

Chapter 14

- 14.2 a. If a jockey's experience increases by one year, then the predicted horse speed increases by b miles per hour, holding all other things (jockey's sex, weight carried on horse and horse's condition) fixed.
- b. When a jockey is male ($\text{sex}=1$), then the predicted horse speed is c miles per hour faster or slower (depending on the sign of c) than with a female jockey, holding all other things (jockey's experience, weight carried on horse and horse's condition) fixed.

14.4 Using the ASCII data file EX14-4.PRN, the ET printouts are:

1st Degree Polynomial Model

```

=====
Dependent Variable      Y          Number of Observations      10
Mean of Dep. Variable  24.0900   Std. Dev. of Dep. Var.     3.131010
Std. Error of Regr.    3.1581    Sum of Squared Residuals   79.7879
R - squared             .09567    Adjusted R - squared       -.01737
F( 1,  8)              .8464     Prob. Value for F          .38449
=====

```

```

=====
Variable  Coefficient  Std. Error  t-ratio  Prob|t|>x  Mean of X  Std.Dev.of X
-----
INTERCEPT  27.1112    3.432      7.899    .00005
X            -.791624   .8605     -.920    .38449    3.81640    1.22338
=====

```

The least-squares regression equation is

$$\hat{y} = 27.1112 - 0.791624 x, r^2 = 0.09567$$

and the slope coefficient is insignificant at the 0.10 Type I error level.

2nd Degree Polynomial Model

```

=====
Dependent Variable      Y          Number of Observations      10
Mean of Dep. Variable  24.0900   Std. Dev. of Dep. Var.     3.131010
Std. Error of Regr.    3.2627    Sum of Squared Residuals   74.5184
R - squared             .15540    Adjusted R - squared       -.08592
F( 2,  7)              .6440     Prob. Value for F          .55371
=====

```

```

=====
Variable  Coefficient  Std. Error  t-ratio  Prob|t|>x  Mean of X  Std.Dev.of X
-----
INTERCEPT  21.2096    9.107      2.329    .05270
X            2.91584    5.344      .546     .60227    3.81640    1.22338
X2           -.518330   .7367     -.704    .50443    15.91189   8.87412
=====

```

The least-squares regression equation is

$$\hat{y} = 21.2096 + 2.91584 x - 0.51833 x^2, R^2 = 0.15540$$

and neither slope coefficient is significant.

3rd Degree Polynomial Model

```

=====
Dependent Variable          Y                Number of Observations          10
Mean of Dep. Variable      24.0900          Std. Dev. of Dep. Var.         3.131010
Std. Error of Regr.        2.5845           Sum of Squared Residuals      40.0767
R - squared                 .54576           Adjusted R - squared           .31865
F( 3,  6)                  2.4030           Prob. Value for F              .16612
=====
Variable  Coefficient  Std. Error  t-ratio  Prob|t|>x  Mean of X  Std.Dev.of X
-----
INTERCEPT  62.2093    19.44      3.200    .01861
X            -38.1160   18.56     -2.054   .08579    3.81640    1.22338
X2           11.7792    5.447     2.163    .07382   15.91189    8.87412
X3           -1.13818   .5012     -2.271   .06360   70.36077   52.93157

```

The least-squares regression equation is

$$\hat{y} = 62.2093 - 38.1160x + 11.7792x^2 - 1.13818x^3, R^2 = 0.546$$

and all three slope coefficient estimates are significant at the 0.1 Type I error level; this third degree polynomial appears to fit the curvature of the data well.

