Linear Models

Cason Wight

11/15/2019

This project is intended as a portfolio of simple applications of linear models under a few different scenarios, with data chosen by me. A linear model is always of the form $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Each application is a special case of this form of model. The five sections of this report include the following analyses:

- SRS Mean
- Two Means
- One Factor Experiment
- Two-Factor Experiment
- One-Factor Analysis of Covariance

For each of the above, this project provides an application description, data details, and an analysis in R. Each step is includes a brief explanation of methods and results, and why each is an example of $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

SRS Mean

A simple random sample (SRS) is a set of random draws from a popuulation of interest. The population has a true mean and true variance. Using the random sample, the mean can be estimated, along with a 95% confidence interval for the true mean using a t-test.

Application Description

The first data analysis will be of fire damage data in my hometown. I am from Washington County, Oregon (just outside of Portland). In the last few years, fires have been a big concern in the Pacific Northwest. The wildfires in Oregon and California have become infamous. Last year a wildfire in Oregon disrupted my family's planned white-water rafting trip.

According to oregon live, "the cost of fighting wildfires in Oregon reached an all-time high [of] 514.6 million in 2018." Statesman Journal reports over 1,800 fires, totaling 846,000 acres during that time. One fire in Klondike affected over 160,000 acres. I am curious as to what the true mean damage (as defined by total acres affected) by fires is in my hometown, along with other inference on that sample.

Data Details

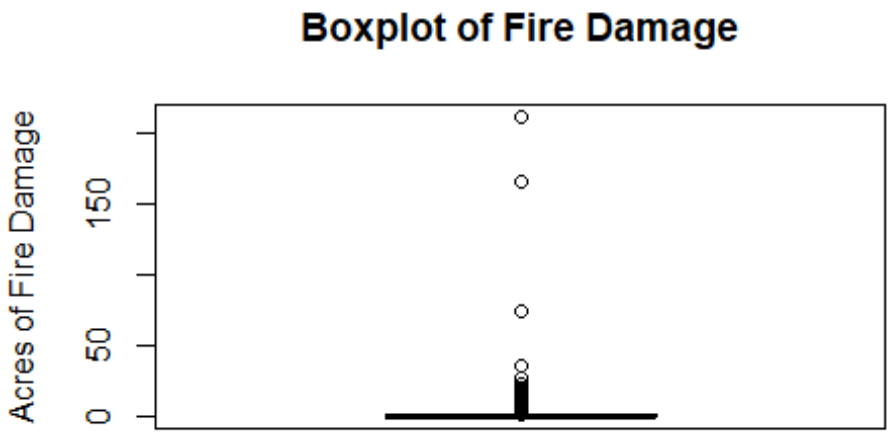
From the Oregon Department of Forestry's (ODF's) website, odf.oregon.gov, I obtained a list of all fires in Washington County reported to the ODF since 1960. This data includes a column

called “Total Acres” that reports how many acres were affected by each fire. A few adjustments were required for the data before analysis, due to two issues.

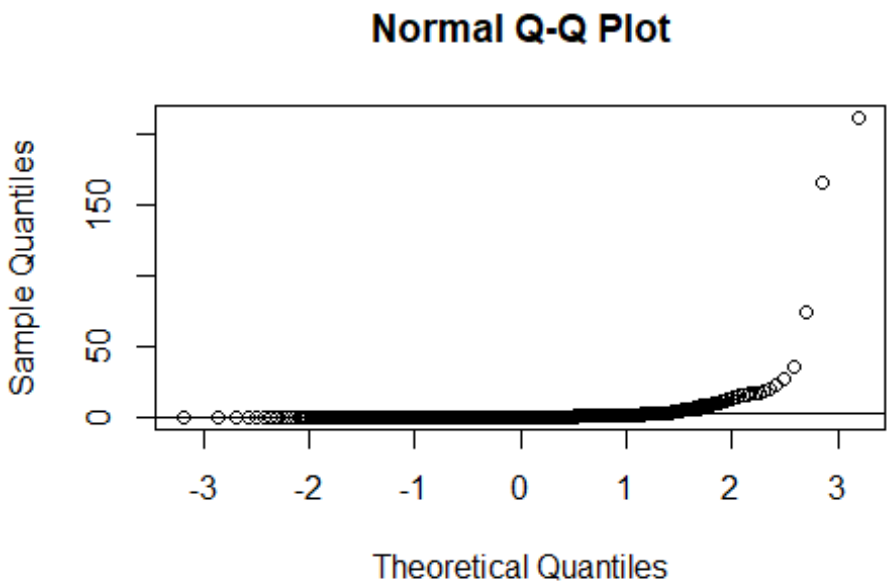
First, 67 of the 770 reported fires either have a reported 0 acres of damage or a blank value. For my analysis, I do not want to look at the reported fires with no damage, because most fires with no damage probably go unreported anyway. Thus, all of these observations are removed.

The other adjustment that was required was a log transformation on the acres damaged in each fire. This is because a t-test on this SRS requires the assumption of normality in the samples. The actual data clearly does not follow this assumption, and a log transformation helps. These two adjustments are labeled in the comments throughout the code.

These observations are reported by the Oregon Department of Forestry. Their website does not provide many details on the data collection, but I assume that every reported fire is included in the data. The following is a boxplot of the data (acres of fire damage).



This boxplot is heavily right-skewed. This may be because non-positive values are not included in this type of data and smaller fires are significantly more likely than bigger fires. This glance at the boxplot shows that the data are not normally distributed, which is an assumption of a SRS analysis. Quantile-quantile plots give a more detailed look into the normality of a dataset. The following is a quantile plot for the fire data.



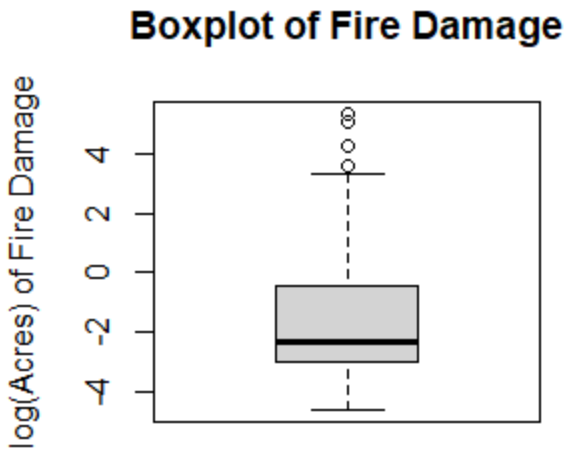
A look at the quantile plot confirms that the data is not normally distributed. Normally distributed data would approximately follow the line included in the quantile plot. A transformation on this data will be necessary for this analysis. Summary statistics of the untransformed data are included below.

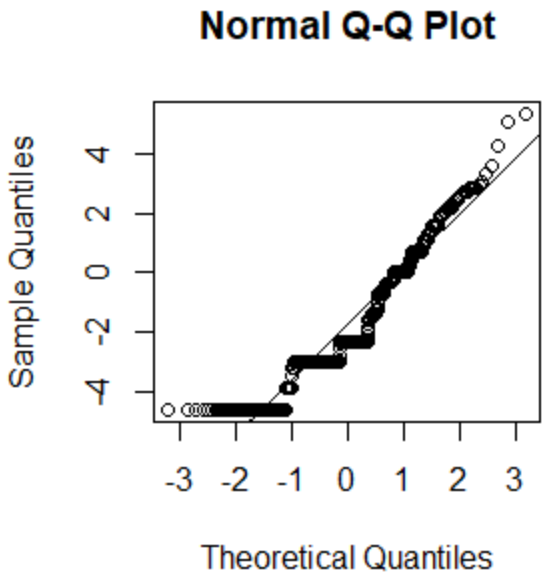
Summary Statistics of Acres

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
0.01	0.05	0.1	1.65	0.62	211	10.86

According to the summary statistics shown in the table, the maximum damage from a fire reported in Washington county is 211 acres, and the minimum is 0.01 acres. Half of the reported fires caused between 0.05 and 0.625 acres of damage.

