Sample Questions for Module 3

- Modules 3 and 4 are about linear models
- Main concepts: theory behind least square estimation for simple and multiple linear regression, properties of $\widehat{\boldsymbol{\beta}}$ and residuals
- Important techniques: hypothesis tests concerning $\boldsymbol{\beta}$ (including contrasts, linear combinations of its elements, *F*-tests)

Sample Question 6

To investigate whether a person's brain size or body size is predictive of his or her intelligence, the performance IQ score (PIQ), the brain size (based on the count obtained from MRI scans given as count/10,000), height (in inches) and weight (in pounds) from a sample of university students were measured.

A multiple regression model was fitted to the data using R and the summary output obtained is given at the end of this question. (Note that some of the entries in the R output are missing and you will be asked to calculate these in the question.).

a) What is the value of the *t*-statistic for $\hat{\beta}_1$?
b) How many observations are in the dataset?
c) What is the estimate of the intercept β_0?
d) Do you reject or fail to reject the null hypothesis $H_0: \beta_2 = 0$ vs $H_a: \beta_2 \neq 0$ at the 5% significance level?

Sample Question 6

e) Consider the hypotheses

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs H_a: at least one $\beta_i \neq 0$, $i = 1,2,3,$

i. State the meaning of the null hypothesis in words.

ii. Write down the value of the appropriate test statistic, the associated degrees of freedom, and P-value, and give your conclusion.

f) Calculate a two-sided 95% confidence interval for β_1, taking into account the other predictor variables.

g) Consider the diagnostic plots from the regression model fitted to the data. Are the assumptions of the satisfied for these data?

