Correlation

The scatterplot gives us a general idea about whether there is a linear relationship between two variables.

More precise: The coefficient of correlation (denoted *r*) is a numerical measure of the strength and direction of the linear relationship between two variables.

Formula for *r* (the correlation coefficient between two variables *X* and *Y*):

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

Most computer packages will also calculate the correlation coefficient.

Interpreting the correlation coefficient:

- **Positive *r* => The two variables are positively associated (large values of one variable correspond to large values of the other variable)**
- **Negative *r* => The two variables are negatively associated (large values of one variable correspond to small values of the other variable)**
- **$r = 0$ => No linear association between the two variables.**

Note: $-1 \leq r \leq 1$ always.

How far *r* is from 0 measures the *strength* of the linear relationship:

- ***r*** **nearly 1 => Strong positive relationship between the two variables**
- ***r*** **nearly -1 => Strong negative relationship between the two variables**
- ***r*** **near 0 => Weak relationship between the two variables**

Pictures:

Example (Drug/reaction time data):

Interpretation?

<u>Notes</u>: (1) Correlation makes no distinction between predictor and response variables.
(2) Variables must be numerical to calculate *r*.

Examples: What would we expect the correlation to be if our two variables were:
(1) Work Experience & Salary?

(2) Weight of a Car & Gas Mileage?

<u>Some Cautions</u>

Example:

Speed of a car (*X*)	20	30	40	50	60
Mileage in mpg (Y)	24	28	30	28	24

Scatterplot of these data:

Calculation will show that $r = 0$ for these data.

Are the two variables related?

<u>Another caution</u>: Correlation between two variables does not automatically imply that there is a cause-effect relationship between them.

<u>Note</u>: The population correlation coefficient between two variables is denoted ρ. To test H_0: $\rho = 0$, we simply use the equivalent test of H_0: $\beta_1 = 0$ in the SLR model. If this null hypothesis is rejected, we conclude there is a significant correlation between the two variables.

The square of the correlation coefficient is called the coefficient of determination, r^2.

<u>Interpretation</u>: r^2 represents the proportion of sample variability in Y that is explained by its linear relationship with X.

$$r^2 = 1 - \frac{SSE}{SS_{yy}}$$ **(r^2 always between 0 and 1)**

For the drug/reaction time example, r^2 =

Interpretation:

Estimation and Prediction with the Regression Model

Major goals in using the regression model:
(1) Determining the linear relationship between Y and X (accomplished through inferences about β_1)

(2) Estimating the mean value of Y, denoted $E(Y)$, for a particular value of X.
Example: Among all people with drug amount 3.5%, what is the estimated mean reaction time?

(3) Predicting the value of Y for a particular value of X.
Example: For a "new" individual having drug amount 3.5%, what is the predicted reaction time?

- **The point estimate for these last two quantities is the same; it is:**

Example:

- **However, the variability associated with these point estimates is very different.**

- **Which quantity has more variability, a single Y-value or the mean of many Y-values?**

This is seen in the following formulas:

$100(1-\alpha)\%$ <u>Confidence Interval</u> for the mean value of Y at $X = x_p$:

where $t_{\alpha/2}$ based on $n-2$ d.f.

$100(1-\alpha)\%$ <u>Prediction Interval</u> for the an individual new value of Y at $X = x_p$:

where $t_{\alpha/2}$ based on $n-2$ d.f.

The extra "1" inside the square root shows the prediction interval is wider than the CI, although they have the same center.

Note: A "Prediction Interval" attempts to contain a random quantity, while a confidence interval attempts to contain a (fixed) parameter value.

The variability in our estimate of E(*Y*) reflects the fact that we are merely estimating the unknown β_0 and β_1.

The variability in our prediction of the new *Y* includes that variability, <u>plus</u> the natural variation in the Y-values.

Example (drug/reaction time data):
95% CI for E(*Y*) with $X = 3.5$:

95% PI for a new *Y* having $X = 3.5$: