in R. We conducted simulation studies to assess the performance of our models and compare them with a naive analysis and their precisely measured counterparts.

We found that if the true data-generating process which generates the measurements with uncertainty from the true values of $\mathbf{y}$ is purely random and normal, then the Bayesian model which incorporates uncertainty can produce accurate and precise (minimal variability)

predictions. This was when the uncertainty was the same for all measurements. However, the model starts to deteriorate when the uncertainty approximately larger than 80% of the population variation.

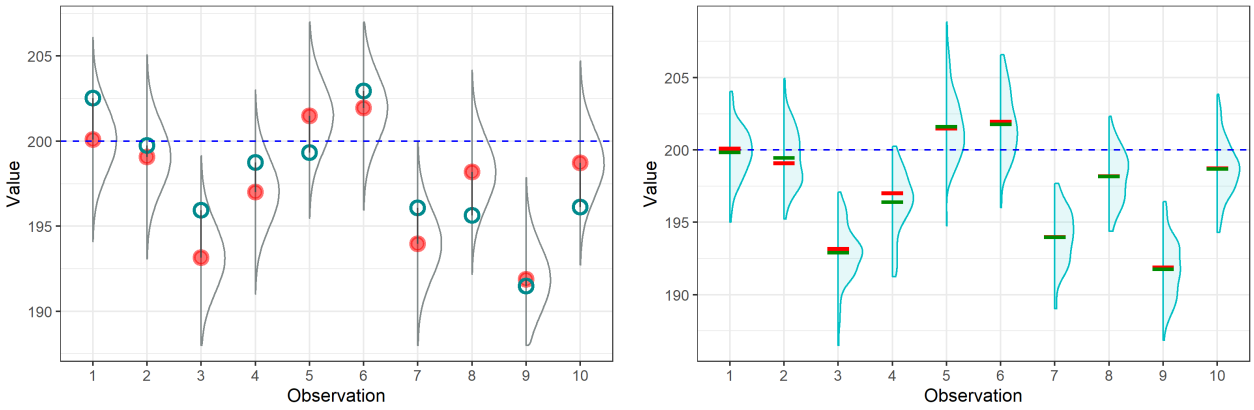
On the other hand, if we have a sample of values instead of a precisely measured data point, and we make no assumptions about the distribution of the data, then the naive analysis produces the best estimators given that the mean of each sample is an unbiased estimate of the actual value.

5.2 Discussion

The reason, when we are given measurements $\mathbf{m}$ and their stated uncertainty $\mathbf{u}$, the naive analysis performs poorly (we observe overcoverage of the population variance and posterior predictive distribution), however a naive analysis performs well when we observe a sample of values instead of a precisely measured value is due to the two different assumed data-generating mechanisms. In Chapter 3 our model for observing our measurements is $m_i \mid y_i, u_i \sim \mathrm{N}(y_i,\ u_i)$. We only observe one m_i for each y_i. The variance of the vector $\mathbf{m}$ would be expected to be larger than the vector $\mathbf{y}$. This is the reason why the naive analysis results in wider prediction intervals. However, in Chapter 4, we observe a sample of values instead of a precisely measured value. The data-generating mechanism for generating our observed sample is $d_{ik} \sim \mathrm{N}(y_i,\ u_i)$. So the expected value of our sample $\mathbf{d_i} = (d_{i1}, \ldots, d_{iK})$ would be $\mathrm{E}[\mathbf{d_i}] = y_i$. Figure 5.1 illustrates these two different assumptions for the simple situation of a sample of values.

5.2.1 Contributions

Our Bayesian models for incorporating uncertainty in the response (assuming normal error) and our study into the implications of ignoring the uncertainty contribute to the field of measurement error. Within the field of measurement error, our focus has been on



(a) Data-generating process for the models in Chapter 3

(b) Data-generating process for the models in Chapter 4

Figure 5.1. An illustration of the two different data-generating processes for the sample of values situation in Chapter 3 (left) and Chapter 4 (right). Both (a) and (b) have the same true values shown in red, however the measured data that is observed shown in blue differs between the two plots. In (a) the filled red circles are the true values and the blue circles are the measured (observed) values. In (b) the red lines are the true values, the densities show the distributions of the samples observed, the green lines represent the means of each sample. In both (a) and (b) the horizontal dashed line represents the true population mean.

