6.5 (a) Find the scatter plot and correlation matrices.

Answer: The scatter plot matrix is seen in figure 1.

The correlation matrix is:

$$\begin{array}{c} \\ Y_i \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{ccc} Y_i & X_1 & X_2 \end{array} \\ \begin{pmatrix} 1 & 0.89 & 0.395 \\ 0.89 & 1 & 0 \\ 0.395 & 0 & 1 \end{pmatrix} \end{array}$$

(b) Fit regression model (6.1) to the data. State the estimated regression function. How is $\hat{\beta}_1$ interpreted here?

Answer: The regression model with two predictor variables is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

which is

$$\hat{Y} = 37.65 + 4.42X_1 + 4.38X_2$$

$\hat{\beta}_1 = 4.42$ indicated the for every 1 unit of moisture content, (X_1) the degree of brand liking increases by 4.42 units.

(c) Obtain the residuals and prepare a box plot of the residuals. What information does this plot provide.

Answer: Figure 3 below. The boxplot shows the residuals are fairly well behaved, centered around 0 without any obvious traces of non-normality.

Table 1: Residuals, stargazer packages

1	2	3	4	5	6	7	8
-0.10	0.15	-3.10	3.15	-0.95	-1.70	-1.95	1.30
9	10	11	12	13	14	15	16
1.20	-1.55	4.20	2.45	-2.65	-4.40	3.35	0.60

(d) Plot the residuals against $\hat{Y}, X_1, X_2$ and X_1X_2 on separate graphs. Also prepare a normal probability plot.

Answer: The plots are displayed below, alongside the other plots from the question.

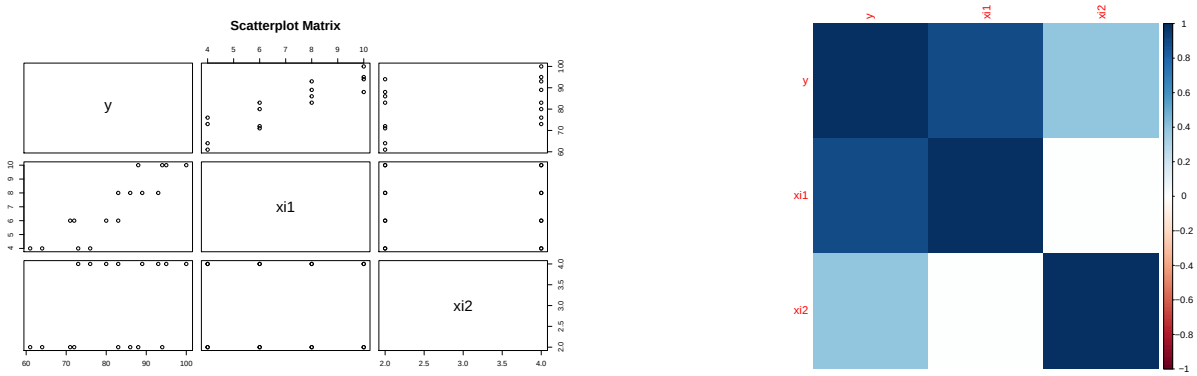

Figure 1: Scatter-plot matrix and correlation plot.