14.6 a. Using the ASCII data file EX14-6.PRN, the printout is:

```

=====
Dependent Variable          Y                Number of Observations          8
Mean of Dep. Variable      9528.7500          Std. Dev. of Dep. Var.         1007.460846
Std. Error of Regr.        986.7039           Sum of Squared Residuals      .584151E+07
R - squared                 .17781           Adjusted R - squared           .04078
F( 1,  6)                  1.2976           Prob. Value for F              .29808
=====
Variable  Coefficient  Std. Error  t-ratio  Prob|t|>x  Mean of X  Std.Dev.of X
-----
INTERCEPT  8130.65    1276.      6.372    .00070
X            29.0514    25.50     1.139    .29808    48.12500    14.62324

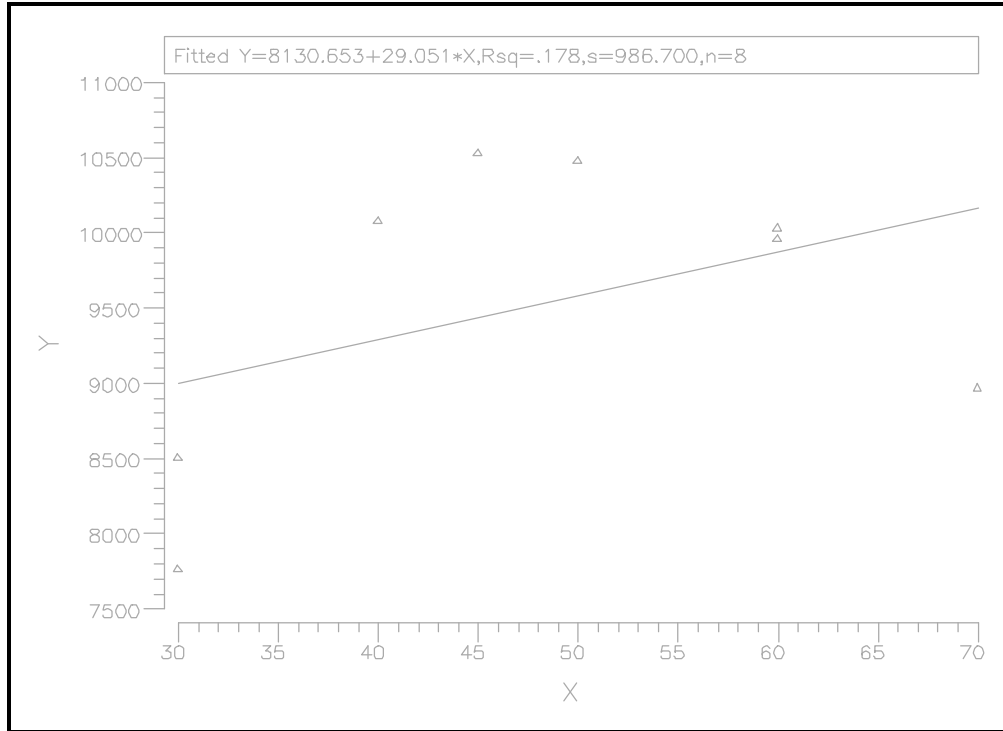
```

1. The least-squares regression equation is

$$\hat{y} = 8130.65 + 29.0514x$$

where y is real earnings and x is age.

2.



P L O T O F R E S I D U A L S				
	-.125E+04	0	.125E+04	
X			Residual
30.0	:	*	:	-507.
30.0	:*	:	:	-.125E+04
40.0	:	:	*	775.
45.0	:	:	*	.108E+04
50.0	:	:	*	884.
60.0	:	:*	:	75.3
60.0	:	: *	:	147.
70.0	:*	:	:	-.121E+04
			

3. This regression equation does not appear to be a good representation of the relationship between two variables. r^2 is only 0.017781 and the residual plot shows a ">" pattern in the following plot of residuals. Thus, a nonlinear relationship is suggested between x and y.

b. Again using the EX14-6.PRN file, after creating the square and cube of x, the regression printout is:

Dependent Variable	Y	Number of Observations	8
Mean of Dep. Variable	9528.7500	Std. Dev. of Dep. Var.	1007.460846

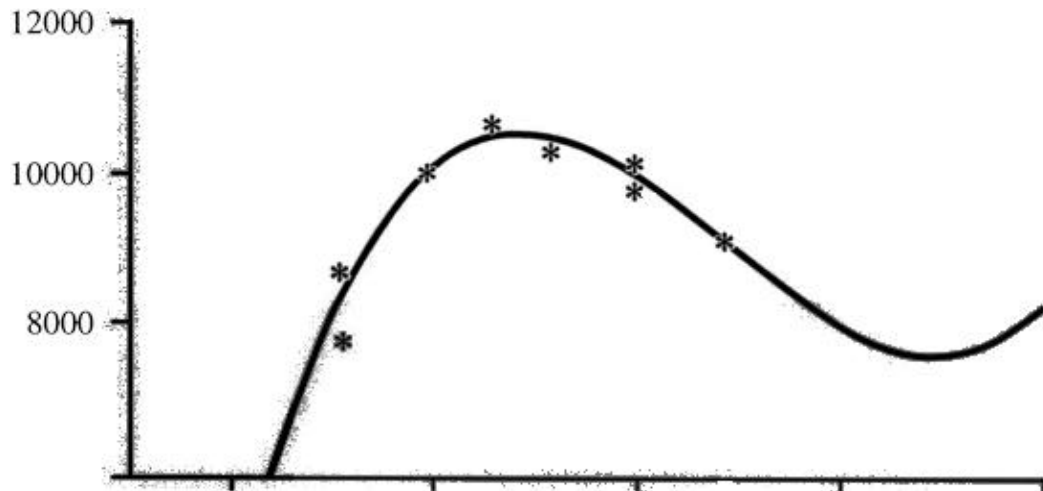
Std. Error of Regr.	265.7941	Sum of Squared Residuals	282586.			
R - squared	.96023	Adjusted R - squared	.93040			
F(3, 4)	32.1896	Prob. Value for F	.00293			
=====						
Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X

