Intro to Regression Analysis

Maria Tackett

2019-05-14

Contents

Chapter 1

Beginning of the Book

This is the introduction to the book.

Chapter 2

Getting Started

This is a chapter about getting started using R and GitHub.

Chapter 3

Simple Linear Regression

3.1 Computing: College Admissions

The primary goal of today's lab is to give you practice with some of the tools you will need to conduct regression analysis using R. An additional goal for today is for you to be introduced to your teams and practice collaborating using GitHub and RStudio.

3.1.1 Packages

We will use the following packages in today's lab.

3.1.2 Data

In today's lab, we will analyze the `scorecard` dataset from the rcfss package. This dataset contains information about 1849 colleges obtained from the Department of Education's College Scorecard. Load the rcfss library into the global R environment and type `?scorecard` in the **console** to learn more about the dataset and variable definitions. Today's analysis will focus on the following variables:

`type`	Type of college (Public, Private - nonprofit, Private - for profit)
`cost`	The average annual cost of attendance, including tuition and feeds, books and supplies, and living expenses, minus the average grant/scholarship aid
`admrate`	Undergraduate admissions rate (from 0 - 100%)

3.1.3 Exercises

3.1.3.1 Exploratory Data Analysis

1. Plot a histogram to examine the distribution of `admrate`. What is the shape of the distribution?
2. To better understand the distribution of `admrate`, we would like calculate measures of center and spread of the distribution. Fill in the code below to use the `skim` function to calculate summary statistics for `admrate`. Report the appropriate measures of center (mean or median) and spread (standard deviation or IQR) based on the shape of the distribution from Exercise 1.

3. Plot the distribution of `cost` and calculate the appropriate summary statistics. Describe the distribution of `cost` (shape, center, and spread) using the plot and appropriate summary statistics.

4. One nice feature of the `skim` function is that it provides information about the number of observations that are missing values of the variable. How many observations have missing values of `admrate`? How many observations have missing values of `cost`?

5. Later in the semester, we will techniques to deal with missing values in the data. For now, however, we will only include complete observations for the remainder of this analysis. We can use the `filter` function to select only the rows that values for both `cost` and `admrate`.

Fill in the code below to create a new dataset called `scorecard_new` that only includes observations with values for both `admrate` and `cost`.

You will use `scorecard_new` for the rest of the lab.

6. Create a scatterplot to display the relationship between `cost` (response variable) and `admrate` (explanatory variable). Use the scatterplot to describe the relationship between the two variables.

7. The data contains information about the type of college, and we would like to incorporate this information into the scatterplot. One way to do this is to use a different color marker for each type of college. Fill in the code below the scatterplot from the previous exercise with the marker colors based on the variable `type`. Describe two new observations from this scatterplot that you didn't see in the previous plot.

3.1.3.2 Simple Linear Regression

8. Fit a regression model to describe the relationship between a college's admission rate and cost. Use the `tidy` function to display the model.

9. Interpret the slope in the context of the problem. Does the intercept have a meaningful interpretation? If so, write the interpretation in the context of the problem. Otherwise, explain why the interpretation is not meaningful.

10. While the `tidy` function is used to display the model, we can obtain a one-row summary of the model using the `glance` function. Use the `glance` function to get a summary of the model fit in the previous exercise. See the documentation for `glance` for the syntax and a list of values output from the function.

11. What is the value of R^2? Interpret this value in the context of the problem. Do you think this is a "good" value of R^2? Explain.

12. What is the value of $\hat{\sigma}$, the residual standard error.

13. What is the 95% confidence interval for the coefficient of `admrate`, i.e. the slope? Interpret the interval in the context of the data.

14. We want to test the following hypotheses about the population slope β_1:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

State what the null and alternative hypotheses mean in terms of the linear relationship between `admrate` and `cost`.

15. Consider the confidence interval from Exercise 13 and the hypotheses in Exercise 14. Is the confidence interval consistent with the null or alternative hypothesis? Briefly explain.

3.2 In-Class Exercise: Advertising Analysis

In this mini analysis, we will work with the `Advertising` data used in Chapters 2 and 3 of *Introduction to Statistical Learning*.

3.2.1 Data and packages

We start with loading the packages we'll use.

We will analyze the advertising and sales data for 200 markets. The variables we'll use are

- `tv`: total spending on TV advertising (in $thousands)
- `radio`: total spending on radio advertising (in $thousands)
- `newspaper`: total spending on newspaper advertising (in $thousands)
- `sales`: total sales (in $millions)

3.2.2 Analysis

We'll begin the analysis by getting quick view of the data:

Next, we can calculate summary statistics for each of the variables in the data set.

1. What type of advertising has the smallest median spending?
2. What type of advertising has the largest variation in spending?
3. Describe the shape of the distribution of `sales`.

We are most interested in understanding how advertising spending affect sales. One way to quantify the relationship between the variables is by calculating the correlation matrix.

1. What is the correlation between `radio` and `sales`? Interpret this value.
2. What type of advertising has the strongest linear relationship with `sales`?

Below are visualizations of `sales` versus each explanatory variable.

Since `tv` appears to have the strongest linear relationship with `sales`, let's calculate a simple linear regression model using these two variables.

1. Write the model equation.
2. Interpret the intercept in the context of the problem.
3. Interpret the slope in the context of the problem.

3.3 In-Class Exercise: Beer Data Analysis

In this analysis, we will analyze the relationship between the amount of alcohol (`PercentAlcohol`) and the caloric content (`CaloriesPer12Oz`) in domestic beers. Let `PercentAlcohol` be the predictor variable and `CaloriesPer12Oz` the response variable.

Due to limited class time, we will not do the exploratory data analysis in this example. In practice, however, you should always start with the exploratory data analysis.

You can add your answers to this R Markdown document.

1. Calculate a regression model to describe the relationship between `PercentAlcohol` and `CaloriesPer12Oz`. Display the model output.

2. Does it make sense to interpret the intercept? Why or why not?

There are non-alocoholic beers, so it is possible to have a meaningful interpretation of the intercept. In our data, however, there are very few beers with less than 3% alcoholic content, so it would not be wise to interpret the intercept. It is not safe to assume the same relationship between `PercentAlcohol` and `CaloriesPer12Oz` hold for beers with 0% alcohol; this would be extrapolation.

3. Interpret the 95% confidence interval for the slope in the context of the data.

We are 95% confident that the interval (26.557, 30.620) contains the true population slope for `PercentAlcohol`. This means we are 95% confident that for every 1% increase in alcohol content, the number of calories (per 12 oz) is expected to increase between 26.557 and 30.620 calories.

4. Find the critical value, t^*, used to calculate the 95% confidence interval. The code below is a guide; uncomment and complete the lines of code to calculate and display the critical value.

The critical value used to calculate the 95% confident interval for the slope is _____.

5. Interpret the test statistic in the context of the data

The estimated slope of 28.577 is 27.78 standard errors above the hypothesized mean of 0, assuming there is no linear relationship between percent alcohol and calories in domestic beers.

6. How was the p-value calculated? Fill in the code below to calculate the p-value. The code below is a guide; uncomment and complete the lines of code to calculate and display the p-value.

The p-value is _____. Given there is no linear relationship between `PercentAlcohol` and `CaloriesPer12Oz`, the probability of obtaining a test statistic with magnitude _____ or more extreme is _____.

7. Fill in the code below to calculate the predicted calories and corresponding 90% interval for a single beer with alcohol content of 4.3%.

8. Fill in the code below to calculate the predicted calories and corresponding 90% interval for the subset of beers with alcohol content of 4.3%.

Chapter 4

Analysis of Variance

This is about analysis of variance.

Chapter 5

Multiple Linear Regression

Chapter 6

Model Selection

Chapter 7

Logistic Regression

Chapter 8

Multinomial Logistic Regression

Chapter 9

Special Topics

Computing Assignments

Chapter 10

Intro to R

10.1 Introduction

```
# R is the name of the programming language itself and RStudio is a convenient interface.
```

The main goal of this lab is to introduce you to R and RStudio, which we will be using throughout the course both to learn the statistical concepts discussed in the course and to analyze real data and come to informed conclusions.

```
# git is a version control system (like "Track Changes" features from Microsoft Word but more powerful)
```

An additional goal is to introduce you to git and GitHub, which is the collaboration and version control system that we will be using throughout the course.

As the labs progress, you are encouraged to explore beyond what the labs dictate; a willingness to experiment will make you a much better programmer and statistician. If you're new to R, you should begin by building some basic fluency in R. Today's lab will focus on fundamental building blocks of R and RStudio: the interface, reading in data, and basic commands. Starting next week, the labs will focus on concepts more specific to regression analysis.

To make versioning simpler, today's lab is individual. This will give you a chance to become more familiar with git and GitHub. Next week you'll learn about collaborating on GitHub and will produce a single lab report as a team.

10.1.1 Topics covered in this lab:

- Exploratory Data Analysis (data visualizations and numerical summaries)
- Simple linear regression
- Writing a lab report using R Markdown
- Tracking changes and submitting work using git and GitHub

10.2 Packages

We will use the following packages in today's lab.

```
library(tidyverse)
library(readr)
```

```
library(skimr)
library(broom)
```

If you need to install any of the packages, you can run the code below in the **console**.

```
install.packages("tidyverse")
install.packages("readr")
install.packages("skimr")
install.packages("broom")
```

10.3 Warm up

Before we introduce the data, let's warm up with some simple exercises.

10.4 Project name:

Currently your project is called *Untitled Project*. Update the name of your project to be "Lab 01 - Intro R".

```
# The top portion of your R Markdown file (between the three dashed lines) is called YAML. It stands fo
```

10.4.1 YAML:

Open the R Markdown (Rmd) file in your project, change the author name to your name, and knit the document.

10.4.2 Commiting changes:

Then go to the Git pane in your RStudio.

If you have made changes to your Rmd file, you should see it listed here. Click on it to select it in this list and then click on **Diff**. This shows you the *diff*erence between the last committed state of the document and its current state that includes your changes. If you're happy with these changes, write "Update author name" in the **Commit message** box and click **Commit**.

You don't have to commit after every change, this would get quite cumbersome. You should consider committing states that are *meaningful to you* for inspection, comparison, or restoration. In the first few assignments we will tell you exactly when to commit and in some cases, what commit message to use. As the semester progresses we will let you make these decisions.

10.4.3 Pushing changes:

Now that you have made an update and committed this change, it's time to push these changes to the web! Or more specifically, to your repo on GitHub. Why? So that others can see your changes. And by others, we mean the course teaching team (your repos in this course are private to you and us, only).

In order to push your changes to GitHub, click on **Push**. This will prompt a dialogue box where you first need to enter your user name, and then your password. This might feel cumbersome. Bear with me... We *will* teach you how to save your password so you don't have to enter it every time. But for this one assignment you'll have to manually enter each time you push in order to gain some experience with it.

10.4.4 Data

Today's data comes from the Capital Bikeshare in Washington D.C. The Capital Bikeshare is a system in which customers can rent a bike for little cost, ride it around the city, and return it to a station near their destination. You can get more information about the bikeshare on their website, https://www.capitalbikeshare.com/. We will read in the data from the file *bikeshare.csv* located in the *data* folder.

```
bikeshare <- read_csv("data/bikeshare.csv")
```

This dataset contains information about the number of bike rentals, environmental conditions, and other information about the each day in 2011 and 2012.

10.5 Exercises

Before doing any analysis, we want to understand the basic structure of the data. One way to do this, is to look at the actual dataset. Type the code below in the **console** to view the entire dataset.

```
View(bikeshare)
```

It is sometimes more useful to view a summary of the data structure rather than view the entire dataset. This is especially true for very large datasets, i.e. those with a very large number of observations and/or rows. We can use the `glimpse()` function to get a general idea about the structure of our dataset. This function can be very useful when importing data from a file such as a .csv file (like in this lab) to ensure that data imported correctly and that we have the number of observations (rows) and variables (columns) we expect. We can also use this function to see each variable's type (e.g. integer, character).

```
# You can  type `??glimpse` in the console to learn more about the function and its syntax.
```

1. Type `glimpse(bikeshare)` in the **console** an overview of the `bikeshare` dataset.

How many observations are in the `bikeshare` dataset? How many variables?

2. In this lab, we will focus the analysis on the following variables:

`season`	1: Winter, 2: Spring, 3: Summer, 4: Fall
`temp`	Temperature (in $°C$) $\div$ 41
`count`	total number of bike rentals

Before fitting any regression models, we want to do an exploratory data analysis (EDA) to summarize the main characteristics of the data. Much of the EDA is visual, which we'll get to in the next exercise. The EDA also consists of calculating summary statistics for the variables in our dataset. It is good practice to examine any variable that may be relevant to the analysis in the EDA, since there may be variables that aren't directly included in the regression model but are still affecting the results. To keep today's lab manageable, we will only examine the three variables `season`, `temp`, and `count`.

There are many ways to calculate summary statistics for each variable, and we will use a few of them throughout the semester. For now, let's use the `skim()` function to calculate basic measures of center and spread along with get a sketch of the distribution.

```
bikeshare %>%
  select(season,temp,count) %>%
  skim()
```

What is the mean number of bike rentals? About 25% of the days in the data have a `count` above what value?

3. Does it make sense to calculate measures of center and spread for the variable `season`? If so, explain

why it makes sense. Otherwise, explain why the `skim()` function calculated these summary statistics for the variable `season` even if they don't make sense.

This is a good place to pause and commit changes with the commit message "Added summary statistics (Ex 1 - 3)", and push.

10.5.1 Visualizing Your Data

4. One important part of EDA is visualizing the data to get a better idea of the shape of the distribution of each variable along with the relationship between variables. There are a lot of ways to make plots in R; we will use the functions available in the `ggplot2` package.