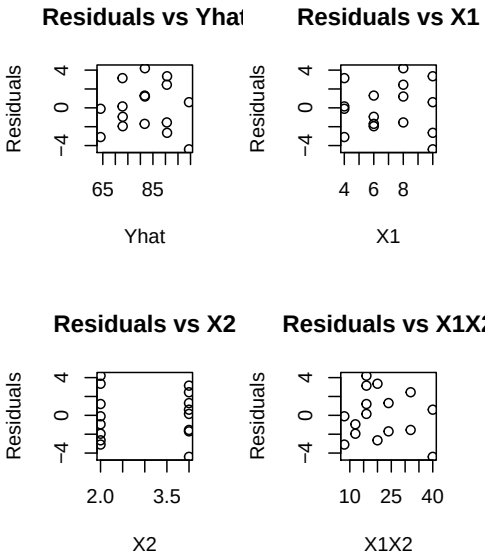

Figure 2

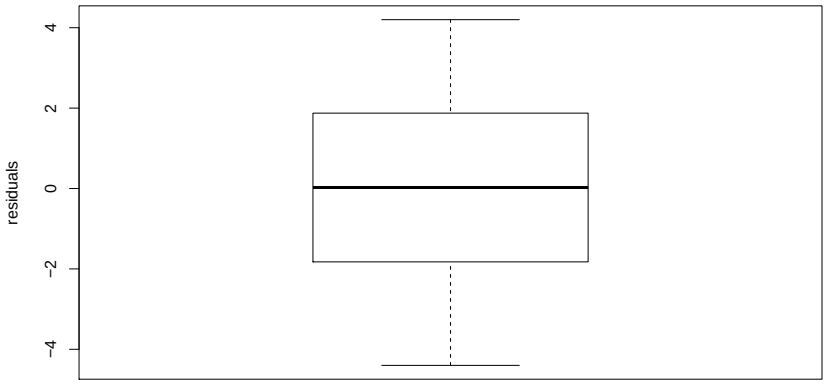

Figure 3: Boxplot of Residuals

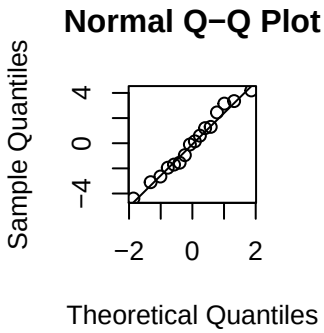

Figure 4

(e) Conduct the Breush-Pagan test for constancy of the error variance, assuming $\log\sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2}$ using $\alpha = 0.01$. State the alternatives, decision rule, and conclusions.

Answer: Breusch-Pagan test. Our null hypothesis is that the error variance is constant, and the alternative is that it isn't. We know $X^2_{BP} = \frac{SSR^*}{2}/(\frac{SSE}{n})^2$ where SSR^* is the regression sum of squares when regressing e^2 on X_1 and X_2 and SSE is the sum of squares for the full regression model, and SSE is the sum of residuals, $\sum_i (Y_i - \hat{Y}_i)^2$. This is easily calculated in matrix form, with $\mathbf{e}^T\mathbf{e}$, where $\mathbf{e}$ is the vector of the residuals.

Now, we compare the test statistic to the critical value determined by $\chi^2(1-\alpha, q)$ where q is the number of predictor variables, in this case, 2. This gives us $X^2_{BP} = 1.04$. At $\alpha = 0.05$, we require $\chi^2(0.99, 2) = 9.21$, so $X^2_{BP} \leq 9.21$, we conclude H_0, that the error variance is constant.

(f) Conduct a formal test for lack of fit of the first-order regression function; use $\alpha = 0.01$. State the alternatives, decision rule, and conclusion.

Answer: Alternative conclusions when testing for lack of fit of a linear regression is that the regression model doesn't hold (for the full model), ie

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
$$H_a : E\{Y\} \neq \beta_0 + \beta_1 X_+ \beta_2 X_2$$

The appropriate test statistic is

$$F^* = \frac{SSLF}{c-p} / \frac{SSPE}{n-c} = \frac{MSLF}{MSPE}$$

the decision rule is

$$\text{if} \quad F^* \leq F(1-\alpha; c-p, n-c), \quad \text{conclude } H_0$$
$$\text{if} \quad F^* > F(1-\alpha; c-p, n-c) \quad \text{conclude } H_a$$

Where in this case c denotes the number of groups with distinct sets of levels for the X variables, ie c is number of coefficients in the full model, which is found by fitting a regression between the Y data and the product of the categorical version of the predictors X_1 and X_2, 8 in this case. This leads to $F^* = 1.04$. The critical value is

$$F(.99; 8-3, 16-8) = 6.632$$

Therefore, since $F^* \leq 6.632$, we conclude H_0.

6.6 Using same data as 6.5.

(a) Test whether there is a regression relation, using $\alpha = 0.01$. State alternatives, decision rule, and conclusion. What does your test imply about β_1 and β_2.

Answer:

$$H_0 : \beta_1 = 0, \beta_2 = 0$$
$$H_a \text{not both } \beta_1 \text{ and } \beta_2 \text{ equal zero}$$

and use test statistic, where $MSR = \frac{SSR}{p-1}$ and $MSE = \frac{SSE}{n-p}$.

$$F^* = \frac{MSR}{MSE} = \frac{936.35}{7.25} = 129.1$$

Meanwhile, we use the critical value $F(1-\alpha, q = 2, n-p = 16-3) = 6.63$, so we conclude H_a (where q is the number of different predictors)

(b) What is the p-value of the test?

Answer: The p-value for X_1 is 1.78×10^{-9}, and for X_2 is 2.01×10^{-5} at $\alpha = 0.05$. At $\alpha = 0.01$, the p-value for the test approaches 0.

(c) Estimate β_1 and β_2 jointly by the Bonferroni procedure, using a 99 percent family confidence coefficient. Interpret your results.

Answer: The $1-\alpha$ confidence limits for β_1 and β_2 for typical regression model by the Bonferroni procedure are

$$\hat{\beta}_1 \pm Bs\{\hat{\beta}_1\} \qquad \hat{\beta}_2 \pm Bs\{\hat{\beta}_2\}$$

where

$$B = t(1-\alpha/4; n-p)$$

In the multiple linear regression model, the estimated variance-covariance matrix is

$$s^2\{\hat{\beta}\} = MSE(X^TX)^{-1}$$

where $s^2\{\hat{\beta}_0\}$ is entry 1,1 of the corresponding 3x3 matrix, $s^2\{\hat{\beta}_1\}$ is the 2,2 entry, and $s^2\{\hat{\beta}_2\}$ is the 3,3 entry, ie the third diagonal. This means

$$s^2\{\hat{\beta}\} = \begin{pmatrix} 8.98 & -.635 & -1.36 \\ -0.63 & 0.091 & 0 \\ -1.36 & 0 & 0.45 \end{pmatrix}$$

So $s\{\hat{\beta}_1\} = 0.30$ and $s\{\hat{\beta}_2\} = 0.67$. Similarly, $B = t(1-\alpha/4 = 1-.01/4, n-p = 16-3) = 3.33$. So,

$$\hat{\beta}_1 \pm Bs\{\hat{\beta}_1\} = 4.43 \pm 3.33 \cdot 0.30 \longrightarrow 3.43 \le \beta_1 \le 5.43$$

$$\hat{\beta}_2 \pm Bs\{\hat{\beta}_2\} = 4.38 \pm 3.33 \cdot 0.67 \longrightarrow 2.15 \le \beta_2 \le 6.61$$

6.7 Using data from 6.5.

(a) Determine coefficient of multiple determination, R^2.

Answer: We calculate R^2 from

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{94.3}{1967} = 0.9521$$

where we measured SSTO and SSE with respect to the set of variables $X_1, \ldots X_{p-1}$, which reduces to the simple case when $p-1=1$. The value of R^2 is 0.9521, more easily found in summary of linear fit.

(b) Calculate the coefficient of simple determination R^2 between Y_i and $\hat{Y}_i$. Does it equal the coefficient of multiple determination?

Answer: We calculate R^2 from (using $\hat{Y}_i$ as a predictor of Y_i)

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{94.3}{1967} = 0.9521$$

the same result as before.

6.8 Using data from 6.5 again.

(a) Obtain an interval estimate of $E\{Y_h\}$ when $X_{h1} = 5$ and $X_{h2} = 4$. Use 99 percent confidence coefficient. Interpret interval estimate.

Answer: The mean response is estimated as

$$E\{Y_h\} = X_h^T \boldsymbol{\beta}$$

From this, we can derive that the $1-\alpha$ confidence limits for $E\{Y_h\}$ are

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{\hat{Y}_h$$

where

$$s^2\{\hat{Y}_h\} = MSE(X_h^T(X^TX)^{-1}X_h) = X_h^T s^2\{\hat{\beta}\}X_h$$

$\hat{Y} = 77.23$ and $s\{\hat{Y}\} = 1.13$. Therefore,

$$77.28 \pm 1.13 \cdot t(1-.01/2; 16-3) = 77.28 \pm 3.012 \cdot 1.13 \rightarrow 73.88 \le E\{Y\} \le 80.68$$

(b) Obtain a prediction interval for a new observation $Y_{h(\textbf{new})}$ when $X_{h1} = 5$ and $X_{h2} = 4$. Use 99% confidence coefficient.

Answer: The $1-\alpha$ prediction limits for a new observation $Y_{h(\text{new})}$ corresponding to X_h, the specified values of the X variables, are:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{\text{pred}\}$$

where

$$s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\}$$

We calculated $s^2\{\hat{Y}_h$ in part a, and the MSE has been calculated from other problems, and is equal to 7.25. Therefore,

$$s^2\{\text{pred}\} = 7.25 + 1.27 \longrightarrow s\{\text{pred}\} = 2.92$$

and the t-value is again 3.012. So,

$$\hat{Y}_h \pm 3.012 \cdot 2.92 \longrightarrow 77.28 \pm 8.79$$
$$68.489 \leq Y_{h(\text{new})} \leq 86.07$$

7.3 Again using the data from problem 6.5.

i. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sum of squares associated with X_1 and with X_2, given X_1.

Answer:

We look for $SSR(X_1)$ and $SSR(X_2|X_1)$. (where $SSR = \sum_i \hat{Y}_i - \overline{Y}_i$).

$$SSR(X_1) = 1566.5$$
$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$$
$$SSR(X_2|X_1) = 1872.7 - 1566.5 = 306.25$$

(c) Test whether X_2 can be dropped from the regression model given that X_1 is retained. Use the F^* test statistic and level of significance 0.01. State the alternatives, decision rule, and conclusion. What is the p-value of the test?

Answer: When we wish to test whether the term $\beta_2 X_2$ can be dropped from a multiple regression model, we are interested in the alternatives

$$H_0 : \beta_2 = 0$$
$$H_a \beta_2 \neq 0$$

The test statistic is

$$t^* = \frac{\hat{\beta}_2}{s\{\hat{\beta}_2\}}$$

Alternatively, we have from the full model that

$$SSE(F) = SSE(X_1, X_2)$$

The reduced model would be

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

And

$$SSE(R) = SSE(X_1)$$

There are $n-3$ degrees of freedom with the full, and $n-2$ with the reduced. The test statistic becomes

$$F^* = \frac{SSE(X_1) - SSE(X_1, X_2)}{(n-2)-(n-3)} \Big/ \frac{SSE(X_1, X_2)}{n-3} = \frac{SSR(X_2|X_1)}{1} \Big/ \frac{SSE(X_1, X_2)}{n-3}$$

So using $SSR(X_2|X_1)$ from part (a), and the SSE of the full model we calculated at many points, 94.3.

$$F^* = \frac{306.25}{94.3/3} = 42.22$$

We could also calculate

$$t^* = \frac{\hat{\beta}_2}{s\{\hat{\beta}_2\}} = \frac{4.375}{0.673} = 6.50$$

so $(t^*)^2 = 42.22$, the same as expected. We compare this value to $F(0.99, q = 1, 16-3) = 9.07$, so since $F^* > 9.07$, we conclude H_a (where $q = 1$ because this is the q of the reduced model in the test where we removed X_2). We should not drop X_2.

7.24 Once again we use the data from question 6.5.

(a) Fit first order simple linear regression model for relating brand liking (Y) to moisture content (X_1). State the fitted regression function.

Answer: The regression equation is

$$\hat{Y} = 50.78 + 4.43X_1$$

(b) Compare the estimated regression coefficient for moisture content obtained in part (a) with the corresponding coefficient obtained in Problem 6.5b. What do you find?

Answer: It is the same.

(c) Does $SSR(X_1)$ equal $SSR(X_1|X_2)$ here? If not, is the difference substantial?

Answer: Recall that

$$\text{SSTO} = \text{SSR} + \text{SSE}$$

where

$$\text{SSR} = \sum_i (\hat{Y}_i - \bar{Y})^2$$
$$\text{SSE} = \sum_i \sum_i (Y_i - \hat{Y}_i)^2$$
$$\text{SSTO} = \sum_i (Y_i - \bar{Y})^2$$

We note that

$$\text{SSR}(X_1|X_2) = \text{SSR}(X_1, X_2) - \text{SSR}(X_2)$$

$$\text{SSR}(X_1) = 1566.45 \qquad \text{SSR}(X_2) = 306.25 \qquad \text{SSR}(X_1|X_2) = 1566.45$$

(d) Refer to the correlation matrix obtained in problem 6.5a. What bearing does this have on your findings in parts (b) and (c)?

Answer: The correlation between X_1 and X_2, denoted r_{12} is given by $\pm\sqrt{R^2} = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}}$, or in alternative form

$$r_{12} = \frac{\sum_i (Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\left[\sum_i (Y_{i1} - \bar{Y}_1)^2 \sum_i (Y_{i2} - \bar{Y}_2)^2\right]^{1/2}}$$

The correlation between X_1 and X_2 is zero. The matrix tells us this as well. There is no linear relation between the X_1 and X_2.