Chapter 7 and 8: Finishing Regression and Dealing with Outliers

Review

Recall from last class that the ____ quantifies the linear relationship between two qunatitative variables. We can calculate R^2 by the equation below:

Example *Suppose we are interested in quantifying the relationship between carat and sales price of diamonds. If we find that the regression between carat and diamond sale price has an R^2 of 0.89. How might we interpret this in context?*

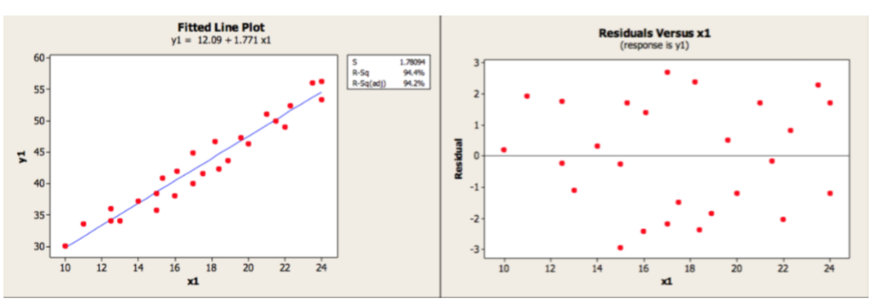
Comparing Residual Plots to R^2

Residual plots and R^2 tell us essentially two different things:

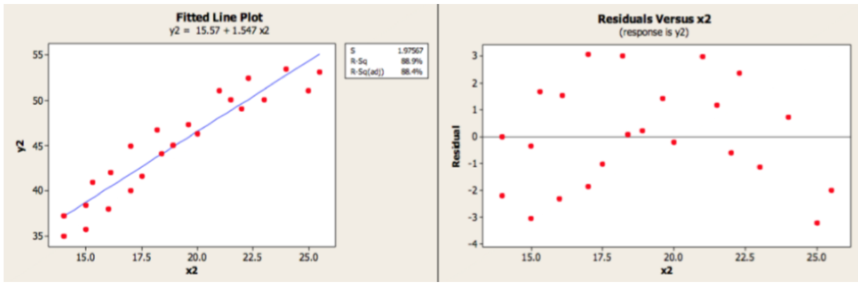
- **Scatterplot:**

- **R^2:**

Example *Below is an example of a dataset where the regression line appears to fit the data really well. This means we would likely see a high R^2. When we look at the scatterplot we see no distinct pattern in the residuals meaning that a linear model is indeed appropriate.*



Example *Once again, we see that the linear regression appears to fit the data well. So we might expect to see a high* R^2. *However, when we look at the residuals plot, we see a distinct pattern in the residuals suggesting that a linear model is not appropriate for this dataset.*



These two examples illustrate that:

-

-

Outliers in Regression

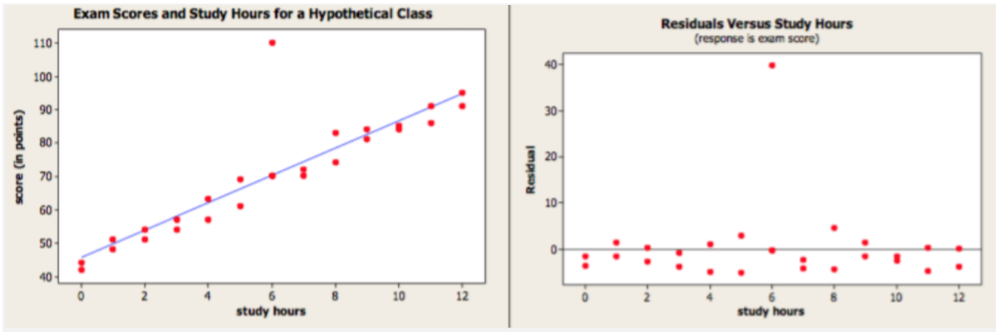
An ____ is an observation that falls outside the overall pattern of the data.

There are different types of outliers that whose names differ depending on the impact they have on the regression line.

-

-
  -
  -

A ________________________________ has a residual that is large compared to the other residuals. It typically is an outlier in the y direction and often DOES NOT affect the placement of regression lines but DOES affect additional analyses.

