

Chapter 13

13.2 As shown in the confidence interval formula

$$\hat{y} \pm t_{n/2} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

when there is a lot of variability in the independent variable x , the value of the denominator tends to be large and thus the margin of error tends to be small.

13.4 $\hat{y} = 17.88$ at $x = 5$.

13.6 The farther the x value is from the mean of x the wider will be the confidence interval for the expected value of y at that x .

13.8 a. The standard error of regression(estimation) is 11.7813. This is the estimate of the standard deviation of population error term g and is calculated to be

$$s_e = \sqrt{\frac{ESS}{n-2}} = \sqrt{\frac{1110.3953}{8}} = 11.7813$$

b. The t statistic makes use of the null hypothesis $H_0: \beta = 0$ and is calculated as follows:

$$t = \frac{b - \beta_0}{s_b} = \frac{0.7747 - 0}{0.4685} = 1.654$$

$$\text{where } s_b = \sqrt{\frac{s_e^2}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{11.7813^2}{(8.3832)^2(9)}} = 0.4685$$

c. $\text{prob} = 0.13678$ is the p -value for a two-tail test ($H_0: \beta = 0$ vs. $H_A: \beta \neq 0$) and is calculated as follows:

$p\text{-value} = 2P(b \geq 0.7747) = 2P(t_{(df=8)} \geq 1.654) = 2(0.068) = 0.13678$. Because we want one-tail test ($H_A: \beta > 0$) the appropriate p -value is half that reported in the regression computer printout; it is 0.0689 or using EXCEL with "=TDIST(1.654,8,1)" it is 0.06836.

d. The null hypothesis $H_0: \beta = 0$ cannot be rejected in favor of the alternative $H_A: \beta > 0$ because the 0.0684 p -value is greater than the 0.01 Type I error level.

13.10 a. Using the ASCII file EX13-10.PRN and a computer program the regression output is:

Dependent Variable	JOBLESS	Number of Observations	12
Mean of Dep. Variable	9.8667	Std. Dev. of Dep. Var.	6.172569
Std. Error of Regr.	4.4812	Sum of Squared Residuals	200.814
R - squared	.52085	Adjusted R - squared	.47294
F(1, 10)	10.8704	Prob. Value for F	.00805

Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X
Intercept	4.38196	2.107	2.079	.06426		
OUTPUT	.300121	.9103E-01	3.297	.00805	18.27500	14.84319

The least-squares regression equation is

$$\widehat{\text{JOBLESS-RATE}} = 4.38196 + 0.300121 (\text{OUTPUT})$$

- b. The two-tail p-value is 0.00805 for the slope coefficient. We can reject the null hypothesis of zero slope value at a Type I error level as low as 0.00805; we can conclude that there is a significant relationship between the jobless rate and the drop in industrial output.

13.12 Using the ASCII file EX10-19.PRN and a computer program the following regression printout is obtained:

Dependent Variable	Y1988	Number of Observations	25
Mean of Dep. Variable	1245480.0000	Std. Dev. of Dep. Var.	1466643.024052
Std. Error of Regr.	29791.5852	Sum of Squared Residuals	.204134E+11
R - squared	.99960	Adjusted R - squared	.99959
F(1, 23)	58143.4901	Prob. Value for F	.00000

Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X
Intercept	-8488.10	7909.	-1.073	.29428		
Y1986	1.01264	.4200E-02	241.130	.000001238320.00000	1448054.60647	

The least-squares regression equation is

$$\widehat{\text{YEAR}1988} = -8488.10 + 1.01264 (\text{YEAR}1986)$$

To test the hypothesis $H_0: \beta = 0$ vs. $H_A: \beta \neq 0$, the test statistic is calculated to be 241.1 and is much greater than the critical $t_{(0.05, df=23)} = 2.069$; thus, H_0 is rejected at 0.05 Type I error. We conclude that there is a significant change in populations between the two years. This is an alternative testing procedure to that undertaken in Chapter 10. In Chapter 12 no testing

procedures were described because we were only interested in describing the sample relationship and not making inferences about the population.

13.14 As n is increased, s_b is decreased and thus t is increased,

$$\text{as shown in the formulas } s_b = \sqrt{\frac{s_e^2}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{ESS/(n-2)}{\sum(x_i - \bar{x})^2}}$$

and $t = \frac{b - \beta_0}{s_b}$. This implies that increasing the sample

size results in an increase in the power of the test, because rejection of the null hypothesis becomes easier with a large t value. That is, increasing n tends to lead to the rejection of the null hypothesis.

