Erik van Werkhoven

Imputation methods for non-response in economical survey data

Master thesis, defended on July 11, 2008

Thesis advisers:
dr. Erik van Zwet (Universiteit Leiden)
dr. Caren Tempelman (Centraal Bureau voor de Statistiek)
dr. Jeroen Pannekoek (Centraal Bureau voor de Statistiek)

Mathematisch Instituut, Universiteit Leiden

Contents

Summary

Economical data collected by Statistics Netherlands usually contains missing items. Various imputation methods are available to fill in these gaps, so that completed datasets can be analyzed using standard statistical tools. One of the methods often used, the ratio imputation method, appears not to perform very well if we want the completed data to satisfy certain restrictions.

This is our motivation to investigate other imputation methods. We look at several methods that we subdivide over two groups. The first group consists of methods based on models that assume a joint distribution for all variables for an individual, and that these variables are all independent. Here we will discuss methods that assumes the data are truncated normally distributed, or exponentially distributed. We propose the proportional variance method, and investigate various possible underlying models.

The second group is made up of methods that only specify certain conditional distributions. Here we will investigate the commonly used ratio imputation method and both the classical and the Bayesian variants of sequential regression imputation methods.

After we have discussed these methods, we repeatedly apply them to a dataset provided by Statistics Netherlands in which we make a missing pattern ourselves. We use the results of these simulations to assess the performance of the methods on several criteria.

Chapter 1

Introduction

Like all other national statistical agencies, Statistics Netherlands collects data about society. These data come from different sources like databases, but also from surveys and questionnaires. One of the most complete datasets of a population is a census, in which all individuals or *units* are contacted, and all their characteristics are fully observed. Recently, agencies have started to use external databases, for instance from local councils or the exchequer.

In practice, almost all datasets will contain missing items. A common way to handle missing data is using *imputation*. Imputation essentially means filling in missing data with estimated values. A very simple form of imputation is to fill in the the average, which preserves the observed sample mean. Yet, this diminishes the variance, misshapes the sample distribution and annihilates all covariances, so examining other methods is worthwhile.

This thesis focuses on imputation methods applied to business surveys. Although in The Netherlands businesses are under legal obligation to fill in these surveys, there usually are enterprises that do not respond to all questions that are being asked. There can be various reasons for this. The answer to a particular question may not be known to the the employee filling in the survey, or it can be a lot of work to find the answer if the information requested by the agency is not exactly the same as the information in the business's accounting system.

Business surveys contain micro-economical data. For us, this means that we will be considering data that satisfy certain linear restrictions. For consistency, imputed values should therefore also satisfy these conditions. A typical scheme is to first apply an imputation method that does not account for the restrictions. After that, the imputed values are adjusted so that the restrictions are are satisfied, but still near their formerly values. More details on this optimization problem can be found in [De Waal 2003]. It seems to make sense however to examine methods that use such conditions in the imputation process itself. In this thesis, we will investigate methods that create imputations which satisfy these conditions directly.

1.1 Notation of survey data

Suppose a survey is carried out, and the results are put in a table, with each column representing a question or variable and each row a record with responses from an individual. In the economical data we will be investigating, the individuals are businesses. In some literature, individuals are called elements or units. Let X be the $n \times k$ matrix with as entries X_{ij} the numerical values individual i answered to question j. We call X a *data matrix*.

In chapter 3, where we discuss regression methods, we will adopt the notation from regression theory. Then y_i will denote the item we wish to impute and $x_i = (x_1, \ldots, x_k)$ will denote the explanatory information for unit i. For n observed units, we get a vector y of length n and a $n \times k$ design matrix which we also denote by X. It will be always be clear from the context if we use X for the data matrix or for a design matrix.

Finally, we define the $n \times k$ *indicator matrix* M where the entry $m_{ij} = 1$ if individual i replies to question j and $m_{ij} = 0$ if he does not. We denote the set of indices for which X_{ij} is missing by mis, and the index set of the observed items by obs. It should be clear from the context whether we mean the column or row indices.

1.2 Errors in survey data

In general, survey data are collected to say something about a population. That is, we wish to estimate a population parameter. There are many different kinds of possible errors in such estimations. *Sampling errors* are errors that occur because not all units or individuals of a population are observed in a survey, but only a sample of them. Therefore the information extracted from the observed units may differ from what we could have known if we would have observed the whole population. There is not much we can do about sampling errors, apart from choosing a suitable sampling design that does not lead to systematical errors, or take a census of a population instead of a sample from it.

In contrast, *non-sampling errors* are errors in sample estimates that cannot be attributed to sampling fluctuations. This type of error can still occur even if all units are in the survey. *Frame errors* occur if the register or database from which the sample is drawn , i.e. the *frame*, does not correspond exactly to the population that was the target of the investigation. If individuals are in the frame but not in the population, we speak of *over-coverage*. The other way round, if there are individuals in the population that are not in the frame we have *under-coverage*. The question of what should be considered as a unit is more difficult in a business survey than for instance in a survey in which the population consists of natural persons.

Another kind of a non-sampling error is *non-response*, i.e. data that are not observed for individuals that were selected in a sample. Generally, we distinguish two different types of non-response. The information missing as a result of an individual's complete failure to respond, is called *unit non-response*. If an

individual has answered at least one question, the missing items are called *item non-response*. In this thesis we will only consider item non-response, in order to be able to use the information of available items to say something about the missing information.

A third type are *measurement errors*. These errors occur if a value is observed, but it is incorrect. Some mistakes can be obvious, for instance if a respondent fills in too many zeros, but others are difficult to find. Finally, *processing errors* are due to the process at the statistical agency itself, e.g. typos or adjusted values that were actually correct.

Detecting errors in surveys is a discipline in itself, and is called *editing*, see e.g. [Chambers 2001]. Once errors have been localized, it is common practice to set them to missing so that an imputation procedure can be used to correct these values. We will not discuss error detection, but only consider non-response.

1.3 Assumptions on the missing data mechanism

In the ideal situation, we would know the mechanism that generates non-response in our surveys and would use that information in our models. In practice however, the mechanism is hardly ever known, so we want to have assumptions under which we do not need that information.

Suppose the distribution of the completely observed data depends on a parameter θ. Write the probability density of an element M_{ij} of the indicator matrix of X as $\pi(m_{ij}|x_{i,mis}, x_{i,obs}, \phi)$ with ϕ as a parameter that does not affect the data directly, but only via the missing data mechanism. We assume that the parameters θ of the distribution of the data and ϕ of the response mechanism are *distinct*. This means the joint parameter space for (θ, ϕ) is the Cartesian product of the individual parameter spaces for θ and ϕ. From a Bayesian point of view, it means the joint prior distribution of (θ, ϕ) can be written as a product of independent marginal prior distributions for θ and ϕ.

Using this notation, Rubin developed three possible assumptions for missing data:

- The data are called *missing at random* (MAR) if the conditional distribution of M_{ij} depends on the observed, but not on the unobserved part of the data. Thus $\pi(m_{ij}|x_{i,mis}, x_{i,obs}, \phi)$ is constant in $x_{i,mis}$ and we may write $\pi(m_{ij}|x_{i,obs}, \phi)$.
- The data are said to be *missing completely at random* (MCAR) if M_{ij} and X_{ij} are mutually independent for all values of i and we may write $\pi(m_{ij}|\phi)$ for $\pi(m_{ij}|x_{i,mis}, x_{i,obs}, \phi)$.
- If M_{ij} depends on both the observed and the missing part of the data, then Robin calls the data *not missing at random* (NMAR). In that case the probability of response depends on the actual value of the response.

It is clear that for NMAR data we cannot ignore the missing mechanism, but need to model it in order to be able to say anything about the distribution of the data. If the combination of MAR with distinct parameters for the data and the

non-response model holds, the missing data mechanism is called *ignorable*. In that case the likelihood of the data factors in a part that depends solely on θ and a part that depends solely on ϕ. So both in maximum likelihood estimation and Bayesian analysis, we do not need to know anything about the non-response mechanism as long as it is ignorable. In this thesis, we will always assume ignorability.

Because of the advantages of ignorability, it is tempting to assume MAR. The risk is that it is adopted without proper reasons. It is well known that in social surveys respondents which are relatively far from the average are more likely not to answer questions about social sensitive variables than others. In business surveys we may expect that these effects play a lesser role. Especially in larger companies, the person who fills in the questionnaire will in general not perceive the information sensitive.

Apart from these considerations that make it seem reasonable to believe we can assume MAR, we would be more confident if we could somehow test the data for MAR. Unfortunately, Gill, Van der Laan and Robins [Gill 1996] show that such a test does not exist. To be more precise, they prove that for a discrete random variable taking values in a finite sample space under a non-parametric model, for any observed outcome there exists a MAR missing data mechanism and a complete data distribution such that the distribution of the observed data is the marginal of the joint distribution of these two.

In this thesis we will consider continuous data with missing items, but in reality, businesses can only fill in integer values for most variables in questionnaires sent out by Statistics Netherlands. The reason we consider them to be continuous, is that we may apply more models.

1.4 Dealing with non-response

The easiest way to handle non-response is to do nothing and simply discard all records that are not completely observed. However, using this *complete case analysis* a lot of information is lost. Another option is to consider only records for which a particular variable of interest is observed, a strategy known as *available case analysis*. This would do slightly better than the complete case analysis, since more information can be used.

Better alternatives are the EM algorithm, which can find maximum likelihood estimates in cases where data are missing, and weighting, a popular method to correct for unit non-response. This method assigns a weight to each unit, compensating for supposedly similar units that did not respond. The similarity is expressed in an auxiliary variable. For valid estimates the data should satisfy the MAR assumption with respect to that auxiliary variable.

Imputation methods, on which this thesis focuses, fill in missing parts of data with estimated values. This is a very common procedure in survey data. There are three major reasons for that. Firstly, the procedure has to be carried out only once, and the data collector can use his knowledge of the data. Secondly, once the dataset is imputed, analysts can use their standard complete data method

to study the data. Thirdly, imputation methods can, in contrast to weighting, use all information available from the partly observed record. Although it might not be necessary for all statistical purposes, we want to impute missing values consistent with the data available, because end-users appreciate it.

1.5 Classification of imputation methods

Over time, a lot of different imputation methods have been developed. An imputation method is called *deterministic* if, for a fixed dataset with missing entries, it will always fill in the same values. *Stochastic* methods, sometimes called also called random methods, use a random draw from a distribution.

The most commonly used imputation methods are

- *Deductive or logical imputation*
  If the value of a missing item can be derived immediately from the observed values in the dataset, the best thing to do is to fill in that value. In this case the imputation is called deductive or logical.
- *Cold deck imputation*
  Sometimes missing items can be replaced by a value found in another source, or in the same data for a previous period. This is called cold deck imputation.
- *Hot deck imputation*
  Hot deck imputation also takes another value, but takes it from within the dataset itself. *Sequential hot deck* just takes a value of the last unit that did respond, so this depends on the way the dataset is sorted. *Nearest neighbour imputation* uses a distance function to find a record from which to copy its values. If the distance function uses a regression analysis, the method is called *predictive mean matching*. There are also stochastic variants of the nearest neighbour imputation method, that do not not take the nearest record but choose randomly from a few nearby records.
- *Mean imputation*
  The mean imputation method fills missing entries with the sample mean of the respondents' values Often the mean of a subset of the respondents is used, to impute the sample mean of a certain class. Mean imputation is a special case of regression imputation, but with no auxiliary information.
- *Regression imputation*
  Regression models use covariates to explain the behaviour of a variable with missing items. The maximum likelihood estimator is then used to infer the missing value, see section 3.1 and further for examples of regression models. This method is sometimes called *predictive regression imputation*, *deterministic regression imputation* or *conditional mean imputation*, but these names all refer to the same method. The special case of regression imputation where the variance of the missing variable is modeled to be proportional to a single explanatory variable is called *ratio imputation*,

see section 3.4. The regression imputation method can be made stochastic by adding an error term to the predictor value found by the regression model. An other way to make it stochastic is to draw the parameters from a posterior distribution instead of calculating their maximum likelihood estimates.

- *imputation with the EM algorithm* The Expectation Maximization (EM) algorithm is a method for finding maximum likelihood estimates for statistics in datasets with missing items, see section 2.2. If you use the method to estimate the parameters of a distribution the data are supposed to follow, you can use it to impute the missing values.
- *sequential regression* The sequential regression method repeatedly applies a regression model for each variable conditioned on all other variables. The process is repeated until it converges. There is however in general no theory guaranteeing convergence. See section 3.3.

Chapter 2

Imputation models assuming independence

In this chapter we will investigate models assuming all items X_{ij} are independently distributed. First we will discuss some well known theory about normally distributed data with missing parts and truncation. Then we describe a model that assumes the data are exponentially distributed. In section 2.5 we suggest a new method, the proportional variances method. We will investigate two models that give rise to this method: one that assumes normality, and another that assumes a Poisson distribution.

2.1 Multivariate normal data with a missing and an observed part

Suppose X is a multivariate normally distributed column vector of length k. We partition X into two parts $X = (X_1^T, X_2^T)^T$. Keeping in mind our missing data problem, X_1 could represent the unobserved part and X_2 the observed part of the variable X. The following theorem, from [Anderson 1971], gives us the conditional distribution of X_1 given X_2.

Theorem 2.1.1. *Let X be a multivariate normally distributed random variable with mean vector μ and covariance matrix Σ, shorthand $X \sim \mathcal{N}_k(\mu, \Sigma)$. Suppose we partition X into X_1 of length m and X_2 of length $k-m$ and partition μ and Σ accordingly, i.e.*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_k \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

then

$$X_1 | X_2 \sim \mathcal{N}_m(\mu_{1.2}, \Sigma_{11.2})$$

where

$$\mu_{1.2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \tag{2.1}$$
$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{2.2}$$

Proof. We would like to find a transformation $Y_1 = X_1 + BX_2$ with B a matrix such that Y_1 and X_2 are uncorrelated. Note that

$$\begin{aligned}
\text{Cov}(Y_1, X_2) &= E[(Y_1 - EY_1)(X_2 - EX_2)^T] \\
&= E[(X_1 + BX_2 - EX_1 - BEX_2)(X_2 - EX_2)^T] \\
&= E[((X_1 - EX_1) + B(X_2 - EX_2))(X_2 - EX_2)^T] \\
&= E[(X_1 - EX_1)(X_2 - EX_2)^T] + BE[(X_2 - EX_2)(X_2 - EX_2)^T \\
&= \Sigma_{12} + B\Sigma_{22}
\end{aligned}$$

so for this to be zero, $B = -\Sigma_{12}\Sigma_{22}^{-1}$ will do. Applying this transformation leads to

$$Y = \begin{pmatrix} Y_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} X$$

which is multivariate normally distributed with mean

$$E(Y) = E\begin{pmatrix} Y_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{pmatrix}$$

and covariance matrix

$$\text{Cov}(Y) = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.$$

We see that it helps to define

$$\tilde{Y} = X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \tag{2.3}$$

because then, according to our previous calculations

$$\begin{pmatrix} \tilde{Y} \\ X_2 \end{pmatrix} \sim \mathcal{N}_k\left(\begin{pmatrix} 0 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}\right)$$

which means $\tilde{Y}$ and X_2 are independent. We calculate the characteristic function of X_1 given X_2:

$$\begin{aligned}
\phi(u)_{X_1|X_2} = E[e^{iu^T X_1}|X_2] &= E\left[e^{iu^T\tilde{Y} + iu^T(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2))}|X_2\right] \\
&= e^{iu^T(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2))}E\left[e^{iu^T\tilde{Y}}|X_2\right] \\
&= e^{iu^T(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2))}E\left[e^{iu^T\tilde{Y}}\right] \\
&= e^{iu^T(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2))}e^{-\frac{1}{2}u^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})u} \\
&= e^{iu^T(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)) - \frac{1}{2}u^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})u} \\
&= e^{iu^T\mu_{1.2} - \frac{1}{2}u^T\Sigma_{11.2}u}
\end{aligned}$$

and see that this is the characteristic function of a $\mathcal{N}_m(\mu_{1.2}, \Sigma_{11.2})$ distribution. □

If we can estimate the parameters $\mu_{1.2}$ and Σ_{12}, theorem 2.1.1 gives us an estimation of the distribution of the unobserved part of a normally distributed vector, given the observations. We can use a sample from this distribution to impute the missing data. We can also fill in the expectation to get a deterministic imputation method.

2.2 EM algorithm

In order to use the formula found in the previous section, we still need to estimate μ and Σ. The usual estimators are not applicable, because of the missing data. Since it is too complicated to calculate the likelihood of the observed data, we need to do something else. The Expectation-maximization (EM) algorithm is suitable for this situation.

The main idea of the EM algorithm is very intuitive. It is to fill in missing values with some estimates, from which we reestimate the parameters, from which we can reestimate the missing values again, and repeat this iteration until the parameter estimates converge. If the underlying distribution is part of an exponential family, we estimate the expectation of the observations themselves, and otherwise we estimate the log-likelihood.

Suppose more generally that we want to estimate a parameter θ of the distribution of a data matrix X of which some part is observed, X_{obs} and some part X_{mis} is not. If in an iteration of the algorithm $\theta^{(t)}$ is the current estimator for parameter θ, the E step calculates the expected log-likelihood as if the $\theta^{(t)}$ would equal the true parameter θ and as described in [Schafer 1997] finds

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}[\ell(X,\theta)|X_{obs}].$$

The M step of the EM algorithm maximizes the expected log-likelihood over θ and gives

$$\theta^{(t+1)} = \text{argmax}_\theta Q(\theta|\theta^{(t)}).$$

In a more general version of the EM algorithm, the M step does not have to find the maximum, but only a better value , i.e. $\theta^{(t+1)}$ such that

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t+1)}|\theta^{(t+1)}).$$

This type of algorithms are called Generalized Expectation-maximization algorithms. We now have to show that with every update of $\theta^{(t)}$, we get a better value of the log-likelihood. The next theorem is from [Schafer 1997].

Theorem 2.2.1. *If we choose a $\theta^{(t+1)}$ such that for all θ*

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) \tag{2.4}$$

then

$$\ell(X_{obs},\theta^{(t+1)}) \geq \ell(X_{obs},\theta^{(t)}). \tag{2.5}$$

Proof. The distribution of the complete data can always be written as

$$P(X|\theta) = P(X_{obs}|\theta)P(X_{mis}|X_{obs},\theta)$$

and by taking the log we see that

$$\ell(X, \theta) = \ell(X_{obs}, \theta) + \log P(X_{mis}|X_{obs}, \theta)$$

so by rearranging the terms we get

$$\ell(X_{obs}, \theta) = \ell(X, \theta) - \log P(X_{mis}|X_{obs}, \theta).$$

Define

$$H(\theta|\theta^{(t)}) = E_{\theta^{(t)}}[\log P(X_{mis}|X_{obs}, \theta)|X_{obs}]$$

and note that $\ell(X_{obs}, \theta)$ does not depend on X_{mis}. Hence,

$$E_{\theta^{(t)}}[\ell(X_{obs}, \theta)|X_{obs}] = \ell(X_{obs}, \theta)$$

and if we take expectations under $\theta^{(t)}$ conditioned on X_{obs} we get

$$\begin{aligned}
\ell(X_{obs}, \theta) &= E_{\theta^{(t)}}[\ell(X, \theta) - \log P(X_{mis}|X_{obs}, \theta)|X_{obs}] \\
&= E_{\theta^{(t)}}[\ell(X, \theta)|X_{obs}] - E_{\theta^{(t)}}[\log P(X_{mis}|X_{obs}, \theta)|X_{obs}] \\
&= Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}).
\end{aligned}$$

Now write

$$\begin{aligned}
\ell(X_{obs}, \theta^{(t+1)}) - \ell(X_{obs}, \theta^{(t)}) = Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) \\
+ H(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}).
\end{aligned}$$

We see that, since we chose $\theta^{(t+1)}$ such that 2.4 holds, it suffices to show that

$$H(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) \geq 0$$

Looking at the opposite we see

$$\begin{aligned}
H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) &= E_{\theta^{(t)}}[\log P(X_{mis}|X_{obs}, \theta)|X_{obs}] \\
&\quad - E_{\theta^{(t)}}[\log P(X_{mis}|X_{obs}, \theta^{(t)})|X_{obs}] \\
&= E_{\theta^{(t)}}[\log \frac{P(X_{mis}|X_{obs}, \theta^{(t+1)})}{P(X_{mis}|X_{obs}, \theta^{(t)})}|X_{obs}] \\
&\leq \log E_{\theta^{(t)}}[\frac{P(X_{mis}|X_{obs}, \theta)}{P(X_{mis}|X_{obs}, \theta^{(t)})}|X_{obs}] \\
&= \log \int [\frac{P(X_{mis}|X_{obs}, \theta^{(t+1)})}{P(X_{mis}|X_{obs}, \theta^{(t)})}] P(X_{mis}|X_{obs}, \theta^{(t)}) dy_{mis} \\
&= \log \int P(X_{mis}|X_{obs}, \theta^{(t+1)}) dy_{mis} \\
&= \log 1 = 0
\end{aligned}$$

which is the the desired result. The inequality (*) uses the fact that the logarithm is a concave function, so Jensen's inequality can be applied. □

Since $\{Q^{(t)}\}_t \in N$ is a positive increasing sequence, the algorithm converges. If the distribution of X with parameter θ is an exponential family, that is, if it can be written in the form

$$f(x) = a(\theta)b(x)\exp\big(c(\theta)t(x)\big),$$

then we only need to find the expected values of the sufficient statistic $t(x)$ of the complete data, instead of the conditional expectation of the likelihood itself. See [Tempelman 2007], p. 42 for more details.

2.3 Truncated normal data

In the previous section we derived the conditional distribution of a missing part given an observed part of a normally distributed random vector. But we want to do more. We would like to get inferences that satisfy restrictions on the data. The way to do that, is to restrict the distribution to a permitted region. In econometrics these conditional distributions are called truncated distributions.

Let X be a multivariate normally distributed k-vector with expectation μ and covariance matrix Σ, shorthand $X \sim \mathcal{N}_k(\mu, \Sigma)$. Then $\tilde{X}$ is called the *truncation* of X to permitted region $G \subset \mathbb{R}^k$ if it is multivariate normally distributed with parameters μ and Σ conditionally on the event $X \in G$. We say $\tilde{X}$ is truncated multivariate normally distributed and we write $\tilde{X} \sim \mathcal{N}_k(\mu, \Sigma)|G$.

