Menopause heralds a complex interplay of hormonal and physiologic changes. Some are temporary discomforts (e.g., hot flashes, sleep disturbances, depression) while others are long-term changes that increase the risk of significant chronic health conditions, bone loss and osteoporosis in particular. Recent observations of an association between depressive symptoms and low bone mineral density (BMD) raise the intriguing possibility that alleviation of depression might confer a risk benefit with respect to bone mineral density loss and osteoporosis. However, the finding of an association in a simple (one predictor) linear regression model analysis has multiple possible explanations, only one of which is causal. Others include, but are not limited to: (1) the apparent association is an artifact of the confounding effects of exercise, body fat, education, smoking, etc; (2) there is no relationship and we have observed a chance event of low probability (it can happen!); (3) the pathway is the other way around (low BMD causes depressive symptoms), albeit highly unlikely; and/or (4) the finding is spurious due to study design flaws (selection bias, misclassification, etc).

In settings where multiple, related predictors are associated with the outcome of interest, multiple predictor linear regression analysis allows us to study the joint relationships among the multiple predictors (depressive symptoms, exercise, body fat, etc) and a single continuous outcome (BMD). In this example, we might be especially interested in using multiple predictor linear regression to isolate the effect of depressive symptoms on BMD, holding all other predictors constant (adjustment). Or, we might want to investigate the possibility of synergism or interaction.
# Table of Contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Learning Objectives</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Simple Linear Regression</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>a. Definition of the Linear Regression Model</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>b. Estimation</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>c. The Analysis of Variance Table</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>d. Assumptions for the Straight Line Regression</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>e. Hypothesis Testing</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>f. Confidence Interval Estimation</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>2. Introduction to Correlation</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>a. Pearson Product Moment Correlation</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>b. Hypothesis Test for Correlation</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>3. Multivariable Regression</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>a. Introduction, Indicator and Design Variables, Interaction Variables</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>b. The Analysis of Variance Table</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>c. The Partial F Test</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>d. Multiple Partial Correlation</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>4. Multivariable Model Development</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>a. Introduction</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>b. Example – Framingham Study</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>c. Suggested Criteria for Confounding and Interaction</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>c. Guidelines for Multivariable Analyses of Large Data Sets</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>5. Goodness-of-Fit and Regression Diagnostics</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>a. Introduction and Terminology</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>b. Assessment of Normality</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>c. Cook-Weisberg Test of Heteroscedasticity</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>d. Method of Fractional Polynomials</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>e. Ramsay Test for Omitted Variables</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>f. Residuals, Leverage, &amp; Cook’s Distance</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>g. Example – Framingham Study</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>
1. Learning Objectives

When you have finished this unit, you should be able to:

- Explain the concepts of association, causation, confounding, mediation, and effect modification;

- Construct and interpret a scatter plot with respect to: evidence of association, assessment of linearity, and the presence of outlying values;

- State the multiple predictor linear regression model and the assumptions necessary for its use;

- Perform and interpret the Shapiro-Wilk and Kolmogorov-Smirnov tests of normality;

- Explain the relevance of the normal probability distribution;

- Explain and interpret the coefficients (and standard error) and analysis of variance tables outputs of a single or multiple predictor regression model estimation;

- Explain and compare crude versus adjusted estimates (betas) of association;

- Explain and interpret regression model estimates of effect modification (interaction);

- Explain and interpret overall and adjusted R-squared measures of association;

- Explain and interpret overall and partial F-tests;

- Draft an analysis plan for a multiple predictor regression model analysis; and

- Explain and interpret selected regression model diagnostics: residuals, leverage, and Cook’s distance.
1. Simple Linear Regression
   
a. Definition of the Linear Regression Model

Simple Linear Regression

A simple linear regression model is a particular model of how the mean $\mu$ (the average value) of one continuous outcome random variable $Y$ (e.g. $Y = $ bone mineral density) varies, depending on the value of a single (usually continuous) predictor variable $X$ (e.g. $X = $ depressive symptoms). Specifically, it says that the average values of the outcome variable, as $X$ changes, lie on a straight line (“regression line”).

For each value of $x$, the values of $y$ are normally distributed around $\mu_{Y|x}$, on the line, with the same variance for all values of $x$, but different means, $\mu_{Y|x}$.

Here, $\sigma^2_{Y|x_1} = \sigma^2_{Y|x_2} = \sigma^2_{Y|x_3} = \sigma^2_{Y|x_4}$

The estimation and hypothesis testing involved are extensions of ideas and techniques that we have already seen. In linear regression,

- we observe an outcome or dependent variable “$Y$” at several levels of the independent or predictor variable “$X$” (there may be more than one predictor “$X$” as seen later).

- A linear regression model assumes that the values of the predictor “$X$” have been fixed in advance of observing “$Y$”.

- However, this is not always the reality. Often “$Y$” and “$X$” are observed jointly and are both random variables.
Correlation

Correlation considers the association of two random variables, Y and X.

- The techniques of estimation and hypothesis testing are the same for linear regression and correlation analyses.
- Exploring the relationship begins with fitting a line to the points.

We develop the linear regression model analysis for a simple example involving one predictor and one outcome.

**Example.**

*Source: Kleinbaum, Kupper, and Muller 1988*

Suppose we have observations of age (days) and weight (ounces) for \(n=11\) chicken embryos. The predictor of interest is \(X=\text{AGE}\). The outcome of interest is weight. For purposes of illustration, suppose we are interested in two models of weight. In one, the outcome variable is \(Y=\text{WT}\). In the other, the outcome is the base 10 logarithm of weight, \(Z=\log_{10}\text{WT}\).

<table>
<thead>
<tr>
<th>WT=Y</th>
<th>AGE=X</th>
<th>LOGWT=Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.029</td>
<td>6</td>
<td>-1.538</td>
</tr>
<tr>
<td>0.052</td>
<td>7</td>
<td>-1.284</td>
</tr>
<tr>
<td>0.079</td>
<td>8</td>
<td>-1.102</td>
</tr>
<tr>
<td>0.125</td>
<td>9</td>
<td>-0.903</td>
</tr>
<tr>
<td>0.181</td>
<td>10</td>
<td>-0.742</td>
</tr>
<tr>
<td>0.261</td>
<td>11</td>
<td>-0.583</td>
</tr>
<tr>
<td>0.425</td>
<td>12</td>
<td>-0.372</td>
</tr>
<tr>
<td>0.738</td>
<td>13</td>
<td>-0.132</td>
</tr>
<tr>
<td>1.13</td>
<td>14</td>
<td>0.053</td>
</tr>
<tr>
<td>1.882</td>
<td>15</td>
<td>0.275</td>
</tr>
<tr>
<td>2.812</td>
<td>16</td>
<td>0.449</td>
</tr>
</tbody>
</table>

**Notation**

- The data are 11 pairs of \((X_1, Y_1)\) where \(X=\text{AGE}\) and \(Y=\text{WT}\)
  \((X_1, Y_1) = (6, 0.029) \cdots (X_{11}, Y_{11}) = (16, 2.812)\) and

- equivalently, 11 pairs of \((X_1, Z_1)\) where \(X=\text{AGE}\) and \(Z=\log_{10}\text{WT}\)
  \((X_1, Z_1) = (6, -1.538) \cdots (X_{11}, Z_{11}) = (16, 0.449)\)
Though simple, it helps to be clear in the research question

- How does weight change with age? Does it change linearly?
- In the language of analysis of variance we are asking the following:
  Can the variability in weight be explained, to a significant extent, by variations in age?
- What is a “good” functional form that relates age to weight?

**Always begin with a scatter plot of the data!** Plot the predictor X on the horizontal and the outcome Y on the vertical. A graph allows you to see things that you cannot see in the numbers alone: range, patterns, outliers, etc. Here, let’s begin with a plot of X=AGE versus Y=WT

![Scatter Plot of WT vs AGE](image)

**What to look for in a scatter plot of the data:**

- The average and median of X
- The range and pattern of variability in X
- The average and median of Y
- The range and pattern of variability in Y
- The nature of the relationship between X and Y
- The strength of the relationship between X and Y
- The identification of any points that might be influential
Example – age (X) and weight (Y) of chicken embryos:

- The plot suggests a relationship between AGE and WT
- A straight line might fit well, but another model might be better
- We have adequate ranges of values for both AGE and WT
- There are no outliers

We might have gotten any of a variety of scatter plots:

No relationship between X and Y

Linear relationship between X and Y
Non-linear relationship between X and Y

Note the arrow pointing to the outlying point
Fit of a linear model will yield
estimated slope that is spuriously
non-zero.

Note the arrow pointing to the outlying point
Fit of a linear model will yield an
estimated slope that is spuriously
near zero.
Note the arrow pointing to the outlying point. Fit of a linear model will yield an estimated slope that is spuriously high.

Example, continuous – age (X) and logweight (Z) of chicken embryos:
The X-Y plot on page 6 is rather “bowl” shaped. Here we consider an X-Z scatter plot. It is much more linear looking, suggesting that perhaps a better model relates the logarithm of WT (Z) to AGE:

We’ll investigate two models.

1) \( WT = \beta_0 + \beta_1 \text{AGE} \)

2) \( \log WT = \beta_0 + \beta_1 \text{AGE} \)
A little review of your high school introduction to straight line relationships

\[ y = \beta_0 + \beta_1 x \]

\( \beta_0 \) = "y-intercept" = value of y when \( x = 0 \)

\( \beta_1 \) = "slope" = \( \frac{\Delta y}{\Delta x} \)

\( \beta_0 \) = “y-intercept” = value of y when \( x = 0 \)

\( \beta_1 \) = “slope” = \( \frac{\Delta y}{\Delta x} \) = (change in y)/(change in x)

<table>
<thead>
<tr>
<th>Slope &gt; 0</th>
<th>Slope = 0</th>
<th>Slope &lt; 0</th>
</tr>
</thead>
</table>
| \[ \Delta y \]
| \[ \Delta x \] |
| \[ \Delta y \]
| \[ \Delta x \] |
### Definition of the Straight Line Model

**Y = β₀ + β₁X**

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = \beta_0 + \beta_1X + \varepsilon )</td>
<td>( Y = \hat{\beta}_0 + \hat{\beta}_1X + e )</td>
</tr>
<tr>
<td>( Y = \beta_0 + \beta_1X ) is the relationship in the population. It is measured with error.</td>
<td>( \hat{\beta}_0, \hat{\beta}_1, ) and ( e ) are our guesses of ( \beta_0, \beta_1 ) and ( \varepsilon )</td>
</tr>
<tr>
<td>( \varepsilon = ) measurement error</td>
<td>( e = ) residual</td>
</tr>
<tr>
<td>We do NOT know the value of ( \beta_0 ) nor ( \beta_1 ) nor ( \varepsilon )</td>
<td>We do have values of ( \hat{\beta}_0, \hat{\beta}_1, ) and ( e )</td>
</tr>
<tr>
<td>The values of ( \hat{\beta}_0, \hat{\beta}_1 ) and ( e ) are obtained by the method of least squares estimation.</td>
<td>To see if ( \hat{\beta}_0 \approx \beta_0 ) and ( \hat{\beta}_1 \approx \beta_1 ) we perform regression diagnostics.</td>
</tr>
</tbody>
</table>

A little notation, sorry!

\( Y = \) the outcome or dependent variable  
\( X = \) the predictor or independent variable

\( \mu_Y = \) The expected value of \( Y \) for all persons in the population  
\( \mu_{Y|X=x} = \) The expected value of \( Y \) for the sub-population for whom \( X=x \)

\( \sigma_Y^2 = \) Variability of \( Y \) among all persons in the population  
\( \sigma_{Y|X=x}^2 = \) Variability of \( Y \) for the sub-population for whom \( X=x \)
b. Estimation

There are a variety of methods for obtaining estimates of $\beta_0$ and $\beta_1$.
In this course, we will consider two of them, maximum likelihood estimation and least squares estimation.

Maximum Likelihood Estimation - This requires use of a probability distribution model. For example, we might assume that the outcome variable Y is distributed normal, with mean values that lie on the regression line. Maximum likelihood estimation chooses estimates of $\beta_0$ and $\beta_1$ that, when applied to the data, gives us the largest value possible for the likelihood of the data that was actually observed.

Least Squares Estimation - NO probability distribution model required here!. Least squares estimation chooses estimates of $\beta_0$ and $\beta_1$ that yield the smallest total of vertical distances (observed to predicted).

When the outcome variable Y is distributed normal,

Maximum Likelihood Estimation = Least Squares Estimation

How Least Squares Estimation works.
Theoretically, we could draw lots of possible lines through the X-Y scatter of the data points. Which one is the “closest”? And what do we mean by “close” anyway? Consider the following:

$Y =$ observed

$\hat{Y} =$ predicted, meaning that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$(Y - \hat{Y})^2 =$ vertical distance between observed outcome and predicted outcome

In least squares estimation, “close” means the following.

♦ We’d like the observed Y and its corresponding prediction $\hat{Y}$ to be as close as possible.

♦ This is the same as wanting $(Y - \hat{Y})^2$ to be as small as possible

♦ It’s not possible to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that it minimizes

$$
(Y_1 - \hat{Y}_1)^2 \quad \text{and minimizes individually}
$$

$$
(Y_2 - \hat{Y}_2)^2 \quad \text{and minimizes individually}
$$


So, instead, we choose values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that makes their total as small as possible

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2$$

How Least Squares Estimation works – a picture.

\[\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X\]

$\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen such that the sum of the squared vertical distances, $\sum_{i=1}^{n}d_i^2$ is minimized.

For each observed value $x_i$, we have an observed $y_i$, and the “predicted” value $\hat{y}_i$, on the line. The vertical distances $d_i = (y_i - \hat{y}_i)$.

The total (a total of squared differences) that we want to minimize has a variety of names.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2$$ is variously called:

- residual sum of squares
- sum of squares about the regression line
- sum of squares due error (SSE)
For the calculus lover, A little calculus yields the solution for the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

- Consider $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2$

- **Step 1:** Differentiate with respect to $\hat{\beta}_1$
  Set derivative equal to 0 and solve.

- **Step 2:** Differentiate with respect to $\hat{\beta}_0$
  Set derivative equal to 0, insert $\hat{\beta}_1$ and solve.

