BIVARIATE REGRESSION ANALYSIS

A. In performing t-tests, one compares two groups on an attribute (e.g., Pepsi vs. Coke drinkers on number of cavities). Let us consider the following variable:

\[ X = \begin{cases} 
1 & \text{if a Coke drinker} \\
2 & \text{if a Pepsi drinker} 
\end{cases} \]

We can place this variable along the horizontal axis below:

\[ Y = \# \text{ cavities in 5 years} \]

Now we can sketch distributions of a dependent variable (e.g., number of cavities in five years) along vertical axes for each type of soft drink. To illustrate, let us assume that \( \mu_2 > \mu_1 \) (i.e., that Pepsi drinkers have more cavities than Coke drinkers). Thus if the data were like this, we might conclude that "taking the Pepsi challenge" (i.e., drinking Pepsi instead of Coke) yields a net increase of 2 cavities per year. We know this because there is an average increase of 10 (25-15) cavities per 5 years and (10/5)-2.

B. Now let's assume that we have done a chemical analysis of Coke and Pepsi...
and find that Coke has 2 mg. sugar whereas Pepsi has 6 mg. sugar (per 16 oz. can). With this new information we can examine the relation between cavities and sugar in sodas:

Instead of 1 = Coke and 2 = Pepsi on the X-axis, we have 2 mg. sugar for Coke drinkers and 6 mg. sugar for Pepsi drinkers. Once this is done we can phrase the sugar-by-cavities association more generally as $2\frac{1}{2}$ cavities in 5 years per additional milligram of sugar (per 16 oz. can). Here's the math:

$$\frac{(25-15)}{(6-2)} = \frac{\text{increase in } Y}{\text{increase in } X}$$

C. Note that we can also use the data to predict the number of cavities per five years for someone who drinks a soda with 4 mg. of sugar per 16 oz. can (i.e., more sugar than Coke has, but less than Pepsi does). To figure this out you must start with a baseline value like the 15 cavities in 5 years that Coke drinkers have on average. We know (from above) that we would expect an additional $2\frac{1}{2}$ cavities for each additional milligram sugar. Thus if the soda has 2 more milligrams than Coke does, one would predict 5 more cavities in 5 years than the average for Coke drinkers. More math:

$$15 + (2\frac{1}{2} \times [4 - 2]) = 20$$

D. Now let's predict the number of cavities per year for someone who drinks a sugar-free brand of soda. Substituting zero for 4 in the above equation, the associated prediction equals 10:

$$15 + (2\frac{1}{2} \times [0 - 2]) = 10$$

1. Given a sugar-free soda's predicted value for the number of cavities ($Y$) per 5 years, one can construct a PREDICTION EQUATION for the predicted value of $Y$ given $X$, where $X$ is $Y$'s corresponding number of milligrams
sugar content per 16 oz. of soda. This equation takes the following form:

\[ \text{predicted } Y\text{-value} = (a \text{ constant}) + 2iX \]

2. Notice that X=0 for any sugar-free soda, leaving the soda's predicted value of Y equal to the constant. More generally, put the constant in the equation equals the predicted value of Y when X equals zero. The resulting \textit{regression equation} is as follows:

\[ \hat{Y} = 10 + 2iX \]

E. \textbf{SOME COMMENTS ABOUT THIS EQUATION:}

1. We can plot this equation as follows:

\begin{align*}
\text{Y axis} = \text{Dependent variable} \\
\text{X axis} = \text{Independent variable} \\
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
10 & 12.5 & 15 & 17.5 & 20 & 22.5 & 25 & 27.5 & 30
\end{align*}

(You might try plotting these X and Y coordinates within the graph on page 127 of these Lecture Notes.)

2. The equation is of the form, \( Y = a + bX \).

a. "a" is called the y-intercept, because it is the point at which the regression line "intercepts" the Y-axis (i.e., when X=0).

b. "b" is called the SLOPE. The larger "b" is, the "steeper" the slope and thus the more units of Y are predicted by a unit-shift in X.

\text{NOTICE} that the "magnitude" of the slope depends upon the \textit{UNITs} you are talking about: We can just as easily consider \( \hat{Y}' = \# \text{ cavities per year, which would render the regression equation as } \hat{Y}' = 10 + iX \). \text{Clearly both equations predict equally well. In this}
regard, consider the following two points:

1) Since a line with ZERO slope suggests no association, it appears that larger slopes imply larger associations (if units are consistent).

2) If the variables had "standard units" (as would standardized variables), then slopes' magnitudes would reflect the magnitudes of associations.

3. The slope of $2\beta$ is POSITIVE. If a slope is positive, the regression line rises to the right; if it is NEGATIVE, the regression line falls to the right. (For example, one might find a negative association between cities' instances of leukemia and the number of miles they are located from toxic waste dump sites.) Thus it is with the sign of the slope that you can express the direction of a linear statistical relationship.

4. There are "floor" and "ceiling" effects.

"Sure your patients have 50% fewer cavities. That's because they have 50% fewer teeth!"
a. It makes no sense to speak of "minus" cavities or "minus" mg. sugar.

b. Sooner or later you have no teeth left to get cavities in.

5. The prediction equation is a formula for a straight line. If the relation between $X + Y$ is not "linear," you might try a curved line. (Later in the course we shall be fitting curved lines to data.) NOTE how a curved line might better take into account the ceiling effects just mentioned:

\[ Y = \# \text{ of cavities in 5 years} \]

\[ X = \text{mg. sugar consumption per day} \]

F. That was easy, right? Just find the means, draw a line, and write down an equation. Well, not quite. Note, for example, that the mean numbers of cavities for drinkers of these and other types of soda are not likely to "line up" into a straight line! Let's add a few other soft drinks into the analysis:

Dr. Pepper (21 mg./16 oz.)
Diet Dr. Pepper (7 mg./16 oz.)
Tab (1 mg./16 oz.)

etc.

And if we were to add "switchers," who drink diet sodas when they're feeling guilty and Pepsi when they are not?
FINALLY, we could work up a measure of sugar consumption and forget sodas altogether. If we did this the scatterplot would probably look like this:

\[ Y = \# \text{ cavities in 5 years} \]

\[ X = \text{mg. sugar consumption per day} \]

G. The problem now is to decide what is the best criterion for choosing one of the many possible prediction lines that could be drawn through these data.

One criterion selected by statisticians is called the Ordinary Least Squares (or OLS) criterion of selecting a regression line. This criterion specifies that you CHOOSE THAT LINE WHICH MINIMIZES THE SQUARED DEVIATIONS OF YOUR OBSERVATIONS FROM THE LINE.

That is, you want

\[ \min \left[ \sum (Y_i - \hat{Y}_i)^2 \right] \]  

where \( \hat{Y}_i = \hat{a} + \hat{b}X_i \).

RECALL (from p. 9 of these Lecture Notes) that \( \bar{X} \) is the number that minimizes \( \sum (X_i - \bar{X})^2 \). What is being minimized here is similar, except in that we are now minimizing the squared deviations from a LINE (i.e., from a "mean" with a value that depends upon the value of \( X \)). Those of you who have had calculus can verify quickly that this criterion requires that two conditions are met:

1. Such a verification is made as follows:
Abducted by an alien circus company, Professor Doyle is forced to write calculus equations in center ring.

a. First let \( \hat{Y}_i = \hat{a} + \hat{b} X_i' \), where \( X_i' = (X_i - \bar{X}) \),

b. then find \( \frac{\partial}{\partial \hat{a}} \Sigma (Y_i - \hat{Y}_i)^2 \) and \( \frac{\partial}{\partial \hat{b}} \Sigma (Y_i - \hat{Y}_i)^2 \).

c. These two derivatives will give you respectively the first and second criteria required by an OLS regression line.

