Determining the Size of a Galaxy's Globular Cluster Population through Imputation of Incomplete Data with Measurement Uncertainty

DETERMINING THE SIZE OF A GALAXY'S GLOBULAR CLUSTER POPULATION THROUGH IMPUTATION OF INCOMPLETE DATA WITH MEASUREMENT UNCERTAINTY

BY

MICHAEL R. RICHARD, Hon. B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2015) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Determining the Size of a Galaxy's Globular Cluster Population through Imputation of Incomplete Data with Measurement Uncertainty
AUTHOR:	Michael R. Richard Hon. B.Sc., (Applied Statistics) University of Toronto, Mississauga, Canada
SUPERVISOR:	Dr. Ben Bolker
NUMBER OF PAGES:	xi, 82

To my parents; Bob and Michele, for their continuous support and encouragement.

Abstract

A globular cluster is a collection of stars that orbits the center of its galaxy as a single satellite. Understanding what influences the formations of these clusters provides understanding of galaxy structure and insight into their early development. We continue the work of Harris *et al.* (2013), who identified a set of predictors that accurately determined the number of clusters N_{GC}, through analysis of an incomplete dataset.

We aimed to improve upon these results through imputation of the missing data. A small amount of precision was gained for the slope of $N_{GC} \sim R_e\Sigma_e$, while the intercept suffered a small loss of precision. Estimates of intrinsic variance also increased with the addition of imputed data.

We also found galaxy morphological type to be a significant predictor of N_{GC} in a model with $R_e\Sigma_e$. Although it increased precision of the slope and reduced the residual variance, its overall contribution was negligible.

KEY WORDS: Missing data, Imputation, Bayesian regression, Predictive mean matching, Measurement uncertainty, Regression with uncertainty, Globular clusters.

Acknowledgements

I wish to express my sincerest thanks to my supervisor Dr. Ben Bolker for his invaluable guidance and continuous support through the entirety of this project.

My thanks also go to Dr. William Harris of the Department of Physics and Astronomy for supplying the data used in this project, for the knowledge and expertise he so willingly provided, and for serving on my examining committee.

I would also like to thank Dr. Roman Viveros-Aguilera for guiding me to a project I so thoroughly enjoyed and for serving on my committee.

Finally I wish to thank my colleagues, my Professors, and the Department of Mathematics and Statistics as a whole for providing such a positive and encouraging environment in which to perform my studies. It has been a wonderful experience that I owe to all those involved.

Notation and abbreviations

This thesis follows the notation defined below.

- Matrices and vectors are denoted in **bold**, with matrices being uppercase and vectors being lowercase so that we can define a data matrix $\mathbf{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_p)$. Single elements of these arrays such as Z_{ij} and $\bar{z}_1$ are not bolded. This data matrix is subsetted in two ways:
  - In discussing theory of missing data, we subset $\mathbf{Z} = (\boldsymbol{Z}^{\text{obs}}, \boldsymbol{Z}^{\text{mis}})$. Here $\boldsymbol{Z}^{\text{obs}}$ and $\boldsymbol{Z}^{\text{mis}}$ represent matrical subsets of the data where the quantities of interest to the current procedure are either observed or missing. This same subsetting applied to a single column of the dataset $\boldsymbol{z_i} = (\boldsymbol{z_i}^{\text{obs}}, \boldsymbol{z_i}^{\text{mis}})$ represents the elements of the column which are observed and missing respectively.
  - When regressing one column of $\boldsymbol{Z}$ on one or more of the remaining columns, we subset $\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X})$ where $\boldsymbol{X}$ contains at least one possible covariate of $\boldsymbol{y}$
- When obtaining values of a statistic, an asterisk denotes a draw from the posterior distribution of interest such that $\hat{\beta}^*$ represents a draw from the distribution of $\hat{\beta}$ rather than its point estimate.

- The intrinsic variance associated with a column $\boldsymbol{y}$ is denoted σ_y^2 while the variance associated with measurement uncertainty is denoted with δ_y^2. In keeping with previous work, uncertainty is frequently referred to by its standard error δ_y.

Contents

List of Figures

Chapter 1

Introduction and Problem Statement

In a perfect world all datasets would be complete and missing data would not be an issue. Unfortunately in the real world scientists drop beakers, plants die before the completion of experiments, and people drop out of trials. Thus arises the problem of missing data that plagues standard statistical methods. Standard methods are not designed to handle missing values, and data with missing information must be altered in order to be processed by these methods.

Missing data comes in many shapes and forms, and there is no single algorithm that can be applied to all situations. In order to handle missing data one requires an understanding of subject matter, as well as information regarding the nature of the missing data. There are countless methods available for dealing with missing data, and in this thesis we discuss some (but not all) of them.

The process employed in this thesis is called *imputation* and involves filling in missing values based on the information present in the dataset. This is applied to

an incomplete dataset provided by Dr. William Harris of the Department of Physics and Astronomy at McMaster University. The quantity of interest is the number of galaxy globular clusters, collections of stars that orbit the center of a galaxy as single units. Harris *et al.* (2013) performed an available case analysis on which we hope to improve.

We perform an analysis similar to that of Harris; however we aim to improve precision through imputation of the missing values. Our goal is to impute missing data in such a way that the power gained from additional observations exceeds the additional variance that accompanies the process.

Chapter 2

Imputation

2.1 The Problem of Missing Data

When dealing with incomplete data there are many remedies that can be applied in preparation for the application of statistical methods. In this chapter we describe the approach to processing incomplete data and discuss some of the methods that can be used.

When presented with incomplete data the obvious solution is to remove cases that contain missing values. This is known as *complete case analysis* and is generally the default method for statistical packages when they receive input with missing values. Although simple, this method throws away useful information, leading to a decrease in statistical power, and possibly biasing results (Little and Rubin, 2002)

Missing values can present themselves in two different ways. The first is as *unit* nonresponse, where no observations are made for a single unit, or *row* of the dataset. In this situation there is no information available about the observation. Remedies for this problem include weighting procedures, and analysis of complete cases. The

second way in which missing data can appear is as *item* nonresponse. In this situation at least some information is collected for all observations, but observations may have missing values for one or more columns. This project will focus solely on the case of item nonresponse. For more information on unit nonresponse see Little and Rubin (2002, chap.3).

With item nonresponse we have some information about each observation, and this affords us more freedom when deciding how to modify the data for analysis. Complete case analysis should always be a last resort for the reasons mentioned above. A slight improvement would be *available case analysis*, where only observations missing the variable(s) of interest are discarded. This method is burdened by the same issues as complete case analysis, though they are often less severe.

Since some information is available for every case, an intuitive solution is to predict missing values from those that are observed. This process is called *imputation* and makes use of information within covariates to fill in missing values and avoid a loss of information. This creates a complete rectangular dataset that can then be analyzed by standard methods.

Imputation can be a very powerful tool that outperforms analysis of observed cases (Heitjan and Little, 1991), but one must be sure to follow the proper procedures as using it blindly may yield wildly invalid inference. The first assumption we must make is that the missing values are placeholders for actual values. For example it would make little sense to impute the value for the quality of life of a deceased patient, or to fill in an opinion for someone who refused to answer due to their legitimate lack of an opinion. If the missing value truly does not exist it makes little sense to impute it. As long as this assumption is reasonable we can move on to inspecting the nature

of the data.

There are two main processes affecting how we handle incomplete data; the missing data *pattern* and the missing data *mechanism*. The missing data pattern describes which values are observed and which are missing. This pattern can help in selection of imputation methods, as some methods work better with specific patterns. Within the scope of this project we will assume the data presents a general case where no discernible pattern is present. Most imputation methods work in the case of a general pattern of missingness, and for this reason different patterns will not be discussed further. More information on imputation methods for specific missing data patterns can be found in Little and Rubin (2002, chap. 7).

The missing data mechanism describes the relationship between missingness and the values of both the observed and unobserved quantities. For an incomplete dataset $\mathbf{Z}$ let us define a missing data matrix $\mathbf{M}$ where

$$M_{ij} = \begin{cases} 0, & \text{if } Z_{ij} \text{ is missing,} \\ 1, & \text{if } Z_{ij} \text{ is observed.} \end{cases}$$

The missing data mechanism is given by the distribution of $\mathbf{M}$ conditional on $\mathbf{Z}$ and some unknown parameters ϕ. There are three different types of mechanisms that missing data can have; missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).

Data is considered to be missing completely at random if the missingness depends only on ϕ, and not on the data at all, ie

$$f(\mathbf{M}|\mathbf{Z}, \phi) = f(\mathbf{M}|\phi) \text{ for all } \mathbf{Z}, \phi.$$

An example of MCAR data would be a tendency for people to accidentally miss a question on a survey such that the question is missed as a mistake, and not for any reasons related to the question.

Data that is MCAR is the ideal situation, as the observed data $\mathbf{Z}^{\text{obs}}$ are then a random sample of the complete data and thus the observed values contain no systematic bias (Huisman, 2009). Therefore a correctly specified model fit to observed data will also correctly fit to the missing data. Although it is theoretically ideal, MCAR data rarely occurs outside of textbooks.

A much more common mechanism is data that is missing at random (MAR). Data that is MAR has missingness that depends on observed values only, and not on the missing values:

$$f(\mathbf{M}|\mathbf{Z}, \phi) = f(\mathbf{M}|\mathbf{Z}^{\text{obs}}, \phi) \text{ for all } \mathbf{Z}, \phi.$$

This can occur in a survey including questions relating to both income and education. Suppose people with higher education are less likely to disclose their income. The probability of a response depends on the observed education, rather than the unobserved income. Because of the relationship between $\mathbf{M}$ and the observed values there is bias in which values of the data are observed; however, this can be controlled for by conditioning on these observed variables.

The last mechanism is when data is not missing at random (NMAR). In this case the missingness depends on values that are not observed $\mathbf{Z}^{\text{mis}}$ (and possibly observed values as well). This will occur if people of higher income are less likely to disclose their income. In this case the chance of a missing income depends on the missing value itself, and not any observed values.

NMAR data is the most realistic situation and is the most difficult to properly

control for. The extent of the bias due to nonresponse cannot be known, and any models applied to the observed values may be invalid for the missing ones. This potentially causes the largest bias as there are systematic differences between the observed and missing values (Huisman, 2009).

An intrinsic part of the imputation process is justifying that the data of interest is in fact MAR, and that the bias created by nonresponse can be properly accounted for by conditioning on the observed data. It is important that imputation is performed while conditioning on observed values (especially those correlated with the variable being imputed), as this has the potential to reduce nonresponse bias as well as variances of estimates. (Little, 1988; Little and Rubin, 2002)

Little (1988, p. 288) provides a list of desirable properties that should be kept in mind when designing an imputation process:

1. Imputations should be based on the predictive distribution of the missing data conditional on the observed data.

2. All observed items for a case should be used in creating imputations.

3. Models used for obtaining imputed values should make use of subject matter knowledge.

4. Predictive models should avoid excessive extrapolation beyond the range of the data.

5. Imputations should be drawn from a distribution and not imputed by use of means (they should be stochastic). Mean imputation will distort marginal distributions as well as relationships between variables.

6. A method should be used to ensure errors are estimated with the understanding that they are drawn from a distribution, and are not observed.

2.2 Single Imputation Methods

This section describes some of the methods used for imputation. We assume we are working in the context of imputing a single incomplete variable $\mathbf{y}$ based on a matrix of complete covariates $\mathbf{X}$ such that these elements make up the dataset $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$. These methods are easily extended to the case when some (or all) of the covariates $\mathbf{X}$ also have missing values. This can be handled through use of the Gibbs sampler and is discussed in Section 2.4.2.

2.2.1 Means and Draws

There is no single nest algorithm for imputing data; and each imputation procedure must be tailored specifically to the dataset while taking as much care as possible to adhere to Little's desirable properties. The simplest and most trivial method is to simply impute the marginal mean $\bar{y}$. Although this is a simple way to fill in missing values the unconditional mean is not a reliable method for imputation. Although the marginal mean is preserved, it creates a spike at the center of the distribution, biasing the variance downwards by inflating the sample size while keeping the sum of squared deviations around the mean constant. If we consider a dataset with n observations, where our incomplete variable $\boldsymbol{y}$ has $n - k$ observed values, then our complete case

estimate for variance is

$$s^2_{\text{obs}} = \frac{1}{n-k-1}\sum_{i=1}^{n-k}(y_i - \bar{y})^2, \tag{2.1}$$

which is a consistent estimate for the true variance under MCAR assumptions (Little and Rubin, 2002). If we impute the mean of the observed values $\bar{y}$ for each missing observation the variance becomes

$$s^2_{\text{mean}} = \frac{1}{n-1}\sum_{i=1}^{n-k}(y_i - \bar{y})^2 + \frac{1}{n-1}\sum_{i=n-k+1}^{n}(\bar{y} - \bar{y})^2. \tag{2.2}$$

Noting that the second term in equation 2.2 goes to zero, this becomes the same as the observed variance in equation 2.1 but with a larger denominator. Therefore the proportion of variance explained by the estimate on the imputed data is $s^2_{\text{mean}}/s^2_{\text{obs}} = (n-k-1)/(n-1)$. Similar calculations will show that covariance is affected in the same way. It may seem logical to simply correct this by multiplying the variance of the imputed data by a factor of $(n-1)/(n-k-1)$, but this yields exactly the complete case estimates, negating any possible information or power gained from the imputation process.

Mean imputation fails Rubin's first and fifth rules as it is neither conditional nor stochastic. We can make a small adjustment to fix the method's lack of stochasticity by adding a random normal deviate to the mean for each imputation so that each imputed value becomes $\bar{y}+\epsilon$ where $\epsilon \sim \text{N}(0, \sigma^2_y)$. This is referred to as a *random draw*. Although a random draw will preserve the marginal distribution of $\mathbf{y}$, it still fails to condition on the observed data and may skew relationships with other variables.

If we adjust the random draw to be conditional on the observed data it will provide

the best methods for imputation. Table 2.1 from Little and Rubin (2002) summarizes the large-sample bias of different parameters for the univariate imputation model for MCAR data for four different classes of imputation. The response $\boldsymbol{y}$ is observed in $\Lambda = \frac{n-k}{n}$ cases, and $\boldsymbol{x}$ is a single complete covariate. Bias is calculated for the mean $\bar{y}$ and variance σ_y^2, as well as the coefficients for regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ (β_{yx}) and regression of $\boldsymbol{x}$ on $\boldsymbol{y}$ (β_{xy}). The conditional means are calculated with standard regression and imputing the predicted values directly, while the conditional draw adds a normal variate ϵ with a mean 0 and a variance equal to the residual variance from the regression.

	Parameter			
Method	μ_y	σ_y^2	β_{yx}	β_{xy}
Unconditional mean	0	$-\Lambda\sigma_y^2$	$-\Lambda\beta_{yx}$	0
Unconditional draw	0	0	$-\Lambda\beta_{yx}$	$-\Lambda\beta_{xy}$
Conditional mean	0	$-\Lambda(1-\rho^2)\sigma_y^2$	0	$\frac{\Lambda(1-\rho^2)}{1-\Lambda(1-\rho^2)}\beta_{xy}$
Conditional draw	0	0	0	0

Table 2.1: Bias of mean, variance, and regression coefficients for univariate imputation methods. Little and Rubin (2002, p. 66)

We can see that all four methods yield unbiased estimates for the mean. Both applications of mean imputation give biased results for the variance, while their draw-based counterparts give consistent estimates, conforming to Rubin's fifth rule of using draws in place of means. Only the conditional draw gives consistent estimates for both regression parameters, as well as the mean and variance of the distribution, further confirming that using a draw rather than a mean while conditioning on the observed data will give the best results.

2.2.2 Bayesian Regression

As implied by the methods used in the table, the obvious way to make use of a conditional draw is by regressing the incomplete variable on its covariates to obtain regression parameters, then using those parameters to predict the missing values of the response variable. In order to ensure the regression is stochastic, we impute a draw from the distribution of the predicted value, rather than the predicted value itself. To keep with the stochastic design, it seems only natural to perform the regression in a Bayesian setting, as this justifies the draws from the posterior distribution in place of means. Furthermore, when a non-informative prior is used, the Bayesian point estimates match those from the non-Bayesian method.

We note the standard form of a posterior distribution

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y}) = p(\boldsymbol{\beta}|\sigma^2, \mathbf{X}, \mathbf{y})p(\sigma^2|\mathbf{y}) \tag{2.3}$$

where factoring is allowed due to the independence of $\boldsymbol{\beta}$ and σ^2. We make use of a non-informative prior for σ^2 and a weak multivariate normal prior for $\boldsymbol{\beta}$ to get

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y}) \propto \sigma^{-2}\text{MVN}(0, \text{Diag}(\Sigma_\beta))$$

where Σ_β is the covariance matrix for $\boldsymbol{\beta}$, and the standard non-informative prior for σ^2 is proportional to σ^{-2}. The use of a normal prior for $\boldsymbol{\beta}$ leads to the method known as ridge regression (Hoerl and Kennard, 1970). Ridge regression applies a penalty to the size of the regression parameters in order to reduce the error of the estimates. It can also help deal with multicollinearity in the predictors however the motivation for its use here is in reducing error.

Ordinary least squares works by giving estimates for the parameters that are firstly unbiased, and secondly have minimum variance. Although this minimizes the variance for the unbiased estimate $\hat{\boldsymbol{\beta}}$, it does not minimize the mean square error defined as

$$\text{MSE}_\beta = \text{Bias}^2(\hat{\boldsymbol{\beta}}) + \text{Var}(\hat{\boldsymbol{\beta}}).$$

Ridge regression shrinks the MSE by sacrificing the unbiasedness of the estimator in order to decrease its variance. The idea is to decrease the variance more than the squared bias increases, reducing the overall error. This is done by adding a small value called a "ridge" to the diagonal of the square of the design matrix so that the point estimate and variance for β becomes

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_n)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.4}$$

and

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma_y^2\mathbf{W}\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{W}, \tag{2.5}$$

where $\mathbf{W} = (\mathbf{X}^\mathbf{T}\mathbf{X} + \lambda\mathbf{I_n})^{-1}$. This has the effect of penalizing the size of the coefficients. While the traditional OLS methods minimize

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2,$$

ridge regression estimates will minimize

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2 + \lambda\sum_{j=1}^{p}\hat{\beta}_j^2.$$

The $\hat{\boldsymbol{\beta}}$ obtained from ridge regression can be expressed in terms of the OLS result where

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = [\mathbf{I}_n + \lambda(\mathbf{X}^T\mathbf{X})^{-1}]\hat{\boldsymbol{\beta}}^{\text{OLS}}. \tag{2.6}$$

The bias and variance of the $\hat{\boldsymbol{\beta}}$ are both functions of the ridge λ. $\text{Var}(\hat{\boldsymbol{\beta}})$ is a decreasing function of λ while $\text{Bias}^2(\hat{\boldsymbol{\beta}}) = (-\lambda\mathbf{W}\boldsymbol{\beta})^2$ is an increasing function of λ. Therefore as $\lambda \to 0$ the estimate of $\hat{\boldsymbol{\beta}}$ reduces to the unbiased estimate provided by ordinary least squares, while simultaneously increasing variance. Note that substituting $\lambda = 0$ into equations 2.4 and 2.5 leads to OLS estimates. When $\lambda \to \infty$ the bias increases as the parameter estimates shrink to zero, while simultaneously decreasing variance (to a point); the goal is to find a value of λ that minimizes the MSE.

Even if all of our typical regression assumptions are perfectly met, there will always exist a λ such that the MSE for the ridge estimate will be smaller than that for ordinary least squares (Hoerl and Kennard, 1970). For this reason ridge regression will always be an improvement (in terms of MSE) upon ordinary least squares methods, and will always increase the precision of estimates.

With our method of obtaining estimates for the regression parameters $\boldsymbol{\beta}$ clearly defined we can now continue through the typical Bayesian regression as described by Gelman *et al.* (2003). We begin by obtaining the estimates $\hat{\boldsymbol{\beta}}$ and $\text{Var}(\hat{\boldsymbol{\beta}})$ as defined in equations 2.4 and 2.5, and we wish to draw $\hat{\boldsymbol{\beta}}^*$ to use for imputation. Recall that * denotes a draw from the distribution of a statistic. By rearranging equation (2.3) we can see that the posterior distribution for σ^2 is

$$p(\sigma^2|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})}{p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y})}$$

which is a scaled inverse Chi-square distribution. Letting $\zeta*$ denote a draw from χ^2_{n-k} we obtain the estimated variance

$$\sigma^{*2} = \sqrt{\frac{(\boldsymbol{y} - \mathbf{X}^{\mathbf{T}}\boldsymbol{\beta})^2}{\zeta*}}$$

and use this to draw our value for $\boldsymbol{\beta}^*$ from

$$\boldsymbol{\beta}|\sigma_y^2, \boldsymbol{y} \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, \mathrm{Var}(\hat{\boldsymbol{\beta}})).$$

With our estimates of β and its variance obtained, we now may draw values from the predictive distribution of the missing data $\boldsymbol{Z}^{\mathrm{mis}}$ conditional on the observed data $\boldsymbol{Z}^{\mathrm{obs}}$ and our parameter estimates. The expected value for the missing observations Z^{mis} is

$$\begin{aligned}
\mathbb{E}(\boldsymbol{y}^{\mathrm{mis}}|\boldsymbol{\beta}, \sigma_y^2, \boldsymbol{y}^{\mathrm{obs}}) &= \mathbb{E}[\mathbb{E}(\boldsymbol{y}^{\mathrm{mis}}|\boldsymbol{\beta}, \sigma_y^2, \boldsymbol{y}^{\mathrm{obs}})|\sigma_y^2, \boldsymbol{y}^{\mathrm{obs}}] && (2.7)\\
&= \mathbb{E}(\mathbf{X}^{\mathrm{mis}}\boldsymbol{\beta}|\sigma_y^2, \boldsymbol{y}^{\mathrm{obs}}) && (2.8)\\
&= \boldsymbol{X}^{\mathrm{mis}}\boldsymbol{\beta}, && (2.9)
\end{aligned}$$

The variance of those predictions is

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{y}^{\mathrm{mis}}|\boldsymbol{\beta}, \sigma_y^2, \boldsymbol{y}^{\mathrm{obs}}) &= \mathbb{E}[\mathrm{Var}(\boldsymbol{y}^{\mathrm{mis}}|\boldsymbol{\beta}, \sigma_y^2, \boldsymbol{Y}^{\mathrm{obs}})|\sigma_y^2, \boldsymbol{Y}^{\mathrm{obs}}] && (2.10)\\
&+ \mathrm{Var}[\mathbb{E}(\boldsymbol{y}^{\mathrm{mis}}|\boldsymbol{\beta}, \sigma_y^2, \boldsymbol{y}^{\mathrm{obs}})|\sigma_y^2, \boldsymbol{y}^{\mathrm{obs}}] && (2.11)\\
&= \mathbb{E}[\sigma_y^2\mathbf{I}|\sigma_y^2, \boldsymbol{y}] + \mathrm{Var}[\boldsymbol{X}^{\mathrm{mis}}\boldsymbol{\beta}|\sigma_y^2, \boldsymbol{y}^{\mathrm{obs}}] && (2.12)\\
&= (\mathbf{I} + \boldsymbol{X}^{\mathrm{mis}}\mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{X}^T)\sigma_y^2. && (2.13)
\end{aligned}$$

Notice the variance is a combination of the intrinsic variance $\sigma_y^2\mathbf{I}$ and the variance due to uncertainty in $\hat{\beta}$.

2.2.3 Implicit methods

There are many types of implicit imputation, and just like imputation in general, each type has its own collection of variations. Implicit methods act on the general principle of replacing missing values with ones observed in some other location. Two common methods of implicit imputation are the *hot deck*, where values are imputed from other observations in the same dataset (Andridge and Little, 2010; Little, 1988), and *predictive mean matching* (Schenker, 1996; Morris *et al.*, 2014; Little, 1988; Heitjan and Little, 1991), a semi-parametric middle ground between the hot deck and traditional explicit methods such as regression.

Implicit methods work by matching each missing value with a close observed value, where closeness is measured by some distance metric within the covariates of the variable being imputed. This "close" value is referred to as a *donor* value, and is often directly filled in for the missing value known as the *recipient*.

The term "hot deck" originates from the computer punch cards that would be substituted for missing values. Since these values come from the deck of cards currently in use, it is referred to as the *hot* deck. A variation of this method called the *cold* deck borrows values from previously processed data such as those from a previous study.

Originally developed for the US Census Bureau for item nonresponse in the Income Supplement of the Current Population Survey (CPS) in 1987, the procedure begins with the assignment of adjustment cells to the data. Adjustment cells are unique

"blocks" of covariate data, where every possible combination of covariates falls into one of the cells. For example if we were imputing the incomplete variable $\boldsymbol{y}$ based on the covariates $\boldsymbol{x}_1$ (income) and $\boldsymbol{x}_2$ (age), we may wish to assign (overly simplified) adjustment cells as they appear in Table 2.2. It is important that each observation is a member of only a single adjustment cell.

	Income	Age
Cell A	$< \$50,000$	Under 35
Cell B	$< \$50,000$	Over 35
Cell C	$> \$50,000$	Under 35
Cell D	$> \$50,000$	Over 35

Table 2.2: Sample adjustment cells

With the adjustment cells in place, the process begins by choosing a starting value for each cell. This value will be referred to as the "on hand" value for a cell. The choice for the starting value is often drawn as a random selection from within the cell. Now the process iterates by moving through the rows of the dataset, while observing whether or not the incomplete variable is observed or missing. If the variable is missing, then the current on hand value for the cell corresponding to that observation is substituted. If the variable is instead observed, then this observed value becomes the new on hand value for that cell. The process continues through the rows of the dataset until all missing values are imputed.

The hot deck has the convenient property that since imputed values are taken from observed ones, only existing values can be imputed. This is very convenient when the data being imputed is binary, or restricted to certain values (such as positive integers

for number of children).

With the proper specification of adjustment cells, all imputed values will be taken from nearby observations, meaning that the hot deck preserves both both linear and nonlinear relationships. It also avoids strong parametric assumptions made by explicit methods. Unfortunately the nature of the adjustment cells can restrict the performance of the hot deck. It may be the case that there very few, or even no donors within an adjustment cell, which can lead to overuse of donors and create bias. This may be especially problematic when sample sizes are small, or when the number of covariates (and thus the number of adjustment cells) becomes large. A common remedy to this is to collapse cells until a donor is found, but this may mean values that are far from the one being imputed are being used, and they may be a very poor fit.

Even when donors are plentiful, overuse of donors can be an issue if sequences of missing values appear in the data. Because the on hand value changes only when the iteration passes an observed value, any sequences of missing values will all be imputed by the same donor. This can be avoided by abandoning the iterative scheme and randomly sampling with replacement from within cells for each missing value, though this can still suffer when there is not a large number of donors available.

As the number of auxiliary variables increases, naturally the number of adjustment cells increases, and can become unmanageable with fewer and fewer donors in each cell. In addition, if some of these variables are continuous, it makes little sense to force them into categorical form in order to define the cells. Some of the problems brought on by use of adjustment cells can be relieved with the more general implicit method of distance matching. Rather than forcing each observation into a discrete

category and making matches within that category, one can match missing values directly with other values based on the distance between their covariates. The hot deck is in fact a special case of the general distance matching method. If we let $C(y_i)$ be the adjustment cell for observation i then the distance metric used by the hot deck for the distance between observations i and j is defined as

$$d(i,j) = \begin{cases} 0, & j \in C(y_i) \\ 1, & j \notin C(y_i). \end{cases}$$

By defining metrics that do not require this categorization of variables, we are making use of what is often referred to as nearest neighbor imputation. A common implementation of this method is the maximum deviation

$$d(i,j) = \max_k |x_{ik} - x_{jk}|,$$

which finds the maximum distance between observations across all covariates, and uses the one with the smallest maximum distance. Some also make use of the Mahalanobis distance

$$d(i,j) = (x_i - x_j)^T \hat{\mathrm{V}}(x_i)^{-1}(x_i - x_j),$$

where $\hat{\mathrm{V}}(x_i)$ is the estimated covariance matrix of x_i. A third method is to use the predicted means

$$d(i,j) = (\hat{y}_i - \hat{y}_j)^2,$$

where $\hat{y}_i$ is the predicted value of y_i. The use of the predictive mean as a matching metric was popularized by Little (1988) who named it predictive mean matching

(PMM). It can be referred to as semi-parametric as it uses predicted values only to define the matching distance, and not actually to create imputed values. It allows for the relaxation of many of the assumptions made by parametric methods, and therefore has the potential of being more robust towards model misspecification.

PMM works by using regression to obtain predicted means for all missing and observed values, then imputing the missing ones with the observed value having the closest predicted mean. In his original work, Little used the point estimate $\hat{\boldsymbol{\beta}}$ to obtain the means for both the observations and the missing values, but in the same paper notes that this does not allow for uncertainty in $\hat{\boldsymbol{\beta}}$ (Little, 1988). A simulation study by Morris *et al.* (2014) suggested that using the predicted value $\hat{\boldsymbol{\beta}}$ for observed values, while using a draw $\hat{\boldsymbol{\beta}}^* \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, V_\beta \sigma^2)$ to obtain predictions for the missing values, fixed this issue.

PMM shares some of the nice properties of hot deck because imputed values are taken from observed ones nearby within the dataset, ensuring only allowable values are imputed, and allowing for the preservation of nonlinear relationships.

Little (1988) defined PMM to directly impute the value with the closest predicted mean, but later publications caution against imputing a single donor directly (Morris *et al.*, 2014; Schenker, 1996; Heitjan and Little, 1991). In order to maintain variability between imputed values, it is better to draw the donor value from a number of close observations collectively referred to as a *donor pool*. This ensures that missing values that are close together do not always receive the same donor value.

Several ways exist of selecting how to populate the donor pool. One method is to include all donors within a fixed distance $\delta = |\hat{y}_i - \hat{y}_j|$. This allows for drawing from any observed values with predicted means close to that of the missing value. The use

of a fixed distance to define the size of the donor pool can easily run into problems relating to the density of observations. A missing value may exist in a location where very few donors lie (or none). This situation can lead to the same issues experienced when imputing a single value directly.

Another option is to use the k nearest donors, where k is defined beforehand. There is no consensus in the literature about the size of the donor pool; it depends on the density of the data. Suggested sizes range from a single donor (Little, 1988) to ten or higher (Morris *et al.*, 2014), though values around 3-5 seem most common (Schenker, 1996; Heitjan and Little, 1991). Even when using a set number of donors PMM is not immune to issues concerning donors. Because the use of k has no restriction on distance it is possible that when the area around a missing value has a low density of observations, that values may be imputed that are far from the predicted mean of the missing value. This can lead to bad imputations and bias.

Schenker (1996) defines a new method for choosing value for the donor pool that uses values within a certain distance, but if the number of values does not reach a certain threshold then the next closest observations are added to the pool to ensure variation. This method only slightly outperformed the use of a fixed size k in simulations. It shows promise, but requires more extensive work in more general situations.

2.3 Multiple Imputation and Controlling for Imputation Uncertainty

2.3.1 Uncertainty from Imputations

Once imputation has been performed and a complete dataset has been obtained, it may seem intuitive (and reasonably so) to then perform analysis on the dataset to obtain inference. This procedure, however, treats the imputed values as if they were observed, and it fails to take into account the uncertainty associated with drawing these values from their predictive distribution. Little and Rubin (2002) provide several possible methods to account for this extra variance.

The first method involves applying an explicit variance formula, and only works for simple methods where the variance of the process can be directly calculated. Another way is to modify the imputations so that valid standard errors can be obtained from a single dataset. Although simple, this method lacks generality and adjustments to the data may unfavorably affect the estimates. Proper estimates of errors can also be obtained through use of resampling methods. The imputation and analysis can be performed multiple times on randomly sampled subsets of the incomplete data. Although straightforward, resampling relies heavily on large sample sizes and can be computationally intensive.

A more desirable method may be to impute a vector of quantities for each missing value, rather than just a single value. This creates multiple independent datasets that contain all the same values for observations, but vary in their imputed values. One can then combine inference across the multiple datasets while noting the associated variation between them. This method is called *multiple imputation*.

All of the above methods rely at least somewhat on a model, and therefore make assumptions about the predictive distribution of the missing data. Resampling methods provide consistent estimates of variance for large samples, but their appropriateness may be questionable for smaller datasets (Little and Rubin, 2002). When performed in a Bayesian framework multiple imputation can perform better in smaller samples, though it is more dependent on model quality. Multiple imputation also helps to improve efficiency of point estimates over single imputation methods because estimates are averaged across multiple datasets, reducing the influence of a "bad" draw.

2.3.2 Multiple Imputation

Multiple Imputation imputes a vector of length $M > 1$ for each missing value. This involves performing single imputation processes on the dataset independently M times to obtain M complete datasets. At this point the imputed values can be treated as observed, and standard analysis can be performed on each dataset. As long as these datasets have been imputed under the same model for nonresponse, they can be combined to obtain parameter values with consistent estimates for variance (Little and Rubin, 2002).

Once estimates of the parameters of interest are obtained from each dataset we may combine them by the rules described by Little and Rubin (2002). Let the parameter estimate of the m^{th} dataset be $\hat{\theta}_m$ and its associated variance be V_m. We then average over all the estimates to obtain a point estimate

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m.$$

We then calculate the average within-imputation variance $\bar{V}_M$ and the between-imputation variance B_M to be

$$\begin{aligned} \bar{V}_M &= \frac{1}{M}\sum_{m=1}^{M} V_m, \text{ and} \\ B_m &= \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2. \end{aligned}$$

Combining the within- and between-imputation variance we obtain the total variance

$$T_M = \bar{V}_M + \frac{M+1}{M}B_M$$

where the factor $(M+1)/M$ is a finite population correction for M. This term is needed to correct the fact that the variance of our distribution is centred about our point estimate $\hat{\theta}$ and not around the true parameter value.

2.4 The MICE package in R

2.4.1 Configuring the Software for Imputation

The package `MICE` (Multivariate Imputation by Chained Equations) (van Buuren and Groothuis-Oudshoorn, 2011) was originally created for S-PLUS in 2000 and updated for use in R in 2001. `MICE` aims to create accurate and feasible multiple imputations for multivariate missing data in various forms, and offers customizability that allows all designs of missing data to be imputed with reasonable values. It performs multiple imputation through use of the chained equations, more commonly referred to as the Gibbs Sampler.

The imputation process within the package can be broken into three distinct steps, each of which should be handled with care to ensure reasonable results. One begins by specifying the imputation model so that the software can fill in the missing values. Once the values are imputed the middle step diagnoses the imputations to ensure they are plausible and meaningful. In the final step results are obtained for each of the M datasets and then pooled according to the rules defined in Section 2.3.

In the first step the user begins with specification of the imputation model. For each incomplete column of the dataset a single imputation method must be chosen. The built-in imputation functions offered by `MICE` include but are not limited to unconditional mean imputation, unconditional draw, Bayesian linear regression, non Bayesian linear regression, predictive mean matching and logistic regression. The user may also specify their own imputation methods.

The user must also choose which predictors are to be used in imputing each incomplete column. This is communicated through use of a predictor matrix $\boldsymbol{P}$. For a dataset with p columns, the predictor matrix is a 0/1 $(p \times p)$ matrix with each column of the dataset corresponding to a single row and a single column of the predictor matrix. If element P_{ij} contains a 1, this indicates that variable $\boldsymbol{z}_j$ will be used to impute variable $\boldsymbol{z}_i$. Suppose we have a dataset containing the variables height, age, and weight, where height and weight have missing values. Let age be complete. A sample predictor matrix may look like that of Table 2.3.

	Age	Height	Weight
Age	0	0	0
Height	1	0	1
Weight	1	0	0

Table 2.3: table

Sample predictor matrix.

Since age contains no missing values it does not need to be imputed and therefore consists of an entire row of zeros as no other variables are being used to predict it. Height is being predicted by age and weight, while weight is predicted only by age. Note that the main diagonal consists only of zeros since a variable cannot predict itself. The matrix need not be symmetric: In this case, for example, weight is used to predict height without having height as a predictor. This matrix can be created with raw R code, or it can be automatically created using the `quickpred()` function, which will create a predictor matrix based on user-specified criteria regarding minimum magnitudes of correlations and proportion of cases observed.

The user may also specify parameters of the imputation such as the number of iterations for the Gibbs sampler (default 5) and the number of multiple imputations to perform (also default 5). Both of these numbers must be large enough. Too few iterations of the sampler will prevent it from converging and give inaccurate results, whereas not enough multiple imputations will lead to undercoverage and overstatement of precision. It may also be useful to the user to specify the visit sequence i.e. the sequence in which the sampler visits the variables. For small to moderate size problems this may not be of concern, but for larger datasets with a greater number of imputations and/or iterations, specification of specific visit sequences can improve

the efficiency of the process and greatly reduce computational burden.

Once imputations are completed it is important to run diagnostics to ensure imputed values are reasonable. These methods include trace plots for the Gibbs sampler, distribution comparisons of imputed and observed values, as well as comparison of imputed values across completed datasets. These methods are not unique to `MICE` and will not be discussed further here.

The last step of the imputation process is pooling the results. As discussed in Section 2.3 multiple imputation yields M parameter estimates and associated variances. These are combined to produce parameter estimates and variances. Again this process will not be detailed here as it follows closely the theory of the previous sections. For a more detailed description of the `MICE` package see van Buuren and Groothuis-Oudshoorn (2011).

2.4.2 The Gibbs Sampler

The `MICE` package uses the Gibbs sampler to perform multiple imputation. The Gibbs sampler is a Markov chain Monte Carlo method of obtaining draws from the joint posterior distribution $P(\boldsymbol{z}_1, ..., \boldsymbol{z}_p)$ when draws from this distribution are very difficult to compute or are not readily available (such as in the case of missing data). As long as we have access to the conditional distributions $P(\boldsymbol{z}_j|\boldsymbol{z}_1, ..., \boldsymbol{z}_{j-1}, \boldsymbol{z}_{j+1}, ..., \boldsymbol{z}_p)$ then we can obtain draws from the joint distribution through the use of the sampler (Little and Rubin, 2002).

The process begins with filling in the missing values with some starting values; often (and in this case) the marginal means $\bar{z}_i^{\text{obs}}$ are used. Then each iteration of the sampler consists of a draw of the parameters θ and a draw of the values from

the posterior predictive distribution of the missing data. At iteration t we begin to impute the missing values of $\boldsymbol{z}_1$ by drawing

$$\theta_1^{*(t)} \sim f(\theta_1|\boldsymbol{z}_1^{\text{obs}}, \boldsymbol{z}_2^{(t-1)}, ..., \boldsymbol{z}_p^{(t-1)}) \tag{2.14}$$

$$\boldsymbol{z}_1^{*(t)} \sim f(\boldsymbol{z}_1|\boldsymbol{z}_2^{(t-1)}, ..., \boldsymbol{z}_p^{(t-1)}, \theta_1^{*(t)}). \tag{2.15}$$

Note that the draw of the parameters here is performed only on the rows of the dataset where $\boldsymbol{z}_1$ is observed, and then these parameters are used to impute the missing values of $\boldsymbol{z}_1$. Now that we have updated the values of $\boldsymbol{z}_1$ within this t^{th} iteration we may use them for filling in the remaining columns, while the variables which have not yet been visited will rely on the values from the $(t-1)^{\text{th}}$ iteration. The process continues

$$\theta_2^{*(t)} \sim f(\theta_2|\boldsymbol{z}_2^{\text{obs}}, \boldsymbol{z}_1^{(t)}, \boldsymbol{z}_3^{(t-1)}, ..., \boldsymbol{z}_p^{(t-1)}) \tag{2.16}$$

$$\boldsymbol{z}_2^{*(t)} \sim f(\boldsymbol{z}_2|\boldsymbol{z}_1^{(t-1)}, \boldsymbol{z}_3^{(t-1)} ..., \boldsymbol{z}_p^{(t-1)}, \theta_2^{*(t)}) \tag{2.17}$$

$$\vdots \tag{2.18}$$

$$\theta_p^{*(t)} \sim f(\theta_p|\boldsymbol{z}_p^{\text{obs}}, \boldsymbol{z}_1^{(t)}, ..., \boldsymbol{z}_{p-1}^{(t)}) \tag{2.19}$$

$$\boldsymbol{z}_p^{*(t)} \sim f(\boldsymbol{z}_p|\boldsymbol{z}_1^{(t)}, ..., \boldsymbol{z}_p^{(t)}, \theta_{p-1}^{*(t)}) \tag{2.20}$$

where at the end all variables have been visited and contain updated values from the current iteration. This concludes the iteration and the process may begin again on the next iteration with new updated values. It may seem that even with updated values the estimates may not change between iterations since θ is calculated using only observed values of the response, but this is not the case since the values of the covariates that θ is being conditioned on are also changing, leading to different

values of θ in each iteration. The exception to this is in the case of monotone missing data (where columns can be ordered so that the rows containing missing values for a column are a subset of the rows with missing values of the previous column), in which case convergence of the sampler is immediate.

Chapter 3

Regression with Uncertainty in X and Y

3.1 The propagation of measurement error

It is nearly universal in observational science for measurements to be accompanied by a quantification of measurement uncertainty. Suppose we have measurement variables x and y with associated standard errors δ_x and δ_y representing uncertainty. Now suppose the quantity of interest is a function of these two variables $g(x, y)$. How can we compute the uncertainty of g?

We make use of the delta method for obtaining this uncertainty as described in Lyons (1991). Let us first define a function as a product of exponents of x and y so that

$$f = Cx^a y^b.$$

We take the logarithm to simplify calculations, and differentiate both sides so that

$$\frac{\delta_f}{f} = a\frac{\delta_x}{x} + b\frac{\delta_y}{y}.$$

Then square both sides and remove the cross term. We make the assumption here that the uncertainties are sufficiently small that the product $\sigma_x\sigma_y$ will be negligible. This gives

$$\left(\frac{\delta_f}{f}\right)^2 = a^2\left(\frac{\delta_x}{x}\right)^2 + b^2\left(\frac{\delta_y}{y}\right)^2.$$

Then isolating the uncertainty of f gives

$$\delta_f = \sqrt{y^2\delta_x^2 + x^2\delta_y^2}.$$

Now we inspect the general case where $f = f(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$ where each $\boldsymbol{x}_i$ has an associated δ_i. The uncertainty in f is equal to the sum of the rate of change times the distance δ for each variable.

$$\delta_f = \frac{\partial f}{\partial x_1}\delta_{x_1} + \frac{\partial f}{\partial x_2}\delta_{x_2} + \cdots + \frac{\partial f}{\partial x_n}\delta_{x_n}$$

Squaring yields

$$\begin{aligned}
\delta_f^2 &= \left[\frac{\partial f}{\partial x_1}\cdots\frac{\partial f}{\partial x_n}\right]\begin{pmatrix} \delta_1^2 & \cdots & \delta_{1,n} \\ \vdots & \ddots & \vdots \\ \delta_{n,1} & \cdots & \delta_n^2 \end{pmatrix}\left[\frac{\partial f}{\partial x_1}\cdots\frac{\partial f}{\partial x_n}\right]^T \\
&= \nabla_f \Sigma_x \nabla_f^T
\end{aligned}$$

which, after cancelling error terms gives

$$\delta_f^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 \delta_i^2.$$

3.2 Type II Regression

With the presence of uncertainty in both the x and y dimensions it seems appropriate to consider type II regression. Typical regression (Type I) works by minimizing the vertical distance between each observed value y_i and its predicted mean $\hat{y}_i = \boldsymbol{x_i\beta}$. This method, however, ignores any variation in the x-coordinate. Type II regression focuses instead on minimizing the perpendicular distance between the points and the fitted line, allowing for variation in both dimensions.

The distance we are interested in minimizing in type II regression is the perpendicular line z shown in Figure 3.1. y' represents the vertical distance between y_i and the fitted line $\boldsymbol{X\hat{\beta}}$, while x' represents the horizontal distance.

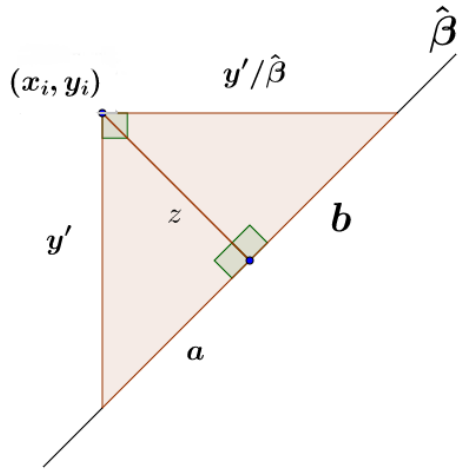


Figure 3.1: Geometry of Type II Regression

We can relate the different axes of the figure through equations

$$\begin{aligned} a^2 + z^2 &= y'^2 \\ b^2 + z^2 &= (y'/\beta)^2 \\ (a+b)^2 &= y'^2 + (y'/\beta)^2 \end{aligned}$$

Using substitution and recalling that $y' = y - \hat{y}$ we obtain the expression

$$z = \frac{y - \hat{y}}{\sqrt{1+\beta^2}}.$$

At this point we note that $y - \hat{y}$ represents the residual of type I regression fit, and therefore if we can assume the residuals to be normal then z has the scaled normal distribution

$$z \sim N\left(\frac{y - (\hat{\beta}_0 + \hat{\beta}_1 x)}{\sqrt{1+\hat{\beta}_1{}^2}}, \frac{\nu_y + \hat{\beta}_1{}^2 \nu_x}{1+\hat{\beta}_1{}^2} \right), \tag{3.1}$$

where ν_x and ν_y represent the total variation in x and y respectively. Since we are dealing with measurement uncertainties, the total variation can be split into an intrinsic element σ^2 and an uncertainty element δ^2 so that $\nu_x = (\sigma_x^2 + \delta_x^2)$ and $\nu_y = (\sigma_y^2 + \delta_y^2)$.

In order to fit estimates of the regression parameters Harris *et al.* (2013) used a Chi-square statistic based on equation 3.1 where the parameters are chosen to minimize the sum

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1(x_i - \bar{x}))^2}{(\sigma_y^2 + \delta_y^2) + \beta^2(\sigma_x^2 + \delta_x^2)}. \tag{3.2}$$

The data is centred in order to minimize the correlation between the intercept and slope. This χ^2 statistic was defined by Press *et al.* (2007) and was tested extensively

by Tremaine *et al.* (2002) and Novak *et al.* (2006). Simulations by both authors supported the use of the χ^2 claiming it provides the most efficient and unbiased estimates of slope.

Without any external constraints it is impossible to solve simultaneously for σ_x^2 and σ_y^2. For this reason we follow the lead of Harris *et al.* (2013) and let $\sigma_x^2 = 0$, assuming that all intrinsic variation is in the y coordinate and that any variation in the x direction is due solely to measurement uncertainty.

Chapter 4

Imputation of Incomplete Data

4.1 Data and Previous Work

This project applies the knowledge of the previous sections to a dataset supplied by Dr. William Harris of the Department of Physics and Astronomy at McMaster University. The dataset is a compilation of information from large surveys as well as 112 published papers obtained through a thorough literature search by Harris and his colleagues in preparation for their work in Harris *et al.* (2013). Because of its synthetic nature, the data lacks homogeneity across observations, and contains many missing values. The data can be found in its entirety at `http://physwww.mcmaster.ca/~harris/GCS_table.txt`.

Globular clusters are spherical collections of stars that orbit the centers of their host galaxies as single satellites. Some galaxies contain as many as 30,000 clusters, while our own Milky Way galaxy contains 160 (Harris *et al.*, 2013). The primary goal of Harris *et al.* (2013), and of this re-analysis, is to understand what determines the size of a galaxy's globular cluster population. The quantity of interest is the number

of these clusters N_{GC}; we wish to find a set of predictors that accurately determine this quantity. Understanding the properties of these clusters provides information on the structure of galaxies as well as insight into their early evolution.

The following is a subset of the data provided by Harris used for analysis in determining the relationships between N_{GC} and other variables. Some information available in the data has been discarded due to lack of usable cases (such as black hole mass, with $> 85\%$ of cases missing), or lack of relevance to the analytic procedures (such as variables relating to the location of galaxies in the sky). This project makes use of the following quantities:

- Galaxy identification.
- Galaxy morphological classification T within the Hubble sequence. The Hubble sequence classifies galaxies into 17 types by shape. T takes on an integer value between -6 and 10. Galaxies with values close to -6 are elliptical, while those nearer the center of the scale are disk-type in structure. A value closer to 10 identifies an irregular shaped galaxy.
- Distance $d \pm \delta_d$ (Mpc) of the galaxy from the Milky Way. (1 parsec $\approx$ 3.3 lightyears.)
- Foreground absorption A_V (unitless) representing the fractional light energy lost to absorbing matter between the galaxy and the observer.
- Absolute visual magnitude $M_V^T \pm \delta_M$ (unitless). Brightness of the galaxy independent of distance.
- Δ_{bv} (unitless), the difference between the amount of light received through the blue filter (B) and the green/yellow filter (V).

- Total number of globular clusters $N_{GC} \pm \delta_N$. This is the response variable and is meristic by nature.

- Stellar velocity dispersion of the galaxy $\Sigma_e \pm \delta_\Sigma$ (m/s).

- Effective radius $R_e \pm \delta_R$ (m). The radius within which half the light of the galaxy is enclosed.

- Dynamical mass of the galaxy $M_{\text{dyn}} \pm \delta_{\text{dyn}}$ (unitless). A function of dispersion and radius. Dynamical mass is defined as

$$M_{\text{dyn}} = \frac{4R_e\Sigma_e^2}{GM_\odot}, \tag{4.1}$$

  where G is the gravitational constant 6.67×10^{-11} $\text{m}^3/\text{kg}\cdot\text{s}^2$ and $M_\odot = 1.9891 \times 10^{30}$kg is the mass of the Sun.

- Total K-band magnitude $K_t \pm \delta_K$ (unitless). The amount of light observed through the near-infrared K-band filter.

In their available case analysis of the data Harris *et al.* (2013) found that N_{GC} correlated well with both the dynamical mass of the galaxy M_{dyn} as well as the product of dispersion and radius $R_e\Sigma_e$. We use these correlations as a starting point for our investigation, following the process performed by Harris but applying it to our complete imputed data.

We notably diverge from Harris' process only in the use of the morphological type T. In his paper Harris subsets the data into four distinct galaxy types: Elliptical, Spiral, Lenticular and Irregular, and obtains results specific to subsets of the data. We instead make use of T as a predictor with 17 values to allow predictions of all

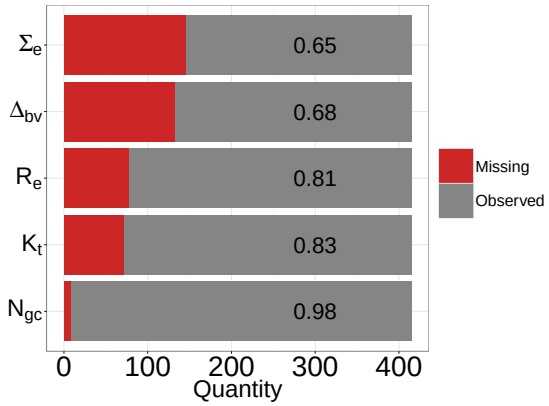
galaxy types from a single model. As is common in the field of astronomy, quantities tend to be linearly related when both are on the log(10) scale. Therefore we treat all quantities as power laws and inspect them under a logarithmic transformation. From this point on all values will be assumed to be under a log(10) transformation, so that R_e refers to $\log_{10}(R_e)$ and is measured in $\log_{10}(m)$.

4.2 Inspecting the Properties of the Missing Data

Before imputation begins we must first inspect the data to understand the nature of the missingness, mainly the missing data pattern and the missing data mechanism. Due to the nature of the data the missing values do indeed represent observable quantities that can be measured, and therefore imputation of these values is reasonable.

Figure 4.1 displays the proportion of observed data for the five incomplete quantities.

Figure 4.1: Proportion of data observed for incomplete columns.



Dispersion Σ_e has the largest amount of missing data with 146 of 415 values missing. Δ_{bv} also has a significant amount with 133 values unobserved. R_e and

K_t both have more than 80% of values observed and our response N_{GC} has only 8 artificially missing values. We label these 8 missing values "artificial" as they exist in the original data as values ≤ 0. Since these values will not transform to the log scale we have allowed them to be imputed as missing values.

It may be of interest to note that the dataset also contains 165 missing values for M_{dyn}. Because this mass is a derived quantity and is not directly observed, it is absent when either Σ_e or R_e are. It will also not be imputed for the same reason, as its values can be calculated directly from the imputed values of Σ_e and R_e.

The quantities absorption (A_v) and distance (d) both have a single missing value as they are not definable for the Milky Way. Because they are missing only in a single case we do not inspect their missingness. This may seem to go against our assumption that all missing values must represent existing values; however these values will not be used in the model determining the response N_{GC}. They will be used only in an auxiliary sense to impute other incomplete variables. The assumption can also be interpreted as a question of an overall understanding of the data, rather than a deterministic requirement for individual observations.

In inspecting the missing data patterns we also ignore the missingness of uncertainty columns for now, as they are missing strictly when their parent variable is missing and therefore follow the same pattern. Figure 4.2 displays this pattern for the first 50 rows of the dataset. Complete data is represented in white while black represents missing data.

