led

VIF

5.2
7.4
4.3

3968
9920

$n = 21$.)

on is

29

of F is 3.20.
reject H_0 and
This also fol-
te that the F

elmingly im-
l evidence of
variables. It
ny indication
erefore, is to

of each par-
iple data in-
e computed
hree things:
or variable,
dictors. The
ng the effect
ularly good

P650

DEFINITION 12.3

Estimated standard error of $\hat{\beta}_j$ in a multiple regression:

$$s_{\hat{\beta}_j} = s_\varepsilon \sqrt{\frac{1}{\Sigma (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}}$$

$x_j = b_0 + b_1 x_1 \cdots x_k$ without x_j

where R_j^2 is the R^2 value obtained by letting x_j be the *dependent* variable in a multiple regression, with all other xs independent variables. Note that s_ε is the residual standard deviation for the multiple regression of y on $x_1, x_2, \ldots, x_k$.

As in simple regression, the larger the residual standard deviation, the larger the uncertainty in estimating coefficients. Also, the less variability there is in the predictor, the larger is the standard error of the coefficient. The most important use of the formula for estimated standard error is to illustrate the **effect of collinearity.** If the independent variable x_j is highly collinear with one or more other independent variables, R_j^2 is by definition very large and $1 - R_j^2$ is near zero. Division by a near-zero number yields a very large standard error. Thus, one important effect of severe collinearity is that it results in very large standard errors of partial slopes and therefore very inaccurate estimates of those slopes.

effect of collinearity

The term $1/(1 - R_j^2)$ is called the **variance inflation factor** (VIF). It measures how much the variance (square of the standard error) of a coefficient is increased because of collinearity. This factor is printed out by some computer packages and is helpful in assessing how serious the collinearity problem is. If the VIF is 1, there is no collinearity at all. If it is very large, such as 10 or more, collinearity is a serious problem.

variance inflation factor

A large standard error for any estimated partial slope indicates a large probable error for the estimate. The partial slope $\hat{\beta}_j$ of x_j estimates the effect of increasing x_j by one unit while all other xs remain constant. If x_j is highly collinear with other xs, when x_j increases, the other xs also vary rather than staying constant. Therefore, it is difficult to estimate β_j, and its probable error is large when x_j is severely collinear with other independent variables.

The standard error of each estimated partial slope $\hat{\beta}_j$ is used in a confidence interval and statistical test for $\hat{\beta}_j$. The confidence interval follows the familiar format of estimate plus or minus (table value) (estimated standard error). The table value is the t table with the error df, $n - (k + 1)$.

DEFINITION 12.4

The confidence interval for β_j is

$$\hat{\beta}_j - t_{\alpha/2} s_{\hat{\beta}_j} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2} s_{\hat{\beta}_j}$$

where $t_{\alpha/2}$ cuts off area $\alpha/2$ in the tail of a t distribution with df $= n - (k + 1)$, the error df.

EXAMPLE 12.13

Calculate a 95% confidence interval for β_3, the coefficient associated with the explanatory variable *INCOME* in the three-predictor model for the data of Example 12.12.

P 654

Solution The least-squares estimator of $\overset{3}{\beta_{\cancel{1}}}$ is $\overset{3}{\hat{\beta}_{\cancel{1}}}$ = .26528 with standard error $s_{\hat{\beta}_3}$ = .10127. The upper .025 percentile of the t distribution with df = $n - (k + 1)$ = 21 − (3 + 1) = 17 is 2.110. The 95% confidence interval on β_3 is computed as

$$\hat{\beta}_3 \pm t_{\alpha/2}s_{\hat{\beta}_3} \Rightarrow .26528 \pm (2.110)(.10127) \Rightarrow .05160 \le \beta_3 \le .47896$$

EXAMPLE 12.14

Locate the estimated partial slope for x_1 and its standard error in the output in Example 12.11. Calculate a 90% confidence interval for β_1.

```
                        Parameter Estimates

                         Parameter    Standard
Variable        DF       Estimate       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
```

