Chapter 10: Inferential Tools for Multiple Regression

Prof. Sharyn O’Halloran
Sustainable Development U9611
Econometrics II
What Is Inference About?

- “Statisticians are people whose aim in life is to be wrong exactly 5% of the time.”

- Inference relates estimation results to the hypotheses being tested.
  - Is the coefficient on a single variable significant?
  - Are the coefficients on a group of variables jointly significant?
  - How much of the variance in the data is explained by a given regression model?

- Regression **interpretation** is about the mean of the coefficients; **inference** is about their **variance**.
Echo location requires more energy in-flight

Is $b_3$ significant? Positive, negative? Magnitude?

$EE_i = b_0 + b_1 \cdot \text{mass}_i + b_2 \cdot \text{bird}_i + b_3 \cdot \text{e-bat}_i + \text{resid}_i$

Echo-locating bats expend more energy while flying per unit body mass

Data: energy expenditures and mass for 4 ne-bats, 4 e-bats, and 12 ne-birds.
Example: Bat Echolocation Data

Q: Do echolocating bats expend more energy than non-echolocating bats and birds, after accounting for mass?
Note: Different Model Parameterizations

- The variable TYPE has 3 levels: birds, e-bats, and ne-bats.
- We have a choice about which of the 3 indicator variables to use
  - If we include 2 indicator variables, the omitted category becomes equal to the constant.
  - i.e. \( \mu(y|x,\text{TYPE}) = \beta_0 + \beta_1 x + (\beta_2 I_{\text{type2}} + \beta_3 I_{\text{type3}}) \)
- Then Type 1 becomes the reference level
  - \( \beta_2 \) and \( \beta_3 \) indicate the difference between Type 1 and Types 2 and 3, respectively.
Generate dummy variables with STATA:

Type category variable:
```
encode type,
genrate(typedum)
```
- Typedum=1 NE bats
- Typedum=2 NE birds
- Typedum=3 E bats

Generate three dummies:
- Type1 NE bats
- Type2 NE birds
- Type3 E bats

```
.tab typedum

<table>
<thead>
<tr>
<th>Typedum</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>20.00</td>
<td>20.00</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>60.00</td>
<td>80.00</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>20.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
```

```
.gen type1=typedum if typedum==1
(16 missing values generated)

.gen type2=typedum if typedum==2
(8 missing values generated)

.gen type3=typedum if typedum==3
(16 missing values generated)
```
Generate dummy variables with STATA:

Label the new dummy variables

label variable type1 "non-echolocating bats"
label variable type2 "non-echolocating birds"
label variable type3 "echolocating bats"

. d type1 type2 type3

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage type</th>
<th>display format</th>
<th>value label</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>type1</td>
<td>float</td>
<td>%9.0g</td>
<td>non-echolocating bats</td>
<td></td>
</tr>
<tr>
<td>type2</td>
<td>float</td>
<td>%9.0g</td>
<td>non-echolocating birds</td>
<td></td>
</tr>
<tr>
<td>type3</td>
<td>float</td>
<td>%9.0g</td>
<td>echolocating bats</td>
<td></td>
</tr>
</tbody>
</table>

. tab type1

<table>
<thead>
<tr>
<th>non-echolocating bats</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16</td>
<td>80.00</td>
<td>80.00</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>20.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

. tab type2

<table>
<thead>
<tr>
<th>non-echolocating birds</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>40.00</td>
<td>40.00</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>60.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

. tab type3

| echolocating |
Dummy variables as shift parameters

$$
\mu(y \mid x, \text{TYPE}) = \beta_0 + \beta_1 \text{mass} + (\beta_2 I_{\text{type2}} + \beta_3 I_{\text{type3}})
$$

![Diagram showing energy vs body mass for different categories: Echococating bats, Birds, Non-Echococating bats, with dummy variables as shift parameters.](image-url)
Dummy variables as shift parameters

In the previous model:

- $\beta_0$ is the intercept for level 1,
- $\beta_2$ is the amount by which the mean of $y$ is greater for level 2 than for level 1 (after accounting for $x$),
- $\beta_3$ is the amount by which the mean of $y$ is greater for level 3 than for level 1 (Display 10.5).

```
.reg  lenergy lmass type2 type3

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29.4214818</td>
<td>3</td>
<td>9.80716059</td>
</tr>
<tr>
<td>Residual</td>
<td>.553317657</td>
<td>16</td>
<td>.034582354</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>

Number of obs = 20
F(  3,    16) = 283.59
Prob > F = 0.0000
R-squared = 0.9815
Adj R-squared = 0.9781
Root MSE = .18596

| lenergy | Coef.  | Std. Err. | t     | P>|t| |  [95% Conf. Interval] |
|---------|--------|-----------|-------|------|----------------------|
| lmass   | .8149575 | .0445414 | 18.30 | 0.000 | .7205338 - .9093811 |
| type2   | .1022618 | .1141827 | 0.90  | 0.384 | -.1397946 - .3443182 |
| type3   | .0786636 | .2026793 | 0.39  | 0.703 | -.3509973 - .5083245 |
| _cons   | -1.57636 | .2872364 | -5.49 | 0.000 | -2.185274 - .9674459 |
```
Another parameterization is:

\[ \mu(y|x, \text{TYPE}) = \beta_1 x + (\beta_2 I_{\text{type1}} + \beta_3 I_{\text{type2}} + \beta_4 I_{\text{type3}}) \]

- In this model, there is no \( \beta_0 \); \( \beta_2 \), \( \beta_3 \) and \( \beta_4 \) are the intercepts for types 1, 2, and 3, respectively.
- We see that the coefficient on \( \beta_2 \) is, indeed, the constant from the previous regression.
  - And the other coefficients are shifted accordingly.
Another parameterization is:

\[
\mu(\ y|\ x,\ \text{TYPE}) = \beta_1 x + (\beta_2 I_{\text{type1}} + \beta_3 I_{\text{type2}} + \beta_4 I_{\text{type3}})
\]

- In this model, there is no \( \beta_0 \); \( \beta_2, \beta_3 \) and \( \beta_4 \) are the intercepts.
- We see that the coefficient on \( \beta_2 \) is, indeed, the constant from the previous regression.
  - And the other coefficients are shifted accordingly.

![Regression Output](image)
Now that we know what the coefficients mean, how do we test hypotheses?

E.g., how can we tell if the value of a coefficient is different from 0?
Simple and Multiple Regression Compared

- Coefficients in a *simple* regression pick up the impact of that variable (plus the impacts of other variables that are correlated with it) and the dependent variable.

- Coefficients in a *multiple* regression account for the impacts of the other variables in the equation.
Simple and Multiple Regression Compared: Example

- Two simple regressions:
  - Oil = $\beta_0 + \beta_1$ Temp + $\varepsilon_i$
  - Oil = $\beta_0 + \beta_1$ Insulation + $\varepsilon_i$

- Multiple regression:
  - Oil = $\beta_0 + \beta_1$ Temp + $\beta_2$ Insulation + $\varepsilon_i$
Least Squares Estimation

\[
\mu(y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad \text{var}(y|X_1, X_2) = \sigma^2
\]

<table>
<thead>
<tr>
<th>Unknown parameters:</th>
<th>Regression coefficients</th>
<th>Variance about regression</th>
</tr>
</thead>
</table>

**Fitted values** (predicted)

\[
\mu(y | x_1, x_2) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \quad i = 1,2,..n
\]

**Residuals**

\[
res_i = y_i - \hat{y}_i
\]

**Least squares estimators**, \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \), are chosen to minimize the sum of squared residuals (matrix algebra formula)

\[
\hat{\sigma}^2 = \frac{\text{(Sum of squared residuals)}}{(n-p)} \quad \text{[p= number of \( \beta \)s]}
\]
t-tests and CI’s for individual $\beta$’s

1. Note: a matrix algebra formula for $SE(\hat{\beta}_j)$ is also available

2. If distribution of Y given X’s is normal, then

$$t\text{-ratio} = \frac{\hat{\beta}_j - \beta_j}{SE (\hat{\beta}_j)}$$

has a t-distribution on n-p degrees of freedom

3. For testing the hypothesis $H_0: \beta_2 = 7$; compare

$$t\text{-stat} = \frac{\hat{\beta}_2 - 7}{SE(\hat{\beta}_2)}$$

to a t-distribution on n-p degrees of freedom.

4. The p-value for the test $H_0: \hat{\beta}_j = 0$ is standard output
5. It’s often useful to think of $H_0: \beta_2 = 0$ (for example) as

Full model: $\mu( y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
Reduced model: $\beta_0 + \beta_1 X_1 + \beta_3 X_3$

Q: Is the $\beta_2 X_2$ term needed in a model with the other $x$’s?

6. 95% confidence interval for $\beta_j$:

$$\hat{\beta}_j \pm t_{n-p} = (.975) \times SE (\hat{\beta}_j)$$

7. The meaning of a coefficient (and its significance) depends on what other $X$’s are in the model (Section 10.2.2)

8. The t-based inference works well even without normality
t-tests and CI’s for Bat Data (From Display 10.6)

1. Question: Do echolocating bats spend more energy than nonecholocating bats?

2. This is equivalent to testing the hypothesis \( H_0: \beta_3=0 \)

![Image of regression output](image-url)
1. Question: Do echolocating bats spend more energy than nonecholocating bats?

2. This is equivalent to testing the hypothesis $H_0: \beta_3 = 0$.

---

```
reg lenergy lmass type2 type3

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29.4214818</td>
<td>3</td>
<td>9.80716059</td>
</tr>
<tr>
<td>Residual</td>
<td>.553317657</td>
<td>16</td>
<td>.034582354</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>

Number of obs = 20  
P(3, 16) = 283.59  
Prob > F = 0.0000  
R-squared = 0.9815  
Adj R-squared = 0.9781  
Root MSE = 0.18596

| lenergy | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|---------|-------|-----------|-------|-----|---------------------|
| lmass   | .8149575 | .0445414  | 18.30 | 0.000 | .7205338 - .9093811 |
| type2   | .1022618 | .1141827  | 0.90  | 0.384 | -.1397946 - .3443182 |
| type3   | .0786636 | .2026793  | 0.39  | 0.703 | -.3509973 - .5083245 |
| _cons   | -1.57636 | .2872364  | -5.49 | 0.000 | -2.185274 - .9674459 |
```
1. Question: Do echolocating bats spend more energy than nonecholocating bats?

2. This is equivalent to testing the hypothesis $H_0: \beta_3 = 0$.
1. Results: The data are consistent with the hypothesis of no energy differences between echolocating and non-echolocating bats, after accounting for body size
   - Confidence interval contains 0
   - 2-sided p-value = .7; i.e., not significant at the 5% level
   - So we cannot reject the null hypothesis that $\beta_3 = 0$

2. However, this doesn’t prove that there is no difference. A “large” p-value means either:
   (i) there is no difference ($H_0$ is true) or
   (ii) there is a difference and this study is not powerful enough to detect it

3. So report a confidence interval in addition to the p-value:
   95% CI for $\beta_3$: .0787 ± 2.12*.2027 = (-.35,.51).
Back-transform:
\[ e^{0.787} = 1.08, \ e^{-0.35} = 0.70 \text{ and } e^{0.51} = 1.67 \]

It is estimated that the median energy expenditure for echolocating bats is 1.08 times the median for non-echolocating bats of the same body weight

(95% confidence interval: 0.70 to 1.67 times).
If we eliminate one of the independent variables (lmass), the other coefficients change
So regression results depend on the model specification
Here, we do not control for body mass, as we did before, and $\beta_3$ becomes negative and significant!
Interpretation Depends…

- Ne-bats are clearly much bigger than e-bats.
- So the they naturally use more energy
  - Not necessarily due to the energy demands of echolocation
Explaining Model Variance

- Instead of examining a single coefficient, analysts often want to know how much variation is explained by all regressors.
  - This is the “coefficient of multiple determination,” better known as $R^2$.
  - Recall that:

$$SST = SSR + SSE$$

- Total Deviation
- Explained Deviation
- Unexplained Deviation
Calculating $R^2$

- Without any independent variables, we would have to predict values of $Y$ by using only its mean:

  Full model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

  Reduced model: $\beta_0$

\[
R^2 = \frac{SSR}{SST} = \frac{\text{Explained Variation}}{\text{Total Variation}}
\]

- $R^2 = \text{proportion of total variability (about } Y)\text{ that is explained by the regression}$

- **Extreme Cases**
  - $R^2 = 0$ if residuals from full and reduced model are the same (the independent variables provide no additional information about $Y$)
  - $R^2 = 1$ if residuals from full model are all zero (the independent variables perfectly predict $Y$)
Calculating $R^2$

