12 Simple Linear Regression and Correlation

12.2 Estimating Model Parameters

Estimating Model Parameters

We will assume in this and the next several sections that the variables x and y are related according to the simple linear regression model.

The values of β_0, β_1, and σ^2 will almost never be known to an investigator. Instead, sample data consisting of n observed pairs $(x_1, y_1), \ldots, (x_n, y_n)$ will be available, from which the model parameters and the true regression line itself can be estimated.

These observations are assumed to have been obtained independently of one another.

Estimating Model Parameters

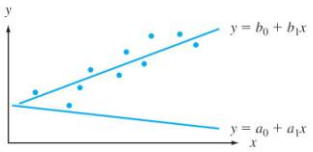
That is, y_i is the observed value of Y_i, where $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and the n deviations $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent rv's.

Independence of $Y_1, Y_2, \ldots, Y_n$ follows from independence of the ϵ_i's.

According to the model, the observed points will be distributed about the true regression line in a random manner.

Estimating Model Parameters

Figure 12.6 shows a typical plot of observed pairs along with two candidates for the estimated regression line.



Two different estimates of the true regression line

Figure 12.6

5

Estimating Model Parameters

Intuitively, the line $y = a_0 + a_1x$ is not a reasonable estimate of the true line $y = \beta_0 + \beta_1x$ because, if $y = a_0 + a_1x$ were the true line, the observed points would almost surely have been closer to this line.

The line $y = b_0 + b_1x$ is a more plausible estimate because the observed points are scattered rather closely about this line.

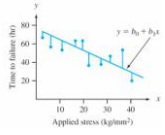
Figure 12.6 and the foregoing discussion suggest that our estimate of $y = \beta_0 + \beta_1x$ should be a line that provides in some sense a best fit to the observed data points.

6

Estimating Model Parameters

This is what motivates the principle of least squares, which can be traced back to the German mathematician Gauss (1777–1855).

According to this principle, a line provides a good fit to the data if the vertical distances (deviations) from the observed points to the line are small (see Figure 12.7).



Deviations of observed data from line $y = b_0 + b_1x$

Figure 12.7

7

Estimating Model Parameters

The measure of the goodness of fit is the sum of the squares of these deviations. The best-fit line is then the one having the smallest possible sum of squared deviations.

Principle of Least Squares
The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1x$ is

height of point – height of line = $y_i - (b_0 + b_1x_i)$

8

Estimating Model Parameters

The sum of squared vertical deviations from the points $(x_1, y_1),\ldots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates,** are those values that minimize $f(b_0, b_1)$.

That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are such that $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ for any b_0 and b_1.

Estimating Model Parameters

The **estimated regression line** or **least squares line** is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero [analogously to $f'(b) = 0$ in univariate calculus], and solving the equations

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

Estimating Model Parameters

Cancellation of the –2 factor and rearrangement gives the following system of equations, called the **normal equations:**

$$nb_0 + (\Sigma x_i) b_1 = \Sigma y_i$$

$$(\Sigma x_i) b_0 + (\Sigma x_i^2) b_1 = \Sigma x_i y_i$$

These equations are linear in the two unknowns b_0 and b_1.

Provided that not all x_i's are identical, the least squares estimates are the unique solution to this system.

Estimating Model Parameters

The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad \textbf{(12.2)}$$

Computing formulas for the numerator and denominator of $\hat{\beta}_1$ are

$$S_{xy} = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n \qquad S_{xx} = \Sigma x_i^2 - (\Sigma x_i)^2/n$$

Estimating Model Parameters

The least squares estimate of the intercept β_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \frac{\Sigma y_i - \hat{\beta}_1 \Sigma x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \textbf{(12.3)}$$

The computational formulas for S_{xy} and S_{xx} require only the summary statistics Σx_i, Σy_i, Σx_i^2 and $\Sigma x_i y_i$ (Σy_i^2 will be needed shortly).

