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 <u>interpretation</u> is about the mean of the coefficients; <u>inference</u> is about their variance.

Example: Bat Echo-Location Data

Echo location requires more energy in-flight

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

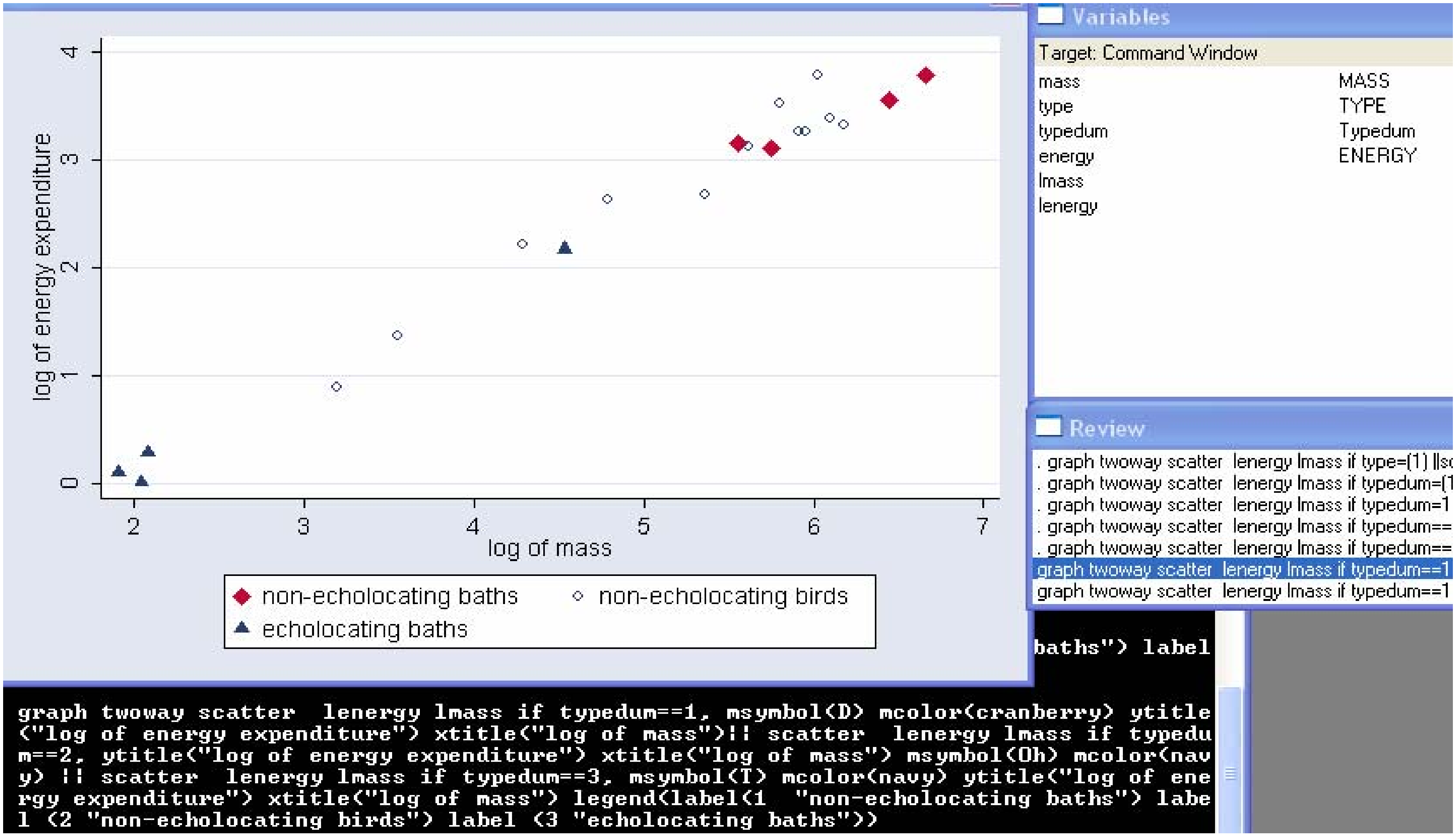
Data: energy expenditures and mass for 4 ne-bats, 4 e-bats, and 12 ne-birds.

$EE_i = b_0 + b_1{*}mass_i + b_2{*}bird_i + b_3{*}e\text{-}bat_i + resid_i$

Is b_3 significant? Positive, negative? Magnitude?

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. $\mu(y|x,\text{TYPE}) = \beta_0 + \beta_1 x + (\beta_2 I_{type2} + \beta_3 I_{type3})$
- Then Type 1 becomes the <u>reference</u> level
  - β_2 and β_3 indicate the <u>difference</u> between Type 1 and Types 2 and 3, respectively.

Generate dummy variables with STATA:

Type category variable:
encode type,
generate(typedum)

- Typedum=1 NE bats
- Typedum=2 NE birds
- Typedum=3 E bats

tab typedum

Typedum	Freq.	Percent	Cum.
1	4	20.00	20.00
2	12	60.00	80.00
3	4	20.00	100.00
Total	20	100.00	

Generate three dummies:

- Type1 NE bats
- Type2 NE birds
- Type3 E bats

```
. 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)
```

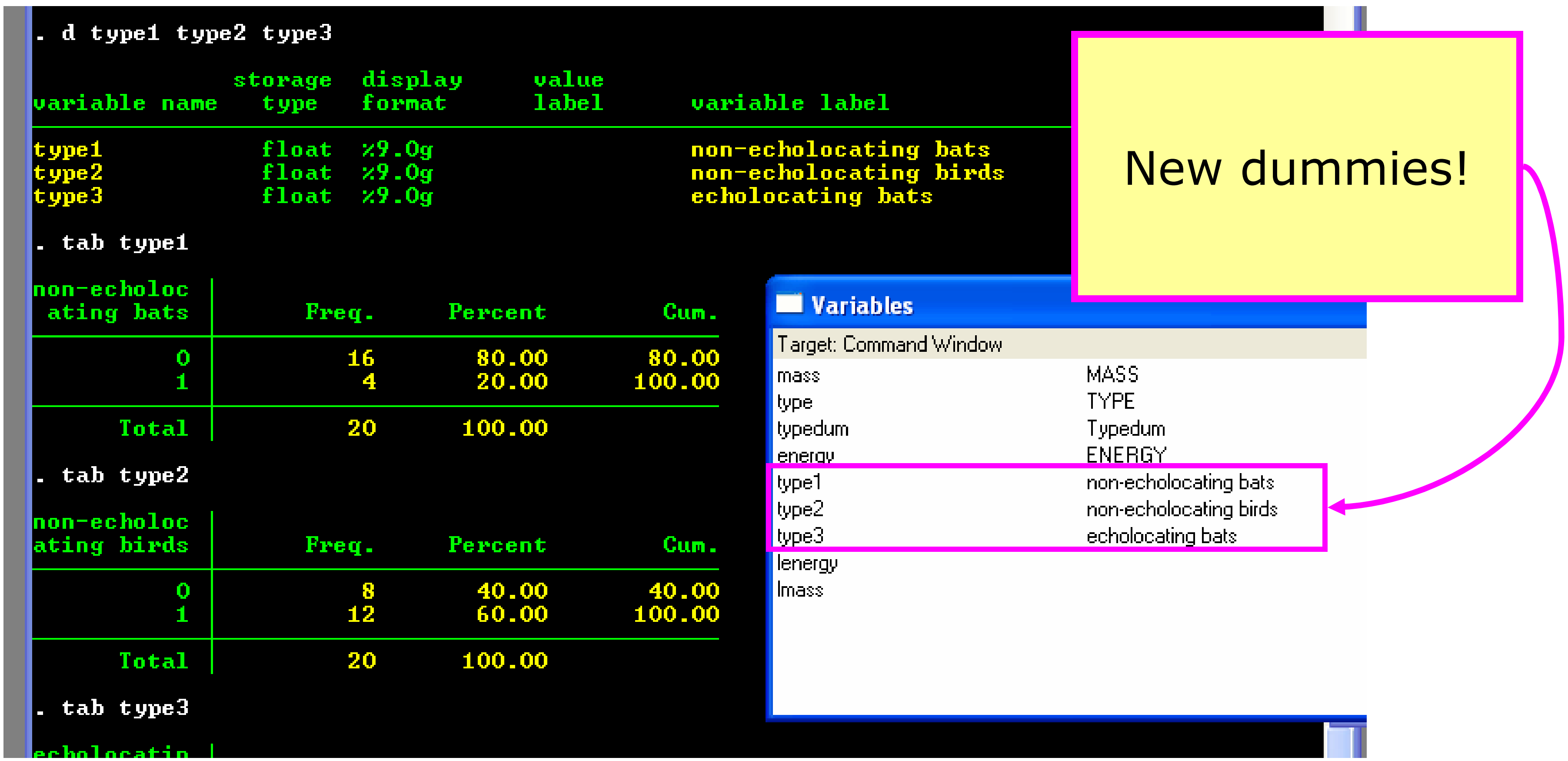
Generate dummy variables with STATA: Continued…

