

Chapter 12

12.2 If $r = \pm 1$, then $r^2 = 1$, and the covariance is the product of the two variances. If $r = 0$, then the covariance is zero.

- 12.4 a. We can complete the matching for all 60 trainees as shown in the next page, and in ASCII file EX12-4.prn.
 b. Using the ASCII data file EX12-4.PRN, the mean of pretest score is calculated to be 56.
 c. Change in score can be calculated for each of 60 students as shown in the next page.
 d. The average change score for those who scored below the average(56) on the pretest is 4.41.
 e. The average change score for those who scored above the average(56) on the pretest is 4.04.

PRETEST SCORE	FRE- QUENCY	PERCENT	...CUMULATIVE...	
			FREQUENCY	PERCENT
25	+))))))))) < 2	3.33	2	3.33
30	*+))))))))) < 3	5.00	5	8.33
35	**+))))))))) < 4	6.67	9	15.00
40	***+))))))))) < 5	8.33	14	23.33
45	****+))))))))) < 6	10.00	20	33.33
50	*****+))))))))) < 6	10.00	26	43.33
55	*****+))))))))) < 8	13.33	34	56.67
60	*****+))))))))) < 7	11.67	41	68.33
65	*****+))))))))) < 4	6.67	45	75.00
70	*****+))))))))) < 4	6.67	49	81.67
75	*****+))))))))) < 3	5.00	52	86.67
80	*****+))))))))) < 3	5.00	55	91.67
85	*****+))))))))) < 2	3.33	57	95.00
90	*****+))))))))) < 2	3.33	59	98.33
95	*****+))))))))) < 1	1.67	60	100.00

MIDTEST SCORE	FRE- QUENCY	PERCENT	...CUMULATIVE...	
			FREQUENCY	PERCENT
25	/3333333333333333) > 2	3.33	2	3.33
30	. 3333333333333333) > 3	5.00	5	8.33
35	. 3333333333333333) > 3	5.00	8	13.33
40	/3333333333333333) > 4	6.67	12	20.00
45	. 3333333333333333) > 5	8.33	17	28.33
50	/3333333333333333) > 5	8.33	22	36.67
55	. 3333333333333333) > 5	8.33	27	45.00
60	/3333333333333333) > 5	8.33	32	53.33
65	. 3333333333333333) > 5	8.33	37	61.67
70	. 3333333333333333) > 7	11.67	44	73.33
75	. 3333333333333333) > 5	8.33	49	81.67

80	. 3323333)	> 2	3.33	51	85.00
85	. 3) 3333)	> 4	6.67	55	91.67
90	.) 2333)	> 2	3.33	57	95.00
95	. 23)	> 2	3.33	59	98.33
100	.) >	1	1.67	60	100.00

Student pretest midtest change

1	25	25	0	S),
2	25	30	5	*
3	30	25	-5	*
4	30	30	0	*
5	30	35	5	*
6	35	30	-5	*
7	35	35	0	*
8	35	40	5	*
9	35	45	10	*
10	40	35	-5	*
11	40	40	0	*
12	40	45	5	*
13	40	50	10	*
14	40	55	15	*
15	45	40	-5	*
16	45	45	0	*
17	45	50	5	/) Q
18	45	55	10	*
19	45	60	15	*
20	45	65	20	*
21	50	45	-5	*
22	50	50	0	*
23	50	55	5	*
24	50	60	10	*
25	50	65	15	*
26	50	70	20	*
27	55	40	-15	*
28	55	45	-10	*
29	55	50	-5	*
30	55	55	0	*
31	55	60	5	*
32	55	65	10	*
33	55	70	15	*
34	55	75	20	S) -
35	60	50	-10	S),
36	60	55	-5	*

