Chapter 17

Inference for Regression

In this chapter we will discuss the following topics:

- We will consider inference for the regression parameters using the R-function **summary()**.
- We will find confidence intervals and Prediction intervals using the R-function **predict()**.
- We will consider model assumptions and check the conditions for inference by plotting the residuals and drawing a Normal QQ-plot of the residuals. We will use the R-function **qqnorm()** to perform a normal QQ-plot.

Suppose we have n distinct points (x, y), where x is the explanatory variable and y is the response variable. For a fixed value of x, we assume that y is Normally distributed with mean $\mu_y = \alpha + \beta x$ and variance σ^2. That is, if we take several measurements at the same x-value, the observations y will vary according to a Normal distribution, $N(\alpha + \beta x, \sigma^2)$.

The variance σ^2 is a measure of the variability of y about the population regression line

$$\mu_y = \alpha + \beta x.$$

The standard deviation is constant for each x. The errors ξ which is defined by $\xi = y - \mu_y$ is the difference between the observed values and the mean response or population regression line. The errors are Normally distributed with mean zero and variance σ^2, that is, ξ is $N(0, \sigma^2)$. We estimate α by $\hat{\alpha}$ and β by $\hat{\beta}$ by minimizing the sum of squared errors, where

$$\hat{\beta} = r\frac{s_y}{s_x} \text{ and } \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Here $\bar{y}$ and $\bar{x}$ are the means and s_y and s_x are the standard deviations of y and x, respectively, and r is their correlation. Thus, the population regression line is estimated by the least squares line,

$$\hat{y} = \hat{\alpha} + \hat{\beta}x.$$

Recall that we defined the residuals as

$$\text{residual} = y - \hat{y}$$

which is the difference between the observed and predicted value of y. We estimate the population standard deviation, σ, by the sample standard deviation of the residuals, $\hat{\sigma}$, which we call the regression standard error. The regression standard error is given by

$$\hat{\sigma} = \sqrt{\frac{1}{(n-2)}\sum(y - \hat{y})^2}. \tag{0.1}$$

Testing Significance of Regression

If the slope β of the population regression line is zero, there is no linear relationship between x and y. We wish to test the hypothesis that the slope is zero:

$$H_0 : \beta = 0.$$

The alternative hypotheses are $H_a : \beta > 0$, $H_a : \beta < 0$ or $H_a : \beta \neq 0$. We compute the test statistics

$$T = \frac{\hat{\beta}}{SE_{\hat{\beta}}},$$

where $SE_{\hat{\beta}}$ is the standard error of β given by

$$SE_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum (x - \bar{x})^2}}$$

and $\hat{\sigma}^2$ is given in (0.1). T has a t distribution with $(n-2)$ degrees of freedom. We reject the null hypothesis at the significance level γ if the p-value is smaller than γ. A level $100(1-\gamma)\%$ confidence level interval for the slope β is

$$\hat{\beta} \pm t^* SE_{\hat{\beta}}.$$

The critical value t^* is defined such that the area under the t-density curve with $(n-2)$ degrees of freedom is $(1-\gamma)$ between $-t^*$ and t^*.

Problem. Again we will consider the following data set [3] that shows the average January minimum temperature in degrees Fahrenheit with the latitude and longitude of 56 U.S. cities from 1931-1960.