To address the heavy skew and non-normality of this data, a log transformation is peformed. The following shows the effects of this transformation on the assumption of normality (through a boxplot and a quantile plot).





The log transformation helps the data fit the normality assumption much better. Although there appears to be some kind of discrete nature to these reports (which is impossible with normally distributed data), the log of the damage (in acreage) appears to be much closer to normally distributed, but less so in the tails. The transformation certainly helps reduce the skew from the original data. We will proceed with an analysis on these transformed data.

Analysis

Now that the data have been transformed, an analysis is performed on this SRS. First a t-test model is fit: $Y_i = \mu + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$ Using this model, an estimate for the true mean (log-transformed) acres of damage by a fire is calculated along with the 95% confidence interval for this true mean.

Estimated Avg. Damage (Acres)

Estimate	Lower	Upper
0.1416	0.1227	0.1633

The estimated true mean fire damage (in acres) for a fire in Washington County is 0.142 acres, with a 95 confidence interval of 0.123 to 0.163. This estimate is surprising to me. I expected the average damage to be a lot bigger, based on what I had heard.

Example of a Linear Model

This case is an example of $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\epsilon}$, with

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\epsilon}$$
$$\begin{bmatrix} \ln y_1 \\ \ln y_2 \\ \vdots \\ \ln y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [\mu] + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad where \quad \epsilon_i \sim N(0, \sigma^2)$$

Although this example includes a log transformation, it is still an example of a linear model. The $\mathbf{X}$ matrix is simply a column of 1s. $\boldsymbol{\beta}$ is a 1×1 matrix with the true mean μ. The $\mathbf{Y}$ matrix is the recorded log(acres) of damage from the data, which has n observations. The $\boldsymbol{\epsilon}$ matrix is an $n \times 1$ matrix of normally distributed error, centered on 0, with σ^2 variance.

The main results of this SRS is that the true mean acres of damage in Washington County, Oregon, is roughly 0.14 acres.

Two Means

A two means analysis compares two groups and tests the hypothesis that the true means of the groups are significantly different from each other. This test assumes that the samples from these two groups are normally distributed and that the two groups have equal (or at least similar) variance.

Application Description

Wrestling has always been a fun sport for me to view and participate in. Unfortunately, college wrestling is not as popular as other college sports. Many universities (like BYU) do not have a team. When I watch, I like to follow a specific team, especially in tournaments. Simply watching hundreds of wrestlers across multiple teams is far less exciting than having a specific team to follow. Because BYU does not have a team, I typically follow Utah Valley University (UVU) or Oregon State University (OSU).

Last year, I watched a few duels at UVU. It was frustrating to root for a losing team, especially one that I didn't know too well. This year, I want to watch the Reno Tournament of Champions (TOC) on December 15. I want to follow the "better" team of UVU and OSU, based on last year's performance at the same tournament.

I will analyze the team points earned or lost for each match of UVU wrestlers from last year's TOC, and compare that to the same for all of the OSU matches. By performing a two means analysis, I will be able to know which wrestlers' matches would have been more "exciting" to follow at last year's tournament between the two teams. This way, I may have a better idea of who to follow this year, if a "better" team exists.

Data Details

The data is from trackwrestling.com, selecting "Oregon State University" and "Utah Valley University" as the teams, and "Reno Tournament of Champions" from the event listing of both schools. The point system for wrestling matches are as follows:

- 6 points for a pin, forfeit, or disqualification
- 5 points for a technical fall (15 point lead)
- 4 points for a major decision (8-14 point lead)
- 3 points for a decision (1-8 point lead)

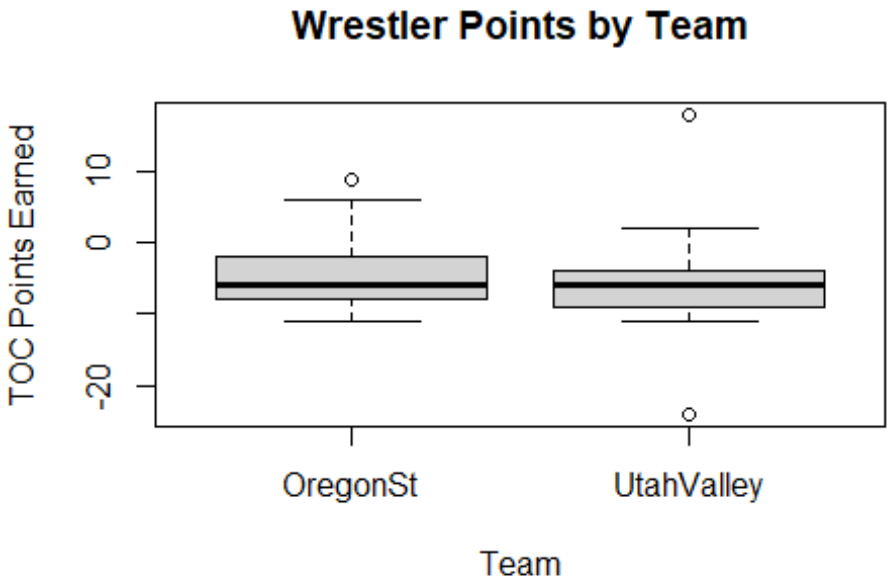
The scores reported in this data set include the total points won and given up in each match of all the wrestlers, summed up across the whole tournament. This means that if one wrestler, say Kaylor, has a total score of 9 team points, it means that over the course of the tournament he won a net of 9 points (earned points minus points earned by opponents). Negative scores mean that that wrestler gave up more team points then he earned during the tournament. The smaller the score, the worse the wrestler performed in the tournament and "less exciting" he was to watch.

The data was manually entered from the above-mentioned website. The scores for each match are summed up for each wrestler to give their total score for the tournament. Combining these two sets will give one dataset of all wrestlers from either team, along with their corresponding team and total scored points.

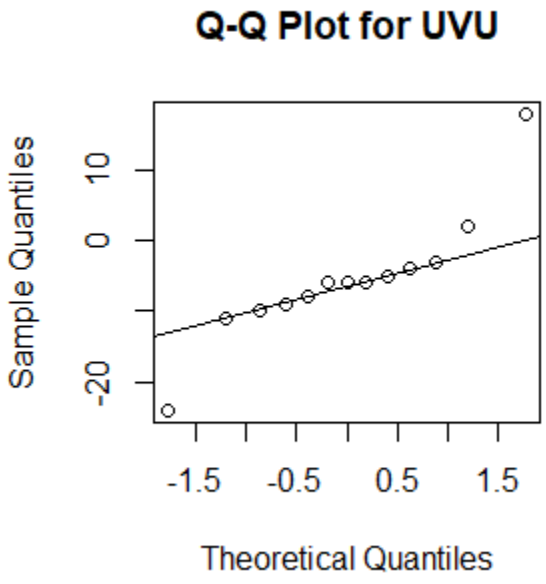
The following table shows the first few rows of the combined dataset, and the subsequent plot is a boxplot of the net points for each wrestler.

