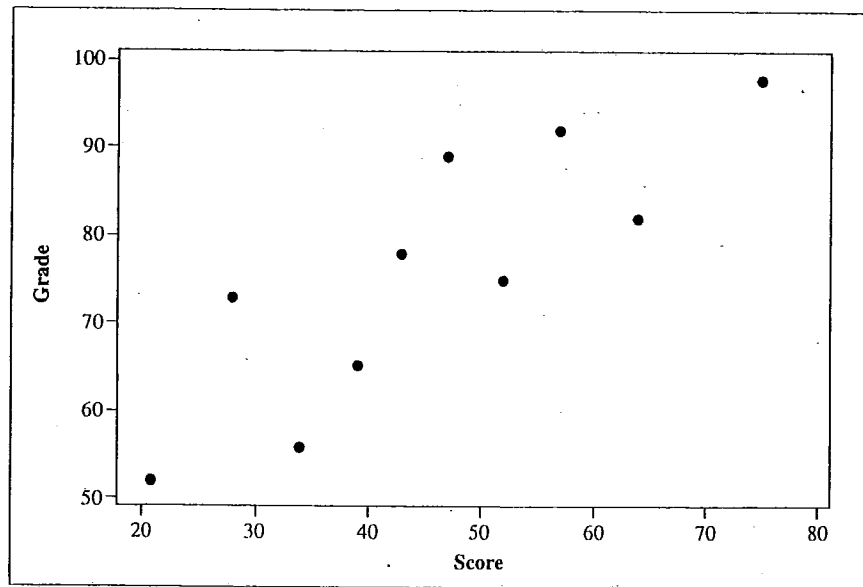


Mathematics Achievement Test Scores and Final Calculus Grades for College Freshmen

Student	Mathematics Achievement Test Score	Final Calculus Grade
1	39	65
2	43	78
3	21	52
4	64	82
5	57	92
6	47	89
7	28	73
8	75	98
9	34	56
10	52	75



Scatter Plot

Calculations for the Data in Table 12.1

	y_i	x_i	x_i^2	$x_i y_i$	y_i^2
	65	39	1521	2535	4225
	78	43	1849	3354	6084
	52	21	441	1092	2704
	82	64	4096	5248	6724
	92	57	3249	5244	8464
	89	47	2209	4183	7921
	73	28	784	2044	5329
	98	75	5625	7350	9604
	56	34	1156	1904	3136
	75	52	2704	3900	5625
Sum	760	460	23,634	36,854	59,816

Then

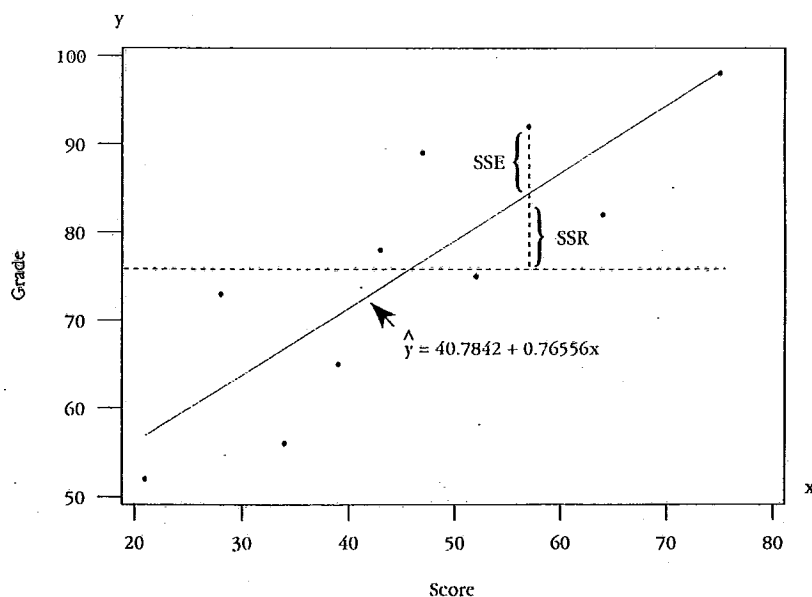
$$b = \frac{S_{xy}}{S_{xx}} = \frac{1894}{2474} = .76556 \quad \text{and} \quad a = \bar{y} - b\bar{x} = 76 - (.76556)(46) = 40.78424$$