The code below is used to create a histogram to visualize the distribution of `count`. Modify the code by writing an informative title and label for the x-axis.

```
ggplot(data=bikeshare, mapping=aes(x=count)) +
  geom_histogram() +
  labs(title="______", x="______")
```

5. Sometimes you may want to customize a plot by changing different features such as the color, marker types, etc. When plotting a histogram, one easy way to customize it is by changing the color the bars. We'll look at two different ways to do this.

First, using a color of your choice, include the option `color="_____"` inside of `geom_histogram()` function. Your code will look similar this. Be sure to also include an informative title and label for the x-axis.

```
ggplot(data=bikeshare, mapping=aes(x=count)) +
  geom_histogram(color="_______") +
  labs(title="______", x="______")
```

```
# This [ggplot2 quick reference](http://sape.inf.usi.ch/quick-reference/ggplot2/colour) has a long list
```

Next, instead of `color="_____"`, use `fill="______"` inside of the `geom_histogram()` function and put the color of your choice inside the blank. You can use the same color as before or use a new one.

What is the difference in the two plots? In other words, what is the difference in the way color is implemented when using `color` versus `fill`?

6. Describe the distribution of `count`. Your description should include comments about the shape, center, spread, and any potential outliers. You should use the histogram and the summary statistics from Exercise 2 in your description.

This is a another good place to pause and commit changes with the commit message "Added data visualization of count (Ex 3 - 6)", and push.

7. Now that we've examined the variables individually, we want to look at the relationship between the variables. To make interpretation easier, we will use the `mutate()` function to create a new variable called `temp_c` that is calculated as `temp * 41`. We will use `temp_c` for the remainder of the analysis, so the temperature can be discussed in terms of degrees Celsius.

```
bikeshare <- bikeshare %>%
  mutate(temp_c = temp * 41)
```

Complete the code below to make a scatterplot of the number of bike rentals versus the temperature.

```
ggplot(data=bikeshare, mapping=aes(x=temp_c,y=count)) +
  -------------------
```

```
# [https://ggplot2.tidyverse.org/](https://ggplot2.tidyverse.org/) is a great resource as you learn `gg
```

Describe the relationship between the temperature and the number of bike rentals.

8. The temperature and number of bike rentals varies greatly depending on the season. Therefore, we would like to create a separate scatterplot of `count` versus `temp_c` for each season. To do so, we will use the `facet_wrap()` function, faceting by `season`. Recall from Exercise 2 that `season` is currently stored as an integer. We need to change it to a factor variable type before using it in the `facet_wrap()` function (you could also change it to a character variable).

```
bikeshare <- bikeshare %>%
  mutate(season = as.factor(season))
```

```
ggplot(data=bikeshare, mapping=aes(x=temp_c,y=count)) +
  geom_point() +
  labs(title = "Number of Bike Rentals vs. Temperature",
       subtitle="Faceted by Season",
       x = "Temperature (Celsius)",
       y = "Number of Bike Rentals") +
  facet_wrap(~season)
```

For which season does the linear relationship between the temperature and the number of bike rentals appear to be the strongest?

This is a another good place to pause and commit changes with the commit message "Added visualization of count vs. temperature (Ex 7 - 8)", and push.

10.6 Simple Linear Regression

We want to fit a least-squares regression using the temperature (`temp_c`) to explain variation in the number of bike rentals (`count`) in the **winter** season. We can use the `filter()` function to create a subset from the data that only includes days during the winter. The `<-` assigns the name `winter_data` to our subset.

```
winter_data <- bikeshare %>%
  filter(season=="1")
```

We will use `winter_data` for the remainder of the lab.

9. Fit a simple linear regression model with using the `lm()` function; assign it the name `winter_model`. Replace *X*, *Y*, and *my.data* in the code below with the appropriate values.

```
winter_model <- lm(Y ~ X, data=my.data)
tidy(model) #output model
```

Interpret the slope.

Does it make sense to interpret the intercept? If so, write the interpretation of the intercept. Otherwise, explain why not.

10. We conclude by checking the assumptions for regression. We use the `mutate()` function to add a new variable called `resid` that is the residual for each observation in `winter_data` data.

```
winter_data <- winter_data %>%
  mutate(resid = residuals(winter_model))
```

The code for the residuals vs. the predictor variable and the Normal QQ plot are below. In addition to these plots, write the code to make a histogram of the residuals. You can reuse code from a previous exercise to plot the histogram.

```
ggplot(data=winter_data, mapping=aes(x=temp_c,y=resid)) +
  geom_point() +
```

```
  geom_hline(yintercept=0,color="red")+
  labs(title="Residuals vs. Temperature",
       x="Temperature",
       y="Residuals")
```

```
ggplot(data=winter_data, mapping=aes(sample=resid)) +
  stat_qq() +
  stat_qq_line()+
  labs(title="Normal QQ Plot of Residuals")
```

Based on the plots of the residuals and the scatterplot, are linearity, normality, and constant variance assumptions met? Briefly explain.

Is the independence assumption met based on the description of the data? Briefly explain.

Optional: Create a plot that could be used to help you assess the independence assumption.

Throughout the semester, we will learn various methods to deal with any violations in regression assumptions. For now, we will just note them.

You're done! Commit all remaining changes, use the commit message "Done with Lab 1!", and push. Before you wrap up the assignment, make sure all documents are updated on your GitHub repo.

10.7 Computing: College Admissions

The primary goal of today's lab is to give you practice with some of the tools you will need to conduct regression analysis using R. An additional goal for today is for you to be introduced to your teams and practice collaborating using GitHub and RStudio.

10.7.1 Packages

We will use the following packages in today's lab.

```
library(tidyverse)
library(skimr)
library(broom)
library(rcfss)
```

10.7.2 Data

In today's lab, we will analyze the `scorecard` dataset from the rcfss package. This dataset contains information about 1849 colleges obtained from the Department of Education's College Scorecard. Load the rcfss library into the global R environment and type `?scorecard` in the **console** to learn more about the dataset and variable definitions. Today's analysis will focus on the following variables:

`type`	Type of college (Public, Private - nonprofit, Private - for profit)
`cost`	The average annual cost of attendance, including tuition and feeds, books and supplies, and living expenses, minus the average grant/scholarship aid
`admrate`	Undergraduate admissions rate (from 0 - 100%)

10.7.3 Exercises

10.7.3.1 Exploratory Data Analysis

1. Plot a histogram to examine the distribution of `admrate`. What is the shape of the distribution?

2. To better understand the distribution of `admrate`, we would like calculate measures of center and spread of the distribution. Fill in the code below to use the `skim` function to calculate summary statistics for `admrate`. Report the appropriate measures of center (mean or median) and spread (standard deviation or IQR) based on the shape of the distribution from Exercise 1.

```
scorecard %>%
  select(admrate) %>%
  skim()
```

3. Plot the distribution of `cost` and calculate the appropriate summary statistics. Describe the distribution of `cost` (shape, center, and spread) using the plot and appropriate summary statistics.

4. One nice feature of the `skim` function is that it provides information about the number of observations that are missing values of the variable. How many observations have missing values of `admrate`? How many observations have missing values of `cost`?

5. Later in the semester, we will techniques to deal with missing values in the data. For now, however, we will only include complete observations for the remainder of this analysis. We can use the `filter` function to select only the rows that values for both `cost` and `admrate`.

Fill in the code below to create a new dataset called `scorecard_new` that only includes observations with values for both `admrate` and `cost`.

```
__________ <- scorecard %>%
  filter(!is.na(admrate),________)
```

```
# Learn more about the `filter` function in [Section 5.2 of R for Data Science] (https://r4ds.had.co.nz/
```

You will use `scorecard_new` for the rest of the lab.

6. Create a scatterplot to display the relationship between `cost` (response variable) and `admrate` (explanatory variable). Use the scatterplot to describe the relationship between the two variables.

7. The data contains information about the type of college, and we would like to incorporate this information into the scatterplot. One way to do this is to use a different color marker for each type of college. Fill in the code below the scatterplot from the previous exercise with the marker colors based on the variable `type`. Describe two new observations from this scatterplot that you didn't see in the previous plot.

```
ggplot(data=scorecard_new, mapping=aes(x=admrate, y=cost, color=type)) +
  ____________________
```

10.7.3.2 Simple Linear Regression

8. Fit a regression model to describe the relationship between a college's admission rate and cost. Use the `tidy` function to display the model.

9. Interpret the slope in the context of the problem. Does the intercept have a meaningful interpretation? If so, write the interpretation in the context of the problem. Otherwise, explain why the interpretation is not meaningful.

10. While the `tidy` function is used to display the model, we can obtain a one-row summary of the model using the `glance` function. Use the `glance` function to get a summary of the model fit in the previous exercise. See the documentation for `glance` for the syntax and a list of values output from the function.

11. What is the value of R^2? Interpret this value in the context of the problem. Do you think this is a "good" value of R^2? Explain.

12. What is the value of $\hat{\sigma}$, the residual standard error.

13. What is the 95% confidence interval for the coefficient of `admrate`, i.e. the slope? Interpret the interval in the context of the data.

14. We want to test the following hypotheses about the population slope β_1:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

State what the null and alternative hypotheses mean in terms of the linear relationship between `admrate` and `cost`.

15. Consider the confidence interval from Exercise 13 and the hypotheses in Exercise 14. Is the confidence interval consistent with the null or alternative hypothesis? Briefly explain.

10.8 ANOVA

The goal of this lab is to use Analysis of Variance (ANOVA) to compare means in multiple groups. Additionally, you will be introduced to new R functions used for wrangling and summarizing data.

10.8.1 Packages

We will use the following packages in today's lab.

```
library(tidyverse)
library(knitr)
library(broom)
```

10.8.2 Data

In today's lab, we will analyze the `diamonds` dataset from the ggplot2 package. Type `?diamonds` in the console to see a dictionary of the variables in the data set. This analysis will focus on the relationship between a diamond's carat weight and its color. Before starting the exercises, take a moment to read more about the diamond attributes on the Gemological Institute of America webpage: https://www.gia.edu/diamond-quality-factor.

10.9 Exercises

The diamonds dataset contains the price and other characteristics for over 50,000 diamonds price from $326 to $18823. In this lab, we will analyze the subset of diamonds that are priced $1200 or less.

1. Create a dataframe called `diamonds_low` that is the subset of diamonds priced $1200 or less. How many observations are in `diamonds_low`?

When using Analysis of Variance (ANOVA) to compare group means, it is ideal to have approximately the same number of observations in each group. Therefore, we will combine the worst two color groups, I and J, and create a new color category called "I/J". Since `color` is an ordinal (`<ord>`) variable, we need to use the `recode_factor` function in the dplyr package to create the new category.

Use the `count` function before and after making the new color category to ensure the recoding worked as expected.

```
## number of observations at each color level
diamonds_low %>%
  count(color)
```

```
#create a new vector of the recoded values
color_recoded <-  recode_factor(diamonds_low$color,
                                `I` = "I/J", `J` = "I/J",
                                .default = levels(diamonds_low$color))

#replace the color variable with the recoded data
diamonds_low <- diamonds_low %>%
  mutate(color = color_recoded)
```

```
# Refer to the [ggplot2 Cheat Sheet](https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsh
```

2. We begin by plotting the relationship between `color` and `carat`. As a group, brainstorm ways to plot the relationship between the two variables, then make one of the plots. Be sure to include informative axes labels and an informative title.

3. Fill in the code below to calculate the mean and variance of `carat` at each level of `color`.

```
# The `group_by` function is used to do calculation in groups. The `summarise` function is used to redu
```

```
diamonds_low %>%
  group_by(_______) %>%
  summarise(n = n(),
            avg_carat = mean(carat),
            var_carat = _______)
```

Based on the plots and summary statistics, does there appear to be a relationship between carat weight the color of diamonds? In other words, does there appear to be a significant difference in the mean carat weight across colors?

4. When using ANOVA to compare means across groups, we make the following assumptions (note how similar they are to the assumptions for regression):

- **Normality**: The distribution of y is approximately normal within each category of x - in the k^{th} category, $y \sim (\mu_k, \sigma^2)$. If the sample size is large, ANOVA is robust to some departures from Normality.
- **Independence**: All observations are independent from one another, i.e. one observation does not affect another.
- **Constant Variance**: The distribution of y within each category of x has a common variance, σ^2. One way to assess if variances are sufficiently equal is to look at the ratio of the maximum group variance to the minimum group variance. If this ratio is less than 2, then we can conclude the variances are approximately equal. This isn't an exact threshold, but rather a commonly used guideline. *Note: There are formal tests for equal variance that are outside the scope of this class.*

Are the assumptions for ANOVA met? Comment on each assumption using the summary statistics and/or plots from previous exercises to support your conclusion. You may also calculate any additional summary statistics or make additional plots as needed.

Regardless of your answer to Excerise 4, We will proceed with the analysis in the remainder of this lab as if

the assumptions are met.

5. Use the code below to calculate the ANOVA table. The `tidy` function from the broom package is used to put the ANOVA output in a dataframe, and with the `kable` function from the knitr package, you can display the results in an easy-to-read table.

```
anova <- aov(carat ~ color, data=diamonds_low)
anova %>%
  tidy() %>%
   kable()
```

6. Use the ANOVA table to calculate the total mean square, i.e. the sample variance of `carat`. Show your calculations. You can put the calculations in a code chunk to use R like a calculator.
7. What is $\hat{\sigma}^2$, the estimated variance of `carat` within each level of `color`.
8. We can use ANOVA to test if the true mean value of `carat` is equal for all levels of `color`, i.e.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_6$$

State the alternative hypothesis is the context of the data.

9. Based on the ANOVA table, what is your conclusion from the test of the hypotheses in the previous question? State the conclusion in the context of the data.
10. Use the code below to plot a 95% confidence interval for the mean carat weight at each level of color. Calculate the value of `sigma` by filling in the estimated variance from Exercise 7.

