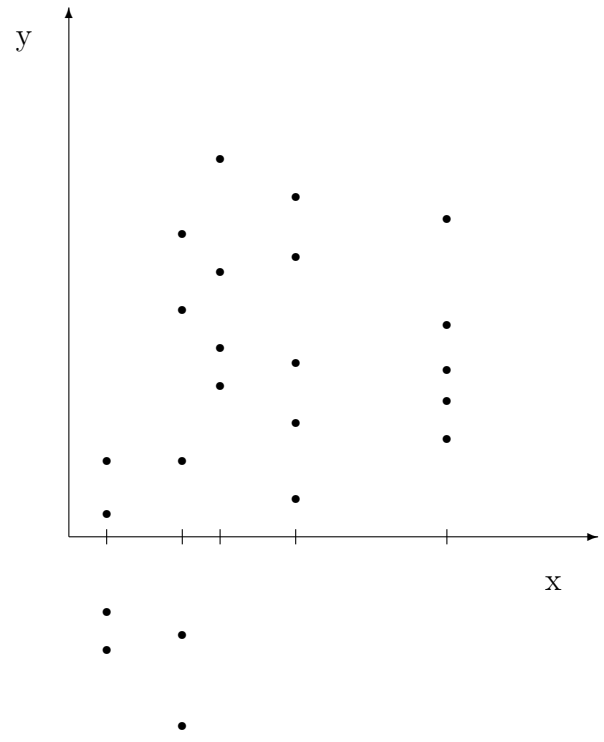
LINEAR REGRESSION AND CORRELATION

Consider `bivariate data` consisting of ordered pairs of numerical values (x, y). Often such data arise by setting an X variable at certain fixed values (which we will call `levels`) and taking a random sample from the population of Y that is assumed to exist at each of the levels of X.

Here we are thinking of X as `not being a random variable`, because we are considering only selected fixed values of X (for sampling purposes).

However, the Y variable is random and we define the random variable Y on the population that exists at each level of X.

Graphically, a scatterplot of the data depicting the y-values obtained by sampling the populations at each of the pre-selected x-values might appear as follows.



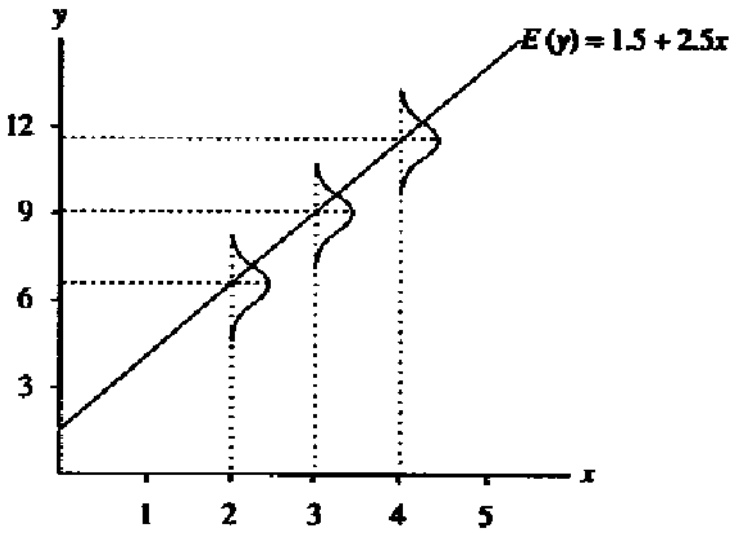
Our objectives given such data are usually to:

1. `Summarize` the characteristics of the Y populations across values of X – `Fit the Model`
2. `Interpolate` between levels of X to estimate parameters of Y populations from which samples were not taken – `Prediction`

The center of our attention is usually on the `means` of the Y populations and especially their `relationship to one another`.

The simple linear regression model says that the populations at each x-value are normally distributed and that the means of these normal distributions all fall on a straight line, called the `regression line.`

FIGURE 11.2
Theoretical distribution of *y* in regression

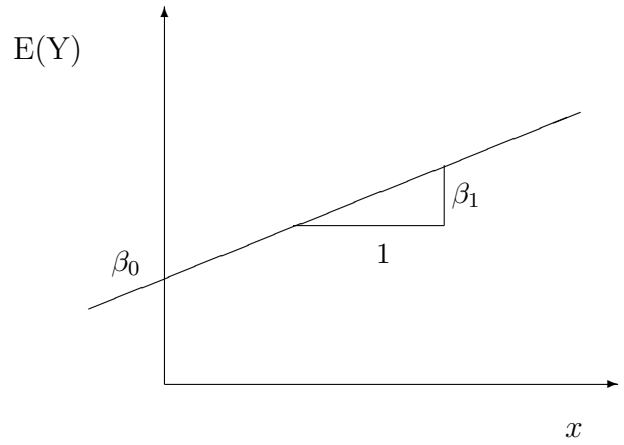


Chapters 11 and 12 are mostly about investigating to what extent the relationship among the population means is `linear`, or some other identifiable mathematical function such as exponential or polynomial of degree > 1.

Let us begin by considering the linear relationship among population means. The equation of a straight line through means $E(Y)$ across x-values can be written as

$$E(Y) = \beta_0 + \beta_1 x \ .$$

Here β_0 is the `intercept` and β_1 is the `slope` of the line.



The y-values observed at each x-value is assumed to be a random sample from a normal distribution with the mean $E(Y) = \beta_0 + \beta_1 x$, i.e., the mean is a linear function of x. The variance of the normal distributions at each x-value is assumed to be the same. Thus the y-values can be related to the x-values through the relationship

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (1)$$

where ϵ is a random variable (called `random error`) with mean zero i.e, $(E(\epsilon) = 0)$, and variance σ_ϵ^2. This model says that sample values are random distances from the line $\mu = \beta_0 + \beta_1 x$ at each x-value. Equation (1) above is called the `simple linear regression model` and β_0, β_1, and σ_ϵ^2 are called the parameters of the model.

The next question we consider is "How do we proceed to derive a `good` approximating line through Y population means, given only samples from some of the Y populations?" In other words, we need to obtain `good` estimates of the parameters of the model using the observed data. The phrase `fitting a line through the data` is used to describe our problem.

It is easy to imagine simply `eye-balling` a line through the points on the scatterplot. It is hard to imagine how this can be a `good` line.