The least-squares regression line is then

$$\hat{y} = a + bx = 40.78424 + .76556x$$

The graph of this line is shown in Figure 12.4. It can now be used given value of x —either by referring to Figure 12.4 or by substituting of x into the equation. For example, if a freshman scored $x = 50$ on test, the student's predicted calculus grade is (using full decimal ac

$$\hat{y} = a + b(50) = 40.78424 + (.76556)(50) = 79.06$$



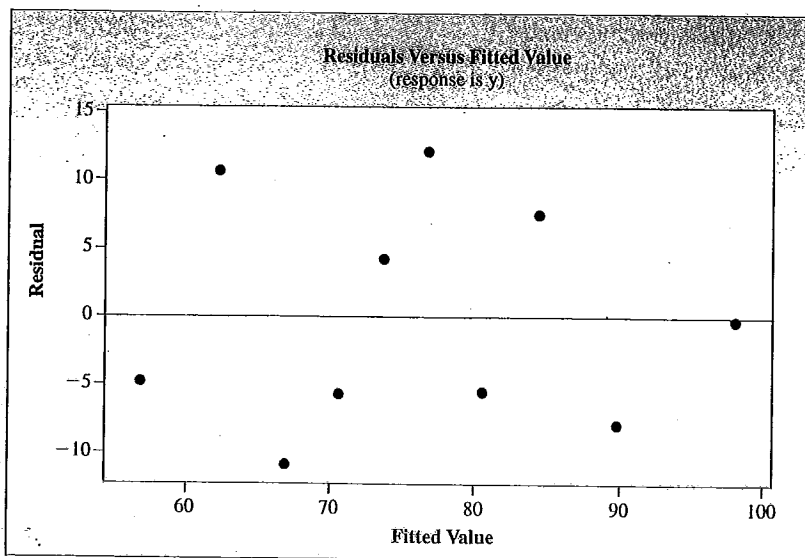
Residuals

y_i	x_i
65	39
78	43
52	21
82	64
92	57
89	47
73	28
98	75
56	34
75	52

$$\hat{y}_i = 40.78424 + 0.76556 x_i$$

70.64108
73.70332
56.86100
89.78008
84.42116
76.76556
62.21992
98.20124
66.81328
80.59336

$e_i = y_i - \hat{y}_i$
-5.64108
4.29668
-4.86100
-7.78008
7.57884
12.23444
10.78008
-0.20124
-10.81328
-5.59336
0



$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i=1, \dots, n.$$

α, β and σ^2 are unknown parameters

$$a = \hat{\alpha} = \bar{y} - b\bar{x}, \quad b = \hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2}$$

$$b = r \cdot \sqrt{\frac{S_{xx}}{S_{yy}}}, \quad \text{where } r = \text{correlation coefficient}$$

$$\text{and } S_{yy} = \sum_1^n (y_i - \bar{y})^2$$

TSS = Total Sum of Squares

$$= \sum_1^n (y_i - \bar{y})^2 \quad (= S_{yy})$$

$$= \sum_1^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2, \quad \text{where } \hat{y}_i = a + bx_i$$

$$= \sum_1^n [(y_i - \bar{y} + b\bar{x} - b\bar{x} + bx_i) + (\bar{y} - b\bar{x} + b\bar{x} - \bar{y})]^2$$

$$= \sum_1^n (y_i - \hat{y}_i)^2 + b^2 \sum_1^n (x_i - \bar{x})^2 + 2 \sum_1^n [(y_i - \bar{y}) - b(x_i - \bar{x})] b(x_i - \bar{x})$$

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{S_{xy}^2}{S_{xx}} + 2 \frac{S_{xy}^2}{S_{xx}} - 2 \frac{S_{xy}^2}{S_{xx}}$$

$$= SSE + SSR$$

ANOVA Table

Source	SS	df	MSS
Regression	SSR	1	SSR
Error	SSE	n-2	$\frac{1}{n-2} SSE \rightarrow$ estimate of σ^2
Total	TSS = S_{yy}	n-1	*

$$F\text{-statistic} = \frac{SSR}{SSE/(n-2)} \sim F_{1, n-2}$$

H_0 : Testing for whether the variable x is significant in predicting y or not

$$H_0: \beta = 0 \quad \text{vs} \quad H_1: \beta \neq 0$$

Under normal assumption,

$$b \sim N\left(\beta, \frac{\sigma^2}{\sum x^2}\right).$$

$$\text{So, } \frac{b - \beta}{\sqrt{\frac{MSSE}{\sum x^2}}} \sim t_{n-2 \text{ d.f.}}$$

Testing $H_0: \beta = 0$ vs $H_1: \beta \neq 0$,

use $\frac{b}{\sqrt{\frac{MSSE}{\sum x^2}}}$ and two-sided CR.

This is the same as F-test in ANOVA table.

But, you can use t to construct CI for β as

$$\left(b - t_{n-2, \frac{\alpha}{2}} \sqrt{\frac{MSSE}{\sum x^2}}, b + t_{n-2, \frac{\alpha}{2}} \sqrt{\frac{MSSE}{\sum x^2}} \right)$$

as a $100(1-\alpha)\%$ CI for β .

Coefficient of Determination. (or strength of linear relationship).

The proportion of total variation in the data ^(in y) that is explained by the linear regression of y on x , viz.,

$$\frac{SSR}{TSS} = \frac{S_{xy}^2 / \sum x^2}{S_{yy}} = \frac{S_{xy}^2}{\sum x^2 S_{yy}} = r^2.$$

This is simply the square of the correlation coeff. R .

If you find $r^2 = 80.1\%$, it means that 80.1% of the variation has been explained by the linear regression,
 \rightarrow in the data on y

Analysis of Variance for Linear Regression

Source	df	SS	MS
Regression	1	$\frac{(S_{xy})^2}{S_{xx}}$	MSR
Error	$n - 2$	$S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$	MSE
Total	$n - 1$	S_{yy}	

For the data in Table 12.1, you can calculate

$$\text{Total SS} = S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 59,816 - \frac{(760)^2}{10} = 2056$$

$$\text{SSR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{(1894)^2}{2474} = 1449.9741$$

so that

$$\text{SSE} = \text{Total SS} - \text{SSR} = 2056 - 1449.9741 = 606.0259$$

and

$$\text{MSE} = \frac{\text{SSE}}{n - 2} = \frac{606.0259}{8} = 75.7532$$

Regression Analysis: y versus x

The regression equation is
 $y = 40.8 + 0.766 x$

Predictor	Coef	SE Coef	T	P
Constant	40.784	8.507	4.79	0.001
x	0.7656	0.1750	4.38	0.002

S = 8.70363 R-Sq = 70.5% R-Sq(adj) = 66.8%

Analysis of Variance	DF	SS	MS	F	P
Regression	1	1450.0	1450.0	19.14	0.002
Residual Error	8	606.0	75.8		
Total	9	2056.0			

Predicting a Particular Value at $x = x_0$

From the fitted regression line $y = a + bx$, suppose we are interested in ~~est~~ predicting a particular value of y at a value $x = x_0$.

It is given by $\hat{y} = a + bx_0$, and its standard error is given by

$$SE(\hat{y}) = \sqrt{MSE \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\}};$$

a 95% CI for the particular value of y at $x = x_0$ is

$$\hat{y} \pm t_{n-2, \alpha/2} \sqrt{MSE \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\}}.$$

Example: Suppose a student took the achievement test and scored 50, but has not taken the calculus test, and we wish to predict her score in calculus test.

Then:

$$\hat{y} = 40.78424 + (0.76556 \times 50) = 79.06$$

and its standard error is

$$SE(\hat{y}) = \sqrt{75.7532 \left\{ 1 + \frac{1}{10} + \frac{(50 - 46)^2}{2474} \right\}} = 9.155.$$

Then, with $t_{8, 0.025} = 2.306$, we obtain a 95% prediction interval for the calculus score to be

$$79.06 \pm (2.306 \times 9.155) = (57.95, 100.17).$$

Note that this prediction interval is much wider than the CI, though the estimates are the same!