Solution $\hat{\beta}_1$ = .01291 with standard error .00283. The tabled t value for $\alpha/2$ = .10/2 = .05 and df = 54 − (4 + 1) = 49 is 1.677. The 90% confidence interval is computed as follows

$$\hat{\beta}_1 \pm t_{\alpha/2}s_{\hat{\beta}_1} \Rightarrow .01291 - (1.677)(.00283) \le \beta_1 \le .01291 + (1.677)(.00283)$$
$$\Rightarrow .00816 \le \beta_1 \le .01766$$

pretation of H_0: $\beta_j = 0$ ✓

The usual null hypothesis for inference about β_j is H_0: $\beta_j = 0$. This hypothesis does not assert that x_j has no predictive value by itself. It asserts that it has no *additional* predictive value over and above that contributed by the other independent variables; that is, if all other xs had already been used in a regression model and then x_j was added last, the prediction would not improve. The test of H_0: $\beta_j = 0$ measures whether x_j has any additional (e.g., unique) predictive value. The t test of this H_0 is summarized next.

or Testing β_j ✓

H_0: **1.** $\beta_j \le 0$ $\quad H_a$: **1.** $\beta_j > 0$

2. $\beta_j \ge 0$ $\quad$ **2.** $\beta_j < 0$

3. $\beta_j = 0$ $\quad$ **3.** $\beta_j \ne 0$

T.S.: $t = \hat{\beta}_j/s_{\hat{\beta}_j}$

R.R.: **1.** $t > t_\alpha$

2. $t < -t_\alpha$

3. $|t| > t_{\alpha/2}$

where t_α cuts off a right-tail area α in the t distribution with df = $n - (k + 1)$.

Check assumptions and draw conclusions.

This test statistic is shown by virtually all multiple regression programs.

EX
Re
Sol

EX
Re
H_a
any
Sol
(IN
the
spe
twc
H_a:
diff
tioı
In s
whı
Reı
adc
var
be s
hav
gro
GD

12.5 Te

F **test for several** β_j**s**

In t
moı
that
no v

EXAMPLE 12.15

Refer to the output given in Example 12.14.

a. Test $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ at the $\alpha = .10$ level.

b. Is the conclusion of the test compatible with the confidence interval?

Solution

a. The test statistic for $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ is

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{.01291}{.00283} = 4.562$$

The .05 upper percentile for the t distribution with df $= 54 - (4 + 1) = 49$ is 1.677. Because the computed value of the test statistic is greater than the tabled value, we conclude there is significant evidence to reject H_0. Thus, x_1 has additional predictive power in the presence of the other three explanatory variables.

b. The 90% confidence interval for β_1 did not include 0, which indicates that $H_0: \beta_1 = 0$ should be rejected at the $\alpha = .10$ level.

EXAMPLE 12.16

Refer to Example 12.12. Locate the t statistic for testing $H_0: \beta_3 \leq 0$ versus $H_a: \beta_3 > 0$ in the output given in Example 12.12. Do the data support $H_a: \beta_3 > 0$ at any of the usual values for α?

Solution The t statistics are shown under the heading *STUDENT'S T*. For x_3 (INCOME), the t statistic is 2.62, which is computed as .26528/.10127. With df = 17, the tabled values from the t distribution are 2.576 and 2.898 for $\alpha = .01$ and .005, respectively. Thus, H_0 would be rejected at the $\alpha = .01$ level but not at the $\alpha = .005$ level.

The output lists a p-value under the column heading *P*. This p-value is for a two-sided alternative hypothesis, $H_a: \beta_3 \neq 0$. The p-value for the 1-sided alternative $H_a: \beta_3 > 0$ is given by p-value $= Pr(t_{17} > 2.62) = .00896 < .01 = \alpha$.

The multiple regression F and t tests that we discuss in this chapter test different null hypotheses. It sometimes happens that the F test results in the rejection of $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$, whereas no t test of $H_0: \beta_j = 0$ is significant. In such a case, we can conclude that there is predictive value in the equation as a whole, but we cannot identify the specific variables that have predictive value. Remember that each t test is testing the unique predictive value. Does this variable add predictive value, given all the other predictors? When two or more predictor variables are highly correlated among themselves, it often happens that no x_j can be shown to have significant, unique predictive value, even though the xs together have been shown to be useful. If we are trying to predict housing sales based on gross domestic product and disposable income, we probably cannot prove that GDP adds value given DI, or that DI adds value given GDP.

12.5 Testing a Subset of Regression Coefficients

F test for several β_js

In the last section, we presented an F test for testing *all* the coefficients in a regression model and a t test for testing *one* coefficient. Another F test of the null hypothesis tests that *several* of the true coefficients are zero—that is, that several of the predictors have no value given the others. For example, if we try to predict the prevailing wage rate in

P653

various geographical areas for clerical workers based on the national minimum wage, national inflation rate, population density in the area, and median apartment rental price in the area, we might well want to test whether the variables related to area (density and apartment price) added anything, given the national variables.

A null hypothesis for this situation would say that the true coefficients of density and apartment price were zero. According to this null hypothesis, these two independent variables together have no predictive value once minimum wage and inflation are included as predictors.

The idea is to compare the SS(Regression) or R^2 values when density and apartment price are excluded and when they are included in the prediction equation. When they are included, the R^2 is automatically at least as large as the R^2 when they are excluded because we can predict at least as well with more information as with less. Similarly, SS(Regression) will be larger for the complete model. The F test for this null hypothesis tests whether the gain is more than could be expected by chance alone. In general, let k be the total number of predictors, and let g be the number of predictors with coefficients not hypothesized to be zero ($g < k$). Then $k - g$ represents the number of predictors with coefficients that are hypothesized to be zero. The idea is to find SS(Regression) values using all predictors (the **complete model**) and using only the g predictors that do not appear in the null hypothesis (the **reduced model**). Once these have been computed, the test proceeds as outlined next. The notation is easier if we assume that the reduced model contains $\beta_1, \beta_2, \ldots, \beta_g$, so that the variables in the null hypothesis are listed last.

nd reduced models

of a Subset f Predictors

H_0: $\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$ $k - g$

H_a: H_0 is not true.

$$\text{T.S.: } F = \frac{[\text{SS(Regression, complete)} - \text{SS(Regression, reduced)}]/(k-g)}{\text{SS(Residual, complete)}/[n-(k+1)]}$$

$\frac{-SSR(R)}{k-g}$

R.R.: $F > F_\alpha$, where F_α cuts off a right-tail of area α of the F distribution with $\text{df}_1 = (k - g)$ and $\text{df}_2 = [n - (k + 1)]$.

Check assumptions and draw conclusions.

EXAMPLE 12.17

A state fisheries commission wants to estimate the number of bass caught in a given lake during a season in order to restock the lake with the appropriate number of young fish. The commission could get a fairly accurate assessment of the seasonal catch by extensive "netting sweeps" of the lake before and after a season, but this technique is much too expensive to be done routinely. Therefore, the commission samples a number of lakes and records y, the seasonal catch (thousands of bass per square mile of lake area); x_1, the number of lakeshore residences per square mile of lake area; x_2, the size of the lake in square miles; $x_3 = 1$ if the lake has public access, 0 if not; and x_4, a structure index. (Structures are weed beds, sunken trees, drop-offs, and other living places for bass.) The data are shown in Table 12.13.

The commission is convinced that residences and size are important variables in predicting catch because they both reflect how intensively the lake has been

TABLE 12.13
Bass catch data

Lake
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

fishe
usefu
four
all fo
porti

```
Fu

Re

Th
ca

Pr
Co
re
si
ac
st

S

An

So
Re
Re
To
```

um wage,
ent rental
d to area

icients of
sis, these
um wage

nd apart-
equation
vhen they
n as with
F test for
by chance
umber of
epresents
The idea
and using
d model)
otation is
: the vari-