- $R^2$ can help, somewhat, with practical significance (bat data)
  - $R^2$ from model with $x_1$, $x_2$ and $x_3$ : .9815
  - $R^2$ from model with $x_2$ and $x_3$ : .5953

- So $x_1$ explains an extra 67% of the variation in $y$ compared to a model with only $x_2$ and $x_3$. 

```
. reg lenergy lmass type2 type3
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29.4214818</td>
<td>3</td>
<td>9.80716059</td>
</tr>
<tr>
<td>Residual</td>
<td>.553317657</td>
<td>16</td>
<td>.034582354</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>

```
. reg lenergy type2 type3
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>17.8444807</td>
<td>2</td>
<td>8.92224034</td>
</tr>
<tr>
<td>Residual</td>
<td>12.1303187</td>
<td>17</td>
<td>.713548161</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>
Limits of $R^2$

- $R^2$ cannot help with
  - Model goodness of fit,
  - Model adequacy,
  - Statistical significance of regression, or
  - Need for transformation.

- It can only help in providing a summary of tightness of fit;
  - Sometimes, it can help clarify *practical* significance.

- $R^2$ can always be made 100% by adding enough terms
Example: Zodiac and Sunshine

- Add two irrelevant variables to bat regression
  - Zodiac sign of month that bat/bird was born
  - Whether they were born on a sunny day
  - (Just to be sure, these were filled in randomly.)
- Even so, $R^2$ increases from 0.9815 to 0.9830

\begin{verbatim}
. reg lenergy lmass type2 type3

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29.4214818</td>
<td>3</td>
<td>9.80716059</td>
</tr>
<tr>
<td>Residual</td>
<td>.553317657</td>
<td>16</td>
<td>.034582354</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>

. reg lenergy lmass type2 type3 zodiac sunshine

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29.4664969</td>
<td>5</td>
<td>5.89329938</td>
</tr>
<tr>
<td>Residual</td>
<td>.508302494</td>
<td>14</td>
<td>.036307321</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>
\end{verbatim}
Adjusted $R^2$

- Proportion of variation in $Y$ explained by all $X$ variables, adjusted for the number of $X$ variables used and sample size

$$r_{adj}^2 = 1 - \left( 1 - r_{Y\cdot12\ldots k}^2 \right) \frac{n - 1}{n - k - 1}$$

- Penalizes Excessive Use of Independent Variables
- Smaller than $R^2$
- Useful in Comparing among Models
Example Regression Output

\[ r_{Y \cdot 12}^2 = \frac{SSR}{SST} \]

Adjusted R²

- Reflects the number of explanatory variables and sample size
- Is smaller than R²
Interpretation of Adjusted $R^2$

- $r_{Y_{12}}^2 = \frac{SSR}{SST} = .5953$
  - 59.53% of the total variation in energy can be explained by types 1 and 2

- $r_{adj}^2 = .5477$
  - 54.77% of the total fluctuation in energy expenditure can be explained by types 1 and 2 after adjusting for the number of explanatory variables and sample size
Example: Zodiac and Sunshine

- Recall that $R^2$ increases from 0.9815 to 0.9830 with the addition of two irrelevant variables.
- But the adjusted $R^2$ falls from 0.9781 to 0.9770

```
. reg lenergy lmass type2 type3

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29.4214818</td>
<td>3</td>
<td>9.80716059</td>
</tr>
<tr>
<td>Residual</td>
<td>0.55317657</td>
<td>16</td>
<td>0.034582354</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>

. reg lenergy lmass type2 type3 zodiac sunshine

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29.4664969</td>
<td>5</td>
<td>5.89329938</td>
</tr>
<tr>
<td>Residual</td>
<td>0.508302494</td>
<td>14</td>
<td>0.036307321</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>
```
The overlap (purple) is the variation in Y explained by independent variable X (SSR).

Think of this as information used to explain Y.
Example: Oil Use & Temperature

Variations in Temp not used in explaining variation in Oil

Variations in Oil explained by Temp, or variations in Temp used in explaining variation in Oil \( (SSR) \)

Variations in Oil explained by the error term \( (SSE) \)
Example: Oil Use & Temperature

\[ R^2 = \frac{SSR}{SSR + SSE} \]
Example: $R^2 = 0$ and $R^2 = 1$
Uncorrelated Independent Variables

Here, two independent variables that are uncorrelated with each other.