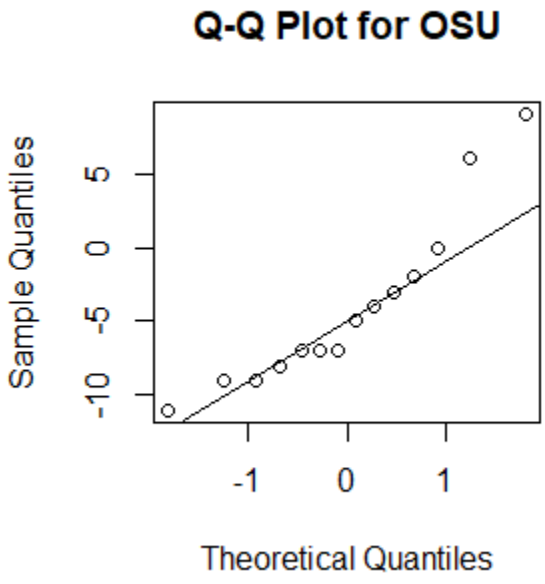
First Few Rows of Combined Data Set

Wrestler	TeamPoints	Team
Allen	-8	OregonSt
Beisley	-7	OregonSt
Bresser	-7	OregonSt
Dematteo	-7	OregonSt
Dixon	-4	OregonSt
Kaylor	9	OregonSt



An initial look at the boxplot gives the impression that the two teams did not have too different results last year. The median score and ranges are roughly equal for the two teams. It does look like the points are approximately normally distributed and that the two teams have approximately equal variance. Quantile plots for the two groups will give a better look at the assumption of normality.





These quantile plots do not bring up any concerns over the normality assumption. Summary statistics on these teams are also reported below, giving a better idea of the center and spread of the teampoints for the two teams at ;ast year's TOC.

Summary Statistics

	Min.	Mean	Max.	SD
OregonSt	-11	-4.07	9	5.77
UtahValley	-24	-5.54	18	9.26

The average team points for the two teams' wrestlers were -4.07 for OSU and -5.54 for UVU. The most team points from any wrestler was 18 (Trussell from UtahValley). The fewest from any wrestler was -24 (Steward from UtahValley). The variance of each wrestler's team points was not too different between the two teams.

Analysis

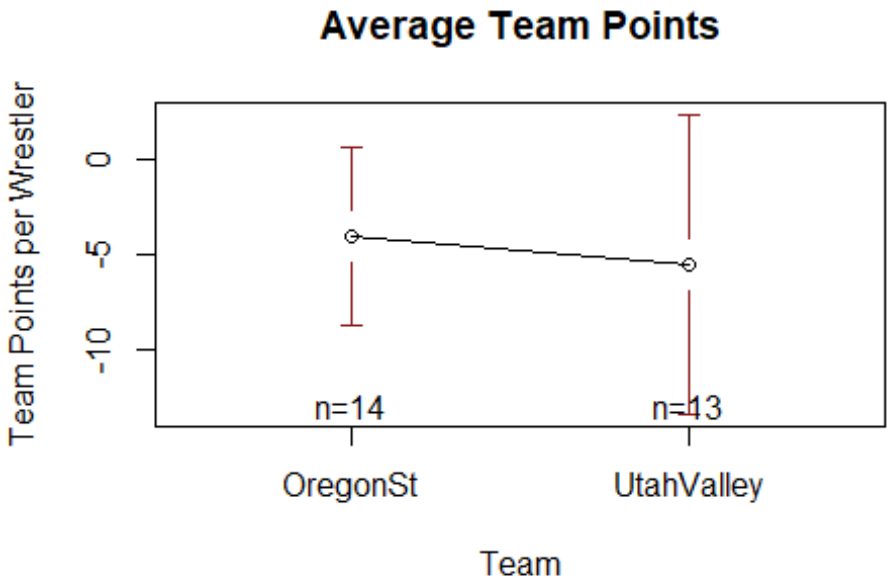
An analysis on these data will tell if there is a statistically significant difference between the means of these 2 groups. First, the model $y_{ij} = \mu_j + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$ is fit. The estimated difference (along with a 95 confidence) of the difference between the two means is determined through a t-test and reported in the table below.

Estimated Difference of Means

Estimated Difference	Lower	Upper
1.467	-4.6	7.534

The estimated difference in mean points scored for a wrestler from either team is 1.467 (95 CI: -4.6 to 7.534). This means that if I were to pick a random wrestler from OSU and from UVU, I would expect the OSU wrestler to score, on average, 1.467 more points than the UVU wrestler. However, the 95 confidence interval for that estimate includes 0, which means that there is insufficient evidence to say that this difference is significantly different from 0. For this year's TOC, there is no clear "better" choice between the two teams for me to watch.

A barplot shows the means for the 2 groups and their standard errors. This plot is shown below.



The plot for the two means confirms what was already concluded. The two team points do not give sufficient evidence to say that there is a significant difference between the true mean points earned by wrestlers from UVU and OSU. The 2 means are well within either's standard error.

Example of a Linear Model

This case is an example of $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\epsilon}$, with

$$\begin{aligned} \mathbf{Y} &= \mathbf{X\beta} + \boldsymbol{\epsilon} \\ \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 - Z_1 & Z_1 \\ 1 - Z_2 & Z_2 \\ & \vdots \\ 1 - Z_n & Z_n \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \text{where} \quad Z_i = \begin{cases} 0, & team = OregonState \\ 1, & team = UtahValley \end{cases} \end{aligned}$$

Each observation (wrestler) is either from OSU or UVU, and never from both. Thus each row should have one 0 and one 1, which is how the $\mathbf{X}$ matrix is defined above. Assuming $n = n_1 + n_2$, this is a $n \times 2$ matrix. The $\mathbf{Y}$ matrix is the observed total tournament points for each wrestler from the data, which has $n = n_1 + n_2$ observations. The $\boldsymbol{\epsilon}$ matrix is an $n \times 1$ matrix of normally distributed error, centered on 0, with σ^2 variance. The $\boldsymbol{\beta}$ here is simply the true mean points scored/lost during the TOC for wrestlers from either team, making a 2×1 matrix. Each observation has its own normally-distributed error term added on, like in most linear models.

One-Factor Experiment

An analysis of variance (ANOVA) can be used to analyze a one-factor experiment. These experiments have one dependent variable of interest, with predetermined levels for this factor.

The goal is to assess the quantitative effect of being in each predefined "group" of the independent variable on the dependent variable.

Application Description

I do not like working on my computer for long periods of time without listening to something as I work. Most people like listening to music as they work. I like listening to podcasts and other forms of media. Some of my favorite podcasts include Radiolab by NPR, Revisionist History by Malcolm Gladwell, and Making Sense with Sam Harris.

I have never liked listening to music while working, and I tend to think that it is more distracting than podcasts. I want to test the effects of listening to podcasts or music on my own productivity. Quantifying productivity is rather difficult, so I decided to measure how many characters per minute (CPM) I could accurately type under different conditions (silence, music, and podcasts) instead. For each of the three conditions, I did a 1-minute typing test to see how many characters I could type.

