

# Chapter 11: Linear Regression and Correlation

January 20, 2009

## 1 The model

The **simple linear regression** model for  $n$  observations can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n. \quad (1)$$

The designation **simple** indicates that there is only one predictor variable  $x$ , and **linear** means that the model is linear in  $\beta_0$  and  $\beta_1$ .

The intercept  $\beta_0$  and the slope  $\beta_1$  are unknown constants, and they are both called **regression coefficients**;  $\epsilon_i$ 's are random errors.

For model (1), we have the following assumptions:

1. The relation is, in fact, linear, so that the errors all have expected value zero:  $E(\epsilon_i) = 0$  for  $i = 1, 2, \dots, n$ , or, equivalently  $E(y_i) = \beta_0 + \beta_1 x_i$ .
2. The errors all have the same variance:  $\text{Var}(\epsilon_i) = \sigma^2$  for  $i = 1, 2, \dots, n$ , or, equivalently,  $\text{var}(y_i) = \sigma^2$ .
3. The errors are independent of each other.
4. The errors are all normally distributed;  $\epsilon_i$  is normally distributed for all  $i$ .

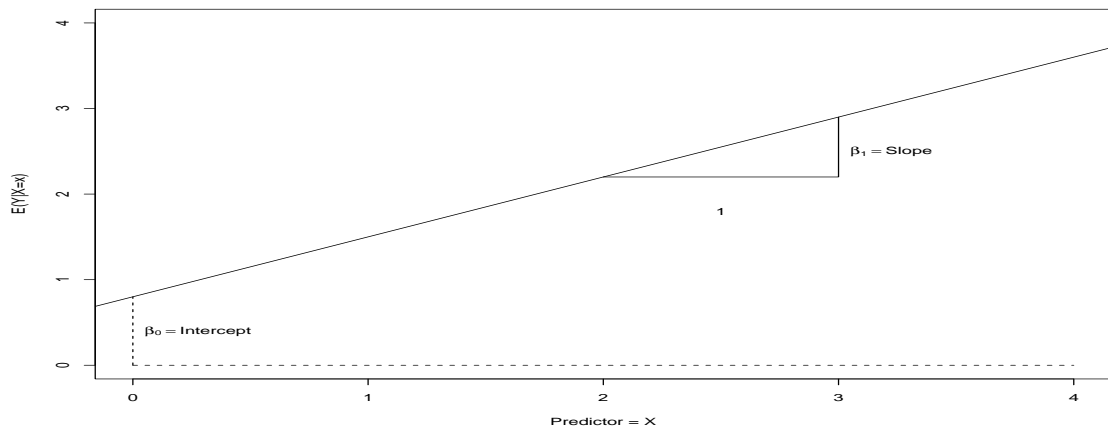


Figure 1: Equation of a straight line  $E(Y|X = x) = \beta_0 + \beta_1 x$ .

## 2 Estimating Model Parameters

The **method of least squares** is to estimate  $\beta_0$  and  $\beta_1$  so that the sum of the squares of the difference between the observations  $y_i$  and the straight line is a minimum, i.e., minimize

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

The least-squares estimators of  $\beta_0$  and  $\beta_1$ , say  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , must satisfy

$$-2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (2)$$

$$-2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \quad (3)$$

Simplifying these two equations yields

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i \end{aligned} \quad (4)$$

Equations (4) are called the **least-squares normal equations**.

The solution to the normal equations is

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned}$$

The difference between the observed value  $y_i$  and the corresponding fitted value  $\hat{y}_i$  is a **residual**, i.e.,

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, 2, \dots, n$$

Pharmacy	Sales Volume, $y$ (in \$1000)	% 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

**Example 2.1** *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 below.*

For this example, we have

$$\bar{x} = 33.8, \quad S_{xx} = 3407.6, \quad S_{xy} = 6714.6,$$

$$\bar{y} = 71.3, \quad S_{xy} = 6714.6.$$

Thus, the parameter estimates are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 1.9704778, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4.6978519.$$

The estimate line, given by either of the equations

$$\hat{E}(y|x) = 4.70 + 1.97x.$$

The fit of this line to the data is excellent as shown in Figure 2.