In computing $\hat{\beta}_0$, use extra digits in $\hat{\beta}_1$ because, if $\bar{x}$ is large in magnitude, rounding will affect the final answer.
In practice, the use of a statistical software package is preferable to hand calculation and hand-drawn plots.

13

Estimating Model Parameters

Once again, be sure that the scatter plot shows a linear pattern with relatively homogenous variation before fitting the simple linear regression model.

14

Example 4

The cetane number is a critical property in specifying the ignition quality of a fuel used in a diesel engine.

Determination of this number for a biodiesel fuel is expensive and time-consuming.

The article "Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study" (*J. of Automobile Engr*., 2009: 565–583) included the following data on x = iodine value (g) and y = cetane number for a sample of 14 biofuels.

15

Example 4

cont'd

The iodine value is the amount of iodine necessary to saturate a sample of 100 g of oil. The article's authors fit the simple linear regression model to this data, so let's follow their lead.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

The necessary summary quantities for hand calculation can be obtained by placing the x values in a column and the y values in another column and then creating columns for x^2, xy, and y^2 (these latter values are not needed at the moment but will be used shortly).

16

Example 4 cont'd

Calculating the column sums gives

$\Sigma x_i = 1307.5$, $\Sigma y_i = 779.2$,

$\Sigma x_i^2 = 128{,}913.93$, $\Sigma x_i y_i = 71{,}347.30$,

$\Sigma y_i^2 = 43{,}745.22$, from which

$S_{xx} = 128{,}913.93 - (1307.5)^2/14 = 6802.7693$

$S_{xy} = 71{,}347.30 - (1307.5)(779.2)/14 = -1424.41429$

17

Example 4 cont'd

The estimated slope of the true regression line (i.e., the slope of the least squares line) is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-1424.41429}{6802.7693} = -.20938742$$

We estimate that the expected change in true average cetane number associated with a 1g increase in iodine value is –.209—i.e., a decrease of .209.

18

Example 4 cont'd

Since $\bar{x} = 93.392857$ and $\bar{y} = 55.657143$, the estimated intercept of the true regression line (i.e., the intercept of the least squares line) is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 55.657143 - (-.20938742)(93.392857)$$

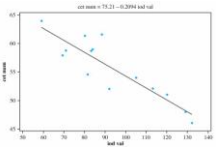
$$= 75.212432$$

The equation of the estimated regression line (least squares line) is $y = 75.212 - .2094x$, exactly that reported in the cited article.

19

Example 4 cont'd

Figure 12.8 displays a scatter plot of the data with the least squares line superimposed.



Scatter plot for Example 4 with least squares line superimposed, from Minitab

Figure 12.8

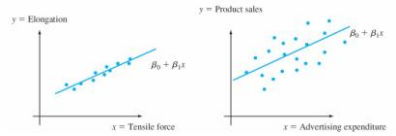
This line provides a very good summary of the relationship between the two variables.

20

Estimating σ^2 and σ

Estimating σ^2 and σ

The parameter σ^2 determines the amount of variability inherent in the regression model. A large value of σ^2 will lead to observed (x_i, y_i)s that are quite spread out about the true regression line, whereas when σ^2 is small the observed points will tend to fall very close to the true line (see Figure 12.9).



Typical sample for σ^2: (a) small; (b) large

Figure 12.9

Estimating σ^2 and σ

An estimate of σ^2 will be used in confidence interval (CI) formulas and hypothesis-testing procedures presented in the next two sections.

Because the equation of the true line is unknown, the estimate is based on the extent to which the sample observations deviate from the estimated line.

Many large deviations (residuals) suggest a large value of σ^2, whereas deviations all of which are small in magnitude suggest that σ^2 is small.

Estimating σ^2 and σ

Definition
The **fitted** (or **predicted**) **values** $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are obtained by successively substituting $x_1, \ldots, x_n$ into the equation of the estimated regression line:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2, \ldots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$$

The **residuals** are the differences $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$ between the observed and fitted y values.