INTERCEPT	-12103.3	5954.	-2.033	.11185		
X	1138.43	393.3	2.895	.04435	48.12500	14.62324
X2	-18.1007	8.254	-2.193	.09337	2503.12500	1431.06089
X3	.876612E-01	.5547E-01	1.580	.18919	138640.62500	111274.42424

1. The least-squares regression equation is

$$\hat{y} = -12103.3 + 1138.43x - 18.1007x^2 + 0.0876612x^3$$

2.



3. This regression equation appears to provide better representation of the relationship between two variables. R^2 is 0.96023 and the residual plot does not show any special pattern in the following plot of residuals:

```

          P L O T   O F   R E S I D U A L S
-370.          0          370.
X
..... Residual
30.0          :          * :          369.
30.0          :*          :          -370.
40.0          :          * :          -15.1
45.0          :          : *          :          56.7
50.0          :          * :          :          -57.1
60.0          :          : *          :          -25.8
60.0          :          : *          :          46.2
70.0          :          * :          :          -4.20
.....

```

14.8 a. Using the ASCII data file EX14-8.PRN, and creating the TOTAL variable as the sum of "Main Fl," "Upper Fl," and "Basement," (because "Fin Base" is included in "Basement"), and arranging the regressors in abutting columns, the computer printout is:

Dependent Variable	PRICE	Number of Observations	13
Mean of Dep. Variable	2238.6154	Std. Dev. of Dep. Var.	249.660282
Std. Error of Regr.	266.9198	Sum of Squared Residuals	641215.
R - squared	.14272	Adjusted R - squared	-.14304
F(3, 9)	.4994	Prob. Value for F	.69187

Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X
INTERCEPT	2719.54	505.1	5.384	.00044		
ROOMS	-4.94027	53.24	-.093	.92810	8.84615	1.62512
TOTAL	-.125848	.1259	-1.000	.34349	3330.76923	669.05968
AGE	-1.33721	3.584	-.373	.71771	13.50000	22.62558

The least-squares regression equation is

$$\hat{PRICE} = 2719.54 - 4.94027ROOMS - 0.125848TOTAL - 1.33721AGE$$

where TOTAL = MAIN-FL + UPPER-FL + BASEMENT

b. At ROOMS = 8, TOTAL = 2500 and AGE = 0, the predicted price is \$236,540 (= 2719.54 - 4.94027(8) - 0.125848(2500) - 1.33721(0)).

14.10 a. For batters, the implied regression equation is

$$\hat{SALARY} = b_1 + 9000HOMERUN + 6000RUN$$

where SALARY is the batter's next year salary, HOMERUN is the number of homeruns and RUN is the number of runs. Silk did not supply the value of the intercept coefficient.

b. For pitchers, the implied regression equation is

$$\hat{SALARY} = b_1 + 38000VICTORY + 16000SAVE + 3000INNING + 12000ERA$$

where SALARY is the pitcher's next year salary, VICTORY is the number of victories, SAVE is the number of saves, INNING is the number of additional innings pitched and ERA is the decrement in e.r.a. measured in the number of one-tenths of point. Silk did not supply the value of the intercept coefficient.

14.12 The death rate of smokers is higher than that of nonsmokers of the same age, but since the dependent variable is measured in natural logarithm units (where the base is $e=2.71828$) assessment of this difference requires the use of the antilog, which is the "=EXP()" command function in EXCEL.

14.14 Those intervals get wider, because the standard error of \hat{y}_* in the formula $\hat{y}_* \pm t_{(n-k)/2} s_{\hat{y}_*}$ becomes larger as the x values are farther from their respective means.

14.16 Although the answer to this question does not require the actual estimation of an interval, using the ASCII data file EX14-8.PRN, the following MINITAB printout can be obtained:

```
MTB > let c10 = c5 + c6 + c7
MTB > name c1 'price' c2 'rooms' c10 'total' c9 'age'
MTB > regress c1 3 c2 c10 c9;
SUBC> predict 8 2500 0.
```

The regression equation is
price = 2720 - 4.9 rooms - 0.126 total - 1.34 age

Predictor	Coef	Stdev	t-ratio	p
Constant	2719.5	505.1	5.38	0.000
rooms	-4.94	53.24	-0.09	0.928
total	-0.1258	0.1259	-1.00	0.343
age	-1.337	3.584	-0.37	0.718

s = 266.9 R-sq = 14.3% R-sq(adj) = 0.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	106748	35583	0.50	0.692
Error	9	641215	71246		
Total	12	747963			

SOURCE	DF	SEQ SS
rooms	1	11388
total	1	85441
age	1	9918

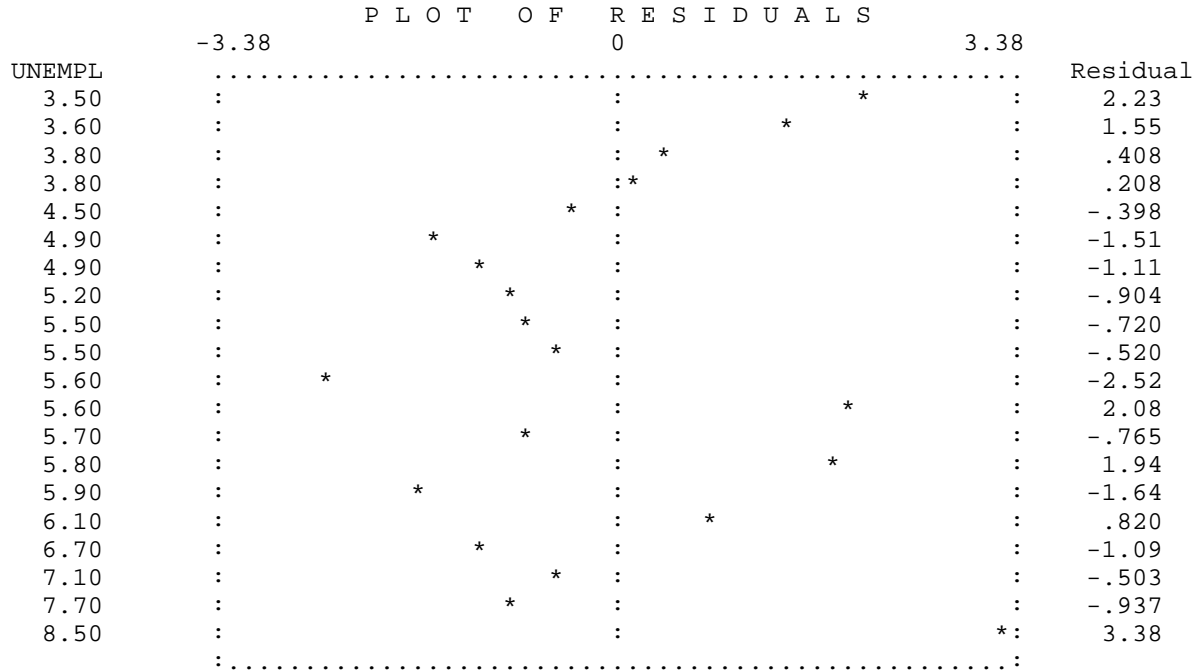
Fit	Stdev.Fit	95% C.I.	95% P.I.
2365.4	128.0	(2075.9, 2654.9)	(1695.6, 3035.2)

Thus, at ROOMS = 8, TOTAL = 2500 and AGE = 0, a 95 percent prediction interval for the individual price is (\$169,560 to \$303,520) and a 95 percent confidence interval for the expected price is \$207,590 to \$265,490. The prediction interval is wider than the confidence interval but they are both centered at the predicted price of \$236,540.

14.18 In Exercise 14.5, the estimated least-squares regression equation was

$$\hat{y} = 3.84521 - 0.277241x + 4.42645\text{DUMMY}$$

where y is the inflation rate and x is the unemployment rate. The dummy variable coefficient has the t value of 5.078, which is significant as low as the 0.00009 Type I error level. The coefficient for the inflation rate has the t value -0.826, which is not significant at typical Type I error levels. The plot of residuals, however, shows a slight "<" pattern, which suggests that the relationship between the inflation rate and the unemployment rate is not linear and that these hypotheses tests are questionable.



14.20 Only the income coefficient is significant at the 0.01 Type I error level because the p-value is 0.00113; p-values are greater than 0.01 for all other regressors.

14.22 We need to test the hypothesis $H_0: \beta_2 \leq 0$ vs. $H_A: \beta_2 > 0$, where β_2 is the SALES coefficient. The p-value of this test is

$$P(b_2 \geq 0.9121) = P(t \geq \frac{0.9121 - 0}{0.7630}) = P(t \geq 1.196) = 0.1772$$

where t has the degrees of freedom 2. Thus, we can reject H_0 at a Type I error level as low as 0.1772, but not at the typical Type I error levels.

14.24 As seen in the EXCEL printout, the coefficient of determination is 0.5446; thus, only 54.46 percent of the variability of price around its mean is explained by this regression. This is not a strong relationship.

East	sqfeet	price	SUMMARY OUTPUT	
0	2,733	1.6		
0	2,241	1.5	Regression Statistics	
0	1,775	1.3	Multiple R	0.73794
0	3,177	2.2	R Square	0.54456
0	2,731	2.2	Adju. R Sq	0.41444
0	1,793	1.4	Std. Error	0.59301
1	3,000	3.6	Observations	10
1	2,575	1.4		
1	2,480	1.4		
1	3,256	2.9		

ANOVA	df	SS	MS	F	Sig. F
Regression	2	2.9434	1.4717	4.1849	0.06375378
Residual	7	2.4616	0.3517		
Total	9	5.4050			

	Coefficient	Std. Error	t Stat	P-value	Lower95%	Upper95%
Intercept	-0.708246	1.03701	-0.68297	0.51659	-3.16039	1.74390
East	0.205599	0.42115	0.48819	0.64034	-0.79026	1.20145
sqfeet	0.001000	0.00042	2.38828	0.04829	0.00001	0.00199

RESIDUAL OUTPUT

Observation	Predicted	Residuals
1	2.02465	-0.42465
2	1.53267	-0.03267
3	1.06669	0.23331
4	2.46864	-0.26864
5	2.02265	0.17735
6	1.08469	0.31531
7	2.49724	1.10276
8	2.07226	-0.67226
9	1.97726	-0.57726
10	2.75323	0.14677

14.26 Below are EXCEL regression results aimed at explaining expected life with wine, beer and liquor consumption

E(life)	Wine	Beer	Liquor
78	63.5	40.1	2.5
78	58.0	25.1	0.9
78	46.0	65.0	1.7
78	15.7	102.1	1.2
77	12.2	100.0	1.5
76	8.9	87.8	2.0
69	2.7	17.1	3.8
73	1.7	140.0	1.0
79	1.0	55.0	2.1
73	0.2	50.4	0.8

SUMMARY OUTPUT E(life) regression Statistics

Multiple R	0.651273764
R Square	0.424157516
Adjusted R Square	0.136236274
Standard Error	2.985960671
Observations	10

ANOVA	df	SS	MS	F
Regression	3	39.40423321	13.1347444	1.473172013
Residual	6	53.49576679	8.915961131	
Total	9	92.9		

	Coefficients	Std. Error	t Stat	P-value
Intercept	74.6174407	4.567088664	16.33807578	3.3477E-06
Wine	0.07672898	0.045242956	1.695932119	0.14082824
Beer	0.01741849	0.033193002	0.524763934	0.61856296
Liquor	-0.86683895	1.295733812	-0.668994624	0.52838504

RESIDUAL OUTPUT

Observation	Predicted E(life)	Residuals
1	78.02111511	-0.021115112
2	78.72477069	-0.724770686
3	77.8055495	0.194450498
4	76.56030682	1.439693179
5	75.99512487	1.004875131
6	75.09599417	0.904005828
7	71.82847711	-2.828477108
8	76.31962965	-3.319629648
9	73.83182484	5.168175159
10	74.81720724	-1.817207241

As seen in this EXCEL printout, the coefficient of determination is 0.4242; thus, only 42.42 percent of the variability of life expectancy around its mean is explained by this regression. This regression does not provide a strong explanation of differences in life expectancy across countries.

