

Testing Hypotheses in Multiple Regression

Applied Regression and Other Multivariable Methods
Sections 9-1 - 9-6

10

F test

- All hypothesis tests can be expressed as a F test
- If testing #2, can also be expressed as a t test

$$F_{1,v} = t_v^2$$

- F test is a ratio of indep estimates of variance

$$F = \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}$$

- The parameter $\hat{\sigma}_0^2$ estimates σ^2 under H_0
- The parameter $\hat{\sigma}_0^2$ will be larger under H_A
- Thus, if $F \gg 1$ will reject H_0
- Will use mean squares as estimates of variance

10-2

Preview

- When performing multivariable regression, there is interest in the following questions:
 - 1 Does the set of indep variables contribute significantly to the prediction of Y ?
 - 2 Does the addition of one variable, all other variables present in the model, help predict Y ?
 - 3 Does the addition of a group of variables, given all others, help predict Y ?

- All can be answered through a hypothesis test
- All can also be considered a comparison of models

Testing #1 (or #3)

$$\begin{aligned} H_0 : Y &= \beta_0 + E \\ H_A : Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E \end{aligned}$$

Testing #2 (or #3)

$$\begin{aligned} H_0 : Y &= \beta_0 + \beta_1 X_1 + E \\ H_A : Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E \end{aligned}$$

- H_0 : variable/variables is/are not helpful

10-1

Testing Overall Significance

- Interested to see if indep variables considered together explain a significant amount of the variation in Y
- Is the set of indep variables helpful in predicting Y ?
- H_0 : These variables **are not** helpful in predicting Y
- In terms of models, interest in comparing

$$\begin{aligned} H_0 : Y &= \beta_0 + E \\ H_A : Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + E \end{aligned}$$

- $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$
- H_0 : Best predictor is simply a constant $\rightarrow \bar{Y}$
- Reduces down to simple ratio of mean squares available in the ANOVA table

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$$

10-3

Testing Overall Significance

- Will use F test

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Error}}} = \frac{(SSY - SSE)/k}{SSE/(n - k - 1)}$$

- If variables not helpful $MS_{\text{Regression}} \approx MS_{\text{Error}}$
- Compare F to $F_{k, n-k-1, \alpha}$ (critical value)

- This is a one-sided directional test

$$F_{1, v} = t_v^2$$

- Recall $R^2 = \frac{SSY - SSE}{SSY}$ so we can write

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

10-4

Example

- Consider the following ANOVA table from Topic 9

The REG Procedure

**** Full model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5081.86742	1016.37348	19.66	<.0001
Error	26	1344.10133	51.69621		
Corrected Total	31	6425.96875			

Root MSE	7.19001	R-Square	0.7908
Dependent Mean	144.53125	Adj R-Sq	0.7506
Coeff Var	4.97471		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	185.69424	73.70952	2.52	0.0182
quet	1	-32.41001	21.84299	-1.48	0.1499
age	1	-1.35438	1.40388	-0.96	0.3436
smk	1	3.46636	20.61748	0.17	0.8678
quet_smk	1	1.80295	5.95779	0.30	0.7646
quet_age	1	0.73888	0.38735	1.91	0.0676

This regression model includes 5 indep variables. The sums of squares and F test are shown in the ANOVA table

$SSY = 6425.97$, $SSE = 1344.10 \rightarrow SSY - SSE = 5081.87$

$$F = \frac{5081.87/5}{1344.1/26} = 19.66 \quad F = \frac{.7908/5}{(1 - .7908)/26} = 19.66$$

10-5

Partial F tests

- Overall fit does not tell use which variables are important
- Partial F tests allow us to look at the importance of a single variable or group of variables
- To understand process, must first understand why order of fit important
- When indep variables are correlated, they may explain the same variation in Y
- If X_1 fit first, this means X_2 may not explain any **additional** variation
- In other words, with X_1 in the model, X_2 does not help better predict Y
- Partial F tests allow us to see which variables contribute extra information when already considering a set of variables in the model