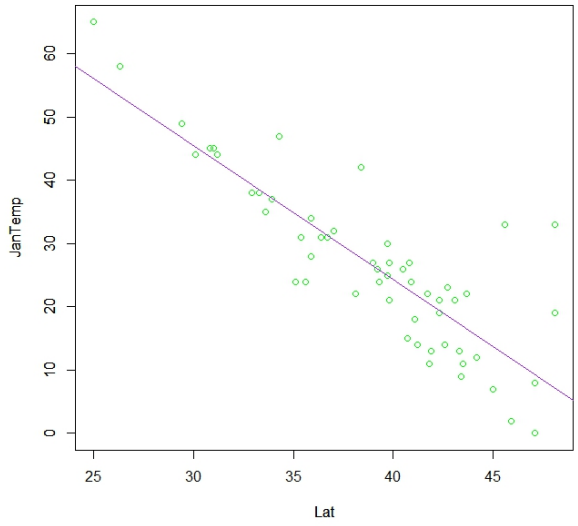
	City	JanTemp	Lat	Long
1	Mobile, AL	44	31.2	88.5
2	Montgomery, AL	38	32.9	86.8
3	Phoenix, AZ	35	33.6	112.5
4	Little Rock, AR	31	35.4	92.8
5	Los Angeles, CA	47	34.3	118.7
..	..	..	..	.. . ..
52	Seattle, WA	33	48.1	122.5
53	Spokane, WA	19	48.1	117.9
54	Madison, WI	9	43.4	90.2
55	Milwaukee, WI	13	43.3	88.1
56	Cheyenne, WY	14	41.2	104.9

Do cities with higher latitudes tend to have lower average January minimum temperature?
(a) Make a scatterplot of the data. Find the regression line and add it to the plot. (This was done in chapter 5).
(b) Find the squared correlation, r^2, also the called coefficient of determination, and test for significance of the slope. Determine a 95% confidence interval for the population slope, β.

Solution to part (a). The explanatory variable, x, is the latitude and the response variable y is the average January minimum temperature.

```
> table=read.delim("US_Temp.txt")
> attach(table)
> plot(Lat,JanTemp,col="green")
> temp.lm=lm(JanTemp~Lat)
> temp.lm

Call:
lm(formula = JanTemp ~ Lat, data = table)

Coefficients:
(Intercept)          Lat
     108.73        -2.11
> abline(temp.lm,col="purple")
```

We obtain the regression line $\hat{y} = 108.73 - 2.11x$. We can see from the scatterplot that there is an approximately linear relationship between x and y. Thus, cities with higher latitudes tend to have lower average January minimum temperature except from the coast cities for which the ocean moderates the temperature.



Solution to part (b). We will test the null hypothesis H_0 that the slope of the regression line is zero. We will do a summary statistics in R and we compute the critical value of the t distribution with 54 degrees of freedom ($n = 56$ and df=n-2=54):

```
> summary(temp.lm)
```

```
Call:
lm(formula = JanTemp ~ Lat, data = table)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6812  -4.5018  -0.2593   2.2489  25.7434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.7277     7.0561   15.41   <2e-16 ***
Lat          -2.1096     0.1794  -11.76   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.156 on 54 degrees of freedom
Multiple R-squared:  0.7192,    Adjusted R-squared:  0.714
F-statistic: 138.3 on 1 and 54 DF,  p-value: < 2.2e-16
> qt(0.975,54)
[1] 2.004879
```

We obtain that $r^2 = 0.7192$ so 71.92% of the variation in average January minimum temperature is explained by latitude. Since the p-value is $< 2 \times 10^{-16}$, we reject the null hypothesis that the slope of the regression line is zero. Thus, there is a significant linear relationship between latitude and January minimum temperature. The critical value is $t^* = 2.004879$ and the 95% confidence interval for the population slope β is

$$\hat{\beta} \pm t^* SE_{\hat{\beta}} = -2.1096 \pm (2.0049)(0.1794) = -2.1096 \pm 0.3597.$$

Thus, we are 95% confident that the population slope β is in the interval $(-2.4693, -1.7499)$. Notice here that the t-statistics is given by

$$\frac{\hat{\beta}}{SE_{\hat{\beta}}} = \frac{-2.1096}{0.1794} = -11.76$$

as given in the summary output.

Explanation. The command **summary(temp.lm)** gives the summary statistics of the least squares regression line *temp.lm*. The outputs are:

- **Residuals**: which is the five number summary of the residuals.
- **Coefficients**: which show the estimates of the intercept and slope with their respective standard errors, observed values of the t-statistics, and p-values.
- **Residual standard error**: which is the regression standard error given in (0.1).
- **Multiple R-squared**: which is the squared correlation, r^2.
- We will not cover **Adjusted R-squared** in this course.

- The last line is a F-test for the null hypothesis that the slope is zero. It returns the exact same value as the squared value of the t-statistics for degrees of freedom equal to one, i.e $(-11.76)^2 = 138.3$. The F-test is not needed for simple linear regression since all the information it provides is given through the t-test. However, the F-test is important for multiple linear regression.

Model assumptions
The following model assumptions need to be satisfied in order to obtain a valid prediction $\hat{y}$ from x: [2]

- There is a linear relationship or an approximately linear relationship between the response y and the regressors x.
- The random error, ϵ, has mean zero.
- The random error, ϵ, has constant variance σ^2.
- The errors are uncorrelated.
- The errors follow a Normal distribution.

The latter condition is necessary for hypothesis testing and the estimate of confidence intervals. One way to diagnose violations of the regression assumptions is to study the **residuals**.

Residuals

- The residuals measure the variation in the response variable not explained by the regression model.
- We can also consider the residuals as the observations of the model errors. Thus, if the model errors do not follow the model assumptions as described above, it will be shown in the residuals.
- In this course, we will analyze residual by plotting them against the regressor x or the predictor $\hat{y}$. If the variance is not constant, the residuals will not be equally scattered above or below the line $y = 0$. A curved residual plot indicates non-linearity.
- It is useful to scale the residuals to locate observations that are outliers. One way to do this is to standardize the residuals so that the resulting residuals have mean zero and variance approximately one. A large value of the standardized residual, say greater than 3, indicates a potential outlier.
- One way to check the normality assumption is to draw a normal probability plot of the residuals. The cumulative normal distribution will be plotted as a straight line and the residuals should lie approximately on the straight line. If the residuals do not lie approximately on the straight line, it indicates that the underlying distribution is not normal.

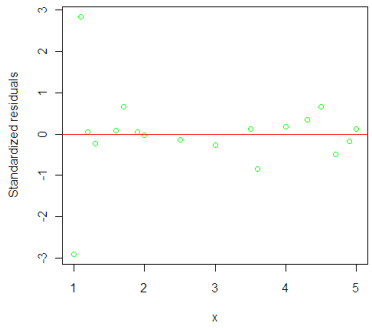


Figure 1: The residuals can be contained in a horizontal band which indicates no model defects.

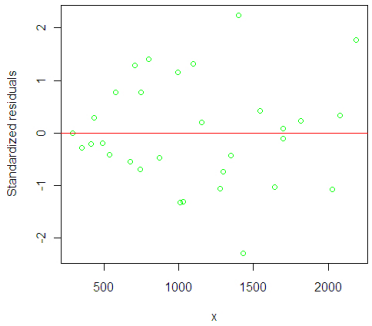


Figure 2: Indicates that the variance is not constant.

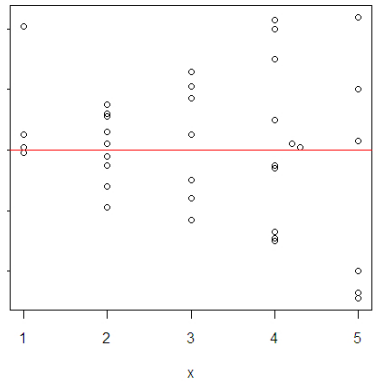


Figure 3: The outward-opening funnel or cone indicates that the variance increases as y increases.

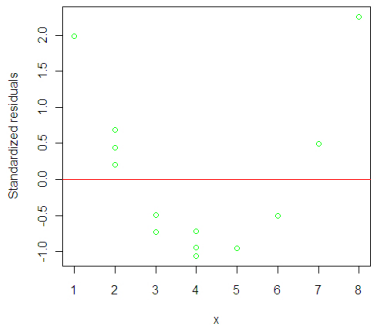


Figure 4: Curvature, indicates non-linearity.

Problem. We will return to the previous problem regarding the relationship between latitude and January temperature. Examine the conditions for regression inference.

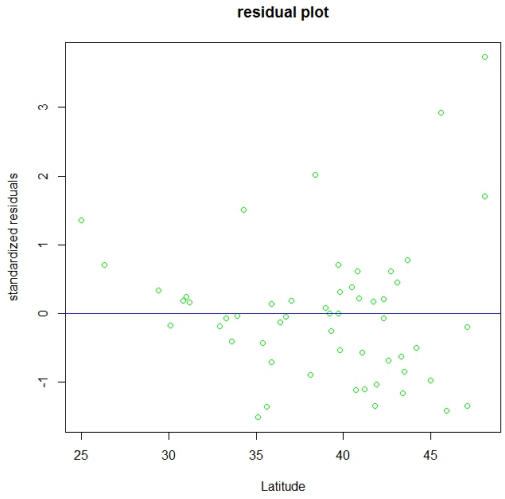
Solution. We will examine the conditions for inference by looking at the residuals. We will standardize the residuals and draw a residual plot, a stem plot, and a Normal QQ-plot of the residuals:

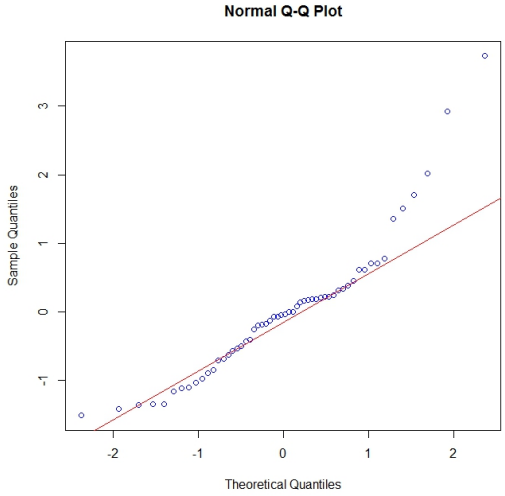
```
> res=rstandard(temp.lm)
> plot(Lat,res,xlab="Latitude",ylab="standardized residuals",
+ main="residual plot",col="green")
abline(0,0,col="blue")
> stem(res)

  The decimal point is at the |

  -1 | 5443321100
  -0 | 98776655443222111000
   0 | 01122222222334566778
   1 | 457
   2 | 09
   3 | 7
```

Here is the command for the Normal QQ-plot of the residuals:

```
> qqnorm(res,col="blue")
> qqline(res,col="red")
```

From the stem plot and the QQ-Normal plot, we see departure from Normality when we include the coast cities. The distribution is skewed. If we exclude the coast cities we do not see any departure from Normality. Again the cost cities are potential outliers, in particular observation 52. The coast cities are responsible for the spread for large and small values of latitudes. If we exclude the coast cities, the residuals are contained in a horizontal band which indicates that the model assumptions are satisfied.





Explanation. The code can be explained as follows:

- The command **qqnorm(res)** returns a QQ-plot which is used to assess whether the residuals comes from a Normal distribution. The observations are ordered and the sample quantiles are plotted against the quantiles of the standard normal distribution. If the plots falls along a straight line, we can conclude that the sample is likely coming from a Normal distribution.

- The command **qqline(res)** draws the cumulative normal distribution as a straight line.

Notice that if you plot average January minimum temperature against longitude, the relationship will not be linear. However, by using multiple regression techniques, it can be shown that both latitude and longitude are significant variables in predicting January temperature. We will return to multiple regression in chapter 20.

Inference about Prediction
We will consider two types of intervals; prediction interval and confidence interval.

- To predict the mean response for a given value x^* of x, we construct a confidence interval for the mean response $\mu_y = \alpha + \beta x^*$.
- To predict an individual response for a given value x^* of x, we construct a prediction interval. A point estimate of the observation y at x^* is given by

$$\hat{y} = \hat{\alpha} + \hat{\beta} x^*.$$

  The value of $\hat{y}$ will change since y changes for different observations of the same x-value x^*.

The margin or error is larger for the prediction interval since individual variation is larger than mean variation. Thus, the prediction interval will be wider than the confidence interval.

A level $100(1-\gamma)\%$ confidence interval for the mean response μ_y when the value of x is x^* is

$$\hat{y} \pm t^* SE_{\hat{\mu}},$$

where the standard error $SE_{\hat{\mu}}$ is given by

$$SE_{\hat{\mu}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}}.$$

A level $100(1-\gamma)\%$ prediction interval for a observation y when the value of x is x^* is

$$\hat{y} \pm t^* SE_{\hat{y}},$$

where the standard error $SE_{\hat{y}}$ is given by

$$SE_{\hat{y}} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}}.$$

The critical value t^* is defined such that the area under the t-density curve, $t(n-2)$, between $-t^*$ and t^* is $(1-\gamma)$.

Problem. A 1965 study [1] discussed the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women. The following data [1], [4] shows the mean annual temperature in degrees Fahrenheit and the Mortality Index for neoplasms of the female breast, where data were obtained from certain regions of Great Britain, Norway, and Sweden:

Mortality	Temperature
102.5	51.3
104.5	49.9

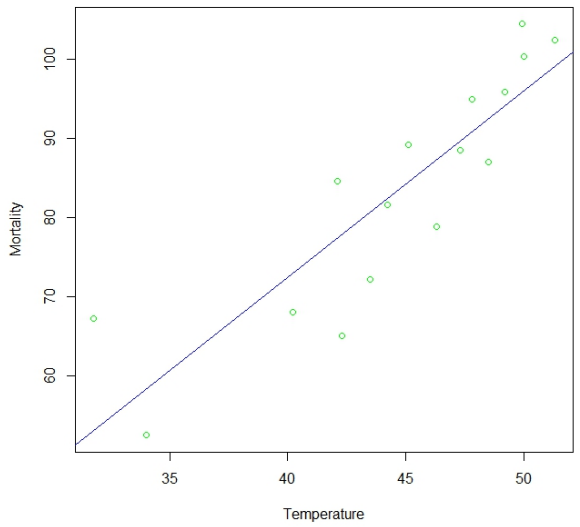
```
100.4     50
95.9     49.2
87       48.5
95       47.8
88.6     47.3
89.2     45.1
78.9     46.3
84.6     42.1
81.7     44.2
72.2     43.5
65.1     42.3
68.1     40.2
67.3     31.8
52.5     34
```

(a) Analyze the relationship between mean annual temperature and the mortality rate for a certain type of breast cancer. Determine if there are any outliers in the data.
(b) We might want to predict the mean Mortality Index for neoplasms with a mean annual temperature of 90.
What is the 95% interval we wish to find?
(c) Find the 95% confidence interval for $x^* = 90$ without the outlier(s).
Find the 95% prediction interval for $x^* = 90$ without the outlier(s).

Solution to part (a). The explanatory variable is the mean annual temperature and the response variable is the Mortality Index for neoplasms. We obtain:

```
> table=read.delim("Breast_cancer.txt")
> attach(table)
> plot(Temperature,Mortality,col="green")
> g=lm(Mortality~Temperature)
> abline(g,col="blue")

> summary(g)

Call:
lm(formula = Mortality ~ Temperature)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8358  -5.6319   0.4904   4.3981  14.1200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.7947    15.6719  -1.391    0.186
Temperature   2.3577     0.3489   6.758  9.2e-06 ***
```

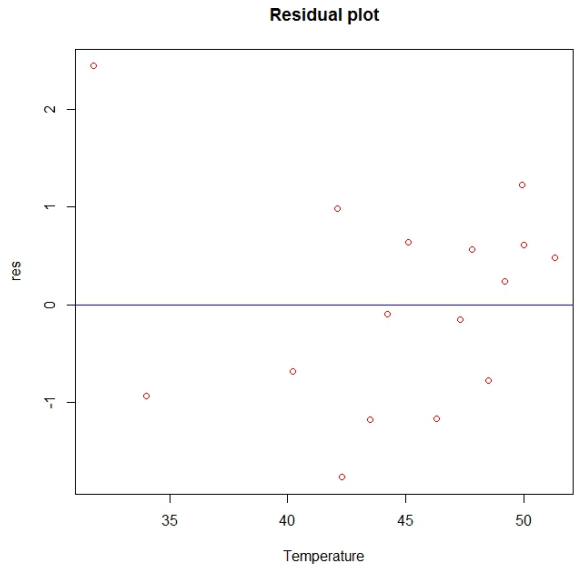
```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.545 on 14 degrees of freedom
Multiple R-squared: 0.7654,     Adjusted R-squared: 0.7486
F-statistic: 45.67 on 1 and 14 DF,  p-value: 9.202e-06
```



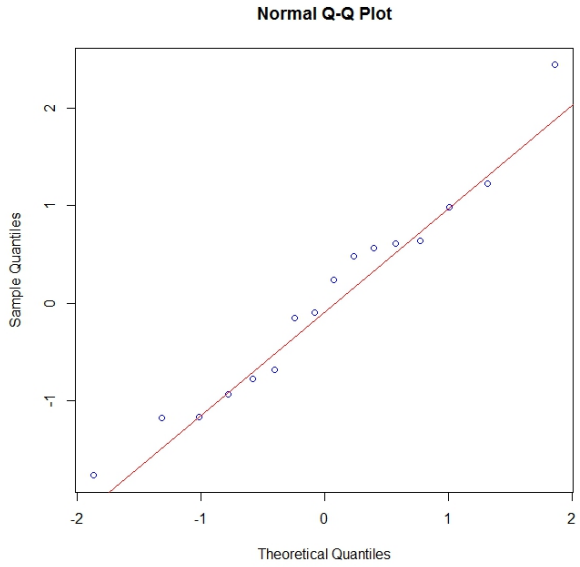
The equation of the regression line is $\hat{y} = -21.7947 + 2.3577x$ and $r^2 = 0.7654$, so 76.54% of the variation in the data can be explained by the regression line. The p-value for testing significance of the slope is < 0.05, so we reject the null hypothesis that the slope is zero at the 5% significant level. Observation number 15 is a potential outlier. Next we will check the residuals and the conditions for inference:

```
> res=rstandard(g)
> plot(Temperature,res,main="Residual plot",col="red")
> abline(0,0,col="blue")
> stem(res)

  The decimal point is at the |

  -1 | 822
  -0 | 98721
   0 | 25666
   1 | 02
   2 | 4
```

```
> qqnorm(res,col="blue")
> qqline(res,col="red")
```



The residuals are contained in a horizontal band except from a large value in observation number 15. From the Normal QQ-plot, we do not see any deviance from the Normal distribution. Hence the model assumptions for inference are satisfied. Next we will



find the regression line for the data without the outlier and plot both lines together:

```
> g2=lm(Mortality~Temperature,subset=-15)
> summary(g2)
```

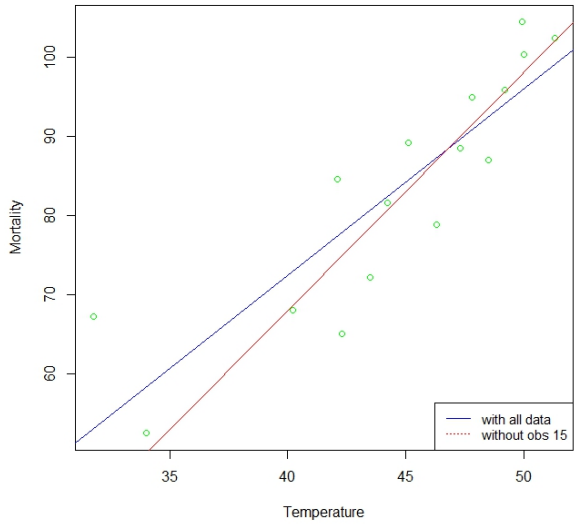
```
Call:
lm(formula = Mortality ~ Temperature, subset = -15)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8255 -3.8726  0.4376  3.0458 10.2776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.6181    15.8239  -3.325  0.00548 **
Temperature   3.0152     0.3466   8.701 8.82e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.933 on 13 degrees of freedom
Multiple R-squared: 0.8534,     Adjusted R-squared: 0.8422
F-statistic:  75.7 on 1 and 13 DF,  p-value: 8.816e-07

> plot(Temperature,Mortality,col="green")
> abline(g,col="blue")
> abline(g2,col="red")
> legend("bottomright",c("with all data","without obs 15"),
+ cex=0.9,lty=c(1,3),col=c("blue","red"))
```

The regression line without observation number 15 is $\hat{y} = -52.6181 + 3.0152x$ with $r^2 = 0.8534$. The slope is significantly different from zero. Omitting observation number 15 resulted in a greater value for r^2 and hence, a greater proportion of the variation in the data can be explained by the regression line. Thus, there is a strong linear relationship between mean annual temperature and the mortality rate for a certain type of breast cancer.

Solution to part (b). The problem asks us to find a 95% interval when the explanatory variable is $x^* = 90$. If we would like to predict the mean Mortality Index for neoplasms for all the persons in the study, we would construct a confidence interval for the mean. If we rather would like to predict the Mortality Index for neoplasms for one individual for $x^* = 90$, we will construct a prediction interval. Since we would like to predict the mean Mortality Index for neoplasms we will want to construct the 95% confidence interval.

Solution to part (c). We obtain:

```
> newdata=data.frame(Temperature=90)
> predict(g2,newdata,interval="confidence")
       fit      lwr     upr
1 218.7512 185.2313 252.271
```




```
> predict(g2,newdata,interval="predict")
       fit      lwr      upr
1 218.7512 182.8646 254.6377
>
```

The 95% confidence interval is $(185.23, 252.27)$ and the 95% prediction interval is $(182.86, 254.64)$.

Explanation. The code can be explained as follows:

- In the argument for the function **predict**, we can choose the option **interval=" confidence"** for confidence interval or the option **interval=" predict"** for prediction interval. Note that we must put the observations under consideration in a data frame.

References

[1] A.J. Lea. *New Observations on Distribution of Neoplasms of Female Breast in Certain Countries.* British Medical Journal, 1, 488-490, 1965

[2] D. C. Montgomery, E. A. Peck, G. G. Vining. *Introduction to Linear Regression Analysis*, Fourth edition, Wiley Series in Probability and Statistics, John Wiley & Sons, INC., New Jersey, 2006

[3] J. L. Peixoto. *A property of well-formulated polynomial regression models.* American Statistician, 44, 26-30, 1990.
Also found in: D. J. Hand, et al. *A Handbook of Small Data Sets*, London:

Chapman & Hall, 208-210, 1994.
The data can also be found on DASL at http://lib.stat.cmu.edu/DASL/DataArchive.html.

[4] P. F. Velleman, D. C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis.* Belmont. CA: Wadsworth, Inc., pp. 127-134, 1981
The data can also be found on DASL at http://lib.stat.cmu.edu/DASL/DataArchive.html.

00