Estimating σ^2 and σ

In words, the predicted value $\hat{y}_i$ is the value of y that we would predict or expect when using the estimated regression line with $x = x_i$; $\hat{y}_i$ is the height of the estimated regression line above the value x_i for which the ith observation was made.

The residual $y_i - \hat{y}_i$ is the vertical deviation between the point (x_i, y_i) and the least squares line—a positive number if the point lies above the line and a negative number if it lies below the line.

25

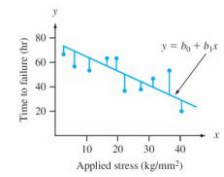
Estimating σ^2 and σ

If the residuals are all small in magnitude, then much of the variability in observed y values appears to be due to the linear relationship between x and y, whereas many large residuals suggest quite a bit of inherent variability in y relative to the amount due to the linear relation.

26

Estimating σ^2 and σ

Assuming that the line in Figure 12.7 is the least squares line, the residuals are identified by the vertical line segments from the observed points to the line.



Deviations of observed data from line $y = b_0 + b_1x$

Figure 12.7

27

Estimating σ^2 and σ

When the estimated regression line is obtained via the principle of least squares, the sum of the residuals should in theory be zero.

In practice, the sum may deviate a bit from zero due to rounding.

28

Example 6

Japan's high population density has resulted in a multitude of resource-usage problems.

One especially serious difficulty concerns waste removal. The article "Innovative Sludge Handling Through Pelletization Thickening" (*Water Research*, 1999: 3245–3252) reported the development of a new compression machine for processing sewage sludge.

An important part of the investigation involved relating the moisture content of compressed pellets (y, in %) to the machine's filtration rate (x, in kg-DS/m/hr).

Example 6

cont'd

The following data was read from a graph in the article:

x	125.3	98.2	201.4	147.3	145.9	124.7	112.2	120.2	161.2	178.9
y	77.9	76.8	81.5	79.8	78.2	78.3	77.5	77.0	80.1	80.2
x	159.5	145.8	75.1	151.4	144.2	125.0	198.8	132.5	159.6	110.7
y	79.9	79.0	76.7	78.2	79.5	78.1	81.5	77.0	79.0	78.6

Relevant summary quantities (*summary statistics*) are

$\Sigma x_i = 2817.9$, $\Sigma y_i = 1574.8$, $\Sigma x^2_i = 415{,}949.85$,

$\Sigma x_i y_i = 222{,}657.88$, and $\Sigma y^2_i = 124{,}039.58$,

from which $x = 140.895$, $y = 78.74$, $S_{xx} = 18{,}921.8295$, $S_{xy} = 776.434$.

Example 6

cont'd

Thus

$$\hat{\beta}_1 = \frac{776.434}{18{,}921.8295} = .04103377 \approx .041$$

$\hat{\beta}_0 = 78.74 - (.04103377)(140.895) = 72.958547 \approx 72.96$

from which the equation of least squares line is $y = 72.96 + .041x$.

For numerical accuracy, the fitted values are calculated from

$\hat{y}_i = 72.958547 + .04103377x_i$:

Example 6

cont'd

$\hat{y}_1 = 72.958547 + .04103377(125.3) \approx 78.100$,

$y_1 - \hat{y}_1 \approx -.200$, etc.

Nine of the 20 residuals are negative, so the corresponding nine points in a scatter plot of the data lie below the estimated regression line.

Example 6

cont'd

All predicted values (fits) and residuals appear in the accompanying table.

