Visva Bharati University

EVEN SEMESTER, 2023-24

STATISTICS

Regression Techniques

Practical Notebook

SNEHANGSHU BHUIN

M.Sc. Sem 2

5 August 2024

1 1) SIMPLE LINEAR REGRESSION

1) Table B. 1 gives data concerning the performance of the 26 National Football League teams in 1976. It is suspected that the number of yards gained rushing by opponents (x 8) has an effect on the number of games won by a team (y).

a. Fit a simple linear regression model relating games won y to yards gained rushing by opponents x8 .
b. Construct the analysis - of - variance table and test for significance of regression.
c. Find a 95%Cl on the slope.
d. What percent of the total variability in y is explained by this model?
e. Find a 95%Cl on the mean number of games won if opponents' yards rushing is limited to **2000** yards.

TABLE B. 1 National Football League 1976 Team Performance

[3]:
```
# Load the necessary library
library("MPV")

# Attach the dataset for easier referencing
attach(table.b1)
```

```
The following objects are masked from table.b1 (pos = 3):

    x1, x2, x3, x4, x5, x6, x7, x8, x9, y
```

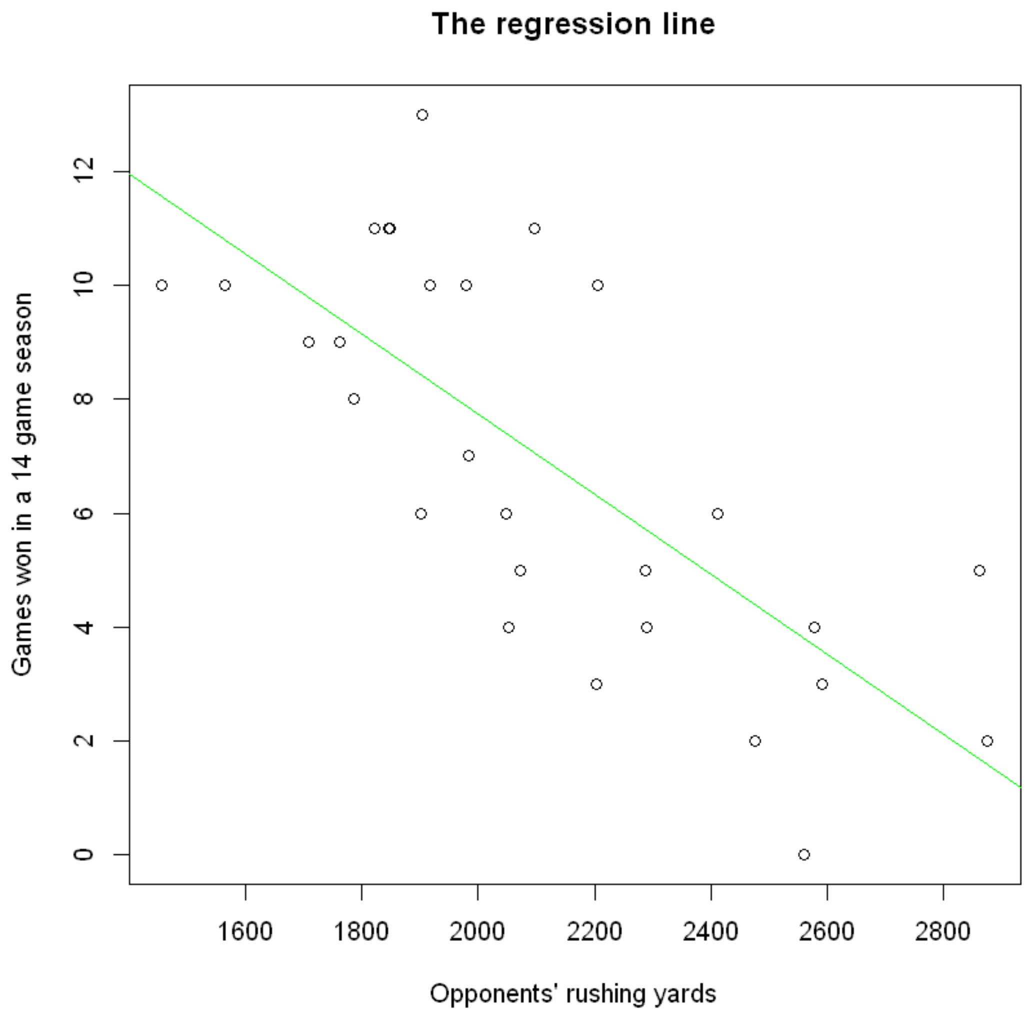
[4]:
```
# Fit a linear model predicting y based on x8
y.lm <- lm(y ~ x8, data = table.b1)

# Summarize the linear model to view key statistics
summary(y.lm)
```

```
Call:
lm(formula = y ~ x8, data = table.b1)

Residuals:
   Min     1Q Median     3Q    Max 
-3.804 -1.591 -0.647  2.032  4.580 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.788251   2.696233   8.081 1.46e-08 ***
x8          -0.007025   0.001260  -5.577 7.38e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.393 on 26 degrees of freedom
Multiple R-squared:  0.5447,	Adjusted R-squared:  0.5272 
F-statistic:  31.1 on 1 and 26 DF,  p-value: 7.381e-06
```

[5]:

```r
# Plot the data with the regression line
plot(table.b1$x8, table.b1$y,
     main = "The regression line",
     xlab = "Opponents' rushing yards",
     ylab = "Games won in a 14 game season")

# Add the regression line to the plot
abline(y.lm, col = "green")
```



[15]:

```r
# Perform the ANOVA
anova_table <- anova(y.lm)
print(anova_table)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x8         1 178.09 178.092  31.103 7.381e-06 ***
Residuals 26 148.87   5.726
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[16]:

```r
# Calculate the 95% confidence interval for the coefficient of x8
conf_interval <- confint(y.lm, 'x8', level = 0.95)
print(conf_interval)
```

```
          2.5 %       97.5 %
x8 -0.009614347 -0.004435854
```

[17]:

```r
# Create a new data frame for prediction
new.dat <- data.frame(x8 = 2000)

# Predict the response for the new data point with confidence interval
predictions <- predict(y.lm, newdata = new.dat, interval = 'confidence')
print(predictions)
```

```
      fit      lwr      upr
1 7.73805 6.765753 8.710348
```

■ **Interpretation:-**

a. In Table B.1, y denotes Games won in a 14 game season and $x8$ denotes Opponents' rushing yards.

According to the question y is the dependent variable and $x8$ is the independent variable. In R programming " lm() " function can fit the linear regression model. So by running the code " $y \cdot \text{Im} = \text{Im}(y \sim x8$, data=table.b1)" and "summary(y.lm)" we get the simple linear regression model as—-

$$\hat{y} = 21.788251 - 0.007025x_8$$

b. In R analysis of variance can be constructed using "anova()" function. So, by the code "anova(y.Im)" the table is as follows—-

Analysis of Variance Table

Response: y

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	8	1178.09	178.092	31.103	7.381e-06

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	26	148.87	5.726		

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

Testing significance of regression—-

The hypotheses for testing significance of regression is $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$. Failing to reject $\mathrm{H}_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y. Alternatively, if $\mathrm{H}_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$t_0 = \frac{\widehat{\beta_1}}{\operatorname{se}\left(\widehat{\beta_1}\right)}$$

The null hypothesis of significance of regression would be rejected if $|t_0| > t_{\alpha/2,n-2}$.

Here, $\widehat{\beta_1} = -0.007025$ and $\operatorname{se}\left(\widehat{\beta_1}\right) = 0.001260$. so, $\mathrm{t}_0 = -5.5754 \& \mathrm{t}_{0.025,26} = 2.055529$. So, the null hypothesis is rejected since $|t_0| > t_{0.025,26}$. That means, $x8$ is of value in explaining the variability in y.

c. The $100(1-\alpha)$ percent Cl on β_1 is $\widehat{\beta_1} - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{MS_{Res}}{\sum_{i=1}^{n} x_i^2}} \leq \beta_1 \leq \widehat{\beta_1} + t_{\frac{\alpha}{2}}, n-1\sqrt{\frac{MS_{Res}}{\sum_{i=1}^{n} x_i^2}}$

The confidence interval can be found in R programming by using "confint()" function. So by the code "confint(y.Im,'x8',level=0.95)" the 95% confidence interval of slope is $(-0.0096, -0.0044)$

d. Coefficient of determination is the proportion of the variation in the response variable that is explained by the regression model.

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

By the output of regression model in R, we get that $R^2 = 0.5447$ that means 54.5% of the variation in y is explained by the regression model.

e. A 100(1- α) percent Cl on the mean response at the point $x = x_0$ I

$$\widehat{\mu_{y|x0}} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq E(y \mid x_0) \leq \widehat{\mu_{y|x0}} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Where, $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$ and $E\widehat{(y \mid x_0)} = \widehat{\mu_{y|x0}} = \widehat{\beta_0} + \widehat{\beta_1} x_0$

To find the confidence interval in R, we create a new dataframe with the desired value i.e., new.dat=data.frame ($x8 = 2000$)

Then the prediction is made by predict() function and 'interval' argument is set to 'confidence' to output the mean interval i.e., predict(y.lm,newdata=new.dat,interval='confidence')

So the 95% confidence interval on the mean number of games won if opponents ' yards rushing is limited to 2000 yards is $(6.765753, 8.710348)$

2) Suppose we would like to use the model developed in Problem 2.1 to predict the number of games a 4 team will win if it can limit opponents ' yards rushing to 1800 yards. Find a point estimate of the number of games won when $x8 = 1800$. Find a 90% prediction interval on the number of games won.

```
[19]: # Create a new data frame for prediction
new.data <- data.frame(x8 = 1800)
# Predict the response for the new data point with a 90% prediction␣
 ↪interval
predictions <- predict(y.lm, newdata = new.data, interval = 'predict',␣
 ↪level = 0.90)
print(predictions)
```

```
      fit      lwr      upr
1 9.14307 4.936392 13.34975
```

■ **Interpretation:-**

To find the point estimate of the number of games won when $x8 = 1800$ we simply put $x8 = 1800$ in the

$$\begin{aligned} \hat{y} &= 21.788251 - 0.007025x_8 \\ \text{regression model} \qquad &= 21.788251 - (0.007025{*}1800) \\ &= 9.143251 \end{aligned}$$

Which also we've got from the output of predict function. So the fitted value is 9.14 .

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

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

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. We now develop a prediction interval for the future observation y0. Note that the random variable $\Psi = y_0 - \widehat{y_0}$ is normally distributed with mean

zero and variance $\operatorname{Var}(\Psi) = \operatorname{Var}\left(y_0 - \widehat{y_0}\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]$

because the future observation y_0 is independent of $\widehat{y_0}$. If we use $\widehat{y_0}$ to predict y 0 , then the standard error of $\Psi = y_0 - \widehat{y_0}$ is the appropriate statistic on which to base a prediction interval. Thus, the 100(1- 1) percent prediction interval on a future observation at x_0 is

$$\widehat{y_0} - t_{\frac{}{2}} \cdot \frac{\alpha}{n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \widehat{y_0} + t_{\frac{\alpha}{2}, n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

To find the prediction interval in R, we create a new data frame with the desired value to predict. The prediction is made with the predict() function. The interval argument is given 'prediction' since prediction interval is required and level is 0.90 since 90% prediction interval is required. So the 90% prediction interval on the number of games won is $(4.936392, 13.34975)$.

3) Table B. 2 presents data collected during a solar energy project at Georgia Tech.

a. Fit a simple linear regression model relating total heat flux y (kilowatts) to the radial deflection of the deflected rays $x4$ (milliradians).

b. Construct the analysis - of - variance table and test for significance of regression.

c. Find a 99%Cl on the slope.

d. Calculate R2 .

e. Find a 95%CI on the mean heat flux when the radial deflection is **16.5** milliradians.

TABLE B. 2 Solar Thermal Energy Test Data

y	x_1	x_2	x_3	x_4	x_5
271.8	783.35	33.53	40.55	16.66	13.20
264.0	748.45	36.50	36.19	16.46	14.11
238.8	684.45	34.66	37.31	17.66	15.68
230.7	827.80	33.13	32.52	17.50	10.53
251.6	860.45	35.75	33.71	16.40	11.00
257.9	875.15	34.46	34.14	16.28	11.31
263.9	909.45	34.60	34.85	16.06	11.96
266.5	905.55	35.38	35.89	15.93	12.58
229.1	756.00	35.85	33.53	16.60	10.66
239.3	769.35	35.68	33.79	16.41	10.85
258.0	793.50	35.35	34.72	16.17	11.41
257.6	801.65	35.04	35.22	15.92	11.91
267.3	819.65	34.07	36.50	16.04	12.85
267.0	808.55	32.20	37.60	16.19	13.58
259.6	774.95	34.32	37.89	16.62	14.21
240.4	711.85	31.08	37.71	17.37	15.56
227.2	694.85	35.73	37.00	18.12	15.83
196.0	638.10	34.11	36.76	18.53	16.41
278.7	77.55	34.79	34.62	15.54	13.10
272.3	75.90	35.77	35.40	15.70	13.63
267.4	753.35	36.44	35.96	16.45	14.51
254.5	704.70	37.82	36.26	17.62	15.38
224.7	666.80	35.07	36.34	18.12	16.10

y	x_1	x_2	x_3	x_4	x_5
181.5	568.55	35.26	35.90	19.05	16.73
227.5	653.10	35.56	31.84	16.51	10.58
253.6	704.05	35.73	33.16	16.02	11.28
263.0	709.60	36.46	33.83	15.89	11.91
265.8	726.90	36.26	34.89	15.83	12.65
263.8	697.15	37.20	36.27	16.71	14.06

[20]:

```
# Load the necessary library
#library("MPV")

# Attach the dataset for easier referencing
attach(table.b2)
```

```
The following objects are masked from table.b1 (pos = 3):

    x1, x2, x3, x4, x5, y


The following objects are masked from table.b1 (pos = 4):

    x1, x2, x3, x4, x5, y
```

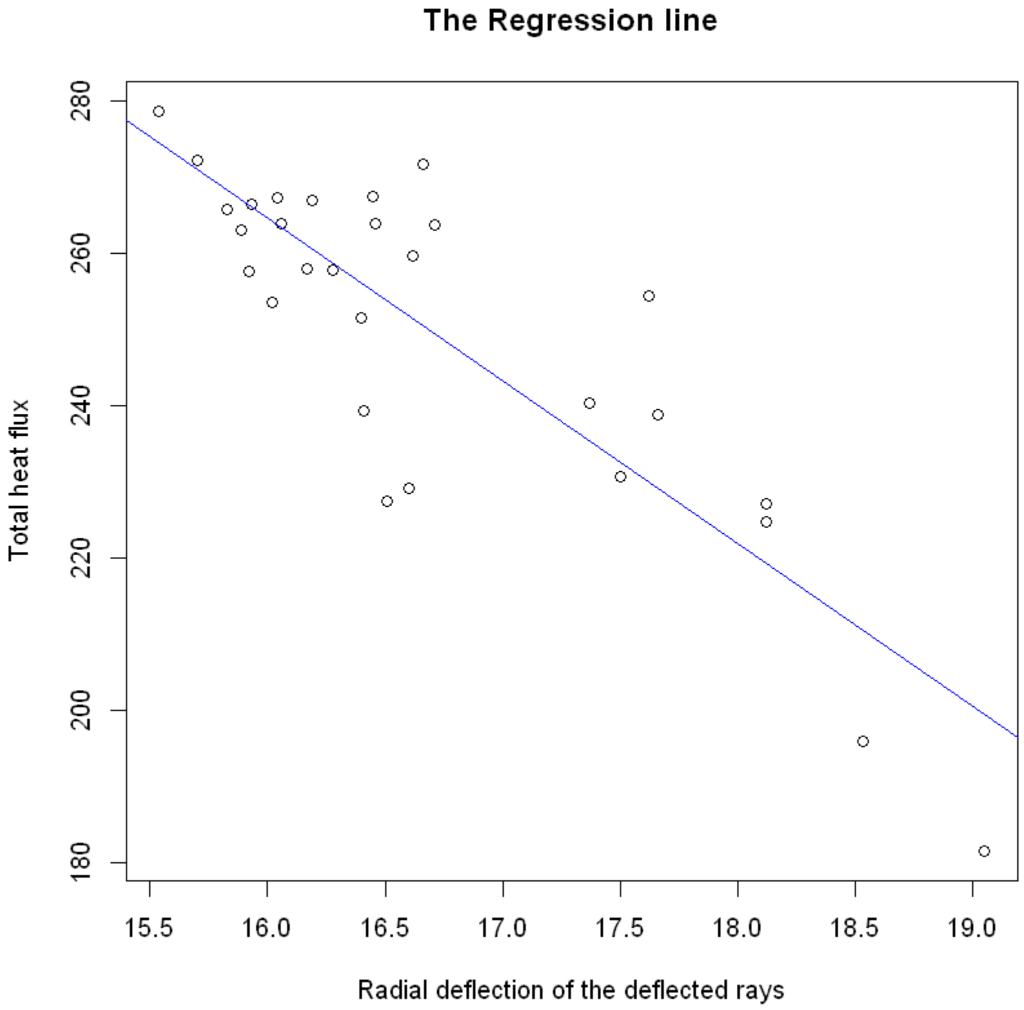
[21]:

```
# Fit a linear model predicting y based on x4
reg <- lm(y ~ x4, data = table.b2)

# Summarize the linear model to view key statistics
summary(reg)
```

```
Call:
lm(formula = y ~ x4, data = table.b2)

Residuals:
     Min       1Q   Median       3Q      Max
-26.2487  -4.5029   0.5202   7.9093  24.5080

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  607.103     42.906  14.150 5.24e-14 ***
x4           -21.402      2.565  -8.343 5.94e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.33 on 27 degrees of freedom
```

```
Multiple R-squared:  0.7205,	Adjusted R-squared:  0.7102
F-statistic: 69.61 on 1 and 27 DF,  p-value: 5.935e-09
```

[22]:
```
# Plot the data with the regression line
plot(table.b2$x4, table.b2$y,
     main = "The Regression line",
     xlab = "Radial deflection of the deflected rays",
     ylab = "Total heat flux")

# Add the regression line to the plot
abline(reg, col = "blue")
```



[25]:
```
# Perform an ANOVA on the linear model
anova_table <- anova(reg)
print(anova_table)
```

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x4         1 10578.7   10579  69.609 5.935e-09 ***
Residuals 27  4103.2     152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[29]:

```
# Calculate the 99% confidence interval for the coefficient of x4
conf_interval <- confint(reg, 'x4', level = 0.99)
print(conf_interval)
```

```
       0.5 %    99.5 %
x4 -28.50995 -14.29497
```

[30]:

```
# Create a new data frame for prediction
d <- data.frame(x4 = 16.5)

# Predict the response for the new data point with confidence interval
predictions <- predict(reg, newdata = d, interval = 'confidence')
print(predictions)
```

```
       fit      lwr      upr
1 253.9627 249.1468 258.7787
```

■ **Interpretation:-**

a. In Table B.2, y denotes total heat flux (in kilowatts) and $x4$ denotes radial deflection of the deflected rays (in milliradians). According to the question y is the dependent variable and $x4$ is the independent variable. In R programming " Im() " function can fit the linear regression model. So by running the code "reg=Im(y x4,data=table.b2)" and "summary(reg)" we get the simple linear regression model as—-

$$\hat{y} = 607.103 - 21.402x_4$$

b. In R, analysis of variance can be constructed using "anova()" function. So, by the code "anova(y.Im)" the table is as follows—Analysis of Variance Table

```
Response: y
                                Df Sum Sq Mean Sq F value Pr(>F)
x4 1 10578.7 10579 69.609 5.935e-09 ***
Residuals 27 4103.2 }15
-_-
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'0.1 ', 1
```

Testing significance of regression—-

The hypotheses for testing significance of regression is $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Failing to reject $H_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y.

Alternatively, if $\mathrm{H}_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$\mathrm{t}_0 = \frac{\widehat{\beta}_1}{\mathrm{se}\left(\widehat{\beta}_1\right)}$$

The null hypothesis of significance of regression would be rejected if $|t_0| > t_{\alpha/2,n-2}$.

Here, $\widehat{\beta_1} = -21.402$ and se $\left(\widehat{\beta_1}\right) = 2.565$. So, $\mathrm{t}_0 = -8.3439\&\mathrm{t}_{0.025,27} = 2.051831$. So, the null hypothesis is rejected since $|t_0| > t_{0.025,27}$. That means, $x4$ is of value in explaining the variability in y.

c. The $100(1-\alpha)$ percent CI on β_1 is $\widehat{\beta_1} - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{MS_{Res}}{\sum_{i=1}^{n} x_i^2}} \leq \beta_1 \leq \widehat{\beta_1} + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{MS_{\mathrm{Res}}}{\sum_{i=1}^{n} x_i^2}}$

The confidence interval can be found in R programming by using "confint()" function. So by the code "confint(y.Im,'x4',level=0.99)" the 99% confidence interval of slope is (-28.50995, -14.29497).

d. Coefficient of determination is the proportion of the variation in the response variable that is explained by the regression model.

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

By the output of regression model in R, we get that $R^2 = 0.7205$ that means 72.1% of the variation in y is explained by the regression model.

e. A 100(1- α) percent Cl on the mean response at the point $x = x_0$ I

$$\widehat{\mu_{y|x0}} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq E\left(y \mid x_0\right) \leq \widehat{\mu_{y|x0}} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Where, $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$ and $\widehat{E(y \mid x_0)} = \widehat{\mu_{y|x0}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$

To find the confidence interval in R, we create a new dataframe with the desired value i.e., $d = data.frame(x4 = 16.5)$

Then the prediction is made by predict() function and 'interval' argument is set to 'confidence' to output the mean interval i.e., predict(reg,newdata=d,interval='confidence')

So the 95% confidence interval on the mean heat flux when the radial deflection is 16.5 milliradians is $(249.1468, 258.7787)$.

4) The purity of oxygen produced by a fractional distillation process is thought to be related to the percentage of hydrocarbons in the main condensor of the processing unit. Twenty samples are shown below.

Purily (%)	Hydrocarbon (%)	Purily (%)	Hydrocarbon (%)
86.91	1.02	96.73	1.46
89.85	1.11	99.42	1.55
90.28	1.43	98.66	1.55
86.34	1.11	96.07	1.55
92.58	1.01	93.65	1.40
87.33	0.95	87.31	1.15
86.29	1.11	95.00	1.01
91.86	0.87	96.85	0.99
95.61	1.43	85.20	0.95
89.86	1.02	90.56	0.98

a. Fit a simple linear regression model to the data.

b. Test the hypothesis HO : $\beta 1 = 0$.

c. Calculate R2 .

d. Find a 95%Cl on the slope.

e. Find a 95%Cl on the mean purity when the hydrocarbon percentage is 1.00 .

[31]:

```r
# Load the necessary library
# library("MPV")

# Attach the dataset for easier referencing
attach(p2.7)
```

[32]:

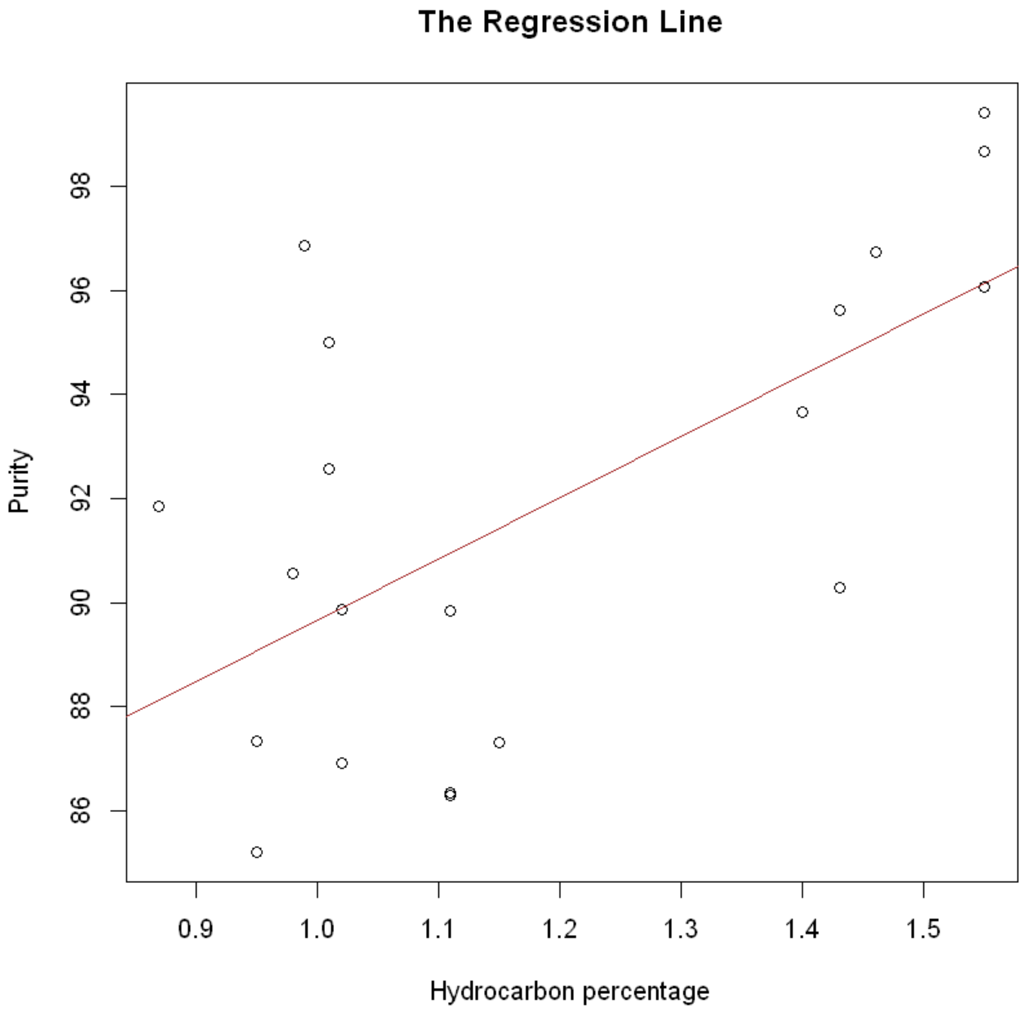
```r
# Fit a linear model predicting purity based on hydro
purity.lm <- lm(purity ~ hydro, data = p2.7)

# Summarize the linear model to view key statistics
summary(purity.lm)
```

```
Call:
lm(formula = purity ~ hydro, data = p2.7)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6724 -3.2113 -0.0626  2.5783  7.3037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.863      4.199  18.544 3.54e-13 ***
hydro         11.801      3.485   3.386  0.00329 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.597 on 18 degrees of freedom
```

```
Multiple R-squared:  0.3891,	Adjusted R-squared:  0.3552 
F-statistic: 11.47 on 1 and 18 DF,  p-value: 0.003291
```

[33]:

```
# Plot the data with the regression line
plot(p2.7$hydro, p2.7$purity,
     main = "The Regression Line",
     xlab = "Hydrocarbon percentage",
     ylab = "Purity")

# Add the regression line to the plot
abline(purity.lm, col = "brown")
```



[34]:

```
# Calculate the 95% confidence interval for the coefficient of hydro
conf_interval <- confint(purity.lm, 'hydro', level = 0.95)
print(conf_interval)
```

```
         2.5 %   97.5 %
hydro 4.479066 19.12299
```

[35]:

```
# Predict the response for hydro = 1.00 with a confidence interval
predictions <- predict(purity.lm, newdata = data.frame(hydro = 1.00),␣
 ↪interval = 'confidence')
print(predictions)
```

```
       fit      lwr      upr
1 89.66431 87.51017 91.81845
```

■ **Interpretation:-**

a. In this given table purity(%) is dependent variable and hydrocarbon(%) is independent variable. In R programming " Im( )" function can fit the linear regression model. So by running the code "purity.Im=Im(purity hydro,data=p2.7)" and "summary(purity.Im)" we get the simple linear regression model as—

$$\hat{y} = 21.788251 - 0.007025x$$

where y = purity, x = hydro

b. Testing significance of regression—-

The hypotheses for testing significance of regression is $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$. Failing to reject $\mathrm{H}_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y. Alternatively, if $H_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$\mathrm{t}_0 = \frac{\widehat{\beta}_1}{\mathrm{se}\left(\widehat{\beta}_1\right)}$$

The null hypothesis of significance of regression would be rejected if $|t_0| > t_{\alpha/2,n-2}$.

Here, $\widehat{\beta}_1 = 11.801$ and $\mathrm{se}\left(\widehat{\beta}_1\right) = 3.485$. So, $\mathrm{t}_0 = 3.0692\&\mathrm{t}_{0.025,18} = 2.100922$. So, the null hypothesis is rejected since $|t_0| > t_{0.025,18}$. That means, hydrocarbon is of value in explaining the variability in purity.

c. Coefficient of determination is the proportion of the variation in the response variable that is explained by the regression model.

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

By the output of regression model in R, we get that $R^2 = 0.3891$ that means 38.9% of the variation in y is explained by the regression model.

d. The $100(1-\alpha)$ percent Cl on β_1 is $\widehat{\beta}_1 - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{MS_{\mathrm{Res}}}{\sum_{i=1}^{n} x_i^2}} \leq \beta_1 \leq \widehat{\beta}_1 + t_{\frac{\alpha}{2}}, n-1\sqrt{\frac{MS_{\mathrm{Res}}}{\sum_{i=1}^{n} x_i^2}}$

The confidence interval can be found in R programming by using “confint()” function. So by the code “confint(purity.Im,‘hydro’,level=0.95)”, the 95% confidence interval of slope is (4.479066,19.12299).

e. A 100(1- α) percent Cl on the mean response at the point $x = x_0$

$$\widehat{\mu_{y|x0}} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq E\left(y \mid x_0\right) \leq \widehat{\mu_{y|x0}} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Where, $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$ and $E\widehat{(y \mid x_0)} = \widehat{\mu_{y|x0}} = \widehat{\beta_0} + \widehat{\beta_1} x_0$

To find the confidence interval in R , we create a new dataframe with the desired value i.e., data.frame(hydro=1.00)

Then the prediction is made by predict() function and ‘interval’ argument is set to ‘confidence’ to output the mean interval i.e., predict(purity. Im, newdata=data.frame(hydro=1.00), interval=‘confidence’)

So the 95% confidence interval on the mean purity when the hydrocarbon percentage is 1.00 is—-

$(81.51017, 91.81845)$.

5) Consider the oxygen plant data in Problem 2.7 and assume that purity and hydrocarbon percentage are jointly normally distributed random variables.

a. What is the correlation between oxygen purity and hydrocarbon percentage?

b. Test the hypothesis that $\rho = 0$.

c. Construct a 95%CI for ρ.

[36]:

```
# Load the necessary library
# library("MPV")

# Attach the dataset for easier referencing
attach(p2.7)
```

```
The following objects are masked from p2.7 (pos = 3):

    hydro, purity
```

[37]:

```
# Fit a linear model predicting purity based on hydro
purity.lm <- lm(purity ~ hydro, data = p2.7)

# Summarize the linear model to view key statistics
summary(purity.lm)
```

```
Call:
lm(formula = purity ~ hydro, data = p2.7)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6724 -3.2113 -0.0626  2.5783  7.3037 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   77.863      4.199  18.544 3.54e-13 ***
hydro         11.801      3.485   3.386  0.00329 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.597 on 18 degrees of freedom
Multiple R-squared:  0.3891,	Adjusted R-squared:  0.3552 
F-statistic: 11.47 on 1 and 18 DF,  p-value: 0.003291
```

[38]:

```
# Extract the R-squared value from the model summary
R <- summary(purity.lm)$r.squared

# Calculate the square root of R-squared (which gives the correlation␣
 ↪coefficient)
sqrt_R <- sqrt(R)
print(sqrt_R)
```

```
[1] 0.6237968
```

[39]:

```
# Perform a Pearson correlation test between hydro and purity
cor_test <- cor.test(p2.7$hydro, p2.7$purity, method = "pearson")
print(cor_test)
```