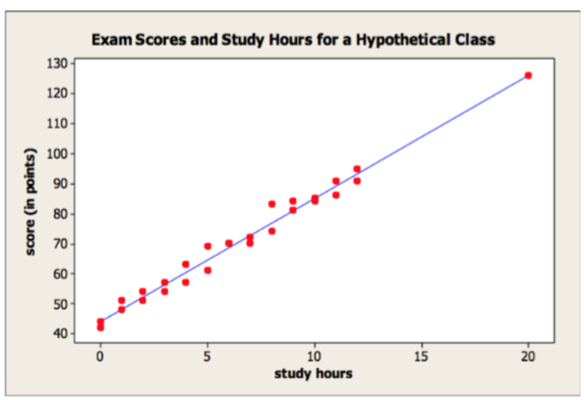


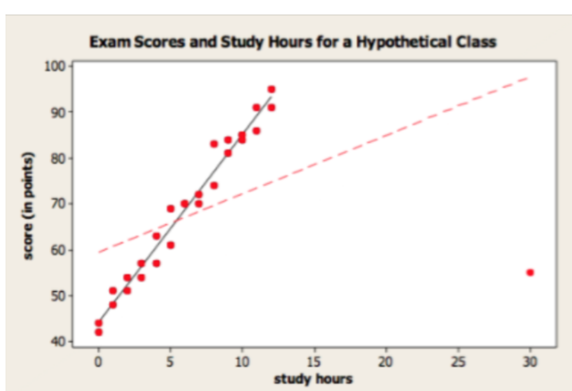
A ________________________________ is an outlier in the x-direction.

Two cases:

- **Influential Point:**

- **Non-Influential Point:**





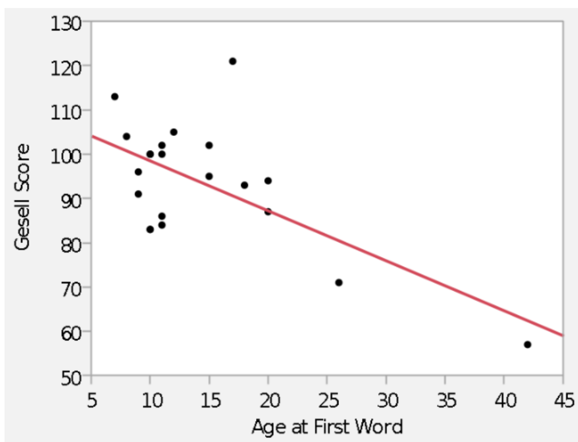
To deal with outliers in our regression data we can take the following steps:

-
-
-

Example *The Gesell test measures the language development of young children. A study recorded Gesell test scores on a sample of children, and the parents were asked at what age their children said their first word. Below is the regression equation and the R^2 value for this dataset.*

$$\hat{y} = 109.87 - 1.13 * x$$

with an $R^2 = 41\%$

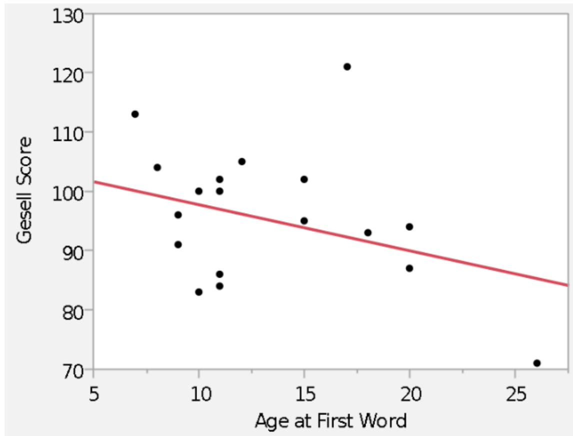


Interpretation of R^2:

Now after removing the outlier we obtain the following equation and R^2

$$\hat{y} = 105.63 - 0.78 * x$$

with an $R^2 = 11.22\%$



Interpretation of Slope:

Interpretation of Intercept:

Interpretation of R^2***:***

Summary of Linear Regression

Goal:

Fitting the Model: Obtain the regression line by calculating the slope and intercept using the following set of equations:

Understanding the Relationship:

-
-
-

Check Model Appropriateness:

-
-
-

Cautions with Regression:

- **Linear Relationships Only:**
  -
  -
  -
- **Extrapolation:**
- **Association/Correlation is not Causeation:**
  -
  -