Tutorial Problems

Question 1

In a study of high-density lipoprotein (HDL, labelled y) in human blood a sample of size 42 was used. Measurements were taken on total cholesterol (x_1), total trigylceride (x_2) as well as noting whether a sticky component, sinking pre-beta (SPB, labelled x_3) was present (coded 1) or absent (coded 0). This data is stored in the data frame `dat` and a partial analysis is given in the R-output at the end of the question. The basis for the analysis is the model

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i,$$

where $\epsilon_i \sim NID(0, \sigma^2)$.

(a) Write down the fitted multiple regression model. What proportion of the total variability in Y is explained by the multiple regression? (Lecture 6)

Solution: The fitted model of `lm1` is

$$\hat{Y} = 48.8645 - 0.0273x_1 + 0.0150x_2 + 8.1483x_3.$$

The model explains only 17.77% of the variability in Y from `Multiple R-squared`.

(b) Provide a 95% confidence interval for β_2 assuming that the model passed all diagnostic checks. (Lectures 5-6)

Solution: Using `Extimate` and `Std. Error` from `lm1`, a 95% C.I. for β_2 is

$$0.01504 \pm t_{38}(0.975) \times 0.02485 = 0.015 \pm 0.050.$$

(c) What is the purpose of the residual versus fitted values plot? Are the above model assumptions reasonable? Justify your answer. (Lecture 6 or any lecture that shows such a plot)

Solution: The residual vs fitted value plot is used to check the common variance assumption. The plot should not show any strong pattern if the model is correct. There should be an even spread of points as the fitted values vary.

The residual vs fitted value plot seems fine here. The normal Q-Q plot is approximately linear (except for the lower extreme point) indicating that the model assumptions (NID$(0, \sigma^2)$ error terms) are reasonable.

(d) Are there any high leverage points in this data set? What characterises a high leverage point in general? (Lecture 8)

Solution: A point is a high leverage point in this case if

$$h_{ii} > 2 \times 4/42 = 0.1905$$

where h_{ii} is from `hat` of `lm.influence(lm1)`. There is one high leverage point, [10] with the largest $h_{ii} = 0.326$. High leverage points have extreme x vectors.

(e) Are there any outliers in the data? (Lecture 8)

Solution: From the Cook's distance we see that [10] has the largest value around 0.25, but this is not close to $F^{-1}_{4,38}(0.5) = 0.854$ so there are no outliers in this data set.

(f) Test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ assuming the model passes the diagnostics checks. (Lecture 9)

Solution: Under H_1, the model is `lm1` and under $H_0 : \beta_1 = \beta_2 = 0$, the model is `lm3` with only `x3`. The statistic is

$$f_0 = \frac{(RSS_{H_0} - RSS_{H_1})/(p-q)}{RSS_{H_1}/(n-p)} = \frac{(3878.414 - 3793.872)/(4-2)}{3793.872/(42-4)} = 0.423$$

where the two RSS are taken from `deviance()` of respective models. If they are not provided, one can evaluate say $RSS_{H_1} = \hat{\sigma}^2(n-p) = 9.992^2(42-4) = 3793.922$ (with some round-off error) where $\hat{\sigma}$ is taken from `Residual standard error`. The p-value $= \Pr(F_{2,38} \geq 0.423) > 0.05$. Since 0.423 is a small value, we can conclude the p-value > 0.05 even though `qf(0.95,2,38)` is not provided for checking. Thus the data are consistent with H_0, i.e. we can drop both x_1 and x_2 from the model.

(g) For the model $Y_i = \beta_0 + \beta_3 x_3 + \epsilon_i, \quad \epsilon_i \sim NID(0, \sigma^2)$, give a 95% confidence interval for the error standard deviation, σ, for the simple regression on x_3 only given that the model assumptions are satified. (Lectures 5-6)

Solution: A 95% C.I. for σ, using `lm3` is

$$\left(\sqrt{\frac{RSS_0}{\chi^{-1}_{40}(0.975)}},\ \sqrt{\frac{RSS_0}{\chi^{-1}_{40}(0.025)}}\right) = \left(\sqrt{\frac{3878.414}{59.342}},\ \sqrt{\frac{3878.414}{24.433}}\right) = (8.084, 12.599).$$

where the two chi-square quantiles are taken from `qchisq(0.975,40)` and `qchisq(0.025,40)` and df $= n - q = 42 - 2 = 40$.

(h) For the model $Y_i = \beta_0 + \beta_3 x_3 + \epsilon_i, \quad \epsilon_i \sim NID(0, \sigma^2)$,, what is the predicted difference in average HDL levels between people with and without SPB? (Assumed knowledge)

Solution: The predicted difference in average HDL levels between people with and without SPB is $\hat{\beta}_3 = 8.377$. This is because x_3 is categorical with 1 to indicate the presence of SPB.