The formula for the confidence interval for the mean of group k is

```
# The critical value $t^*$ is calculated using the *t* distribution with $n-K$ degrees of freedom.
```

```
# The standard error of the mean is calculated using $\hat{\sigma}$, the square root of the variance wi
```

$$\bar{y}_k \pm t^* \frac{\hat{\sigma}}{\sqrt{n_k}}$$

```
n.groups <- diamonds_low %>% distinct(color) %>% count()
crit.val <- qt(0.975, (nrow(diamonds_low)-n.groups$n))
sigma <- sqrt(_________)

conf.intervals <- diamonds_low %>%
  group_by(color) %>%
  summarise(mean_carat = mean(carat), n = n(),
            lower = mean_carat - crit.val * sigma/sqrt(n),
            upper = mean_carat + crit.val * sigma/sqrt(n))
```

```
ggplot(data=conf.intervals,aes(x=color,y=mean_carat)) +
  geom_point() +
  geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.1) +
  labs(title="95% confidence interval for the mean value of carat",
       subtitle="by Color") +
  coord_flip()
```

11. For what color level is the mean carat weight the most different from all the others?
12. Based on this analysis, describe the relationship between the color and the mean carat weight in diamonds that cost $1200 or less. *Refer to the diamond documentation to recall what the color scale means.*

You're done! Commit all remaining changes, use the commit message "Done with Lab 3!", and push. Before you wrap up the assignment, make sure the .Rmd, .html, and .md documents are all updated on your GitHub repo.

10.10 Multiple Linear Regression

The goal of this lab is to use multiple linear regression to understand the variation in the selling price of houses in King County, Washington. You will also gain practice using special predictors, such as categorical predictors and interaction effects, in the model, and you will be introduced to variable transformations.

10.10.1 Packages

We will use the following packages in today's lab.

```
library(tidyverse)
library(knitr)
library(broom)
```

10.10.2 Data

The for today's lab contains the price and other characteristics of over 20,000 houses sold in King County, Washington (the county that includes Seattle). The dataset includes the following variables:

- `price`: selling price of the house
- `date`: date house was sold, measured in days since January 1, 2014
- `bedrooms`: number of bedrooms
- `bathrooms`: number of bathrooms
- `sqft`: interior square footage
- `floors`: number of floors
- `waterfront`: 1 if the house has a view of the waterfront, 0 otherwise
- `yr_built`: year the house was built
- `yr_renovated`: 0 if the house was never renovated, the year the house was renovated if else

```
houses <- read_csv("data/KingCountyHouses.csv")
```

10.11 Exercises

1. Use data visualization and summary statistics to examine the distribution of `bedrooms`. What is the maximum value? Does this value make sense? If not, what is this an indication of, i.e. how did this value get recorded in the data? Briefly explain.

```
# See the [documentation](https://dplyr.tidyverse.org/reference/summarise.html) for more information ab
```

2. We want to remove observations that have extreme values for bedrooms, i.e. those with values for `bedrooms` above the 95th percentile in the data. What is the 95th percentile for `bedrooms`? Use the `summarise` function to help you calculate this value.

3. Fill in the code below to filter the data so that the extreme observations are removed. How many observations are in the updated dataset?

```
houses <- houses %>% filter(bedrooms <= ____)
```

We will use this dataset for the remainder of the analysis.

4. Fit a regression model using square feet to explain variation in the price. Plot the residuals versus the predicted values. Based on this plot, what regression assumption appears to be violated? Briefly explain.

Plot the histogram and Normal QQ-plot of the residuals. Based on these plots, what regression assumption appears to be violated? Briefly explain.

5. One way to deal with violations in regression assumptions is to transform the response variable and use that transformed variable when fitting the regression model. (We will talk about this in class next week). Some common transformations used in regression are the natural log (`log(y)`), the square root (`sqrt(y)`), and the reciprocal (`1/y`).

Each transformation is applied to the response variable `price`, and the distributions of the transformed data are shown below.

Which transformation should we use to fix the violations of the model assumptions observed in the previous exercise? Briefly explain your choice.

6. Add the variable `logprice`, the log-transformed version of `price`, to the data frame. Fit a regression model with `logprice` as the response and `sqft` as the predictor variable. Create the residuals plots (residuals vs. predicted, histogram of residuals, Normal QQ-plot). Briefly comment on whether or not using the transformed variable improved on the model assumptions.

7. Though we can explain about 48% of the variation in a house prices by the square footage, we would like to incorporate some of the other available house characteristics in the model.

Before fitting the model, use the code below to add the variable`floorsCat` that is the categorical version of the variable `floors`. Discuss with your group why it may make sense to treat `floors` as categorical, even though it represents a count.

```
houses <- houses %>%
  mutate(floorsCat = as.factor(floors))
```

```
# See the [documentation](https://dplyr.tidyverse.org/reference/tally.html) for more information about
```

Use the `count` function to see the number of observations at each level of `floorsCat`. What is the most common number of floors?

8. Use the code below to calculate the mean-centered versions of the variables `sqft`, `bedrooms`, and `bathrooms` and add them to the data frame.

```
houses <- houses %>%
  mutate(sqftCent = sqft - mean(sqft),
         bedroomsCent = bedrooms - mean(bedrooms),
         bathroomsCent = bathrooms-mean(bathrooms))
```

It is not appropriate to calculate the mean-centered version of the variable `waterfront`. Briefly explain why it isn't.

9. Fit a regression model with `logprice` as the response variable, and the mean-centered variables from the previous exercise along with `waterfront` and `floorsCat` as the predictor variables. Display the model output.

10. What is the baseline level for the variable `floorsCat`?

11. Interpret the intercept of the model in the context of the data. Write the interpretation in terms of the **price**.

12. What is the intercept of the model for the subset of houses with 3 floors that are not on the waterfront? Write the intercept in terms of the **log(price)**.

13. We would like to consider potential interactions for the model. A significant **interaction** occurs when the relationship of a predictor variable with the response depends on the value of another predictor variable.

Fill in the code below to plot the relationship between `logprice` and `bedrooms` by `waterfront`. Based on this plot, do you think there is a significant interaction effect between `bedrooms` and `waterfront`? In other words, do you think the relationship between the logprice and the number of bedrooms differs based on whether or not a house is on the waterfront? Briefly explain.

```
ggplot(data=houses,mapping=aes(x=_____,y=_____,color=as.factor(waterfront))) +
  geom_smooth(method="lm", se=FALSE) +
  labs(title="___________________",
       x="Number of bedrooms",
       y="Log Price",
       color="Waterfront")
```

We will talk more about interaction effects in Monday's lecture. In HW 03, you explore potential interaction effects using this housing data.

You're done! Commit all remaining changes, use the commit message "Done with Lab 4!", and push. Before you wrap up the assignment, make sure the .Rmd and .md documents are updated in your GitHub repo. There is a 10% penalty if the .Rmd file has to be knitted to display graphs, i.e. the graphs are not showing in the .md file on GitHub.

10.11.1 Acknowledgement

The data used in this lab was obtained from https://github.com/proback/BYSH.

10.12 Data Wrangling & Multiple Linear Regression

When doing statistical analyses in practice, there is often a lot of time spent on cleaning and preparing the data. The goal of today's lab is to practice cleaning messy data, so it can be used in a regression analysis. You will also practice interpreting the results from a regression model that has numeric and categorical predictors and a log-transformed response variable.

10.12.1 Packages

We will use the following packages in today's lab.

```
library(tidyverse)
library(knitr)
library(skimr)
library(broom)
```

10.12.2 Data

Today's data is about Airbnb listings in Asheville, NC. The data was obtained from http://insideairbnb.com/; it was originally scraped from airbnb.com.

You can see a visualization of some of the data used in today's lab at http://insideairbnb.com/asheville/.

```
basic_info <- read_csv("data/airbnb_basic.csv")
details <- read_csv("data/airbnb_details.csv")
```

We will use the following variables in this lab:

- `price`: Cost per night (in U.S. dollars)
- `cleaning_fee`: Cleaning fee (in U.S. dollars)
- `property_type`: Type of dwelling (House, Apartment, etc.)
- `room_type`:
  - *Entire home/apt* (guests have entire place to themselves)
  - *Private room* (Guests have private room to sleep, all other rooms shared)
  - *Shared room* (Guests sleep in room shared with others)
- `number_of_reviews`: Total number of reviews for the listing
- `review_scores_rating`: Average review score (0 - 100)

10.13 Exercises

10.13.1 Data wrangling

1. We would like to use variables from both the `basic_info` and `details` data frames in this analysis. Both dataframes have the variable `id` that uniquely identifies each Airbnb listing. Because we need data from `basic_info` and `details`, we only want to include observations that are in both the `basic_info` and `details` datasets. Therefore, we will use an `inner_join` to combine the two data sets. (Note: Both data frames include a variable called `id` that uniquely identifies each Airbnb listing. R will use this variable to join the two data frames.)

```
# See [Section 13.4 of *R for Data Science*](https://r4ds.had.co.nz/relational-data.html#mutating-joins
```

```
airbnb <- inner_join(basic_info,details)
```

How many observations are in `airbnb`? How many variables?

2. Some Airbnb rentals have cleaning fees, and we want to include the cleaning fee when we calculate the total rental cost. Use the code below to see how the data in the column `cleaning_fee` is currently stored in the `airbnb` data frame.

```
typeof(airbnb$cleaning_fee)
```

The column `cleaning_fee` currently contains what type of data? Why do you think the data is stored this way even though `cleaning_fee` is a quantitative variable?

3. Since `cleaning_fee` is a quantitative variable, we need to make sure it is stored as numeric data in the dataframe. To do so, we will first use the `extract` function in `tidyr` package to create a column of cleaning fees that don't have the dollar sign. Then, we will use the `as.numeric()` function to make the extracted data the numeric data type `double`.

```
# See [https://tidyr.tidyverse.org/reference/extract.html) for more information about the `extract` fun
```

```
airbnb <- airbnb %>%
  extract(cleaning_fee, "cleaning_fee") %>%
  mutate(cleaning_fee = as.numeric(cleaning_fee))
```

Use the `typeof` function to confirm that `cleaning_fee` is now stored as a `double` data type.

4. Use the `skim` function to view a summary of the `cleaning_fee` data. How many observations have missing values for `cleaning_fee`? What do you think is the most likely reason for the missing observations of `cleaning_fee`? In other words, what does a missing value of `cleaning_fee` indicate?

5. Fill in the code below to impute the missing values of `cleaning_fee` with an appropriate numeric value. Then use the `skim` function to confirm that there are no longer missing values of `cleaning_fee`.

```
# See [https://dplyr.tidyverse.org/reference/case_when.html](https://dplyr.tidyverse.org/reference/case_
```

```
airbnb <- airbnb %>%
  mutate(cleaning_fee = case_when(
    is.na(cleaning_fee) ~ ______,
    TRUE ~ cleaning_fee
  ))
```

This is an example of data that is missing not at random, since there is a specific pattern/explanation to the misisng data. We will talk more about dealing with missing data later in the semester.

```
# See [Section 5.6.3 of *R for Data Science*](https://r4ds.had.co.nz/transform.html#counts) for more in
```

6. Next, we look at the variable `property_type`.
   - Use the `count` function to determine how many categories are in the variable `property` and the frequency of each category.
   - What are the top 4 most common property types? These make up what proportion of the observations?
7. Since an overwhelming majority of the observations in the data are one of the top 4 property types, we would like to create a simplified version of the `proprety_type` variable that has 5 categories: *House*, *Apartment*, *Guest suite*, *Bungalow*, and *Other*. Fill in the code below to create `prop_type_simp`.

```
airbnb <- airbnb %>%
  mutate(prop_type_simp = case_when(
    property_type %in% c("House","______", "______","______") ~ property_type,
    TRUE ~ "Other"
  ))
```

Use the code below to check that `prop_type_simp` was correctly made.

```
airbnb %>%
  count(property_type, prop_type_simp) %>%
  arrange(desc(n))
```

8. Airbnb is most commonly used for travel purposes, i.e. as an alternative to traditional hotels. We only want to include Airbnb listings in our regression analysis that are intended for travel purposes. What are the 5 most common values for the variable `minimum_nights`? Which value in the top 5 stands out? What is the likely intended purpose for Airbnb listings with this seemingly unusual value for `minimum_nights`?

Filter the `airbnb` data so that it only includes observations with `minimum_nights <= 3`. You will use this filtered dataset for the remainder of the lab.

10.13.2 Regression Analysis

9. For the response variable, will use the cost to stay at an Airbnb location for 3 nights. Create a new variable called `price_3_nights` that uses `price` and `cleaning_fee` to calculate the total cost to stay at the Airbnb property for 3 nights. Be sure to add this variable to your dataframe.
10. Use histograms to examine the distributions of `price_3_nights` and `log(price_3_nights)`. Based on the histograms, which variable should you use for the regression model? Briefly explain.

Use this variable as the response for the remainder of the lab.

11. Fit a regression model called `model1` with the response variable from the previous question and the following predictor variables: `prop_type_simp`, `number_of_reviews`, and `review_scores_rating`. Display the model output.

12. Interpret the coefficient `review_scores_rating` in terms of `price_3_nights`.

13. Interpret the coefficient of `prop_type_simpGuest suite` in terms of `price_3_nights`.

14. We want to determine if `room_type` is a significant predictor of the cost for 3 nights, given everything else in the model. Fit a regression model called `model2` that includes all of the predictor variables in `model1` and `room_type`. Display the model output.