Data Details

This data comes from an experiment I conducted, the layout of which is as follows:

Response Variable Accurately typed characters per minute (CPM) from a typing test

Factor (Levels) Factor: Sound condition (nothing, music, podcast)

Experimental Unit For each trial, set the sound condition for a few minutes (on headphones), and then perform an online, 1-minute typing test while maintaining the conditions. The CPM of the test is the result for the observation.

Replication Each factor level combination had 5 replicates.

Randomization The order of the tests were randomly assigned. Several website provide typing speed tests. I decided to use livechatinc.com, which has a good format. For this experiment, the music I used was from Miike Snow (indie pop), an artist that I like. The podcast that I listened to was NPR's Radiolab. On each observation, I allowed myself to listen for 1 minute before taking the test to "get into" whatever I was listening to. Each observation of data includes a sound condition (Nothing, Music, or Podcast) and a recorded CPM from a test.

The following is a glimpse at the first few rows of the data and boxplots of the CPM data for the different groups. Quantile plots also are provided to show the normality of the data from the three groups, as in the previous analysis.

First Rows of Typing Test Data

cpm	Condition
243	Music

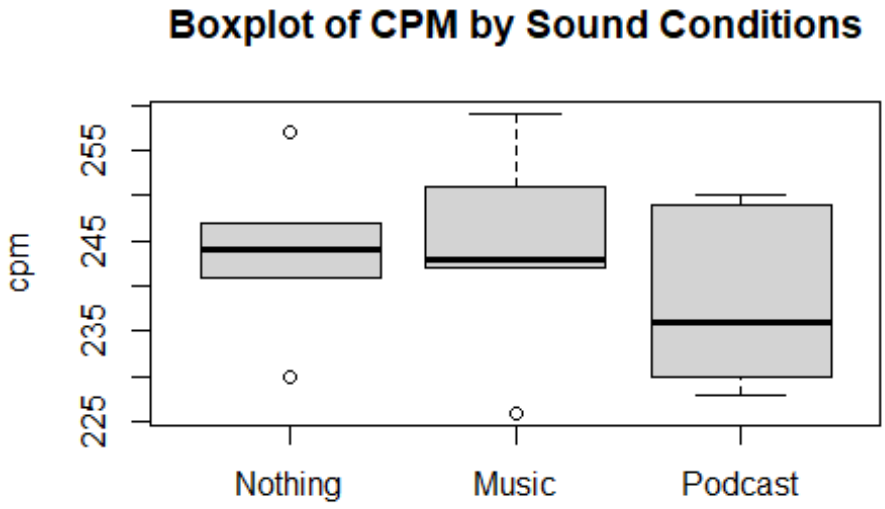
242 Music

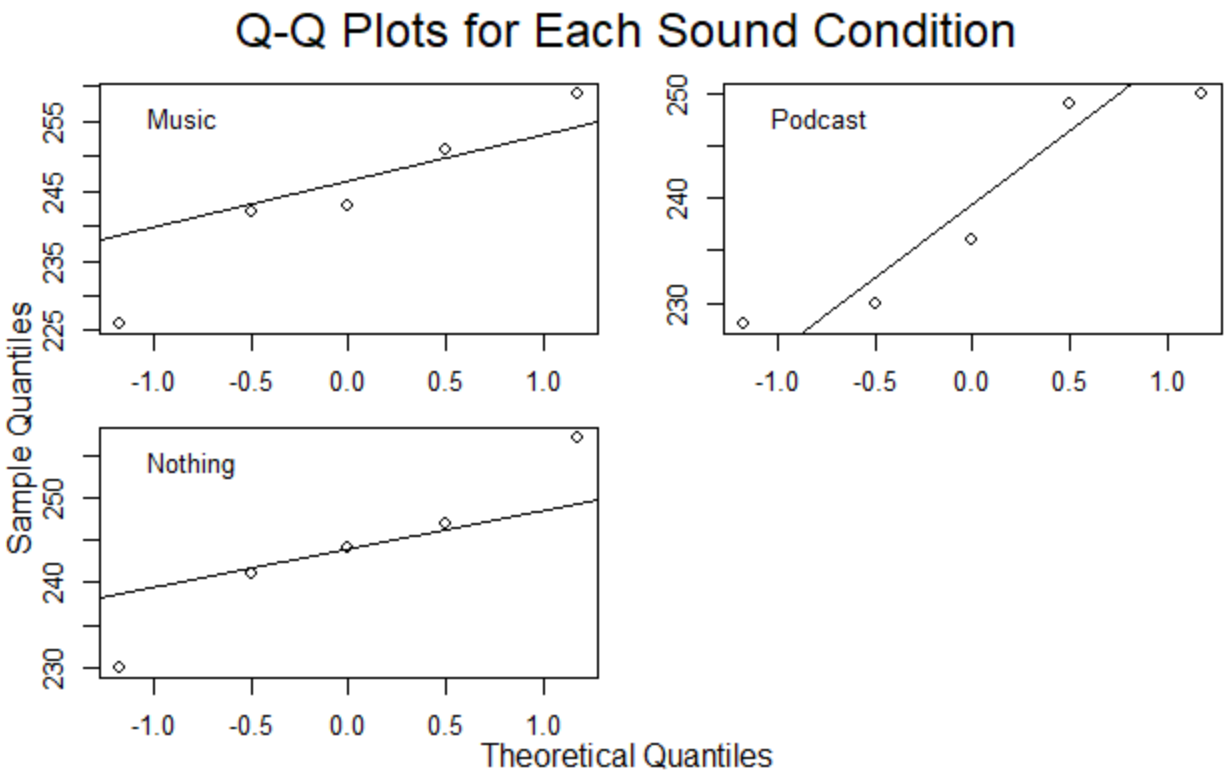
250 Podcast

249 Podcast

241 Nothing

244 Nothing





From the above boxplots and quantile plots, the assumption of normality for these samples does not appear to be violated. The music group may have a little bit of a skew, but nothing to cause concern, especially with only 5 samples. The other 2 groups are approximately symmetrically distributed. The summary statistics of these data by group are included below.

Summary Statistics

Condition	Mean CPM	Std Dev	N
Nothing	243.8	9.78	5
Music	244.2	12.28	5
Podcast	238.6	10.38	5

The mean CPM for listening to nothing, music, and a podcast are 243.8, 244.2, and 238.6 respectively. The standard deviations are relatively similar to each other. By the looks of the summary statistics, the three groups do not have obvious differences in the CPM typed, but an ANOVA will give more details.

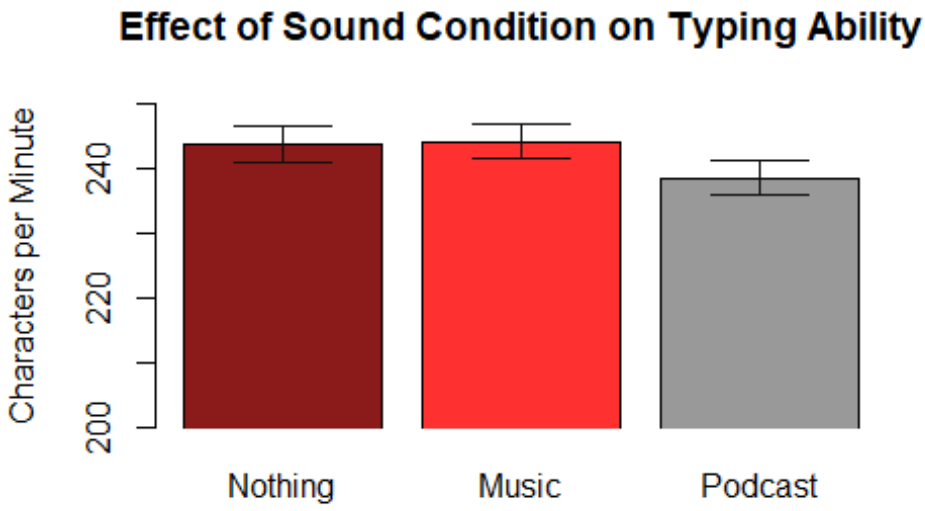
Analysis

The first step for this analysis is fitting the ANOVA model: $y_{ij} = \mu + \alpha_i * condition + \epsilon, \epsilon \sim N(0, \sigma^2)$. The output of ANOVA is included in the table below.