But both affect oil prices.

Then $R^2$ is just the sum of the variance explain by each variable.
Uncorrelated Independent Variables

\[ R^2 = \frac{SSR}{SSR + SSE} \]
Now each explains some of the variation in $Y$

But there is some variation explained by both $X$ and $W$ (the Red area)
Venn Diagrams and Explanatory Power of Regression

Variation *NOT* explained by Temp nor Insulation ($SSE$)

Variation explained by Temp and Insulation ($SSR$)
Venn Diagrams and Explanatory Power of Regression

\[ r_{Y \cdot 12}^2 = \frac{SSR}{SSR + SSE} \]
F-tests: Overall Model Significance

To calculate the significance of the entire model, use an F-test

This compares the added variance explained by including the model’s regressors, as opposed to using only the mean of the dependent variable:

Full model: \( \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \)
Reduced model: \( \beta_0 \)

i.e. in full model, \( H_0: \beta_1=\beta_2=\beta_3=0 \)

Extra SS = (SSR from full model) - (SSR from reduced model)

\[
F - statistic = \frac{[\text{Extra SS} / \text{Extra \# of } \beta\text{'s}]}{\hat{\sigma}^2_{\text{full}}}
\]
1. **Fit full model:**

\[
\mu(y|x, \text{TYPE}) = \beta_0 + \beta_1 \text{mass} + \beta_2 \text{I}_{\text{type2}} + \beta_3 \text{I}_{\text{type3}}
\]

- This is the ANOVA section of the regression output
- It has all the information needed to calculate the F-statistic
1. **Fit full model:**

\[
\mu(y|x, \text{TYPE}) = \beta_0 + \beta_1 \text{mass} + \beta_2 \text{I}_{\text{type2}} + \beta_3 \text{I}_{\text{type3}}
\]

Sum of Squared Residuals = 0.55332

F-tests: Example
F-tests: Example

1. Fit full model:
   \[
   \mu(y|x, \text{TYPE}) = \beta_0 + \beta_1 \text{ mass} + \beta_2 I_{\text{type2}} + \beta_3 I_{\text{type3}}
   \]

   Sum of Squared Residuals = 0.55332
   Degrees of freedom = 16
1. Fit full model:

\[ \mu(y|x,\text{TYPE}) = \beta_0 + \beta_1 \text{mass} + \beta_2 \text{I}_{\text{type2}} + \beta_3 \text{I}_{\text{type3}} \]

Sum of Squared Residuals = 0.55332
Degrees of freedom = 16
Mean Squared Error = 0.03458
F-tests: Example

2. Fit reduced model:
   \( \mu(y|x, \text{TYPE}) = \beta_0 \)

Notice that the coefficient on the constant
2. Fit reduced model:
\[ \mu(y|x,\text{TYPE}) = \beta_0 \]

Notice that the coefficient on the constant = mean of Y
2. Fit reduced model:
\[ \mu(y|x,\text{TYPE}) = \beta_0 \]

```
. reg lenergy constant, nocons
```

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>123.226402</td>
<td>1</td>
<td>123.226402</td>
<td>F( 1, 19) = 78.11</td>
</tr>
<tr>
<td>Residual</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>153.201201</td>
<td>20</td>
<td>7.6600606</td>
<td>R-squared = 0.8043</td>
</tr>
</tbody>
</table>

| lenenergy | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|-----------|-------|-----------|---|------|---------------------|
| constant  | 2.482201 | 0.2808577 | 8.84 | 0.000 | 1.894359 | 3.070043 |

```

```
. sum lenergy
```

```
<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>lenergy</td>
<td>20</td>
<td>2.482201</td>
<td>1.256034</td>
<td>0.0198026</td>
<td>3.777348</td>
</tr>
</tbody>
</table>
```

Sum of Squared Residuals = 29.97
F-tests: Example

2. Fit reduced model:
   \[ \mu(y|x,\text{TYPE}) = \beta_0 \]