Use the code below to conduct a Nested F test to determine if `room_type` is a significant predictor of the minimum cost. What is your conclusion from the Nested F test?

```
anova(model1, model2)
```

15. Suppose you are planning to visit Asheville over spring break, and you want to stay in an Airbnb. You find an Airbnb that is an apartment with a private room, has 10 reviews, and an average rating of 90. Use `model2` to predict the total cost to stay at this Airbnb for 3 nights. Include the appropriate 95% interval with your prediction. Report the prediction and interval in terms of `price_3_nights`.

You're done! Commit all remaining changes, use the commit message "Done with Lab 5!", and push. Before you wrap up the assignment, make sure the .Rmd and .md documents are updated in your GitHub repo. There is a 10% penalty if the .Rmd file has to be knitted to display graphs, i.e. the graphs are not showing in the .md file on GitHub.

10.13.3 Acknowledgement

The data from this lab is from insideairbnb.com

10.14 Model Selection

The goal of today's lab is to practice forward and backward model selection. In addition to practice with model selection functions in R, you will manually conduct a backward selection procedure to better understand what occurs when you use model selection functions.

10.14.1 Packages

You will need the following packages for today's lab:

```
library(tidyverse)
library(knitr)
library(broom)
library(leaps)
library(Sleuth3) #case1201 data
library(ISLR) #Hitters data
```

10.14.2 Data

There are two datasets for this lab.

10.14.2.1 Part I

The dataset for Part I contains the SAT score (out of 1600) and other variables that may be associated with SAT performance for each of the 50 states in the U.S. The data is based on test takers for the 1982 exam. The following variables are in the dataset:

- **SAT**: average total SAT score
- **State**: U.S. State
- **Takers**: percentage of high school seniors who took exam
- **Income**: median income of families of test-takers ($ hundreds)
- **Years**: average number of years test-takers had formal education in social sciences, natural sciences, and humanities
- **Public**: percentage of test-takers who attended public high schools
- **Expend**: total state expenditure on high schools ($ hundreds per student)
- **Rank**: median percentile rank of test-takers within their high school classes

10.14.2.2 Part II

The dataset for Part II contains the performance statistics and salaries of Major League Baseball players in the 1986 and 1987 seasons. The data is in the `Hitters` dataset in the `ISLR` package. Type `?Hitters` in the console to see the variables names and their definitions.

10.15 Exercises

10.15.1 Part I

For the first part of the lab, you will return to the model selection activity you started in class using the SAT data. The data is in the `case1201` data frame in the `Sleuth3` package.

```
sat_scores <- case1201 %>%
  select(-State)
```

1. Manually perform backward selection using $Adj.R^2$ as the selection criterion. To help you get started, the full model and the code for the first set of models to test are below. Show each step of the selection process. Display the coefficients and $Adj.R^2$ of your final model.

```
full_model <- lm(SAT ~ ., data = sat_scores)
```

```
m1 <- lm(SAT ~ Income + Years + Public + Expend + Rank, data = sat_scores)
```

```
m2 <- lm(SAT ~ Takers + Years + Public + Expend + Rank, data = sat_scores)
```

```
m3 <- lm(SAT ~ Takers + Income + Public + Expend + Rank, data = sat_scores)
```

```
m4 <- lm(SAT ~ Takers + Income + Years + Expend + Rank, data = sat_scores)
```

```
m5 <- lm(SAT ~ Takers + Income + Years + Public + Rank, data = sat_scores)
```

```
m6 <- lm(SAT ~ Takers + Income + Years + Public + Expend, data = sat_scores)
```

2. What is the best 5-variable model? Display the model output.
3. Use the `regsubsets` function to perform backward selection. What is the final model when $Adj.R^2$ is the selection criterion? Display the coefficients and the $Adj.R^2$ of the final model. *This should be the same result you got in Exercise 1.*

4. What is the final model when BIC is the selection criterion? Display the coefficients and the BIC of the final model.

5. Compare the final models selected by $Adj.R^2$ and BIC.

   - Do the models have the same number of predictors? Briefly explain.
   - Are the same predictor variables in each model? Briefly explain.

6. Consider the comparisons made in the previous exercise. Are these differences what you would expect given the selection criteria used? Briefly explain.

10.15.2 Part II

The data for this part of the lab is the `Hitters` dataset in the `ISLR` package. Your goal is to fit a regression model that uses the performance statistics of baseball players to predictor their salary. There are 19 potential predictor variables, so you will use the `regsubsets` function to conduct forward selection to choose a final model.

7. Read through the data dictionary for the `Hitters` dataset. You can access it by typing `?Hitters` in the console. What is the difference between the variables `HmRun` and `CHmRun`?

8. Some observations have missing values for `Salary`. Filter the data, so only observations that have values for `Salary` are included. You will use this filtered data for the remainder of the lab.

9. Fill in the code below to conduct forward selection and save the results in an object called `sel_summary` (selection summary).

```
# The `nvmax` option indicates the maximum-sized variable subsets to consider in the model selection.
```

```
regfit_forward <- regsubsets(_______, ________, method="forward", nvmax = 19)
sel_summary <- summary(_______)
```

10. The object `sel_summary` contains the summary statistics for the best fit model containing k predictors, where $k = 1, \ldots, 19$. The object `sel_summary` is a list object, so it is cumbersome to extract the relevant summary statistics. Therefore, you can create a data frame called `summary_stats` such that each row represents the best fit model with k predictors and each column is a summary statistic. For example, the second row contains the summary statistics of the best fit model that contains 2 variables.

Fill in the code below to create the data frame `summary_stats` that includes the BIC, R^2, $Adj.R^2$, and residual sum of squares (RSS) for each model in `sel_summary`. The data frame `summary_stats` will also include the column `np`, the number of predictors in the model represented on each row.

```
summary_stats <- data.frame(bic = sel_summary$bic,
                            adjr2 = _______,
                            rsq = _______,
                            rss = _______) %>%
  mutate(np = row_number()) #number of variables
```

```
# See the [ggplot2 documentation](https://ggplot2.tidyverse.org/reference/geom_abline.html#arguments) f
```

11. Use the data in the `summary_stats` data frame to plot BIC versus the number of predictors. Include a vertical line on your plot that shows the number of predictors for the overall final model you would select based on BIC. Be sure your plot has clear and informative title and axes labels.
    - How does BIC change as the number of predictors increases?
    - How many predictors are in the final model selected based on BIC?

You can fill in the code below with either ***max*** *or* ***min*** *to find the number of predictors in the final model selected based on BIC.*

```
np_bic <- summary_stats %>%
  filter(bic == _____(bic)) %>%
  select(np) %>%
  pull()
```

12. Use the data in the `summary_stats` data frame to plot $Adj.R^2$ versus the number of predictors. Include a vertical line on your plot that shows the number of predictors for the final model you would select based on $Adj.R^2$. Be sure your plot has clear and informative title and axes labels.
    - How does $Adj.R^2$ change as the number of predictors increases?
    - How many predictors are in the final model selected based on $Adj.R^2$?
13. Use the data in the `summary_stats` data frame to plot R^2 versus the number of predictors. Include a vertical line on your plot that shows the number of predictors for the final model selected based on R^2. Be sure your plot has clear and informative title and axes labels.
    - How does R^2 change as the number of predictors increases?
    - How many predictors are in the final model selected based on R^2?
14. Should R^2 be used as a model selection criterion? Briefly explain why or why not using your answers to Exercises 11 - 13.
15. Choose a final model to predict a baseball player's `Salary` from his performance statistics. Display the variables, their coefficients, and the summary statistics from the `summary_stats` data frame for this model.
16. Briefly explain why you chose the model in the previous exercise.
    - Which model selection criteria did you use (BIC, $Adj.R^2$, R^2)? Why?
    - What other factors did you consider besides the value of the model selection criteria?

10.15.3 Acknowledgements

Part II of this lab was inspired by Lab 6.5 in An Introduction to Statistical Learning and Variable Selection in Regression.

10.16 Logistic Regression

Over the past ten years, recommendation systems have become increasingly popular as more companies strive to offer customized user experiences. Amazon recommends products you may like based on your browse and purchase history, Netflix recommends movies and TV shows based on your viewing history, and music platforms like Spotify recommend songs you may like based on your listening history. While these recommendation systems are built using a variety of algorithms, they are all trying to achieve the same goal: use the characteristics of the products/movies/music a user is known to like to figure out the products/movies/music the user may like but hasn't discovered yet.

```
# See ["How Does Spotify Know You So Well?"](https://medium.com/s/story/spotifys-discover-weekly-how-ma
```

In today's lab, we will focus on using the characteristics of songs a user previously played to determine whether or not a user will like a new song. We will use logistic regression to build a model that predicts the probability a user likes a song using the relevant characteristics of that song.

10.16.1 Packages

You will need the following packages for today's lab:

```
library(tidyverse)
library(broom)
## Fill in other packages as needed
```

10.16.2 Data

The data in this lab is from the Spotify Song Attributes data set in Kaggle. This data set contains song characteristics of 2017 songs played by a single user and whether or not he liked the song. Since this dataset contains the song preferences of a single user, the scope of the analysis is limited to this particular user.

You will use data `spotify.csv` to build the logistic regression model and test the performance of the model using the songs in `test_songs.csv`. Click **here** to download the dataset `spotify.csv`, and click **here** to download the dataset `test_songs.csv`. Upload both files to the to the `data` folder in your `lab-07` project.

The Spotify documentation page contains a description of the variables included in this dataset.

10.17 Exercises

10.17.1 Exploratory Data Analysis

1. Read through the Spotify documentation page to learn more about the variables in the dataset. The response variable for this analysis is `like`, such that 1 indicates that the user likes the song and 0 otherwise. The remaining will be considered as predictor variables in the model.

```
# Part of the code to make `x` a factor.

# mutate(x = factor(x))
```

```
- Which potential predictor variables are categorical? You only need to include the variables that are
- Recode the each of the categorical predictors so they are a `factor` variable type.
```

2. Choose a quantitative predictor variable. Make the appropriate plot of the response versus this predictor variable. Describe the relationship between the two variables.

3. Choose a categorical predictor variable. Make the appropriate plot of the response versus this predictor variable. Describe the relationship between the two variables.

4. Let's consider a potential interaction effect between the variables you choose in Exercises 2 and 3. Make the appropriate plots to examine the potential interaction effect. Do these plots suggest there is a significant interaction effect? Briefly explain.

In practice, you should do exploratory data analysis for all potential explanatory variables. We did an abbreviated exploratory data analysis to make the assignment more manageable.

10.17.2 Part II: Logistic Regression Model

5. Fit the full model and display the model output. The main objective for the model is to predict whether the user will like a song. Should we use this model for this objective? Briefly explain.

6. Use the `step` function to perform backward selection. Display the output for the selected model.

7. Briefly describe the criteria used by `step` to select the final model.

For the remainder of this lab, you will use the model chosen by model selection . In practice; however, you would not just stop with the results from the automated model selection procedure and would examine the model further to see if there are any significant interactions, higher-order terms, or if it could even be simplified.

8. Consider the variable `duration_ms`.
    - Interpret the coefficient of `duration_ms` and its 95% confidence interval in terms of the odds of the user liking a song.
    - Suppose instead of `duration_ms`, we use the variable `duration_s`, the duration of a song in seconds. What would be the effect of `duration_s` on the odds of the user liking a song? Include the updated coefficient and corresponding 95% confidence interval for `duration_s`. *Assume all other variables in the model are unchanged.*
9. Interpret `mode` and its 95% confidence interval in terms of the odds of the user liking a song.
    - Based on this model, is there evidence of a significant difference in the user's preference between songs in a major key versus those in a minor key?

10.17.3 Part III: Model Assessment

In the next few questions, we will do an abbreviated analysis of the residuals.

10. Create a binned plot of the residuals versus the predicted probabilities. *You will first need to use the* `augment` *function with the options* `type.predict = "response"` *and* `type.residuals = "response"` *to get the predicted probabilities and corresponding residuals.*
11. Choose a quantitative predictor in the final model. Make the appropriate table or plot to examine the residuals versus this predictor variable.
12. Choose a categorical predictor in the final model. Make the appropriate table or plot to examine the residuals versus this predictor variable.

In practice, you should examine plots of residuals versus every predictor variable to make a complete assessment of the model fit. For the sake of time on the lab, you will use these three plots to help make the assessment in Exercise 14.

13. Plot the ROC curve and find the area under the curve.
14. Based on the residual plots and the ROC curve, is this logistic model a good fit for the data? Briefly explain.

10.17.4 Part IV: Prediction

15. You are part of the data science team at Spotify, and your model will be used to make song recommendations to users. The goal is to recommend songs the user has a high probability of liking.

As a group, choose a threshold value to distinguish between songs the user will like and those the user won't like. What is your threshold value? Use the ROC curve to help justify your choice.

16. Now let's put your model and decision threshold to the test! Use your model to calculate the predicted probability that the user will like the following two songs:
    - "Sign of the Times" by Harry Styles
    - "Hotline Bling" by Drake

The data for the songs can be found in *test_songs.csv*.

17. Using your decision threshold from Question 15, would you recommend "Sign of the Times" to this user? Would your recommend "Hotline Bling" to this user? Briefly explain your decision.

18. The user likes "Hotline Bling" but doesn't like "Sign of the Times". How good were your recommendations based on these two songs?

    - If they were good recommendations, explain how the model and threshold helped you distinguish between songs the user would like and those he wouldn't.
    - If they were not good recommendations, explain the limitations in your model and/or threshold.

10.18 Multinomial Logistic Regression

The General Social Survey (GSS) has been used to measure trends in attitudes and behaviors in American society since 1972. In addition to collecting demographic information, the survey includes questions used to gauge attitudes about government spending priorities, confidence in institutions, lifestyle, and many other topics. A full description of the survey may be found here.