Bayesian approaches that deal with response variable error assuming a classical measurement error model with a focus on performance in terms of predictive inference compared with an emphasis on estimating population parameters in the work by Gustafson (2003) and Carroll et al. (2006). We investigated the implications of ignoring uncertainty (the naive analysis) in the response variable in a Bayesian framework. We explored in detail how large the uncertainty can be before the model breaks down when the model describing the relationship between $\mathbf{m}$ and $\mathbf{y}$ is a normal distribution, which appears to be 80% of the population variation. We specify MCMC algorithms, specifically Gibbs samplers to implement our models that are straightforward to implement and are not computationally intensive. We have investigated the impact of the naive analysis on predictive inference. When the generation of the data with uncertainty is described by a probabilistic model,

ignoring the error can lead to an overestimation of the population variance and less precise predictions. We found the full conditional distribution can be analytically derived when the distribution assumed for $p(\mathbf{m} \mid \mathbf{y})$ is normal and the likelihood for the model, $p(\mathbf{y} \mid \theta)$ is normal. We also investigated the impact of a naive analysis when there is uncertainty in the response. Our simulation studies showed that a naive analysis results in an unbiased estimate of the population mean, however there is a positive bias in the estimate of the population variance and our predictions also have greater uncertainty. This agrees with Gustafson (2003) where he states that in the context of a simple linear regression with variables $\mathbf{x}$ and $\mathbf{y}$, adding noise to $\mathbf{y}$ "does not shift the estimated regression relationship in a systematic manner, though it does increase the inferential uncertainty about the ($\mathbf{x}$, $\mathbf{y}$) relationship" (Gustafson, 2003, p. 4).

5.2.2 Implications

Our findings have implications for the development of sensors which measure the mass of grape bunches for grape yield prediction. If there is systematic error in the sensor output then the model will result in inaccurate predictions. We recommend that this should be considered in the development of sensors. Another implication is that if the true data generating process of the sensor is a normal random error, then the uncertainty cannot be too large or our Bayesian model which incorporates uncertainty will result in undercoverage of the predictions. From our simulation studies, we found that when the data generating process is a normal random error and the uncertainty is larger than 80% of the population variation, then performance of our Bayesian model which incorporates uncertainty deteriorates and we observe undercoverage. This could be a used as a guideline for the development of the sensor. However, if the sensor produces a sample of values for the mass of a single grape bunch, and the mean of the sample is an unbiased estimate of the true bunch mass, then we can just use the mean of the sample and it does not matter how large the uncertainty is; the model can still produce accurate and precise predictions.

Our recommendation for those who are wanting to incorporate sensor data with uncertainty into a regression model is to think carefully about the data-generating mechanism and choose a model based on the most realistic mechanism for generating the observed data. The two mechanisms presented in this thesis are: i) a random measurement error in Chapter 3 and ii) a distribution of values whose mean is relatively close to the actual value in Chapter 4. The recommended model from Chapter 3 is the Bayesian model which incorporates uncertainty presented and the recommended model from Chapter 4 is the naive analysis which takes the mean of each observed sample. Both these recommended models produce the most accurate (unbiased) and precise (minimal variance) predictions.

The modelling framework developed here is a step towards the integration of sensor data with the grape yield growth model described in Ellis et al. (2020) and on the way to a fully developed tool that can help NZ grape growers and winemakers. Having tool for the wine industry to predict grape yield early on in the season to help with optimising human resources and equipment during the harvest (Henry et al., 2019). The ability to use sensors instead of manually measuring grape bunches could mean there is less destructive sampling of grape bunches. In the long term, the use of sensors can improve profitability of vineyards and wineries. Being able to incorporate additional data in the form of measurements with uncertainties can be beneficial to the grape yield prediction model or other crop growth models as more data can improve the accuracy of the predictions (Liu et al., 2019).

The modelling framework described here could be useful for other fields such as horticulture or agriculture where sensors are being developed to measure fruit e.g. apples and kiwifruit. They could also be useful in in other areas of research where sensors are used to take measurements on a continuous scale, such as in medicine or chemistry.

