Lab #5, Part #2 – Regression with MLB Attendance and Performance

In this lab, we'll be seeking to try and understand the relationship between team performance and average attendance in North American professional team sports leagues. The data will include information on team performance and attendance for the NHL, NBA and MLB. As this is the 10th lab assignment of the semester, you'll be left on your own a bit more for code that we've gone over before. But most of the code I leave out here will look nearly identical to your past labs (summaries of data, histograms, t-tests, and so on). The hope is that at this point in the semester, you have built up some solid coding skills as you finish up the course.

The data for this lab come from Rodney Fort's Sports Business Data website, a common website for various types of sports league data (link: https://sites.google.com/site/rodswebpages/codes), and I've combined multiple files together for a clean file for this lab. You can download the clean file on Cavnas called `attdat.csv`. I've provided a brief key for this data set at the end of this lab document.

Please be sure to include all code and all visuals and tables in your submission on Canvas. Note that all responses to questions should be in complete sentences and provide details and explanations when asked.

Activity #1: Open R-Studio and Load in Data.
Go ahead and download the files (`attdat.csv`) from the Canvas Lab #5 page and save them to a new lab folder in your class folder directory (Lab 5 would be a good choice for your new folder) as .csv files. Then open up R-Studio, set your working directory with `setwd()`, and load the data. Name the data `attdat`.

```
#set working directory and load up the data.
setwd("c:/Users/…")
attdat <- read.csv(file = "attdat.csv", h = T)
head(attdat)
```

Be sure that you've told R-Studio that the data has variable names (headers) with `h = T`, and confirm that they show up correctly. You should see that you have 5 variables, 3 categorical and 2 numeric. The data should also have 91 rows (you can check this with the `nrow()` function).

Activity #2. Summarize Your Data.
Using the `aggregate()` function, summarize all variables in your data in a neat table by league. Report the mean, standard deviation for each numeric variable in the data for each league separately. I have the first line of code, calculating the mean for `Attendance`, below:

```
#summarize the data
aggregate(attdat$Attendance ~ attdat$League, FUN = mean)
```

2.1 Make a table of means and standard deviations by league. Be sure to make it nice and neat. What do you notice about NBA and NHL `Attendance` levels? How about when you compare them to MLB. Can you think of any reason why this might be happening? Finally, is the comparison of means of `WinPercent` useful in any way? Tell me what you see and explain why or why not.

Activity #3. Visualize Your Data.
Start here by making a histogram of each league's `Attendance` levels using `hist()`. Draw them side-by-side with the `par(mfrow = c())` function (3 across) and be sure to title them appropriately and label the axes.

3.1 Describe each of the distributions of `Attendance`. Do they data look normally distributed, skewed, or bimodal?

Now make a histogram of each league's `WinPercent` levels using `hist()`. Draw them side-by-side with the `par(mfrow = c())` function (3 across) and be sure to title them appropriately and label the axes.

3.2 Describe each of the distributions of `WinPercent`. Do they data look normally distributed, skewed, or bimodal?

Now create 3 side-by-side scatterplots of your data, one for each league. On the y-axis, place Attendance, and on the x-axis, WinPercent.

3.3 Include these scatterplots in your lab report and describe what you see for each one. Specifically, directly discuss **1)** the ***direction*** of the relationship (if any) you see for each league, **2)** the ***strength*** of the relationship you see (comparing across leagues), **3)** whether each relationship (if any) looks linear, **4)** whether the variance of `Attendance` looks equal across levels of `WinPercent`, and **5)** whether there look to be any major outlying observations.

Activity #4. Research Question and Hypotheses.
The question we want to think about for this lab is whether there is an association between `Attendance` and `WinPercent`. We'll treat the `Attendance` variable as the dependent variable and the `WinPercent` variable as the independent variable.

4.1 State the Null and Alternative hypotheses related to this question. Do so with mathematical notation ***and*** in words in the context of the data. You can do a single null/alternative hypothesis for all leagues, but note that we will also test each league separately with a least squares regression estimation.

4.2 Are the variables assigned randomly? Given the answer to that question, does our research question imply a causal relationship between winning and the number of fans that come to games? Why?

Activity #5. Correlation of Attendance and Win Percent.
The function `cor()` can give us the correlation between two variables. Go ahead and use it to calculate the correlation between Attendance and WinPercent in the data overall, and then separately for each of the leagues. I'll get you started below:

```
#correlation of attendance and winning across all leagues
cor(attdat$Attendance, attdat$WinPercent)
```

5.1 Report each correlation. Which league has the strongest relationship? The weakest? Does this confirm your visual description from **3.3**?