In today's lab, we will use multinomial logistic regression to understand the relationship between a person's political views and their attitudes towards government spending on mass transportation projects. To do so, we will use data from the 2010 GSS survey. Refer to the Multinomial Logistic Regression notes for help with concepts and code.

10.18.1 Packages

You will need the following packages for today's lab:

```
library(tidyverse)
library(nnet)
library(knitr)
library(broom)
## Fill in other packages as needed
```

10.18.2 Data

The data for this lab is from the 2016 General Social Survey. The original data set contains 2867 observations and 935 variables. Given the size of the dataset, we will handle it differently in our workflow than we've handled data in previous assignments.

```
# [Working with large files](https://help.github.com/en/articles/working-with-large-files)
```

The size of this dataset is 34.3 MB. Compare that to the Spotify dataset from last weeks' lab which was 149 KB (0.149 MB)! GitHub will not allow you to push files larger than 100 MB and will give you a warning when you push files as large as 50 MB. Though we could push the file we're working with today to GitHub, it's large enough that we'd still prefer not to.

You have may noticed that each repo contains a file called `.gitignore`. It contains a list of the files you don't want commit or push to GitHub. If you look at the `.gitignore` file for today's lab, you will notice that `gss2016.csv` is listed at the bottom.

- Click **here** to download `gss2016.csv`.
- Upload `gss2016.csv` into the data folder of your project.
- Notice that `gss2016.csv` does not appear in your Git pane. This is because it is being ignored by git, since it is listed in the `.gitignore` file.

You will use the following variables in the lab:

- `natmass`: Respondent's answer to the following prompt:

  "We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount…are we spending too much, too little, or about the right amount on mass transportation?"

- `age`: Age in years.
- `sex`: Sex recorded as *male* or *female*
- `sei10`: Socioeconomic index from 0 to 100
- `region`: Region where interview took place
- `polviews`: Respondent's answer to the following prompt:

  "We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal - point 1 - to extremely conservative - point 7. Where would you place yourself on this scale?"

Use the code below to read in the data.

```
gss <- read_csv("data/gss2016.csv",
  na = c("", "Don't know", "No answer",
         "Not applicable"),
         guess_max = 2867) %>%
  select(natmass, age, sex, sei10, region, polviews) %>%
  drop_na()
```

The argument `guess_max = 2867` tells the `read_csv` function to use all of the observations in a column to determine its data type. Without this argument, only the first 1,000 observations would be used to make this determination. This becomes important for a variable like `age`; though `age` is coded as numeric data for most of the observations, there are some in which `age` is coded as `"89 or older"`. Without the `guess_max` argument, you will get warnings when loading the data.

Note also that only the variables of interest will be loaded, not the entire dataset. This will make for faster computation and knitting as you work on the lab.

10.19 Exercises

10.19.1 Part I: Exploratory Data Analysis

```
# See [Reorder factor levels by hand](https://forcats.tidyverse.org/reference/fct_relevel.html) for doc
```

1. The variable `natmass` will be the response variable in the model, and you want to compare more opinionated views to the moderate position. Recode `natmass` so it is a factor variable with `"About right"` as the baseline.

2. Recode `polviews` so it is a factor variable type with levels that are in an order that is consistent with question on the survey. *Note how the categories are spelled in the data.*

Make a plot of the distribution of `polviews`. Which political view occurs most frequently in this data set?

3. Make a plot displaying the relationship between `natmass` and `polviews`. Use the plot to describe the relationship between a person's political views and their views on mass transportation spending.

4. You want to use `age` as a quantitative variable in your model; however, it is currently a character data type because some observations are coded as `"89 or older"`. Recode `age` so that is a numeric variable.

Note: Before making the variable numeric, you will need to replace the values `"89 or older"` *with a single value.*

10.19.2 Part II: Multinomial Logistic Regression Model

5. You plan to fit a model using `age`, `sex`, `sei10`, and `region` to understand variation in opinions about spending on mass transportation. Briefly explain why you should fit a multinomial logistic model.
6. Fit the model described in the previous exercise and display the model output. Make any necessary adjustments to the variables so the intercept will have a meaningful interpretation. Be sure `About Right` is the baseline level. Be sure the full model displays in the knitted document.
7. Interpret the intercept associated with odds of having an opinion of "Too much" versus "About right".
8. Consider the relationship between age and one's opinion about spending on mass transportation.
   - Interpret the coefficient of age in terms of the log odds of having an opinion of "Too little" versus "About right".
   - Interpret the coefficient of age in terms of the odds of having an opinion of "Too little" versus "About right".
9. In general, what is the relationship between an person's age and their opinions on mass transportation spending?

Now that you have adjusted for some demographic factors, let's examine whether a person's political views has a significant impact on their attitude towards spending on mass transportation.

10. Conduct the appropriate test to determine if `polviews` is a significant predictor of attitude towards spending on mass transportation. State the null and alternative hypothesis, display all relevant code and output, and state your conclusion in the context of the problem.

Choose the appropriate model based on the results from the test. Use this model for the next part of the lab.

10.19.3 Part III: Model Fit

11. Calculate the predicted probabilities and residuals from your model.
12. Plot the binned residuals versus the predicted probabilities for each category of `natmass`. *You will have three plots.*

```
# You can change the size of your plots, so you can fit multiple plots on a single page. Include the ar

# See [Using R Markdown](https://rstudio.github.io/dygraphs/r-markdown.html) for an example.
```

13. Use binned residual plots to examine the residuals versus each of the quantitative variables.
    - Create binned plots of the residuals for each category of `natmass` versus `age`. *You will have three plots.*
    - Create binned plots of the residuals for each category of `natmass` versus `sei10`. *You will have three plots.*
14. To examine the residuals versus each categorical predictor, you will look at the average residuals for each each category of the categorical variables.
    - For each level of `natmass`, calculate the average residuals across categories of `sex`.
    - For each category of `natmass`, calculate the average residuals across categories of `region`.
    - For each category of `natmass`, calculate the average residuals across categories of `polviews`.

15. Based on the analysis of the residuals in Exercises 12 - 14, is the model an appropriate fit for the data? Explain.

Regardless of your asssesment of the residuals, use your model for the remainder of the lab.

10.19.4 Part IV: Using the Model

16. Use your model to describe the relationship between one's political views and their attitude towards spending on mass transportation.
17. Use your model to predict the category of `natmass` for each observation in your dataset. Display a table of the actual versus the predicted `natmass`. What is the misclassification rate?

10.19.5 Acknowledgements

The "Data" section is largely inspired by datasciencebox.org.

10.20 Putting It All Together

In this lab, you will put together everything you've learned thus far. Unlike previous lab assignments, your lab write up will be in the form of a small report (rather than numbered exercises). Though this analysis will not be as in-depth as your analysis in the final project, this assignment will give your group practice organizing the results of a statistical analysis to tell a complete narrative.

You will also practice imputing missing data and using k-fold cross validation to assess your model's performance on test data.

10.20.1 Packages

You will need the following packages for today's lab:

```
library(tidyverse)
library(dslabs)
## Fill in other packages as needed
```

10.20.2 Data

The data for this lab is the `gapminder` dataset in the **dslabs** package. This dataset contains health and income data for 184 countries during the years 1960 to 2016. After loading the dslabs package, you can type `?gapminder` in the console to to see the variables in the dataset.

You will only use data from 2011 in this lab.

10.20.3 Exercises

The goal of this analysis is to build a regression model that could be used to predict a country's gross domestic product (`gdp`) using the other characteristics included in the data.

Introduction

Brief introduction of the data and the research question

Exploratory Data Analysis

At a minimum, your exploratory data analysis should include the following:

- Analysis of each variable
- Dealing with missing values using imputation methods
- Analysis of the relationships between variables
- Discussion of any potential transformations, if needed

Regression Model

At a minimum, the discussion for the final regression model should include the following:

- Brief discussion about the type of model you used (multiple linear regression, logistic, multinomial logistic regression) and why
- Discussion of any transformations on the response and/or explanatory variables, if applicable
- Display of the final model
- Test of interesting interactions
- Conclusions drawn from the model, including any interesting insights based on the model coefficients

Assumptions

At a minimum, the discussion of model assumptions should include the following:

- Appropriate residual plots
- Check for influential points
- Check for multicollinearity
- Discussion of whether or not assumptions are met and how any issues may affect conclusions drawn from the model

Model Validation

At a minimum, the discussion of the model validation should include the following:

- Results and discussion from a 5-fold cross validation

Conclusion

Brief summary of the conclusions drawn from the analysis.

10.21 Movies Analysis

We will look at the relationship between budget and revenue for movies made in the United States in 1986 to 2016. The data is from the Internet Movie Database (IMDB).

```
library(readr)
library(tidyverse)
library(DT)
```

10.22 Data

The `movies` data set includes basic information about each movie including budget, genre, movie studio, director, etc. A full list of the variables may be found here.

```
movies <- read_csv("https://raw.githubusercontent.com/danielgrijalva/movie-stats/master/movies.csv")
```

```
movies <- movies %>%
  filter(country=="USA",
         !(genre %in% c("Musical","War","Western"))) #remove genres with < 10 movies
```

```
movies
```

10.23 Analysis

10.23.1 Part 1

We begin by looking at how the gross revenue (`gross`) has changed over time. Since we want to visualize the results, we will choose a few genres of interest for the analysis.

```
genre_list <- c("Horror", "Drama", "Action", "Animation")
```

```
movies %>%
  filter(genre %in% genre_list) %>%
  group_by(genre,year) %>%
  summarise(avg_gross = mean(gross)) %>%
  ggplot(mapping = aes(x = year, y = avg_gross, color=genre)) +
    geom_line() +
    ylab("Average Gross Revenue (in US Dollars)") +
    ggtitle("Gross Revenue Over Time")
```

10.23.2 Part 2

Next, let's see the relationship between a movie's budget and its gross revenue. Because there is a large range of values for budget and revenue, we will plot the log-transformed version of each variable to more easily visualize the relationship. We will talk more about variable transformations later in the semester.

```
movies %>%
  filter(genre %in% genre_list, budget > 0) %>%
  ggplot(mapping = aes(x=log(budget), y = log(gross), color=genre)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE) +
  xlab("Log-transformed Budget")+
  ylab("Log-transformed Gross Revenue") +
  facet_wrap(~ genre)
```

10.24 Next Steps

1. Put your name in the author field at the top of the file (in the `yaml` – we will discuss what this is at a later date). Knit again.
2. Change the genre names in parts 1 and 2 to genres that interest you. The spelling and capitalization must match what's in the data, so you can use the Appendix to see the correct spelling and capitalization. Knit again.

You have made your first data visualization!

10.25 Discussion Questions

1. Consider the plot in Part 1.
   - Describe how movie revenue has changed over time.
   - Suppose we use revenue as a measure of popularity. How has the popularity of each genre changed over time? In other words, are the genres that were most popular in 1986 still the most popular today?
2. Consider the plot in Part 2.
   - Which genre(s) tend to have the highest budgets?
   - In general, what is the relationship between a movie's budget and its total revenue? Are there any genres that show a different relationship between budget and revenue?

10.26 References

1. https://github.com/danielgrijalva/movie-stats
2. Internet Movie Database

10.27 Appendix

Below is a list of genres in the data set:

```
movies %>%
  arrange(genre) %>%
  select(genre) %>%
  distinct() %>%
  datatable()
```

10.28 In-Class Exercise: Advertising Analysis

In this mini analysis, we will work with the `Advertising` data used in Chapters 2 and 3 of *Introduction to Statistical Learning*.

10.28.1 Data and packages

We start with loading the packages we'll use.

```
library(readr)
library(tidyverse)
library(skimr)
library(broom)
```

```
advertising <- read_csv("data/advertising.csv")
```

We will analyze the advertising and sales data for 200 markets. The variables we'll use are

- `tv`: total spending on TV advertising (in $thousands)
- `radio`: total spending on radio advertising (in $thousands)
- `newspaper`: total spending on newspaper advertising (in $thousands)
- `sales`: total sales (in $millions)

10.28.2 Analysis

We'll begin the analysis by getting quick view of the data:

```
glimpse(advertising)
```

Next, we can calculate summary statistics for each of the variables in the data set.

```
advertising %>% skim()
```

1. What type of advertising has the smallest median spending?
2. What type of advertising has the largest variation in spending?
3. Describe the shape of the distribution of `sales`.

We are most interested in understanding how advertising spending affect sales. One way to quantify the relationship between the variables is by calculating the correlation matrix.

```
advertising %>%
  cor()
```

1. What is the correlation between `radio` and `sales`? Interpret this value.
2. What type of advertising has the strongest linear relationship with `sales`?

Below are visualizations of `sales` versus each explanatory variable.

```
advertising %>%
  ggplot(mapping = aes(x=tv,y=sales)) +
  geom_point(alpha=0.7) +
  geom_smooth(method="lm",se=FALSE,color="blue") +
  labs(title = "Sales vs. TV Advertising",
       x= "TV Advertising (in $thousands)",
       y="_____") #fill in the Y axis label
```

```
advertising %>%
  ggplot(mapping = aes(x=radio,y=sales)) +
  geom_point(alpha=0.7) +
  geom_smooth(method="lm",se=FALSE,color="red") +
  labs(title = "Sales vs. TV Advertising",
       x= "Radio Advertising (in $thousands)",
       y="Sales (in $millions)")
```

```
advertising %>%
  ggplot(mapping = aes(x=newspaper,y=sales)) +
  geom_point(alpha=0.7) +
  geom_smooth(method="lm",se=FALSE,color="purple") +
  labs(title = "Sales vs. Newspaper Advertising",
       x= "Newspaper Advertising (in $thousands)",
       y="Sales (in $millions)")
```

Since `tv` appears to have the strongest linear relationship with `sales`, let's calculate a simple linear regression model using these two variables.

```
ad_model <- lm(sales ~ tv, data=advertising)
ad_model
```

1. Write the model equation.
2. Interpret the intercept in the context of the problem.

3. Interpret the slope in the context of the problem.

10.29 In-Class Exercise: Beer Data Analysis

```
library(tidyverse)
library(readr)
library(broom)
```

```
beer <- read_csv("data/beer.csv")
```

In this analysis, we will analyze the relationship between the amount of alcohol (`PercentAlcohol`) and the caloric content (`CaloriesPer12Oz`) in domestic beers. Let `PercentAlcohol` be the predictor variable and `CaloriesPer12Oz` the response variable.

