Chapter 7 - Linear Regression

September 16, 2014

7.1 Least Squares: The Line of "Best Fit"

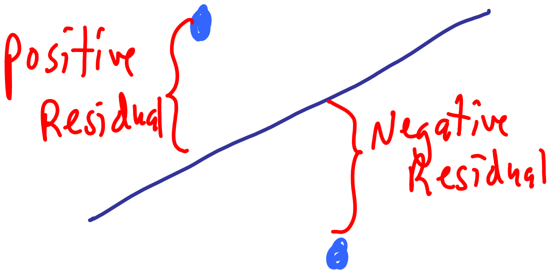
In this Chapter we look at a Linear Model for data that has a linear relationship. This **Linear Model** is an equation of a line that goes through the data.

In math models will be wrong since it cannot match exactly the real world.

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. 179): 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: 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.

7.2 The Linear Model & 7.3 Finding the Least Squares Line

	Ours	Book
Regression Line	$\hat{y} = a + bx$	$\hat{y} = b_0 + b_1 x$
Slope	$b = r\dfrac{s_y}{s_x}$	$b_1 = r\dfrac{s_y}{s_x}$
y-intercept	$a = \bar{y} - b\bar{x}$	$b_0 = \bar{y} - b_1\bar{x}$

$\hat{y} = a + bx$ $\quad b = r\frac{s_y}{s_x}$ $\quad a = \bar{y} - b\bar{x}$

p. 203: 26 a)

$\bar{x} = 30$, $s_x = 4$, $\bar{y} = 18$, $s_y = 6$, $r = -.2$

$b = r\frac{s_y}{s_x} = -.2\frac{6}{4} = -.3$ ← slope

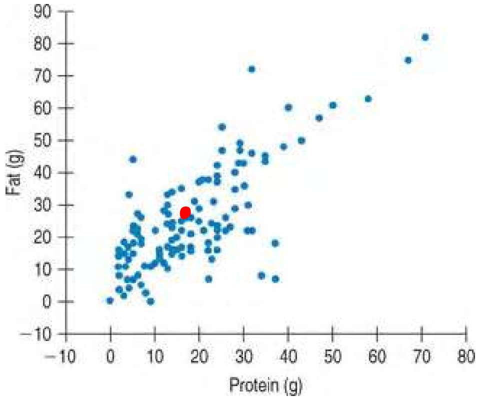
$a = \bar{y} - b\bar{x} = 18 - (-.3)30 = 27$ ← y-int.

$\hat{y} = a + bx \Rightarrow \hat{y} = 27 + (-.3)x$ or $\hat{y} = 27 - .3x$

y-int: $\hat{y} = 27$ when $x = 0$

Slope: For every unit increase in x
there is a .3 unit decrease in y.

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.



p.178

Figure 7.1 Total *Fat* versus *Protein* for 122 items on the BK menu. The Triple Whopper (71 grams of protein and 82 grams of fat) is in the upper right corner. It's extreme, but is it out of line?

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

$\widehat{Fat} = 8.4 + 0.91\,Protein$ $\qquad \hat{y} = 8.4 + .91x$

Protein = 20 $\Rightarrow \widehat{fat} = 8.4 + .91 \times 20 = 26.6$ g of fat

y-int: If protein is 0g then $\widehat{fat}$ is 8.4 g.

Slope: For every 1g increase in protein the fat will increase .91g.

Use the given Regression Line to predict the fat content for a sandwich at Burger King that contains 100 grams of protein.

protein = 100 g $\widehat{fat} = 8.4 + .91 \times 100 = 99.4$ grams of fat.

Data only goes from 0 grams to 71 grams of protein.

we do not know if regression model will fit value outside data range.

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.

On Calculator

$\hat{y} = 86.73 + 5.01x$ $\widehat{yield} = 86.73 + 5.01$ Bushels

y-int: When no Rain ($x=0$) the yield is 86.73 bushels.

Slope: For every additional day of rain the yield increases 5.01 bushels.

$x = 25 \Rightarrow \hat{y} = 211.98$ bushels

$x = 35 \Rightarrow \hat{y} = 262.08$ bushels

7.5 Examining the Residuals

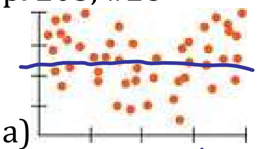
Regression Line: $\hat{y} = a + bx$ where $b = r\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$

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

Residual Plot (p. 188): 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.

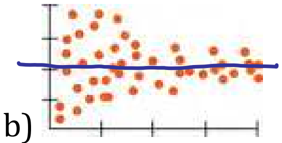
p. 203, #28

a)



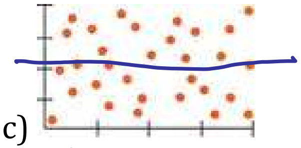
Linear model not appropriate because of the curve.

b)



Linear model not appropriate since data is closer for far x-values

c)



Linear model is appropriate.

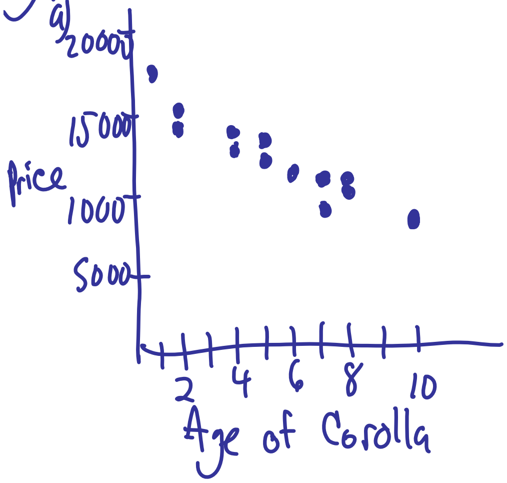
7.6 R^2 – The Variation Accounted For by the Model

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. $1 - R^2$ is the fraction of the original variation left in the residuals or that cannot be explained by the linear model. R^2 is always between 0% and 100%.

p. 208, #59 and 61:

59

a)



b) Strong, Negative, Linear association

c) Linear model is appropriate

d) $R^2 = 89.1\%$ $r = -\sqrt{.891} = -.944$

e) Age accounts for 89.1% of the variation in price as determined by the linear model.

f) 10.9% of price comes from other factors like mileage, condition, location, History, color, options

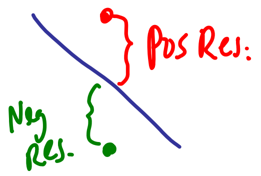
61 a) $\hat{y} = 17766.98 - 862{,}05x$ or $\hat{y} = 17767 - 862x$

b) For every additional year of age the price decreases \$862.

c) The average price of a New corolla is \$17,767 (Age = 0)

d) Age: $\hat{y} = 17767 - 862(7) = \$11{,}733$

e)



Negative Residual for buyer would be better because of the lower price.

f) Asking price of 10 yr old Corolla is \$8500.

Age = 10 $\hat{y} = 17767 - 862(10) = \$9147$

Residual $= y - \hat{y} = 8500 - 9147 = -647$

g) Extrapolation is used since data goes from 1 to 10 yrs. Linear Model will not be appropriate at 25 years.

In #59 and 61 on p. 208, 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.

7.7 Regression Assumptions and Conditions

Remember to check the Conditions

- “Quantitative Variable Condition” - variables must be quantitative.
- “Straight Enough Condition” - association must be linear.
- “Outlier Condition” - Exclude outliers
- “Does the Plot Thicken? Condition” – Does the data spread seem to be consistent for all values of *x*?