2. The first criterion is that \( \Sigma (Y_i - \hat{Y}_i) = 0 \). (RECALL [from p. 9 of these Lecture Notes] the parallel equivalence that \( \Sigma (X_i - \bar{X}) = 0 \).)

a. Thus \( 0 = \Sigma (Y_i - [\hat{a} + \hat{b}X_i]) \)

\[ = \Sigma Y_i - n\hat{a} - \hat{b} \cdot \Sigma X_i \]

\[ = \frac{\Sigma Y_i}{n} - \hat{a} - \hat{b} \cdot \frac{\Sigma X_i}{n} \]

\[ = \bar{Y} - \hat{a} - \hat{b}\bar{X} \]

OR \( \hat{a} = \bar{Y} - \hat{b}\bar{X} \) \( \Leftarrow \) which is the computation formula for \( \hat{a} \) !!!

133
b. Notice also that $Y = \hat{a} + \hat{b}X$ implies that the point, $(\bar{X}, \bar{Y})$, must fall on the regression line. This is a good check to make when plotting a regression line.

c. IMPLICATIONS of the first OLS requirement:

1) Say we are comparing income and education and we have the following two pairs of observations:

(12-years education, $15,000$) and (16-years education, $25,000$)

For the first person in the sample, we can choose a predicted value (viz., $\hat{Y}_1$) and then determine what $\hat{Y}_2$ would be under the constraint, $\Sigma (Y_i - \hat{Y}_1) = 0$.

Since at this point we can take any value, let's arbitrarily take the value of $\hat{Y}_1 = 5,000$. According to the constraint,

$(Y_1 - 5,000) + (Y_2 - \hat{Y}_2) = 0$

OR $(15,000 - 5,000) + (25,000 - \hat{Y}_2) = 0$.

Thus $\hat{Y}_2 = 35,000 \leftarrow$ a BAD prediction!

We can GRAPH the observations and predictions as follows:

![Graph showing linear regression](image)
Choosing another prediction for $Y_1$, say $20,000$, and evoking the constraint gives a different result:

$$(Y_1 - 20,000) + (Y_2 - \hat{Y}_2) = 0.$$  

OR $$(15,000 - 20,000) + (25,000 - \hat{Y}_2) = 0.$$  

Solving for $\hat{Y}_2$, we get $\hat{Y}_2 = 20,000$.

NOTICE that if $\bar{Y}$ is the predictor for any value of $X$ (other than $\bar{X}$), then it is the predictor for ALL $Y$s. In this example this suggests that no matter what one's education is, the best predictor of income is the overall mean, $20,000$. In other words, education is of no value in predicting income. This uninformative relationship is depicted in the graph as line 2. Also be sure to note that the point at which the two lines intersect is $(\bar{X}, \bar{Y})$. As a matter of fact, ANY line that passes through the point, $(\bar{X} = 14$-years education, $\bar{Y} = 20,000)$ satisfies the first OLS constraint!!!

2) CONCLUSION: The constraint that $\Sigma (Y - \hat{Y}) = 0$ only requires that the regression line passes through the point, $(\bar{X}, \bar{Y})$.

3) NOTICE also that since $\hat{a} = \bar{Y} - \hat{b}\bar{X}$, that if we transform $X$ such that $X' = X - \bar{X}$, then $X' = 0$ and $\hat{a} = \bar{Y}$. This can be a useful transformation, because you then know that when $\hat{b} > 0$ the predicted value, $\hat{Y}$, will be above the mean, $\bar{Y}$, when $X'$ is positive and will be below the mean when $X'$ is negative. (When $\hat{b} < 0$, the converse will be true.)

3. The second OLS criterion is that $\Sigma (X - \bar{X})(Y - \hat{Y}) = 0$. Let's investigate what this means:

a. Whereas the first criterion (viz., $\Sigma (Y - \hat{Y}) = 0$) does no more than
require that the regression line pass through the point, \((\bar{X}, \bar{Y})\), the second criterion (viz., \(\Sigma (X - \bar{X})(Y - \hat{Y}) = 0\)) determines the "best" regression line from all possible lines through this point.

b. Consider the three lines discussed previously:

<table>
<thead>
<tr>
<th></th>
<th>line 1</th>
<th>line 2</th>
<th>line 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(X - \bar{X})</td>
<td>(Y - \hat{Y})</td>
<td>((X - \bar{X})(Y - \hat{Y}))</td>
</tr>
<tr>
<td>Person 1:</td>
<td>-2</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>Person 2:</td>
<td>2</td>
<td>25</td>
<td>-10</td>
</tr>
<tr>
<td>Total ((\Sigma)):</td>
<td>0</td>
<td>0</td>
<td>-40</td>
</tr>
</tbody>
</table>

↑→ closer to 0 ← Zero!

Recall that \(\Sigma (X_i - \bar{X}) = 0\). (And it makes sense too!)

4. O.K. Now that we understand what the second condition means, let's see how to use it:

\[
\begin{align*}
\Sigma (X - \bar{X})(Y - \hat{Y}) &= 0 \\
\Sigma (X - \bar{X})(Y - [\hat{a} + \hat{b}X]) &= 0 \\
\Sigma (X - \bar{X})(Y - (\bar{Y} - \hat{b}\bar{X})) &= 0 \\
\Sigma (X - \bar{X})(Y - \bar{Y}) - \hat{b} \Sigma (X - \bar{X})^2 &= 0 \\
\text{OR} \quad \hat{b} &= \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (X - \bar{X})^2}
\end{align*}
\]

This can be simplified as follows:

a. Numerator:

\[
\begin{align*}
\Sigma (X - \bar{X})(Y - \bar{Y}) &= \Sigma XY - \Sigma X \bar{Y} - \Sigma \bar{X}Y + \Sigma \bar{X} \bar{Y} \\
&= \Sigma XY - \bar{Y} \Sigma X - \bar{X} \Sigma Y + n \Sigma \bar{Y} \\
&= \Sigma XY - \frac{\Sigma Y}{n} \Sigma X - \frac{\Sigma X}{n} \Sigma Y + n \left(\frac{\Sigma X}{n} \cdot \frac{\Sigma Y}{n}\right)
\end{align*}
\]
\[ \Sigma XY - \frac{(\Sigma Y)(\Sigma X)}{n} = \frac{(\Sigma X)(\Sigma Y)}{n} + \frac{(\Sigma Y)(\Sigma X)}{n} \]

\[ = \Sigma XY - \frac{(\Sigma Y)(\Sigma X)}{n} = \Sigma XY - n\bar{X}\bar{Y} \]

b. Denominator:

\[ \Sigma (x - \bar{x})^2 = \Sigma \left( x^2 - 2x\bar{x} + \bar{x}^2 \right) \]

\[ = \Sigma x^2 - 2\bar{x}\Sigma x + \Sigma x^2 \]

\[ = \Sigma x^2 - 2 \cdot \frac{\Sigma x}{n} \cdot \frac{\Sigma x}{n} + \frac{\Sigma x}{n} \cdot \frac{\Sigma x}{n} \]

\[ = \Sigma x^2 - 2 \cdot \frac{\Sigma x}{n} \cdot \frac{\Sigma x}{n} + \frac{\Sigma x}{n} \cdot \frac{\Sigma x}{n} \]

\[ = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = \Sigma x^2 - nx^2 \]

c. Thus the computation formula for \( \hat{b} \) is as follows:

\[ \hat{b} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma x^2 - \frac{(\Sigma X)^2}{n}} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma x^2 - nx^2} \]

H. This formula becomes even simpler if we use STANDARDIZED VARIABLES.

1. Recall that standardized variables are obtained by subtracting a variable from its mean and dividing this difference by its standard deviation. That is,

\[ Z = \frac{x - \bar{x}}{\hat{\sigma}_x} \]

where \( \hat{\sigma}_x = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}} \).

