My amazing title

Your R. Name
April DD, 20YY

Submitted to the Department of
Mathematics and Statistics
of Amherst College in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with honors.

Advisor:
Advisor F. Name

Abstract

The abstract should be a short summary of your thesis work. A paragraph is usually sufficient here.

Acknowledgments

Use this space to thank those who have helped you in the thesis process (professors, staff, friends, family, etc.). If you had special funding to conduct your thesis work, that should be acknowledged here as well.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

In standard undergraduate curricula, there is a strong focus on cross sectional data, and thus no emphasis on how time-sequence data is analyzed. However, a significant portion of data that we encounter in the real world is dependent on time. If we want to track trends and changes over time, such as an effect of a certain drug on the body or growth of a company, longitudinal data and analysis will help us examine those points of interest. For example, the Chinese Longitudinal Healthy Longevity Survey from Duke University assessed physical and mental well-being of Chinese elders for over almost 2 decades and re-interviewed survivors every few year. This follow up in data collection allowed researchers to investigate the aging process over time and identify risk factors and causes leading up to death.

Not only can we observe change over time in individuals, but we can look at higher-level grouping, such as change in schools, counties, and organizations. It should be emphasized that only longitudinal data can capture changes within a subject or group; cross-sectional data contain responses that are captured at only one occasion that are then compared to other subjects. Ultimately, it cannot provide information about changes over time.

One key aspect of longitudinal data is that there needs to be repeated measurements of the same individuals across multiple periods of time. If there aren't repeated observations, then it is not possible to make any comparisons between two or more time points. Having repeated measurements of the same individual allows for re-

moval of potential confounding effects, such as gender or socioeconomic status, from the analysis. Since we assume that these confounding variables are fixed effects that do not vary from measurement to measurement, all changes from an individual cannot be attributed to these effects.

The measure that captures the observed changes within an individual is referred to as a response trajectory. There are different ways of comparing response trajectories. For example, it is possible to compare the post-treatment vs baseline changes across multiple treatment groups, or it is also possible to compare the rate of change. The method chosen depends on the specific question of the study.

Apart from comparing just the response trajectories, it is also of interest to compare individual differences in the relationship between covariates and the response trajectory. This can be captured using various different statistical models. The choice of model depends on several characteristics of the data.

1.1 Characteristics of longitudinal data

While the only requirement of longitudinal data is that there is more than one observation for a given individual, there are other components that affect the model chosen. Data can be unbalanced or balanced: *balanced* data refers to when all individuals have the same number of repeated measurements taken at the same occasions. In addition, data can also be missing, resulting in automatically unbalanced data. This affects the accuracy of how changes over time are analyzed depending on if there are any patterns to the missing data or not.

Another unique characteristic of longitudinal data is that repeated measurements of each individual are typically positively correlated. This feature violates conditions of other common statistical methods such as linear regression, where measurements

are assumed to be independent. This positive correlation allows for more accurate estimates of the model coefficients and response trajectories since there is reduced uncertainty knowing that a previous measurement can help predict the next one.

In longitudinal analysis, a covariance matrix is calculated for each individual and all of their measurements. The diagonals of this matrix represent the variance of each of the measurements, which are not constant over time. The off-diagonals of the matrix are non-zero to account for the lack of independence between measurements, and are usually not constant because correlations between measurements decrease over time. While these values are rarely 0, they are also rarely 1. There are different covariance pattern structures that are imposed that account for these features.

These features of the covariance of longitudinal data serve as the underlying premise to the idea that variation can be separated into three distinct parts: 1) between-individual variation, 2) within-individual variation, and 3) measurement error.

Between-individual variation helps explain why measurements from the same individual are more likely to be positively correlated than measurements to a different individual. Within-individual variation helps explain why correlations decrease with increasing time differences, and measurement error explains why correlations are never one. These three types of variation may contribute to total variation in unequal amounts, but may not need to be differentiated depending on the type of longitudinal analysis desired.

1.1.1 Notation

Throughout the rest of the text, we will use a standard set of notation for all parameters and variables. Y_{ij} represents the response variable for the i^{th} individual at the j^{th} measurement. When we have repeated n_i measurements for an individual, we can

construct a vector,

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ .. \\ Y_{in_i} \end{pmatrix}.$$

We use μ_{ij} as the conditional mean response at the j^{th} measurement, where conditional entails a dependence of the mean response on the covariates.

1.2 Estimation and Inference

Regression coefficient values β and the covariance matrix Σ_i can be estimated using maximum likelihood estimation, which identifies values of β and Σ_i that maximize the joint probability of the response variable occurring based on the observed data; the probability is known as the likelihood function. These values are estimates that are denoted by $\hat{\beta}$ and $\hat{\Sigma}_i$. When observations are independent of one another, maximizing the likelihood function for β is equivalent to finding a value of $\hat{\beta}$ that minimizes the sum of the squares of the residuals. However, since there are repeated measurements of each individual that are not independent of one another we use the generalized least squares (GLS) estimator:

$$\hat{\beta} = \{\Sigma_{i=1}^{N}(X_i'\Sigma_i^{-1}X_i)\}^{-1}\Sigma_{i=1}^{N}(X_i'\Sigma_i^{-1}y_i).$$

In addition, the sampling distribution of $\hat{\beta}$ has mean β and covariance:

$$\hat{Cov}(\hat{\beta}) = \{\Sigma_{i=1}^{N}(X_i'\Sigma_i^{-1}X_i)\}^{-1}.$$

The GLS estimator assumes that Σ_i is known. However, since this isn't usually

the case, we can substitute Σ_i with a maximum likelihood estimate $\hat{\Sigma}_i$.. It can be shown that the properties of $\hat{\beta}$ still hold using an estimate of the covariance.