```

	Pearson's product-moment correlation

data:  p2.7$hydro and p2.7$purity
t = 3.3861, df = 18, p-value = 0.003291
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2503961 0.8356439
sample estimates:
      cor 
0.6237968 
```

■ **Interpretation:-**

a. According to the question purity and hydrocarbon percentage are jointly normally distributed random variables. Then the estimator of correlation coefficient (ρ) is the sample correlation coefficient

$$r=\frac{\sum_{i=1}^{n} y_i\left(x_i-\bar{x}\right)}{\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\right]^{1/2}}=\frac{S_{xy}}{\left[S_{xx} SS_T\right]^{1/2}}$$

Note that,

$$\widehat{\beta_1}=\left(\frac{SS_T}{S_{xx}}\right)^{1/2} r$$

so that the slope $\widehat{\beta_1}$ is just the sample correlation coefficient r multiplied by a scale factor that is the square root of the spread of the y' s divided by the spread of the $x's$. Thus, $\widehat{\beta_1}$ and r are closely related, although they provide somewhat different information. The sample correlation coefficient r is a measure of the linear association between y and x, while $\widehat{\beta_1}$ measures the change in the mean of y for a unit change in x. In the case of a controllable variable x, r has no meaning because the magnitude of r depends on the choice of spacing for x. We may also write,

$$r^2=\widehat{\beta}_1^2 \frac{S_{xx}}{SS_T}=\frac{\widehat{\beta_1} S_{xy}}{SS_T}=\frac{SS_R}{SS_T}=R^2$$

That is, the coefficient of determination R^2 is just the square of the correlation coefficient between y and x .

So, in R programming first we compute the regression model and then extract R-squared

“summary(purity.Im)$r.squared” then store it in a variable R. since $r^2 = R^2$ so correlation coefficient (r) = root of R^2. i.e.,sqrt (R^2) so $r = 0.624$.

b. While regression and correlation are closely related, regression is a more powerful tool in many situations. Correlation is only a measure of association and is of little use in prediction. However, regression methods are useful in developing quantitative relationships between variables, which can be used in prediction. It is often useful to test the hypothesis that the correlation coefficient equals zero, that is,

$$H_0: \rho=0, H_1: \rho \neq 0$$

The appropriate test statistic for this hypothesis is

$$t_0=\frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$$

which follows the t distribution with n − 2 degrees of freedom if $H_0 : \rho = 0$ is true. Therefore, we would reject the null hypothesis if $|t_0| > t_{\frac{\alpha}{2},n-2}$.

In R, it can be computed from “cor.test()” function which gives $t_0 = 3.3861$ and p-value = 0.003291 which is < 0.05. So, the null hypothesis is rejected at 5% level of significance i.e., there is a non-zero correlation between oxygen purity and hydrocarbon.

c. It is also possible to construct a $100(1-\alpha)$ percent CI for ρ. The $100(1-\alpha)$ percent CI is

$$\tanh\left(\operatorname{arctanh} r - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right)$$

In R, by the output of cor.test(p2.7$hydro,p2.7$purity,method="pearson") we've got the confidence interval too which is, $(0.25, 0.84)$.

Note that, this 95% confidence interval does not contain 0 , which is consistent with the decision to reject the null hypothesis.

6) The weight and systolic blood pressure of 26 randomly selected males in the age group **25** – **30** are shown below. Assume that weight and blood pressure (BP) are jointly normally distributed.

a. Find a regression line relating systolic blood pressure to weight.

b. Estimate the correlation coefficient.

c. Test the hypothesis that $\rho = 0$.

d. Test the hypothesis that $\rho = 0.6$.

e. Find a 95%CI for ρ.

Subject	Weight	Symbolic BP	Subject	Weight	Systolic BP
1	165	130	14	172	153
2	167	133	15	159	128
3	180	150	16	168	132
4	155	128	17	174	149
5	212	151	18	183	158
6	175	146	19	215	150
7	190	150	20	195	163
8	210	140	21	180	156
9	200	148	22	143	124
10	149	125	23	240	170
11	158	133	24	235	165
12	169	135	25	192	160
13	170	150	26	187	159

[40]:

```
# Attach the dataset for easier referencing
attach(p2.10)
```

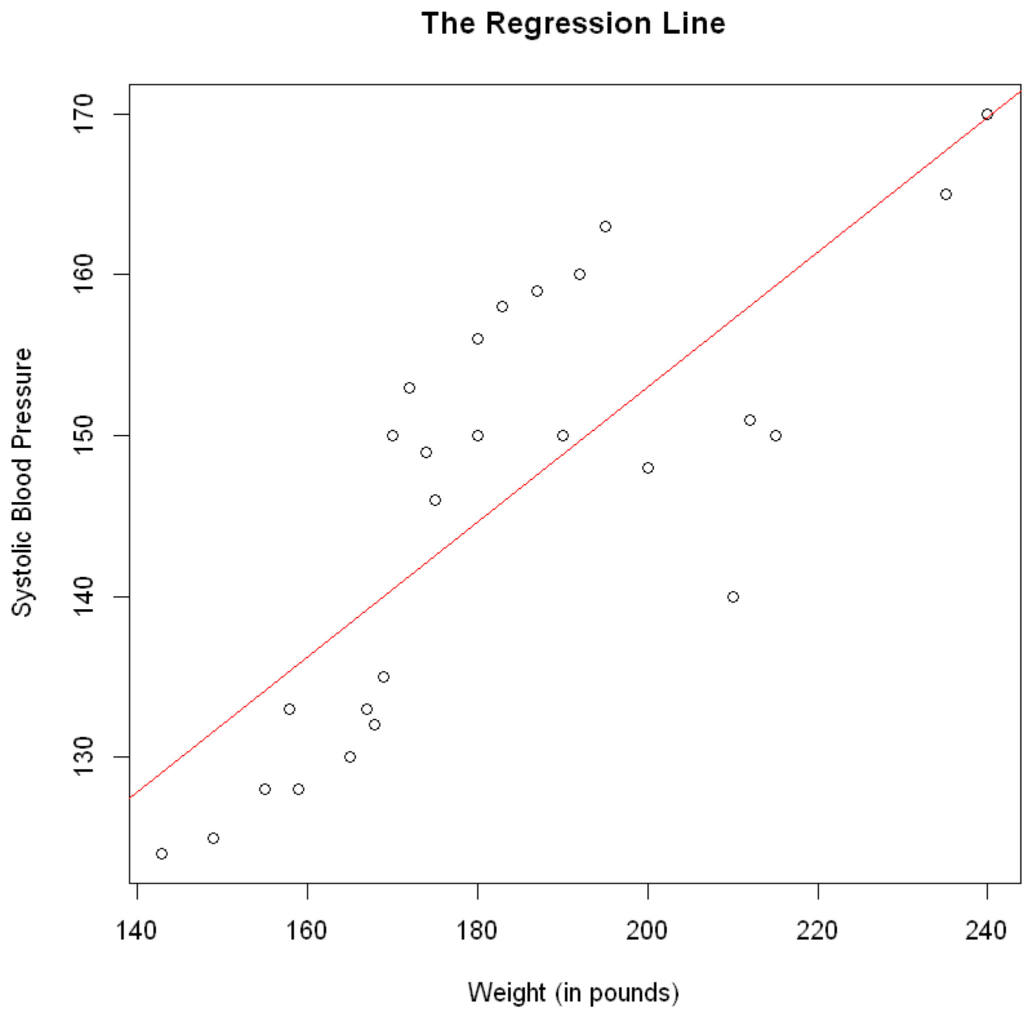
[41]:

```
# Fit a linear model predicting systolic blood pressure (sysbp) based␣
 ↪on weight
model <- lm(sysbp ~ weight, data = p2.10)

# Summarize the linear model to view key statistics
summary(model)
```

```
Call:
lm(formula = sysbp ~ weight, data = p2.10)

Residuals:
    Min      1Q  Median      3Q     Max
-17.182  -6.485  -2.519   8.926  12.143

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 69.10437   12.91013   5.353 1.71e-05 ***
weight       0.41942    0.07015   5.979 3.59e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.681 on 24 degrees of freedom
Multiple R-squared:  0.5983,	Adjusted R-squared:  0.5815
F-statistic: 35.74 on 1 and 24 DF,  p-value: 3.591e-06
```

[42]:

```
# Plot the data with the regression line
plot(p2.10$weight, p2.10$sysbp,
     main = "The Regression Line",
     xlab = "Weight (in pounds)",
     ylab = "Systolic Blood Pressure")

# Add the regression line to the plot
abline(model, col = "red")
```



[43]:
```
# Extract the R-squared value from the model summary
R <- summary(model)$r.squared

# Calculate the square root of R-squared (which gives the correlation␣
↪coefficient)
sqrt_R <- sqrt(R)
print(sqrt_R)
```

```
[1] 0.7734903
```

[44]:
```
# Perform a Pearson correlation test between weight and systolic blood␣
↪pressure
cor_test <- cor.test(p2.10$weight, p2.10$sysbp, method = "pearson")
print(cor_test)
```

```
        Pearson's product-moment correlation
```

```
data:  p2.10$weight and p2.10$sysbp
t = 5.9786, df = 24, p-value = 3.591e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5513214 0.8932215
sample estimates:
      cor
0.7734903
```

■ **Interpretation:-**

a. According to the question weight is the independent variable and systolic blood pressure is dependent variable. In R programming “Im()” gives the regression model. So, by the code

“model = Im (sysbp ~ weight, data = p2.10)

summary(model)” where sysbp=systolic blood pressure

we get the regression equation $\hat{y} = 69.10437 + 0.41942x$

where y = systolic blood pressure, x = weight.

b. According to the question weight and systolic blood pressure are jointly normally distributed random variables. Then the estimator of correlation coefficient (ρ) is the sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\left[\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2\right]^{1/2}} = \frac{S_{xy}}{\left[S_{xx} SS_T\right]^{1/2}}$$

Note that,

$$\widehat{\beta_1} = \left(\frac{SS_T}{S_{xx}}\right)^{1/2} r$$

so that the slope $\widehat{\beta_1}$ is just the sample correlation coefficient r multiplied by a scale factor that is the square root of the spread of the y' s divided by the spread of the $x's$. Thus, $\widehat{\beta_1}$ and r are closely related, although they provide somewhat different information. The sample correlation coefficient r is a measure of the linear association between y and x, while $\widehat{\beta_1}$ measures the change in the mean of y for a unit change in x. In the case of a controllable variable x, r has no meaning because the magnitude of r depends on the choice of spacing for x. We may also write,

$$r^2 = \widehat{\beta_1}^2 \frac{S_{xx}}{SS_T} = \frac{\widehat{\beta_1} S_{xy}}{SS_T} = \frac{SS_R}{SS_T} = R^2$$

That is, the coefficient of determination R^2 is just the square of the correlation coefficient between y and x .

So, in R programming first we compute the regression model and then extract R-squared "summary(model)$r.squared" then store it in a variable R. since $r^2 = R^2$ so correlation coefficient (r) =root of R^2. i.e.,sqrt(R^2) so $r = 0.7734903$.

c. While regression and correlation are closely related, regression is a more powerful tool in many situations. Correlation is only a measure of association and is of little use in prediction. However, regression methods are useful in developing quantitative relationships between variables, which can be used in prediction. It is often useful to test the hypothesis that the correlation coefficient equals zero, that is,

$$H_0 : \rho = 0, H_1 : \rho \neq 0$$

The appropriate test statistic for this hypothesis is

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which follows the t distribution with n $-$ 2 degrees of freedom if $H_0 : \rho = 0$ is true. Therefore, we would reject the null hypothesis if $|t_0| > t_{\frac{\alpha}{2},n-2}$.

In R, it can be computed from "cor.test()" function which gives $t_0 = 5.9786$ and p - value=3.591e-06 which is < 0.05. So, the null hypothesis is rejected at 5% level of significance i.e., there is a non-zero correlation between weight and systolic blood pressure.

d. The test procedure for the hypotheses

$$H_0 : \rho = \rho_0, H_1 : \rho \neq \rho_0$$

For moderately large samples (e.g., n $\geq$ 25) the statistic

$$Z = \operatorname{arctanh} r = \frac{1}{2} \ln \frac{1+r}{1-r}$$

is approximately normally distributed with mean

$$\mu_Z = \operatorname{arctanh} \rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

And variance $\quad \sigma_Z^2 = (n-3)^{-1}$

Therefore to test the hypothesis $H_0 : \rho = \rho_0$ we may compute the statistic

$$Z_0 = (\operatorname{arctanh} r - \operatorname{arctanh} \rho_0)(n-3)^{1/2}$$

And reject $H_0 : \rho = \rho_0$ if $|Z_0| > Z_{\alpha/2}$.

Here, $r = 0.7734903$ and $\rho_0 = 0.6$ (given) so,

$$
\begin{aligned}
Z_0 &= (\operatorname{arctanh} 0.7734903 - \operatorname{arctanh} 0.6)(26-3)^{\frac{1}{2}} \\
&= (1.0289 - 0.6932)(23)^{\frac{1}{2}} \\
&= 1.60996
\end{aligned}
$$

Since, the rejection region is $|Z_0| > Z_{\alpha/2} = 1.96$. So, we fail to reject H_0 at 5% level of significance.

e. It is also possible to construct a $100(1-\alpha)$ percent Cl for ρ. The $100(1-\alpha)$ percent Cl is

$$
\tanh\left(\operatorname{arctanh} r - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right)
$$

In R, by the output of cor.test(p2.10$weight,p2.10$sysbp,method="pearson") we've got the confidence interval too which is, $(0.55, 0.89)$.

7) The number of pounds of steam used per month is thought to be related to the average monthly ambient temperature. The past year's usages and temperatures follow:

Month	Temperature	Usage/1000	Month	Temperature	Usage/1000
Jan.	21	185.79	Jul.	68	621.55
Feb.	24	214.47	Aug.	74	675.06
Mar.	32	288.03	Sep.	62	562.03
Apr.	47	424.84	Oct.	50	452.93
May	50	454.68	Nov.	41	369.95
Jun.	59	539.03	Dec.	30	273.98

a. Fit a simple linear regression model to the data.

b. Test for significance of regression.

c. Plant management believes that an increase in average ambient temperature of 1 degree will increase average monthly steam consumption by 10,000lb. Do the data support this statement?

d. Construct a 99% prediction interval on steam usage in a month with average ambient temperature of 58°.

[45]:

```r
# Attach the dataset for easier referencing
attach(p2.12)
```

[46]:

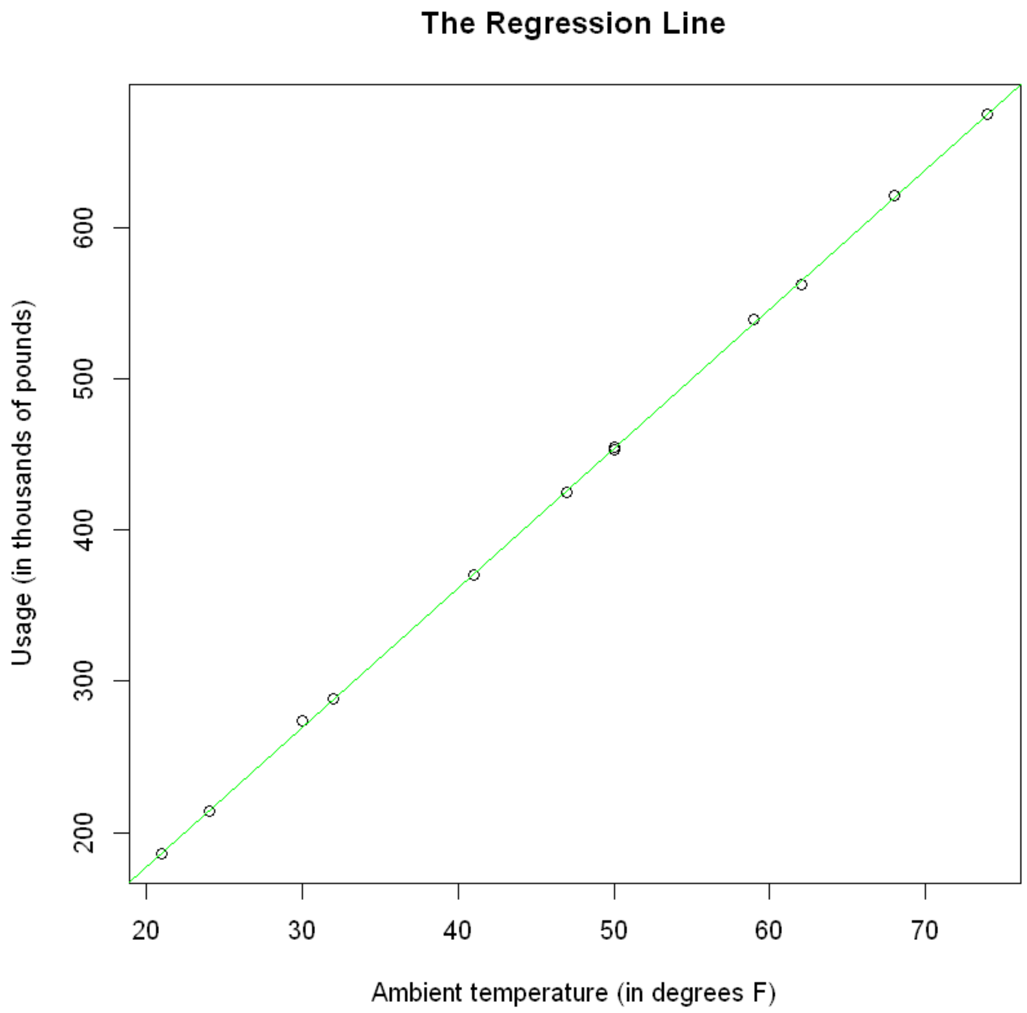
```r
# Fit a linear model predicting usage based on temperature
model <- lm(usage ~ temp, data = p2.12)

# Summarize the linear model to view key statistics
summary(model)
```

```
Call:
```

```
lm(formula = usage ~ temp, data = p2.12)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5629 -1.2581 -0.2550  0.8681  4.0581 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.33209    1.67005  -3.792  0.00353 ** 
temp         9.20847    0.03382 272.255  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.946 on 10 degrees of freedom
Multiple R-squared:  0.9999,	Adjusted R-squared:  0.9999 
F-statistic: 7.412e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

[47]:

```r
# Plot the data with the regression line
plot(p2.12$temp, p2.12$usage,
     main = "The Regression Line",
     xlab = "Ambient temperature (in degrees F)",
     ylab = "Usage (in thousands of pounds)")

# Add the regression line to the plot
abline(model, col = "green")
```



[48]:
```
# Predict the usage for a temperature of 58°F with a 99% prediction␣
 ↪interval
predictions <- predict(model, newdata = data.frame(temp = 58), interval␣
 ↪= 'predict', level = 0.99)
print(predictions)
```

```
       fit      lwr      upr
1 527.759 521.2237 534.2944
```

■ **Interpretation:-**

a. According to the question temperature is the independent variable and usage/1000 is the dependent variable. Let, temp=temperature. In R programming, “Im()” gives the regression model. So by the code “model = Im (usage $^{\text{temp}}$,data = p2.12)

Summary(model)”

We get the regression equation as $\hat{y} = -6.33209 + 9.20847x$

Y = usage/1000, x = temperature.

b. Testing significance of regression—-

The hypotheses for testing significance of regression is $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$. Failing to reject $\mathrm{H}_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y. Alternatively, if $\mathrm{H}_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$t_0 = \frac{\widehat{\beta_1}}{\operatorname{se}\left(\widehat{\beta_1}\right)}$$

The null hypothesis of significance of regression would be rejected if $|t_0| > t_{\alpha/2,n-2}$.

Here, $\widehat{\beta_1} = 9.20847$ and se $\left(\widehat{\beta_1}\right) = 0.03382$. So, $\mathrm{t}_0 = 272.2788 \& \mathrm{t}_{0.025,10} = 2.228139$. So, the null hypothesis is rejected since $|t_0| > t_{0.025,18}$. That means, temperature is of value in explaining the variability in the number of pounds of steam used per month.

c. An increase in average ambient temperature of 1 degree will increase average monthly steam consumption by 10, 000lb.

Then, per 1000 average monthly steam consumption will increase by 10 lb .

So, here we have to test the hypothesis that the slope of regression equation is a constant, say β_{10}. The appropriate hypotheses are

$$H_0 : \beta_1 = \beta_{10}, \quad H_1 : \beta_1 \neq \beta_{10}$$

where we have specified a two - sided alternative. Since the errors ε_i are NID $(0, \sigma^2)$, the observations y_i s are NID $(\beta_0 + \beta_1 x_i, \sigma^2)$. Now $\widehat{\beta_1}$ is a linear combination of the observations, so $\widehat{\beta_1}$ is normally distributed with mean β_1 and variance σ^2 / S_{xx}.

Therefore, the statistic

$$Z_0 = \frac{\widehat{\beta_1} - \beta_{10}}{\sqrt{\sigma^2 / S_{xx}}}$$

is distributed $N(0,1)$ if the null hypothesis $H_0 : \beta_1 = \beta_{10}$ is true. If σ^2 were known, we could use Z_0 to test the hypotheses. Typically, σ^2 is unknown. We have already seen that MS_{Res} is an unbiased estimator of σ^2. It is established that (n – 2)$MS_{\text{Res}} / \sigma^2$ follows a χ^2_{n-2} distribution and that MS_{Res} and σ^2 are independent. By the definition of a t statistic

$$t_0 = \frac{\widehat{\beta_1} - \beta_{10}}{\sqrt{MS_{\text{Res}} / S_{xx}}}$$

follows a t_{n-2} distribution if the null hypothesis $H_0 : \beta_1 = \beta_{10}$ is true. The degrees of freedom associated with t_0 are the number of degrees of freedom associated with MS_{Res} . Thus, the ratio t_0 is the test statistic used to test $H_0 : \beta_1 = \beta_{10}$. The test procedure computes t_0 and compares the observed value of t_0 with the upper $\alpha/2$ percentage point of the t_{n-2} distribution $\left(t_{\alpha/2,n-2}\right)$. This procedure rejects the null hypothesis if

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

Alternatively, a P - value approach could also be used for decision making. The denominator of the test statistic, t_0 is often called the estimated standard error, or more simply, the standard error of the slope. That is,

$$\operatorname{se}\left(\widehat{\beta_1}\right) = \sqrt{\frac{MS_{Res}}{S_{xx}}}$$

Therefore, we often see t_0 written as,

$$t_0 = \frac{\widehat{\beta_1} - \beta_{10}}{\operatorname{se}\left(\widehat{\beta_1}\right)}$$

Here, according to the question the appropriate hypotheses are, $H_0 : \beta_1 = 10, \quad H_1 : \beta_1 \neq 10$

And by the output of the summary of regression model, we get $\widehat{\beta_1} = 9.20847$ and se $\left(\widehat{\beta_1}\right) = 0.03382$. So, $\mathrm{t}_0 = -23.4$.

Since, $|t_0| > \mathrm{t}_{0.025,10} = 2.228139$ we reject H_0 and claim that the usage increase is less than 10,000 .

d. 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

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

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. We now develop a prediction interval for the future observation y_0. Note that the random variable $\Psi = y_0 - \widehat{y_0}$ is normally distributed with mean zero

and variance

$$\operatorname{Var}(\Psi) = \operatorname{Var}\left(y_0 - \widehat{y_0}\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{s_{xx}}\right]$$

because the future observation y_0 is independent of $\widehat{y_0}$. If we use $\widehat{y_0}$ to predict y 0 , then the standard error of $\Psi = y_0 - \widehat{y_0}$ is the appropriate statistic on which to base a prediction interval. Thus, the 100($1 - \alpha$) percent prediction interval on a future observation at x_0 is

$$\widehat{y_0} - t_{\frac{\alpha}{2}n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le y_0 \le \widehat{y_0} + t_{\frac{\alpha}{2}n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

To find the prediction interval in R, we create a new data frame with the desired value to predict. The prediction is made with the predict() function. The interval argument is given ‘prediction’ since prediction interval is required and level is 0.99 since 99% prediction interval is required. So the 99% prediction interval on steam usage in a month with average ambient temperature of 58° is ($521.22, 534.29$).

8) Davidson (“Update on Ozone Trends in California’s South Coast Air Basin” Air and Waste, 43, 226, 1993) studied the ozone levels in the South Coast Air Basin of California for the years 1976-1991. He believes that the number of days the ozone levels exceeded 0.20 ppm (the response) depends on the seasonal meteorological index, which is the seasonal average 850 - millibar temperature (the regressor). The following table gives the data.

Year	Days	Index
1976	91	16.7
1977	105	17.1
1978	106	18.2
1979	108	18.1
1980	88	17.2
1981	91	18.2
1982	58	16.0
1983	82	17.2
1984	81	18.0
1985	65	17.2
1986	61	16.9
1987	48	17.1
1988	61	18.2
1989	43	17.3
1990	33	17.5
1991	36	16.6

a. Make a scatterplot of the data.

b. Estimate the prediction equation.

c. Test for significance of regression.

d. Calculate and plot the 95% confidence and prediction bands.

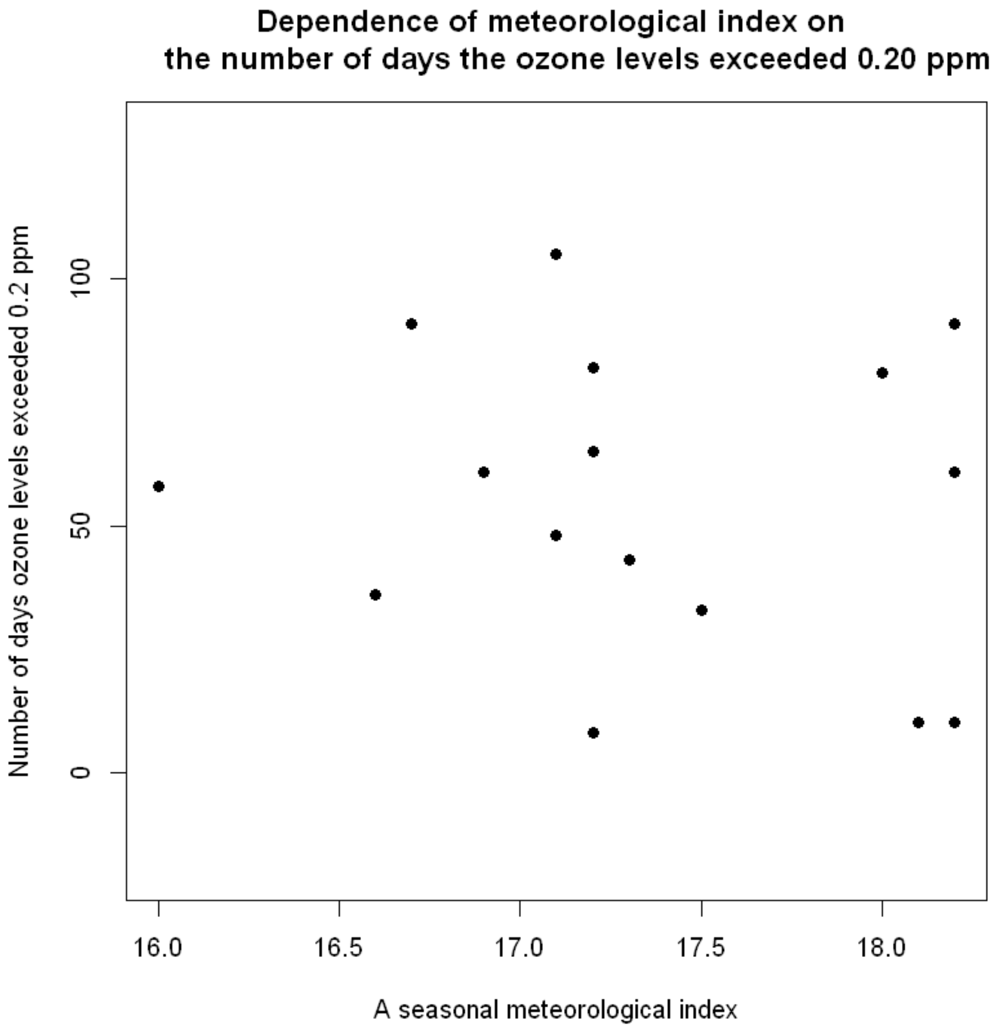
```
[75]: # Attach the dataset for easier column access
      attach(p2.13)
```

```
The following objects are masked from p2.13 (pos = 3):


    days, index
```

```
The following objects are masked from p2.13 (pos = 4):

    days, index


The following objects are masked from p2.13 (pos = 6):

    days, index
```

[90]:

```
# Create a scatter plot with a linear model fit
# Set up a rectangular plotting area
par(mar = c(5, 5, 4, 2) + 0.1)  # Adjust margins if needed
par(pty = "m")  # Make the plot rectangular

plot(index, days,
     main = "Dependence of meteorological index on
     the number of days the ozone levels exceeded 0.20 ppm",
     xlab = "A seasonal meteorological index",
     ylab = "Number of days ozone levels exceeded 0.2 ppm",
     ylim = c(-20, 130),
     pch = 19)
```



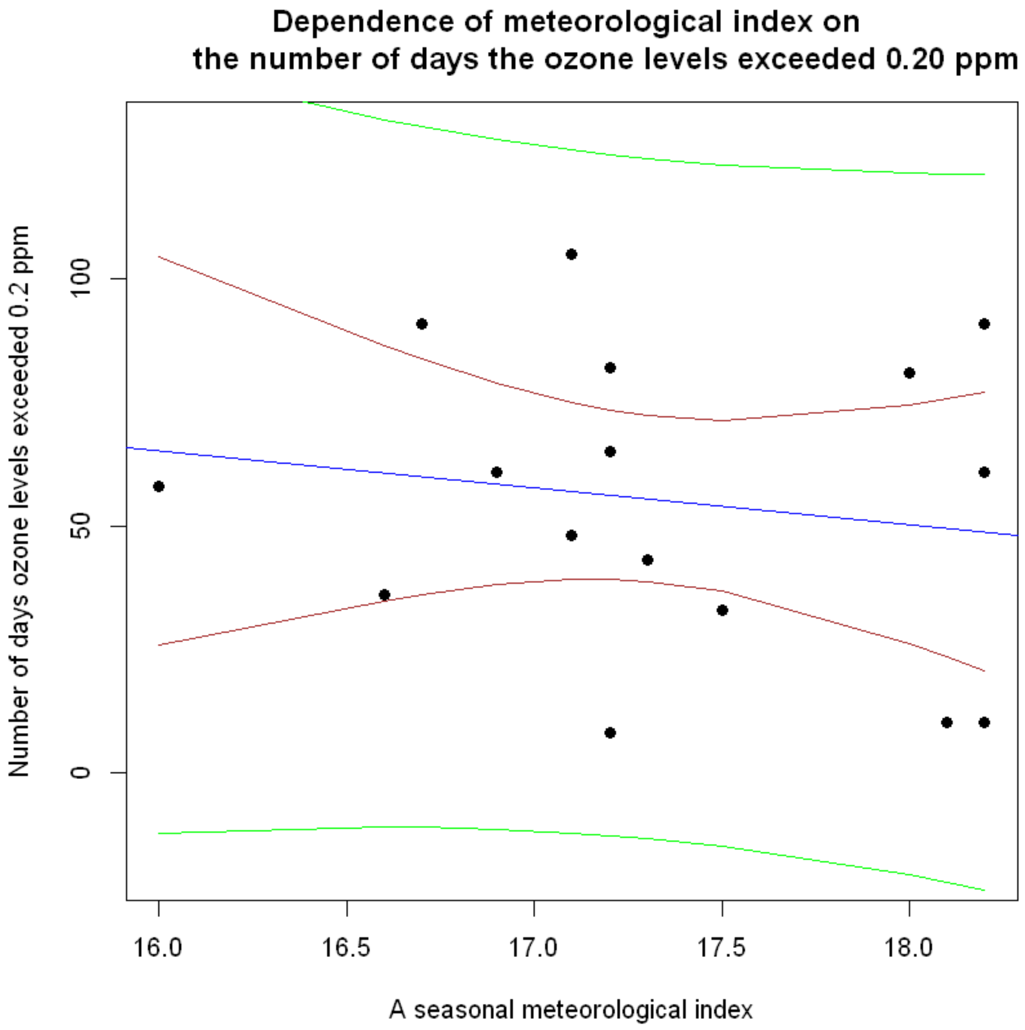
[85]:
```
# Fit a linear model
ozone.lm = lm(days ~ index, data = p2.13)
# Display the summary of the linear model
summary(ozone.lm)
```

```
Call:
lm(formula = days ~ index, data = p2.13)

Residuals:
    Min      1Q  Median      3Q     Max
-48.252 -21.947  -2.305  26.979  48.008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  183.596    214.359   0.856    0.406
index         -7.404     12.351  -0.599    0.558
```

```
Residual standard error: 31.2 on 14 degrees of freedom
Multiple R-squared:  0.02502,	Adjusted R-squared:  -0.04462
F-statistic: 0.3593 on 1 and 14 DF,  p-value: 0.5585
```

[93]:

```r
# Fit a linear model
ozone.lm <- lm(days ~ index, data = p2.13)

# Plot the data with appropriate labels and limits
plot(p2.13$index, p2.13$days,
     main = "Dependence of meteorological index on
        the number of days the ozone levels exceeded 0.20 ppm",
     xlab = "A seasonal meteorological index",
     ylab = "Number of days ozone levels exceeded 0.2 ppm",
     ylim = c(-20, 130),
     pch = 19)

# Add the regression line to the plot
abline(ozone.lm, col = "blue")

# Predict confidence intervals for the linear model
ozone.conf = predict(ozone.lm, interval = 'confidence')

# Add confidence interval lines to the plot
lines(sort(index), ozone.conf[order(index), 2], col = "brown")
lines(sort(index), ozone.conf[order(index), 3], col = "brown")

# Predict prediction intervals for the linear model
ozone.pred = predict(ozone.lm, newdata = p2.13, interval = 'predict',␣
 ↪level = 0.95)

# Add prediction interval lines to the plot
lines(sort(index), ozone.pred[order(index), 2], col = "green")
lines(sort(index), ozone.pred[order(index), 3], col = "green")
```



■ **Interpretation:-**

a. Scatter diagram is a graph in which the values of two variables are plotted along two axes. It is a chart which shows the relationship between two variables. For the given table of meteorological index and the number of days the ozone level exceeded by 0.20 ppm, the scatter plot is easily drawn in R by “plot()” function.