Label the new dummy variables

label variable type1 "non-echolocating bats"

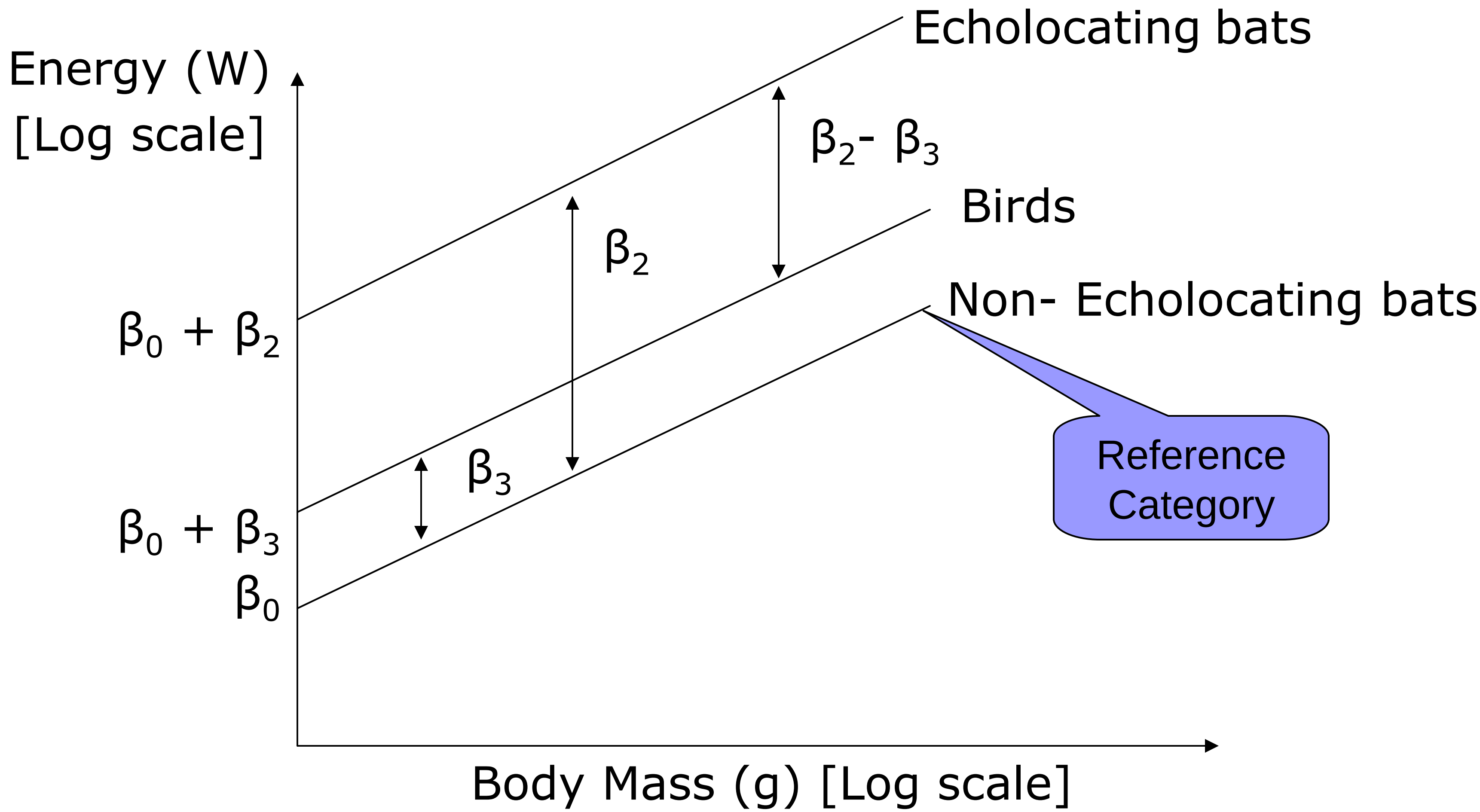
label variable type2 "non-echolocating birds"

label variable type3 "echolocating bats"



Dummy variables as shift parameters

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



Dummy variables as shift parameters

In the previous model:

- β_0 is the intercept for level 1,
- β_2 is the amount by which the mean of y is greater for level 2 than for level 1 (after accounting for x),
- β_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

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           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:

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

- In this model, there is no β_0; β_2, β_3 and β_4 are the intercepts for types 1, 2, and 3, respectively
- We see that the coefficient on β_2 is, indeed, the constant from the previous regression
  - And the other coefficients are shifted accordingly

```
. reg lenergy lmass type1 type2 type3, nocons

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  4,    16) = 1103.51
       Model |  152.647883     4  38.1619709           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9964
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  153.201201    20  7.66006006           Root MSE      =  .18596

------------------------------------------------------------------------------
     lenergy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lmass |   .8149575   .0445414    18.30   0.000     .7205338    .9093811
       type1 |   -1.57636   .2872364    -5.49   0.000    -2.185274   -.9674459
       type2 |  -1.474098   .2390155    -6.17   0.000    -1.980788    -.967408
       type3 |  -1.497696    .149869    -9.99   0.000    -1.815405   -1.179988
------------------------------------------------------------------------------
```

Another parameterization is:

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

- In this model, there is no β_0; β_2, β_3 and β_4 are the intercepts
- We see that the coefficient on β_2 is, indeed, the constant from the previous regression
  - And the other coefficients are shifted accordingly