Estimation of Average Value at $x = x_0$

From the fitted regression line $y = a + bx$, suppose we are interested in estimating the average value of y at a value $x = x_0$.

It is given by $\hat{y} = a + bx_0$, and its standard error is given by

$$SE(\hat{y}) = \sqrt{MSE \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\}};$$

a 95% CI for the average value of y at $x = x_0$ is

$$\hat{y} \pm t_{n-2, \alpha/2} \sqrt{MSE \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\}}.$$

Example: Suppose we are interested in estimating the average calculus grade for students whose achievement score is 50.

Then:

$$\hat{y} = 40.78424 + (0.76556 \times 50) = 79.06$$

and its standard error is

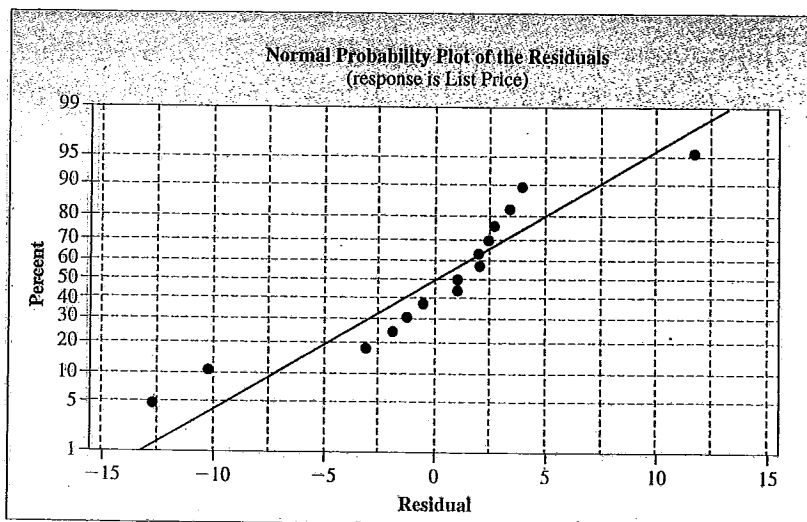
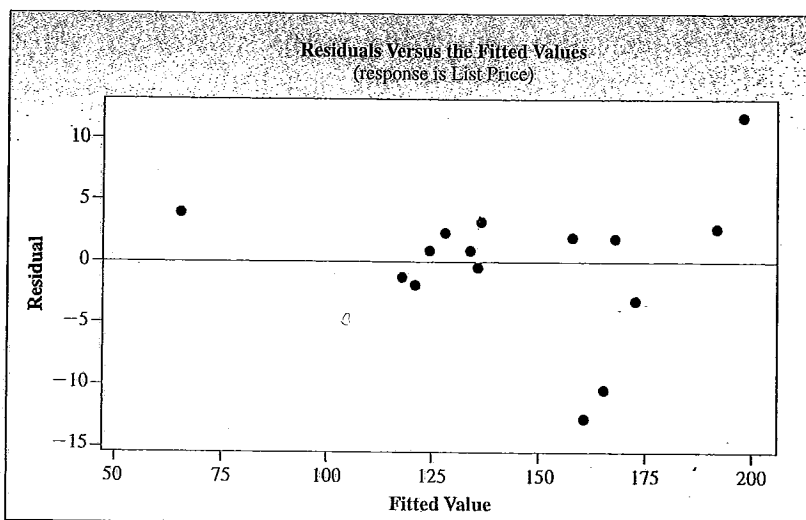
$$SE(\hat{y}) = \sqrt{75.7532 \left\{ \frac{1}{10} + \frac{(50 - 46)^2}{2474} \right\}} = 2.840.$$