The models described here can be used in any field where there is measurement uncertainty in a continuous outcome variable. Some examples of variables on a continuous scale are height, income and dietary intake. We hope that the models described here will be useful for practitioners seeking to fit regression models with uncertainty in a continuous outcome variable.

5.2.3 Limitations

Simulation studies are experiments which are a useful tool in statistics allowing us to obtain empirical results about the performance of statistical methods in certain scenarios (Morris et al. (2019)). Since we only assess the resilience of our models under a finite number of situations, we cannot make complete generalisations. Thus, for the case that a normal error distribution is assumed it is difficult to recommend an exact threshold where the size of the uncertainty in the measurement would cause the model's performance to deteriorate.

We did not explore the effects of model misspecification and its impact on parameter estimation and thus prediction. For example, if the model describing the relationship between m_i and y_i was another parametric distribution such as a t-distribution. For the Bayesian nonparameteric model, we could investigate its performance when, for example, the data-generating process is a mixture of normal distributions.

In Section 3.7, when we applied our approach to incorporating uncertainty for the double sigmoidal growth model, we simulated measurements with uncertainty and analysed the data. We provided a possible explanation that the reason in Figure 3.14 the envelope for the model incorporating uncertainty is more narrow than the precise measurements is because of undercoverage of the model. We suspected this is highly likely, because this phenomenon was observed in the results for our simulation studies for the situation of a sample of measurements of uncertainty and the simple linear regression case. However, we did not conduct a simulation study to assess the performance of our model for the double sigmoidal growth curve, due considerations of computation cost, so we cannot make a definitive statement about whether there is undercoverage.

5.2.4 Future directions

The development of models which incorporate aggregated data with uncertainty would be the next step. This will enable the incorporation of sensor data where the combined mass of grape bunches on an interval has been measured.

Another possible area for further research could include a detailed exploration of other parametric error distributions, such as a t-distribution or triangular distribution, and an evaluation of their performance in terms of estimating population parameters and predictive inferences.

Future work could explore the implications of model misspecification. That is, how well would the model estimate the true parameters if the true data-generating process differed from the model. For example, if the true relationship between the measured value m_i and the true value y_i is a t-distribution, but the model assumes a normal distribution. We have considered the special case where the size of the uncertainty is the same for all measurements, that is $u_i = u$. Further research could explore impacts and adjustments for of measurement specific uncertainty. It could involve seeing if logging the response is an adequate solution since typically in regression models taking the log of the response reduces heteroscedasticity. In Chapter 4, the performance of the model when the simulated data comes from a mixture of normal distributions or another distribution, e.g. a t-distribution or uniform distribution could be explored in detail.

We have shown in Chapter 3 when applying our Bayesian model which incorporates uncertainty to a double sigmoidal curve, that for a single simulation, when the uncertainty is 50% of the actual grape bunch weight, then our model appears to show undercoverage in its prediction, when we compared the 95% posterior predictive envelopes to the Bayesian model with precise measurements. Future work could include conducting simulation studies to evaluate our model by generating data by repeated sampling with replacement of the dataset to further assess the impact of uncertainty on the performance of our model.

Bibliography

Abrevaya, J., & Hausman, J. A. (2004). Response error in a transformation model with an application to earnings-equation estimation. *The Econometrics Journal*, *7*(2), 366–388.

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. Boca Raton, CRC Press.

Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement error in nonlinear models* (1st, Vol. 63). London, Chapman & Hall.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Boca Raton, FL: Chapman & Hall/CRC press.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data analysis: An introduction for scientists and statisticians*. Boca Raton, FL : CRC Press.

Eccleston, K. W., Platt, I. G., Jafari, A., Werner, A., Bateman, C., Woodhead, I. M., Fourie, J., Hsiao, J. W. H., & Carey, P. (2019). Observations from radar scans of grape vines conducted over a growing season, In *2019 IEEE Conference on Antenna Measurements & Applications (CAMA)*. IEEE.

Eccleston, K. W., Platt, I. G., & Tan, A. E.-C. (2018). SAR for grape bunch detection in vineyards, In *2018 Australian Microwave Symposium (ams)*. IEEE.