The `method of least squares` provides a more sound and clearly defined procedure. As an example, consider the data in Section 11.2:

`Example: Road Surfacing Data`

Project	1	2	3	4	5
Cost y_i(in \$1000's)	6.0	14.0	10.0	14.0	26.0
Mileage (in miles)	1.0	3.0	4.0	5.0	7.0

In this example, as well as in other examples in this Chapter, for simplicity we will assume that ony one y-value has been observed at each of the x-values. To explain this method we must first define the terms `predicted value` (denoted as $\hat{y}$) and `residual` $(y - \hat{y})$.

The residual is the difference between an observed value y at a given value of x, and $\hat{y}$, the value of y predicted by the model at that particular value of x.

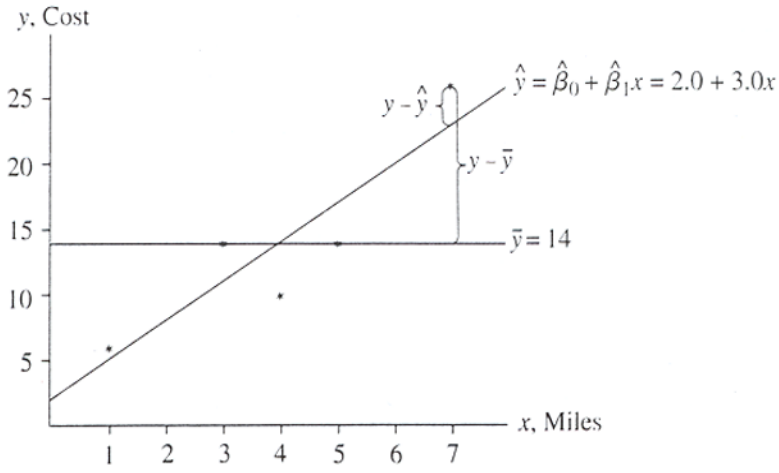
The method of least squares selects a specific line which is claimed to be good. It does so by estimating a value $\hat{\beta}_0$ for β_0 and $\hat{\beta}_1$ for β_1 using the data.

The `least squares` line then has equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ is the point estimate of the mean of the population that exists at x. $\hat{y}$ is called the `predicted value` of y at the specified x. For any one of the sample values y observed from the population existing at x, the difference $y - \hat{y}$ is called a `residual`.

FIGURE 11.9
Deviations from the least-squares line from the mean



The residual $y - \hat{y}$ is the estimate $\hat{\epsilon}$ of ϵ, i.e., it is a point estimate of the sampling error in y under the assumption that the population means lie on a straight line.

Now, the method of least squares selects that line which produces the smallest value of the sum of squares of all residuals (hence the name least squares) i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

where (x_i, y_i) $i = 1, 2, \ldots, n$ are pairs of observations, $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the `least squares estimates` (L.S. estimates) of β_0 and β_1, respectively.

The L.S. estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ given data pairs (x, y) are found as follows:

$$\begin{aligned} \hat{\beta}_1 &= S_{xy}/S_{xx} \qquad\qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \text{where} \quad S_{xx} &= \sum (x - \bar{x})^2 \\ S_{xy} &= \sum (x - \bar{x})(y - \bar{y}) \end{aligned}$$

To simplify matters initially, we considered only sample size 1 from each population. The more general case of larger sample sizes is described as follows.

Let $x_1, x_2, \ldots, x_k$ be the given levels of X. Let $y_{i1}, y_{i2}, \ldots, y_{in_i}$ be the sample of size n_i from the population indexed by x_i. Now

$$\bar{x} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} n_i x_i / n, \qquad \bar{y} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}/n$$

where $n = \sum_{i=1}^{k} n_i$. In this case to compute L.S. estimates we use the following formulas:

$\hat{\beta}_1 = S_{xy}/S_{xx} \qquad\qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$S_{xx} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_i - \bar{x})^2 = \sum_{i=1}^{k} n_i (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_i - \bar{x})(y_{ij} - \bar{y})$$

EXAMPLE 11.2

Data from a sample of 10 pharmacies are used to examine the relation between prescription sales volume and the percentage of prescription ingredients purchased directly from the supplier. The sample data are shown here:

Pharmacy	Sales Volume, y (in \$1,000)	% of Ingredients Purchased Directly, x
1	25	10
2	55	18
3	50	25
4	75	40
5	110	50
6	138	63
7	90	42
8	60	30
9	10	5
10	100	55

a. Find the least-squares estimates for the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
b. Predict sales volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.
c. Plot the (x, y) data and the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
d. Interpret the value of $\hat{\beta}_1$ in the context of the problem.

Solution

To see how the computer does the calculations, you can obtain the least-squares estimates from the following table:

	y	x	$y - \bar{y}$	$x - \bar{x}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	25	10	−46.3	−22.8	1,101.94	566.44
	55	18	−16.3	−15.8	257.54	249.64
	50	25	−21.3	−8.8	187.44	77.44
	75	40	3.7	6.2	22.94	38.44
	110	50	38.7	16.2	626.94	262.44
	138	63	66.7	29.2	1,947.64	852.64
	90	42	18.7	8.2	153.34	67.24
	60	30	−11.3	−3.8	42.94	14.44
	10	5	−61.3	−28.8	1,765.44	829.44
	100	55	28.7	21.2	608.44	449.44
Totals	713	338	0	0	6,714.60	3,407.60
Means	71.3	33.8				

$S_{xx} = \sum (x - \bar{x})^2 = 3{,}407.6$

$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = 6{,}714.6$

Substituting into the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$,

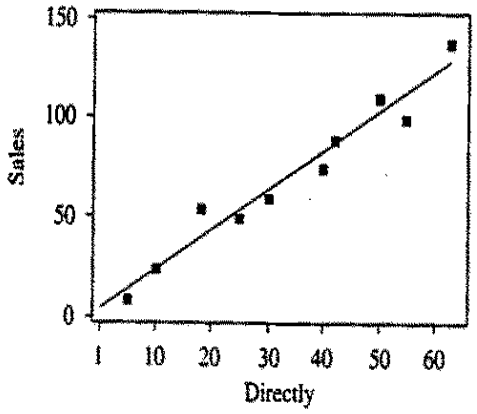
$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{6{,}714.6}{3{,}407.6} = 1.9704778$ rounded to 1.97

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 71.3 - 1.9704778(33.8) = 4.6978519$ rounded to 4.70

b. When $x = 15\%$, the predicted sales volume is $\hat{y} = 4.70 + 1.97(15) = 34.25$ (that is, \$34,250).

c. The (x, y) data and prediction equation are shown in Figure 11.10.

FIGURE 11.10
Sample data and least-squares prediction equation



d. From $\hat{\beta}_1 = 1.97$, we conclude that if a pharmacy would increase by 1% the percentage of ingredients purchased directly, then the estimated increase in average sales volume would be \$1,970.