$\beta_1$ is the unknown slope in the population
- Its estimate is denoted $\hat{\beta}_1$ or $b_1$

$\beta_0$ is the unknown intercept in the population
- Its estimate is denoted $\hat{\beta}_0$ or $b_0$

How to use some summation calculations to obtain these estimates
Calculate
- $S_{xx} = \sum X^2 - N\bar{X}^2$
- $S_{yy} = \sum Y^2 - N\bar{Y}^2$
- $S_{xy} = \sum XY - N\bar{X}\bar{Y}$

Review. These expressions make use of a special notation called the “summation notation”.
The capitol “$S$” indicates “summation”.
In $S_{xy}$, the first subscript “$x$” is saying $(x-\bar{x})$.
The second subscript “$y$” is saying $(y-\bar{y})$.

$$S_{xy} = \sum(X-\bar{X})(Y-\bar{Y})$$

Nature Population/ Sample Observation/ Data Relationships/ Modeling Analysis/ Synthesis
## Formulae for Estimated Slope and Intercept

<table>
<thead>
<tr>
<th>Slope</th>
<th>[ \hat{\beta}<em>1 = \frac{\sum</em>{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}} ]</th>
<th>[ \hat{\beta}<em>1 = \frac{S</em>{xy}}{S_{xx}} ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \frac{(n-1)}{(n-1)} = \frac{\text{côv}(X,Y)}{\text{vâr}(X)} ]</td>
<td>[ \hat{\beta}<em>1 = \frac{S</em>{xy}}{S_{xx}} ]</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>[ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} ]</td>
<td>[ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} ]</td>
</tr>
<tr>
<td>Prediction of Y</td>
<td>[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = b_0 + b_1 X ]</td>
<td>[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = b_0 + b_1 X ]</td>
</tr>
</tbody>
</table>

### Do these estimates make sense?

<table>
<thead>
<tr>
<th>Slope</th>
<th>[ \hat{\beta}<em>1 = \frac{\sum</em>{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} ] [ = \frac{\text{côv}(X,Y)}{\text{vâr}(X)} ]</th>
<th>The linear movement in Y with linear movement in X is measured relative to the variability in X.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \hat{\beta}_1 = 0 ] says: With a unit change in X, overall there is a 50-50 chance that Y increases versus decreases.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[ \hat{\beta}_1 \neq 0 ] says: With a unit increase in X, Y increases also (( \hat{\beta}_1 &gt; 0 )) or Y decreases (( \hat{\beta}_1 &lt; 0 )).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>[ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} ]</td>
<td>If the linear model is incorrect, or, if the true model does not have a linear component, we obtain ( \hat{\beta}_1 = 0 ) and ( \hat{\beta}_0 = \bar{Y} ) as our best guess of an unknown Y.</td>
</tr>
</tbody>
</table>
Illustration in Stata

Command.

Notes: (1) In Stata, comments begin with an asterisk; (2) You do not type the leading period.

```
* regress yvariable xvariable
.regress wt age
```

Partial listing of output – annotations in red

|       | Coef.     | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-------|-----------|-----------|-------|-------|----------------------|
| age   | .2350727  | .0459425  | 5.12  | 0.001 | .1311437 .3390018   |
| _cons | -1.884527 | .5258354  | -3.58 | 0.006 | -3.07405 -.695005  |

The fitted line is therefore \( \hat{WT} = -1.88453 + 0.23507 \times \text{AGE} \)

**Interpretation:** (1) \( b_1 = 0.2350727 \) The estimated slope, \( b_1 \), says that, associated with each 1 day increase in age, the estimated average change in weight is a +0.23507 ounces (Hooray – chicken embryos get bigger with advancing age!) that is statistically significant (\( P>|t| = .001 \)). With 95% confidence, the estimated average increase in weight associated with a 1 day increase in age is between +0.1311 ounces and +0.339 ounces; and (2) \( b_0 = -1.884527 \) The estimated intercept, \( b_0 \), says that a 0 day old chicken embryo (\( \text{AGE}=0 \)) is estimated to have an average weight of -1.88453 ounces. Clearly, the latter makes no sense and illustrates the inappropriateness of extending the interpretation of the fitted line beyond the range of the data.

Overlay of Least Squares Line on the X-Y Scatter Plot

As we might have guessed, the straight line model may not be the best choice.

However … it’s worth noting that the “bowl” shape of the scatter plot does have a linear component.

\( \Rightarrow \) So … without the plot, we might have believed the straight line fit is okay.
Illustration of straight line model fit to $Z=$LOGWT versus $X=$AGE.

Command.

```
.* regress yvariable xvariable
.regress logwt age
```

Partial Listing of Output – Annotations in red.

|        | Coef.       | Std. Err.  | t    | P>|t|   | [95% Conf. Interval] |
|--------|-------------|------------|------|-------|---------------------|
| age    | .1958909 = slope = $b_1$ | .0026768  | 73.18| 0.000 | .1898356 .2019462 |
| _cons  | -2.689255 = intercept = $b_0$ | .030637  | -87.78| 0.000  | -2.75856 -2.619949 |

Thus, the fitted line is \( \text{LOGWT} = -2.68925 + 0.19589 \times \text{AGE} \)

**Interpretation:** This fitted line, because it is a model of the logarithm of weight, does not have a straightforward interpretation. The estimated slope, $b_1=0.19589$ says that, associated with each 1 day increase in age, the estimated average change in the logarithm of weight is $a + 0.1958$ that is highly statistically significant ($P>|t| = .0001$ which we often write as $p < .0001$). The estimated intercept, $b_0$, says that a 0 day old chicken embryo ($\text{AGE}=0$) is estimated to have an average logarithm of weight equal to -2.6892. All is not lost. Even though the estimated slope and intercept are not themselves straightforward, we are still able to assess whether an association exists (as opposed to none at all).
Now You Try …

Prediction of Weight from Height
Source: Dixon and Massey (1969)

<table>
<thead>
<tr>
<th>Individual</th>
<th>Height (X), inches</th>
<th>Weight (Y), lbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60</td>
<td>110</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>135</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>120</td>
</tr>
<tr>
<td>4</td>
<td>62</td>
<td>120</td>
</tr>
<tr>
<td>5</td>
<td>62</td>
<td>140</td>
</tr>
<tr>
<td>6</td>
<td>62</td>
<td>130</td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>135</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td>150</td>
</tr>
<tr>
<td>9</td>
<td>64</td>
<td>145</td>
</tr>
<tr>
<td>10</td>
<td>70</td>
<td>170</td>
</tr>
<tr>
<td>11</td>
<td>70</td>
<td>185</td>
</tr>
<tr>
<td>12</td>
<td>70</td>
<td>160</td>
</tr>
</tbody>
</table>

Some preliminary calculations have been done for you

\[
\overline{X} = 63.833 \\
\sum X_i^2 = 49,068 \\
\sum X_i Y_i = 109,380 \\
S_{yy} = 5,266.667 \\
S_{xy} = 863.333 \\
\overline{Y} = 141.667 \\
\sum Y_i^2 = 246,100 \\
S_{xx} = 171.667
\]
The fitted line here is \( \text{WEIGHT} = -179.3573 + 5.0291 \times \text{HEIGHT} \)

**Interpretation**: This fitted line says that associated with each 1 inch increase in height is an estimated average change in weight equal to \( b_1 = 5.02 \) pounds and that an individual with height equal to 0 inches is estimated to have an average weight of \( b_0 = -179.3573 \) pounds. Here, too, we see that the linearity of the relationship does not extend beyond the range of the data.
c. The Analysis of Variance Table

**Question** – Has the fitted line explained a lot? Or, is the amount explained by the fitted line not so much? The analysis of variance table is used to assess the explanatory power of the model just fit. Analysis of variance calculations include sums of squares, degrees of freedom and mean squares.

The **starting point** in an analysis of variance (actually the bottom row of the analysis of variance) is a “**total variability**” that is also called a “**total sum of squares**”. Note – If you look closely at the TSS, you might recognize this as the numerator of the sample variance of the Y’s. It is the sum of all the squared deviations of the individual Y_i about their sample mean. Viewed this way, it can be appreciated as a “total variability”. This is the starting point of the analysis of variance.

**TOTAL SUM OF SQUARES, TSS = \( \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \)**

- The TSS measures the total variability of the observed outcomes Y about their mean (“the average”).
- TSS is thus 100% of what we are trying to explain (the whole pie) using the model just fit.

In a simple linear regression analysis of variance, the TSS is split into just two components:

\[
TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]

The **MODEL SUM OF SQUARES** is the portion (wedge of the pie) of the variability in outcome that **is explained** by the model just fit,

\[
\text{MSS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2
\]

while the **RESIDUAL SUM OF SQUARES** is the portion (the remainder of the pie) of the variability in outcome that remains as leftover as **unexplained** by the model just fit.

\[
\text{RSS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]
Question, again – Has the fitted line explained a lot? Or, is the amount explained by the fitted line not so much? A picture helps here. If our fitted model explains a lot, then the amount of variability that is left over and is still unexplained will be small. Here is a picture of that. The TSS is the total variability and represents the totality of what we want to explain (schematically represented by the large double blue arrow at left); it is the overall variability of the individual data points. The fitted line is the red line. The RSS is the variability of the individual data points about the fitted line. It is thus, the left over variability (the “residual” variability) that remains unexplained (schematically represented by the smaller double blue arrow at right). This picture illustrates the setting where the fitted line has explained a lot!
Yes. The total sum of squares (TSS) is, conveniently, the sum of two portions, a residual sum of squares (RSS) + a regression sum of squares (MSS). Here is the partition (Note – Look closely and you’ll see that both sides are the same because I have added and subtracted the predicted Y)

\[
(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})
\]

Some algebra (not shown) confirms the partition of the total sum of squares into its two components.

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

Total sum of squares Residual sum of squares Model sum of squares

**Introduction to R^2, also called Coefficient of Determination.** Consider what it is we want to explain is the variability in the outcome Y, namely TSS = total sum of squares. The portion of the TSS that is explained is the 2\textsuperscript{nd} term in the expression above, namely MSS = model sum of squares. With these, we can now compute the proportion of total variance in Y that is explained by the model:

\[
R^2 = \frac{MSS}{TSS} = \frac{\text{Model sum of squares}}{\text{Total sum of squares}}\text{ an easy proportion!}
\]

<table>
<thead>
<tr>
<th>Fitted line explains a lot</th>
<th>Fitted line does NOT explain a lot</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TSS = same fixed total →</strong></td>
<td><strong>TSS = same fixed total →</strong></td>
</tr>
<tr>
<td>If RSS (residual) is small</td>
<td>If RSS (residual) is big</td>
</tr>
<tr>
<td>Then MSS (model) is relatively big →</td>
<td>Then MSS (model) is relatively small →</td>
</tr>
</tbody>
</table>
| \[
R^2 = \frac{\text{Model ssq}}{\text{Total ssq}} = \text{large proportion}
\]| \[
R^2 = \frac{\text{Model ssq}}{\text{Total ssq}} = \text{small proportion}
\]
We’re not quite done. Tests of statistical significance in analysis of variance make use of mean squares, not sums of squares. To obtain mean squares, division by degrees of freedom are required.

**Introduction to degrees of freedom (df).** Each of the three sums of squares (TSS, RSS, and MSS) is a calculation that utilizes every data point, all “n” of them. They differ, however, in the constraints that were also utilized. Tip! – Every time a constraint is placed on the data, a degree of freedom is lost.

- To start with, the data are a random sample of **mutually independent** outcomes, sample size = n
  The key here is “mutually independent”, because it means “free to vary”
  Thus, to start with, and before any constraints, **degrees of freedom = sample size = n**.

- **TSS:** “1 degree of freedom is lost because there is 1 constraint on the data”
  In computing the **total sum of squares**, squared deviations are measured about the sample mean $\bar{Y}$. There is 1 constraint on the data in fixing $\bar{Y}$. Thus,
  TSS degrees of freedom = (n-1)

- **RSS:** “2 degrees of freedom are lost because there are 2 constraints on the data”
  In computing the **residual sum of squares**, squared deviations are measured about the predicted values $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Now there are 2 constraints on the data, one for fixing $\hat{\beta}_0$ and the second for fixing $\hat{\beta}_1$. Thus,
  RSS degrees of freedom, simple linear regression = (n-2)

- **MSS:** Tip – Now we have to think about this as follows: “count one degree of freedom for each regression parameter AFTER the intercept”
  In simple linear regression there are two regression **model** parameters, one for the slope and one for the intercept.
  Thus, after the intercept, there is just the regression parameter and it is for the slope.
  MSS degrees of freedom = (1)
**Introduction to mean squares.** A sum of squares by itself is not a variance estimate because it is a measure of all the variability; eg “all the variability about the mean (TSS)” or “all the variability of the model about the mean (MSS)” or “all the variability of the observations about their associated predicted values (RSS)” . Instead, mean squares are variance estimates. They are defined:

\[
\text{mean square} = \text{variance estimate} = \frac{\text{sum of squares}}{\text{degrees of freedom}}
\]

The analysis of variance in simple linear regression compares the two variance estimates, “due model” versus “due residual” to assess the amount of variability in the outcomes \(Y_1 \ldots Y_n\) that is explained by the model just fit.

<table>
<thead>
<tr>
<th>The relationship between (X) and (Y) has a <strong>linear component</strong> with a non-zero slope: (\beta_1 \neq 0)</th>
<th>The relationship between (X) and (Y) (if any) has <strong>no linear component.</strong> (\beta_1 = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A good prediction of (Y) is the fitted line: (\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X)</td>
<td>A better prediction of (Y) is the average of the (Y)’s: (\hat{Y} = \hat{\beta}_0 = \bar{Y})</td>
</tr>
<tr>
<td>Consider the “<strong>due model</strong>” deviations: ((\hat{Y} - \bar{Y}) = (\hat{\beta}_0 + \hat{\beta}_1 X - \bar{Y})) (= \bar{Y} - \hat{\beta}_0 + \hat{\beta}_1 X - \bar{Y}) (= \hat{\beta}_1 (X - \bar{X}))</td>
<td>Here, consider the due “<strong>due residual</strong>” deviations: ((Y - \hat{Y}) = (Y - \hat{\beta}_0) = (Y - \bar{Y}))</td>
</tr>
<tr>
<td>A straight line relationship is helpful. (\text{MSS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2) is relatively large</td>
<td>A straight line relationship is not helpful (\text{MSS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2) is relatively small</td>
</tr>
<tr>
<td>(\text{RSS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2) is relatively small</td>
<td>(\text{RSS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2) is relatively large</td>
</tr>
<tr>
<td>(\frac{\text{MODEL mean square}}{\text{RESIDUAL mean square}}) will be large</td>
<td>(\frac{\text{MODEL mean square}}{\text{RESIDUAL mean square}}) will be small (close to 1)</td>
</tr>
</tbody>
</table>

| Nature | Population/ Sample | Observation/ Data | Relationships/ Modeling | Analysis/ Synthesis |
Summary of Analysis of Variance Terms

1. **TSS:** The “total” or “total, corrected” refers to the variability of $Y$ about $\bar{Y}$
   
   - sum of squares: $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$ is called the “total sum of squares”
   - degrees of freedom: $df = (n-1)$
   - mean square = (sum of squares)/(df): “total mean square”

2. **RSS:** The “residual” or “due error” refers to the variability of $Y$ about $\hat{Y}$
   
   - sum of squares: $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ is called the “residual sum of squares”
   - degrees of freedom: $df = (n-2)$
   - mean square = (sum of squares)/(df): “residual mean square”.