Then, with $t_{8, 0.025} = 2.306$, we obtain a 95% confidence interval to be

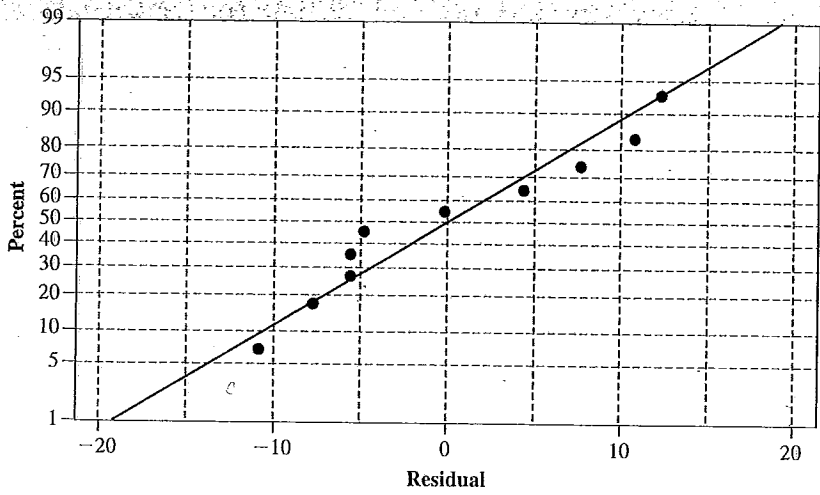
$$79.06 \pm (2.306 \times 2.840) = (72.51, 85.61).$$

Checking the Regression Assumptions

Before using the regression model for its main purpose—estimation and prediction of y —you should look at computer-generated **residual plots** to make sure that all the regression assumptions are valid. The *normal probability plot* and the *plot of residuals versus fit* are shown in Figure 13.5 for the real estate data. There appear to be three observations that do not fit the general pattern. You can see them as outliers in both graphs. These three observations should probably be investigated; however, they do not provide strong evidence that the assumptions are violated.



Normal Probability Plot of the Residuals
(response is y)



Data on 15 Condominiums

Observation	List Price, y	Living Area, x_1	Floors, x_2	Bedrooms, x_3	Baths, x_4
1	69.0	6	1	2	1
2	118.5	10	1	2	2
3	116.5	10	1	3	2
4	125.0	11	1	3	2
5	129.9	13	1	3	1.7
6	135.0	13	2	3	2.5
7	139.9	13	1	3	2
8	147.9	17	2	3	2.5
9	160.0	19	2	3	2
10	169.9	18	1	3	2
11	134.9	13	1	4	2
12	155.0	18	1	4	2
13	169.9	17	2	4	3
14	194.5	20	2	4	3
15	209.9	21	2	4	3

13.3 A MULTIPLE REGRES

Regression Analysis: List Price versus Square Feet, Number of Floors, Bedrooms, Baths

The regression equation is

$$\text{ListPrice} = 18.8 + 6.27 \text{ Square Feet} - 16.2 \text{ Number of Floors} - 2.67 \text{ Bedrooms} + 30.3 \text{ Baths}$$

Predictor	Coef	SE Coef	T	P
Constant	18.763	9.207	2.04	0.069
Square Feet	6.2698	0.7252	8.65	0.000
Number of Floors	-16.203	6.212	-2.61	0.026
Bedrooms	-2.673	4.494	-0.59	0.565
Baths	30.271	6.849	4.42	0.001

S = 6.84930 R-Sq = 97.1% R-Sq(adj) = 96.0%

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	4	15913.0	3978.3	84.80	0.000
Residual Error	10	469.1	46.9		
Total	14	16382.2			

Source	DF	Seq SS
Square Feet	1	14829.3
Number of Floors	1	0.9
Bedrooms	1	166.4
Baths	1	916.5

Predictor	Coef	SE Coef	T	P
Constant	18.763	9.207	2.04	0.069
Square Feet	6.2698	0.7252	8.65	0.000
Number of Floors	-16.203	6.212	-2.61	0.026
Bedrooms	-2.673	4.494	-0.59	0.565
Baths	30.271	6.849	4.42	0.001

Salary Versus Gender and Years of Experience

Years of Experience, x_1	Salary for Men, y	Salary for Women, y
1	\$50,710	\$49,510
2	—	50,440
3	53,160	51,340
3	53,210	51,760
4	54,140	52,750
5	55,760	53,200
5	55,590	—

Solution The *MINITAB* regression printout for the data in Table 13.3 is shown in Figure 13.12. You can use a step-by-step approach to interpret this regression analysis, beginning with the fitted prediction equation, $\hat{y} = 48,593 + 969x_1 + 867x_2 + 260x_1x_2$. By substituting $x_2 = 0$ or 1 into this equation, you get two straight lines—one for women and one for men—to predict the value of y for a given x_1 . These lines are

$$\text{Women: } \hat{y} = 48,593 + 969x_1$$

$$\text{Men: } \hat{y} = 49,460 + 1229x_1$$

Regression Analysis: y versus x1, x2, x1x2

The regression equation is
 $y = 48593 + 969 x1 + 867 x2 + 260 x1x2$

Predictor	Coef	SE Coef	T	P
Constant	48593.0	207.9	233.68	0.000
x1	969.00	63.67	15.22	0.000
x2	866.7	305.3	2.84	0.022
x1x2	260.13	87.06	2.99	0.017

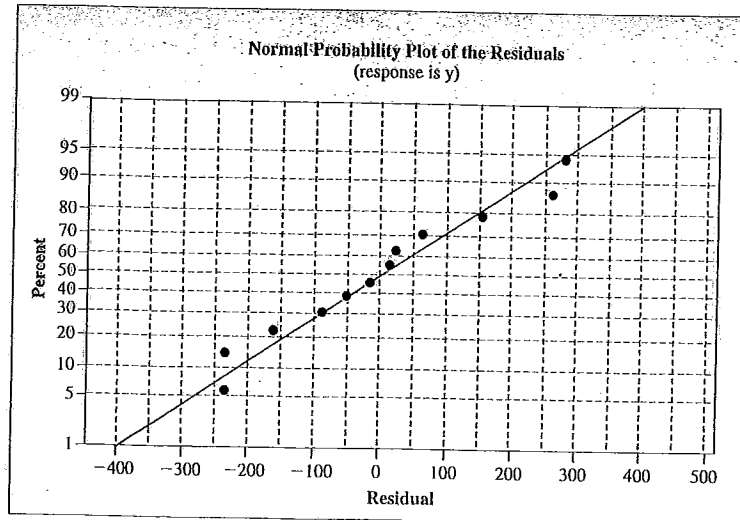
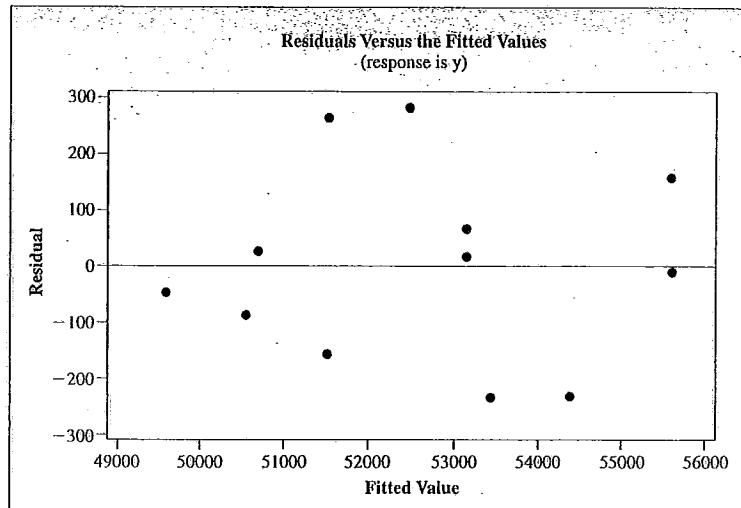
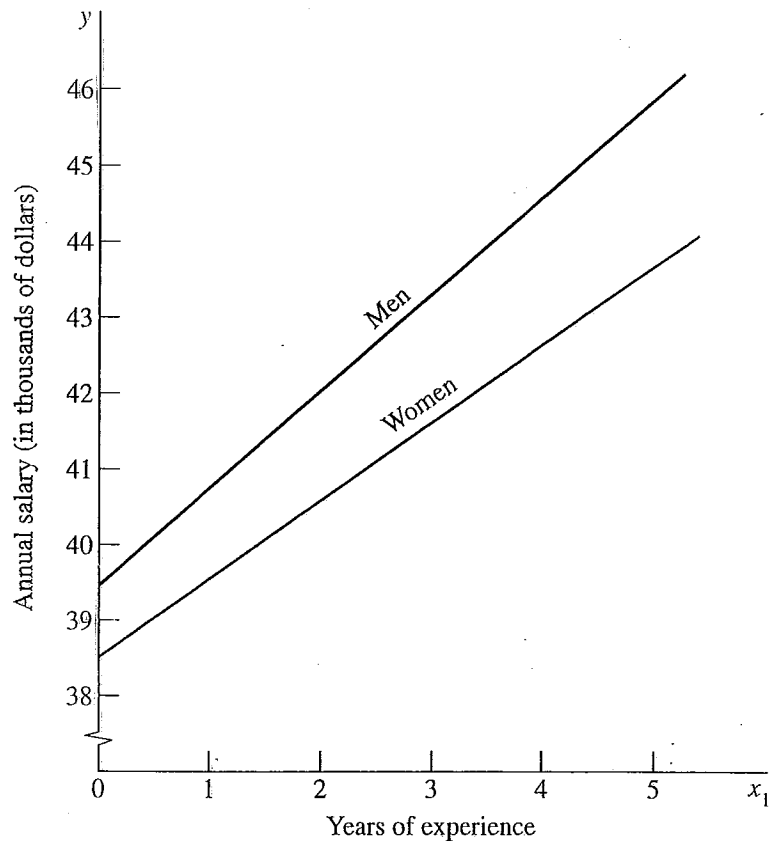
S = 201.344 R-Sq = 99.2% R-Sq(adj) = 98.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	42108777	14036259	346.24	0.000
Residual Error	8	324315	40539		
Total	11	42433092			

