Quantitative Exploratory Data Analysis

Learning Objectives

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables
- Use scatterplots to assess the relationship between two quantitative variables
- Find the correlation coefficient
- Find the estimated line of regression using summary statistics and R Linear Model Output
- Understand what the slope coefficient represents
- Understand what the coefficient of determination is

Movies Released in 2016

We will revisit the data set used last week collected on Movies released since 1916 to 2016. Here is a reminder of the variables collected on these movies.

- Year: Year the movie was released
- Budget: The amount of money (in US $ millions) budgeted for the production of the movie
- Revenue: The amount of money (in US $ millions) the movie made after release
- Duration: The length of the movie (in minutes)
- Content Rating: Rating of the movie (G, PG, PG-13, R, Not Rated)
- IMDb Score: User rating score from 1 to 10
- Genre: Category the movie falls into
- Movie Facebook Likes: Number of likes a movie receives on Facebook

Terminology Review

In today's activity we will review summary measures and plots for two quantitative variables. Some terms covered in this activity are

- Scatterplot
- Correlation
- Slope
- Line of Regression
- Coefficient of determination

To review these concepts see Chapter 3 in the textbook.

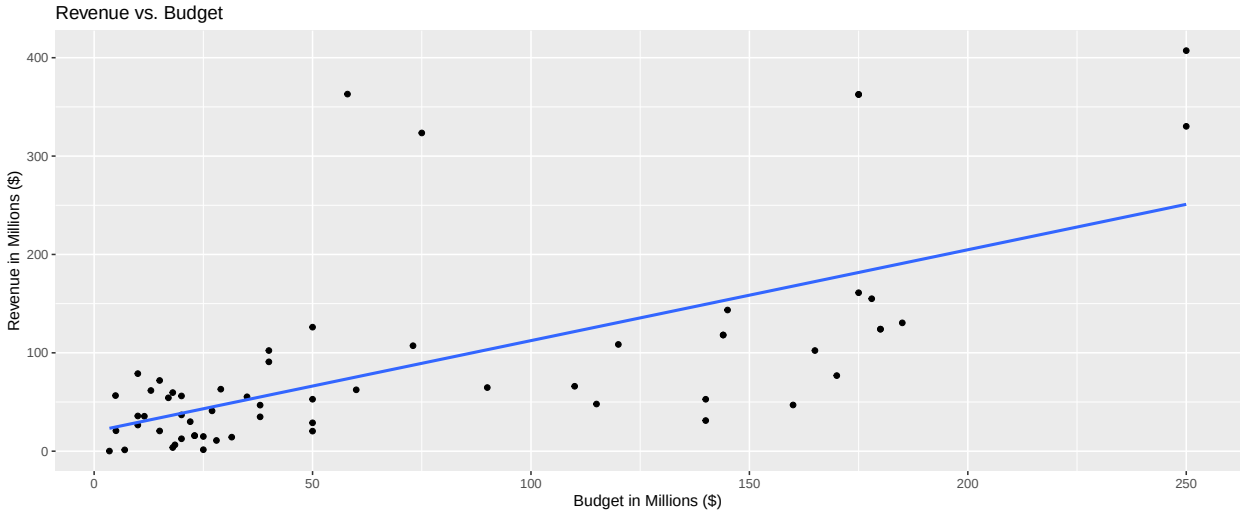
Vocabulary Review

1. What type of plot is used to display two quantitative variables?

2. What summary statistics are used to describe the relationship between two quantitative variables?

We will look at the relationship between 'Budget' and 'Revenue' for movies released in 2016. This shows a scatterplot of 'Budget' as a predictor of 'Revenue' (note: both variables are measures in "millions of dollars".

```
ggplot(data = moviesa,    #This is the data set
       aes(x = budget_mil, y = revenue_mil))+  #Specify variables
  geom_point() +  #Add scatterplot of points
  labs(x = "Budget in Millions ($)",  #Label x-axis
       y = "Revenue in Millions ($)",  #Label y-axis
       title = "Revenue vs. Budget") + #Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  #Add regression line
```



3. Assess the four features of the scatterplot that describe this relationship.

* Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual Observations or Outliers

4. Does there appear to be an association between 'Budget' and 'Revenue'? Explain.

Correlation

Correlation measures the strength and the direction between two quantitative variables. The closer the value of correlation to + or - 1 the stronger the linear relationship. Values close to zero indicate a very weak linear relationship between the two variables. The following output shows a correlation matrix between several pairs of quantitative variables.

```
moviesb <- moviesa[,c("budget_mil", "revenue_mil", "duration",
                      "imdb_score", "movie_facebook_likes")]
round(cor(moviesb,use="pairwise.complete.obs"),4)
```

```
##                      budget_mil revenue_mil duration imdb_score
## budget_mil               1.0000      0.6466   0.5274     0.3081
## revenue_mil              0.6466      1.0000   0.2516     0.4876
## duration                 0.5274      0.2516   1.0000     0.2362
## imdb_score               0.3081      0.4876   0.2362     1.0000
## movie_facebook_likes     0.6481      0.6710   0.5619     0.3462
##                      movie_facebook_likes
## budget_mil                         0.6481
## revenue_mil                        0.6710
## duration                           0.5619
## imdb_score                         0.3462
## movie_facebook_likes               1.0000
```

5. Using the output above, which two variables have the strongest correlation?

6. What is the value of correlation between 'Budget' and 'Revenue'?

7. Based on the value of correlation what would the sign of the slope be? Positive or negative? Explain.

8. Does your answer to question 13 match the direction you choose in question 3?

9. Explain why the correlation values on the diagonal are equal to 1.0.

Slope

The slope measures the change in y for each increase in x by 1. In other words, as the x variable increases by 1 unit, the y variable changes (increase/decreases) by the value of slope.

