Chapter 16
Simple Linear Regression

In this chapter we will discuss the following topics:

- How to find the least-squares regression line using the R-function **lm()** (*l*inear *m*odel).
- How to plot the regression line using the R-function **abline()**.
- How to calculate and plot the residuals using the R-function **resid()**.

The Least-Squares Regression Line

We use the regression line to describe the relationship between two variables where one variable (the explanatory variable) helps explain or predict the other (the response variable). The least-squares regression line is given by the equation

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

where $\hat{\alpha}$ is the intercept and $\hat{\beta}$ is the slope. The function $\hat{y}$ gives the predicted response for the explanatory variable x, where y is the observed response of x. The method of least squares computes the parameters $\hat{\alpha}$ and $\hat{\beta}$ by minimizing the sum of squared errors. The parameters are given by

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

where $\bar{y}$ 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.

Problem. Find the equation of the least-squares regression line with explanatory variable x and response variable y, where x and y are given by

```
 x  y
 2  1
 3  3
 5  6
 7  5
 8  7
10  8
13 10
```

Solution. We have

```
> x=c(2,3,5,7,8,10,13)
> y=c(1,3,6,5,7,8,10)
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.6038       0.7453
```

The equation of the regression line is: $y = 0.6038 + 0.7453x$.

Explanation. The explanation of the code is as follows:

- The symbol $\sim$ in the model formula for **lm** should be read as *described by*. The format for the command is, **lm(response variable $\sim$ explanatory variable)**.

Problem. In the next five problems we will consider the following data set [1] 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
```

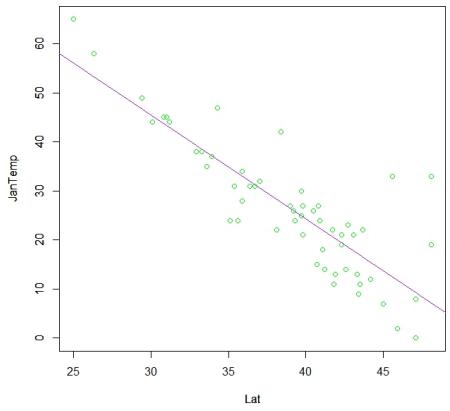
Make a scatterplot of the data and find the equation of the least-squares regression line for predicting January temperatures from changes in latitude. Draw the regression line on the same plot. Compute the correlation coefficient between latitude and average January minimum temperature.

Solution. The explanatory variable is *latitude* and the response variable is *January temperature*. We first import the data in a table which we name **table** and then make a scatterplot and compute the correlation:

```
> table=read.delim("US_Temp.txt")
> attach(table)
> plot(Lat,JanTemp,col="green")
> cor(Lat,JanTemp)
[1] -0.8480352
```

We can see from the scatterplot and the correlation coefficient ($r = -0.848$), that there is a strong negative association between latitude and temperature except for the coasts where the temperatures are moderate due to the ocean. From the scatter plot and the correlation coefficient, it looks like there is approximately a linear relationship between these two variables. We next find the least squares regression line and add the regression line to the plot:

```
> temp.lm=lm(JanTemp~Lat)
> temp.lm

Call:
```




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

Coefficients:
(Intercept)          Lat
     108.73        -2.11

> 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

> abline(temp.lm,col="purple")
```

The equation of the regression line is: JanTemp $= 108.73 - 2.11\times$ latitude.

Explanation. The explanation of the code is as follows:

- We name the regression line **temp.lm**.

- The command **summary(temp.lm)** is used for inference for the regression parameters and will be explained in chapter 17.
- The command **abline(temp.lm)** adds the regression line to the scatterplot.
- Notice that the **plot()** function must be called before the **abline()** function.

Residuals

A residual is defined as

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

which is the difference between an observed value y of the response variable and the predicted value $\hat{y}$ given by the regression line.

Problem. We will consider the previous problem, where the latitude was measured along with the average January minimum temperature for 56 U.S. cities from 1931-1960. Compute the residuals and plot the residuals against the explanatory variable, latitude.

Solution. The residuals are given by

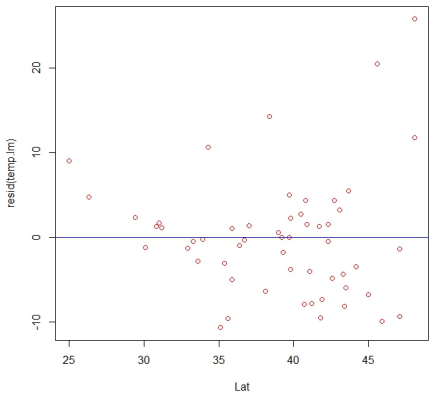
```
> res=resid(temp.lm)
> res
          1           2           3           4           5           6
 1.09139870 -1.32230196 -2.84559046 -3.04833234 10.63112103 14.28043121
          7           8           9          10          11          12
-7.86751674  1.24207111  2.71056569  5.02289541  1.66948113  9.01195404
         13          14          15          16          17          18
 4.75441825 -0.21271411  5.46124680 -0.49217618 -3.76614580 -9.54697011
         19          20          21          22          23          24
-6.35244515  0.54618392  1.24756356 -3.48395927  0.02289541  4.35165895
         25          26          27          28          29          30
 3.19549409 -9.89765993 -1.82093973 -1.36615451 -7.33601132 -5.96067077
         31          32          33          34          35          36
 2.23385420 -10.68120869 -4.85929983  4.34344204  1.00646159 -0.93874449
         37          38          39          40          41          42
-9.36615451 -0.03189851  1.50782382 -4.99353841 20.46946371  1.55440083
         43          44          45          46          47          48
 1.55440083 -0.47846682 -0.30586813 -9.62641477  2.29414058 -1.22914793
         49          50          51          52          53          54
-4.02368160 -6.79628899  1.32700822 25.74343333 11.74343333 -8.17162955
         55          56
-4.38258834 -7.81272282
```

Next we plot the residuals against the explanatory variable, latitude.

```
> plot(Lat,resid(temp.lm),col="red")
> abline(0,0,col="blue")
```

We will analyze residual plots more carefully in chapter 17.