ANOVA Output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
condition	2	97.6	48.8000	0.4133	0.6705
Residuals	12	1416.8	118.0667	NA	NA

As predicted from a look at the summary statistics, the model shows no evidence of any effect of any of the listening conditions on how many characters I can type in a minute (p-value: 0.6705). The following barplot shows just how similar the three groups are, with bands of their standard errors.



Although my experiment of typing tests may not fully reflect how listening conditions can affect productivity, it is likely a good indicator that the impact is less severe than I originally thought. At a minimum, there is no evidence that listening to music or a podcast while typing would slow me down or affect my productivity. Also, my bias against listening to music while typing remains without evidence.

Example of a Linear Model

This model is also linear, but needs some adjustment from the intuitive approach to an ANOVA model. A first approach to this model would likely give

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{X\beta} + \boldsymbol{\epsilon} \\
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & Z_{11} & Z_{21} & Z_{31} \\ 1 & Z_{12} & Z_{22} & Z_{32} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & Z_{1n} & Z_{2n} & Z_{3n} \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \\
\text{where } Z_{1i} &= \begin{cases} 1, & sound = Nothing \\ 0, & otherwise \end{cases}, Z_{2i} = \begin{cases} 1, & sound = Music \\ 0, & otherwise \end{cases}, \\
\text{and } Z_{3i} &= \begin{cases} 1, & sound = Podcast \\ 0, & otherwise \end{cases}
\end{aligned}
$$

Unfortunately, this naive model produces linearly dependent columns in the X matrix. One solution is to remove the second column, which affects the interpretation of $\boldsymbol{\beta}$ but fits the same model. The correct linear interpretation of the one-factor analysis is as follows:

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{X\beta} + \boldsymbol{\epsilon} \\
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & Z_{11} & Z_{21} \\ 1 & Z_{12} & Z_{22} \\ \vdots & \vdots & \vdots \\ 1 & Z_{1n} & Z_{2n} \end{bmatrix} \begin{bmatrix} \mu + \alpha_1 \\ \alpha_2 - \alpha_1 \\ \alpha_3 - \alpha_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \\
\text{where } Z_{1i} &= \begin{cases} 1, & sound = Music \\ 0, & otherwise \end{cases}, \text{ and } Z_{2i} = \begin{cases} 1, & sound = Podcast \\ 0, & otherwise \end{cases}
\end{aligned}
$$

In this model, the elements of $\boldsymbol{\beta}$ represent the true mean plus the effect of having no sound in the first element, the difference between the effects of no sound and having music in the second, and the difference between the effects of no sound and having podcast in the third. The $\mathbf{Y}$ matrix is the recorded CPM from the experiment, which has n observations. The $\boldsymbol{\epsilon}$ matrix is an $n \times 1$ matrix of normally distributed error, centered on 0, with σ^2 variance.

Two-Factor Experiment

A two-factor experiment brings another variable into the previous type of analysis. On top of another effect to review, these experiments also allow for a look at the potential interaction effect between the two variables. An interaction means that the effect of one of the factors changes based on the level of the other factor.

Application Description

A few years ago in Stat 230, I tested the effects of the type of "mug", type of liquid, and any interaction between the two on how hot a microwave can get hot chocolate after 90 seconds in the microwave. With Todd Okeson and Eric McGill's permission, this data is included as an example of a two-factor experiment. In this experiment, hot chocolate powder was mixed into one of three different liquids, and into different types of cups. These hot chocolate mixtures were heated in microwaves for 90 seconds. Thermometers measured the temperatures immediately following the 90 seconds.

Data Details

This data comes from an experiment I conducted with my group, the layout of which is as follows:

Response Variable Temperature of hot chocolate (degrees Farenheit)

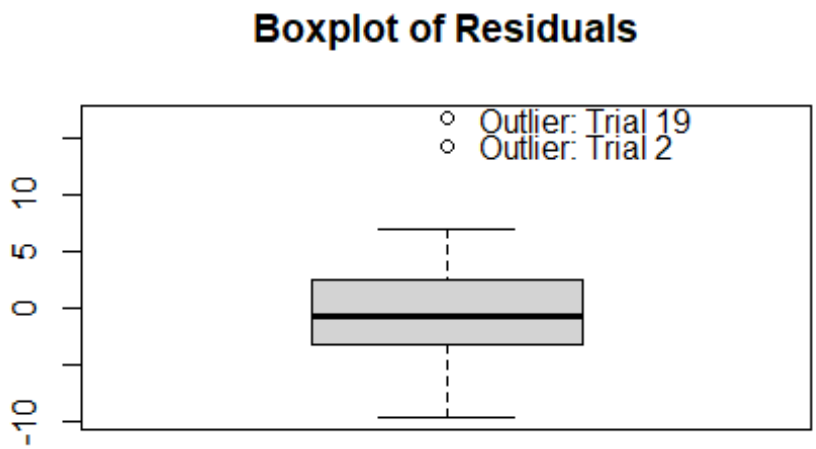
Factor (Levels) Factor 1: Cup (plastic cup, ceramic mug) Factor 2: Liquid (water, 2% milk, almond milk)

Experimental Unit For each liquid-cup combination, mix the hot chocolate for 15 seconds before microwaving for 90 seconds. The liquid is then measured with a thermometer (not touching the sides of the cup) for 20 seconds.

Replication Each factor level combination had 6 replicates.

Randomization We took several precautions to avoid potential lurking variables. For example, we collected several mugs of the different types from varying locations, and then randomly selectewd out of those a sample to experiment on. We also bought our milks and got our waters from different sources. We stored the liquids in the same type of jugs overnight in the same fridge so that they would all start at similar temperatures. The trials were randomized and we used room-temperature water baths for the hot chocolate whisks in between trials. Other measures were also taken to minimize latent effects.

A boxplot of the results of this experiment are included below.



The two outliers from this experiment are trials 19 and 2. They are both from trials with plastic cups, so they will not be excluded from this analysis, in case they represent an important trend. Summary statistics of these data are included below.