The mean of the truncated normal distribution If the mean of the original non-truncated variable μ is not in the middle of the permitted region, the truncation shifts the mean. We can express the mean of the truncated version in terms of the original μ and σ. To simplify notation, we look at the univariate case, where a variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is truncated to the interval $(a, b) \subset \mathbb{R}$.

$$\begin{aligned}
E(\tilde{X}) = E(X|X \in G) &= \frac{\int_a^b \sigma^{-1}\phi\big(\frac{x-\mu}{\sigma}\big)x dx}{\Phi(z_b) - \Phi(z_a)} \\
&= \frac{\int_{z_a}^{z_b} \phi(z)(\mu + \sigma z)dz}{\Phi(z_b) - \Phi(z_a)} = \mu\frac{\int_{z_a}^{z_b} \phi(z)dz}{\Phi(z_b) - \Phi(z_a)} + \sigma\frac{\int_{z_a}^{z_b} z\phi(z)dz}{\Phi(z_b) - \Phi(z_a)} \\
&= \mu + \sigma\frac{\frac{1}{\sqrt{2\pi}}\int_{z_a}^{z_b} z\exp(-\frac{1}{2}z^2)dz}{\Phi(z_b) - \Phi(z_a)} = \mu + \sigma\frac{\frac{1}{\sqrt{2\pi}}\big[-\exp(-\frac{1}{2}z^2)\big]_{z_b}^{z_a}}{\Phi(z_b) - \Phi(z_a)} \\
&= \mu + \sigma\frac{-\big(\phi(z_b) - \phi(z_a)\big)}{\Phi(z_b) - \Phi(z_a)} = \mu - \sigma\frac{\phi(z_b) - \phi(z_a)}{\Phi(z_b) - \Phi(z_a)}
\end{aligned}$$

where $z_a = (a - \mu)/\sigma$ and $z_b = (b - \mu)/\sigma$.

The next theorem, from [Tempelman 2007] links truncation with linear transformations. As you would expect, it does not matter whether you first truncate and then apply such a transformation, or the other way round, first transform and then truncate with respect to the transformed permitted region.

Theorem 2.3.1. *Let $G \subset \mathbb{R}^k$, $\tilde{X} \sim \mathcal{N}_k(\mu, \Sigma)|G$ and D a nonsingular $k \times k$ matrix. Then*

$$D\tilde{X} \sim \mathcal{N}_k(D\mu, D\Sigma D^T)|T,$$

with $T = DG = \{D\tilde{X}, \tilde{X} \in G\}$.

Proof. Take L from the definition of the multivariate normal distribution such that $\Sigma = LL^T$. We know that for the non-truncated version X of $\tilde{X}$ we can write $X = \mu + LZ$ with Z a vector with independent $\mathcal{N}(0, 1)$ distributed random variables. So for any $b \in \mathbb{R}^k$

$$P(\mu + LZ \leq b) = \int_{z:\mu+Lz\leq b} \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} dz.$$

Note that $Z = L^{-1}(X - \mu)$, so since $\tilde{X}$ takes values in G, Z should be truncated to the set $\tilde{G} = L^{-1}(G - \mu)$. Therefore,

$$\begin{aligned} P(D\tilde{X} \leq b) &= \frac{1_{\{Z \in \tilde{G}\}}}{C} P(D\mu + DLZ \leq b) \\ &= \frac{1_{\{Z \in \tilde{G}\}}}{C} \int_{z: D\mu + DLz \leq b} \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} dz \end{aligned}$$

where C is the normalization constant from the definition of truncation.

Now, apply the change of variables $\xi = D\mu + DLz$. This is a linear transformation with Jacobian $\partial z / \partial x = L^{-1} D^{-1}$ which has determinant $\det(D\Sigma D^T)^{-\frac{1}{2}}$.

$$\begin{aligned} \det(L^{-1}D^{-1}) &= \det(L^{-1})\det(D^{-1}) = (\det\Sigma)^{-\frac{1}{2}} \det(D^{-1}) \\ &= (\det\Sigma)^{-\frac{1}{2}} \det(D^{-\frac{1}{2}} D^{-\frac{1}{2}}) = (\det\Sigma)^{-\frac{1}{2}} \det(D^{-\frac{1}{2}}) \det(D^{-\frac{1}{2}}) \\ &= (\det\Sigma)^{-\frac{1}{2}} (\det D)^{-\frac{1}{2}} (\det D^T)^{-\frac{1}{2}} \\ &= \det(D\Sigma D^T)^{-\frac{1}{2}}. \end{aligned}$$

Also note that

$$\begin{aligned} \sum z_i^2 = z^T z &= ((DL)^{-1}(\xi - \mu))^T ((\xi - \mu)(DL)^{-1}) \\ &= (\xi - D\mu)^T (L^{-1}D^{-1})^T (\xi - D\mu)(L^{-1}D^{-1}) \\ &= (\xi - D\mu)^T (D^T)^{-1} (L^{-1})^T L^{-1} D^{-1} (\xi - D\mu) \\ &= (\xi - D\mu)^T (D^T)^{-1} (LL^T)^{-1} D^{-1} (\xi - D\mu) \\ &= (\xi - D\mu)^T (D^T)^{-1} (\Sigma)^{-1} D^{-1} (\xi - D\mu) \\ &= (\xi - D\mu)^T (D\Sigma D^T)^{-1} (\xi - D\mu). \end{aligned}$$

Finally, note that if $Z \in \tilde{G}$, this means $\xi \in D\mu + DL\tilde{G} = D\mu + DLL^{-1}(G - \mu) = DG = T$. Therefore also $P(X \in G) = P(DX \in T)$ and the normalization constant C remains the same. Hence,

$$P(D\tilde{X} \leq b) = \frac{1_{\{\xi \in T\}}}{C} \int_{\xi: \xi \leq b} \frac{1}{(2\pi)^{k/2} \sqrt{\det(D\Sigma D^T)}} e^{-\frac{1}{2}(\xi - D\mu)^T (D\Sigma D^T)^{-1} (\xi - D\mu)} d\xi.$$

□

Example: MLE for the left truncated normal distribution

In this example, we suppose $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. random variables. Now let be given that $X_i \geq 0$ for all $i = 1, \ldots, n$. What is the maximum likelihood estimator (MLE) for μ and σ?

Each x_i has density function

$$f(x) = \frac{\sigma^{-1} \phi\big((x - \mu)/\sigma\big)}{1 - \Phi(-\mu/\sigma)} \tag{2.6}$$

and the variables $X_1, X_2, \ldots, X_n$ are all independent, so their joint density is

$$f(x_1, \ldots, x_n) = \frac{\sigma^{-n} \prod_{i=1}^n \phi\big((x_1 - \mu)/\sigma\big)}{\big(1 - \Phi(-\mu/\sigma)\big)^n}$$

which gives log-likelihood

$$\begin{aligned}
\ell(\mu, \sigma) &= -n \log \sigma + \sum_{i=1}^n \log \phi\big((x_n - \mu)/\sigma\big) - n \log\big(1 - \Phi(-\mu/\sigma)\big) \\
&= -n \log \sigma + \sum_{i=1}^n \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \big(\frac{x_i - \mu}{\sigma}\big)^2 \right) - n \log\big(1 - \Phi(-\mu/\sigma)\big) \\
&= -n \log \sigma - \frac{n}{2} \log(2\pi) - \sum_{i=1}^n \frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 - n \log\big(1 - \Phi(-\mu/\sigma)\big).
\end{aligned}$$

The derivatives with respect to μ and σ are

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^n \big(\frac{x_i - \mu}{\sigma^2}\big) - \frac{n}{\sigma} \frac{\phi(-\mu/\sigma)}{(1 - \Phi(-\mu/\sigma))} \tag{2.7}$$

$$\frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^3} + \frac{n}{\sigma^2} \frac{\mu \phi(-\mu/\sigma)}{1 - \Phi(-\mu/\sigma)}. \tag{2.8}$$

We can use numerical methods to set the equations (2.7) and (2.8) to zero and find the maximum likelihood estimators for μ and σ.

2.3.1 Truncated normal with a missing and an observed part

Partition X in an observed part X_1 and a missing part X_2, like we did in section 2.1. Assume $X \sim \mathcal{N}_k(\mu, \Sigma)|G$ and, and apply theorem 2.3.1 on X with

$$D = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix}$$

chosen as in the proof of theorem 2.1.1 so that

$$D(X - \mu) = \begin{pmatrix} X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ X_2. \end{pmatrix}$$

Theorem 2.3.1 now tells us $D(X - \mu)$ is also truncated multivariate normal, with a vector of zeros as mean and covariance matrix

$$D\Sigma D^T = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \tag{2.9}$$

so just like in the proof of theorem 2.1.1, we see the two parts X_1 and X_2 are independent. This means $X_1 - \mu_1|X_2$ is truncated multivariate normal too, but with mean vector $\Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and we conclude

$$X_1|X_2 \sim \mathcal{N}(\mu_{1.2}, \Sigma_{11.2})|T. \tag{2.10}$$

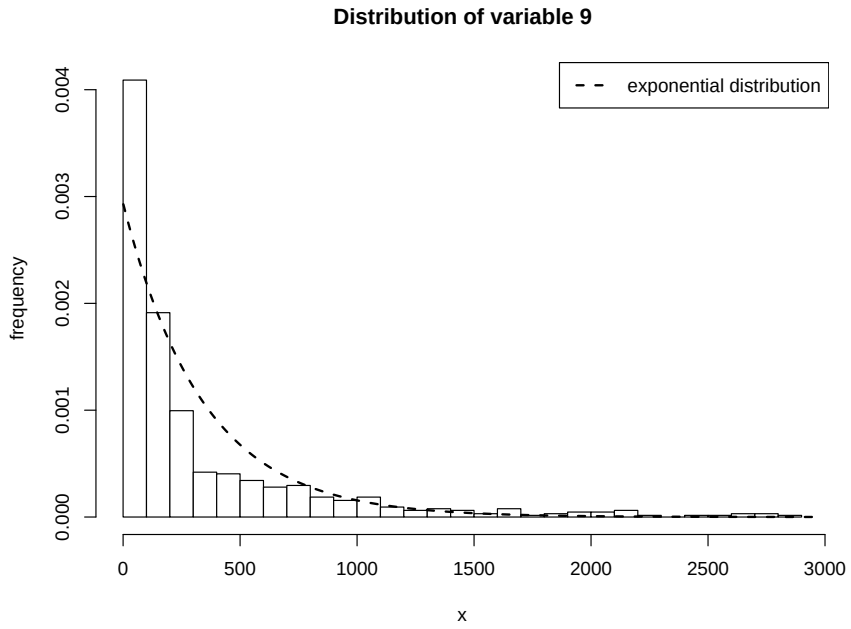


Figure 2.1: A histogram of variable `bt34600t`. The dashed line shows an exponential distribution fitted to the data.

where T is the region dependent on X_2 in which X_1 must take values to make sure X takes values in G.

The fact that the conditional distribution of a partition of a truncated normal is again truncated normal is a nice property. But you have to take into account that the permitted region may also need to be transformed. Horrace shows that if you do not do that, the only transformation that leads to a truncated normal distribution again is the identity [Horrace 2005].

In our missing data problem, we can use truncated normal distributions to model the missing items. Due to the structure of the datasets we will be looking at, we will always be able determine a region in which the missing values must lie in.

2.4 Exponential distributions

In the previous sections, we considered normally distributed data. Let us now have a look at the distribution of a single variable. Figure 2.1 shows a typical histogram of a column of the dataset we will investigate. It seems appropriate to model the distribution of a single column with an exponential distribution. Since in our dataset we know the some of a record, our aim is to find the conditional expectation of a single exponentially distributed random variable, given a sum of exponentials in which it is part.

In the following lemma, from [Balazs 2005], we calculate the distribution of the sum of independently distributed exponential random variables.

Lemma 2.4.1. *Let X_i for $i = 1, \dots n$ with $n \geq 2$ be independent exponentially distributed random variables with pairwise distinct parameters $\lambda_i > 0$. Then the density of the sum $X_1 + \ldots + X_n$ is*

$$f_{X_1+\ldots+X_n}(z) = \left(\prod_{i=1}^{n} \lambda_i\right) \sum_{j=1}^{n} \frac{e^{-\lambda_j z}}{\prod_{k=1,k\neq j}^{n} (\lambda_k - \lambda_j)} \tag{2.11}$$

for $z \geq 0$.

Proof. We prove the lemma using induction with respect to n. For $n = 2$ we see

$$\begin{aligned}
f_{X_1+X_2}(z) &= \big(f_{X_1} * f_{X_2}\big)(z) = \int_0^z f_{X_1}(x) f_{X_2}(z-x)dx \\
&= \int_0^z \lambda_1 e^{-\lambda_1 x} \lambda_2 e^{-\lambda_2 (z-x)} dx \\
&= \lambda_1\lambda_2 e^{-\lambda_2 z} \int_0^z e^{(\lambda_2-\lambda_1)x} dx = \lambda_1\lambda_2 e^{-\lambda_2 z} \left[\frac{1}{(\lambda_2-\lambda_1)} e^{(\lambda_2-\lambda_1)x}\right]_{x=0}^{x=z} \\
&= \lambda_1\lambda_2 e^{-\lambda_2 z} \left(\frac{e^{-\lambda_2 z} e^{-\lambda_1 z}}{(\lambda_2-\lambda_1)} - \frac{1}{(\lambda_2-\lambda_1)}\right) \\
&= \lambda_1\lambda_2 \left(\frac{e^{-\lambda_1 z}}{(\lambda_2-\lambda_1)} - \frac{e^{-\lambda_2 z}}{(\lambda_2-\lambda_1)}\right) \\
&= \lambda_1\lambda_2 \left(\frac{e^{-\lambda_1 z}}{(\lambda_2-\lambda_1)} + \frac{e^{-\lambda_2 z}}{(\lambda_1-\lambda_2)}\right)
\end{aligned}$$

and therefore (2.11) holds. Now let $n \geq 3$ and suppose that statement (2.11) is

true for $n-1$. Then

$$\begin{aligned}
f_{X_1+X_2+\ldots X_n}(z) &= \big(f_{X_1+X_2+\ldots X_{n-1}} * f_{X_n}\big)(z) \\
&= \left(\left[\prod_{i=1}^{n-1}\lambda_i\right]\sum_{j=1}^{n-1}\frac{e^{-\lambda_j x}}{\prod_{k\neq j,k=1}^{n-1}(\lambda_k-\lambda_j)} * \lambda_n e^{-\lambda_n x}\right)(z) \\
&= \int_0^z \left(\prod_{i=1}^{n-1}\lambda_i\right)\sum_{j=1}^{n-1}\frac{e^{-\lambda_j x}\lambda_n e^{-\lambda_n x}}{\prod_{k\neq j,k=1}^{n-1}(\lambda_k-\lambda_j)}dx \\
&= \left(\prod_{i=1}^{n}\lambda_i\right)\sum_{j=1}^{n-1}\frac{1}{\prod_{k\neq j,k=1}^{n-1}(\lambda_k-\lambda_j)}\int_0^z e^{-\lambda_j x}e^{-\lambda_n(z-x)}dx \\
&= \left(\prod_{i=1}^{n}\lambda_i\right)\sum_{j=1}^{n-1}\frac{1}{\prod_{k\neq j,k=1}^{n-1}(\lambda_k-\lambda_j)}e^{-\lambda_n z}\int_0^z e^{(\lambda_n-\lambda_j)x}dx \\
&= \left(\prod_{i=1}^{n}\lambda_i\right)\sum_{j=1}^{n-1}\frac{1}{\prod_{k\neq j,k=1}^{n-1}(\lambda_k-\lambda_j)}e^{-\lambda_n z}\left[\frac{1}{(\lambda_n-\lambda_j)}e^{(\lambda_n-\lambda_j)x}\right]_{x=0}^{x=z} \\
&= \left(\prod_{i=1}^{n}\lambda_i\right)\sum_{j=1}^{n-1}\frac{1}{\prod_{k\neq j,k=1}^{n-1}(\lambda_k-\lambda_j)}e^{-\lambda_n z}\left(\frac{e^{(\lambda_n-\lambda_j)z}}{(\lambda_n-\lambda_j)}-\frac{1}{(\lambda_n-\lambda_j)}\right) \\
&= \left(\prod_{i=1}^{n-1}\lambda_i\right)\sum_{j=1}^{n-1}\frac{1}{\prod_{k\neq j,k=1}^{n-1}(\lambda_k-\lambda_j)}e^{-\lambda_n z}\left(\frac{e^{\lambda_j z}}{(\lambda_n-\lambda_j)}-\frac{e^{-\lambda_n z}}{(\lambda_n-\lambda_j)}\right) \\
&= \left(\prod_{i=1}^{n-1}\lambda_i\right)\sum_{j=1}^{n-1}\frac{e^{-\lambda_j z}-e^{-\lambda_n z}}{(\lambda_n-\lambda_j)\prod_{k\neq j,k=1}^{n-1}(\lambda_k-\lambda_j)}.
\end{aligned}$$

We have to show that

$$\sum_{j=1}^{n-1}\frac{e^{-\lambda_j z}}{\prod_{k\neq j,k=1}^{n}(\lambda_k-\lambda_j)}-\sum_{j=1}^{n-1}\frac{-e^{-\lambda_n z}}{\prod_{k\neq j,k=1}^{n}(\lambda_k-\lambda_j)}=\sum_{j=1}^{n}\frac{e^{-\lambda_j z}}{\prod_{\substack{k=1\\k\neq j}}^{n}(\lambda_k-\lambda_j)}. \quad (2.12)$$

Subtracting the first sum on both sides gives

$$-\sum_{j=1}^{n-1}\frac{-e^{-\lambda_n z}}{\prod_{k\neq j,k=1}^{n}(\lambda_k-\lambda_j)}=\frac{e^{-\lambda_n z}}{\prod_{k=1}^{n-1}(\lambda_k-\lambda_n)}.$$

We see that the expression on the right hand side would be th n-th term of the sum on the left. So (2.12) holds if and only if

$$-\sum_{j=1}^{n}\frac{1}{\prod_{k\neq j,k=1}^{n}(\lambda_k-\lambda_j)}=0$$

Multiply numerator and denominator by $\prod_{\substack{1\leq k,l\leq n\\k\neq l\neq j}}(\lambda_k-\lambda_l)$

$$\sum_{j=1}^{n}\frac{1}{\prod_{k\neq j,k=1}^{n}(\lambda_k-\lambda_j)}=\sum_{j=1}^{n}\frac{\prod_{k\neq l\neq j,k,l=1}^{n}(\lambda_k=\lambda_l)}{\prod_{\substack{1\leq k,l\leq n\\l\neq l}}(\lambda_k-\lambda_l)}.$$

This is zero if and only if

$$\sum_{j=1}^{n} \prod_{\substack{k\neq l\neq j \\ k,l=1}}^{n} (\lambda_k = \lambda_l) = 0. \tag{2.13}$$

Rewriting this expression gives

$$\begin{aligned}
&\sum_{j=1}^{n} \prod_{\substack{k\neq l\neq j \\ k,l=1}}^{n} (\lambda_k - \lambda_l) \\
&= \sum_{j=1}^{n} \prod_{\substack{1\leq k,l\leq n \\ j\neq k\neq l\neq j}} (\lambda_k - \lambda_l) \prod_{\substack{1\leq l\leq n \\ k=j, l\neq j}} (\lambda_k - \lambda_l) \\
&= \sum_{j=1}^{n} \prod_{\substack{1\leq k,l\leq n \\ j\neq k\neq l\neq j}} (\lambda_k - \lambda_l) \prod_{\substack{1\leq l<j \\ k=j}} (\lambda_k - \lambda_l) \prod_{\substack{j<l\leq n \\ k=j}} (\lambda_k - \lambda_l) \\
&= \pm \sum_{j=1}^{n} \prod_{\substack{1\leq l<k \\ 1\leq k\leq n \\ k\neq j, l\neq j}} (\lambda_k - \lambda_l)^2 \prod_{\substack{1\leq l<j \\ k=j}} (\lambda_k - \lambda_l) \prod_{\substack{j<l\leq n \\ k=j}} (\lambda_k - \lambda_l) \\
&= \pm \sum_{j=1}^{n} \prod_{\substack{1\leq l<k \\ 1\leq k\leq n \\ k\neq j, l\neq j}} (\lambda_k - \lambda_l)^2 \prod_{\substack{1\leq l<j \\ k=j}} (\lambda_k - \lambda_l) \prod_{\substack{j<k\leq n \\ l=j}} (\lambda_k - \lambda_l)(-1)^{n-j} \\
&= \pm \sum_{j=1}^{n} \prod_{\substack{1\leq l<k \\ 1\leq k\leq n \\ k\neq j, l\neq j}} (\lambda_k - \lambda_l) \Bigg(\prod_{\substack{1\leq l<k \\ 1\leq k\leq n \\ j\neq k\neq l\neq j}} (\lambda_k - \lambda_l) \prod_{\substack{1\leq l<j \\ k=j}} (\lambda_k - \lambda_l) \prod_{\substack{j<k\leq n \\ l=j}} (\lambda_k - \lambda_l) \Bigg) (-1)^{n-j} \\
&= \pm \sum_{j=1}^{n} \prod_{\substack{1\leq l<k \\ 1\leq k\leq n \\ k\neq j, l\neq j}} (\lambda_k - \lambda_l) \Bigg(\prod_{\substack{1\leq l<k \\ 1\leq k\leq n}} (\lambda_k - \lambda_l) \Bigg) (-1)^{n-j} \\
&= \pm \prod_{\substack{1\leq l<k \\ 1\leq k\leq n}} (\lambda_k - \lambda_l) \sum_{j=1}^{n} \prod_{\substack{1\leq l<k \\ 1\leq k\leq n \\ k\neq j, l\neq j}} (\lambda_k - \lambda_l)(-1)^{n-j}
\end{aligned}$$

which is zero if and only if

$$\sum_{j=1}^{n} \prod_{\substack{1\leq l<k \\ 1\leq k\leq n \\ k\neq j, l\neq j}} (\lambda_k - \lambda_l)(-1)^{j} \tag{2.14}$$

is zero. The product in (2.14) is the determinant of a Vandermonde matrix of

order n from which the j-th row and the $(n-1)$-th column have been removed:

$$\begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \dots & \lambda_1^{n-2} \\ 1 & \lambda_2 & \lambda_2^2 & \dots & \lambda_2^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{j-1} & \lambda_{j-1}^2 & \dots & \lambda_{j-1}^{n-2} \\ 1 & \lambda_{j+1} & \lambda_{j+1}^2 & \dots & \lambda_{j+1}^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \dots & \lambda_n^{n-2} \end{vmatrix}.$$

Consequently, (2.14) is the expansion with respect to its second column of a Vandermonde determinant of order n with the last column removed and an extra column of ones added in front.

$$\begin{vmatrix} 1 & 1 & \lambda_1 & \lambda_1^2 & \dots & \lambda_1^{n-2} \\ 1 & 1 & \lambda_2 & \lambda_2^2 & \dots & \lambda_2^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \lambda_n & \lambda_n^2 & \dots & \lambda_n^{n-2} \end{vmatrix}$$

This determinant is zero because its columns are linearly dependent. Therefore, (2.14) is zero too and (2.13) holds. Now we have completed the induction step.