13.16 Although the coefficient of determination ($r^2 = 0.9802$) is very high and the slope coefficient has a relatively large t ratio (21.1), the residual plot shows a pattern and suggests that years are not linearly related to the returns. Ambitious students may try a semilog linear model: $\text{LnReturn} = \beta_0 + \beta_1(\text{Yr}) + \epsilon$, where LnReturn is the natural logarithm of the return. Using the ASCII file EX13-16.PRN and the LIMDEP computer regression program the output is

Dependent Variable	LnReturn	Number of Observations	11			
Mean of Dep. Variable	5.0881	Std. Dev. of Dep. Var.	.325492			
Std. Error of Regr.	.0137	Sum of Squared Residuals	.168814E-02			
R - squared	.99841	Adjusted R - squared	.99823			
F(1, 9)	5639.2614	Prob. Value for F	.00000			
=====						
Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X

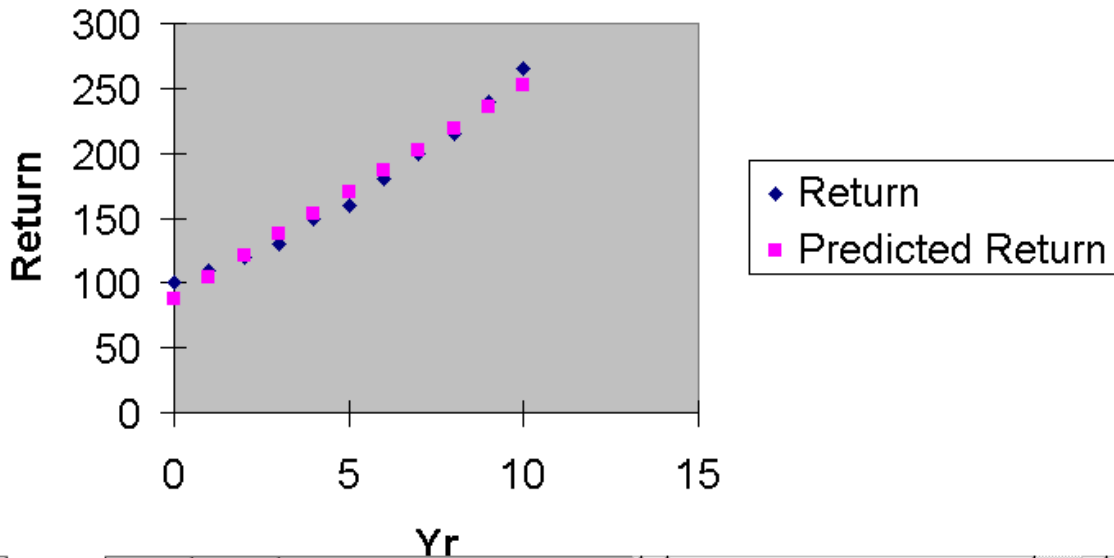
Intercept	4.59776	.7725E-02	595.149	.00000		
Yr	.980614E-01	.1306E-02	75.095	.00000	5.00000	3.31662

P L O T O F R E S I D U A L S

Observ.#	Residual
1	.741E-02
2	.466E-02
3	-.639E-02
4	-.244E-01
5	.206E-01
6	-.129E-01
7	.683E-02
8	.141E-01
9	-.116E-01
10	.323E-03
11	.135E-02

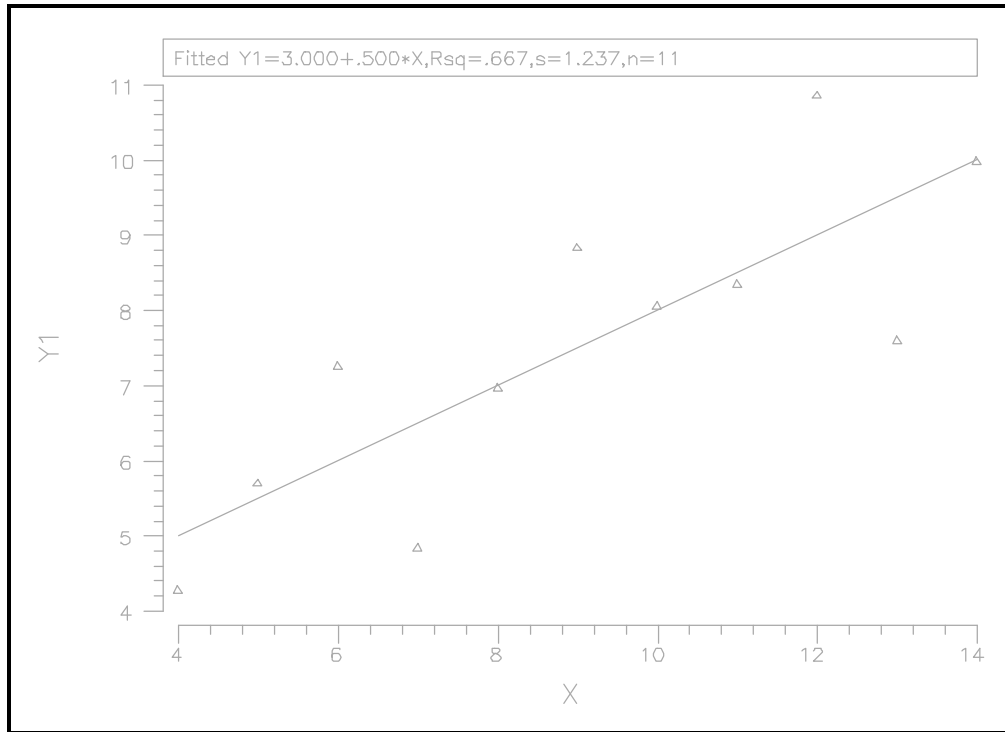
This new regression equation appears to have a better description of the relationship. The coefficient of determination ($r^2 = 0.99841$) is higher and the slope coefficient has a larger t ratio (75.1). Moreover, the residual plot does not show "<" or ">" patterns.

13.18

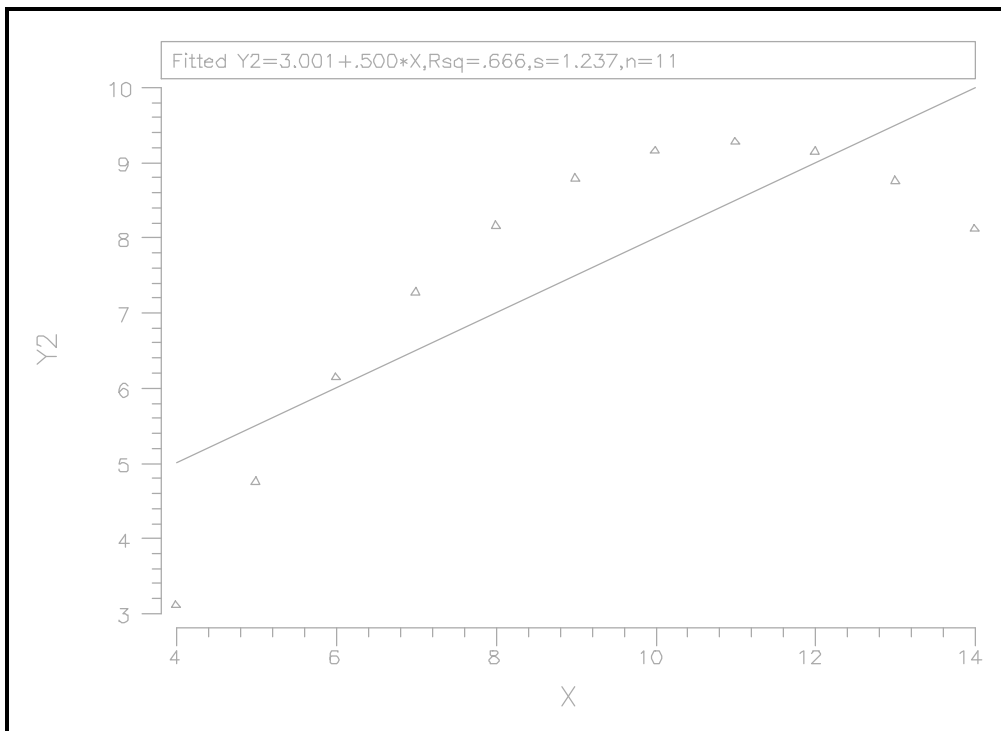


As in Exercise 13.16 the exponential nature of the returns can be seen in this scatterplot where the actual returns are underpredicted at the early and later years and overpredicted in the middle years.

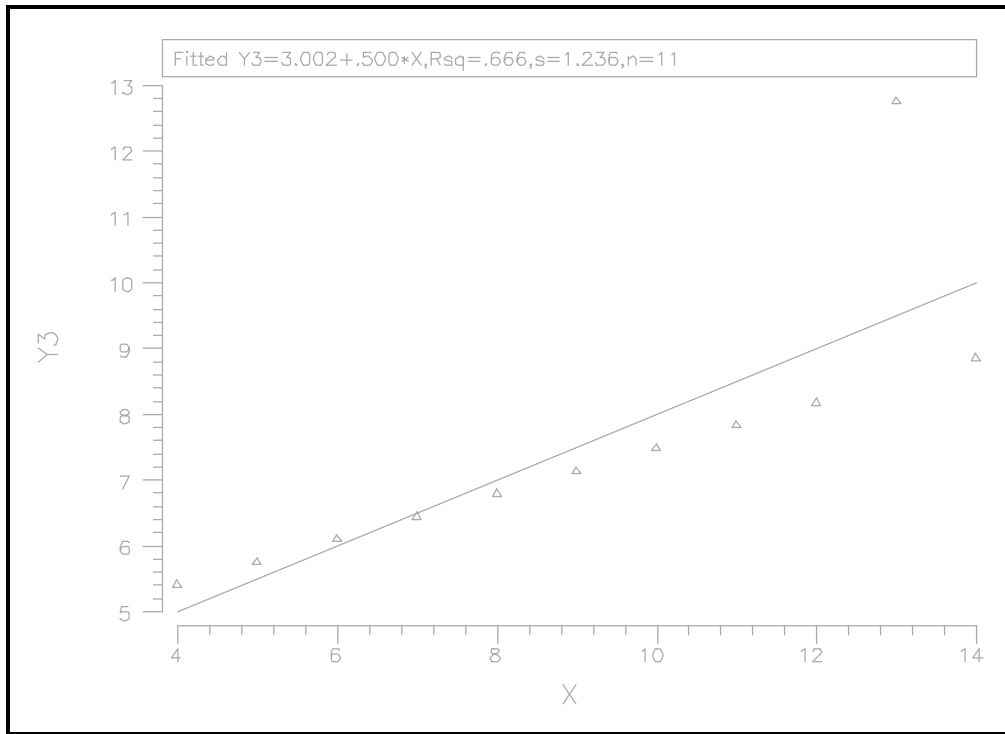
13.20 y1 on x



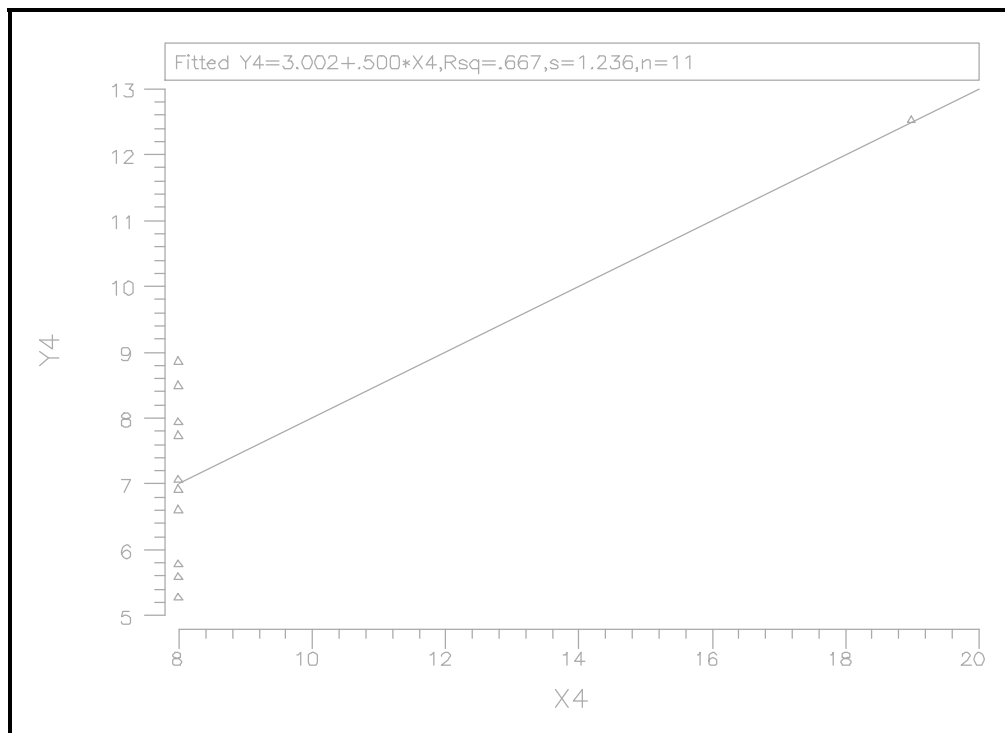
y2 on x



y3 on x



y4 on x4



Only the regression y_1 on x seems to satisfy the required assumptions. The regression y_2 on x violates the assumption of linearity; the second residual plot suggests a polynomial relationship between y_2 and x . The regression y_3 on x violates the assumption of no correlation between the disturbance term and the independent variable x ; the third residual plots suggests a patterned relationship between the disturbance term and x . The regression y_4 on x_4 violates the assumption of homoscedastic disturbance; the variance of disturbance terms seems to be unequal over the range of x values.

- 13.22 a. True.
 b. True.
 c. True.
 d. False. This confidence interval of $10 < E(y|x=20) < 30$ does not say that the mean of y is between 10 and 30 with a probability of 0.95. We can say only that under repeated sampling, 95 percent of similarly constructed intervals would contain the true mean of y and the interval 20 ± 10 may (or may not) be one of those correct intervals.
 e. False. A 95 percent confidence interval for $E(y|x)$ could not be wider than a 95 percent prediction interval for an individual value of y .
 f. False. The decision to reject the null hypothesis depends on the probability of a Type I error.
 g. False. There is no reason to believe that an estimated b equals the parameter β . We, however, can say that the mean of b equals β , i.e., $E(b) = \beta$.
 h. True.
 i. True.
 j. True.
 k. False. The purpose of regression analysis is not necessarily limited to the prediction of the dependent variable; it may also be used for testing, for example.

13.24 In Exercise 12.17, $b = -0.86059$, $ESS = 26.7799$ and

$$\sum(x_i - \bar{x})^2 = 314.9 \text{ were calculated.}$$

