ANOVA & Testing in the Multiple Linear Regression Models

(EPI-809)

ANOVA in the Multiple Regression Model

- Variance decomposition
- Degrees of freedom (by factor and for the model)
- Type-I and Type-III SS
- Testing multiple effects jointly (Long and Short regression and the F-test)
- Type-I error control and sequential testing

ANOVA in the Multiple Regression Model

Model equation (Ha): $y_i = \beta_0 + \beta_1 ED_i + \beta_2 W_i + \beta_3 B_i + \varepsilon_i$

Null Hypothesis (H0): $\beta_1 = \beta_2 = \beta_3 = 0$ or $y_i = \beta_0 + \varepsilon_i$

Predictions & Residuals from Ha:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 ED_i + \hat{\beta}_2 W_i + \hat{\beta}_3 B_i \qquad \hat{\varepsilon}_i = y_i - \hat{y}_i = \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 ED_i + \hat{\beta}_2 W_i + \hat{\beta}_3 B_i\right)$$

Predictions & Residuals from H0: $\hat{\varepsilon}_{i(0)} = (y_i - \bar{y})$

Variance partition:

$$\left.\begin{aligned} \text{Total}: \quad & SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \\ Un-Explained: \quad & \text{SSE} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \\ Explained\ \text{(regression)}: \quad & \text{SSR} = SST - SSE \end{aligned}\right\} \quad R^2 = \frac{SSR}{SSE + SSR}$$

Degrees of freedom

Model equation (Ha): $y_i = \beta_0 + \beta_1 ED_i + \beta_2 W_i + \beta_3 B_i + \varepsilon_i$ p=4 parameters

Null Hypothesis (H0): $y_i = \beta_0 + \varepsilon_i$ q=1 parameter

Residual degrees of freedom: n-p=n-4

Model degrees of freedom: p-q=3

Total degrees of freedom (that is residual df of H0): n-q=model df+res. df

Example in R :

https://github.com/gdlc/EPI809/blob/master/ANOVA_MLR.md

Testing in the Multiple Regression Model

- t-test for 1-df tests
- F-test for more than 1-df tests
- Testing groups of predictors
- Error control

Multiple Regression Model

Regression of, Wages (y) on Region, Sex, Experience and Education.

Model

$$y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + Exp_{3i}\beta_3 + Edu_{4i}\beta_4 + \varepsilon_i$$

Dummy: South (1) ; North(0)

Dummy: Female (1) ; Male (0)

Years of Experience

Years of Education

Examples of tests of hypothesis that we can do once we fit the model

Type of Test	Example/Description	Set of Hypothesis	Stat. Test
Single-parameter	Are there differences between South & North?	$\begin{cases} H_0 & \beta_1 = 0 \\ Ha & \beta_1 \neq 0 \end{cases}$	t-test (Parameter Estimates Table)
Model as a whole	Do Region, Sex, Experience and Education together have any effect on wages?	$\begin{cases} H_0 & \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\ Ha & \text{At least one of them} \neq 0. \end{cases}$	F-test (ANOVA table)
Groups of predictors	Do Experience and Education together have any effect on wages?	$\begin{cases} H_0 & \beta_3 = \beta_4 = 0 \\ Ha & \text{At least one of the two} \neq 0. \end{cases}$	F-test (we will calculate it)

Single-Effect Test

$$\begin{cases} H_0 \quad \beta_j = 0 \\ Ha \quad \beta_j \neq 0 \end{cases}$$

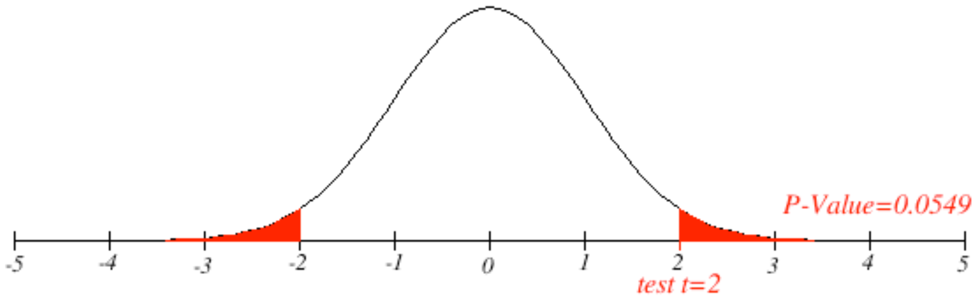
Standard Errors & t-statistic

Standard Error of Estimates $$SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$$

Standardized Estimates (t-statistic) $$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

pValue: We compare the t-statistic against a reference distribution (the distribution of the t-statistic under the null hypothesis (H0: b1=0). If the observed statistic is 'unusual' for this reference distribution that provides evidence against H0.

