General F tests in regression
STT 422: Summer, 2004
Vince Melfi

The text doesn't cover general F testing in linear models, which is important both in regression models and in ANOVA models. This handout covers the basics of F testing. For more, see a book on linear models.

The general setting

Consider a data set with n observations on a response variable y and p predictor variables $x_1, x_2, \ldots, x_p$. This includes simple linear regression as the special case with $p = 1$. Consider two models: A "small" model, and a larger model which contains all of the predictors in the small model plus at least one other predictor. The general F testing procedure will allow us to test the null hypothesis that a the small model is adequate versus the alternative hypothesis that the larger model is required.

Here are several examples of the models we can compare using the general F testing procedure; there are many other examples.

- **Test of all predictors** The null hypothesis specifies the no-predictor model $y_i = \beta_0 + \epsilon_i$. The alternative hypothesis specifies the "full" model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$. Another way to write the hypotheses is

  $$H_0 \colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

  versus

  $$H_a \colon \text{at least one of the } \beta_i \text{ is non zero.}$$

- **Test of one predictor** The null and alternative hypotheses differ by just one predictor, say x_k. We can informally write the hypotheses in the form

  $$H_0 \colon \beta_k = 0$$

  versus

  $$H_0 \colon \beta_k \neq 0.$$

  But we must be careful to understand that this can mean different things depending on the larger model being considered. One extreme is the following:

  $$H_0 \colon y_i = \beta_0 + \epsilon_i$$

  versus

  $$H_a \colon y_i = \beta_0 + \beta_k x_{ik} + \epsilon_i,$$

  where we're assessing the significance of the predictor x_k by itself. The other extreme is

  $$H_0 \colon y_i = \beta_0 + \cdots + \beta_k x_{i,k-1} + \beta_k x_{i,k+1} + \epsilon_i$$
  $$H_a \colon y_i = \beta_0 + \cdots + \beta_k x_{i,k-1} + \beta_k x_{ik} + \beta_k x_{i,k+1} + \epsilon_i,$$

  where we're assessing the value of adding x_k to a model already containing all the other predictors.

- **Testing a pair of predictors** As a third example, we can compare a model with several predictors to the model where two of the predictors, say x_j and x_k, have been removed. Note that this is different from testing the two predictors separately.

The test statistic and its distribution

Recall that we're comparing a small model to a larger model. If we fit each using least squares, we'll have two error sums of squares, one for the small model, denoted SSE_S, and one for the larger model denoted SSE_L. It makes sense to decide between the models by comparing these. Clearly $\text{SSE}_S \geq \text{SSE}_L$. If SSE_S is "a lot" larger than SSE_L, then we should decide in favor of the larger model, since we've reduced the amount of error "a lot." On the other hand, if SSE_S isn't much larger than SSE_L, then the larger model doesn't reduce the error by much, and we should not reject the small model.

More specifically, we'll compute

$$F = \frac{(\text{SSE}_S - \text{SSE}_L)/(df_S - df_L)}{\text{SSE}_L/df_L}.$$

Here df_S and df_L denote the degrees of freedom associated with the error sums of squares. This is always the sample size n minus the number of predictors in the model minus 1. For example, for the full model containing all p predictors, the degrees of freedom are $n-p-1$. If the null hypothesis is true, the F statistic has an F distribution with $df_S - df_L$ numerator degrees of freedom and df_L denominator degrees of freedom.

An example

We'll use the general F testing procedure in the context of the fuel consumption dataset from *Applied Linear Regression* by Sanford Weisberg. There are $n = 48$ observations from the 48 contiguous states in the United States. The response variable `fuel` gives the motor fuel consumption in gallons per person per year in 1972. The predictors `tax, inc, road` and `dlic` give the tax rate in cents per gallon on motor fuel, the per-capita income in thousands of dollars, the number of thousands of miles of federal highways, and the percentage of the population holding drivers licenses, respectively. The first fifteen observations are given in Figure 1. Various regression models were fit in SAS. The SAS program is in Figure 2. The output (with some parts removed to save paper) is in Figures 3–7.