Thus, the standard error of regression is

$$s_e = \sqrt{\frac{ESS}{n-2}} = \sqrt{\frac{26.7799}{8}} = 1.8296$$

and the standard error of b is

$$s_b = \sqrt{\frac{s_e^2}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{1.8296^2}{314.9}} = 0.1031$$

To test the hypothesis $H_0: \beta = 0$ vs. $H_A: \beta \neq 0$, the test statistic is calculated as follows:

$$t = \frac{b - \beta_0}{s_b} = \frac{-0.86059 - 0}{0.1031} = -8.347$$

This t value is beyond the critical $-t_{(0.025, df=8)} = -2.306$; H_0 is rejected at $\alpha = 0.05$. We can conclude that there is a significant relationship between two variables.

13.26 The t of 21.0995 was calculated by $(16.3636 - 0)/0.7755$. The P-value may be determined in EXCEL by the command `"=TDIST(21.0995,9,2)"`.

13.28 Using the given data in Exercise 12.10 and a computer program the regression output is:

Dependent Variable	Y	Number of Observations	5			
Mean of Dep. Variable	2.0000	Std. Dev. of Dep. Var.	2.738613			
Std. Error of Regr.	.2226	Sum of Squared Residuals	.148649			
R - squared	.99505	Adjusted R - squared	.99339			
F(1, 3)	602.4545	Prob. Value for F	.00015			
=====						
Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X

Intercept	2.00000	.9955E-01	20.091	.00027		
X	-.635135	.2588E-01	-24.545	.00015	.00000	4.30116

The standard error of regression can be calculated as

$$s_e = \sqrt{\frac{ESS}{n-2}} = \sqrt{\frac{0.148649}{3}} = 0.2226$$

and the standard error of b can be calculated as

$$s_b = \sqrt{\frac{s_e^2}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{0.2226^2}{(4.30116)^2(4)}} = 0.02588$$

These quantities are used in the calculation of test statistics. For example, 0.02588 is the denominator of the calculated t statistic (-24.545) in the above printout for the slope coefficient.