Source	DF	Seq SS
x1	1	33294036
x2	1	8452797
x1x2	1	361944

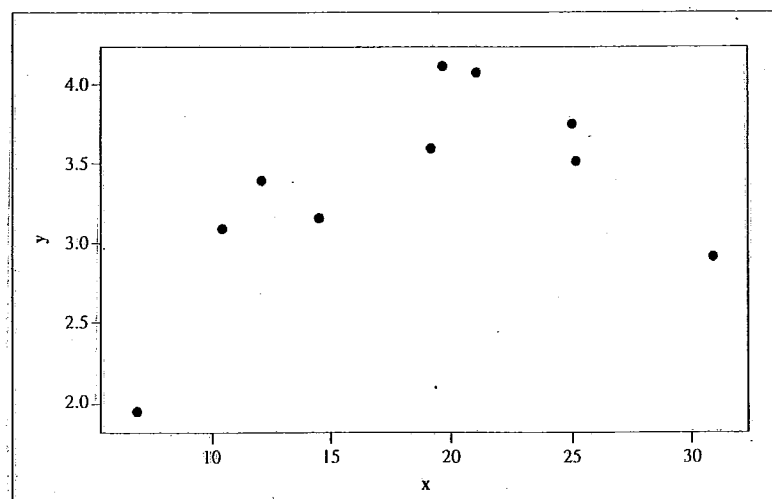
Next, consider the overall fit of the model using the analysis of variance F test. Since the observed test statistic in the ANOVA portion of the printout is $F = 346.24$ with $P = .000$, you can conclude that at least one of the predictor variables is contributing information for the prediction of y . The strength of this model is further measured by the coefficient of determination, $R^2 = 99.2\%$. You can see that the model appears to fit very well.



In a study of variables that affect productivity in the retail grocery trade, W.S. Good uses value added per work-hour to measure the productivity of retail grocery outlets.¹ He defines "value added" as "the surplus [money generated by the business] available to pay for labor, furniture and fixtures, and equipment." Data consistent with the relationship between value added per work-hour y and the size x of a grocery outlet described in Good's article are shown in Table 13.2 for 10 fictitious grocery outlets. Choose a model to relate y to x .