Dependence of meteorological index on the number of days the ozone levels exceeded 0.20 ppm

b. According to the question, the seasonal meteorological index is the independent variable and the number of days ozone level exceeded by 0.20 ppm is the dependent variable. In R the regression model or prediction equation can be computed by using “Im()” function. So by the code “ozone.Im=Im(days $^{\text{index, data=}}$ p2.13)” we get the required equation,

$$\hat{y} = 183.596 - 7.404x$$

In the code, days= the number of days ozone level exceeded by 0.20 ppm Index= seasonal meteorological index

And in the equation, $y =$ index and $x =$ days.

c. Testing significance of regression—-

The hypotheses for testing significance of regression is $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$. Failing to reject $\mathrm{H}_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y. Alternatively, if $\mathrm{H}_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$\mathrm{t}_0 = \frac{\widehat{\beta_1}}{\mathrm{se}\left(\widehat{\beta_1}\right)}$$

The null hypothesis of significance of regression would be rejected if $|t_0| > t_{\alpha/2,n-2}$.

Here, $\widehat{\beta_1} = -7.404$ and se $\left(\widehat{\beta_1}\right) = 12.351$. So, $\mathrm{t}_0 = -0.599 \& \mathrm{t}_{0.025,14} = 2.144787$. So, the null hypothesis can't be rejected since $|t_0| < t_{0.025,14}$. That means, there is no linear relationship between the seasonal meteorological index and the number of days ozone level exceeded by 0.20 ppm .

d. Confidence interval:-

A 100(1- α) percent CI on the mean response at the point $x = x_0$ I

$$\widehat{\mu_{y|x0}} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq E\left(y \mid x_0\right) \leq \widehat{\mu_{y|x0}} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Where, $S_{xx} = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$ and $\widehat{E(y \mid x_0})} = \widehat{\mu_{y|x0}} = \widehat{\beta_0} + \widehat{\beta_1} x_0$

Note that the width of the Cl for $E\left(y \mid x_0\right)$ is a function of x_0. The interval width is a minimum for $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases. Intuitively this is reasonable, as we would expect our best estimates of y to be made at x values near the center of the data and the precision of estimation to deteriorate as we move to the boundary of the x space.

Prediction interval:-

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

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

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. We now develop a prediction interval for the future observation y_0. Note that the random variable $\Psi = y_0 - \widehat{y_0}$ is normally distributed with mean zero

and variance

$$\operatorname{Var}(\Psi) = \operatorname{Var}\left(y_0 - \widehat{y_0}\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]$$

because the future observation y_0 is independent of $\widehat{y_0}$. If we use $\widehat{y_0}$ to predict y 0 , then the standard error of $\Psi = y_0 - \widehat{y_0}$ is the appropriate statistic on which to base a prediction interval. Thus, the 100(1- α) percent prediction interval on a future observation at x_0 is

$$\widehat{y_0} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right)} \leq y_0 \leq \widehat{y_0} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right)}$$

In R, confidence and prediction band can be drawn. First, we need to calculate confidence and prediction intervals using “predict()” function. Then by “lines()” function the bands can be drawn as follows:

Dependence of meteorological index on the number of days the ozone levels exceeded 0.20 ppm

This above figure shows that green lines are the 95% prediction limits and brown lines are 95% confidence limits for ozone model. This graph illustrates the point that the prediction interval is wider than confidence interval.

9) Hsuie, Ma, and Tsai (“Separation and Characterizations of Thermotropic Co-polyesters of pHydroxybenzoic Acid, Sebacic Acid, and Hydroquinone,” Journal of Applied Polymer Science, 56, 471 - 476, 1995) study the effect of the molar ratio of sebacic acid (the regressor) on the intrinsic viscosity of copolyesters (the response). The following table gives the data:

Ratio	Viscosity
1.0	0.45
0.9	0.20
0.8	0.34
0.7	0.58
0.6	0.70
0.5	0.57
0.4	0.55
0.3	0.44

a. Make a scatterplot of the data.

b. Estimate the prediction equation.

c. Perform a complete, appropriate analysis (statistical tests, calculation of R2, and so forth).

d. Calculate and plot the 95% confidence and prediction bands.

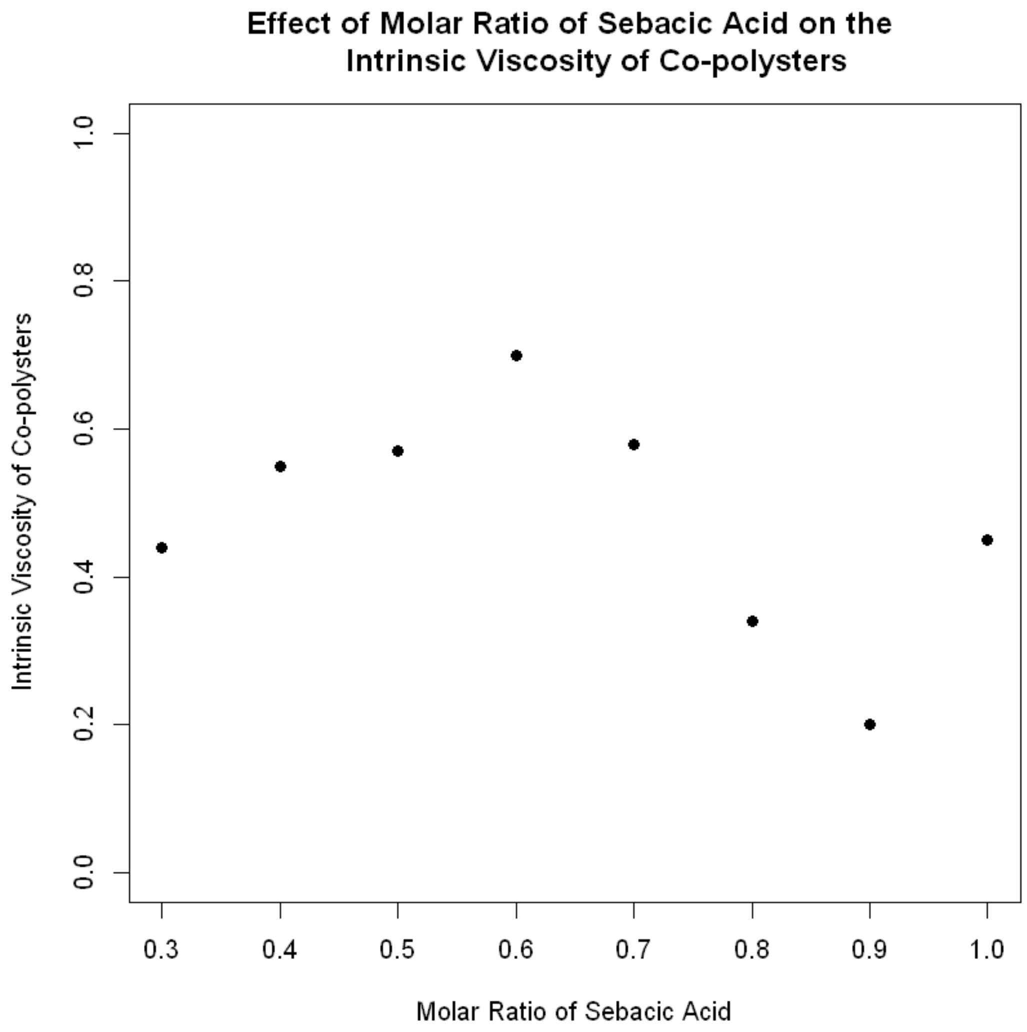
[96]:

```
# Attach the dataset for easier referencing
attach(p2.14)
```

```
The following objects are masked from p2.14 (pos = 3):

    ratio, visc


The following objects are masked from p2.14 (pos = 4):

    ratio, visc
```

[98]:

```
# Plot the data with appropriate labels and limits
plot(p2.14$ratio, p2.14$visc,
     main = "Effect of Molar Ratio of Sebacic Acid on the
     Intrinsic Viscosity of Co-polysters",
     xlab = "Molar Ratio of Sebacic Acid",
     ylab = "Intrinsic Viscosity of Co-polysters",
     pch = 19,
     ylim = c(0, 1))
```



[99]:

```
# Fit a linear model predicting intrinsic viscosity (visc) based on␣
 ↪molar ratio (ratio)
visc.lm <- lm(visc ~ ratio, data = p2.14)

# Summarize the linear model to view key statistics
summary(visc.lm)
```

```
Call:
lm(formula = visc ~ ratio, data = p2.14)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20464 -0.10634  0.02196  0.08527  0.20643

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   0.6714     0.1595   4.209  0.00563 **
ratio        -0.2964     0.2314  -1.281  0.24754
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.15 on 6 degrees of freedom
Multiple R-squared:  0.2147,	Adjusted R-squared:  0.08382
F-statistic:  1.64 on 1 and 6 DF,  p-value: 0.2475
```

[107]:

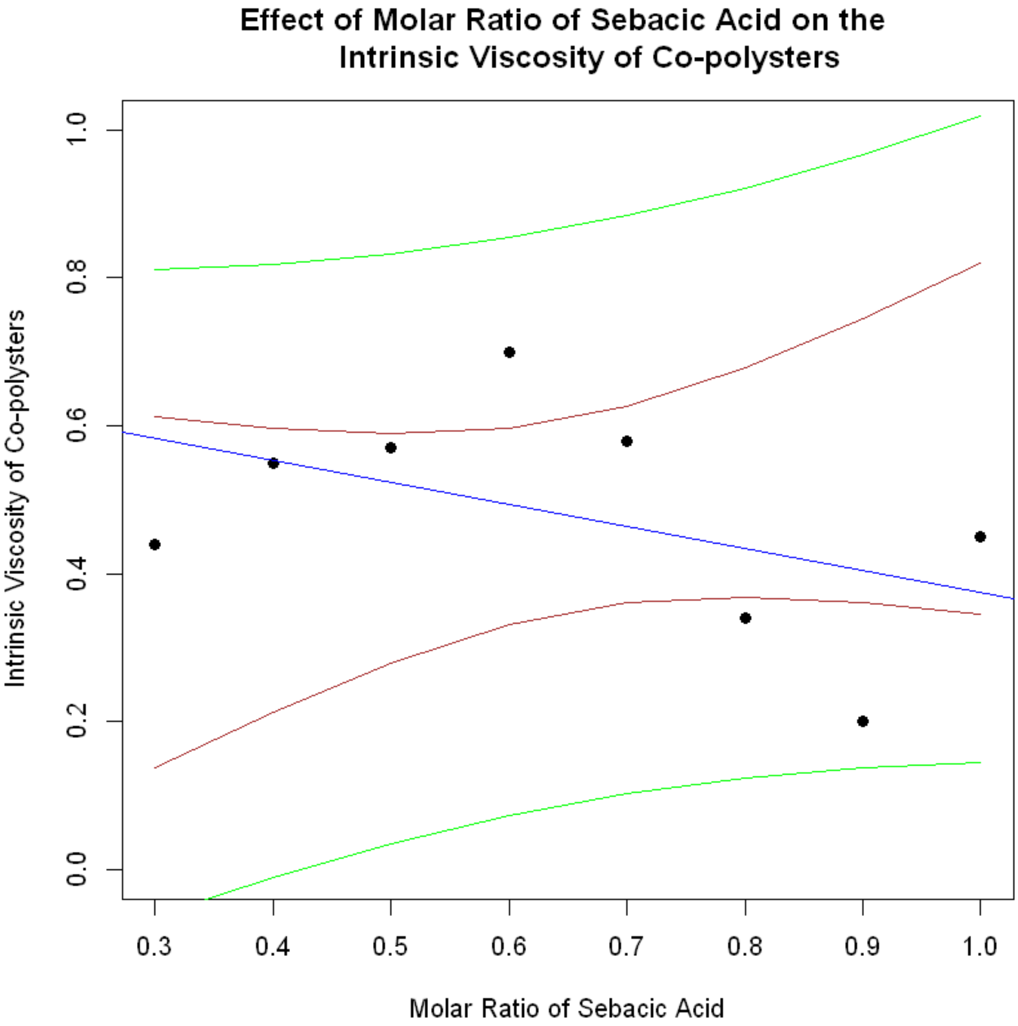
```
# Perform ANOVA on the fitted linear model to assess the significance␣
 ↪of the predictors
print(anova(visc.lm))
```

```
Analysis of Variance Table

Response: visc
          Df   Sum Sq  Mean Sq F value Pr(>F)
ratio      1 0.036905 0.036905  1.6405 0.2475
Residuals  6 0.134982 0.022497
```

[106]:

```
plot(p2.14$ratio, p2.14$visc,
     main = "Effect of Molar Ratio of Sebacic Acid on the
     Intrinsic Viscosity of Co-polysters",
     xlab = "Molar Ratio of Sebacic Acid",
     ylab = "Intrinsic Viscosity of Co-polysters",
     pch = 19,
     ylim = c(0, 1))
# Add the regression line to the existing plot
abline(visc.lm, col = "blue")

# Predict confidence intervals for the fitted values
visc.conf <- predict(visc.lm, interval = 'confidence')

# Add confidence interval lines to the plot
sorted_ratio <- sort(p2.14$ratio)
lines(sorted_ratio, visc.conf[order(p2.14$ratio),␣
 ↪2][match(sorted_ratio, p2.14$ratio)], col = "brown")  # Lower bound
lines(sorted_ratio, visc.conf[order(p2.14$ratio),␣
 ↪3][match(sorted_ratio, p2.14$ratio)], col = "brown")  # Upper bound

# Predict prediction intervals for new data
visc.pred <- predict(visc.lm, newdata = p2.14, interval = 'predict')

# Add prediction interval lines to the plot
lines(sorted_ratio, visc.pred[order(p2.14$ratio),␣
 ↪2][match(sorted_ratio, p2.14$ratio)], col = "green")  # Lower bound
```

```
lines(sorted_ratio, visc.pred[order(p2.14$ratio),␣
↪3][match(sorted_ratio, p2.14$ratio)], col = "green")  # Upper bound
```



■ Interpretation:-

a. Scatter diagram is a graph in which the values of two variables are plotted along two axes. It is a chart which shows the relationship between two variables. The scatterplot of effects of molar ratio of sebacic acid on the intrinsic viscosity of co-polysters is,

Effect of molar ratio of sebacic acid on the intrinsic viscosity of co-polysters

b. According to the question, molar ratio of sebacic acid is the independent variable (regressor) and intrinsic viscosity of co-polysters is dependent variable(response variable). In R programming regression model or prediction equation can be computed from "Im()" function. So by the code “visc. Im $=$ Im $(vi\mathrm{I}^{\sim}$ ratio, data $=$ p2.14)” we get the required equation,

$$\hat{y}=0.6714-0.2964x$$

Where in the code, visc=intrinsic viscosity of co-polysters, ratio=molar ratio of sebacic acid.

In the equation, $y=visc, x=$ ratio.

c. Testing significance of regression—-

The hypotheses for testing significance of regression is $\mathrm{H}_0: \beta_1=0$ vs. $\mathrm{H}_1: \beta_1 \neq 0$. Failing to reject $\mathrm{H}_0: \beta_1=0$ implies that there is no linear relationship between x and y. Alternatively, if $\mathrm{H}_0: \beta_1=0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$\mathrm{t}_0=\frac{\widehat{\beta_1}}{\operatorname{se}\left(\widehat{\beta_1}\right)}$$

The null hypothesis of significance of regression would be rejected if $|t_0|>t_{\alpha/2,n-2}$.

Here, $\widehat{\beta_1}=-0.2964$ and $\operatorname{se}\left(\widehat{\beta_1}\right)=0.2314$. So, $\mathrm{t}_0=-1.2809\&\mathrm{t}_{0.025,6}=2.446912$. So, the null hypothesis can't be rejected since $|t_0|<t_{0.025,6}$. That means, there is no linear relationship between the molar ratio of sebacic acid and the intrinsic viscosity of co-polysters.

We can also use the usual analysis - of - variance F test to test the hypothesis $\mathrm{H}_0: \beta_1=0$. We know that

(1) $SS_{\text{Res}}=(n-2)MS_{\text{Res}}/\sigma^2$ follows a χ_{n-2}^2 distribution.

(2) If the null hypothesis $\mathrm{H}_0: \beta_1=0$ is true, then SS_R/σ^2 follows a χ_1^2 distribution

(3) SS_{Res} and SS_R are independent.

By the definition of an F statistic,

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

follows the $F_{1,n-2}$ distribution. We know that the expected values of these mean squares are

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

These expected mean squares indicate that if the observed value of F_0 is large, then it is likely that the slope $\beta_1 \neq 0$. If $\beta_1 \neq 0$, then F_0 follows a noncentral F distribution with 1 and $n-2$ degrees of freedom and a noncentrality parameter of

$$\lambda=\frac{\beta_1^2 S_{xx}}{\sigma^2}$$

This noncentrality parameter also indicates that the observed value of F_0 should be large if $\beta_1 \neq 0$. Therefore, to test the hypothesis $\mathrm{H}_0 : \beta_1 = 0$, compute the test statistic F_0 and reject H_0 if

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

To compute the analysis-of-variance in R, “anova(visc.Im)” gives

Source on variation	sum of squares	df	mean square	F_0
Regression (ratio)	0.036905	1	0.036905	0.036905/0.022497
Residual	0.134982	6	0.022497	= 1.64

Here $F_0 = 1.64$ and $F_{0.05;1,6} = 5.99$ by the code in R ” $qf\,(0.05, 1, 6,$ lower.tail=FALSE)“. So, we can’t reject H_0. That means, there is no linear relationship between the molar ratio of sebacic acid and the intrinsic viscosity of copolysters. Also, from the output of”Im()” function p-value $= 0.2475$ which is not less than 0.05 which means we can’t reject the null hypothesis.

Coefficient of determination is the proportion of the variation in the response variable that is explained by the regression model.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \tag{23}$$

By the output of regression model in R, we get that $R^2 = 0.2147$ that means 21.5% of the variation in y is explained by the regression model.

d. Confidence interval:-

A 100(1- α) percent Cl on the mean response at the point x = x$_0$ I

$$\widehat{\mu_{y|x0}} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)} \leq E\left(y \mid x_0\right) \leq \widehat{\mu_{y|x0}} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)}$$

Where, $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$ and $\widehat{E(y \mid x_0)} = \widehat{\mu_{y|x0}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ Note that the width of the Cl for $E\left(y \mid x_0\right)$ is a function of x$_0$. The interval width is a minimum for $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases. Intuitively this is reasonable, as we would expect our best estimates of y to be made at x values near the center of the data and the precision of estimation to deteriorate as we move to the boundary of the x space.

Prediction interval:-

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

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

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. We now develop a prediction interval for the future observation y_0. Note that the random variable $\Psi = y_0 - \widehat{y_0}$ is normally distributed with mean zero

and variance $\quad \text{Var}(\Psi) = \text{Var}\left(y_0 - \widehat{y_0}\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}\right]$

because the future observation y_0 is independent of $\widehat{y_0}$. If we use $\widehat{y_0}$ to predict y 0 , then the standard error of $\Psi = y_0 - \widehat{y_0}$ is the appropriate statistic on which to base a prediction interval. Thus, the 100($1-\alpha$) percent prediction interval on a future observation at x_0 is

$$\widehat{y_0} - t_{\frac{\alpha}{2}n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le y_0 \le \widehat{y_0} + t_{\frac{\alpha}{2}n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

In R, confidence and prediction band can be drawn. First, we need to calculate confidence and prediction intervals using “predict()” function. Then by “lines()” function the bands can be drawn as follows:

Effect of molar ratio of sebacic acid on the intrinsic viscosity of co-polysters

This above figure shows that green lines are the 95% prediction limits and brown lines are 95% confidence limits for viscosity model. This graph illustrates the point that the prediction interval is wider than confidence interval.

10) Carroll and Spiegelman (“The Effects of Ignoring Small Measurement Errors in Precision Instrument Calibration,” Journal of Quality Technology, 18, 170 -173, 1986) look at the relationship between the pressure in a tank and the volume of liquid. The following table gives the data. Use an appropriate statistical software package to perform an analysis of these data. Comment on the output produced by the software routine.

Volume	Pressure	Volume	Pressure	Volume	Pressure
2084	4599	2842	6380	3789	8599
2084	4600	3030	6818	3789	8600
2273	5044	3031	6817	3979	9048
2273	5043	3031	6818	3979	9048
2273	5044	3221	7266	4167	9484
2463	5488	3221	7268	4168	9487
2463	5487	3409	7709	4168	9487
2651	5931	3410	7710	4358	9936
2652	5932	3600	8156	4358	9938

Volume	Pressure	Volume	Pressure	Volume	Pressure
2652	5932	3600	8158	4546	10377
2842	6380	3788	8597	4547	10379

[108]:

```r
# Load the necessary library
# library("MPV")

# Attach the dataset for easier referencing
attach(p2.16)

# View the first few rows of the dataset
head(p2.16)

# Get a summary of the dataset
summary(p2.16)

# Check the dimensions of the dataset
dim(p2.16)

# Count the number of missing values in each column
sapply(p2.16, function(x) sum(is.na(x)))
```

```
The following object is masked from package:datasets:

    pressure
```

A data.frame: 6 × 2

	volume	pressure
	<dbl>	<dbl>
1	2084	4599
2	2084	4600
3	2273	5044
4	2273	5043
5	2273	5044
6	2463	5488

```
     volume         pressure
 Min.   :2084   Min.   : 4599
 1st Qu.:2652   1st Qu.: 5932
 Median :3221   Median : 7268
 Mean   :3295   Mean   : 7441
 3rd Qu.:3979   3rd Qu.: 9048
 Max.   :4547   Max.   :10379
```

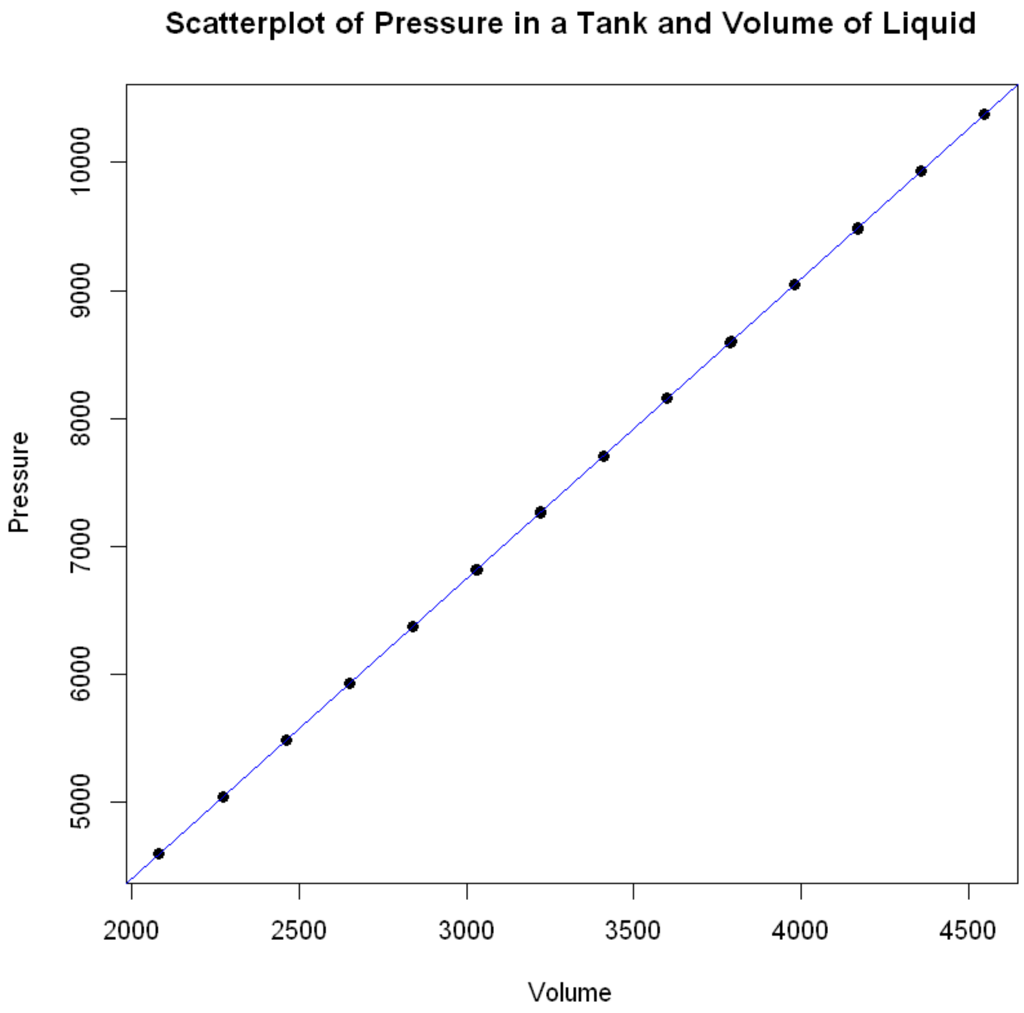
1. 33 2. 2

volume 0 **pressure** 0

[111]:

```
# Create a scatterplot of pressure vs. volume
plot(p2.16$volume, p2.16$pressure,
     main = "Scatterplot of Pressure in a Tank and Volume of Liquid",
     xlab = "Volume",
     ylab = "Pressure",
     pch = 16)
# Fit a linear model predicting pressure based on volume
pressure.lm <- lm(pressure ~ volume, data = p2.16)

# View a summary of the linear model
summary(pressure.lm)

# Add the regression line to the scatterplot
abline(pressure.lm, col = "blue")
```

```
Call:
lm(formula = pressure ~ volume, data = p2.16)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3276 -0.9227  0.0773  1.2676  2.9577 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.907e+02  1.355e+00  -214.6   <2e-16 ***
volume       2.346e+00  4.007e-04  5855.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.741 on 31 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 3.429e+07 on 1 and 31 DF,  p-value: < 2.2e-16
```



[113]:

```
# Perform ANOVA on the fitted linear model
print(anova(pressure.lm) )
```

```
Analysis of Variance Table

Response: pressure
          Df    Sum Sq   Mean Sq  F value    Pr(>F)
volume     1 103947022 103947022 34286009 < 2.2e-16 ***
Residuals 31        94         3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[114]:

```
# Calculate the correlation between volume and pressure
cor(p2.16$volume, p2.16$pressure)

# Perform Pearson correlation test
cor.test(p2.16$volume, p2.16$pressure, method = "pearson")
```

0.999999547920861

```
        Pearson's product-moment correlation

data:  p2.16$volume and p2.16$pressure
t = 5855.4, df = 31, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9999991 0.9999998
sample estimates:
      cor
0.9999995
```

[116]:
```
# Calculate the 95% confidence intervals for the model coefficients
print(confint(pressure.lm, 'volume', level = 0.95))
```