3. **MSS:** The “model” or “due regression” refers to the variability of $\hat{Y}$ about $\bar{Y}$
   
   - sum of squares: $\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})^2$ is called the “regression sum of squares”
   - degrees of freedom: $df = 1$
   - mean square = (sum of squares)/(df): “regression mean square” or “model mean square”.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>MSS = $\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$</td>
<td>MSS/1 = SS(model) /1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$SS(\text{model}) = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>(n-2)</td>
<td>RSS = $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$</td>
<td>RSS/(n-2) = SS(residual)/(n-2)</td>
</tr>
<tr>
<td>Total, corrected</td>
<td>(n-1)</td>
<td>TSS = $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$</td>
<td></td>
</tr>
</tbody>
</table>

*Hint – The entry in the “mean square” column is always the sum of squares divided by the degrees of freedom*
Be careful!

Analysis of variance answers a **limited** question.

Does the fit of the straight line model explain a significant portion of the variability of the individual \( Y \) about \( \bar{Y} \)? Is this better than using \( \bar{Y} \) alone?

In simple linear regression analysis of variance, the 2 models being compared are:

\[
Y = \beta_0 + \beta_1 x + \text{error} \quad \text{versus} \quad Y = \beta_0 + \text{error}
\]

**Analysis of Variance does **NOT** address:**
- Is the choice of the straight line model correct?
- Would another functional form be a better choice?

**Illustration in Stata**

**Command.**

```
* regress yvariable xvariable
.regress logwt age
```

**Partial listing of output (now I’m showing you the analysis of variance portion)**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4.22105734 = MSS/1</td>
<td>1</td>
<td>4.22105734 = MSS/1</td>
<td>F( 1, 9) = 5355.60 = MS (Model)/MS (Residual)</td>
</tr>
<tr>
<td>Residual</td>
<td>.007093416 = RSS/9</td>
<td>9</td>
<td>.000788157 = RSS/9</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>4.22815076 = RSS/10</td>
<td>10</td>
<td>.422815076 = RSS/10</td>
<td>R-squared = 0.9983 = SS(Model)/SS(total)</td>
</tr>
</tbody>
</table>

**Adj R-squared = 0.9981 = R^2 adjusted for n, #predictors**

Root MSE = .02807 = \( \sqrt{\text{MS(Residual)}} \)
d. Assumptions for a Straight Line Regression Analysis

See again, page 12. Least squares estimation does not require a probability model. However, if we want to do hypothesis tests or confidence interval estimation or both, then we do need a probability model.

**Assumptions of Simple Linear Regression**

1. The outcomes \( Y_1, Y_2, \ldots, Y_n \) are independent.

2. The values of the predictor variable \( X \) are fixed and measured without error.

3. At each value of the predictor variable \( X=x \), the distribution of the outcome \( Y \) is normal with mean \( \mu_{Y|X=x} = \beta_0 + \beta_1 X \) and common variance \( \sigma^2_{Y|x} \).

Assumptions 1-3 also mean that, for each individual “i”,

\[
Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{where}
\]

\( \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \) are independent and identically distributed Normal with

- mean \( \mu_\varepsilon = 0 \) and
- variance \( \sigma^2_\varepsilon = \sigma^2_{Y|x} \).

For each value of \( x \), the values of \( y \) are normally distributed around \( \mu_{Y|x} \) on the line, with the same variance for all values of \( x \), but different means, \( \mu_{Y|x} \).

Here, \( \sigma^2_{Y|x_1} = \sigma^2_{Y|x_2} = \sigma^2_{Y|x_3} = \sigma^2_{Y|x_4} \).
With these assumptions, the comparison of the “due model” versus “due residual” variance estimates is an F-statistic under the null hypothesis of zero slope.

\[ F = \frac{\text{mean square (due model)}}{\text{mean square (due residual)}} \quad \text{with df} = 1, (n-2) | \text{null true.} \]

<table>
<thead>
<tr>
<th>Null Hypothesis true</th>
<th>Null Hypothesis not true</th>
</tr>
</thead>
<tbody>
<tr>
<td>the model does NOT explain a lot</td>
<td>the model explains a lot</td>
</tr>
<tr>
<td>( \beta_1 = 0 )</td>
<td>( \beta_1 \neq 0 )</td>
</tr>
</tbody>
</table>

Due model mean square has expected value
\[ \sigma_{Y|X}^2 \]
Due residual mean square, MS(residual), has expected value
\[ \sigma_{Y|X}^2 \]
\[ F = \text{MS(model)/MS(residual)} \] will be close to 1

Due model means square has expected value
\[ \sigma_{Y|X}^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \]
Due residual mean square, MS(residual), has expected value
\[ \sigma_{Y|X}^2 \]
\[ F = \text{MS(model)/MS(residual)} \] will be LARGER than 1

Illustration in Stata for the model of \( Y = \text{LOGWT} \) to \( X = \text{AGE} \):

```
.* regress yvariable xvariable
.regress logwt age
```

Output (another partial listing) - Annotations in red.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4.22105734</td>
<td>1</td>
<td>4.22105734</td>
<td>F( 1, 9) = 5355.60 = MS (Model)/MS (Residual)</td>
</tr>
<tr>
<td>Residual</td>
<td>.007093416</td>
<td>9</td>
<td>.000788157</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>4.22815076</td>
<td>10</td>
<td>.422815076</td>
<td>R-squared = 0.9983 = SS(Model)/SS(total)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.9981 = ( R^2 ) adjusted for n and # predictors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = .02807 = ( \sqrt{MS(Residual)} )</td>
</tr>
</tbody>
</table>
This output corresponds to the following.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due model</td>
<td>1</td>
<td>( \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}<em>1 \sum</em>{i=1}^{n} (X_i - \bar{X})^2 )</td>
<td>MSS/1 = 4.22106</td>
</tr>
<tr>
<td>Due residual</td>
<td>(n-2) = 9</td>
<td>( \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 0.00709 )</td>
<td>RSS/(n-2) = 0.00078816</td>
</tr>
<tr>
<td>Total, corrected</td>
<td>(n-1) = 10</td>
<td>( \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 4.22815 )</td>
<td></td>
</tr>
</tbody>
</table>

Stata provides other information, too:

- **R-SQUARED** = (MSS)/(TSS) = (Model sum of squares) / (Total sum of squares)
  See again page 22. \( R^2 \) = coefficient of determination is a proportion, the percent of the total sum of squares TSS that is explained by the fit of the current model (in this case, the straight line model).

- **Looking ahead …** We will see that, as predictors are added to the model, R-SQUARED can only increase. Eventually, we need to “adjust” this measure to take this into account. We’ll then learn about the ADJUSTED R-SQUARED.

- **F( 1, 9) = \left[ \text{mean square(model)} \right] / \left[ \text{mean square(residual)} \right]**
  This is the overall F test introduced on page 26.
  \[
  = \left[ \frac{4.22106}{0.00078816} \right] 
  = 5355.60 \text{ with df = 1, 9}
  \]

- **Prob > F = achieved significance level (p-value)**
  This is the result of the p-value calculation for the F test. Thus, it is the probability of an F statistic value as extreme or more extreme than the value attained for the observed data under the null hypothesis assumption. In this example, assumption of the null hypothesis model (\( Y = \beta_0 + \text{error} \)) has led to an extremely unlikely result (p-value < 0.0001), prompting rejection of the null hypothesis \( H_0 \). We conclude that the fitted line (\( Y = \beta_0 + \beta_1 X + \text{error} \)) is a statistically significant improvement.

- **Root MSE = \sqrt{ \text{mean square (residual)} }**
  This is used in many hypothesis test and confidence interval calculations.
e. Hypothesis Testing

Simple Linear Regression Model: \( Y = \beta_0 + \beta_1 X \)

When the linear regression is simple (meaning that there is just one predictor \( X \)), we might ask:

1. Relative to the null hypothesis model with no predictors, does the fit of a straight line model in \( X \) explain statistically significantly more of the variability in outcomes? It turns out that there are 3 equivalent procedures for addressing this question:

   Overall F-test
   t-test of zero slope
   t-test of zero correlation

   \( \text{Tip! These are all equivalent!} \)

2. Given the fit of a straight line relationship, is the intercept statistically significantly different from zero; that is, does the line pass through the origin?

   t-test of zero intercept

**Overall F-Test**

**Research Question:** Does the fitted regression model explain statistically significantly more of the variability among the outcomes \( Y \) than is explained by the average of the \( Y \)’s?

**Assumptions:** As before (see page 26).

**\( H_O \) and \( H_A \):**

\[
\begin{align*}
H_O : & \quad \beta_1 = 0 \\
H_A : & \quad \beta_1 \neq 0
\end{align*}
\]

**Test Statistic:**

\[
F = \frac{\text{mean square(model)}}{\text{mean square(residual)}}
\]

\( \text{df} = 1, (n-2) \)

**Evaluation rule:**

Under the null hypothesis, F-statistic values will be close to 1.
Under the alternative hypothesis, \( \beta_1 \neq 0 \), F-statistic values will tend to be larger than 1.

Thus, our p-value calculation answers: “What are the chances of obtaining our value of the F or one that is larger if we believe the null hypothesis that \( \beta_1 = 0 \)?”
Calculations:

For our data, we obtain p-value =

\[
\Pr \left( \frac{F_{1,(n-2)} \geq \frac{\text{mean square(model)}}{\text{mean square(residual)}}}{\beta_1 = 0} \right) = \Pr \left( F_{1,9} \geq 5355.60 \right) < .0001
\]

Evaluate:
Under the null hypothesis that \( \beta_1 = 0 \), the chances of obtaining an F-statistic value as (or more) extreme as 5355.60 were less than 1 chance in 10,000. This is a very small likelihood! \( \Rightarrow \) Statistical rejection.

Interpret:
The fitted straight line model explains statistically significantly more of the variability in Y=LOGWT than is explained by the average of LOGWT alone

... Stay tuned. Later, we'll see that the analysis does not stop here …
**T-test of Zero Slope**

Preliminaries: (1) The overall F test and the test of the slope are equivalent; (2) the test of the slope uses a t-score approach to hypothesis testing; and (3) it can be shown that \( t \text{-score for slope} \) \( = \) \( \sqrt{\text{overall F}} \)

**Research Question:** Is the slope \( \beta_1 = 0 \)?

**Assumptions:** As before.

**H₀ and Hₐ:**

\[
H_0: \quad \beta_1 = 0 \\
H_{A}: \quad \beta_1 \neq 0
\]

**Test Statistic:**

To compute the t-score, we need an estimate of the standard error of \( \hat{\beta}_1 \)

\[
\text{S}\hat{E}(\hat{\beta}_1) = \sqrt{\text{mean square(residual)}} \left[ \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right]
\]

Our t-score is therefore:

\[
t\text{-score} = \left[ \frac{(\text{observed}) - (\text{expected})}{\text{s}\hat{e}(\text{observed})} \right] = \left[ \frac{(\hat{\beta}_1) - (0)}{\text{s}\hat{e}(\hat{\beta}_1)} \right]
\]

\[\text{df} = (n-2)\]
Illustration in Stata

| logwt  | Coef.   | Std. Err. | t   | P>|t| | [95% Conf. Interval] |
|--------|---------|-----------|-----|------|---------------------|
| age    | 0.19589 | 0.00268   | 73.18 | 0.000 | 0.1898356 - 0.2019462 |
| _cons  | -2.689255 | 0.030637 | -87.78 | 0.000 | -2.75856 - 2.619949 |

Review - Recall what we mean by a t-score:

T=73.18 says “the estimated slope is estimated to be 73.18 standard error units away from its null hypothesis model expected value of zero”.

Check that \( t \)-score \( \chi^2 = \chi \text{ Overall F} \):

\[ 73.18^2 = 5355.31 \text{ which is close.} \]

Evaluation rule:

The p-value calculation answers: “Assuming null hypothesis model (\( \beta_1 = 0 \)), what were the chances of obtaining an estimated slope (0.1959) that is as extreme as 73.18 standard error units away (in either direction!) from its expected value of 0?

Calculations:

For our data, we obtain the two sided p-value =

\[
2 \text{pr} \left[ t_{n-2} \geq \left| \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} \right| \right] = 2 \text{pr} \left[ t_9 \geq 73.18 \right] \ll .0001
\]

Evaluate:

Under the null hypothesis that \( \beta_1 = 0 \), the chances of obtaining a t-statistic value as (or more) extreme as 73.18 were less than 1 chance in 10,000. This is a very small likelihood! → Statistical rejection.

Interpret:

The interpretation is the same as for the overall F-test.
T-test of Zero Intercept

*Tip:* This is rarely of interest

**Research Question:** Is the intercept $\beta_0 = 0$?

**Assumptions:** As before.

**$H_0$ and $H_A$:**

- $H_0$: $\beta_0 = 0$
- $H_A$: $\beta_0 \neq 0$

**Test Statistic:**

To compute the t-score for the intercept, we need an estimate of the standard error of $\hat{\beta}_0$

$$SE(\hat{\beta}_0) = \sqrt{\frac{\text{mean square(residual)}}{n}} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Our t-score is therefore:

$$t = \frac{(\text{observed})-(\text{expected})}{\hat{\text{se}}(\text{observed})} = \frac{(\hat{\beta}_0)-(0)}{\text{se}(\hat{\beta}_0)}$$

$$df=(n-2)$$

**Illustration in Stata**

| logwt | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|-------|-----------|-------|------|----------------------|
| age   | -.1958909 | .0026768 | 73.18 | 0.000 | .1898356 -.2019462 |
| _cons | -.2.689255 | .030637 | -87.78 | 0.000 | -2.75856 -2.619949 |

$$t = \frac{(\text{Coef.})}{(\text{Std. Err.})}$$

$$t = \frac{-.1958909}{.0026768} = -87.78$$

**Evaluation rule:**

The p-value calculation answers: “Assuming null hypothesis model ($\beta_0 = 0$), what were the chances of obtaining an estimated intercept (-2.6893) that is as extreme as 87.78 standard error units away (in either direction!) from its expected value of 0?”
Calculations:
For these data the two sided p-value =

\[
2 \text{pr} \left[ t_{(n-2)} \geq \frac{\hat{\beta}_0 - 0}{\hat{\text{se}}(\hat{\beta}_0)} \right] = 2 \text{pr} \left[ t_9 \geq 87.78 \right] \ll .0001
\]

Evaluate:
Under the null hypothesis that \( \beta_0 = 0 \), the chances of obtaining a t-statistic value as (or more) extreme as 87.78 were less than 1 chance in 10,000. This is a very small likelihood! → Statistical rejection.