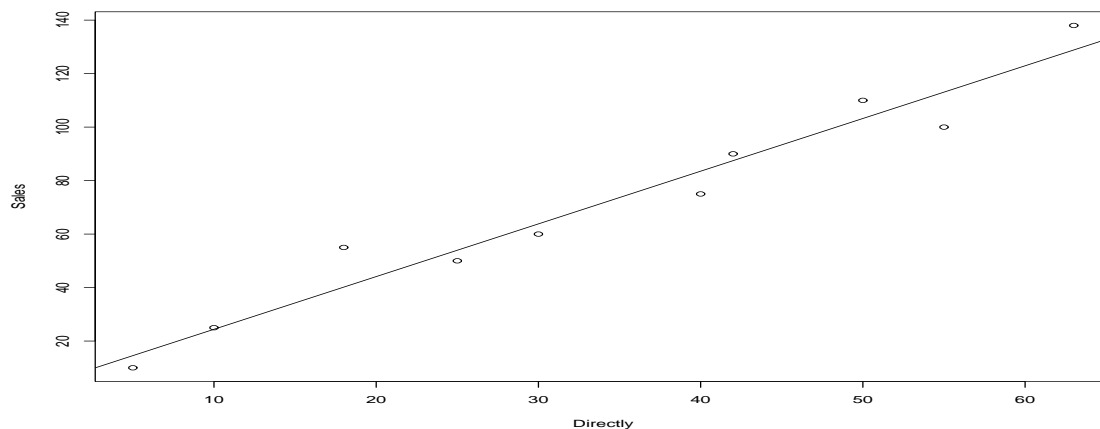


Figure 2: Sample data and least-squares prediction equation.

Some interpretations on the results:

- When  $x = 15\%$ , the predicted sales volume is  $\hat{y} = 4.7 + 1.97(15) = 34.25$ .
- 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 \$1970.

Case Number	Temperature	Pressure	Lpres=100×log( <i>Pressure</i> )
1	194.5	20.79	131.79
2	184.3	20.79	131.79
3	197.9	22.40	135.02
4	198.4	22.67	135.55
5	199.4	23.15	136.46
6	199.9	23.35	136.83
7	200.9	23.89	137.82
8	201.1	23.99	138.0
9	201.4	24.02	138.06
10	201.3	24.01	138.04
11	203.6	25.14	140.04
12	204.6	26.57	142.44
13	209.5	28.49	145.47
14	208.6	27.76	144.34
15	210.7	29.04	146.30
16	211.9	29.88	147.54
17	212.2	30.06	147.80

**Example 2.2** *In an 1857 article, a Scottish physicist named James D. Forbes discussed a series of experiments that he had done at 17 locations in the Alps and Scotland concerning the relationship between atmospheric pressure and the boiling point of water. The data are as follows.*

Using Forbe’s data, we have

$$\begin{aligned} \bar{x} &= 202.95294, & S_{xx} &= 530.78235, & S_{xy} &= 475.31224, \\ \bar{y} &= 139.60529, & S_{yy} &= 427.79402. \end{aligned}$$

Thus, the parameter estimates are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.895, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -42.138.$$

The estimate line, given by either of the equations

$$\hat{E}(Lpress|temp) = -42.138 + 0.895Temp.$$

The fit of this line to the data is excellent as shown in Figure 3.