Ellis, R., Moltchanova, E., Gerhard, D., Trought, M., & Yang, L. (2020). Using Bayesian growth models to predict grape yield. *OENO One*, *54*(3), 443–453. https://doi.org/10.20870/oeno-one.2020.54.3.2972

Gelman, A. B., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. (2014). *Bayesian data analysis (3rd ed.)* Boca Raton, FL: CRC press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. CRC Press.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Henry, D., Aubert, H., & Véronèse, T. (2019). Proximal radar sensors for precision viticulture. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(7), 4624–4635.

JCGM. (2008). Evaluation of measurement data—guide to the expression of uncertainty in measurement. *JCGM*, *100*(2008), 1–116.

Kirkup, L., & Frenkel, B. (2006). *An introduction to uncertainty in measurement using the GUM (guide to the expression of uncertainty in measurement)* (Vol. 9780521844284). Cambridge, Cambridge University Press.

Liu, C.-A., Chen, Z.-X., Yun, S., Chen, J.-S., Hasi, T., & Pan, H.-Z. (2019). Research advances of SAR remote sensing for agriculture applications: A review. *Journal of Integrative Agriculture*, *18*(3), 506–525.

Martin, S., Dunstone, R., & Dunn, G. (2003). How to forecast wine grape deliveries using grape forecaster excel workbook version 7. *GWRDC, Adelaide, Australia*, 100.

McElreath, R. (2015). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, CRC Press/Taylor & Francis Group.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan (2nd ed.)* CRC press.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Parr, B., Legg, M., Alam, F., & Bradley, S. (2020). Acoustic identification of grape clusters occluded by foliage, In *2020 IEEE Sensors Applications Symposium (sas)*. IEEE.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rabinovich, S. G. (2005). *Measurement errors and uncertainties: Theory and practice (3rd ed.)* New York, AIP Press.

Scaccia, L., & Green, P. J. (2003). Bayesian growth curves using normal mixtures with nonparametric weights. *Journal of Computational and Graphical Statistics*, *12*(2), 308–331.

Stan Development Team. (2015). *Stan: A C++ library for probability and sampling, version 2.5.0*. https://mc-stan.org/

Tan, A. E.-C., Riding, P., Eccleston, K. W., Platt, I. G., Werner, A., & Woodhead, I. M. (2019). Cavity backed antenna for microwave estimation of grape bunches in vineyard, In *2019 IEEE Asia-Pacific microwave conference (APMC)*.

Appendix A

Gibbs sampler for a sample of values with a normal error in R

This appendix provides an implementation of the Gibbs sampler in R to sample from the posterior distribution for a sample of values measured with uncertainty under the assumption of a normal error.

We will start by simulating some data for the Gibbs sampler.

```
# Setting the random number seed for reproducible results
set.seed(10)

n = 100
true_mu <- 200
true_sd <- 5
true_tau <- 1/((true_sd)^2)
u_sd = 2
u <- 1/(u_sd^2)

y <- rnorm(n, mean = true_mu, sd = true_sd)
m <- rnorm(n, mean = y, sd = u_sd)

# Prior parameters
```

```
tau_0 <- 1/10000
mu_0 <- 0
```

The Gibbs sampler code starts here.

```
# Set the number of iterations for the Gibbs sampler
N <- 5000

# Initialising an empty matrix to store results
posterior_store <- matrix(numeric(0), nrow = N, ncol = 4)
colnames(posterior_store) <- c("mu", "tau", "y_1", "y_2")

# Initial values
mu <- mean(m)
tau <- 1/var(m)
y <- rep(200, times = n)

for (i in 1:N){

  # Sample the vector of y values
  y <- rnorm(n, mean = ((u * m) + (tau * mu))/(u + tau),
             sd = 1 / sqrt(u + tau))

  # Sample mu from its full conditional distribution
  mu <- rnorm(1, mean = (tau*n*mean(y) + tau_0*mu_0)/(tau*n+tau_0),
              sd = 1/sqrt(tau*n + tau_0))

  # Sample tau from its full conditional distribution
  tau <- rgamma(1, shape = 0.001 + n/2, rate = 0.001 + 0.5*sum((y-mu)^2))

  posterior_store[i,] <- c(mu, tau, y[1], y[2])

}

# Checking for convergence
plot(mcmcr::as.mcmc(posterior_store[,c(1,2)]))
plot(mcmcr::as.mcmc(posterior_store[,c(3,4)]))

# Discard burn-in period
```

```r
posterior_sample <- posterior_store[-c(1:500),]

# Checking for convergence
plot(mcmcr::as.mcmc(posterior_sample[,c(1,2)]))
plot(mcmcr::as.mcmc(posterior_sample[,c(3,4)]))

# Check posterior means
posterior_means <- apply(posterior_sample, 2, mean)
posterior_means
```