NOTE!

```
. reg lenergy lmass type1 type2 type3, nocons

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  4,    16) = 1103.51
       Model |  152.647883     4  38.1619709           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9964
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  153.201201    20  7.66006006           Root MSE      =  .18596

------------------------------------------------------------------------------
     lenergy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lmass |   .8149575   .0445414    18.30   0.000     .7205338    .9093811
       type1 |   -1.57636   .2872364    -5.49   0.000    -2.185274   -.9674459
       type2 |  -1.474098   .2390155    -6.17   0.000    -1.980788    -.967408
       type3 |  -1.497696    .149869    -9.99   0.000    -1.815405   -1.179988
------------------------------------------------------------------------------
```

Statistical Inference

- 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?

```
. reg lenergy lmass type1 type2 type3, nocons

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  4,    16) = 1103.51
       Model |  152.647883     4  38.1619709           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9964
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  153.201201    20  7.66006006           Root MSE      =  .18596

------------------------------------------------------------------------------
     lenergy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lmass |   .8149575   .0445414    18.30   0.000     .7205338    .9093811
       type1 |   -1.57636   .2872364    -5.49   0.000    -2.185274   -.9674459
       type2 |  -1.474098   .2390155    -6.17   0.000    -1.980788    -.967408
       type3 |  -1.497696    .149869    -9.99   0.000    -1.815405   -1.179988
------------------------------------------------------------------------------
```

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:
  - $\text{Oil} = \beta_0 + \beta_1 \text{ Temp} + \varepsilon_i$
  - $\text{Oil} = \beta_0 + \beta_1 \text{ Insulation} + \varepsilon_i$

- Multiple regression:
  - $\text{Oil} = \beta_0 + \beta_1 \text{ Temp} + \beta_2 \text{ Insulation} + \varepsilon_i$

Least Squares Estimation

$\mu(y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ var$(y|X_1, X_2) = \sigma^2$

Unknown parameters: Regression coefficients Variance about regression

Fitted values (predicted) $\mu(\mathrm{y} \,|\, \mathrm{x}_1, \mathrm{x}_2) = \hat{\mathrm{y}} = \hat{\beta}_0 + \hat{\beta}_1 \, \mathrm{x}_{i1} + \hat{\beta}_2 \, \mathrm{x}_{i2} \quad \mathrm{i} = 1,2,..n$

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)

$\hat{\sigma}^2 =$ (Sum of squared residuals) / (n-p) [p= number of βs]

t-tests and CI's for individual β'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

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

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

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

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

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 β_j:

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

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: $\beta_3=0$

. reg lenergy lmass type2 type3

Source	SS	df	MS
Model	29.4214818	3	9.80716059
Residual	.553317657	16	.034582354
Total	29.9747994	19	1.57762102

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

t-tests and CIs 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: $\beta_3 = 0$

t-statistic

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           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
------------------------------------------------------------------------------
```

t-tests and CIs 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: $\beta_3 = 0$

t-statistic

Confidence interval

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           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
------------------------------------------------------------------------------
```

t-tests and CIs for Bat Data (From Display 10.6)

1. Results: The data are consistent with the hypothesis of <u>no</u> 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 <u>cannot</u> 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 β_3: $.0787 \pm 2.12*.2027 = (-.35,.51)$.

Interpretation

- Back-transform:

  $e^{.0787} = 1.08$, $e^{-.35} = .70$ and $e^{.51} = 1.67$

  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: .70 to 1.67 times).

Interpretation Depends…

- 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 β_3 becomes <u>negative</u> and significant!

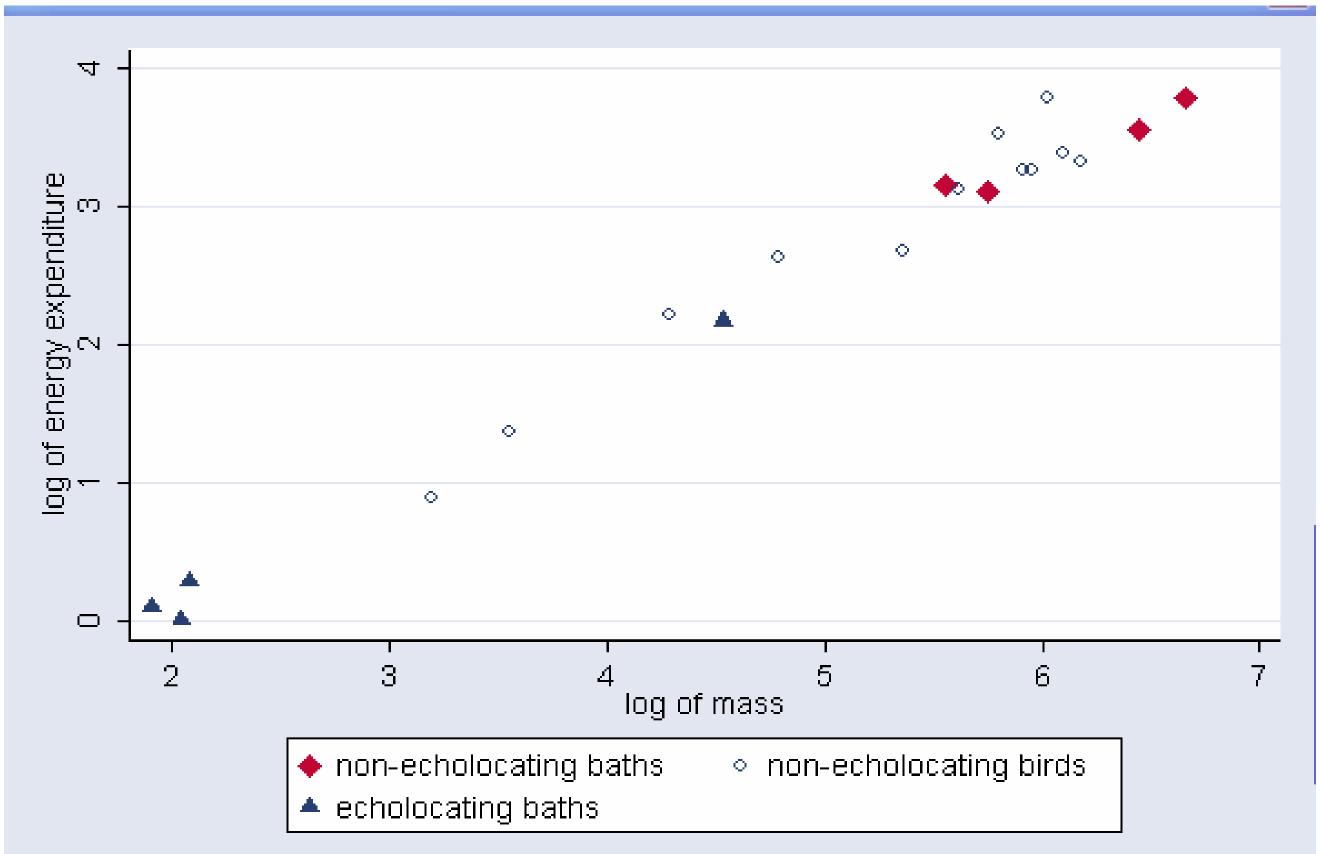
```
. reg lenergy type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   12.50
       Model |  17.8444807     2  8.92224034           Prob > F      =  0.0005
    Residual |  12.1303187    17  .713548161           R-squared     =  0.5953
-------------+------------------------------           Adj R-squared =  0.5477
       Total |  29.9747994    19  1.57762102           Root MSE      =  .84472

------------------------------------------------------------------------------
     lenergy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       type2 |  -.6087747    .487698    -1.25   0.229    -1.637728    .4201783
       type3 |  -2.743272   .5973057    -4.59   0.000    -4.003477   -1.483067
       _cons |    3.39612   .4223589     8.04   0.000     2.505021    4.287219
------------------------------------------------------------------------------
```