Due to limited class time, we will not do the exploratory data analysis in this example. In practice, however, you should always start with the exploratory data analysis.

You can add your answers to this R Markdown document.

1. Calculate a regression model to describe the relationship between `PercentAlcohol` and `CaloriesPer12Oz`. Display the model output.

```
model <- lm(CaloriesPer12Oz ~ PercentAlcohol, data=beer)
model %>%
  tidy(conf.int=TRUE)
```

2. Does it make sense to interpret the intercept? Why or why not?

There are non-alocoholic beers, so it is possible to have a meaningful interpretation of the intercept. In our data, however, there are very few beers with less than 3% alcoholic content, so it would not be wise to interpret the intercept. It is not safe to assume the same relationship between `PercentAlcohol` and `CaloriesPer12Oz` hold for beers with 0% alcohol; this would be extrapolation.

3. Interpret the 95% confidence interval for the slope in the context of the data.

We are 95% confident that the interval (26.557, 30.620) contains the true population slope for `PercentAlcohol`. This means we are 95% confident that for every 1% increase in alcohol content, the number of calories (per 12 oz) is expected to increase between 26.557 and 30.620 calories.

4. Find the critical value, t^*, used to calculate the 95% confidence interval. The code below is a guide; uncomment and complete the lines of code to calculate and display the critical value.

```
n <- nrow(beer)

df <- n-2
crit_val <- qt(0.975,df)
```

The critical value used to calculate the 95% confident interval for the slope is _____.

5. Interpret the test statistic in the context of the data

The estimated slope of 28.577 is 27.78 standard errors above the hypothesized mean of 0, assuming there is no linear relationship between percent alcohol and calories in domestic beers.

6. How was the p-value calculated? Fill in the code below to calculate the p-value. The code below is a guide; uncomment and complete the lines of code to calculate and display the p-value.

```
test_statistic <- 27.778990

prob <- 1 - pt(abs(test_statistic),df)

p_value <- 2 * prob
```

The p-value is _____. Given there is no linear relationship between `PercentAlcohol` and `CaloriesPer12Oz`, the probability of obtaining a test statistic with magnitude _____ or more extreme is _____.

7. Fill in the code below to calculate the predicted calories and corresponding 90% interval for a single beer with alcohol content of 4.3%.

```
x0 <-  data.frame(PercentAlcohol=4.3)
predict.lm(model,x0,interval="prediction",conf.level=0.9)
```

8. Fill in the code below to calculate the predicted calories and corresponding 90% interval for the subset of beers with alcohol content of 4.3%.

```
x0 <-  data.frame(PercentAlcohol=4.3)
predict.lm(model,x0,interval="confidence",conf.level= 0.9)
```

10.30 Analyzing Wages

```
library(tidyverse)
library(knitr)
library(broom)
library(Sleuth3)
```

```
wages <- case1202 %>%
  mutate(Female = ifelse(Sex=="Female",1,0)) %>%
  select(-Sal77,-Sex)
```

10.30.1 Initial model

```
model <- lm(Bsal ~ Senior + Age + Educ + Exper + Female,
            data=wages)
tidy(model,conf.int=TRUE)
```

10.30.2 Model with mean-centered variables

```
wages <- wages %>%
  mutate(SeniorCent = Senior - mean(Senior),
         AgeCent = Age-mean(Age),
         EducCent = Educ - mean(Educ),
         ExperCent = Exper - mean(Exper))
```

- Calculate the regression model using the mean-centered variables.
- How did the model change?

10.30.3 Model with indicator variables

- Use the code below to create a categorical variable for `Educ`.

```
wages <- wages %>%
  mutate(EducCat = as.factor(Educ))
```

- Calculate the regression model using `EducCat` instead of `Educ`.

10.31 Exam 01 Review

```
library(tidyverse)
library(broom)
library(rms)
library(knitr)
```

```
set.seed(12)
diamonds_samp <- ggplot2::diamonds %>%
  filter(carat < 1.1) %>%
  sample_n(300) %>%
  mutate(log_price = log(price),
         caratCent = carat - mean(carat),
         caratCent_sq = caratCent^2,
        color = factor(as.character(color)), # to fix variable format in model output
        clarity = factor(as.character(clarity)) # to fix variable format in model output
)
```

10.31.1 Main Effects Model

- **Why should we use `logprice` instead of `price` as the response variable? In other words, what is an example of previous analysis that could have been done to help us determine whether to use `logprice` or `price`?**

To determine if a transformation on the response variable is needed, we can examine the following:

- The distribution of the response variable to see if there is extreme skewness
- The plot of the residuals vs. predicted to check for non-constant variance
- The histogram and QQ-plot of the residuals to see if there is extreme skewness in the residuals

Below is the model with `log_price` as the response and `caratCent`, `color`, and `clarity` as the predictor variables.

```
model_orig <- lm(log_price ~ caratCent +  color + clarity, data=diamonds_samp)
kable(tidy(model_orig),format="markdown")
```

- **What is the baseline level of `color`? What is the baseline level of `clarity`?**

The baseline level of `color` is `D`. The baseline level of `clarity` is `I1`.

- **Interpret the intercept in terms of `price`.**

```
coef <-  model_orig$coefficients
```

We expect the median price of diamonds with color D, clarity I1, and the mean carat weight (_____) to be approximately exp(_____) = _____.

- **Describe the difference in the typical prices of diamonds that are color E and diamonds that are color D, holding all else constant.**

The difference in terms of the log(price) is the coefficient of `colorE`, _____. Therefore, the difference in terms of the price is

```
(diff_e_d <- exp(coef[3]))
```

Therefore, holding all else constant, diamonds that are color E are expected to have a median price that is _____ times the median price of diamonds that are color D.

- **Describe the difference in the typical prices of diamonds that are color E and diamonds that are color G, holding all else constant.**

```
(diff_e_g_log <- coef[3] - coef[5])
```

Therefore, the difference in terms of the price is

```
(diff_e_g <- exp(diff_e_g_log))
```

Therefore, holding all else constant, diamonds that are color E are expected to have a median price that is _____ times the median price of diamonds that are color G.

- **What is the predicted price of a single diamond that has color E, clarity VS2 and is 0.3 carats? Finish the code below the predicted value and the corresponding interval.**

```
x0 <- data.frame(color="E", clarity="VS2", carat=0.3)
x0 <- x0 %>% mutate(
  caratCent = carat - mean(diamonds_samp$carat),
  caratCent_sq = caratCent^2
)

(exp(predict(model_orig,x0,interval="prediction"))) #interval to predict for single observation
```

- **Suppose we wish to find the predicted median price of subset of all diamonds with color E, clarity VS2, and 0.3 carats. How do you expect the predicted price to change? How do you expect the corresponding interval to change?**
  - The predicted price won't change, but the interval will be more narrow.
- **Write code to find the predicted price and corresponding interval for the median price for the subset of all diamonds with color E, clarity VS2 and 0.3 carats.**

```
(exp(predict(model_orig,x0,interval="confidence"))) #interval to predict typical price for subset
```

Use the code below to obtain the ANOVA table for this model.

```
anova(model_orig)
```

- **What is the estimated regression variance?**

The estimated regression variance is the Residual Mean Square, 0.0204.

- **Compare R^2 and $Adj.R^2$. What does this comparison tell you about the predictors in the model?**

R^2 is _____ and Adjusted R^2 is _____. These values are very close, indicating that the predictors in the model are important for understanding variation in price. There aren't a lot of predictors in the model that aren't significant.

- Use the code below to calculate the VIF for this model.

```
vif(model_orig)
```

- **This model has potential problems with multicollinearity. How did we come to this conclusion? Which variables are highly collinear?**

We know this model has potential problems with multicollinearity, because there are multiple predictors with VIFs close or above 10. The variables with high collinearity are the indicator variables for `clarity`.

- **Why do you think this multicollinearity is occurring?** ***Hint: Examine the distribution of the variable(s) that have high multicollinearity.***

Let's look at the distribution of `clarity`.

```
diamonds_samp %>%
  count(clarity)
```

The baseline level is `I1`, and there are only 2 observations out of _____ with this level for clarity. Because there are so few observations at the baseline level, it is almost as if we have no baseline level for the categorical predictor `clarity` in the model. Remember, if we have no baseline level for a categorical variable in the model and there is an intercept, then the indicator variables are just linear combinations for one another. In this case, the indicator variables aren't exact linear combinations of one another, but they are highly collinear.

This multicollinearity is reduced when the baseline level is changed to a different level of `clarity`. Below is the VIF for a model with `IF`, the highest level of clarity as the baseline.

```
diamonds_samp %>%
  mutate(clarity = fct_rev(clarity)) %>% #reverse the factoring order of clarity
  lm(log_price ~ caratCent +  color + clarity, data=.) %>%
  vif()
```

10.32 Model with Interactions

- Below is the model that includes an interaction term between `caratCent` and `color`. What is the appropriate method to determine if the interaction is significant?
- We should use the nested F test.

```
model_int <- lm(log_price ~ caratCent + color +
                  clarity + caratCent*color, data=diamonds_samp)
```

- Conduct the test you listed above. What is your conclusion?

```
anova(model_orig, model_int)
```

The p-value is 0.4104, indicating that there is not sufficient evidence that the interaction between carat and color is significant.

Regardless of your answer to the previous question, use ***`model_int`*** *to answer the next two questions:*

```
kable(tidy(model_int),format="markdown")
```

```
coef_int <- model_int$coefficients
```

- **What is the estimated slope of `caratCent` for a diamond with `color==D`?**
- **What is the estimated slope of `caratCent` for a diamond with `color==J`?**

10.33 Model Selection

```
library(tidyverse)
library(knitr)
library(broom)
library(Sleuth3)
library(leaps)
```

```
sat_scores <- case1201 %>%
  select(-State) #remove the state variable
```

10.34 Backward selection "manually"

- Manually perform backward selection using Adj. R^2 as the selection criteria. Show each step of the selection process. To help you get started, the full model and the code for the first set of models to test are below. You will need to find Adj. R^2 for each model.

```
full_model <- lm(SAT ~ ., data = sat_scores)
```

```
m1 <- lm(SAT ~ Income + Years + Public + Expend + Rank, data = sat_scores)
```

```
m2 <- lm(SAT ~ Takers + Years + Public + Expend + Rank, data = sat_scores)
```

```
m3 <- lm(SAT ~ Takers + Income + Public + Expend + Rank, data = sat_scores)
```

```
m4 <- lm(SAT ~ Takers + Years + Income + Expend + Rank, data = sat_scores)
```

```
m5 <- lm(SAT ~ Takers + Years + Public + Income + Expend, data = sat_scores)
```

```
m6 <- lm(SAT ~ Takers + Years + Public + Income + Rank, data = sat_scores)
```

Continue the model selection until you have a final model. Show each step of the model selection process.

10.35 Backward selection using regsubsets

- Use the `regsubsets` function to perform backward selection using Adj. R^2 as the selection criteria. Are the variables the same as the ones at you chose? Is the Adj. R^2 the same?

10.36 Changing selection criteria

- Use the `regsubsets` function to perform backward selection using BIC as the selection criteria. What variables were chosen for the follow model? How does this model compare to the one selected using Adj. R^2?
- Use the `step` function to perform backward selection using AIC as the selection criteria. What variables were chosen for the follow model? How does this model compare to the models chosen from the other selection criteria?

10.37 Different selection procedure

- Use forward or stepwise selection to choose a model. Choose the criteria you will use to select the model.
- How does this model compare to the previous selected models?

10.38 Choosing a final model

- You likely have at least 2 different models chosen by the various model selection procedures. Which variables will you include in your final model? Why did you choose this to be your final model?

10.39 Logistic Regression

The goal of this exercise is to walk through a logistic regression analysis. It will give you a basic idea of the analysis steps and thought-process; however, due to class time constraints, this analysis is not exhaustive.

```
library(tidyverse)
library(broom)
library(rms)
## add any other packages as needed
```

This data is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The goal is to predict whether a patient has a 10-year risk of future coronary heart disease. The dataset includes the following:

- `male`: 0 = Female; 1 = Male
- `age`: Age at exam time.
- `education`: 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = College
- `currentSmoker`: 0 = nonsmoker; 1 = smoker
- `cigsPerDay`: number of cigarettes smoked per day (estimated average)
- `BPMeds`: 0 = Not on Blood Pressure medications; 1 = Is on Blood Pressure medications
- `prevalentStroke`
- `prevalentHyp`
- `diabetes`: 0 = No; 1 = Yes
- `totChol`: total cholesterol (mg/dL)
- `sysBP`: systolic blood pressure (mmHg)
- `diaBP`: diastolic blood pressure (mmHg)
- `BMI`: BodyMass Index calculated as: Weight (kg) / Height(meter-squared)
- `heartRate` Beats/Min (Ventricular)
- `glucose`: total glucose mg/dL
- `TenYearCHD`: 0 = Patient doesn't have 10-year risk of future coronary heart disease; 1 = Patient has 10-year risk of future coronary heart disease;

```
fram_data <- read_csv("data/framingham.csv") %>%
  drop_na() %>%
  mutate(education = case_when(
    education == 1 ~ "Some HS",
    education == 2 ~ "HS or GED",
    education == 3 ~ "Some College",
    education == 4 ~ "College"
  ),
```

```
currentSmoker = if_else(currentSmoker == 0, "nonsmoker", "smoker"),
diabetes = if_else(diabetes == 0,"No", "Yes"),
male = factor(male)
)
```

- Fit a full model (main effects only) with `TenYearCHD` as the response. Display the model output.
- Based on the goal of the analysis, should the full model be the final model? Why or why not?
- Use the `step` function to conduct backward model selection. What is selection criteria used by the `step` function?
  - Display the final model.
- There is reason to believe that the factors related to coronary heart disease may have different effects for men and women. We would like to include this information in the model. Use the drop-in-deviance test to test at least three interactions with `male`.
  - Which interactions did you choose? Why?
  - Include the output from the tests.
- Use the results from model selection and the drop-in-deviance test to select a final model. Display the model below.
- Plot and analyze the binned residuals for the final model. Include all appropriate plots. What is your assessment on the model fit based on these plots?
- Plot and analyze the ROC curve. Based on the ROC curve, does the model fit the data well?
- A doctor plans to use the results from your model to help select patients for a new heart disease prevention program. She asks you which threshold would be best to select patients for this program. What threshold would you recommend to the doctor? Why?