After we obtain a least squares fitted line, we are then usually interested in seeing how well the line seems to go through Y population means. One way to investigate this is to look at relative magnitudes of certain sums of squares. The reason is derived as follows.

The `Total Variation` in all the sample y values is measured by

$$\frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2, \quad n = \sum_{i=1}^{k} n_i$$

Let us deal only with the numerator

$$\text{Total SS} = \text{SSTot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2,$$

A little algebraic manipulation will result in the `algebraic identity`:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i}\sum_{j}(y_{ij} - \hat{y}_i)^2 + \sum_{i}\sum_{j}(\hat{y}_i - \bar{y})^2$$

We interpret this by noting that the measure of total variation, which is the left part of the equation, is expressible as the sum of two parts which constitute the right side. The first part of the right side is the sum of squared residuals.

We would expect residuals to be close to zero if the Y population means lie close to the estimated least squares line. Thus the smaller the value of

$$\text{Residual SS} = \text{SSE} = \sum \sum (y_{ij} - \hat{y}_i)^2$$

more closer the regression line will be to the data. The other term in the right side of the algebraic identity is

$$\text{Regression SS} = \text{SSReg} = \sum_i \sum_j (\hat{y}_i - \bar{y})^2$$

The identity is the basis for `analysis of variance for regression` summarized below:

Source	df	Sum of Squares	Mean Square
Regression	1	SSReg	MSReg=SSReg/1
Error	n-2	SSE	MSE=SSE/(n-2)
Total	n-1	SSTot	

Interpretation of the Slope and Intercept Parameters

In any straight line equation $y = a + bx$, the slope b measures the change in the y-value for a unit change in the x-value (rate of change in y). If b is positive y would increase as x increases and if b is negative y would decrease as x increases.

In the fitted regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, the slope $\hat{\beta}_1$ is the change y-value for a unit change in the x-value predicted by the fitted model.

As in the case when we estimated μ in a single sample case or $\mu_1 - \mu_2$ in the 2 sample case, we need to obtain the standard error of the estimate of $\hat{\beta}_1$ (and of $\hat{\beta}_0$). These indicate how accurate our estimates are and help construct confidence intervals and perform tests of hypotheses about the true parameter values β_0 and β_1.

The standard deviation of $\hat{\beta}_1$, the slope parameter is given by

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_\epsilon}{\sqrt{S_{xx}}}$$

If the error variance σ_ϵ is large, then $\sigma_{\hat{\beta}_1}$ would be large, which says that the slope parameter is estimated with high variability. That is, our estimate of the rate of change in y will be less accurate, which will result in, say, a wider confidence interval for $\hat{\beta}_1$.

By the above definition, we also see that the standard deviation of $\hat{\beta}_1$ is also affected by S_{xx}: smaller S_{xx} the larger $\sigma_{\hat{\beta}_1}$ would be. S_{xx} measures the spread of the x-values around its mean. This says, that if we have not selected enough x-values to cover the range of possible y-values we want to predict, then the model we built will not be able to predict changes in those y's with enough accuracy.

The standard deviation of $\hat{\beta}_0$, the intercept parameter is given by

$$\sigma_{\hat{\beta}_0} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

The intercept estimate is the predicted value of y at $x = 0$. If all of the x's were large in magnitude (all of the same sign) $\bar{x}^2$ would be large compared to S_{xx}, then the standard error of $\hat{\beta}_0$ would be large, as seen by the above formula. Thus, if all of the x's were large, extrapolation of the model to predict at $x = 0$ will be not accurate.

To estimate the above standard deviations we need an estimate of σ_ϵ. Since σ_ϵ^2 is the variance of the random errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, we would construct estimate of σ_ϵ^2 based on the residuals $\hat{\epsilon}_i = y_i - \hat{y}_i,\ i = 1, 2, \ldots, n$. The estimator of σ_ϵ^2 is

$$s_\epsilon^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2} = MSE$$

Recall that for the sample variance s^2 of a sample $y_1, y_2, \ldots, y_n$, we divide $\sum_i (y_i - \bar{y})^2$ by $n-1$ because with $\bar{y}$ we were estimating a single parameter μ. In the estimate s_ϵ^2 for the sample variance of the residuals, the divisor is $n-2$ because in $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ we are estimating 2 parameters: β_0 and β_1. We say that the Residual SS has $n-2$ degrees of freedom.

Using the estimate s_ϵ of σ_ϵ as defined above, the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_0$ are, respectively,

$$s_{\hat{\beta}_1} = \frac{s_\epsilon}{\sqrt{S_{xx}}}$$

and

$$s_{\hat{\beta}_0} = s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

where $s_\epsilon = \sqrt{MSE}$

`Refer to SAS Analysis of Example 11.2: Pharmacy Data`

Computations Associated with the Simple Linear Regression Model

In Example 11.2, the quantities S_{xx}, S_{xy} were computed by first computing the deviations $(x-\bar{x})$, $(y-\bar{y})$ and the products of the deviations $(x-\bar{x})(y-\bar{y})$ and then forming the sums of squares of these quantities. In practice, however, the following formulas can be used in hand computations, making the computation of the deviations unnecessary:

$$\begin{aligned} S_{xx} &= \sum x^2 - \frac{(\sum x)^2}{n} \\ S_{xy} &= \sum xy - \frac{(\sum x)(\sum y)}{n} \\ S_{yy} &= \sum y^2 - \frac{(\sum y)^2}{n} \end{aligned}$$

In Example 11.2, the quantities needed are

$$\sum x = 338, \quad \sum y = 713, \quad \sum x^2 = 14,832\,,$$

$$\sum xy = 30,814, \quad \sum y^2 = 64,719, \quad n = 10$$

$$\begin{aligned} S_{xx} &= \sum x^2 - \frac{(\sum x)^2}{n} = 14,832 - \frac{338^2}{10} = 3,407.6 \\ S_{xy} &= \sum xy - \frac{(\sum x)(\sum y)}{n} = 30,814 - \frac{(338)(713)}{10} \\ &= 6,714.6 \\ S_{yy} &= \sum y^2 - \frac{(\sum y)^2}{n} = 64,719 - \frac{713^2}{10} = 13,882.1 \end{aligned}$$

These could be used to obtain the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ as before.

$$\begin{aligned} \hat{\beta}_1 &= S_{xy}/S_{xx} = 6,714.6/3,407.6 = 1.97048 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = 71.3 - (1.97048)(33.8) = 4.6979 \end{aligned}$$

In addition, the following formulas are needed to compute the quantities for an `analysis of variance` or `anova` table.

$$\text{SSTot} = S_{yy} = 13,882.1$$

$$\begin{aligned} \text{SSReg} &= \frac{S_{xy}^2}{S_{xx}} = 13,230.97 \\ \text{SSE} &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 13,882.1 - 13,230.97 = 651.13 \end{aligned}$$

This gives the following anova table:

Source	df	Sum of Squares	Mean Square
Regression	1	13,230.97	13,230.97
Error	8	651.13	81.39
Total	9	13,882.1	

Coefficient of Determination

$$r^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{\text{SSReg}}{\text{SSTot}} = \frac{13,230.97}{13,882.1} = .9531 \approx 95\%$$

This is a measure of how much better the regression model does in predicting y than just using $\bar{y}$ to predict y.

INFERENCES ABOUT $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

We are still considering the model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

and the least squares fit using a random sample (x_i, y_{ij}), $i = 1, 2, \ldots k;\ j = 1, 2, \ldots, n_i$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

is the prediction equation, and the L.S. estimates have form

$$\hat{\beta}_1 = S_{xy}/S_{xx} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\ .$$

We have assumed that the Y population at each value of x is Normal with mean $\beta_0 + \beta_1 x$ and the same variance σ_ϵ^2 for all populations. Under this assumption, the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are each normally distributed:

$$\begin{aligned} \hat{\beta}_1 &\sim N\left(\beta_1, \sigma_{\hat{\beta}_1}^2\right) \\ \hat{\beta}_0 &\sim N\left(\beta_0, \sigma_{\hat{\beta}_0}^2\right) \end{aligned}$$

We have earlier shown that the estimators of the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_0$ are

$$\hat{\sigma}_{\hat{\beta}_1} = s_{\hat{\beta}_1} = \frac{s_\epsilon}{\sqrt{S_{xx}}}$$

and

$$\hat{\sigma}_{\hat{\beta}_0} = s_{\hat{\beta}_0} = s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}},$$

respectively, where $s_\epsilon = \sqrt{MSE}$

Using the above results, confidence intervals and tests about the parameters β_1 (and β_0) can be obtained.

A $100(1-\alpha)\%$ Confidence Interval for β_1 :

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot s_{\hat{\beta}_1} \quad \text{giving} \quad \hat{\beta}_1 \pm t_{\alpha/2} \cdot \frac{s_\epsilon}{\sqrt{S_{xx}}}$$

where $t_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the student's t distribution with $(n-2)$ degrees of freedom.

Tests of Hypotheses About β_1

H_0 :	1. $\beta_1 \leq 0$	H_a :	1. $\beta_1 > 0$
	2. $\beta_1 \geq 0$		2. $\beta_1 < 0$
	3. $\beta_1 = 0$		3. $\beta_1 \neq 0$

T.S.: $t = \frac{\hat{\beta}_1 - 0}{s_\epsilon/\sqrt{S_{xx}}}$

R.R:

1. Reject H_0 if $t > t_{\alpha,(n-2)}$
2. Reject H_0 if $t < -t_{\alpha,(n-2)}$
3. Reject H_0 if $|t| > t_{\alpha/2,(n-2)}$

where $t_{\alpha,(n-2)}$ is the $100(1-\alpha)$ percentile of the student's t distribution with $(n-2)$ degrees of freedom.

For a hypothesis like $H_0 : \beta_1 = 3$, the test statistic is modified as,

$$t = \frac{\hat{\beta}_1 - 3}{s_\epsilon/\sqrt{S_{xx}}}$$

An F-test from the analysis of variance

An alternative test of

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0\,,$$

which is more important in the multiple regression case than our simple linear regression models, comes from the `analysis of variance table` given below.

Source	df	Sum of Squares	Mean Square	F
Regression	1	SSReg	MSReg	F=MSReg/MSE
Error	n-2	SSE	MSE	
Total	n-1	SSTot		

The F test statistic computed above is used for an F-distribution-based test with $df_1 = 1$ and $df_2 = n - 2$. Intuitively, large values of this ratio do indicate that the slope β_1 is not zero.

$H_0 : \ \beta_1 = 0$ against $H_a : \beta_1 \neq 0$

T.S.: $F = \frac{\text{MSReg}}{\text{MSE}}$

R.R: Reject H_0 if $F > F_\alpha$

where F_α is the $100(1-\alpha)$ percentile of the F distribution with $df_1 = 1$ and $df_2 = n - 2$

Example 11.6

A simple linear regression model was fitted to the mean age, x, of executives of 15 firms in the food industry and the previous year's percentage increase in earning per share of the firms, y.

Mean Age	38.2	40.0	42.5	43.4	44.6	44.9	45.0	45.4
% Change(in earnings per share)	8.9	13.0	4.7	-2.4	12.5	18.4	6.6	13.5
Mean Age	46.0	47.3	47.3	48.0	49.1	50.5	51.6	
% Change(in earnings per share)	8.5	15.3	18.9	6.0	10.4	15.9	17.1	

The quantities needed are

$$\sum x = 683.8, \quad \sum y = 167.3, \quad \sum x^2 = 31,358.58\,,$$

$$\sum xy = 7,741.74, \quad \sum y^2 = 2,349.61$$

$$S_{xx} \ = \ \sum x^2 - \frac{(\sum x)^2}{n} = 31,358.58 - \frac{683.8^2}{15} = 186.4173$$

$$\begin{aligned} S_{xy} &= \sum xy - \frac{(\sum x)(\sum y)}{n} = 7,741.74 - \frac{(683.8)(167.3)}{15} \\ &= 115.0907 \\ S_{yy} &= \sum y^2 - \frac{(\sum y)^2}{n} = 2,349.61 - \frac{167.3^2}{15} = 483.6573 \end{aligned}$$

These could be used to obtain the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $s_\epsilon = \sqrt{MSE}$ as before.

$$\begin{aligned} \hat{\beta}_1 &= S_{xy}/S_{xx} = 115.0907/186.4173 = 0.617382 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 11.153 - (0.617382)(45.5867) = -16.991 \\ \text{SSE} &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 483.6573 - \frac{115.0907^2}{186.4173} = 412.60236 \\ \text{MSE} &= \text{SSE}\ /(n-2) = 412.60236/13 = 31.7386 \\ s_\epsilon &= \sqrt{MSE} = 5.634 \end{aligned}$$

Thus a 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{.025,13} \cdot \frac{s_\epsilon}{\sqrt{S_{xx}}}$$

$$0.617382 \pm (2.16)(\frac{5.634}{\sqrt{186.4173}}) \quad \text{or} \quad 0.617382 \pm 0.89130$$

i.e., $(-0.27392, 1.5087)$. In this problem, to determine if executive age has any predictive value for predicting change in earnings, we need to test

$$H_0 :\ \beta_1 = 0 \quad \text{vs.} \quad H_a :\ \beta_1 \neq 0$$

We chose the two-sided research hypothesis because, if executive age was a good predictor, we do not know whether it would have a negative or a positive effect on change in earnings. We use $\alpha = .05$ for the test

T.S.: $t = \frac{\hat{\beta}_1 - 0}{s_\epsilon/\sqrt{S_{xx}}} = \frac{(0.617382-0)}{5.634/\sqrt{186.4173}} = \frac{0.617382}{0.412642} = 1.496$

R.R.: $|t| > t_{.025,13} = 2.16$

Since the computed t-statistic is not in the rejection region we fail to reject H_0 and there is no evidence to conclude that change in earnings can be predicted by executive age using a regression model.

We can also use an F-test to test the above hypothesis. The calculations above gives the following Anova table:

Source	df	Sum of Squares	Mean Square	F
Regression	1	71.0549	71.0549	2.24
Error	13	412.6024	31.7386	
Total	14	483.6573		

The rejection region for the F-test at $\alpha = .05$ is $F > F_{.05,1,13}$ i.e., $F > 4.67$ from Table 8. Thus we fail to reject $H_0 : \beta_1 = 0$ again at $\alpha = .05$. The p-value is between .10 and .25 from Table 8.

Coefficient of determination is

$$r^2 = \frac{\text{SSReg}}{\text{SSTot}} = \frac{71.0549}{483.6573} = .1469 = 14.7\%$$

This says that using executive age as a predictor of change in earnings in a straight line model is only 14.7% better than using the sample mean of change in earnings.

Another interpretation of r^2 is that it is the proportion or percentage of variation in y that is explained by $\hat{y}$. In multiple regression models, this interpretation is affected the number of x variables in the model.

`Refer to SAS Analysis of Example 11.6`

Predicting New y Values Using Regression

There are two possible interpretations of a y prediction at a specified value of x.

Recall that the prediction equation for the highway construction problem was $\hat{y} = 2.0 + 3.0\,x$, where y = cost of highway construction contract and x = miles of highway. The highway director substitutes $x = 6$ in this equation and gets the value $\hat{y} = 20$.

This `predicted value` of y can be interpreted as either.

> `The average or mean cost` $E(y)$ `of all resurfacing contracts for 6 miles of road will be $20,000.`

or

> `The cost` y `of a specific resurfacing contract for 6 miles of road will be $20,000.`

The difference in the two predictions is that the standard error of predictions (and therfore the confidence intervals associated with them) will be different. Since it is easier to more accurately predict a mean than an individual value, the first type of prediction will have less error than the the second type.

Predicting the mean $E(Y)$ at a given x

For any Y population, $E(Y)$ is the population mean. According to our model, the expression for $E(Y)$ in terms of x and the parameters β_0 and β_1 is

$$E(Y) = \beta_0 + \beta_1 x\,.$$

The least squares (point) estimate of $E(Y)$ for a given population at a new value of x (call it x_{n+1}) is

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}\,.$$

Using our assumptions about ϵ in the model description, the standard deviation of $\hat{y}_{n+1}$ is

$$\sigma_\epsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

We estimate this by

$$\text{s.e.}(\hat{y}_{n+1}) = s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

where $s_\epsilon^2 = \text{SSE}/(\text{n} - 2)$ Since we assume normally distributed data we have that a $100(1 - \alpha)\%$ confidence interval for $E(Y)$ is

$$\hat{y}_{n+1} \pm t_{\alpha/2} \cdot s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

where $t_{\alpha/2}$ is based on $df = n - 2$.

Example: (Example 11.2 continued)

The prediction equation in the pharmacy example is

$$\hat{y} = 4.70 + 1.97\,x$$

If the % of ingredients purchased directly by a pharmacy is 15, i.e., $x_{n+1} = 15$, obtain a 95% confidence interval for the mean sales volume $E(Y_{n+1})$ for similar pharmacies.

The point estimate of $E(Y_{n+1})$ at $x_{n+1} = 15$ is

$$\hat{y}_{n+1} = 4.70 + (1.97)(15) = 34.25$$

as we have seen before.

The 95% confidence interval for the mean sales volume at $x_{n+1} = 15$ is

$$\hat{y}_{n+1} \pm t_{.025,8} \cdot s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

$$34.25 \pm (2.306)(9.022)\sqrt{\frac{1}{10} + \frac{(15 - 33.8)^2}{3407.6}}$$

$$34.25 \pm 9.39$$

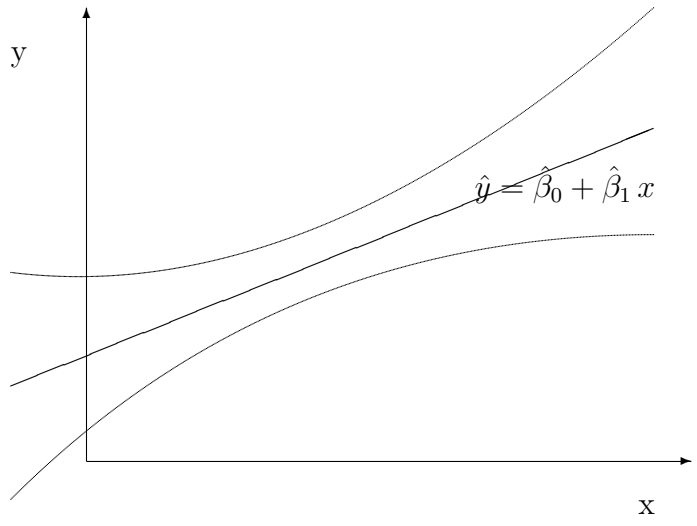
giving (24.86, 43.64) or ($24,860, $43,640).

The confidence interval for $E(Y)$ becomes wider as x_{n+1} gets further away from $\bar{x}$ because the term

$$\frac{(x_{n+1} - \bar{x})^2}{S_{xx}}$$

gets larger. This is called the `extrapolation penalty`.

Since the above interval has endpoints that are a function of x_{n+1} it yields a `100(1-`α`)% confidence band` for $E(Y)$ at all possible x_{n+1} values.