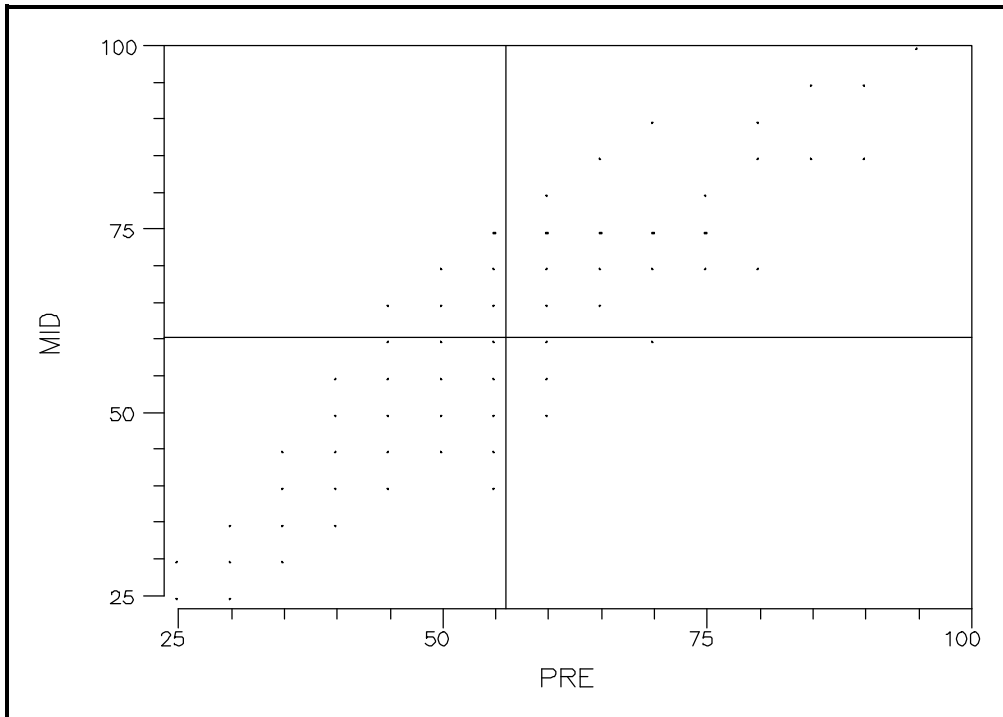
Mean change is
4.41 for these
students.

37	60	60	0	*
38	60	65	5	*
39	60	70	10	*
40	60	75	15	*
41	60	80	20	*
42	65	65	0	*
43	65	70	5	*
44	65	75	10	*
45	65	85	20	*
46	70	60	-10	*
47	70	70	0	*) Q
48	70	75	5	*
49	70	90	20	*
50	75	70	-5	*
51	75	75	0	*
52	75	80	5	*
53	80	70	-10	*
54	80	85	5	*
55	80	90	10	*
56	85	85	0	*
57	85	95	10	*
58	90	85	-5	*
59	90	95	5	*
60	95	100	5	S)-

Sum	3360	3615		
Average	56	60.25		

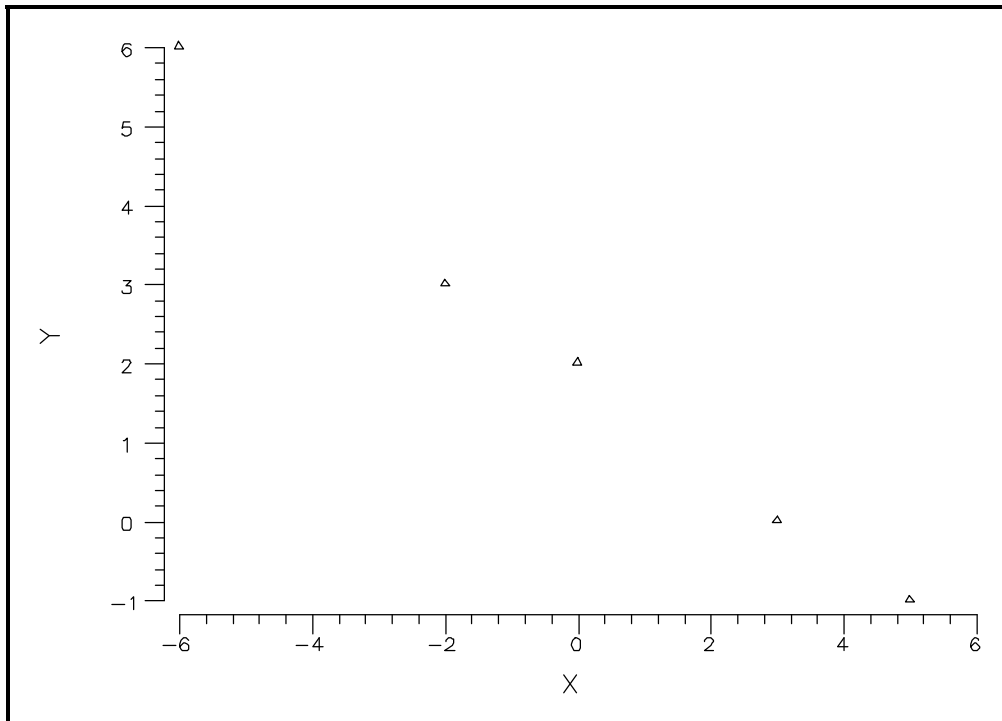
Mean change is 4.04 for these students.