Appendix B

Gibbs sampler for a simple linear regression with a normal error in R

This appendix provides an implementation of the Gibbs sampler in R to sample from the posterior distribution for simple linear regression with values measured with uncertainty under the assumption of a normal error.

We will start by simulating some data for the Gibbs sampler.

```
# Setting the random number seed for reproducible results
set.seed(10)

n <- 100
x <- seq(from = 10, to = 200, length.out = n)
alpha_true = 0
beta_true = 2
sd_true = 10

y <- rnorm(n = length(x), mean = alpha_true + beta_true * x,
           sd = sd_true)

u_sd <- 5
u_tau <- 1/(u_sd)^2 # uncertainty defined by the precision
```

```
m <- rnorm(n, mean = y, sd = u_sd)

# Vague priors for mu and tau

mu_alpha <- mu_beta <- 0
tau_alpha <- tau_beta <- 1 / 10000
a <- 0.001 ; b <- 0.001
```

The Gibbs sampler code starts here.

```
# Initial values for the Gibbs sampler using estimates from
# frequentist linear regression

alpha_initial <- summary(lm(m~x))$coef[1]
beta_initial  <- summary(lm(m~x))$coef[2]
tau_initial <- 1 / (summary(lm(m~x))$sigma)^2

alpha <- alpha_initial
beta <- beta_initial
tau <- tau_initial
y <- rep(mean(m), n)

number_of_iterations <- 2000

# Initialising an empty matrix to store results

number_columns <- 5
posterior_result <- matrix(data = numeric(number_columns),
                           nrow = number_of_iterations,
                           ncol = number_columns,
                           byrow = TRUE)
colnames(posterior_result) <- c("alpha", "beta", "tau", "y_1", "y_n")


for (i in 1:number_of_iterations){

  # Sample the vector of y values
  y <-rnorm(n, mean = ((u_tau) * m + (tau*(alpha + beta * x))) /
              (u_tau + tau),
```

```
            sd = 1/sqrt(u_tau + tau))

  z <- y - (beta * x)
  alpha <- rnorm(n = 1,
                 mean = ((tau * sum(z)) + (tau_alpha * mu_alpha)) /
                   ((n * tau) + tau_alpha),
                 sd = sqrt(1/((n * tau) + tau_alpha))
  )

  # Sample beta from its full conditional distribution
  w <- y - alpha
  beta <- rnorm(n = 1,
                mean = (tau * sum (x * w) + (tau_beta * mu_beta)) /
                  (tau * sum (x ^ 2) + tau_beta),
                sd = sqrt(1/(tau * sum (x ^ 2) + tau_beta)))


  # Sample tau from its full conditional distribution
  shape_input <- a + (n / 2)
  rate_input <- b + (1 / 2)*sum((y - alpha - (beta * x))^2)
  tau <- rgamma(n = 1, shape = shape_input, rate = rate_input)

  posterior_result[i,] <- c(alpha, beta, tau, y[1], y[n])


}

# Discard burn-in period of 500 iterations
posterior_sample <- as.data.frame(posterior_result[-c(1:500),])

# Check for convergence
plot(mcmcr::as.mcmc(posterior_sample))

# Analyse results

# Posterior means
posterior_means <- sapply(posterior_sample, mean)
posterior_means

# Posterior standard deviations
```

```r
posterior_sds <- sapply(posterior_sample, sd)
posterior_sds

# 95% Credible intervals
credible_intervals <- sapply(posterior_sample, function(x){
  quantile(x, probs = c(0.025, 0.975))
})
credible_intervals
```

Appendix C

Gibbs sampler for the Bayesian nonparametric model in R

This appendix provides an implementation of the Gibbs sampler in R to sample from the posterior distribution for the Bayesian nonparametric model described in Section 4.3.