- 14.28 $R = 1$ implies that all the observations are exactly on the regression line, showing a perfect fit; all the y values are correctly predicted by the fitted line, so $\text{ErrorSS} = 0$. But, R and thus R^2 are not related to the percentage of correct predictions except in the extreme case where $R = 1$. If $R = 0.9$ then $R^2 = 0.81$, which means that 81 percent of the total variation in y values around this mean is explained by the fitted regression line.
- 14.30 The standard error of the regression, 2665.51038, is equal to the square root of the error sum of squares divided by its degrees of freedom, $\sqrt{0.78154 \cdot 10^8 / 11}$. R^2 is one minus the ratio of the error sum of squares to the total sum of squares, $1 - (0.78154 \cdot 10^8 / 5.91355 \cdot 10^8)$. Thus both depend on the ErrorSS .
- 14.32 $R = 0.424 = \sqrt{39.4 / 92.9}$
- 14.34 The R^2 and r^2 are the ratio of the Regression Sum of Squares ($=\text{TotSS} - \text{ErrorSS}$) to Total Sum of Squares (TotSS) regardless of the number of regressors.
- 14.36 The dummy coefficient estimate -0.987623 means that the enactment of 55 mph speed limit is associated with an estimated 0.987623 person decrease in the death rate, implying a \$1,975,246 savings per 100 million vehicle miles ($= 0.987623(2,000,000)$).
- 14.38 a. We can use the ASCII data file EX14-38.PRN or EX14-38M.PRN or EX14-38A.PRN (notice that EX14-38M.PRN is modification of EX14-38.PRN with variable RYR deleted and cases 157, 164, 168, and 174 deleted because of missing values, and EX14-38A.PRN is the data set for those 15 faculty members to be analyzed in part a). Using one of these, the printouts are:

First Model

```

=====
Ordinary Least Squares
Dependent Variable      INCR%SAL      Number of Observations      15
Mean of Dep. Variable   .2971      Std. Dev. of Dep. Var.      .220964
Std. Error of Regr.     .1715      Sum of Squared Residuals    .323450
R - squared              .52681     Adjusted R - squared         .39776
F( 3, 11)               4.0822     Prob. Value for F            .03562
=====
Variable Coefficient Std. Error t-ratio Prob|t|>x Mean of X Std.Dev.of X
-----
INTERCEPT -.620553E-02 .1085 -.057 .95540
PHDYR .274607E-01 .7942E-02 3.458 .00536 11.26667 5.86109
BOOK -.600495E-02 .6443E-01 -.093 .92742 .60000 .73679
SEX -.758088E-02 .9583E-01 -.079 .93837 .33333 .48795

```

The least-squares regression equation is

$$\widehat{\text{INCR}\% \text{SAL}} = -0.0062 + 0.0275\text{PHDYR} - 0.0060\text{BOOK} - 0.0076\text{SEX}$$

where $\text{INCR}\% \text{SAL} = (\text{CURSAL} - \text{STSAL}) / \text{STSAL}$.

The SEX coefficient estimate is negative, but is not significant at typical Type I error levels.

Second Model

```

=====
Ordinary Least Squares
Dependent Variable      DELTASAL      Number of Observations      15
Mean of Dep. Variable   11946.6667     Std. Dev. of Dep. Var.      8021.410338
Std. Error of Regr.     5873.2515     Sum of Squared Residuals    .379446E+09
R - squared              .57877     Adjusted R - squared         .46389
F( 3, 11)               5.0380     Prob. Value for F            .01947
=====
Variable Coefficient Std. Error t-ratio Prob|t|>x Mean of X Std.Dev.of X
-----
INTERCEPT 511.460 3715. .138 .89299
PHDYR 1042.74 272.0 3.833 .00278 11.26667 5.86109
BOOK -198.069 2207. -.090 .93010 .60000 .73679
SEX -582.326 3282. -.177 .86241 .33333 .48795

```

The least-squares regression equation is

$$\widehat{\Delta \text{SAL}} = 511.460 + 1042.74\text{PHDYR} - 198.069\text{BOOK} - 582.326\text{SEX}$$

where $\Delta \text{SAL} = \text{CURSAL} - \text{STSAL}$.

The SEX coefficient estimate is negative, but is not significant at typical Type I error levels.

