SDS358: Applied Regression Analysis

Day 5: Simple Linear Regression

Dr. Michael J. Mahometa

Agenda for Today:

- Simple Regression
  - Intro & Similarity/Difference to r
  - The Linear equation and interpretation
  - Diagnostics

Today's Question:

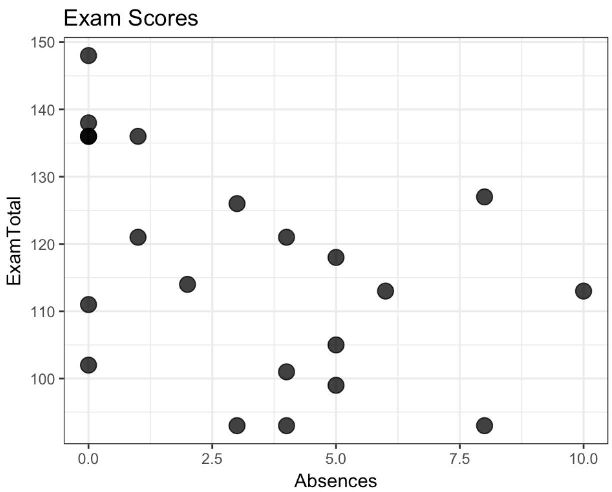
What is the Simple Linear Model predicting cost from carat weight in a sample of diamonds? *Are there any outliers in the data?*

Recap: Correlation

- We use correlation (Pearson Correlation) to:
- Determine the relationship between two *quantitiative* variables.
- Give that relationship definition.
  - Size of r
- However, only *linear* relationships are correctly captured by r
  - That's why we use the scatterplot

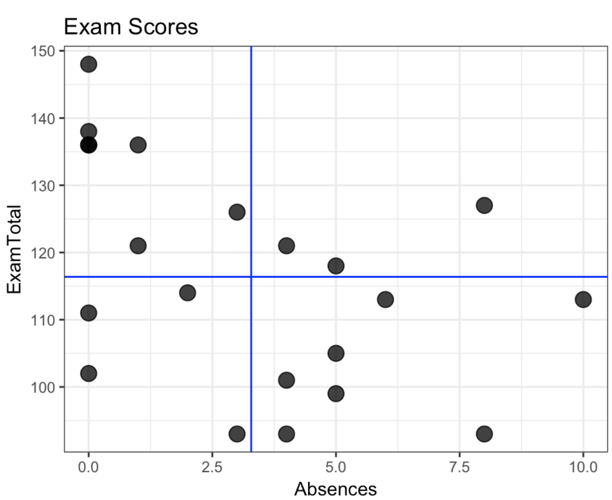
Since we know r…

What happens when we see this?



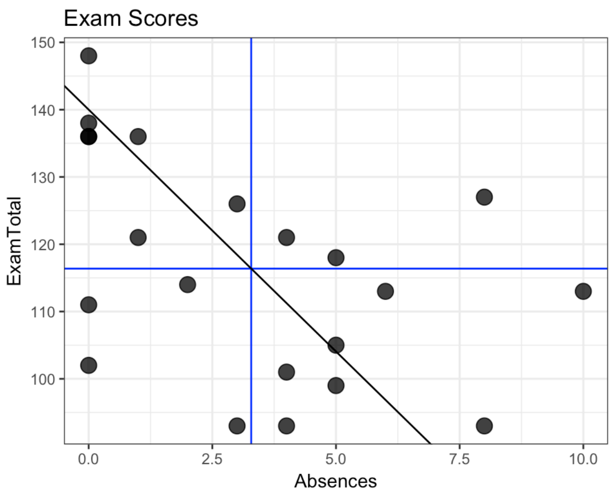
Since we know r…

What happens when we see this?



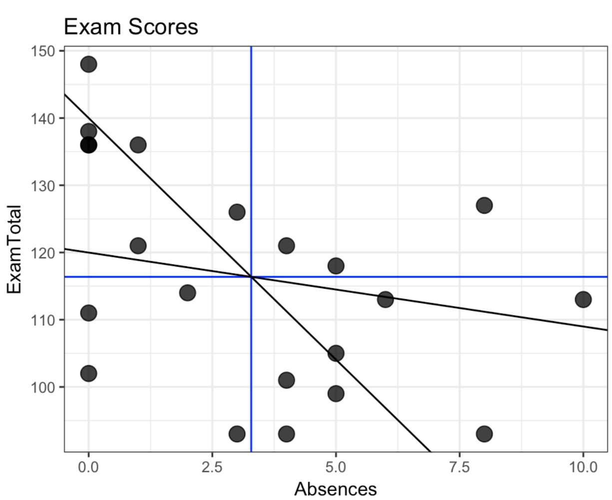
Since we know r…

What happens when we see this?



