ACTIVITY 5

Movie Profits

5.1 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 (`lm`) output
- Interpret the slope coefficient in context of the problem
- Interpret the coefficient of determination in context of the problem

5.2 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
- Least-squares line of regression
- Coefficient of determination (r-squared)

To review these concepts, see Chapter 3 in the textbook.

5.3 Movies released in 2016

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

Variable	**Description**
`budget_mil`	Amount of money (in US $ millions) budgeted for the production of the movie
`revenue_mil`	Amount of money (in US $ millions) the movie made after release
`duration`	Length of the movie (in minutes)
`content_rating`	Rating of the movie (`G`, `PG`, `PG-13`, `R`, `Not Rated`)
`imdb_score`	IMDb user rating score from 1 to 10
`genres`	Categories the movie falls into (e.g., Action, Drama, etc.)
`movie_facebook_likes`	Number of likes a movie receives on Facebook

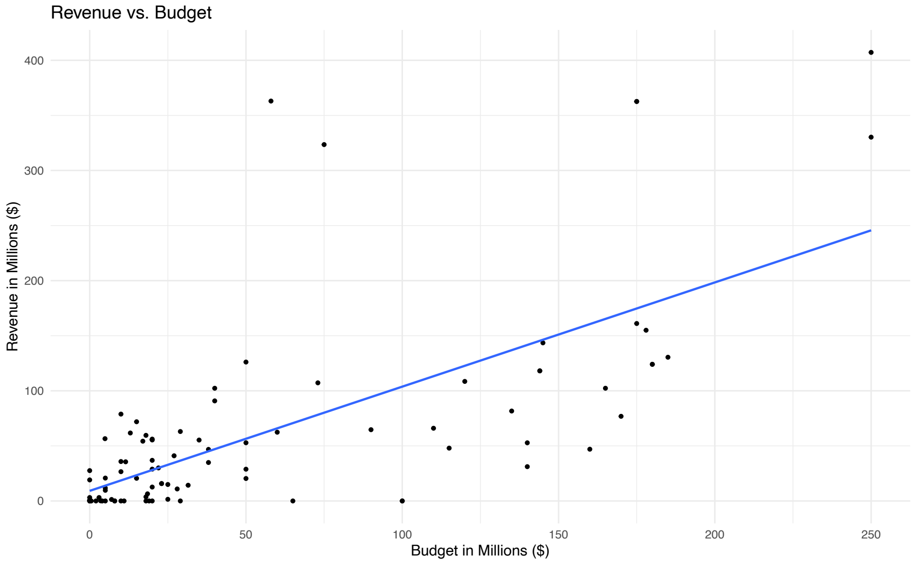
Vocabulary review

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").

```
movies <- read.csv("data/Movies2016.csv") #Reads in data set
movies %>% #Data set pipes into...
ggplot(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. Describe each feature using a complete sentence!

- 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.

```
# Take subset of variables
movies %>%  #Data set pipes into...
  select(c("budget_mil", "revenue_mil",  # Take subset of variables
           "duration", "imdb_score",
           "movie_facebook_likes")) %>%
  cor(use="pairwise.complete.obs") %>% # Calculate correlation matrix
  round(3)  # Round to 3 decimals
```

```
#>                      budget_mil revenue_mil duration imdb_score
#> budget_mil                1.000       0.686    0.463      0.292
#> revenue_mil               0.686       1.000    0.227      0.398
#> duration                  0.463       0.227    1.000      0.261
#> imdb_score                0.292       0.398    0.261      1.000
#> movie_facebook_likes      0.678       0.723    0.438      0.309
#>                      movie_facebook_likes
#> budget_mil                          0.678
#> revenue_mil                         0.723
#> duration                            0.438
#> imdb_score                          0.309
#> movie_facebook_likes                1.000
```

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 7 match the direction you choose in question 3?

9. Explain why the correlation values on the diagonal are equal to 1.

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_mil` is the value of b_1, the slope.

```r
# Fit linear model: y ~ x
revenueLM <- lm(revenue_mil ~ budget_mil, data=movies)
summary(revenueLM)$coefficients # Display coefficient summary
```

```
#>              Estimate Std. Error  t value     Pr(>|t|)
#> (Intercept) 9.1693054  9.0175499 1.016829 3.119606e-01
#> budget_mil  0.9460001  0.1056786 8.951670 4.339561e-14
```

You may remember from middle and high school that slope $= \frac{\text{rise}}{\text{run}}$.

Using b_1 to represent slope, we can write that as the fraction $\frac{b_1}{1}$.

Therefore, the slope predicts the how much the line will *rise* for each *run* of $+1$. In other words, as the x variable increases by 1 unit, the y variable is expected to change (increase/decrease) by the value of slope.

10. Write out the least squares line using the summary statistics provided in proper statistical notation.

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 (squared correlation)

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

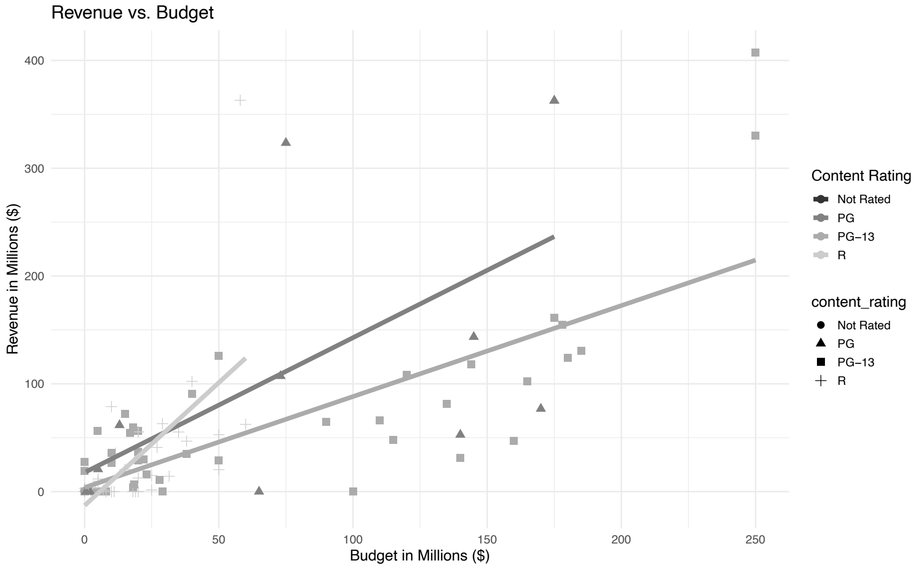
15. Use the correlation, r, to calculate the coefficient of determination between 'Budget' and 'Revenue', r^2.

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

Multivariate plots

What if we wanted to see if the relationship between 'Budget' and 'Revenue' differs if we add another variable into the picture? The following plot visualized three variables, creating a **multivariate** plot.

```
movies %>% #Data set pipes into...
ggplot(aes(x = budget_mil, y = revenue_mil, color = content_rating)) +  #Specify variables
  geom_point(aes(shape = content_rating), size = 3) +  #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, lwd = 2) + #Add regression lines
  scale_color_grey() #Make black and white
```



25. Identify the three varables plotted in this graph.

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

5.4 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity.