Stat 501 Oct. 8

"Extra" Sum of Squares

In multiple regression, we may have interest in how much one set of variables (set 2) reduces the sum of squared errors given that another set of variables (set 1) already is in the model .The authors or our text refers to this as the "extra sum of squares" for the variables in set 2. The notation and calculation is

SSR (set 2 of X variables | set 1 of X variables) = SSE(set 1) – SSE(set 1, set 2)

This could be read as the "reduction in SSE due to set 2 after controlling for set 1."

Examples:

SSR(X4|X1,X2,X3) = SSE(X1, X2, X3)-SSE(X1,X2,X3,X4)

This is the reduction in SSE due to X4 after controlling for X1, X2, and X3. Put another way, this is how much SSE will decrease when X4 is added to a model that already includes X1, X2, X3.

SSR(X1, X3| X2) =SSE(X2)-SSE(X1, X2, X3)

This is how much SSE will decrease when X4 is added to a model that already includes X2 and X3.

For the home sale price example, let Y = log (price), X1= log of home square area, X2= 1 if air conditioning present and 0 if not, X3 = number of cars that can fit in the garage, X4 = number of bedrooms.

Following are results for the model that includes X1, X2, X3, and X4. Note that SSE(X1,X2,X3,X4) = 23.120.

Regression Analysis: logprice versus logarea, Air, CarGarage, Bedrooms

Predictor	Coef	SE Coef	T	P
Constant	3.7959	0.3013	12.60	0.000
logarea	1.08432	0.04370	24.82	0.000
Air	0.08585	0.02656	3.20	0.001
CarGarage	0.11163	0.01723	6.48	0.000
Bedrooms	0.00175	0.01135	0.15	0.878

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	73.409	18.352	409.59	0.000
Residual Error	516	23.120	0.045		
Total	520	96.529			

Following are results for the model that includes X1, X2, X3. Note that SSE(X1,X2,X3) = 23.121, so SSR(X4|X1,X2,X3) = SSE(X1, X2, X3)-SSE(X1,X2,X3,X4) = 23.121 – 23.120 = 0.001.
Not very big, which is whyt bedrooms was not significant above.

Regression Analysis: logprice versus logarea, Air, CarGarage

Predictor	Coef	SE Coef	T	P
Constant	3.7764	0.2733	13.82	0.000
logarea	1.08760	0.03812	28.53	0.000
Air	0.08538	0.02644	3.23	0.001
CarGarage	0.11166	0.01722	6.49	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	73.408	24.469	547.14	0.000
Residual Error	517	23.121	0.045		
Total	520	96.529			

Following are results for a model with only X2 included. Note that SSE = 84.526.

So, SSR(X1, X3| X2) =SSE(X2)-SSE(X1, X2, X3) = 84.256-23.121 = 61.135.
Adding X1 and X3 to a model that had only X2 will decrease the SSE by 72.135.

Regression Analysis: logprice versus Air

Predictor	Coef	SE Coef	T	P
Constant	12.0965	0.0430	281.18	0.000
Air	0.40512	0.04719	8.58	0.000

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	12.003	12.003	73.70	0.000
Residual Error	519	84.526	0.163		
Total	520	96.529			

Testing Whether One or More Variables Can Be Dropped From the Model

A statistical test of whether a set of variables can be simultaneously eliminated as predictor variables is based on the F-statistic

$$F = \frac{SSE(\text{Reduced}) - SSE(\text{Full})}{\dfrac{\text{error df for reduced} - \text{error df for full}}{MSE(\text{full})}}$$

with df = (error df for reduced – error df for full) and error df for full.

.

The precise null hypothesis is that some specified set of β coefficients is 0.
The Reduced model is the model that results if this null hypothesis is true. The Full model is the model with all variables being considered included.

Example:

Suppose a potential model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$. This is the "full" model.

If we wish to test

$H_0: \beta_1 = \beta_3 = 0$

the reduced model is $y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i$

Basically, we're testing to see if variables x_1 and x_3 are "significant" given that x_2 will be in the model. The relevant F-statistic is

$$F = \frac{\dfrac{SSE(x_2) - SSE(x_1, x_2, x_3)}{\text{error df reduced} - \text{error df full}}}{MSE(x_1, x_2, x_3)} = \frac{\dfrac{SSR(x_1, x_3 \mid x_2)}{\text{error df reduced} - \text{error df full}}}{MSE(x_1, x_2, x_3)}$$

Tasks we'll complete in class:

1. Consider the home price example with a full model of X1,X2,X3. Calculate an F for testing $H_0: \beta_1 = \beta_3 = 0$

2. Consider the home price example with a full model of X1,X2,X3, X4. Calculate an F for testing $H_0: \beta_4 = 0$