Note that the interval is narrowest at the point $x = \bar{x}$ and gets wider as x move away from $\bar{x}$ and the prediction becomes less accurate.

Predicting a future observation y at a given x

Often it is more relevant to ask a question like "If I take an observation at $x = x_{n+1}$, what y value am I likely to get?"

In other words we are asking what y should we predict at $x = x_{n+1}$. This is different from estimating the average (mean) $E(Y)$ at $x = x_{n+1}$. We now want to predict the value of a future observation, not estimating the population mean $E(Y)$ at $x = x_{n+1}$.

The least squares (point) estimate of y at a new value of x_{n+1}

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 \, x_{n+1} \, .$$

the same as the estimate of $E(Y)$. However the standard error is now different. As we did for a confidence interval for $E(Y)$ we can derive a `prediction interval` for the future y_{n+1}.

A $100(1-\alpha)\%$ Prediction Interval for a future y_{n+1} at x_{n+1} is

$$\hat{y}_{n+1} \pm t_{\alpha/2} \cdot s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

where $s_\epsilon^2 = \text{SSE}/(\text{n-2})$, and $t_{\alpha/2}$ is based on $df = n - 2$.

Note that a 1 has been added to the square root part of the standard error of $\hat{y}_{n+1}$. This means that there is greater error in predicting a future observation compared to estimating a mean, as discussed earlier.

Example: (Example 11.2 continued)

If the % of ingredients purchased directly by a pharmacy is 15, i.e., $x_{n+1} = 15$, obtain a 95% prediction interval for the sales volume y for that pharmacy.

The 95% prediction interval for the sales volume at $x_{n+1} = 15$ is

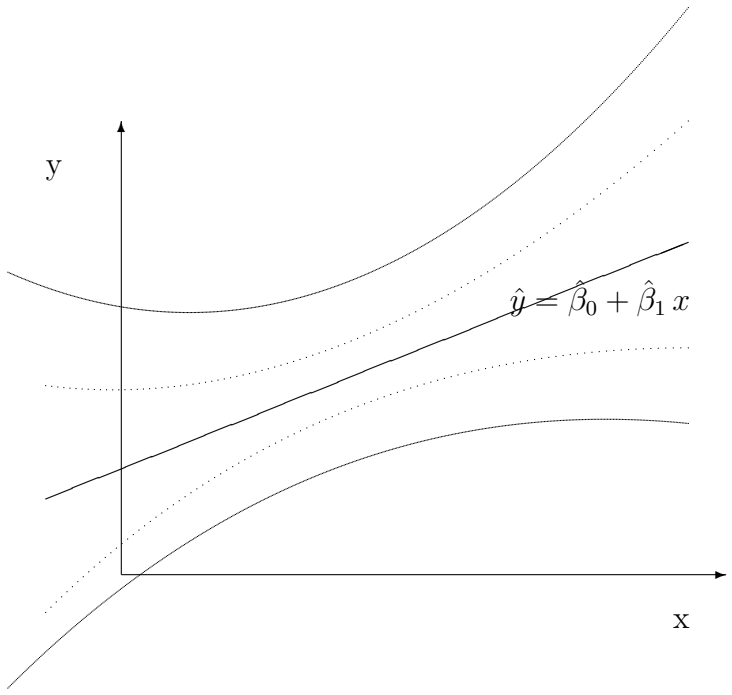
$$\hat{y}_{n+1} \pm t_{.025,8} \cdot s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

$$34.25 \pm (2.306)(9.022)\sqrt{1 + \frac{1}{10} + \frac{(15 - 33.8)^2}{3407.6}}$$

$$34.25 \pm 22.83$$

giving (11.43, 57.08) or ($11,430, $57,080). As you will notice this is a much wider interval than the 95% confidence interval for $E(Y)$ the mean sales volume at $x_{n+1} = 15$.

Since the endpoints of the above prediction interval are a function of x_{n+1}, this is actually a `prediction band`. This band will contain the confidence band for $E(Y)$.



A Statistical Test for Lack of Fit of the Linear Model

The assumptions we have made about the distribution of ϵ's in our linear regression model permit us to derive a test for lack of fit under certain conditions which we will describe.

Whenever the data contain more than one observation at one or more levels of x, we can partition SSE into two parts. This is another algebraic identity like we have seen for partitioning total variability into SSReg and SSE.

Let the data be:

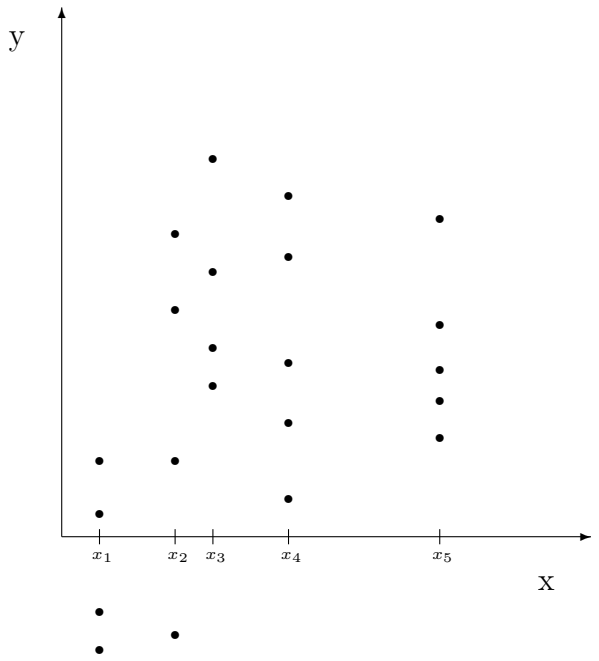
$$(x_i,\, y_{ij}) \quad i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, n_i$$

Thus we imagine k levels of x and at each x_i there are n_i observations $y_{ij}, j = 1, 2, \ldots, n_i$. Graphically we envision a situation like the one shown on the next page. Note that n_i may be one (1) in some cases. If $n_i = 1$ in all cases then we have no repeated observations at any x and we cannot test for lack of fit.

The algebraic identity is:

$$\sum_i \sum_j (y_{ij} - \hat{y}_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i \sum_j (\bar{y}_i - \hat{y}_i)^2$$

SSE	SSE_{exp}	SS_{Lack}
(Sum of squares of residuals)	(SS due to pure experimental error)	(SS due to lack of fit)



Note that the last term in the right hand side of the above equation:

$$\sum \sum (\bar{y}_i - \hat{y}_i)^2 .$$

If indeed there is a linear relationship $E(Y) = \beta_0 + \beta_1 x$ among Y population means, then this sum of squares should not be large because $\bar{y}_i$ is a point estimate of $E(Y_i)$ at x_i, and $\hat{y}_i$

is a point estimate of the same mean $E(Y_i) = \beta_0 + \beta_1 x_i$.