We start by simulating some data for the Gibbs sampler.

```
# Setting the random number seed for reproducible results
set.seed(10)

# Using a true population mean of 200 and sd of 5 (precision of 0.4)
n <- 500
mean_true  <- 200
sd_true <- 5
precision_true <- 1/(sd_true)^2
mu_prior_mean <- 0
mu_prior_precision <- 10^-5
y_true_vector <-rnorm(n, mean = mean_true, sd = sd_true)

uncertainty <- 2

# Let there be K values in each sample
```

```
K <- 1000

# Create the list of samples
data_list <- lapply(y_true_vector, function(y_value){
  sample_values <- rnorm(K, mean = y_value, sd = uncertainty)
  return(sample_values)
})

names_list <- sapply(i <-1:n, function(number){
  return(paste0("sample_", number))
})

names(data_list) <- names_list
```

The Gibbs sampler code starts here.

```
# Set the number of iterations for the Gibbs sampler
number_of_iterations <- 10000

# Initialising an empty matrix to store the results
posterior_store <- matrix(numeric(0),
                          nrow = number_of_iterations,
                          ncol = 4)
colnames(posterior_store) <- c("mu", "tau", "y_1", "y_2")

# Initial values:
mu <- mean(unlist(data_list, use.names = FALSE))
y <- sapply(data_list, mean)
tau <- 1/var(unlist(data_list, use.names=FALSE))

for (i in 1:number_of_iterations){

  # 1. Sampling from p(y_i|d_i, mu, tau) which is the
  # equal to the normalised likelihood for a single y_i.
  # For each sample, get the likelihood for each value
  # in the sample then sample from the likelihood
  y <- vapply(data_list, function(mass_sample){

    probabilities <- dnorm(mass_sample, mean = mu,
```

```
                              sd = 1/sqrt(tau))
    y_single_bunch <-sample(mass_sample,
                             size = 1, prob = probabilities)

  }, FUN.VALUE = numeric(length = 1))

  # 2. Sample mu from its full conditional distribution
  mu <- rnorm(1, mean = (tau*n*mean(y) +
                              mu_prior_mean*mu_prior_precision)/
                (tau*n+mu_prior_precision),
              sd = 1/sqrt(tau*n + mu_prior_precision))

  # 3. Sample tau from its full conditional distribution
  tau <- rgamma(1, shape = 0.001 + n/2,
                rate = 0.001 + 0.5*sum((y-mu)^2))

  posterior_store[i, ] <- c(mu, tau, y[1], y[2])

}

# Checking for convergence
plot(mcmcr::as.mcmc(posterior_store[ ,1:2]))
plot(mcmcr::as.mcmc(posterior_store[ ,3:4]))

# Discard burn-in period of 1000 iterations
posterior_sample <- posterior_store[-(1:1000),]

# Make the burn-in 1000 iterations
plot(mcmcr::as.mcmc(posterior_sample[ ,1:2]))
plot(mcmcr::as.mcmc(posterior_sample[ ,3:4]))

summary(posterior_store[-(1:1000),])
```

Appendix D

Metropolis-Hastings algorithm for the double sigmoidal curve for measurements with uncertainty (normal error) in R