Interpret:
Conclude that the intercept is statistically significantly different from zero or, equivalently, that the straight line relationship does not pass through the origin.
f. Confidence Interval Estimation

Simple Linear Regression Model: \( Y = \beta_0 + \beta_1 X \)

Confidence intervals are helpful in providing information about the range of possible parameter values that are consistent with the observed data. A simple linear regression analysis might include confidence interval estimation of four parameters: (1) slope: \( \beta_1 \); (2) intercept: \( \beta_0 \); (3) mean of population for whom \( X=x_0 \): \( \beta_0 + \beta_1 x_0 \) and (4) predicted response for an individual with \( X=x \): \( \beta_0 + \beta_1 x \).

In all instances, the confidence coefficient is a percentile of the student t-distribution with \( df = (n-2) \).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>SE (Estimate)</th>
<th>Confidence Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope: ( \beta_1 )</td>
<td>( \hat{\beta}_1 )</td>
<td>( \sqrt{\frac{\text{mean square(residual)}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} )</td>
<td>Percentile</td>
</tr>
<tr>
<td>Intercept: ( \beta_0 )</td>
<td>( \hat{\beta}_0 )</td>
<td>( \sqrt{\frac{\text{mean square(residual)}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} )</td>
<td>Percentile</td>
</tr>
<tr>
<td>Mean: ( \beta_0 + \beta_1 x_0 )</td>
<td>( \hat{Y}_{x=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 )</td>
<td>( \sqrt{\frac{\text{mean square(residual)}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} )</td>
<td>Percentile</td>
</tr>
<tr>
<td>Prediction: ( \beta_0 + \beta_1 x_0 )</td>
<td>( \hat{Y}_{x=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 )</td>
<td>( \sqrt{\frac{\text{mean square(residual)}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} )</td>
<td>Percentile</td>
</tr>
</tbody>
</table>

*Review of PubHlth 540!* (1) for a 95% CI, the correct percentile is the 97.\text{th} percentile; and more generally (2) for a \((1-\alpha)100\%\) CI, the correct percentile is the \((1-\alpha/2)100\%\) percentile.
Stata illustration for the model which fits \( Y = \text{LOGWT} \) to \( X = \text{AGE} \).
How nice – Stata pretty much gives it to you!

|      | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|------|-------|-----------|------|-----|----------------------|
| age  | 0.1958909 | 0.0026768 | 73.18 | 0.000 | 0.1898356 - 0.2019462 |
| _cons | -2.689255 | 0.030637 | -87.78 | 0.000 | -2.75856 - -2.619949 |

**95% Confidence Interval for the Slope, \( \beta_1 \)**
1) Estimate = \( \hat{\beta}_1 = 0.19589 \)
2) SE (Estimate) = \( se(\hat{\beta}_1) = 0.00268 \)
3) Confidence coefficient = 97.5\(^{th}\) percentile of Student \( t = t_{0.975, df=9} = 2.26 \)
95% Confidence Interval for Slope \( \beta_1 \) = Estimate ± (confidence coefficient)*SE
\[
= 0.19589 ± (2.26)(0.00268)
\]
\[
= (0.1898, 0.2019)
\]

**95% Confidence Interval for the Intercept, \( \beta_0 \)**
1) Estimate = \( \hat{\beta}_0 = -2.68925 \)
2) SE (Estimate) = \( se(\hat{\beta}_0) = 0.03064 \)
3) Confidence coefficient = 97.5\(^{th}\) percentile of Student \( t = t_{0.975, df=9} = 2.26 \)
95% Confidence Interval for Intercept \( \beta_0 \) = Estimate ± (confidence coefficient)*SE
\[
= -2.68925 ± (2.26)(0.03064)
\]
\[
= (-2.7585, -2.6200)
\]
Confidence Intervals for Predictions

Stata code. Green=comment, black = command, blue=output

```
. * Confidence Intervals for Fit of Y=LOGWT to X=AGE
. * Obtain conf coeff as 97.5th percentile of Student t w df=9
. display invttail(9,.025)
2.2621572

. regress logwt age
<output not shown>

. * Obtain predicted values yhat
. predict yhat, xb

. * Obtain se for predicted invidual sei
. predict sei, stdf

. * Obtain se for predicted mean semean
. predict semean, stdp

. * 95% Confidence Intervals for Individual Predictions
. generate cllowi = yhat - (2.2621572*sei)
. generate cluppi = yhat + (2.2621572*sei)
. list logwt yhat cllowi cluppi
<output shown below>

. * 95% Confidence Intervals for Mean Predictions
. generate cllowm = yhat - (2.2621572*semean)
. generate cluppm = yhat + (2.2621572*semean)
. list logwt yhat cllowm cluppm
<output shown below>
```

<table>
<thead>
<tr>
<th>logwt</th>
<th>yhat</th>
<th>cllowi</th>
<th>cluppi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1538</td>
<td>-1.513909</td>
<td>-1.586824</td>
<td>-1.440994</td>
</tr>
<tr>
<td>1.1284</td>
<td>-1.318018</td>
<td>-1.388634</td>
<td>-1.247402</td>
</tr>
<tr>
<td>1.102</td>
<td>-1.122127</td>
<td>-1.190902</td>
<td>-1.053353</td>
</tr>
<tr>
<td>.903</td>
<td>-.9262364</td>
<td>-.9936649</td>
<td>-.8588079</td>
</tr>
<tr>
<td>.742</td>
<td>-.7303454</td>
<td>-.7969533</td>
<td>-.6637375</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>logwt</th>
<th>yhat</th>
<th>cllowm</th>
<th>cluppm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1538</td>
<td>-1.513909</td>
<td>-1.549733</td>
<td>-1.478086</td>
</tr>
<tr>
<td>1.1284</td>
<td>-1.318018</td>
<td>-1.348894</td>
<td>-1.287142</td>
</tr>
<tr>
<td>1.102</td>
<td>-1.122127</td>
<td>-1.148522</td>
<td>-1.095733</td>
</tr>
<tr>
<td>.903</td>
<td>-.9262364</td>
<td>-.9488931</td>
<td>-.9035797</td>
</tr>
<tr>
<td>.742</td>
<td>-.7303454</td>
<td>-.7504284</td>
<td>-.7102624</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>logwt</th>
<th>yhat</th>
<th>cllowm</th>
<th>cluppm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1538</td>
<td>-1.513909</td>
<td>-1.5344545</td>
<td>-1.5536029</td>
</tr>
<tr>
<td>1.1284</td>
<td>-1.318018</td>
<td>-.3385637</td>
<td>-.3586467</td>
</tr>
<tr>
<td>1.102</td>
<td>-1.122127</td>
<td>-.2010103</td>
<td>-.1653294</td>
</tr>
<tr>
<td>.903</td>
<td>-.9262364</td>
<td>-.05532182</td>
<td>-.0268239</td>
</tr>
<tr>
<td>.742</td>
<td>-.7303454</td>
<td>-.7491091</td>
<td>-.7182332</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>logwt</th>
<th>yhat</th>
<th>cllowm</th>
<th>cluppm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1538</td>
<td>-1.513909</td>
<td>.449</td>
<td>.445</td>
</tr>
<tr>
<td>1.1284</td>
<td>-1.318018</td>
<td>.4552182</td>
<td>.4091766</td>
</tr>
</tbody>
</table>
2. Introduction to Correlation

a. Pearson Product Moment Correlation

Quick review – Recall the distinction between “linear” and “monotone.” A positive linear association is a special case of a monotone increasing association.

<table>
<thead>
<tr>
<th>Linear</th>
<th>Monotone increasing</th>
<th>Monotone increasing</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Linear" /></td>
<td><img src="image2.png" alt="Monotone increasing" /></td>
<td><img src="image3.png" alt="Monotone increasing" /></td>
</tr>
</tbody>
</table>

What is a correlation coefficient?
A correlation coefficient is a measure of the association between two paired random variables (e.g. height and weight).

The **Pearson product moment correlation**, in particular, is a measure of the strength of the **straight line** relationship between the two random variables (left panel in picture).

The **Spearman** correlation is a measure of the strength of the **monotone increasing (or decreasing)** relationship between the two random variables (all panels in picture).

**Formula for the Pearson Product Moment Correlation ρ**
- The population parameter designation is rho, written as \( \rho \)
- The estimate of \( \rho \), based on information in a sample is represented using \( r \).
- Some preliminaries:
  1. Suppose we are interested in the correlation between X and Y
  2. \[ \text{cov}(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)} = \frac{S_{xy}}{(n-1)} \]
    This is the covariance(X,Y)
  3. \[ \text{var}(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)} = \frac{S_{xx}}{(n-1)} \]
    and similarly
  4. \[ \text{var}(Y) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{(n-1)} = \frac{S_{yy}}{(n-1)} \]
Sample Pearson Product Moment Correlation

\[
\hat{\rho} = r = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}
\]

Tip! If you absolutely have to do it by hand, an equivalent (more calculator friendly formula) is

\[
\hat{\rho} = r = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right) \left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}}
\]

- The correlation \( r \) can take on values between 0 and 1 only
- Thus, the correlation coefficient is said to be dimensionless – it is independent of the units of \( x \) or \( y \).
- Sign of the correlation coefficient (positive or negative) = Sign of the estimated slope \( \hat{\beta}_1 \).
### Relationship between slope $\hat{\beta}_1$ and the sample correlation $r$

Because 

$$ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} $$

A little algebra reveals the following interrelationships:

$$ r = \left[ \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} \right] \hat{\beta}_1 \quad \text{and} \quad \hat{\beta}_1 = \left[ \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} \right] r $$

**Thus, beware!!!**

- It is possible to have a very large (positive or negative) $r$ might accompanying a very non-zero slope, inasmuch as
  - A very large $r$ might reflect a very large $S_{xx}$, all other things equal
  - A very large $r$ might reflect a very small $S_{yy}$, all other things equal.
b. Hypothesis Test of Zero Correlation

Recall (see page 29) -
The null hypothesis of zero correlation $\rho = 0$ is equivalent to the null hypothesis of zero slope.

**Research Question:** Is the correlation $\rho = 0$? Is the slope $\beta_1 = 0$?

**Assumptions:** As before. See page 25.

$H_0$ and $H_A$:

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

**Test Statistic:**
A little algebra (not shown) yields a very nice formula for the $t$-score that we need.

\[
t - \text{score} = \left[ \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \right] \\
df = (n-2)
\]

We can find this information in our output. Recall the first example and the model of $Y=$LOGWT to $X=$AGE:

**Stata illustration for the model which fits $Y=$LOGWT to $X=$AGE.**

**Tip!** The Pearson Correlation, $r$, is the $\sqrt{R\text{-squared}}$ in the output.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4.22105734</td>
<td>1</td>
<td>4.22105734</td>
<td>F( 1, 9) = 5355.60</td>
</tr>
<tr>
<td>Residual</td>
<td>0.007093416</td>
<td>9</td>
<td>0.000788157</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>4.22815076</td>
<td>10</td>
<td>.422815076</td>
<td>R-squared = 0.9983</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.9981</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 0.02807</td>
</tr>
</tbody>
</table>

Pearson Correlation, $r = \sqrt{0.9983} = 0.9991$
Substitution into the formula for the t-score yields

\[ t - score = \left[ \frac{r \sqrt{(n-2)}}{\sqrt{1-r^2}} \right] = \left[ \frac{.9991\sqrt{9}}{\sqrt{1-.9983}} \right] = \left[ \frac{2.9974}{.0412} \right] = 72.69 \]

Note: The value .9991 in the numerator is \( r = \sqrt{R^2} = \sqrt{.9983} = .9991 \)

This is very close to the value of the t-score (73.18) that was obtained for testing the null hypothesis of zero slope. The discrepancy is probably rounding error. I did the calculations on my calculator using 4 significant digits. Stata probably used more significant digits - cb.
3. Multivariable Linear Regression

a. Definition, Indicator Variables and Design Variables, Interaction Variables

It is possible to consider multiple predictors in a linear regression model and these can be any mix of continuous or discrete. There is still one outcome variable $Y$ that is continuous and assumed distributed normal.

**When there are multiple predictors, the least squares fit is multi-dimensional.** In the setting of just 2 predictors, it’s possible to show a schematic of the fitted plane that results from least squares estimation. In this illustration, the outcome is $Y=\text{body length}$ and there are two predictors: $X_1=\text{glabella length}$ and $X_2=\text{glabella width}$. The purple ellipse is the least squares fit and is a 2-dimensional plane in 3-dimensional space; it is the analogue of the straight line fit that was explained in simple linear regression.

Source: www.palass.org

**Definition**

A multiple linear regression model is a particular model of how the mean $\mu$ (the average value) of one continuous outcome random variable $Y$ (e.g. $Y=\text{length of hospital stay}$) varies, depending on the value of two or more (these can be a mixture of continuous and discrete) predictor variables $X$ (e.g. $X_1=\text{age}$, $X_2=0/1 \text{ history of vertebral fractures, etc.}$) Specifically, it says that the average values of the outcome variable, as the profiles of predictors $X_1$, $X_2$, … etc change, lie on a “plane” (“regression plane”).

| Nature | Population/ Sample | Observation/ Data | Relationships/ Modeling | Analysis/ Synthesis |
Example
P53 is a tumor suppressor gene that has been extensively studied in breast cancer research. Suppose we are interested in understanding the correlates of p53 expression, especially those that are known breast cancer risk variables. We might hypothesize that p53 expression is related to, among other things, number of pregnancies and age at first pregnancy.

\[ Y = \text{p53 expression level} \]
\[ X_1 = \text{number of pregnancies (coded 0, 1, 2, etc)} \]
\[ X_2 = \text{age at first pregnancy} \leq 24 \text{ years} \ (1=\text{yes}, \ 0=\text{no}) \]
\[ X_3 = \text{age at first pregnancy} > 24 \text{ years} \ (1=\text{yes}, \ 0=\text{no}) \]

A multivariable linear model that relates \( Y \) to \( X_1 \) and \( X_2 \) is the following

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \text{error} \]

The General Multivariable Linear Model
Similarly, it is possible to consider a multivariable model that includes \( p \) predictors:

\[ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \text{error} \]

- \( p = \# \) predictors, \textit{apart} from the intercept
- Each \( X_1 \cdots X_p \) can be either discrete or continuous.
- Data are comprised of \( n \) data points of the form \((Y_i, X_{1i}, \ldots, X_{pi})\)
- For the \( i^{th} \) individual, we have a vector of predictor variable values that is represented \( X'_i = [X_{1i}, X_{2i}, \ldots, X_{pi}] \)

Assumptions
The assumptions required are an extension of the ones we saw previously.

1. The separate observations \( Y_1, Y_2, \cdots, Y_n \) are independent.
2. The values of the predictor variables \( X_1 \cdots X_p \) are fixed and measured without error.
3. For each vector value of the predictor variable \( X = x \), the distribution of values of \( Y \) follows a normal distribution with mean equal to \( \mu_{Y|X=x} \) and common variance equal to \( \sigma_{Y|X=x}^2 \).
4. The separate means \( E[Y \text{ at } X = x] = \mu_{Y|X=x} \) lie on the line with definition

\[ \mu_{Y|X=x} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \]
Indicator Variables (also called “dummy variables”) and Design Variables
Indicator variables are commonly used as predictors in multivariable regression models. We let

\[\begin{align*}
1 &= \text{value of indicator when “trait” is present} \\
0 &= \text{value of indicator when “trait” is not present}
\end{align*}\]

♦ The estimated regression coefficient \(\beta\) associated with an indicator variable has a straightforward interpretation, namely:

\[\beta = \text{predicted change in outcome Y that accompanies presence of “trait”}\]

Examples of Indicator Variables

SEXF = 1 if individual is female
0 otherwise

TREAT = 1 if individual received experimental treatment
0 otherwise

Design Variables (Meaningful “sets” of 0/1 predictor variables)
What do you do if you have a nominal predictor with more than 2 possible values? Answer – design variables! Design variables are sets of indicator variables that together define values of nominal variables. If a nominal variable has \(k\) possible values, \((k-1)\) indicator variables are needed to distinguish the entire range of possibilities.

Examples of Design Variables

Suppose a randomized trial seeks to compare medical therapy versus angioplasty versus bypass surgery for the treatment of myocardial infarction. Thus, the original treatment variable TREAT is nominal with 3 possible values:

\[\begin{align*}
\text{TREAT} &= 1 \text{ if treatment is medical therapy} \\
&= 2 \text{ if treatment is angioplasty} \\
&= 3 \text{ if treatment is bypass surgery}
\end{align*}\]

We cannot put TREAT into a regression model as is because the estimated regression coefficient would be uninterpretable. So TREAT is replaced with a set of 2 design variables. For example, we might include the following set:

\[\begin{align*}
\text{TR}_\text{ANG} &= 1 \text{ if treatment is angioplasty} \\
&= 0 \text{ otherwise} \\
\text{TR}_\text{SUR} &= 1 \text{ if treatment is bypass surgery} \\
&= 0 \text{ otherwise}
\end{align*}\]
A set of design variables comprised of \((3-1) = 2\) indicator variables summarize three possible values of treatment. The reference category is medical therapy.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Value of TR ANG</th>
<th>Value of TR SUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREAT=1 (“medical”)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREAT=2 (“angioplasty”)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TREAT=3 (“surgery”)</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Guidelines for the Definition of Indicator and Design Variables**

1) Consider the choice of the reference group. Often this choice will be straightforward. It might be one of the following categories of values of the nominal variable:

- The unexposed
- The placebo
- The standard
- The most frequent

2) \(K\) levels of the nominal predictor \(\rightarrow (K-1)\) indicator variables
When the number of levels of the nominal predictor variable = \(k\), define \((k-1)\) indicator variables that will identify persons in each of the separate groups, apart from the reference group.

3) In general (this is not hard and fast), treat the \((k-1)\) design variables as a set.

   - Enter the set together
   - Remove the set together
   - In general, retain all \((k-1)\) of the indicator variables, even when only a subset are significant.

**Interaction Variables**
Sometimes the nature of an \(X\)-\(Y\) relationship is *different*, depending on the level of some third variable which, for now, we’ll call \(Z\). This is interaction. To capture how an \(X\)-\(Y\) relationship is “different” (or “modified by”), depending on the level of \(Z\), we can define an *interaction variable* and then incorporate it as an additional predictor in the model.

Interaction of predictor \(X\) with third variable \(Z = XZ = X*Z\)

**Example:** \(X = \text{age (years)}\) \(\quad Z = 0/1\) indicator of history of vertebral fracture \(\quad Y = \text{length of hospital stay}\).

Interaction \(XZ = X*Z \rightarrow \beta_{XZ} = \text{extra change in } y=\text{length of stay per year of } X=\text{age for persons with vertebral fracture} \)
b. The Analysis of Variance Table

The ideas of the analysis of variance table introduced previously (see page 20) apply here, as well.

1. **TSS**: “Total” or “total, corrected”
   - TSS = \( \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \) is the variability of \( Y \) about \( \bar{Y} \)
   - Degrees of freedom = df = (n-1).

2. **MSS**: “Regression” or “due model”
   - MSS = \( \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \) is the variability of \( \hat{Y} \) about \( \bar{Y} \)
   - Degrees of freedom = df = p = # predictors apart from intercept

3. **RSS**: “Residual” or “due error” refers to the
   - RSS = \( \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \) is the variability of \( Y \) about \( \hat{Y} \)
   - Degrees of freedom = df = (n-1) – (p)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>p</td>
<td>MSS = ( \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 )</td>
<td>(MSS)/p</td>
</tr>
<tr>
<td>Residual</td>
<td>(n-1) - p</td>
<td>RSS = ( \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 )</td>
<td>(RSS)/(n-1-p)</td>
</tr>
<tr>
<td>Total, corrected</td>
<td>(n-1)</td>
<td>TSS = ( \sum_{i=1}^{n} (Y_i - \bar{Y})^2 )</td>
<td></td>
</tr>
</tbody>
</table>
Overall F Test

The overall F test introduced previously (see page 29) also applies, yielding an overall F-test to assess the significance of the variance explained by the model. Note that the degrees of freedom is different here; this is because there are now “p” predictors instead of 1 predictor.

\[ H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0 \]
\[ H_A: \text{At least one } \beta_i \neq 0 \]

\[ F = \frac{\text{mean square(model)}}{\text{mean square(residual)}} \text{ with df} = p, (n-1-p) \]

Stata illustration

**Example** - Consider again a multiple linear regression analysis of the relationship of \( Y = \text{p53 expression} \) to \( \text{age at first pregnancy (pregnum)}, \text{ 1st pregnancy at age } \leq 24 \text{ (early)}, \text{ and 1st pregnancy at age } > 24 \text{ (late)}. \) The variables \text{early} \text{ and } \text{late} \text{ are each } 0/1. \text{ The referent group is nulliparous.}

```
. use http://people.umass.edu/biep691f/data/p53paper_small.dta, clear
. regress p53 pregnum early late

Source |       SS       df       MS              Number of obs =      67
       |-------------+---------------------------------------------F(  3,    63) =  5.35
Model |  14.8967116     3  4.96557054                  Prob > F      =  0.0024
Residual |  58.486889    63  .928363317           R-squared     =  0.2030
---------+---------------------------------------------Adj R-squared =  0.1650
Total |  73.3836006    66  1.11187274          Root MSE      =  .96352

F_{3,63} = \frac{\text{msq(Model)}}{\text{msq(Residual)}} = \frac{4.96557054}{0.92836317} = 5.3487
```

The null hypothesis is rejected (p-value of overall F-test = .0024). At this point, all we can say is that this model explains statistically significantly more of the variability in \( Y = \text{p53} \) then is explained by no model at all.
c. The Partial F Test

What if we want to compare and choose between two models?
The partial F test is a statistical technique for comparing two models that are “hierarchical.” It permits the assessment of associations while controlling for confounding.

What are hierarchical models?

- Hierarchical models are two models of a particular type. One model is called “smaller” or “reduced” or “reference”. The other model is called “larger” or “comparison”

- “Hierarchical” means that all of the predictors in the smaller (reduced, reference) are contained in the larger (comparison) model.

- In the $Y = p53$ example, we might be interested in comparing the following two hierarchical models:
  
  Predictors in smaller model = \{ pregnum \}
  Predictors in larger model = \{ pregnum + early + late \}

- “Hierarchical” is satisfied because all of the predictors (here there is just one - pregnum) that are contained in the smaller model are contained in the larger model.

- The important point to note is this. The comparison of these two models is an analysis of the nature and significance of the extra predictors, (here - early and late) for the prediction of $Y = p53$, adjusting for (controlling for) all of the variables in the smaller model (pregnum).

Thus, the comparison of the hierarchical models is addressing the following question:

What is the significance of early and late for the prediction of $Y = p53$, after controlling for the effects of pregnum?
Statistical Definition of the Partial F Test

**Research Question:** Does inclusion of the “extra” predictors explain significantly more of the variability in outcome compared to the variability that is explained by the predictors that are already in the model?

### Partial F Test

**H₀:** Addition of \( X_{p+1} \cdots X_{p+k} \) is of no statistical significance for the prediction of \( Y \) after controlling for the predictors \( X_1 \cdots X_p \) meaning that:

\[
\beta_{p+1} = \beta_{p+2} = \ldots = \beta_{p+k} = 0
\]

**Hₐ:** Not

**F⁡_{\text{PARTIAL}} =**

\[
\frac{\text{\{ Extra regression sum of squares \}}}{\text{\{ Extra regression df \}}} \div \frac{\text{\{ Residual sum of squares larger model \}}}{\text{\{ Residual df larger model \}}}
\]

\[
= \frac{\left[ \text{MSS}(X_1 \cdots X_p X_{p+1} \cdots X_{p+k}) - \text{MSS}(X_1 \cdots X_p) \right]}{\left[ \text{RSS}(X_1 \cdots X_p X_{p+1} \cdots X_{p+k}) \right]} \div \frac{\left[ (n-1) - (p+k) \right]}{\left[ (p+k) - p \right]}
\]

Numerator df = \( (p+k) - (p) = k \)
Denominator df = \( (n-1) - (p+k) \)

**H₀ true:**
The extra predictors are not significant in adjusted analysis

**H₀ false:**
The extra predictors are significant in adjusted analysis

**Stata illustration Example – continued.**

```stata
. regress p53 pregnum early late
. test early late
( 1) early = 0
( 2) late = 0

F(  2,    63) =    0.31
Prob > F =    0.7381
```

**The null hypothesis is NOT rejected (p-value = .74).** Conclude that early and late are not predictive after adjustment for pregnum. Specifically, their addition to the model does not explain statistically significantly more of the variability in \( Y=p53 \) beyond that explained by pregnum.
d. Multiple Partial Correlation

The concept of a partial correlation is related to that of a partial F test.

- **“To what extent are two variables, say X and Y, correlated after accounting for a control variable, say Z?”**

  - **Preliminary 1**: Regress X on the control variable Z
    - Obtain the residuals
    - These residuals represent the information in X that is independent of Z

  - **Preliminary 2**: Now regress Y on the control variable Z
    - Obtain the residuals
    - These residuals represent the information in Y that is independent of Z

  - These two sets of residuals permit you to look at the relationship between X and Y, independent of Z.

| Partial correlation (X,Y | controlling for Z) |
|-------------------------|
| = Correlation (residuals of X regressed on Z, residuals of Y regressed on Z) |

If there is more than one control variable Z, the result is a multiple partial correlation

**A nice identity** allows us to compute a partial correlation by hand from a multivariable model development

- Recall that $R^2 = \frac{\text{[model sum of squares]}}{\text{[total sum of squares]}} = \frac{\text{MSS}}{\text{TSS}}$

- A **partial correlation** is also a **ratio of sums of squares**.

  **Tip!** – *A partial F statistic is a ratio of mean squares.*

Nature | Population/ Sample | Observation/ Data | Relationships/ Modeling | Analysis/ Synthesis
The hypothesis test of a zero partial correlation is the partial F test introduced previously.

**Research Question:** Controlling for $Z$, is there a linear correlation between $X$ and $Y$?

$H_0$: $\rho_{X,Y|Z} = 0$

$H_A$: Not

$$F_{\text{PARTIAL}} = \frac{[\text{MSS}(X, Z) - \text{MSS}(X)] / (2 - 1)}{[\text{RSS}(X, Z)] / [(n-1) - (2)]}$$

$$= \frac{\{\text{Extra regression sum of squares}\}}{\{\text{Extra regression df} = 1\}} / \{\text{SS Residual larger model}\} / \{\text{df Residual larger model}\}$$

Numerator df = (2) – (1) = 1
Denominator df = (n – 1) – (2)

**BEWARE!** Notice that the denominator of the partial F test contains the residual sum of squares (RSS) for the larger model, whereas the denominator of the partial correlation contains the residual sum of squares (RSS) for the smaller model!
4. Multivariable Model Development

a. Introduction

George E.P. Box, a very famous statistician once said, “All models are wrong, but some are useful.” Incorrectness notwithstanding, we do statistical modeling for a very good reason: we seek an understanding of the natures and strengths of the relationships (if any) that might exist in a set of observations that vary. For any set of observations, theoretically, lots of models are possible. So, how to choose? The goal of statistical modeling is to obtain a model that is simultaneously minimally adequate and a good fit. The model should also make sense.

Minimally adequate

- Each predictor is “important” in its own right
- Each extra predictor is retained in the model only if it yields a significant improvement (in fit and in variation explained).
- The model should not contain any redundant parameters.

Good Fit

- The amount of variability in the outcomes (the Y variable) explained is a lot
- The outcomes that are predicted by the model are close to what was actually observed.

The model should also make sense

- A preferred model is one based on “subject matter” considerations
- The preferred predictors are the ones that are simply and conveniently measured.

It is not possible to choose a model that is simultaneously minimally adequate and a perfect fit. Model estimation and selection, therefore, must achieve an appropriate balance.
General Guidelines for Multivariable Regression Model Estimation

**Preliminary** – Be sure you have: (1) checked, cleaned and described your data, (2) screened the data for multivariate associations, and (3) thoroughly explored the bivariate relationships.

