Math 316

Section 12.3 The Fitted Regression Line

$slope = \frac{rise}{run} = \frac{change\ in\ y}{change\ in\ x}$, so $change\ in\ y = slope * change\ in\ x.$

For *perfectly* linear data $(x_1, y_1), \ldots, (x_n, y_n)$ where $y_i = mx_i + b$, it turns out that the slope is $m = \frac{rise}{run} = \frac{s_y}{s_x}$ where s_x and s_y are the standard deviations of the x and y values in the sample data. If the data are not perfectly linear (which is normally the case), then the least squares (a.k.a. linear regression) line that best fits the sample data is

$$Y = b_0 + b_1 X$$

where slope

$$b_1 = r\frac{s_y}{s_x}$$

and

$$b_0 = \bar{y} - b_1\bar{x} \quad (\text{from } b_0 = Y - b_1 X).$$

The SD line (in general, not very useful) has slope $\frac{s_y}{s_x}$. See **Figures 12.3.2** and **12.3.3** on page 526. Don't stress a lot if the book's discussion about the SD vs. linear regression line isn't completely clear. It's most important to understand what we're actually going to be using (the linear regression line) rather than what we are not (the SD line). The solid line (the linear regression line) in each figure better fits the data within each of the shaded regions for the X value. Each triangle shows the mean y-value in each region.

Consider **Example/Figure 12.3.5** on page 529. Rather than predict a specific y-value for a given x-value, it is more appropriate to predict an average Y-value for a given X-value:

$$\mu_{Y|X} = b_0 + b_1 X$$

So in **Example 12.3.6** on page 530 (Example 12.3.5 continued) with regression line (based on the sample) $Y = 99.3 - 9.01X$, so at $X = 2.5$ we have

$$\mu_{Y|X=2.5} = 99.3 - 9.01(2.5) = 76.78$$

So based on this sample, we estimate that the *average* amount of food consumption after an amphetamine dose of 2.5 mg/kg would be 76.78 mg/kg. (Notice the units.)

Speaking of units, what does the slope and its units mean? Recall that

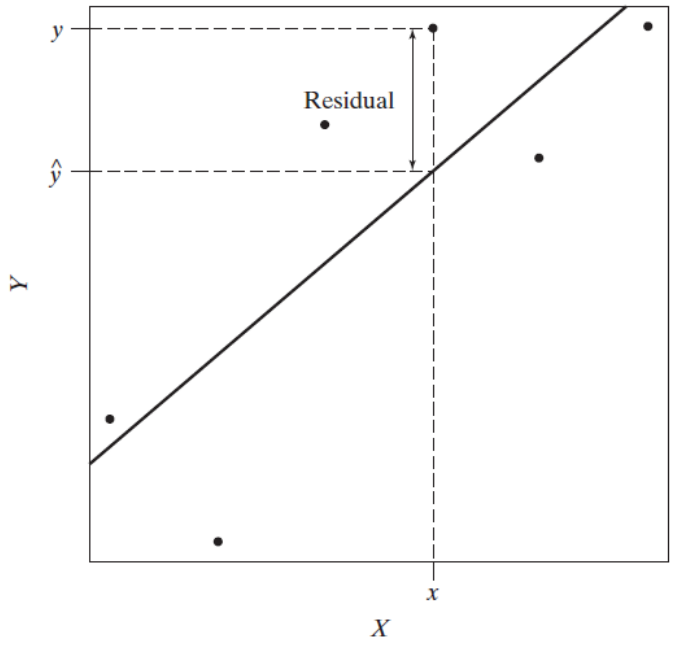
$$change\ in\ y = slope * change\ in\ x.$$

For **Example 12.3.5**, notice the units in the slope

$$-\frac{9.01\ g/kg\ \ food\ consumption}{mg/kg\ amphetamine} = -\frac{9.01\ g\ food\ consumption}{mg\ amphetamine}$$

Interpretation: for each increase of 1 milligram of amphetamine administered to the rat, there is a decrease of 9.01 grams in food consumption.

So we can find the least squares line that best fits the given sample data. So where does the "least squares" name come from? First, residual means "what is left over" or in this case "error." Given some data and the line that fits the data, for a given value of x, the residual or error is the difference between y (the actual value of the data for this x) and $\hat{y}$ (the value that the line would predict for us).



So at a particular x_i the residual is $e_i = y_i - \hat{y}_i$, and the total error between the line and all of the data points is

Residual Sum of Squares

$$\text{SS(resid)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Don't try to connect this to SS in ANOVA. SS simply means sum of squares.

Least-Squares Criterion

The "best" straight line is the one that minimizes the residual sum of squares.

So the best fit line simply is the line which minimizes the sum of these errors, a pretty simple idea. Next is the standard deviation of the residuals:

Residual Standard Deviation

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{\text{SS(resid)}}{n-2}}$$

This is quite similar to the formula for the standard deviation of sample values $y_1, \ldots, y_n$.

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Next, recall that for normally distributed data (as in Figure 4.3.5), approximately

68% of data are within 1 SD of the mean

95% of data are within 2 SD of the mean

99.7% of data are within 3 SD of the mean

Similarly, for least squares lines, approximately:

68% of data are within 1 SD of the line

95% of data are within 2 SD of the line

99.7% of data are within 3 SD of the line

See **Figure 12.3.8** on page 533.

Finally:

Fact 12.3.1: Approximate Relationship of r to s_e and s_y

The correlation coefficient r obeys the following approximate relationship:

$$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

This would be exact if in finding s_e we divided by $n-1$ (as we do in finding s_y) rather than $n-2$. There will always be variation in the Y values of the data since the Y values will change as the X values change. (Otherwise the data would be flat, which simply doesn't happen.) And then there will be some variation due to the line not perfectly fitting the data. s_Y^2 is the total amount of variation in the Y values in the data, and s_e^2 is how much variation there is due to the line and data not perfectly fitting each other. So $s_Y^2 - s_e^2$ is the *amount* of s_Y^2 which is simply due to the linearly changing nature of the data. So $r^2 \approx \frac{s_Y^2 - s_e^2}{s_Y^2}$ can be thoughts of as the *fraction* of the total variance s_Y^2 in the Y values due to the linear nature of the data. Two examples:

Figure	Page	r	r^2	% of variance in Y values due to linear relationship
–	–	1	1	100%
12.3.7	533	0.94	0.88	88%
12.3.8	533	−0.57	0.32	32%
–	–	0	0	0%

And look at **Figure 12.2.3** on page 516. r^2 is the coefficient of determination. Again, it tells us what proportion of the variance in the Y values is due to the (non-zero slope) linear relationship between X and Y.