2. We have shown earlier (Lecture Notes, p. 25) that

\[ Z = \frac{\Sigma \frac{x - \bar{x}}{\hat{\sigma}_x}}{\frac{1}{\hat{\sigma}_x} \Sigma (x - \bar{x})} = \frac{\hat{\sigma}_x}{\Sigma x} = 0 \]

and that

137
\[ \hat{\sigma}_Z^2 = \frac{\sum (Z - \bar{Z})^2}{n - 1} = \frac{\sum Z^2}{n - 1} = \frac{\sum \left( \frac{X - \bar{X}}{\bar{\sigma}_X} \right)^2}{n - 1} = \frac{1}{\bar{\sigma}_X^2} \frac{\sum (X - \bar{X})^2}{n - 1} = \frac{1}{\bar{\sigma}_X^2} \bar{\sigma}_Z^2 = 1. \]

That is, if Z is a standardized random variable, \( \bar{Z} = 0 \) and \( \bar{\sigma}_Z^2 = 1 \).

I. Notice what happens to \( \hat{\alpha} \) and \( \hat{\beta} \) (in \( \hat{Z}_Y = \hat{\alpha} + \hat{\beta}Z_X \)) when \( X \) and \( Y \) have been standardized:

\[ \hat{\alpha} = \bar{Z}_Y - \hat{\beta}\bar{Z}_X = 0 \]
\[ \hat{\beta} = \frac{\sum (Z_X - \bar{Z}_X)(Z_Y - \bar{Z}_Y)}{\sum (Z_X - \bar{Z}_X)^2} = \frac{\sum Z_XZ_Y}{\sum Z_X^2} \]

1. When variables are standardized, we shall no longer use the letter "b", but switch to the Greek letter beta (\( \beta \)) which stands for the standardized regression coefficient. (On SPSS regression output, you will notice that there is a column for "b" and one for "beta". Notice that there is no constant in the beta-column. Now you know why.)

2. It is unfortunate that the symbol, \( \beta \), has two meanings. That is, you will recall that \( \beta \) also stands for the probability of a Type II error (Lecture Notes, p. 95). Be sure not to confuse these two uses of the same symbol.

J. NOTATION on regression coefficients

<table>
<thead>
<tr>
<th>unstandardized</th>
<th>standardized</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate</td>
<td>( \hat{\beta} )</td>
</tr>
<tr>
<td>parameter</td>
<td>( \beta )</td>
</tr>
</tbody>
</table>

138
Note how this retains the convention to designate an estimate by placing a "hat" over the parameter that it estimates.

K. Interpreting constants and unstandardized regression slopes in bivariate regression equations

1. Computers are adept at generating the regression coefficients you request. Yet it is up to you, the statistician, to interpret these numbers. Possibly the best way to explain how a regression slope should be interpreted is by giving an incorrect rendering and then by pointing out what is missing.

   a. An incorrectly rendered regression slope: "The slope indicates that a change of three units on the income measure would be estimated for each 1 unit change on the education measure."

   b. What is missing:

      1) No indication is given whether the slope is positive or negative.

      2) No indication is given of what the units are on the independent and dependent variables.

      3) No indication is given of what the unit of analysis (here, U.S. residents) is.

   c. A correctly rendered regression slope: "The slope indicates that an increase of three thousand dollars of annual income would be estimated for each additional year of a U.S. resident's education."

2. Likewise, the constant in a regression equation is more than "the y-intercept." Thus, it would be incomplete to refer to a constant as, for example, "the value on the income variable when the education measure
equals zero." A **correct** rendering would be one like the following:

"One would estimate U.S. residents with no educations to have annual incomes of $1,000."

L. **IMPORTANT:** A regression line explains part of the "variability" (or variance) in the dependent variable.

1. Please note the distinction here between the variance "in" a variable (i.e., \( SS_Y = \Sigma (Y - \bar{Y})^2 \)) and the variance "of" a variable (i.e., \( \sigma^2 = \frac{\Sigma (Y - \bar{Y})^2}{n-1} \)).

2. When statisticians speak of the variance "in" a variable, they are usually referring to something called a **SUM OF SQUARES**. For example, the variance in the variable \( Y \) is \( \Sigma (Y - \bar{Y})^2 \). Whenever \( Y \) is the dependent variable, this variance is called the "Total Sum of Squares."

3. If you think of scientific research as a game with specific rules that specify (among other things) which statistics are appropriate under which circumstances, the **GOAL** of this game is to find evidence of causal relations among phenomena. This evidence always consists (in part) of a covariation between measures of cause and effect. Thus, it is reasonable to think of the goal in the "statistics game" as being one of explaining as much of the variance in the dependent variable as you can. This is done by **partitioning** the total sum of squares into an explained part (that covaries with some or all of the independent variables) and an unexplained part (the **residual** variance—the variance left over).

4. The partitioning of the total sum of squares begins as follows:

\[
\Sigma (Y - \bar{Y})^2 = \Sigma [(\hat{Y} - \bar{Y}) + (Y - \hat{Y})]^2
\]

\[
= \Sigma (\hat{Y} - \bar{Y})^2 + \Sigma (Y - \hat{Y})^2 + 2 \Sigma (\hat{Y} - \bar{Y})(Y - \hat{Y})
\]
Now just consider the term \( 2 \sum (\hat{Y} - \bar{Y})(Y - \hat{Y}) \).

This term can be shown to equal zero. To demonstrate this, first recall that \( \hat{Y} = \hat{\alpha} + \hat{\beta}X \) and \( Y = \bar{Y} + \hat{\beta}X \).

Substituting these equalities into the above term yields the following:

\[
\sum (\hat{Y} - \bar{Y})(Y - \hat{Y}) = \sum \left( [\hat{\alpha} + \hat{\beta}X] - [\bar{Y} + \hat{\beta}X] \right)(Y - \hat{Y})
\]

\[
= \sum (\hat{\alpha} + \hat{\beta}X - \hat{\alpha} - \hat{\beta}X)(Y - \hat{Y})
\]

\[
= \sum (\hat{\beta}X - \hat{\beta}X)(Y - \hat{Y})
\]

\[
= \hat{\beta} \sum (X - \bar{X})(Y - \hat{Y})
\]

But according to the 2nd OLS criterion, \( \sum (X - \bar{X})(Y - \hat{Y}) = 0 \). Thus, the proof concludes as follows:

\[
\hat{\beta} \cdot 0 = 0.
\]

Consequently, the total sum of squares can be partitioned as follows:

\[
\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2
\]

\[
SS_{\text{TOTAL}} = SS_{\text{REGRESSION}} + SS_{\text{ERROR}}
\]

Total variance in Y \hspace{1cm} Variance explained \hspace{1cm} Variance unexplained

\[ df = n - 1 \hspace{2cm} df = 1 \hspace{2cm} df = n - 2 \]

M. AN ASIDE: The distance to the regression line is always drawn parallel to the Y-axis. That is, deviations of \( Y \) from \( \hat{Y} \) are always for fixed values of \( X \). It is to convey this conditionality of \( Y \) on \( X \) that one refers to the regression of \( Y \) on \( X \), never to the regression of \( X \) on \( Y \).

N. A demonstration of the nonintuitive nature of the inequality, \( \sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2 \):

If you calculate \( \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \) and \( \hat{\beta} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \), then not only
for each person is \([Y - \bar{Y}] = [\hat{Y} - \bar{Y}] + [Y - \hat{Y}]\), BUT the squares of the 3 parts of this equivalence add up to an equivalence as well.