While the maximum likelihood estimate of Σ_i is adequate, a modified method known as restricted maximum likelihood (REML) estimation is suggested to reduce bias in finite samples. The bias originates from the fact that β itself is also estimated from data, but is not accounted for when estimating covariance. In REML estimation of Σ_i, β is removed from the likelihood function. This REML estimation of Σ_i can be used in the GLS estimator for $\hat{\beta}$ mentioned above, and is recommended in place of the ML estimator.

Now that we have estimates for β, we can make inferences through construction of confidence intervals and hypothesis testing. For example, using the ML estimate $\hat{\beta}$ and $\hat{Cov}(\hat{\beta})$, we can construct a Wald statistic to test for significance of $\hat{\beta}_k$:

$$Z = \frac{\hat{\beta}_k}{\sqrt{\hat{Var}(\hat{\beta}_k)}}.$$

One crucial assumption when conducting inference using the ML estimate for β is that the sample size is sufficient enough where it does not affect the estimate for Σ_i. However, what happens when the sample size is too small? This causes $\hat{\Sigma}_i$ to underestimate the true variance, which in turn causes $\hat{Cov}(\hat{\beta})$ to be too small since it relies on covariance estimator. If $\hat{Cov}(\hat{\beta})$ is too small, the denominator of the test statistic is inflated, leading to increased Type I error. One can see that the bias of the covariance estimator weakens the entire foundation of estimation and inference.

How can this be fixed? Both Satterthwaite and Kenward and Roger have proposed reductions to the degrees of freedom when conducting tests in order to account for this uncertainty of the covariance estimator. Kenward and Roger go one step forward to also adjust the test statistic itself. In subsequent chapters, we will go into further

detail about these two methods, and compare their performance on longitudinal data with different sample sizes and distributions of the response variable.

1.3 Linear models for longitudinal data

As mentioned previously, there are multiple ways to model longitudinal data. When the response variable is continuous, we can consider a model that relates the mean response and the covariates in a linear way. In a linear model all components can be represented using vectors and matrices. The most general form of the linear model can be represented as:

$$E(Y|X_i) = X_i\beta$$

, where β is a vector of regression coefficients and X_i is a vector of covariates. We will discuss three methods for linear models: 1) response profile analysis, 2) parametric time model, 3) linear mixed effect model.

1.3.1 Response profile analysis

In response profile analysis, we allow for arbitrary patterns in the mean response over time. A sequence of means over time is known as the mean response profile. The main goal of this analysis is to identify differences in pattern of change in mean response profile among 2 or more groups. This method requires that the data be balanced.

There are three effects of interest when analyzing response profiles in longitudinal analysis: 1. $group \times time$ interaction effect (are the mean response profiles different in groups over time?) 2. time effect (assuming mean response profiles are parallel between groups, are the means changing over time?) 3. Group effect (do the mean response profiles differ?)

However, the first question is the primary interest. The goal is to find whether

the change in mean response over time differs across groups.

To test for significance of the $group \times time$ effect, we have a null hypothesis that the difference in means between the n groups is constant over time, which in other words entails that mean response profiles between the groups have parallel slopes. We can implement the general linear model $\mu_i = X_i\beta$ to test our hypotheses, using comparison of β slope parameters to determine whether there is a $group \times time$ effect.

For example, to express the model for response profile analysis for G groups and n occasions of measurement, we have $G \times n$ parameters for the G mean response profiles. For two groups measured at three occasions, we have 6 slope parameters. if $\beta_1 - \beta_3$ represent slope parameters for mean responses in group 1 and $\beta_4 - \beta_6$ represent slope parameters for mean responses in group 2, our null hypotheses would be that $(\beta_1 - \beta_4) = (\beta_2 - \beta_5) = (\beta_3 - \beta_6)$.

An unstructured covariance model is typically assumed for response profile analysis. "Unstructured" means that there is no explicit structure or pattern imposed on the covariance for the repeated measures, so each of the variances and covariance pairs are estimated using restricted maximum likelihood estimation (REML). For n repeated measures, there are n variances and $n \times (n-1)/2$ covariances to be estimated. In a study where there are 10 repeated measurements, there 55 total covariance parameters to be estimated, which can become computationally intensive.

One other aspect to consider when conducting analysis on mean response profiles is how to adjust for the baseline measurement. The baseline value is important when we want to calculate measures that compare mean response to the baseline. How we adjust depends on whether the study is randomized or observational. When the study is randomized and baseline measurement is taken before treatment assignment, the mean response at occasion 1 is independent of the group, and assumed to be equal. One possible method is to treat the baseline measurement as a covariate, and

use response measurements 2 through n as the dependent measures. This is referred to as the analysis of covariance approach. Additionally, this method only works for randomized studies because using the baseline measurement as a covariate for observational studies may produce confounding effects. For an observational study, it is recommended to subtract the baseline response to create a change score. For both types of longitudinal studies there are various methods to account for the baseline value, and should be considered carefully before implementing the method.

Overall, response profile analysis is a straightforward method in investigating differences between groups for longitudinal data. Since both the covariance and mean responses have no imposed structure, the analysis is more robust and immune to inaccurate results due to model misspecification. However, there are drawbacks as well. Response profile analysis does not consider time-order of the measurements and does not distinguish between between-individual variation and within-individual variation. In addition, it can only provide a broad analysis of whether there are differences across groups and time, but does not provide the amount of detail usually needed to answer research questions, such as how exactly measurements taken towards the end of the study compare to measurements taken at the beginning. In this method, time is treated as a categorical covariate rather than a continuous one. Another method that addresses the issue of examining time order of the data is parametric time models.

1.3.2 Parametric Time Models

Parametric time models are able to capture time order of the data by fitting linear or quadratic curves to capture an increasing or decreasing pattern over time. Time is treated as a continuous covariate rather than a categorical one. In addition, unlike response profile analysis, parametric time models are able to handle unbalanced and missing data. Rather than fitting a complex and perfect model onto the observed

mean response profile, parametric time models fit simple curves that produce covariate effects of greater power. This is because in mean response profile we are testing a wider range of hypotheses since we are looking for inequality between two groups; however, in parametric time models, we are testing more specifically whether the data follow a linear trend, which results in more power.