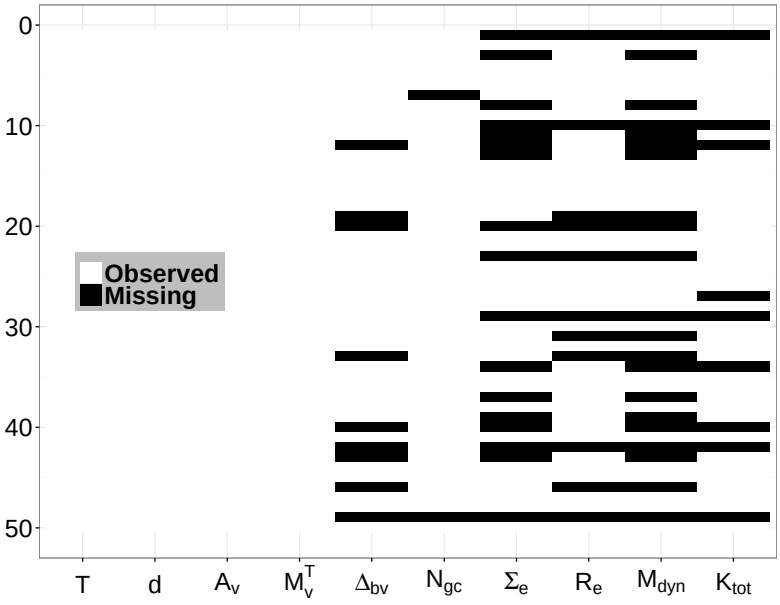


Figure 4.2: Sample of the missing data pattern.

The figure shows no clear patterns other than the tendency for some quantities to be missing simultaneously. This pattern is further inspected in Table 4.1, which summarizes Figure 4.2 numerically, identifying the patterns of the missing data by looking at which variables are missing for each row. The interior columns indicate whether a variable is missing (0) or observed (1). The first column represents the number of observations of that pattern, and the last column represents the number of missing values in each of those observations.

The first row of the table tells us there are 213 complete rows, while the second shows that there are 28 rows where only Δ_{bv} is missing. The last row shows the total number of missing values for each variable (which we inspected in Figure 4.1), for a grand total of 439 missing observations. From this table we further conclude that there is no obvious pattern to the missingness of the data, and although the 4 rightmost quantities perhaps tend to be missing as a group (in 31 cases), there is no

pattern here that warrants any special treatment from the methods of the previous sections.

n Rows	Galaxy	T	M_v^T	d	A_v	N_{GC}	K_t	R_e	Δ_{bv}	Σ_e	Missing/row
213	1	1	1	1	1	1	1	1	1	1	0
28	1	1	1	1	1	1	1	1	0	1	1
2	1	1	1	1	1	0	1	1	1	1	1
37	1	1	1	1	1	1	1	1	1	0	1
8	1	1	1	1	1	1	1	0	1	1	1
2	1	1	1	1	1	1	0	1	1	1	1
22	1	1	1	1	1	1	1	1	0	0	2
1	1	1	1	1	1	0	1	1	1	0	2
10	1	1	1	1	1	1	1	0	0	1	2
7	1	1	1	1	1	1	1	0	1	0	2
4	1	1	1	1	1	1	0	1	0	1	2
5	1	1	1	1	1	1	0	1	1	0	2
14	1	1	1	1	1	1	1	0	0	0	3
1	1	1	1	1	1	0	1	0	1	0	3
22	1	1	1	1	1	1	0	1	0	0	3
1	1	1	1	1	1	1	0	0	0	1	3
6	1	1	1	1	1	1	0	0	1	0	3
1	1	1	1	0	0	1	0	1	0	1	4
27	1	1	1	1	1	1	0	0	0	0	4
4	1	1	1	1	1	0	0	0	0	0	5
Totals	0	0	0	1	1	8	72	78	133	146	439

Table 4.1: Missing pattern.

When variables are being imputed, it is of interest to see the distribution of missing values across the missingness of other variables. For example Figure 4.1 identified 78 missing values of R_e and 72 missing values for K_t. This could indicate any pattern between the extreme cases of there being 150 rows where one of them is missing, or 72 rows where both are missing and 6 where only R_e is missing.

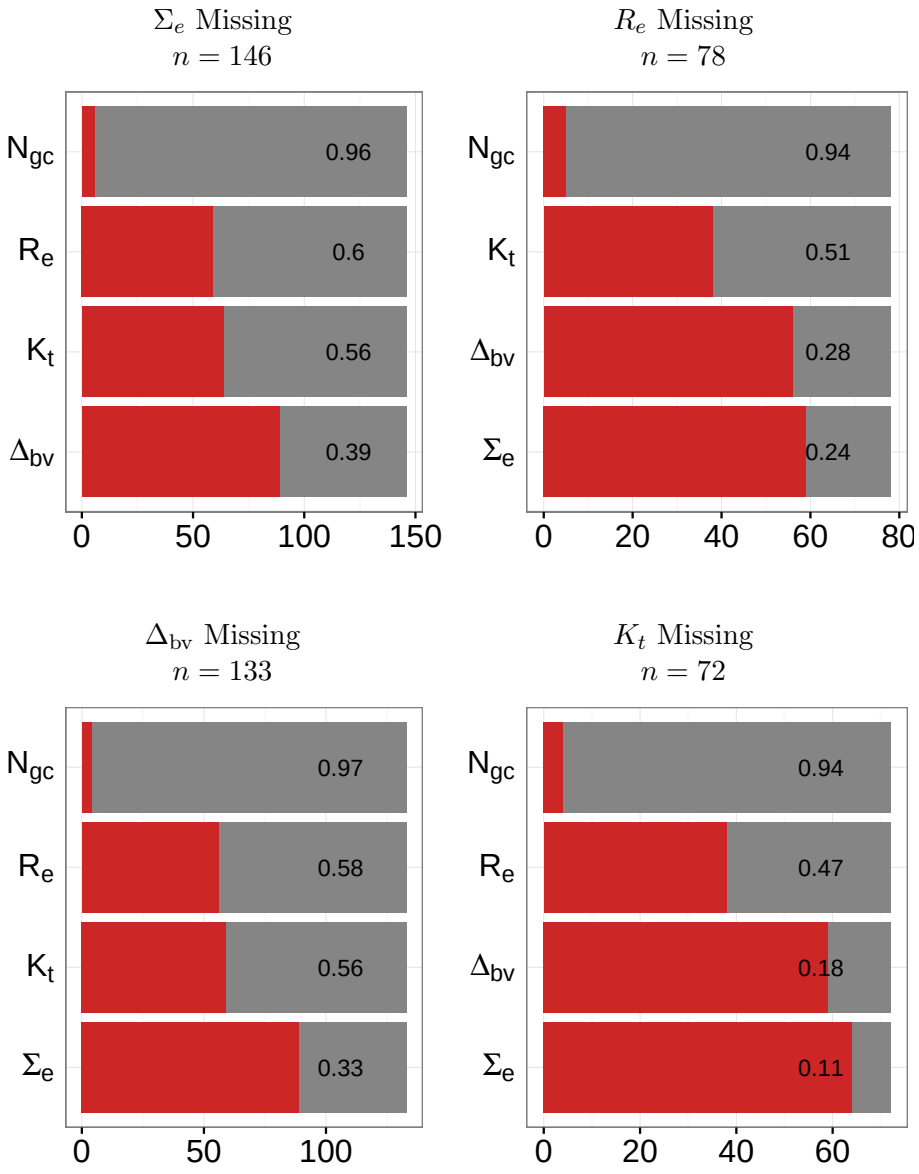


Figure 4.3: Missingness of other variables when a single variable is missing.

Figure 4.3 allows to investigate this by looking at the missingness within subsets of the data. Similar to 4.1 these graphs show the proportion of observed data per variable, however, each plot here represents a subset of data where a single variable

is missing. This allows us to compare the missingness of variables when one quantity is observed vs. when it is missing. Each plot shows the missingness of other variables on a subset of the data where the named variable is missing. The size of the subset is indicated by n.

The subsets where Σ_e and Δ_{bv} are missing show the quantities tend to be missing as a pair. The remaining plots show that Δ_{bv} and Σ_e are almost always missing when R_e is missing, and that all incomplete quantities tend to have a significant proportion of missing data when K_t is not observed.

The information from Figure 4.3 is useful for predictor selection within the imputation process. If both a variable and its predictor are missing then the imputations may be based on a large number of imputed values themselves, which can lead to a large variation in values across imputations. This error is accounted for through the procedures of section 2.3, and although the Gibbs sampler does a good job of drawing values from the joint distribution even when multiple quantities are missing, it is better to minimize this variance as much as possible by encouraging predictions to be based on observed values rather than imputed ones.

Next we investigate the missing data mechanism. We do so through the use of select correlation plots that indicate the location of missing values. Of course we cannot know the values of the missing observations in order to judge the mechanism by which they are missing; however, through the use of correlation plots we may gain understanding of how the quantities likely behave. In each plot the variable whose mechanism is being inspected is placed on the y axis, with the x axis being reserved for the variables sharing a considerable correlation with the variable of interest. The reason for plotting against correlated variables is that MAR assumptions can be

more easily justified when the auxiliary variables being conditioned on have greater predictive power. We can be more confident that a missing value y_i is a reasonable imputation if it is predicted by a highly correlated x rather than a less correlated one.

The plots in Figures 4.4-4.7 show the relationship between incomplete variables and their potential predictors. The locations of the missing values are represented by a red histogram, giving an impression of where these missing values lie on their covariate axes. Ideally we would view these relationships in a multivariate setting, however, with many correlated covariates for each incomplete variable we must inspect multiple univariate relationships.

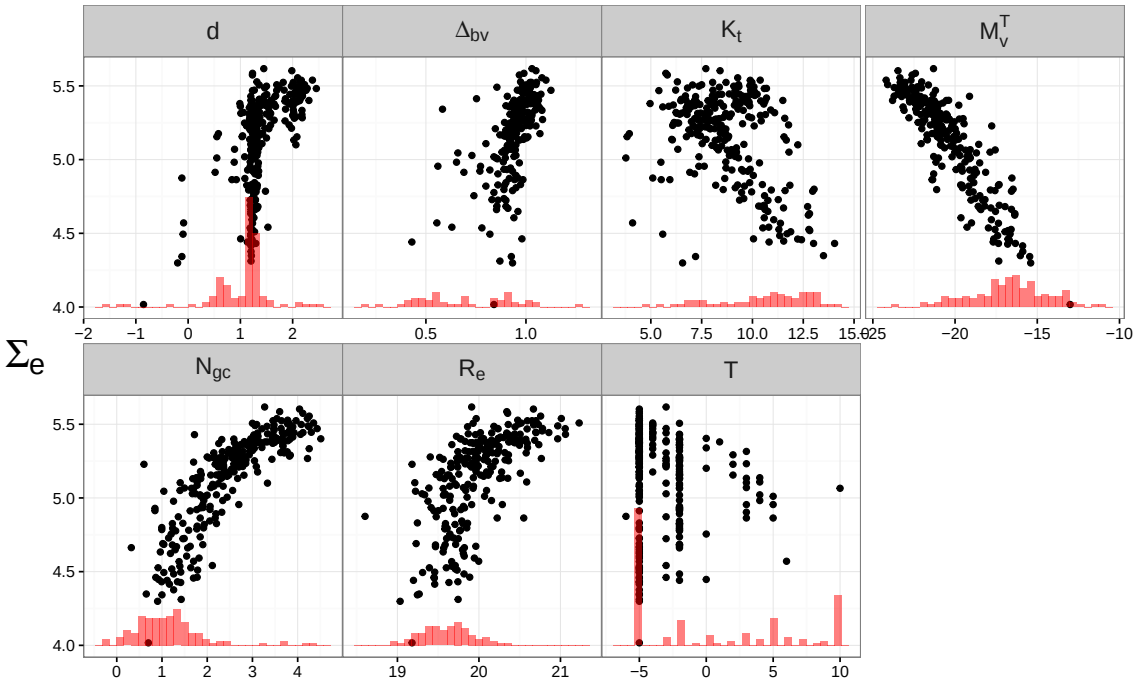


Figure 4.4: Σ_e against correlated quantities.

Figure 4.4 shows Σ_e plotted against 7 potential predictors. Almost every plot shows at least some missing values of Σ_e falling in an extreme range of its covariates. The plots against R_e and K_t show missing values somewhat mixed with observed

ones, but $M_v^T, \Delta_{\text{bv}}$ and N_{GC} show concentrations of covariate values beyond those of the complete cases.

R_e shows a much nicer distribution of missing values across its covariates. As seen in Figure 4.5 the majority of missing data lies within the observed range of values. M_v^T and N_{GC} show collections of missing values near the end of the observed values with a few falling outside the observed range. Values of R_e are clearly not missing completely at random, as the values are not dispersed evenly throughout the covariates, but they do lie mostly within the observed ranges of the data.

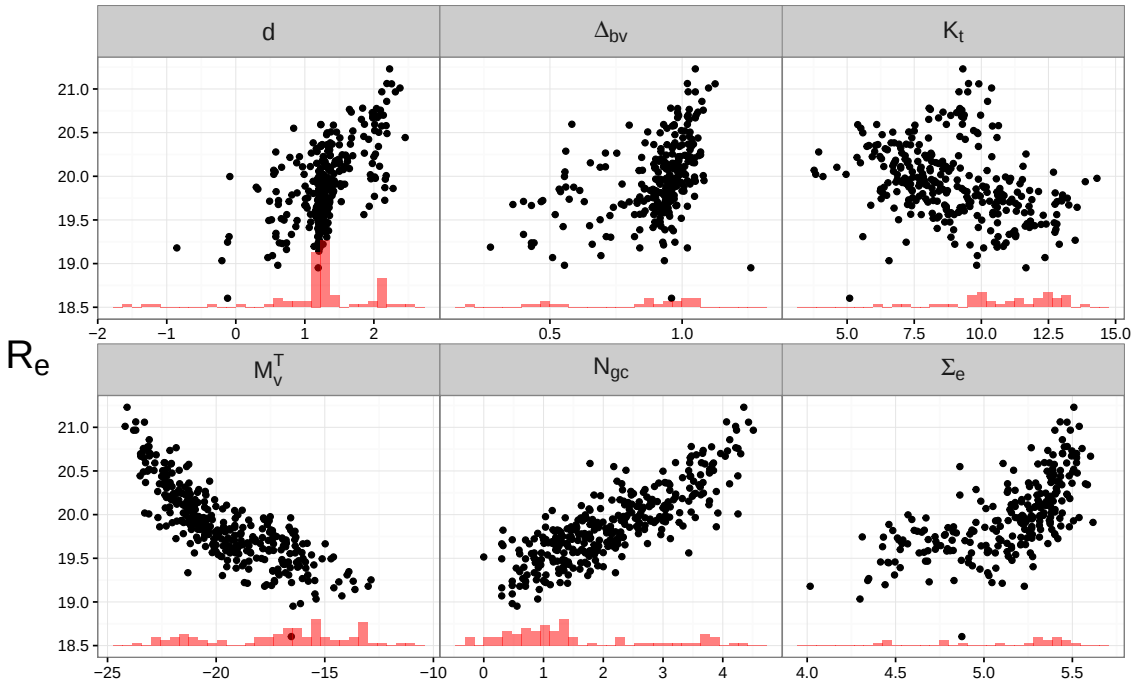


Figure 4.5: R_e against correlated quantities.

The distribution of missing values for Δ_{bv} is shown in Figure 4.6. We see that missing values generally occur within the range of the observed values with the exception of M_v^T, Σ_e and perhaps N_{GC} which have high concentrations of missing data at extreme values. The remaining covariates have missing data at concentrated points,

but these points fall within the range of observed data.

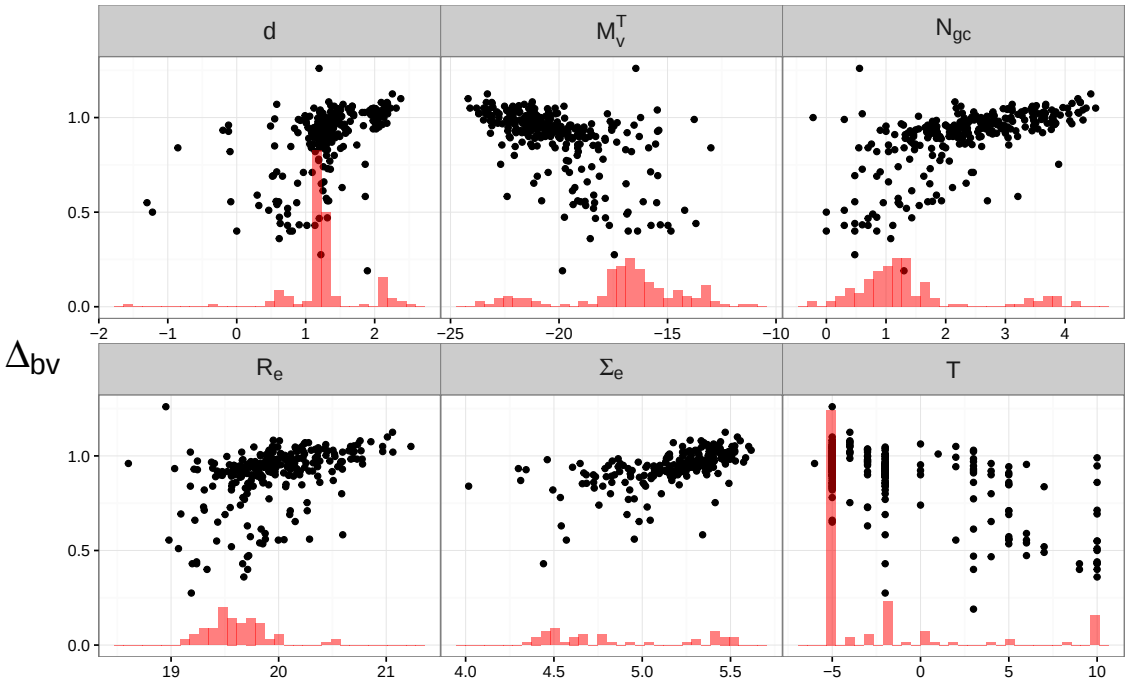


Figure 4.6: Δ_{bv} against correlated quantities.

The predictors of K_t in Figure 4.7 are probably the greatest concern to our assumptions. Most of the missing data lie at or beyond the edge of the observed data. R_e is the only predictor not having missing data beyond the range of the majority of the data. The correlations for K_t are also weak in general ($|\rho| \leq 0.5$).