**Step 1 – Fit the “maximal” model.**
The maximal model is the large model that contains all the explanatory variables of interest as predictors. This model also contains all the covariates that might be of interest. It also contains all the interactions that might be of interest. Note the amount of variation explained.

**Step 2 – Begin simplifying the model.**
Inspect each of the terms in the “maximal” model with the goal of removing the predictor that is the least significant. Drop from the model the predictors that are the least significant, beginning with the higher order interactions (Tip -interactions are complicated and we are aiming for a simple model). Fit the reduced model. Compare the amount of variation explained by the reduced model with the amount of variation explained by the “maximal” model.

If the deletion of a predictor has little effect on the variation explained
Then leave that predictor out of the model.
And inspect each of the terms in the model again.
If the deletion of a predictor has a significant effect on the variation explained
Then put that predictor back into the model.

**Step 3 – Keep simplifying the model.**
Repeat step 2, over and over, until the model remaining contains nothing but significant predictor variables.

**Beware of some important caveats**

- Sometimes, you will want to keep a predictor in the model regardless of its statistical significance (an example is randomization assignment in a clinical trial)
- The order in which you delete terms from the model matters
- You still need to be flexible to considerations of biology and what makes sense.
b. Example

Framingham Study

Source:
Framingham Heart Study

Description:
Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. In 1948, the Framingham Heart Study - under the direction of the National Heart Institute (now known as the National Heart, Lung, and Blood Institute or NHLBI) was initiated. The objective of the Framingham Heart Study was to identify the common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke.

Here we use a subset of the data in a subset of n=1000.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>Codings</th>
</tr>
</thead>
<tbody>
<tr>
<td>sbp</td>
<td>Systolic Blood Pressure (mm Hg)</td>
<td></td>
</tr>
<tr>
<td>ln_sbp</td>
<td>Natural logarithm of sbp</td>
<td>ln_sbp=ln(sbp)</td>
</tr>
<tr>
<td>age</td>
<td>Age, years</td>
<td></td>
</tr>
<tr>
<td>bmi</td>
<td>Body Mass index (kg/m2)</td>
<td></td>
</tr>
<tr>
<td>ln_bmi</td>
<td>Natural logarithm of bmi</td>
<td>ln_bmi=ln(bmi)</td>
</tr>
<tr>
<td>sex</td>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>Female Indicator</td>
<td></td>
</tr>
<tr>
<td>scl</td>
<td>Serum Cholesterol (mg/100 ml)</td>
<td></td>
</tr>
<tr>
<td>ln_scl</td>
<td>Natural logarithm of scl</td>
<td>ln_scl=ln(scl)</td>
</tr>
</tbody>
</table>

Multiple Regression Variables:
Outcome Y = ln_sbp
Predictor Variables: ln_bmi, ln_scl, age, sex

Research Question:
From among these 4 “candidate” predictors, what are the important “risk” factors and what is the nature of their association with Y=ln_sbp?
Stata illustration

. * ----- Preliminary: Check variables for completeness, adequacy of range, etc.
. codebook sex sbp scl age bmi id, compact

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Unique</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>1000</td>
<td>2</td>
<td>1.557</td>
<td>1</td>
<td>2</td>
<td>Sex</td>
</tr>
<tr>
<td>sbp</td>
<td>1000</td>
<td>87</td>
<td>132.35</td>
<td>80</td>
<td>270</td>
<td>Systolic Blood Pressure</td>
</tr>
<tr>
<td>scl</td>
<td>996</td>
<td>182</td>
<td>227.84</td>
<td>115</td>
<td>493</td>
<td>Serum Cholesterol</td>
</tr>
<tr>
<td>age</td>
<td>1000</td>
<td>36</td>
<td>45.922</td>
<td>30</td>
<td>66</td>
<td>Age in Years</td>
</tr>
<tr>
<td>bmi</td>
<td>998</td>
<td>186</td>
<td>25.57</td>
<td>16.4</td>
<td>43.4</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>id</td>
<td>1000</td>
<td>1000</td>
<td>2410.031</td>
<td>1</td>
<td>4697</td>
<td></td>
</tr>
</tbody>
</table>

. * ----- Numerical descriptives: Explore the data for shape, range, outliers and completeness.
. tabstat sbp ln_sbp age bmi ln_bmi scl, statistics(n mean sd min q max) columns(statistics) format(%8.2f)

<table>
<thead>
<tr>
<th>variable</th>
<th>N</th>
<th>mean</th>
<th>sd</th>
<th>min</th>
<th>p25</th>
<th>p50</th>
<th>p75</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>sbp</td>
<td>1000</td>
<td>132.35</td>
<td>23.04</td>
<td>80</td>
<td>116</td>
<td>128</td>
<td>144</td>
<td>270</td>
</tr>
<tr>
<td>ln_sbp</td>
<td>1000</td>
<td>4.87</td>
<td>0.16</td>
<td>4.38</td>
<td>4.75</td>
<td>4.85</td>
<td>4.97</td>
<td>5.60</td>
</tr>
<tr>
<td>age</td>
<td>1000</td>
<td>45.92</td>
<td>8.55</td>
<td>30.00</td>
<td>38.50</td>
<td>45.00</td>
<td>53.00</td>
<td>66.00</td>
</tr>
<tr>
<td>bmi</td>
<td>998</td>
<td>25.57</td>
<td>3.85</td>
<td>16.40</td>
<td>23.00</td>
<td>25.10</td>
<td>27.80</td>
<td>43.40</td>
</tr>
<tr>
<td>ln_bmi</td>
<td>998</td>
<td>3.23</td>
<td>0.15</td>
<td>2.80</td>
<td>3.14</td>
<td>3.22</td>
<td>3.33</td>
<td>3.77</td>
</tr>
<tr>
<td>scl</td>
<td>996</td>
<td>227.85</td>
<td>45.09</td>
<td>115</td>
<td>197</td>
<td>225</td>
<td>255</td>
<td>493</td>
</tr>
</tbody>
</table>

. fre sex

<table>
<thead>
<tr>
<th>Sex</th>
<th>Freq.</th>
<th>Percent</th>
<th>Valid</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>1000</td>
<td>100.00</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>1 Men</td>
<td>443</td>
<td>44.30</td>
<td>44.30</td>
<td>44.30</td>
</tr>
<tr>
<td>2 Women</td>
<td>557</td>
<td>55.70</td>
<td>55.70</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>100.00</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

. * ----- Assess normality of "candidate" dependent variable Y=sbp
. * sfancia test of normality (Null: distribution is normal)
. sfancia sbp

Shapiro-Francia W' test for normal data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>W'</th>
<th>V'</th>
<th>z</th>
<th>Prob&gt;z</th>
</tr>
</thead>
<tbody>
<tr>
<td>sbp</td>
<td>1000</td>
<td>0.92135</td>
<td>52.674</td>
<td>9.088</td>
<td>0.00001</td>
</tr>
</tbody>
</table>

Interpretation: The null hypothesis of normality of the distribution of sbp is rejected (p=.00001)
Interpretation: This confirms what the sfrancia test suggests. The null hypothesis of normality of the distribution of sbp is not supported.
. * Tip - command gladder to explore appropriate transformations of Y=sbp
. * NOTE – You may need to issue the command findit gladder and download the routine sed2

. gladder sbp

Interpretation: Comparison of these plots suggests that the log transformation is reasonable. This is the distribution that “looks” the most normal.
* ----- Create "regression-friendly" variables
* ----- 0/1 Indicator of FEMALE gender
. generate female=sex
. recode female (1=0) (2=1)
(female: 1000 changes made)

. tab2 sex female

<table>
<thead>
<tr>
<th>female</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>Men</td>
<td>443</td>
<td>0</td>
</tr>
<tr>
<td>Women</td>
<td>0</td>
<td>557</td>
</tr>
</tbody>
</table>

. label variable female "Female (0/1)"

* ----- INTERACTIONS

* Interaction age * female sex
. generate age_female=age*female
(0 missing values generated)

* Interaction ln(scl) * female sex
. generate lnsc1_female=ln_scl*female
(4 missing values generated)

* Interaction ln(bmi) * female sex
. generate lnbmi_female=ln_bmi*female
(2 missing values generated)

. label variable age_female "Age x Female Interaction"
. label variable lnsc1_female "ln(scl) x Female Interaction"
. label variable lnbmi_female "ln(bmi) x Female Interaction"
. * ----- Bivariate Relationships – Numerical and Graphical Assessments
. * Command pwcorr gives a quick and succinct look at the correlations of Y with each X
   pwcorr ln_sbp age ln_bmi ln_scl sex, obs sig
   | ln_sbp   age   ln_bmi   ln_scl   sex
   |--------------------------------------------------------
   | 1.0000   age   ln_bmi   ln_scl   sex
   |---------------------------------------------
   | ln_sbp |   1.0000
   |        | 1000   correlation(ln_sbp, age) = .4122  (Thus, R-squared = .1699)
   | age    | 0.4122  1.0000
   |        | 2.0000  p-value for Null: zero correlation < .0001 ➔ Reject null.
   | 1000   | 1000
   | ln_bmi | 0.3465  0.1961  1.0000
   |        | 0.0000  0.0000
   |        | 998     998     998
   | ln_scl | 0.2509  0.3066  0.2358  1.0000
   |        | 0.0000  0.0000  0.0000
   |        | 996     996     994     996
   | sex    | 0.0164  0.0239  -0.0692  0.0077  1.0000
   |        | 0.6047  0.4501  0.0288  0.8074
   |        | 1000    1000    998     996    1000

. * Command graph matrix gives a quick and succinct look at the pairwise scatter plots.
   graph matrix ln_sbp age ln_bmi ln_scl, half msize(vsmall)
**Tip** - Want a closer look? Use graph twoway with 3 overlays: scatter, linear, and lowess.

```
graph twoway (scatter ln_sbp ln_bmi, symbol(d) msize(vsmall)) (lfit ln_sbp ln_bmi) (lowess ln_sbp ln_bmi), title("Bivariate Association") ylabel4(.5)6) ytitle("Y = ln_sbp") xtitle("X = ln(bmi)"") legend(off)
```

---

**MODEL DEVELOPMENT**

**FIT of maximal model**

```
. reg ln_sbp ln_bmi ln_scl age female lnbmi_female lnscl_female age_female
```

```
Source |       SS           df       MS      Number of obs   =       994
-------------+------------------------------------------------------------------
Model | 7.01711933         7  1.00244562   Prob > F        =    0.0000
Residual | 19.3006631       986  .019574709   R-squared       =    0.2666
-------------+------------------------------------------------------------------
Total | 26.3177825       993  .026503306   Root MSE        =    .13991

```

| ln_sbp | Coef.  | Std. Err.  | t     | P>|t|    | [95% Conf. Interval] |
|--------|--------|------------|-------|--------|----------------------|
| ln_bmi | .303811| .0549107   | 5.53  | 0.000  | .1960557             | .4115663 |
| ln_scl | .0591585| .0368291   | 1.61  | 0.109  | -.013114            | .131431 |
| age    | .003694| .0008046   | 4.59  | 0.000  | .002115             | .0052729 |
| female | -.0109333| .0368291 | -0.28 | 0.777  | -.0748536          | .0530863 |
| lnbmi_female | -.0507228| .0674812 | -0.75 | 0.452  | -.1831461          | .0817005 |
| lnscl_female | -.0091802| .0498751 | -0.18 | 0.854  | -.1070538         | .0886934 |
| age_female | .0050381| .0011343 | 4.44  | 0.000  | .0028121            | .0072641 |
| _cons  | 3.396028| .233872 | 14.52 | 0.000  | 2.937084            | 3.854972 |

**Interpretation:** Maximal model performs better than no model at all, with approximately 27% of the variability in Y=ln(sbp) explained. However, a number of predictors are not significant suggesting that a more parsimonious model is possible.
... 3 df Partial F test of Interactions (NULL: interactions are not significant, controlling for main effects)
  . testparm lnbmi_female lnscl_female age_female

( 1)  lnbmi_female = 0
( 2)  lnscl_female = 0
( 3)  age_female = 0

F(  3,   986) =    6.89
Prob > F =    0.0001

Interpretation: Together, these 3 interactions are significant (p=.0001), but this may be driven by a subset.

... 2 df Partial F test of Interactions (NULL: interactions are not significant, controlling for main effects)
  . testparm lnbmi_female lnscl_female

( 1)  lnbmi_female = 0
( 2)  lnscl_female = 0

F(  2,   986) =    0.34
Prob > F =    0.7144

Interpretation: okay to DROP lnbmi_female and lnscl_female

... Fit of reduced multiple predictor model
  . regress ln_sbp ln_bmi ln_scl age female age_female

Source | SS           df       MS      Number of obs   =       994
-------------+--------------------------------------------------
Model |  7.00394663         5  1.40078933   Prob > F        =    0.0000
Residual |  19.3138358       988  .019548417   R-squared       =    0.2661
-------------+--------------------------------------------------
Total |  26.3177825       993  .026503306   Root MSE        =    .13982

ln_sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+--------------------------------------------------
ln_bmi |   .2707647   .0318537     8.50   0.000      .208256    .3332734
ln_scl |   .0559982    .024711     2.27   0.024     .0075061    .1044902
age |   .0036879   .0008017     4.60   0.000     .0021147    .0052612
female |  -.2169167   .0508166    -4.27   0.000    -.3166377    -.1171957
age_female |  -.0000000   .0000000     0.00   1.000     .0000000    .0000000
_cons |   3.520535   .1586124    22.20   0.000     3.209279    3.831791

Interpretation: Not bad. Previously 27% of the variance was explained. This has not changed much and we have a simpler model.