Additionally, while in the mean response profile analysis an unstructured covariance pattern is assumed, here there is flexibility in choice of the covariance model; there are several options such as Toeplitz or compound symmetric that impose various structures on the model. For example, a Toeplitz model:

$$Cov(Y_i) = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \dots & \dots \\ \rho_3 & & & & \end{pmatrix}$$

structures the covariance matrix such that any pair of responses that are equally separated in time have the same correlation.

It is possible to choose an unstructured covariance model as well, but can be computationally intense if there are a large number of measurements.

We can use parametric time models in two ways: through polynomial trends and linear spines.

1.3.3 Polynomial Trends

Using polynomial trends such as linear or quadratic, we can model longitudinal data as a function of time. Linear trends are the most common and interpretable ways to model change in mean over time. In an example comparing a treatment group to a

control group, we can fit a linear trend using the following equation:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Group_i + \beta_4 Time_{ij} \times Group_i.$$

If $\beta_4 = 0$, then the two groups do not differ in terms of changes in the mean response over time.

For quadratic trends, the changes in mean are no longer constant since the rate of change depends on the time. Thus, we fit an additional parameter to express the rate of change. Using the previous example of treatment vs. control group, we have the model:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Time_{ij}^2 + \beta_4 Group_i + \beta_5 Time_{ij} \times Group_i + \beta_6 Time_{ij}^2 \times Group_i.$$

As we can see from the models above, the inclusion of an additional parameter $Time_{ij}^2$ changes the mean response rate. One problem that may arise from using quadratic trends is that there is collinearity between $Time_{ij}$ and $Time_{ij}^2$, which can affect the estimation of β. To account for this, we can center the $Time_{ij}$ variable around the mean time value for all individuals, instead of centering it around zero as done in normal analysis. For example if we have a set of times $Time = 0, 1, 2, ...10$, then the mean time value is five. Thus time zero would be recentered as -5. The interpretation of the intercept changes to represent the mean response at that recentered mean time value.

1.3.4 Linear splines

In instances where responses cannot be adequately fit by polynomial trends, such as when the responses fluctuate between increasing and decrease at different extents, we can employ a linear spline model. This model consists of piece-wise line segments

that have unique slopes for a given set of time measurements. The point at which different line segments meet are called knots, and the number of knots depends on the context of the data and researcher discretion.

Drawing again from our treatment vs control group design, a linear model for the mean responses of the control group is:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 (Time_{ij} - t^*)_+.$$

The $()_+$ indicates a truncated line function and is positive when $Time_{ij} - t^*$ is greater than 0, and otherwise is equal to 0. In this case, the function depends on the specified time t^*. If the mean response is before t^*, then the mean response is modeled by:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij}.$$

If the mean response is after t^*, it is modeled by

$$E(Y_{ij}) = (\beta_1 - \beta_3 t^*) + (\beta_2 + \beta_3) Time_{ij}.$$

There are benefits to parametric models that make them a more appealing choice compared to response profile analysis. Parametric time models are able to capture time order, and can be used with unbalanced data. However, they do not differentiate between subject and within subject variation. If further analysis of individual variation is desired, linear mixed effects models can be employed.

1.3.5 Linear Mixed Effects

In both response profile analysis and parametric time models, the regression parameters are considered to be universal for each population group. However, in instances

where we want to account for heterogeneity within a population, we can use a linear mixed effects model and consider a subset of the regression parameters to be random. This model distinguishes between fixed effects, which are population characteristics shared by all individuals, and subject specific effects, also known as random effects, which pertain to each individual. These subject specific effects mean that parameters are random, which induces a structure onto the covariance model.

In addition, distinguishing between fixed and random effects allows for differentiation between within-subject and between-subject variation.

One example of the linear mixed effects model is the random intercept model, which is the simplest version of the linear mixed effects model:

$$Y_{ij} = X'_{ij}\beta + b_i + \epsilon_{ij}$$

This model is very similar to the general linear model with a few additions. b_i is the random subject effect and ϵ is the measurement error. Both effects are random, with mean 0 and $\text{Var}(b_i) = \sigma_b^2, \text{Var}(\epsilon_{ij}) = \sigma^2$.

$X'_{ij}\beta$ is the population mean, and b_i represents the differing subject effect that is unique to each individual. b_i is interpreted as how the subject deviates from the population mean while accounting for covariates.

As mentioned previously, the random effects are responsible for inducing a structure on the covariance model. This structure is not to be confused with the covariance structures that can be chosen when using parametric time models. For a given individual, it can be shown that variance of each response is:

$$\text{Var}(Y_{ij}) = \sigma_b^2 + \sigma^2$$

and the covariance between two measurements Y_{ij} and Y_{ik} is equal to σ_b^2. The resulting

covariance matrix $\begin{pmatrix} \sigma_b^2+\sigma^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma^2 & \sigma_b^2+\sigma^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma^2 & ... & ... \end{pmatrix}$\$

implies correlation between measurements, and also highlights the role played by the random effects in determining the covariance.

Extending beyond the random intercept model, multiple random effects can be incorporated.

A linear mixed effects model can expressed as

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i.$$

Where: β is a $p \times 1$ vector of fixed effects b_i is a $q \times 1$ vector of random effects X_i is a $n \times p$ matrix of covariates Z_i is a $n \times q$ matrix of covariates

The subset of regression covariates that vary randomly are found in Z_i. We assume that b_i comes from a multivariate normal distribution with mean 0 and covariance matrix G. We also assume that ϵ_i are independent of b_i, and come from multivariate normal distribution with mean 0 and covariance matrix R_i.

The covariance of Y_i can be modeled by

$$\text{Cov}(Z_i b_i) + Cov(\epsilon_i) = Z_i G Z_i' + R_i.$$

This model, which outlines a distinction between G and R_i, allows for separate analysis of between subject and within subject variation. Unlike other covariance models, in linear mixed effects models the covariance is a function of the times of measurement. This allows for unbalanced data to be used for the model since each individual can have their unique set of measurement times. Lastly, the model allows for variance and covariance to change as a function of time. To illustrate, consider the following

model:

In an example where individuals can vary both in their baseline response and their rate of change, we have:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

where both X_i and $Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \dots & \dots \\ 1 & t_{in} \end{pmatrix}$. For the i^{th} subject at the j^{th} measurement, the equation is as follows:

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}.$$

If $\text{Var}(b_{1i}) = g_{11}$, $\text{Var}(b_{2i}) = g_{22}$, and $Cov(b_{1i}, b_{2i}) = g_{12}$ where these three components represent the G covariance for b_i, then it can be shown that $Cov(Y_{ij}, Y_{ik}) = g_{11} + (t_{ij} + t_{ik})g_{12} + t_{ij}t_{ik}g_{22}$.

Here in the covariance matrix we can see the dependence of the covariance on time. In this example there are four covariance parameters that arise from the two random effects of intercept and time. The number of covariance parameters is represented by $q \times (q+1)/2+1$, where q is the number of random effects. To choose the most optimal model for covariance, we compare two nested models, one with $q+1$ random effects and one with q random effects. We use the likelihood ratio test to make a decision for which model to use.

One additional analysis that is possible with linear mixed effects models is predicting subject-specific responses. Given that b_i is a random variable, we can predict

it using:

$$E(b_i|Y_i) = GZ_i(\Sigma)_i^{-1}(Y_i - X_i\hat{\beta}).$$

Because the covariance of Y_i is unknown, we can estimate both G and $(\Sigma)_i^{-1}$ using REML, creating $\hat{b}_i$, also known as the empirical best linear unbiased prediction (BLUP). Thus, the equation for predicting the response profile is:

$$\hat{Y}_i = X_i\hat{\beta} + Z_i\hat{b}_i.$$

This equation to estimate the mean response profile can be extended to incorporate R_i, which represents within-subject variability. From this extension, we see that the equation and the empirical BLUP account for the weighting of both the within-subject variability and between-subject variability. If there is more within-subject variability, then more weight is assigned to $X_i\hat{\beta}$, the population mean response profile, in comparison to the subject's individual responses, and vice versa.

1.4 Choosing the best model

After presenting three methods of evaluating longitudinal data, the natural question arises of how to choose the most appropriate model. While there is no definite correct answer, there are several factors to consider. If data are unbalanced, response profile analysis should not be considered; rather, parametric time model or linear mixed effect model would be more optimal. If time order is important to the analysis, then only parametric time model and linear mixed effect model should be used. If there is a need to distinguish between the two types of variation that can occur, then only linear mixed effect models are appropriate. The model should ultimately be chosen based on the characteristics and constraints of the data, as well as the specificity of

the research question at hand.

1.5 Conclusion

Longitudinal analysis is a valuable method to analyze changes over time. It is important to understand the unique characteristics that come with this analysis and to choose the best model that can capture the salient patterns that arise from the data. In subsequent chapters we will dive more deeply into how inference in longitudinal analysis is affected when sample sizes are not efficient through both simulation and application.

Chapter 2 R Markdown Basics

Be careful with your spacing in *Markdown* documents. While whitespace largely is ignored, it does at times give *Markdown* signals as to how to proceed. As a habit, try to keep everything left aligned whenever possible, especially as you type a new paragraph. In other words, there is no need to indent basic text in the Rmd document (in fact, it might cause your text to do funny things if you do).

2.1 Lists

It's easy to create a list. It can be unordered like

- Item 1
- Item 2

or it can be ordered like

1. Item 1
2. Item 2

Notice that I intentionally mislabeled Item 2 as number 4. *Markdown* automatically figures this out! You can put any numbers in the list and it will create the list. Check it out below.

To create a sublist, just indent the values a bit (at least four spaces or a tab). (Here's one case where indentation is key!)

1. Item 1
2. Item 2
3. Item 3
    - Item 3a
    - Item 3b

2.2 Line breaks

Make sure to add white space between lines if you'd like to start a new paragraph. Look at what happens below in the outputted document if you don't:

Here is the first sentence. Here is another sentence. Here is the last sentence to end the paragraph. This should be a new paragraph.

Now for the correct way:

Here is the first sentence. Here is another sentence. Here is the last sentence to end the paragraph.

This should be a new paragraph.

2.3 R chunks

When you click the **Knit** button above a document will be generated that includes both content as well as the output of any embedded **R** code chunks within the document. You can embed an **R** code chunk like this (`cars` is a built-in **R** dataset):

```
summary(cars)
```

```
     speed           dist
 Min.   : 4.0   Min.   :  2
 1st Qu.:12.0   1st Qu.: 26
 Median :15.0   Median : 36
```

```
Mean   :15.4   Mean   : 43
3rd Qu.:19.0   3rd Qu.: 56
Max.   :25.0   Max.   :120
```

2.4 Inline code

If you'd like to put the results of your analysis directly into your discussion, add inline code like this:

> The `cos` of 2π is 1.

Another example would be the direct calculation of the standard deviation:

> The standard deviation of `speed` in `cars` is 5.288.