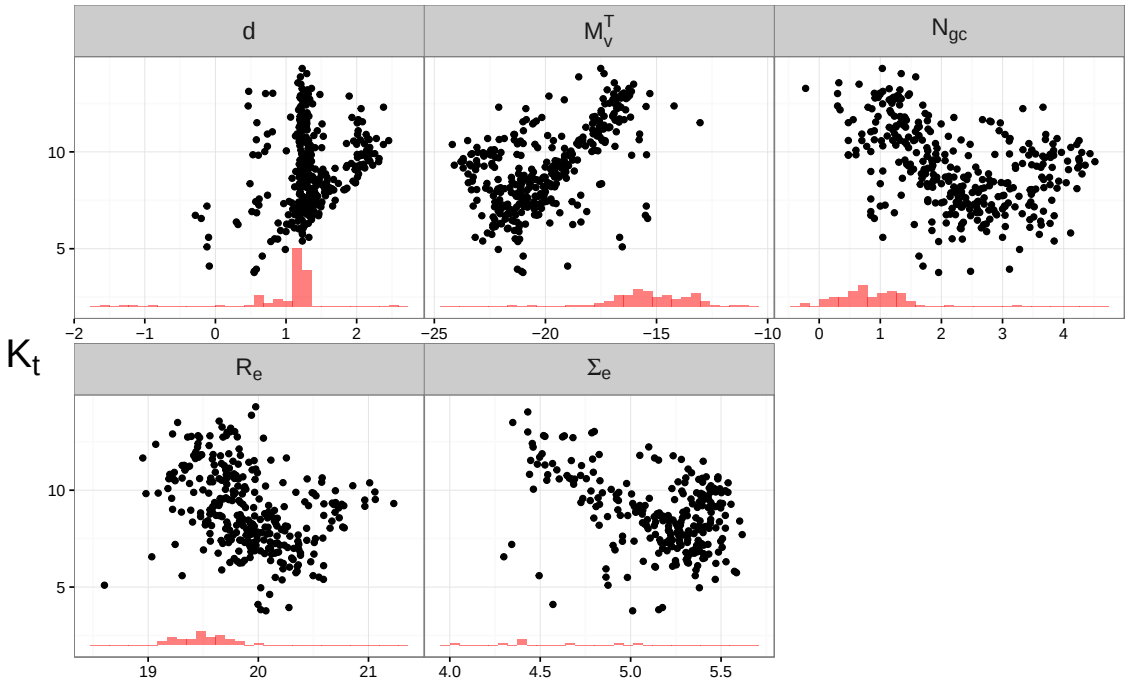


Figure 4.7: K_t against correlated quantities.

It is clear from the inspection of the missing data that to claim a MAR scenario would be tenuous. Although missing values generally have a reasonable spread across the values of the covariates, this distribution extends beyond the range of the observed values, and if we are to rely on the correlations, depend on the magnitude of the missing values themselves. This is the definition of NMAR. The issue with data that is NMAR is that extreme values are more likely to be unobserved, and we cannot say that the relationship that exists for the observed values can be extended to those extreme cases.

In a general observational sense it seems reasonable that missing values would tend to be those that are more extreme. It is likely more difficult to measure characteristics of a galaxy that is smaller in size, or emits less light. There may be thresholds at which values can no longer be observed because of the values of the covariates

such as distance, size or amount of light observed. Although it seems many of the values are missing due to these reasons, it is also likely that many are missing due to the heterogeneous nature of the dataset, and that many of the studies from which information was gathered, were simply not interested in some of the quantities that we are interested in. This may imply that a subset of the missing values are missing at random but we cannot claim this for the data as a whole.

In order to validate any imputation on these missing values we must ignore Little's suggestion to avoid extrapolating beyond the range of the covariates. We make a leap here and assume that the relationships between variables extend to the extreme values where the missing data lies. Without any information about the missing quantities in this range we have little choice but to extrapolate and assume the relationships will hold. With this assumption we can now shift the cause of the missing values from the incomplete variables themselves onto the complete variables they are correlated with. With the cause of the missingness assigned to the complete variables we can now justify treating the data as MAR, allowing the base assumption for the imputation process to hold.

To inspect the mechanism by which the *uncertainties* are missing it would be ideal to view their missing values on the scale of their parent variable, similar to the way the parent variables were viewed in Figures 4.4-4.7. Due to the pattern by which the uncertainties are missing this is not possible. Uncertainties are missing only when their parent variable is missing (and vice versa). This means there is no y value at which to plot the missing values. It makes little sense to use another variable for correlation since one can only assume that the observational uncertainty relates only to the quantity it represents (if there is any pattern at all). For this reason

the only inspection of these variables we perform is simply compute their correlation with their parent variables. Though this does not provide any real information on the mechanism behind the missingness, it helps to express how well missing values could be imputed by their parent.

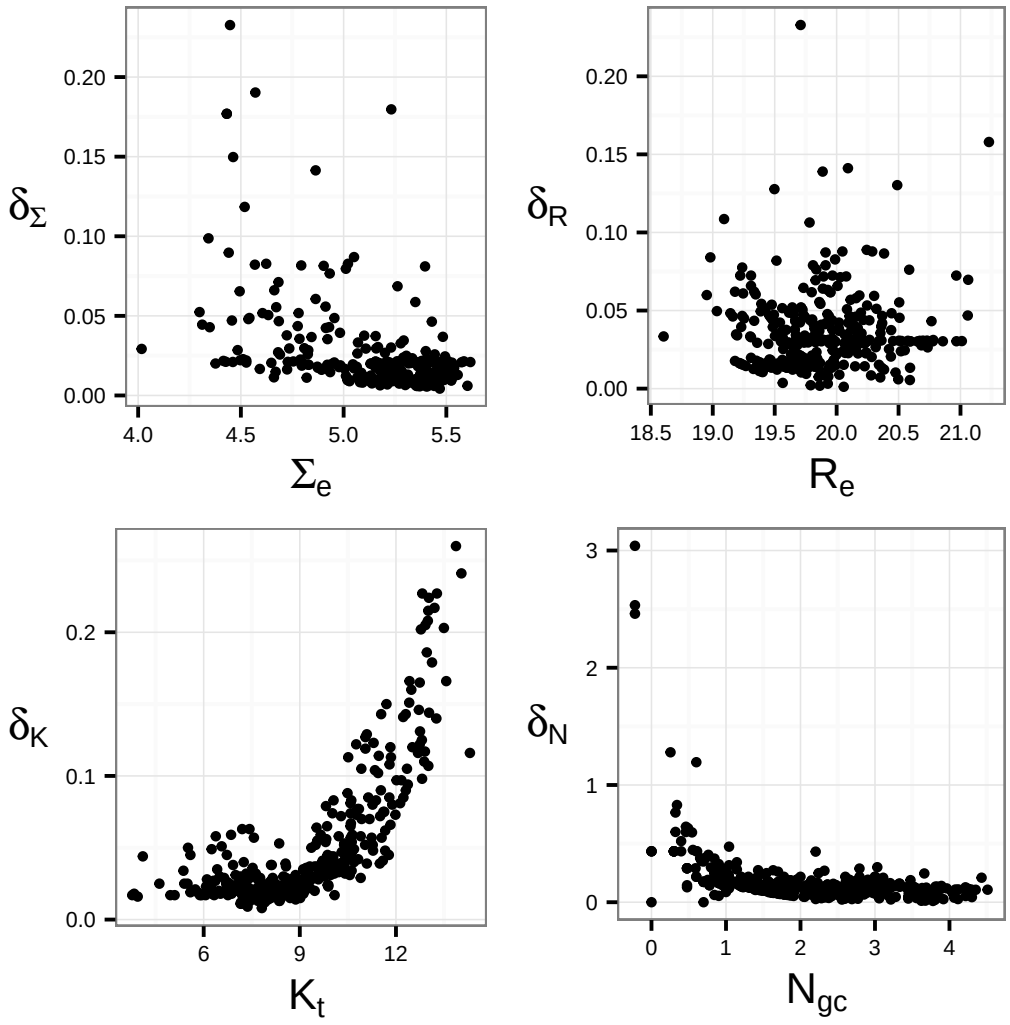


Figure 4.8: Correlation of uncertainties with their parent variables

Figure 4.8 shows clearly that the uncertainties generally do not correlate nicely (at least in a linear sense) with their parent variables. The correlations for K_t and

N_{GC} show strong nonlinear relationships, while those for Σ_e and R_e consist mostly of scatter. This scatter is possibly due to the heterogeneous nature of the dataset. The observations come from many different sources and thus may lack any homogeneity in the methods used to obtain the values. It is possible the precision of the measurements depends on the observer or their methods, and not any observable quantity, leading to a lack of consistency in determination of uncertainty.

Naturally we also investigated distance as a possible predictor for the uncertainty of measurements; however, this yielded even less desirable results, often with subsets of points creating a strong correlation of infinite slope.

4.3 Imputation of Missing Values

The imputation process begins with selection of single imputation methods for each variable. The columns in the dataset are classified into two separate groups: parent variables and associated uncertainties. We treat these two types of quantities differently, as as they possess distinct relationships. Intuition tells us that uncertainty columns should only be related to their parent variable, and therefore will be predicted by only the parent they correspond to. Parent variables will not be predicted by uncertainties as we assume that they depend on other quantities within the data, and not on the uncertainties associated with them.

For imputing parent columns we use Bayesian regression. With almost all information being continuous (with the exception of T), and strong linear relationships between variables, regression seems like the obvious choice. As discussed in Section 2.3 a Bayesian framework will help to propagate uncertainty about imputed quantities.

It is harder to define a method to use for imputing uncertainties. As seen in Figure 4.8 they do not correlate strongly even with the variable they represent. For this reason we chose an implicit method. Without the presence of strong correlations, explicit methods are likely a bad choice as they depend heavily on proper model specification which is unlikely here. Although the correlations of uncertainties with their parent variables were less than ideal, with no other information relating to the nature of the errors we have little choice but to make use of this small amount of information and predict uncertainties from only their parent variables.

While the hot deck was considered, we decided to use predictive mean matching for the uncertainty columns as both the response and predictor are continuous, and PMM avoids forcing continuous variables into categories. With only a single covariate predicting missing values PMM and hot deck give very similar results.

To use the imputation methods above we must define the predictor matrix, determining which covariates will be used for imputation. The predictor matrix will define how well the imputation model fits the data, and properly setting it up can make the difference between good and bad imputations. The matrix is built around the correlations between variables, while considering the number of usable cases for each quantity. In complete data analysis regression may be based strictly on the correlations between variables. In a situation with missing values the procedure is slightly more complicated. If a predictor with a strong correlation tends to be missing when the response is missing, then it may not be as useful as its correlation suggests. This is why we must also consider the number of usable cases and the relationship of missingness between variables.

In a predictive imputation setting the proportion of usable cases is calculated on a

subset of the data where the variable being imputed is missing, rather than the entire dataset. We denote this value Λ', and ensuring this value is high enough avoids the scenario where a variable and its predictor are consistently missing as a pair. This avoids heavily imputing values based on imputed values.

The process of selecting variables for predicting is still not as simple as selecting values for ρ and Λ'. There is an interaction between these two quantities that requires an attention to detail. Suppose we set $\rho = 0.3$ and $\Lambda' = 0.6$. If we were to plug these numbers in and go we could be sure that all predictors would meet these requirements. It may be the case however, that some predictors are very highly correlated with slightly lower Λ', or that they are completely observed with a correlation barely below the threshold. For this reason once we set the minimum requirements we must inspect the predictor matrix and tweak it where necessary.

For a minimum requirement of correlation we chose to use 0.2. It may be reasonable to use no minimum requirement at all, and thus to use all available information, though with 9 variables included in the imputation process we felt that the majority of information would be available through those variables with at least a small amount of correlation. The minimum correlation ensures that the information in highly predictive variables does not get swamped by a large number of variables presenting small amounts of information.

With a questionable MAR assumption supporting our procedures it seemed helpful to consider a higher value for Λ'. We initially considered $\Lambda' = 0.7$, but inspection of the predictor matrix showed that this excluded many relationships with significant correlations such as $\Sigma_e \sim R_e$ ($\rho = 0.65, \Lambda' = 0.59$) and $\Delta_{\text{bv}} \sim R_e$ ($\rho = 0.57, \Lambda' = 0.57$). Lowering Λ' to 0.55 included numerous useful predictors while maintaining a

reasonable amount of observed information within each regression.

It is important to note that using predictors with large amounts of missing information will not necessarily create wrong imputations. Because of the nature of multiple imputation all uncertainty of imputed values (including those predicted from other imputed values) is accounted for, and the result of using imputed values for predictions is primarily an issue of noisy relationships with large uncertainties, rather than one of incorrect inference.

The resulting predictor matrix can be seen in Figure 4.9. Both axes contain all variables involved in the imputation process. The y axis represents the variable being imputed, while the x axis contains the predictors. This matrix is similar to Figure 2.3, though the numbers now represent the correlations between variables while the 0/1 nature is presented through color. Red squares indicate the variable on the y axis is being imputed by the variable on the x axis. For example distance d is being imputed from a regression on $T, M_v^T, N_{GC}, \Sigma_e$ and R_e. The quantities T, M_v^T and σ_M are not being predicted at all since they are complete. As stated above all uncertainties are predicted by only by their parent variable.

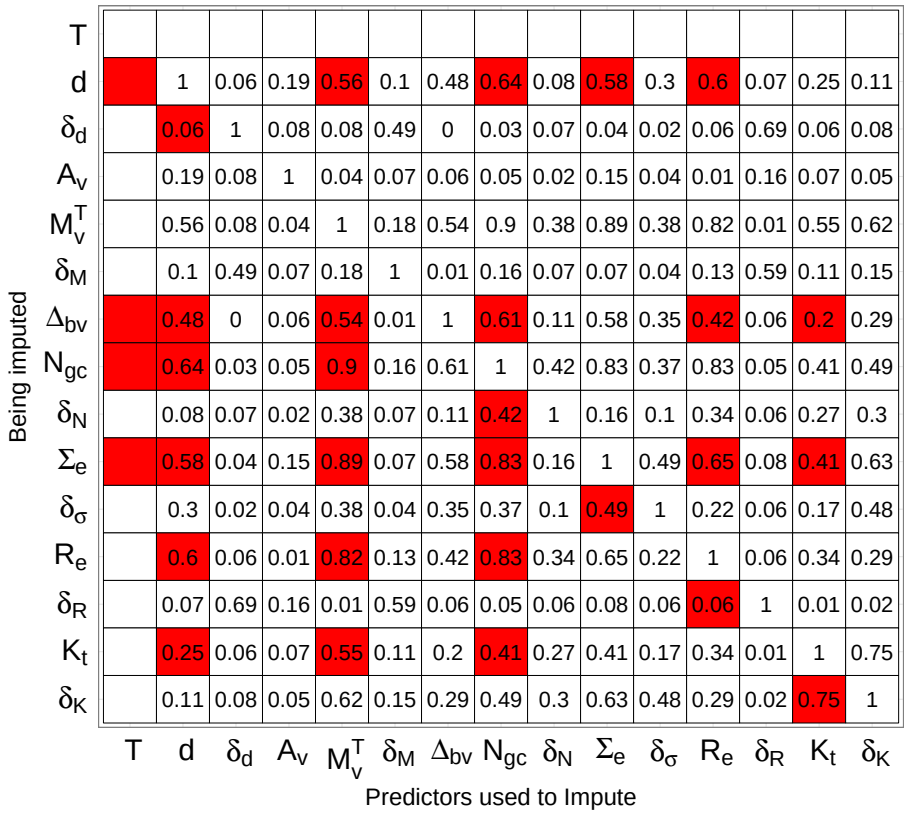


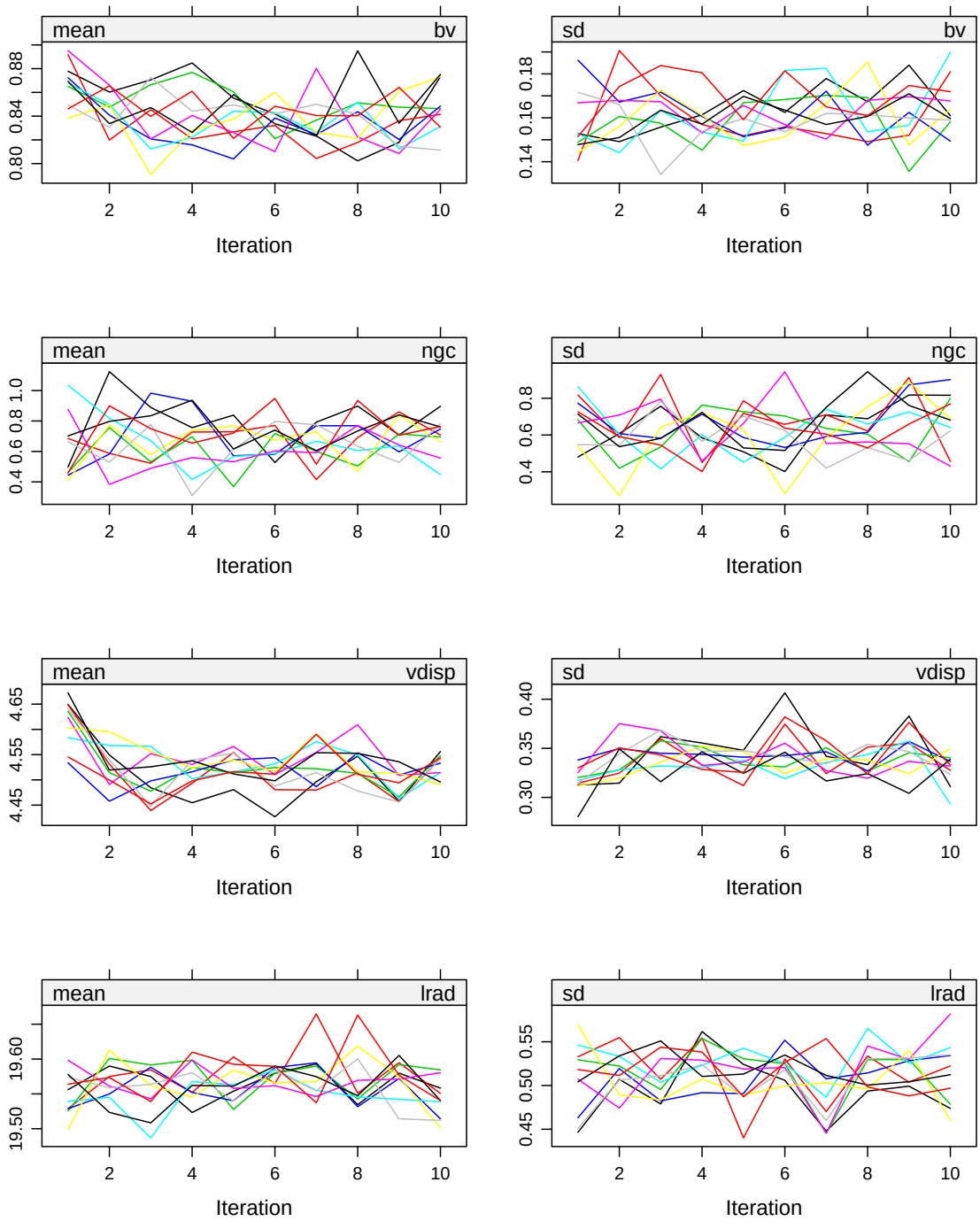
Figure 4.9: Predictor matrix for the imputation process.