In both models, we cannot find significant evidence for a sex discrimination in the department four for those

15 faculty members who are not administrators. Model choice does not seem to affect the conclusion.

b. Using the ASCII data file EX14-38M.PRN, the printout is:

Semilog Linear Model

```

=====
Ordinary Least Squares
Dependent Variable      LCURSAL      Number of Observations      182
Mean of Dep. Variable   10.8969      Std. Dev. of Dep. Var.      .099636
Std. Error of Regr.     .0650        Sum of Squared Residuals    .736193
R - squared              .59028      Adjusted R - squared        .57380
F( 7, 174)              35.8121     Prob. Value for F           .00000
=====
Variable Coefficient Std. Error t-ratio Prob|t|>x Mean of X Std.Dev.of X
-----
INTERCEPT 10.7920      .1109E-01 973.110 .00000
SEX          -.281161E-01 .1233E-01 -2.281 .02377 .20879 .40757
PROYR        .840654E-02 .1037E-02 8.106 .00000 7.22527 5.23258
ADMIN        .895241E-01 .1465E-01 6.110 .00000 .17033 .37696
SERV         .487507E-01 .1120E-01 4.353 .00002 .29670 .45806
ART          .314283E-02 .1340E-02 2.346 .02010 3.73626 4.36418
BOOK         .130432E-01 .7150E-02 1.824 .06985 .35714 .77881
DUMMY        .445299E-01 .1766E-01 2.522 .01256 .08791 .28395

```

The least-squares regression equation is

$$\hat{LCURSAL} = 10.792 - 0.028SEX + 0.008PROYR + 0.090ADMIN + 0.049SERV + 0.003ART + 0.013BOOK + 0.045DUMMY$$

where LCURSAL = Log(CURSAL) and DUMMY = 1 for a person from the department 4, 0 otherwise.

Simple Linear Model

```

=====
Ordinary Least Squares
Dependent Variable      CURSAL      Number of Observations      182
Mean of Dep. Variable   54278.5549  Std. Dev. of Dep. Var.      5500.867612
Std. Error of Regr.     3516.1706   Sum of Squared Residuals    .215124E+10
R - squared              .60722      Adjusted R - squared        .59142
F( 7, 174)              38.4282     Prob. Value for F           .00000
=====
Variable Coefficient Std. Error t-ratio Prob|t|>x Mean of X Std.Dev.of X
-----
INTERCEPT 48397.3      599.5      80.730 .00000
SEX          -1367.80     666.4      -2.053 .04160 .20879 .40757
PROYR        466.010     56.06      8.312 .00000 7.22527 5.23258
ADMIN        5185.94     792.0      6.548 .00000 .17033 .37696
SERV         2685.11     605.3      4.436 .00002 .29670 .45806
ART          165.454     72.42      2.285 .02353 3.73626 4.36418
BOOK         741.397     386.5      1.918 .05674 .35714 .77881
DUMMY        2693.88     954.4      2.823 .00532 .08791 .28395

```

The least-squares regression equation is

$$\hat{CURSAL} = 48397.3 - 1367.8SEX + 466.0PROYR + 5185.9ADMIN + 2685.1SERV + 165.5ART + 741.4BOOK + 2693.9DUMMY$$

1. The semilog salary model is considered by labor economists to be better than the salary level model because salary increases from year to year tend to be based on percentage increases, which yields a nonlinear relationship between salary levels and time dependent regressors. The log transformation of salaries has a greater effect on high salaries than on low salaries thus tending to pull down extremely high salaries to give a linear relationship between log(salary) and regressors.
2. In the semilog model, the DUMMY coefficient estimate has the t value of 2.522, which is significant at an " level as low as 0.01256; we can conclude that the fourth department is significantly different from the others in explaining salary.
3. In the semilog model, the SEX coefficient estimate is negative and has the t value of -2.281, which is significant at an " level 0.05; at " = 0.05, we can conclude that the gender variable plays a significant role in explaining salary, which may imply sex discrimination.

14.40 a. Using the ASCII data file EX14-40.PRN, the printout is:

Dependent Variable	SALES	Number of Observations	22			
Mean of Dep. Variable	491.9091	Std. Dev. of Dep. Var.	119.023371			
Std. Error of Regr.	81.5824	Sum of Squared Residuals	119802.			
R - squared	.59730	Adjusted R - squared	.53018			
F(3, 18)	8.8994	Prob. Value for F	.00078			
=====						
Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X
-----	-----	-----	-----	-----	-----	-----
INTERCEPT	-30.5857	162.5	-.188	.85279		
AD	2.41994	.9249	2.616	.01749	155.59091	21.29757
NUM	26.6538	12.21	2.183	.04251	8.59091	1.59341
COMP	-18.2615	10.40	-1.755	.09624	4.54545	1.76547

The least-squares regression equation is

$$\hat{SALES} = -30.5857 + 2.4199AD + 26.6538NUM - 18.2615COMP$$

- b. At AD = 165(hundreds of dollars), NUM = 10 and COMP = 4, the predicted SALES is \$562,196(= -30.5857 + 2.41994(165) + 26.6538(10) - 18.2615(4)).

c. The AD coefficient estimate has a one-tail p-value of 0.00875. Thus, the AD variable is a significant positive explanatory variable at the 0.01 Type I error level.