```
. reg lenenergy constant, nocons

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>133.226402</td>
<td>1</td>
<td>123.226402</td>
<td>F(1, 19) = 78.11</td>
</tr>
<tr>
<td>Residual</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>153.201201</td>
<td>20</td>
<td>7.66006006</td>
<td>R-squared = 0.8043</td>
</tr>
</tbody>
</table>

| lenenergy | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----------|--------|-----------|-------|------|-----------------------|
| constant  | 2.482201 | .2808577  | 8.84  | 0.000 | 1.894359 3.070043 |

. sum lenenergy

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>lenenergy</td>
<td>20</td>
<td>2.482201</td>
<td>1.256034</td>
<td>.0198026</td>
<td>3.777348</td>
</tr>
</tbody>
</table>
F-tests: Example

3. The extra sum of squares is the difference between the two residual sum of squares

Extra SS = 29.97 - 0.5533 = 29.42

4. Numerator degrees of freedom: \( \# \) of \( \beta \)'s in the full model - \( \# \) of \( \beta \)'s in the reduced model

Numerator d.f. = 19 - 16 = 3

5. Calculate the F-statistic

\[
F - \text{statistic} = \frac{29.42}{0.03458} = 283.56
\]

6. Find \( \text{Pr}(F_{3,16} > 283.56) \) from table or computer

P-value = 0.0000
F-tests: Example

Check against regression output:

Sure enough, the results agree!
Contribution of a Subset of Independent Variables

- We often want to test the significance of a subset of variables, rather than one or all.
  - For instance, does the type of animal (e-bat, ne-bat, bird) have any impact on energy use?
- Let $X_S$ be the subset of independent variables of interest
  - Then the extra variation explained by $X_S$ is:
    \[
    SSR\left(X_s \mid \text{all others except } X_s\right)
    = SSR\left(\text{all}\right) - SSR\left(\text{all others except } X_s\right)
    \]
So we want to test whether $X_s$ explains a significant amount of the variation in $Y$

Hypotheses:

1. $H_0$: Variables $X_s$ do not significantly improve the model given all others variables included
2. $H_1$: Variables $X_s$ significantly improve the model given all others included

Note: If $X_S$ contains only one variable, then the F-test is equivalent to the t-test we performed before.
Example: Bat Data

For the bat data, to test whether type of animal makes a difference, we have:

Full model: \[ \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \]

Reduced model: \[ \beta_0 + \beta_1 x_1 \]

\[ H_0: \beta_2 \& \beta_3 \text{ are not jointly significant} \]

\[ H_0: \beta_2 \& \beta_3 \text{ are jointly significant} \]

The test statistic is essentially the same as before:

\[ F - statistic = \frac{[\text{Extra SS} / \text{Extra # of } \beta's]}{\hat{\sigma}^2_{\text{full}}} \]

The only difference is that the Extra SS comes from adding \( x_2 \) and \( x_3 \) to the reduced model.
1. Fit full model:

\[ \mu(y|x, \text{TYPE}) = \beta_0 + \beta_1 \text{mass} + \beta_2 I_{\text{type2}} + \beta_3 I_{\text{type3}} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29.4214818</td>
<td>3</td>
<td>9.80216059</td>
</tr>
<tr>
<td>Residual</td>
<td>0.55331765?</td>
<td>16</td>
<td>0.034582354</td>
</tr>
<tr>
<td>Total</td>
<td>29.9747994</td>
<td>19</td>
<td>1.57762102</td>
</tr>
</tbody>
</table>

Sum of Squared Residuals = 0.55332
Degrees of freedom = 16
Mean Squared Error = 0.03458
2. Fit reduced model:

\[ \mu(y|x, \text{TYPE}) = \beta_0 + \beta_1 \text{ mass} \]

Sum of Squared Residuals = 0.5829
Degrees of freedom = 18
Testing Subsets: Example

3. The extra sum of squares is the difference between the two residual sum of squares

Extra SS = .5829 - .5533 = .0296

4. Numerator degrees of freedom: # of β’s in the full model - # of β’s in the reduced model

Numerator d.f. = 18 – 16 = 2

5. Calculate the F-statistic

\[ F - statistic = \frac{.0296}{.03458} = 0.43 \]

6. Find Pr(F_{2,16}>0.43) from table or computer

P-value = 0.659
Testing Subsets: Example

Check against regression output:

The results agree again…
Testing Subsets: Example

Check against regression output:

Note that this is easy to do in Stata