This appendix provides the Metropolis-Hastings algorithm in R for the double sigmoidal curve with an added Gibbs step for measurements with uncertainty. This has been adapted from previous work by Elena Moltchanova who kindly provided her algorithm for me to use.

```
# Date: 1 March 2020
# Author: Elena Moltchanova and Marina Chen

# MCMC sampler for the double logistic growth curve
# with reparametrised priors

library(ggplot2)
library(dplyr)
```

```
# Auxiliary functions ------------------------------------------------

# Expit
expit <- function(x){1/(1+exp(-x))}

# Log-likelihood
loglik <- function(x,y,a0,da,b0,db,g0,g1,tau){
  mu <- a0*expit(g0*(x-b0))+da*expit(g1*(x-b0-db))
  LL <- sum(dnorm(y,mu,1/sqrt(tau),log=T))
  return(LL)
}

# Sum of squared errors
ss <- function(x,y,a0,da,b0,db,g0,g1){
  mu <- a0*expit(g0*(x-b0))+da*expit(g1*(x-b0-db))
  SS <- sum((y-mu)^2)
  return(SS)
}

# MH sampler ---------------------------------------------------------

MH <- function(x, m, u_tau, ID,n,ITER,
               mu.a0, sd.a0, mu.da, sd.da ,
               mu.b0, sd.b0, mu.db, sd.db,
               mu.g0, mu.g1, a.tau, b.tau,
               sd.g0, sd.g1
               ){

  # NB. ID must be numeric
  # Here, N is the number of vineyards
  N <- max(ID)

  mon.a0 <- mon.da <- mon.b0 <- mon.db <-
    mon.g0 <- mon.g1 <- mon.tt <- numeric(ITER)

  # Initialising
  a0 <- rep(mu.a0,N)
  da <- rep(mu.da,N)
  b0 <- rep(mu.b0,N)
  db <- rep(mu.db,N)
```

```
g0 <- rep(mu.g0,N)
g1 <- rep(mu.g1,N)
tau <- rep(a.tau/b.tau,N)
y <- rep(mean(m), n)

# Here, x is the date, y is the bunch weight and n is the number of bunches

# Metropolis-Hastings sampler
for(iter in 1:ITER){print(iter)

  # Gibbs step for sampling y
  mu <- a0*expit(g0*(x-b0))+da*expit(g1*(x-b0-db))
  y <-rnorm(n, mean = ((u_tau) * m + (tau * mu))/(u_tau + tau),
            sd = 1/sqrt(u_tau + tau))

  # Sampling a0n (and a1n as a result)
  a0n <- rnorm(N,a0,.005);

  logR <- loglik(x,y,a0n,da,b0,db,g0,g1,tau)-
    loglik(x,y,a0 ,da,b0,db,g0,g1,tau)+
    sum(dnorm(a0n,mu.a0,sd.a0,log=T))-
    sum(dnorm(a0 ,mu.a0,sd.a0,log=T))

  logU <- log(runif(1,0,1))

  if(logR>logU){a0 <- a0n}

  # Sampling delta.a (and a1n as the result)
  dan <- exp(rnorm(N,log(da),.01));

  logR <- loglik(x,y,a0,dan,b0,db,g0,g1,tau)-
    loglik(x,y,a0,da,b0,db,g0,g1,tau)+
    sum(dnorm(dan,mu.da,sd.da,log=T))-
    sum(dnorm(da ,mu.da,sd.da,log=T))-
    sum((dnorm(log(dan),log(da ),.05,log=T)-log(dan)))+
    sum((dnorm(log(da ),log(dan),.05,log=T)-log(da )))

  logU <- log(runif(1,0,1))

  if(logR>logU){da <- dan}
```

```
# Sampling b0n (and b1n as the result)
b0n <- rnorm(N,b0,.05)

logR <- loglik(x,y,a0,da,b0n,db,g0,g1,tau)-
  loglik(x,y,a0,da,b0 ,db ,g0,g1,tau)+
  sum(dnorm(b0n,mu.b0,sd.b0,log=T))-
  sum(dnorm(b0 ,mu.b0,sd.b0,log=T))

logU <- log(runif(1,0,1))

if(logR>logU){b0 <- b0n}

# Sampling db (and b1n as the result)
dbn <- exp(rnorm(N,log(db),.005))

logR <- loglik(x,y,a0,da,b0,dbn,g0,g1,tau)-
  loglik(x,y,a0,da,b0,db ,g0,g1,tau)+
  sum(dnorm(dbn,mu.db,sd.db,log=T))-
  sum(dnorm(db ,mu.db,sd.db,log=T))-
  sum((dnorm(log(dbn),log(db ),.025,log=T)-log(dbn)))+
  sum((dnorm(log(db ),log(dbn),.025,log=T)-log(db )))

logU <- log(runif(1,0,1))

if(logR>logU){db <- dbn}