One last neat feature is the use of the `ifelse` conditional statement which can be used to output text depending on the result of an **R** calculation:

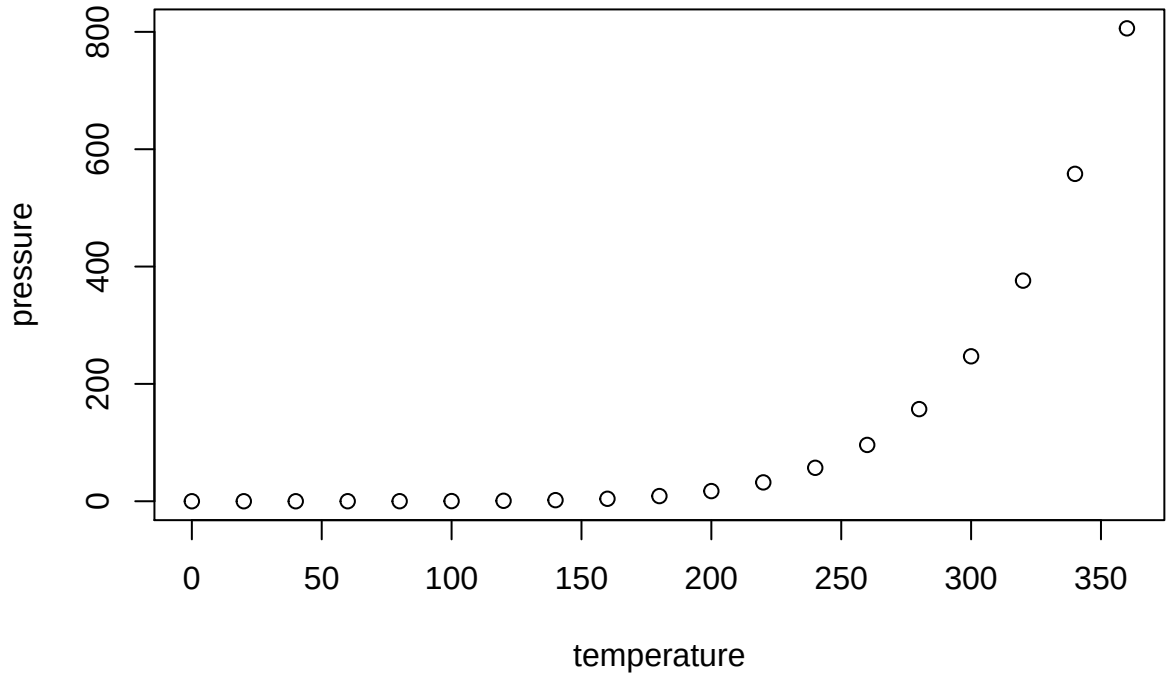
> The standard deviation is less than 6.

Note the use of `>` here, which signifies a quotation environment that will be indented.

As you see with `$2 \pi$` above, mathematics can be added by surrounding the mathematical text with dollar signs. More examples of this are in Mathematics and Science if you uncomment the code in Math.

2.5 Including plots

You can also embed plots. For example, here is a way to use the base **R** graphics package to produce a plot using the built-in `pressure` dataset:



Note that the `echo=FALSE` parameter was added to the code chunk to prevent printing of the **R** code that generated the plot. There are plenty of other ways to add chunk options. More information is available at http://yihui.name/knitr/options/.

Another useful chunk option is the setting of `cache=TRUE` as you see here. If document rendering becomes time consuming due to long computations or plots that are expensive to generate you can use knitr caching to improve performance. Later in this file, you'll see a way to reference plots created in **R** or external figures.

2.6 Loading and exploring data

Included in this template is a file called `flights.csv`. This file includes a subset of the larger dataset of information about all flights that departed from Seattle and Portland in 2014. More information about this dataset and its **R** package is available at http://github.com/ismayc/pnwflights14. This subset includes only Portland flights and only rows that were complete with no missing values. Merges were also done with the `airports` and `airlines` data sets in the `pnwflights14` package to get more descriptive airport and airline names.

We can load in this data set using the following command:

```
flights <- read.file("data/flights.csv")
```

```
Reading data with read.csv()
```

The data is now stored in the data frame called `flights` in **R**. To get a better feel for the variables included in this dataset we can use a variety of functions. Here we can see the dimensions (rows by columns) and also the names of the columns.

```
dim(flights)
```

```
[1] 52808    16
```

```
names(flights)
```

```
 [1] "month"        "day"          "dep_time"     "dep_delay"
 [5] "arr_time"     "arr_delay"    "carrier"      "tailnum"
 [9] "flight"       "dest"         "air_time"     "distance"
[13] "hour"         "minute"       "carrier_name" "dest_name"
```

```r
# read long paragraph file
longtext <- readLines("data/paragraphs.txt")
```

```
Warning in readLines("data/paragraphs.txt"): incomplete final
line found on 'data/paragraphs.txt'
```

```r
# display text as vector
longtext
```

```
[1] "Lorem ipsum dolor sit amet, consectetur adipiscing elit,
sed do eiusmod tempor incididunt ut labore et dolore magna
aliqua. Ut enim ad minim veniam, quis nostrud exercitation
ullamco laboris nisi ut aliquip ex ea commodo consequat. "
[2] ""
[3] "Duis aute irure dolor in reprehenderit in voluptate
velit esse cillum dolore eu fugiat nulla pariatur. Excepteur
sint occaecat cupidatat non proident, sunt in culpa qui
officia deserunt mollit anim id est laborum."
```

```r
# display text as paragraphs
cat(longtext)
```

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed
do eiusmod tempor incididunt ut labore et dolore magna
aliqua. Ut enim ad minim veniam, quis nostrud exercitation
ullamco laboris nisi ut aliquip ex ea commodo consequat.
Duis aute irure dolor in reprehenderit in voluptate velit
esse cillum dolore eu fugiat nulla pariatur. Excepteur sint
occaecat cupidatat non proident, sunt in culpa qui officia
deserunt mollit anim id est laborum.
```

```r
# display text without linewidth option specified
longtext
```

```
[1] "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor
[2] ""
[3] "Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu
```

Another good idea is to take a look at the dataset in table form. With this dataset having more than 50,000 rows, we won't explicitly show the results of the command here. I recommend you enter the command into the Console ***after*** you have run the **R** chunks above to load the data into **R**.

