Linear Regression Reference Sheet

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. Each individual in the data appears as a point.

A **response variable (y)** measures the outcome of a study.

An **explanatory variable (x)** helps explain or influences change in a response variable.

Direction: Rising is called a positive association. Falling is called a negative association.

Form: Is the pattern linear or does it follow another type of function?

Strength: How close does the data follow the given function/form? Are there outliers or striking deviations from the pattern? What are the r and r^2 values?

Correlation r

The correlation measures the strength and direction of the **linear** relationship between two quantitative variables:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\cdot\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Value of r	Strength of relationship
-1.0 to -0.5 or 1.0 to 0.5	Strong
-0.5 to -0.3 or 0.3 to 0.5	Moderate
-0.3 to -0.1 or 0.1 to 0.3	Weak
-0.1 to 0.1	None or very weak

Reversing the variables does NOT change r

r^2 Coefficient of Determination

The coefficient of determination gives us *the proportion of variation in the values of y that is explained by least-squares regression of y on* x. The coefficient of determination turns out to be r^2 (correlation coefficient squared).

AP Statement:

"_____% of the variation in _____ is accounted for by the regression line."
(r^2 ; response variable (y))

Or

"_____% of the variation in _____ can be attributed to the variation in _____."
(r^2 ; response variable (y) ; Explanatory variable (x))

To turn r^2 into r: square root and then match the sign of the slope (pos/neg)

Remember: Correlation does NOT imply causation. AKA Association does NOT imply causation.

Least-Squares Regression

Least Squares Regression A.K.A. linear regression allows you to fit a line to a scatter diagram in order to be able to predict what the value of one variable will be based on the value of another variable.

$$\hat{y} = a + bx$$

Slope

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

r = correlation
S_y = standard deviation of y
S_x = standard deviation of x

Y-Intercept

$$a = \bar{y} - b\bar{x}$$

b = slope
x-bar: mean of x's
y-bar: mean of y's

Describing slope:

_____ will change by _____ as _____ increases by 1.
Response variable ; *slope* ; *explanatory variable*

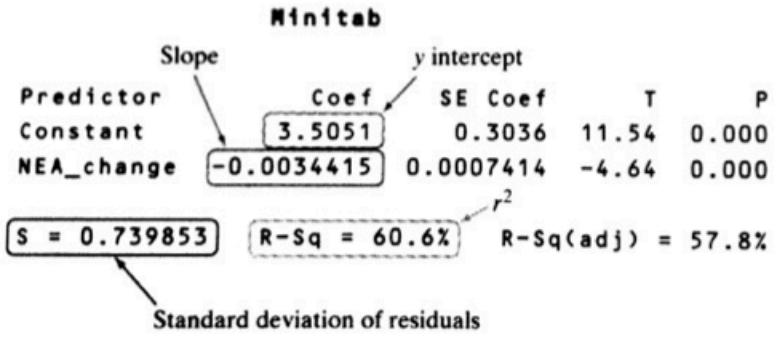
Describing y-intercept:

_____ will be _____ when _____ is 0.
Response variable ; *y-intercept* ; *explanatory variable*

Facts about least-squares regression

The distinction between explanatory and response variables is essential in regression. Least-squares regression looks at the distances of the data points from the line only in the *y* direction. If we reverse the roles of the two variables, we get different least-squares regression line (both slope and y-intercept change)

The LSRL always passes through the point $(\bar{x}, \bar{y})$



SEE = Standard Error of Estimate

- The estimate is y hat
- The error is the residual
- Standard means Standard Deviation
- So SSE is the standard deviation of the residuals.
- The Letter S = tells this on the computer printout (near R-sq)
- On the calculator: Run the AP program, select 4 Lin Reg. then run "residual scatterplot"

Residual: the vertical distance between the observed data values and a trend line. This is a measure of error. A positive residual means the trendline value is an under-estimate. A negative residual means the trendline value is an over-estimate.

Residual = y – y-hat or residual = observed – predicted

Residual Plots: A residual plot is a scatter diagram that plots the residuals on the y-axis and their corresponding x values on the x-axis. Positive residuals are graphed above the x-axis and negative residuals are graphed below the x-axis. We are looking for patterns, including: residuals form a parabola or another type of graph, residuals shrink or grow as the x-value increases.

How do you know the relationship is linear?

- The scatterplot is relatively straight
- There is NOT a pattern in the residual plot (random)
- The correlation and coefficient of determination are strong

Outlier: An observation that lies outside the overall pattern in the scatter plot (either in the x or y direction) Outliers affect the direction, form, and strength of the line.

Influential point: A point is influential if removing it would markedly change the position of the regression line. Points that are outliers in the x direction are often influential. Influential points often have small residuals because they tend to pull the line towards themselves. Therefore, you might miss influential points if you only look at residuals.

Lurking variable: A variable not shown in the regression equation that affects both x and y. Many times, there is no association between x and y, but it appears there is an association because they are both affected by the third variable (the lurking variable). This is called a nonsense correlation.

Extrapolation: using the regression line for x values beyond the data used to create the regression line. This is considered dangerous and is to be avoided.
Example the regression line was created for x values 0 to 60. Avoid using x values over 60.