For instance, the above graphic depicts a case in which \(\bar{Y} = 7\) and where for the \(i\)th observation, \(\hat{Y}_i = 10\) and \(Y_i = 5\). Note, for example, that when \(\hat{Y}_i - \bar{Y} = 3\) and \(Y_i - \hat{Y}_i = -5\), then \((Y_i - \bar{Y}) = 3 + [-5] = -2\).

Yet, it is clear that
\[
(Y_i - \bar{Y})^2 = (-2)^2 = 4 \neq (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 = 3^2 + (-5)^2 = 34.
\]

BUT if you add up all of these squares you DO find that
\[
\Sigma (Y - \bar{Y})^2 = \Sigma (\hat{Y} - \bar{Y})^2 + \Sigma (Y - \hat{Y})^2.
\]

THIS IS NOT INTUITIVELY TRUE. Nonetheless it IS true!

0. One important CONSEQUENCE of this equivalence is an improved estimate of the "true" variance of \(Y\). The idea here is that the variance in \(Y\) explained by \(X\) makes the variance of \(Y\) appear much larger than it "really" is. This improved estimate is called the MEAN SQUARE ERROR (or MSE) and is assigned the symbol, \(\sigma^2\). (Note the absence of a subscript on this lower-case sigma.)
\[ \sigma^2 = \text{MSE} = \frac{\text{SS}_{\text{ERROR}}}{n - k - 1} \]

NOTE: This is the general formula for the MSE in which \( k \) equals the number of independent variables in the regression equation. Since there is only one independent variable in a bivariate regression, the denominator equals \( n - 2 \) in this case.

AN ASIDE: Regression output from SPSS (e.g., Lecture Notes, p. 164) labels the MSE as the MEAN SQUARE RESIDUAL and labels its square root as the "Std. Error of the Estimate"—an estimate of the "true" standard deviation of \( Y \). This is a misleading use of the term, standard error!!! The standard error of a statistic is the standard deviation estimate for the sampling distribution of that statistic. What is labeled a standard error on your regression output is a standard deviation estimate for the distribution of average residuals (or, if you prefer, of "true" deviations of \( Y \) from their mean after their adjustment for the effects of the independent variable[s]).

P. A second important CONSEQUENCE of the fact that variance can be partitioned into explained and unexplained components is that regression analysis affords a measure of LINEAR association between two variables, namely the proportion of variance in the dependent variable linearly explained by the independent variable:

\[ \text{Coefficient of determination} = r^2 = \frac{\text{SS}_{\text{REGRESSION}}}{\text{SS}_{\text{TOTAL}}} \]

THE CORRELATION COEFFICIENT (or Pearson correlation), "\( r \)", is the square root of the coefficient of determination with the same sign as the slope (\( b \)) between the correlated variables. Two useful formulas:
\[ r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{SS_X * SS_Y}} = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right) \left(\sum_{i=1}^{n}y_i^2 - n\bar{y}^2\right)}} \]

1. The relation between \( \hat{b} \) and \( r \):
   
   a. Recall that \( r^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} \).

   b. And that \( \hat{y}_i = \hat{a} + \hat{b}x_i \)
   
   and \( \bar{y} = \hat{a} + \hat{b}\bar{x} \).

   c. Taking the numerator of the formula for \( r^2 \), we get
   
   \( \sum (\hat{y} - \bar{y})^2 = \sum (\hat{a} + \hat{b}x - \hat{a} - \hat{b}\bar{x})^2 = \sum (\hat{b} [x - \bar{x}]^2) = \hat{b}^2 \sum (x - \bar{x})^2 = \hat{b}^2 \sum (x - \bar{x})^2 \).

   d. This implies that
   
   \[ r^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{\hat{b}^2 \sum (x - \bar{x})^2}{\sum (y - \bar{y})^2} = \hat{b}^2 \frac{SS_X}{SS_Y} = \hat{b}^2 \frac{\sigma_X^2}{\sigma_Y^2} . \]

   Therefore \( \hat{b}^2 \) is proportional to \( r^2 \)—the proportion of variance explained in one variable by another.

   e. Moreover, recalling that \( r \) and \( \hat{b} \) take the same sign, it follows that
   
   \[ r = \hat{b} \frac{\sigma_X}{\sigma_Y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{SS_X} \times \sqrt{\frac{SS_X}{SS_Y} \left(\frac{SS_X}{n-1}\right)} = r_{XY} \]

2. Given this relation between \( r \) and \( \hat{b} \), you will note that when \( X \) and \( Y \) are standardized (i.e., when \( \sigma_X = \sigma_Y = 1 \)) the correlation between them is
the same thing as their slope. Thus, for example, if within a random sample of Canadians the correlation between education and income were .3, one could express this in words as, "One would estimate a .3 standard deviation increase in a Canadian's income for each 1 standard deviation increase in his or her education." (WARNING: The next section of these Lecture Notes introduces partial correlations, which unlike correlation coefficients may NOT be interpreted as slopes.)

3. It is with the correlation coefficient and its square—the coefficient of determination—that we reach the boundary between a world understood by "normal people" and a world only understood by statisticians. With some effort most "normal people" can come to understand the meaning of standardized units, and with it the idea of a slope between two standardized variables. However, it is only statisticians who understand the phrase, "Education linearly explains 9% of the variance in income." To understand what this means, one must know what is meant by the variance in a variable and what it means to "explain" a proportion of this. Once you understand the meanings of both the correlation coefficient (a slope between two standardized variables) and the coefficient of determination (the proportion of variance in one variable that is linearly explained by another variable), you have begun to understand a language that is exclusively shared among statisticians.

Q. The assumptions of linear regression

1. HOMOSCEDASTICITY: The variance of Y is the same within different categories of X.

2. LINEARITY: The means of Y are lie in a straight line across different categories of X.
3. RANDOMNESS: The random variables $Y_i$ are statistically independent.

4. NORMALITY: For any fixed value of $X$, $Y$ is normally distributed. There are TWO WAYS of expressing this last assumption:

a. $Y \sim N(\alpha + \beta X, \sigma^2)$, where $\sigma^2$ is the variance of $Y$ around the regression line.

b. $Y = \beta X + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.

NOTE: The distributions of $Y$ and $\epsilon$ are identical, except for their means. Also note that in each case the estimate of the variance is the MEAN SQUARE ERROR term from the regression.

R. Using a BOX PLOT to check the assumptions of linear regression.

1. What is a box plot?

   a. A box plot is a plot of the data that has the dependent variable ($Y$) on the vertical axis and the independent variable ($X$) on the horizontal axis.

   b. Values on the dependent variable are in terms of the units it has. (For example, the illustration on the next page was generated only after an income measure ($\text{rincome}$) was recoded into units of dollars.)

   c. Values on the independent variable should be collapsed into groups of at least 30 data points. (In any case, roughly equal numbers of data points should belong to each group.)

   d. A box is drawn above that point on the X-axis that corresponds to each group. The box’s top and bottom delineate the interquartile range of values on the dependent variable for units of analysis.
within the box's corresponding group. A line across the box indicates the median value within the group. "Whiskers" out each end of the box show the range of values within the group (excluding outliers).

e. Outliers (namely, values within a group that exceed the box's interquartile bounds by 1.5 box lengths) are indicated by circles and their corresponding subject numbers.

f. The plot below illustrates a box plot for the linear relation between RINCOME (Y) and PRESTIGE (X) from the 1984 NORC survey of the U.S. adult population.
The program used to generate this output:

temporary.
select if ((rincome ne 13) and (rincome ne 99)).
recode rincome (1=500)(2=2000)(3=3500)(4=4500)(5=5500)(6=6500)
recode prestige(12 thru 16=17)(18 thru 23=24)(25 thru 27=28)
(29,30,31=32)(33=34)(35 thru 36)(37 thru 39=40)(41 thru 44=45)(46=47)
(48 thru 49)(51 thru 56=57)(58 thru 60=61)(62 thru 78=82).
examine variables=rincome by prestige/plot=boxplot/statistics=none/nototal.