Interpretation Depends…

- Ne-bats are clearly much bigger than e-bats.
- So the they naturally use more energy
  - <u>Not</u> 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:

$$\boxed{SST = SSR + SSE}$$

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: β_0

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

R^2 = proportion of total variability (about Y) 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

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

```
. reg lenergy type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   12.50
       Model |  17.8444807     2  8.92224034           Prob > F      =  0.0005
    Residual |  12.1303187    17  .713548161           R-squared     =  0.5953
-------------+------------------------------           Adj R-squared =  0.5477
       Total |  29.9747994    19  1.57762102           Root MSE      =  .84472
```

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

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

```
. reg lenergy lmass type2 type3 zodiac sunshine

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =  162.32
       Model |  29.4664969     5  5.89329938           Prob > F      =  0.0000
    Residual |  .508302494    14  .036307321           R-squared     =  0.9830
-------------+------------------------------           Adj R-squared =  0.9770
       Total |  29.9747994    19  1.57762102           Root MSE      =  .19054
```

Adjusted R^2

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

$$r_{adj}^2 = 1 - \left[\left(1 - r_{Y \bullet 12 \cdots k}^2\right) \frac{n-1}{n-k-1} \right]$$

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

Example Regression Output

```
. reg lenergy type1 type2

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   12.50
       Model |  17.8444807     2  8.92224034           Prob > F      =  0.0005
    Residual |  12.1303187    17  .713548161           R-squared     =  0.5953
-------------+------------------------------           Adj R-squared =  0.5477
       Total |  29.9747994    19  1.57762102           Root MSE      =  .84472

------------------------------------------------------------------------------
     lenergy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       type1 |  -2.134498    .487698    -4.38   0.000    -3.16345   -1.105545
       type2 |   .6087747    .487698     1.25   0.229   -.4201783    1.637728
       _cons |   2.787345    .243849    11.43   0.000    2.272869    3.301822
------------------------------------------------------------------------------

.
```

$$r_{Y \bullet 12}^{2} = \frac{SSR}{SST}$$

Adjusted R^2

- reflects the number of explanatory variables and sample size
- is smaller than R^2

Interpretation of Adjusted R^2

- $$r^2_{Y\bullet 12} = \frac{SSR}{SST} = .5953$$
  - 59.53% of the total variation in energy can be explained by types 1 and 2
- $$r^2_{\text{adj}} = .5477$$
  - 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

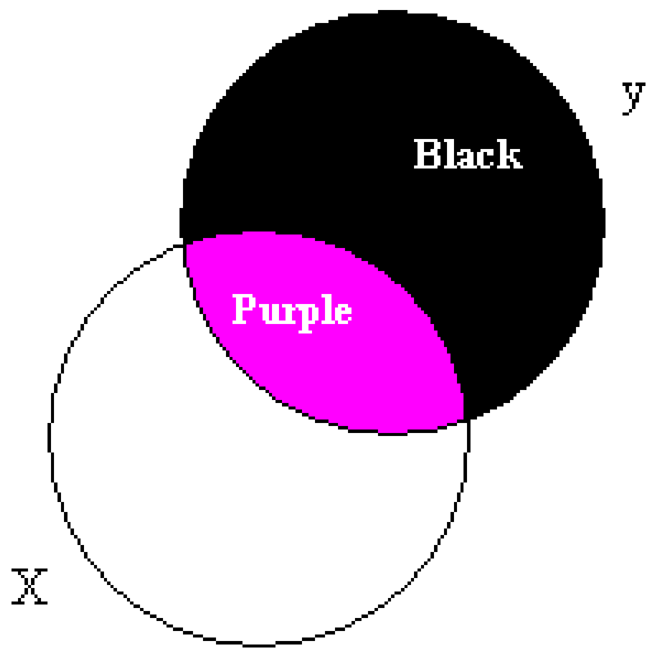
```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