... Produce side-by-side comparison of models
  . NOTE – You may need to issue the command findit eststo and download
  . TIP – I choose to suppress all the output (we’re not needing this right now) using prefix quietly:

  . eststo model1

  . *-- model 1 – Initial “maximal” model
  . quietly: regress ln_sbp ln_bmi ln_scl age female age_female lnbmi_female lnscl_female
  . eststo modell
*. *--- model 2 - Candidate final multiple predictor model
. quietly: regress ln_sbp ln_bmi ln_scl age female age_female
. eststo model2

*. * --- model 3 - Single Predictor model, X=ln(bmi)
. quietly: regress ln_sbp ln_bmi
. eststo model3

*. * ---- model 4 - Single Predictor model, X=ln(scl)
. quietly: regress ln_sbp ln_scl
. eststo model4

*. * ---- model 5 - Two Predictor model + Interaction: age, female, and [age x female]
. quietly: regress ln_sbp age female age_female
. eststo model5

*. * -- Show comparison of models #1 - #5
. esttab, r2 se scalar(rmse)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln_sbp</td>
<td>0.304***</td>
<td>0.271***</td>
<td>0.384***</td>
<td>0.0549</td>
<td>0.0319</td>
</tr>
<tr>
<td>ln_bmi</td>
<td>0.0592</td>
<td>0.0560*</td>
<td>0.210***</td>
<td>0.0368</td>
<td>0.0247</td>
</tr>
<tr>
<td>ln_scl</td>
<td>0.00369***</td>
<td>0.00369***</td>
<td>0.00366***</td>
<td>0.000805</td>
<td>0.000802</td>
</tr>
<tr>
<td>age</td>
<td>-0.0109</td>
<td>-0.217***</td>
<td>-0.331***</td>
<td>0.00113</td>
<td>0.00109</td>
</tr>
<tr>
<td>female</td>
<td>0.00918</td>
<td>0.00918</td>
<td>0.00918</td>
<td>0.0675</td>
<td>0.0499</td>
</tr>
<tr>
<td>age_female</td>
<td>0.00504***</td>
<td>0.00487***</td>
<td>0.00726***</td>
<td>0.00726***</td>
<td>0.00109</td>
</tr>
<tr>
<td>lnbmi_female</td>
<td>-0.0507</td>
<td>-0.0507</td>
<td>-0.0507</td>
<td>0.0675</td>
<td></td>
</tr>
<tr>
<td>lnscl_female</td>
<td>-0.00918</td>
<td>-0.00918</td>
<td>-0.00918</td>
<td>0.0499</td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>3.396***</td>
<td>3.521***</td>
<td>3.630***</td>
<td>3.736***</td>
<td>4.702***</td>
</tr>
</tbody>
</table>

| N        | 994 | 994 | 998 | 996 | 1000 |
| R-sq     | 0.267 | 0.266 | 0.120 | 0.063 | 0.205 |
| rmse     | 0.140 | 0.140 | 0.153 | 0.158 | 0.146 |

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

*. * model 2 is our "tentative" final model

Further work, regression diagnostics, are needed next (See, section 4. Checking Model Assumptions and Fit).
c. Suggested Criteria for Confounding and Interaction

A Suggested Statistical Criterion for Determination of Confounding

A variable Z might be judged to be a confounder of an X-Y relationship if BOTH of the following are satisfied:

1) Its inclusion in a model that already contains X as a predictor has adjusted significance level < .10 or < .05; and

2) Its inclusion in the model alters the estimated regression coefficient for X by 15-20% or more, relative to the model that contains only X as a predictor.

A Suggested Statistical Criterion for Assessment of Interaction

A “candidate” interaction variable might be judged to be worth retaining in the model if BOTH of the following are satisfied:

1) The partial F test for its inclusion has significance level < .05; and

2) Its inclusion in the model alters the estimated regression coefficient for the main effects by 15-20% or more.
d. Guidelines for Multivariable Analysis of Large Data Sets

#1. State the Research Questions.
Aim for a focus that is explicit, complete, and focused, including:

♦ Statement of population
♦ Definition of outcome
♦ Specification of hypotheses (predictor-outcome relationships)
♦ Identification of (including nature of) hypothesized covariate relationships

#2. Define the Analysis Variables.
For each research question, note for each analysis variable, its hypothesized role.

♦ Outcome
♦ Predictor
♦ Confounder
♦ Effect Modifier
♦ Intermediary (also called intervening)

#3. Prepare a “Clean” Data Set Ready for Analysis (Data Management)
For each variable, check its distribution, especially:

♦ Completeness
♦ Occurrence of logical errors
♦ Within form consistency
♦ Between form consistency
♦ Range
#4. Describe the Analysis Sample

This description serves three purposes:

1) Identifies the population actually represented by the sample
2) Defines the range(s) of relationships that can be explored
3) Identifies, tentatively, the function form of the relationships

Methods include:
- Frequency distributions for discrete variables
- Mean, standard deviation, percentiles for continuous variables
- Bar charts
- Box and whisker plots
- Scatter plots

#5. Assessment of Confounding

The identification of confounders is needed for the correct interpretation of the predictor-outcome relationships. Confounders need to be controlled in analyses of predictor-outcome relationships.

Methods include:
- Cross-tabulations and single predictor regression models to determine whether suspected confounders are predictive of outcome and are related to the predictor of interest.
- This step should include a determination that there is a confounder-exposure relationship among controls.

#6. Single Predictor Regression Model Analyses

The fit of these models identifies the nature and magnitude of crude associations. It also permits assessment of the appropriateness of the assumed functional form of the predictor-outcome relationship.

- Cross-tabulations
- Graphical displays (Scatter plots)
- Estimation of single predictor models
5. Goodness-of-Fit and Regression Diagnostics

a. Introduction and Terminology

Neither prediction nor estimation has meaning when the estimated model is a poor fit to the data:

- Our eye “tells” us:
  - A better fitting relationship between X and Y is quadratic
  - We notice different sizes of discrepancies
  - Some observed Y are close to the fitted $\hat{Y}$ (e.g. near X=1 or X=8)
  - Other observed Y are very far from the fitted $\hat{Y}$ (e.g. near X=5)

Poor fits of the data to a fitted line can occur for several reasons and can occur even when the fitted line explains a large proportion ($R^2$) of the total variability in response:

- The wrong functional form (link function) was fit.
- Extreme values (outliers) exhibit uniquely large discrepancies between observed and fitted values.
- One or more important explanatory variables have been omitted.
- One or more model assumptions have been violated.
Consequences of a poor fit include:

♦ We learn the wrong biology.

♦ Comparison of group differences aren’t “fair” because they are unduly influenced by a minority.

♦ Comparison of group means aren’t “fair” because we used the wrong standard error.

♦ Predictions are wrong because the fitted model does not apply to the case of interest.

Available techniques of goodness-of-fit assessment are of two types:

1. **Systematic** - those that explore the appropriateness of the model itself

   *Have we fit the correct model?*
   
   *Should we fit another model?*

2. **Case Analysis** – those that investigate the influence of individual data points

   *Are there a small number of individuals whose inclusion in the analysis influences excessively the choice of the fitted model?*
Goodness-of-Fit Assessment

Some Terminology

The Multiple Linear Regression Model, again:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \text{error} \]

\[ \text{Observed} = \text{systematic} + \text{error} \]

this is the mean of \( Y \)
at \( X_1, X_2, \ldots, X_p \)
\[ = E[Y \text{ at } \bar{X}] = \]

Systematic Component

<table>
<thead>
<tr>
<th>Link:</th>
<th>The functional form (and the assumed underlying distribution of the errors) is sometimes called the link.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example:</td>
<td>When ( \mu ) is the mean of a normal distribution, we model ( \mu_{Y</td>
</tr>
<tr>
<td>Example:</td>
<td>When ( \mu ) is a proportion, we might model ( \ln \left( \frac{\mu_{Y</td>
</tr>
<tr>
<td>Normality:</td>
<td>In the linear model regression analysis, we assume that the errors ( E ) follow a ( \text{Normal}(0, \sigma^2_{Y</td>
</tr>
<tr>
<td>Recall:</td>
<td>The errors ( \varepsilon ) are estimated by the residuals ( e ).</td>
</tr>
<tr>
<td>Heteroscedasticity:</td>
<td>If the assumption of constant variance of the errors ( E ) is not true, we say there is heteroscedasticity of errors, or non-homogeneity of errors.</td>
</tr>
</tbody>
</table>
### Goodness-of-Fit Assessment

#### Some Terminology - continued

<table>
<thead>
<tr>
<th>Case Analysis</th>
</tr>
</thead>
</table>
| **Residual:** | The residual is the difference between the observed outcome Y and the fitted outcome \( \hat{Y} \).
\[
e = Y - \hat{Y}
\]
It estimates the unobservable error \( \epsilon \).
| **Outlier:** | An outlier is a residual that is unusually large.
*Note:* As before, we will rescale the sizes of the residuals via standardization so that we can interpret their magnitudes on the scale of SE units.
| **Leverage:** | The leverage is a measure of the unusualness of the value of the predictor X.

Leverage = distance (observed X, center of X in sample)

Predictor values with high leverages have, potentially, a large influence on the choice of the fitted model.
| **Influence:** | Measures of influence gauge the change in the fitted model with the omission of the data point.

*Example: Cook’s Distance* |
A Feel for Residual, Leverage, Influence
Large residuals may or may not be influential

Large residual
Low leverage

The large residual effects a large influence.

Large residual
Low leverage

Despite its size, the large residual effects only small influence.
A Feel for Residual, Leverage, Influence

High leverage may or may not be influential

Thus, case analysis is needed to discover all of:

♦ high leverage
♦ large residuals
♦ large influence
## Overview of Techniques of Goodness-of-Fit Assessment

### Linear Model

<table>
<thead>
<tr>
<th>Question Addressed</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systematic Component</strong></td>
<td></td>
</tr>
<tr>
<td>Error Distribution: Is it reasonable to assume a normal distribution of errors with a constant variance?</td>
<td>Shapiro-Wilk test of normality</td>
</tr>
<tr>
<td>H₀: E ~ Normal (0, σ²)</td>
<td>Cook-Weisberg test of heteroscedasticity</td>
</tr>
<tr>
<td>Functional Form: Is the choice of functional form relating the predictors to outcome a “good” one?</td>
<td>Method of fractional polynomials.</td>
</tr>
<tr>
<td>Systematic Violation: Have we failed to include any important explanatory (predictor) variables?</td>
<td>Ramsey Test for omitted variables.</td>
</tr>
<tr>
<td><strong>Case Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Are there outliers with respect to the <strong>outcome</strong> values?</td>
<td>Studentized residuals</td>
</tr>
<tr>
<td>Are there outliers with respect to the <strong>predictor</strong> variable values?</td>
<td>Leverage</td>
</tr>
<tr>
<td>Are there individual observations with unduly large influence on the fitted model?</td>
<td>Cook’s distance (influence)</td>
</tr>
</tbody>
</table>
b. Assessment of Normality

Recall what we are assuming with respect to normality:

- **Simple Linear Regression:**
  At each level \( x \) of the predictor variable \( X \), the outcomes \( Y_X \) are distributed normal with mean \( \mu_{Y|x} = \beta_0 + \beta_1 x \) and constant variance \( \sigma^2_{Y|x} \).

- **Multiple Linear Regression:**
  At each vector level \( \mathbf{x} = [x_1, x_2, \ldots, x_p] \) of the predictor vector \( X \), the outcomes \( Y_X \) are distributed normal with mean \( \mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \) and constant variance \( \sigma^2_{Y|x} \).

*This is what it looks like (courtesy of a picture on the web!)*

Violations of Normality are sometimes, but not always, a serious problem

- **When not to worry:** Estimation and hypothesis tests of regression parameters are fairly robust to modest violations of normality.

- **When to worry:** Predictions are sensitive to violations of normality.

- **Beware:** Sometimes the cure for violations of normality is worse than the problem.
Some graphical assessments of normality and what to watch out for:

<table>
<thead>
<tr>
<th>Method</th>
<th>What to watch out for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Histogram of outcome variable Y and/or</td>
<td>Look for normal shape of the histogram.</td>
</tr>
<tr>
<td>Histogram of residuals</td>
<td></td>
</tr>
<tr>
<td>2. Histogram of residuals (or studentized or</td>
<td>Look for normal shape of the histogram.</td>
</tr>
<tr>
<td>jackknife residuals)</td>
<td></td>
</tr>
<tr>
<td>3. Quantile quantile plot of the quantiles of the residuals</td>
<td>Normally distributed residuals will appear,</td>
</tr>
<tr>
<td>versus the quantiles of the assumed</td>
<td>approximately, linear.</td>
</tr>
<tr>
<td>normal distribution of the residuals</td>
<td></td>
</tr>
</tbody>
</table>

Stata Illustration  
(note – This example uses a data set from another source, not this lecture)

Histgram with overlay normal

```
. histogram weight, normal title("Histogram with Overlay Normal")
```

Quantile Quantile Plot w reference = Normal

```
. qnorm weight, title("Simple Normal QQ-Plot for Y=Weight")
```

![Histogram with Overlay Normal](image1)

![Simple Normal QQ-Plot for Y=Weight](image2)
Skewness and Kurtosis Statistics for Assessing Normality:

<table>
<thead>
<tr>
<th><strong>Skewness - symmetry</strong> of the curve</th>
<th><strong>What to watch out for:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardization of the 3rd sample moment about the mean</td>
<td></td>
</tr>
<tr>
<td>$m_2 = E \left[ (Y-\mu)^2 \right]$</td>
<td></td>
</tr>
<tr>
<td>$m_3 = E \left[ (Y-\mu)^3 \right]$</td>
<td></td>
</tr>
<tr>
<td>What is actually examined is $a_3 = \frac{m_3}{(m_2)^{3/2}}$ because it is unitless</td>
<td></td>
</tr>
<tr>
<td>$a_3 = 0$ indicates symmetry</td>
<td></td>
</tr>
<tr>
<td>$a_3 &lt; 0$ indicates left-hand skew (tail to left)</td>
<td></td>
</tr>
<tr>
<td>$a_3 &gt; 0$ indicates right-hand skew (tail to right)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Kurtosis - flatness versus peakedness</strong> of the curve</th>
<th><strong>What to watch out for:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardization of the 4th sample moment about the mean</td>
<td></td>
</tr>
<tr>
<td>$m_2 = E \left[ (Y-\mu)^2 \right]$</td>
<td></td>
</tr>
<tr>
<td>$m_4 = E \left[ (Y-\mu)^4 \right]$</td>
<td></td>
</tr>
<tr>
<td>Pearson kurtosis is $a_4 = \frac{m_4}{(m_2)^2}$</td>
<td></td>
</tr>
<tr>
<td>$a_4 = 3$ when distribution is normal</td>
<td></td>
</tr>
<tr>
<td>$a_4 &lt; 3$ is “leptokurtic” is too little in the tails</td>
<td></td>
</tr>
<tr>
<td>$a_4 &gt; 3$ is “platykurtic” is too much in the tails</td>
<td></td>
</tr>
</tbody>
</table>

* Stata command
  . summarize `yvariable', detail

or

* Stata command
  . tabstat `yvariable', statistics(skewness)

When `yvariable` is distributed normal:

Skewness = 0

**Look for** skewness between -2 and +2, roughly.

* Stata command
  . summarize `yvariable', detail

or

* Stata command
  . tabstat `yvariable', statistics(kurtosis)

When `yvariable` is distributed normal:

Kurtosis = 3
Hypothesis Tests of Normality and what to watch out for:

<table>
<thead>
<tr>
<th><strong>Test Statistic</strong></th>
<th><strong>What to watch out for:</strong></th>
</tr>
</thead>
</table>
| **1. Shapiro Wilk (W)** | *Stata command*  
`. swilk yvariable*  

W is a measure of the correlation between the values in the sample and their associated normal scores (for review of Normal Scores, see BE540 Topic 5 – Normal Distribution)  

W = 1 under normality  

Null Hypothesis \( H_0\): \( y\text{variable} \) is distributed normal:  

Alternative Hypothesis \( H_A\): Not.  