```
          2.5 %   97.5 %
volume 2.345614 2.347249
```

■ **Interpretation:-**

In R programming first we load a package called “MPV” by the code “library(”MPV”)“. Then we attach the above table by the code”attach(p2.16)“. Then by the code”head(p2.16)” we can see the first 6 rows of the data. Then by the code “summary(p2.16)”, we get more information about the data.

In the volume column, minimum value=2084, first quantile=2652, median=3221, mean=3295, third quantile=3979 and maximum value=4547. In the pressure column, minimum value=4599, first quantile=5932, median=7268, mean=7441, third quantile = 9048, maximum value = 10379.

By the code “dim(p2.16)” we get the dimension of the data, that is, number of rows and columns. Here by the output it means p2.16 data has 33 rows and 2 columns.

The “sapply()” function helps us in applying functions on a list, vector or data frame and returns an array or matrix object of the same length. Here by the code “sapply(p2.16, function(x) sum is.na(x))” means in each column we try to find the sum of missing values. As a result of the output, volume and pressure column has no missing values.

Then we draw a scatter plot where x-axis is the volume of the liquid and the y-axis is pressure in a tank. By the diagram we can see that there is a positive correlation between volume and pressure.

Now, according to the question, volume of liquid is the independent variable and pressure in a tank is the dependent variable. In R by “Im()” function we can compute the regression model, so y the code “Im(pressure ∼ volume,data=p2.16)” we get the regression equation as

$$\hat{y} = -290.7 + 2.346x$$

Where, y = pressure in a tank, x = volume of liquid.

Testing for regression:-

The hypotheses for testing significance of regression is $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$. Failing to reject $\mathrm{H}_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y. Alternatively, if $\mathrm{H}_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$\mathrm{t}_0 = \frac{\widehat{\beta_1}}{\mathrm{se}\left(\widehat{\beta_1}\right)}$$

The null hypothesis of significance of regression would be rejected if $|t_0| > t_{\alpha/2,n-2}$.

Here, $\widehat{\beta_1} = 2.346$ and $\mathrm{se}\left(\widehat{\beta_1}\right) = 0.0004007$. So, $\mathrm{t}_0 = 5854.7542\&\mathrm{t}_{0.025,31} = 2.039513$. So, the null hypothesis is rejected since $|t_0| > t_{0.025,6}$. That means, there is a linear relationship between the pressure in a tank and volume of liquid.

Analysis of variance:-

We can also use the usual analysis - of - variance F test to test the hypothesis $\mathrm{H}_0 : \beta_1 = 0$. We know that

(1) $SS_{\text{Res}} = (n-2)MS_{\text{Res}} / \sigma^2$ follows a χ^2_{n-2} distribution.

(2) If the null hypothesis $\mathrm{H}_0 : \beta_1 = 0$ is true, then SS_R/σ^2 follows a χ^2_1 distribution

(3) SS_{Res} and SS_R are independent.

By the definition of an F statistic,

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

follows the $F_{1,n-2}$ distribution. We know that the expected values of these mean squares are

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

These expected mean squares indicate that if the observed value of F_0 is large, then it is likely that the slope $\beta_1 \neq 0$. If $\beta_1 \neq 0$, then F_0 follows a noncentral F distribution with 1 and $n-2$ degrees of freedom and a noncentrality parameter of

$$\lambda = \frac{\beta_1^2 S_{xx}}{\sigma^2}$$

This noncentrality parameter also indicates that the observed value of F_0 should be large if $\beta_1 \neq 0$. Therefore, to test the hypothesis $\mathrm{H}_0 : \beta_1 = 0$, compute the test statistic F_0 and reject H_0 if

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

To compute the analysis-of-variance in R, “anova(visc.Im)” gives

Source on variation	sum of squares	df	mean square	F_0
Regression (ratio)	103947022	1	103947022	103947022/3
Residual	94	31	3	= 34286009

Here F$_0$ = 34286009 and $F_{0.05;1,31}$ = 4.159615 by the code in R ” qf (0.05, 1, 6 ,lower.tail=FALSE)“. So, we reject H$_0$. That means, there is a linear relationship between the pressure in a tank and volume of liquid. Also, from the output of”anova(pressure.Im)” function p-value is $< 2.2\text{e} - 16$ which is so much less than 0.05 which means we will reject the null hypothesis.

Even from the output of the Im() function in R we can interpret the validation of model.

Coefficient:-

In this section we can see Pr($>$ $|t|$). For both intercept and volume there are three stars which represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis at 5% level of significance which allows us to conclude that there is a relationship between volume and pressure.

Residual standard error:-

Residual standard error is a measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term. Due to the presence of this error term we are not capable of perfectly predicting our response variable(pressure) from the predictor(volume) one. The residual standard error is the average amount that the response(pressure) will deviate from the true regression line.

In this output, pressure in a tank can deviate from true regression line by approximately 1.741 on average with 31 degrees of freedom. Degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters(restriction).

Multiple R-squared, Adjusted R-squared:-

The R-squared statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. It is a measure of the linear relationship between our predictor variable and our response variable. It always lies between 0 and 1. A number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable. Adjusted R-squared is calculated by dividing the residual mean square error by the total mean square error. Here adjusted R-squared is 1 . It indicates that the response variable(pressure) can be perfectly explained without error by the predictor variable(volume).

F-statistic:-

F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. How much larger the F-statistic needs to be depend on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis. In this data, F -statistic is 3.429*10^7 which is sufficiently larger than 1 . So there is a relationship between volume of liquid and pressure in tank.

Correlation test:-

According to the question pressure and volume are jointly normally distributed random variables. Then the estimator of correlation coefficient(ρ) is the sample correlation coefficient

$$r=\frac{\sum_{i=1}^{n} y_i\left(x_i-\bar{x}\right)}{\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\right]^{1/2}}=\frac{S_{xy}}{\left[S_{xx} SS_T\right]^{1/2}}$$

Note that,

$$\widehat{\beta_1}=\left(\frac{SS_T}{S_{xx}}\right)^{1/2} r$$

so that the slope $\widehat{\beta_1}$ is just the sample correlation coefficient r multiplied by a scale factor that is the square root of the spread of the $y's$ divided by the spread of the $x's$. Thus, $\widehat{\beta_1}$ and r are closely related, although they provide somewhat different information. The sample correlation coefficient r is a measure of the linear association between y and x, while $\widehat{\beta_1}$ measures the change in the mean of y for a unit change in x. In the case of a controllable variable x, r has no meaning because the magnitude of r depends on the choice of spacing for x. We may also write,

$$r^2=\widehat{\beta_1}^2 \frac{S_{xx}}{SS_T}=\frac{\widehat{\beta_1} S_{xy}}{SS_T}=\frac{SS_R}{SS_T}=R^2$$

That is, the coefficient of determination R^2 is just the square of the correlation coefficient between y and x .

So, in R programming first we compute the regression model and then extract R-squared “summary(model)$r.squared” then store it in a variable R. since $r^2=R^2$ so correlation coefficient (r) =root of R^2. i.e.,sqrt (R^2) so $r=0.999995$.

While regression and correlation are closely related, regression is a more powerful tool in many situations. Correlation is only a measure of association and is of little use in prediction. However, regression methods are useful in developing quantitative relationships between variables, which can be used in prediction. It is often useful to test the hypothesis that the correlation coefficient equals zero, that is,

$$H_0: \rho=0, H_1: \rho \neq 0$$

The appropriate test statistic for this hypothesis is

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which follows the t distribution with n − 2 degrees of freedom if $H_0 : \rho = 0$ is true. Therefore, we would reject the null hypothesis if $|t_0| > t_{\frac{\alpha}{2},n-2}$.

In R, it can be computed from "cor.test()" function which gives $t_0 = 5855.4$ and p -value $< 2.2^{*10^{-6}}$ which is < 0.05. So, the null hypothesis is rejected at 5% level of significance i.e., there is a non-zero correlation between weight and systolic blood pressure.

It is also possible to construct a $100(1-\alpha)$ percent Cl for ρ. The $100(1-\alpha)$ percent Cl is

$$\tanh\left(\operatorname{arctanh} r - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(\operatorname{arctanh} r + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right)$$

In R, by the output of cor.test(p2.16$volume,p2.16$pressure,method="pearson")

we've got the confidence interval too which is, (0.9999991,0.9999998).

confidence interval:-

The 100(1 $-\alpha$) percent Cl on β_1 is $\widehat{\beta_1} - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{MS_{\text{Res}}}{\sum_{i=1}^{n} x_i^2}} \le \beta_1 \le \widehat{\beta_1} + t\frac{\alpha}{2}, n-1\sqrt{\frac{MS_{\text{Res}}}{\sum_{i=1}^{n} x_i^2}}$

The confidence interval can be found in R programming by using "confint()" function. So by the code "confint(pressure.lm,'volume',level=0.95)" the 95% confidence interval of slope is (2.3456,2.3472). Since, the confidence interval of the slope(volume) does not contain 0 , so it can be concluded there is a significant evidence of a linear relationship between volume of liquid and pressure of tank.

By the above tests and scatterplot we can conclude that there is almost a perfect linear fit of the data.

11) On March 1, 1984, the Wall Street Journal published a survey of television advertisements conducted by Video Board Tests, Inc., a New York ad - testing company that interviewed 4000 adults. These people were regular product users who were asked to cite a commercial they had seen for that product category in the past week. In this case, the response is the number of millions of retained impressions per week. The regressor is the amount of money spent by the firm on advertising. The data follow.

Firm	Amount spent(millions)	Returned impressions per week(millions)
Miller Lite	50.1	32.1
Pepsi	74.1	99.6
Stroh's	19.3	11.7
Federal Express	22.9	21.9
Burger King	82.4	60.8
Coca-Cola	40.1	78.6
McDonald's	185.9	92.4
MCI	26.9	50.7
Diet Cola	20.4	21.4

Firm	Amount spent(millions)	Returned impressions per week(millions)
Ford	166.2	40.1
Levi's	27	40.8
Bud Lite	45.6	10.4
ATT Bell	154.9	88.9
Calvin Klein	5	12
Wendy's	49.7	29.2
Polaroid	26.9	38
Shasta	5.7	10
Meow Mix	7.6	12.3
Oscar Meyer	9.2	23.4
Crest	32.4	71.1
Kibbles N bits	6.1	4.4

a. Fit the simple linear regression model to these data.

b. Is there a significant relationship between the amount a company spends on advertising and retained impressions? Justify your answer statistically.

c. Construct the 95% confidence and prediction bands for these data.

d. Give the 95% confidence and prediction intervals for the number of retained impressions for MCI.

[120]:

```r
# Load necessary library for reading Excel files
library(readxl)

# Get the current working directory (optional, for reference)
getwd()

# Read the Excel file into a data frame
d <- read_excel("C:/Users/SNEHANGSHU/Desktop/R code/Book1.xlsx")

# View the first few rows of the dataset
head(d)

# Extract the relevant columns for analysis
amount <- d$`Amount spent(millions)`
impression <- d$`Returned impressions per week(millions)`
```

```
'C:/Users/SNEHANGSHU/Desktop/R code'
```

A tibble: 6 × 3

Firm	Amount spent(millions)	Returned impressions per week(millions)
<chr>	<dbl>	<dbl>
Miller Lite	50.1	32.1
Pepsi	74.1	99.6
Stroh's	19.3	11.7
Federal Express	22.9	21.9
Burger King	82.4	60.8
Coca-Cola	40.1	78.6

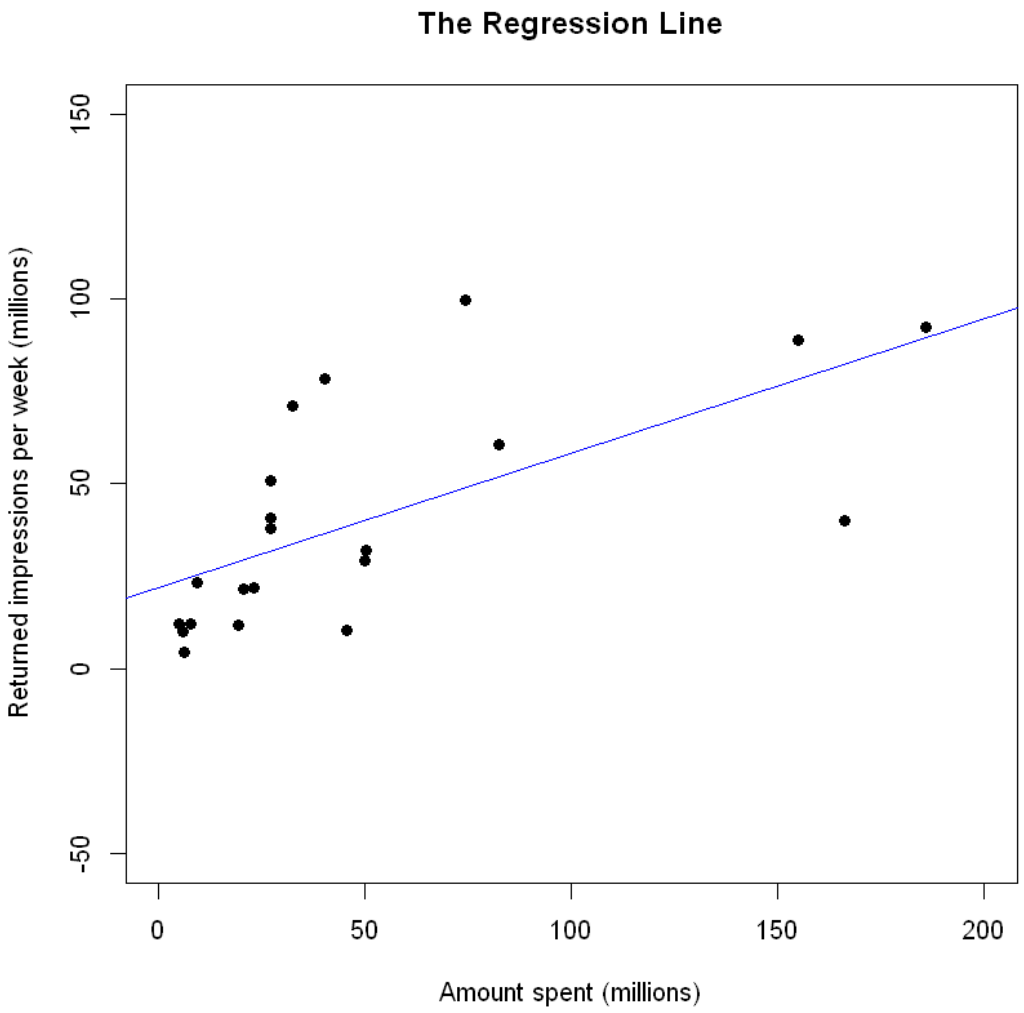
[121]:

```r
# Fit a linear model predicting returned impressions based on the amount spent
amount.lm <- lm(impression ~ amount, data = d)

# View a summary of the linear model
summary(amount.lm)
```

```
Call:
lm(formula = impression ~ amount, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.422 -12.623  -8.171   8.832  50.526 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 22.16269    7.08948   3.126  0.00556 **
amount       0.36317    0.09712   3.739  0.00139 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.5 on 19 degrees of freedom
Multiple R-squared:  0.424,	Adjusted R-squared:  0.3936 
F-statistic: 13.98 on 1 and 19 DF,  p-value: 0.001389
```

[122]:

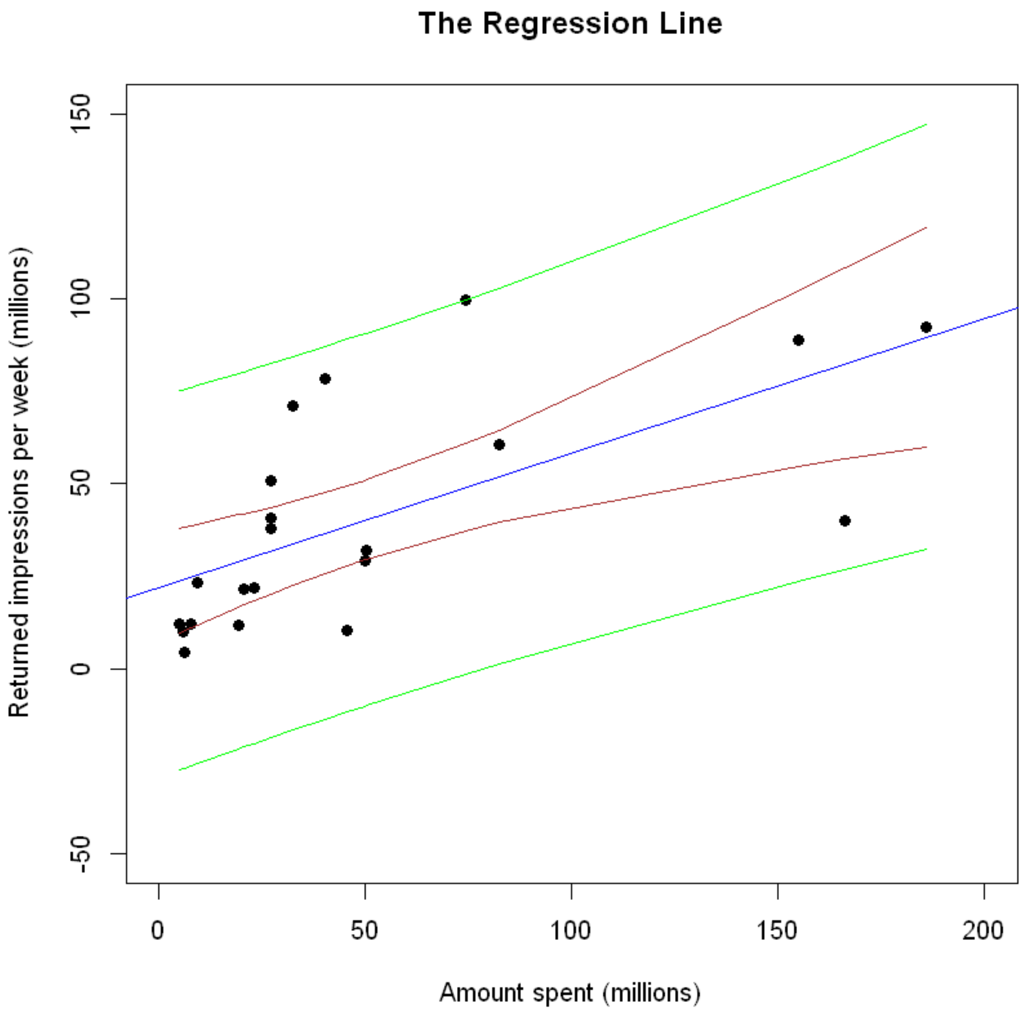
```r
# Create a scatterplot of impressions vs. amount spent
plot(amount, impression,
     main = "The Regression Line",
     xlab = "Amount spent (millions)",
     ylab = "Returned impressions per week (millions)",
     pch = 19,
     ylim = c(-50, 150),
     xlim = c(0, 200))

# Add the regression line to the plot
abline(amount.lm, col = "blue")
```



[124]:
```
# Perform ANOVA on the fitted linear model
print(anova(amount.lm))
```

```
Analysis of Variance Table

Response: impression
          Df  Sum Sq Mean Sq F value   Pr(>F)
amount     1  7723.3  7723.3  13.983 0.001389 **
Residuals 19 10494.1   552.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[137]:
```
# Create a scatterplot of impressions vs. amount spent
plot(amount, impression,
     main = "The Regression Line",
     xlab = "Amount spent (millions)",
     ylab = "Returned impressions per week (millions)",
     pch = 19,
     ylim = c(-50, 150),
     xlim = c(0, 200))

# Add the regression line to the plot
abline(amount.lm, col = "blue")

# Predict confidence intervals for the fitted values
amount.conf <- predict(amount.lm, interval = 'confidence')

# Add confidence interval lines to the plot
lines(sort(amount), amount.conf[order(amount), 2], col = "brown")  #␣
 ↪Lower bound
lines(sort(amount), amount.conf[order(amount), 3], col = "brown")  #␣
 ↪Upper bound
# Predict prediction intervals for the dataset
amount.pred <- predict(amount.lm, newdata = d, interval = 'predict')

# Add prediction interval lines to the plot
lines(sort(amount), amount.pred[order(amount), 2], col = "green")  #␣
 ↪Lower bound
lines(sort(amount), amount.pred[order(amount), 3], col = "green")  #␣
 ↪Upper bound
```



[132]:

```
# Predict confidence interval for a specific value of amount
print(predict(amount.lm, newdata = data.frame(amount = 26.9), interval␣
 ↪= 'confidence', level = 0.95))
```

```
       fit      lwr      upr
1 31.93208 20.18314 43.68102
```

[133]:

```
# Predict prediction interval for a specific value of amount
print(predict(amount.lm, newdata = data.frame(amount = 26.9), interval␣
 ↪= 'predict', level = 0.95))
```

```
       fit       lwr      upr
1 31.93208 -18.64084 82.50499
```

■ **Interpretation:-**

First we have to read the given data in R. because it is a excel file we have to load “readxl” by “library(readxl)” then we saw the working directory of recent workspace. Then we've read the data by the code “d=read_excel(”C:/Users/Hp/Desktop/p2.18.xlsx”)“. The column”Amount spent(millions)” is saved in a variable named amount and the column “Returned impressions per week (millions)” is saved in a variable named impression.

a. According to the question, the regressor is amount of money spent by the firm on advertising and returned impressions per week(millions) is the response variable. In R, by the function ” Im( )” we can get the regression model as follows:-

$$\hat{y} = 22.163 + 0.36317x$$

Where $y =$ returned impressions per week(millions), $x =$ amount of money spent by the firm on advertising(millions).

b. Testing for regression:-

The hypotheses for testing significance of regression is $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$. Failing to reject $\mathrm{H}_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y. Alternatively, if $\mathrm{H}_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$\mathrm{t}_0 = \frac{\widehat{\beta_1}}{\operatorname{se}\left(\widehat{\beta_1}\right)} \tag{31}$$

The null hypothesis of significance of regression would be rejected if $|t_0| > t_{\alpha/2,n-2}$.

Here, $\widehat{\beta_1} = 0.36317$ and $\operatorname{se}\left(\widehat{\beta_1}\right) = 0.09712$. So, $\mathrm{t}_0 = 3.739 \& \mathrm{t}_{0.025,19} = 2.093024$. So, the null hypothesis is rejected since $|t_0| > t_{0.025,19}$. That means, there is a linear relationship between amount spent and returned impression per week.

Analysis of variance:-

We can also use the usual analysis - of - variance F test to test the hypothesis $\mathrm{H}_0 : \beta_1 = 0$. We know that

(1) $SS_{\text{Res}} = (n-2)MS_{\text{Res}} / \sigma^2$ follows a χ^2_{n-2} distribution.

(2) If the null hypothesis $\mathrm{H}_0 : \beta_1 = 0$ is true, then SS_R/σ^2 follows a χ^2_1 distribution

(3) SS_{Res} and SS_R are independent.

By the definition of an F statistic,

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

follows the $F_{1,n-2}$ distribution. We know that the expected values of these mean squares are

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

These expected mean squares indicate that if the observed value of F_0 is large, then it is likely that the slope $\beta_1 \neq 0$. If $\beta_1 \neq 0$, then F_0 follows a noncentral F distribution with 1 and $n-2$ degrees of freedom and a noncentrality parameter of

$$\lambda = \frac{\beta_1^2 S_{xx}}{\sigma^2}$$

This noncentrality parameter also indicates that the observed value of F_0 should be large if $\beta_1 \neq 0$. Therefore, to test the hypothesis $\mathrm{H}_0 : \beta_1 = 0$, compute the test statistic F_0 and reject H_0 if

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

To compute the analysis-of-variance in R, "anova(amount.Im)" gives

Source on variation	sum of squares	df	mean square	F_0
Regression (ratio)	7723.3	1	7723.3	7723.3/552.3
Residual	10494.1	19	552.3	= 13.983

Here $F_0 = 13.983$ and $F_{0.05;1,31} = 4.38075$ by the code in R "qf($0.05, 1, 19$,lower.tail=FALSE)". So, we reject H_0. That means, there is a linear relationship between amount spent and returned impression per week. Also, from the output of "anova(amount.Im)" function p-value is 0.001389 which is less than 0.05 which means we will reject the null hypothesis.

Even from the output of the Im() function in R we can interpret the validation of model.

Coefficient:-

In this section we can see Pr(> $|t|$). For both intercept and volume there are three stars which represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis at 5% level of significance which allows us to conclude that there is a relationship between amount spent and impression.

Residual standard error:-

Residual standard error is a measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term. Due to the presence of this error term we are not capable of perfectly predicting our response variable(impression) from the predictor(amount) one. The residual standard error is the average amount that the response(impression) will deviate from the true regression line.

In this output, impression can deviate from true regression line by approximately 23.5 on average with 19 degrees of freedom. Degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters(restriction).

Multiple R-squared, Adjusted R-squared:-

The R-squared statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. It is a measure of the linear relationship between our predictor variable and our response variable. It always lies between 0 and 1 . A number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable. Adjusted R-squared is calculated by dividing the residual mean square error by the total mean square error. Here adjusted R-squared is 0.424 . It indicates that 42.4% of the variance found in the response variable(impression) can be explained by the predictor variable(amount).

F-statistic:-

F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. How much larger the F-statistic needs to be depend on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis. In this data, F-statistic is 13.98 which is sufficiently larger than 1 . So there is a significant relationship between amount spent on advertising and impressions returned per week.

By the above tests and output we can conclude there is a significant relationship between the amount a company spends on advertising and retained impressions though by $R^2 = 42.4$ there is still a lot of unexplained variation in this model.

c.

The Regression Line

This above figure shows that green lines are the 95% prediction limits and brown lines are 95% confidence limits for amount model. This graph illustrates the point that the prediction interval is wider than confidence interval.

d. Confidence interval:-

A 100(1 $-\alpha$) percent Cl on the mean response at the point $x = x_0$ I

$$\widehat{\mu_{y|x0}}-t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq E\left(y \mid x_0\right) \leq \widehat{\mu_{y|x0}}+t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Where, $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$ and $\widehat{E(y \mid x_0)} = \widehat{\mu_{y|x0}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$

Note that the width of the Cl for $E\left(y \mid x_0\right)$ is a function of x_0. The interval width is a minimum for $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases. Intuitively this is reasonable, as we would expect our best estimates of y to be made at x values near the center of the data and the precision of estimation to deteriorate as we move to the boundary of the x space.

Prediction interval:-

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

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

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. We now develop a prediction interval for the future observation y_0. Note that the random variable $\Psi = y_0 - \widehat{y_0}$ is normally distributed with mean zero and variance $\operatorname{Var}(\Psi) = \operatorname{Var}(y_0 - \widehat{y_0}) = \sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right]$ because the future observation $\mathrm{y_0}$ is independent of $\widehat{y_0}$. If we use $\widehat{y_0}$ to predict y 0 , then the standard error of $\Psi = y_0 - \widehat{y_0}$ is the appropriate statistic on which to base a prediction interval. Thus, the 100(1- α) percent prediction interval on a future observation at x_0 is

$$\widehat{y_0} - t^{\alpha}_{2,n-2}\sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le y_0 \le \widehat{y_0} + t_{2,n-2}^{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

In R, we can calculate confidence interval and prediction interval for a given point of independent variable. The 95% confidence interval for $x = 26.9$ is $(20.18, 43.68)$

The 95% prediction interval for x = 26.9 is $(-18.64, 82.50)$.

12) Byers and Williams (" Viscosities of Binary and Ternary Mixtures of Polynomatic Hydrocarbons, " Journal of Chemical and Engineering Data , 32,349 - 354, 1987) studied the impact of temperature on the viscosity of toluene tetralin blends. The following table gives the data for blends with a 0.4 molar fraction of toluene.

Temperature (°C)	Viscosity (mPa · s)
24.9	1.1330
35.0	0.9772
44.9	0.8532
55.1	0.7550
65.2	0.6723
75.2	0.6021
85.2	0.5420
95.2	0.5074

a. Estimate the prediction equation.

b. Perform a complete analysis of the model.

c. Calculate and plot the 95% confidence and prediction bands.

[2]:

```r
# Load the necessary library
library("MPV")

# Attach the dataset for easier referencing
```

```r
attach(p2.15)
```

```
Loading required package: lattice

Loading required package: KernSmooth

KernSmooth 2.23 loaded
Copyright M. P. Wand 1997-2009

Loading required package: randomForest

randomForest 4.7-1.1

Type rfNews() to see new features/changes/bug fixes.
```

[3]:

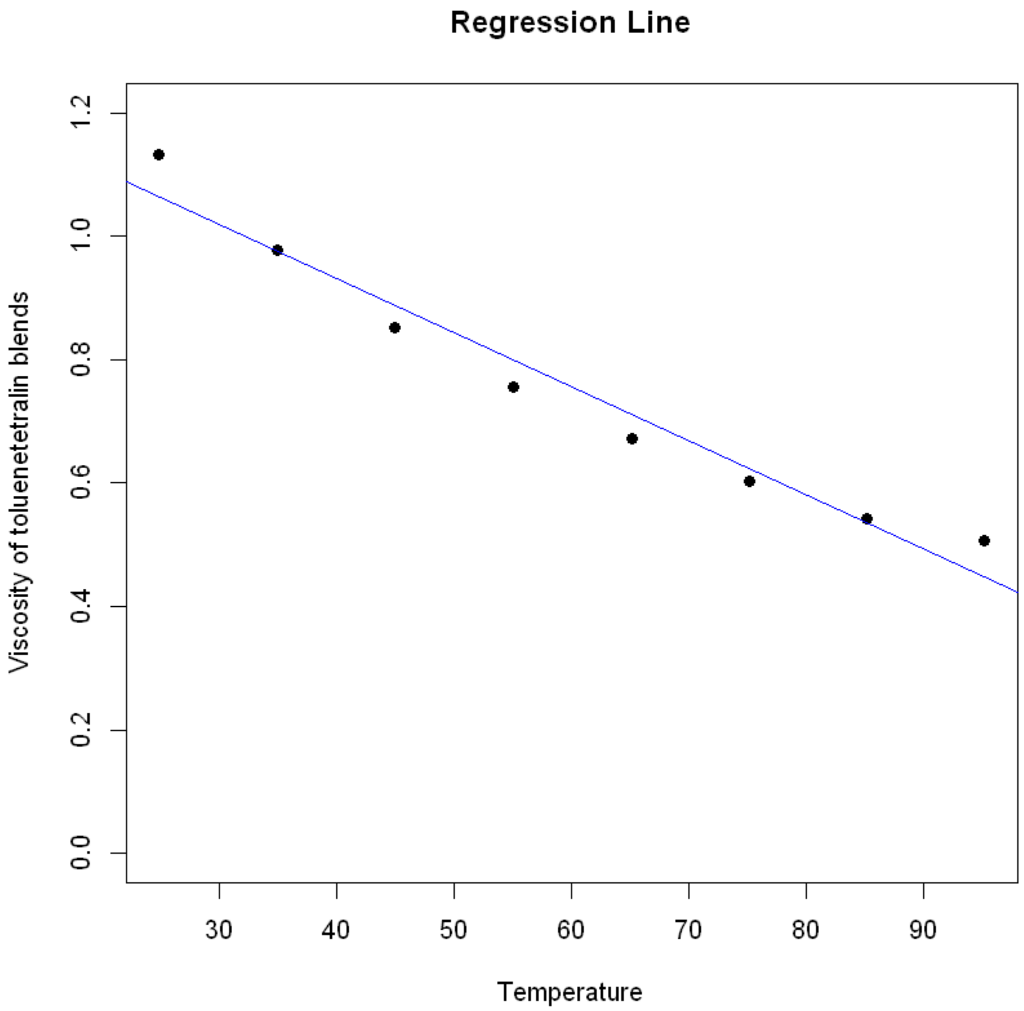
```r
# Fit a linear model predicting viscosity based on temperature
visc.lm <- lm(visc ~ temp, data = p2.15)

# View a summary of the linear model
summary(visc.lm)
```

```
Call:
lm(formula = visc ~ temp, data = p2.15)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.043955 -0.035863 -0.009305  0.019900  0.069559 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.2815107  0.0468683   27.34 1.58e-07 ***
temp        -0.0087578  0.0007284  -12.02 2.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04743 on 6 degrees of freedom
Multiple R-squared:  0.9602,	Adjusted R-squared:  0.9535 
F-statistic: 144.6 on 1 and 6 DF,  p-value: 2.007e-05
```

[4]:

```r
# Create a scatterplot of viscosity vs. temperature
plot(p2.15$temp, p2.15$visc,
     main = "Regression Line",
     xlab = "Temperature",
     ylab = "Viscosity of toluenetetralin blends",
     pch = 16,
     ylim = c(0, 1.2))
```

```
# Add the regression line to the plot
abline(visc.lm, col = "blue")
```



[5]:
```
# Perform ANOVA on the fitted linear model
print(anova(visc.lm))
```

```
Analysis of Variance Table

Response: visc
          Df  Sum Sq Mean Sq F value    Pr(>F)    
temp       1 0.32529 0.32529  144.58 2.007e-05 ***
Residuals  6 0.01350 0.00225                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

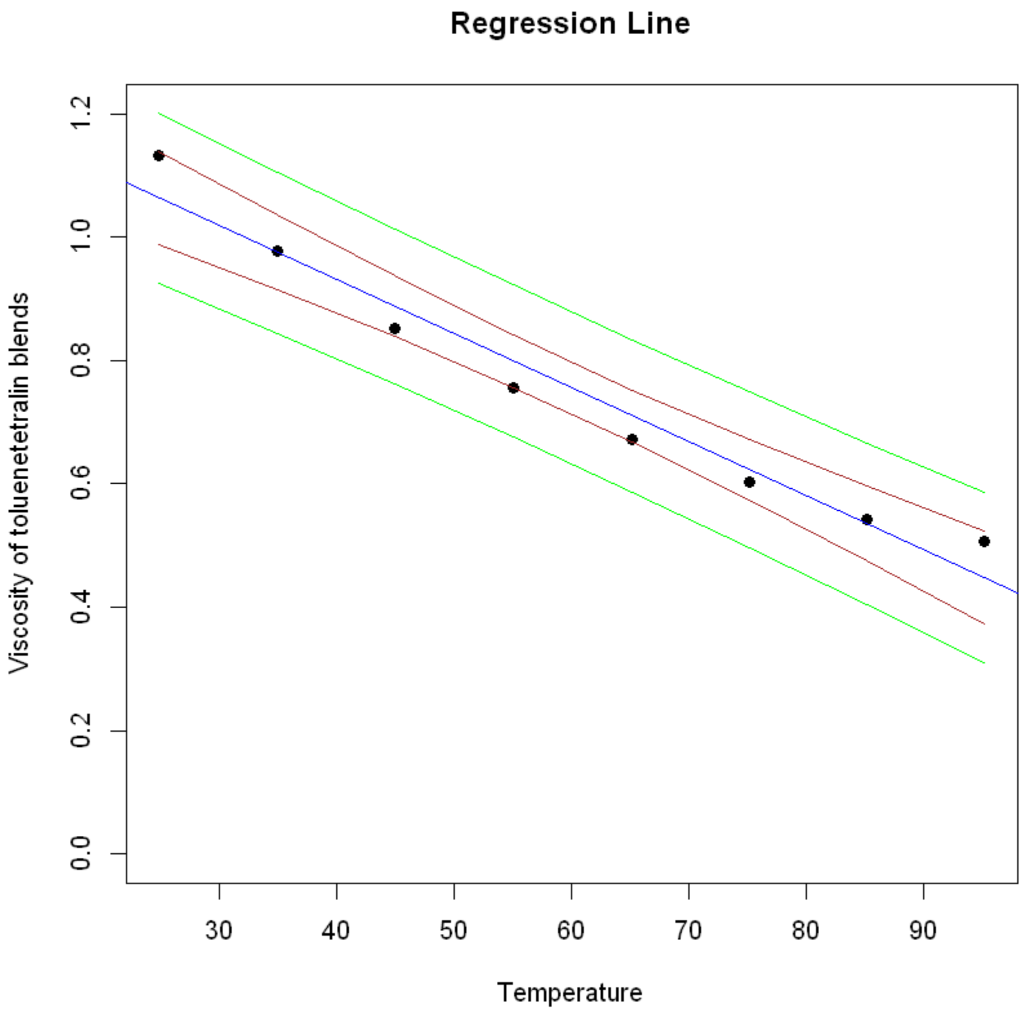
[6]:

```
# Perform Pearson correlation test between temperature and viscosity
cor.test(p2.15$temp, p2.15$visc, method = "pearson")
```

```

	Pearson's product-moment correlation

data:  p2.15$temp and p2.15$visc
t = -12.024, df = 6, p-value = 2.007e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9964840 -0.8891539
sample estimates:
       cor
-0.9798742
```

[163]:

```
# Create a scatterplot of viscosity vs. temperature
plot(p2.15$temp, p2.15$visc,
     main = "Regression Line",
     xlab = "Temperature",
     ylab = "Viscosity of toluenetetralin blends",
     pch = 16,
     ylim = c(0, 1.2))