- **Test of all predictors** Here we're comparing the model with all four predictors (the alternative hypothesis) to the model with no predictors (the null hypothesis). In this case SSE_L is the error sum of squares from the full model, 189050, given in Figure 4, and SSE_S is the total sum of squares, 588366, since the total sum of squares is the sum of squares with no predictors. The relevant degrees of freedom are $df_S = 47$ and $df_L = 43$. These can be read off the SAS output, or computed via $n - 1 = 48 - 1 = 47$ and $n - p - 1 = 48 - 4 - 1 = 43$. So the F statistic is

$$F = \frac{(588366 - 189050)/(47 - 43)}{189050/43} \approx 22.71.$$

Compare this to the F distribution with 4 and 47 degrees of freedom to compute the p-value. Using the table in the text with 4 and 40 degrees of freedom (there isn't an entry for 4 and 47 degrees of freedom), we find the p-value is less than 0.001.

This is the test that's reported in the Analysis of Variance table. There SAS reports that the p-value is less than 0.0001. Using statistical software, the p-value is actually about 1.72×10^{-10}.

- **Test of one predictor** We'll focus on the predictor `inc`. First, we compare the full model to the model including all the predictors except `inc`. The appropriate sums of squares can be found in Figure 4 and Figure 5. The F statistic is

$$F = \frac{(254779 - 189050)/(44 - 43)}{189050/43} \approx 14.95.$$

From the table in the text with 1 and 40 degrees of freedom, the p-value is seen to be less than 0.001. Using statistical software, the p-value is computed to be approximately 0.00037.

The test is equivalent to the t test reported in the output for the full model. From Figure 4 the t statistic is seen to be -3.87, with p-value 0.0004.

Next, we compare the model including predictors `tax` and `inc` to the model including only `tax`. The F statistic is

$$F = \frac{(468543 - 434889)/(46 - 45)}{434889/45} \approx 3.48.$$

From the table in the text with 1 and 40 degrees of freedom, the p-value is bounded between 0.05 and 0.1. Using statistical software, the p-value is computed to be approximately 0.0686.

Again, there is an equivalent t test reported in Figure 8. The t statistic is reported to be -1.87, with p-value 0.0685.

- **Testing a pair of predictors** Now we'll compare the model including all four predictors to the model including only `dlic` and `road`. So we're testing whether the pair of predictors `inc, tax` can be removed from the full model.

The F statistic is

$$F = \frac{(298509 - 189050)/(45 - 43)}{189050/43} \approx 12.45.$$

Comparing to the F distribution with 2 and 40 degrees of freedom from the text leads to a p-value less than 0.001. Using statistical software, the p-value is computed to be approximately 5.425×10^{-5}.

Obs	state	tax	inc	road	dlic	fuel
1	ME	9.00	3.571	1.976	52.5	541
2	NH	9.00	4.092	1.250	57.2	524
3	VT	9.00	3.865	1.586	58.0	561
4	MA	7.50	4.870	2.351	52.9	414
5	RI	8.00	4.399	0.431	54.4	410
6	CN	10.00	5.342	1.333	57.1	457
7	NY	8.00	5.319	11.868	45.1	344
8	NJ	8.00	5.126	2.138	55.3	467
9	PA	8.00	4.447	8.577	52.9	464
10	OH	7.00	4.512	8.507	55.2	498
11	IN	8.00	4.391	5.939	53.0	580
12	IL	7.50	5.126	14.186	52.5	471
13	MI	7.00	4.817	6.930	57.4	525
14	WI	7.00	4.207	6.580	54.5	508
15	MN	7.00	4.332	8.159	60.8	566

Figure 1: First fifteen observations in the fuel data set

```
options ls = 70;

data fuel;
  infile 'u:\msu\course\stt\422\summer04\fuel.dat' DLM='09'x;
  input state $ tax inc road dlic fuel;

proc corr data = fuel;

proc corr data = fuel;

proc reg data = fuel;
  dlic_road_inc_tax: model fuel = dlic road inc tax;
  dlic_road_tax: model fuel = dlic road tax;
  tax: model fuel = tax;
  dlic_road: model fuel = dlic road;
  inc_tax: model fuel = inc tax;
run;
run;
```