Next we must decide on the number of iterations for the Gibbs sampler. We must ensure that the sampler converges to reasonable values for the draws. A nice quality about this is that aside from computational requirements there is essentially no downside to choosing a large number of iterations. A larger number of iterations can help to ensure that the process has indeed converged.

Brand (1999) performed simulations showing 5 iterations were sufficient for 31% missing data across four columns. With no real computational restrictions we increase this to 10 iterations for our data containing 21% missing data across 5 columns.

Although there is no absolute way of identifying convergence, a general method

is to accept convergence when the variance between imputations is not substantially larger than the variance for each individual sequence. The different chains should be nicely intermingled and show no identifiable trend.


mean
bv
0.80 0.84 0.88
2 4 6 8 10
Iteration
sd
bv
0.14 0.16 0.18
2 4 6 8 10
Iteration
mean
ngc
0.4 0.6 0.8 1.0
2 4 6 8 10
Iteration
sd
ngc
0.4 0.6 0.8
2 4 6 8 10
Iteration
mean
vdisp
4.45 4.55 4.65
2 4 6 8 10
Iteration
sd
vdisp
0.30 0.35 0.40
2 4 6 8 10
Iteration
mean
lrad
19.50 19.60
2 4 6 8 10
Iteration
sd
lrad
0.45 0.50 0.55
2 4 6 8 10
Iteration

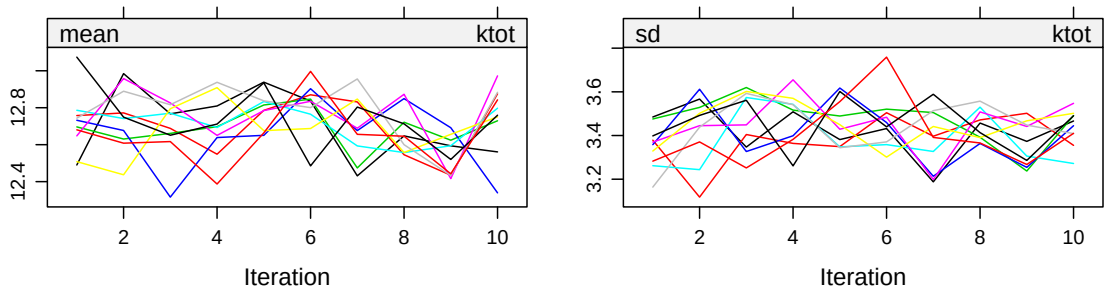


Figure 4.11: Gibbs sampler trace plots for imputed quantities.

The trace plots in Figure 4.11 show that the sampler converges very quickly as expected. Most variables even lack a burn in period, which is surprising since the mean of the observed data (which `MICE` uses as a starting point) is substantially different from the mean of the imputed data. Convergence here can be reasonably be claimed after approximately 5 iterations, though one could likely claim it in as few as 3 or 4.

The plots in Figure 4.12 show the trace plots for the uncertainties.

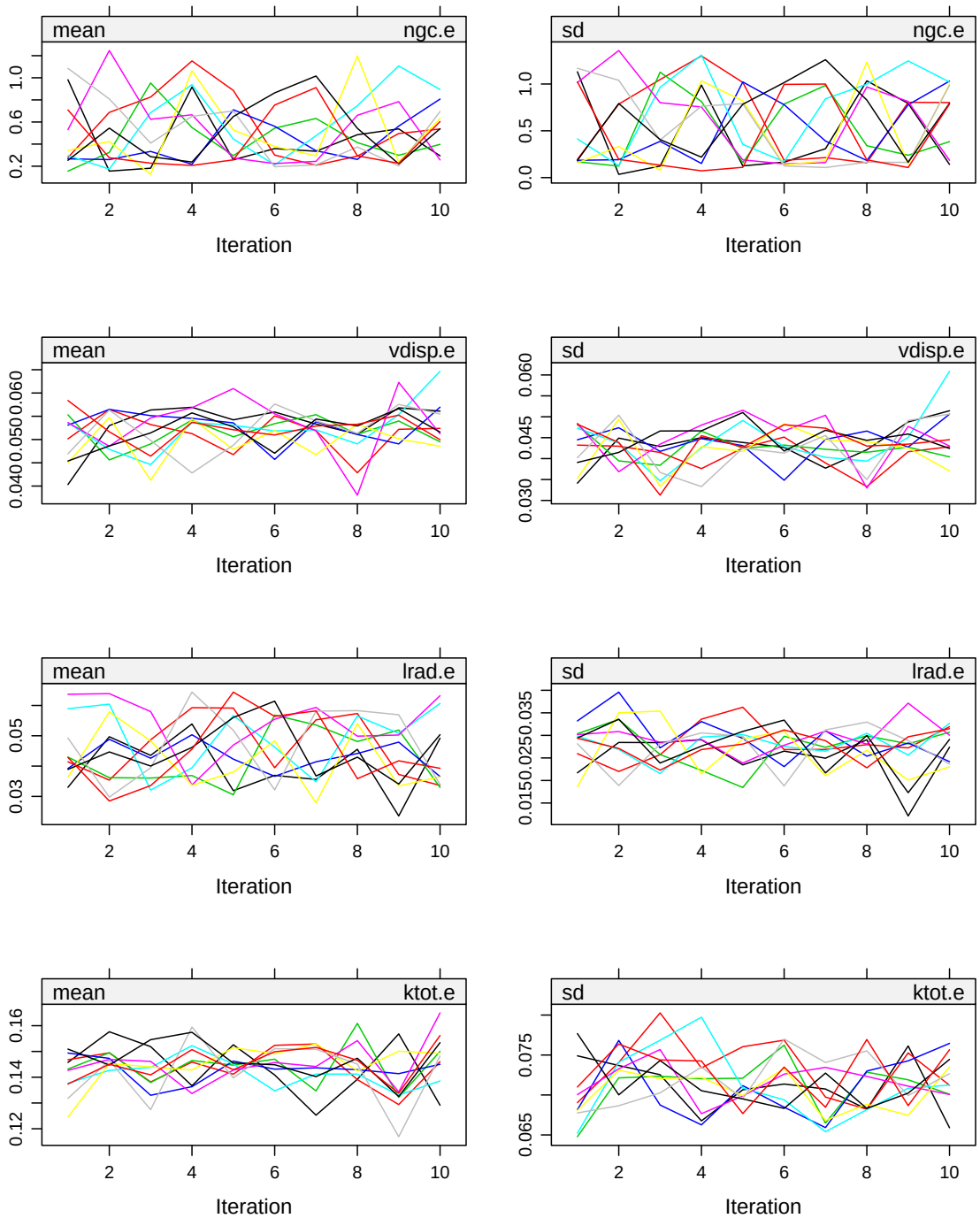


Figure 4.12: Gibbs sampler trace plots for imputed uncertainties.

Again the uncertainties converge nicely; perhaps even quicker than the parent columns. All trace plots show no sign of trends or patterns and present with the randomness we wish to see in a trace plot. This indicates the sampler has converged and that imputed values should be reasonable.

The number of multiple imputations must be chosen to ensure that the imputed values have enough of an opportunity to vary, allowing us to get a reasonable estimate of their variance. Similar to the selection of iteration size for the Gibbs sampler, the process will not suffer from an increased number of imputations aside from the computational power required. Larger numbers of iterations will only help to fill in the distribution of the parameters obtained from inference on the individual completed datasets. We used $M = 10$ as this seems large enough to capture the distributions of interest while still being quick to compute.

The last issue of concern is the sequence in which variables are visited by the sampler. With all variables involved in the process being unique (none are functions of others) the values imputed by the process are not affected by the sequence. Because of this we allowed the process to follow the default of `MICE` and visit variables in the order they appear in the dataset.

4.4 Diagnosing the Imputations

Now that we have imputed 10 independent datasets we inspect them to ensure that the values appear reasonable and that inference will be valid. In an ideal case where data is MCAR diagnosing imputed values would be somewhat straightforward, perhaps even deterministic. Because the missing values are a random subset of the entire dataset any comparisons between observed and imputed values should show strong

similarities. In our case missing values are often extreme and have covariates that range beyond those of complete cases, so diagnostics showing that imputed values fall close to observed values would likely imply that something has gone wrong.

The first diagnostics we perform are to ensure that imputations are consistent. We plot the distributions of imputed values across all ten datasets to see how the imputed data varies across datasets.

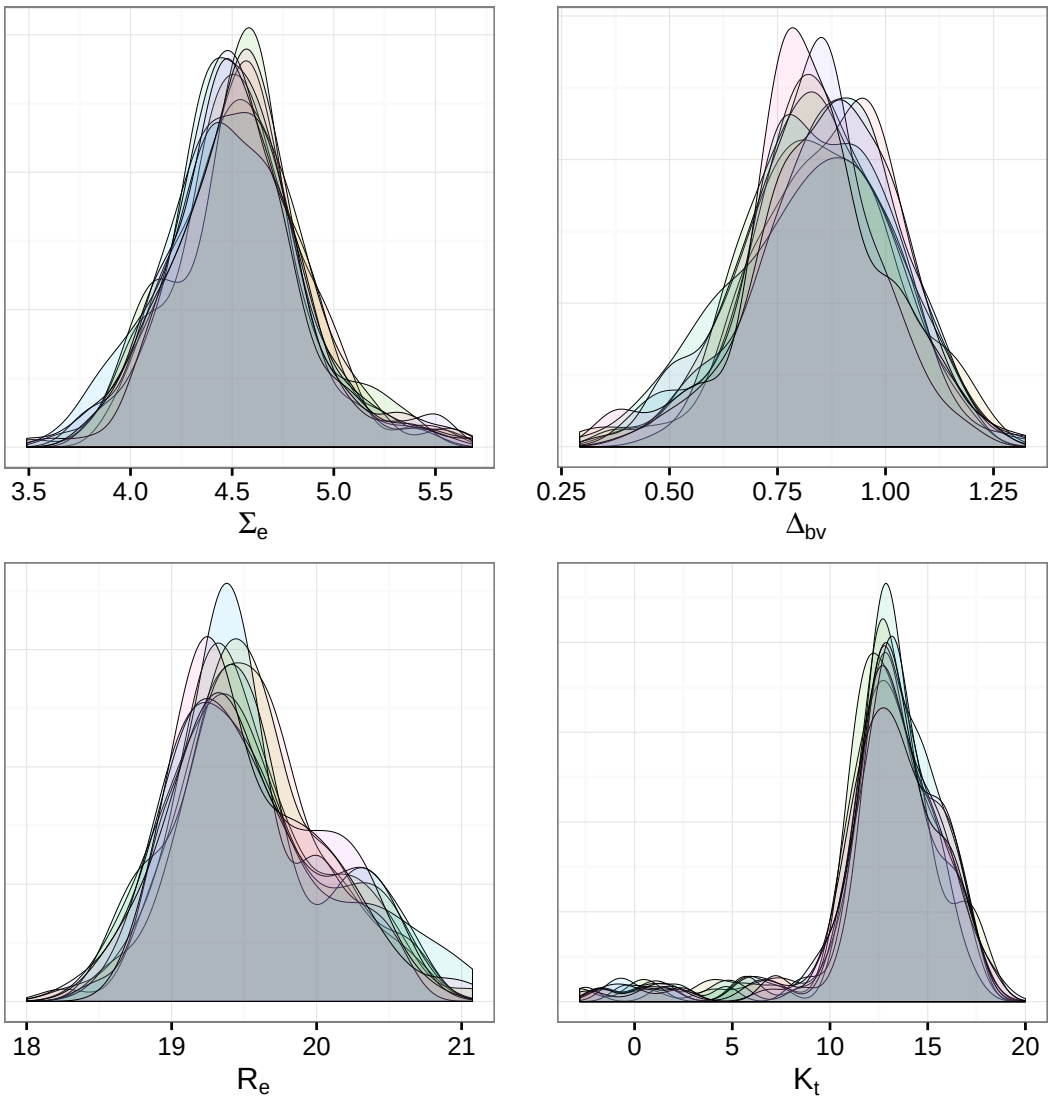


Figure 4.13: Distribution of imputed values across imputations.

Figure 4.13 shows that the distributions of all imputed variables are consistent with no imputations breaking substantially from the overall distribution. It is clear that the use of 10 imputations was sufficient to accurately cover the variation required, as almost all areas of the distributions are overlapped by multiple datasets. The only exception to this may be the peaks of Σ_e and Δ_{bv} where several datasets seem to vary

from the general trend in terms of the location of their peaks, though they quickly conform to the rest of the data as they move outward from the peak.

Next we inspect how the imputed values relate to the observed values. We collapse all the imputed values into a single distribution and compare this with the observed values for each incomplete variable. Because the imputed values are expected to differ from the observed values, we present these distributions along with the plots of the incomplete variables against their most highly correlated variables. These plots have observations as black dots, and imputed values as red crosses. Because imputations are reasonably consistent across datasets we display only a single imputation for each covariate plot, as displaying all imputed values for ten imputations becomes excessively cluttered without providing substantial information.

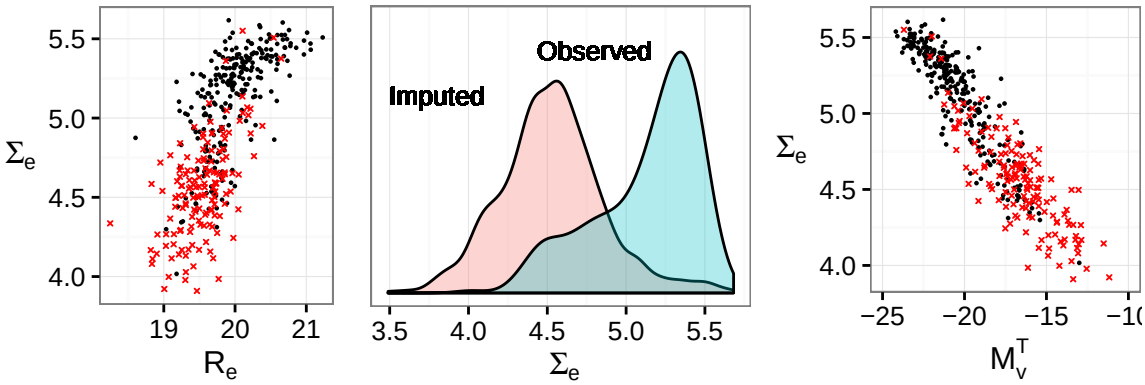


Figure 4.14: Inspection of Imputed data for Σ_e.

Figure 4.14 shows the behavior of the imputed values of Σ_e. We can see from the scatter plots the imputed values follow the trend of the observed values nicely. This explains the distinction between imputed and observed values in the distribution plot. On its own the distribution may seem worrying as the imputed values are completely different from the observed ones, but this is explained by the covariate plots and

asserts that the imputed values are indeed reasonable.

A similar yet less extreme pattern can be seen in the plots for R_e in Figure 4.15. Here the distributions are closer together as the imputed values fall more within the range of the observed data. The imputations again trend nicely with the covariates and the slight offset between distributions of observed and imputed data matches with the covariate plots.

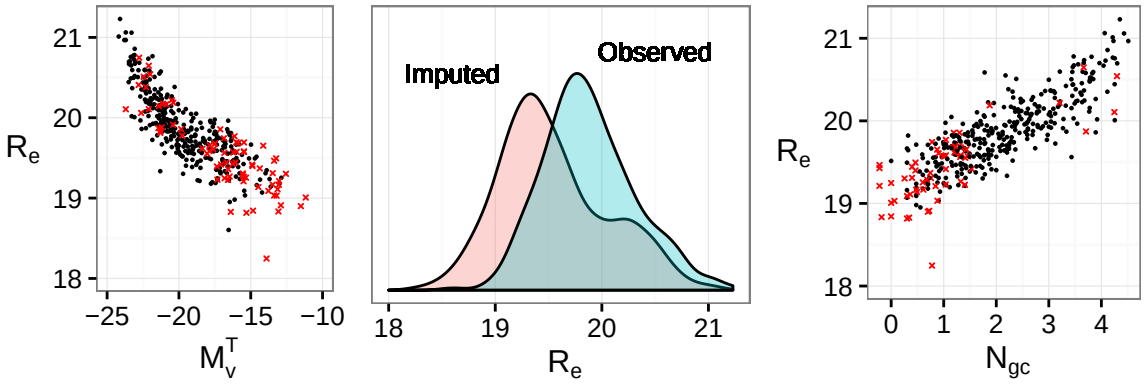


Figure 4.15: Inspection of Imputed data for R_e.

Inspecting Δ_{bv} in Figure 4.16 we immediately notice that it lacks the strong relationships present with Σ_e and R_e. Without strong linear correlations (all correlations have $|\rho| \leq 0.57$) the imputed values become quite scattered. All covariate plots show a single core of correlated values with a substantial number of values scattered around the central correlation. With imputed values falling at the extreme ends of these central relationships it is impossible to know whether the missing values should fall in line with tight correlation, or if they belong to the scatter.

The imputed values do however fall within the range of the observed values, implying they are indeed reasonable imputations that likely represent the unobserved values. The exception to this is the covariate plots for K_t and Σ_e where the imputed

values concentrate beyond the observed values. In these two plots especially it is impossible to know whether the imputed values should follow the tight correlation or the scatter as they do.

It seems strange that there is not more of a distinction here between the distributions of observed and imputed values given the extreme locations of the imputed data. This may be explained by the tendency for extreme values of a covariate being cancelled out by extreme values in another covariate. It is for this reason the distribution plot seems reasonable.