Specifically, the p-value is the probability of observing a t-statistic at least as extreme as the one observed if the null hypothesis holds.



Let's discuss this from the JMP output

Testing the Model as a whole (all effects together)

$$y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + Exp_{3i}\beta_3 + Edu_{4i}\beta_4 + \varepsilon_i$$

$$\begin{cases} H_0 \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\ Ha \quad At\text{ least one of them} \neq 0. \end{cases}$$

If the NULL Hypothesis holds…

$$y_i = \beta_0 + \varepsilon_i$$

Mean of Y

To test the model as a whole we will use the ANOVA Table…

ANOVA and F-test

$$\mathrm{H}_0: \quad y_i = \beta_0 + \varepsilon_i$$

$$\mathrm{H}_a: \quad y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + Exp_{3i}\beta_3 + Edu_{4i}\beta_4 + \varepsilon_i$$

Mean Square=Sum Sq. / DF

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	3707.363	926.841	47.3474
Error	523	10237.889	19.575	Prob > F
C. Total	527	13945.252		<.0001*

F-statistic: MS.Model/MS.Error

SS Under H0

SS Under Ha

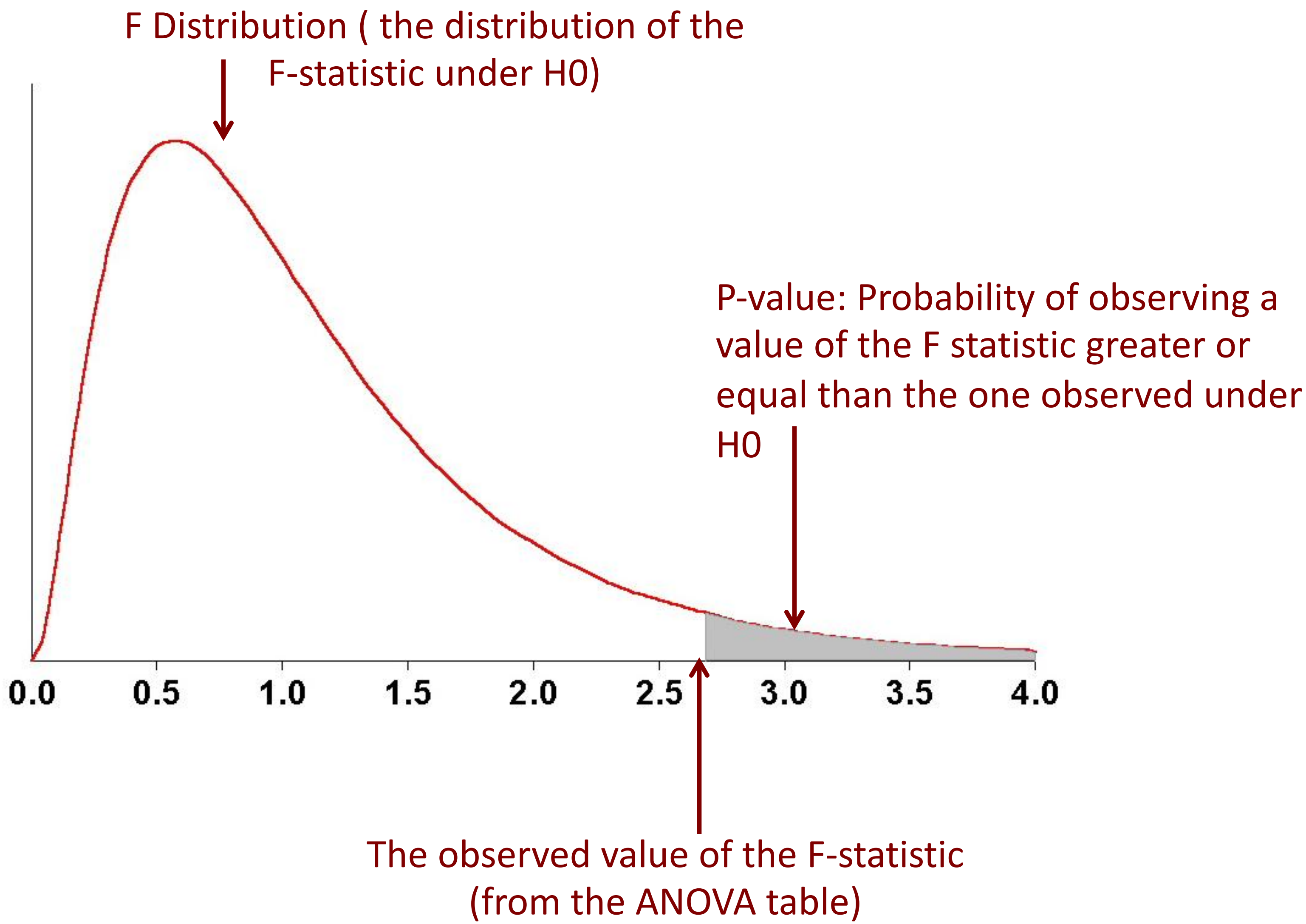
Is the F-ratio large enough to reject the NULL (see H0 above)?

Note:

The F-statistic is directly related to the R-squared of the model.

The larger the R-sq. the larger the F-stat.

ANOVA and F-test



Testing Groups of Predictors

(e.g., Do Education and Experience Have Any Effect on Wages?)

Null and Alternative

$$y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + Exp_{3i}\beta_3 + Edu_{4i}\beta_4 + \varepsilon_i$$

Short Regression

$$\begin{cases} H_0 \quad y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + \varepsilon_i \\ Ha \quad y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + Exp_{3i}\beta_3 + Edu_{4i}\beta_4 + \varepsilon_i . \end{cases}$$

Long Regression

Implementing the above test boils down to comparing the short and long regressions.

The R-squared of the Long Regression will always be $\geq$ than that of the Short Regression

How can we test whether the reduction in R^2 we observed is large enough to reject the null hypothesis?

F-test

F Test for the Long Vs Short Regression

Long Regression $y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + Exp_{3i}\beta_3 + Edu_{4i}\beta_4 + \varepsilon_i$

$$\begin{cases} H_0: & \beta_3 = \beta_4 = 0 \\ H_a: & \text{At least one} \neq 0 \end{cases}$$

Short Regression (H_0) $y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + \varepsilon_i$

ANOVA Table of the Short Regression (k=3)

Source	DF	SS	MS	R^2	*f*
Model	2	SSR_0	MSR_0	SSR_0 /SST	MSR_0 / MSE_0
Error	N-3	SSE_0	MSE_0		
Total	N-1	SST	MST		

ANOVA Table of Long Regression (k=5)

Source	DF	SS	MS	R^2	*f*
Model	4	SSR_A	MSR_A	SSR_A /SST	MSR_A / MSE_A
Error	N-5	SSE_A	MSE_A		
Total	N-1	SST	MST		

F-Test

Source	DF	SS	MS	R^2	F
Model (Ha)	2	$SSE_0 - SSE_A$	$(SSE_0 - SSE_A)/2$	$(SSE_0 - SSE_A)/SSE_0$	$\frac{(SSE_0 - SSE_A)/2}{SSE_A/(N-5)}$
Error	N-5	SSE_A	$SSE_A/(N-5)$		
Total (H_0)	N-3	SSE_0			

F Test for the Long Vs Short Regression

Long Regression $y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + Exp_{3i}\beta_3 + Edu_{4i}\beta_4 + \varepsilon_i$

$$\begin{cases} H_0: & \beta_3 = \beta_4 = 0 \\ H_a: & \text{At least one} \neq 0 \end{cases}$$

Short Regression (H_0) $y_i = \beta_0 + South_{1i}\beta_1 + Fe_{2i}\beta_2 + \varepsilon_i$

ANOVA Short Regression (H_0)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	950.000	475.000	19.1897
Error	525	12995.252	24.753	Prob > F
C. Total	527	13945.252		<.0001*

ANOVA Long Regression (H_A)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	3707.363	926.841	47.3474
Error	523	10237.889	19.575	Prob > F
C. Total	527	13945.252		<.0001*

	ANOVA Table For Testing H_0 Vs H_A					
	DF	SS	MS	R2	f	P(F>f)
Model	2	2757.4	1378.7	0.2122	70.43	8.22909E-28
Error	523	10237.9	19.6			
Total	525	12995.3				

Sequential Variance Partition and the F-test

	Model	Res. DF	Residual Sum of Squares
"Null Model"	$y_i = \beta_0 + \varepsilon_i$	N-1	$TSS : \sum_{i=1}^{n} (y_i - \bar{y})^2$
Short Regression	$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$	N-3	$RSS_S : \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x_{1i}\hat{\beta}_1 - x_{2i}\hat{\beta}_2)^2$
Long Regression	$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \varepsilon_i$	N-4	$RSS_L : \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x_{1i}\hat{\beta}_1 - x_{2i}\hat{\beta}_2 - x_{3i}\hat{\beta}_3)^2$

$$TSS \geq RSS_S \geq RSS_L$$

Model Sum of Squares

Short Regression: $MSS_s = TSS - RSS_S$

Long Regression: $MSS_L = TSS - RSS_L$

R-Squared

$$R_S^2 = \frac{TSS - RSS_S}{TSS}$$

$$R_L^2 = \frac{TSS - RSS_L}{TSS}$$

$$R_L^2 \geq R_S^2$$

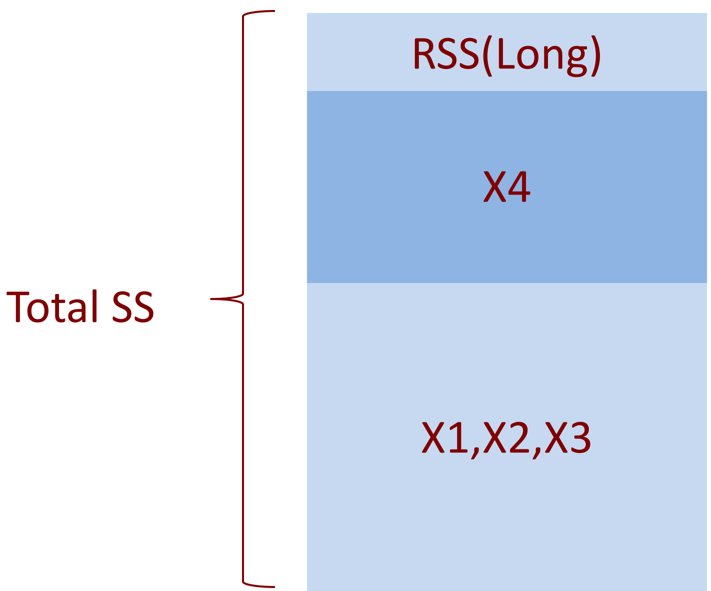
By construction, because OLS minimizes the RSS, the R-sq. of the Long regression will be larger than that of the short regression.

Partial R-squared

What proportion of variance of Y can be Explained by X4, after we account for by differences due to X1, X2 and X3?

$$R_S^2 = \frac{TSS - RSS_S}{TSS} \qquad R_L^2 = \frac{TSS - RSS_L}{TSS} \qquad R_{L|S}^2 = \frac{RSS_S - RSS_L}{RSS_S}$$

(Partial R-squared)



Suppose we observe a given increase in R-squared between the Short and the Long Regression.

Is this statistically significant?

F-TEST

Special Cases of the F-test

(1) Suppose that the Short regression is the Null model. Then this is a test of the model as a whole. We are testing whether at least one predictor has an effect different than zero.

(2) Suppose that the Short and Long Regression differ only in 1 predictor (1df). Then the F-test is equivalent to the t-test (actually the F-ratio equals the squared of the t-statistic).

Let's verify this.....