# Sampling g0 (NB. proposal is not symmetric)

g0n <- exp(rnorm(N,log(g0),.01))

logR <- loglik(x,y,a0,da,b0,db,g0n,g1,tau)-
  loglik(x,y,a0,da,b0,db,g0 ,g1,tau)+
  sum(dnorm(g0n,mu.g0,sd.g0,log=T))-
  sum(dnorm(g0 ,mu.g0,sd.g0,log=T))-
  sum((dnorm(log(g0n),log(g0 ),.05,log=T)-log(g0n)))+
  sum((dnorm(log(g0 ),log(g0n),.05,log=T)-log(g0 )))

logU <- log(runif(1,0,1))
```

```
    if(logR>logU){g0 <- g0n}


    # Sampling g1 (NB. proposal is not symmetric)

    g1n <- exp(rnorm(N,log(g1),.05))

    logR <- loglik(x,y,a0,da,b0,db,g0,g1n,tau)-
      loglik(x,y,a0,da,b0,db,g0,g1 ,tau)+
      sum(dnorm(g1n,mu.g1,sd.g1,log=T))-
      sum(dnorm(g1 ,mu.g1,sd.g1,log=T))-
      sum((dnorm(log(g1n),log(g1 ),.2,log=T)-log(g1n)))+
      sum((dnorm(log(g1 ),log(g1n),.2,log=T)-log(g1 )))

    logU <- log(runif(1,0,1))

    if(logR>logU){g1 <- g1n}

    # Sampling tau (Gibbs step)
    tau <- rgamma(N,a.tau+n/2,b.tau+.5*ss(x,y,a0,da,b0,db,g0,g1))


    mon.a0[iter] <- a0; mon.b0[iter] <- b0;
    mon.g0[iter] <- g0; mon.g1[iter] <- g1;
    mon.da[iter] <- da; mon.db[iter] <- db;
    mon.tt[iter] <- tau
  } # end of iterations

  return(data.frame(mon.a0=mon.a0, mon.da=mon.da,
                    mon.b0=mon.b0, mon.db=mon.db,
                    mon.g0=mon.g0, mon.g1=mon.g1, mon.tt=mon.tt))

} # end of MH function

# Additional arguments for the MH sampler --------------------

ITER <- 1200000

ID <- 1
```

```
# Prior parameters
mu.a0 <- 4.09
sd.a0 <- sqrt(1/0.11)
mu.da <- 0.69
sd.da <- 1
mu.b0 <- 40
sd.b0 <- 1/sqrt(0.02)
mu.db <- 30
sd.db <- 1/sqrt(0.11)
mu.g0 <- 0.3
mu.g1 <- 0.3
a.tau <- 4
b.tau <- 1

sd.g0 <- 1/sqrt(44.44)
sd.g1 <- 1/sqrt(44.44)

# Run algorithm --------------------------------------
posterior_result <- MH(x, y, u_tau, ID, n, ITER,
                       mu.a0, sd.a0, mu.da, sd.da ,
                       mu.b0, sd.b0, mu.db, sd.db,
                       mu.g0, mu.g1, a.tau, b.tau,
                       sd.g0, sd.g1)

# Discarding burn-in period of 20,000 iterations
posterior_sample <- posterior_result[-(1:20000),]

# Thinning the sample
posterior_sample <- posterior_sample[seq(1, nrow(posterior_sample),
    length.out = 5000),]

# Checking posterior means
sapply(posterior_sample, mean)

# Obtain posterior predictive means and 95% credible intervals ----------

x_values <- seq(0, 120, by = 1)

post_pred_y_result <- lapply(x_values, function(x){
```

```
  post_pred_sample <- rnorm(10000,
                            mean = posterior_sample$mon.a0*expit(
                              posterior_sample$mon.g0*(
                              x-posterior_sample$mon.b0))+
                              posterior_sample$mon.da*expit(
                                posterior_sample$mon.g1*(
                                  x-posterior_sample$mon.b0-
                                    posterior_sample$mon.db)),
                            sd = 1/sqrt(posterior_sample$mon.tt))

  return(c(mean(post_pred_sample), quantile(post_pred_sample,
                                            probs = c(0.025, 0.975))))

})

post_pred_y_data_inc_unc <- as.data.frame(do.call(rbind, post_pred_y_result))
colnames(post_pred_y_data_inc_unc) <- c("mean", "lower_limit", "upper_limit")
post_pred_y_data_inc_unc <- cbind(data.frame(x = x_values),
                          post_pred_y_data_inc_unc)
```