```
. reg lenergy lmass type2 type3 zodiac sunshine

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =  162.32
       Model |  29.4664969     5  5.89329938           Prob > F      =  0.0000
    Residual |  .508302494    14  .036307321           R-squared     =  0.9830
-------------+------------------------------           Adj R-squared =  0.9770
       Total |  29.9747994    19  1.57762102           Root MSE      =  .19054
```

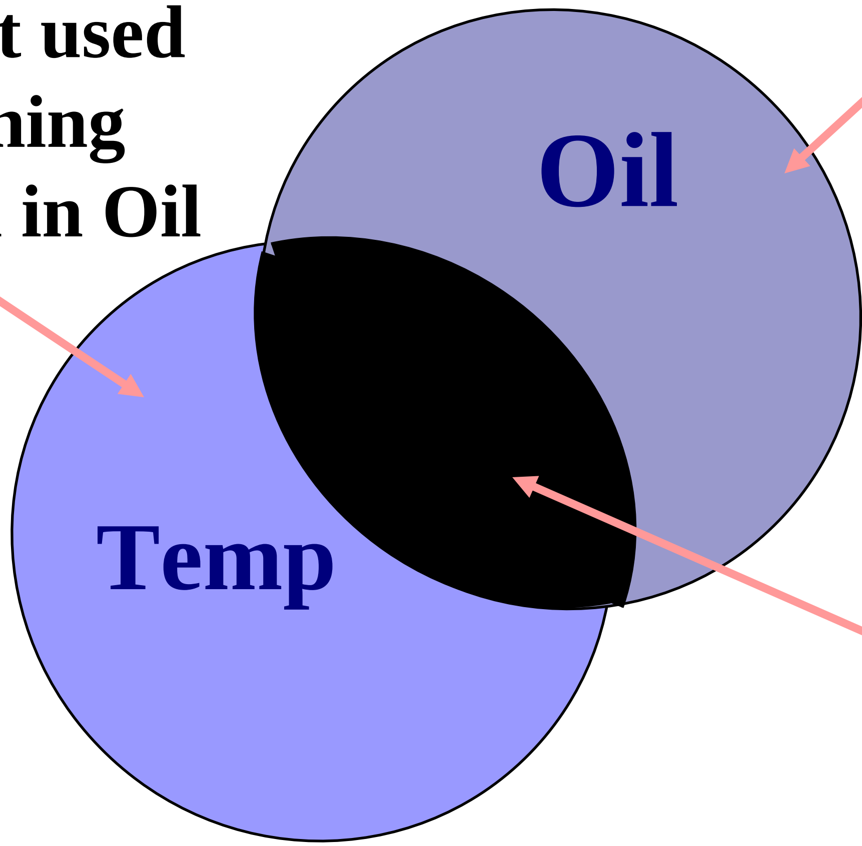
Venn Diagram Representation



- The overlap (purple) is the variation in Y explained by independent variable X (SSR).
- Think of this as <u>information</u> used to explain Y.

Example: Oil Use & Temperature

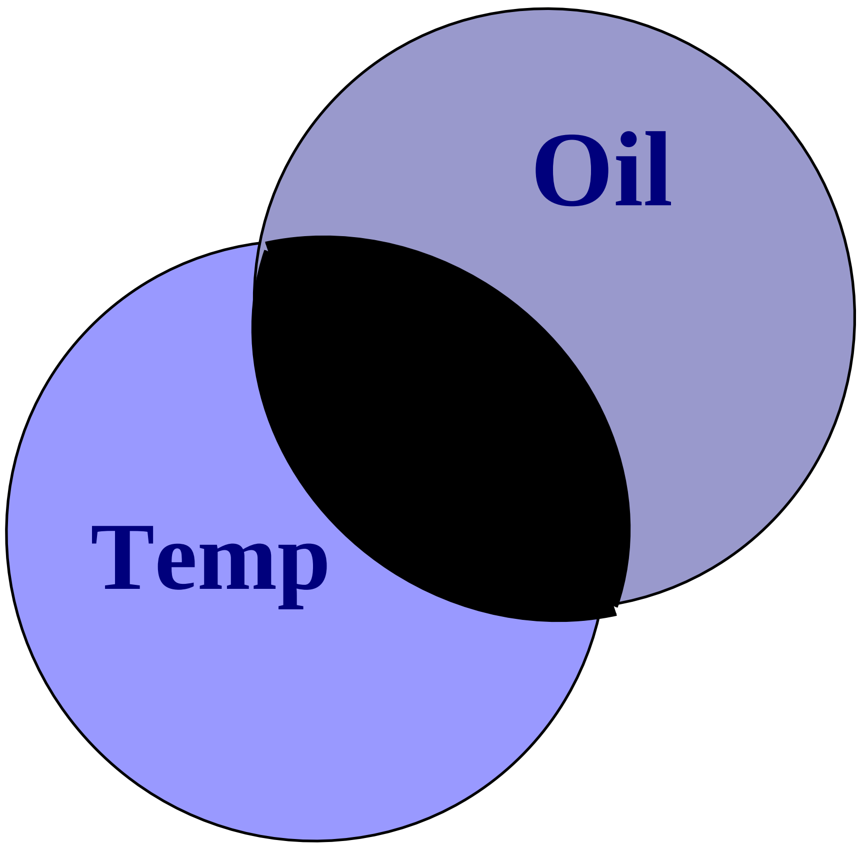
Variations in Temp not used in explaining variation in Oil

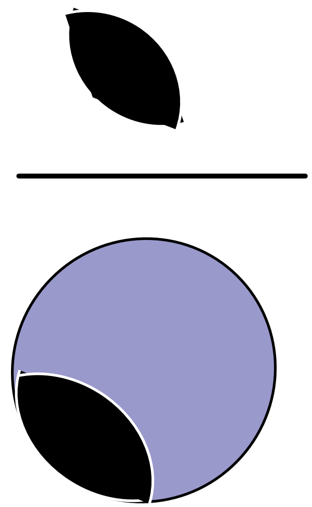


Variations in Oil explained by the error term (SSE)

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

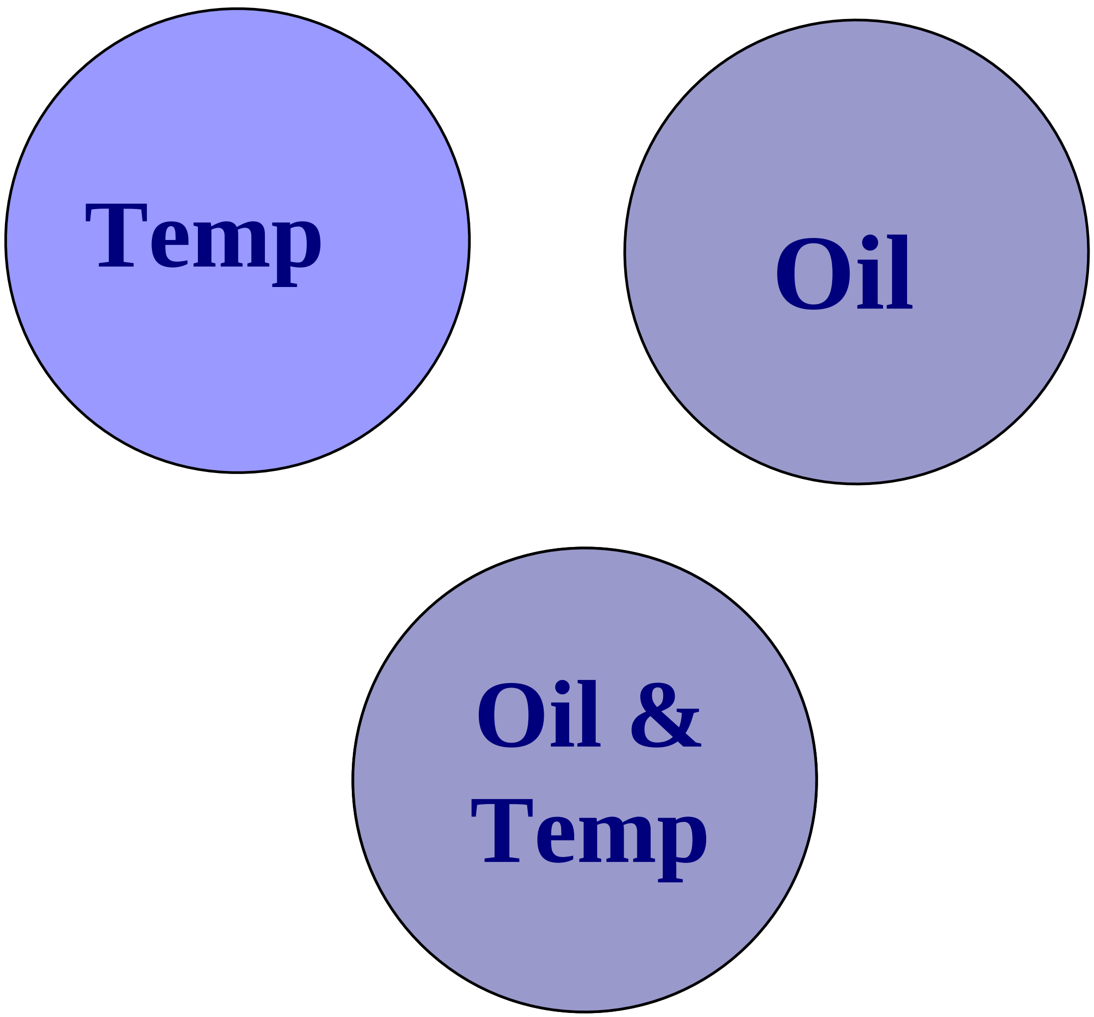
Example: Oil Use & Temperature



$R^2 =$ 

$$= \frac{SSR}{SSR + SSE}$$