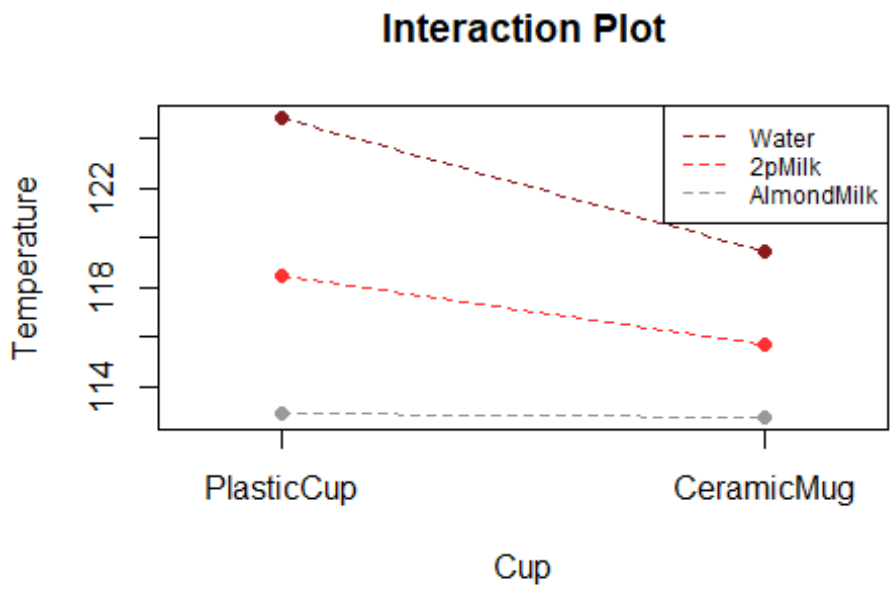
Summary Statistics of Hot Chocolate Data

Liquid	Cup	Mean Temp	Min Temp	Max Temp	Std Dev of Temp
Water	PlasticCup	124.8167	118.2	130.3	3.9867
2pMilk	PlasticCup	118.4500	114.3	132.8	7.1023
AlmondMilk	PlasticCup	112.9167	104.0	129.7	9.5690
Water	CeramicMug	119.4667	115.7	122.2	2.8232
2pMilk	CeramicMug	115.6500	108.0	120.5	5.2558
AlmondMilk	CeramicMug	112.7667	103.2	119.8	5.4416

```
getName <- function(cup, liquid) {
  cupL <- substr(cup,1,1)
  liquidL <- substr(liquid,1,1)
  cup.out <- ifelse(cupL=="P", "plastic cup", "ceramic mug")
  liquid.out <- ifelse(liquidL=="W","water", ifelse(liquidL=="2", "2% milk", "almond milk"))
  c("c"=cup.out,"l"=liquid.out)
}
```

From the two factors, 6 factor level combinations are possible. The overall minimum temperature was 103.2 (from almond milk in a ceramic mug). The overall maximum temperature was 132.8 (from 2% milk in a plastic cup). The different groups have relatively similar variance of temperatures and the means for the different groups range between 112.77 and 124.82.

An interaction plot displays the potential for an interaction effect. This type of plot shows the means of all of the different factor-level combinations. If lines cross, this is typically a strong indication of an interaction effect. The interaction plot from this experiment is shown below.



Although the lines of these effects are not parallel, the lines do not cross. This most likely means that evidence for an interaction effect is inconclusive.

Analysis

To get a further look at main and interaction effects, first the ANOVA model is fit: $y_{ijk} = \mu + \beta_{1i} * Liquid + \beta_{2j} * Cup + \beta_{3ij} * Liquid * Cup + \epsilon_k, \epsilon \sim N(0, \sigma^2)$. Output from this model is included in the folowing table.

ANOVA Output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Liquid	2	520.5006	260.2503	6.9989	0.0032
Cup	1	68.8900	68.8900	1.8527	0.1836
Liquid:Cup	2	40.5650	20.2825	0.5455	0.5852
Residuals	30	1115.5333	37.1844	NA	NA

The effect of liquid on temperature appears to be significant (p-value: 0.0032). There is no evidence for a significant effect of cup type on temperature (p-value: 0.1836) and no evidence of an interaction effect between liquid and cup type on temperature (p-value: 0.5852). Because there is no evidence of an effect, we will exclude the interaction term from the model to fit the new model: $y_{ijk} = \mu + \beta_{1i} * Liquid + \beta_{2j} * Cup + \epsilon_k, \epsilon \sim N(0, \sigma^2)$.

ANOVA Output on Simplified Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Liquid	2	520.5006	260.2503	7.2035	0.0026
Cup	1	68.8900	68.8900	1.9068	0.1769
Residuals	32	1156.0983	36.1281	NA	NA

This new model yields the same conclusions of strong evidence of an effect of liquid on temperature (p-value: 0.0026) and no significant evidence of cup type on temperature (p-value: 0.1769). A Tukey pairwise comparison test looks at which levels of the liquid factor are significantly different. A table of these pairwise comparisons is shown below.

Tukey Pairwise Comparisons

	diff	lwr	upr	p adj
2pMilk-Water	-5.09167	-11.12167	0.93834	0.11115
AlmondMilk-Water	-9.30000	-15.33000	-3.27000	0.00177
AlmondMilk-2pMilk	-4.20833	-10.23834	1.82167	0.21523

From the table, the only pair that have evidence of a significant difference is almond milk against water. The temperature of hot chocolate made with water is an estimated 9.3 degrees hotter than hot chocolate made with almond milk (p-value: 0.0018). Although not statistically significant, the temperature of hot chocolate made with water is an estimated 5.1 degrees hotter than hot chocolate made with 2% milk (p-value: 0.1112).

Example of a Linear Model

This case is an example of $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\epsilon}$, with

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & Z_{11} & Z_{21} & Z_{31} & Z_{11}*Z_{31} & Z_{21}*Z_{31} \\ 1 & Z_{12} & Z_{22} & Z_{32} & Z_{12}*Z_{32} & Z_{22}*Z_{32} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & Z_{1n} & Z_{2n} & Z_{3n} & Z_{1n}*Z_{3n} & Z_{2n}*Z_{3n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

$$\text{where } Z_{1i} = \begin{cases} 1, & liquid = 2\%Milk \\ 0, & otherwise \end{cases}, Z_{2i} = \begin{cases} 1, & liquid = AlmondMilk \\ 0, & otherwise \end{cases},$$

$$\text{and } Z_{3i} = \begin{cases} 1, & Cup = CeramicMug \\ 0, & otherwise \end{cases}$$

The linear model for two-factor experiments are like those of one-factor experiments, with a few additional columns in the $\mathbf{X}$ matrix and a few additional coefficients in the $\boldsymbol{\beta}$ matrix. First of all, columns are added for the second factor (one column for every level after the first). Interaction columns are added, as products of the columns that correspond to each factor-level combination. In the $\boldsymbol{\beta}$ matrix, the elements are as follows:

- β_0 is the true mean temperature plus the effect of water in a plastic cup (and it's interaction)
- β_1 is the difference between the effect of water and the effect of 2% Milk (and the interaction with water)
- β_2 is the difference between the effect of water and the effect of almond milk (and the interaction with water)
- β_3 is the difference between the effect of a plastic cup and the effect of a ceramic mug (and the interaction with water)
- β_4 is the difference between the interaction effect of water and a plastic cup and the interaction effect of 2% milk and a ceramic mug
- β_5 is the difference between the interaction effect of water and a plastic cup and the interaction effect of almond milk and a ceramic mug

Thus for any row of the X matrix, the effect of a single level of liquid, of a single level of cup, and the interaction are included. The $\mathbf{Y}$ matrix is the recorded temperature of each hot chocolate from the experiment, which has n observations. The $\boldsymbol{\epsilon}$ matrix is an $n \times 1$ matrix of normally distributed error, centered on 0, with σ^2 variance.