```
View(flights)
```

While not required, it is highly recommended you use the `dplyr` package to manipulate and summarize your data set as needed. It uses a syntax that is easy to understand using chaining operations. Below I've created a few examples of using `dplyr` to get information about the Portland flights in 2014. You will also see the use of the `ggplot2` package, which produces beautiful, high-quality academic visuals.

The example we show here does the following:

- Selects only the `carrier_name` and `arr_delay` from the `flights` dataset and then assigns this subset to a new variable called `flights2`.
- Using `flights2`, we determine the largest arrival delay for each of the carriers.

```
flights2 <- flights %>%
  select(carrier_name, arr_delay)
max_delays <- flights2 %>%
  group_by(carrier_name) %>%
  summarize(max_arr_delay = max(arr_delay, na.rm = TRUE))
```

A useful function in the `knitr` package for making nice tables in *R Markdown* is called `kable`. It is much easier to use than manually entering values into a table by copying and pasting values into Excel or LaTeX. This again goes to show how nice reproducible documents can be! (Note the use of `results="asis"`, which will produce the table instead of the code to create the table.) The `caption.short` argument is used to include a shorter title to appear in the List of Tables.

```r
kable(max_delays,
      col.names = c("Airline", "Max Arrival Delay"),
      caption = "Maximum Delays by Airline",
      caption.short = "Max Delays by Airline",
      longtable = TRUE,
      booktabs = TRUE)
```

Table 2.1: Maximum Delays by Airline

Airline	Max Arrival Delay
Alaska Airlines Inc.	338
American Airlines Inc.	1539
Delta Air Lines Inc.	651
Frontier Airlines Inc.	575
Hawaiian Airlines Inc.	407
JetBlue Airways	273
SkyWest Airlines Inc.	421
Southwest Airlines Co.	694
United Air Lines Inc.	472
US Airways Inc.	347
Virgin America	366

The last two options make the table a little easier-to-read.

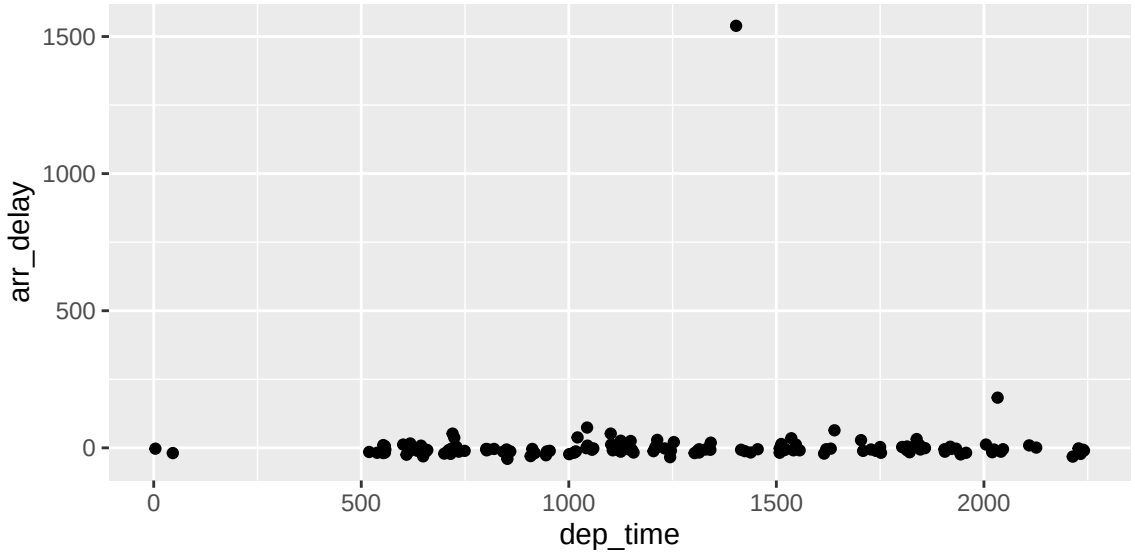
We can further look into the properties of the largest value here for American Airlines Inc. To do so, we can isolate the row corresponding to the arrival delay of 1539 minutes for American in our original `flights` dataset.

```
flights %>% filter(arr_delay == 1539,
                   carrier_name == "American Airlines Inc.") %>%
  select(-c(month, day, carrier, dest_name, hour,
            minute, carrier_name, arr_delay))
```

```
  dep_time dep_delay arr_time tailnum flight dest air_time
1     1403      1553     1934  N595AA   1568  DFW      182
  distance
1     1616
```

We see that the flight occurred on March 3rd and departed a little after 2 PM on its way to Dallas/Fort Worth. Lastly, we show how we can visualize the arrival delay of all departing flights from Portland on March 3rd against time of departure.