The hypotheses are:

H_0 : A linear model is appropriate

H_a : A linear model is not appropriate

The test for lack of fit is an F test. The F statistic is the ratio of `mean squares` for lack of fit and pure experimental error.

The mean squares are sums of squares divided by their degrees of freedom (this is the definition of a mean square).

$$\text{MS}_{\text{Lack}} = \frac{\text{SS}_{\text{Lack}}}{(n-2) - \sum_i (n_i - 1)} \equiv \frac{\sum_i \sum_j (\bar{y}_i - \hat{y}_i)^2}{(n-2) - \sum_i (n_i - 1)}$$

$$\text{MS}_{\text{exp}} = \frac{\text{SSE}_{\text{exp}}}{\sum_i (n_i - 1)} \equiv \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{\sum_i (n_i - 1)}$$

The F statistic is:

$$F = \frac{\text{MS}_{\text{Lack}}}{\text{MS}_{\text{exp}}},$$

We reject H_0 at level α whenever the computed value of the F-statistic exceeds the $100(1-\alpha)$ percentile from the F table with degrees of freedom $df_1 = n - 2 - \sum_i (n_i - 1)$ and $df_2 = \sum_i (n_i - 1)$.

Failure to reject H_0 implies that there is not enough evidence to declare the linear model inappropriate. (We will never declare the linear model as appropriate because we always view it as an approximation to some unknown relationship among Y population means.)

Correlation

We have proceeded under the assumption that Y population means fall on a straight line. We computed the least squares line as an approximation to this straight line. We also looked at the sum of squared residuals as an indicator of relative success in explaining variation in Y.

There is a measure of the strength of the linear relationship between two variables X and Y.

It is called `correlation coefficient` ρ. Its estimate is called the `sample correlation coefficient` r. For n pairs of observations (x_i, y_i) we define

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} \\ \text{where} S_{xx} &= \sum (x_i - \bar{x})^2, \quad S_{yy} = \sum (y_i - \bar{y})^2 \\ S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

Properties of r are:

1. $-1 \leq r \leq 1$. $r = 0$ indicates no linear relationship between x and y. $r = 1$ indicates a perfect linear relationship between x and y, and the line has positive slope. $r = -1$ also indicates a perfect linear relationship, but with negative slope. Strength is measured, relatively, by how far $|r|$ is from 0 and 1.

   We imagine a true correlation existing as a parameter ρ, and r is the estimate of ρ based on a sample.
2. It can be shown that
   $$r^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$
   i.e. r^2 is the ratio of the `sum of squares due to regression` to the `total sum of squares`. This is the same as the `coefficient of determination` defined earlier.

Its interpretation is that r^2 is the proportion of total variability in Y accounted for by the model.

Since r only indicates the strength of the `linear` relationship between x and y, its value is not useful when there is a strong `curvature` relationship. Another statistic, the `Spearman rank order correlation coefficient` can be used to help us in such situations. The rank order correlation coefficient measures the strength of the `monotonic` association between x and y. This means whether y tends to increase (decrease) with x.

To compute the rank order correlation coefficient you assign ranks to the x and y values separately and then compute the ordinary correlation coefficient for the ranks.

Diagnosing the Fitted Model: Residual Analysis

We review first the consequences of the assumptions of normality, homogeneity of variance,

and independence of errors ϵ_i in the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, \ldots, n$.

Recall that each y_i is a normal random variable because it equals a constant plus a normal random variable, and that the y_i's are independent because the ϵ_i's are independent. We also assumed that the variances of the ϵ_i's for the populations at each of the x_i's are the same and is equal to σ_ϵ^2. This is called the `homogeneity of variance assumption`.

The consequences of this are the results concerning the distributions of $\hat{\beta}_1$, $\hat{\beta}_0$, and $\hat{y}_i$ that we have already used in the inference procedures so far discussed.

That is

$$t = \frac{\hat{\beta}_0 - \beta_0}{s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}},$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_\epsilon \sqrt{S_{XX}}},$$

and,

$$t = \frac{\hat{y} - E(y)}{s_\epsilon \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}}}$$

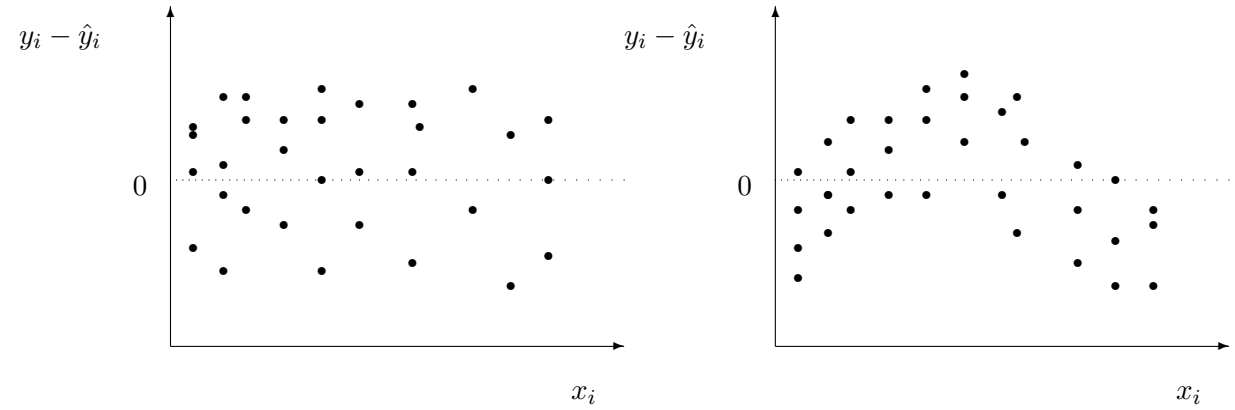
are each distributed as a Student's t random variable.