□

Let X_i for $i = 1, \dots n$ with $n \geq 2$ be independent exponentially distributed random variables with pairwise distinct parameters λ_i and let $z > 0$. We use lemma 2.4.1 to calculate the conditional expectation that we were looking for.

$$\begin{aligned} &E(X_1 | X_1 + \dots + X_n = z) \\ &= \frac{1}{f_{X_1+\dots+X_n}(z)} \int_0^z x\lambda_1 e^{-\lambda_1 x} f_{X_2+\dots+X_n}(z-x)dx \\ &= C(z) \int_0^z x\lambda_1 e^{-\lambda_1 x} \Big(\prod_{i=2}^n \lambda_i \Big) \sum_{j=2}^n \frac{e^{-\lambda_j(z-x)}}{\prod_{\substack{k=2 \\ k\neq j}}^n (\lambda_k - \lambda_j)} dx \\ &= C(z) \int_{x=0}^z x e^{-\lambda_1 x} \sum_{j=2}^n \frac{e^{-\lambda_j(z-x)}}{\prod_{\substack{k=2 \\ k\neq j}}^n (\lambda_k - \lambda_j)} dx \\ &= C(z) \sum_{j=2}^n \frac{1}{\prod_{\substack{k=2 \\ k\neq j}}^n (\lambda_k - \lambda_j)} e^{-\lambda_j z} \int_{x=0}^z x e^{(\lambda_j - \lambda_1)x} dx \\ &= C(z) \sum_{j=2}^n \frac{1}{\prod_{\substack{k=2 \\ k\neq j}}^n (\lambda_k - \lambda_j)} e^{-\lambda_j z} \left(\frac{-e^{-(\lambda_1+\lambda_j)z}\big(z(\lambda_1+\lambda_j)+1\big)+1}{(\lambda_1+\lambda_j)^2} \right) \end{aligned} \tag{2.15}$$

with

$$C(z) = \left[\Big(\prod_{i=1}^n \lambda_i \Big) \sum_{j=1}^n \frac{e^{-\lambda_j z}}{\prod_{\substack{k=1 \\ k\neq j}}^n (\lambda_k - \lambda_j)} \right]^{-1} = \left[\sum_{j=1}^n \frac{e^{-\lambda_j z}}{\prod_{\substack{k=1 \\ k\neq j}}^n (\lambda_k - \lambda_j)} \right]^{-1}$$

To get equation (2.15) we evaluated the integral

$$\begin{aligned}
\int_0^z xe^{-(\lambda+\mu)x}dx &= \left[\frac{-x}{(\lambda+\mu)}e^{-(\lambda+\mu)x}\right]_{x=0}^{x=z} - \int_0^z \frac{-1}{(\lambda+\mu)}e^{-(\lambda+\mu)x}dx \\
&= \left(\frac{-z}{(\lambda+\mu)}e^{-(\lambda+\mu)z} - 0\right) + \frac{1}{(\lambda+\mu)}\int_0^z e^{-(\lambda+\mu)x}dx \\
&= \frac{-ze^{-(\lambda+\mu)z}}{(\lambda+\mu)} + \frac{1}{(\lambda+\mu)}\left[\frac{-1}{(\lambda+\mu)}e^{-(\lambda+\mu)x}\right]_{x=0}^{x=z} \\
&= \frac{-ze^{-(\lambda+\mu)z}}{(\lambda+\mu)} + \frac{1}{(\lambda+\mu)}\left(\frac{-e^{-(\lambda+\mu)z}}{(\lambda+\mu)} - \frac{-1}{(\lambda+\mu)}\right) \\
&= \frac{-z(\lambda+\mu)e^{-(\lambda+\mu)z} - e^{-(\lambda+\mu)z} + 1}{(\lambda+\mu)^2} \\
&= \frac{-e^{-(\lambda+\mu)z}\big(z(\lambda+\mu)+1\big)+1}{(\lambda+\mu)^2}.
\end{aligned}$$

2.5 The proportional variances method

Suppose again we partition a record $X_{i\cdot}$ into a missing and an observed part $(X_{i,mis}^T, X_{i,obs}^T)^T$ with the restriction $X_{ik} = \sum_{j=1}^{k-1} X_{ij}$, and let $l \in mis$. We wish to infer X_{il}, which is a missing item. Suppose that the total $\sum_{j=1}^{k-1} X_{ij}$ is observed. If more than one item is missing, we do not know $X_{i,l}$, but since we know the total, we do know some sum where X_{ij} is part in. Namely, $\sum_{j\in mis} X_{ij} = \sum_{j=1}^{k-1} X_{ij} - \sum_{j\in obs} X_{ij}$, and those last two terms on the right hand side are observed.

During our investigations, we thought about an estimator for missing value X_{il}:

$$\frac{\mu_l}{\sum_{j\in mis}\mu_j}\sum_{j\in mis} X_{ij}$$

where we used the means over the observed values in a column j for μ_j. This turned out to lead to good imputations. In this section, we will be looking for models that give rise to this estimator. That is, we want models such that

$$E\big(X_{il}|\sum_{j\in mis} X_{ij}\big) = \frac{\mu_l}{\sum_{j\in mis}\mu_j}\sum_{j\in mis} X_{ij}. \tag{2.16}$$

The first model we will consider assumes normality. This is a bit of a strange assumption, but it leads to the insight that estimator (2.16) is related to models in which the variance is proportional to the mean. Hence the name *proportional variances* method. This insight then brings us to the idea of using Hachemeister and Stanard's Poisson model, which is widely used in actuarial science [Mack 1999].

2.5.1 The normal model with proportional variances

Let the variable $X_{.j}$ for $j = 1, \ldots, k$ and i fixed be independently normally distributed, each with its own mean μ_j and variance σ_j^2. Then the sum of the missing items is again normally distributed and we can write

$$\begin{pmatrix} X_{il} \\ \sum_{j \in mis} X_{ij} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \mu_l \\ \sum_{j \in mis} \mu_j \end{pmatrix}, \begin{pmatrix} \sigma_l^2 & \sigma_l^2 \\ \sigma_l^2 & \sum_{j \in mis} \sigma_j^2 \end{pmatrix} \right) \tag{2.17}$$

where the cross elements of the covariance matrix are just σ_l^2 because we assumed $X_{.j}$ for $j = 1, \ldots, k$ to be independent.

Now theorem 2.1.1 gives us the conditional expectation of the missing item given the observed sum it is a part of

$$E\big(X_{il} | \sum_{j \in mis} X_{ij}\big) = \mu_l + \frac{\sigma_l^2}{\sum_{j \in mis} \sigma_j^2} \big(\sum_{j \in mis} X_{ij} - \sum_{j \in mis} \mu_j \big) \tag{2.18}$$

and we could choose this as the value to impute for X_{ij}.

For all items in a fixed record i except the total we assume that their variances are linear in their means with a common factor α

$$X_j \sim \mathcal{N}(\mu_j, \alpha\mu_j) \text{ independently for } j = 1, \ldots, k-1 \tag{2.19}$$

and we consider the record as a whole to be a multivariate normally distributed vector

$$X_{i,\cdot} \sim \mathcal{N}(\mu, \Sigma)$$

with

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{k-1} \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \alpha\mu_1 & & & 0 \\ & \alpha\mu_2 & & \\ & & \ddots & \\ 0 & & & \alpha\mu_{k-1} \end{pmatrix}.$$

We call this model *normal with proportional variances*. As we see later in section 3.4, this model shows some similarities with the model on which the ratio imputation method is based, but here without dependency between the variables $X_{i,1}, \ldots, X_{i,k-1}$.

The model may very well not stand as far from reality as might appear at first sight. Suppose as an illustration, the data X are from a shop which sells fixed quantities (of different kinds of products $j = 1, \ldots, k-1$. The shop can for example be supplied weekly, and each week sell its whole stock. If X_{ij} is the revenue of all products of kind j sold week i, it may be modeled as the sum of independent and identically distributed normal variables with variance σ^2. The randomness can for instance be due to bargaining. Every kind of product can have its own μ_j, which stands for the average revenue, but if all σ_j^2 are the same and we call it α, we get model (2.19).

The extra assumptions of (2.19) simplify (2.18) to

$$E\big(X_{il}|\sum_{j\in mis} X_{ij}\big) = \mu_l + \frac{\mu_l}{\sum_{j\in mis}\mu_j}\big(\sum_{j\in mis} X_{ij} - \sum_{j\in mis}\mu_j\big)$$
$$= \frac{\mu_l}{\sum_{j\in mis}\mu_j}\sum_{j\in mis} X_{ij}. \tag{2.20}$$

We now want to find the maximum likelihood estimators for μ and α. In order to be able to implement an EM algorithm, we need to calculate the likelihood for the data without any information missing.

If we fully observe the first $k-1$ items of a record, observing the k-th item, which is the sum of the other ones, does not add any additional information. So the likelihood function for n records with all items observed is

$$L_{\mu,\alpha}(X) = \prod_{i=1}^{n}\prod_{j=1}^{k-1}\frac{1}{\sqrt{2\pi\alpha\mu_j}}\exp\big(-\frac{(x_{ij}-\mu_j)^2}{2\alpha\mu_j}\big)$$

and the log-likelihood

$$\ell_{\mu,\alpha}(X) = -\frac{n}{2}\sum_{j=1}^{k-1}\log(2\pi\alpha\mu_j) - \sum_{j=1}^{k-1}\frac{\sum_{i=1}^{n}(x_{ij}-\mu_j)^2}{2\alpha\mu_j}. \tag{2.21}$$

Differentiation of (2.21) with respect to μ_j for $j = 1\dots k-1$ gives

$$\begin{aligned}
\frac{\partial}{\partial\mu_j}\ell_{\mu,\alpha}(X) &= -\frac{n}{2\mu_j} - \frac{\partial}{\partial\mu_j}\frac{\sum_{i=1}^{n}(x_{ij}-\mu_j)^2}{2\alpha\mu_j}\\
&= -\frac{n}{2\mu_j} - \sum_{i=1}^{n}\frac{-2\alpha\mu_j 2(x_{ij}-\mu_j) - 2\alpha(x_{ij}-\mu_j)^2}{(2\alpha\mu_j)^2}\\
&= -\frac{n}{2\mu_j} + \sum_{i=1}^{n}\frac{2\mu_j(x_{ij}-\mu_j) + (x_{ij}-\mu_j)^2}{2\alpha\mu_j^2}\\
&= -\frac{n}{2\mu_j} + \sum_{i=1}^{n}\frac{2\mu_j(x_{ij}-\mu_j) + x_{ij}^2 - 2\mu_j x_{ij} + {\mu_j}^2}{2\alpha\mu_j^2}\\
&= -\frac{n}{2\mu_j} + \sum_{i=1}^{n}\frac{x_{ij}^2 - {\mu_j}^2}{2\alpha\mu_j^2}\\
&= -\frac{n\alpha\mu_j}{2\alpha\mu_j^2} + \frac{\sum_{i=1}^{n}(x_{ij}^2-\mu_j^2)}{2\alpha\mu_j^2}
\end{aligned}$$

which is zero if

$$-n\alpha\mu_j + \sum_{i=1}^{n}(x_{ij}^2 - \mu_j^2) = 0.$$

Differentiation of (2.21) to α leads to

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ell_{\mu,\alpha}(X) &= -\frac{n}{2}\sum_{j=1}^{k-1}\frac{1}{\alpha}+\sum_{j=1}^{k-1}\sum_{i=1}^{n}\frac{(x_{ij}-\mu_j)^2}{2\mu_j\alpha^2}\\ &= -\frac{n(k-1)}{2\alpha}+\sum_{j=1}^{k-1}\sum_{i=1}^{n}\frac{(x_{ij}-\mu_j)^2}{2\mu_j\alpha^2}\end{aligned}$$

and this expression is zero if we set

$$\alpha = \frac{1}{n(k-1)}\sum_{j=1}^{k-1}\sum_{i=1}^{n}\frac{(x_{ij}-\mu_j)^2}{\mu_j}. \tag{2.22}$$

We use an implementation of the Newton-Raphson method to find the roots of these equations. Write

$$\nabla\ell = \begin{pmatrix} \frac{\partial\ell}{\partial\mu_1}\\ \frac{\partial\ell}{\partial\mu_2}\\ \vdots\\ \frac{\partial\ell}{\partial\mu_{k-1}}\\ \frac{\partial\ell}{\partial\alpha}\end{pmatrix} = \begin{pmatrix} \frac{-n\alpha\mu_1+\sum_{i=1}^{n}(x_{i1}^2-\mu_1^2)}{2\alpha\mu_1^2}\\ \frac{-n\alpha\mu_2+\sum_{i=1}^{n}(x_{i2}^2-\mu_2^2)}{2\alpha\mu_2^2}\\ \vdots\\ \frac{-n\alpha\mu_{k-1}+\sum_{i=1}^{n}(x_{i,k-1}^2-\mu_{k-1}^2)}{2\alpha\mu_{k-1}^2}\\ -\frac{n(k-1)}{2\alpha}+\sum_{j=1}^{k-1}\sum_{i=1}^{n}\frac{(x_{ij}-\mu_j)^2}{2\mu_j\alpha^2}\end{pmatrix} \tag{2.23}$$

for the gradient of the log-likelihood. In order to be able top use the Newton-Raphson method to find the roots of this function, we need to calculate its Jacobian matrix. That is, we calculate the Hessian of the log-likelihood

$$H_\ell = \begin{pmatrix} \frac{\partial^2\ell}{\partial\mu_1^2} & \cdots & \frac{\partial^2\ell}{\partial\mu_1\partial\mu_{k-1}} & \frac{\partial^2\ell}{\partial\mu_1\partial\alpha}\\ \vdots & \ddots & \vdots & \vdots\\ \frac{\partial^2\ell}{\partial\mu_{k-1}\partial\mu_1} & \cdots & \frac{\partial\ell}{\partial\mu_{k-1}^2} & \frac{\partial\ell}{\partial\mu_{k-1}\partial\alpha}\\ \frac{\partial^2\ell}{\partial\alpha\partial\mu_1} & \cdots & \frac{\partial^2\ell}{\partial\alpha\partial\mu_{k-1}} & \frac{\partial^2\ell}{\partial\alpha^2}\end{pmatrix}$$

which has at the first $k-1$ diagonal elements

$$\begin{aligned}\frac{\partial^2\ell}{\partial\mu_j^2} &= \frac{\partial}{\partial\mu_j}\Big(-\frac{n\alpha\mu_j+\sum_{i=1}^{n}(x_{ij}^2-\mu_j^2)}{2\alpha\mu_j^2}\Big)\\ &= -\frac{2\alpha\mu_j^2\times(n\alpha+\sum_{i=1}^{n}-2\mu_j)-\big(n\alpha\mu_j+\sum_{i=1}^{n}(x_{ij}^2-\mu_j^2)\big)4\alpha\mu_j}{(2\alpha\mu_j^2)^2}\\ &= -\frac{2n\alpha^2\mu_j^2-4n\alpha\mu_j^3-4n\alpha^2\mu_j^2-4\alpha\mu_j\sum_{i=1}^{n}(x_{ij}-\mu_j^2)}{4\alpha^2\mu_j^4}\\ &= \frac{2n\alpha^2\mu_j^2+4n\alpha\mu_j^3+4\alpha\mu_j\sum_{i=1}^{n}(x_{ij}^2-\mu_j^2)}{4\alpha^2\mu_j^4}\\ &= \frac{n\alpha\mu_j+2n\mu_j^2+2\sum_{i=1}^{n}(x_{ij}^2-\mu_j^2)}{2\alpha\mu_j^3}\end{aligned}$$

and in the first $(k-1) \times (k-1)$ block outside of the diagonal are only zeros. The first $k-1$ elements of the last row are

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \alpha \partial \mu_j} &= \frac{\partial}{\partial \alpha}\big(-\frac{n\alpha\mu_j + \sum_{i=1}^n (x_{ij}^2 - \mu_j^2)}{2\alpha\mu_j^2}\big) \\
&= -\frac{2\alpha\mu_j^2 \times n\mu_j - \big(n\alpha\mu_j + \sum_{i=1}^n (x_{ij}^2 - \mu_j^2)\big) \times 2\mu_j^2}{4\alpha^2\mu_j^2} \\
&= -\frac{n\alpha\mu_j^3 - n\alpha\mu_j^3 + \mu_j^2 \sum_{i=1}^n (x_{ij}^2 - \mu_j^2)}{2\alpha^2\mu_j^4} \\
&= -\frac{\sum_{i=1}^n (x_{ij}^2 - \mu_j^2)}{2\alpha^2\mu_j^2} = \frac{\sum_{i=1}^n (\mu_j^2 - x_{ij}^2)}{2\alpha^2\mu_j^2}
\end{aligned}$$

and the first $k-1$ elements of the last column are the same, because the Hessian is symmetrical around the diagonal.

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu_j \partial \alpha} &= \frac{\partial}{\partial \mu_j}\left(-\frac{n(k-1)}{2\alpha} + \sum_{j=1}^{k-1}\sum_{i=1}^{n} \frac{(x_{ij} - \mu_j)^2}{2\mu_j\alpha^2}\right) \\
&= \frac{\partial}{\partial \mu_j} \sum_{i=1}^{n} \frac{(x_{ij} - \mu_j)^2}{2\mu_j\alpha^2} \\
&= \sum_{i=1}^{n} \frac{2\mu_j\alpha^2 \times 2(x_{ij} - \mu_j) \times -1 - (x_{ij} - \mu_j)^2 \times 2\alpha^2}{4\mu_j^2\alpha^4} \\
&= \sum_{i=1}^{n} \frac{-4\mu_j\alpha^2(x_{ij} - \mu_j) - 2\alpha^2(x_{ij} - \mu_j)^2}{4\mu_j^2\alpha^4} \\
&= \sum_{i=1}^{n} \frac{-2\mu_j(x_{ij} - \mu_j) - (x_{ij} - \mu_j)^2}{2\mu_j^2\alpha^2} \\
&= \sum_{i=1}^{n} \frac{-2\mu_j x_{ij} + 2\mu_j^2 - (x_{ij}^2 - 2x_{ij}\mu_j + \mu_j^2)}{2\mu_j^2\alpha^2} \\
&= \frac{\sum_{i=1}^n (\mu_j^2 - x_{ij}^2)}{2\mu_j^2\alpha^2}
\end{aligned}$$

Finally, we calculate the lower right entry

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \alpha^2} &= \frac{\partial}{\partial \alpha}\left(-\frac{n(k-1)}{2\alpha} + \sum_{j=1}^{k-1}\sum_{i=1}^{n} \frac{(x_{ij} - \mu_j)^2}{2\mu_j\alpha^2}\right) \\
&= \frac{n(k-1)}{2\alpha^2} + \sum_{j=1}^{k-1}\sum_{i=1}^{n} -\frac{(x_{ij} - \mu_j)^2}{\mu_j\alpha^3}.
\end{aligned}$$

The proportional variances method where more sums are known It might be interesting to note that the proportional variance method method can be extended to situations where the missing value is part of more than one observed sum. Without loss of generality, suppose the missing value X_{il} is part

two sums $\sum_{j\in A} X_{ij}$ and $\sum_{j\in B} X_{ij}$ with different index sets A and B respectively. We need to write down the distribution of $(X_{il}, \sum_{j\in A} X_{ij}, \sum_{j\in B} X_{ij})$ as we did in (2.17).

It is obvious what the mean vector should be, so consider the covariance matrix. Since we assume the $X_{\cdot j}$ to be independently normally distributed for $j = 1, \ldots, k$, we have $\text{cov}(X_{il}, \sum_{j\in A} X_{ij}) = \sigma_l^2$ and $\text{cov}(X_{il}, \sum_{j\in B} X_{ij}) = \sigma_l^2$ too. On the diagonal, we get the the variance of the sums, which equal are the sums of the variances. The tricky ones are the cross terms of the two sums. Here we have to consider which terms the sums have in common and take the sum of the variances accordingly. So we get

$$\begin{pmatrix} X_{i,l} \\ \sum_{j\in A} X_{i,j} \\ \sum_{j\in B} X_{i,j} \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{pmatrix} \mu_l \\ \sum_{j\in A} \mu_j \\ \sum_{j\in B} \mu_j \end{pmatrix}, \begin{pmatrix} \sigma_l^2 & \sigma_l^2 & \sigma_l^2 \\ \sigma_l^2 & \sum_{j\in A} \sigma_j^2 & \sum_{j\in A\cap B} \sigma_j^2 \\ \sigma_l^2 & \sum_{j\in A\cap B} \sigma_j^2 & \sum_{j\in B} \sigma_j^2 \end{pmatrix} \right)$$

and theorem 2.1.1 tells us that

$$E(X_{i,l} | \sum_{j\in A} X_{i,j}, \sum_{j\in B} X_{i,j}) =$$
$$\mu_l + (\sigma_l^2, \sigma_l^2) \begin{pmatrix} \sum_{j\in A} \sigma_j^2 & \sum_{j\in A\cap B} \sigma_j^2 \\ \sum_{j\in A\cap B} \sigma_j^2 & \sum_{j\in B} \sigma_j^2 \end{pmatrix}^{-1} \left(\begin{pmatrix} \sum_{j\in A} \sigma_j^2 \\ \sum_{j\in B} \sigma_j^2 \end{pmatrix} - \begin{pmatrix} \sum_{j\in A} \mu_j \\ \sum_{j\in B} \mu_j \end{pmatrix} \right).$$

2.5.2 Poisson models

Another model with the property that the variance is proportional to the mean is the following. It takes some of the assumptions of the model Hachemeister and Stanard used in actuarial science, see [Mack 1999]. They used their model in a missing data problem in insurance mathematics. Insurance companies often have to deal with insured losses that have occurred but that not have been reported yet. This type of losses is called *IBNR*: incurred but not reported. Usually, tables are created displaying all the claims assigned to the year the corresponding insured event occurred. In such tables, X_{ij} stands for the amount claimed after j years related to events in year i . Actuaries call i the accident year and $i + j$ the development year.

In table 2.1, we see the structure of the missing data pattern. In for example the year 2006, the insurance company did not know yet how much damage occurred in 2000 would be claimed after 7 years. The lower triangle with missing values refers to future claims regarding past events. This kind of tables are called *IBNR triangles*, and Hachemeister and Stanard's model can be used to fill in the unknown part of the triangle, given the observed part. In our datasets, the missing pattern will not look this nice, but we will be able to observe the sum of every row.

Assume every item X_{ij} is an independently Poisson distributed random variable with parameter $\lambda_{ij} = \alpha_i \mu_j$, where both $\alpha_i > 0$ and $\mu_j > 0$ for all $i =$

accident year	development year 1	2	3	4	5	6	7	8	9
1997	142	1570	2400	3150	3500	3775	4300	4400	4200
1998	225	1250	1850	2675	3000	3100	3300	3200	?
1999	350	1550	2500	2775	3400	3550	3750	?	?
2000	170	700	1170	2610	3015	3195	?	?	?
2001	240	970	2025	3250	3750	?	?	?	?
2002	300	3000	4750	6250	?	?	?	?	?
2003	140	2310	3200	?	?	?	?	?	?
2004	400	950	?	?	?	?	?	?	?
2005	180	?	?	?	?	?	?	?	?

Table 2.1: Example of an IBNR triangle. The data are taken from the *Reinsurance* magazine, issue September 2006.

$1, \dots, n$ and $j = 1, \dots, k$. To make the model identifiable we add the restriction $\sum_{i=1}^{n} \alpha_i = 1$. In this model, we have

$$E(X_{ij}) = \text{var}(X_{ij}) = \lambda_{ij} = \alpha_i \mu_j$$

so the expected value of an entry is the product of a parameter μ_j which is the same throughout the column j, multiplied by a parameter α_i corresponding to row i. It is interesting to see the connection between the INBR background of the model, and our missing data problem. In the IBNR context, the parameter α_i models the amount of damage occurred in year i. Since every row in our dataset corresponds to an individual business, that the parameter α_i can be seen as a way to model the size of a company.

To find the maximum likelihood estimators for the parameters α_i and μ_j, we calculate the log-likelihood for the case where all values X_{ij} are observed:

$$\begin{aligned} \ell_{\alpha,\mu}(X) &= \log\Big(\prod_{i=1}^{n}\prod_{j=1}^{k} \frac{e^{-\lambda_{ij}}\lambda_{ij}^{x_{ij}}}{x_{ij}!}\Big) = \sum_{i=1}^{n}\sum_{j=1}^{k} \log\Big(\frac{e^{-\alpha_i\mu_j}(\alpha_i\mu_j)^{x_{ij}}}{x_{ij}!}\Big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k}\Big[-\alpha_i\mu_j + x_{ij}\log(\alpha_i\mu_j) - \log(x_{ij}!)\Big]. \end{aligned} \tag{2.24}$$

For fixed j, taking derivative with respect to μ_j gives

$$\begin{aligned} \frac{\partial}{\partial\mu_j}\ell_{\alpha,\mu}(X) &= \frac{\partial}{\partial\mu_j}\sum_{i=1}^{n}\sum_{j=1}^{k}\Big[-\alpha_i\mu_j + x_{ij}\log(\alpha_i\mu_j) - \log(x_{ij}!)\Big] \\ &= \sum_{i=1}^{n}\Big[-\alpha_i + x_{ij}\frac{1}{\alpha_i\mu_j}\alpha_i\Big] \\ &= \sum_{i=1}^{n}\Big[\frac{x_{ij}}{\mu_j} - \alpha_i\Big] = \sum_{i=1}^{n}\frac{x_{ij}}{\mu_j} - 1 \end{aligned}$$

where we use our assumption $\sum_{i=1}^{n}\alpha_i = 1$ in the last equality. Setting this

expression equal to zero gives us the maximum likelihood estimator for μ_j

$$\hat{\mu}_j = \sum_{i=1}^{n} x_{ij}. \tag{2.25}$$

Since (2.24) is symmetrical in μ_j and λ_i, we have

$$\frac{\partial}{\partial \alpha_i} \ell_{\alpha,\mu}(X) = \sum_{j=1}^{k} \left[\frac{x_{ij}}{\alpha_i} - \mu_j \right]$$

and if we also fix i and fill in $\mu_j = \hat{\mu}_j$ we can set this expression to zero

$$\sum_{j=1}^{k} \left[\frac{x_{ij}}{\alpha_i} - \hat{\mu}_j \right] = \sum_{j=1}^{k} \left[\frac{x_{ij}}{\alpha_i} - \sum_{i=1}^{n} x_{ij} \right] = 0$$

$$\frac{1}{\alpha_i} \sum_{j=1}^{k} x_{ij} - \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} = 0$$

to get the maximum likelihood estimator for α_i

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{k} x_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}}. \tag{2.26}$$

After we have estimated the parameters, we want to know the expectation of a single item if we observe the sum of the whole record which the item is part of.

Theorem 2.5.1. *Suppose X_j are independently Poisson distributed random variables with parameters $\lambda_j > 0$ for $j = 1, \ldots, k$ and let $z > 0$ an integer. Then for all $i = 1, \ldots, k$ fixed*

$$X_i | \sum_{j=1}^{k} X_j = z \sim \text{binom}(z, p_i) \tag{2.27}$$

with

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}.$$

Proof. Without loss of generality, we assume $i = 1$. We use the fact that the sum of independently Poisson distributed random variables is again Poisson distributed, i.e. $X_2 + \ldots + X_m \sim \text{Poisson}(\lambda_2 + \ldots + \lambda_k)$.

$$P(X_1 = x, \sum_{j=1}^{k} X_j) = P(X_1 = x) P(X_2 + \ldots + X_k = z - x)$$

$$= \frac{\lambda_1^x \exp -\lambda_1}{x!} \frac{(\lambda_2 + \ldots + \lambda_k)^{(z-x)} \exp -(\lambda_2 + \ldots \lambda_k)}{(z-x)!}$$

and

$$
\begin{aligned}
P(X_1 = x | \sum_{j=1}^{k} X_j = z) &= \frac{P(X_1 = x, X_1 + \ldots + X_k = z)}{P(X_1 + \ldots + X_k = z)} \\
&= \frac{\lambda_1^x (\lambda_2 + \ldots \lambda_k)^{(z-x)} \exp -(\lambda_1 + \ldots \lambda_k)/k! (z-x)!}{(\lambda_1 + \ldots + \lambda_k)^z \exp -(\lambda_1 + \ldots + \lambda_k)/z!} \\
&= \frac{z!}{x!(z-x)!} \frac{\lambda_1^x}{(\lambda_2 + \ldots + \lambda_k)^x} \frac{(\lambda_2 + \ldots \lambda_k)^z}{(\lambda_1 + \ldots + \lambda_k)^z} \\
&= \binom{z}{x} \frac{\lambda_1^x}{(\lambda_2 + \ldots + \lambda_k)^x} \frac{(\lambda_1 + \ldots + \lambda_k)^x}{(\lambda_1 + \ldots + \lambda_k)^x} \frac{(\lambda_2 + \ldots + \lambda_k)^z}{(\lambda_1 + \ldots + \lambda_k)^z} \\
&= \binom{z}{x} \frac{\lambda_1^x}{(\lambda_1 + \ldots + \lambda_k)^x} \frac{(\lambda_2 + \ldots + \lambda_k)^{(z-x)}}{(\lambda_1 + \ldots + \lambda_k)^{(z-x)}} \\
&= \binom{z}{x} p_1^x (1 - p_1)^{z-x}.
\end{aligned}
$$

□

Note that theorem 2.5.1 implies that

$$
E(X_i | X_1 + X_2 + \ldots + X_k = z) = \frac{\lambda_i z}{\lambda_1 + \lambda_2 + \ldots + \lambda_k}. \tag{2.28}
$$

Let us now return to our missing data problem, where the i-th record $X_{i,\cdot}$ satisfies the restriction $X_{ik} = \sum_{j=1}^{k-1} X_{ij}$, and we know some sum where a missing item X_{il} is part in. For the expectation of an item given the sum we get from (2.28)

$$
\begin{aligned}
E\big(X_{il} | \sum_{j \in mis} X_{ij}\big) &= \frac{\lambda_{il}}{\sum_{j \in mis} \lambda_{ij}} \sum_{j \in mis} X_{ij} \\
&= \frac{\alpha_i \mu_l}{\sum_{j \in mis} \alpha_i \mu_j} \sum_{j \in mis} X_{ij} \\
&= \frac{\mu_l}{\sum_{j \in mis} \mu_j} \sum_{j \in mis} X_{ij}
\end{aligned} \tag{2.29}
$$

which is the same as (2.20).

We will need to estimate μ_l in the presence of missing data. A naïve estimator would be $\hat{\mu}_l$ from (2.25), where we let i run over all values for which item X_{ij} is observed. To get more accurate estimates for these parameters we implement the EM algorithm. In the initial step we estimate the parameters μ_j with the naïve estimator, and calculate the conditional expectation of the missing values given the sum. In every next iteration, we calculate the maximum likelihood estimator (2.25) and use it to update the values for the missing entries.

Remark: a Poisson model without row parameter Since we saw in (2.29) that the parameters α_i fall out of the conditional expectation, the model $X_{ij} \sim \text{Poisson}(\lambda_j)$ leads to the same conditional expectation. If we look at a histogram

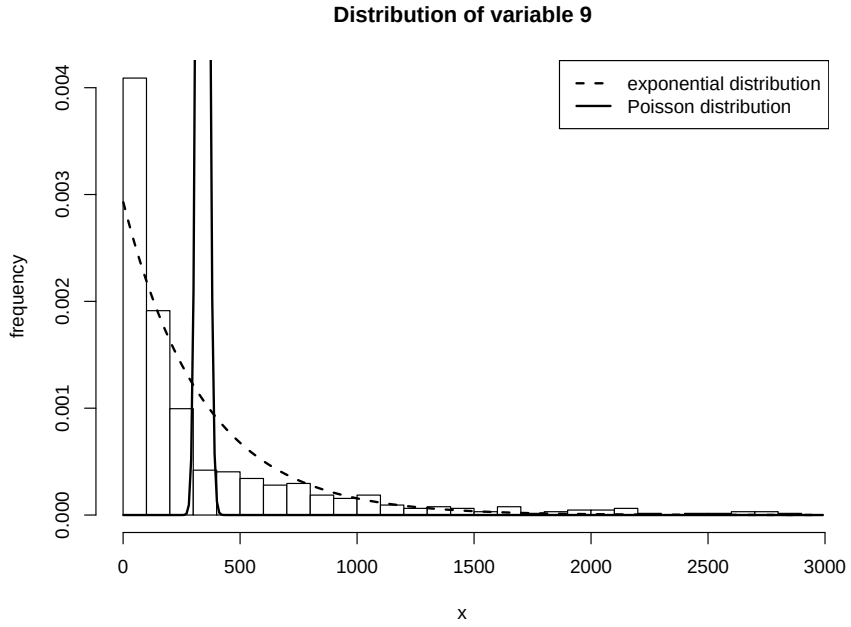


Figure 2.2: A histogram of variable `bt34600t`. The solid line shows a Poisson distribution fitted to the data. The dashed line shows an exponential distribution fitted to the data.

of the data in a column however, we see that a Poisson distribution does not fit to the data. In figure 2.2 it is clear that variable `bt34600t` is more likely to be exponentially distributed. For the other variables, the picture looks almost the same.

Chapter 3

Imputation methods using linear regression

One of the most common statistical methods used in econometrics is linear regression. The main goal is always to find a relation between a variable y_i and an explanatory variable $x_i = (x_1, \dots x_k)$. In order to do that, we observe the variables n times, so that we get a vector y and a matrix X. We assume that the n observations $(x, y)_i$ are exchangeable and use i to index the units or subjects. For the items, the k components of x, we use the index j, which is consistent with the notation we introduced for survey variables. The only difference lays in the $n \times k$ design matrix $X = (x_{ij})_{i,j}$ we get now. The variable we want to explain has been left out, and has been renamed y instead. In some literature, y is called X_j and X_{-j} is used instead of our X to stress that the design matrix is the original data matrix with one column left out. We wish to obtain the conditional distribution of y given X, which in the regression model is parameterized as $p(y|X, \theta)$. Since we will consider X fixed and not random, we will suppress all notation stating we are conditioning on X from now on.

In this chapter we will first recall some basic regression theory. Then we will extend it with a Bayesian approach and use truncated normal distributions to be able to handle restrictions on the data. In the last sections of this chapter we will investigate the ratio imputation method and the sequential regression method. Both methods use models that do not make any assumptions on a joint distribution, but only specify some conditional distributions.

3.1 Linear regression models

In the *normal linear regression model* the distribution of y given X is normal with mean linear in the parameter β:

$$E(y|\beta, X) = X\beta.$$

A constant term that does not depend on an actual explanatory variable x_i might be added by augmenting a column with only ones to the matrix X.

For all regression models, we assume that x has rank k. We have to do that to make sure the columns of X are linearly independent, so that the k coefficients of β are uniquely determined by the observed data. For the estimators and distributions, this assumption has the advantage that it implies $X^T X$ is invertible. This assumption is not completely necessary. Much of the regression theory still applies if one replaces the matrix inverse by a more general notion of inverse, so called pseudo inverses. For instance, Tempelman [Tempelman 2007] shows that the maximum likelihood estimates for a multivariate singular normal distribution using the Moore-Penrose pseudo-inverse agree with the estimates for the nonsingular case. However, we are in this thesis not only interested in the estimation of the parameters of regression models. We also want to analyse these models in a Bayesian fashion. To be able to actually draw from the posterior distributions we find, we need them to be non singular.

The other assumption we make is that $n > k$. This means the number of observed units should be at least as much as the number of parameters to estimate.

In *ordinary* linear regression models the conditional variances $\text{var}(y_i|\beta, X)$ are equal to a constant $\sigma^2 > 0$, and the observations y_i are for all i independent given $\theta = (\beta_1, \ldots, \beta_k, \sigma^2)$ and X.

It is important to note that in terms of our survey notation, we have implicitly assumed that each column x_j of the data matrix has its own error term $e_j \sim \mathcal{N}(0, \sigma_j^2)$ and that all the variances σ_j^2 may be different for each column. Here, the index j does not denote the element in the vector, but stresses the fact that all columns have their own individual regression model. In some practical cases it might be more realistic to model a common variance mechanism applying for more columns simultaneously, but this would lead to cumbersome calculations. Moreover, the approach we use now is in line with the idea of sequential regression, where a regression model is fitted to the data column by column.

There are in general two ways to implement a regression model in an imputation method. The first consists of obtaining a value for β, say $\hat{\beta}$ using the observed values in the data matrix. Then, for a missing value y_i, we fill in $\hat{y}_i = (X\hat{\beta})_i$. This regression method is sometimes called *deterministic*, because there is no account for the value of the variance of the error term σ^2. We simply leave the error term out, because by our modeling assumptions its expectation is zero. Note that the word deterministic is a bit suggestive here, since the way $\hat{\beta}$ is obtained can involve same random process. So strictly speaking, we cannot call this method a deterministic imputation method in the sense of section 1.5. Therefore, we will not use the word deterministic for these methods, but instead say that these fill in the expectation of a distribution. The key disadvantage of this approach is that the variance in the items filled in is smaller than what we would expect from the regression model, since we fill in the expectation.

The second method overcomes this problem by adding a variance term to the

imputed items. In this approach, we not only estimate β but also σ, say with $\hat{\sigma}$ and for a missing value y_i we fill in $\hat{y}_i = (X\hat{\beta})_i + e_i$, where we draw e_i from a $\mathcal{N}(0, \hat{\sigma})$ distribution. Now the values we filed in get the right distribution, that is the distribution of the model. In practice however, we may expect that the mean of the imputed items will be not as close to the real mean as with the first approach. Regression methods that use a variance term are often called stochastic in literature. But again, in order not to be confused with the notions of section 1.5, we will say these methods use a draw from a distribution.

3.1.1 The classical ordinary normal linear regression model

In this model we suppose $y = X\beta + e$ with $e \sim \mathcal{N}_n(0, \sigma^2 I)$. This means $y \sim \mathcal{N}_n(X\beta, \sigma^2 I)$ so that indeed $E(y) = X\beta$ and Cov$(y) = \sigma^2 I$. The maximum likelihood estimator for β is $\hat{\beta} = (X^T X)^{-1} X^T y$, and the maximum likelihood estimator of σ^2 is the average of the squared residuals

$$\hat{\sigma}^2 = \frac{1}{n}(y - \hat{y})^T (y - \hat{y}) \tag{3.1}$$

were $\hat{y} = X\hat{\beta}$. Since $\hat{\beta} = (X^T X)^{-1} X^T y$, it is convenient to define a projection matrix $P_X = X(X^T X)^{-1} X^T$, so that $P_X y = X\hat{\beta} = \hat{y}$. The matrix P_X is sometimes called the the *hat matrix*, and the t^{th} diagonal element of it is denoted by h_t. Similarly, define the complementary projection $M_X = I - P_X$, so that $M_X y = y - P_X y = y - X\hat{\beta}$. We call $\hat{e} := y - X\hat{\beta}$ the vector of least squared residuals. Note that $M_X X\beta = 0$, so

$$\hat{e} = M_X y = M_X X\beta + M_X e = M_X e.$$

Furthermore,

$$X^T \hat{e} = X^T y - X^T X\hat{\beta} = X^T y - X^T X(X^T X)^{-1} = 0 \tag{3.2}$$

and we see that $\hat{e}$ is orthogonal to X. In order to calculate var$(\hat{e})$, look at the whole covariance matrix of $\hat{e}$ and use the fact that $E(M_X e) = 0$ and that M_X is a symmetric matrix, so $M_X^T = M_X$,

$$\begin{aligned} Cov(\hat{e}) = Cov(M_X e) = E(M_X e e^T M_X) = M_X E(ee^T) M_X = \\ = M_X Cov(e) M_X = M_X(\sigma^2 I) M_X = \\ = \sigma^2 M_X M_X = \sigma^2 M_X. \end{aligned}$$

Since $E\hat{e} = 0$ and the t^{th} diagonal element of M_X is $1 - h_t$,

$$var(\hat{e}_t) = E(\hat{e}^2) = (1 - h_t)\sigma^2.$$

Looking at the diagonal elements of the hat matrix and using the cyclic property of trace we get

$$\begin{aligned} \sum_{t=1}^{n} h_t = Tr(P_X) = Tr(X(X^T X)^{-1} X^T) \\ = Tr((X^T X)^{-1} X^T X) = Tr(I_k) = k. \end{aligned}$$

Now we calculate the expectation

$$E(\hat{\sigma}^2) = \frac{1}{n}E\big[(y-\hat{y})^T(y-\hat{y})\big] = \frac{1}{n}\sum_{t=1}^{n}E(\hat{e}_t^2)$$
$$= \frac{1}{n}\sum_{t=1}^{n}var(\hat{e}_t) = \frac{1}{n}\sum_{t=1}^{n}(1-h_t)\sigma^2 = \frac{n-k}{n}\sigma^2.$$

Therefore, an alternative to the maximum likelihood estimator (3.1) is the unbiased estimator

$$s^2 = \frac{1}{n-k}(y-\hat{y})^T(y-\hat{y}). \tag{3.3}$$

3.1.2 Bayesian analysis of the ordinary normal linear regression model

We remain with the ordinary linear model but now write the model in terms of the distribution of y given the parameters and the design

$$y|\beta,\sigma^2 \sim \mathcal{N}(X\beta,\sigma^2 I).$$

Prior distributions

Since we do not have any real prior information for the parameters β and σ, we want to give them a non informative prior density. For β, this obviously means that we choose $p(\beta)$ proportional to a constant.

A widely used approach for the prior for σ^2 comes from Jeffreys as explained in [Zellner 1971]. The key idea is that if the parameter ϕ is a function of another parameter $\phi = h(\theta)$, the prior for ϕ should reflect the same information as the prior for θ, using the change of variables formula

$$p(\phi) = p(\theta)|\frac{\partial\theta}{\partial\phi}| = p(\theta)|h'(\theta)|^{-1}. \tag{3.4}$$

For parameters that may assume values in $(-\infty,\infty)$, for instance the mean μ or the parameter β from our ordinary regression model, he suggests to choose $p(\mu)$ so that

$$p(\mu)d\mu \propto d\mu, -\infty < \mu < \infty \tag{3.5}$$

and he takes (3.5) as a formal notion of ignorance. Note that (3.5) implies $p(\mu)$ is proportional to a constant. Obviously , this is an improper density because $\int_{-\infty}^{\infty}p(\mu)d\mu = \infty$. This is no problem however, as long as the integral with respect to the parameter μ of the likelihood $p(y|\mu)$ is finite. For in that case we can normalize the posterior

$$\frac{p(\mu,y)}{p(y)} = \frac{p(y|\mu)p(\mu)}{\int_{\mu'}p(y|\mu')p(\mu')d\mu'} = \frac{p(y|\mu)}{\int_{\mu'}p(y|\mu')d\mu'} \tag{3.6}$$

so that it integrates to one and therefore is a proper distribution. We used Jeffreys' (3.5) in the last equality in equation (3.6).