```
flights %>% filter(month == 3, day == 3) %>%
  ggplot(aes(x = dep_time, y = arr_delay)) + geom_point()
```



This is a proof. There is a proof environment in which you can create equations

$$\hat{\beta}_0 + \hat{\beta}_1 x$$

□

2.7 Additional resources

- *Markdown* Cheatsheet - https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet
- *R Markdown* Reference Guide - https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf
- Introduction to `dplyr` - https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html
- `ggplot2` Documentation - http://docs.ggplot2.org/current/

Chapter 3 Mathematics and Science

3.1 Math

TEX is the best way to typeset mathematics. Donald Knuth designed TEX when he got frustrated at how long it was taking the typesetters to finish his book, which contained a lot of mathematics. One nice feature of *R Markdown* is its ability to read LaTeX code directly.

If you are doing a thesis that will involve lots of math, you will want to read the following section.

$$\sum_{j=1}^{n}(\delta\theta_j)^2 \leq \frac{\beta_i^2}{\delta_i^2+\rho_i^2}\left[2\rho_i^2+\frac{\delta_i^2\beta_i^2}{\delta_i^2+\rho_i^2}\right] \equiv \omega_i^2$$

From Informational Dynamics, we have the following (Dave Braden):

After n such encounters the posterior density for θ is

$$\pi(\theta|X_1 < y_1, \ldots, X_n < y_n) \propto \pi(\theta)\prod_{i=1}^{n}\int_{-\infty}^{y_i}\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)\,dx$$

Another equation:

$$\det \begin{vmatrix} c_0 & c_1 & c_2 & \dots & c_n \\ c_1 & c_2 & c_3 & \dots & c_{n+1} \\ c_2 & c_3 & c_4 & \dots & c_{n+2} \\ \vdots & \vdots & \vdots & & \vdots \\ c_n & c_{n+1} & c_{n+2} & \dots & c_{2n} \end{vmatrix} > 0$$

3.2 Statistics Symbols and Expressions

Exponent or Superscript: x^2

Subscript: $x_1, x_2, \dots, x_n$

Both combined: x_1^{k+1}.

Our favorite Greeks: σ, ϵ, μ

Defining a normally distributed random variable: $X \sim N(\mu, \sigma)$

How do we compute sample variance again?

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Sometimes you'll need to consider asymptotics, that is, what happens as $n \to \infty$.

3.3 Additional information

Many of the symbols you will need can be found on Reed College's math page http://web.reed.edu/cis/help/latex/math.html and the Comprehensive LaTeX Symbol Guide (http://mirror.utexas.edu/ctan/info/symbols/comprehensive/symbols-letter.pdf).

Chapter 4 Tables, Graphics, References, and Labels

4.1 Tables

In addition to the tables that can be automatically generated from a data frame in **R** that you saw in R Markdown Basics using the `kable` function, you can also create tables using *pandoc*. (More information is available at http://pandoc.org/README.html#tables.) This might be useful if you don't have values specifically stored in **R**, but you'd like to display them in table form. Below is an example. Pay careful attention to the alignment in the table and hyphens to create the rows and columns.

Table 4.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child	Inherited
Education	-0.49	Yes
Socio-Economic Status	0.28	Slight
Income	0.08	No
Family Size	0.18	Slight
Occupational Prestige	0.21	Slight

We can also create a link to the table by doing the following: Table 4.1. If you go back to Loading and exploring data and look at the `kable` table, we can create a

reference to this max delays table too: Table 2.1. The addition of the `(\#tab:inher)` option to the end of the table caption allows us to then make a reference to Table `\@ref(tab:label)`. Note that this reference could appear anywhere throughout the document after the table has appeared.

4.2 Figures

If your thesis has a lot of figures, *R Markdown* might behave better for you than that other word processor. One perk is that it will automatically number the figures accordingly in each chapter. You'll also be able to create a label for each figure, add a caption, and then reference the figure in a way similar to what we saw with tables earlier. If you label your figures, you can move the figures around and *R Markdown* will automatically adjust the numbering for you. No need for you to remember! So that you don't have to get too far into LaTeX to do this, a couple **R** functions have been created for you to assist. You'll see their use below.

In the **R** chunk below, we will load in a picture stored as `amherst.png` in our main directory. We then give it the caption of "Amherst logo", the label of "amherstlogo", and specify that this is a figure. Make note of the different **R** chunk options that are given in the R Markdown file (not shown in the knitted document).

```
include_graphics(path = "figures/amherst.png")
```

Figure 4.1: Amherst logo

Here is a reference to the Amherst logo: Figure 4.1. Note the use of the `fig:` code here. By naming the **R** chunk that contains the figure, we can then reference that figure later as done in the first sentence here. We can also specify the caption for the figure via the R chunk option `fig.cap`.

Below we will investigate how to save the output of an **R** plot and label it in a way similar to that done above. Recall the `flights` dataset from Chapter 2. (Note that we've shown a different way to reference a section or chapter here.) We will next explore a bar graph with the mean flight departure delays by airline from Portland for 2014. Note also the use of the `scale` parameter which is discussed on the next page.