2. Checks made using a box plot.

a. Heteroscedasticity (i.e., the absence of homoscedasticity) is evident
   when the IQRs (i.e., the boxes) in a box plot are of different heights.

b. The linearity assumption is violated if no monotonic increase (or
decrease) in the Ys can be detected for increasingly larger Xs. This
can be checked by seeing if a single (but not a horizontal) straight
line can be drawn between upper and lower quartile bounds (i.e., the
top and bottom) of all boxes within the plot.

c. Randomness is usually built into one's research design. Statistical
independence among observations is ensured by random sampling
subjects. Nonetheless, if your data were collected over TIME, the
assumption is probably not met. (E.g., the GNP of the United States
in 1999 is not independent of the U.S. GNP in 1998.) If you plot the
data by time (that is, let time be the X-variable), such a violation
of the randomness assumption may be detected as TRACKING of the data.
Tracking is evident when similar values on the dependent variable are
found for observations that follow each other in sequence.

d. Violations of the normality assumption may result in biased
regression coefficients as well as in misleading tests of
significance. However, the data must depart radically from normality for these tests to be noticeably effected. Sometimes an extremely aberrant observation (a.k.a. an "outlier") can skew the conditional distribution of $Y$ for a particular $X$-value. When this happens one usually identifies the outlier, determines why it is atypical of the rest of the data, drops the outlier from the analysis, and acknowledges the outlier and its atypical nature in the written results (usually in a footnote).

"He says we've ruined his positive association between height and weight."

S. Testing hypotheses about $\hat{b}$

A derivation of the distribution of $\hat{b}$ is beyond the scope of the course, so here it is:

$$\hat{b} \sim N(b, \sigma_b^2),$$

where

$$\sigma_b^2 = \frac{\text{MSE}}{\sum (X - \bar{X})^2}.$$

1. Note that the variance of $\hat{b}$ is the MEAN SQUARED ERROR standardized for the amount of variation in $X$. Thus if a sample has large deviations from $\bar{X}$ and a small corresponding error in estimates of $Y$, the variance
of $\hat{b}$ is small. The converse of this is also true.

2. Because the units of a slope are "units Y per unit X" (e.g., dollars income per year of education), the units of this variance are the square of this (e.g., squared dollars per squared year of education).

3. This variance can be used to test $H_0: b = 0$
   $H_A: b \neq 0$ at a given significance level (e.g., $\alpha=.05$). Here your rejection rule would be as follows:

   $$\text{Reject } H_0 \text{ when } |\hat{b}| > 1.96 \sqrt{\frac{\text{MSE}}{SS_X}}$$

   (NOTE: You should use $t_{n-2,\alpha/2}$ instead of 1.96 when the degrees of freedom associated with MSE [namely, $df = n-2$] are less than 30.)

T. Finding a 95% confidence interval for $\hat{b}$

Similarly (and again assuming a large enough sample such that the MSE's degrees of freedom equal 30 or more), finding a 95% confidence interval about $\hat{b}$ is done using the formula,

$$\hat{b} \pm 1.96 \sqrt{\frac{\text{MSE}}{SS_X}}$$

U. AN ILLUSTRATION: Imagine that you wish to know whether greater use of knee braces by football players has an effect upon the number of serious football-related injuries suffered by members of college football teams. Because football teams usually have policies concerning knee braces, you find that use of knee braces is NOT a random variable among members of the same football team. Since use of knee braces is (less arguably) random among football teams, you use "the football team" as your unit of analysis.

You have data on thirty football teams. Your SPSS data set includes two variables, namely "pbrace" (proportion of the football team comprised of
players that regularly use a knee brace) and "injuries" (number of serious football-related injuries per 100 players suffered by the football team last year). You run the following SPSS instructions:

compute j = pbrace * injuries.
frequencies general = pbrace,injuries,j / statistics = mean, variance.

Parts of your output look as follows:

<table>
<thead>
<tr>
<th></th>
<th>PBRACE</th>
<th>INJURIES</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>.682</td>
<td>2.97</td>
<td>2.23</td>
</tr>
<tr>
<td>Variance</td>
<td>.013</td>
<td>28.17</td>
<td>18.84</td>
</tr>
</tbody>
</table>

1. What is known at this point:

\[ n = 30 \quad \bar{x} = .682 \quad \sigma_x^2 = .013 \quad \bar{y} = 2.97 \quad \sigma_y^2 = 28.17 \quad \bar{j} = 2.23 \]

\[ SS_X = (n-1) \cdot \sigma_x^2 = 29 \cdot .013 = .377 \]

\[ SS_Y = SS_{TOTAL} = (n-1) \cdot \sigma_y^2 = 29 \cdot 28.17 = 816.93 \]

\[ \Sigma XY = n \cdot \bar{j} = 30 \cdot 2.23 = 66.9 \]

\[ r_{XY} = \frac{\Sigma XY - \bar{x} \bar{y} n}{\sqrt{SS_X \cdot SS_Y}} = \frac{66.9 - (.682 \cdot 2.97 \cdot 30)}{\sqrt{.377 \cdot 816.93}} = \frac{6.134}{17.549} = .35 \]

\[ r_{XY}^2 = (.35)^2 = .122 \]

\[ SS_{REGRESSION} = r_{XY}^2 \cdot SS_{TOTAL} = .122 \cdot 816.93 = 99.797 \]

\[ SS_{ERROR} = (1 - r_{XY}^2) \cdot SS_{TOTAL} = .878 \cdot 816.93 = 717.133 \]

\[ MSE = \frac{SS_{ERROR}}{n - k - 1} = \frac{717.133}{30 - 2} = 25.612 \]

\[ t_{28,.05} = 1.701 \quad t_{28,.025} = 2.048 \]

2. Give the unstandardized regression equation appropriate to your research question and say in words what each regression coefficient means.

\[ \hat{b} = \frac{\Sigma XY - n \bar{x} \bar{y}}{SS_X} = \frac{66.9 - (30 \cdot .682 \cdot 2.97)}{.377} = 16.27 \]

In words: For every additional 10% of a football team's players who
regularly used a knee brace, one would estimate an additional 1.627 serious football-related injuries per 100 players suffered by the team last year.

\[ \hat{a} = \bar{Y} - \hat{b} \bar{X} = 2.97 - (16.27 \times .682) = -8.126 \]

In words: One would estimate minus (8.126) serious football-related injuries per 100 players suffered last year by a football team on which none of the team's players regularly used a knee brace.

**NOTE:** An alternative formula for finding the unstandardized regression coefficient is (with rounding error) as follows:

\[ \hat{b} = r \frac{\hat{a}_Y}{\hat{a}_X} = .35 \sqrt{\frac{28.17}{.013}} = 16.29 \]

3. Would you recommend use of a knee brace to prevent serious football-related injuries? (Justify your answer by referring to your data.)

No. The sign of the slope suggests that rather than preventing serious football-related injuries, knee brace use enhances such injuries.

4. Test the null hypothesis that the proportion of football teams comprised of players that regularly use a knee brace (linearly) influences the number of serious football-related injuries suffered by the teams. (Use the .05 level of significance.)

\[ H_0: b = 0 \]
\[ H_A: b \neq 0 \]

Note that the standard error of \( \hat{b} \) is

\[ \sigma_{\hat{b}} = \sqrt{\frac{MSE}{SS_X}} = \sqrt{\frac{25.612}{.377}} = 8.242 \]

The rejection rule is thus
"Reject $H_0$ if $|\hat{b}| > t_{28, .025} \sigma_b^2 = 2.048 \times 8.242 = 16.88$ ."

