Solution to HW 5

Guanhua FANG

November 2, 2017

1

First, we recall the one dimension case. If $Y = cX$ and the variance of X is σ^2, then it is easy to see that

$$\mathrm{var}(Y) = \mathrm{var}(cX) = \mathrm{E}(c^2X^2) - [\mathrm{E}(cX)]^2 = c^2\mathrm{var}(X).$$

and

$$\mathrm{cov}(c_1X_1, c_2X_2) = \mathrm{E}(c_1c_2X_1X_2) - \mathrm{E}(c_1X_1)\mathrm{E}(c_2X_2) = c_1c_2\mathrm{cov}(X_1, X_2)$$

Then, suppose $A = (a_1, \ldots, a_n)$ is a 1 by n vector,

$$\mathrm{var}(Y) = \mathrm{var}(A^TX) = \mathrm{var}(\sum a_ix_i) = \sum a_ia_j\mathrm{cov}(X_i, X_j) = \sum a_ia_j\Sigma_{ij} = A\Sigma A^T.$$

Furthermore, we assume A is k by n matrix. What we only need to consider is the covariance between A_iX and A_jX where A_i is the ith row of A. We repeat the above display and get $\mathrm{cov}(A_iX, A_jX) = A_i\Sigma A_j^T$. Lastly, we put all i, j together and get $\mathrm{AX} = A\Sigma A^T$.

2

We know that $s^2(\mathbf{b}_1) = MSE[(X^TX)^{-1}]_{11}$. $\mathbf{b}_1 = [(X^TX)^{-1}X^TY]_1$. We also know that $t = \mathbf{b}_1/s(\mathbf{b}_1)$, then $t^2 = \mathbf{b}_1^2/s^2(\mathbf{b}_1)$. Since we know that $MSE \sim \chi^2(n-p-1)$ and $\mathbf{b}_1 \perp MSE$, what we only need to show is that $\mathbf{b}_1/\sqrt{[(X^TX)^{-1}]_{11}}$ is a standard normal distribution. (Here, we assume $\sigma = 1$ without loss of generality.) Notice that, $\mathbf{b}_1$ is the linear combination of normal distribution, then it is still a normal distribution. Under null, $\mathrm{E}\mathbf{b}_1 = \beta_1 = 0$ and $\mathrm{var}(\mathbf{b}_1) = (X^TX)_{11}/(X^TX)_{11} = 1$. Thus, $t^2 \sim F(1, n-p-1)$.

You can also work on this problem by directly showing that $t^2 = F$, that is,

$$(\frac{b_1}{sd(b_1)})^2 = \frac{SSE^{(1)} - SSE}{MSE} \tag{1}$$

where SSE^1 is SSE of the regression without first column of X. Write $X = (a, X_1)$, a is the first column of X and X_1 is the rest block of X. Then, we can compute

$$
\begin{aligned}
(X^T X)^{-1} &= \left(((a, X_1))^T (a, X_1)\right)^{-1} \\
&= \begin{pmatrix} p^{-1} & -a^T X_1 (X_1^T X_1)^{-1} p^{-1} \\ -(X_1^T X_1)^{-1} X_1^T a p^{-1} & (X_1^T X_1)^{-1} + p^{-1} (X_1^T X_1)^{-1} X_1^T a a^T X_1 (X_1^T X_1)^{-1} \end{pmatrix} \qquad (2)
\end{aligned}
$$

where we let $p^{-1} = (X^T X)^{-1}$. After we get the inverse matrix, then the following computation is quite standard. What we just to do it to plug this to the formula $SSE = Y^T(I - H)Y$ and $SSE^{(1)} = Y^T(I - H^{(1)})Y$. The computation detail is omitted here, you can try by yourself.

6.5

a. We can see that both X_1 and X_2 are categorical variables and there is a linear relationship