Obs	Filtrate	Moistcon	Fit	Residual
1	125.3	77.9	78.100	−0.200
2	98.2	76.8	76.988	−0.188
3	201.4	81.5	81.223	0.277
4	147.3	79.8	79.003	0.797
5	145.9	78.2	78.945	−0.745
6	124.7	78.3	78.075	0.225
7	112.2	77.5	77.563	−0.063
8	120.2	77.0	77.891	−0.891
9	161.2	80.1	79.573	0.527
10	178.9	80.2	80.299	−0.099
11	159.5	79.9	79.503	0.397
12	145.8	79.0	78.941	0.059
13	75.1	76.7	76.040	0.660
14	151.4	78.2	79.171	−0.971
15	144.2	79.5	78.876	0.624
16	125.0	78.1	78.088	0.012
17	198.8	81.5	81.116	0.384
18	132.5	77.0	78.396	−1.396
19	159.6	79.0	79.508	−0.508
20	110.7	78.6	77.501	1.099

33

Estimating σ^2 and σ

In much the same way that the deviations from the mean in a one-sample situation were combined to obtain the estimate $s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$, the estimate of σ^2 in regression analysis is based on squaring and summing the residuals.

We will continue to use the symbol s^2 for this estimated variance, so don't confuse it with our previous s^2.

34

Estimating σ^2 and σ

Definition
The **error sum of squares** (equivalently, residual sum of squares), denoted by SSE, is

$$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = \sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

35

Estimating σ^2 and σ

The divisor $n - 2$ in s^2 is the number of degrees of freedom (df) associated with SSE and the estimate s^2.

This is because to obtain s^2, the two parameters β_0 and β_1 must first be estimated, which results in a loss of 2 df (just as μ had to be estimated in one sample problems, resulting in an estimated variance based on $n - 1$df).

Replacing each y_i in the formula for s^2 by the rv Y_i gives the estimator S^2.

It can be shown that S^2 is an unbiased estimator for σ^2 (though the estimator S is not unbiased for σ).

36

Estimating σ^2 and σ

An interpretation of *s* here is similar to what we suggested earlier for the sample standard deviation:

Very roughly, it is the size of a typical vertical deviation within the sample from the estimated regression line.

37

Example 7

The residuals for the filtration rate–moisture content data were calculated previously.

The corresponding error sum of squares is

$$SSE = (-.200)^2 + (-.188)^2 + \cdots + (1.099)^2 = 7.968$$

The estimate of σ^2 is then $\hat{\sigma}^2 = s^2 = 7.968/(20 - 2) = .4427$, and the estimated standard deviation is

$$\hat{\sigma} = s = \sqrt{.4427} = .665.$$

38

Example 7 cont'd

Roughly speaking, .665 is the magnitude of a typical deviation from the estimated regression line—some points are closer to the line than this and others are further away.

39

Estimating σ^2 and σ

Computation of SSE from the defining formula involves much tedious arithmetic, because both the predicted values and residuals must first be calculated.

Use of the following computational formula does not require these quantities.

$$SSE = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

This expression results from substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into $\sum (y_i - \hat{y}_i)^2$, squaring the summand, carrying through the sum to the resulting three terms, and simplifying.

40

Estimating σ^2 and σ

This computational formula is especially sensitive to the effects of rounding in $\hat{\beta}_0$ and $\hat{\beta}_1$, so carrying as many digits as possible in intermediate computations will protect against round-off error.

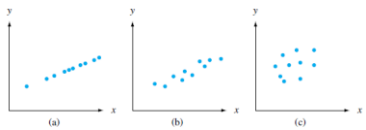
41

The Coefficient of Determination

42

The Coefficient of Determination

Figure 12.10 shows three different scatter plots of bivariate data. In all three plots, the heights of the different points vary substantially, indicating that there is much variability in observed *y* values.



Using the model to explain y variation: (a) data for which all variation is explained; (b) data for which most variation is explained; (c) data for which little variation is explained

Figure 12.10

43

The Coefficient of Determination

The points in the first plot all fall exactly on a straight line. In this case, all (100%) of the sample variation in *y* can be attributed to the fact that *x* and *y* are linearly related in combination with variation in *x*.

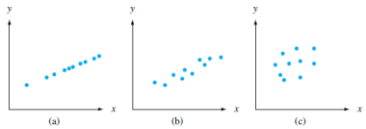
The points in Figure 12.10(b) do not fall exactly on a line, but compared to overall *y* variability, the deviations from the least squares line are small.

It is reasonable to conclude in this case that much of the observed *y* variation can be attributed to the approximate linear relationship between the variables postulated by the simple linear regression model.

44

The Coefficient of Determination

When the scatter plot looks like that of Figure 12.10(c), there is substantial variation about the least squares line relative to overall *y* variation, so the simple linear regression model fails to explain variation in *y* by relating *y* to *x*.



Using the model to explain y variation: (a) data for which all variation is explained; (b) data for which most variation is explained; (c) data for which little variation is explained

Figure 12.10

45

The Coefficient of Determination

The error sum of squares SSE can be interpreted as a measure of how much variation in *y* is left unexplained by the model—that is, how much cannot be attributed to a linear relationship.

In Figure 12.10(a), SSE = 0, and there is no unexplained variation, whereas unexplained variation is small for the data of Figure 12.10(b) and much larger in Figure 12.10(c).

A quantitative measure of the total amount of variation in observed *y* values is given by the **total sum of squares**

$$\text{SST} = S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2/n$$

46

The Coefficient of Determination

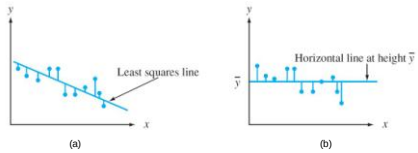
Total sum of squares is the sum of squared deviations about the sample mean of the observed *y* values.

Thus the same number $\bar{y}$ is subtracted from each y_i in SST, whereas SSE involves subtracting each different predicted value $\hat{y}_i$ from the corresponding observed y_i.

47

The Coefficient of Determination

Just as SSE is the sum of squared deviations about the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$, SST is the sum of squared deviations about the horizontal line at height (since then vertical deviations are $y_i - \bar{y}$), as pictured in Figure 12.11.



Sums of squares illustrated: (a) SSE = sum of squared deviations about the least squares line; (b) SSE = sum of squared deviations about the horizontal line

Figure 12.11

48

The Coefficient of Determination

Furthermore, because the sum of squared deviations about the least squares line is smaller than the sum of squared deviations about *any* other line, SSE < SST unless the horizontal line itself is the least squares line.

The ratio SSE/SST is the proportion of total variation that cannot be explained by the simple linear regression model, and 1 – SSE/SST (a number between 0 and 1) is the proportion of observed *y* variation explained by the model.

49

The Coefficient of Determination

Definition
The **coefficient of determination**, denoted by r^2, is given by

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

It is interpreted as the proportion of observed *y* variation that can be explained by the simple linear regression model (attributed to an approximate linear relationship between *y* and *x*).

The higher the value of r^2, the more successful is the simple linear regression model in explaining *y* variation.

50

The Coefficient of Determination

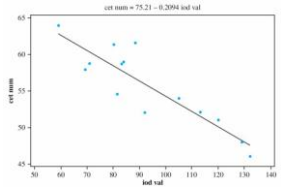
When regression analysis is done by a statistical computer package, either r^2 or $100r^2$ (the percentage of variation explained by the model) is a prominent part of the output.

If r^2 is small, an analyst will usually want to search for an alternative model (either a nonlinear model or a multiple regression model that involves more than a single independent variable) that can more effectively explain *y* variation.

51

Example 9

The scatter plot of the iodine value–cetane number data in Figure 12.8 portends a reasonably high r^2 value.



Scatter plot for Example 4 with least squares line superimposed, from Minitab

Figure 12.8

52

Example 9

cont'd

With

$\hat{\beta}_0 = 75.212432$ $\hat{\beta}_1 = -.20938742$ $\Sigma y_i = 779.2$

$\Sigma x_i y_i = 71{,}347.30$ $\Sigma y^2_i = 43{,}745.22$

we have

SST = 43,745.22 – (779.2)²/14 = 377.174

SSE = 43,745.22 – (75.212432)(779.2)
– (–.20938742)(71,347.30)

= 78.920

Example 9

cont'd

The coefficient of determination is then

$r^2 = 1 - SSE/SST = 1 - (78.920)/(377.174) = .791$

That is, 79.1% of the observed variation in cetane number is attributable to (can be explained by) the simple linear regression relationship between cetane number and iodine value (r^2 values are even higher than this in many scientific contexts, but social scientists would typically be ecstatic at a value anywhere near this large!).