For parameters in the interval $(0, \infty)$, for instance a standard deviation σ, Jeffreys advises to take the logarithm uniform, so that if we write $\theta = \log \sigma$

$$p(\theta)d\theta \propto d\theta$$

which is consistent with the concept of ignorance (3.5), since $-\infty < \theta < \infty$. In terms of the actual parameter σ this leads to

$$p(\sigma)d\sigma \propto \frac{1}{\sigma}d\sigma. \tag{3.7}$$

Suppose we use Jeffrey's choice (3.7), and another researcher uses a model with parameterization σ^2 or precision parameter $1/\sigma^2$. More generally, consider a transformation of the form $\phi(\sigma) = \sigma^n$. Then $d\phi = n\sigma^{n-1}d\sigma$ and $\frac{1}{\phi}d\phi = n\sigma^{-1}d\sigma$ which leads to

$$\frac{1}{\phi}d\phi \propto \frac{1}{\sigma}d\sigma \tag{3.8}$$

and we see these other parameters bring about priors of the same form, which is a nice property.

Keeping in mind these considerations we choose the prior distribution for $(\beta, \log \sigma)$to be uniform on $\mathbb{R}^2$. Furthermore we assume that β and $\log \sigma$ are independent, and consequently we may state that

$$p(\beta, \sigma^2) = p(\beta)p(\sigma^2) \propto p(\sigma^2) \propto \sigma^{-2}. \tag{3.9}$$

Posterior distributions

To find the posterior distribution, we look at the factorization

$$p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)p(\sigma^2|y) \tag{3.10}$$

which tells us that we can first derive the posterior for β conditional on σ^2 and then the marginal posterior distribution for σ^2. For the first derivation, look at

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

We chose the prior for β to be the improper uniform distribution on $\mathbb{R}$, independent of σ^2, so $p(\beta|\sigma^2) = p(\beta)$ and is a constant. Furthermore, the densities $p(\sigma^2)$ and $p(\sigma^2, y)$ do not involve β, so

$$p(\beta|\sigma^2, y) \propto p(y|\beta, \sigma^2). \tag{3.11}$$

By our model assumption

$$p(y|\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \sigma^2 I}} e^{-\frac{1}{2}(y-X\beta)^T(\sigma^2 I)^{-1}(y-X\beta)}$$

and filling this in in equation (3.11) we get

$$p(\beta|\sigma^2, y) \propto e^{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)} \tag{3.12}$$

in which we recognize a normal distribution.

We now want to know the mean and the variance of this distribution.
Proposition 3.1.1. *Define (suggestive notation)* $\hat{\beta} = (X^T X)^{-1} X^T y$. *Then*

$$(y - X\beta)^T (y - X\beta) = (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) + (y - X\hat{\beta})^T (y - X\hat{\beta}). \quad (3.13)$$

Proof. Consider

$$y - X\beta = y - X\hat{\beta} + X\hat{\beta} - X\beta = \hat{e} - X(\beta - \hat{\beta})$$

to see

$$(y - X\beta)^T (y - X\beta) = \hat{e}^T \hat{e} - \hat{e}^T X(\beta - \hat{\beta}) - (\beta - \hat{\beta})^T X^T \hat{e} + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}).$$

and from the orthogonality relation (3.2) the terms $\hat{e}^T X(\beta - \hat{\beta})$ and $(\beta - \hat{\beta})^T X^T \hat{e}$ are zero

$$(y - X\beta)^T (y - X\beta) = (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) + \hat{e}^T \hat{e}.$$

□

The second part of the right hand side of (3.13) does not involve the actual β, which means that (3.12) reduces to

$$p(\beta|\sigma^2, y) \propto e^{\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})} \quad (3.14)$$

and we conclude that the mean of the distribution of $\beta|\sigma^2, y$ is indeed $\hat{\beta} = (X^T X)^{-1} X^T y$ and the variance $(X^T X)^{-1}\sigma^2$.

For the second part we use (3.10) and write the marginal posterior distribution of σ^2 as

$$p(\sigma^2|y) = \frac{p(\beta, \sigma^2|y)}{p(\beta|\sigma^2, y)}. \quad (3.15)$$

We know

$$p(\beta|\sigma^2, y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(X^T X)^{-1}\sigma^2}} e^{\frac{1}{2}(\beta - \hat{\beta})^T ((X^T X)^{-1}\sigma^2)^{-1} (\beta - \hat{\beta})}$$

and we assumed the prior was $p(\beta, \sigma^2) \propto \sigma^{-2}$ so

$$p(\beta, \sigma^2|y) \propto p(\beta, \sigma^2) p(y|\beta, \sigma^2) \propto \sigma^{-2} p(y|\beta, \sigma^2)$$

and using proposition (3.13) we can fill in (3.15)

$$\begin{aligned} p(\sigma^2|y) &= \frac{\sigma^{-2}\big((2\pi)^{n/2}\sqrt{\det \sigma^2 I}\big)^{-1} \exp\big(-\frac{1}{2}(y - X\beta)^T (\sigma^2 I)^{-1} (y - X\beta)\big)}{\big((2\pi)^{k/2}\sqrt{\det(X^T X)^{-1}\sigma^2}\big)^{-1} \exp\big(-\frac{1}{2}(\beta - \hat{\beta})^T ((X^T X)^{-1}\sigma^2)^{-1} (\beta - \hat{\beta})\big)} \\ &= \frac{\sigma^{-2}\big((2\pi)^{n/2}\sqrt{\det \sigma^2 I}\big)^{-1}}{\big((2\pi)^{k/2}\sqrt{\det(X^T X)^{-1}\sigma^2}\big)^{-1}} \exp\left(-\frac{1}{2\sigma^2}(y - X\hat{\beta})^T (y - X\hat{\beta})\right) \end{aligned}$$

where I is the $n \times n$ identity matrix. Note that $X^T X$ has dimension k, so $\det((X^T X)^{-1}\sigma^2) = \sigma^{2k}\det((X^T X)^{-1})$, and $\det(\sigma^2 I) = \sigma^{2n}$, so

$$\frac{\sigma^{-2}\left((2\pi)^{n/2}\sqrt{\det \sigma^2 I}\right)^{-1}}{\left((2\pi)^{k/2}\sqrt{\det(X^T X)^{-1}\sigma^2}\right)^{-1}} = \sigma^{k-n-2}(2\pi)^{(k-n)/2}\sqrt{\det(X^T X)^{-1}}$$

and we see that

$$p(\sigma^2|y) \propto \sigma^{-(n-k+2)} e^{-(y-X\hat{\beta})^T(y-X\hat{\beta})/2\sigma^2}$$

in which we recognize the density of a scaled inverse χ^2 distribution:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-k, s^2)$$

with

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})^T(y - X\hat{\beta}). \tag{3.16}$$

3.2 Regression models with restrictions

In this chapter, we revisit the regression models described in section 3.1. We still assume $y = X\beta + e$, but now we want y to take values only in a certain region. The boundaries of the permitted area are given by a set of linear equations that may depend on X. In the notation we introduced for survey variables, the restrictions take the form $l \leq BX \leq u$, where X is the data matrix containing all variables, including y. For simplicity, choose l^* and u^* so that we can describe the permitted region for y as an interval $l^* \leq y \leq u^*$.

To restrict the range of y, we will have to impose some extra properties on the errors e. We want to truncate e such that y becomes truncated to the permitted region, which means $l^* - X\beta \leq e \leq u^* - X\beta$.

$$e_i \sim \mathcal{N}(0, \sigma^2)|G \text{ with } G = [l^* - X\beta, u^* - X\beta]. \tag{3.17}$$

3.2.1 The classical truncated normal regression model

The classical way of looking at model (3.17) is to fill in estimators for the parameters β and σ. In this model, the expectation of the error terms is in general no longer zero. This means the least squares estimator is not consistent any more. Instead, we should calculate the maximum likelihood estimators.

In the permitted region, the density function for an individual value y_i is still normal, but divided by the probability the non truncated version would take values in that region

$$f(y_i) = \frac{1/\sigma\phi\left(\frac{(y-X\beta)_i}{\sigma}\right)}{\Phi\left(\frac{(u^*-X\beta)_i}{\sigma}\right) - \Phi\left(\frac{(l^*-X\beta)_i}{\sigma}\right)} \tag{3.18}$$

where Φ and ϕ denote the standard normal distribution and its density function respectively. Outside that region the density is zero. So the log-likelihood function of item y_i is

$$\ell_i(\beta,\sigma) = -\log\sigma + \log\phi\big(\frac{(y-X\beta)_i}{\sigma}\big) - \log\bigg(\Phi\big(\frac{(u^*-X\beta)_i}{\sigma}\big) - \Phi\big(\frac{(l^*-X\beta)_i}{\sigma}\big)\bigg).$$

Because we also want to use information from the other items in the vector y, we take a look at the multivariate likelihood function

$$\begin{aligned}\ell(\beta,\sigma) = & -n\log\sigma + \sum_{i=1}^{n}\log\phi\big(\frac{y_i-(X\beta)_i}{\sigma}\big) - \\ & + n\log\bigg(\Phi\big(\frac{(u^*-X\beta)_i}{\sigma}\big) - \Phi\big(\frac{(l^*-X\beta)_i}{\sigma}\big)\bigg).\end{aligned}$$

To find the values for β and σ for which this expression is maximal, we differentiate to both variables

$$\begin{aligned}\frac{\partial\ell(\beta,\sigma)}{\partial\beta} &= -\frac{n}{\sigma}\sum_{i=1}^{n}\bigg(\frac{(y_i-X\beta)_i}{\sigma^2} + \frac{1}{\sigma}\frac{\phi\big((u^*-X\beta)_i/\sigma\big)-\phi\big((l^*-X\beta)_i/\sigma\big)}{\Phi\big((u^*-(X\beta)_i/\sigma\big)-\Phi\big((l^*-X\beta)_i/\sigma\big)}\bigg) \\ \frac{\partial\ell(\beta,\sigma)}{\partial\sigma} &= -\frac{n}{\sigma} + \sum_{i=1}^{n}\bigg(\frac{(y-X\beta)^2}{\sigma^3} \\ &\quad + \frac{1}{\sigma^2}\frac{(u^*-X\beta)\phi\big(\frac{u^*-X\beta}{\sigma}\big)-(l^*-X\beta)\phi\big(\frac{l^*-X\beta}{\sigma}\big)}{\Phi\big(\frac{u^*-X\beta}{\sigma}\big)-\Phi\big(\frac{l^*-X\beta}{\sigma}\big)}\bigg)\end{aligned}$$

and may use a numerical method to find the roots of these expressions.

To get an idea what the likelihood for β looks like, we run a small simulation, roughly based on Raghunatan's smoking case study [Raghunatan 2001]. We will omit the context here, but the idea is that we simulate the end of a time period (the number of years someone has been smoking) as an exponential random variable. This time period is then the permitted region for a truncated normally distributed random variable. In our simulation, we

1. generate a random vector u of 10 exponentially distributed random variables with mean 20;
2. generate a random vector x of 10 standard normally distributed random variables truncated to the permitted region bounded by the zero vector and u;
3. calculate the log-likelihood for a lot of values of β, using the actual value $\sigma = 1$;
4. take for $\hat{\beta}$ the value of β for which the log-likelihood is the highest;
5. also calculate the sample mean $\bar{x}$ of x. This would be the maximum likelihood estimator if we would not have known the data had been truncated.

A picture for the log-likelihood as a function of β is plotted in figure 3.1. We see that the top of the plot is close to the true value $\beta = 0$.

We see the likelihood is more steep on the right of zero than on the left. This is due to the truncation on the left which we fixed at zero. On the right, we have

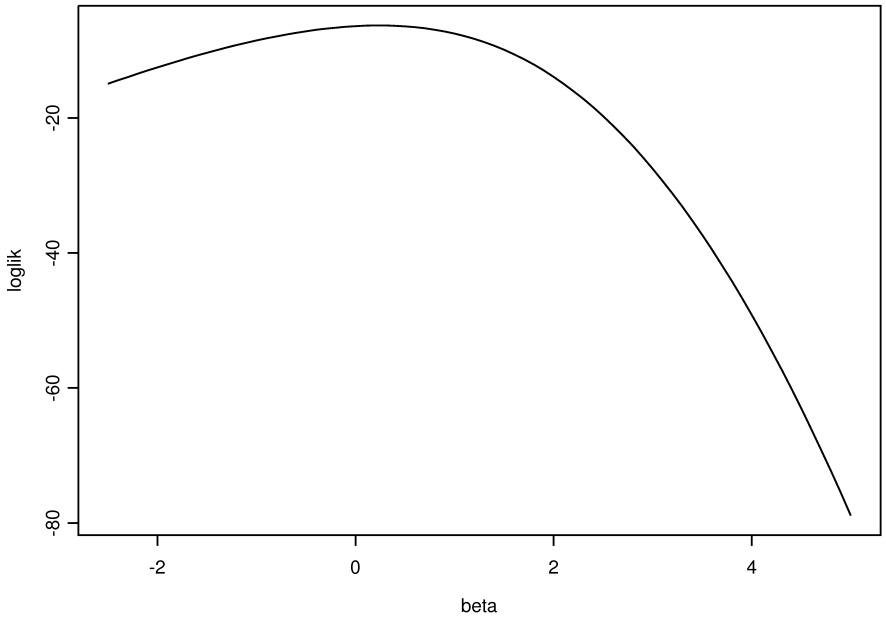


Figure 3.1: The log-likelihood of standard normally distributed variables truncated below zero as a function of β.

used different bounds for each observation. As a result, more information is available around the right side of the permitted region. For the likelihood this means that is easier to discriminate between values on the right of zero than on the left.

We repeat our simulation program 500 times, and compare the sample means for the data drawn in each simulation with the maximum likelihood estimator. Figure 3.2 shows the results. On the x-axis it shows the sample means of the data generated in each iteration and on the y axis the corresponding maximum likelihood estimate. These values along the y-axis are clustered because of the (lack of) numerical precision in our estimation of β.

It seems that for larger sample means, $\hat{\beta}$ is getting closer to the sample mean. In the picture we see this because the data points are getting closer to the line $y = x$, which we have drawn for convenience. A possible explanation is the following. If the sample mean is a bit low, it could be caused by a short permitted interval. A large part of the right tail of the $\mathcal{N}(0, 1)$ distribution is then truncated, so only points near zero are being observed. This results in the likelihood giving too much weight to values of β that are lower than zero. For lager sample means, this effect diminishes, and the estimate $\hat{\beta}$ comes closer to the value of the sample mean. This also explains why almost all data points are below the line $y = x$.

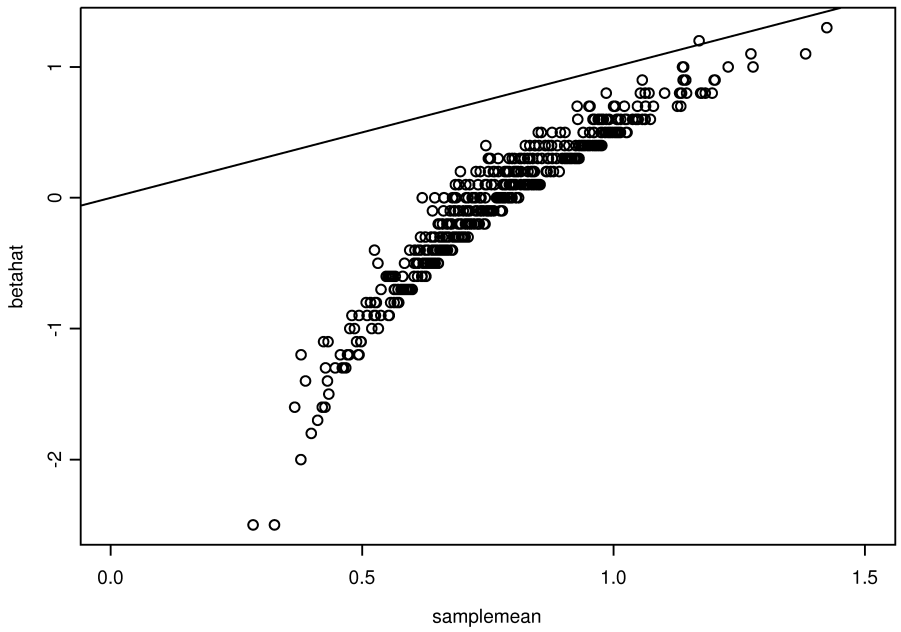


Figure 3.2: The value sample means of standard normally distributed variables truncated below zero plotted against the value $\hat{\beta}$ for which the log-likelihood is maximal.

3.2.2 Bayesian analysis of the truncated normal regression model

For the Bayesian analysis of the model, we take the same prior (3.9) on β and σ^2 and use (3.10) again. Although we need to be somewhat more careful as now e depends on β, we can still say that $p(y)$ does not depend on β so

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

still holds. But our model assumption on y given β and σ^2

$$p(y|\beta, \sigma^2) = \begin{cases} \frac{((2\pi)^{n/2}\sqrt{\det \sigma^2 I})^{-1}\exp\left(-\frac{1}{2}(y-X\beta)^T(\sigma^2 I)^{-1}(y-X\beta)\right)}{\int_G ((2\pi)^{n/2}\sqrt{\det \sigma^2 I})^{-1}\exp\left(-\frac{1}{2}(y-X\beta)^T(\sigma^2 I)^{-1}(y-X\beta)\right)dy} & \text{if } y \in G \\ 0 & \text{if } y \notin G \end{cases}$$

is now a truncated distribution. Looking only at the case $y \in G$ we get

$$\begin{aligned} &\frac{((2\pi)^{n/2}\sqrt{\det \sigma^2 I})^{-1}\exp\left(-\frac{1}{2}(y-X\beta)^T(\sigma^2 I)^{-1}(y-X\beta)\right)}{\int_G ((2\pi)^{n/2}\sqrt{\det \sigma^2 I})^{-1}\exp\left(-\frac{1}{2}(y-X\beta)^T(\sigma^2 I)^{-1}(y-X\beta)\right)dy} \\ &= \frac{\exp\left(-\frac{1}{2}(y-X\beta)^T(\sigma^2 I)^{-1}(y-X\beta)\right)}{\int_G \exp\left(-\frac{1}{2}(y-X\beta)^T(\sigma^2 I)^{-1}(y-X\beta)\right)dy} \\ &= \frac{\exp(y-X\beta)^T(y-X\beta)}{\int_G \exp(y-X\beta)^T(y-X\beta)dy}. \end{aligned}$$

If we calculate $p(\sigma^2|y)$ we get

$$p(\sigma^2|y) = \frac{p(\beta, \sigma^2|y)}{p(\beta|\sigma^2, y)} \propto \frac{\sigma^{-2} p(y|\beta, \sigma^2)}{p(\beta|\sigma^2, y)}. \tag{3.19}$$

Unfortunately however, in this expression we dot recognize a nice distribution from which we could easily draw inferences. Therefore, as an alternative, we suggest to use the Bayesian version of the linear model described in section 3.1.2, to get a posterior for β and σ^2. We can then draw inferences for y from a multivariate normal distribution with parameters $X\beta$ and $\sigma^2 I$. To make sure the inferences for each y_i satisfy the restrictions, we could reject such a draw if it does not lay in the appropriate interval, and repeat drawing until it does.

Again, we implemented two versions of the imputation process using this regression method. The first one uses the posterior $p(y|\beta, \sigma^2)$ to draw values y_i for items that are missing. The second one still draws β from its posterior, but calculates the expectation of th error term e_i instead of drawing it from a truncated distribution.

3.3 Sequential regression

Sequential regression methods were developed by Van Buuren and Ranghunatan. The idea is to use a separate regression model for the univariate distribution of each variable conditioned on all other variables. By doing that, we can consider the restrictions for each variable separately. In the initial step of the method, all missing values must be imputed by a guess that need not necessarily be adequate, but it should satisfy the conditions. Next, all variables are imputed one by one by the different regression models, using the most recent updates of the imputed explanatory variables. The imputations are drawn from the posterior distribution that the model brings forward, which is truncated to the correct region so that the draw satisfies the restrictions. This sequence of iterations is then repeated, hence the name of the method.

In this method, all variables can have separate regression models. Most often used are ([Raghunatan 2001], p. 87) normal linear models for continuous variables, logistic models for binary variables, generalized logit models for categorial variables, Poisson models for non-negative integer variables and for semi-continuous variables a two stage model, which first uses a logistic regression to model the zero or non zero status and conditional on non-zero status a normal linear regression to impute non-zero values. The problem with the linearity of the restriction is also solved, because for each variable the restrictions can be rewritten separately.

In our simulations we will be considering continuous data from a business survey. That means that for us a normal model would be convenient, because we can use truncation to take the restrictions into account. But there is a problem. In economical data where the normality assumption is inappropriate, we can often use some transformation to make the data more suitable for the model. In our case however, we want linear restrictions to hold, of the form

$X_1 + X_2 + \ldots + X_m = X_{tot}$. If the transformation, say ϕ, is not linear, then $\phi(X_1) + \phi(X_2) + \ldots + \phi(X_m) \neq \phi(X_{tot})$, so it is not clear what the restriction transforms to. Therefore we cannot apply such transformations to our data.

The main disadvantage of sequential regression is the possibility of lack of convergence of the methods. The freedom to use different regression models for each variable comes with the drawback that is is not clear under what conditions the distributions of the imputations converge to a multivariate distribution. In that case, the conditional distributions estimated by the univariate regression models are said to be *incompatible*.

Although the possibility of incompatibility is a very serious problem on which more research should be focused, both Raghunatan [Raghunatan 2001] and Van Buuren [Van Buuren 2007] claim to have achieved good results by applying sequential regression. Therefore it seems worthwhile to apply the method in simulation studies on real data.

For our simulation studies, we implemented a sequential regression algorithm in R. In this implementation, we use that the specific dataset we will be investigating must satisfy a linear balance equation of the form $\sum_i X_i = X_{tot}$.