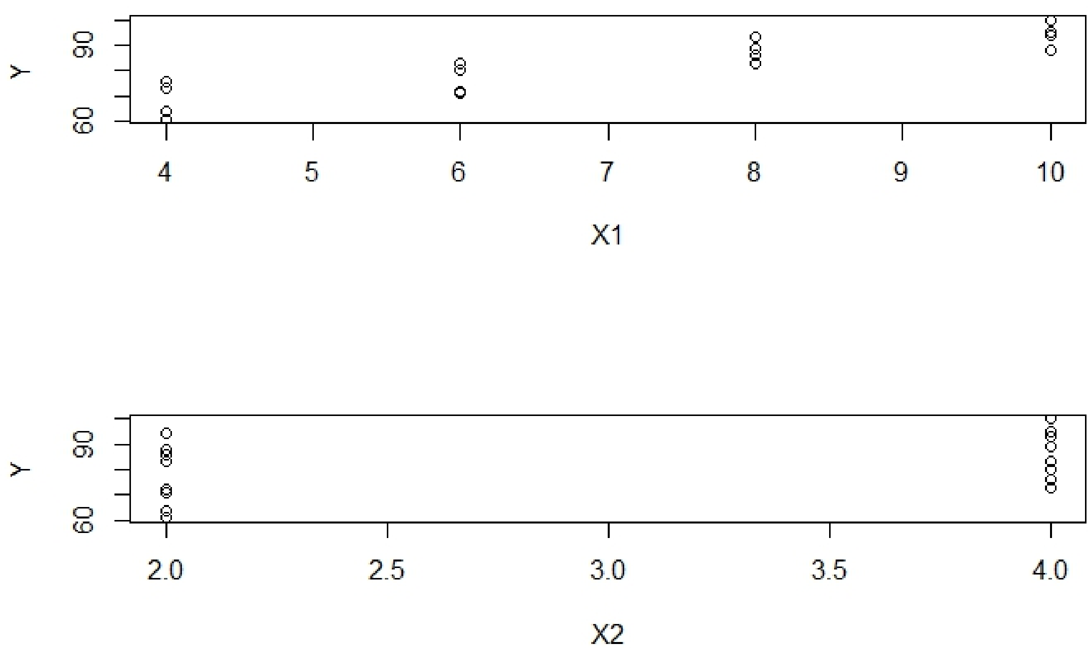

Figure 1: scatter plot for 6.5.a

between X_1 and Y.

b. The estimated regression function is $Y = 37.65 + 4.42X_1 + 4.38X_2$. The degree of brand liking will increases by 4.42 if moisture content is increased by 1.

c. From figure, we can see that the residual are quite symmetric about zero and there is no outlier.

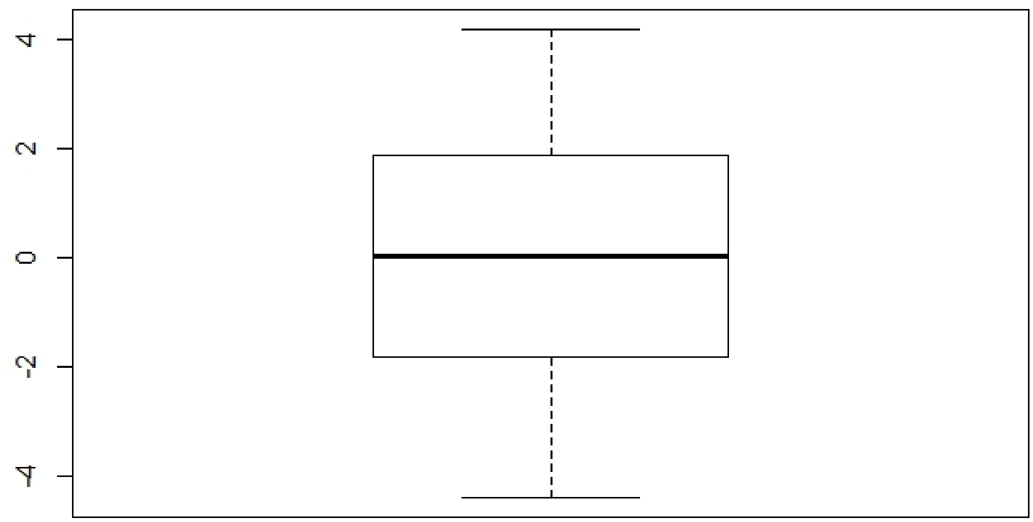

Figure 2: Box plot for 6.5.c

f. $H_0 : \mathrm{E}Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and $H_a : \mathrm{E}Y \neq \beta_0 + \beta_1 X_1 + \beta_2 X_2$. There are $c = 8$ groups. The $SSPE$ equals $2 * 1.5^2 + 2 * 1.5^2 + 2 * 0.5^2 + 2 * 1.5^2 + 2 * 1.5^2 + 2 * 2^2 + 2 * 3^2 + 2 * 2.5^2 = 57$. $SSE = 2.693^2 * 13 = 94.3$. Then, $SSLF = 94.3 - 57 = 37.3$. Hence, $F = \frac{SSLF/(8-3)}{SSPE/(16-8)} = 1.04$. We know that $F_{0.99}(5, 8) = 6.6$. We fail to reject. The model is not lack of fit.

6.7

a. $R^2 = 0.95$, we know that $R^2 = 1 - \frac{SSE}{SSTO}$. Hence ,the high value means the model fits the data well.

b. The $r^2 = 0.95$, it is equal to the value in part (a) by using the fact $JH = J$, $HJ = J$, $\frac{1}{n}J\frac{1}{n}J = \frac{1}{n}J$ and $H^2 = H$. The computation detail is omit, it is quite standard.

6.8

a. $Y_h = 77.27$ when $X_1 = 5, X_2 = 4$. The MSE is 2.693^2. $s^2(Y_h) = MSE(1, 5, 4)(X^T X)^{-1}(1, 5, 4)^T = 2.693^2 * 0.175 = 1.27$. Hence, the 99% confidence interval is $(77.27 + t_{0.005}(13) * \sqrt{1.27}, 77.27 + t_{0.995}(13) * \sqrt{1.27}) = (73.9, 80.7)$.

b. $Y_new = 77.27$ when $X_1 = 5, X_2 = 4$. $s^2(Y_new) = MSE(1 + (1, 5, 4)(X^T X)^{-1}(1, 5, 4)^T) = 8.52$. Then, 99% prediction interval is $(77.27 + t_{0.005}(13) * \sqrt{8.52}, 77.27 + t_{0.995}(13) * \sqrt{8.52}) = (68.5, 86.1)$.

6.25

If we know that $\beta_2 = 4$, we then define $\tilde{y}_i = y_i - \beta_2 X_{i2}$. We could fit model $\tilde{Y} \sim X_{i1}, X_{i3}$ and get the other estimates.