One-Factor Analysis of Covariance

A one-factor analysis of covariance is simply a one-factor ANOVA, after accounting for a second, continuous variable.

Application Description

Many interesting characteristics could possibly affect the education of young students. Many think that stay-at-home moms make for the best environment for kids. The theory is that moms who are at home (not working) have more time to teach and nurture kids, who will then do better in school. I would love to explore this theory.

Another interesting characteristic to look at is the effect of absences on student performance. An intuitive guess is that students who don't come to class have their grades negatively impacted, but the actual effect needs to be explored. An interaction between number of absences and the students' mother's occupation will also be addressed.

Data Details

This data was originally collected by Dr. Paulo Cortez, but was obtained in 2019 from UC Irvine's Machine Learning Repository. This particular data set was released in 2014. It includes roughly 30 attributes about a group of about 650 Portuguese middle-school students. Out of these attributes, 2 will be analyzed (the two variables Mjob and absences). Datasets on results for Math and Portuguese classes are given, but this analysis focuses only on the grades for math.

The student's grades are reported per semester as G1, G2, or G3. These grades are scored on a 0-20 scale.

We want only the columns of interest: Mother's job, number of absences, and trimester grades. We want the average grade across the 3 trimesters instead of 3 separate outcomes across the different trimesters. The mother's job will be releveled to compare each level to at-home mothers. For simplicity, the data is subset to only rows with 3 of the possible options of mother's jobs: at-home mothers, mothers working in health, and teaching mothers.

A glimpse at a few rows of these updated data are shown in the folowing table.

First Few Rows of Subset Student Data

	Mjob	absences	avg.grades
1	at-home	6	5.6667
2	at-home	4	5.3333
3	at-home	10	8.3333
4	health	2	14.6667
13	health	2	14.0000
14	teacher	2	10.3333

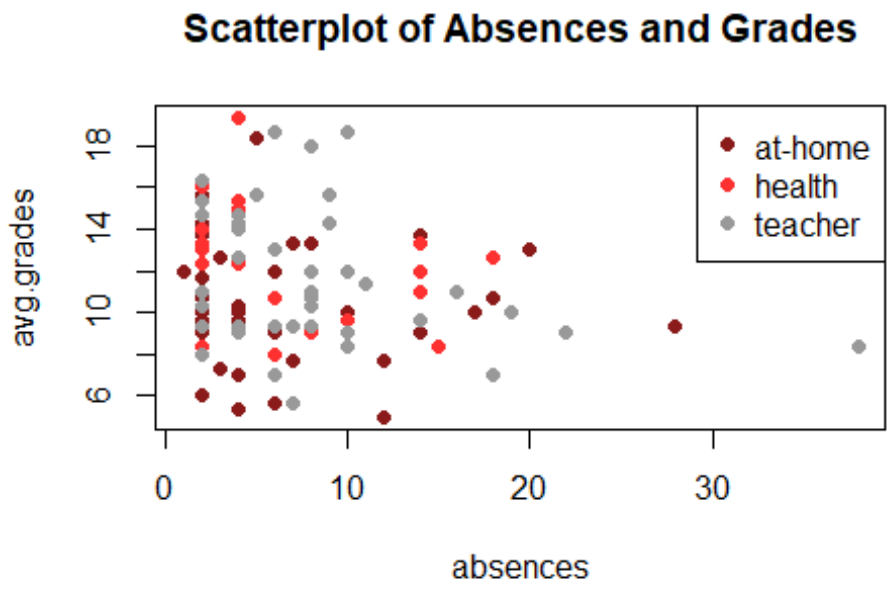
The next step is to look at the summary statistics of grades, split by mother's work. These statistics are summarized in the following table.

Summary Statistics of Grades

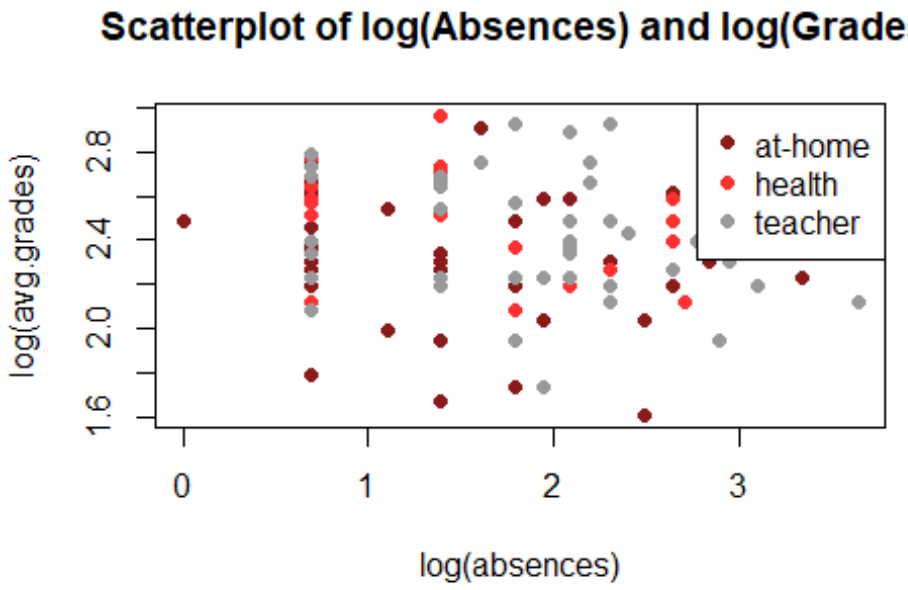
Mother's Job	Mean	SD	Min	Max	n
at-home	10.37209	2.914305	5.000000	18.33333	43
health	12.20000	2.896358	8.000000	19.33333	25
teacher	11.75194	3.278189	5.666667	18.66667	43

The 3 group means are similar (around 10-12) and the standard deviations are also comparable (around 3). Maximum possible grade. The overall minimum average grade is 5 and the overall maximum grade is 19.33.

The following scatterplot summarizes all of the grades, by number of absences and shows the need for transformations, due to non-linearity.

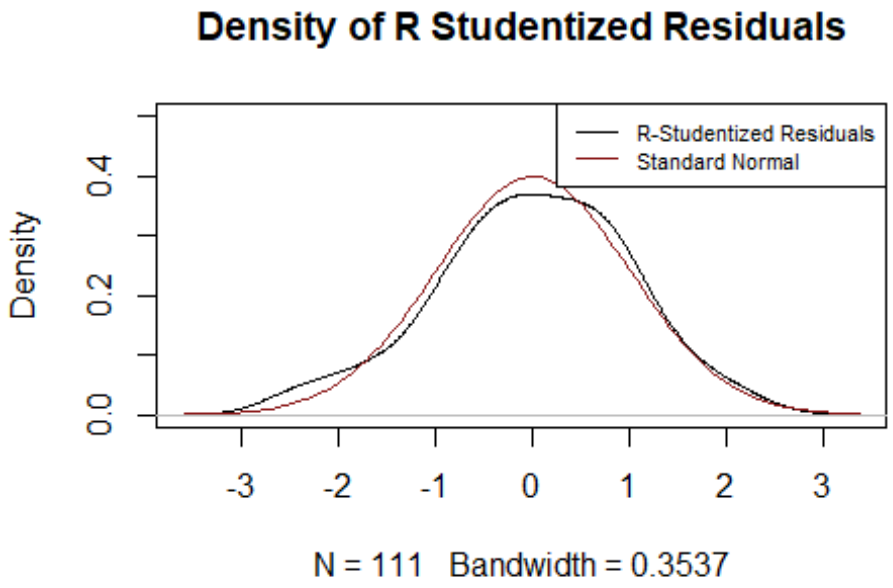


This scatterplot shows that a log transformation of both absences and average grades is required. The data should look like a linear scatterplot, but instead looks like it has a multiplicative effect. The following is the same plot on log scale, labeled by mother's job. Because $\log(0) = -\infty$, the students with 0 absences are excluded from this analysis.