Figure 2: The SAS program

Pearson Correlation Coefficients, N = 48
Prob > |r| under H0: Rho=0

	tax	inc	road	dlic	fuel
tax	1.00000	0.01267	-0.52213	-0.28804	-0.45128
		0.9319	0.0001	0.0471	0.0013
inc	0.01267	1.00000	0.05016	0.15707	-0.24486
	0.9319		0.7349	0.2863	0.0935
road	-0.52213	0.05016	1.00000	-0.06413	0.01904
	0.0001	0.7349		0.6650	0.8978
dlic	-0.28804	0.15707	-0.06413	1.00000	0.69897
	0.0471	0.2863	0.6650		<.0001
fuel	-0.45128	-0.24486	0.01904	0.69897	1.00000
	0.0013	0.0935	0.8978	<.0001	

Figure 3: Correlation matrix for the fuel data set

The REG Procedure
Model: dlic_road_inc_tax
Dependent Variable: fuel

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	399317	99829	22.71	<.0001
Error	43	189050	4396.51089		
Corrected Total	47	588366			

Root MSE	66.30619	R-Square	0.6787
Dependent Mean	576.77083	Adj R-Sq	0.6488
Coeff Var	11.49611		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	377.29115	185.54119	2.03	0.0482
dlic	1	13.36449	1.92298	6.95	<.0001
road	1	-2.42589	3.38917	-0.72	0.4780
inc	1	-66.58875	17.22175	-3.87	0.0004
tax	1	-34.79015	12.97020	-2.68	0.0103

Figure 4: Regression output for the full model with all four predictors

The REG Procedure
Model: dlic_road_tax
Dependent Variable: fuel

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	333588	111196	19.20	<.0001
Error	44	254779	5790.42566		
Corrected Total	47	588366			

Root MSE	76.09485	R-Square	0.5670
Dependent Mean	576.77083	Adj R-Sq	0.5374
Coeff Var	13.19325		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	229.71300	208.37824	1.10	0.2763
dlic	1	11.93318	2.16559	5.51	<.0001
road	1	-3.95045	3.86310	-1.02	0.3121
tax	1	-40.62749	14.78379	-2.75	0.0087

Figure 5: Regression output for the model including predictors `dlic, road, tax`

The REG Procedure
Model: tax
Dependent Variable: fuel

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	119823	119823	11.76	0.0013
Error	46	468543	10186		
Corrected Total	47	588366			

Root MSE	100.92435	R-Square	0.2037
Dependent Mean	576.77083	Adj R-Sq	0.1863
Coeff Var	17.49817		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	984.00763	119.62361	8.23	<.0001
tax	1	-53.10630	15.48359	-3.43	0.0013

Figure 6: Regression output for the model including the predictor `inc`

The REG Procedure
Model: dlic_road
Dependent Variable: fuel

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	289858	144929	21.85	<.0001
Error	45	298509	6633.52609		
Corrected Total	47	588366			

Root MSE	81.44646	R-Square	0.4926
Dependent Mean	576.77083	Adj R-Sq	0.4701
Coeff Var	14.12111		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-243.47745	125.61175	-1.94	0.0589
dlic	1	14.18137	2.14614	6.61	<.0001
road	1	2.05505	3.40961	0.60	0.5497

Figure 7: Regression output for the model including predictors `dlic` and `road`

The REG Procedure
Model: inc_tax
Dependent Variable: fuel

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	153478	76739	7.94	0.0011
Error	45	434889	9664.19230		
Corrected Total	47	588366			

Root MSE	98.30662	R-Square	0.2609
Dependent Mean	576.77083	Adj R-Sq	0.2280
Coeff Var	17.04431		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1179.16917	156.57064	7.53	<.0001
inc	1	-46.65322	25.00009	-1.87	0.0685
tax	1	-52.74981	15.08319	-3.50	0.0011

Figure 8: Regression output for the model including predictors `inc` and `tax`