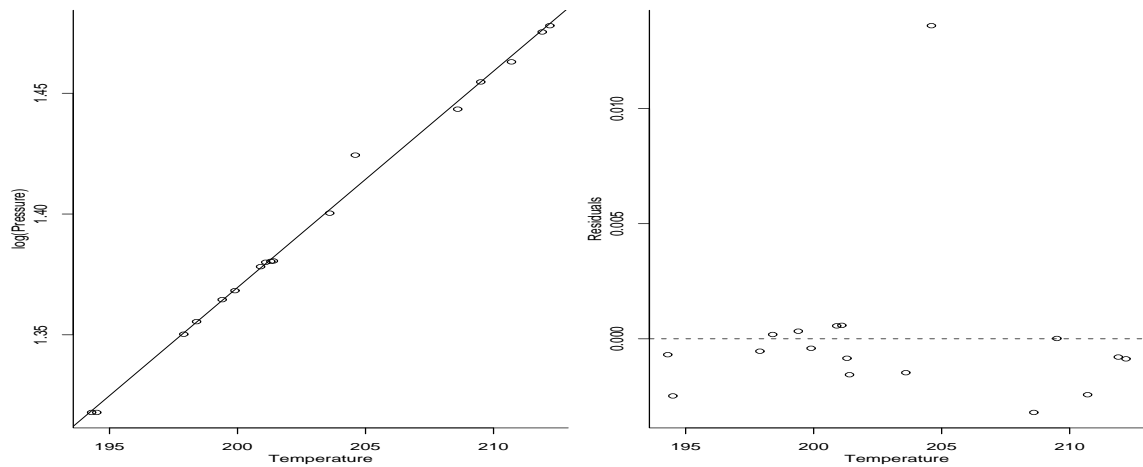


Figure 3: Regression for log(pressure) versus temp.

If the assumptions in section 1 hold, then the least squares estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are unbiased and have minimum variance among all linear unbiased estimates (best linear unbiased estimators).

$$E(\hat{\beta}_1) = \beta_1,$$

$$E(\hat{\beta}_0) = \beta_0,$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\text{var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

### 3 Estimation of $\sigma^2$

The estimate of  $\sigma^2$  is obtained from the residual sum of squares ( $SS_{Res}$ ) or sum of squared error (SSE),

$$SS_{Res} = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

The related formulas are regression sum of squares ( $SS_R$ ) and total sum of squares ( $SS_T$ )

$$SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy},$$
$$SS_T = \sum_{i=1}^n (y_i - \bar{y})^2.$$

And they satisfy the following equation,

$$SS_T = SS_R + SS_{Res}.$$

An unbiased estimate of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n - 2} = MS_{Res}.$$

#### 4 The models in the centered form

Suppose we redefine the regressor variable  $x_i$  as the deviation from its own average, say  $x_i - \bar{x}$ . The regression model then becomes

$$\begin{aligned}y_i &= \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + \epsilon_i \\&= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + \epsilon_i \\&= \beta'_0 + \beta_1(x_i - \bar{x}) + \epsilon_i\end{aligned}$$

It is easy to show that  $\hat{\beta}'_0 = \bar{y}$ , the estimator of the slope is unaffected by the transformation, and  $\text{Cov}(\hat{\beta}'_0, \hat{\beta}_1) = 0$ .

## 5 Hypothesis testing on the slope and intercept

Hypothesis testing and confidence intervals (next section) require the fourth assumption that the model errors  $\epsilon_i$  are normally distributed.

Suppose that we wish to test the hypothesis that the slope equals a constant, say  $\beta_{10}$ . The appropriate hypotheses are

$$\begin{aligned} H_0 : \beta_1 &= \beta_{10} \\ H_1 : \beta_1 &\neq \beta_{10} \end{aligned} \tag{5}$$

Since  $\epsilon_i \sim N(0, \sigma^2)$ , we have  $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$  and  $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$ . Therefore,

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}} \sim N(0, 1)$$

if the null hypothesis  $H_0 : \beta_1 = \beta_{10}$  is true. If  $\sigma^2$  were known, we could use  $Z_0$  to test the hypothesis (5).

If  $\sigma^2$  is unknown, we know that (1)  $MS_{Res}$  is an unbiased estimator of  $\sigma^2$ ; (2)  $(n - 2)MS_{Res}/\sigma^2$  follows a  $\chi_{n-2}^2$  distribution; and (3)  $MS_{Res}$  and  $\hat{\beta}_1$  are independent. Therefore,

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta}_1)}$$

follows a  $t_{n-2}$  distribution if the null hypothesis  $H_0 : \beta_1 = \beta_{10}$  is true. The null hypothesis is rejected if

$$|t_0| > t_{\alpha/2, n-2}.$$

To test

$$\begin{aligned}H_0 : \beta_0 &= \beta_{00} \\H_1 : \beta_0 &\neq \beta_{00},\end{aligned}\tag{6}$$

we could use the test statistic

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_{00}}{se(\hat{\beta}_0)}.$$

For example, in Forbes' data, consider testing the null hypothesis  $\beta_0 = -35$  against the alternative that  $\beta_0 \neq -35$ . The statistic is

$$t = \frac{-42.138 - (-35)}{3.340} = 2.137$$

which has a  $p$ -value near 0.05, providing some evidence against the null hypothesis.

A very important special case of the hypothesis in (5) is

$$\begin{aligned}H_0 : \beta_1 &= 0 \\H_1 : \beta_1 &\neq 0,\end{aligned}\tag{7}$$

Failing to reject the null hypothesis implies that there is no linear relationship between  $x$  and  $y$ .

## 6 The analysis of variance

We may also use an analysis of variance approach to test significance of regression. The analysis of variance is based on the fundamental analysis of variance identity for a regression model, i.e.,

$$SS_T = SS_R + SS_{Res}.$$

$SS_T$  has  $df_T = n - 1$  degrees of freedom because one degree of freedom is lost as a result of constraint  $\sum_{i=1}^n (y_i - \bar{y})$  on the deviations  $y_i - \bar{y}$ ;  $SS_R$  has  $df_R = 1$  degree of freedom because  $SS_R$  is completely determined by one parameter, namely,  $\hat{\beta}_1$ ;  $SS_{Res}$  has  $df_{Res} = n - 2$  degrees of freedom because two constraints are imposed on the deviations  $y_i - \hat{y}_i$  as a result of estimating  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . Note that the degrees of freedom have an additive property:

$$\begin{aligned} df_T &= df_R + df_{Res} \\ n - 1 &= 1 + (n - 2) \end{aligned}$$

We can show: (1) that  $SS_{Res}/\sigma^2 = (n - 2)MS_{Res}/\sigma^2$  follows a  $\chi_{n-2}^2$  distribution; (2) that if the null hypothesis  $H_0 : \beta_1 = 0$  is true, then  $SS_R/\sigma^2$  follows a  $\chi_1^2$  distribution; and (3) that  $SS_{Res}$  and  $SS_R$  are independent. By the definition of an  $F$  statistic,

$$F_0 = \frac{SS_R/df_R}{SS_{Res}/df_{Res}} = \frac{SS_R/1}{SS_{Res}/(n - 2)} = \frac{MS_R}{MS_{Res}}$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	$F_0$
Regress	$SS_R = \hat{\beta}_1 S_{xy}$	1	$MS_R$	$\frac{MS_R}{MS_{Res}}$
Residual	$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$	$n - 2$	$MS_{Res}$	
Total	$SS_T$	$n - 1$		

Table 1: Analysis of Variance (ANOVA) for testing significance of regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	$F_0$
Regress	425.639	1	425.639	2962.79
Residual	2.155	15	0.144	

Table 2: Analysis of Variance (ANOVA) for Forbes' data.

follows the  $F_{1,n-2}$  distribution. If

$$F_0 > F_{\alpha,1,n-2}$$

we reject the null hypothesis  $H_0 : \beta_1 = 0$ . The rejection region is single-sided, due to that

$$E(MS_{Res}) = \sigma^2, \quad E(MS_R) = \sigma^2 + \beta_1^2 S_{xx},$$

that is, it is likely that the slope  $\beta_1 \neq 0$  if the observed value of  $F_0$  is large.

The analysis of variance is summarized in the following table.

The analysis of variance for Forbes' data is given in Table 2.

## 7 Coefficient of determination

The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

is called the coefficient of determination. For Forbes' data,

$$R^2 = \frac{425.63910}{427.79402} = 0.995,$$

and thus about 99.5% of the variability in the observed values is explained by boiling point.

In the below, we list some properties of  $R^2$ .

1. The range of  $R^2$  is  $0 \leq R^2 \leq 1$ . If all the  $\hat{\beta}_j$ 's were zero, except for  $\hat{\beta}_0$ ,  $R^2$  would be zero. (This event has probability zero for continuous data.) If all the  $y$ -values fell on the fitted surface, that is, if  $y_i = \hat{y}_i$ ,  $i = 1, 2, \dots, n$ , then  $R^2$  would be 1.
2. Adding a variable  $x$  to the model increases (cannot decrease) the value of  $R^2$ .
3.  $R^2$  is invariant to a scale change on  $x$  and  $y$ .
4.  $R^2$  does not measure the appropriateness of the linear model, for  $R^2$  will often be large even though  $y$  and  $x$  are nonlinearly related.

## 8 Interval estimation in simple linear regression

### 8.1 Confidence intervals on $\beta_0$ , $\beta_1$ and $\sigma^2$

The width of these confidence intervals is a measure of the overall quality of the regression line.

If the errors are normally and independently distributed, then the sampling distribution of both

$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \quad \text{and} \quad \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)}$$

is  $t$  with  $n - 2$  degrees of freedom. Therefore, a  $100(1 - \alpha)$  percent confidence interval on the slope  $\beta_1$  is given by

$$\hat{\beta}_1 - t_{\alpha/2, n-2} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} se(\hat{\beta}_1)$$

and a  $100(1 - \alpha)$  percent confidence interval on the intercept  $\beta_0$  is

$$\hat{\beta}_0 - t_{\alpha/2, n-2} se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2, n-2} se(\hat{\beta}_0)$$

**Frequency interpretation:** If we were to take repeated samples of the same size at the sample  $x$  levels and construct, for example, 95% confidence intervals on the slope for each sample, then 95% of those intervals will contain the true value of  $\beta_1$ .

For Forbes' data,  $se(\hat{\beta}_0) = 0.37903(1/17 + (202.95294)^2/530.78235)^{1/2} = 3.340$ , and  $se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{S_{xx}} = 0.0164$ . For a 90% confidence inter-

val,  $t(0.05, 15) = 1.753$ , and the interval is

$$-42.138 - 1.753(3.340) \leq \beta_0 \leq -42.138 + 1.753(3.340),$$

$$-47.993 \leq \beta_0 \leq -36.282.$$

A 95% confidence interval for the slope is

$$0.8995 - 2.131(0.0164) \leq \beta_1 \leq 0.8995 + 2.141(0.0164),$$

$$0.867 \leq \beta_1 \leq 0.930.$$

If the errors are normally and independently distributed, then the sampling distribution of

$$(n - 1)MS_{Res}/\sigma^2$$

is chi-square with  $(n - 2)$  degrees of freedom. Thus,

$$P\left\{\chi_{1-\alpha/2, n-2}^2 \leq \frac{(n - 2)MS_{Res}}{\sigma^2} \leq \chi_{\alpha/2, n-2}^2\right\} = 1 - \alpha$$

and consequently a  $100(1 - \alpha)$  percent confidence interval on  $\sigma^2$  is

$$\frac{(n - 2)MS_{Res}}{\chi_{\alpha/2, n-2}^2} \leq \sigma^2 \leq \frac{(n - 2)MS_{Res}}{\chi_{1-\alpha/2, n-2}^2}.$$

## 8.2 Interval estimation of the mean response

Let  $x_0$  be the level of the regressor variable for which we wish to estimate the mean response, say  $E(y|x_0)$ . We assume that  $x_0$  is any value of the regressor variable within the range of the original data on  $x$  used to fit the model. An unbiased point estimator of  $E(y|x_0)$  is

$$\widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

The variance of  $\hat{\mu}_{y|x_0}$  is

$$\text{Var}(\hat{\mu}_{y|x_0}) = \text{Var}[\bar{y} + \hat{\beta}_1(x_0 - \bar{x})] = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

since  $\text{cov}(\bar{y}, \hat{\beta}_1) = 0$ . Thus, the sampling distribution of

$$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MS_{Res}(1/n + (x_0 - \bar{x})^2/S_{xx})}}$$

is  $t$  with  $n-2$  degrees of freedom. Consequently, a  $100(1-\alpha)$  percent confidence interval on the mean response at the point  $x = x_0$  is

$$\begin{aligned} \hat{\mu}_{y|x_0} - t_{\alpha/2, n-2} \sqrt{MS_{Res}(1/n + (x_0 - \bar{x})^2/S_{xx})} &\leq E(y|x_0) \\ &\leq \hat{\mu}_{y|x_0} + t_{\alpha/2, n-2} \sqrt{MS_{Res}(1/n + (x_0 - \bar{x})^2/S_{xx})} \end{aligned} \quad (8)$$

## 9 Prediction of new observations

An important application of the regression model is prediction of new observations  $y$  corresponding to a specified level of the regressor variable  $x$ . If  $x_0$  is the value of the regressor variable of interest, then

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

is the point estimate of the new value of the response  $y_0$ . Now consider obtaining an interval estimate of this future observation  $y_0$ . The confidence interval on the mean response at  $x = x_0$  is inappropriate for this problem because it is an interval estimate on the mean of  $y$  (a parameter), not a probability statement about future observations from the distribution.

Let  $\psi = y_0 - \hat{y}_0$  is normally distributed with mean 0 and variance

$$Var(\psi) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right].$$

Thus, the  $100(1 - \alpha)\%$  percent prediction interval on a future observation at  $x_0$  is

$$\begin{aligned} \hat{y}_0 - t_{\alpha/2, n-2} \sqrt{MS_{Res} \left( 1 + 1/n + (x_0 - \bar{x})^2 / S_{xx} \right)} &\leq y_0 \\ &\leq \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{MS_{Res} \left( 1 + 1/n + (x_0 - \bar{x})^2 / S_{xx} \right)} \end{aligned} \quad (9)$$

By comparing (8) and (9), we observe that the prediction interval at  $x_0$  is always wider than the confidence interval at  $x_0$  because the

prediction interval depends on both the error from the fitted model and the error associated with future observations.

We may generalize (9) somewhat to find a  $100(1 - \alpha)$  percent prediction interval on the mean of  $m$  future observations on the response at  $x_0$ . The  $100(1 - \alpha)\%$  prediction interval on  $\bar{y}_0$  is

$$\begin{aligned} \hat{y}_0 - t_{\alpha/2, n-2} \sqrt{MS_{Res}(1/m + 1/n + (x_0 - \bar{x})^2/S_{xx})} &\leq \bar{y}_0 \\ &\leq \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{MS_{Res}(1/m + 1/n + (x_0 - \bar{x})^2/S_{xx})}. \end{aligned} \quad (10)$$

For prediction of  $100 \times \log(\text{Pressure})$  for a location with  $x_0 = 200$ , the point prediction is  $\hat{y}_0 = -42.13778 + 0.89549(200) = 136.961$ , with standard error of prediction

$$0.37903 \left( 1 + \frac{1}{17} + \frac{(200 - 202.95294)^2}{530.78235} \right)^{1/2} = 0.393.$$

Thus, a 99% predictive interval is

$$\begin{aligned} 136.961 - 2.95(0.393) &\leq \hat{y}_0 \leq 136.961 + 2.95(0.393), \\ 135.803 &\leq \hat{y}_0 \leq 138.119. \end{aligned}$$

A 99% predictive interval for Pressure is

$$10^{135.803/100} \leq \text{Pressure} \leq 10^{138.119/100}, \quad , i.e., 22.805 \leq \text{Pressure} \leq 24.$$

## 10 Model Diagnostic

### 10.1 High Influence points

The estimate of the regression slope can potentially be greatly affected by **high leverage points**. These are points that have very high or very low values of the independent variable—outliers in the  $x$  direction. They carry great weight in the estimate of the slope.

A high leverage point that also happens to correspond to a  $y$  outlier is a **high influence point**. It will alter the slope and twist the line badly. A point has high influence if omitting it from the data will cause the regression line to change substantially. To have high influence, a point must first have high leverage and, in addition, must fall outside the pattern of the remaining points.

Most computer programs that perform regression analyses will calculate one or another of several diagnostic measures of leverage and influence. We won't try to summarize all of these measures. We only note that very large values of any of these measures correspond to very high leverage or influence points.

## 10.2 Residuals

Plots of residuals versus other quantities are used to find failures of model assumptions. The most common plot, especially useful in simple regression, is the plot of residuals versus the fitted values.

- A null plot indicate no failure of assumptions.
- Curvature might indicate that the fitted mean function is inappropriate.
- Residuals that seem to increase or decrease in average magnitude with the fitted values might indicate nonconstant residual variance.
- A few relatively large residuals may be indicative of outliers, case for which the model is somehow inappropriate.

The plot of residuals versus fitted values for the pharmacy data is shown in Figure 4. This is a null plot.

The fitted values and residuals for Forbes' data are plotted in Figure 5. This plot indicates that case 12 is an outlier. Delete this point from the dataset. Refitting the model resulting in the following results (Table 3):

Table 3: Summary statistics for Forbes' data with all data and with case 12 deleted.

Quantity	All data	Delete case 12
$\hat{\beta}_0$	-42.138	-41.308
$\hat{\beta}_1$	0.895	0.891
$se(\hat{\beta}_0)$	3.340	1.001
$se(\hat{\beta}_1)$	0.016	0.005
$\hat{\sigma}$	0.379	0.113
$R^2$	0.995	1.000

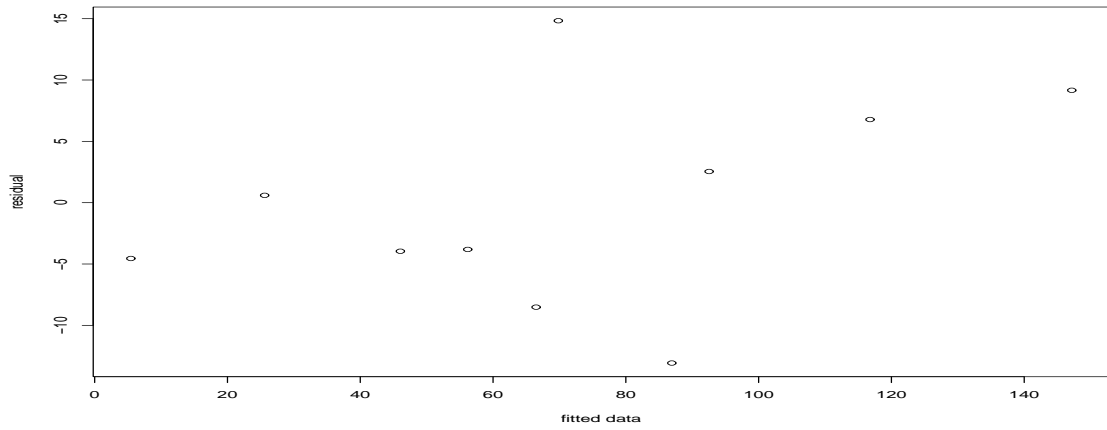


Figure 4: Residuals versus fitted values for the pharmacy data.

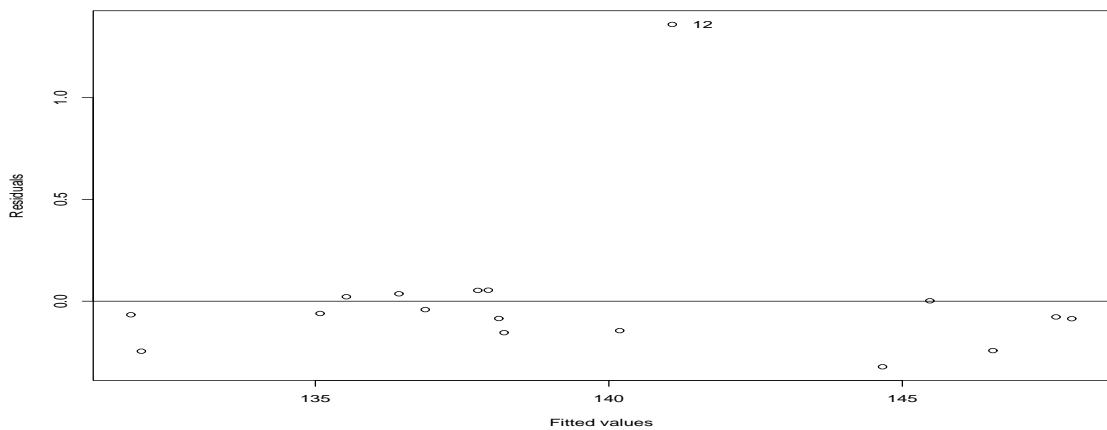


Figure 5: Residual plot for Forbes' data.

Some remarks on outliers:

- Cases with large residuals are candidates for outliers. Not all large residual cases are outliers, since large errors will occur with the frequency prescribed by the generating probability distribution. A simultaneous testing procedure is often used to protect against declaring too many cases to be outliers.
- Outlier identification is done relative to a specified model. If the form of the model is modified, the status of individual cases as outliers may change.
- Not all outliers are bad. For example, a geologist searching for oil deposit may be looking for outliers, if the oil is in places where the fitted model does not match the data.