14.42 One example of a regression model that might be estimated is:

$$\text{death} = \beta_1 + \beta_2(\text{age}) + \beta_3(\text{health}) + \beta_4(\text{operation}) + \epsilon$$

where age is the average age of those operated on in a year at each of the n hospitals in Pennsylvania; health is an average measure of a health index of those operated on in a year at each of the n hospitals in Pennsylvania; and operation is the number of operations per year performed at each of the n hospitals in Pennsylvania.

We assume that ϵ is distributed normally but death is a count (0, 1, 2, ..., 12, ...); thus, it is not continuous as implied by the assumption that ϵ is normal.

14.44 a. $n = 119 (= 105 + 14)$ where 105(=n-k) is the degrees of freedom of the t statistic and 14(=k) is number of regressors including the intercept.

b. The multiple coefficient of determination(R^2) of 0.388 means that 38.8 percent of the total variability in the profit is explained by the fitted regression line.

14.46 Yes, the inclusion of a dummy variable may be able to cure the problem. If we use the dummy variable of DUMMY = 1 for women and 0 for men, we could separate the gender effect on salary. See Exercise 14.5 as an example. In that exercise, the simple linear regression of the inflation rate on the unemployment rate results in a positive slope coefficient. But with an introduction of the dummy variable of DUMMY = 0 for 1960s and 1 for 1970s, we can obtain a negative slope coefficient.

14.48 Using the ASCII data set EX12-49.PRN) or ROADTEST.PRN, we obtain the following EXCEL printout. (Notice that the ASCII file consists of 125 records, but the records of 4 cars are deleted in the calculations because they have

missing data. Also notice that you must extend the default fixed width of the price column to capture the \$471,375 prices of the Ferraris of interest here.)

```
SUMMARY OUTPUT      Regression Statistics
                    Multiple R      0.787118404
                    R Square      0.619555381
                    Standard Error  34235.17742
                    Observations   121
```

```
ANOVA      df  SS      MS      F      Signif F
Regression  6  217589886177  36264981030  30.94157  8.00233E-22
Residual   114  133613400549  1172047373
Total     120  351203286726
```

```
                Coefficients StandardError  t Stat  P-value  Lower95%  Upper95%
Intercept  -439071.8804  211203.6648  -2.0789  0.03987  -857464.90 -20678.87
to60       13243.13296  5437.0019  2.4357  0.01641  2472.47  24013.80
quarter   -17138.88094  9747.2627  -1.7583  0.08138  -36448.14  2170.38
top        1247.477812  257.0141  4.8537  0.00000  738.33  1756.62
brake      1344.42749  271.9429  4.9438  0.00000  805.71  1883.14
mpg        -1613.595796  645.4710  -2.4999  0.01385  -2892.27  -334.92
grip       332880.8595  98565.3416  3.3773  0.00100  137623.59  528138.13
```

Ferrari F40

```
                price  to60  quarter  top  brake  mpg  grip
Actual          471375   4.2    12.1  197   218   12  1.01
Predicted       264854
Error           206521
```

The least-squares regression equation is

$$\hat{\text{PRICE}} = -439072 + 13243 \text{ TO60} - 17139 \text{ QUARTER} + 1248 \text{ TOP} + 1344 \text{ BRAKING} - 1614 \text{ FUEL} + 332881 \text{ HOLD}$$

The 30th complete record is for the Ferrari F40, which has the largest residual, which is calculated in the EXCEL printout to be 206,521: Actual price - Predicted price = 471375 - (-439072 + 13243(4.2) - 17139(12.1) + 1248(197) + 1344(218) - 1614(12) + 332881(1.01)) = 471375 - 264854. Thus, according to this regression, the Ferrari F40 is overpriced by \$206,521. It is not a good buy unless the buyer places a value on some factor (e.g., prestige) not included in this regression.