10.40 References

- Data obtained from https://www.kaggle.com/neisha/heart-disease-prediction-using-logistic-regression/data

10.41 Multinomial Logistic Regression

The main objective of this analysis is to understand how encouragement affects the frequency that children watch *Sesame Street*. We will use the following variables:

Response:

- `viewcat`
  - 1: rarely watched show
  - 2: once or twice a week
  - 3: three to five times a week
  - 4: watched show on average more than five times a week

Predictors:

- `age`: child's age in months
- `prenumb`: score on numbers pretest (0 to 54)
- `prelet`: score on letters pretest (0 to 58)
- `viewenc`: 1: encouraged to watch, 2: not encouraged

- site:
  - 1: three to five year old from disadvantaged inner city area
  - 2: four year old from advantaged suburban area
  - 3: from advantaged rural area
  - 4: from disadvantaged rural area
  - 5: from Spanish speaking home

```
# read in dataset
sesame <- read_csv("data/sesame.csv")
```

```
# mean-center relevant continuous variables, make categorical variables factors
sesame <- sesame %>%
  mutate(viewcat = as.factor(viewcat),
         site = as.factor(site),
         prenumbCent = prenumb - mean(prenumb),
         preletCent = prelet - mean(prelet),
         ageCent = age - mean(age),
         viewenc = ifelse(viewenc == 1, "Encouraged", "Not Encouraged"))
```

10.41.1 Questions

1. We will build a model to predict how often a child in this study watched *Sesame Street*. What type of model should we build? Why?

2. Describe how you would conduct exploratory data analysis. What plots and/or summary statistics would you include? What information would you learn from the exploratory data analysis?

```
model1 <- multinom(viewcat ~ site + viewenc + prenumbCent + preletCent + ageCent,
                   data = sesame)
kable(tidy(model1, conf.int=TRUE, exponentiate = FALSE),
      format = "markdown")
```

3. Interpret the intercept associated with the odds of `viewcat == 2` versus `viewcat == 1`.

4. Interpret the effect of the numbers pretest score on the odds of viewership.

5. The primary objective of the experiment was to understand the effect of encouragement `viewenc` on viewership. Does encouragement have a significant effect on viewership? If so, describe the effect. Otherwise, explain why not.

6. We want to test if there are any significant interactions with `viewenc` and the pretests. We create a model that includes the variables from `model1` along with `viewenc*preletCent` and `viewenc*prenumbCent`.

```
model2 <- multinom(viewcat ~ site + viewenc + prenumbCent + preletCent + ageCent +
                     viewenc*preletCent + viewenc*prenumbCent,
                   data = sesame)
```

The results from the drop-in-deviance test are shown below. Is there evidence of a significant interaction effect? Explain.

```
anova(model1, model2, test = "Chisq")
```

7. How would you assess the appropriateness of the model flit? Describe the plots, tables, and/or calculations you would create to assess model fit.

10.41.2 References

Data from http://www2.stat.duke.edu/~jerry/sta210/sesamelab.html

10.42 Dealing with Missing Data

```
library(tidyverse)
library(knitr)
library(broom)
library(skimr)
library(nnet)
```

```
nhanes <- mice::nhanes2
nhanes
```

1. Explore missingness

```
library(nnet)
m <- multinom(age ~ bmi + hyp + chl, data = mice::nhanes2)
knitr::kable(tidy(m, conf.int = T),format = "markdown")
```

2. Complete-case analyis

```
complete <- nhanes %>% drop_na()
m <- multinom(age ~ bmi + hyp + chl, data = complete)
knitr::kable(tidy(m, conf.int = T),format = "markdown")
```

3. Single imputation
4. Indicator variables

10.43 Matrix Form of Linear Regression

This document provides the details for the matrix form of multiple linear regression. We assume the reader has familiarity with some matrix alegbra. Please see Chapter 1 of *An Introduction to Statistical Learning* for a brief review of matrix algebra.

10.44 Introduction

Suppose we have n observations. Let the i^{th} be $(x_{i1}, \ldots, x_{ip}, y_i)$, such that $x_{i1}, \ldots, x_{ip}$ are the explanatory variables (predictors) and y_i is the response variable. We assume the data can be modeled using the least-squares regression model, such that the mean response for a given combination of explanatory variables follows the form in (10.23).

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{10.1}$$

We can write the response for the i^{th} observation as shown in (10.24)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{10.2}$$

such that ϵ_i is the amount y_i deviates from $\mu\{y|x_{i1}, \ldots, x_{ip}\}$, the mean response for a given combination of explanatory variables. We assume each $\epsilon_i \sim N(0, \sigma^2)$, where σ^2 is a constant variance for the distribution of the response y for any combination of explanatory variables $x_1, \ldots, x_p$.

10.45 Matrix Form for the Regression Model

We can represent the (10.23) and (10.24) using matrix notation. Let

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{10.3}$$

Thus,

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

Therefore the estimated response for a given combination of explanatory variables and the associated residuals can be written as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{10.4}$$

10.46 Estimating the Coefficients

The least-squares model is the one that minimizes the sum of the squared residuals. Therefore, we want to find the coefficients, $\hat{\boldsymbol{\beta}}$ that minimizes

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{10.5}$$

where $\mathbf{e}^T$, the transpose of the matrix $\mathbf{e}$.

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - (\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}) \tag{10.6}$$

Note that $(\mathbf{Y^T X}\hat{\boldsymbol{\beta}})^T = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y}$. Since these are both constants (i.e. 1×1 vectors), $\mathbf{Y^T X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y}$. Thus, (10.6) becomes

$$\mathbf{Y}^T\mathbf{Y} - 2\mathbf{X}^T\hat{\boldsymbol{\beta}}^T\mathbf{Y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \tag{10.7}$$

Since we want to find the $\hat{\boldsymbol{\beta}}$ that minimizes (10.5), will find the value of $\hat{\boldsymbol{\beta}}$ such that the derivative with respect to $\hat{\boldsymbol{\beta}}$ is equal to 0.

$$\begin{aligned}
\frac{\partial \mathbf{e}^T\mathbf{e}}{\partial \hat{\boldsymbol{\beta}}} &= \frac{\partial}{\partial \hat{\boldsymbol{\beta}}}(\mathbf{Y}^T\mathbf{Y} - 2\mathbf{X}^T\hat{\boldsymbol{\beta}}^T\mathbf{Y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}) = 0 \\
&\Rightarrow -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \\
&\Rightarrow 2\mathbf{X}^T\mathbf{Y} = 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&\Rightarrow \mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&\Rightarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&\Rightarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{I}\hat{\boldsymbol{\beta}}
\end{aligned} \tag{10.8}$$

Thus, the estimate of the model coefficients is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

10.47 Variance-covariance matrix of the coefficients

We will use two properties to derive the form of the variance-covarinace matrix of the coefficients:

1. $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I$
2. $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\epsilon$

First, we will show that $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I$

$$\begin{aligned}
E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] &= E\begin{bmatrix}\epsilon_1 & \epsilon_2 & \dots & \epsilon_n\end{bmatrix}\begin{bmatrix}\epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n\end{bmatrix} \\
&= E\begin{bmatrix}\epsilon_1^2 & \epsilon_1\epsilon_2 & \dots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & \dots & \epsilon_2\epsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & \dots & \epsilon_n^2\end{bmatrix} \\
&= \begin{bmatrix}E[\epsilon_1^2] & E[\epsilon_1\epsilon_2] & \dots & E[\epsilon_1\epsilon_n] \\ E[\epsilon_2\epsilon_1] & E[\epsilon_2^2] & \dots & E[\epsilon_2\epsilon_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[\epsilon_n\epsilon_1] & E[\epsilon_n\epsilon_2] & \dots & E[\epsilon_n^2]\end{bmatrix}
\end{aligned} \tag{10.9}$$

Recall, the regression assumption that the errors $\epsilon_i's$ are Normally distributed with mean 0 and variance σ^2. Thus, $E(\epsilon_i^2) = Var(\epsilon_i) = \sigma^2$ for all i. Additionally, recall the regression assumption that the errors are uncorrelated, i.e. $E(\epsilon_i\epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$ for all i, j. Using these assumptions, we can write (10.9) as

$$E[\epsilon\epsilon^T] = \begin{bmatrix}\sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2\end{bmatrix} = \sigma^2\mathbf{I} \tag{10.10}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

Next, we show that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\epsilon$.

Recall that the $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. Then,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon
\end{aligned} \tag{10.11}$$

Using these two properties, we derive the form of the variance-covariance matrix for the coefficients. Note that the covariance matrix is $E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T]$

$$\begin{aligned}
E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] &= E[(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} - \boldsymbol{\beta})(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} - \boldsymbol{\beta})^T] \\
&= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned} \tag{10.12}$$

10.48 Log Transformations in Linear Regression

This document provides details about the model interpretion when the predictor and/or response variables are log-transformed. For simplicity, we will discuss transformations for the simple linear regression model:

$$y = \beta_0 + \beta_1 x \tag{10.13}$$

All results and interpretations can be easily extended to transformations in multiple regression models.

Note: *log* refers to the natural logarithm.

10.49 Log-transformation on the response variable

Suppose we fit a linear regression model with $\log(y)$, the log-transformed y, as the response variable. Under this model, we assume a linear relationship exists between x and $\log(y)$, such that $\log(y) \sim N(\beta_0 + \beta_1 x, \sigma^2)$ for some β_0, β_1 and σ^2. In other words, we can model the relationship between x and $\log(y)$ using the model in (10.14).

$$\log(y) = \beta_0 + \beta_1 x \tag{10.14}$$

If we interpret the model in terms of $\log(y)$, then we can use the usual interpretations for slope and intercept. When reporting results, however, it is best to give all interpretations in terms of the original response variable y, since interpretations using log-transformed variables are often more difficult to truly understand.

In order to get back on the original scale, we need to use the exponential function (also known as the anti-log), $\exp\{x\} = e^x$. Therefore, we use the model in (10.14) for interpretations and predictions, we will use (10.15) to state our conclusions in terms of y.

$$\begin{aligned} \exp\{\log(y)\} &= \exp\{\beta_0 + \beta_1 x\} \\ \Rightarrow y &= \exp\{\beta_0 + \beta_1 x\} \\ \Rightarrow y &= \exp\{\beta_0\}\exp\{\beta_1 x\} \end{aligned} \tag{10.15}$$

In order to interpret the slope and intercept, we need to first understand the relationship between the mean, median and log transformations.

10.49.1 Mean, Median, and Log Transformations

Suppose we have a dataset `y` that contains the following observations:

```
y <- c(3,5,6,7,8)
y
```

If we log-transform the values of `y` then calculate the mean and median, we have

```
log_y <- tibble(log_y = log(y))
summary <- log_y %>%
  summarise(mean_log_y = mean(log_y), median_log_y = median(log_y))
kable(summary,digits=5)
```

If we calculate the mean and median of `y`, then log-transform the mean and median, we have

```
centers <- tibble(y) %>% summarise(mean_y = mean(y), median_y = median(y))
summary2 <- centers %>%
  summarise(log_mean = log(mean_y), log_median = log(median_y))
kable(summary2,digits=5)
```

This is a simple illustration to show

1. Mean[$\log(y)$] $\neq$ log[Mean(y)] - the mean and log are not commutable
2. Median[$\log(y)$] = log[Median(y)] - the median and log are commutable

10.49.2 Interpretaton of model coefficients

Using (10.14), the mean $\log(y)$ for any given value of x is $\beta_0 + \beta_1 x$; however, this does **not** indicate that the mean of $y = \exp\{\beta_0 + \beta_1 x\}$ (see previous section). From the assumptions of linear regression, we assume that for any given value of x, the distribution of $\log(y)$ is Normal, and therefore symmetric. Thus the median of $\log(y)$ is equal to the mean of $\log(y)$, i.e Median($\log(y)$) $= \beta_0 + \beta_1 x$.

Since the log and the median are commutable, Median($\log(y)$) $= \beta_0 + \beta_1 x \Rightarrow$ Median(y) $= \exp\{\beta_0 + \beta_1 x\}$. Thus, when we log-transform the response variable, the interpretation of the intercept and slope are in terms of the effect on the **median** of y.

Intercept: The intercept is expected median of y when the predictor variable equals 0. Therefore, when $x = 0$,

$$\begin{aligned} \log(y) &= \beta_0 + \beta_1 \times 0 = \beta_0 \\ \Rightarrow y &= \exp\{\beta_0\} \end{aligned} \tag{10.16}$$

Interpretation: When $x = 0$, the median of y is expected to be $\exp\{\beta_0\}$.