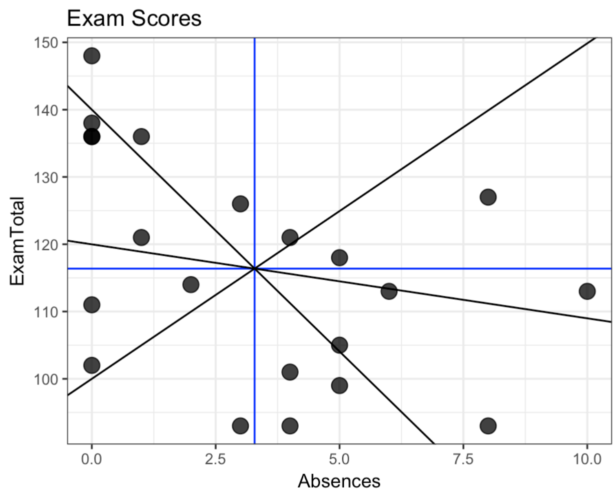
Since we know r…

What happens when we see this?



Since we know r…

What happens when we see this?

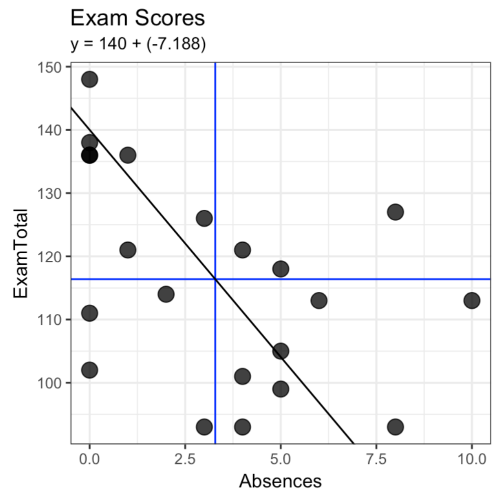


The Line of Best Fit

- "Find me the line that goes through the mean of x and the mean of y, that best fits the data."
- What's your definition of "fit?"
    - Residuals

A few things to notice:

- Each point does not fall directly on the line (any line).
- This "miss" is called the residual.
- Here's a *random line*



The Line

- We'll formally define a line as:

$$\hat{y} = b_0 + b_1 x_1$$

The Residual(s)

- We have two choices:
- Calculate **vertical** offset of each point to the line.
- Vertical offset provides a fit that *estimates y for a given x*

$$residual = e_i = (y - \hat{y})$$

- Calculate the **perpencicular** offset of each point to the line.

$$d_i = \frac{|y_i - (a + bx_i)|}{\sqrt{1 - b^2}}$$

The Line of Best fit

- The best fitting line is the one that is closest to the data points.
- As statisticians think:

Because the line can be used to predict a value of Y based on any X (with a corresponding Y), the best fitting line is a line that has the lowest amount of error from the predicted Y and the actual Y.

- In other words, the best line is the one that minimizes the residuals for all points (the *squared* residuals)

Lest see this in action!

R Script for residuals

This will make it a whole lot easier….

- Instead of doing this for every set of data (and we could)…
- We can simply turn to some basic algebra:

$$b_1 = r\left(\frac{S_y}{S_x}\right)$$

$$b_0 = \bar{y} - b_1\bar{x}$$

An Important Difference

- With r, we just get a relationship
  - It's symmetrical

$$COR_{(x,y)} = COR_{(y,x)}$$

- But symmetry fails with the linear equation of simple regression (and rightly so)
  - There are *two* possible lines:
    - x predicting y
    - y predicting x

Two Lines (and separate parameters)

- We can have R help us to find the "line of best fit" with the `lm()` function:
- So, we can *technically* have *two* lines:

```
lm(ExamTotal ~ Absences, data=school)
```

```
##
## Call:
## lm(formula = ExamTotal ~ Absences, data = school)
##
## Coefficients:
## (Intercept)     Absences
##     124.409       -2.443
```

