34. TESTING WHETHER THE MODEL IS USEFUL FOR PREDICTING *Y*

It is always possible that our "explanatory variables" are completely useless for predicting *y*. We can formulate this as the null hypothesis that all regression parameters (except the intercept) are zero, that is,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 .$$

Under H_0, the true regression function does not depend at all on our explanatory variables, although our estimates $\hat{\beta}_1, \cdots, \hat{\beta}_k$ will almost certainly be different from zero, due to natural variability, i.e., "noise", as opposed to "signal".

To determine if the model is useful, we would like to perform a hypothesis test of H_0.

The alternative hypothesis H_A is that H_0 is false.

So H_A says that at least one of $\beta_1, ..., \beta_k$ is nonzero.

In other words, under H_A, the model is not completely useless.

(But notice that H_A does not say that *all* coefficients are nonzero).

You might think that you could test H_0 versus H_A by checking to see if any of the estimated coefficients is significantly different from zero. The problem with this approach is that you would be performing *k* different hypothesis tests at the same time. Even if H_0 is true, the probability of finding *at least one* significant *t*-statistic at (let's say) the 5% level of significance is actually *much larger* than 0.05.

The larger the value of *k* is, the worse the problem becomes.

The level of significance of the *t*-test gives the probability of a Type I error for that test by itself. But in the scenario above, we have *many chances* to find a significant *t*-statistic. This affects the overall probability of a Type I error.

Here's an analogy: If we draw a card at random from a shuffled deck, we only have a 1/52 chance of obtaining the Ace of Spades. But if we repeat this experiment 100 times, the probability that we will draw the Ace of Spades at least once is much larger. In fact, it's almost 1.

A better way to test H_0 above is to use the "*F*-test".

This allows you to test H_0 versus H_A, with an actual Type I error rate of α .

The test statistic is $F = \dfrac{SSR/k}{SSE/(n-k-1)}$.

To understand the *F*-statistic better, we need to return to the Analysis of Variance part of the Minitab output.

Regression Analysis: Price versus Size, Age, Lot Size

```
The regression equation is
Price = - 161 + 41.5 Size - 2.36 Age + 48.3 Lot Size

Predictor    Coef  SE Coef      T      P
Constant   -160.6    190.7  -0.84  0.418
Size       41.462    7.512   5.52  0.000
Age        -2.361    8.812  -0.27  0.794
Lot Size   48.309    9.011   5.36  0.000

S = 68.9399   R-Sq = 91.6%   R-Sq(adj) = 89.3%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       3  570744  190248  40.03  0.000
Residual Error  11   52280    4753
Total           14  623024
```

We can write the F-statistic as

$$F = \frac{MSR}{MSE}, \text{ where } MSR = \frac{SSR}{k} \text{ and } MSE = \frac{SSE}{n-k-1}.$$

Here, MS stands for "Mean Square".

So MSE is the "Mean Square for Residuals", that is, the ratio of *SSE* to the (Residual) degrees of freedom, $n–k–1$. Note that *MSE* is the same as s^2.

In the housing example, from the Minitab output, we find that *MSE* = 4753, which is the ratio *SSE*/DF(Residual) = 52280/11.

Minitab gives three entries for degrees of freedom in the DF column: DF(Regression), DF(Residual), and DF(Total).

We have already mentioned DF(Residual) = $n-k-1$. This is the DF we have been using to calculate confidence intervals and tests on the individual regression parameters.

In fact, DF(Residual) is the number of degrees of freedom available for estimating σ^2. The estimator, s^2, is based on the residuals, which have lost $k + 1$ degrees of freedom from the original n because of the need to estimate $k + 1$ parameters in order to calculate the residuals.

DF(Regression) is k, the number of regression parameters, not counting the intercept.

DF(Total) is $n-1$, the same as the df we have used in estimating the variance of the y's.

For the housing example, DF(Regression) = 3, DF(Residual) = 11, and DF(Total) = 14.

Note: The value of n does not appear anywhere in the output, but we can get it by adding one to DF(Total).

The Mean Square for Regression is defined as
$MSR = SSR/k = SSR/\text{DF(Regression)}$

MSR represents the "average" squared fluctuation of the $\hat{y}$ values about their mean, $\bar{y}$.

In the housing example, $MSR = 190248 = 570744 / 3$.

The F statistic is the ratio $F = MSR/MSE$.

In the housing example, $F = 40.03 = 190248 / 4753$.

It is clear from its definition that the F statistic is always positive.

If $F = MSR/MSE$ is sufficiently large, we can reject H_0 in favor of H_A .

ANOVA					
Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$	p-value
Residual Error	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$		
Total	$n-1$	SST= SSR+SSE			

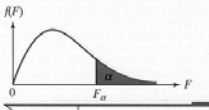
If H_0 is true, then the F-statistic has an F distribution with k, and $n–k–1$ degrees of freedom.

Note that the F distribution has two different df: one for the numerator (ν_1), one for the denominator (ν_2).

We have $\nu_1 = k$, $\nu_2 = n–k–1$.

The F distribution is skewed to the right, since the F-statistic can never be negative. If the F-statistic exceeds F_α (Table 9), then we reject H_0 at level α. This test is inherently right-tailed, since we are looking for a large value of the F statistic as evidence that the regression is useful. Thus, we use the critical value F_α, and not $F_{\alpha/2}$.

Table 9 Critical Values for the F Statistic: $F_{.05}$



Denominator Degrees of Freedom (ν_2) \ Numerator Degrees of Freedom (ν_1)	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Source: From M. Merrington and C. M. Thompson, "Tables of Percentage Points of the inverted Beta (F)-Distribution." *Biometrika*, 1943, 33, pp. 73–88. Reproduced by permission of the *Biometrika* trustees.

(continued)

Fortunately, we can avoid using tables here by just looking at the *p*-value reported by Minitab.

In the housing example, Minitab gives a *p*-value for the *F*-test of $p = 0.000$. So we can reject H_0 at level 0.05, and also at level 0.01. The model does seem to have predictive power.

For a manual calculation, suppose we wanted to do a formal test at level 0.05. From Table 9 with $\nu_1 = 3$, $\nu_2 = 11$, we find $F_{0.05} = 3.59$. Since the observed *F* statistic exceeds 3.59, we reject H_0 at level 0.05. (We already knew this had to happen.)

Practical Interpretation of the F-Test

Most books recommend that in analyzing multiple regression computer output we look first at the *F*-test. If the *F*-statistic is not significant, we do not continue with the analysis, since the regression is not useful for prediction.

From a practical point of view, there are some problems with the *F*-test. If the *F*-statistic exceeds the critical value, then we have some indication that *at least one* of the β_i is nonzero. However, the test gives us no clue as to *which* of the β_i is (are) nonzero.

Unfortunately, this is precisely what we will typically want to know in practice.

Going back to the individual *t*-statistics for each parameter and picking those which are significant does not solve the problem, due partly to the multiple testing issues we discussed earlier.

It is tempting to conclude that if H_0 is rejected, the model, with all *k* variables, must be "good". But this is not necessarily true.

In the housing example, the coefficient for age is not significant (p=0.794).

So even though we have a highly significant *F*-statistic, (p=0.000), it is not clear that the age variable has predictive power, given that the other two variables are in the model. Perhaps the age variable should be deleted. The *F*-statistic tells us nothing about this.

So even if H_0 is rejected, we need to worry about the possibility that we have too many variables in our model. We also may be missing some variable that is of great importance.

To recap: A significant *F*-statistic simply suggests that the model is not completely useless. It does *not* indicate that the model is "correct", or even "adequate".