The linear model function in R gives us the summary for the least squares regression line. The estimate for (Intercept) is the y-intercept for the line of least squares and the estimate for budget is the value of b_1, the slope.

```
revenueLM <- lm(revenue_mil~budget_mil,data=moviesa)
summary(revenueLM)$coefficients
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 20.0362329 14.3458255 1.396659 1.677479e-01
## budget_mil   0.9236972  0.1418579 6.511426 1.806269e-08
```

10. Write out the least squares line using the summary statistics provided.

11. Interpret the value of slope in context of the problem.

12. Using the least squares line from Question 10, predict the revenue for a movie with a budget of 165 million.

Residuals:

The model we are using assumes the relationship between the two variables follows a straight line. The residuals are the errors, or the part that hasn't been modeled by the line.

$$\text{Data} = \text{Model} + \text{Residual}$$

$$\text{Residual} = \text{Data - Model}$$

$$e_i = y_i - \hat{y}_i$$

13. The movie, *Independence Day: Resurgence*, had a budget of 165 million and revenue of 102.315 million. Find the residual for this movie.

14. Did the line of regression overestimate or underestimate the revenue for this movie?

Coefficient of Determination

The coefficient of determination, R^2, can also be used to describe the strength of the linear relationship between two quantitative variables. R^2 describes the amount of variation in the response that is explained by the least squares line with the explanatory variable.

15. Calculate the coefficient of determination between 'Budget' and 'Revenue'.

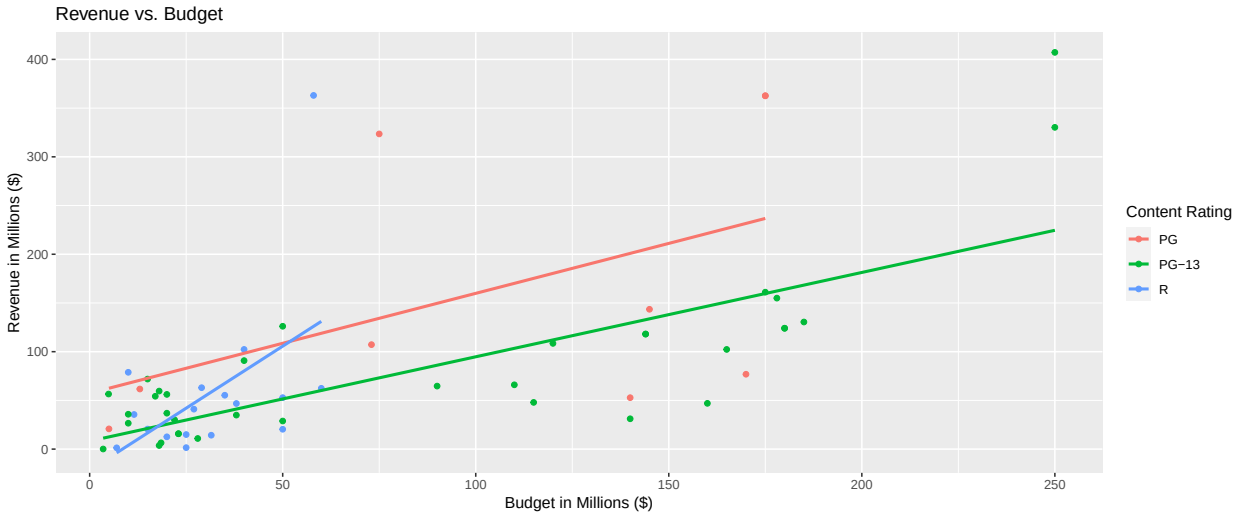
16. Interpret the coefficient of determination in context of the problem.

Multivariate Plot

In the next plot we are graphing three variables.

```
ggplot(data = moviesa,   #This is the data set
       aes(x = budget_mil, y = revenue_mil, color = content_rating))+  #Specify variables
```

```
  geom_point() +  #Add scatterplot of points
  labs(x = "Budget in Millions ($)",  #Label x-axis
       y = "Revenue in Millions ($)",  #Label y-axis
       color = "Content Rating",  #Label legend
       title = "Revenue vs. Budget") + #Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  #Add regression line
```



25. Identify the three varables plotted in this graph.

26. Does the relationship between 'Budget' and 'Revenue' differ among the different content ratings? Explain.