Residual plotting to look for possible violation of assumptions about ϵ

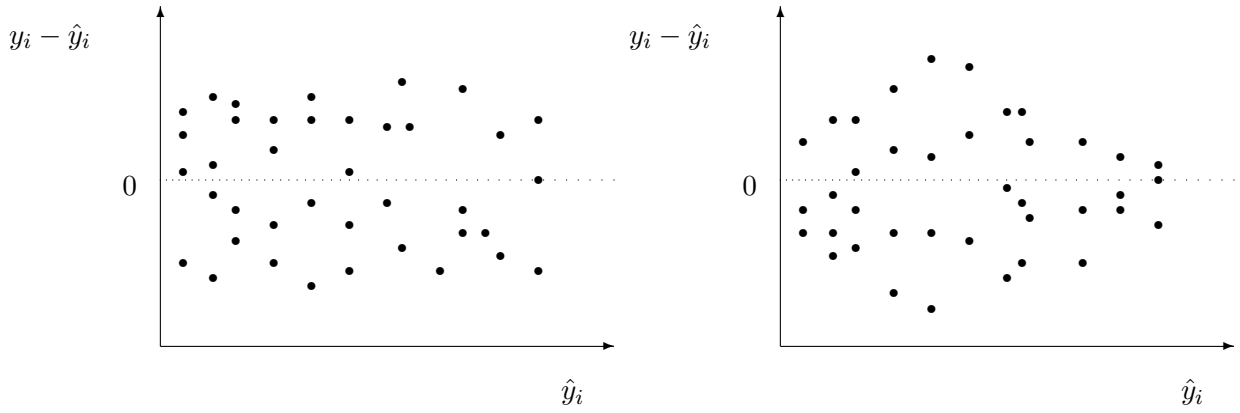
We can use graphics to help identify cases where assumptions about the distribution of ϵ's are not valid. The plots suggested are all based on the fact that the residuals from fitting the model, $(y_i - \hat{y}_i)$, $i = 1, \ldots, n$ are estimates of the ϵ's in the model.

Recommended plots are:

1. Plot residuals $y_i - \hat{y}_i$ vs. x_i. If the model is correct, we would expect the residuals to scatter evenly and randomly around zero as the value of x_i changes. If a curved or nonlinear pattern is apparent, it usually indicates a need for a higher order or a nonlinear model (model inadequacy or nonlinearity). It will also highlight outliers, if any. This plot may also show violation of the homogeneity of variance assumption, as a marked decrease or increase of the spread of the residuals around zero, if the variance depends on the actual value of x_i.



2. Plot of residuals $y_i - \hat{y}_i$ vs. the predicted values $\hat{y}_i$. This scatterplot should show no pattern, and should indicate random scatter of residuals around the zero value. If the homogeneity of variance assumption is violated, a pattern indicating an increase/decrease in spread of the residuals as $\hat{y}_i$ increases. This pattern may show up along with the curved pattern in both this and the previous plot if nonlinearity is also present.



3. A Normal probability plot of the `studentized residuals`

$$\frac{(y_i - \hat{y}_i)}{\text{s.e. of } (y_i - \hat{y}_i)}$$

versus percentiles from the standard normal distribution. The plot will show linearity if the normality assumption about ϵ's is not seriously violated. This plot may also identify one or two outliers if they stand out from a well-defined straight line pattern.

Other Case Statistics

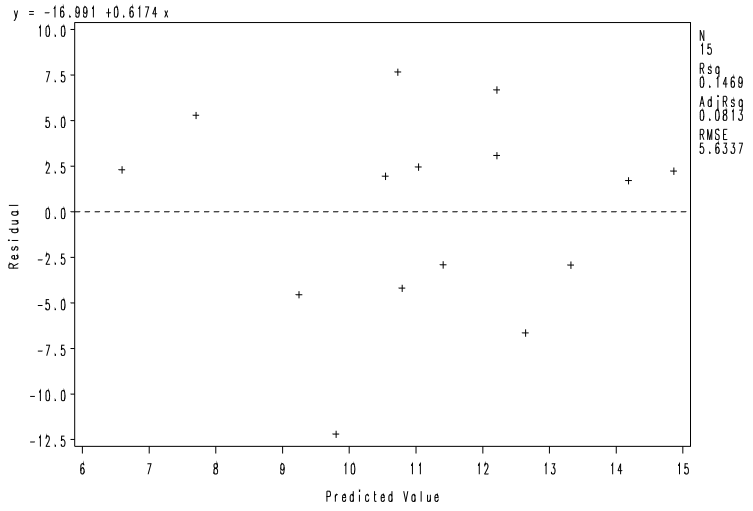
In addition to the residuals and studentized residuals several other statistics related to each observation (or case, as they are commonly called) are computed by computer programs and are printed on request. For the REG procedure in SAS will output case statistics labelled

as Cook's D and Hat Diag, respectively. These statistics measure how well a specific data point fits the regression line. If the point is a distance away from the center of the fitted line in the x-direction it is said to be a `high leverage point` and is called an x-outlier. A high leverage point will show up as having a comparatively large value for the Hat Diag statistic.

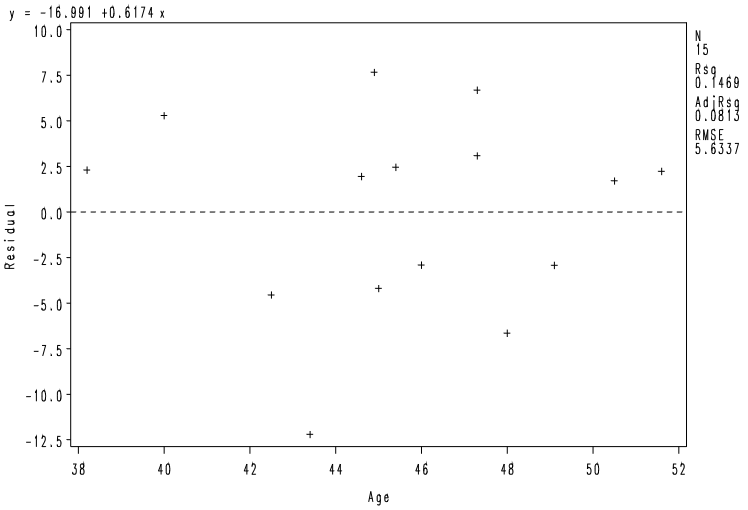
If the point is a distance away from the fitted line in the y-direction, it will have a large residual or large studentized residual. A statistical test procedure is available to check is a studentized residual is significantly large for it to be declared an outlier. The Cook's D case statistic measures `influence` a data point will have on the estimated parameters. That is, it measures whether the deletion of a data point will markedly change the estimated value of the slope parameter β_1. If this happens, then that single data point is said to be highly influential.

A high leverage point that is also a y-outlier will most likely be a high influence point and will have to be examined for correctness by the experimenter, because it may affect the predictions drastically. Read Section 11.2 of the text for a more detailed discussion.

Example 11.6: Earnings data for Food Firms



Example 11.6: Earnings data for Food Firms



Example 11.6: Earnings data for Food Firms