# Add the regression line to the plot
abline(visc.lm, col = "blue")

# Predict confidence intervals for the fitted values
visc.conf <- predict(visc.lm, interval = 'confidence')

# Add confidence interval lines to the plot
lines(sort(temp), visc.conf[order(temp), 2], col = "brown")  # Lower␣
 ↪bound
lines(sort(temp), visc.conf[order(temp), 3], col = "brown")  # Upper␣
 ↪bound
# Predict prediction intervals for the dataset
visc.pred <- predict(visc.lm, newdata = p2.15, interval = 'predict',␣
 ↪level = 0.95)


# Add prediction interval lines to the plot
lines(sort(temp), visc.pred[order(temp), 2], col = "green")  # Lower␣
 ↪bound
lines(sort(temp), visc.pred[order(temp), 3], col = "green")  # Upper␣
 ↪bound
```



```
[15]: # Display the confidence intervals
      visc.conf
```

A matrix: 8 × 3 of type dbl

	fit	lwr	upr
1	1.0634409	0.9884960	1.1383857
2	0.9749869	0.9142988	1.0356750
3	0.8882844	0.8391259	0.9374430
4	0.7989546	0.7569675	0.8409418
5	0.7105006	0.6684657	0.7525355
6	0.6229224	0.5738373	0.6720075
7	0.5353442	0.4746233	0.5960651
8	0.4477660	0.3729329	0.5225990

```
[16]: # Display the prediction intervals
      visc.pred
```

A matrix: 8 × 3 of type dbl

	fit	lwr	upr
1	1.0634409	0.9252815	1.2016002
2	0.9749869	0.8440124	1.1059614
3	0.8882844	0.7622374	1.0143314
4	0.7989546	0.6755277	0.9223815
5	0.7105006	0.5870575	0.8339438
6	0.6229224	0.4969041	0.7489408
7	0.5353442	0.4043545	0.6663339
8	0.4477660	0.3096672	0.5858647

■ **Interpretation:-**

a. According to the question, the regressor is temperature and viscosity of toluene tetralin blends is the response variable. In R, by the function ” lm()″ ” we can get the regression model as follows:-

$$\hat{y} = 1.2815 - 0.0088x$$

Where $y =$ viscosity of toluene tetralin blends, $x =$ temperature.

b. Testing for regression:-

The hypotheses for testing significance of regression is $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$. Failing to reject $\mathrm{H}_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y. Alternatively, if $\mathrm{H}_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. The test procedure is

$$t_0 = \frac{\widehat{\beta_1}}{\operatorname{se}\left(\widehat{\beta_1}\right)}$$

The null hypothesis of significance of regression would be rejected if $|t_0| > t_{\alpha/2,n-2}$.

Here, $\widehat{\beta_1} = -0.0087578$ and $\operatorname{se}\left(\widehat{\beta_1}\right) = 0.0007284$. So, $\mathrm{t}_0 = -12.0233\&\mathrm{t}_{0.025,6} = 2.446912$. So, the null hypothesis is rejected since $|t_0| > t_{0.025,6}$. That means, there is a linear relationship between temperature and viscosity of toluene tetralin blelnds.

Analysis of variance:-

We can also use the usual analysis - of - variance F test to test the hypothesis $\mathrm{H}_0 : \beta_1 = 0$. We know that

(1) $SS_{\text{Res}} = (n-2)MS_{\text{Res}}/\sigma^2$ follows a χ^2_{n-2} distribution.

(2) If the null hypothesis $\mathrm{H}_0 : \beta_1 = 0$ is true, then SS_R/σ^2 follows a χ^2_1 distribution

(3) SS_{Res} and SS_R are independent.

By the definition of an F statistic,

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

follows the $F_{1,n-2}$ distribution. We know that the expected values of these mean squares are

$$E\left(MS_{\text{Res}}\right)=\sigma^{2}, \quad E\left(MS_{R}\right)=\sigma^{2}+\beta_{1}^{2} S_{xx}$$

These expected mean squares indicate that if the observed value of F_0 is large, then it is likely that the slope $\beta_1 \neq 0$. If $\beta_1 \neq 0$, then F_0 follows a noncentral F distribution with 1 and $n-2$ degrees of freedom and a noncentrality parameter of

$$\lambda=\frac{\beta_{1}^{2} S_{xx}}{\sigma^{2}}$$

This noncentrality parameter also indicates that the observed value of F_0 should be large if $\beta_1 \neq 0$. Therefore, to test the hypothesis $\mathrm{H}_0 : \beta_1 = 0$, compute the test statistic F_0 and reject H_0 if

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

To compute the analysis-of-variance in R, "anova(amount.Im)" gives

Source on variation	sum of squares	df	mean square	F_0
Regression (ratio)	0.32529	1	0.32529	0.32529/0.00225
Residual	0.01350	6	0.00225	= 144.58

Here $F_0 = 144.58$ and $F_{0.05;1,6} = 5.987378$ by the code in R " $qf(0.05,1,6,$ lower.tail=FALSE)". So, we reject H_0. That means, there is a linear relationship between temperature and viscosity of toluene tetralin blends. Also, from the output of"anova(visc. Im)" function p-value is 0.00002 which is less than 0.05 which means we will reject the null hypothesis.

Even from the output of the Im() function in R we can interpret the validation of model.

Coefficient:-

In this section we can see $\Pr(> |t|)$. For both intercept and volume there are three stars which represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis at 5% level of significance which allows us to conclude that there is a relationship between temperature and viscosity of toluene tetralin blends.

Residual standard error:-

Residual standard error is a measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term. Due to the presence of this error term we are not capable of perfectly predicting our response variable(visc) from the predictor(temp) one. The residual standard error is the average amount that the response(visc) will deviate from the true regression line.

In this output, viscosity of toluene tetralin blends can deviate from true regression line by approximately 0.05 on average with 6 degrees of freedom. Degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters(restriction).

Multiple R-squared, Adjusted R-squared:-

The R-squared statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. It is a measure of the linear relationship between our predictor variable and our response variable. It always lies between 0 and 1. A number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable. Adjusted R-squared is calculated by dividing the residual mean square error by the total mean square error. Here adjusted R-squared is 0.9535 . It indicates that 95.4% of the variance found in the response variable(viscosity) can be explained by the predictor variable(temperature).

F-statistic:-

F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. How much larger the F-statistic needs to be depend on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis. In this data, F-statistic is 144.58 which is sufficiently larger than 1 . So there is a significant relationship between amount spent on advertising and impressions returned per week.

By the above tests and output we can conclude there is a significant relationship between temperature and viscosity of toluene tetralin blends.

c. Confidence interval:-

A 100(1- α) percent Cl on the mean response at the point $x = x_0$ I

$$\widehat{\mu_{y|x0}} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq E\left(y \mid x_0\right) \leq \widehat{\mu_{y|x0}} + t\frac{\alpha}{2}, n-2\sqrt{MS_{\text{Res}}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Where, $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$ and $E\widehat{(y \mid x_0)} = \widehat{\mu_{y|x0}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$

Note that the width of the Cl for $E\left(y \mid x_0\right)$ is a function of x_0. The interval width is a minimum for $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases. Intuitively this is reasonable, as we would expect our best estimates of y to be made at x values near the center of the data and the precision of estimation to deteriorate as we move to the boundary of the x space.

Prediction interval:-

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

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

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. We now develop a prediction interval for the future observation y_0. Note that the random variable $\Psi = y_0 - \widehat{y_0}$ is normally distributed with mean zero

$$\text{and variance} \qquad \text{Var}(\Psi) = \text{Var}\left(y_0 - \widehat{y_0}\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]$$

because the future observation y_0 is independent of $\widehat{y_0}$. If we use $\widehat{y_0}$ to predict y 0 , then the standard error of $\Psi = y_0 - \widehat{y_0}$ is the appropriate statistic on which to base a prediction interval. Thus, the 100(1- α) percent prediction interval on a future observation at x_0 is

$$\widehat{y_0} - t^{\alpha}_{\frac{1}{2},n-2} \sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \widehat{y_0} + t^{\alpha}_{\frac{1}{2},n-2} \sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

In R, confidence and prediction band can be drawn. First, we need to calculate confidence and prediction intervals using “predict()” function. Then by “lines()” function the bands can be drawn as follows:

Regression line

This above figure shows that green lines are the 95% prediction limits and brown lines are 95% confidence limits for amount model. This graph illustrates the point that the prediction interval is wider than confidence interval.

2 2)MULTIPLE LINEAR REGRESSION

13) The concentration of NbOCl in a tube-flow reactor as a function of several controllable variables is shown in Table B.6.

Run No.	y	x	x	x	x	Run No.	y	x	x	x	x
1	0.000450	0.0105	90.9	0.0164	0.0177	15	0.001364	0.0110	516.4	0.0190	0.0161
2	0.000450	0.0110	84.6	0.0165	0.0172	16	0.001254	0.0117	488.0	0.0189	0.0149
3	0.000473	0.0106	88.9	0.0164	0.0157	17	0.001396	0.0110	534.5	0.0189	0.0163
4	0.000507	0.0116	488.7	0.0187	0.0082	18	0.001575	0.0104	542.3	0.0189	0.0164
5	0.000457	0.0121	454.4	0.0187	0.0070	19	0.001615	0.0067	98.8	0.0163	0.0379
6	0.000452	0.0123	439.2	0.0187	0.0065	20	0.001733	0.0066	84.8	0.0162	0.0360
7	0.000453	0.0122	447.1	0.0186	0.0071	21	0.002753	0.0044	69.6	0.0163	0.0327

Run No.	y	x	x	x	x	Run No.	y	x	x	x	x
8	0.000426	0.0122	451.6	0.0187	0.0062	22	0.003186	0.0073	436.9	0.0189	0.0263
9	0.001215	0.0123	487.8	0.0192	0.0153	23	0.003227	0.0078	406.3	0.0192	0.0200
10	0.001256	0.0122	467.6	0.0192	0.0129	24	0.003469	0.0067	447.9	0.0192	0.0197
11	0.001145	0.0094	95.4	0.0163	0.0354	25	0.001911	0.0091	58.5	0.0164	0.0331
12	0.001085	0.0100	87.1	0.0162	0.0342	26	0.002588	0.0079	394.3	0.0177	0.0674
13	0.001066	0.0101	82.7	0.0162	0.0323	27	0.002635	0.0068	461.0	0.0174	0.0770
14	0.001111	0.0099	87.0	0.0163	0.0337	28	0.002725	0.0065	469.2	0.0173	0.0780

- **y**: $NbOCl_3$ concentration (g – mol/l)
- **x**: $COCl_2$ concentration (g – mol/l)
- **x**: Space time (sec)
- **x**: Molar density (g – mol/l)
- **x**: Mole fraction CO_2

a. Fit a multiple regression model relating concentration of $NbOCl_3$ (y) to concentration of COCl (x) and mole fraction (x).

b. Test for the significance of regression.

c. Calculate R^2 and adjusted R^2 for this model.

d. Using t-tests, determine the contribution of x and x to the model. Are both regressors x and x necessary?

e. Is multicollinearity a potential concern in this model?

[18]:

```r
library(MPV)
attach(table.b6)
model=lm(y~x1+x4,data=table.b6)
summary(model)
```

```
Call:
lm(formula = y ~ x1 + x4, data = table.b6)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0009015 -0.0003526 -0.0001538  0.0003847  0.0010874 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.0048333  0.0008142   5.936 3.39e-06 ***
x1          -0.3449837  0.0673963  -5.119 2.74e-05 ***
x4          -0.0001430  0.0078151  -0.018    0.986    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0005804 on 25 degrees of freedom
```

```
Multiple R-squared:  0.6636,       Adjusted R-squared:  0.6367
F-statistic: 24.66 on 2 and 25 DF,  p-value: 1.218e-06
```

■ **Interpretation:-**

a. From the R code, it can be seen that the multiple regression relating concentration of NbOCl3(y) to concentration of COCl 2 (x_1) and mole fraction (x_4) is

$$y = 0.0048 - 0.345x_1 - 0.000143x_4$$

b. F-statistic = 24.66 on 2 and 25 df with p -value = 1.218e − 06 which is greater than $F_{0.05;2,25} = 3.3852$ implies that the regression is significant.

c. It can be seen that

$$R^2 = 0.6636 \text{ and } R^2_{Adj} = 0.6367$$

d. From the summary of the fit, we have **** for slope of $x1$ which means it is significant but slope for $x4$ is not significant. Here, regressor $x1$ is necessary but regressor $x4$ is not necessary at all.

e. The model is significant with $R2 = 0.6636$ which is not very high. Also the correlation between x 1 and $x4$ is -0.6865 . So, multicollinearity does not seems to be a potential concern in this model.

14) The kinematic viscosity of a certain solvent system depends on the ratio of the two solvents and the temperature. The table 7 on the next page summarizes a set of experimental results.

a. Fit a multiple linear regression model relating the viscosity to these regressors.

b. Test for significance of regression. What conclusions can you draw?

c. Use t-test to access the contribution of each regressor to the model. Discuss your findings.

d. Calculate R^2 and ${R_{Adj}}^2$. Compare these values to the R^2 and ${R_{Adj}}^2$ for the simple linear regression model relating viscosity to temperature only. Discuss your results.

e. Find a 99%Cl for the regression coefficient for x_2 for both models in part d. Discuss any differences.

TABLE B. 10 Kinematic Viscosity Data

x_1	x_2	y
0.9189	-10	3.128
0.9189	0	2.427
0.9189	10	1.94
0.9189	20	1.586
0.9189	30	1.325
0.9189	40	1.126
0.9189	50	0.9694

x_1	x_2	y
0.9189	60	0.8473
0.9189	70	0.7481
0.9189	80	0.6671
0.7547	-10	2.27
0.7547	0	1.819
0.7547	10	1.489
0.7547	20	1.246
0.7547	30	1.062
0.7547	40	0.916
0.7547	50	0.8005
0.7547	60	0.7091
0.7547	70	0.6345
0.7547	80	0.5715
0.5685	-10	1.593
0.5685	0	1.324
0.5685	10	1.118
0.5685	20	0.9576
0.5685	30	0.8302
0.5685	40	0.7282
0.5685	50	0.647
0.5685	60	0.5784
0.5685	70	0.5219
0.5685	80	0.4735
0.361	-10	1.161
0.361	0	0.9925
0.361	10	0.8601
0.361	20	0.7523
0.361	30	0.6663
0.361	40	0.594
0.361	50	0.5338
0.361	60	0.4804
0.361	70	0.4361
0.361	80	0.4016

y : Kinematic viscosity (10^{-6} m^2/s).

x_1 : Ratio of 2-methoxyethanol to 1,2-dimethoxyethane (dimensionless).

x_2 : Temperature (°C).

Source: "Viscosimetric Studies on 2-Methoxyethanol + 1, 2-Dimethoxyethane Binary Mixtures from -10 to 80°C," Canadian Journal of Chemical Engineering, 75, 494-501.

[20]:

```
# Load the necessary library
library(MPV)

# Attach the dataset for easier referencing
```

```
attach(table.b10)
```

```
The following objects are masked from table.b10 (pos = 3):

    x1, x2, y


The following objects are masked from table.b6:

    x1, x2, y
```

[24]:
```
# Fit a linear model predicting y based on both x1 and x2
model <- lm(y ~ x1 + x2, data = table.b10)

# View a summary of the linear model
print(summary(model))
```

```
Call:
lm(formula = y ~ x1 + x2, data = table.b10)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.22179 -0.18102 -0.08439  0.09111  0.99908 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.679439   0.143532   4.734 3.20e-05 ***
x1           1.407331   0.196925   7.147 1.81e-08 ***
x2          -0.015629   0.001428 -10.948 3.67e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2593 on 37 degrees of freedom
Multiple R-squared:  0.822,	Adjusted R-squared:  0.8124 
F-statistic: 85.46 on 2 and 37 DF,  p-value: 1.351e-14
```

[25]:
```
# Fit a linear model predicting y based only on x2
model1 <- lm(y ~ x2, data = table.b10)

# View a summary of the reduced linear model
print(summary(model1))
```

```
Call:
lm(formula = y ~ x2, data = table.b10)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.60279 -0.27348 -0.02503  0.19194  1.37642 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.595295   0.098404  16.212  < 2e-16 ***
x2          -0.015629   0.002173  -7.191 1.36e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3948 on 38 degrees of freedom
Multiple R-squared:  0.5764,	Adjusted R-squared:  0.5653 
F-statistic: 51.71 on 1 and 38 DF,  p-value: 1.358e-08
```

```
[26]: # Calculate the 99% confidence interval for the coefficient of x2 in␣
       ↪the full model
      conf_interval_full <- confint(model, 'x2', level = 0.99)
      print(conf_interval_full)

      # Calculate the 99% confidence interval for the coefficient of x2 in␣
       ↪the reduced model
      conf_interval_reduced <- confint(model1, 'x2', level = 0.99)
      print(conf_interval_reduced)
```

```
         0.5 %      99.5 %
x2 -0.01950537 -0.01175233
        0.5 %       99.5 %
x2 -0.0215221 -0.009735601
```

■ **Interpretation:-**

a. It is seen that a multiple linear regression model relating this dimensionless number to these regressors willbe $y = 0.679439 + 1.407331 \times x_1 - 0.015629 \times x_2$

b. F-statistic $= 85.46$ with 2 and 37 degree of freedom and p-value $= 1.351\text{e} - 14$ implies that the regression is highly significant.

c. Both the regressor x_1 and x_2 contribute significantly to the model.

d. For multiple linear regressor using all of the regressor,

$$R^2 = 0.822 \text{ and } {R_{Adj}}^2 = 0.8124$$

The simple linear regression relating viscosity to temperature will be (from Fig. 24)

$$y = 1.595 - 0.0156 \times x_2$$

and the corresponding R-squared values are $R^2 = 0.5764$ and ${R_{Adj}}^2 = 0.5653$

In this case, the R-squared values differ significantly suggesting the need of the other regressor for better fit.

e. 99% confidence interval for temperature under multiple linear regression is $(-0.01950537, 0.01175233\ )$

99% confidence interval for temperature in the simple linear regression is $(-0.0215221, -0.009735601\ )$. Clearly, the 99%Cl for temperature is wider under simple linear regression model than that of multiple regression model.

15) The table below presents the test - firing results for 25 surface - to - air antiaircraft missiles at targets of varying speed. The result of each test is either a hit ($y = 1$) or a miss ($y = 0$).

Test	Target Speed, x (knots)	y	Test	Target Speed, x (knots)	y
1	400	0	14	330	1
2	220	1	15	280	1
3	490	0	16	210	1
4	210	1	17	300	1
5	500	0	18	470	1
6	270	0	19	230	0
7	200	1	20	430	0
8	470	0	21	460	0
9	480	0	22	220	1
10	310	1	23	250	1
11	240	1	24	200	1
12	490	0	25	390	0
13	420	0			

a. Fit a logistic regression model to the response variable y. Use a simple linear regression model as the structure for the linear predictor.

b. Does the model deviance indicate that the logistic regression model from part a is adequate?

c. Provide an interpretation of the parameter $\beta 1$ in this model.

d. Expand the linear predictor to include a quadratic term in target speed. Is there any evidence that this quadratic term is required in the model?

[29]:

```r
# Load the necessary library
library(MPV)

# Attach the dataset for easier referencing
attach(p13.1)
```

```
The following objects are masked from p13.1 (pos = 3):
    x, y


The following objects are masked from p13.1 (pos = 4):

    x, y


The following object is masked from table.b10 (pos = 5):

    y


The following object is masked from table.b10 (pos = 6):

    y


The following object is masked from table.b6:

    y
```

[38]:

```r
# Fit a binomial logistic model with x as the predictor
model <- glm(y ~ x, family = "binomial")

# View a summary of the model
print(summary(model))
```

```
Call:
glm(formula = y ~ x, family = "binomial")

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.070884   2.108996   2.879  0.00399 **
x           -0.017705   0.006076  -2.914  0.00357 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.617  on 24  degrees of freedom
Residual deviance: 20.364  on 23  degrees of freedom
AIC: 24.364

Number of Fisher Scoring iterations: 4
```

[33]:

```
# Display the model coefficients
print(model$coefficients)
```

```
(Intercept)           x
  6.0708839  -0.0177047
```

[34]:

```
# Calculate and display the deviance of the model
dev <- model$deviance
print(dev)
```

```
[1] 20.36366
```

[35]:

```
# Calculate degrees of freedom
df <- length(x) - 2

# Calculate and display the deviance per degree of freedom
print(dev / df)
```

```
[1] 0.8853764
```

[36]:

```
# Calculate and display the p-value using the chi-square test
p_value <- pchisq(dev, df, lower.tail = FALSE)
print(p_value)
```

```
[1] 0.6198867
```

[37]:

```
# Calculate and display the odds ratio for x
odds_ratio <- exp(model$coefficients[2])
print(odds_ratio)
```

```
        x
0.9824511
```

[39]:

```
# Fit a binomial logistic model with a polynomial term (quadratic)
model1 <- glm(y ~ poly(x, 2), family = "binomial")

# Display the model coefficients
print(model1$coefficients)
```

```
(Intercept) poly(x, 2)1 poly(x, 2)2
 0.07622163 -9.72828252  0.04030119
```

[40]:

```
# Calculate and display the deviance of the model
dev1 <- model1$deviance
print(dev1)
```

```
[1] 20.36346
```

[41]:

```
# Calculate degrees of freedom for the quadratic model
df1 <- length(x) - 3
```

```
# Calculate and display the deviance per degree of freedom for the␣
 ↪quadratic model
print(dev1 / df1)
```

```
[1] 0.925612
```

[42]:
```
# Calculate and display the p-value using the chi-square test for the␣
 ↪quadratic model
p_value1 <- pchisq(dev1, df1, lower.tail = FALSE)
print(p_value1)
```

```
[1] 0.5603156
```

■ **Interpretation:-**

a. The fitted logistic model is to the response y is

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{-6.07 + 0.0177 \times x}}$$

where x is the target speed in knots and the linear predictor can be expressed as

$$\eta = x\beta = \beta_0 + \beta_1 x$$

b. Model deviance, $D = 20.36366$ with $df = n - p = 25 - 2 = 23$. The p-value is 0.6198867 and the ratio D/(n − p) is 0.8853764 (less than unity), so there is no apparent reason to doubt the adequacy of the fit.

c. We obtain the odds ratio as

$$O^\wedge{}_R = e^{6^\wedge 1} = e^{-0.0177} = 0.98245$$

This implies that for every additional knot in the speed of the target, the odds of hitting the target is decreased by 1.755%. So if the speed of the target is increased by 10 knots, then the odd ratio becomes $e^{-0.0177} = 0.8377$. This indicates that the odd of hitting the target is reduced by 16.225%.

d. Now let us consider the model

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{\beta_0 + \beta_1 x + \beta_2 x^2}}$$

where the linear predictor has quadratic term in target speed, expressed as $\eta = \mathbf{x}b = b_0 + b_1 x + b_2 x^2$. From the R output, we find that the deviance for the full model is $D(B) = 20.36366$ with df $= 22$. The reduced model was originally fit whose deviance was $D(B1) = 20.36366$ with $df = 23$. Since the difference of the deviance is basically zero, it indicates that the quadratic term is not required in the model.

16) A study was conducted attempting to relate home ownership to family income. Twenty households were selected and family income was estimated, along with infor-

mation concerning home ownership (y = 1 indicates yes and y = 0 indicates no). The data are shown below.

Household	Income	Home Ownership Status	Household	Income	Home Ownership Status
1	38,000	0	11	38,700	1
2	51,200	1	12	40,100	0
3	39,600	0	13	49,500	1
4	43,400	1	14	38,000	0
5	47,700	0	15	42,000	1
6	5,000	0	16	54,000	1
7	4,500	1	17	51,700	1
8	4,800	0	18	39,400	0
9	45,400	1	19	40,900	0
10	52,400	1	20	52,800	1

a. Fit a logistic regression model to the response variable y. Use a simple linear regression model as the structure for the linear predictor.

b. Does the model deviance indicate that the logistic regression model from part a is adequate?

c. Provide an interpretation of the parameter $\beta 1$ in this model.

d. Expand the linear predictor to include a quadratic term in income. Is there any evidence that this quadratic term is required in the model?

[43]:

```
# Load the necessary library
library(MPV)

# Attach the dataset for easier referencing
attach(p13.2)
```

```
The following objects are masked from p13.1 (pos = 3):

    x, y


The following objects are masked from p13.1 (pos = 4):

    x, y


The following objects are masked from p13.1 (pos = 5):

    x, y


The following object is masked from table.b10 (pos = 6):
    y


The following object is masked from table.b10 (pos = 7):

    y


The following object is masked from table.b6:

    y
```

[46]:

```
# Fit a binomial logistic model with x as the predictor
model <- glm(y ~ x, family = "binomial")

# View a summary of the model
print(summary(model))
```

```
Call:
glm(formula = y ~ x, family = "binomial")

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.7395139  4.4394326  -1.969   0.0490 *
x            0.0002009  0.0001006   1.998   0.0458 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.526  on 19  degrees of freedom
Residual deviance: 22.435  on 18  degrees of freedom
AIC: 26.435

Number of Fisher Scoring iterations: 4
```

[47]:

```
# Display the model coefficients
print(model$coefficients)
```

```
  (Intercept)             x
-8.7395139021  0.0002009056
```

[48]:

```
# Calculate and display the deviance of the model
dev <- model$deviance
```

```
print(dev)
```

```
[1] 22.43492
```

[49]:
```
# Calculate degrees of freedom
df <- length(x) - 2

# Calculate and display the deviance per degree of freedom
print(dev / df)
```

```
[1] 1.246385
```

[50]:
```
# Calculate and display the p-value using the chi-square test
p_value <- pchisq(dev, df, lower.tail = FALSE)
print(p_value)
```

```
[1] 0.2132469
```

[45]:
```
# Calculate and display the odds ratio for x
odds_ratio <- exp(model$coefficients[2])
print(odds_ratio)
```

```
       x
1.000201
```

[51]:
```
# Fit a binomial logistic model with a polynomial term (quadratic)
model1 <- glm(y ~ poly(x, 2), family = "binomial")

# Display the model coefficients
print(model1$coefficients)
```

```
(Intercept) poly(x, 2)1 poly(x, 2)2
   0.224427    5.127813   -2.561793
```

[52]:
```
# Calculate and display the deviance of the model
dev1 <- model1$deviance
print(dev1)
```

```
[1] 21.32632
```

[53]:
```
# Calculate degrees of freedom for the quadratic model
df1 <- length(x) - 3

# Calculate and display the deviance per degree of freedom for the␣
 ↪quadratic model
print(dev1 / df1)
```

```
[1] 1.25449
```

[54]:
```
# Calculate and display the p-value using the chi-square test for the␣
 ↪quadratic model
```

```
p_value1 <- pchisq(dev1, df1, lower.tail = FALSE)
print(p_value1)
```

```
[1] 0.2120329
```

■ **Interpretation:-**

a. The fitted logistic model is to the response y is

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{8.74-0.00}}$$

where x is the target speed in knots and the linear predictor can be expressed as

$$\eta = \mathbf{x}\beta = \beta_0 + \beta_1 x$$

b. Model deviance, $D = 22.43492$ with $df = n - p = 20 - 2 = 18$. The p-value is 0.2132469 and the ratio $D/(n-p)$ is 1.246385 , so there is a reason to doubt the adequacy of the fit.

c. We obtain the odds ratio as

$$O^{\wedge}_{R} = e^{B^{\wedge}1} = e^{-0.0002} = 1.002001$$

This implies that for every additional knot in the family income, the odds of home ownership is decreased by 0.002%. So if the family income is increased by 10 knots, then the odd ratio becomes $e^{-0.002} = 1.002$. This indicates that the odd of being the home ownership is reduced by 16.225%.

d. Now let us consider the model

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{\beta_0 + \beta_1 x + \beta_2 x^2}}$$

where the linear predictor has quadratic term in target speed, expressed as $\eta = \mathbf{x}\beta = b_0 + b_1 x + b_2 x^2$. From the R output, we find that the deviance for the full model is $D(B) = 21.32632$ with df $= 17$ The reduced model was originally fit whose deviance was $D(61) = 22.43492$. with $df = 18$. Since the difference of the deviance is very little, it indicates that the quadratic term is not required in the model.

17) Using Hald cement data , find the eigenvector associated with the smallest eigen value of $X'X$. Interprete the elements of this vector. What can you say about the source of multicollinearity in these data?

[55]:
```
# Load the necessary library
library(corrplot)

# Define the data vectors
y <- c(78.5, 74.3, 104.3, 87.6, 95.9, 109.2, 102.7, 72.5, 93.1, 115.9,␣
 ↪83.8, 113.3, 109.4)
x1 <- c(7, 1, 11, 11, 7, 11, 3, 1, 2, 21, 1, 11, 10)
```

```
x2 <- c(26, 29, 56, 31, 52, 55, 71, 31, 54, 47, 40, 66, 68)
x3 <- c(6, 15, 8, 8, 6, 9, 17, 22, 18, 4, 23, 9, 8)
x4 <- c(60, 52, 20, 47, 33, 22, 6, 44, 22, 26, 34, 12, 12)
```

```
corrplot 0.92 loaded
```

[57]:

```
# Create a data frame
Hald_cement_data <- data.frame(y, x1, x2, x3, x4)

# Print the data frame
print(Hald_cement_data)
```

```
       y x1 x2 x3 x4
1   78.5  7 26  6 60
2   74.3  1 29 15 52
3  104.3 11 56  8 20
4   87.6 11 31  8 47
5   95.9  7 52  6 33
6  109.2 11 55  9 22
7  102.7  3 71 17  6
8   72.5  1 31 22 44
9   93.1  2 54 18 22
10 115.9 21 47  4 26
11  83.8  1 40 23 34
12 113.3 11 66  9 12
13 109.4 10 68  8 12
```

[58]:

```
# Combine the x vectors into a single vector
x <- c(x1, x2, x3, x4)

# Print the combined vector
print(x)
```

```
 [1]  7  1 11 11  7 11  3  1  2 21  1 11 10 26 29 56 31 52 55 71 31 54␣
 ↪47 40 66
[26] 68  6 15  8  8  6  9 17 22 18  4 23  9  8 60 52 20 47 33 22  6 44␣
 ↪22 26 34
[51] 12 12
```

[59]:

```
# Reshape the combined vector into a matrix
X_matrix <- matrix(x, nrow = 13, ncol = 4, byrow = TRUE)

# Print the matrix
print(X_matrix)
```

```
      [,1] [,2] [,3] [,4]
 [1,]    7    1   11   11
 [2,]    7   11    3    1
 [3,]    2   21    1   11
 [4,]   10   26   29   56
 [5,]   31   52   55   71
 [6,]   31   54   47   40
 [7,]   66   68    6   15
 [8,]    8    8    6    9
 [9,]   17   22   18    4
[10,]   23    9    8   60
[11,]   52   20   47   33
[12,]   22    6   44   22
[13,]   26   34   12   12
```

[60]:

```
# Calculate and print the transpose of the matrix multiplied by the␣
 ↪matrix
XtX <- t(X_matrix) %*% X_matrix
print(XtX)
```

```
      [,1]  [,2]  [,3]  [,4]
[1,] 11226 10861  8210  9129
[2,] 10861 13704  8753 10481
[3,]  8210  8753 10955 10903
[4,]  9129 10481 10903 15659
```

[61]:

```
# Perform eigen decomposition on the matrix
eg <- eigen(XtX)

# Print the eigenvalues and eigenvectors
print(eg)
```

```
eigen() decomposition
$values
[1] 42290.928  5624.922  2253.720  1374.430

$vectors
           [,1]       [,2]       [,3]       [,4]
[1,] -0.4647793  0.4736991 -0.2590947  0.7017545
[2,] -0.5191874  0.5299015  0.3484843 -0.5728940
[3,] -0.4586550 -0.3385608 -0.7434423 -0.3497225
[4,] -0.5514166 -0.6165956  0.5086475  0.2388038
```

[62]:

```
# Calculate and print the condition number (kappa)
kappa <- max(eg$values) / min(eg$values)
print(kappa)
```

```
[1] 30.7698
```

[63]:

```
# Print the minimum eigenvalue
print(min(eg$values))
```

```
[1] 1374.43
```

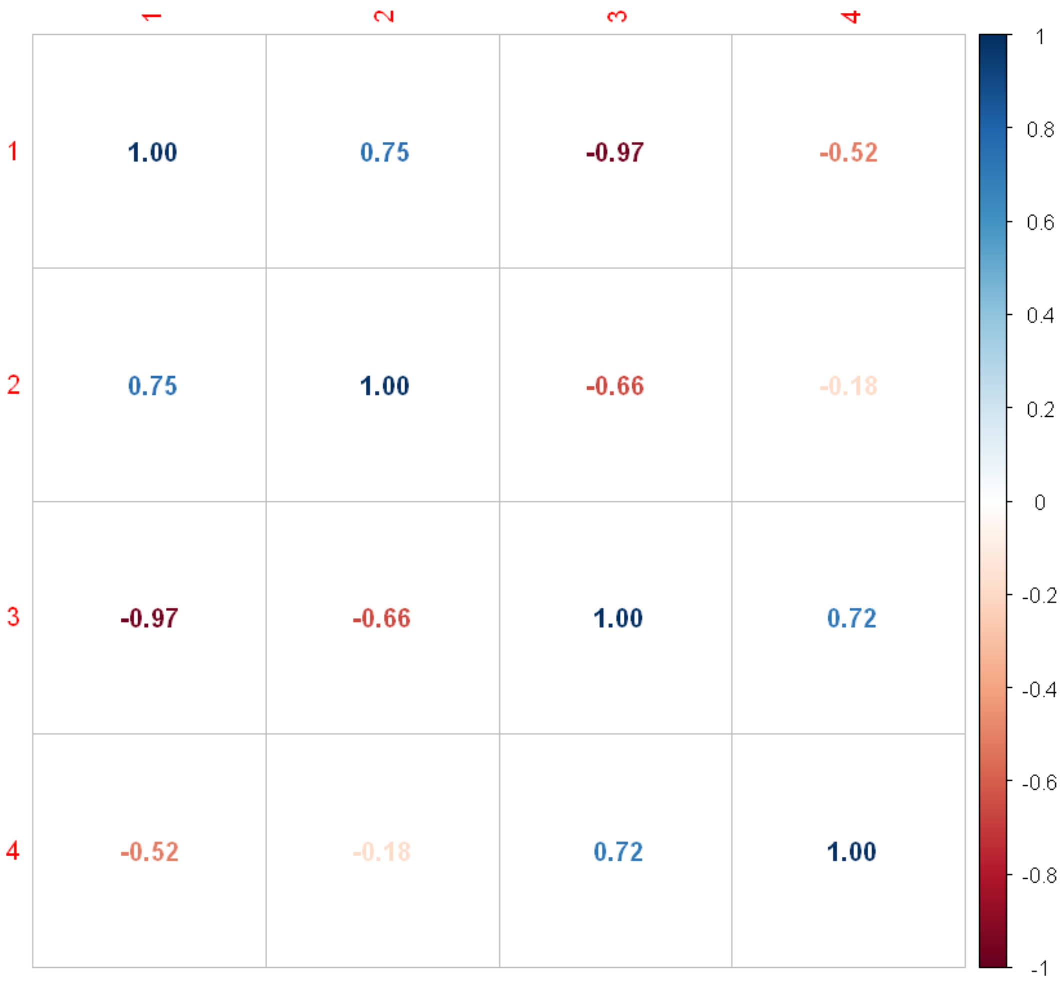
```
[64]: # Print the fourth eigenvector
      print(eg$vectors[,4])
```

```
[1]  0.7017545 -0.5728940 -0.3497225  0.2388038
```

```
[65]: # Calculate and print the correlation matrix
      cor_matrix <- cor(XtX)
      print(cor_matrix)
```

```
           [,1]       [,2]       [,3]       [,4]
[1,]  1.0000000  0.7504830 -0.9676217 -0.5174374
[2,]  0.7504830  1.0000000 -0.6580870 -0.1845232
[3,] -0.9676217 -0.6580870  1.0000000  0.7163816
[4,] -0.5174374 -0.1845232  0.7163816  1.0000000
```

```
[66]: # Plot the correlation matrix
      corrplot(cor_matrix, method = "number")
```



■ **Interpretation:-**

We have obtained that the lowest eigenvalue of the matrix $X'X$ is 1374.430 and the corresponding eigenvector is

$(0.7017, -0.5728, -0.3497, 0.2388)$.

Hence from these elements of the eigenvectors we can construct the equation

$$x_1 = 0.8163x_2 - 0.4948x_3 + 0.2388x_4$$

Hence we can say that $x1$ and x 2 are correlated. Also from the correlation matrix we can see that x 1 and x2 are correlated positively.

The source of multicollinearity arise from the fact of correlation between those predictor variables which causes the ill construction of $X'X$.

18) Analyze the Patient Satisfaction data for multicollinearity.

[78]:

```
# Load necessary library
library(corrplot)
library(car)

# Define the variables
satisfaction <- c(68, 77, 96, 80, 43, 44, 26, 88, 75, 57, 56, 88, 88,␣
 ↪102, 88, 70,
                  82, 43, 46, 56, 59, 26, 52, 83, 75)

age <- c(55, 46, 30, 35, 59, 61, 74, 38, 27, 51, 53, 41, 37, 24, 42,␣
 ↪50, 58, 60,
         62, 68, 70, 79, 63, 39, 49)

severity <- c(50, 24, 46, 48, 58, 60, 65, 42, 42, 50, 38, 30, 31, 34,␣
 ↪30, 48, 61,
              71, 62, 38, 41, 66, 31, 42, 40)

surgical_medical <- c(0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1,␣
 ↪1, 1, 0, 0,
                      1, 1, 1, 0, 1)

anxiety <- c(2.1, 2.8, 3.3, 4.5, 2, 5.1, 5.5, 3.2, 3.1, 2.4, 2.2, 2.1,␣
 ↪1.9, 3.1, 3,
             4.2, 4.6, 5.3, 7.2, 7.8, 7, 6.2, 4.1, 3.5, 2.1)

# Create the data frame
Patient_satisfaction_data <- data.frame(satisfaction, age, severity,␣
 ↪surgical_medical, anxiety)
print(Patient_satisfaction_data)
```

```
   satisfaction age severity surgical_medical anxiety
1            68  55       50                0     2.1
2            77  46       24                1     2.8
3            96  30       46                1     3.3
4            80  35       48                1     4.5
5            43  59       58                0     2.0
6            44  61       60                0     5.1
7            26  74       65                1     5.5
8            88  38       42                1     3.2
9            75  27       42                0     3.1
10           57  51       50                1     2.4
11           56  53       38                1     2.2
12           88  41       30                0     2.1
13           88  37       31                0     1.9
14          102  24       34                0     3.1
15           88  42       30                0     3.0
16           70  50       48                1     4.2
17           82  58       61                1     4.6
18           43  60       71                1     5.3
19           46  62       62                0     7.2
20           56  68       38                0     7.8
21           59  70       41                1     7.0
22           26  79       66                1     6.2
23           52  63       31                1     4.1
24           83  39       42                0     3.5
25           75  49       40                1     2.1
```

[79]:

```
# Prepare the predictor variables
predictor_vector <- data.frame(age, severity, surgical_medical, anxiety)
predictor_matrix <- as.matrix(predictor_vector)
```

[80]:

```
# Calculate and print the matrix product
XtX <- t(predictor_matrix) %*% predictor_matrix
print(XtX)
```

```
                     age severity surgical_medical anxiety
age              69881.0  60814.0            756.0 5387.10
severity         60814.0  56790.0            671.0 4760.60
surgical_medical   756.0    671.0             14.0   57.40
anxiety           5387.1   4760.6             57.4  461.21
```

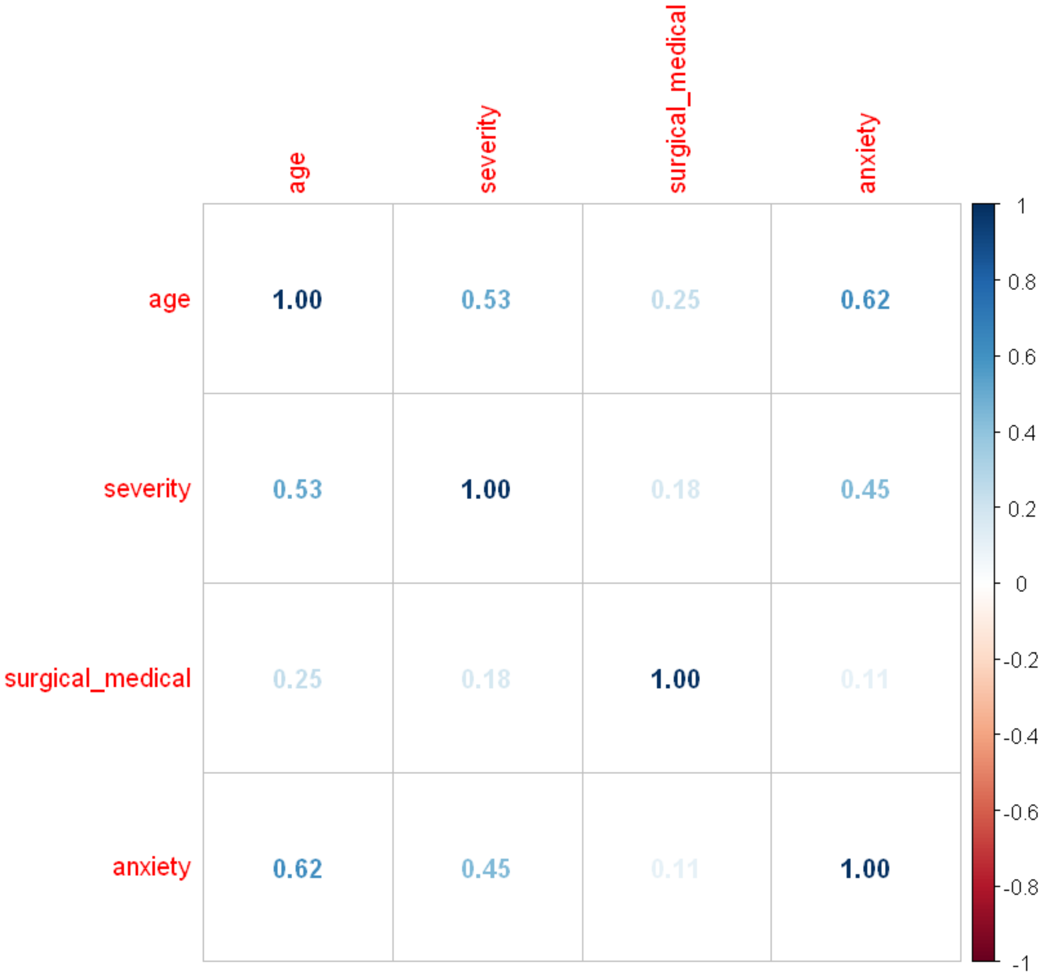
[81]:

```
# Perform eigen decomposition
eg1 <- eigen(XtX)

# Calculate and print the condition number (kappa)
kappa <- max(eg1$values) / min(eg1$values)
print(kappa)
```

```
[1] 21749.6
```

[82]:
```
# Plot the correlation matrix
cor_matrix <- cor(predictor_vector)
print(cor_matrix)
corrplot(cor_matrix, method = "number")
```

```
                      age  severity surgical_medical   anxiety
age             1.0000000 0.5290246        0.2456932 0.6212453
severity        0.5290246 1.0000000        0.1775101 0.4471567
surgical_medical 0.2456932 0.1775101        1.0000000 0.1096486
anxiety         0.6212453 0.4471567        0.1096486 1.0000000
```



[83]:
```
# Fit the linear model
model <- lm(satisfaction ~ age + severity + surgical_medical + anxiety,␣
 ↪data = Patient_satisfaction_data)
print(summary(model))
```

```
Call:
lm(formula = satisfaction ~ age + severity + surgical_medical +
    anxiety, data = Patient_satisfaction_data)

Residuals:
    Min      1Q  Median      3Q     Max
-18.506  -5.096   1.306   4.738  28.722

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      140.1689     8.3191  16.849 2.77e-13 ***
age               -1.1428     0.1904  -6.002 7.22e-06 ***
severity          -0.4699     0.1866  -2.518   0.0204 *
surgical_medical   2.2259     4.1402   0.538   0.5968
anxiety            1.2673     1.4922   0.849   0.4058
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.921 on 20 degrees of freedom
Multiple R-squared:  0.8183,    Adjusted R-squared:  0.7819
F-statistic: 22.51 on 4 and 20 DF,  p-value: 3.611e-07
```

[84]:

```
# Calculate and print the variance inflation factors
print(vif(model))
```

```
             age         severity surgical_medical          anxiety
        1.939128         1.441055         1.072782         1.689768
```

[85]:

```
# Calculate and print the variance inflation factor for each predictor
k <- max(eg1$values) / eg1$values
print(k)
```

```
[1]     1.00000    57.51972  2812.59891 21749.60378
```

■ **Interpretation:-**

In R output, We have obtained kappa more than 1000 hence we can say that there is presence of strong multicollinearity among the predictors.

we can see that there age and severity is connected with correlation 0.53 and between the anxiety and age the correlation is 0.62 , and also between severity and anxiety the correlation is 0.45 .

We are getting the VIF are close to 1 hence we can say that there is no problem with the multicollinearity in the model.

Here we can see from the kappa is 21749.6 which is greater than 1000 and hence we can say that there is the presence of multicollinearity.

Looking at the VIF we can say there in the model Multicollinearity is not making any such problem.

hence we have two contradictory results but as the kappa is more beneficial and determining each of the incidences we are getting:

k <- max(eg1$values)/eg$1values

print(k)

[1] 1.00000   57.51972   2812.59891   21749.60378

Hence we can see that two of the incidences are greater than 1000 so there is surely a near linear relationship between the predictors. Hence multicollinearity exists in this particular model.

19) Analyse the fuel consumption data in Table b. 18 for multicollinearity.

TABLE B. 18 Fuel Consumption Data

y	x1	x2	x3	x4	x5	x6	x7	x8
343	0	52.8	811.7	2.11	220	261	87	1.8
356	1	52.8	811.7	2.11	220	261	87	1.8
344	0	50.0	821.3	2.11	223	260	87	16.6
356	1	50.0	823.3	2.11	223	260	87	16.6
352	0	47.2	833.0	2.09	221	261	92	23.0
361	1	47.2	832.0	2.09	221	261	92	23.0
372	0	47.0	831.3	2.26	190	323	75	25.1
355	1	47.0	831.3	2.26	190	323	75	25.1
375	0	48.3	83.8	2.47	180	364	71	26.1
359	1	48.3	836.8	2.47	180	364	71	26.1
364	0	44.7	808.3	1.41	180	300	64	20.0
357	1	44.7	808.3	1.41	180	300	64	20.0
368	0	55.7	808.7	1.44	176	299	64	20.5
360	1	55.7	808.7	1.44	176	299	64	20.5
372	0	52.8	813.2	1.96	175	301	75	17.3
352	1	52.8	813.2	1.96	175	301	75	17.3

y : fuel consumption (g/km)

x1 : vehicle (0 - bus, 1− truck)

$x2$: cetane number

x 3 : density (g/L, 15°C)

x 4 : viscosity (KV, 40°C)

$x5$: initial boiling point (degrees C)

$x6$: final boiling point (degrees C)

x7 : flash point (degrees C)

$x8$: total aromatics (percent)

Code and interpretation:

[93]:

```
# Load the necessary library
library(readxl)

# Read the Excel file
data1 <- read_excel("Analyse the fuel consumption data in Table b. 18␣
 ↪for multicollinearity.xlsx")

# View the data (for interactive use; not needed in scripts)
View(data1)
```

A tibble: 16 × 9

y <dbl>	x1 <dbl>	x2 <dbl>	x3 <dbl>	x4 <dbl>	x5 <dbl>	x6 <dbl>	x7 <dbl>	x8 <dbl>
343	0	52.8	811.7	2.11	220	261	87	1.8
356	1	52.8	811.7	2.11	220	261	87	1.8
344	0	50.0	821.3	2.11	223	260	87	16.6
356	1	50.0	823.3	2.11	223	260	87	16.6
352	0	47.2	833.0	2.09	221	261	92	23.0
361	1	47.2	832.0	2.09	221	261	92	23.0
372	0	47.0	831.3	2.26	190	323	75	25.1
355	1	47.0	831.3	2.26	190	323	75	25.1
375	0	48.3	83.8	2.47	180	364	71	26.1
359	1	48.3	836.8	2.47	180	364	71	26.1
364	0	44.7	808.3	1.41	180	300	64	20.0
357	1	44.7	808.3	1.41	180	300	64	20.0
368	0	55.7	808.7	1.44	176	299	64	20.5
360	1	55.7	808.7	1.44	176	299	64	20.5
372	0	52.8	813.2	1.96	175	301	75	17.3
352	1	52.8	813.2	1.96	175	301	75	17.3

[92]:

```
# Convert data to matrix
X_matrix <- as.matrix(data1)

# Calculate eigenvalues and eigenvectors
eg <- eigen(t(X_matrix) %*% X_matrix)

# Calculate the condition number (kappa)
kappa <- max(eg$values) / min(eg$values)

# Print the kappa value
print(kappa)
```

```
[1] 105972083
```

Interpretation:

We get condition number of $X'X$ as,

$$\kappa = 105972083 >> 1000$$

which defines the severe multicollinearity between the variables.

[90]:
```
model<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8,data=data1)
vif_results<-car::vif(model)
print(vif_results)
```

```
       x1        x2        x3        x4        x5        x6        x7  ␣
↪     x8
 1.126957  1.840089  1.818914 41.118985 12.755788 73.289319 59.507844  2.
↪823316
```

■ **Interpretation:-**

We have used the "car" package to interpret results determining the VIFs of the variables.

The output is showing the VIF values for each predictor variable. VIF values greater than 5 or 10 may indicate significant multicollinearity. Let's interpret the results:

- The variables $x1$ and $x2$ have a VIF close to 1 , which suggests that there is little to no multicollinearity between that variable and the other predictors. The variables $x3, x4, x5, x6, x7$ and $x8$ have VIF greater than 5 , which indicates high multicollinearity.

Thus we get the corresponding results using condition number and VIF resulting in high multicollinearity.

20) Estimate model parameters for the Hald cement data (Table B.21) using principal component regression.

a. What is the loss in **R2** for this model compared to least squares?

b. How much shrinkage in the coefficient vector has resulted?

c. Compare the principal - component model with the ordinary ridge model developed in Problem 9.17. Comment on any apparent differences in the models.

TABLE B. 21 Hald Cement Data

	Observation				
i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Source: Hald, A. [1952], Statistical Theory with Engineering Applications, Wiley, New York.

[99]:

```
# Load the necessary libraries
library(readxl)
library(pls)

# Read the Excel file
data <- read_excel("Table B.21.xlsx")

# Display the first few rows of the data
head(data)
```

```
Attaching package: 'pls'


The following object is masked from 'package:corrplot':

    corrplot


The following object is masked from 'package:stats':

    loadings
```

A tibble: 6 × 6

i	yi	xi1	xi2	xi3	xi4
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22

[100]:

```
# Prepare the data by removing the first column
data_1 <- data[,-1]

# Fit the PCR model
fit1 <- pcr(yi ~ xi1 + xi2 + xi3 + xi4, data = data_1, scale = TRUE,␣
 ↪validation = "CV")

# Print the model summary
print(summary(fit1))
```

```
Data:    X dimension: 13 4
         Y dimension: 13 1
```

```
Fit method: svdpc
Number of components considered: 4

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps
CV           15.66    3.164    3.484    2.716    2.951
adjCV        15.66    2.897    3.427    2.661    2.886

TRAINING: % variance explained
    1 comps  2 comps  3 comps  4 comps
X     55.89    95.29    99.96   100.00
yi    96.49    96.50    98.21    98.24
NULL
```

■ Interpretation:-

(a) Multiple Linear Regression uses a method of least squared residual to minimize the sum of squared residuals. However, when the predictor variables are highly correlated then multicollinearity can become a problem which can be avoided by principal component regression. We have used a package called “pls”.

Here we have used pcr with yi as the response variable and the $xi1, xi2.xi3, xi4$ as the predictor variables. The above table tells us the test RMSE calculated by the k fold cross validation.

- If we only use the intercept term in the model,the test RMSE is 15.66
- If we add in the first principal component, the test RMSE drops to 3.242
- If we add in the second principal component, the test RMSE rises to 3.453

We can see that adding additional principal components actually leads to an increase in test RMSE. Thus, it appears that it would be optimal to use only one principal component in the final model.

Training: Tells the percentage of the variance in the response variance explained by the principal components.

- By using the 1$^{\text{st}}$ principal component 55.89% of variation is explained.
- By adding in the 2$^{\text{nd}}$ principal component, we can explain 95.29% of the variation in the response variable.

Adding more variables will increase the variation explained by the prediction variables but our aim is to reduce the size. So we use only one Principal Component.

R Code:

[101]:

```
fit1<-lm(yi~xi1+xi2+xi3+xi4,data=data_1)
summary(fit1)
```

```
Call:
lm(formula = yi ~ xi1 + xi2 + xi3 + xi4, data = data_1)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.1750 -1.6709  0.2508  1.3783  3.9254 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  62.4054    70.0710   0.891   0.3991  
xi1           1.5511     0.7448   2.083   0.0708 .
xi2           0.5102     0.7238   0.705   0.5009  
xi3           0.1019     0.7547   0.135   0.8959  
xi4          -0.1441     0.7091  -0.203   0.8441  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,	Adjusted R-squared:  0.9736 
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

■ Interpretation:-

From the above result of multiple linear regression we concluded that $R\hat{}2 = 0.9824$ which implies that 98.24% of variability of response variable is explained by the other regressors but in pcr we can see tha t using only one regressor will explain 95.29% of variability so it will be less time consuming to work wit h one variable rather than all the variables in the model.

(b) By using principal component regression shrinkage in the coefficient vector resulted to “one”.

(c) R Code:

```
[188]: # Load the necessary packages
library(readxl)
library(glmnet)

# Read the Excel file
data <- read_excel("Table B.21.xlsx")

# Prepare the data
data_1 <- data[,-1]  # Remove the first column
y <- data$yi  # Response variable
x <- data.matrix(data_1[, c('xi1', 'xi2', 'xi3', 'xi4')])  # Predictor␣
 ↪variables

# Fit a ridge regression model
fit1 <- glmnet(x, y, alpha = 0)

# Perform cross-validation to find the optimal lambda
cv_fit1 <- cv.glmnet(x, y, alpha = 0)
```

```
# Get the best lambda value
best_lambda <- cv_fit1$lambda.min

# Print the best lambda value
print(best_lambda)
```

```
Warning message:
"Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per␣
↪fold"
```

```
[1] 1.187077
```

[107]:

```
best_model<-glmnet(x,y,alpha=0,lambda=best_lambda)
coef(best_model)
```

```
5 x 1 sparse Matrix of class "dgCMatrix"
                    s0
(Intercept) 86.3899965
xi1          1.1298726
xi2          0.2906186
xi3         -0.2546536
xi4         -0.3464944
```

[108]:

```
y_predicted<-predict(fit1,s=best_lambda,newx=x)
sst<-sum((y-mean(y))^2)
sse<-sum((y_predicted-y)^2)                                                    ␣
 ↪
rsq<-1-sse/sst
rsq
```

0.97931868802768

■ **Interpretation:-**

From the above results we can conclude that the lambda value that minimizes the test MSE turns out to be 1.187077 and the r-squared turns out to be 0.9793187 . That is, the best model was able to explain 9 7.93% of the variation in the response values of the training data.

Ridge regression shrinks everything, but it never shrinks anything to zero but in PCR either it doesn't shrink or shrink to 0 and another advantage of ridge regression is it avoids the problem of overfitting. In ridge regression variability explained is 97.93% and in pcr variability explained is 95.29% so there will be less amount of loss in R-squared value for ridge regression. So we prefer Ridge Regression here.

21) Myers Montgomery and Anderson - Cook (Response Surface Methodology 3 rd edition, Wiley, New York, 2009) discuss an experiment to determine the influence of five factors:

x1 - acid bath temperature

x2 - cascade acid concentration

x3 - water temperature

x4 - sulfide concentration

x5 - amount of chlorine bleach on an appropriate measure of the whiteness of rayon (y).

The engineers conducting this experiment wish to minimize this measure. The experimental result *f* ollow.

Acid				Amount of	
Temp.	Acid Conc.	Water Temp.	Sulfide Conc.	Bleach	y
35	0.3	82	0.2	0.3	76.5
35	0.3	82	0.3	0.5	76.0
35	0.3	88	0.2	0.5	79.9
35	0.3	88	0.3	0.3	83.5
35	0.7	82	0.2	0.5	89.5
35	0.7	82	0.3	0.3	84.2
35	0.7	88	0.2	0.3	85.7
35	0.7	88	0.3	0.5	99.5
55	0.3	82	0.2	0.5	89.4
55	0.3	82	0.3	0.3	97.5
55	0.3	88	0.2	0.3	103.2
55	0.3	88	0.3	0.5	108.7
55	0.7	82	0.2	0.3	115.2
55	0.7	82	0.3	0.5	111.5
55	0.7	88	0.2	0.5	102.3
55	0.7	88	0.3	0.3	108.1
25	0.5	85	0.25	0.4	80.2
65	0.5	85	0.25	0.4	89.1
45	0.1	85	0.25	0.4	77.2
45	0.9	85	0.25	0.4	85.1
45	0.5	79	0.25	0.4	71.5
45	0.5	91	0.25	0.4	84.5
45	0.5	85	0.15	0.4	77.5
45	0.5	85	0.35	0.4	79.2
45	0.5	85	0.25	0.2	71.0
45	0.5	85	0.25	0.6	90.2

a. Perform a thorough analysis of the results including residual plots.

b. Perform the appropriate test for lack of fit.

[137]:

```r
library(MPV)
dataD <- p4.20  # Load the dataset into a variable
```

[152]:

```r
# Fit models with different sets of predictors and calculate AIC
model <- lm(y ~ x1 + x2 + x3 + x4 + x5, data = dataD)

# Print the summary of the model
print(summary(model))
```

```r
# Calculate and print AIC of the model
aic <- AIC(model)
print(aic)
```

```

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = dataD)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8929  -7.7179   0.0696   5.9581  19.9946 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -35.2626    62.0105  -0.569  0.57592   
x1            0.7454     0.2101   3.548  0.00202 **
x2           20.2292    10.5045   1.926  0.06847 . 
x3            0.7931     0.7003   1.132  0.27084   
x4           25.5833    42.0181   0.609  0.54947   
x5           17.2083    21.0090   0.819  0.42239   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.29 on 20 degrees of freedom
Multiple R-squared:  0.4822,	Adjusted R-squared:  0.3527 
F-statistic: 3.724 on 5 and 20 DF,  p-value: 0.01518

[1] 202.1958
```

[153]:

```r
# Fit the model
model1 <- lm(y ~ x1 + x2 + x3 + x4, data = dataD)

# Print the summary of the model
print(summary(model1))

# Calculate and print AIC of the model
aic1 <- AIC(model1)
print(aic1)
```

```

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = dataD)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.0846  -7.7179   0.8654   6.0862  18.2737 

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.3793    60.9551  -0.466  0.64631
x1            0.7454     0.2084   3.576  0.00178 **
x2           20.2292    10.4219   1.941  0.06580 .
x3            0.7931     0.6948   1.141  0.26654
x4           25.5833    41.6875   0.614  0.54601
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.21 on 21 degrees of freedom
Multiple R-squared:  0.4648,    Adjusted R-squared:  0.3628
F-statistic: 4.559 on 4 and 21 DF,  p-value: 0.008298

[1] 201.0537
```

[154]:

```r
# Fit the model
model2 <- lm(y ~ x1 + x2 + x3, data = dataD)

# Print the summary of the model
print(summary(model2))

# Calculate and print AIC of the model
aic2 <- AIC(model2)
print(aic2)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3, data = dataD)

Residuals:
     Min       1Q   Median       3Q      Max
-18.0846  -7.9554   0.8487   6.1581  16.9946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.9834    59.2006  -0.371  0.71394
x1            0.7454     0.2055   3.628  0.00149 **
x2           20.2292    10.2732   1.969  0.06166 .
x3            0.7931     0.6849   1.158  0.25930
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.07 on 22 degrees of freedom
Multiple R-squared:  0.4552,    Adjusted R-squared:  0.3809
F-statistic: 6.127 on 3 and 22 DF,  p-value: 0.003433

[1] 199.5159
```

[155]:

```
# Fit the model
model3 <- lm(y ~ x1 + x2, data = dataD)

# Print the summary of the model
print(summary(model3))

# Calculate and print AIC of the model
aic3 <- AIC(model3)
print(aic3)
```

```

Call:
lm(formula = y ~ x1 + x2, data = dataD)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.0846  -4.3867   0.5696   5.9966  16.2071 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  45.4263    10.8389   4.191  0.00035 ***
x1            0.7454     0.2070   3.601  0.00150 ** 
x2           20.2292    10.3490   1.955  0.06288 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.14 on 23 degrees of freedom
Multiple R-squared:  0.422,	Adjusted R-squared:  0.3717 
F-statistic: 8.395 on 2 and 23 DF,  p-value: 0.00183

[1] 199.0541
```

[157]:

```
# Fit the model
model4 <- lm(y ~ x1, data = dataD)

# Print the summary of the model
print(summary(model4))

# Calculate and print AIC of the model
aic4 <- AIC(model4)
print(aic4)
```

```

Call:
lm(formula = y ~ x1, data = dataD)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.085  -6.762   1.038   6.502  18.661 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  55.5409    10.0682   5.516 1.13e-05 ***
x1            0.7454     0.2188   3.407  0.00232 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.72 on 24 degrees of freedom
Multiple R-squared:  0.326,	Adjusted R-squared:  0.2979 
F-statistic: 11.61 on 1 and 24 DF,  p-value: 0.002319

[1] 201.0499
```

Interpretation

From the above results, we can see that the lowest AIC is 199.0541, which corresponds to model 3, where (y) is the predicted variable and (x_1) & (x_2) are the independent variables. Therefore, model 3 is the best for prediction purposes.

b) R Code:

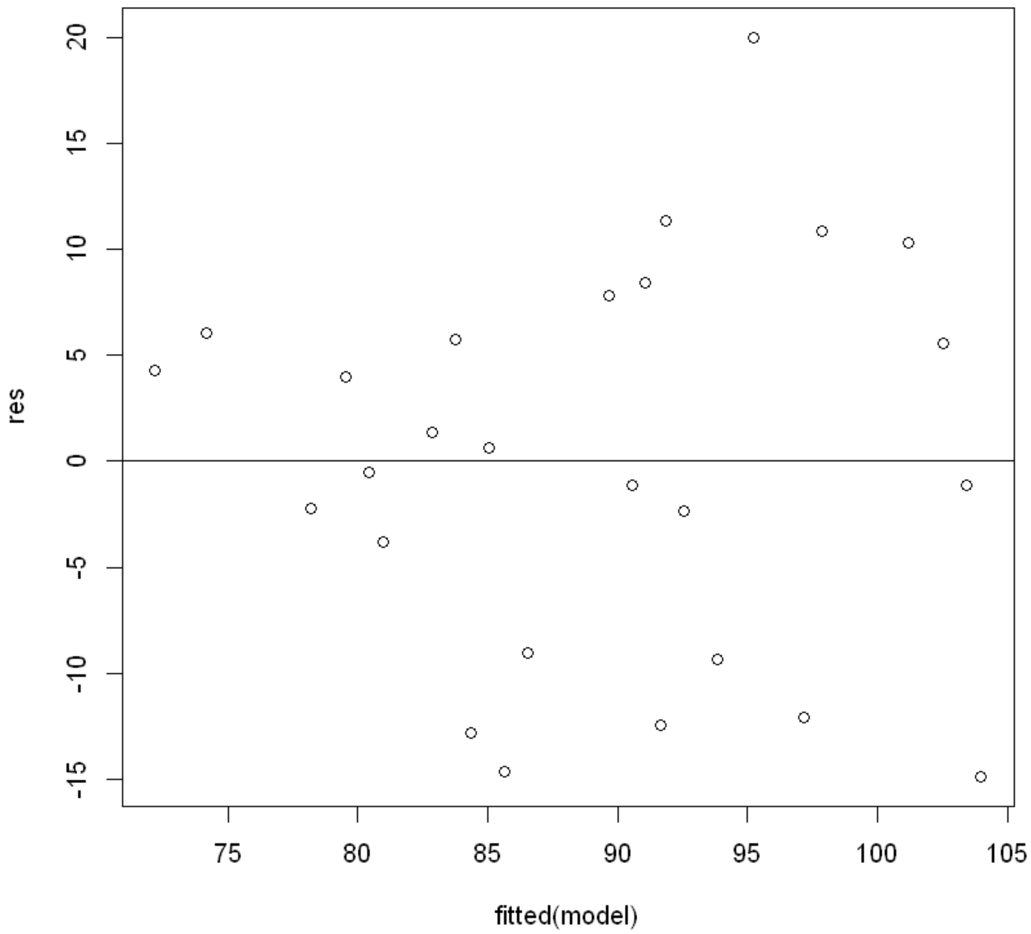
[159]:

```
res<-resid(model)
print(res)
```