3.3.1 Initial step

In the first step of the method we need to fill in initial values for the missing entries. We choose the proportional variances method from section 2.5 for this, because it is a fast method that gives quite good imputations, as we will see later. We do not make any further assumptions on the underlaying model, but just take ad hoc estimators for the values of μ_j for each variable j in equation 2.20. Start by taking the column sums s_j of those records in the simulation dataset for which all variables are observed. Calculate the proportions of each value s_j to the value s_{tot} corresponding to the variable X_{tot}, and use it as an ad hoc estimator for μ_j

$$\mu_j = s_j / s_{tot}. \tag{3.20}$$

Let mis denote the set of indices for which the value is missing in a record, and obs the indices for which the value is observed. We will always assume that the total X_{tot} is observed. Then we know that the sum of the missing values must be $\sum_{j \in mis} X_{ij} = X_{tot} - \sum_{j \in obs} X_{ij}$. So if the k-th position in record i is missing we can fill in

$$\frac{\mu_k}{\sum_{j \in mis} \mu_j} \Big(\sum_{j \in mis} X_{ij} \Big). \tag{3.21}$$

If we do that for all missing entries we have filled the record in such a manner that it satisfies the restrictions. This turns out to lead to pretty good estimates, which are later spoiled in the sequential regression process.

3.3.2 Regression iteration

In every next iteration, take the column containing the data for the first variable. Take this variable as y for the regression model. The problem is now that

y is completely determined by the balance restriction, since all the other variables in the dataset are considered as fixed covariates. It makes therefore no sense at all to try to find a better value for y to update it.

The solution we suggest is to leave one variable out of the model. This variable, say X_{out}, should not be the variable of the totals X_{tot}. In our simulations we choose to leave out the variable with the most missing items. Now the equality restriction reduces to an inequality. If we draw y so that it satisfies this new inequality, we can plug it in without violating the original restriction. After we updated y, we can calculate the values of X_{out} from the original equality.

This method is sound because if we suppose Y and the total $Z = Y + X_{out}$ are normally distributed, the imputed value of X_{out} is also normal. Moreover, according to theorem 2.3.1 this holds for truncated normal distributions too. Since the two regions in $\mathbb{R}^2$

$$\{(x, y)|x \geq 0, y \geq 0, x + y = z\} \text{ and } \{(x, y)|0 \leq x \leq y, x + y = z\}$$

are identical, we conclude that the distributions

$$P(Y|X_{out} \geq 0, Y \geq 0, X_{out} + Y = z) \text{ and } P(Y|0 \leq Y \leq z, X_{out} + Y = z)$$

are the same. Here we see that we cannot leave out more than one variable. With two or more variables left out, the distributions

$$\begin{aligned} P(Y|X_{out_1}, \ldots, X_{out_m} \geq 0, Y + \sum_j X_{out_j} = z) \\ \neq P(Y|0 \leq Y \leq z, Y + \sum_j X_{out_j} = z) \end{aligned}$$

are no longer the same. In the second one, some of the variables left out can be negative, while that is impossible in the first one.

To find inferences for the missing values of y, we apply the Bayesian regression model described in section 3.1.2. Draw the parameters β and σ^2 once. These parameters completely fix the distribution of y, so for each entry that was missing in the simulation dataset we can draw a value from that distribution. We repeat drawing until we get a value that satisfies the inequality. If that takes too long, we uniformly draw a value in the allowed interval between zero and the sum of all missing items in the record, and count it as a *misfit*. Note that this is not a very accurate approximation of a draw from a truncated normal distribution, but since a misfit is an indication that the normal model is not appropriate anyway, we do not want to increase the computer time with importance sampling.

After we insert the column X_{out} again we had left out earlier and calculate the values from the equality restriction, we follow the same procedure for all the other variables with missing items in the dataset. We repeat the whole iteration a fixed amount of times.

For the sequential regression method too, there are two ways to make sure an inputed value comes in the appropriate interval. The first is a variant which draws the error e_i from a truncated distribution, and the second is a variant,

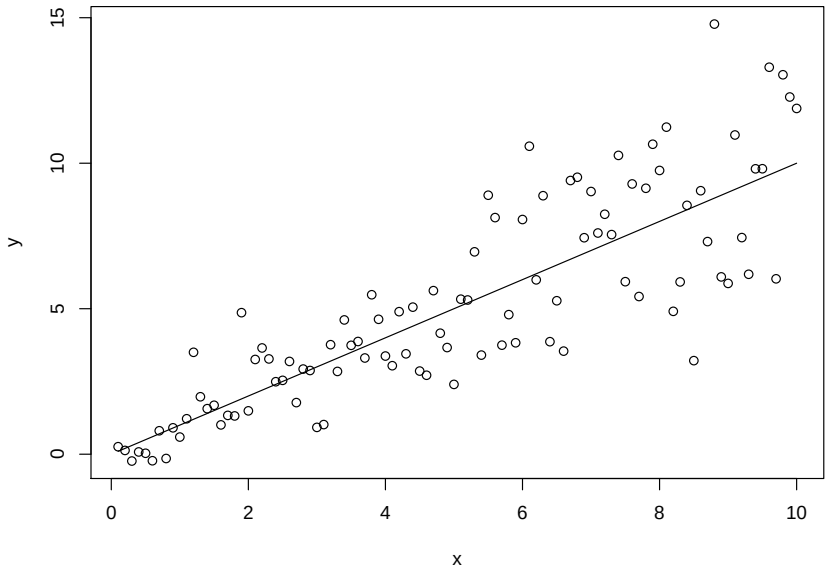


Figure 3.3: In economical data, data points corresponding to larger values values often get more dispersed.

which calculates the expectation $E(e_i)$ of he error term. Also, we can choose to draw from distributions estimated by a maximum likelihood method instead of a Bayesian posterior distribution. In total, we now have four variants of the sequential regression method.

3.4 Ratio imputation

In some economical data, the model assumption that the error terms e have constant variance is not appropriate. Larger data values are often more dispersed than smaller values, as illustrated in the picture below. Suppose for instance, a variable x indicates the number of items sold, and another variable, say y denotes the revenue. If we assume the prices of the item are i.i.d. random variables with a variance σ^2, then the variance of the revenue, which is the sum of the prices, is $\sigma^2 x$.

In cases like this, it is appropriate to use the following model

$$y = x\beta + e \tag{3.22}$$

where x and y are n-vectors, β a k-vector, and e a vector of n dependently distributed error terms, but now with variance proportional to x:

$$E(e_i) = 0 \text{ and } \text{var}(e_i) = \sigma^2 x_i \tag{3.23}$$

with $\sigma^2 > 0$.

In the ordinary regression models described in chapter (3.1) and (3.2) we used a method to find an estimate $\hat{\beta}$ or a posterior distribution for β treating all observations equally. However, such an approach is unsuitable for model (3.22). Since for large values of x the variance of e and thereby the variance of y becomes larger and larger, we may expect that larger values of y will have far more influence on the regression estimator $\hat{\beta}$ than smaller values. But because these smaller values have a smaller variance, these are more reliable observations. That is why we want to use a weighted average of the squared residuals to estimate β in this model

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} \frac{(y_i - x_i\beta)^2}{x_i} \tag{3.24}$$

taking the derivative with respect to β gives

$$\begin{aligned} \frac{\partial}{\partial\beta} \sum_{i}^{n} \frac{(y_i - \beta x_i)^2}{x_i} &= \sum_{i}^{n} \frac{-2x_i(y_i - x_i\beta)}{x_i} \\ &= -2\sum_{i}^{n}(y_i - x_i\beta) \end{aligned}$$

which is zero if $\sum_i^n y_i = \beta \sum_i^n x_i$, so that is at $\hat{\beta} = \sum_i^n y_i / \sum_i^n x_i$ and the estimator becomes the ratio of the sample means

$$\hat{\beta} = \bar{y}/\bar{x}. \tag{3.25}$$

We call $\hat{\beta}$ the *ratio estimator* of β.

Now let us assume the error term e is normally distributed. For the ordinary linear regression model, i.e. with $\text{Cov}(e) = \sigma^2 I$, we know the maximum likelihood estimator is $\hat{\beta} = (X^T X)^{-1} X^t y$. But in model 3.22, not all error terms have the same variance, so the Gauss-Markov conditions are not satisfied. Setting $\Sigma = xI$ We can transform model (3.22) into an ordinary form using multiplication by $\Sigma^{-1/2}$

$$\Sigma^{-1/2} y = \Sigma^{-1/2} x\beta + \Sigma^{-1/2} e$$

to get a new model

$$y^* = x^*\beta + e^*. \tag{3.26}$$

Looking at $\text{Cov}(e^*) = \text{Cov}(\Sigma^{-1/2}e) = \Sigma^{-1/2}\text{Cov}(e)\Sigma^{-1/2} = \Sigma^{-1/2}\sigma^2\Sigma\Sigma^{-1/2} = \sigma^2 I$ we see that (3.26) is indeed an ordinary linear model.

In the next calculation, we use the fact that $\Sigma^{-1} = (xI)^{-1}$ is the $n \times n$ diagonal matrix with fractions $1/x_i$ on the ith diagonal entry and zeros elsewhere, so

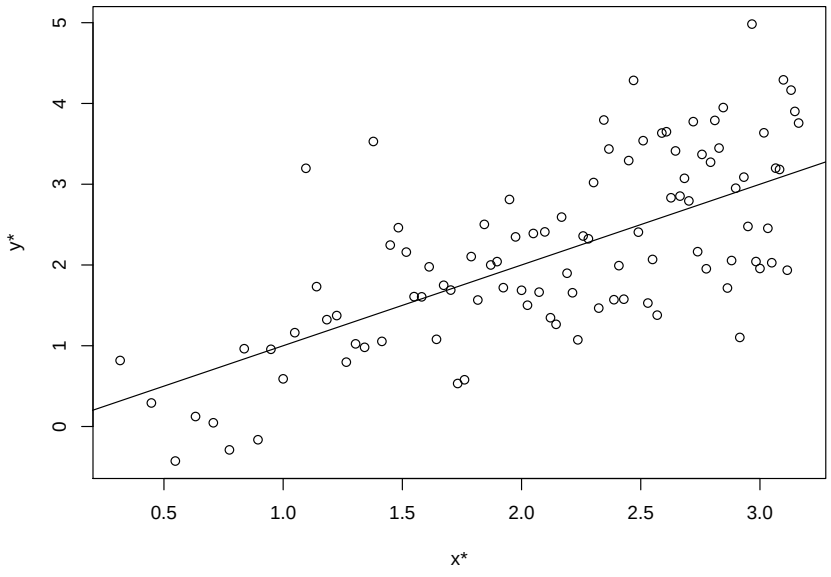


Figure 3.4: The same data, but now transformed.

$x^t\Sigma^{-1}$ is a row vector of length n with all entries 1, which we denote by $\mathbf{1}$.

$$\begin{aligned}\hat{\beta} &= (x^{*T}x^*)^{-1}x^{*T}y^* = \left((\Sigma^{-1/2}x)^T\Sigma^{-1/2}x\right)^{-1}(\Sigma^{-1/2}x)^T\Sigma^{-1/2}y \\ &= (x^T(\Sigma^{-1/2})^T\Sigma^{-1/2}x)^{-1}x^T(\Sigma^{-1/2})^T\Sigma^{-1/2}y \\ &= (x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}y = (\mathbf{1}x)^{-1}\mathbf{1}y \\ &= \left(\sum_i x_i\right)^{-1}\sum_i y_i.\end{aligned}$$

Hence, the maximum likelihood estimator

$$\hat{\beta} = \bar{y}/\bar{x} \tag{3.27}$$

is indeed the ratio estimator we found in (3.25).

Figure 3.4 shows the same data as figure 3.3, but after application of the linear transformation $\Sigma^{-\frac{1}{2}}$. Note that $\Sigma^{-\frac{1}{2}}$ is the $n \times n$ diagonal matrix with $1/\sqrt{x_i}$ as the i-th diagonal element. We clearly see that after transformation the data are more equally spread around the regression line.

As with all regression methods, the ratio imputation methods comes in two variants: one that fills in the expected value for an imputation, and one that uses a draw. The fist one just calculates the ratio estimator (3.25) and fills in $\hat{\beta}x_i$. This we call the deterministic ratio method. The second option is to fill in $\hat{\beta}x_i + e_i$, where e_i is drawn form a $\mathcal{N}(0, \hat{\sigma}^2 x_i)$ distribution. The parameter σ^2

should be estimated with

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{x_i} \tag{3.28}$$

and this gives the stochastic variant of the ratio imputation method.

Both variants of the ratio method can be adjusted so that they can handle restrictions, like we did in section 3.2. That means that we fill in the expectation of a truncated normal distribution, or a draw from it. We choose the permitted region so that the restrictions hold. The flaw of this approach is that in that case the ratio estimator (3.25) is no longer the maximum likelihood estimator. As we will see in the simulation results, the ratio method is not performing very well on the dataset with restrictions. Still, it is one of the most widespread imputation methods, and also much applied to data sets with restrictions at Statistics Netherlands, although in a different variant than we discuss here.

In our implementations of the ratio imputation method, we always use the total X_{tot} variable as regressor x. If for items missing a subtotal was available, we used that instead. This would seem a logical way to use information available, and it also gives a realistic method, since in practice it does happen often that a total is filled in in a survey but not completely specified.

We still have to handle the problem that at some point a value y_i that is missing may be completely determined by the balance restriction. So as we go through the row of the dataset, we constantly check if that is the case. If not, we impute the value using one of the ratio methods. If so, we fill in what it should be and are done.

As we saw, there is no straight forward way to make the ratio method respect linear equality restrictions. For each variable we calculate a permitted region that does not prevent the restriction from holding. Bu at some point, if there is only one missing item left in a record, we have no choice and the value we have to impute is completely determined by the restriction. What makes the problem worse, is that it is somewhat arbitrary which variable that leftover will be. Sequential regression at least deals with that last point. It can change every variable by leaving out another variable X_{out}. Which variable to choose as x_{out} may be arbitrary too, but at least this variable can be changed in the next iteration round.

Chapter 4

Simulation studies

In our simulations, we take a completely observed part of a dataset and remove some of the data entries and consider them missing. We then apply each of our imputation methods. Since we know what the actual values were for the missing part, we can then analyze the behaviour of these methods.

The idea behind this simulation strategy shows some similarities to the bootstrap method for estimating the variance of an estimator. However, the difference is that we repeatedly create missing items in the observed part of the dataset, whereas the bootstrap method would take a dataset with missing data, and repeatedly draw records without replacement from it. See appendix B for a more detailed description of the bootstrap method in the presence of missing data.

4.1 Dataset

The data on which we perform our studies are provided by Statistics Netherlands and consists of 15 continuous variables $X_1, \ldots, X_{15}$, one variable containing reference numbers for the records and four categorial variables. The data are part of a retail trade survey, which has 108 variables and 800 records in total. See appendix C for a description of the data variables. Of these 800 records, 684 have no missing items in the 15 variables we will deal with. The restrictions that must hold for this dataset are $X_1 + X_2 = X_3$ and $X_3 + X_4 + \ldots + X_{14} = X_{15}$ and $X_i \geq 0$ for $i = 1, \ldots 15$. To make the restrictions easier to handle, we leave X_3 out of our simulation and use the alternative restriction $X_1 + X_2 + X_4 + X_5 + \ldots X_{14} = X_{15}$. This will be our simulation dataset. After the imputation process is done, we can easily calculate X_3 from X_1 and X_2. For notational convenience, X_{15} will be written as X_{tot}.

Apart from these 15, we will consider three more variables: an identification number, a size class, which is a categorial variable indicating the size of the business and a categorial variable indicating the branch of trade, the so called SBI code. These three variables are observed for all businesses in the original dataset.

4.2 Missing data mechanism

The easiest missing data mechanism to implement is just to set each value missing with a certain probability, or alternatively with different probabilities for each variable. This mechanism is MAR, because it does not depend on the unobserved part of the data, and it is even MCAR since it is also independent of the observations.

A more sophisticated missing data mechanism developed at Statistics Netherlands is so called *hot deck amputation* [*sic*]. At the start, the dataset is divided into two parts. In the first part all records are stored that have no items missing in the variables we are interested in. The second part consists of the records that do have missing items. The missing mechanism now does the following: for each record in the first part, called receiver, it takes a record in the second part that is the nearest with respect to a distance function. This record is called the donor. Once the donor is chosen, the mechanism looks which items of the donor are missing. In the receiver record, the items for which the corresponding item in the donor are missing, are also set to missing. That way, the receiver takes over the missing pattern from the donor. The receiver records then form the dataset to simulate with.

The best thing about this mechanism is that it respects patterns in the nonresponse. It seems likely that in real situations, the probability that an item misses is not independent of the rest of the missing pattern in the record. In that sense the method is more realistic. Note that this mechanism is MAR if the real missing items in the second part of the dataset were MAR, *and* the distance function does not involve variables in which missingness is created.

We suggest a slight modification of the amputation method. In our simulation, we do not use a distance function. Instead, we stratify the whole dataset according to an auxiliary variable. Next, for each record in the dataset for which all the variables are observed, the receiver records on which we will perform our simulation studies, we randomly draw with replacement a donor from the same stratum as the receiver was in. Then, like in the amputation method, we give the receiver the missing pattern of the donor. The difference is that we can also choose donors that do not have any missing items.

By doing this, we take two things into account:

- we take over patterns in the missing data in the original dataset as far as these are related to the variable business size;
- for each variable, the expectation of the fraction of items that is missing is the same as it was in the original dataset.

In our simulation datasets, we always let the total X_{15} be observed. This is realistic, because in practice it often happens that someone does not know the answers to some specific questions in a survey, but nevertheless fills in the total. Also, we do not leave out the subtotal X_3.

To make the missing mechanism more realistic, we stratify the dataset with respect to the variable size class and apply the hot deck amputation to the

strata separately. That means that we take over different missing data patterns for different company sizes.

The dataset we use contains no missing items that can be deductively imputed, because this has already been done. The advantage of this is that the simulation datasets also do not contain missing values that can be deduced from the observations. Therefore, this gives a fair dataset to test imputation methods on.

4.3 Evaluation criteria

In this section we mention a few criteria which may be used to asses different imputation methods, if the true values are known. The criteria are mainly inspired on [Chambers 2001]. Let $\tilde{x}_{ij}$ denote the imputed value corresponding to the true value x_{ij}, and $mis_j = \{i | x_{ij} \text{ is missing}\}$ the index set of missing items in column j.

The first option is to measure the ability of an imputation method to reproduce individual values. We use the average of the absolute difference between the true and the imputed values

$$L_1(j) = \frac{1}{\#mis_j} \sum_{i \in mis_j} |x_{ij} - \tilde{x}_{ij}| \tag{4.1}$$

and the average squared differences

$$L_2(j) = \frac{1}{\#mis_j} \sum_{i \in mis_j} (x_{ij} - \tilde{x}_{ij})^2. \tag{4.2}$$

Secondly, we might be interested in the average of the differences for each individual value in relation to the true values

$$R_1(j) = \frac{1}{\#mis_j} \sum_{i \in mis_j} \frac{|x_{ij} - \tilde{x}_{ij}|}{x_{ij}}, \tag{4.3}$$

but this quantity is not defined if there is some i for which $x_{ij} = 0$. Therefore, in our implementation, in cases where $x_{ij} = 0$ we check if also $\tilde{x}_{ij} = 0$. If that is the case, we set its contribution to the sum to nil, because that means the correct value was imputed. If $\tilde{x}_{ij}$ is not zero, we still cannot define R_1. In our simulation dataset, a lot of values x_{ij} are zero. Moreover, the methods we evaluate most of the times do not impute $\tilde{x}_{ij} = 0$, so the R_2 criterion is undefined very often. Therefore we do not include this criterion in the numerical results section.

We would like to use a third criterion to asses the ability of an imputation method to reconstruct means and totals of variables. We could use the difference between the mean of the true values and the values after imputation. A drawback of that idea is that it depends very much on the number of items missing. It would be a good criterion for the quality of the dataset after imputation, but

less suitable for the evaluation of the imputation process itself. An alternative is

$$R_2(j) = \left| \sum_{i \in mis_j} (x_{ij} - \tilde{x}_{ij}) / \sum_{i \in mis_j} x_{ij} \right| \tag{4.4}$$

which takes the relative difference in mean over the imputed values only. This criterion can handle situations where some x_{ij} are zero. If all are zero, the denominator of (4.4) is zero and R_2 is undefined.

Fourthly, we take the total residual sum of squares (RSS). It looks similar to (4.1), but does not take an average.

$$RSS(j) = \sum_{i \in mis_j} (x_{ij} - \tilde{x}_{ij})^2 \tag{4.5}$$

Finally, we would also like to look at the standard deviation of the distribution of the imputed data and compare that with the one of the original data. Therefore, for each variable j, we compute the standard deviation of both the original and the imputed data and look at the difference between the two. Because these values as such do not say a lot, we could also give the differences as a percentage of the original standard deviation.

Chapter 5

Results

In this chapter we will present the results of our simulation studies. We apply the scheme described in chapter 4 and calculate the criteria from section 4.3 for each method separately. The methods we investigate are:

1 the Bayesian sequential regression method that draws from the posterior;

2 the Bayesian sequential regression method that uses the expectation of the posterior;

3 the classical sequential regression method that uses the expectation of the estimated distribution;

4 the classical sequential regression method that draws from the estimated distribution;

5 the deterministic ratio method;

6 the stochastic ratio method;

7 the normal proportional variance method with ad hoc parameter estimation;

8 the normal proportional variance method with the EM algorithm to estimate the parameters;

9 the imputation method fitting a normal distribution using the likelihood of a left truncated distribution to estimate the parameters;

10 the imputation method fitting an exponential distribution;

11 the Poisson proportional variance method using the EM algorithm to estimate the parameters.

The numbers in the figures refer to the numbers of the methods as listed above.