: − g)

ion

ught in a
propriate
ssment of
id after a
herefore,
nal catch
hore resi-
es; x_3 = 1
tures are
data are

variables
has been

TABLE 12.13
Bass catch data

	Y	X_1	X_2	X_3	X_4
Lake	Catch	Residence	Size	Access	Structure
1	3.6	92.2	.21	0	81
2	.8	86.7	.30	0	26
3	2.5	80.2	.31	0	52
4	2.9	87.2	.40	0	64
5	1.4	64.9	.44	0	40
6	.9	90.1	.56	0	22
7	3.2	60.7	.78	0	80
8	2.7	50.9	1.21	0	60
9	2.2	86.1	.34	1	30
10	5.9	90.0	.40	1	90
11	3.3	80.4	.52	1	74
12	2.9	75.0	.66	1	50
13	3.6	70.0	.78	1	61
14	2.4	64.6	.91	1	40
15	.9	50.0	1.10	1	22
16	2.0	50.0	1.24	1	50
17	1.9	51.2	1.47	1	37
18	3.1	40.1	2.21	1	61
19	2.6	45.0	2.46	1	39
20	3.4	50.0	2.80	1	53

fished. However, the commission is uncertain whether access and structure are useful as additional predictor variables. Therefore, two regression models (with all four predictor variables entered linearly) are fitted to the data, the first model with all four variables and the second model without access and structure. The relevant portions of the Minitab output follow:

```
Full Model:

Regression Analysis: catch versus residenc, size, access, structur

The regression equation is
catch = - 2.78 + 0.0268 residenc + 0.504 size + 0.743 access + 0.0511 structur


Predictor       Coef     SE Coef       T      P
Constant     -2.7840      0.8157   -3.41  0.004
residenc    0.026794    0.009141    2.93  0.010
size          0.5035      0.2208    2.28  0.038
access        0.7429      0.2021    3.68  0.002
structur    0.051129    0.004542   11.26  0.000

S = 0.389498   R-Sq = 91.4%   R-Sq(adj) = 89.1%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       4  24.0624  6.0156  39.65  0.000
Residual Error  15   2.2756  0.1517
Total           19  26.3380
```

P655

```
Reduced Model:

Regression Analysis: catch versus residenc, size

The regression equation is
catch = - 0.87 + 0.0394 residenc + 0.828 size

Predictor      Coef     SE Coef       T       P
Constant     -0.871       2.409   -0.36   0.722
residenc    0.03941     0.02733    1.44   0.168
size         0.8280      0.6372    1.30   0.211

S = 1.17387   R-Sq = 11.1%   R-Sq(adj) = 0.6%

Analysis of Variance

Source           DF        SS       MS      F      P
Regression        2     2.913    1.456   1.06  0.369
Residual Error   17    23.425    1.378
Total            19    26.338
```

a. Write the complete and reduced models.

b. Write the null hypothesis for testing that the omitted variables have no (incremental) predictive value.

c. Perform an F test for this null hypothesis.

Solution

a. The complete and reduced models are, respectively,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

and

$\beta_3 = 0$ $\beta_4 = 0$

Reduced

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

The corresponding multiple regression least-squares equations based on the sample data are

$$\text{Complete: } \hat{y} = -2.78 + .0268x_1 + .504x_2 + .743x_3 + .0511x_4$$
$$\text{Reduced: } \hat{y} = -.87 + .0394x_1 + .828x_2$$

b. The appropriate null hypothesis of no predictive power for x_3 and x_4 is $H_0: \beta_3 = \beta_4 = 0$.

c. The test statistic for the H_0 of part (b) makes use of SS(Regression, complete) = 24.0624, SS(Regression, reduced) = 2.913, SS(Residual, complete) = 2.2756, $k = 4$, $g = 2$, and $n = 20$:

$$\text{T.S.:} \quad F = \frac{[\text{SS(Regression, complete)} - \text{SS(Regression, reduced)}]/(4-2)}{\text{SS(Residual, complete)}/(20-5)}$$

$$= \frac{(24.0624 - 2.913)/2}{2.2756/(20-5)} = 69.705$$

The tabled value $F_{.01}$ for 2 and 15 df is 6.36. The value of the test statistic is much larger than the tabled value, so we have conclusive evidence that the access and structure variables add predictive value ($p < .0001$).

12.6 Forec

One of
given ce
just sub
section,

A
two way
long-ru
of y wh
this is t
standar
require

C
This inf
most co
a foreca
priate p
the stan
progran

EXAMPL

An adve
an equa
evision
monthly
inflation
output:

Do