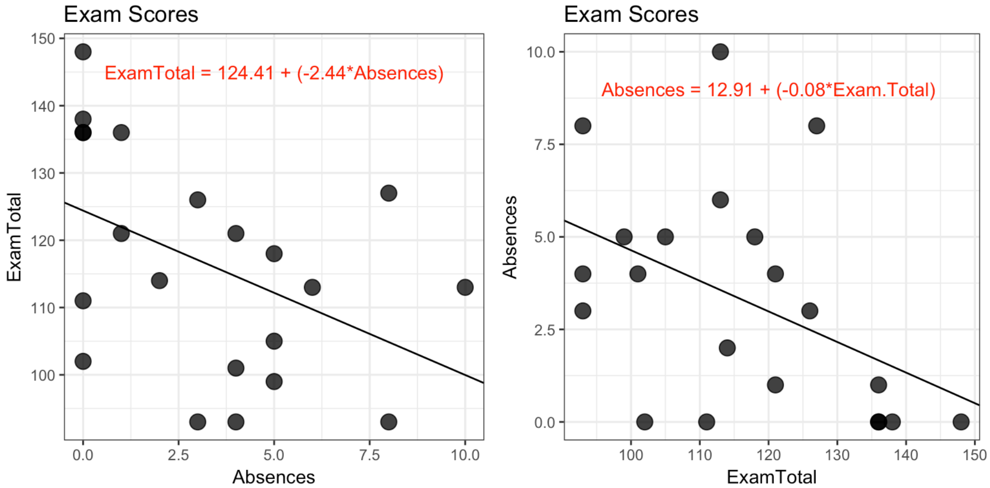
Two Lines (and separate parameters)

- We can have R help us to find the "line of best fit" with the `lm()` function:
- So, we can *technically* have *two* lines:

```
lm(Absences ~ ExamTotal, data=school)
```

```
##
## Call:
## lm(formula = Absences ~ ExamTotal, data = school)
##
## Coefficients:
## (Intercept)    ExamTotal
##     12.9102      -0.0827
```

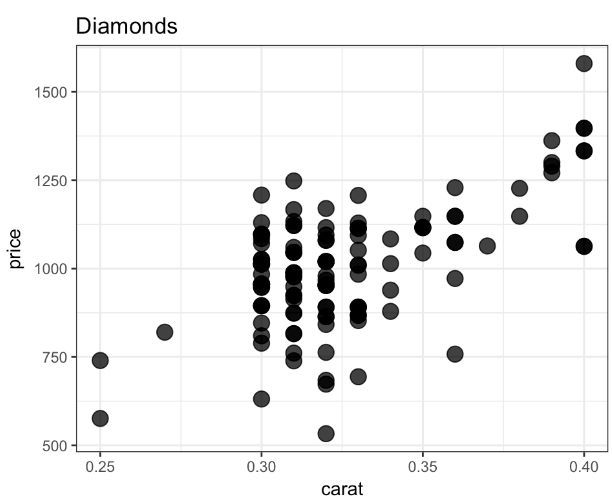

Two Lines (and separate parameters)



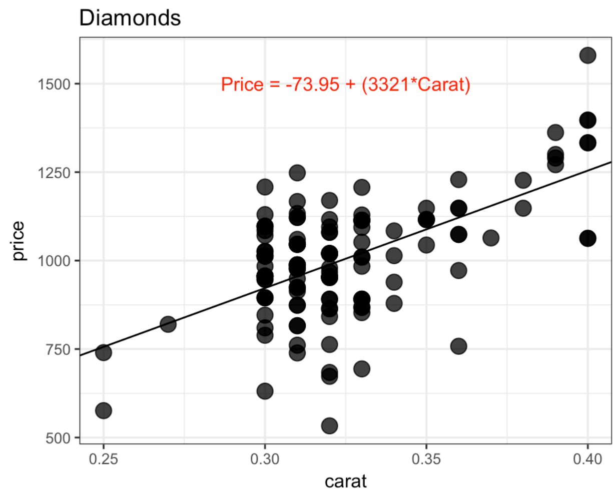
Interpretation

- The intercept and the slope have *very* specific interpretations.
  - Intercept: Value of y at an x of **zero**.
  - Slope: How much the predicted value of y changes given a **single** unit change in x.
- Let's try some new "enlightening" data

Diamonds



Diamonds



Interpretation:
- Slope
- Intercept

Assumptions of Simple Linear Regression

- Quantitative Data for both variables.
- The relationship is linear in nature.
- Residuals
  - Independence: Error associated with each data point is independent of every other value.
  - For a given value of x, e has a normal distribution, meaning:
    - The population mean of e is 0.
    - For a given value of x, the population variance of e is σ_e^2
  - Homoscedasticity
  - No Outliers

Summary:

- How Pearson r and Simple Linear Regression are the same and different.
- Definition of residuals
- The true meaning of the line of best fit.
- Interpretation for the Intercept and the Slope