Explanation. The code can be explained as follows:



- The function **resid(temp.lm)**, where the argument is the **lm** function, returns the residuals.
- Adding the command **abline(0,0)** results in the drawing of the line $y = 0$ in the same plot as the residual plot.

Problem. Compute the correlation coefficient with observation number 52 deleted from the observations in the previous problem.

Solution. We are deleting the outlier with observation number 52. We obtain

```
> table_New=table[c(-52),]
> JanTemp_New=table_New$JanTemp
> Lat_New=table_New$Lat
> cor(JanTemp_New,Lat_New)
[1] -0.8891845
```

The new correlation coefficient is -0.89 which is a little smaller than the previous one so it shows a stronger association between latitude and January temperature when we remove the outlier.

Explanation. The code can be explained as follows:

- The command **table**$[c(-52),]$ removes rows 52 in **table**.
- We rename the table and the variables.

Problem. Find the least-squares regression line for predicting January temperatures from changes in latitude without observation 52. Add both lines to the scatterplot.

Solution. We have

```
> temp.lm2=lm(JanTemp~Lat,subset=c(-52))
> temp.lm2

Call:
```

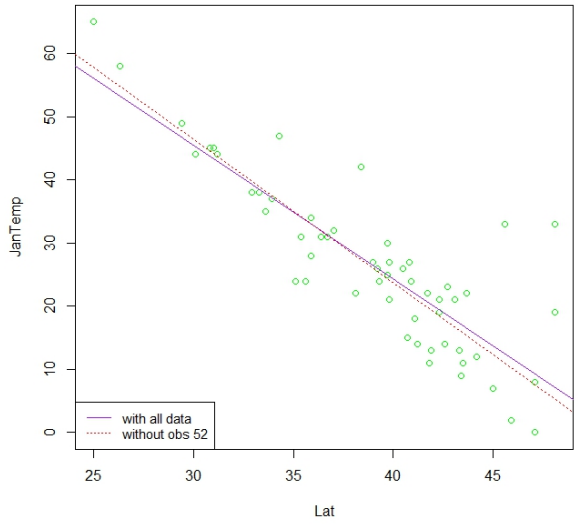
```
lm(formula = JanTemp ~ Lat, subset = c(-52))

Coefficients:
(Intercept)          Lat
    114.425       -2.268
```

so the regression line without observation number 52 is

$$JanTemp = 114.425 - 2.268 \times latitude.$$

We plot the lines in the scatterplot:

```
> plot(Lat,JanTemp,col="green")
> abline(temp.lm,col="purple")
> abline(temp.lm2,col="red",lty=3)
> legend("bottomleft",c("with all data","without obs 52"),
+ cex=0.9,lty=c(1,3),col=c("purple","red"))
```



Explanation. The code can be explained as follows:

- In order to calculate the regression line for the data without observation 52, we added the entry **subset=c(-52)** in the argument for **lm()**. The negative number removes the observation indexed 52.
- We plot the regression line without observation 52, which we named **temp.lm2**, with type **lty=3**.

Problem. Use both lines to predict the January temperature at latitude 49 degrees north of the equator.

Solution. We have

```
> newdata=data.frame(Lat=49)
> predict(temp.lm,newdata)
       1
5.357938
> predict(temp.lm2,newdata)
       1
3.269773
```

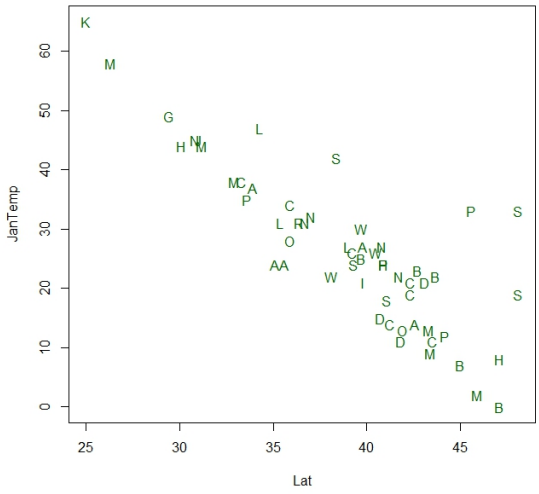
so with all the data, the predicted temperature is 5.4 Fahrenheit degrees and without observation 52 the temperature is 3.3 Fahrenheit degrees at latitude 49 degrees north of the equator.

Explanation. The explanation of the code is as follows:

- The command **newdata=data.frame(Lat=49)** puts the value 49 of the explanatory variable *Lat* in a data frame which we store as **newdata**.
- The command **predict(temp.lm, newdata)** returns the predicted values for the data stored in the data frame **newdata**.

To see the first initials of the cities in the plot, do the following:

```
> plot(Lat,JanTemp,col="darkgreen",pch=as.character(City))
```



References

[1] 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.

00