Slope: The slope is the expected change in the median of y when x increases by 1 unit. The change in the median of y is

$$\exp\{[\beta_0+\beta_1(x+1)]-[\beta_0+\beta_1 x]\} = \frac{\exp\{\beta_0+\beta_1(x+1)\}}{\exp\{\beta_0+\beta_1 x\}} = \frac{\exp\{\beta_0\}\exp\{\beta_1 x\}\exp\{\beta_1\}}{\exp\{\beta_0\}\exp\{\beta_1 x\}} = \exp\{\beta_1\} \tag{10.17}$$

Thus, the median of y for $x+1$ is $\exp\{\beta_1\}$ times the median of y for x.

Interpretation: When x increases by one unit, the median of y is expected to multiply by a factor of $\exp\{\beta_1\}$.

10.50 Log-transformation on the predictor variable

Suppose we fit a linear regression model with $\log(x)$, the log-transformed x, as the predictor variable. Under this model, we assume a linear relationship exists between $\log(x)$ and y, such that $y \sim N(\beta_0 + \beta_1 \log(x), \sigma^2)$ for some β_0, β_1 and σ^2. In other words, we can model the relationship between $\log(x)$ and y using the model in (10.18).

$$y = \beta_0 + \beta_1 \log(x) \tag{10.18}$$

Intercept: The intercept is the mean of y when $\log(x) = 0$, i.e. $x = 1$.

Interpretation: When $x = 1$ ($\log(x) = 0$), the mean of y is expected to be β_0.

Slope: The slope is interpreted in terms of the change in the mean of y when x is mutliplied by a factor of C, since $\log(Cx) = \log(x) + \log(C)$. Thus, when x is multiplied by a factor of C, the change in the mean of y is

$$\begin{aligned} -[\beta_0 + \beta_1 \log(x)] &= \beta_1[\log(Cx) - \log(x)] \\ &= \beta_1[\log(C) + \log(x) - \log(x)] \\ &= \beta_1 \log(C) \end{aligned} \tag{10.19}$$

Thus the mean of y changes by $\beta_1 \log(C)$ units.

Interpretation: When x is multiplied by a factor of C, the mean of y is expected to change by $\beta_1 \log(C)$ units. For example, if x is doubled, then the mean of y is expected to change by $\beta_1 \log(2)$ units.

10.51 Log-transformation on the the response and predictor variable

Suppose we fit a linear regression model with $\log(x)$, the log-transformed x, as the predictor variable and $\log(y)$, the log-transformed y, as the response variable. Under this model, we assume a linear relationship

exists between $\log(x)$ and $\log(y)$, such that $\log(y) \sim N(\beta_0 + \beta_1 \log(x), \sigma^2)$ for some β_0, β_1 and σ^2. In other words, we can model the relationship between $\log(x)$ and $\log(y)$ using the model in (10.20).

$$\log(y) = \beta_0 + \beta_1 \log(x) \tag{10.20}$$

Because the response variable is log-transformed, the interpretations on the original scale will be in terms of the median of y (see the section on the log-transformed response variable for more detail).

Intercept: The intercept is the mean of y when $\log(x) = 0$, i.e. $x = 1$. Therefore, when $\log(x) = 0$,

$$\begin{aligned} \log(y) &= \beta_0 + \beta_1 \times 0 = \beta_0 \\ \Rightarrow y &= \exp\{\beta_0\} \end{aligned} \tag{10.21}$$

Interpretation: When $x = 1$ ($\log(x) = 0$), the median of y is expected to be $\exp\{\beta_0\}$.

Slope: The slope is interpreted in terms of the change in the median y when x is mutliplied by a factor of C, since $\log(Cx) = \log(x) + \log(C)$. Thus, when x is multiplied by a factor of C, the change in the median of y is

$$\begin{aligned} \exp\{[\beta_0 + \beta_1 \log(Cx)] - [\beta_0 + \beta_1 \log(x)]\} &= \exp\{\beta_1[\log(Cx) - \log(x)]\} \\ &= \exp\{\beta_1[\log(C) + \log(x) - \log(x)]\} \\ &= \exp\{\beta_1 \log(C)\} = C^{\beta_1} \end{aligned} \tag{10.22}$$

Thus, the median of y for Cx is C^{β_1} times the median of y for x.

Interpretation: When x is multiplied by a factor of C, the median of y is expected to multiple by a factor of C^{β_1}. For example, if x is doubled, then the median of y is expected to multiply by 2^{β_1}.

10.52 Details about Model Diagnostics

This document discusses some of the mathematical details of the model diagnostics - leverage, standardized residuals, and Cook's distance. We assume the reader knowledge of the matrix form for multiple linear regression.Please see Matrix Form of Linear Regression for a review.

10.53 Introduction

Suppose we have n observations. Let the i^{th} be $(x_{i1}, \ldots, x_{ip}, y_i)$, such that $x_{i1}, \ldots, x_{ip}$ are the explanatory variables (predictors) and y_i is the response variable. We assume the data can be modeled using the least-squares regression model, such that the mean response for a given combination of explanatory variables follows the form in (10.23).

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{10.23}$$

We can write the response for the i^{th} observation as shown in (10.24)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{10.24}$$

such that ϵ_i is the amount y_i deviates from $\mu\{y|x_{i1}, \ldots, x_{ip}\}$, the mean response for a given combination of explanatory variables. We assume each $\epsilon_i \sim N(0, \sigma^2)$, where σ^2 is a constant variance for the distribution of the response y for any combination of explanatory variables $x_1, \ldots, x_p$.

10.54 Matrix Form for the Regression Model

We can represent the (10.23) and (10.24) using matrix notation. Let

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{10.25}$$

Thus,

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

Therefore the estimated response for a given combination of explanatory variables and the associated residuals can be written as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{10.26}$$

10.55 Hat Matrix & Leverage

Recall from the notes **Matrix Form of Linear Regression** that $\hat{\boldsymbol{\beta}}$ can be written as the following:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{10.27}$$

Combining (10.26) and (10.27), we can write $\hat{\mathbf{Y}}$ as the following:

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \end{aligned} \tag{10.28}$$

We define the **hat matrix** as an $n \times n$ matrix of the form $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Thus (10.28) becomes

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \tag{10.29}$$

The diagonal elements of the hat matrix are a measure of how far the predictor variables of each observation are from the means of the predictor variables. For example, h_{ii} is a measure of how far the values of the predictor variables for the i^{th} observation, $x_{i1}, x_{i2}, \ldots, x_{ip}$, are from the mean values of the predictor variables, $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$. In the case of simple linear regression, the i^{th} diagonal, h_{ii}, can be written as

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

We call these diagonal elements, the **leverage** of each observation.

The diagonal elements of the hat matrix have the following properties:

- $0 \leq h_i i \leq 1$
- $\sum_{i=1}^{n} h_{ii} = p + 1$, where p is the number of predictor variables in the model.
- The mean hat value is $\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p+1}{n}$.

Using these properties, we consider a point to have **high leverage** if it has a leverage value that is more than 2 times the average. In other words, observations with leverage greater than $\frac{2(p+1)}{n}$ are considered to be **high leverage** points, i.e. outliers in the predictor variables. We are interested in flagging high leverage points, because they may have an influence on the regression coefficients.

When there are high leverage points in the data, the regression line will tend towards those points; therefore, one property of high leverage points is that they tend to have small residuals. We will show this by rewriting the residuals from (10.26) using (10.29).

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{HY} \\ &= (1 - \mathbf{H})\mathbf{Y} \end{aligned} \tag{10.30}$$

Note that the identity matrix and hat matrix are **idempotent**, i.e. $\mathbf{II} = \mathbf{I}$, $\mathbf{HH} = \mathbf{H}$. Thus, ($\mathbf{I} - \mathbf{H}$ is also idempotent. These matrices are also symmetric. Using these properties and (10.30), we have that the variance-covariance matrix of the residuals $\boldsymbol{e}$, is

$$\begin{aligned} Var(\mathbf{e}) &= \mathbf{e}\mathbf{e}^T \\ &= (1 - \mathbf{H})Var(\mathbf{Y})^T(1 - \mathbf{H})^T \\ &= (1 - \mathbf{H})\hat{\sigma}^2(1 - \mathbf{H})^T \\ &= \hat{\sigma}^2(1 - \mathbf{H})(1 - \mathbf{H}) \\ &= \hat{\sigma}^2(1 - \mathbf{H}) \end{aligned} \tag{10.31}$$

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1}$ is the estimated regression variance. Thus, the variance of the i^{th} residual is $Var(e_i) = \hat{\sigma}^2(1-h_{ii})$. Therefore, the higher the leverage, the smaller the variance of the residual. Because the expected value of the residuals is 0, we conclude that points with high leverage tend to have smaller residuals than points with lower leverage.

10.56 Standardized Residuals

In general, we standardize a value by shifting by the expected value and rescaling by the standard deviation (or standard error). Thus, the i^{th} standardized residual takes the form

$$std.res_i = \frac{e_i - E(e_i)}{SE(e_i)}$$

The expected value of the residuals is 0, i.e. $E(e_i) = 0$. From (10.31), the standard error of the residual is $SE(e_i) = \hat{\sigma}\sqrt{1 - h_{ii}}$. Therefore,

$$std.res_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{10.32}$$

10.57 Cook's Distance

Cook's distance is a measure of how much each observation influences the model coefficients, and thus the predicted values. The Cook's distance for the i^{th} observation can be written as

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})^T(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{(p+1)\hat{\sigma}} \tag{10.33}$$

where $\hat{\mathbf{Y}}_{(i)}$ is the vector of predicted values from the model fitted when the i^{th} observation is deleted. Cook's Distance can be calculated without deleting observations one at a time, since (10.34) below is mathematically equivalent to (10.33).

$$D_i = \frac{1}{p+1} std.res_i^2 \left[\frac{h_{ii}}{(1 - h_{ii})}\right] = \frac{e_i^2}{(p+1)\hat{\sigma}^2(1 - h_{ii})}\left[\frac{h_{ii}}{(1 - h_{ii})}\right] \tag{10.34}$$

10.58 Model Selection Criteria: AIC & BIC

This document discusses some of the mathematical details of Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC). We assume the reader knowledge of the matrix form for multiple linear regression.Please see Matrix Form of Linear Regression for a review.

10.59 Maximum Likelihood Estimation of β and σ

To understand the formulas for AIC and BIC, we will first briefly explain the likelihood function and maximum likelihood estimates for regression.

Let $\mathbf{Y}$ be $n \times 1$ matrix of responses, $\mathbf{X}$, the $n \times (p+1)$ matrix of predictors, and $\boldsymbol{\beta}$, $(p+1) \times 1$ matrix of coefficients. If the multiple linear regression model is correct then,

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2) \tag{10.35}$$

When we do linear regression, our goal is to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2 from (10.35). In Matrix Form of Linear Regression, we showed a way to estimate these parameters using matrix alegbra. Another approach for estimating $\boldsymbol{\beta}$ and σ^2 is using *maximum likelihood estimation.*

A **likelihood function** is used to summarise the evidence from the data in support of each possible value of a model parameter. Using (10.35), we will write the likelihood function for linear regression as

$$L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^T(Y_i - \mathbf{X}_i\boldsymbol{\beta})\right\} \tag{10.36}$$

where Y_i is the i^{th} response and $\mathbf{X}_i$ is the vector of predictors for the i^{th} observation. One approach estimating $\boldsymbol{\beta}$ and σ^2 is to find the values of those parameters that maximize the likelihood in (10.36), i.e. **maximum likelhood estimation**. To make the calculations more manageable, instead of maximizing the likelihood function, we will instead maximize its logarithm, i.e. the log-likelihood function. The values of the parameters that maximize the log-likelihood function are those that maximize the likelihood function. The log-likelihood function we will maximize is

$$\begin{aligned}\log L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^T(Y_i - \mathbf{X}_i\boldsymbol{\beta}) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\end{aligned} \tag{10.37}$$

[–insert details MLES–]

The maximum likelihood estimate of $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \qquad \hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{n}RSS \tag{10.38}$$

where RSS is the residual sum of squares. Note that the maximum likelihood estimate is not exactly equal to the estimate of σ^2 we typically use $\frac{RSS}{n-p-1}$. This is because the maximum likelihood estimate of σ^2 in (10.38) is a *biased* estimator of σ^2. When n is much larger than the number of predictors p, then the differences in these two estimates are trivial.

10.60 AIC

Akaike's Information Criterion (AIC) is

$$AIC = -2\log L + 2(p+1) \tag{10.39}$$

where $\log L$ is the log-likelihood. This is the general form of AIC that can be applied to a variety of models, but for now, let's focus on AIC for mutliple linear regression.

$$\begin{aligned}AIC &= -2\log L + 2(p+1) \\ &= -2\bigg[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\bigg] + 2(p+1) \\ &= n\log\big(2\pi\frac{RSS}{n}\big) + \frac{1}{RSS/n}RSS \\ &= n\log(2\pi) + n\log(RSS) - n\log(n) + 2(p+1)\end{aligned} \tag{10.40}$$

10.61 BIC

[—]

Chapter 11

Data Sets

Data Set	Description	Chapter	Original Source
advertising.csv	test	test	test
airbnb_basic.csv	test	test	test
airbnb_details.csv	test	test	test
beer.csv	test	test	test
evals_mod.csv	test	test	test
fivethirtyeight-recent-grads.R	test	test	test
framingham.csv	test	test	test
gss2016.csv	test	test	test
KingCountyHouses.csv	test	test	test
ncbreweries.csv	test	test	test
recent-grads.csv	test	test	test
sesame.csv	test	test	test
sis.csv	test	test	test
spotify.csv	test	test	test
test_songs.csv	test	test	test

Chapter 12

References