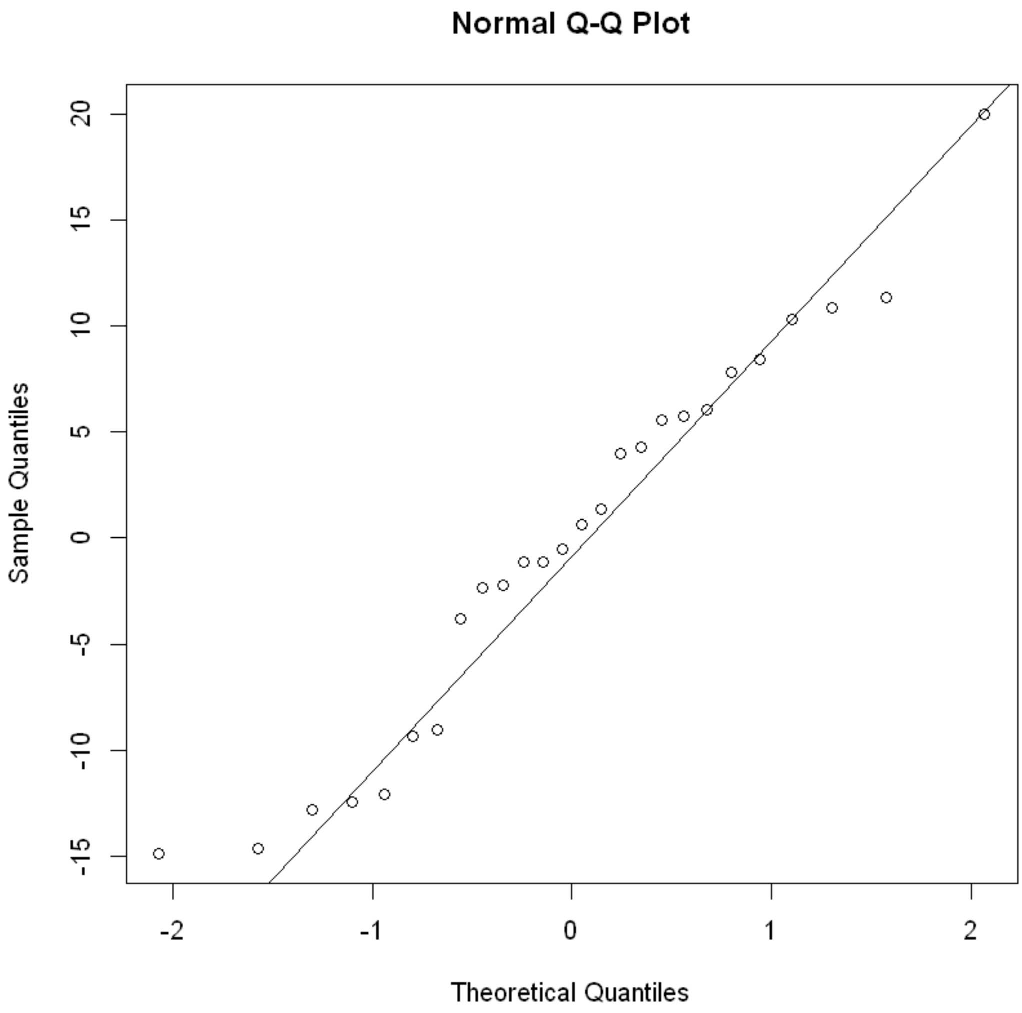
```
          1           2           3           4           5           6 
  4.2945513  -2.2054487  -0.5054487   3.9778846   5.7612179   1.3445513 
          7           8           9          10          11          12 
  0.6445513   8.4445513  -1.1554487   7.8278846  11.3278846  10.8278846 
         13          14          15          16          17          18 
 19.9945513  10.2945513  -1.1054487   5.5778846   6.0237179 -14.8929487 
         19          20          21          22          23          24 
 -3.7929487 -12.0762821 -12.8262821  -9.3429487  -9.0262821 -12.4429487 
         25          26 
-14.6429487  -2.3262821 
```

[160]:

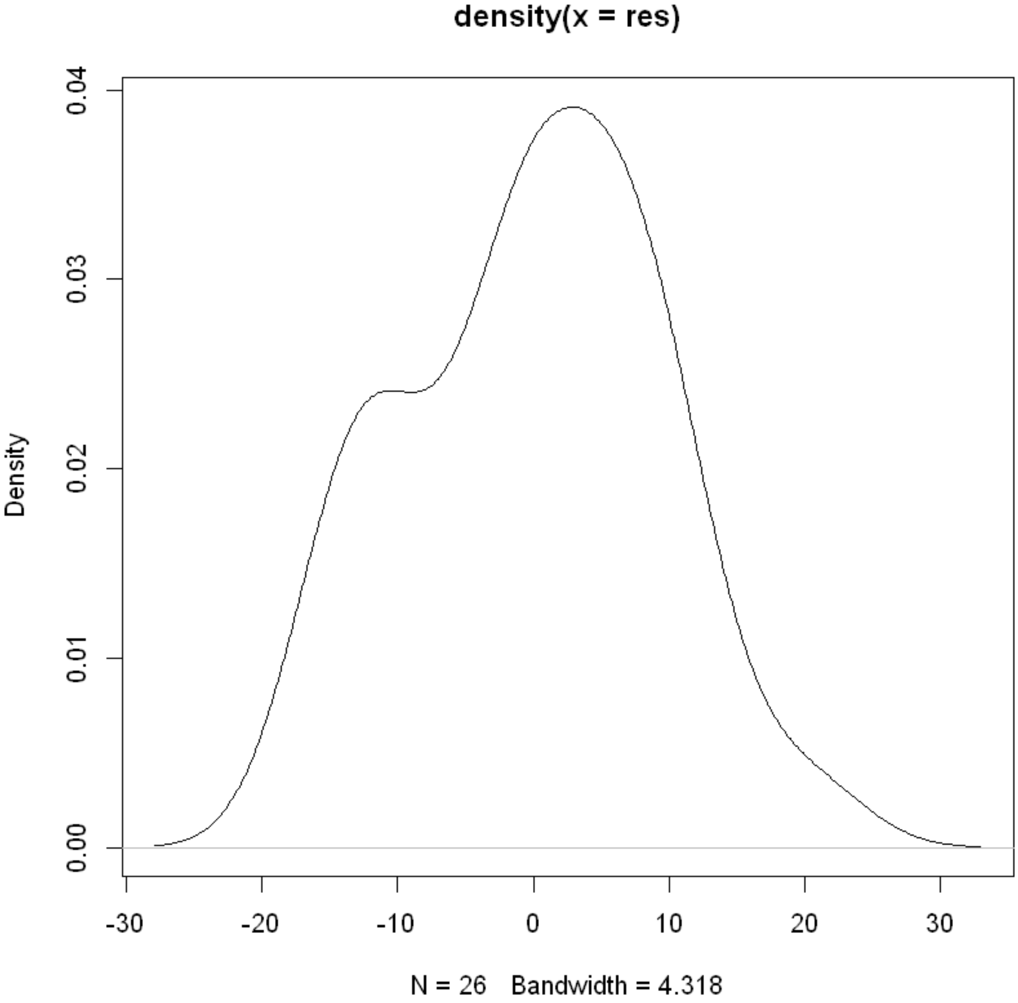
```
plot(fitted(model),res)
abline(0,0)
```



[161]:
```
qqnorm(res)                                                                    ␣
 ↪                                                                            ␣
 ↪
qqline(res)
```



```
[162]: plot(density(res))
```



■ **Interpretation:-**

Here we see that linearity yseems to hold reasonably well, as the residual points are close to the centre line. We can also note the heteroskedasticity: as we move to the right on the x-axis, the spread of the residuals seems to be increasing, which implies that the data is not normally distributed which has also been seen from normal qqplot and density plot.

3 3)MODEL ADEQUACY CHECKING

22) Derringer and Suich ("Simultaneous Optimization of Several Response Variables," *Journal of Quality Technology*, 1980) studied the relationship of an abrasion index for a tire tread compound in terms of three factors: (x_1) (hydrated silica level), (x_2) (silane coupling agent level), and (x_3) (sulfur level). The following table gives the actual results.

(x_1)	(x_2)	(x_3)	(y)
-1	-1	1	102
1	-1	-1	120
-1	1	-1	117
1	1	1	198
-1	-1	-1	103
1	-1	1	132
-1	1	1	132
1	1	-1	139
0	0	0	133
0	0	0	133
0	0	0	140
0	0	0	142
0	0	0	145
0	0	0	142

a. Perform a thorough analysis of the results including residual plots.

b. Perform the appropriate test for lack of fit.

[166]:

```
library("MPV")
#4.19(a)
data1=attach(p4.19)
```

```
The following objects are masked _by_ .GlobalEnv:

    x1, x2, x3, y


The following objects are masked from p4.19 (pos = 3):

    x1, x2, x3, y


The following object is masked from p4.20 (pos = 4):

    y


The following object is masked from p4.20 (pos = 5):

    y


The following object is masked from p4.20 (pos = 6):

    y
```

```
The following object is masked from p4.20 (pos = 7):

    y

The following object is masked from p4.20 (pos = 8):

    y

The following object is masked from p13.2:

    y

The following object is masked from p13.1 (pos = 17):

    y

The following object is masked from p13.1 (pos = 18):

    y

The following object is masked from p13.1 (pos = 19):

    y

The following objects are masked from table.b10 (pos = 20):

    x1, x2, y

The following objects are masked from table.b10 (pos = 21):

    x1, x2, y

The following objects are masked from table.b6:

    x1, x2, x3, y
```

[171]:

```
model=lm(data1$y~x1+x2+x3,data=p4.19)
summary(model)
```

```

Call:
lm(formula = data1$y ~ x1 + x2 + x3, data = p4.19)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.518  -8.768  -1.143   7.857  20.232 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  134.143      3.406  39.382 2.66e-12 ***
x1            16.875      4.506   3.745  0.00382 ** 
x2            16.125      4.506   3.579  0.00502 ** 
x3            10.625      4.506   2.358  0.04009 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.74 on 10 degrees of freedom
Multiple R-squared:  0.7641,	Adjusted R-squared:  0.6933 
F-statistic:  10.8 on 3 and 10 DF,  p-value: 0.001772
```

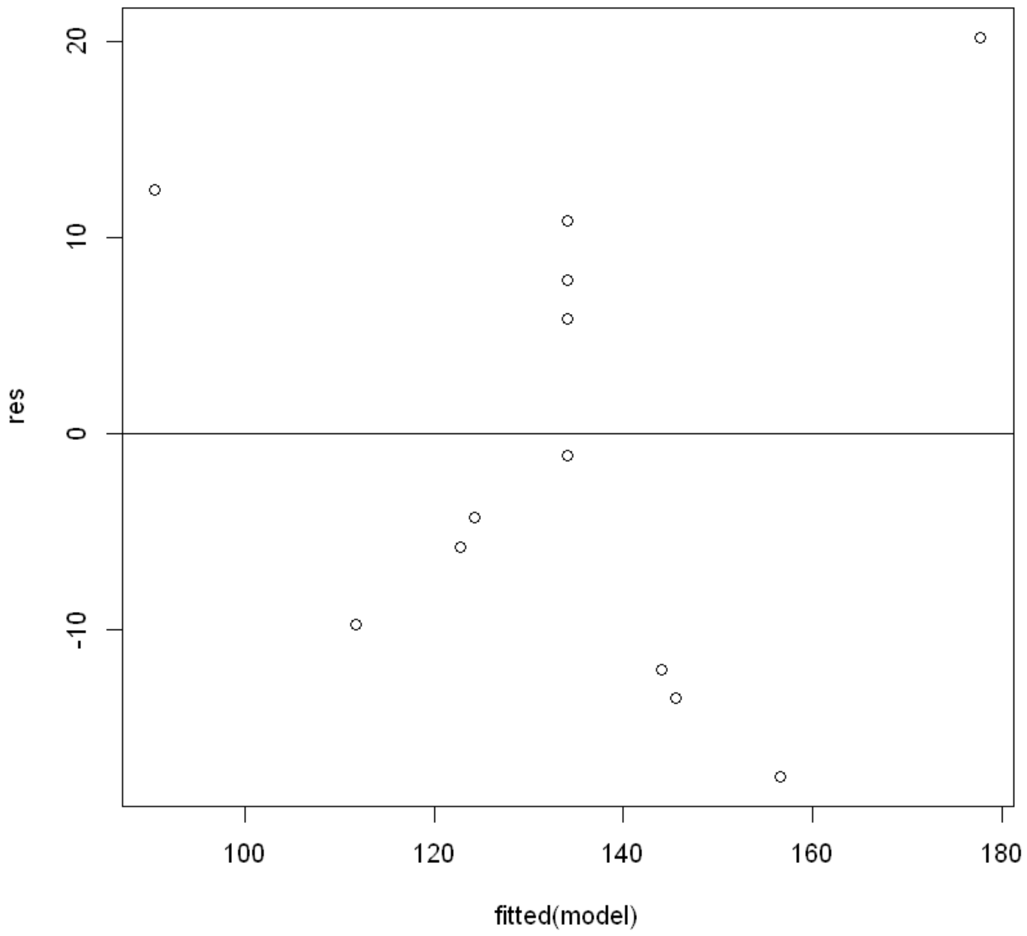
[172]:

```
res<-resid(model)
print(res)
```

```
         1          2          3          4          5          6          ␣
 ↪  7 
 -9.767857  -4.267857  -5.767857  20.232143  12.482143 -13.517857 -12.
 ↪017857 
         8          9         10         11         12         13          ␣
 ↪ 14 
-17.517857  -1.142857  -1.142857   5.857143   7.857143  10.857143   7.
 ↪857143 
```

[173]:

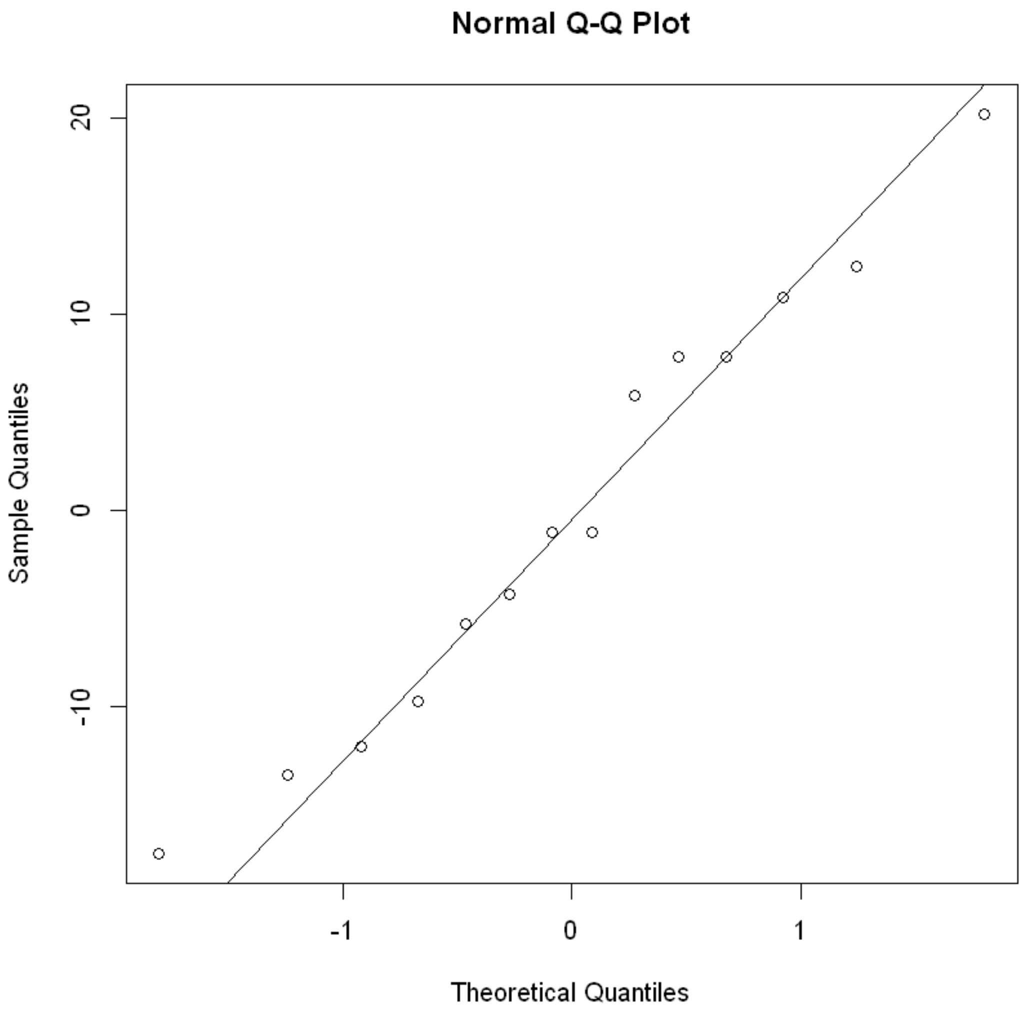
```
plot(fitted(model),res)
abline(0,0)
```



a. This plot seems to indicate that the residuals and the fitted values are uncorrelated, as they should be in homoscedastic linear model with normally distributed errors.

[174]:

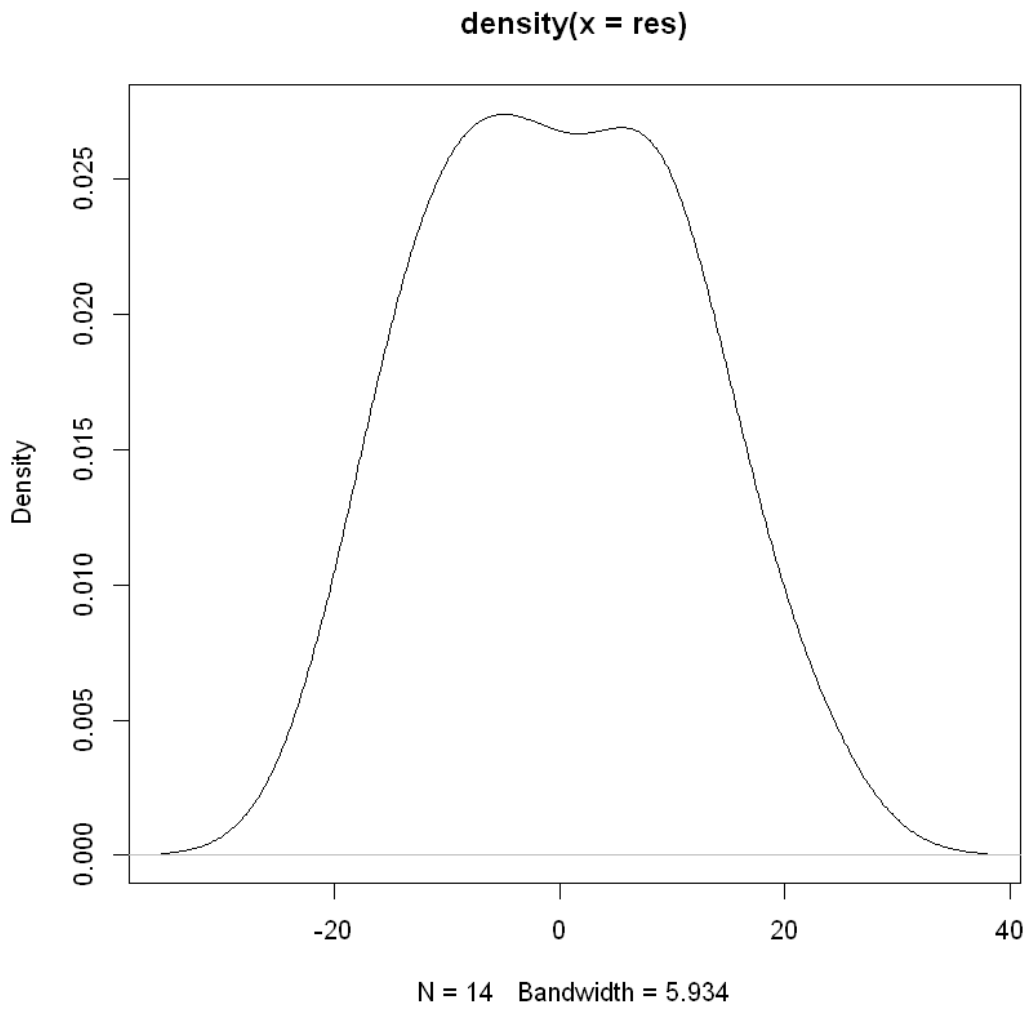
```
qqnorm(res)
qqline(res)
```



■ **Interpretation:-**

As the data values in the plot fall along a roughly straight line at a 45 -degree angle using the qqline() function, it can be concluded that the data is normally distributed.

[175]:
```
plot(density(res))
```



■ **Interpretation:-**

As we can see in the picture below that its almost a bell shaped so we can say that the residuals were normally distributed.

R Code:

[178]:
```
model<-lm(y~x1+x2+x3,data=data1)
summary(model)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-17.518  -8.768  -1.143   7.857  20.232
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  134.143      3.406  39.382 2.66e-12 ***
x1            16.875      4.506   3.745  0.00382 **
x2            16.125      4.506   3.579  0.00502 **
x3            10.625      4.506   2.358  0.04009 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.74 on 10 degrees of freedom
Multiple R-squared:  0.7641,	Adjusted R-squared:  0.6933
F-statistic:  10.8 on 3 and 10 DF,  p-value: 0.001772
```

[179]:

```
aic<-AIC(model)
aic
```

116.283466176577

[180]:

```
model1<-lm(y~x1+x2,data=data1)
summary(model1)
```

```
Call:
lm(formula = y ~ x1 + x2, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-28.1429  -2.5179  -0.1429   7.3571  30.8571

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  134.143      4.051  33.112 2.28e-12 ***
x1            16.875      5.359   3.149  0.00926 **
x2            16.125      5.359   3.009  0.01189 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.16 on 11 degrees of freedom
Multiple R-squared:  0.6329,	Adjusted R-squared:  0.5662
F-statistic: 9.484 on 2 and 11 DF,  p-value: 0.004037
```

[181]:

```
aic1<-AIC(model1)
aic1
```

120.473081802308

[182]:

```
model2<-lm(y~x1,data=data1)
summary(model2)
```

```
Call:
lm(formula = y ~ x1, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.018 -13.705  -0.705   7.857  46.982 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  134.143      5.237  25.614 7.63e-12 ***
x1            16.875      6.928   2.436   0.0314 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.6 on 12 degrees of freedom
Multiple R-squared:  0.3308,	Adjusted R-squared:  0.2751 
F-statistic: 5.933 on 1 and 12 DF,  p-value: 0.0314
```

[184]:

```
aic2<-AIC(model2)
aic2
```

126.879917262903

■ **Interpretation:-**

AIC value for model is 116.2835

AIC value for model 1 is 120.4731

AIC value for model 2 is 126.8799

Lower the values of AIC indicates a better model. As here for model its AIC is comparatively less so we prefer model for best fit.

23) Consider the soft drink delivery time data.

The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time (y) are the number of cases of product stocked ((x_1)) and the distance walked by the route driver ((x_2)). The engineer has collected 25 observations on delivery time, which are shown in the table below.

Obs No	Delivery Time (in min)	No of Cases ((x_1))	Distance ((x_2))	Obs No	Delivery Time (in min)	No of Cases ((x_1))	Distance ((x_2))
1	16.68	7	560	14	19.75	6	462
2	11.5	3	220	15	24	9	448
3	12.03	3	340	16	29	10	776
4	14.88	4	80	17	15.35	6	200
5	13.75	6	150	19	19	7	132
6	18.11	7	330	20	95.5	3	36

Obs No	Delivery Time (in min)	No of Cases ((x_1))	Distance ((x_2))	Obs No	Delivery Time (in min)	No of Cases ((x_1))	Distance ((x_2))
7	8	2	110	21	17.9	17	770
8	17.83	7	210	22	52.32	10	140
9	79.24	30	1460	23	18.75	96	810
10	21.5	5	605	24	19.83	9	450
11	40.33	16	688	25	10.75	8	635
12	21	10	215				
13	13.5	4	255				

a. Find the simple correlation between cases ((x_1)) and distance ((x_2)).

b. Find the variance inflation factors (VIFs).

c. Find the condition number of (X'X). Is there evidence of multicollinearity in these data?

[65]:

```
library(readxl)

# Load the dataset
mydata <- read_excel("deliverydata.xlsx")

# Display the first few rows of the data
head(mydata)
```

A tibble: 6 × 4

Obs No <dbl>	y <dbl>	x1 <dbl>	x2 <dbl>
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330

[66]:

```
# Extract variables
x1 <- mydata$x1
x2 <- mydata$x2

# Compute correlation
correlation <- cor(x1, x2)
print(correlation)
```

```
[1] 0.4913436
```

[70]:

```
library(car)

# Fit the linear model
model <- lm(y ~ x1 + x2, data = mydata)
```

```
summary(model)
```

```

Call:
lm(formula = y ~ x1 + x2, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.557  -9.112  -6.144  -1.841  76.822 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 18.19475    7.10417   2.561   0.0178 *
x1          -0.05776    0.26864  -0.215   0.8317  
x2           0.01823    0.01551   1.175   0.2524  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.22 on 22 degrees of freedom
Multiple R-squared:  0.06602,	Adjusted R-squared:  -0.01889 
F-statistic: 0.7775 on 2 and 22 DF,  p-value: 0.4718
```

[71]:

```
# Compute VIFs
vif_values <- vif(model)
print(vif_values)
```

```
     x1      x2 
1.31825 1.31825 
```

[68]:

```
# Extract design matrix X
X <- cbind(1, mydata$x1, mydata$x2)

# Compute X'X
XtX <- t(X) %*% X

# Compute condition number
condition_number <- kappa(XtX)
print(condition_number)
```

```
[1] 756664.7
```

■ **Interpretation:-**

a. Simple Correlation Between Cases $(x_1)*andDistance*(x_2 $)

Correlation Coefficient Calculation: The correlation coefficient measures the strength and direction of the linear relationship between two variables. It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear relationship,

- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear relationship.

Theoretical Equation:

$$\text{Correlation} = \frac{\text{Cov}(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}}$$

Where: - $\text{Cov}(x_1, x_2)$ is the covariance between x_1 and x_2, - σ_{x_1} and σ_{x_2} are the standard deviations of x_1 and x_2, respectively.

Result Interpretation: The correlation coefficient between x_1 and x_2 is approximately 0.49, indicating a moderate positive relationship. This suggests that as the number of cases increases, the distance walked tends to increase as well.

b. Variance Inflation Factors (VIFs)

VIF Calculation: Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors.

Theoretical Equation for VIF:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

Where R_j^2 is the R^2 value obtained when the j-th predictor is regressed against all other predictors.

Result Interpretation: The VIF values for both predictors are approximately 1.32, which is well below the common threshold of 10. This indicates that there is no significant multicollinearity in the dataset.

c. Condition Number of (X'X)

Condition Number Calculation: The condition number measures the sensitivity of the solution of a linear system to perturbations in the input data. A high condition number indicates potential multicollinearity.

Theoretical Equation for Condition Number:

$$\text{Condition Number} = \frac{\text{Largest Eigenvalue}}{\text{Smallest Eigenvalue}}$$

Result Interpretation: The condition number is approximately 756,664.7, which is very high. This suggests that there is significant multicollinearity in the dataset, as large condition numbers often indicate numerical instability and multicollinearity problems.

Summary:

- **Correlation:** Shows a moderate positive relationship between cases and distance.
- **VIFs:** Low VIF values indicate minimal multicollinearity among predictors.
- **Condition Number:** High value suggests severe multicollinearity.

24) Analysis of Multicollinearity in Chemical Process Data To analyze the chemical process data for evidence of multicollinearity, we will use two methods: Variance Inflation Factors (VIFs) and the Condition Number of (X'X).