13.30 Using the ASCII file EX12-14.PRN and the EXCEL computer program the following regression output is obtained:

Primary	General	SUMMARY OUTPUT	
1346	1509		
979	1125	Regression Statistics	
1378	1278	Multiple R	0.569094107
1125	1978	R Square	0.323868102
874	378	Adjusted R Square	0.154835128
1002	1200	Standard Error	482.3080327
		Observations	6

ANOVA	df	SS	MS	F
Regression	1	445703.1798	445703.1798	1.916005461
Residual	4	930484.1536	232621.0384	
Total	5	1376187.333		

	Coefficients	Standard Error	t Stat	P-value
Intercept	-375.6344872	1187.014817	-0.316453074	0.76748506
Primary	1.450150197	1.047646134	1.38419849	0.23851455

Thus, at x (number of votes in primary election) = 1200, a 95 percent confidence interval for the expected value of y (number of votes in general election) is

$$\hat{y} \pm t_{(0.025, df=4)} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$$= 1365 \pm 2.776(482.308) \sqrt{\frac{1}{6} + \frac{(1200 - 1117.3333)^2}{211943}}$$

$$= 1365 \pm 597$$

where $\hat{y} = a + bx^* = -375.634 + 1.45015(1200) = 1365$ and

$$\sum(x_i - \bar{x})^2 = (n-1)s_x^2 = 5(205.88508)^2 = 211943$$

Where the standard deviation of x is obtain from EXCEL with the command "=STDEV()"

The 95 percent confidence interval means that, for a candidate who receives 1,200 votes in the primary, if we repeat the process of estimation of regression equation and the construction of a confidence interval for the mean

number of votes that the candidate will receive in the general election, 95 percent of similarly constructed intervals would contain the true mean number of votes in the general election. The interval 1365 ± 597 above may (or may not) be one of those correct intervals.

13.32 a. Using the ASCII file EX12-34.PRN and a computer program the following regression printout is obtained:

Dependent Variable	SPEED	Number of Observations	7			
Mean of Dep. Variable	37.1429	Std. Dev. of Dep. Var.	1.925147			
Std. Error of Regr.	.5229	Sum of Squared Residuals	1.36725			
R - squared	.93852	Adjusted R - squared	.92622			
F(1, 5)	76.3208	Prob. Value for F	.00033			
=====						
Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X

Intercept	33.7130	.4395	76.700	.00000		
EXPRNCE	.727536	.8328E-01	8.736	.00033	4.71429	2.56348
=====						

The slope coefficient estimate 0.727536 is significant at $\alpha = 0.1$. We can conclude that there is a significant relationship between horse speed and jockey experience.

b. At $x(\text{jockey experience}) = 3$, a 90 percent confidence interval for the expected value of $y(\text{horse-speed})$ is

$$\begin{aligned} \hat{y} \pm t_{(0.05, df=5)} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \\ = 35.9 \pm 2.015(0.5229) \sqrt{\frac{1}{7} + \frac{(3 - 4.71429)^2}{39.43}} \\ = 35.9 \pm 0.49 \end{aligned}$$

where $\hat{y} = a + bx^* = 33.7130 + 0.727536(3) = 35.9$ and $\sum(x_i - \bar{x})^2 = (n-1)s_x^2 = 6(2.56348)^2 = 39.43$

c. At $x = 3$, a 90 percent prediction interval for an individual value of y is

$$\hat{y} \pm t_{(0.05, df=5)} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2} + 1}$$

$$= 35.9 \pm 2.015(0.5229) \sqrt{\frac{1}{7} + \frac{(3 - 4.71429)^2}{39.43} + 1}$$

$$= 35.9 \pm 1.16$$

13.34 The P-value is 0.18514; thus, the null hypothesis could be rejected only if the Type I error level was set higher than 0.18514. At typical Type I error levels portfolio returns are not significantly related to the S&P market returns.

13.36 a. The slope coefficient implies that a one percentage point increase in industrial production is associated with a 0.5317 percentage point increase in employment.

b. We need to test the hypothesis $H_0: \beta = 0$ vs. $H_A: \beta \neq 0$ where β is the population slope parameter. At a typical " level 0.05, we can reject H_0 because the calculated t 4.895 is greater than the critical $t_{(0.025, df=5)} = 2.571$, or because the p-value 0.00449 is less than " = 0.05. We can conclude that there is a significant relationship between a change in industrial production and change in employment in the population.

