Chapter 8 - Linear Regression

In this Chapter we look at a Linear Model for data that has a linear relationship. This **Linear Model** is an equation of a line that goes through the data.

y is the observed or true value from the data.

$\hat{y}$ (y-hat) is the predicted or estimated value from the linear model.

Residual (p. 195): The vertical distance or difference between the observed value and the predicted value. Calculated by taking $y-\hat{y}$.

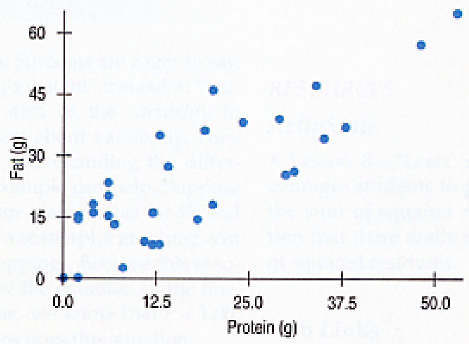
Least Squares Regression Line or Regression Line (p. 197): The line that minimizes the sum of the squared residuals.

We use the square of the residuals for the same reason we used the squares in the standard deviation. If they are summed up without the squares the positive and negative residuals will cancel each other out. Also the square magnifies the larger residuals.

Ours	Book
$\hat{y}=a+bx$	$\hat{y}=b_0+b_1x$
$b=r\frac{s_y}{s_x}$	$b_1=r\frac{s_y}{s_x}$
$a=\overline{y}-b\overline{x}$	$b_0=\overline{y}-b_1\overline{x}$

p. 216: 12 a)

We can use the Regression line to estimate what a y-value is for a given x-value by simply plugging x-value into the Regression Line.



Total *fat* versus *protein* for 30 items on the Burger King menu. The Double Whopper is in the upper right corner. It's extreme, but is it out of line? **Figure 8.1**

Use the given Regression Line to predict the fat content for a sandwich at Burger King that contains 20 grams of protein.

$$\widehat{fat} = 6.8 + 0.97\,protein$$

Use the given Regression Line to predict the fat content for a sandwich at Burger King that contains 80 grams of protein.

We want to avoid **extrapolation**, which is attempting to use the regression line to make predictions for x values outside the range of data.   The Linear Model may not be correct outside the range of data.

Regression Line: $\hat{y} = a + bx$ where $b = r\frac{s_y}{s_x}$ and $a = \overline{y} - b\overline{x}$

Residual: The vertical distance or difference between the observed value and the predicted value. Calculated by taking $y - \hat{y}$.

Residual Plot (p.203): A scatterplot of the residuals versus the x-values. Residuals will be plotted on the y-axis. This plot makes the regression line the x-axis ($y = 0$). This graph magnifies the distance between the residual and the regression line. The residual plot should have no interesting features like shape, direction, or extreme values. The points should stretch horizontally with the same spread throughout the data.

p. 216, #13

Square of the correlation, r^2, is the fraction of variation in the values of y that can be explained by the regression line of y on x. r^2 measures the variation along the line as a fraction of the total variation.

p. 221, #45 and 47:

In #45 and 47 on p. 221, there are two sources of variation in y in the regression setting.

> The first source can be explained from the line. As the Age of the car goes up the Price of the car will decrease. This is the part the line explains.
>
> The Prices of the cars do not lie exactly on the line because they are scattered above and below the regression line. This is the second source, which the regression line tells nothing about.

Remember to check the Conditions

- "Quantitative Variable Condition" - variables must be quantitative.
- "Straight Enough Condition" - association must be linear.
- "Outlier Condition" - Exclude outliers