Because $\hat{b} = 16.27 < 16.88$ we fail to reject $H_0$.

5. Find the 95% confidence interval for $\hat{b}$.

The 95% confidence interval for $\hat{b}$ is $16.27 \pm t_{28, .025} \sigma_b^2$

or $16.27 \pm 2.048 \times 8.242$ or $16.27 \pm 16.88$ or $(-0.61, 33.15)$.

V. Testing hypotheses about $r$

1. We know that

$$\hat{b} \sim N( b, \sigma_b^2 )$$, where $\sigma_b^2 = \frac{\text{MSE}}{SS_X}$.

2. Referring back to the earlier discussion [Lecture Notes, p. 24] regarding conversions among normal distributions, you may recall that if $X \sim N(\mu, \sigma_X^2)$ then the distribution of $X' = k \times X$ is $X' \sim N( k\mu, k^2 \sigma_X^2 )$. For example, if $X$ is in units of feet then $X' = 12X$ is in units of inches, and if $X \sim N(5, 10)$ then $X' \sim N(60, 1440)$.

3. Taking into account that $\hat{b} \sim N( b, \sigma_b^2 )$ and setting $k = \frac{\sigma_X}{\sigma_Y}$, it follows that

$$\hat{b} \frac{\sigma_X}{\sigma_Y} \sim N( b \frac{\sigma_X}{\sigma_Y}, \sigma_b^2 \frac{\sigma_X^2}{\sigma_Y} )$$.

4. If we approximate $\hat{b} \frac{\sigma_X}{\sigma_Y}$ using $r = \hat{b} \frac{\sigma_X}{\sigma_Y}$, the formula for estimating the variance of $r$ is as follows:

$$\text{Var}(r) = \frac{\hat{b}^2 \sigma_X^2}{\sigma_Y^2} \frac{\text{MSE}}{SS_X} = \frac{\text{MSE}}{SS_X}$$

153
5. Now consider this last quotient:

\[
\frac{\text{MSE}}{\text{SS}_Y} = \frac{\sum (Y - \hat{Y})^2}{(n-2) \sum (Y - \bar{Y})^2}
\]

Recalling that \(\sum (Y - \bar{Y})^2 = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2\),

\[
= \frac{\sum (Y - \bar{Y})^2 - \sum (\hat{Y} - \bar{Y})^2}{(n-2) \sum (Y - \bar{Y})^2}
\]

\[
= \frac{1}{n-2} \left( \frac{\sum (Y - \bar{Y})^2}{\sum (Y - \bar{Y})^2} - \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} \right)
\]

\[
= \frac{1}{n-2} \left( 1 - \frac{\text{SS}_{\text{Regression}}}{\text{SS}_{\text{Total}}} \right) = \frac{1 - r^2}{n-2}
\]

Thus the numerator equals the proportion of the variance NOT explained and we are led to the conclusion that

\[r \sim N(\rho, \frac{1 - \rho^2}{n-2})\].

I.e., we would conclude that the variance around \(\rho\) depends on the magnitude of \(\rho\). Consequently, the estimated variance of \(r\) will not be constant for different estimates of \(\rho\).

6. AN ASIDE: Although \(\hat{\sigma}_r^2\) is dependent on \(\rho\), the variance of \(\hat{b}\) is not dependent on \(b\). The former dependence is introduced when the above approximation of \(\frac{\sigma_X}{\hat{\sigma}_Y}\) with \(\frac{\hat{\sigma}_X}{\hat{\sigma}_Y}\) is made.

7. The dependence of \(\hat{\sigma}_r^2\) on \(\rho\) presents no problem in tests of null hypotheses of the form, \(H_0: \rho = 0\). A PROBLEM does result as one attempts to set up confidence intervals around estimates of \(\rho\), however.

   a. A confidence interval around \(r\) is an interval that includes the
estimated value of \( \rho \) between an upper and a lower confidence bound.

b. The larger the absolute value of \( \rho \), the smaller the variance estimate that should be used.

c. These two statements imply that "the bound of a confidence interval that is further from zero" should be closer to the estimate of \( \rho \) than the other bound, since it requires a smaller variance estimate.

d. In contrast, this problem does not exist when testing the null hypothesis, \( H_0: \rho = 0 \), since in a two-tailed test each critical value will be equidistant from zero. Now to an illustration.

8. Let's now return to the knee brace illustration.

a. First, we can ask how much of the variance in injuries is explained by use of knee braces. (Please remember that all proportions of variance explained are some form of "squared r." Here it is the correlation coefficient that is squared.)

\[
r_{XY}^2 = (0.35)^2 = 0.122
\]

b. Second, we can test if the amount of variance found in part a is a statistically significant amount of variance at the .05 level.

\[
H_0: \rho^2 = 0 \\
H_A: \rho^2 > 0
\]

is equivalent to testing

\[
H_0: \rho = 0 \\
H_A: \rho \neq 0
\]

\[
t_{28} = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{0.35}{\sqrt{\frac{1 - 0.122}{30 - 2}}} = 1.977 < 2.048 = t_{28,.025}
\]

Conclusion: No, \( r_{XY}^2 = 0.122 \) is not significantly large at \( \alpha = 0.05 \).

9. A FEW COMMENTS:

a. The degrees of freedom are \( n - 2 \), since a degree of freedom is
lost when the mean of EACH variable is calculated.

b. A correlation of ±.35 is not very likely to occur by chance! (In this sample of only 30 cases it was almost sufficiently large to reject $H_0$.) Thus if one variable explains (linearly) about 15% of the variance in another, a sample of around 30 will likely be large enough to detect it. This should give you some intuitive "feel" for the relative strength of $|r| = .35$.

W. Finding a 95% confidence interval for $r$

RECALL that $\sigma_r^2 = \frac{1 - \rho^2}{n - 2}$.

Thus when $|\rho|$ is small, you would have a larger variance than when $|\rho|$ is large! As a result, your confidence interval must be "lopsided" with a larger deviation from $r$ toward zero and a smaller deviation from $r$ toward ±1. To estimate these deviations we use Table E, which can be found at the end of the Syllabus. Here are THE MECHANICS:

1. Table E provides a transformation (i.e., a "mapping") of $r$ to a variable with a normal distribution—a variable with a distribution unrelated to $\rho$’s magnitude. Specifically,

$$T(r) \sim N\left(T(\rho), \frac{1}{n - 3}\right).$$

2. In the knee brace problem we have $r = .35$ and $n = 30$, thus the 95% confidence interval is

$$T(r) \pm z_{.025} \cdot \sqrt{\text{Var}[T(r)]}$$

OR

$$T(r) \pm 1.96 \cdot \sqrt{\frac{1}{n - 3}}.$$
3. From Table E we find that $T(.35) = .3654$. Thus the confidence interval is

$$.3654 \pm 1.96 \sqrt{\frac{1}{n - 3}}$$

or $.3654 \pm 1.96 * \sqrt{\frac{1}{30 - 3}}$ or $.3654 \pm .3772$ or $(-.012, .743)$

4. BUT this is a confidence interval around $T(r)$! So we must convert these two values back to $r$'s. This yields $T^{-1}(-.012) = -.012$ and $T^{-1}(.743) = .631$. Thus the 95% confidence interval for $r = .35$ is $(-.012, .631)$.

5. Notice how these confidence bounds look on a number line:

In particular, notice that the bounds of the interval are "lopsided." That is, note that $|.35 - (-.012)| = .362 > .281 = |.35 - .631|$. X. The distribution of $\hat{a}$ when $\bar{X} = 0$.

1. Let $\hat{a}_c$ equal the constant in a bivariate regression equation in which the mean on the independent variable equals zero. When $X$ is transformed by subtracting out its mean (i.e., when $\bar{X} = 0$), then $\hat{a}_c = \bar{Y}$ and thus $\hat{a}_c$ is an estimate of $\mu_Y$.