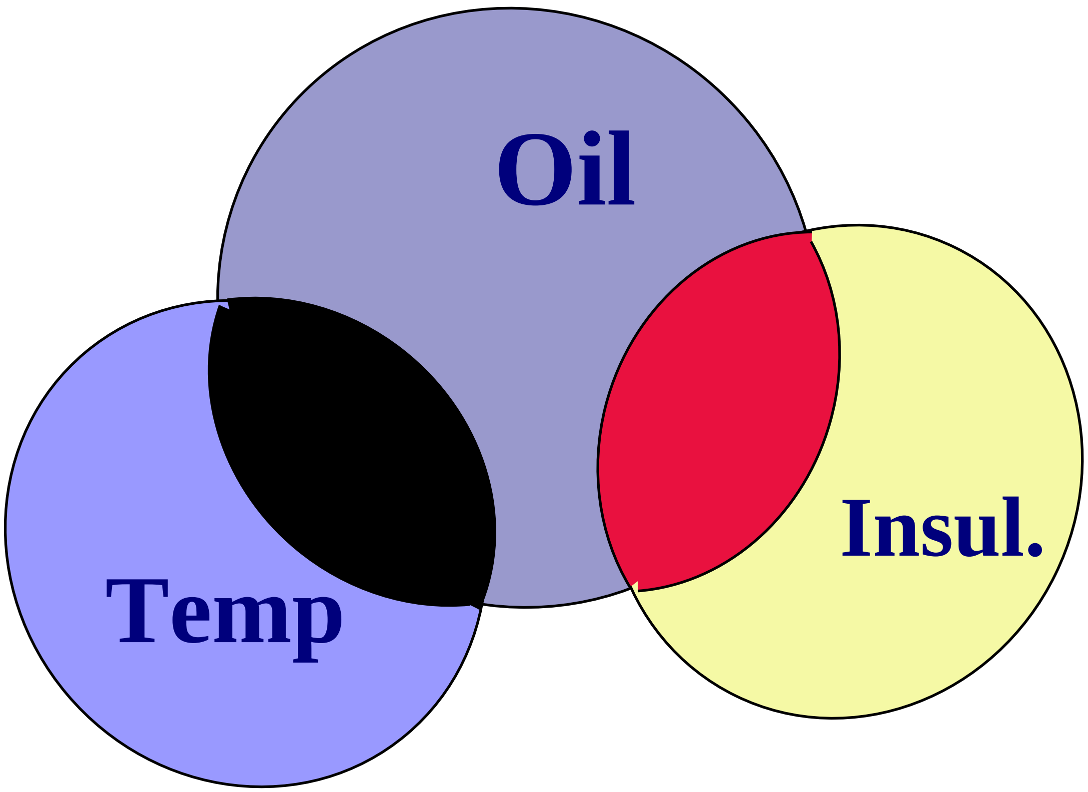
Example: R^2=0 and R^2=1



$$R^2 = 0$$

$$R^2 = 1$$

Uncorrelated Independent Variables

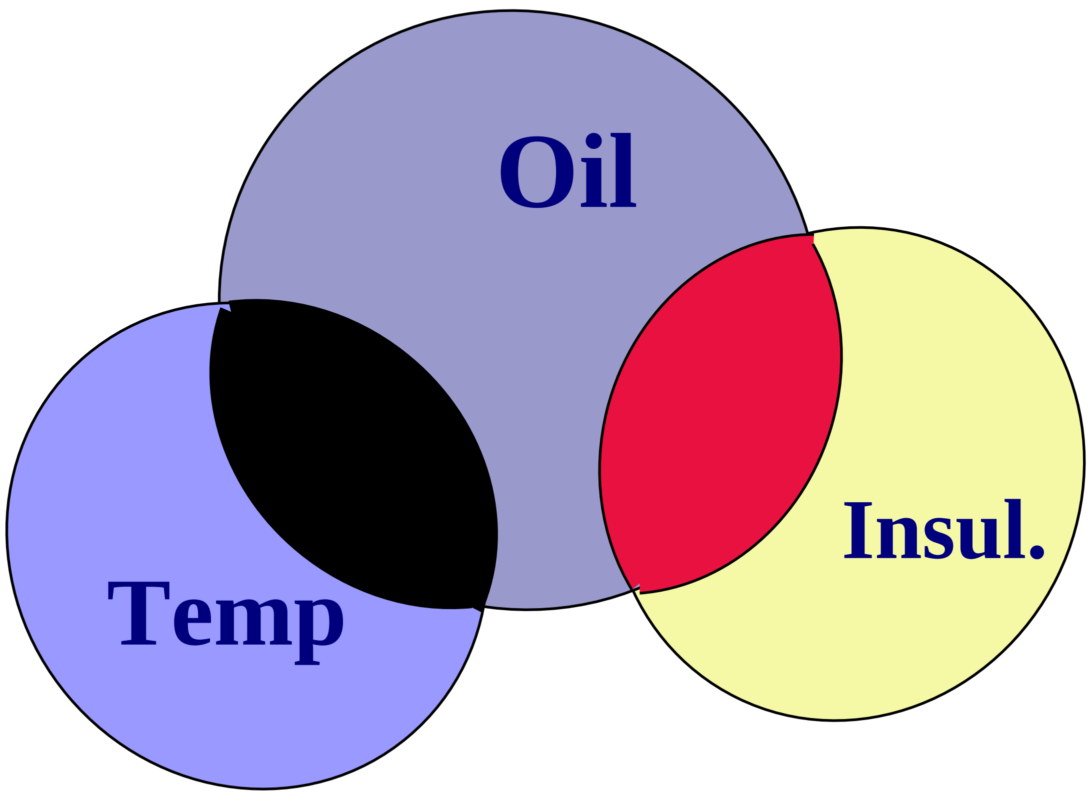


Here, two independent variables that are <u>uncorrelated</u> with each other.

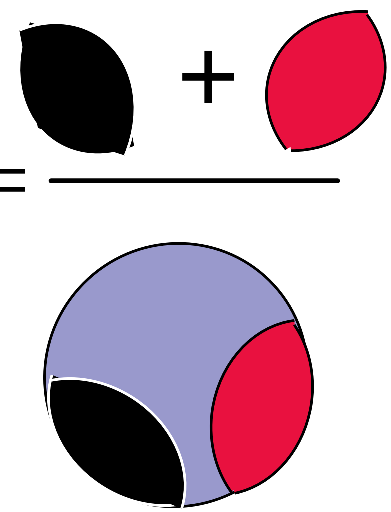
But both affect oil prices.

Then R^2 is just the <u>sum</u> of the variance explain by each variable.

Uncorrelated Independent Variables

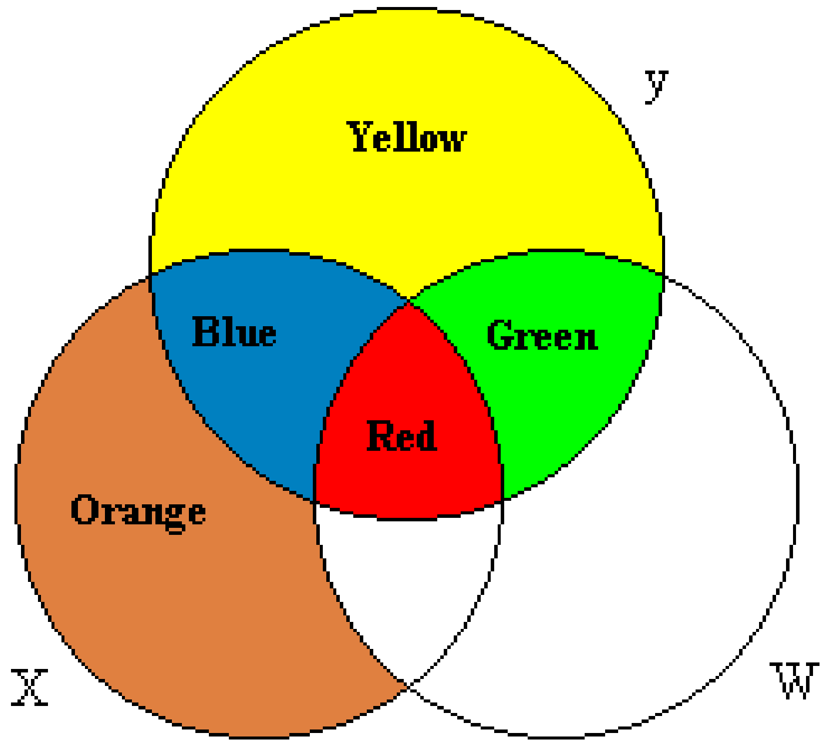


$R^2 =$



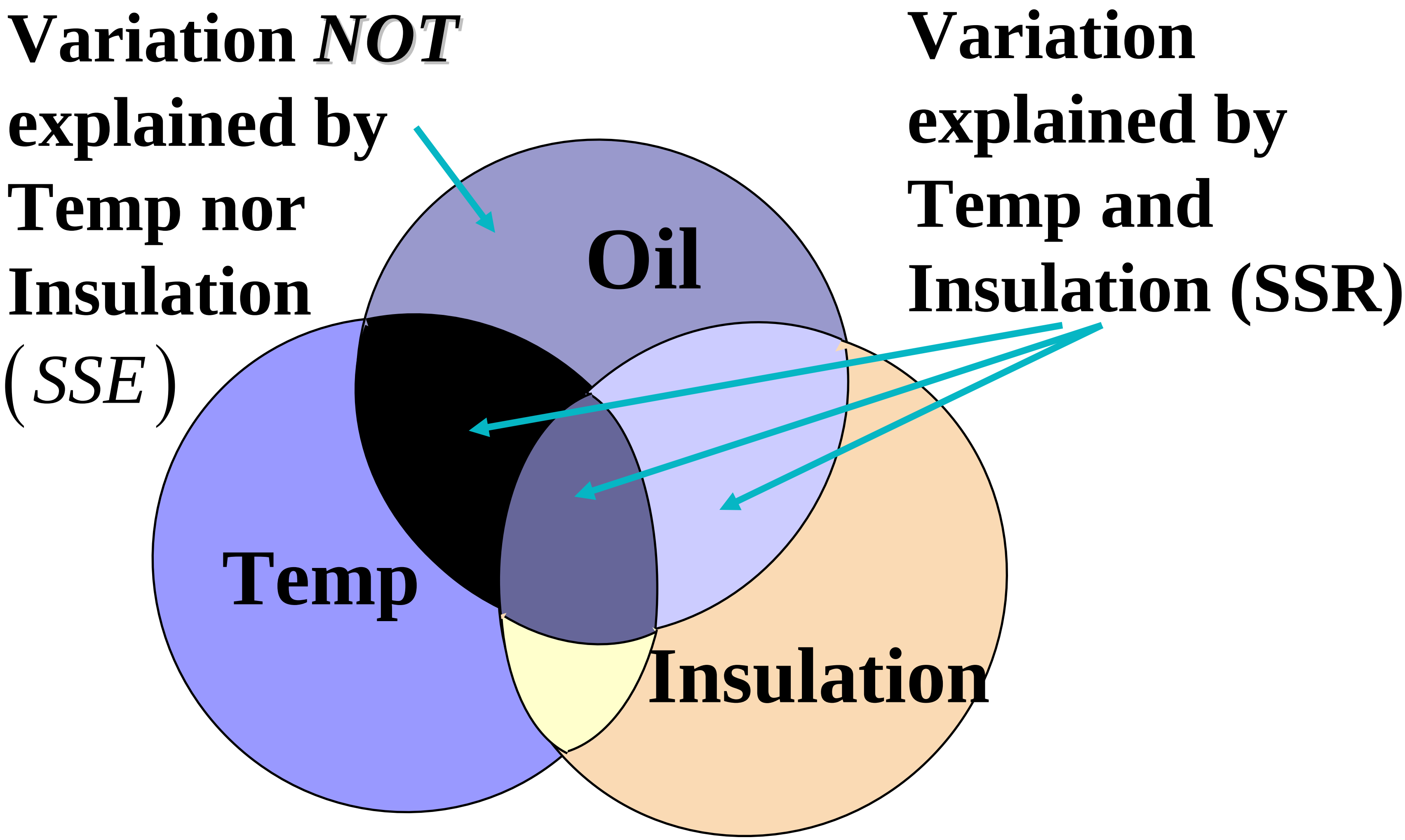
$$= \frac{SSR}{SSR + SSE}$$

Correlated Independent Variables

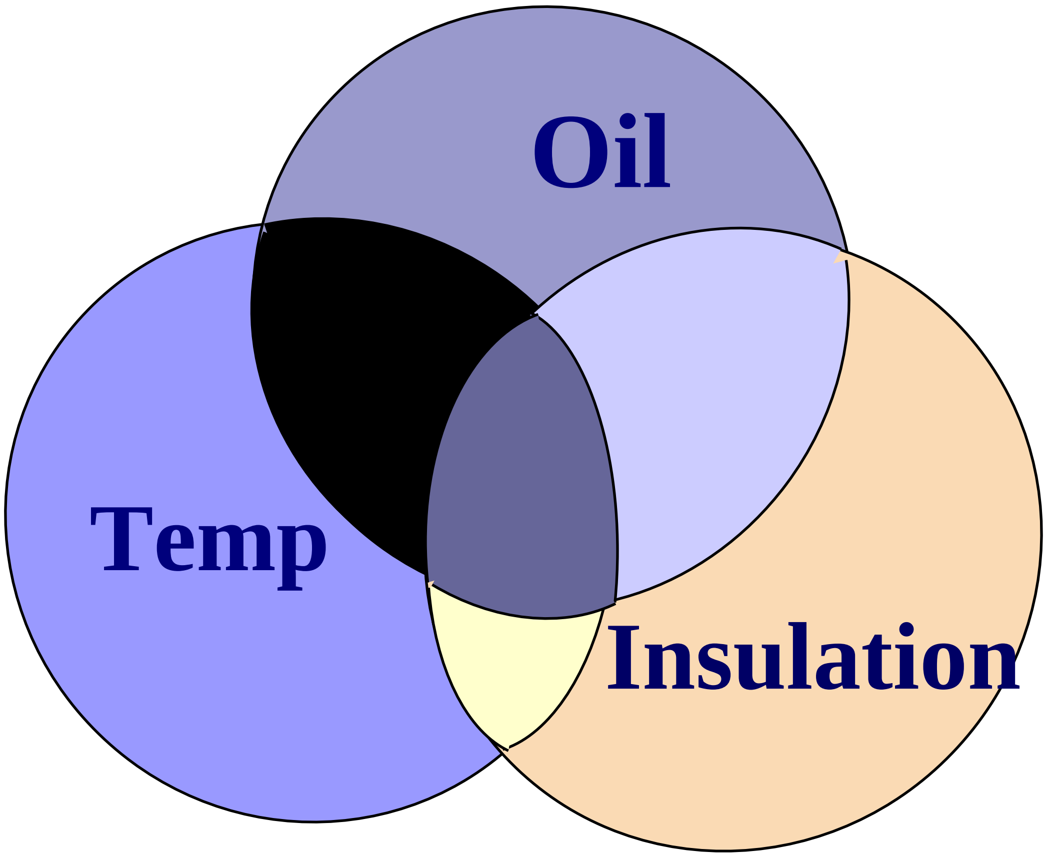


- Now each explains some of the variation in Y
- But there is some variation explained by <u>both</u> X and W (the Red area)

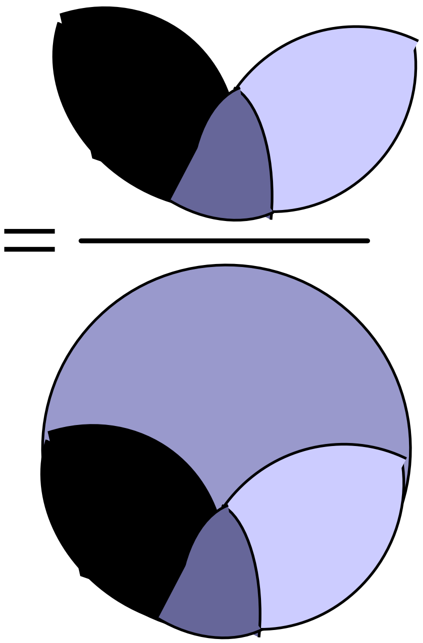
Venn Diagrams and Explanatory Power of Regression



Venn Diagrams and Explanatory Power of Regression



$r^2_{Y \bullet 12} =$



$$= \frac{SSR}{SSR + SSE}$$

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: β_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)

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

F-tests: Example

1. Fit full model:

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

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

- This is the ANOVA section of the regression output
- It has all the information needed to calculate the F-statistic

F-tests: Example

1. Fit full model:

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

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

Sum of Squared Residuals = 0.55332

F-tests: Example

1. Fit full model:

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

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

Sum of Squared Residuals = 0.55332
Degrees of freedom = 16

F-tests: Example

1. Fit full model:

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

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

Sum of Squared Residuals = 0.55332

Degrees of freedom = 16

Mean Squared Error = 0.03458

F-tests: Example

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

