Unlike SS(Total) and SS(Residual), we don't interpret SS(Regression) in terms of prediction error. Rather, it measures the extent to which the predictions $\hat{y}_i$ vary. If SS(Regression) = 0, the predicted y values ($\hat{y}$) are all the same. In such a case, information about the xs is useless in predicting y. If SS(Regression) is large relative to SS(Residual), the indication is that there is real predictive value in the independent variables $x_1, x_2, \ldots, x_k$. We state the test statistic in terms of mean squares rather than sums of squares. As always, a mean square is a sum of squares divided by the appropriate df.

F Test of H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_a: At least one $\beta \neq 0$.

T.S.: $F = \dfrac{\text{SS(Regression)}/k}{\text{SS(Residual)}/[n - (k + 1)]} = \dfrac{\text{MS(Regression)}}{\text{MS(Residual)}}$

R.R.: With $df_1 = k$ and $df_2 = n - (k + 1)$, reject H_0 if $F > F_\alpha$.

Check assumptions and draw conclusions.

EXAMPLE 12.11

The following SAS output is provided for fitting the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ to the maximal oxygen uptake data of Example 12.6.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6.10624	1.52656	17.02	<.0001
Error	49	4.39376	0.08967		
Corrected Total	53	10.50000			

Root MSE	0.29945	R-Square	0.5815
Dependent Mean	2.00000	Adj R-Sq	0.5474
Coeff Var	14.97236		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.58767	1.02985	5.43	<.0001
x1	1	0.01291	0.00283	4.57	<.0001
x2	1	-0.08300	0.03484	-2.38	0.0211
x3	1	-0.15817	0.02658	-5.95	<.0001
x4	1	-0.00911	0.00251	-3.64	0.0007

Use this information to answer the following questions.

a. Locate SS(Regression).
b. Locate the F statistic.
c. Is there substantial evidence that the four independent variables x_1, x_2, x_3, x_4 as a group have at least some predictive power? That is, does the evidence support the contention that at least one of the βs is

Solution

a. SS(Regression) is shown in the Analysis of Variance table as **SS(Model)** with a value of 6.10624.

b. The MS(Regression) is given as MS(Model) = 1.52656, which is just SS(Regression)/df = SS(Model)/df = 6.10624/4. MS(Residual) is given as MS(Error) = .08967, which is just MS(Residual)/df = SS(Error)/df = 4.39376/49 = .08967.

The F statistic is given as 17.02, which is computed as follows

$$F = \frac{\text{MS(Regression)}}{\text{MS(Residual)}} = \frac{1.52656}{.08967} = 17.02$$

c. For $df_1 = 4$, $df_2 = 49$, and $\alpha = .01$, the tabled F value is 3.73. The computed F is 17.02 which is much larger than 3.73. Therefore, there is strong evidence (p-value $< .0001$, much smaller than $\alpha = .01$) in the data to reject the null hypothesis and conclude that the four explanatory variables collectively have at least some predictive value.

This F test may also be stated in terms of R^2. Recall that $R^2_{y \cdot x_1 \cdots x_k}$ measures the reduction in squared error for y attributed to how well the xs predict y. Because the regression of y on the xs accounts for a proportion $R^2_{y \cdot x_1 \cdots x_k}$ of the total squared error in y,

$$\text{SS(Regression)} = R^2_{y \cdot x_1 \cdots x_k}\text{SS(Total)}$$

The remaining fraction, $1 - R^2$, is incorporated in the residual squared error:

$$\text{SS(Residual)} = (1 - R^2_{y \cdot x_1 \cdots x_k})\text{SS(Total)}$$

F and R^2

The overall F test statistic can be rewritten as

$$F = \frac{\text{MS(Regression)}}{\text{MS(Residual)}} = \frac{R^2_{y \cdot x_1 \cdots x_k}/k}{(1 - R^2_{y \cdot x_1 \cdots x_k})/[n - (k + 1)]}$$

This statistic is to be compared with tabulated F values for $df_1 = k$ and $df_2 = n - (k + 1)$.

EXAMPLE 12.12

A large city bank studies the relation of average account size in each of its branches to per capita income in the corresponding zip code area, number of business accounts, and number of competitive bank branches. The data are analyzed by Statistix, as shown here:

```
CORRELATIONS (PEARSON)

          ACCTSIZE  BUSIN    COMPET

BUSIN     -0.6934
COMPET     0.8196  -0.6527
INCOME     0.4526   0.1492   0.5571

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF ACCTSIZE
```

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P	VIF
CONSTANT	0.15085	0.73776	0.20	0.8404	
BUSIN	-0.00288	8.894E-04	-3.24	0.0048	5.2
COMPET	-0.00759	0.05810	-0.13	0.8975	7.4
INCOME	0.26528	0.10127	2.62	0.0179	4.3

R-SQUARED	0.7973	RESID. MEAN SQUARE (MSE)	0.03968
ADJUSTED R-SQUARED	0.7615	STANDARD DEVIATION	0.19920

SOURCE	DF	SS	MS	F	P
REGRESSION	3	2.65376	0.88458	22.29	0.0000
RESIDUAL	17	0.67461	0.03968		
TOTAL	20	3.32838			

a. Identify the multiple regression prediction equation.
b. Use the R^2 value shown to test H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. (*Note:* $n = 21$.)

Solution

a. From the output, the multiple regression forecasting equation is

$$\hat{y} = 0.15085 - 0.00288x_1 - 0.00759x_2 + 0.26528x_3$$

b. The test procedure based on R^2 is

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_a: At least one β_j differs from zero.

T.S.: $$F = \frac{R^2_{y \cdot x_1 x_2 x_3}/3}{(1 - R^2_{y \cdot x_1 x_2 x_3})/(21 - 4)} = \frac{.7973/3}{.2027/17} = 22.29$$

R.R.: For $df_1 = 3$ and $df_2 = 17$, the critical .05 value of F is 3.20

Because the computed F statistic, 22.29, is greater than 3.20, we reject H_0 and conclude that one or more of the x values has some predictive power. This also follows because the p-value, shown as .0000, is (much) less than .05. Note that the F value we compute is the same as that shown in the output.

Rejection of the null hypothesis of this F test is not an overwhelmingly impressive conclusion. This rejection merely indicates that there is good evidence of *some* degree of predictive value *somewhere* among the independent variables. It does not give any direct indication of how strong the relation is, nor any information of which individual independent variables are useful. The next task, therefore, is to make inferences about the individual partial slopes.

To make these inferences, we need the estimated standard error of each partial slope. As always, the standard error for any estimate based on sample data indicates how accurate that estimate should be. These standard errors are calculated and shown by most regression computer programs. They depend on the sample size, the residual standard deviation, the amount of variation in the predictor variable, and the degree of correlation between that predictor and the other predictors. The expression that we present for the standard error is useful in considering the effects of collinearity (correlated independent variables), but it is *not* a particularly efficient way to do the computation. Let a computer program do the arithmetic.