12.6 As seen below, for positive (negative) deviations from the mean of the pretest score, there tends to be positive (negative) deviations from the mean of the midtest score.



12.8 $r^2 = 0.791$; thus, there is a strong linear relationship between earnings and sales.

12.10 $r = -0.99752$



- 12.12 $r = 0.73321$ for those selected for executive positions
 $r = 0.75903$ for those not selected for executive positions
 $r = 0.88809$ for the total group

Two correlation coefficients for the subsamples are less than the correlation for the entire sample. As shown in the scatterplot in Exercise 12.6, around the mean of pretest score are there relatively many observations whose midtest scores are wide ranged. If we divide the entire data set into two subsamples, those wide ranged observations play crucial roles in determining the correlation coefficients for both subsamples and thus result in smaller correlation coefficients.

- 12.14 Using the ASCII data file EX12-14.PRN, the correlation coefficient $r = 0.569$ shows a positive, but not a strong linear relationship between the number of votes obtained by candidates in the primary and the general election.

- 12.16 a. True.
 b. False. The slope of the sample regression line indicates how the predicted value of y changes as x changes.
 c. True.
 d. False. A residual can be positive or zero or negative. The sum of all residuals is zero.
 e. False. The least square method minimizes the sum of squared residuals.
 f. True.
 g. True.
 h. True.
 i. False. The residuals always sum to zero whether the y intercept is zero or not.

- 12.18 a. $b = -1.06098$, $a = 0$ and the regression is
 $\hat{y} = -1.06098x$
 b. Both the mean of x and mean of y are zero and thus the intercept is zero.
 c. The slope coefficient -1.06098 means that one unit increase in x is associated with a 1.06098 unit decrease in y .

- 12.20 a. The slope coefficient is

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{655}{1305} = 0.5019$$

and the intercept coefficient is

$$a = \bar{y} - b\bar{x} = \frac{\sum y_i}{n} - b \frac{\sum x_i}{n} = \frac{405}{8} - 0.5019 \frac{475}{8} = 20.8247$$

As a result, we have the regression equation

$$\hat{y} = 20.8247 + 0.5019x$$

b. $\hat{y} = 20.8247 + 0.5019(40) = 40.9007$

12.22 The slope coefficient is

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{1710}{3640} = 0.4698$$

and the intercept coefficient is

$$a = \bar{y} - b\bar{x} = \frac{\sum y_i}{n} - b \frac{\sum x_i}{n} = \frac{720}{20} - 0.4698 \frac{1350}{20} = 4.2885$$

As a result, we have the regression equation

$$\hat{y} = 4.2885 + 0.4698x$$

12.24 Using the ASCII data file EX12-14.PRN, the regression equation can be calculated to be $\hat{\text{GENERAL}} = -375.634 + 1.45015\text{PRIMARY}$. This implies that each additional vote in the primary is associated with 1.45015 votes in the general election.

12.26 Considering $a \approx 0$ and $b \approx 27$, Fisher's regression can be written to be $\hat{y} = 0 + 27x$ where y is the TV station advertising revenue and x is the audience size.

Using the relationship $a = \bar{y} - b\bar{x}$, the slope coefficient b can be calculated as

$$b = \frac{\bar{y}}{\bar{x}} - \frac{a}{\bar{x}} = \frac{\bar{y}}{\bar{x}} = \frac{\sum y/n}{\sum x/n} = \frac{\sum y}{\sum x} \text{ since } a \approx 0.$$

$\sum y$ is nothing but the sum of all television revenues and

Σx is the total audience size. The total audience size equals the number of television households multiplied by the fraction (60 percent) of prime time that the average household watches television. Thus, the estimation of b should result in the same coefficient value 27.