```
#if(!exists("flights")) flights <- read.csv("data/flights.csv")
flights %>% group_by(carrier) %>%
  summarize(mean_dep_delay = mean(dep_delay)) %>%
  ggplot(aes(x = carrier, y = mean_dep_delay)) +
  geom_bar(position = "identity", stat = "identity", fill = "red")
```

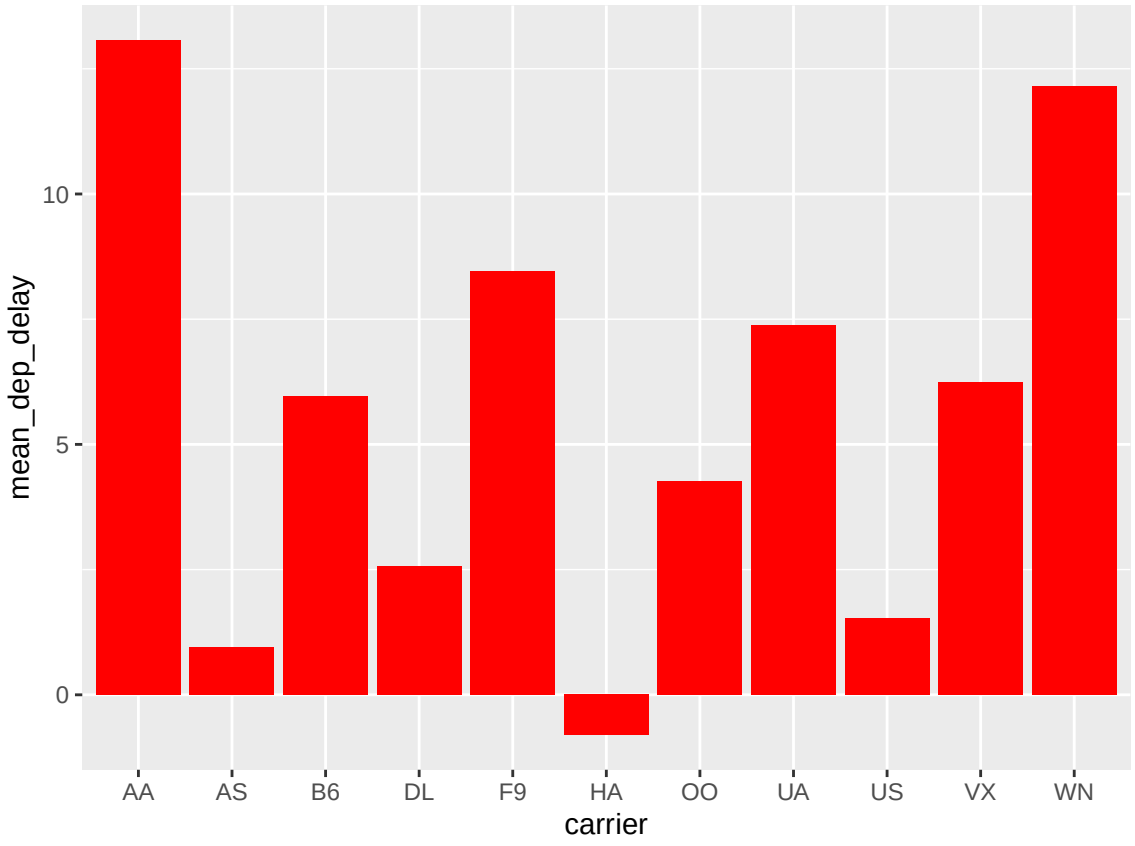


Figure 4.2: Mean Delays by Airline

Here is a reference to this image: Figure 4.2.

A table linking these carrier codes to airline names is available at https://github.com/ismayc/pnwflights14/blob/master/data/airlines.csv.

Next, we will explore the use of the `out.extra` chunk option, which can be used to shrink or expand an image loaded from a file by specifying `"scale= "`. Here we use the mathematical graph stored in the "subdivision.pdf" file.

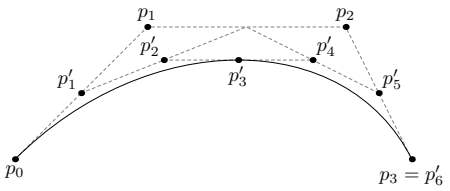


Figure 4.3: Subdiv. graph

Here is a reference to this image: Figure 4.3. Note that `echo=FALSE` is specified so that the **R** code is hidden in the document.

More Figure Stuff

Lastly, we will explore how to rotate and enlarge figures using the `out.extra` chunk option. (Currently this only works in the PDF version of the book.)

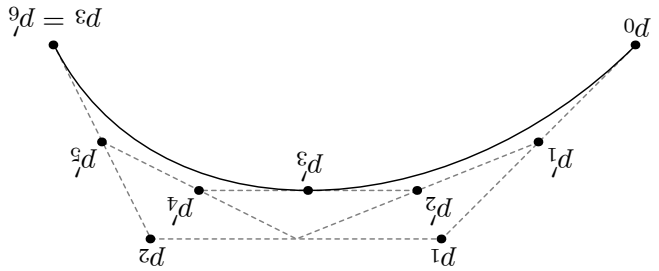


Figure 4.4: A Larger Figure, Flipped Upside Down

As another example, here is a reference: Figure 4.4.

4.3 Footnotes and Endnotes

You might want to footnote something.[1] The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be

[1] footnote text

found about both on the Reed Thesis site https://www.reed.edu/cis/help/latex/thesis.html or feel free to reach out to Prof. Bailey at bebailey@amherst.edu.

4.4 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Amherst librarians have created Zotero documentation at https://www.amherst.edu/library/find/citation/zotero. In addition, a tutorial is available from Middlebury College at http://sites.middlebury.edu/zoteromiddlebury/.

R Markdown uses *pandoc* (http://pandoc.org/) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: (Molina & Borkovec, 1994). This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see the Reed College CUS site (http://web.reed.edu/cis/help/latex/index.html)[2]. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at http://web.reed.edu/cis/help/latex/bibtex.html), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at http://web.reed.edu/cis/help/

[2]Reed College (2007)

latex/bibtexstyles.html) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at http://web.reed.edu/cis/help/latex/bibman.html). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at https://www.zotero.org/styles. Make sure to download the file into the csl folder.

Tips for Bibliographies

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.
- The cite key (a citation's label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.
- Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.
- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.
- You can add a Reed Thesis citation[3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

[3]Noble (2002)

4.5 Anything else?

If you'd like to see examples of other things in this template, please contact Professor Bailey (email bebailey@amherst.edu) with your suggestions.

Conclusion

If we don't want the conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

A.1 In the main file 4:

A.2 In Chapter 4:

Appendix B The Second Appendix

R code

Corrections

A list of corrections after submission to department.

Corrections may be made to the body of the thesis, but every such correction will be acknowledged in a list under the heading "Corrections," along with the statement "When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected." This list will be given on a sheet or sheets to be appended to the thesis. Corrections to spelling, grammar, or typography may be acknowledged by a general statement such as "30 spellings were corrected in various places in the thesis, and the notation for definite integral was changed in approximately 10 places." However, any correction that affects the meaning of a sentence or paragraph should be described in careful detail. The files samplethesis.tex and samplethesis.pdf show what the "Corrections" section should look like. Questions about what should appear in the "Corrections" should be directed to the Chair.

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl*. Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime*. Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.

Molina, S. T., & Borkovec, T. D. (1994). The Penn State worry questionnaire: Psychometric properties and associated characteristics. In G. C. L. Davey & F. Tallis (Eds.), *Worrying: Perspectives on theory, assessment and treatment* (pp. 265–283). New York: Wiley.

Noble, S. G. (2002). *Turning images into simple line-art* (Undergraduate thesis). Reed College.

Reed College. (2007). LaTeX your document. Retrieved from http://web.reed.edu/cis/help/LaTeX/index.html