   NOTE: Variables are "centered" if their means equal zero. The "c" subscripts on $\hat{a}_c$ and $X_c$ are reminders that $\bar{X}_c = 0$.

2. The standard error of $\hat{a}_c$ is usually much smaller than that associated...
with the usual estimate of $\gamma$, however. In particular, 
\[ \hat{a}_c \sim N\left( a_c, \frac{\sigma^2}{n} \right), \]
but in this case $\hat{\sigma}^2 = \text{MSE}$. And the
MSE is a better estimate of the "true" variance of $Y$—a variance net of
the effects of $X$.

3. Note that $\hat{a}_c$ and $\hat{b}$ are independent.

   a. Recall that the first OLS criterion fixes $(\bar{x}, \bar{y})$ as a point on
   the regression line. Once one's independent variable has been
   centered such that $\bar{x}_c = 0$ (and consequently, such that $\hat{a}_c = \bar{y}$),
   this point is fixed at $(0, \hat{a}_c)$.

   b. This point will be on the regression line independent of the value of
   $\hat{b}$ as determined by the second OLS criterion.

   c. Since $\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{a}$ is dependent upon $\hat{b}$ under any other condition
      (namely, whenever $\bar{x} \neq 0$).

Y. The distribution of $\hat{y}_e$ (the estimated value of $Y$ when $X = X_e$):

1. We begin with knowledge of the following four things:

   a. If $X_c = X - \bar{X}$, then $\bar{x}_c = 0$.

   b. $\hat{a}_c = (\hat{a} \mid \bar{x} = 0) \sim N\left( a_c, \frac{\sigma^2}{n} \right)$

   c. $\hat{b} \sim N\left( b, \frac{\sigma^2}{SS_X} \right)$

   d. $\hat{a}_c$ and $\hat{b}$ are statistically independent.

2. If $X$ and $Y$ are normally distributed and statistically independent, the
distribution of $W = X + Y$ is $N\left( \mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 \right)$. (See the earlier
discussion [Lecture Notes, p. 107] of the distribution of the difference
between two proportions.)

3. Conclusion:
\[
\hat{\theta}_c = \hat{\alpha}_c + \hat{b}X_c - N\left( \alpha_c + bX_c, \frac{\sigma^2}{n} + \frac{\chi^2_c \sigma^2}{SS_X} \right), \text{ where } \sigma^2 = \text{MSE}.
\]

NOTE: Since \( \hat{b} \) is multiplied by the constant, \( X_c \), the variance of this product equals the variance of \( \hat{b} \) times \( X_c^2 \).

4. In the more general case in which data on the independent variable have not been centered, the distribution of \( \hat{\gamma} \) would be as follows:
\[
\hat{\gamma} = \hat{a}_c + \hat{b}(X - \bar{X}) - N\left( \alpha_c + b(X - \bar{X}), \sigma^2 \left[ \frac{1}{n} + \frac{(X - \bar{X})^2}{SS_X} \right] \right),
\]
which, after "uncentering" the data by letting \( \hat{a} = \hat{a}_c - \hat{b}\bar{X} \) and \( a = a_c - b\bar{X} \), can be rewritten as follows:
\[
\hat{\gamma} = \hat{a} + \hat{b}X - N\left( a + bX, \sigma^2 \left[ \frac{1}{n} + \frac{(X - \bar{X})^2}{SS_X} \right] \right).
\]
Yet note that this is really a class of distributions, since the distribution if \( \hat{\gamma} \) is different for every value of \( X \). Accordingly, it now becomes meaningful to speak of the distribution of \( \hat{\gamma}_e \) (i.e., the distribution of \( \hat{\gamma} \) for a fixed value of \( X = X_e \)). For example, if \( Y \) measures "starting income on one's first job" and if \( X \) measures "last year of education completed," one might wish to estimate the first-job-starting-income of high school graduates" (for whom \( X_e = 12 \)). And so,
\[
\hat{\gamma}_e = \hat{a} + \hat{b}X_e - N\left( a + bX_e, \sigma^2 \left[ \frac{1}{n} + \frac{(X_e - \bar{X})^2}{SS_X} \right] \right).
\]

5. This formula yields a more general formula for the standard error of \( \hat{a} \) than the one just given—a formula that applies when \( \bar{X} \neq 0 \):
Since (as always) \( \hat{a} = \hat{\gamma}_e \mid X_e = 0 \) (i.e., \( \hat{a} \) equals the estimated value

159
of \( Y_e \) when \( X_e = 0 \), the formula for the standard error of the
can be written as:

\[
\hat{\sigma}_a = \sqrt{\text{MSE} \left( \frac{1}{n} + \frac{(0 - \bar{X})^2}{SS_X} \right)}.
\]

Note that this formula is just for the special case in which \( X_e = 0 \).
The standard error for the more general case in which \( X_e = k \) is found
by substituting \( k \) for \( 0 \) in the formula.

Z. The distribution of \( \hat{Y}_e \) allows you to estimate the parameter, \( \mu_Y | X_e \). For
example, you may wish an interval estimate for the mean income (\( \mu_Y \)) of
Americans with a Ph.D. degree (\( X_e \)). That is, you might want a confidence
interval around \( \hat{Y}_e \).

Until now \( \hat{Y} \) has been rather freely referred to as a "prediction" instead of
as an estimate of a population parameter. In fact, you may use \( \hat{Y} \) as both
predictor (\( \hat{Y}_p \)) and estimator (\( \hat{Y}_e \)). You will find only a slight difference
in the formula for its variance when it is used in these two ways. The
variance introduced above is the variance of \( \hat{Y}_e \)—the variance of \( \hat{Y} \) when it
is used as an estimator.

When \( \hat{Y} \) is used to make a prediction based upon data previously collected, a
different variance estimate is needed. Let's say that you wish to predict
how much money you would earn if you earned a Ph.D. If you wish to make an
interval estimate of a data point for a particular unit of analysis, your
PREDICTION INTERVAL must take into account the additional variance due to
the introduction of this new datum. Like \( \hat{Y}_e \), \( \hat{Y}_p = \hat{a} + \hat{b}X_p \). However,

\[
\hat{Y}_p \sim N( \hat{a} + \hat{b}X_p, \sigma^2 \left[ \frac{1}{n} + \frac{(X_p - \bar{X})^2}{SS_X} + 1 \right] )
\]

Adding \( \sigma^2 \) into the variance term increments the variance by the variation
contributed by a particular unit of analysis (in this case, you—a potential Ph.D. recipient). NOTE: At issue in distinguishing a prediction from an estimate is that in the former case \( \hat{Y} \) is being viewed as the dependent variable's value for a particular unit of analysis (i.e., one with a name: Nebraska, the Ford Foundation, you, etc.) having a specific value(s) on an independent variable(s). When \( \hat{Y} \) is an estimate, it is viewed as the dependent variable's value for a type of unit of analysis (again, having a specific value[s] on the independent variable[s]) within the population from which the sample was drawn. In the former case one is predicting a single data point; in the latter case one is estimating a conditional mean.

AA. AN ILLUSTRATION: You have a random sample of 100 American cities and have fit the following regression model to data on ROBBEREZ (the annual number of robberies in a city) and POPPTHOU (city population in thousands):

\[
ROBBEREZ = -1 + .1 \times POPPTHOU + \epsilon
\]

Also assume that \( \sigma^2 = MSE = 4 \) and that you have the following output:

<table>
<thead>
<tr>
<th>Statistics</th>
<th>ROBBEREZ</th>
<th>POPPTHOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>9.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Variance</td>
<td>4.21</td>
<td>25.00</td>
</tr>
</tbody>
</table>

1. Assuming that Ames has a population of 60,000, one can use these data to predict the number of robberies that will occur in Ames next year to equal 5 (i.e., \(-1 + 5\)).