Since we use the same simulation datasets for all methods, we can for each variable in each dataset compare the results for the different criteria and decide which method performed best. By doing so, we can make a bar plot for each variable which shows on how many simulation datasets each method was best performing. In figure 5.1 for example, we see that method number 8,

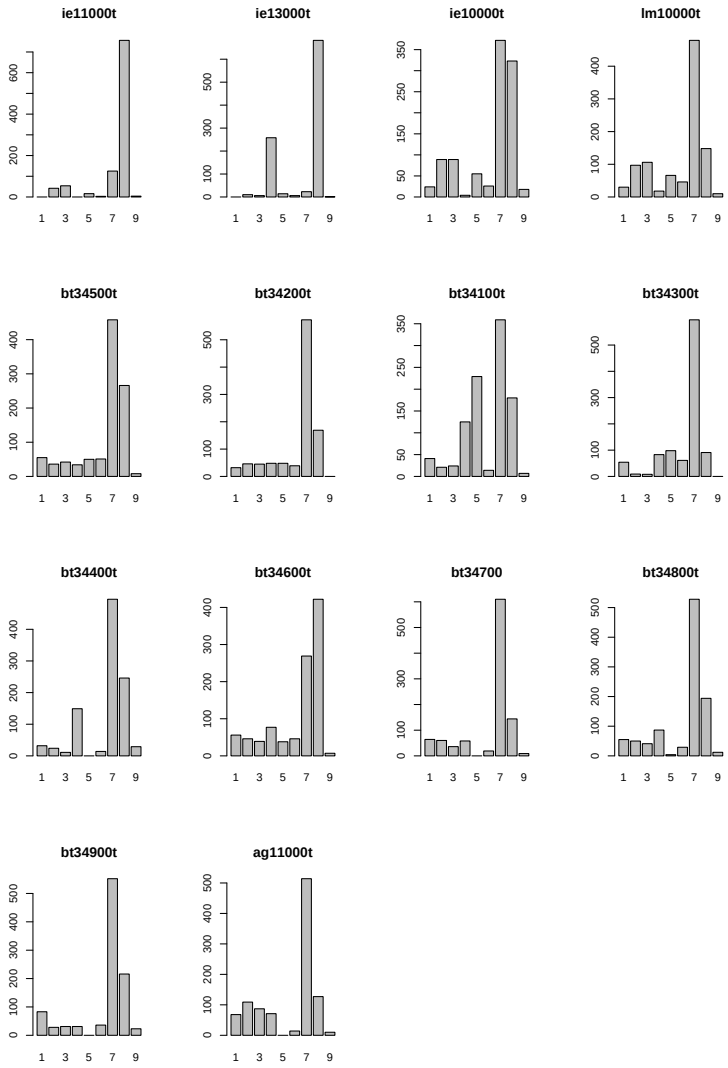
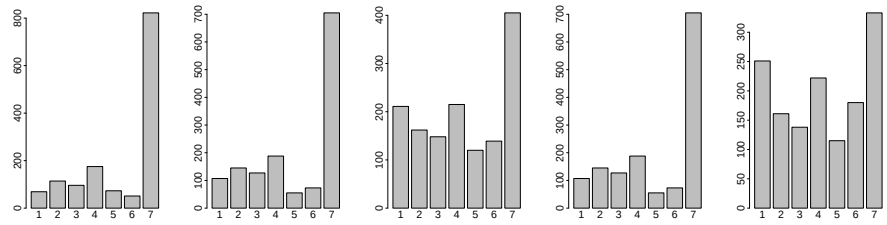


Figure 5.1: Bar plots comparing methods 1 up to 9 on the L1 criterion for each variable separately. The height of the bars indicate on how many simulation datasets the methods we best performing.



(a) L1 (b) L2 (c) R2 (d) RSS (e) variance

Figure 5.2: Bar plots comparing methods 1 up to 7 on the different criteria. The height of the bars indicates how many times the corresponding methods we best performing over all datasets and variables for the corresponding criterion.

the normal proportional variance method with the EM algorithm to estimate the parameters, is most often the best for variable `ie11000t`, but for variable `lm100000t`, method number 7, the same model but with the ad hoc estimators, is most often performing best. To get a more clearly arranged view, we can add the counts over the different variables together for each method. that gives us figures like 5.2.

A glance at figure 5.2 suggests that the sequential regression methods and the ratio method are performing less than the proportional variance method. However, it may still be the case that the methods 1 up to 6 are each better than number 7. They could be competing amongst each other in the cases where number 7 is not performing well. Therefore, apart from a first indication of the performance of these methods, these pictures have very little significance. Instead, we should compare the methods pairwise.

In the next section, we will compare the sequential regression methods and the ratio methods individually to the proportional variance method. After that, we will compare the methods that based on models assuming independence.

5.1 Methods using linear regression

Sequential regression imputation

To apply a sequential regression method to our data, we first have to fill in the missing entries in the first step. We take the column sums of the records in the simulation dataset for which all 14 variables are observed. We calculate the values $s_1, \ldots, s_{15}$ from equation 3.20 but not s_3 since we left out X_3. Then we fill in the missing values using equation 3.21 with $X_{tot} = X_{15}$.

In every next iteration, we apply the regression iteration of section 3.3. If drawing from the posterior or estimated distribution of y does not give a in the permitted interval after we tried a 100 times, give a warning, count it as a *misfit* and pick a value uniformly random from the interval. The misfits reported in table 5.1 for the sequential regression methods are the counts from the last iteration round of the methods.

	Bayesian sequential regression method	classical sequential regression method	ratio method
ie11000t	0.011	0.029	0
ie13000t	0.016	0.001	0
lm10000t	0.514	0.072	0.004
bt34500t	0.031	0.003	0.010
bt34200t	0.010	0.000	0.003
bt34100t	0.022	0.005	0.013
bt34300t	0.191	0.035	0.045
bt34400t	0.025	0.007	0.020
bt34600t	0.034	0.011	0.054
bt34700t	0.026	0.004	0.034
bt34800t	0.165	0.029	0.070
bt34900t	0.012	0.000	0.014

Table 5.1: The number of misfits of the Bayesian sequential regression method, the classical one and the ratio method, divided by the number of missing values for each variable separately. The numbers shown are averages over all simulation datasets.

We repeat this iteration 20 times. That does not seem much, but increasing the number of iterations can increase the computer time a lot. Furthermore, Raghunatan [Raghunatan 2002] claims 10 iterations should be sufficient for most applications.

Bayesian sequential regression method using draws from the posterior

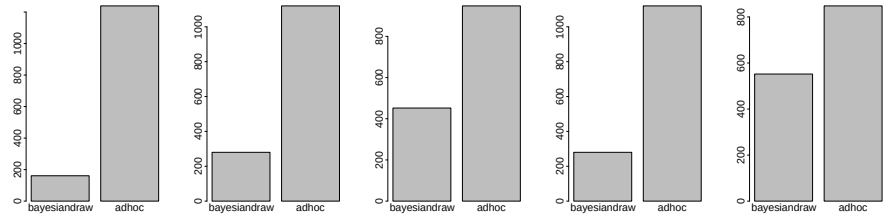
For this sequential regression method, we use the posterior distributions from section 3.1.2 and truncate them to the appropriate interval. The imputed values are draws from these truncated normal distributions.

During our simulations we counted the misfits and came to surprisingly high numbers, see table 5.1 The high fractions for some variables indicate that the truncated normal distribution is not very suitable for the data. For variable `lm10000t` for instance, we count a misfit in on average more than 50 percent of the cases .

In figure 5.3, we compare the Bayesian sequential regression method using draws from the posterior with the proportional variance method. We see that on all criteria, it is performing less than the deterministic proportional variance method using the ad hoc estimators.

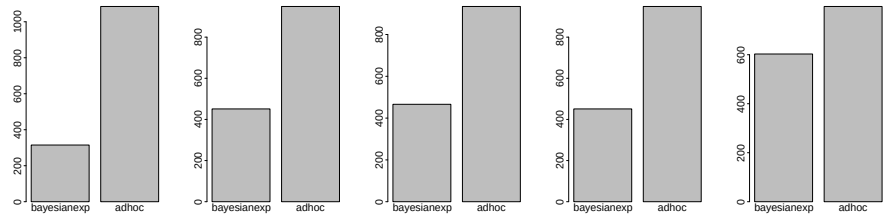
Bayesian sequential regression method using the expectation of the posterior

For this sequential regression method, we use the posterior distributions from section 3.1.2 and truncate them to the appropriate interval. However, now the imputed values are not draws but the expected values of these truncated normal distributions.



(a) L1 (b) L2 (c) R2 (d) RSS (e) variance

Figure 5.3: Bar plots comparing the sequential regression method that imputes draws from a posterior with the deterministic proportional variance method.



(a) L1 (b) L2 (c) R2 (d) RSS (e) variance

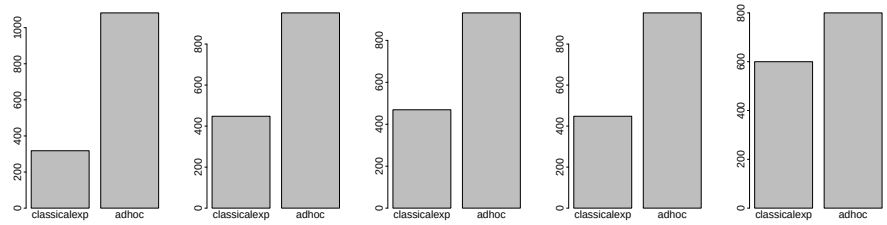
Figure 5.4: Bar plots comparing the sequential regression method that imputes the expectation of a posterior with the deterministic proportional variance method.

In figure 5.4, we compare the this method with the proportional variance method. We see that on all criteria, it is performing less than the deterministic proportional variance method using the ad hoc estimators.

Classical sequential regression method using the expectation of the estimated distribution

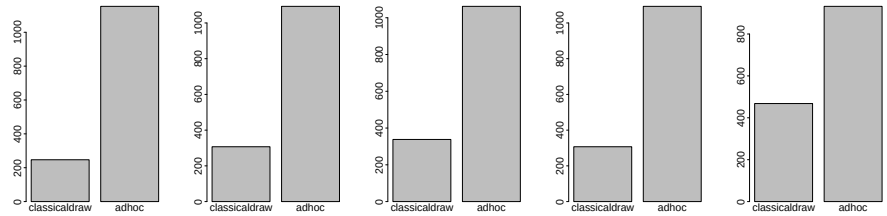
For the classical sequential regression method, we use estimators from section 3.1.1. We do not use the maximum likelihood estimators for β and σ for from section 3.2.1 because it takes too much computer time to calculate these estimates in every iteration round for every variable. So our implementations of both the classical and the Bayesian methods do not use the bounds on the variables to generate the parameters. The imputed values are again the expected values of the truncated versions of the estimated normal distributions.

In figure 5.5, we compare the this method with the proportional variance method using the ad hoc estimators. We see that on all criteria, it is performing less than the deterministic proportional variance method.



(a) L1 (b) L2 (c) R2 (d) RSS (e) variance

Figure 5.5: Bar plots comparing the classical regression method that imputes the expectation of the estimated distribution with the deterministic proportional variance method.



(a) L1 (b) L2 (c) R2 (d) RSS (e) variance

Figure 5.6: Bar plots comparing the classical regression method that imputes draws from the estimated distribution with the deterministic proportional variance method.

Classical sequential regression method using draws from the estimated distribution

For this method, we use the same procedure as for the previous method, but now the imputed values are draws from the truncated versions of the estimated normal distributions.

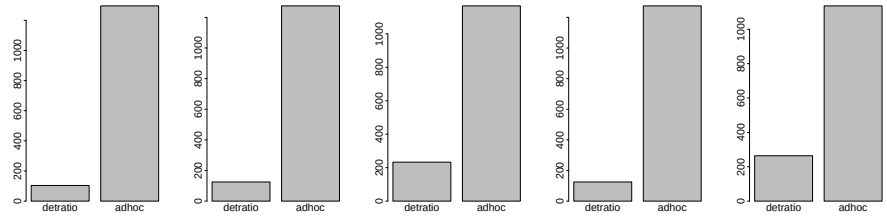
Here too we counted the misfits. In table 5.1, we see the average over the simulation datasets of the number of misfits counted per variable. We count much less misfits for the classical regression than for the the Bayesian version.

In figure 5.6, we compare the this method with the proportional variance method using the ad hoc estimators. We see that on every criterion, it is performing less than the deterministic proportional variance method.

Ratio methods

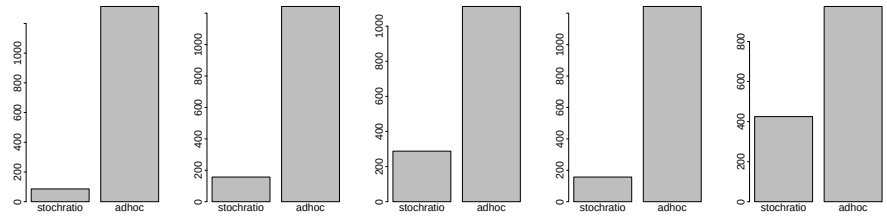
The next method we test is the ratio method, as described in section 3.4. We implemented the deterministic and the stochastic variant, and compare these with the proportional variance method that uses the ad hoc estimators too.

For the stochastic variant, we counted the number of misfits just like with the



(a) L1 (b) L2 (c) R2 (d) RSS (e) variance

Figure 5.7: Bar plots comparing the deterministic ratio method that with the deterministic proportional variance method.



(a) L1 (b) L2 (c) R2 (d) RSS (e) variance

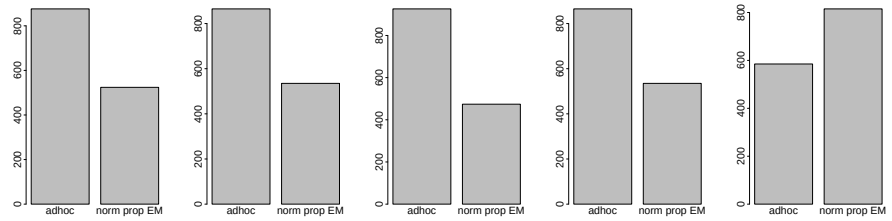
Figure 5.8: Bar plots comparing the stochastic ratio method that with the deterministic proportional variance method.

sequential regression methods, see table 5.1 for the results. Note that for the first two variables, `ie110000t` and `ie130000t`, no misfits were counted at all.

We see that both versions of the ratio method are doing worse than the proportional variance method.

5.2 Methods based on models assuming independence

In the previous section we saw that the sequential regression and ratio methods are not working very well. Therefore we now focus on the methods that are based on models which assume independence between variables. First we look at the proportional variance method, to see that the normal model with proportional variances does not fit the data. After that, we compare the four remaining methods: the proportional variance method with ad hoc parameter estimation, the imputation method fitting a normal distribution, the imputation method fitting an exponential distribution and the imputation method that uses the EM algorithm to fit a Poisson distribution.



(a) L1 (b) L2 (c) R2 (d) RSS (e) variance

Figure 5.9: Bar plots comparing the proportional variance method with the version that uses the EM algorithm to find the parameters of the normal proportional variance model for every imputed variable in each simulation dataset. The height of the bar indicates the number of variables for which the method was performing better than the other.

The normal proportional variance model

In figure 5.9, we see that the method uses the EM algorithm to find the parameters of the normal proportional variance model is performing worse than the proportional variance method that uses the ad hoc parameters. A possible explanation is that the normal proportional variance model is inappropriate. The method is doing all right because it uses the given sum of each record, and is therefore not too much handicapped by the model. Since the EM variant is trying to get parameter estimates that fit better to the inappropriate model, it is works less well than the ad hoc variant.

We can also test if the normal proportional variance model fits the data provided by Statistics Netherlands. We test the null hypothesis

$$H_0 : \text{the data } X_{ij} \text{ follow a } \mathcal{N}(\mu_j, \alpha\mu_j) \text{ distribution}$$

with the same α for all j, against the alternative

$$H_A : \text{the data } X_{ij} \text{ follow a } \mathcal{N}(\mu_j, \sigma_j) \text{ distribution.}$$

We use a likelihood ratio test statistic

$$L_\theta(X) = \frac{max_{\theta \in H_0} \log \ell_\theta(X)}{max_{\theta \in H_A} \log \ell_\theta(X)}$$

so that we know that asymptotically,

$$2L_\theta(X) \sim \chi^2_{k-1}. \tag{5.1}$$

Here the degrees of freedom of the chi square distribution equals the difference in the number of parameters between the models in H_0 and H_A. Instead of a parameter σ_j for each column, the model of the null hypothesis only has the parameter α. So the number of degrees of freedom equals the number of columns of the data set minus one.

We use a computer simulation to find a c such that $P_{H_0}(L > c) = 0.05$. The test tells us that we reject the null hypotheses. Using the asymptotic distribution

(5.1), we find a p-value of $9.682852e-05$, so we have indeed reason to believe that the model is inappropriate for this dataset.

To get an idea of the difference between the estimators for α that both methods use, we can apply the estimation process to a simulation dataset with missing items. Since α is a measure for the fraction of the variance divided by the mean, we compare the results in a graph where we plot the variance of each variable in the completely observed dataset against its mean, see figure 5.10(a). The solid and dashed lines are lines through the origin that have the estimated α as slope coefficient. The dotted line is a regression line on the points. Note that there is a large cluster of points in the bottom left corner, and only two points outside the cluster. In figure 5.10(b) we zoom into the cluster, and redraw the regression line, now using only points in the cluster.

We see in figure 5.10(a) that the EM estimator for α is closer to the slope of the regression line than the ad hoc estimator. That is strange, because we saw the ad hoc estimator did better. However, it is the other way around in the cluster shown in figure 5.10(b). Here, we see that the line corresponding to the ad hoc estimator, is closer to the regression line than the line of the EM estimator. The regression line in figure 5.10(b) was fitted only to the variables in the cluster. Since most variables lie in the cluster, that could explain why the ad hoc estimator did do better nevertheless.

The other methods based on models assuming independence

The four other methods that use models assuming independence are

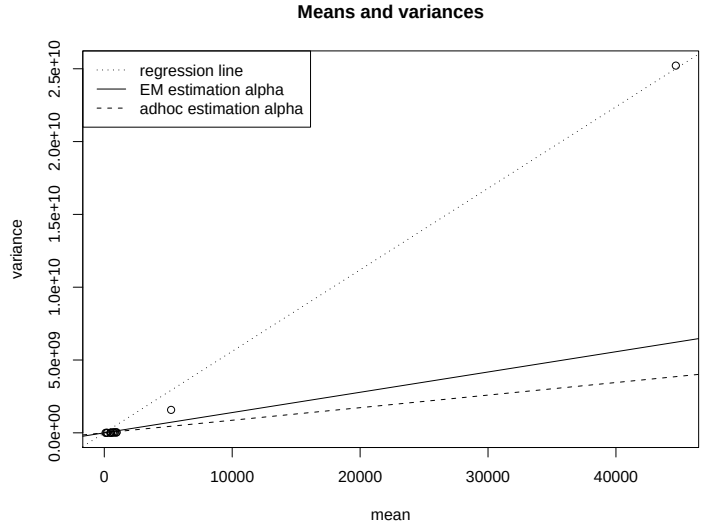
- the proportional variance method with ad hoc parameter estimator (ad hoc);
- the imputation method fitting a normal distribution. We use the likelihood of a left-truncated normal distribution, see section 2.3 to estimate the parameters. (normal);
- the imputation method fitting an exponential distribution (exp)
- the Poisson proportional variance method the EM algorithm to find parameter estimates (poisson)

In the next figures these methods are referred to by the keywords in brackets. We compare these four methods pairwise for each criterion separately.

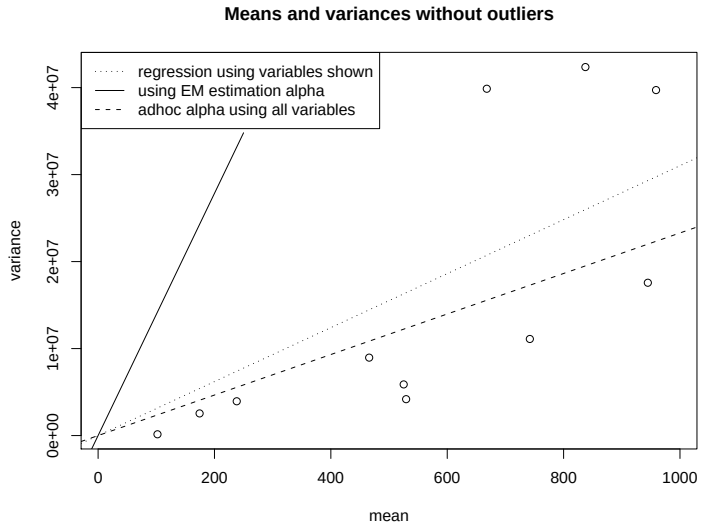
In figure 5.11, we see that we can order the methods according to their performance on the L1 criterion; First comes the proportional model with ad hoc estimators, then the Poisson model, after that the exponential model and finally the normal model.

For the L2 criterion, we see almost the same pattern in figure 5.12. Here the Poisson model is performing best, the the proportional model with ad hoc estimators comes second and the exponential model is slightly better than the normal model, which is last again.

In figure 5.13 we get the same order of succession for the R2 criterion, but the differences between most methods are larger. The exponential model is just

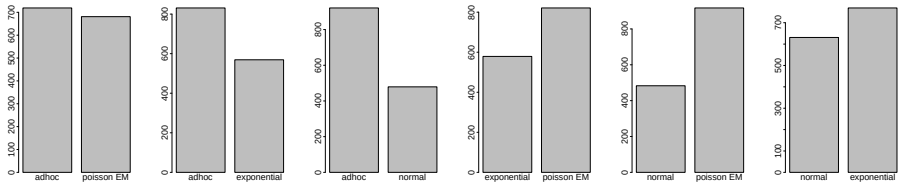


(a) The means of the fully observed variables plotted against their variances. The total and the subtotal variable have been omitted



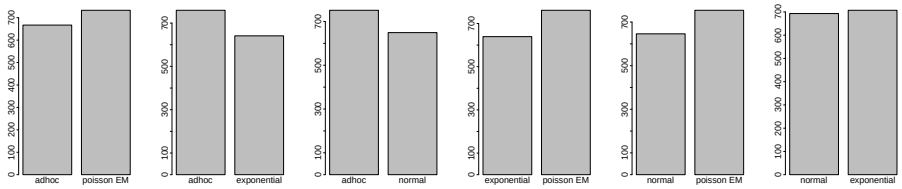
(b) The same picture but now zoomed into the cluster of points on the left.

Figure 5.10: The means of the fully observed variables plotted against their variances.



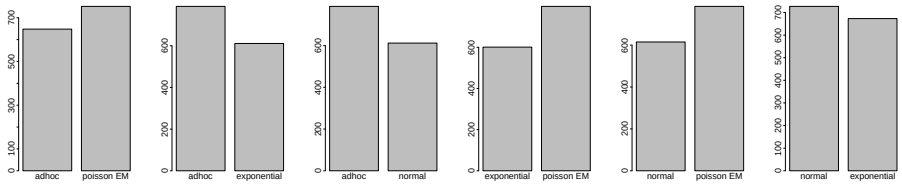
(a) (b) (c) (d) (e) (f)

Figure 5.11: Bar plots comparing the methods pairwise on the L1 criterion for every imputed variable in each simulation dataset. The height of the bar indicates the number of variable for which the method was performing better than the other.



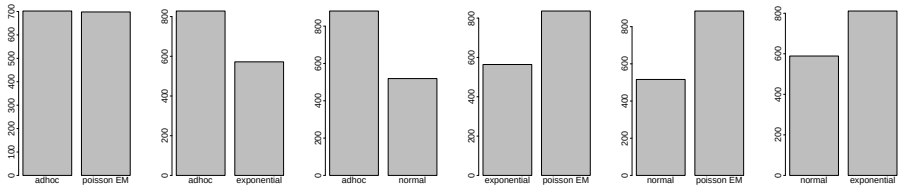
(a) (b) (c) (d) (e) (f)

Figure 5.12: Bar plots comparing the methods pairwise on the L2 criterion for every imputed variable in each simulation dataset. The height of the bar indicates the number of variable for which the method was performing better than the other.



(a) (b) (c) (d) (e) (f)

Figure 5.13: Bar plots comparing the methods pairwise on the R2 criterion for every imputed variable in each simulation dataset. The height of the bar indicates the number of variable for which the method was performing better than the other. There were three datasets out of the 100 where all four methods gave the same R2 value for the variable `ie130000t`, but that does not significantly change the picture.



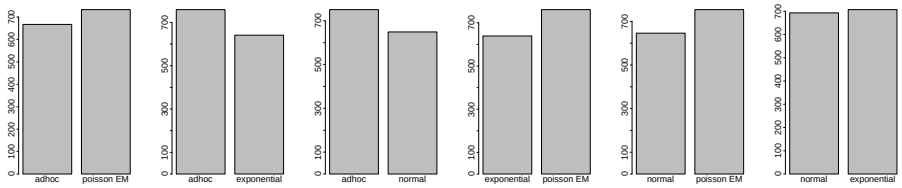
(a) (b) (c) (d) (e) (f)

Figure 5.14: Bar plots comparing the methods pairwise on the difference in variance for every variable in each simulation dataset. The height of the bar indicates the number of variables for which the method was performing better than the other.

lagging behind the Poisson model is best and the the proportional model with ad hoc estimators. For the R2 criterion there were three datasets out of the 100 where all four methods gave the same value for the variable `ie130000t`. Since the differences in this picture are clearly all larger than three, that does not significantly change the picture.

For the difference in variance, the Poisson model and the the proportional model with ad hoc estimators are coming very close to each other. Behind those two, the exponential model clearly is better than the normal model.

The last criterion we investigated was the total residual sum of squares (RSS). Here we see an already familiar pattern: the Poisson method comes first, than the proportional model with ad hoc estimators, next the exponential model and just after that the normal model.



(a) (b) (c) (d) (e) (f)

Figure 5.15: Bar plots comparing the methods pairwise on the residual sum of squares (RSS) for every variable in each simulation dataset. The height of the bar indicates the number of variables for which the method was performing better than the other.

Chapter 6

Conclusions

6.1 Discussion and suggestions for future research

In the previous chapter, we gathered the results from our simulations. We saw that the sequential regression methods did not do a good job. There are several possible explanations for that. The most obvious one is that the truncated normal model used in these methods does not fit to the data. This is backed by the numbers in table 5.1 where we counted the amount of imputations for which the mean of the estimated normal distribution was far outside the permitted truncation region.

We also did not check if the regression sequence converged after 20 iteration rounds. If we would like to check this ,we could for instance take one missing item and follow it during the iteration rounds of the sequential method. This would give a sequence of imputed values for the single missing item. However, we should not look for convergence of the imputations themselves, but at the distributions they come from. A possible strategy would be to cut the sequence of imputations into pieces of a certain length, and calculate the mean and variance of these pieces. If these numbers are coming closer to each other along the sequence, that could be an indication that the distributions are converging.

Another interesting experiment could be to vary the regression order of the variables in each iteration. We did not use any specific order, but just took the ordering of the variables in the data matrix. It could make sense to update fist the variable with the least number of missing items, than the variable with the second least number and so on. On the other hand, if the order really matters, one can wonder if convergence of the sequential method makes any sense.

There might be opportunities to improve the sequential regression methods. We could use other regression models. We could use other truncated distributions, or for instance semi-continuous random variables that are zero with a positive probability and or truncated normally distributed otherwise. Alternatively, if we wish to stick to truncated normal distributions, we could try to improve the Bayesian analysis of the truncated normal regression model. We

used the Bayesian linear model to find the posterior and then truncated to the correct interval given by the restrictions, but that is only an approximation.

Apart from using other models, we could think about which variable X_{out} we leave out in each iteration. We always left out the variable which had the most missing items, because that leaves as much observed information in as possible. A disadvantage is that we often leave out the same variable, and only if we are busy imputing this variable with most missing items itself we use the information in it. There might be other strategies which could perhaps give better results, for instance leave out a variable randomly, or with some probability according to the number of missing items in a variable. Also, as mentioned before, we could improve the way we draw from the truncated normal distributions using importance sampling. Finally, we could improve the imputation process by using multiple imputation, see appendix A.

Currently, the ratio imputation method is the most commonly used method at Statistics Netherlands, although in a different version than we implemented. In our implementation we changed the method by drawing or filling in the expectation of a truncation of the estimated distribution. We saw that on our simulation data, our implementation leads to very poor imputations. We counted several misfits, see page 55. For some variables up to 7 percent of the imputations was a misfit. This could indicate that the ratio model is not appropriate for this kind of data with restrictions. Of course, our simulation dataset did not give any specific reason to use the ratio model, and our choice of taking the total variable `bt310000t` as regressor is somewhat arbitrary.

The best methods we investigated all use models that assume the items are independently distributed with a variance proportional to their mean. There are different models that have this property, and we looked at two of them. There could certainly be many more, and it would be worth to find and them and test them too. In our test, there was not much difference between the results if we estimated the parameters based on a normal model, a Poisson model or even used an ad hoc estimator.

At first sight it is a little bit surprising that an exponential model does not lead to better imputations. In the histograms of single variables the exponential distribution seems to fit the data much better, see figure 2.2 on page 31. Apparently it is not necessary to find a good model for the data themselves. Since we observe the sum of each row in the dataset, it is better to use models that lead to a good conditional distribution of the data given these sums.

We see that for all criteria, the results of the Poisson model and the proportional model with ad hoc estimators lie very close to each other, which is not surprising since we saw in section 2.5.2 that the two are related. But both are each better than the exponential and the normal model. This gives reason to believe that the Poisson model is actually the model that gives the best conditional distribution of the data given the observed row sums.

6.2 Conclusion

We advise Statistics Netherlands not to use sequential regression methods for imputation of economical micro-data without doing further research first. The ratio imputation method, now the most commonly used method, is also not working very well. Although Statistics Netherlands uses a version that does not take the restrictions into account and changes the data afterwards, it might be a good idea to examine this method more closely. The proportional variance method based on Hachemeister and Stanard's Poisson model worked best in our simulation studies. It would be interesting to test the hypothesis that the data are following a Poisson distribution using datasets from several time periods.

Acknowledgements

I would like to thank a number of people who have helped and encouraged me to write this thesis. My thesis adviser Erik van Zwet always gave valuable comments and suggestions and he never ceased to support me. My internship supervisors Caren Tempelman and Jeroen Pannekoek gave me the opportunity of working on this project at Statistics Netherlands. Caren's PhD thesis was a comprehensive introduction to the theory behind imputation methods. When Jeroen took over the task of supervising me, he was always patient to listen to me talking about my failures and successes.

I would also like to thank my other colleagues at the methodology department of Statistics Netherlands, especially Léander for helping me to stay motivated, Nino for being my roommate, Paul for his comments on my presentation and his humour during lunches, and Fannie and Mark for inviting me to some of the department's day trips and dinners, even after I already had left the department.

At the Mathematical Institute of the University of Leiden I would like to thank my friends Daniël, Martin and Bart for their support and understanding, and the teachers with whom I worked together or who taught me something during my studies.

Finally, I thank my parents for their support and efforts to understand what I was doing all the time, and Yvonne for her suggestions on writing this thesis in English.

Bibliography

[Anderson 1971] Anderson, T.W., 1971. *An introduction to multivariate statistical analysis*, John Wiley & Sons, New York. Second edition, 1984.

[Balazs 2005] Balázs, M., 2005. *Sum of independent exponentials*. Unpublished. Available at http://www.math.bme.hu/ balazs/sumexp.pdf

[Brockwell] Brockwell, P.J.; Davis, R.A. *Time series: theory and methods* second edition, Springer Verlag

[Chambers 2001] Chambers, R., 2001. *Evaluation Criteria for Statistical Editing and Imputation*, National Statistics Methodological Series No. 28, HMSO, 2001. Available at http://www.cs.york.ac.uk/euredit/

[De Waal 2003] De Waal, A.G., 2003. *Processing of erroneous and unsafe data*. PhD thesis Erasmus University Rotterdam

[Effron 1993] Efron, B., Tibshirani, R.J., 1993. *An introduction to the Bootstrap*, Monographs on statistics and probability 57, Chapman & Hall / CRC, London.

[Gelman 2004] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. *Bayesian data analysis*, 2nd edition, Chapman & Hall/CRC, London

[Gill 1996] Gill, R.D., Van der Laan, M.J., Robins, J.M., 1996. Coursening at Random: Characterizations, Conjectures and Counterexamples. In: *Proceedings of the First Seattle Symposium on Survival Analysis*, pp. 255–294.

[Greene 1993] Greene, W.H., 1993. *Econometric Analysis*, 2nd edition, Prentice Hall international editions, London.

[Greenlees 1982] Greenlees, J.S., Reece, W.S., Zieschang, K.D., 1982. Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American statistical association*, vol. 77, pp. 251-261

[Horrace 2005] Horrace, William C, 2005. *Journal of Multivariate Analysis* vol 94, pp. 209-221.

[Little 1987] Little, R.J.A.;Rubin, D.B., 1987. *Statistical analysis with missing data*, John Wiley & Sons, New York.

[Mack 1999] Mack, T., Venter, G., 1999, *A Comparison of Stochastic Models that Reproduce Chain Ladder Reserve Estimates*, presented at ASTIN

Colloquium, Tokyo, Japan. Available for download on the web at http://www.actuaries.org/ASTIN/Colloquia/Tokyo/Mack_Venter.pdf

[R language] R Development Core Team, 2005. *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[Raghunatan 2001] Raghunatan, T.E., J.M. Lepkowski, J. van Hoewyk and P. Solenberger, 2001. A multivariate technique for multiply imputing missing values using a sequence of regerssion models, *Survey Methodology*, Vol.27, No.1, pp 85-95.

[Raghunatan 2002] Raghunatan, T.E., P. Solenberger, J. van Hoewyk, 2002. *IVEware: Imputation and Variance Estimation Software User Guide* Survey Research Center, Institute for Social Research University of Michigan, available for download at http://www.isr.umich.edu/src/smp/ive/

[Raghunatan 2004] Raghunatan, T.E., 2004. What do we do with missing data? Some options for analysis of incomplete data. *Annual review of public health*, Vol. 25, pp. 99-117

[Robins 1997] Robins, Gill, 1997. Non response models for the analysis of non monotone ignorable missing data, *Statistics in medicine*, Vol. 16, 39-56

[Rubin 1987] Rubin, D.B., 1987. *Multiple imputation for non-response in surveys*. John Wiley & Sons, New York.

[Rubin 1996] Rubin, D.B., 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, Vol. 91, No. 434. (Jun., 1996), pp. 473-489

[Schafer 1997] Schafer, J.L.(1997). *Analysis of incomplete multivariate data*, Monographs on statistics and probability 72, Chapman & Hall, London.

[Schafer 1999] Schafer, J.L., 1999. Multiple imputation: a primer. *Statistical methods in medical research* **8**, 3-15

[Shao 1996] Shao, J.; Sitter, R.R., 1996. *Bootstrap for Imputed Survey Data* Journal of the American Statistical Association, Vol. 91, No. 435. (Sep., 1996), pp. 1278-1288.

[Tempelman 2007] Tempelman, D.C.G., 2007. *Imputation of restricted data*, PhD thesis Rijksuniversiteit Groningen. Statistics Netherlands, Voorburg.

[Tempelman 2004] Tempelman, D.C.G., Steerneman, T., 2004. *Imputation for economic data under linear restrictions*, discussion paper 04002 Statistics Netherlands, Voorburg, march 2004.

[Wonnacott 1970] Wonnacott, R.J., Wonnacott, T.H., 1970. *Econometrics*. John Wiley & sons, New York.

[Van Buuren 2007] Van Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, Vol. 16, pp. 219 - 242.

[Zellner 1971] Zellner, A., 1971. *An introduction to Bayesian Inference in Econometrics*. John Wiley & sons, New York.

Appendix A

Multiple imputation

A major disadvantage of imputation methods is that they fill in only a single value for each missing item. That single value can never reflect the variability due to the fact that the value was actually missing. Besides, it also does not take the variance into account that is due to uncertainty the non-response mechanism.

For these reasons Rubin proposed to use multiple imputation [Rubin 1987]. The idea is to calculate for each missing value not one, but several, say m possible values to impute. This gives m complete data sets instead of one. Each of these data sets can be analyzed as if it were an original complete dataset with no non-response at all.

Let Q be a statistic. Suppose that $\hat{Q} = \hat{Q}(X_{obs}, X_{mis})$ is an estimator for Q such that if we would have a complete dataset without non-response, the difference between Q and its estimator is normally distributed

$$(Q - \hat{Q}) \sim \mathcal{N}_k(0, U) \tag{A.1}$$

where U is a statistic that gives the variance or if Q is a vector the covariance matrix of $(Q - \hat{Q})$. Suppose that repeated imputations have led to m completed data sets, with corresponding statistics $\hat{Q}_{*1}, \ldots \hat{Q}_{*m}$ and $U_{*1}, \ldots, U_{*m}$ for q and U respectively. Now, Let

$$\bar{Q}_m = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_{*l} \text{ and } \bar{U}_m = \frac{1}{m}\sum_{l=1}^{m} U_{*l}$$

denote the the averages of the statistics Q and U over the complete data estimates and and variances and

$$B_m = \frac{1}{m-1}\sum_{l=1}^{m} (\hat{Q}_{*l} - \bar{Q}_m)^T (\hat{Q}_{*l} - \bar{Q}_m)$$

the variance between the m complete-data estimates. Then the total variance of $(Q - \bar{Q}_m)$ is

$$T_m = \bar{U}_m + (1 + \frac{1}{m})B_m.$$

Appendix B

Estimating variances using the bootstrap

The bootstrap re-sampling method, developed by Efron, is a method to estimate variance of a population estimator. The term bootstrap comes from the phrase *to pull oneself up by one's own bootstrap*. Suppose we draw a completely observed random sample $x = (x_1, \ldots x_n)$ from an unknown distribution function F population. We define the *empirical distribution function* $\hat{F}$ to be the distribution that puts probability $1/n$ on each value X_i for $i = 1, \ldots, n$. So to a set A in the sample space of x it assigns the *empirical probability*

$$P_{\hat{F}}(A) = \#\{x_i \in A\}/n,$$

which is the proportion of the observed sample x that lays in A. Suppose we wish to use these observations to estimate a parameter of interest $\theta = t(F)$ with an estimator $\hat{\theta}$. An example could be the *plug-in estimator* which is defined as $\hat{\theta} = t(\hat{F})$. For now however we suppose we would like to estimate θ using a more general statistic $\hat{\theta} = Q(x)$, which may be but is not necessary the plug-in estimator. The idea is to use the empirical distribution $\hat{F}$ of x to estimate the variance of $\hat{\theta} = t(\hat{F})$.

For this end we define a *bootstrap sample* $x^* = (x_1^*, \ldots, x_{n^*}^*)$ to be a simple random sample with replacement from the originally observed sample x. This way, we can draw a random sample from the empirical distribution $\hat{F}$. In most applications n^* is taken equal to n, as we will do too. For each bootstrap sample x^* there is a corresponding *bootstrap replication* of $\hat{\theta}$

$$\hat{\theta}^* = Q(x^*).$$

The *ideal bootstrap estimates* of the standard error $\text{se}_F(\hat{\theta})$ and variance $\text{var}_F(\hat{\theta})$ are the plug-in estimators that use $\hat{F}$ in place of the unknown F: $\text{se}_{\hat{F}}(\hat{\theta}^*)$ and $\text{var}_{\hat{F}}(\hat{\theta}^*)$ respectively. Since there are almost no $\hat{\theta}$ for which nice formulas exist for these ideal bootstrap estimators, except for the mean, we must be satisfied

with an approximation. The obvious approximation uses the empirical standard deviation.

Summarizing the bootstrap algorithm to estimate the variance of $\hat{\theta}$:

1 draw say B bootstrap samples $x^{*1}, \ldots, x^{*B}$ from x independently;

2 calculate for $b = 1, \ldots, B$ the bootstrap replication corresponding to each bootstrap samples x_b^*

$$\hat{\theta}^*(b) = Q(x^{*b}).$$

3 estimate $\text{var}_F(\hat{\theta})$ by calculating

$$\widehat{\text{var}}_B = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right)^2 \tag{B.1}$$

where

$$\bar{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b)$$

or take the square root of (B.1) to get an estimate of the standard error $\text{se}_F(\hat{\theta})$

The question remains how large we should choose the number of bootstrap replications B. As B gets bigger, the bootstrap estimator becomes a better approximation of the variance of $\hat{\theta}$

$$\lim_{B \to \infty} \widehat{\text{se}}_B = \text{se}_{\hat{F}} = \text{se}_{\hat{F}}(\hat{\theta}^*).$$

For a measure for the increased variability due to stopping after only B bootstrap replications, Efron [Effron 1993] suggests the increase of the coefficient of variation of $\widehat{\text{se}}_B$, which is the ratio of its standard deviation to its expectation. But he also states as a rule of thumb that "very seldom" more than $B = 200$ replications are needed to estimate a standard error.

The ideal bootstrap estimate $\text{se}_{\hat{F}}(\hat{\theta}^*)$ and its approximation $\widehat{\text{se}}_{\hat{F}}(\hat{\theta}^*)$ are called *non-parametric* bootstrap estimates because $\hat{F}$ is a non-parametric estimate of the actual distribution function F. There are also *parametric* bootstrap estimators, which use parametric models to find an estimate $\hat{F}_{par}$ for the distribution F. The parametric bootstrap is useful if we wish to make assumptions about the distribution F that make available formulas for variances. However, in bootstrap simulations we do not need such formulas, so we will only use the non-parametric bootstrap.

So far we have assumed the sample x was completely observed, so let us now look at what to do if part of the data is missing. In that case, we want to use the bootstrap method to estimate the variance of parameters in datasets subject to imputation. We can use the outcomes to asses the imputation procedure used. The simplest way would be to impute the data x using some imputation method to get x_I, and then apply the algorithm stated above to x_I.

A major disadvantage of this method is that it treats the imputed data as if it were the completely observed data. So it does not take the variance due to

the imputation process and the missing of the data into account, and leads to underestimates of the variance. This why Shao and Sitter [Shao 1996] call it the naive method. They propose a better alternative. Instead of using the imputed data to create bootstrap samples, they suggest to draw bootstrap samples from the data with missing entries. These bootstrap samples should then be imputed the same way and after that the bootstrap estimators can be calculated. So to reckon with the imputation process, the bootstrap scheme changes to

1 draw B bootstrap samples $x^{*1}, \ldots, x^{*B}$ from the dataset with missing items x independently;

2 apply the imputation procedure to each bootstrap sample x^{*1} to get B imputed versions $x_I^{*1}, \ldots, x_I^{*B}$

3 calculate for $b = 1, \ldots, B$ the bootstrap replication corresponding to each imputed bootstrap sample x_I^{*b}

$$\hat{\theta}_I^*(b) = Q(x_I^{*b}).$$

4 estimate $\text{var}_F(\hat{\theta}_I)$ by calculating

$$\widehat{\text{var}}_{IB} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_I^*(b) - \hat{\theta}_I^*(\cdot) \right)^2 \tag{B.2}$$

where

$$\bar{\theta}_I^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_I^*(b)$$

or take the square root of (B.2) to get an estimate of the standard error $\text{se}_F(\hat{\theta}_I)$

Note that in practice most users will have an already imputed dataset available. If they want to estimate the variance of some parameter using a bootstrap method, they should have the original dataset with missing items at their disposal and use the same imputation method for their bootstrap samples as the one that was used for the original dataset.

Appendix C

Simulation data

The data we analyse are a selection from the dataset `Productiestatistiek Detailhandel` with `SBI` from 52321 up to 5233 and `GK` from 4 up to 9. These data come from Statistics Netherlands' questionnaire `Vragenlijst DHgroot VL 11 deel E: Bedrijfslasten`.

variable name	abbreviation	variable description (in Dutch)
INKWRDE110000	ie11000t	Inkoopwaarde handelsgoederen
INKWRDE130000	ie13000t	Totaal overige inkoopwaarde
INKWRDE100000	ie10000t	**Totaal inkoopwaarde**
LOONSOM100000	lm10000t	Totaal arbeidskosten
BEDRLST345000	bt34500t	Totaal andere personeelskosten
BEDRLST342000	bt34200t	Totaal kosten vervoermiddelen
BEDRLST341000	bt34100t	Totaal energiekosten
BEDRLST343000	bt34300t	Totaal huisvestingskosten
BEDRLST344000	bt34400t	Totaal kosten machines, apparatuur, installaties, kantoorinventaris
BEDRLST346000	bt34600t	Totaal verkoopkosten
BEDRLST347000	bt34700t	Kosten communicatie
BEDRLST348000	bt34800	Totaal kosten dienstverlening t door derden
BEDRLST349000	bt34900t	bedrijfslasten niet elders genoemd
AFSCHRG110000	ag11000t	Afschrijvingen op materiële en immateriële vaste activa
BEDRLST310000	bt31000t	**Totaal bedrijfslasten**

Table C.1: The variables in the simulation dataset. The variables that are a sum of other variables are bold faced.