Data on Store Size and Value Added

Store	Value Added Per Work-Hour, y	Size of Store (thousand square feet), x
1	\$4.08	21.0
2	3.40	12.0
3	3.51	25.2
4	3.09	10.4
5	2.92	30.9
6	1.94	6.8
7	4.11	19.6
8	3.16	14.5
9	3.75	25.0
10	3.60	19.1



Scatter Plot

The regression equation is
 $Y = -0.159 + 0.392x - 0.00949x^2$

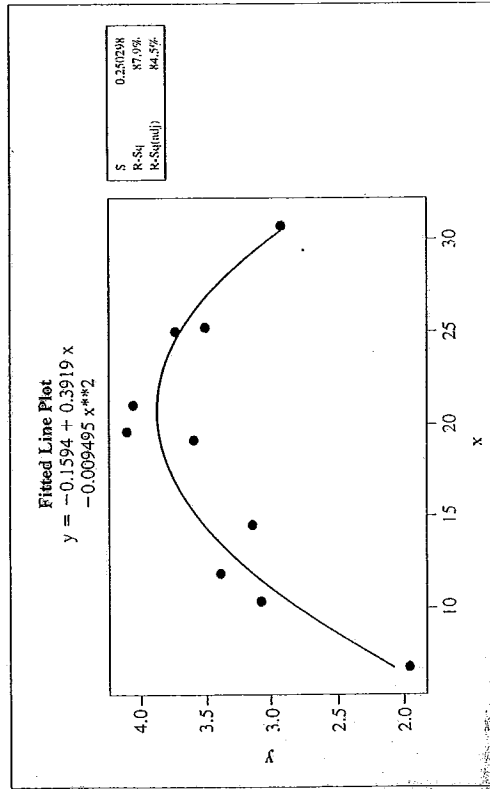
Predictor	Coef	St Coef	T	P
Constant	-0.1594	0.5006	-0.32	0.760
x	0.39193	0.05801	6.76	0.000
x-sq	-0.009495	0.001535	-6.19	0.000

S = 0.250298 R-Sq = 87.9% R-Sq(adj) = 84.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	3.1989	1.5994	25.53	0.001
Residual Error	7	0.4385	0.0626		
Total	9	3.6374			

Source	DF	Seq SS
x	1	0.8003
x-sq	1	2.3986



To assess the adequacy of the quadratic model, the test of

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

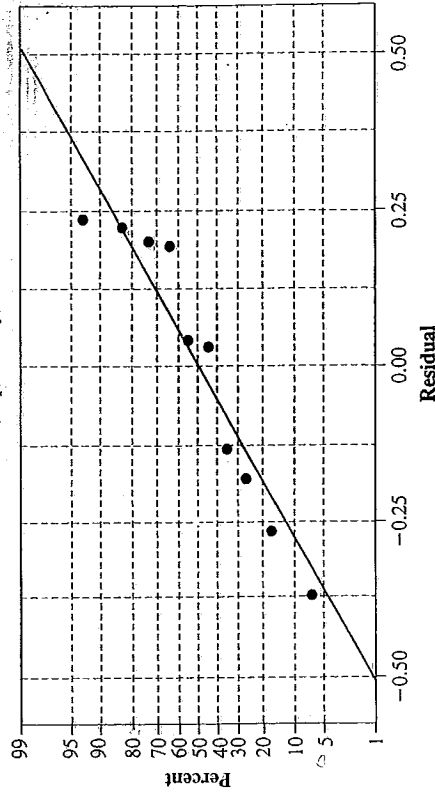
VERSUS

$$H_a: \text{Either } \beta_1 \text{ or } \beta_2 \text{ is not } 0$$

is given in the printout as

$$F = \frac{MSR}{MSE} = 25.53$$

Normal Probability Plot of the Residuals
 (response is y)



Residuals Versus the Fitted Values
 (response is y)

