**Lab Assignment 10**

**I. Objectives**: To learn the key components of simple regression analysis, particularly the interpretation of regression coefficients, and to merge data sets.

**II. Datasets**: “Cruise.csv” and “cpi1983.csv”

**III. Packages**: none

**IV. Preparation**

1) Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.

2) Clear any data in memory by typing rm(list=ls())

3) Download datasets “Cruise.csv” and “cpi1983.csv” and place them in your working directory

4) Download R script “POLS 6480 Lab 10.R” and place it in your working directory.

5) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.

**V. Instructions for Lab 10**

The dataset you will use examines the effects of movie reviews on box-office receipts. It focuses on the career of actor Tom Cruise, who has had lead roles in 33 movies (plus supporting roles in 6 other movies) over the past 35 years. The variable that you will explain is domestic box office receipts. The key explanatory variable that you will use is Rotten Tomatoes’ “freshness” rating which summarizes how positively reviewers perceive the movie. The dataset also includes the year in which the movie was released, the genre, the MPAA rating, and Tom Cruise’s character and role in the movie.

Data on revenues is from ***http://www.the-numbers.com/person/540401-Tom-Cruise***

Data on freshness is from ***http://www.rottentomatoes.com/celebrity/tom\_cruise/***

For some background: ***http://fivethirtyeight.com/datalab/the-four-types-of-tom-cruise-movies/***. [[1]](#footnote-1)

1. Load the dataset, typing the following line, changing the directory if needed:

> data <- read.csv("C:/Cruise.csv")

> leading <- data[data$Role == "Lead", ]

The second line creates a new data frame that is a subset in which Tom Cruise had a leading role.

Let us begin by examining the MPAA ratings of Tom Cruise’s movies. To create a bar plot, R needs numerical data, but Rating is a factor variable (type factor(leading$Rating)to see its values, and type is.factor(leading$Rating) to ask whether the variable is a factor). You will need to run line 3 to create a table of data based on the values of the variable, and then run line 4 to plot the bar graph, which shows that most of his movies are PG-13 or R rated.

To warm up, recall that we used ANOVA in Lab 9, which told us whether knowing the value of an explanatory variable helps us account for some of the variance of a response variable. Line 6 shows the box plot revenues by rating; there is a great deal of overlap between the box office revenues of Tom Cruise’s R and PG-13 rated movies. Lines 7–8 run the ANOVA model and display the results. Does rating appear to have a statistically significant effect on box office?

2. Our main interest is on the effects of reviewers’ ratings of movies. To see a histogram of the Rotten Tomatoes’ freshness ratings for Tom Cruise’s movies, you can type:

> hist(leading$Freshness)

Then to see the association between US box office revenues and freshness ratings, you can type:

> plot(leading$Domestic ~ leading$Freshness, pch = 19)

It would be reasonable to expect that higher ratings would translate into higher box-office revenues; does this relationship appear to hold in the data?

In the R script, these commands are shown in lines 10–11.

3. In the last few labs, difference-of-means *t*-tests and ANOVA F-tests were used to assess whether an explanatory variable had statistically significant effect on a response variable. Recall that the numerator of the *t-*test was the difference between two groups’ means; the squared value of this difference would be the numerator of a one-way ANOVA F-test for two groups.

In this lab, the main explanatory variable is not binary or categorical, so we cannot form groups in any obvious way. Instead, when we have a continuous explanatory variable, we typically estimate a linear relationship between two variables. A line has two elements: the intercept tells us the value that the response variable is predicted to take when the explanatory variable equals zero; the slope tells us how the response variable changes when we increase the value of the explanatory variable by one unit.

To find the best-fitting line representing the relationship between freshness ratings and revenues, you will use the lm command. It would be wise to type ?lm in the Console window and examine the help file. When using the lm command, you need to name the model, then specify the response variable, explanatory variables, and location of the data. For example, you could type:

> model <- lm(dataset$yvar ~ dataset$xvar + dataset$control)

> model <- lm(yvar ~ xvar + control, data = dataset)

These two lines will estimate the same model; you can also eliminate any reference to dataset by typing attach(dataset) before running the models (but then you might also need to run detach(dataset) later.)

Line 13 estimates the simple regression model in which Domestic is the response variable and Freshness is the explanatory variable, and then Line 14 asks R to display just the coefficients from the model. Remember that every line is defined by its intercept and slope.

The intercept, \_\_\_\_\_\_\_\_ , tells us that the expected domestic box office if Freshness was equal to 0 is roughly \_\_\_\_\_ million dollars.

The Freshness coefficient, \_\_\_\_\_\_\_\_ , tells us that for each point that Freshness increases, we would expect domestic box office to increase by roughly \_\_\_\_\_ million dollars.

If you run line 15, you can add the regression line to the scatterplot; the upward slope of the line represents that increasing Freshness leads to increased expected domestic box office revenues.

There is a great deal more that you can learn from estimating the linear regression. If you run line 17, R will tell you the names of all the items that it stores as the model summary and which could be displayed if you typed summary(reg.simple). For instance, you can see the F statistic and d.f. by typing: summary(reg.simple)[10] or by typing anova(reg.simple).

4. If you type summary(reg.simple)[10], as shown in line 18, then R will show you the coefficients (both intercept and slope), their standard errors, their *t-*statistics, and *p*-values. Recall from lab 8 that the *t*-statistic for a two-sample difference-of-means test equaled the difference between the samples’ means divided by the standard error of the difference of sample means; when conducting a *t*-test for a regression coefficient, *t* equals the estimated coefficient divided by the standard error of the coefficient. What is the *t* statistic for Freshness? \_\_\_\_\_\_ Is *p* less than .05, indicating a statistically significant effect at the 95% confidence level?

5. Recall that we can also perform significance tests by comparing the *t* statistic to a *t* critical value, for a given degrees of freedom. In a linear regression analysis, the degrees of freedom typically will equal the number of observations minus the number of explanatory variables minus 1 (the last –1 is for the intercept). With 33 observations and just one explanatory variable, you ought to have \_\_\_\_\_ degrees of freedom.

You can double-check degrees of freedom using R. If you type names(reg.simple), you will find a list of items that R stores when it estimates the regression model. The items are numbered. The eighth item in that list is the degrees of freedom, so if you type reg.simple[8], as is shown in line 20, then R will show you the model degrees of freedom.

The *t* critical value for the can be found using the code in line 21: \_\_\_\_\_. Notice that the second entry inside the parentheses is the degrees of freedom. Is the calculated value for *t* greater than the critical value?

To summarize, the effect of increasing the Freshness rating of a Tom Cruise movie is to increase the expected domestic box office by \_\_\_\_\_ million dollars, and this relationship \_\_\_\_\_ (is/is not) statistically significant at the 95% confidence level.

6. If one takes a hard look at the list of Tom Cruise’s movies, one can see an important pattern of his recent movies yielding higher revenues. Of the ten movies Cruise has made in the last ten years (between 2005 and 2015), six exceeded $100 million. Of the eleven movies that Cruise made in his first ten years as a leading man (between 1983 and 1992), only two – *Top Gun* and *Rain Man* – broke the $100 million mark. Indeed, If you run line 24, you will see that domestic box office revenues are positively correlated to the year in which the movie was released.

There are two possible explanations: inflation and skill. The inflation explanation is that admission to movies costs a great deal more now than it did in the 1980s, and so looking at box office revenues is misleading. The skill explanation is either that Cruise is a more skilled actor now, after three decades of acting, and so his movies make more money irrespective of inflation, or that Cruise is more skilled at identifying projects that will become blockbusters.

If you run line 23, it will show you the correlation between the freshness rating and the year; the positive correlation would seem to indicate a positive association, so his more recent movies are more highly rated.

When confronted with such a situation, a frequently employed strategy is to control for the troublesome variable by including it in the analysis. Lines 26–27 estimate the multiple regression model using R’s built-in lm command, and then show the coefficients from that model. You interpret the slope coefficient for Freshness as before: a one-unit increase in Freshness translates into a \_\_\_\_\_ million dollar increase in box office revenue, after controlling for the Year.

7. Another approach is to take inflation into account by modifying the revenue variable. If you run line 30, R will load a second data frame which lists the Consumer Price Index per year (1982 = 100). Running line 31 will merge the two data frames, using the variable Year to match movies and CPI values; the code in the R script creates a new data frame named combined.

Running line 32 will create a new variable, constant, in the data frame. This variable divides the domestic box office revenues by the CPI value, so that revenues are in constant values, controlling for inflation. Actually, there is an extra term in the denominator, multiplying CPI (by 10,000) so that the variable constant is coded in *millions* of dollars. R will show you the average constant-dollar revenue earned by Tom Cruise’s movies, which is roughly $64.6 million, if you type mean(combined$constant).

If you run line 33, R will show you the correlation between constant-dollar revenues (in millions of dollars) and the year in which the movie was released; this correlation is now very close to zero.

8. Since we no longer need to worry about inflation, line 35 will run the simple regression model with Freshness as the only explanatory variable. This model is called reg.new. Since the dependent variable was rescaled, the coefficient now indicates by how many *millions* of dollars the domestic box office revenue changes in response to a one-unit increase in the Freshness rating. You can see the coefficient by typing reg.new$coefficients; you can see the coefficients plus standard errors, *t*-statistics and *p*-values by typing summary(reg.new)[4]

If you wish to perform your null hypothesis significance test using the *t* critical values, then you can run line 38 to find the degrees of freedom, and then run line 39 to find the critical value at the 95% confidence level. Do we reject or retain the null hypothesis that movie ratings have no effect on constant-dollar box office revenues?

9. One of the advantages of linear regression is the ability to make predictions. Since the intercept and slope define a line, one can enter a value for the explanatory variable into the equation and generate a predicted value of the response variable, given that prediction.

If you type mean(combined$Freshness), then R will show you the average Rotten Tomatoes Freshness rating earned by Tom Cruise’s movies, which is roughly 66. So, suppose we wished to predict domestic box office revenues for three different levels of Freshness ratings. Line 41 in the R script provides code using the predict() command. The first entry is the name of the regression model that you are using to generate the prediction. The second entry is a list of values for Freshness; I chose 33, 66, and 100; the prediction for a Freshness rating of 0 is, naturally, the intercept. The code for making the predictions is:

> predict(reg.new, data.frame(Freshness = c(33,66,100)))

10. Finally, R is good at incorporating factor variables into the analysis. Suppose you were interested in how different MPAA ratings affected revenues. You can run line 43 to add the variable, Rating, which we used at the beginning of the lab. Note that even though the variable is a factor rather than numerical, R estimates what can be called an “intercept shift” for the PG-13 and R categories (relative to PG).

11. To clear the **Environment**, type rm(list=ls()) or click on the broom icon.

To clear the **Console** window, type Ctrl-*l*

1. Similar analyses of Will Ferrell, Adam Sandler, Anne Hathaway, and Sandra Bullock can be found at the ‘Hollywood Taxonomy’ page at fivethirtyeight.com. [↑](#footnote-ref-1)