- 12.28 a. Assuming there is a positive relationship between x and y , the correlation coefficient is

$$r = +\sqrt{r^2} = +\sqrt{0.6968} = 0.83475$$

Thus, the slope coefficient b is

$$b = r \frac{s_y}{s_x} = 0.83475 \frac{6.5056}{5.3135} = 1.0220$$

And the intercept coefficient a is

$$a = \bar{y} - b\bar{x} = 12.9 - 1.0220(9.7) = 2.9866$$

- b. The total sum of squares is

$$\begin{aligned} TSS &= \sum (y_i - \bar{y})^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} (n-1) = s_y^2 (n-1) \\ &= (6.5056)^2 (10-1) = 380.905 \end{aligned}$$

and using $r^2 = \frac{RSS}{TSS} = \frac{TSS - ESS}{TSS} = 0.6968$

the residual sum of squares is

$$\begin{aligned} ESS &= TSS - TSS(r^2) \\ &= 380.905 - (380.905)(0.6968) = 115.490 \end{aligned}$$

- 12.30 a. The estimated regression equation is

$$\hat{y} = 15.9374 + 17.5466x$$

where x is the October heating bills (in thousands of dollars) and y (in thousands of dollars) is the winter heating bills. Thus, if the October heating bill increases by \$1,000, the best point estimate of increase in the winter heating bill would be \$17,546.6

- b. If the October heating bill is \$5,000, the predicted heating bill for the winter would be \$103,670.4 (= 15.9374 + 17.5466(5)).

c. $r = +\sqrt{r^2} = +\sqrt{0.31413} = 0.56047$

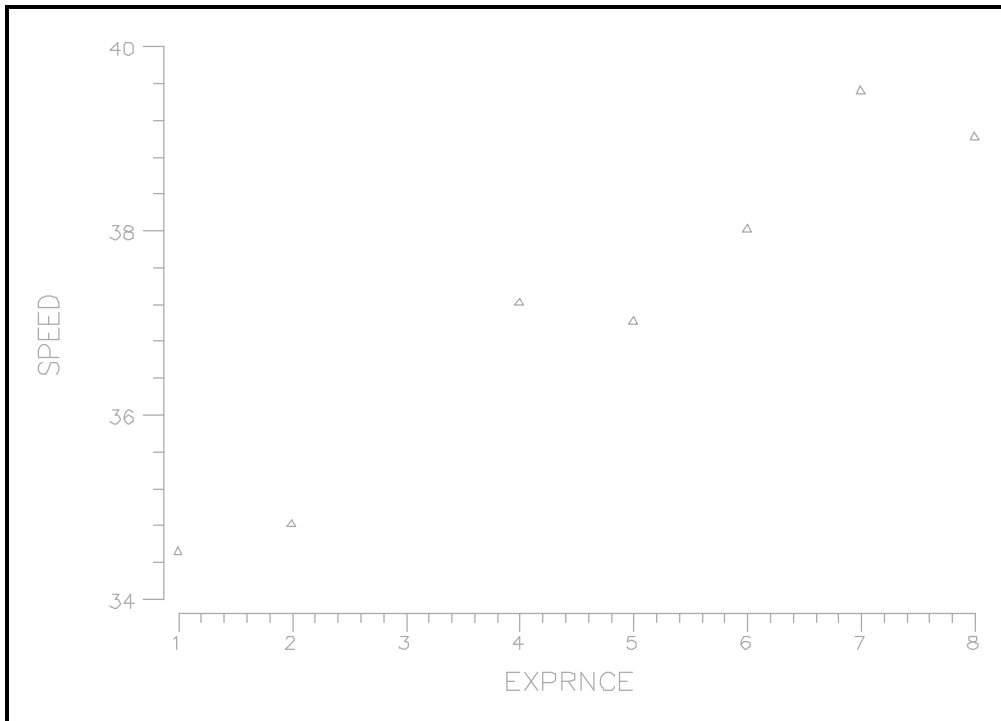
d. 11584.0 is the smallest possible sum of squared residuals, calculated using the least square method.

12.32 $r^2 = 0.997$ in the computer printout in Exercise 12.31. This says that 99.7 percent of the variability in objective probabilities around their mean is explained by the regression.

12.34 a/b. Using the ASCII data file EX12-34.PRN and a computer program, the following regression results and scatterplot can be 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 least-squares regression equation is

$$\widehat{\text{SPEED}} = 33.713 + 0.727536\text{EXPERIENCE} \text{ and } r^2 \text{ is } 0.93852$$

c. If a jockey has three years of experience, then the predicted value of horse speed is 35.90 mph (= 33.713 +

0.727536(3)).

12.36 Using the ASCII data file EX12-36.PRN, we a computer program we obtain:

Dependent Variable	Y1989	Number of Observations	20
Mean of Dep. Variable	31051.3000	Std. Dev. of Dep. Var.	16179.734186
Std. Error of Regr.	3057.9153	Sum of Squared Residuals	.168315E+09
R - squared	.96616	Adjusted R - squared	.96428
F(1, 18)	513.9190	Prob. Value for F	.00000

Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X
INTERCEPT	104.422	1527.	.068	.94623		
Y1985	1.83920	.8113E-01	22.670	.00000	16826.30000	8647.04579

The least-squares regression equation is

$\hat{Y}_{1989} = 104.422 + 1.83920 \text{ YEAR}_{1985}$; thus, the slope coefficient 1.83920 implies that the average rate of growth between 1985 and 1989 is approximately 83.9 percent.

12.38 If sales are \$120,000,000, then predicted earnings are \$21,579,820, and if sales are \$125,000,000 then predicted earnings are \$23,464,660. The increase in earnings is \$1,884,830. (Remember that x and y in this regression were measured in \$1,000 units as stated in Exercise 12.8; thus, the increase in earnings can be seen to be 5000 times the slope 0.376967.)

y:Earnings	x:Sales	SUMMARY OUTPUT	
21455	128070	Regression Statistics	
18655	118580	Multiple R	0.889412331
25882	128263	R Square	0.791054295
36703	161522	Adjusted R Sq.	0.721405726
29120	127008	Standard Error	3715.886227
		Observations	5

ANOVA	df	SS	MS
Regression	1	156826306.6	156826306.6
Residual	3	41423431.4	13807810.5
Total	4	198249738	

A	B	C	D	E
	Coefficients	Standard Error	t Stat	P-value

17	Intercept	-23656.16106	14934.6379	-1.5840	0.2114
18	x:Sales	0.37696653	0.1119	3.3701	0.0434

RESIDUALS:	Observation	Predicted y:Earnings	Residuals
	1	24621.94239	-3166.942386
	2	21044.53002	-2389.53002
	3	24694.69693	1187.303074
	4	37232.22674	-529.2267356
	5	24221.60393	4898.396068

If x:sales are

120000	21579.82	=D17+D18*C30	
125000	23464.66	=D17+D18*C31	
	1884.83	=D31-D30	=5000*(D18)

12.40 Approximately 73 percent of the variability in excess returns around their mean is explained by the regression:

Predicted excess returns = 4.35 + 10.57(book/market ratio).

Predicted excess returns = 10.17 percent if b/m ratio =.55.

Returns	b/m ratio	
7	0.3	SUMMARY OUTPUT
10	0.4	Regression Statistics
8	0.5	Multiple R 0.853719873
11	0.6	R Square 0.728837622
13	0.7	Adjusted R Sq 0.661047028
12	0.8	Standard Error 1.348720734
		Observations 6

ANOVA	df	SS	MS	F
Regression	1	19.55714286	19.55714286	10.7513089
Residual	4	7.276190476	1.819047619	
Total	5	26.83333333		

	Coefficients	Standard Error	t Stat	P-value
Intercept	4.352380952	1.856752096	2.344082962	0.079025
b/m ratio	10.57142857	3.224059215	3.278918862	0.030532

RESIDUAL	Observation	Predicted Returns	Residuals
	1	7.523809524	-0.523809524
	2	8.580952381	1.419047619
	3	9.638095238	-1.638095238
	4	10.6952381	0.304761905
	5	11.75238095	1.247619048
	6	12.80952381	-0.80952381
		10.16666667	=D17+D18*0.55

12.42 Using the ASCII data file EX12-42.PRN, and a computer program, the correlation coefficient (r) is 0.77044.

12.44

Diagram A

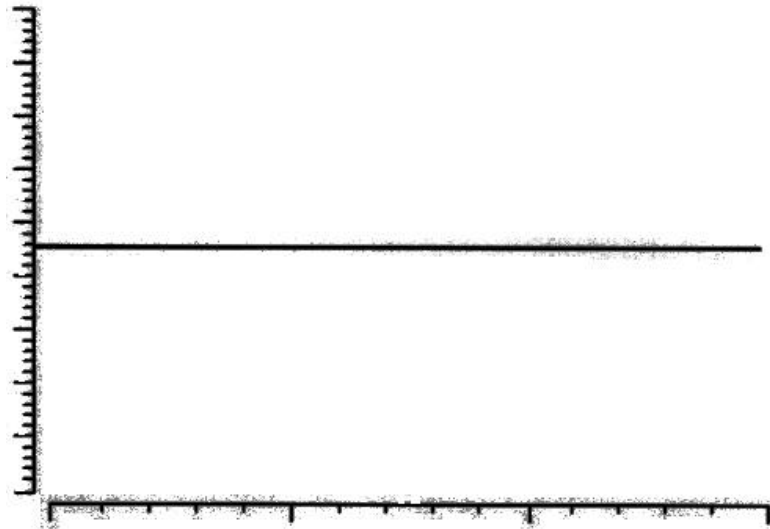


Diagram B

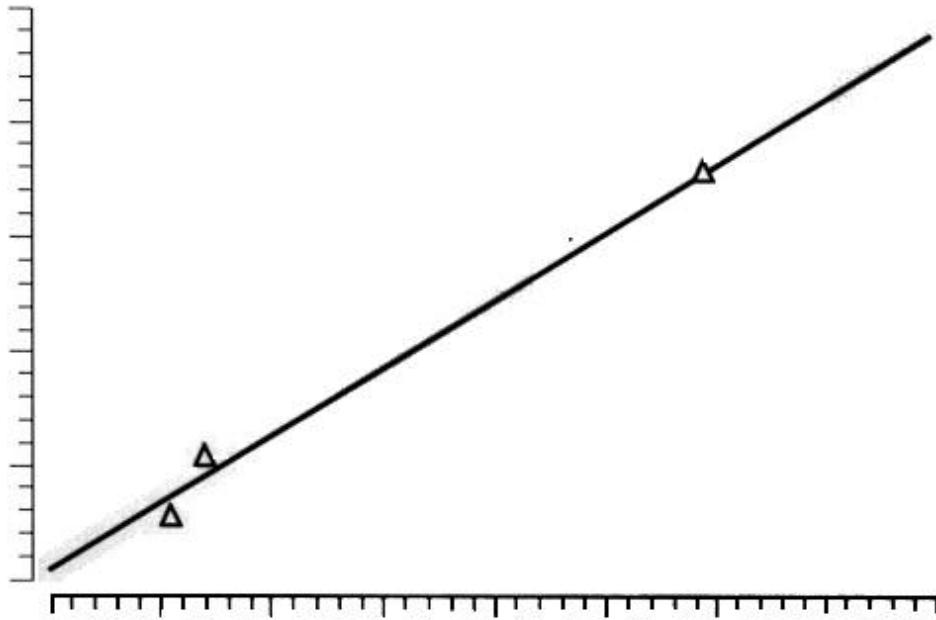


Diagram C

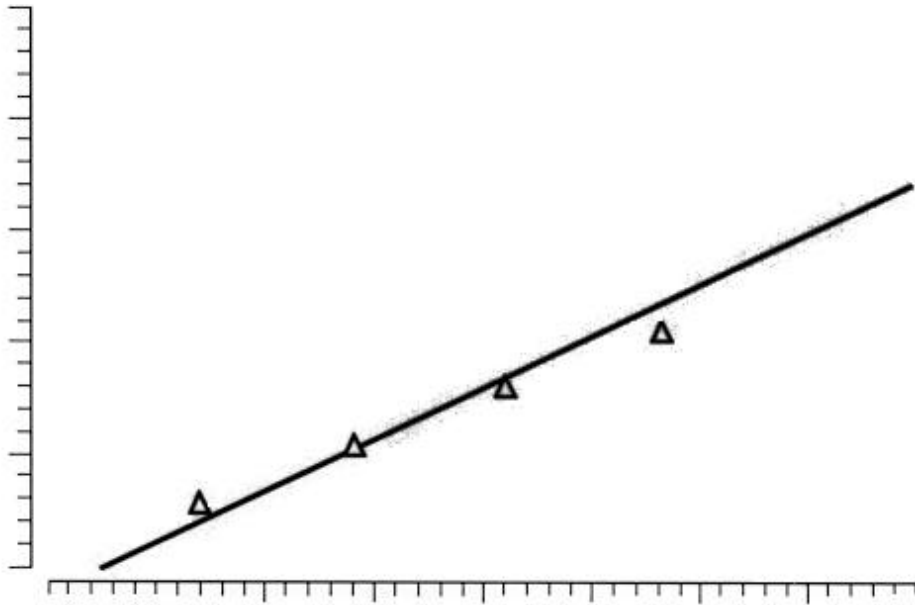
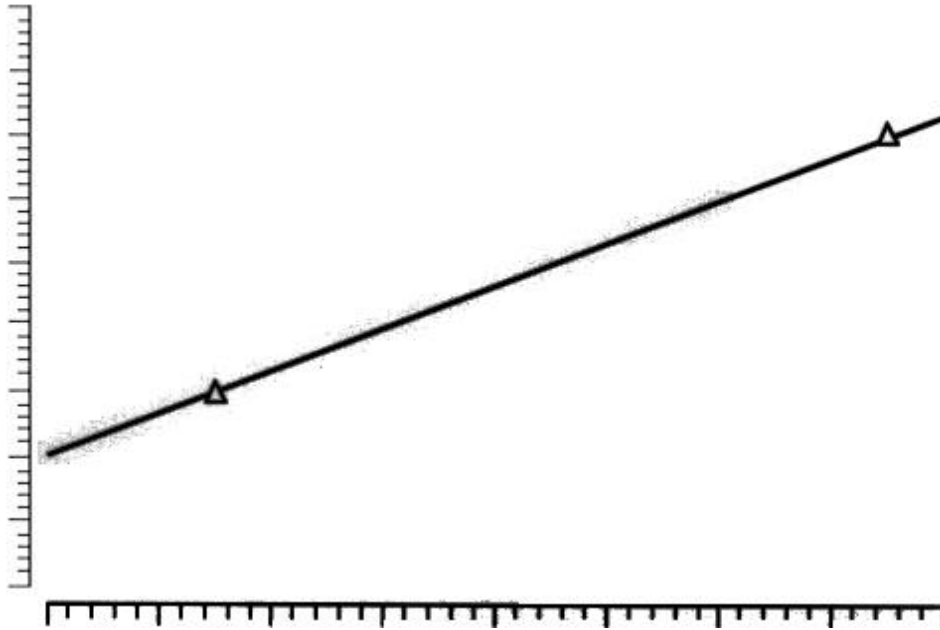


Diagram D



12.46 Using the ASCII data file EX12-46.PRN and a computer program, the following regression output may be obtained:

Dependent Variable	MPG	Number of Observations	12
Mean of Dep. Variable	39.6167	Std. Dev. of Dep. Var.	1.626951
Std. Error of Regr.	.8321	Sum of Squared Residuals	6.92360
R - squared	.76221	Adjusted R - squared	.73843
F(1, 10)	32.0542	Prob. Value for F	.00021

Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X
INTERCEPT	63.1571	4.165	15.164	.00000		
WEIGHT	-.876440E-02	.1548E-02	-5.662	.00021	2685.91667	162.06534

The least-squares regression equation is

$$\hat{MPG} = 63.1571 - 0.0087644WEIGHT;$$

thus, for a car of 2,910 lbs., the predicted MPG is

$$37.65(= 63.1571 - 0.0087644(2910)).$$

The regression equation gives a better prediction than just using the average mileage of the two cars because the regression makes use of relationship between the mileage and weight of all the cars. In fact 76 percent of the variability in mpg around its mean (39.6 mpg) is explained

by the regression.

12.48 The implied least-squares regression equation is

$$\widehat{\text{CANCER-FACTOR}} = 2.5\text{CIGARETTE}$$

where $\widehat{\text{CANCER-FACTOR}}$ is the predicted incidence of lung cancer and CIGARETTE is daily cigarettes consumed.