13.38 Total sum of squares(TSS) = RegSS + ErrorSS
 $= 4.6286 + 0.9657$
 $= 5.5943$

Error sum of squares or Sum of squared residuals(ErrorSS)
 $= 0.9657$

Coefficient of determination(r^2) = 0.8274, which indicates that 82.74 percent of the variability in employment around its mean is explained by the regression.

13.40 Using the ASCII file EX12-42.PRN and a computer program the following regression printout is obtained:

Dependent Variable	Y	Number of Observations	7
Mean of Dep. Variable	471025.8571	Std. Dev. of Dep. Var.	113755.802496
Std. Error of Regr.	79442.5144	Sum of Squared Residuals	.315556E+11
R - squared	.59358	Adjusted R - squared	.51229
F(1, 5)	7.3025	Prob. Value for F	.04267
=====			
Variable	Coefficient	Std. Error	t-ratio Prob t >x Mean of X Std.Dev.of X

Intercept	404495.	.3883E+05	10.417 .00014
X	4902.26	1814.	2.702 .04267 13.57143 17.87789

The slope coefficient estimate 4902.26 is significant at $\alpha = 0.05$; we can conclude that there is a significant relationship between CEO compensation(y) and the number of headquartered companies(x).

13.42 Using the ASCII file EX13-42.PRN (the same as EX11-33.PRN or EMERG1.PRN or LOTUS spreadsheet file EX11-33.WK1) the computer regression printout is given below. [The wait time in the emergency room, for those patients whose stay extended from one day to the next, 1440 minutes(24 hours) was added to TIMEOUT (time of leaving)].

Dependent Variable	WAITTIME	Number of Observations	196			
Mean of Dep. Variable	95.3316	Std. Dev. of Dep. Var.	130.295669			
Std. Error of Regr.	127.7801	Sum of Squared Residuals	.316759E+07			
R - squared	.04317	Adjusted R - squared	.03824			
F(1, 194)	8.7533	Prob. Value for F	.00348			
=====						
Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X

Intercept	61.7337	14.57	4.237	.00003		
AGE	1.06059	.3585	2.959	.00348	31.67857	25.52613

The slope coefficient estimate 1.06059 is significant at a typical 0.05 Type I error level; we can conclude that there is a significant relationship between time in the emergency room and patient age. Staff should warn older patients to expect a long visit to E.R.

13.44 Using the ASCII file EX13-43.PRN and a computer program the following regression results may be obtained:

Dependent Variable	MISERY	Number of Observations	20			
Mean of Dep. Variable	13.2100	Std. Dev. of Dep. Var.	6.080850			
Std. Error of Regr.	3.1867	Sum of Squared Residuals	182.793			
R - squared	.73982	Adjusted R - squared	.72536			
F(1, 18)	51.1822	Prob. Value for F	.00000			
=====						
Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X

Intercept	-.613529	2.059	-.298	.76918		
POLITICS	5.52941	.7729	7.154	.00000	2.50000	.94591

The slope coefficient estimate 5.52941 is significant at the typical Type I error levels, which means that there is a significant relationship between the misery index and political stability.

13.46 The regression printout in Exercise 12.47 reports that the t value for the slope coefficient is 85.762, which is significant at the 0.01 Type I error level; we can conclude that there is a statistically significant relationship between two variables.

13.48 Answer will depend on data selected.

13.50 a. Using the regression result in Exercise 12.52, the MACHDIF coefficient estimate is significant at an " level as low as 0.00042; we can conclude that there is a significant relationship between absentee votes and machine votes.

b. A 95 percent prediction interval for an individual value of ABSDIF at MACHDIF = -564 is

$$\begin{aligned} \hat{y} \pm t_{(0.025, df=19)} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2} + 1} \\ = -133.07 \pm 2.093(324.838) \sqrt{\frac{1}{21} + \frac{(-564 - 30066.7)^2}{11880139800} + 1} \\ = -133.07 \pm 721.64 \end{aligned}$$

$$\begin{aligned} \text{where } \sum(x_i - \bar{x})^2 &= (n-1)s_x^2 = 20(24372.25862)^2 \\ &= 11880139800 \end{aligned}$$

c. The observed absentee votes 1025 is not inside this 95 percent of prediction interval, which may suggest there was fraud in the election.

13.52 First, there is heteroscedasticity visible in the scatter plot. In particular at lower inflation rates there is more variability in the S&P 500 P/E Ratio. Second, the standard error of prediction does not remain fixed as predictions are made away from the mean of x. The standard error of prediction increases as we go to more extreme inflation rates.