Activity #6. Regression Estimation for Each League.
Let's start by looking at the data for MLB. We'll estimate a linear model through least squares regression with the `lm()` function ("lm" for "linear model"). We'll tell R-Studio that we want to know the level of `Attendance` as a function of `WinPercent`. Remember "as a function of" often necessitates the ~ symbol. We'll use it here just like we do in the `aggregate()` function. We'll also make use of `summary()`, as just using `lm()` doesn't give us the output we want. We'll assign `lm()` as an object with a unique name, and then use `summary()` on that object to get our coefficients, standard errors, and p-values. Here is the code for MLB below (named `modMLB`):

```
#fit model for Major League Baseball
modMLB <- lm(Attendance ~ WinPercent, data = subset(attdat, League
     == "MLB"))
summary(modMLB)
```

6.1 Report the results of the MLB model in a neat table in your lab report and answer the following: **1)** Do you reject or fail to reject the null hypothesis? Give an answer in the context of the data and strength of evidence, **2)** given what you saw in your plots and our common understanding of professional sports, respond as to whether you are surprised at the results, and **3)** For each regression, report what the $\hat{\beta}_1$ estimated coefficient means in words and in the context of the scales of data.

6.2 Take a close look at the intercept for each league. Do you find them meaningful? Explain the meaning of the intercept in this context. Further, if you don't find an intercept very meaningful, explain what the problem is in this context.

Activity #7. Regression Estimation for All Leagues.
Since there are relatively small samples (numbers of teams in a league) in a single season, it might be useful to pool all the leagues together and estimate a regression on all the data.

7.1 Start with a summary table of means and standard deviations of `Attendance` and `WinPercent` across the entire data set. Include this in your lab report in a neat and tidy table.

7.2 Create a histogram of `Attendance` and `WinPercent` for the data as a whole. Include this in your lab report.

7.3 Crate a scatterplot of Attendance by Win Percent for the entire data set combined. Include this in your lab report and directly discuss **1)** the ***direction*** of the relationship (if any), **2)** the ***strength*** of the relationship you see, **3)** whether the relationship (if any) looks linear, **4)** whether the variance of `Attendance` looks equal across levels of `WinPercent`, and **5)** whether there look to be any major outlying observations.

7.4 Estimate a regression for the entire data set together. Report the results of this regression just like in Activity #6. Do you reject or fail to reject the null hypothesis from Activity #4? Put this in words and in the context of the data, and describe what the estimate of the $\hat{\beta}_1$ coefficient means in the context of the scales here.

7.5 If the results seem different from those in Activity #6, explain what you think is going on.

Activity #8. Regression with a Two-Category Variable.
In the lectures, we noted that with a two-category variable, we can use least squares regression to fit a linear model. For this, we'll estimate a regression testing the difference in Attendance between MLB and the NBA and NHL. For our purposes, we'll combine the NBA and NHL, making this effectively a test between Summer and Winter sports attendance. Use the code below to create a new variable called `baseball`. This is a 0/1 variable, and equal to 1 when the league is MLB.

```
#create dichotomous baseball variable
attdat$baseball <- ifelse(attdat$League == "MLB", 1, 0)
```

8.1 state the null and alternative hypotheses for this regression using mathematical notation and in words in the context of the data.

8.2 Use this variable as the independent variable in a regression for `Attendance`. Report the results in a nice table in your lab report and tell me whether you reject or fail to reject your null hypothesis. Do so in words in the context of the data and interpret the meaning of the $\hat{\beta}_1$ coefficient.

Activity #9. Repeat Activity #8 with `t.test()`.
Test the differences in Summer and Winter sports using the `t.test()` function. In this case, be sure to tell R-Studio that you are assuming that there are equal variances across groups.

9.1 Report the results of your t-test and compare it to **8.2**. Do you arrive at similar results?

9.2 Notice that this function reports the mean in group 0 and the mean in group 1 at the end of its output. Discuss how these two values relate to the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ from **8.2**.

Data Variable Key

`League` = Sports League Team Plays in

- "`MLB`" = Major League Baseball
- "`NBA`" = National Basketball Association
- "`NHL`" = National Hockey League

`Location` = Home Location of Team

`TeamName` = Name of Team

`WinPercent` = Percentage of Games Won in the 2019 (2018-19) Season (**Note:** ties in NHL counted as a non-win)

`Attendance` = Average Game Attendance for the 2019 (2018-19) Season