Chapter 8: Linear Regression

Least-Squares Regression Line

The LSRL is a model used to represent a set of _____ data. Suppose you find the distance from each point in the data to the linear model, then square those distances and find the sum. This is called the _____. The Least-Squares Regression Line (LSRL) is the line that _____ this sum. The equation of the LSRL is $\hat{y} = b_0 + b_1 x$.

x represents _____.
$\hat{y}$ represents _____.
b_0 represents _____.
b_1 represents _____.

Given a set of data, you can calculate the LSRL (without using your calculator!). Knowing the correlation makes this task even easier. Use the following formulas:

$$b_1 = r\left(\frac{s_y}{s_x}\right) \qquad r = \frac{\sum z_x z_y}{n-1} \qquad s_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} \qquad b_0 = \bar{y} - b_1\bar{x}$$

Let's try one!... The following ordered pairs represent the scores for five former statistics students on the Unit I test and the semester exam, respectively: (76, 82), (78, 75), (84, 88), (65, 71), and (79, 85). Calculate the LSRL (without using your calculator) for predicting semester exam grades.

x	y	z_x	z_y	$z_x z_y$	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$

sum: _____ sum: _____ sum: _____

r = _____
s_x = _____
s_y = _____ define x: _____
b_1 = _____
b_0 = _____ define y: _____

LSRL: _____
(Now verify your results using the linear regression option on your calculator).

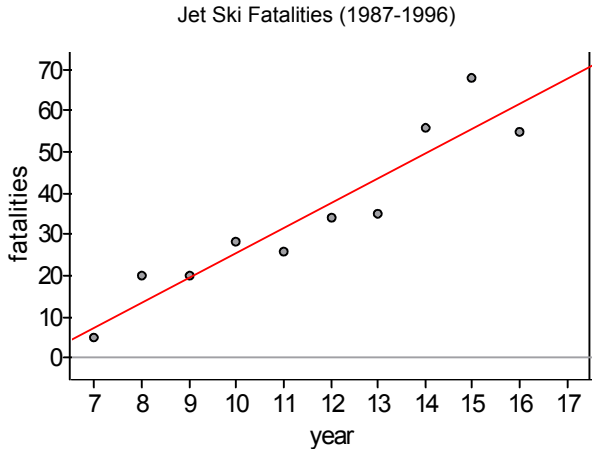
Coefficient of Determination

The coefficient of determination, also called R^2, is the square of the _____. The R^2 value tells how much of the variation in the response variable is accounted for by the linear regression model. For example, if $R^2 = 1$, then _____% of the variability in the response variable is accounted for by the linear model. In other words, the relationship between the two variables is perfectly linear. If $R^2 = 0.95$, we can conclude that _____ % of the variability in the response variable is accounted for by the linear relationship with the explanatory variable.

1. Given the following set of data, find the equations of the LSRL, then find and interpret both the correlation and the coefficient of determination.

Jet Ski Use

	year	fatalities
1	7	5
2	8	20
3	9	20
4	10	28
5	11	26
6	12	34
7	13	35
8	14	56
9	15	68
10	16	55



   a. LSRL: _____ (use meaningful variables in your equation rather than *x* and *y*, and use proper statistical notation!)

   b. Correlation (*r-value*): _____. A correlation of _____ indicates that there is a _____, _____, _____ relationship between _____ ___ and _____ _____.

   c. Coefficient of determination (R^2): _____. An R^2 value of _____ indicates that _____% of the _____ in _____ is accounted for by the _____ relationship with _____.

2. A study of class attendance and grades earned among first-year students at a state university showed that in general students who attended a higher percent of their classes earned higher grades. Class attendance explained 16% of the variation in grades among the students. What is the numerical correlation between percent of classes attended and grades earned? _____

Residual Plots

A residual is the difference between the observed *y*-value and the _____ y-value for a given *x*-value.

residual = $y - \hat{y}$

The _____ (SSR) is used to determine the Least-Squares Regression Line for a given set of data.

A _____ is a scatterplot which graphs the residuals on the _____ axis and the values of the explanatory variable on the _____ axis for each data point, $\left(x_i, y_i - \hat{y}_i\right)$.

The residual plot gives a visual representation of the amount of error in the model. The closer the residuals are to _____, the smaller the error and the more accurate the model.

The LSRL is a good model if the residual plot shows random _____ relatively close to the horizontal axis (zero). The horizontal axis represents the _____.

Points in the residual plot that lie directly on the horizontal axis lie directly on the _____.

Points in the residual plot that lie above the horizontal axis lie _____ the LSRL. Therefore, the model gives an underestimate at that point. Therefore _____ residuals represent underestimates.

Points in the residual plot that lie below the horizontal axis, lie _____ the LSRL. Therefore the model gives an overestimate at that point. Therefore _____ residuals represent overestimates.

The LSRL is not a good model if the residual plot shows _____.

3. Construct a well-labeled residual plot using the data on jet ski fatalities from #7. What can you conclude about the appropriateness of the linear model based on the residual plot?