10-6

Extra Sums of Squares

- Consider three indep variables X_1, X_2 , and X_3
- We will break down the regression SS as follows
 - $SS(X_1)$: SS explained by only using X_1
 - $SS(X_1, X_2)$: SS explained by using both X_1 and X_2
 - $SS(X_1, X_2, X_3)$: SS explained by using all three variables
- Because of correlation, $SS(X_1, X_2) \neq SS(X_1) + SS(X_2)$
- Will introduce extra SS
 - $SS(X_2|X_1)$: extra SS explained by including X_2 given that X_1 has already been fit in model
 - $SS(X_3|X_1, X_2)$: extra SS explained by including X_3 given that X_1 and X_2 have already been fit in model
- Using these extra SS

$$SS(X_1, X_2) = SS(X_1) + SS(X_2|X_1)$$

$$SS(X_1, X_2, X_3) = SS(X_1, X_2) + SS(X_3|X_1, X_2)$$

10-7

Example

- Consider the following models from Topic 9

$$Y = \beta_0 + \beta_1 \text{QUET}$$

$$Y = \beta_0 + \beta_1 \text{QUET} + \beta_2 \text{AGE}$$

$$Y = \beta_0 + \beta_1 \text{QUET} + \beta_2 \text{AGE} + \beta_3 \text{SMK}$$

- Will compute extra sum of squares using SAS

- This can be obtained using the `ss1` option

```
options nocenter;
data problem81;
infile 'i:\www\datasets502\EX0502.DAT' firstobs=2 dlm='09'x;
input person sbp quet age smk;

/* Investigate different models */
proc reg;
model sbp = quet;
model sbp = quet age;
model sbp = quet age smk /ss1;
run;
```

10-8

MODEL 1					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3537.94574	3537.94574	36.75	<.0001
Error	30	2888.02301	96.26743		
Corrected Total	31	6425.96875			

Root MSE	9.81160	R-Square	0.5506
Dependent Mean	144.53125	Adj R-Sq	0.5356
Coeff Var	6.78856		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	70.57640	12.32187	5.73	<.0001
quet	1	21.49167	3.54515	6.06	<.0001

MODEL 2					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4120.59225	2060.29612	25.92	<.0001
Error	29	2305.37650	79.49574		
Corrected Total	31	6425.96875			

Root MSE	8.91604	R-Square	0.6412
Dependent Mean	144.53125	Adj R-Sq	0.6165
Coeff Var	6.16893		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	55.32344	12.53475	4.41	0.0001
quet	1	9.75073	5.40246	1.80	0.0815
age	1	1.04516	0.38606	2.71	0.0113

10-9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4889.82570	1629.94190	29.71	<.0001
Error	28	1536.14305	54.86225		
Corrected Total	31	6425.96875			

Root MSE	7.40691	R-Square	0.7609
Dependent Mean	144.53125	Adj R-Sq	0.7353
Coeff Var	5.12478		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	45.10319	10.76488	4.19	0.0003	668457
quet	1	8.59245	4.49868	1.91	0.0664	3537.94574
age	1	1.21271	0.32382	3.75	0.0008	582.64651
smk	1	9.94557	2.65606	3.74	0.0008	769.23345

The last column of the Parameters Estimates gives the extra SS fit in the order stated in the model statement

Let $X_1 = \text{QUET}$, $X_2 = \text{AGE}$, and $X_3 = \text{SMK}$

$$SS(\text{QUET}) = 3537.95$$

$$SS(\text{AGE}|\text{QUET}) = 582.65$$

$$SS(\text{SMK}|\text{QUET}, \text{AGE}) = 769.23$$

We will now use these extra SS to test hypotheses

Example Continued

- Can use extra SS to construct SS for other models

- Consider Model 1: $Y = \beta_0 + \beta_1 \text{QUET}$

– The model SS will be the SS due solely to QUET. Since this variable was fit first in the full model, the model SS equals $SS(\text{QUET}) = 3537.95$.