This plot looks much more linear, which is an assumption of an analysis of covariance of this type. There is no clear trend between the three types of mothers' working situations. The following plot shows the R studentized residuals of a model on these data, as compared to a

standard normal distribution. If the two densities are roughly equal, than this model's assumtion of normality is filled.



Because the two lines are roughly equal, the above plot shows that the assumption of normality has no red flags.

Analysis

As always, the first step of the analysis after inspecting the data is to fit the model. We will fit the model: $\log(grade) = \beta_0 + \beta_1 * health + \beta_2 * teacher + \beta_3 * \log(absences) + \beta_4 * health * \log(absences) + \beta_5 * health * \log(absences) + \epsilon_{ij}$. The following table reports the resulting model output.

Model Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.3693	0.0891	26.5863	0.0000
log(absences)	-0.0461	0.0517	-0.8913	0.3748
Mjobhealth	0.2222	0.1546	1.4371	0.1537
Mjobteacher	0.2326	0.1363	1.7069	0.0908
log(absences):Mjobhealth	-0.0302	0.0902	-0.3354	0.7380
log(absences):Mjobteacher	-0.0548	0.0747	-0.7330	0.4652

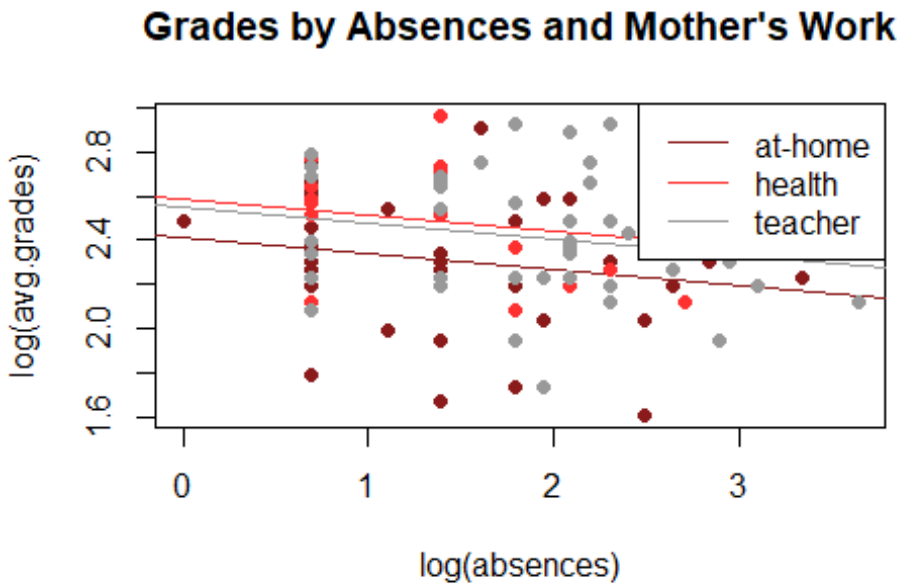
There is no evidence of any interaction effect between the mother's job and absences on students' grades (p-values of 0.738 and 0.4652 for jobs in health or teaching, respectively), so those terms will be excluded from the model. We will now fit the model $\log(grade) = \beta_0 + \beta_1 * health + \beta_2 * teacher + \beta_3 * \log(absences) + \epsilon_{ij}$ excluding the interaction terms. The output from this model is shown in the following table.

Model Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.4104	0.0653	36.9205	0.0000
log(absences)	-0.0731	0.0331	-2.2092	0.0293
Mjobhealth	0.1761	0.0685	2.5711	0.0115
Mjobteacher	0.1430	0.0592	2.4157	0.0174

For every 1% increase in absences, a student's grades will decrease by an estimated 7.3% (p-value: 0.0293). The grades of a student whose mother works in health (as opposed to a stay-at-home mom) has an estimated increase of 17.6% (p-value: 0.0115). The grades of a student whose mother works as a teacher (as opposed to a stay-at-home mom) has an estimated increase of 14.3% (p-value: 0.0174).

The following plot shows the effects of absences and mothers' job on students' grades.



A look at the three distinct lines of the effect plot shown makes it seem that students with moms working in health or working as teachers have significantly better grades, but not too different

from eah other. Because there is no significant evidence of any interaction effect between mother's job and number of absences on grades, the slopes of the three lines are the same. The slope of the three lines representes the incremental effect of more absences on grades. The shifts between the lines represents the relative effects of students having a mom in one of the three working situations.

A reduced model is fit, and compared to the full model in an ANOVA. This type of fit shows if the extra terms included in the full model are statistically significant. The following table reports the model output.

Results of ANOVA Testing Effect of Mother's Work

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
109	8.5825	NA	NA	NA	NA
107	7.9345	2	0.6479	4.3688	0.015

The ANOVA comparing a model with just absences to one with absences and mother's job reveals that the additive effect of a students' mother's job has a significant effect on grades (p-value: 0.015). This analysis gives compelling evidence that students with at-home mothers do not necessarily get better grades (at least in math) than those with moms working in health or as teachers. A lot could be done to perform a more rigorous analysis, and an experiment would be necessary to describe any causal relationships.

Example of a Linear Model

This case is an example of $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\epsilon}$, with

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} \ln y_1 \\ \ln y_2 \\ \vdots \\ \ln y_n \end{bmatrix} = \begin{bmatrix} 1 & \ln X_1 & Z_{11} & Z_{21} & \ln X_1 * Z_{11} & \ln X_1 * Z_{21} \\ 1 & \ln X_2 & Z_{12} & Z_{22} & \ln X_2 * Z_{12} & \ln X_2 * Z_{22} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & \ln X_n & Z_{1n} & Z_{2n} & \ln X_n * Z_{1n} & \ln X_n * Z_{2n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

where $Z_{1i} = \begin{cases} 1, & health \\ 0, & otherwise \end{cases}$ and $Z_{2i} = \begin{cases} 1, & teacher \\ 0, & otherwise \end{cases}$

This example of a linear model includes log transformations of both the repsonse variable (average grades) and the continuous explanatory variable (absences). The $\mathbf{X}$ matrix in this case is extremely similar to that of the two-factor experiment. One difference is that this $\mathbf{X}$ matrix has columns with continuous values for one of the explanatory variables instead of all indicator functions. The $\mathbf{Y}$ matrix is the recorded average grades of each of the n students. The $\boldsymbol{\epsilon}$ matrix is an $n \times 1$ matrix of normally distributed error, centered on 0, with σ^2 variance. The $\boldsymbol{\beta}$ is also similar.

Conclusion

Linear models are extremely useful for analysing and interpreting real data. Experimental designs are the most useful to infer causal relationships, but inferences can be drawen even from observational data using linear models. The examples included in this application portfolio are just a few simple examples, but many more exist. Many times, a model may not initially appear to fit the assumption of a linear model (like non-normality or an apparent multiplicative effect), but may be transformed in such a way that a linear model is still an effective tool.