12.50 Answer will depend on data and years selected.

12.52 a. Using the ASCII data file EX12-52.PRN and a computer program the following regression output is obtained:

Dependent Variable	ABSDIF	Number of Observations	21			
Mean of Dep. Variable	256.0476	Std. Dev. of Dep. Var.	442.835576			
Std. Error of Regr.	324.8380	Sum of Squared Residuals	.200488E+07			
R - squared	.48882	Adjusted R - squared	.46192			
F(1, 19)	18.1690	Prob. Value for F	.00042			
=====						
Variable	Coefficient	Std. Error	t-ratio	Prob t >x	Mean of X	Std.Dev.of X

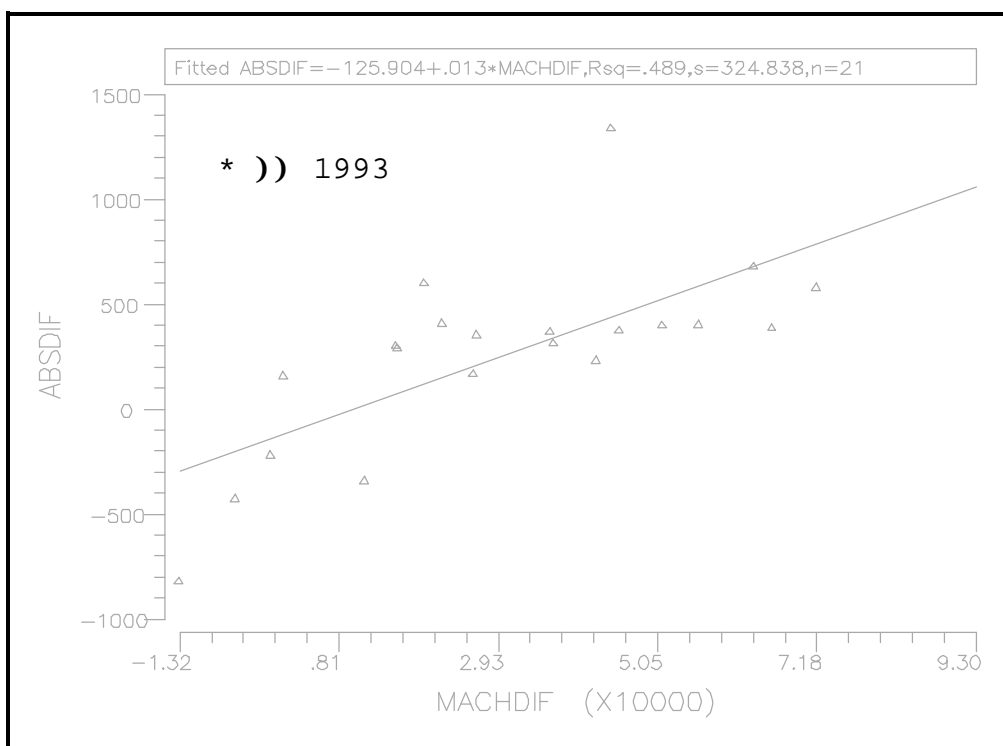
INTERCEPT	-125.904	114.3	-1.102	.28425		
MACHDIF	.127035E-01	.2980E-02	4.263	.00042	30066.71429	24372.25862

The least-squares regression equation is

$$\widehat{\text{ABSDIF}} = -125.904 + 0.0127\text{MACHDIF}$$

where ABSDIF is the difference in absentee votes and MACHDIF is the difference in machine votes. The slope OF 0.0127 indicates that an additional machine vote is associated with 0.0127 additional absentee votes.

b. The 22nd observation must be excluded because the 1993 election is in question.



- c. The correlation coefficient (r) is 0.699; there is a positive and somewhat strong relationship between the difference in absentee votes and the difference in machine votes.
- d. At MACHDIF = -564, the predicted ABSDIF is -133 [= -125.904 + 0.0127(-564)]; the difference between the actual and the predicted ABSDIF in 1993 is 1158 [= 1025 - (-133)].

12.54 Using the formula

predicted lace length = 10.28 + 3.429(number of eyes)

the error sum of squares is 553.429 which is greater than the 242.648 error sum of squares from the least squares regression line reported in the answers to Query 12.7 and Query 12.8.

Lace	Eyes	predicted	error	error^2
45	8	37.712	7.288	53.11494
54	10	44.570	9.430	88.92490
26	4	23.996	2.004	4.01602
63	14	58.286	4.714	22.22180
63	12	51.428	11.572	133.91118
36	8	37.712	-1.712	2.93094
54	12	51.428	2.572	6.61518
24	4	23.996	0.004	0.00002
72	14	58.286	13.714	188.07380
54	12	51.428	2.572	6.61518
72	16	65.144	6.856	47.00474
72	18	72.002	-0.002	0.00000
			59.012	553.42870