Evidence of violation of normality is reflected in  

\( W < 1 \)  

small p-value |
| **2. Kolmogorov-Smirnov (D). See also Lilliefors (K-S)** | Evidence of violation of normality is reflected in  

\( D > 0 \)  

\( K-S > 0 \)  

small p-value |

Guidelines

In practice, the assessment of normality is made after assessment of other model assumption violations. The linear model is often more robust to violations of the assumption of normality. The cure, is often worse than the problem. (e.g. – transformation of the outcome variable)

Consider doing a scatterplot of the residuals. Look for

♦ Bell shaped pattern  
♦ Center at zero  
♦ No gross outliers
c. **Cook-Weisberg Test of Heteroscedasticity**

Recall what we are assuming with respect to homogeneity of variance:

- **In Simple Linear Regression:**
  At each level “x” of the predictor variable X, the outcomes Y are distributed normal with mean \( \mu_{Y|x} = \beta_0 + \beta_1 x \) and constant variance \( \sigma^2_{Y|x} \)

Evidence of a violation of homogeneity (this is heteroscedasticity) is seen when

- There is increasing or decreasing variation in the residuals with fitted \( \hat{Y} \)
- There is increasing or decreasing variation in the residuals with predictor X

**Some graphical assessments of homogeneity of variance and what to watch out for:**

<table>
<thead>
<tr>
<th>Method</th>
<th>What to watch out for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Plot Residuals or standardized residuals or studentized residuals on the vertical – versus - Predicted outcomes ( \hat{Y} ) on the horizontal</td>
<td>Look for even band at zero</td>
</tr>
<tr>
<td>2. Plot Residuals or standardized residuals or studentized residuals on the vertical – versus - Predictor values X</td>
<td>Look for even band at zero</td>
</tr>
</tbody>
</table>

**Hypothesis Test of homogeneity of variance is Cook-Weisberg**

<table>
<thead>
<tr>
<th>Cook-Weisberg Test</th>
<th>What to watch out for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>This test is based on a model of the variance as a function of the fitted values (or the predictor X). Specifically, it is a chi square test of whether the squared standardized residuals are linearly related to the fitted values (or the predictor X).</td>
<td>Evidence of violation of homogeneity of variance is reflected in Large test statistic &gt; 0 small p-value</td>
</tr>
</tbody>
</table>
d. The Method of Fractional Polynomials

This method is beyond the scope of this course. However, it’s helpful to understand its theory.

Goal: The goal is to select a “good” functional form that relates Y to X from a collection of candidate models. Candidates are lower polynomials and members of the Box-Tidwell family.

Fractional Polynomials: Instead of \( Y = \beta_0 + \beta_1 X \), we consider the following:

\[
Y = \beta_0 + \beta_1 X + \beta_2 X^{p_2} + \beta_3 X^{p_3} + \ldots + \beta_m X^{p_m}
\]

where

\[
m \quad \text{number of powers (“degree”)}
\]

\[
p_1, p_2, p_3, \ldots, p_m \quad \text{are choices from a special set of 8 candidate powers} = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}
\]

And where, when powers repeat,

E.g. - when \( p_2 = p_1 \) we consider \( \beta_1 X^{p_1} + \beta_2 X^{p_1} \log(X) \).

Example: Suppose \( m=1 \) with \( p_1 = 1 \). This yields

\[
Y = \beta_0 + \beta_1 X
\]

Example: Next, suppose \( m=2 \) with \( p_1 = 0.5 \) and \( p_2 = 0.5 \). Because \( p_2 = p_1 \) this yields

\[
Y = \beta_0 + \beta_1 \sqrt{X} + \beta_2 \sqrt{X} \ln(X)
\]
The Method of Fractional Polynomials - Continued

Guidelines

Competing models are assessed using a chi square statistic that compares the likelihoods of the data under each of the two models using what is called a “deviance” statistic.

Don’t worry: We will learn about the “deviance” statistic in Unit 5 in the context of the logistic regression model.

Search begins with examination of all models for which m=1. We choose the one model in this class that has the smallest deviance.

♦ We compare the best m=1 model to the specific model for which m=1 and p1 = 1 because the latter is the simple linear model.

♦ Thus, we are asking whether it is really necessary to abandon the simple linear model.

Next, we compare the best m=1 model to the best m=2 model. And so on …

♦ In general, we must choose between two costs:

1) A smaller model has a lower goodness-of-fit but more generalizability

2) A larger model has a higher goodness-of-fit but less generalizability

♦ Our goal is to choose the smallest model for which the goodness-of-fit is acceptable.
e. Ramsey Test for Omitted Variables

A fitted model that fails to include an important explanatory variable is problematic.

- Our understanding of the outcomes is incomplete.
- Estimated associations may be biased due to confounding.
- Model assumptions may be violated.

Method of the Ramsey Test

- H₀: Predicted values from the fitted model are unrelated to powers of the fitted model, after adjustment for the predictor variables in the model.

\[
\text{corr} (\hat{Y}, \hat{Y}^p) = 0
\]

- For example, we might fit the model \( \hat{Y} = \beta_0 + \beta_1 \hat{Y} + \beta_2 \hat{Y}^2 + \beta_3 X + \text{error} \) and test the significance of \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \).

- The test statistic is an F statistic.

Guidelines

Evidence of a failure to include one or more explanatory variables is reflected in a large F statistic value.

As a suggestion, do also a scatterplot of the squared standardized residuals versus the leverage values. Omission of an important explanatory variables is suggested by

- Extreme values
- Any systematic pattern
f. Residuals, Leverage, and Cook’s Distance

Residuals - There are multiple measures of “residual”.

<table>
<thead>
<tr>
<th>Ordinary residual</th>
<th>Standardized residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e = (Y - \hat{Y}) )</td>
<td>( e^* = \frac{e}{\sqrt{ms(residual)}} = \frac{e}{\sqrt{\hat{\sigma}^2_{Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Studentized residual</th>
<th>Jacknife residual, also called Studentized deleted residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e^* = \frac{e}{\sqrt{ms(residual)} \sqrt{1-h}} = \frac{e}{\sqrt{\hat{\sigma}^2_{Y</td>
<td>x}} \sqrt{1-h}} )</td>
</tr>
</tbody>
</table>

Which one or ones should we use?

- Standardized residuals can be appreciated as we do z-scores.
- Studentized residuals are distributed Student’s t (df=n-p-1) when regression assumptions hold.
- Jacknife residuals are distributed Student’s t (df=n-p-2) when regression assumptions hold. These also have the advantage of correcting the magnitude of the \( \sqrt{MS(residual)} \) when it is otherwise too big because of the effects of influential points.

Leverage, \( h \):

Leverage is the distance of a predictor value \( X=x \) from the center of the values of the predictor value \( \bar{X} = \bar{x} \).

This distance is denoted \( h_i \).

For simple linear regression, \( h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \)

For simple linear regression, a “large” leverage value is \( h_i \geq \frac{4}{n} \)
Cook’s Distance, $d$

Recall from our pictures that neither a large residual alone nor a high leverage determine the influence of an individual data point.

Cook’s distance provides a measure of the influence of an individual data point on the fitted model and is a function of the values of both the residual and leverage:

$$\text{Cook’s Distance} = \text{Change in estimated regression coefficient value, expressed in standard error units.}$$

1) For simple linear regression

$$d = \frac{e^2 h}{2s^2 (1 - h)^2}$$

2) For multivariable linear regression models

$$d_i = \frac{(\hat{\beta}_{-i} - \hat{\beta})' (X'X)(\hat{\beta}_{-i} - \hat{\beta})}{p' s_{Y|x}^2}$$

where

- $i$ indexes the individual for which measure of influence is sought
- $\hat{\beta}$ = vector of estimated regression coefficients using the entire sample
- $\hat{\beta}_{-i}$ = vector of estimated regression coefficients with omission of the $i^{th}$ data point
- $X$ = matrix of values of the predictor variables
- $p'$ = rank ($X$) = number of predictors + 1

**Guidelines**

- For the linear regression model, a “noteworthy” influential data point is one for which $d \geq 1$.

For a multivariable regression model, a “noteworthy” influential data point is one for which $d \geq 2(p+1)/n$ where $p=$# predictors.
Example - continued

Framingham Study

* * * * * 
Plot of Observed v Predicted (LOOK FOR: Points along a straight line)  
* Command predict to create a new variable=ypredicted that contains the predicted Y values  
. predict ypredicted  
(option xb assumed; fitted values)  
(6 missing values generated)

. graph twoway (scatter ypredicted ln_sbp, symbol(d) msize(vsmall)) (lfit ypredicted ln_sbp)  
(lfitci ypredicted ln_sbp), title("Model Assessment") subtitle("Plot of Observed versus Predicted")  
xtitle("Observed Y = ln(sbp)") ytitle("Predicted Y=ln(sbp)") xlabel(4.5(.25)5.5) ylabel(4.5(.25)5.5)

Interpretation – Not bad! Ideally the scatter lies on the line defined by 45 degrees. We expect some widening  
of the confidence intervals at the ends of the range but not too much. What we see here is reasonable.
--- Normality of residuals (LOOK FOR: Points lying on the line)
* Command predict with option resid to create yresidual that contains the residuals
predict yresidual, resid
(6 missing values generated)

* pnorm yresidual, msize(vsmall) title("Model Assessment") subtitle("Std Normal Plot of Residuals")

Model Assessment
Std Normal Plot of Residuals

Interpretation – Here too, we hope to see a scatter on the 45 degree line. Not bad!

* sfrancia test of normality (NULL: residuals are distributed normal)
sfrancia yresidual

Shapiro-Francia W' test for normal data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>W'</th>
<th>V'</th>
<th>z</th>
<th>Prob&gt;z</th>
</tr>
</thead>
<tbody>
<tr>
<td>yresidual</td>
<td>994</td>
<td>0.97683</td>
<td>15.434</td>
<td>6.271</td>
<td>0.00001</td>
</tr>
</tbody>
</table>

--- 2 Tests of Model Misspecification
* --- LINK test (Null: No misspecification. _htsq is NOT significant)
linktest

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 994</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7.02566005</td>
<td>2</td>
<td>3.51283003</td>
<td>F(2, 991) = 180.45</td>
</tr>
<tr>
<td>Residual</td>
<td>19.2921224</td>
<td>991</td>
<td>.019467328</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>26.3177825</td>
<td>993</td>
<td>.026503306</td>
<td>R-squared = 0.2670</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.2655</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = .13953</td>
</tr>
</tbody>
</table>

| ln_sbp     | Coef.     | Std. Err. | t     | P>|t|  [95% Conf. Interval] |
|------------|-----------|-----------|-------|-------|------------------------|
| _ht | -3.398892 | 4.165516  | -0.82 | 0.415 | -11.57314 4.775353     |
| _htsq | .4506253  | .4266839  | 1.06  | 0.291 | -.3866825 1.287933     |
| _cons | 10.73201  | 10.16508  | 1.06  | 0.291 | -9.215532 30.67956     |

Interpretation – Predictor htsq is NOT significant (p=.29) suggesting we’re okay!
. * ---- OMITTED VARIABLES TES (NULL: no variables omitted. All is well)  
. ovtest

Ramsey RESET test using powers of the fitted values of ln_sbp  
Ho: model has no omitted variables  
\[ F(3, 985) = 1.81 \]  
Prob > F = 0.1442

**Interpretation** - Ramsey test is also NOT significant (p=.14) suggesting we’re okay!

. * ---- Assessment of Multicollinearity (All is well if VIF <10 OR 1/VIF > 0.10)  
. vif

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
<th>1/VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_female</td>
<td>34.12</td>
<td>0.029311</td>
</tr>
<tr>
<td>female</td>
<td>32.39</td>
<td>0.030869</td>
</tr>
<tr>
<td>age</td>
<td>2.38</td>
<td>0.420495</td>
</tr>
<tr>
<td>ln_scl</td>
<td>1.18</td>
<td>0.850679</td>
</tr>
<tr>
<td>ln_bmi</td>
<td>1.12</td>
<td>0.896450</td>
</tr>
</tbody>
</table>

Mean VIF | 14.24

**Interpretation** - age_female and female appear to be collinear suggesting some concern about the extent to which there is adequacy of range of age in the 2 genders.

. * ---- Cooks Distances (LOOK FOR: all to be less than 4/N)  
. * Command predict with option cooksd to create cook that contains the Cook’s distances  
. predict cook, cooksd  
(6 missing values generated)

. * Command generate subject=_n to create subject id for nice plotting on x-axis  
. generate subject=_n

. graph twoway (scatter cook subject, symbol(d) msize(vsmall)), title("Model Assessment")  
subtitle("Plot of Cook Distances")

**Interpretation** - Looks fine!
. * ----- Predictions
. * Again -- this assumes that you have just fit your "final" model

. * ----- Predicted Y=ln(sbp) with increasing X = ln(bmi)
. margins, at(ln_bmi=(2.6(.2)3.8)) vsquish

Predictive margins Number of obs = 994
Model VCE : OLS
Expression : Linear prediction, predict()
1._at : ln_bmi =  2.6
2._at : ln_bmi =  2.8
3._at : ln_bmi =  3.0
4._at : ln_bmi =  3.2
5._at : ln_bmi =  3.4
6._at : ln_bmi =  3.6
7._at : ln_bmi =  3.8

|            Delta-method
|  Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
|-----------------------------------------------------------------------
_at |                      |            |            |            |
 1 |  4.701072   .0205758   228.48   0.000     4.660695    4.741449
 2 |  4.755225   .0144203   329.76   0.000     4.726927    4.783523
 3 |  4.809378   .0085848   560.22   0.000     4.792531    4.826224
 4 |  4.863531   .0045417  1070.87   0.000     4.854618    4.872443
 5 |  4.917664   .009805   704.49   0.000     4.903985    4.931382
 6 |  4.971836   .0186667   269.25   0.000     4.94262   5.001052
 7 |  5.025989   .0206667   269.25   0.000     4.989359    5.062622

.marginsplot, recast(line) recastci(rarea)