Example 9

cont'd

Figure 12.12 shows partial Minitab output from the regression of cetane number on iodine value.

```
The regression equation is
cet num = 75.2 - 0.209 iod val

                              β̂0      β̂1
Predictor          Coef      SE Coef        T        P
Constant         75.212        2.984    25.21    0.000
iod val        -0.20939      0.03109    -6.73    0.000

s = 2.56450  R-sq = 79.1%     R-sq(adj) = 77.3%
                              100r²

Analysis of Variance                  SSE

SOURCE         DF        SS        MS        F        P
Regression      1    298.25    298.25    45.35    0.000
Error          12     78.92      6.58
Total          13    377.17                      SST
```

Minitab output for the regression of Examples 4 and 9

Figure 12.12

Example 9

cont'd

The software will also provide predicted values, residuals, and other information upon request.

The formats used by other packages differ slightly from that of Minitab, but the information content is very similar. Regression sum of squares will be introduced shortly.

The Coefficient of Determination

The coefficient of determination can be written in a slightly different way by introducing a third sum of squares—**regression sum of squares,** SSR—given by

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = SST - SSE.$$

Regression sum of squares is interpreted as the amount of total variation that *is* explained by the model.

Then we have

$r^2 = 1 - SSE/SST = (SST - SSE)/SST = SSR/SST$

the ratio of explained variation to total variation.

57

The Coefficient of Determination

The ANOVA table in Figure 12.12 shows that SSR = 298.25, from which $r^2 = 298.25/377.17 = .791$ as before.

```
The regression equation is
cet num = 75.2 - 0.209 iod val

                          β̂0      β̂1
Predictor          Coef          SE Coef          T          P
Constant         75.212          2.984      25.21      0.000
iod val       -0.20939         0.03109      -6.73      0.000
                                        100r²
s = 2.56450  R-sq = 79.1%  R-sq(adj) = 77.3%

Analysis of Variance                SSE

SOURCE         DF          SS          MS          F          P
Regression      1      298.25      298.25      45.35      0.000
Error          12       78.92        6.58
Total          13      377.17                              SST
```

Minitab output for the regression of Examples 4 and 9

Figure 12.12

58

Terminology and Scope of Regression Analysis

59

Terminology and Scope of Regression Analysis

The term *regression analysis* was first used by Francis Galton in the late nineteenth century in connection with his work on the relationship between father's height *x* and son's height *y*.

After collecting a number of pairs (x_i, y_i), Galton used the principle of least squares to obtain the equation of the estimated regression line, with the objective of using it to predict son's height from father's height.

60

Terminology and Scope of Regression Analysis

In using the derived line, Galton found that if a father was above average in height, the son would also be expected to be above average in height, *but not by as much as the father was*.

Similarly, the son of a shorter-than-average father would also be expected to be shorter than average, but not by as much as the father.

Thus the predicted height of a son was "pulled back in" toward the mean; because regression means a coming or going back, Galton adopted the terminology *regression line*.

61

Terminology and Scope of Regression Analysis

This phenomenon of being pulled back in toward the mean has been observed in many other situations (e.g., batting averages from year to year in baseball) and is called the **regression effect.**

Our discussion thus far has presumed that the independent variable is under the control of the investigator, so that only the dependent variable Y is random.

This was not, however, the case with Galton's experiment; fathers' heights were not preselected, but instead both X and Y were random.

62

Terminology and Scope of Regression Analysis

Methods and conclusions of regression analysis can be applied both when the values of the independent variable are fixed in advance and when they are random, but because the derivations and interpretations are more straightforward in the former case, we will continue to work explicitly with it.

63