– The other two variables are not considered. This means we can compute the error SS by adding the remaining extra SS terms together with error SS for this model. Thus, the SSE for Model 1 is $1536.14 + 582.65 + 769.23 = 2888.02$.

- Consider Model 2: $Y = \beta_0 + \beta_1 \text{QUET} + \beta_2 \text{AGE}$

Since $SS(\text{AGE}|\text{QUET}) = 582.65$, the Regression SS for Model 2 is $SS(\text{QUET}) + SS(\text{AGE}|\text{QUET}) = 3537.95 + 582.65 = 4120.60$.

The SSE for Model 2 is $1536.14 + 769.23 = 2305.37$.

- Recall $SSY = \text{Regression SS} + \text{SSE}$

- Because SSY constant, changing the number of indep variables is simply a redistribution of SS into SSE and Regression SS

10-10

10-11

Partial F test

- Interest in whether X^* can further help predict Y
- Consider extra SS of X^* as %age of full model MSE
- The partial F test for variable X^*

$$F = \frac{SS(X^*|X_1, X_2, \dots, X_p)}{MSE(X^*, X_1, X_2, \dots, X_p)}$$

- Compare F to $F_{1, n-p-2, \alpha}$ (critical value)

Example

Consider the test to see if $X_2=AGE$ can help better predict SBP when QUET is already in the model. From the output, we know $SS(AGE|QUET) = 582.65$. For the denominator, we computed the SSE for MODEL 2 to be 2305.38. Thus,

$$F = \frac{582.65}{2305.38/(32 - 1 - 2)} = 7.33$$

From Table A-4, the P-value is between .01 and .025.

10-12

The Partial t Test

- Recall the relationship between F and t
- Since the partial F has 1 df in numerator,

$$\sqrt{F} = T$$

- Compare T to $t_{n-p-2, \alpha/2}$

Example

Consider the test to see if $X_2=AGE$ can help better predict SBP when QUET is already in the model. The F test statistic was 7.08 which means the T test statistic is $\sqrt{7.08} = 2.66$. From Table A-2, the P-value is between 2(.005) and 2(.01). This turns out to be the same P-value as the F test.

NOTE: This t-test, is the same t-test for the parameter AGE that is shown in the output for Model 2. The difference is due to my rounding. We'll talk more about the t-tests that are shown in the SAS output in the next Topic.

10-13

Multiple Partial F Test

- So far, we've considered the inclusion of one variable at a time
- Could also do test for group of variables, $\{X_1^*, X_2^*, \dots, X_k^*\}$
- The partial F test generalizes single variable test

$$\begin{aligned} F &= \frac{SS(X_1^*, X_2^*, \dots, X_k^*|X_1, X_2, \dots, X_p)/k}{MSE(X_1^*, X_2^*, \dots, X_k^*, X_1, X_2, \dots, X_p)} \\ &= \frac{(SSE(\text{reduced}) - SSE(\text{full}))/k}{MSE(\text{full})} \end{aligned}$$

- Compare F to $F_{k, n-p-k-1, \alpha}$ (critical value)
- This test used when variables naturally grouped

Example: AGE_QUET, AGE², and QUET² are of order 2

10-14

Examples

- 1 Consider the test to see if the combination of AGE and SMK is significant when QUET is already in the model.

$$F = \frac{(2888.02 - 1536.14)/2}{1536.14/28} = 12.32$$

Comparing this to $F_{2,28}$, the P-value is between .001 and .005. This means it is important to include these two variables.

- 2 Consider the test to see if QUET and AGE should be included in the model when nothing else is already in the model. This test is similar to the F-test for Model 2. When no terms are in the model, the SSE is the SSY. This is because $SSE = SSY - \text{Regression SS}$.

$$F = \frac{(6425.97 - 2305.38)/2}{2305.38/29} = 25.92$$

Comparing this to $F_{2,29}$, the P-value is less than .001.

10-15