```
. reg lenergy constant, nocons

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   78.11
       Model |  123.226402     1  123.226402           Prob > F      =  0.0000
    Residual |  29.9747994    19  1.57762102           R-squared     =  0.8043
-------------+------------------------------           Adj R-squared =  0.7940
       Total |  153.201201    20  7.66006006           Root MSE      =   1.256

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

. sum lenergy

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     lenergy |        20    2.482201    1.256034   .0198026   3.777348
```

Notice that the coefficient on the constant

F-tests: Example

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

```
. reg lenergy constant, nocons

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   78.11
       Model |  123.226402     1  123.226402           Prob > F      =  0.0000
    Residual |  29.9747994    19  1.57762102           R-squared     =  0.8043
-------------+------------------------------           Adj R-squared =  0.7940
       Total |  153.201201    20  7.66006006           Root MSE      =   1.256

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

. sum lenergy

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+-----------------------------------------------------
     lenergy |        20    2.482201    1.256034   .0198026   3.777348
```

Notice that the coefficient on the constant = mean of Y

F-tests: Example

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

```
. reg lenergy constant, nocons

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   78.11
       Model |  123.226402     1  123.226402           Prob > F      =  0.0000
    Residual |  29.9747994    19  1.57762102           R-squared     =  0.8043
-------------+------------------------------           Adj R-squared =  0.7940
       Total |  153.201201    20  7.66006006           Root MSE      =   1.256

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

. sum lenergy

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     lenergy |        20    2.482201    1.256034   .0198026   3.777348
```

Sum of Squared Residuals = 29.97

F-tests: Example

2. Fit reduced model:

$\mu(y|x,TYPE) = \beta_0$

```
. reg lenergy constant, nocons

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   78.11
       Model |  123.226402     1  123.226402           Prob > F      =  0.0000
    Residual |  29.9747994    19  1.57762102           R-squared     =  0.8043
-------------+------------------------------           Adj R-squared =  0.7940
       Total |  153.201201    20  7.66006006           Root MSE      =   1.256

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

. sum lenergy

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     lenergy |        20    2.482201    1.256034   .0198026   3.777348
```

Sum of Squared Residuals = 29.97

Degrees of freedom = 19

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 β's in the full model - # of β's in the reduced model

Numerator d.f. = 19 – 16 = 3

5

Calculate the F-statistic

$$F-statistic = \frac{\frac{29.42}{3}}{.03458} = 283.56$$

6

Find Pr($F_{3,16}$>283.56) from table or computer

P-value = 0.0000

F-tests: Example

Check against regression output:

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

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:

$$SSR\left(X_s \mid \text{ all others except } X_s\right)$$

$$= SSR(\text{all}) - SSR\left(\text{all others except } X_s\right)$$

Testing Portions of Model

- So we want to test whether X_s explains a significant amount of the variation in Y
- Hypotheses:
  - H_0: Variables X_s do not significantly improve the model given all others variables included
  - 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: β_2 & β_3 are <u>not</u> jointly significant

H_0: β_2 & β_3 are jointly significant

The test statistic is essentially the same as before:

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

The only difference is that the Extra SS comes from adding x_2 and x_3 to the reduced model

Testing Subsets: Example

1. Fit full model:

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

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           Root MSE      =  .18596
```

Sum of Squared Residuals = 0.55332
Degrees of freedom = 16
Mean Squared Error = 0.03458

Testing Subsets: Example

2. Fit reduced model:

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

```
. reg  lenergy lmass

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =  907.64
       Model |  29.3919082     1  29.3919082           Prob > F      =  0.0000
    Residual |  .582891195    18  .032382844           R-squared     =  0.9806
-------------+------------------------------           Adj R-squared =  0.9795
       Total |  29.9747994    19  1.57762102           Root MSE      =  .17995
```

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 →

$$F - statistic = \frac{\frac{.0296}{2}}{.03458} = 0.43$$

6 Find Pr($F_{2,16}$>0.43) from table or computer → P-value = 0.659

Testing Subsets: Example

Check against regression output:

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           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
------------------------------------------------------------------------------

. test type2 type3

 ( 1)  type2 = 0
 ( 2)  type3 = 0

       F(  2,    16) =    0.43
            Prob > F =    0.6593
```

The results agree again…

Testing Subsets: Example

Check against regression output:

```
. reg  lenergy lmass type2 type3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =  283.59
       Model |  29.4214818     3  9.80716059           Prob > F      =  0.0000
    Residual |  .553317657    16  .034582354           R-squared     =  0.9815
-------------+------------------------------           Adj R-squared =  0.9781
       Total |  29.9747994    19  1.57762102           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
------------------------------------------------------------------------------

. test type2 type3

 ( 1)  type2 = 0
 ( 2)  type3 = 0

       F(  2,    16) =    0.43
            Prob > F =    0.6593
```

Note that this is easy to do in Stata