Chapter 7 and 8: Introduction to Linear Regression Part Two

Review

The goal of linear regression is to find a straight line that best represents the relationship between two quantitative variables.

To quantify this relationship we take a response variable y and an explanatory variable x, yielding the following equation:

where $\hat{y}$ is the predicted value of y. Recall that we can compute the ________________________ ____ using the following formulas.

b_o : b_1 :

Introduction to Prediction

If there is a linear relationship between two quantitative variables, we can use regression to make predictions.

Prediction: For some value of the explanatory variable, x', that is of interest. We can plug in x' into the regression line to predict the response variable:

$$\hat{y} = b_o + b_1 * x'$$

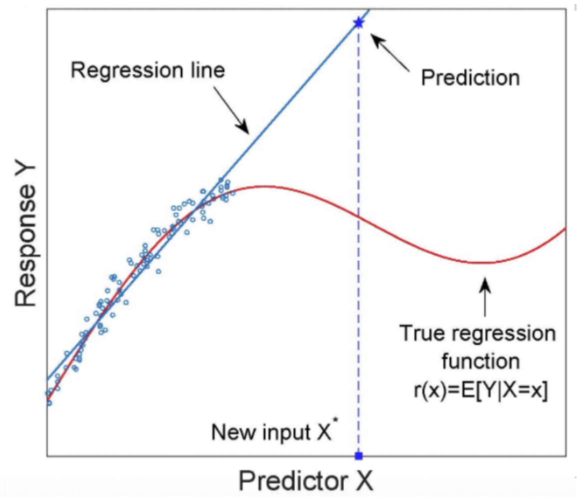
Example *Suppose we are interested in studying corn yields(in hecters) and fertilizer(in pounds) amounts. After some intense research we find that the yield of corn can be characterized with the following linear regression:*

$$\widehat{Yield} = 4.32 + 1.05 * fertilizer$$

Calculate the predicted yield when we apply 5 pounds of fertilizer to our field of corn.

Caution: Extrapolation

It is generally not recommended to make predictions far outside the range of the explanatory variable. This is known as ____.



For each of the following situations, determine whether each situations would be extrapolation.

1. Predicting boiling point of water at a high altitude, using regression created from data at sea level.

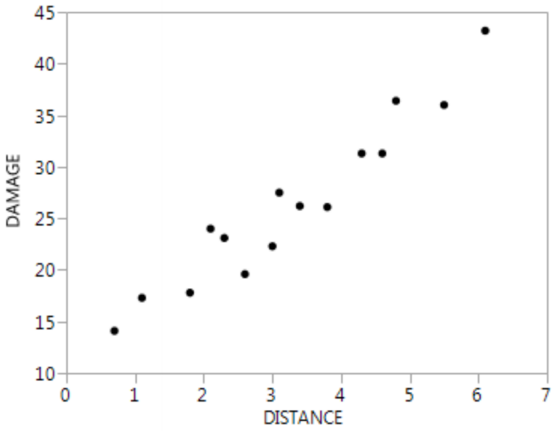
2. Predicting test scores for middle schoolers at one school when the regression was created from data at a different school.

3. Predicting video game sales from data collected the year previous.

4. Predicting Usain Bolt's mile time from data collected from middle schooler mile run times.

Let's return to our fire damage example. Recall that the relationship between amount of damage in thousands of dollars and the distance from a fire station in miles is characterized by the following regression equation:

$$\widehat{Damage} = 10.31 + 4.91 * Distance$$



Calculate the predicted damage for each situation and determine whether there is extrapolation.

- Distance of 5.5 miles

- Distance of 1.2 miles

- Distance of 15 miles

- Distance of 3 miles

Residuals

The ______________________________ is the difference between the observed value and predicted value of the response variable. The residual value tells us how inaccurate our model's prediction is at that particular point.

Notation: e

The formula for calculating residuals for a particular observation (x_i, y_i)

$$e = y_i - \hat{y}$$

where $\hat{y} = b_o + b_1 * x_1$ (the linear regression line evaluate at x_1)

On a scatterplot, a residual is the ____.

-
-

Example *Once again let's look at our fire damage dataset. Compute the residual for each situation.*

$$\widehat{Damage} = 10.31 + 4.91 * Distance$$

- *A house that is 2.1 miles from the fire station experienced $24,000 in damages.*

- *A house that is 5.5 miles from the fire station experienced $36,000 in damages.*

What does least squares estimate mean? Let's illustrate by drawing a picture.

Another thing we can do with residuals is to compute their standard deviation. We can compute the standard deviation of the residuals the same way we calculate the standard deviation of other data.

Notation: s_e = standard deviation of the residuals

Interpretation: s_e is a measure of the variation of the points around the regression line.

- large $s_e \implies$ lots of variation around the line
- small $s_e \implies$ little variation around the line