2. Assuming that Los Angeles has a population of 10 million, one would predict the number of robberies that will occur there next year to equal 999 (i.e., \(-1 + 1,000\)).

3. Now, let us find a 95% prediction interval around the prediction that 5
robberies would occur in Ames next year. We begin with the general formula,

\[ \hat{\gamma}_p \pm 1.96 \times \hat{\sigma} \left( \frac{1}{n} + \frac{(\hat{x}_p - \bar{x})^2}{SS_x} \right) + 1 \]

OR

\[ 5 \pm 1.96 \times 2 \left( \frac{1}{100} + \frac{(60 - 100)^2}{(100 - 1) \times 25} \right) + 1 \]

OR

\[ 5 \pm 5.05 \quad \text{OR} \quad (-0.05, 10.05) \]

I.e., assuming that the factors that determined the number of robberies in cities within our sample are the same as those in play in Ames next year, we predict that next year from 0 to 11 robberies will take place in Ames. Thus if there were 12 robberies next year, you might suspect that the high rate indicated something peculiar about that year in Ames.

One can begin graphing this and other prediction intervals as follows:

![Graph showing # of robberies per year and city size (in thousands)]

\[ \hat{\gamma}_{LA} = 999 \]

\[ \bar{\gamma} = 9 \]

\[ \hat{\gamma}_{Ames} = 5 \]
4. We can now add to this graph the 95% prediction interval around the prediction of 999 robberies in LA next year:

\[ 999 \pm 1.96 \times 2 \sqrt{\frac{1}{100} + \frac{(10,000 - 100)^2}{99 \times 25} + 1} = 999 \pm 780.08 \]

OR (218 to 1780) ← a VERY wide prediction interval!

5. Finally, note that the smallest prediction interval is at \( X = \bar{X} \):

\[ 9 \pm 1.96 \times 2 \sqrt{\frac{1}{100} + 1} = 9 \pm 3.94 \]

OR (5.06, 12.94) OR from 5 to 13 robberies

**AB. CONCLUSIONS:**

a. The further the basis of your prediction (i.e., \( X_p \)) is from \( \bar{X} \) AND

b. the smaller the sample size AND

c. the smaller the estimated population variance of \( X \),

**THE LESS PRECISE YOUR PREDICTION will be.**

These conclusions follow as a direct consequence of the formula for a prediction interval:

\[
\hat{y}_p \pm t_{\alpha/2} \times \hat{\sigma} \left[ \frac{1}{n} + \frac{(X_p - \bar{X})^2}{SS_X} + 1 \right]^{1/2}
\]

**NOTE:** A critical value of \( t \) (rather than \( Z \)) will be needed in computing both confidence and prediction intervals whenever \( n - k - 1 < 30 \), where \( k \) = the number of independent variables in the regression equation. Thus in the bivariate case, the number of degrees of freedom for \( t \) is \( n - 2 \).
Some SPSS Regression Output

The program:

```plaintext
DATA LIST RECORDS=1 / attend 1-2 prejud 4-5.
COMPUTE newx = (attend - 28)**2.
BEGIN DATA.
11 36
46 33
3 6
16 42
41 49
21 51
23 61
10 23
34 57
48 18
28 65
55 3
END DATA.
REGRESSION VARIABLES=newx, prejud/DEPENDENT=prejud/ENTER.
```

The output:

### Regression

#### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.714*</td>
<td>0.5466</td>
<td>0.5303</td>
<td>5.20688</td>
</tr>
</tbody>
</table>

* Predictors: (Constant), NEWX

#### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>4644.67558</td>
<td>1</td>
<td>4644.67558</td>
<td>167.49969</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>271.32442</td>
<td>10</td>
<td>27.13244</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>4816.00000</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Predictors: (Constant), NEWX
b Dependent Variable: PREJUD

#### Coefficients

```
<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std Error</td>
</tr>
<tr>
<td>1</td>
<td>59.198150</td>
<td>2.280982</td>
</tr>
<tr>
<td></td>
<td>NEWX</td>
<td>.085542</td>
</tr>
</tbody>
</table>
```

* Dependent Variable: PREJUD
Assumptions Underlying Regression Analysis (continued)

A. Assumption 5: The design matrix, $X$, is fixed, or measured without error.

1. A researcher “fixes” values on an independent variable when subjects are assigned to particular groups (e.g., as in a psychological experiment), or when “fixed” numbers of subjects are sampled within strata (e.g., 50 males and 50 females).

2. When $X$ has been fixed, one speaks of its variables as having “fixed effects.” If the values of the independent variables in $X$ are the unforeseen results of randomly sampled cases, these variables are said to have “random effects” on the dependent variable. Even though you may not have fixed $X$, it is important to be aware that you are assuming $X$ to be fixed (or measured without error) when you do regression analysis.

3. When a researcher has less control over the values of the independent variables (i.e., when their effects are random ones), she must assume that these variables are measured without error. If one thinks of $e_x$ as the total influences that lead to inaccurate measurement of the independent variable, $x$, one might depict this assumption as follows:

$$e_x = 0$$

4. Note that we shall NOT deal with issues of measurement error in Stat 404.
B. Assumption 6: $X$ is not correlated with errors in the measurement of $Y$.

1. Put differently, this assumption suggests that variances in $X$ and $Y$ are not related to any third effect (e.g., time) such that they covary due to this third effect. A depiction of this assumption for a particular independent variable, $x$, would be as follows:

\[
\begin{align*}
0 &= e_Y \\
&\downarrow \\
x &\rightarrow Y
\end{align*}
\]

2. For our purposes the matrix, $Y$, is simply a vector of values for a single dependent variable, $y$. Thus, $e_y$ is simply the error terms (i.e., the $e_i$) from the regression of $y$ on $x$.

3. What does it mean to say that the $x_i$ are uncorrelated with the $e_i$?

a. To answer this question, let’s consider a theory of cognitive development according to which children learn self-confidence by following the example of (or by “modeling”) self-confident parents.

b. You assemble data to test this theory. However, only after collecting your data do you realize (say, after spending a bit more time perusing other relevant literature) that self-confidence is also associated with physical stature. The taller the child, the more self-confident she is. Thus a child’s self-confidence has both psychological and physiological origins. Given that physiology is genetically inherited, we might sketch out the relations among parents’ and child’s stature and self-confidence as follows:
c. Unfortunately, you have no data on parents’ or child’s physiology, so you are left with a causal model of the following form:

\[
\begin{align*}
\text{Parents’ self-confidence} & \rightarrow \text{Child’s physical stature} \\
\text{Parents’ self-confidence} & \rightarrow \text{Child’s self-confidence} \\
\text{Other causes (e}_Y) & \rightarrow \text{Child’s self-confidence (Y)} \\
0 \neq & \\
\text{Parents’ self-confidence (x)} & \rightarrow \text{Other causes (e}_Y)
\end{align*}
\]

\[\text{Parents’ physical stature} \rightarrow \text{Child’s physical stature} \]

d. Since here child’s physical stature is among the “other causes” excluded from the model and since this stature and parents’ self-confidence have a common origin in parents’ physical stature, you could NOT assume that parents’ self-confidence is uncorrelated with the errors from a regression of child’s self-confidence on parents’ self-confidence.

C. Assumption 7: No independent variable (i.e., no column in the design matrix, X) may have all of its variance explained by any subset of the remaining independent variables. The assumption can be put mathematically as follows:
\[ R_{x_j . x_1 \ldots x_{j-1} x_{i+1} \ldots x_k} < 1, \forall i \]

We shall soon discover that when this assumption is not met, your statistics program will be unable to calculate any regression coefficients.