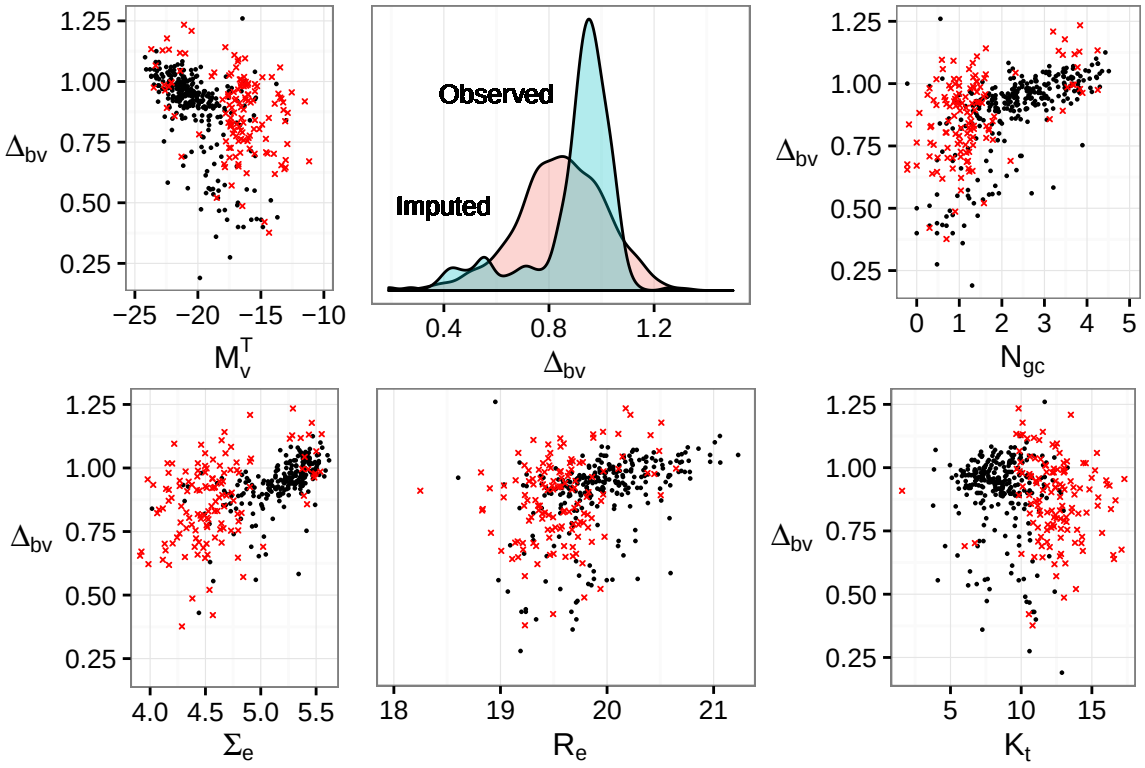


Figure 4.16: Inspection of Imputed data for Δ_{bv}.

The last quantity we inspect is K_t in Figure 4.17. Although the data is again beyond the range of the observed values, the imputed data follows trends nicely and

as expected the distribution of missing values is a very similar but slightly translated version of the observed data.

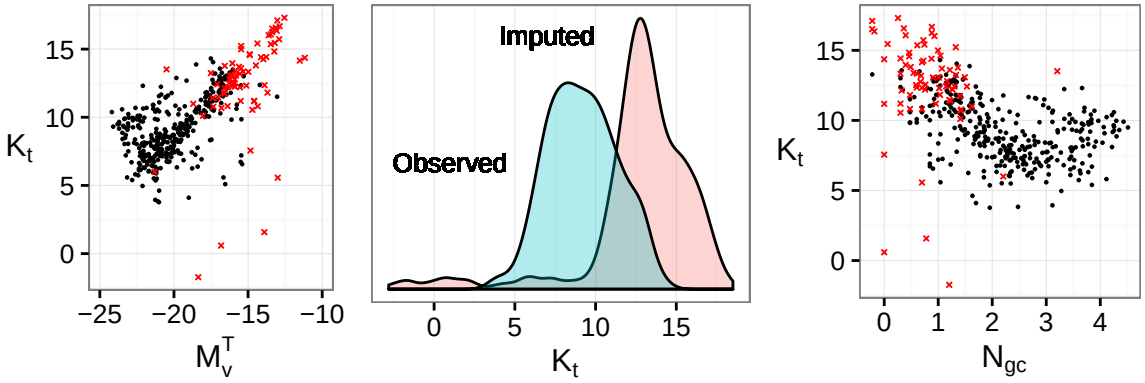


Figure 4.17: Inspection of Imputed data for K_t.

Next we inspect the imputation of the uncertainty columns. We first compare the imputed values across imputations. The uncertainties δ_Σ and δ_k in Figure 4.18 are nicely overlapped, with the latter having a little more spread across imputation. Both appear that the distribution was captured by the imputations and that the imputations were acceptably consistent. The distribution of δ_k shows more of a discrepancy between imputations, though the general shape of the distribution is maintained. This is likely due to the lack of correlation between the uncertainty and it's parent column. Unfortunately without more information about the nature of this uncertainty we are restricted to using these imputed values.

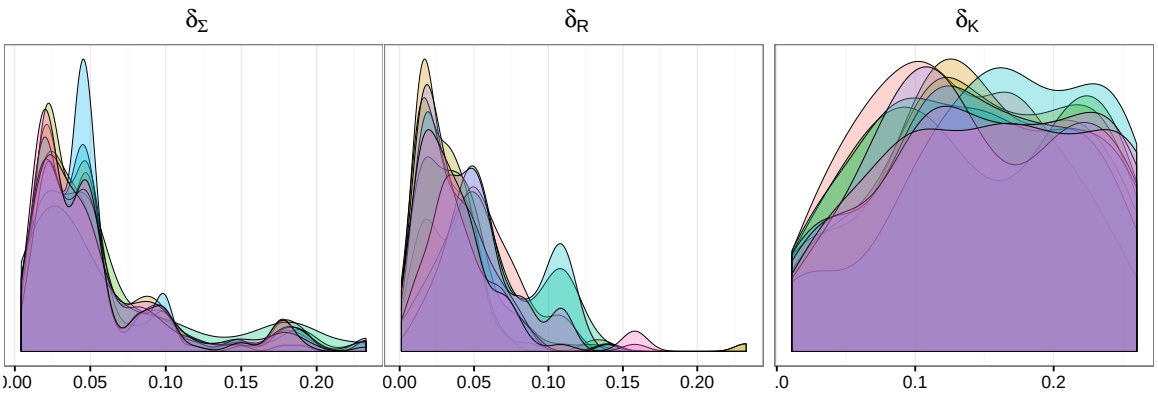


Figure 4.18: Distribution of Imputed uncertainty across imputations.

Now we compare the imputed uncertainties to the observed ones in Figure 4.19. There is a clear difference in style from the predictions of the main quantities as we are now using PMM to impute values. We can see that values are being assigned to nearby values rather than along a fitted line. This causes all uncertainties to be imputed with existing values. It is especially noticeable for δ_k where the missing values have been imputed by the observations with the highest values of K_t. This is because they will have the closest predictive mean as discussed in Section 2.2.3.

These imputations of δ_k really show the magic of PMM (and implicit methods in general). There is a clear quadratic relationship here, and were we to use regression here (especially second order), the imputed uncertainties would be incredibly large, and although they would be following the trend in the data, it is more likely that the uncertainties belong withing the range of the observed data, as PMM has done. Thus, the imputed values are allowed to follow the trend of the data, as long as they remain reasonable.

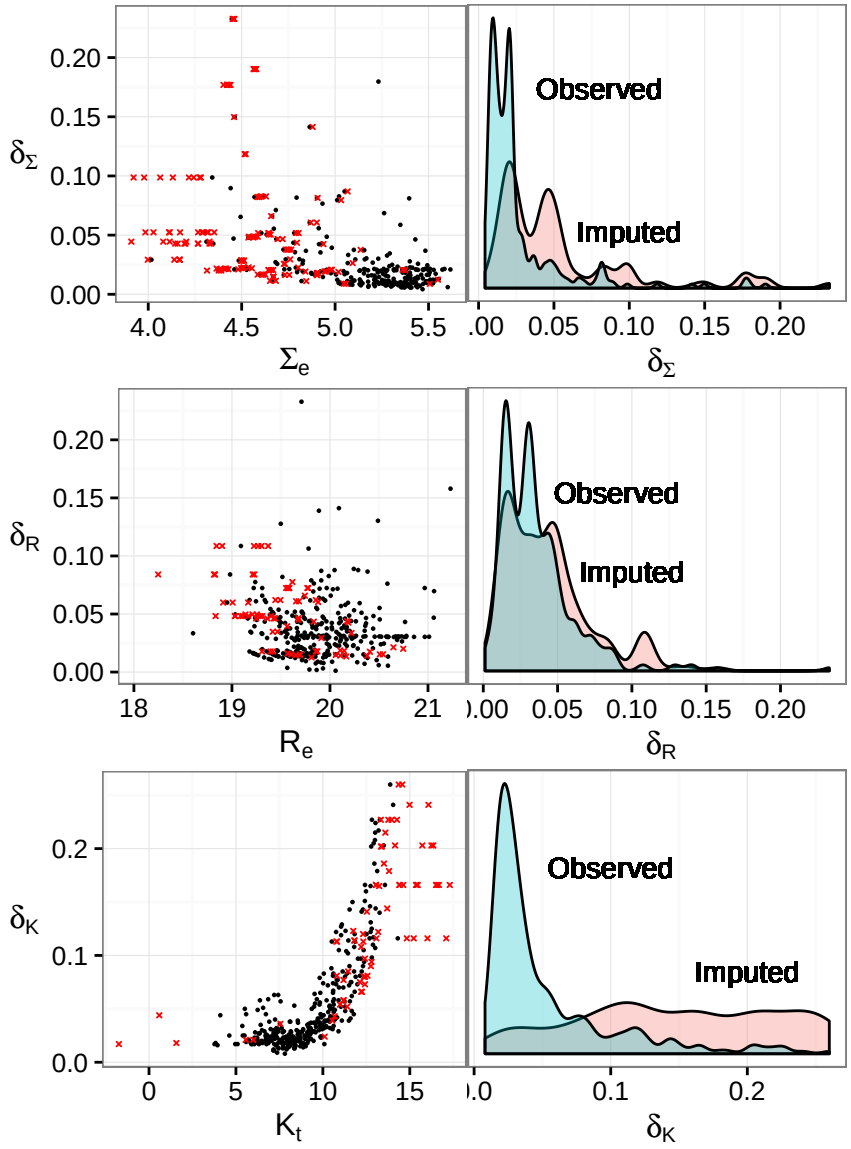
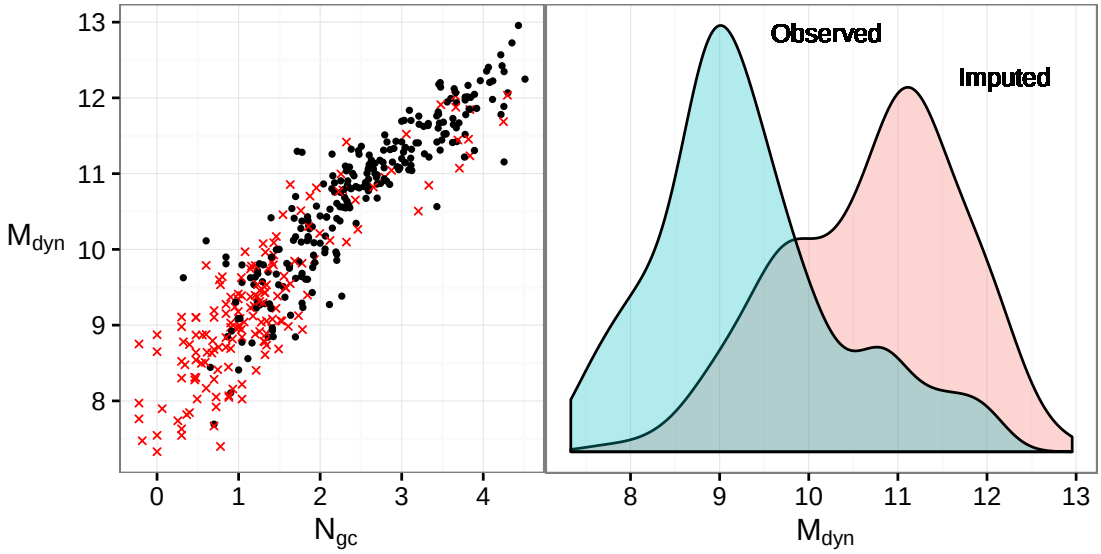


Figure 4.19: Inspection of Imputed uncertainty.

The last part of the data we wish to inspect is the values of M_{dyn}, which is the main predictor of our response. Although M_{dyn} was not imputed directly as it is a function of Σ_e and R_e we consider any rows where either Σ_e or R_e are imputed to be imputed values of M_{dyn}, since it is calculated from imputed values. We compare the

imputed to observed values in Figure 4.20



(a) M_{dyn} vs N_{GC} (b) Density

Figure 4.20: Comparison of imputed and observed values of M_{dyn}.

Much like the other quantities, M_{dyn} appears to have reasonable imputed values. They again extrapolate beyond the range of observed data but they do so in a manner consistent with the relationship seen in the observed data.

After inspection of the data we conclude that the imputed data is adequate for our purposes of analysis. Although the imputations for Δ_{bv} may be questionable, without more information at extreme values there is little that can be done. All other variables present favorable imputations that follow the trends of their covariates nicely. The uncertainties are also acceptable.

We conclude that the imputed data is acceptable and believe it accurately represents the missing values they replace. Now that we have ten completed datasets we may move on to analyzing them individually before combining their results.

Chapter 5

Analysis of Complete Data

5.1 Analysis of Individual Datasets

With a number of complete datasets at our disposal, we now wish to perform identical analyses on each of them. These results can then be pooled as described in Section 2.3 to obtain our final results. We begin by replicating the results obtained by Harris *et al.* (2013). This involves fitting the model $N_{GC} = \beta_0 + \beta_1 M_{\rm dyn}$. Harris fitted this model for 139 elliptical galaxies having $M_{\rm dyn} > 10^{10} M_{\odot}$. He also regressed $N_{GC} \sim R_e \Sigma_e$ for 158 elliptical galaxies.

We fit both of these correlations; first with an available case (AC) analysis of 248 galaxies, then with complete imputed data containing 415 galaxies. Initially we ignore galaxy type, and then investigate its potential as a predictor. We begin with $M_{\rm dyn}$ as the only predictor.

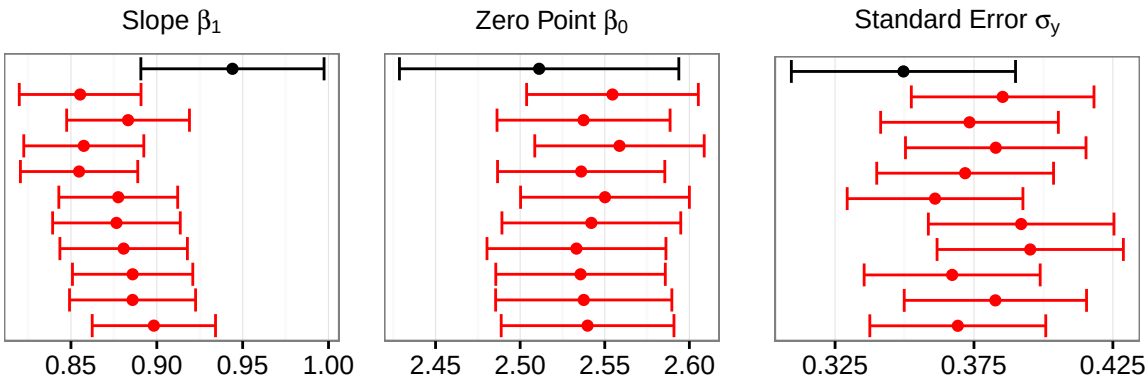


Figure 5.1: Estimates of regression parameters across imputations for $N_{GC} \sim M_{\text{dyn}}$.

In Figure 5.1 we see the estimates and uncertainty for β_0, β_1, and ϵ_y across the $m = 10$ imputations. Available case results are black, while results of the individual imputations are shown in red.

Estimates of slope for the imputed datasets are very consistent. They are slightly smaller than the slope of the AC result, though not significantly different. The zero point shows very high consistency between imputations, with all estimates falling almost completely within the AC interval. The estimate of standard error σ_y has increased slightly over the AC results, but again shows great consistency across imputations.

The resulting models are compared in Figure 5.2. The coefficient plots show the combined results of the imputation process compared with those from the AC analysis. The slope has shallowed slightly while the estimate of σ_y have increased. The intercept has remained essentially unchanged.

The plot on the right shows data from all ten imputations. The results of the AC analysis are shown as a dotted line. The imputed values show a large amount of variation relative to the observed data, contributing to the increase in σ_y. There is a

noticeable pattern here of multiple imputed datapoints falling on the same y value. This is due to the inclusion of all ten imputed datasets on a single plot. Since each missing value of M_{dyn} is imputed ten times, this gives ten imputed values at a single value of N_{GC}. With some values of N_{GC} existing in multiple incomplete observations, this leads to multiple imputed points associated with individual values of N_{GC}.

Overall none of the parameters show a significant difference from the AC results. The final model was determined to be $N_{GC} = (2.540 \pm 0.065) + (0.876 \pm 0.053) M_{\text{dyn}}$ with an estimated standard error of $\sigma_y = 0.378 \pm 0.044$.

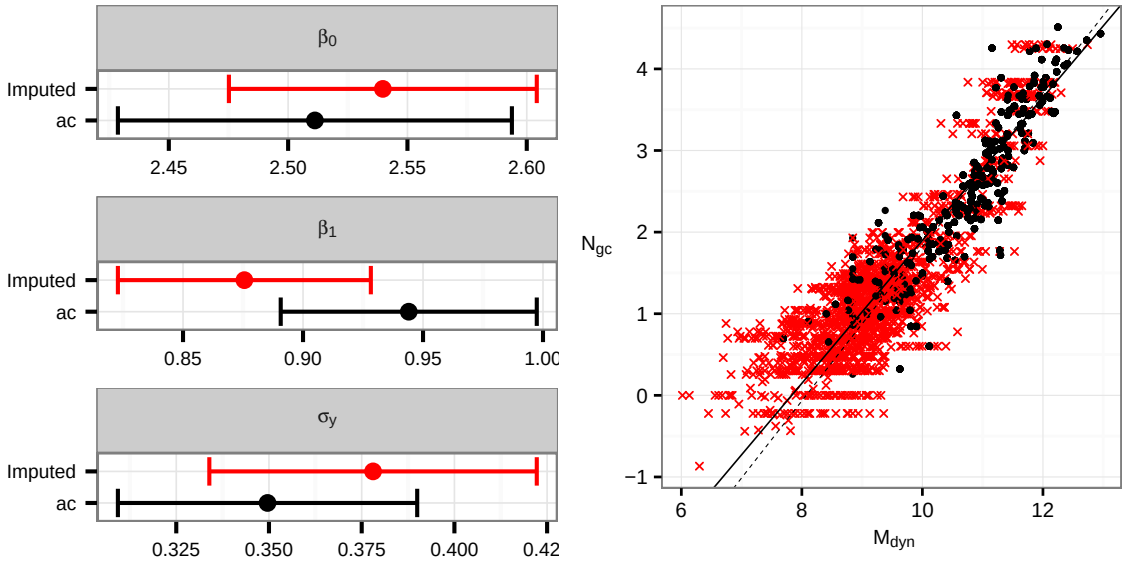


Figure 5.2: Fit for $N_{GC} \sim M_{\text{dyn}}$.

We now inspect the correlation $N_{GC} \sim R_e \Sigma_e / 1000$. Our AC analysis includes the same 248 galaxies as the analysis of M_{dyn}. Estimates from the individual imputations, as well as AC analysis are shown in Figure 5.3.

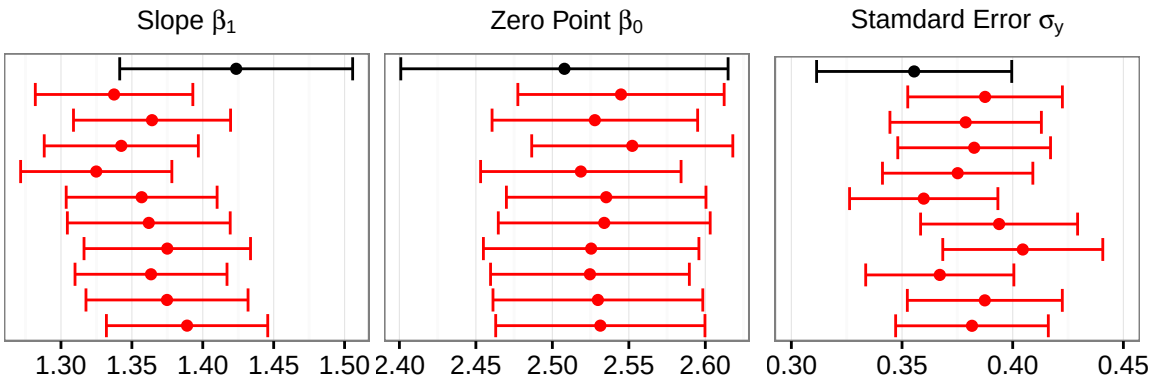


Figure 5.3: Estimates of regression parameters across imputations for $N_{GC} \sim R_e \Sigma_e / 1000$.

The results here are very similar to those from the correlation with $M_{\rm dyn}$, which was expected as the two relationships differ only by a factor proportional to Σ_e. Again intervals for slope and variance share substantial overlap, while estimates of the intercept for imputed data lie almost completely within the AC intervals. The combined results are shown in Figure 5.4, again compared to the AC estimates.

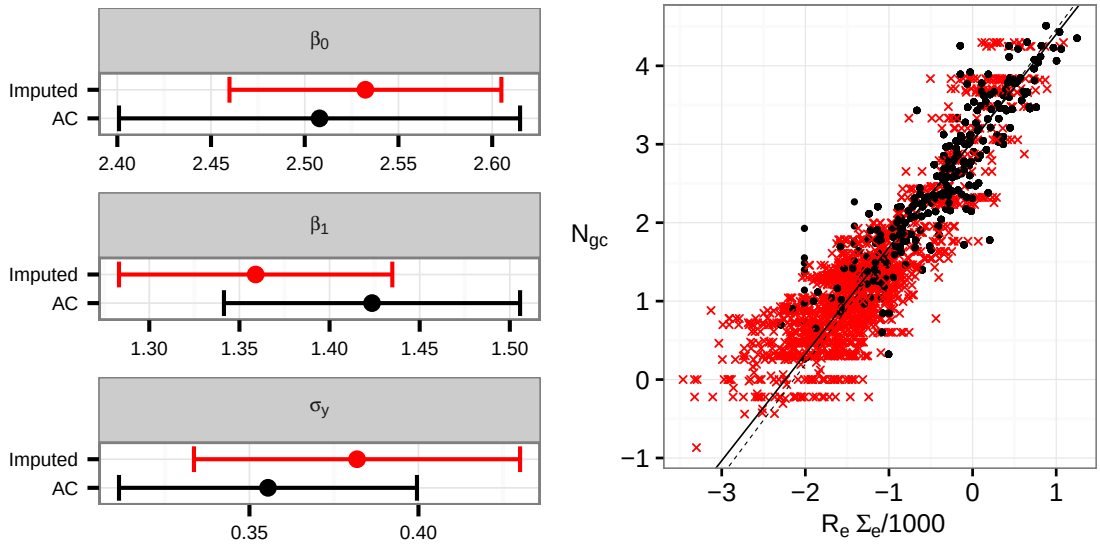


Figure 5.4: Fit for $N_{GC} \sim R_e \Sigma_e / 1000$.

Similar to the previous regression, the slope has decreased slightly and the standard error has marginally increased. Both intervals overlap with those from the AC analysis. The complete results of both correlations are shown in Figure 5.1.

Correlation	Source	N	Zero Point β_0	Slope β_1	σ_y
$N_{GC} \sim M_{\text{dyn}}$	Available case	248	2.511 ± 0.003	0.944 ± 0.053	0.349 ± 0.040
	Imputed	415	2.540 ± 0.065	0.876 ± 0.053	0.378 ± 0.044
$N_{GC} \sim R_e\Sigma_e$	Available case	248	2.508 ± 0.003	1.424 ± 0.082	0.355 ± 0.044
	Imputed	415	2.532 ± 0.073	1.359 ± 0.076	0.382 ± 0.048

Table 5.1: Comparison of regression coefficients

In both cases the slopes decreased slightly (though not significantly), and in fitting $R_e\Sigma_e$ a small amount of precision was gained. The fit to M_{dyn} did not gain or lose any precision. Overall the amount of variance increased slightly, which can likely be explained by the spread of the imputed data.

5.2 Galaxy Type as a Predictor

In this section we investigate fitting galaxy morphological type T as a predictor. We make this an addition to the model already containing $R_e\Sigma_e$, as it is of greater scientific interest. Although both quantities contain the same variables (Σ_e and R_e), M_{dyn} is a derived quantity, while $R_e\Sigma_e$ is directly observed. For this reason we move forward with a model including $R_e\Sigma_e$, and not one containing M_{dyn}.

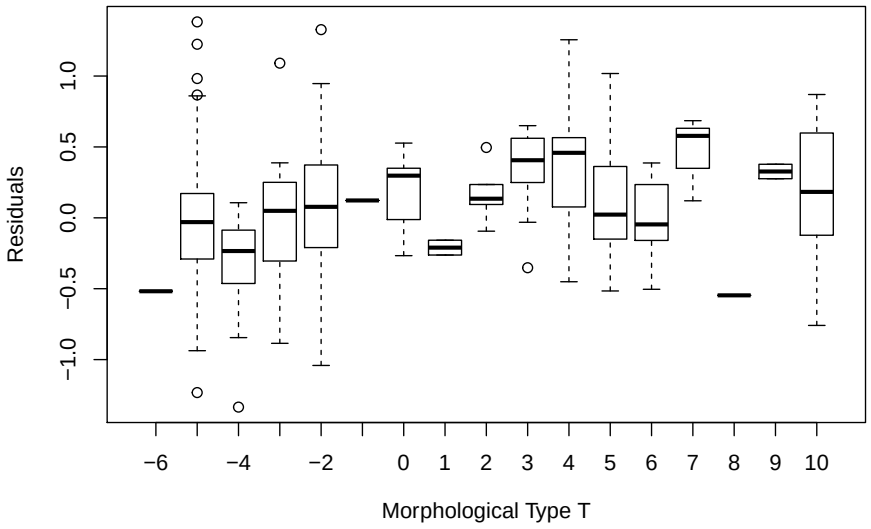


Figure 5.5: Galaxy type T against residuals of $N_{GC} \sim R_e\Sigma_e$

Figure 5.5 shows clearly that galaxy type provides some information, though it is somewhat sporadic. It would be most conservative to represent type on an ordinal scale as one would assume that galaxies of similar shapes would have similar effects on the response but that the magnitudes of differences among adjacent galaxy types might differ. This would require including 16 additional coefficients in the model, and is perhaps excessive. For this reason we will treat T as a numerical predictor. This approach assumes that the difference in the predicted response between any two adjacent levels of galaxy type is the same. Changing galaxy type from -4 to –5 will present the same change in the predicted value as moving from galaxy type 12 to 13. Though this may not be perfectly ideal, it will still provide information to the final model without the need for an excessive number of parameters.

In order to fit a model including galaxy type we extend the Chi-squared fit from

Equation 3.2 to include an extra variable. We do not center T here as this would only shift the "base level" for the predictor, which is arbitrarily defined. Without centering we allow the galaxy type corresponding to $T = 0$ to be the base level.

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1(x_i - \bar{x}) - \beta_2(T_i))^2}{(\sigma_y^2 + \delta_y^2) + \beta_1^2(\sigma_x^2 + \delta_x^2)}. \tag{5.1}$$

The extension of the numerator is straightforward, as it remains the residuals of the regression of y on its predictors. The denominator, however, is somewhat counter-intuitive. It may seem the added predictor should carry with it a variance term, adding a $\beta_2^2(\sigma_T^2 + \delta_T^2)$. However, this is incorrect if we inspect the meaning of these terms. $\delta_T = 0$ since there is no uncertainty associated with galaxy type, which leaves the variance σ_T^2. As discussed in Section 3.2 it is not possible (without introducing additional assumptions or information) to determine the variances for more than a single dimension. For this reason in Section 3.2 we allowed all variation to be attributed to y, and we will do the same here by letting $\sigma_T^2 = 0$.

We found that galaxy type was a significant predictor ($p < 10^{-4}$) in a model containing $R_e\Sigma_e$, though it provided only a small amount of information. The results from this fit are compared to the model without T in Table 5.2. Both analyses here are of the complete imputed data.

Correlation	N	Zero Point β_0	Slope $\beta_{R_e\Sigma_e}$	Slope β_T	σ_y
$N_{GC} \sim R_e\Sigma_e$	415	2.016 ± 0.074	1.359 ± 0.076	NA	0.382 ± 0.048
$N_{GC} \sim R_e\Sigma_e + T$	415	4.360 ± 0.165	1.316 ± 0.067	-0.024 ± 0.016	0.361 ± 0.045

Table 5.2: Comparison of regression coefficients

Adding T to the model greatly increased the intercept and slightly decreased the

effect of $R_e\Sigma_e$ on the response. The effect of T is negative and almost negligible. While precision in the estimate of the intercept has decreased, the effect of $R_e\Sigma_e$ is slightly more precise, and the model estimates less intrinsic variance.

This model is a slight improvement on one without galaxy type as long as the intercept is not the dominant scientific interest. The slope and variance both saw improvements and will provide slightly more accurate results.

Chapter 6

Conclusion and Future Directions

6.1 Conclusion

The goal of imputation was to increase the precision by having the power from additional observations outweigh the variance added from the process itself. Design of the imputation process is of utmost importance but success also depends directly on the amount of missing data, as even imputed data from a perfectly designed imputation model cannot replace legitimate observations. The entire process requires maintaining a balance between the precision gain from additional observations, and the variation added with their inclusion.

We found that for this data the imputation procedure had a very slight effect on the precision of estimates, as well as a small effect on the point estimates of the parameters. Only slope saw a reduction in the width of its interval. The estimate of the zero point saw a substantial increase in both its estimate and uncertainty, while variance showed just a minor change in the same direction.

Because the results show promise without an actual gain in precision, it is likely

that this data lies on the tipping point of having too much missing information for the imputation process to perform to its potential. Had the data set been slightly more complete it is likely that results would have presented estimates with greater precision than the complete case analysis. Conversely, any more missing information might have lead to a loss of precision when compared to the available case analysis.

6.2 Future Work

6.2.1 Imputation Procedures

Imputation of the parent columns is based on Bayesian linear regression models. Many correlations (such as those in Figures 4.4-4.8) suggest possible quadratic relationships between quantities (even when both are log-transformed). It may be beneficial to consider quadratic predictors for some incomplete columns. This would require extensive inspection of each relationship within the dataset, complete with thorough diagnostics of each model. If specified correctly, models employing quadratic predictors would likely reduce the residual variance of the regressions, allowing for more precise imputed values and greater precision in the final results.

The Bayesian regression used by the imputation process also fails to account for the observational uncertainty provided with the data. It is possible that the Chi-square statistic used in the final model could be applied to the the steps of the imputation process, allowing for better predictions that account for measurement uncertainty.

6.2.2 Quadratic final model

The relationship between N_{GC} and M_{dyn} for the observed data as well as the imputed data suggests that a quadratic fit may be appropriate, as can be seen in Figure 6.1. Although the focus of this analysis was the relationship $N_{GC} \sim R_e \Sigma_e$, the relationship with M_{dyn} may still be worth investigating.

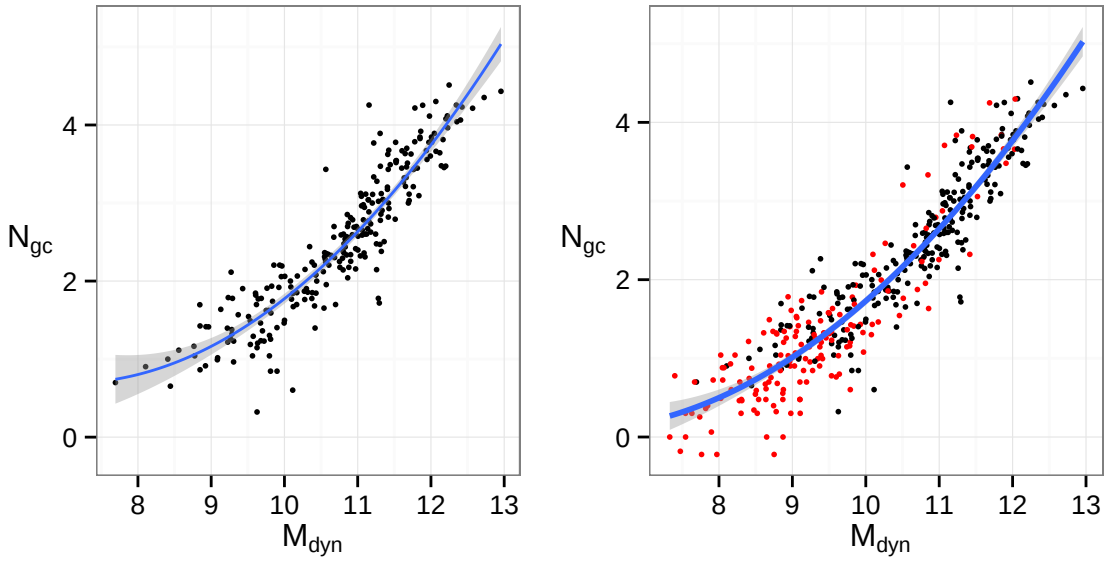


Figure 6.1: Possible quadratic relationships.

A brief attempt to incorporate a quadratic fit ran into technical troubles due to the inclusion of the uncertainty terms, though this was not thoroughly explored and could possibly lead to a viable model relating N_{GC} with M_{dyn}.

6.2.3 Simulations

As stated in the conclusion, the amount of missing data plays a key role in the success of the imputation procedure. It would be of future interest to perform simulations on data with varying amounts of missing values. This could help understand the effect of

the quantity of missing information on the precision of estimates. Simulations could also include data missing in different locations within the response and predictors. Perhaps missing values located within a single covariate have a different effect on the response than the same amount of missing data spread across multiple covariates.

Bibliography

Andridge, R. R. and Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International statistical review*, **78**(1), 40–64.

Brand, J. P. L. (1999). Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. CRC Press, second edition.

Harris, W. E., Harris, G. L. H., and Alessi, M. (2013). A Catalog of Globular Cluster Systems: What Determines the Size of a Galaxy's Globular Cluster Population? *The Astrophysical Journal*, **772**(2), 82.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple Imputation for the Fatal Accident Reporting System. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **40**(1), 13–29.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55–67.

Huisman, M. (2009). Imputation of missing network data: some simple procedures. *Journal of Social Structure*.

Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, **6**(3), 287–296.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley-Interscience, Hoboken, N.J, second edition.

Lyons, L. (1991). *A practical guide to data analysis for physical science students*. Cambridge University Press, Cambridge.

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, **14**(1), 1–13.

Novak, G. S., Faber, S. M., and Dekel, A. (2006). On the Correlations of Massive Black Holes with Their Host Galaxies. *The Astrophysical Journal*, **637**(1), 96.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK ; New York.

Schenker, N. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, **22**(4), 425–446.

Tremaine, S., Gebhardt, K., Bender, R., Bower, G., Dressler, A., Faber, S. M., Filippenko, A. V., Green, R., Grillmair, C., Ho, L. C., Kormendy, J., Lauer, T. R., Magorrian, J., Pinkney, J., and Richstone, D. (2002). The Slope of the Black Hole Mass versus Velocity Dispersion Correlation. *The Astrophysical Journal*, **574**, 740–753.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**(3), 1–67.