[1]:

```
library(MPV)
table.b5
```

```
Loading required package: lattice

Loading required package: KernSmooth

KernSmooth 2.23 loaded
Copyright M. P. Wand 1997-2009

Loading required package: randomForest

randomForest 4.7-1.1

Type rfNews() to see new features/changes/bug fixes.
```

A data.frame: 27 × 8

	y	x1	x2	x3	x4	x5	x6	x7
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	36.98	5.1	400	51.37	4.24	1484.83	2227.25	2.06
2	13.74	26.4	400	72.33	30.87	289.94	434.90	1.33
3	10.08	23.8	400	71.44	33.01	320.79	481.19	0.97
4	8.53	46.4	400	79.15	44.61	164.76	247.14	0.62
5	36.42	7.0	450	80.47	33.84	1097.26	1645.89	0.22
6	26.59	12.6	450	89.90	41.26	605.06	907.59	0.76
7	19.07	18.9	450	91.48	41.88	405.37	608.05	1.71
8	5.96	30.2	450	98.60	70.79	253.70	380.55	3.93
9	15.52	53.8	450	98.05	66.82	142.27	213.40	1.97
10	56.61	5.6	400	55.69	8.92	1362.24	2043.36	5.08
11	26.72	15.1	400	66.29	17.98	507.65	761.48	0.60
12	20.80	20.3	400	58.94	17.79	377.60	566.40	0.90
13	6.99	48.4	400	74.74	33.94	158.05	237.08	0.63
14	45.93	5.8	425	63.71	11.95	130.66	1961.49	2.04
15	43.09	11.2	425	67.14	14.73	682.59	1023.89	1.57
16	15.79	27.9	425	77.65	34.49	274.20	411.30	2.38
17	21.60	5.1	450	67.22	14.48	1496.51	2244.77	0.32
18	35.19	11.7	450	81.48	29.69	652.43	978.64	0.44
19	26.14	16.7	450	83.88	26.33	458.42	687.62	8.82
20	8.60	24.8	450	89.38	37.98	312.25	468.38	0.02
21	11.63	24.9	450	79.77	25.66	307.08	460.62	1.72
22	9.59	39.5	450	87.93	22.36	193.61	290.42	1.88
23	4.42	29.0	450	79.50	31.52	155.96	233.95	1.43
24	38.89	5.5	460	72.73	17.86	1392.08	2088.12	1.35
25	11.19	11.5	450	77.88	25.20	663.09	994.63	1.61
26	75.62	5.2	470	75.50	8.66	1464.11	2196.17	4.78
27	36.03	10.6	470	83.15	22.39	720.07	1080.11	5.88

[2]:

```
str(table.b5)
```

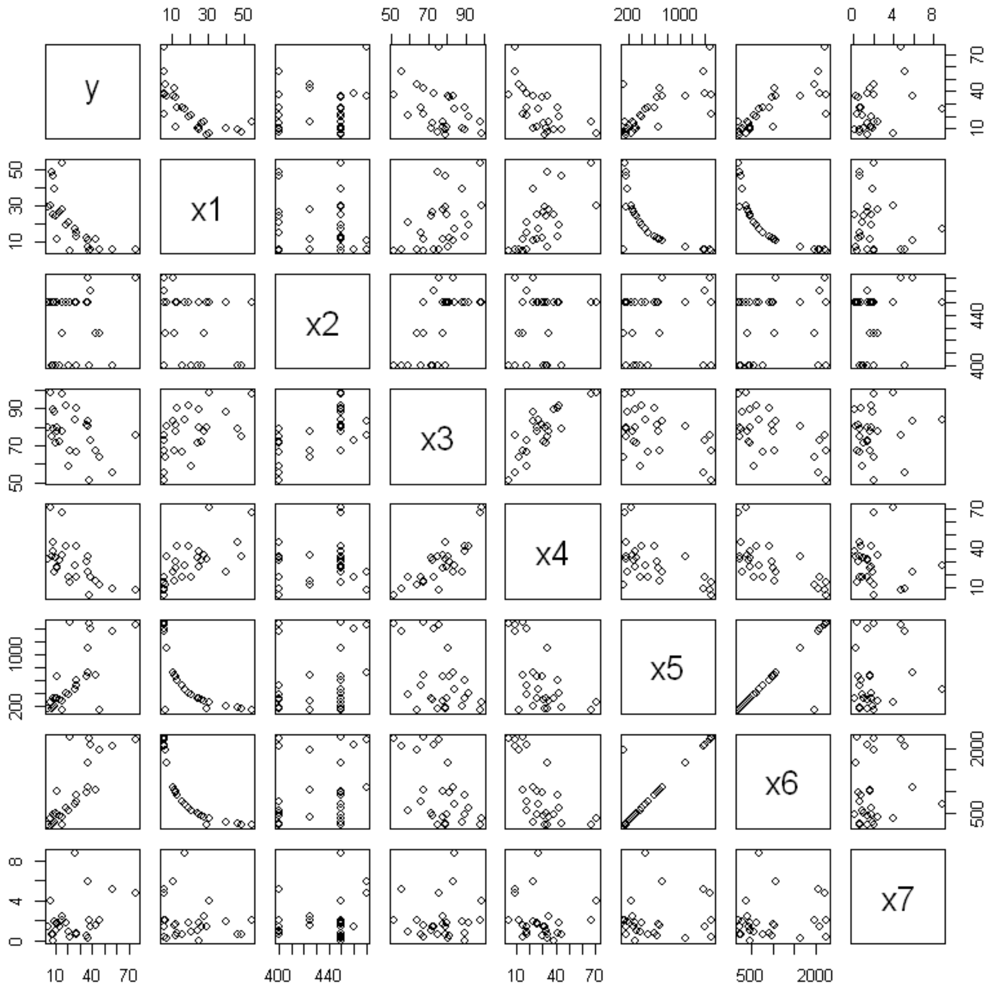
```
'data.frame':	27 obs. of  8 variables:
 $ y : num  36.98 13.74 10.08 8.53 36.42 ...
 $ x1: num  5.1 26.4 23.8 46.4 7 12.6 18.9 30.2 53.8 5.6 ...
 $ x2: num  400 400 400 400 450 450 450 450 450 400 ...
 $ x3: num  51.4 72.3 71.4 79.2 80.5 ...
 $ x4: num  4.24 30.87 33.01 44.61 33.84 ...
 $ x5: num  1485 290 321 165 1097 ...
 $ x6: num  2227 435 481 247 1646 ...
 $ x7: num  2.06 1.33 0.97 0.62 0.22 0.76 1.71 3.93 1.97 5.08 ...
```

[3]:

```
attach(table.b5)
```

[4]:

```
plot(table.b5)
```



```
[5]: cor(table.b5)
```

```
                                 y           x1          x2          x3           x4          x5
                           y     1.0000000   -0.7048039  0.1323896   -0.43305581  -0.6074955  0.69
                           x1    -0.7048039  1.0000000   -0.1919017  0.48865334   0.6729092   -0.7
                           x2    0.1323896   -0.1919017  1.0000000   0.63018845   0.1765413   0.16
A matrix: 8 × 8 of type dbl  x3  -0.4330558  0.4886533   0.6301884   1.00000000   0.8136508   -0.4
                           x4    -0.6074955  0.6729092   0.1765413   0.81365077   1.0000000   -0.5
                           x5    0.6964487   -0.7360404  0.1658431   -0.46339900  -0.5774790  1.00
                           x6    0.7977784   -0.8167299  0.1271191   -0.55860771  -0.6645804  0.88
                           x7    0.3616423   -0.1779805  0.2551410   0.07328754   -0.1140676  0.14
```

```
[6]: colnames(table.b5)
```

1. 'y' 2. 'x1' 3. 'x2' 4. 'x3' 5. 'x4' 6. 'x5' 7. 'x6' 8. 'x7'

```
[7]: data<-data.frame(table.b5)
```

[8]:

```
data1<-lm(y˜., data=data)
```

[9]:

```
summary(data1)
```

```
Call:
lm(formula = y ˜ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.035  -4.681  -1.144   4.072  21.214

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.937016  57.428952   0.939   0.3594
x1          -0.127653   0.281498  -0.453   0.6553
x2          -0.229179   0.232643  -0.985   0.3370
x3           0.824853   0.765271   1.078   0.2946
x4          -0.438222   0.358551  -1.222   0.2366
x5          -0.001937   0.009654  -0.201   0.8431
x6           0.019886   0.008088   2.459   0.0237 *
x7           1.993486   1.089701   1.829   0.0831 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.61 on 19 degrees of freedom
Multiple R-squared:  0.728,	Adjusted R-squared:  0.6278
F-statistic: 7.264 on 7 and 19 DF,  p-value: 0.0002674
```

[12]:

```
# library(car)

print(vif(data1) )
```

```
       x1        x2        x3        x4        x5        x6        x7
 3.674572  7.726406 19.203390  7.463641  4.697002  7.731523  1.119451
```

[19]:

```
CN <- function(x) {
  y <- t(x) %*% x
  l <- eigen(y)$values
  return(max(l) / min(l))
}
```

[20]:

```
CN(cbind(x1, x2, x3, x4, x5, x6, x7))
```

602937.480466244

[21]:

```
detach(table.b5)
```

■ **Interpretation:-**

1. Variance Inflation Factors (VIFs)

VIF Calculation:

The Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is increased due to multicollinearity. Higher VIF values indicate higher multicollinearity.

VIF Values: - x_1: 3.67 - x_2: 7.73 - x_3: 19.20 - x_4: 7.46 - x_5: 4.70 - x_6: 7.73 - x_7: 1.12

Interpretation:

- **VIF $>$ 10**: Indicates significant multicollinearity. In your case, x_3 has a VIF value of 19.20, which is quite high, suggesting it might be contributing to multicollinearity.
- **VIF between 5 and 10**: Indicates moderate multicollinearity. Variables like x_2, x_4, and x_6 fall into this category.
- **VIF $<$ 5**: Generally indicates acceptable levels of multicollinearity. x_1, x_5, and x_7 have lower VIFs, suggesting they are less problematic.

2. Condition Number of $X'X$

Condition Number Calculation:

The condition number of $X'X$ helps to measure the sensitivity of the regression model to multicollinearity. A high condition number indicates potential multicollinearity issues.

Condition Number: 602,937.48

Interpretation:

- **Condition Number $<$ 30**: Indicates no multicollinearity issues.
- **Condition Number between 30 and 1000**: Indicates moderate multicollinearity.
- **Condition Number $>$ 1000**: Suggests severe multicollinearity.

With a condition number of 602,937.48, this suggests severe multicollinearity in your dataset, as it's extremely high. This condition number implies that the predictors are highly collinear, which can lead to instability in the regression coefficients and difficulties in interpreting them.

Summary of Findings

1. **VIF Analysis:** Variables x_3, x_2, x_4, and x_6 show moderate to high multicollinearity, with x_3 being particularly problematic.
2. **Condition Number:** Extremely high, indicating severe multicollinearity.

25) The market research department of a soft drink manufacturer is investigating the effectiveness of a price discount coupon on the purchase of a two-liter beverage product. A sample of 5500 customers was given coupons for varying price discounts between 5 and 25 cents. The response variable was the number of coupons in each price discount category redeemed after one month. The data are shown below.

Discount, x	Sample Size, n	Number Redeemed, r
5	500	100
7	500	122

Discount, x	Sample Size, n	Number Redeemed, r
9	500	147
11	500	176
13	500	211
15	500	244
17	500	277
19	500	310
21	500	343
23	500	372
25	500	391

Tasks

a. Fit a logistic regression model to the data. Use a simple linear regression model as the structure for the linear predictor.

b. Does the model deviance indicate that the logistic regression model from part (a) is adequate?

c. Draw a graph of the data and the fitted logistic regression model.

d. Expand the linear predictor to include a quadratic term. Is there any evidence that this quadratic term is required in the model?

e. Draw a graph of this new model on the same plot that you prepared in part (c). Does the expanded model visually provide a better fit to the data than the original model from part (a)?

f. For the quadratic model in part (d), find Wald statistics for each individual model parameter.

g. Find approximate 95% confidence intervals on the model parameters for the quadratic logistic regression model from part (d).

[27]:

```
# Load the MPV package
library(MPV)

# Attach the dataset from the MPV package
attach(p13.4)
```

[29]:

```
# Define the data
data <- p13.4
data
```

A data.frame: 11 × 3

	x	n	r
	<dbl>	<dbl>	<dbl>
1	5	500	100
2	7	500	122
3	9	500	147
4	11	500	176
5	13	500	211
6	15	500	244
7	17	500	277
8	19	500	310
9	21	500	343
10	23	500	372
11	25	500	391

[30]:

```
# Calculate the proportion redeemed
y <- data$r / data$n

# Fit the logistic regression model
model_a <- glm(cbind(r, n - r) ~ x, data = data, family = binomial)

# Print the summary of the model
print(summary(model_a))
```

```

Call:
glm(formula = cbind(r, n - r) ~ x, family = binomial, data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.084754   0.080398  -25.93   <2e-16 ***
x            0.135727   0.004957   27.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 871.21962  on 10  degrees of freedom
Residual deviance:   0.29426  on  9  degrees of freedom
AIC: 75.708

Number of Fisher Scoring iterations: 3
```

[31]:

```
# Output the coefficients
print(model_a$coefficients)
```

```
(Intercept)           x
 -2.0847543   0.1357274
```

■ Interpretation for Part (a)

We have fit a logistic regression model to the data with the following equation:

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

where: - p denotes the probability of redeeming the coupon. - β_0 (Intercept) is -2.0848. - β_1 (slope for x) is 0.1357.

Coefficients: - **Intercept** (β_0): -2.0848 - This represents the log-odds of redeeming the coupon when the discount is 0. Although a 0 discount is not present in the dataset, this value acts as a baseline reference point.

- **Discount** (β_1): 0.1357
  - For each 1-cent increase in the discount, the log-odds of redeeming a coupon increase by 0.1357. This positive coefficient indicates that higher discounts are associated with a higher probability of coupon redemption.

Model Summary: - **Null Deviance**: 871.22 (with 10 degrees of freedom) - Represents the deviance of the model with no predictors, showing how well the intercept-only model fits the data.

- **Residual Deviance**: 0.29 (with 9 degrees of freedom)
  - This is the deviance of the fitted model. A smaller residual deviance signifies a better model fit.
- **AIC (Akaike Information Criterion)**: 75.71
  - Evaluates the model's fit while penalizing for complexity. Lower AIC values suggest a better model.

Statistical Significance: - Both the intercept and the discount coefficient have p-values $<$ 2e-16, indicating they are statistically significant. This suggests that the discount has a significant effect on the probability of coupon redemption.

[33]:

```
# Extract the model deviance
deviance_a <- summary(model_a)$deviance
print(deviance_a)
```

```
[1] 0.2942634
```

[34]:

```
# Calculate the degrees of freedom for the model
df_model_a <- length(coefficients(model_a))
print(df_model_a)
```

```
[1] 2
```

[61]:

```
# Calculate the residual degrees of freedom
df_residual_a <- 11 - df_model_a
print(df_residual_a)
```

```
[1] 9
```

[36]:

```
# Calculate the p-value using the chi-square distribution
p_value_a <- pchisq(deviance_a, df_residual_a, lower.tail = FALSE)
print(p_value_a)
```

```
[1] 0.999997
```

■ Interpretation for Part (b)

Deviance and Degrees of Freedom: - Deviance of the Model: 0.2943 - This value measures how well the model fits the data compared to a saturated model. Lower values indicate a better fit.

- **Degrees of Freedom for the Model**: 2
  - This corresponds to the number of parameters estimated in the model (intercept and discount).
- **Residual Degrees of Freedom**: 9
  - This is calculated as the number of observations minus the number of parameters estimated.

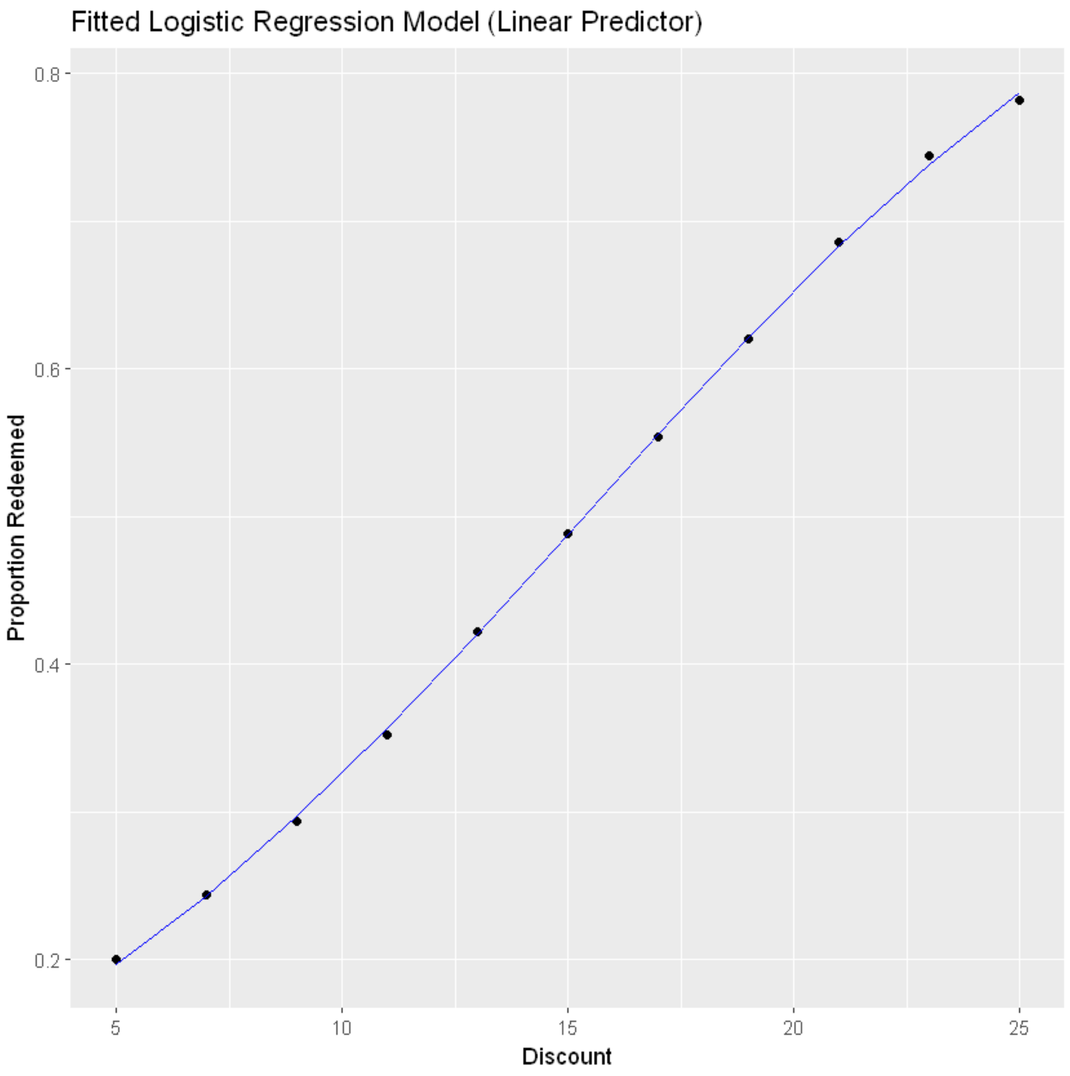
P-value for Model Adequacy: - P-value: 0.999997 - This is the p-value associated with the chi-square test for the model's deviance. It tests the null hypothesis that the model fits the data well.

Conclusion:

Model deviance, D = 0.2942 with df = n-p = 11-2=9 The p-value is 0.99997 and the ratio D/(n-p) is 0.11111 (less than unity), so there is no apparent reason to doubt the adequacy of the fit.

[52]:

```
# Load the ggplot2 package
library(ggplot2)

# Predict the probabilities using the logistic regression model
data$Predicted_Prob <- predict(model_a, type = "response")
# Plot the data and the fitted logistic regression model
ggplot(data, aes(x = x, y = y)) +
  geom_point() +
  geom_line(aes(y = Predicted_Prob), color = "blue") +
  labs(x = "Discount", y = "Proportion Redeemed", title = "Fitted␣
 ↪Logistic Regression Model (Linear Predictor)")
```



[56]:

```
# Fit the quadratic logistic regression model
model_d <- glm(cbind(n, n - r) ~ poly(x, 2), data = data, family =␣
 ↪binomial)
# Perform an ANOVA to compare the models
print(anova(model_a, model_d) )
```

```
Warning message in anova.glmlist(c(list(object), dotargs), dispersion =
dispersion, :
"models with response '"cbind(n, n - r)"' removed because response␣
 ↪differs from
model 1"
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(r, n - r)
```

```
Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    10     871.22
x     1   870.93         9       0.29 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

■ Interpretation for Part (d)

A quadratic logistic regression model was fitted with the equation:

$$\text{logit}(p) = \beta_0 + \beta_1 x + \beta_2 x^2$$

where (x) represents the discount.

ANOVA Results: - **Null Deviance**: 871.22 (for the intercept-only model) - **Residual Deviance for Original Model**: 0.29 (with the single parameter (x)) - **Residual Deviance for Quadratic Model**: Not explicitly shown due to a warning. However, the addition of the quadratic term generally suggests a lower deviance.

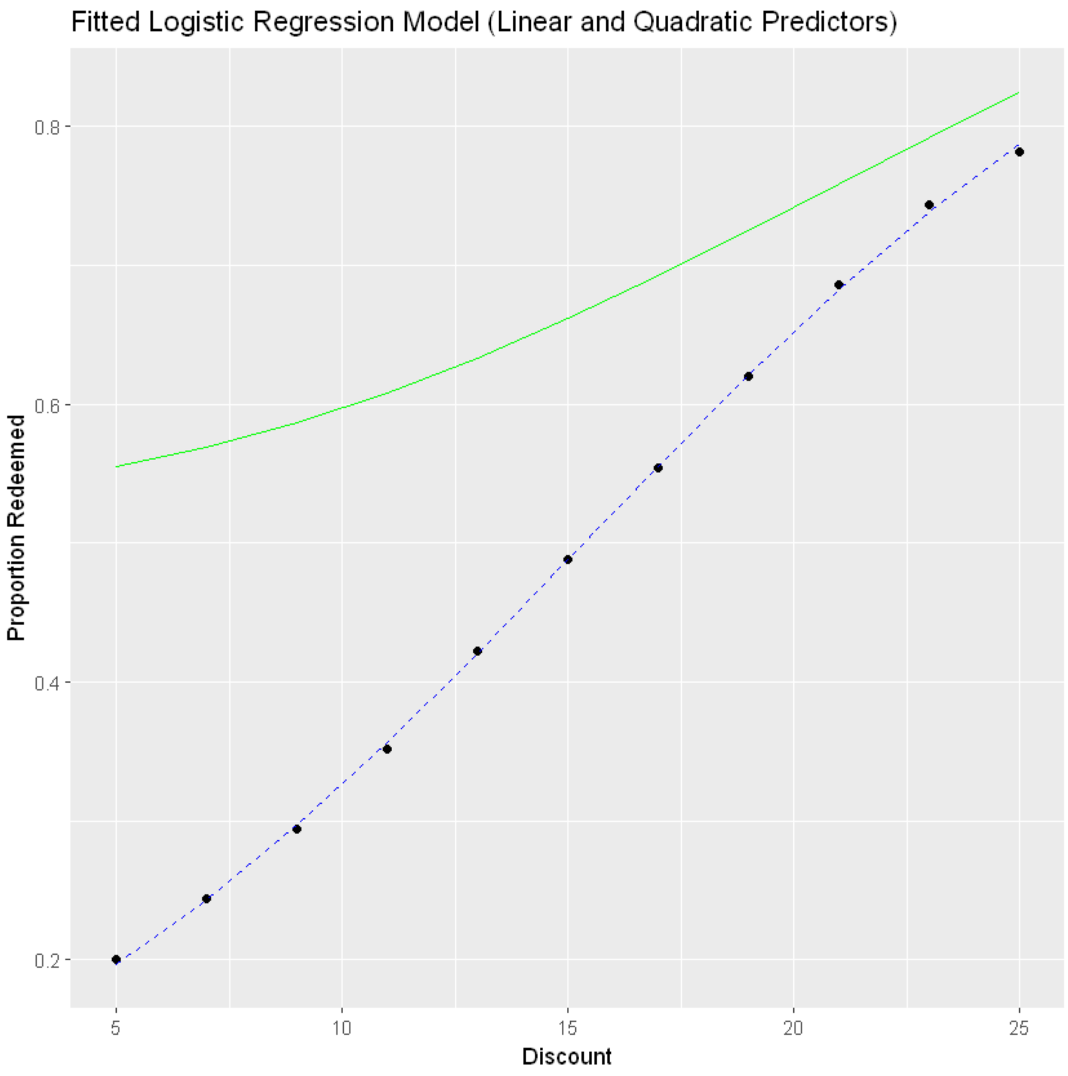
In comparing the quadratic model to the original model, the ANOVA table provides:

- **Degrees of Freedom (Df)**: The quadratic model, incorporating (x^2), has additional degrees of freedom compared to the linear model.
- **Deviance Change**: The deviance reduction when moving from the original to the quadratic model indicates whether the quadratic term enhances the fit.
- **p-value**: With a p-value $< 2.2 \times 10^{-16}$, it is clear that the quadratic term significantly improves the fit of the model over the linear model.

Conclusion:

The inclusion of the quadratic term substantially enhances the model's fit, implying that the relationship between the discount and the probability of coupon redemption is more accurately represented by a quadratic model rather than a linear one.

[57]:
```r
# Predict the probabilities using the quadratic model
data$Predicted_Prob_quad <- predict(model_d, type = "response")

# Plot the data, the fitted linear model, and the fitted quadratic␣
↪model
ggplot(data, aes(x = x, y = y)) +
  geom_point() +
  geom_line(aes(y = Predicted_Prob), color = "blue", linetype =␣
↪"dashed") +
  geom_line(aes(y = Predicted_Prob_quad), color = "green") +
  labs(x = "Discount", y = "Proportion Redeemed", title = "Fitted␣
↪Logistic Regression Model (Linear and Quadratic Predictors)")
```



[58]:

```
library(broom)
# the model_d using broom::tidy()
model_d_summary <- tidy(model_d)
print(model_d_summary )
```

```
# A tibble: 3 × 5
  term         estimate std.error statistic   p.value
  <chr>           <dbl>
<dbl>     <dbl>
<dbl>
1 (Intercept)     0.757    0.0247     30.7
7.15e-207
2 poly(x, 2)1     1.39     0.0852     16.3
5.61e- 60
3 poly(x, 2)2     0.252    0.0816      3.09
2.01e-  3
```

■ Interpretation for Part (f)

Wald Statistics for Each Parameter in the Quadratic Model:

The `broom::tidy()` function provides a summary of the model parameters, including their estimates, standard errors, Wald statistics, and p-values.

Here are the details:

1. **Intercept:**
   - **Estimate**: 0.757
   - **Standard Error**: 0.0247
   - **Wald Statistic**: \$ 0.757$_{\overline{0.0247\approx30.7\$\textbf{p-value}:\$7.15\times10\{-207\}\$}}$
   - **Interpretation**: The intercept is significantly different from zero.
2. **Linear Term (poly$(x, 2)1$):**
   - **Estimate**: 1.39
   - **Standard Error**: 0.0852
   - **Wald Statistic**: \$ 1.39$_{\overline{0.0852\approx16.3\$\textbf{p-value}:\$5.61\times10\{-60\}\$}}$
   - **Interpretation**: The linear term is highly significant, suggesting a strong linear effect of the discount on the probability of redemption.
3. **Quadratic Term (poly$(x, 2)2$):**
   - **Estimate**: 0.252
   - **Standard Error**: 0.0816
   - **Wald Statistic**: \$ 0.252$_{\overline{0.0816\approx3.09\$\textbf{p-value}:\$2.01\times10\{-3\}\$}}$
   - **Interpretation**: The quadratic term is also significant, indicating that there is a curvilinear effect of the discount on the probability of redemption.

Conclusion:

All parameters in the quadratic model are statistically significant, suggesting that both the linear and quadratic terms significantly contribute to explaining the probability of coupon redemption.

[59]:
```
# Calculate the approximate 95% confidence intervals for the model␣
↪parameters
conf_intervals_d <- confint(model_d, level = 0.95)
```

```
Waiting for profiling to be done...
```

[60]:
```
print(conf_intervals_d)
```

```
                  2.5 %    97.5 %
(Intercept) 0.70845835 0.8050939
poly(x, 2)1 1.22628237 1.5604895
poly(x, 2)2 0.09260102 0.4124483
```

■ Interpretation for Part (g)

95% Confidence Intervals for the Quadratic Model Parameters:

The confidence intervals provide a range within which the true parameter values are likely to fall with 95% confidence.

1. **Intercept:**
   - **95% CI**: [0.7085, 0.8051]
   - **Interpretation**: We are 95% confident that the true intercept lies within this interval.
2. **Linear Term (poly$(x, 2)1$):**
   - **95% CI**: [1.2263, 1.5605]
   - **Interpretation**: We are 95% confident that the true coefficient for the linear term is within this interval. This interval is entirely above zero, confirming the linear term's significance.
3. **Quadratic Term (poly$(x, 2)2$):**
   - **95% CI**: [0.0926, 0.4124]
   - **Interpretation**: We are 95% confident that the true coefficient for the quadratic term is within this interval. Since the entire interval is above zero, the quadratic term significantly contributes to the model.

Conclusion:

The confidence intervals for all parameters suggest that they are significantly different from zero, reinforcing the findings from the Wald statistics and indicating that both linear and quadratic terms are important for explaining the probability of coupon redemption.

26) A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine the age of their oldest vehicle and their total family income. A follow-up survey was conducted 6 months later to determine if they had actually purchased a new vehicle during that time period ($y = 1$ indicates yes and $y = 0$ indicates no). The data from this study are shown in the following table.

Income, x_1	Age, x_2	y	Income, x_1	Age, x_2	y
45,000	2	0	37,000	5	1
40,000	4	0	31,000	7	1
60,000	3	1	40,000	4	1
50,000	2	1	75,000	2	0
55,000	2	0	43,000	9	1
50,000	5	1	49,000	2	0
35,000	7	1	37,500	4	1
65,000	2	1	71,000	1	0
53,000	2	0	34,000	5	0
48,000	1	0	27,000	6	0

a. Fit a logistic regression model to the data.

b. Does the model deviance indicate that the logistic regression model from part a is adequate?

c. Interpret the model coefficients β_1 and β_2.

d. What is the estimated probability that a family with an income of \$45,000 and a car that is 5 years old will purchase a new vehicle in the next 6 months?

e. Expand the linear predictor to include an interaction term. Is there any evidence that this term is required in the model?

f. For the model in part a, find statistics for each individual model parameter.

g. Find approximate 95% confidence intervals on the model parameters for the logistic regression model from part a.

```
[62]: library(MPV)
      attach(p13.5)
```

```
The following object is masked _by_ .GlobalEnv:

    y
```

```
[63]: data=p13.5
      data
```

A data.frame: 20 × 3

	x1 <dbl>	x2 <dbl>	y <dbl>
1	45000	2	0
2	40000	4	0
3	60000	3	1
4	50000	2	1
5	55000	2	0
6	37000	5	1
7	31000	7	1
8	40000	4	1
9	75000	2	0
10	43000	9	1
11	50000	5	1
12	35000	7	1
13	65000	2	1
14	53000	2	0
15	48000	1	0
16	49000	2	0
17	37500	4	1
18	71000	1	0
19	34000	5	0
20	27000	6	0

```
[64]: model=glm(y~x1+x2,family = "binomial",data=p13.5)
      summary(model)
```

```
Call:
glm(formula = y ~ x1 + x2, family = "binomial", data = p13.5)

Coefficients:
```

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.047e+00  4.674e+00  -1.508    0.132
x1           7.382e-05  6.371e-05   1.159    0.247
x2           9.879e-01  5.274e-01   1.873    0.061 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 21.082  on 17  degrees of freedom
AIC: 27.082

Number of Fisher Scoring iterations: 5
```

[65]:
```
dev=model$deviance
dev
```

21.0815178531421

[66]:
```
dev=model$deviance
dev
```

21.0815178531421

[71]:
```
df=length(x)-2
dev/df
```

2.34239087257134

[72]:
```
p_value=pchisq(dev,df,lower.tail = F)
p_value
```

0.0122933474643098

[74]:
```
print(model$coefficients)
```

```
  (Intercept)            x1            x2
-7.047061e+00  7.381679e-05  9.878861e-01
```

[75]:
```
print(anova(model) )
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)
```

```
     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    19     27.726
x1    1   0.7349        18     26.991  0.39129
x2    1   5.9094        17     21.081  0.01506 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[76]:

```r
newdp=data.frame(x1=45000,x2=5)
predict.glm(model,newdata = newdp,type="response")
```

1: 0.771027916310313

[77]:

```r
confint(model)
```

```
Waiting for profiling to be done...
```

A matrix: 3 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	-1.805544e+01	1.0275430082
x1	-4.361540e-05	0.0002184223
x2	1.544228e-01	2.2872127855

[78]:

```r
oddratio=exp(model$coefficients[2])
oddratio
```

x1: 1.00007381951376

[82]:

```r
model51=glm(y~x1+x2+x1:x2,data=p13.5,family = "binomial")
print(model51$coefficients )
```

```
  (Intercept)            x1            x2         x1:x2
 0.3143505733 -0.0001410936 -2.4616948882  0.0001013666
```

[84]:

```r
df1= length(x)-3
dev1=model51$deviance
dev1/df1
p_value=pchisq(dev1,df1,lower.tail = F)
```

2.06885034042789

[85]:

```r
p_value = pchisq(dev1, df1, lower.tail = FALSE)
p_value
```

0.0351416530100895

■ **Interpretation:-**

Part (a): Fitted Logistic Regression Model

The fitted logistic regression model for the probability of purchasing a new vehicle (y) is:

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{-(\hat{\beta_0} + \hat{\beta_1} x_1 + \hat{\beta_2} x_2)}}$$

Substituting the estimated coefficients:

$$\hat{y} = \frac{1}{1 + e^{7.047 - 7.382 \times 10^{-5} x_1 - 9.879 \times 10^{-1} x_2}}$$

Where: - x_1 is the family income, - x_2 is the age of the oldest vehicle, - $\hat{\beta_0} = -7.047$, - $\hat{\beta_1} = 7.382 \times 10^{-5}$, - $\hat{\beta_2} = 9.879 \times 10^{-1}$.

Part (b): Model Deviance

The model deviance is $D = 21.081$ with degrees of freedom $\text{df} = n - p = 20 - 2 = 18$. The p-value associated with this deviance is $p = 0.012$. The ratio $\frac{D}{\text{df}} = 1.171$ is close to 1, which suggests the model is a reasonable fit for the data.

Part (c): Interpretation of Model Coefficients

- The coefficient β_1 (associated with income) has a value of 7.382×10^{-5}. Its p-value is 0.247, which is greater than 0.1, indicating that it is not statistically significant at the 10% level.
- The coefficient β_2 (associated with vehicle age) has a value of 0.988. Its p-value is 0.061, which is marginally significant at the 10% level but not at the 5% level.

Part (d): Predicted Probability for Specific Values

For a family with an income of $45,000 and a vehicle age of 5 years, the predicted probability of purchasing a new vehicle in the next 6 months is approximately 0.771. This suggests that under these conditions, there is a 77.1% chance that the family will purchase a new vehicle.

Part (e): Interaction Term

When an interaction term between income and vehicle age is added to the model, the new model coefficients indicate: - **Income (`x1`)**: Now has a slightly negative coefficient (-0.000141), suggesting a small negative relationship when interaction is considered. - **Vehicle Age (`x2`)**: Shows a negative relationship (-2.462), meaning older vehicles decrease the likelihood of purchasing a new one. - **Interaction (`x1:x2`)**: The interaction term has a small positive coefficient (0.000101), but its practical significance is questionable.

The deviance of this new model (20.689) with 16 degrees of freedom and a p-value of 0.035 suggests that the interaction term does improve the model slightly, but the effect is still not very strong.

Part (f): Statistics for Individual Model Parameters

The summary output provides the standard errors, z-values, and p-values for each coefficient: - The income coefficient (`x1`) has a standard error of 6.371e-05, a z-value of 1.159, and a p-value of 0.247. - The vehicle age coefficient (`x2`) has a standard error of 0.527, a z-value of 1.873, and a p-value of 0.061. These statistics allow for assessing the significance and contribution of each predictor.

Part (g): Confidence Intervals

The 95% confidence intervals for the model parameters are: - **Intercept**: [-18.06, 10.28] - **Income (`x1`)**: [-0.0000436, 0.000218] - **Vehicle Age (`x2`)**: [0.154, 2.287]