Sample Question 6

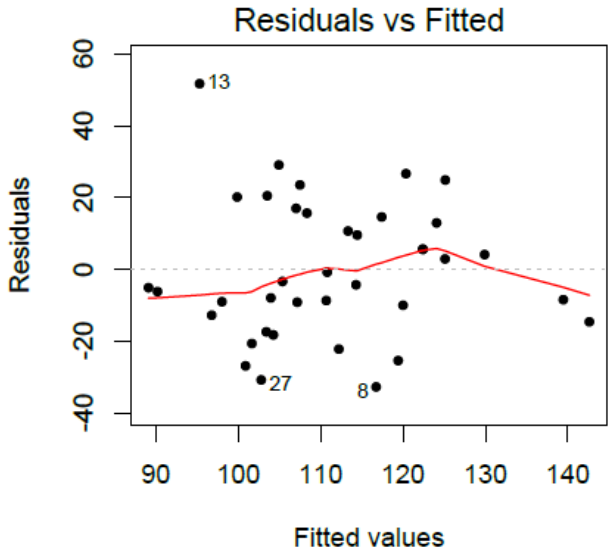
```
Call:
lm(formula = PIQ ~ Brain + Height + Weight, data = piq)

Residuals:
    Min      1Q  Median      3Q     Max
-32.74 -12.09   -3.84   14.17   51.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) ????????? 6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01 ????? 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

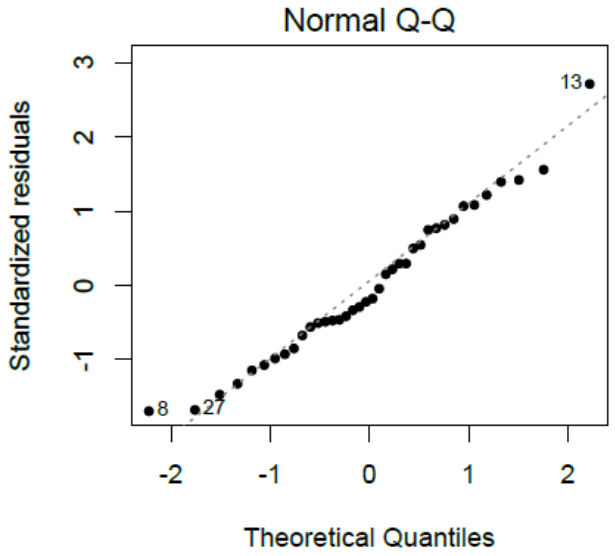
Part f)

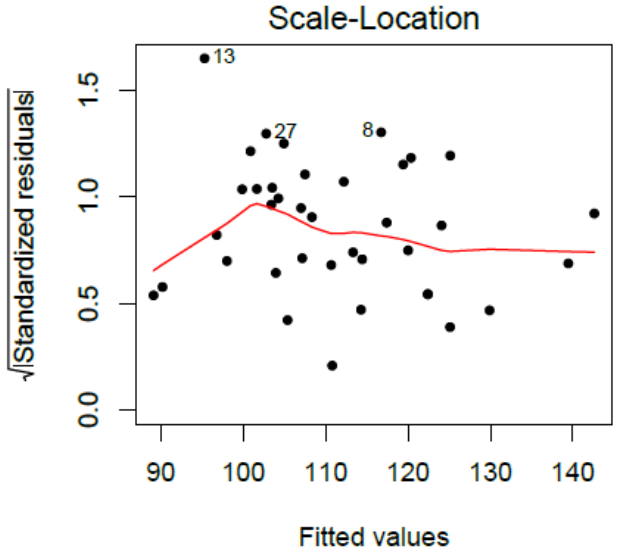
Part d)

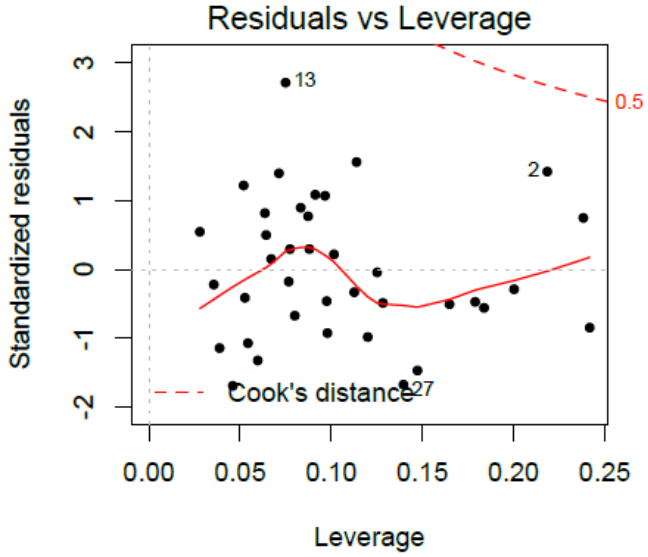
Part e)

Sample Question 6









Sample Question 6 Solution

(a) $t_{\hat{\beta}_1} = \frac{2.06}{0.5634} \approx 3.6564$ $\left(t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)$

(b) $n - p = 34$, $p = 4$

$\therefore \quad n = 34 + 4 = 38$

(c) $\hat{\beta}_0 = t_{\hat{\beta}_0} \times SE(\hat{\beta}_0) = 1.768 \times 62.97 = 111.331$

(d) β_2 corresponds to the row 'Height'

The P-value is 0.033034.

Since this is less than the level of significance (0.05), we reject H_0.

Sample Question 6 Solution

e) i) H_0: there is no significant linear relationship between the three predictors (Brain, Height, Weight) and the response variable (PIQ).

ii) The observed test statistic is $F = 4.741$.
Under H_0, we know that $F \sim F_{3,34}$.
The P-value is 0.007215.
Since this is less than 0.05, we reject H_0 at 5% significance level.
Hence, we conclude that there is a statistically significant (linear) relationship between the three predictors (Brain, Height, Weight) and the response variable (PIQ), at 5% level of significance.

Sample Question 6 Solution

f) $CI = \hat{\beta}_1 \pm t_{0.025, n-p} \, SE(\hat{\beta}_1)$

$= 2.06 \pm 2.032245 \times 0.5634$

$\approx (0.915, 3.205)$

qt(0.975, 34)

≈ 2.032245

g) Linearity: (Check residuals vs fitted plot)

The random scatter looks reasonable, but there is evidence of mild curvature.

Normality: (Check QQ plot)

Apart from two observations in the tails, the residuals are approximately normally distributed.

Constant variance: (Check scale-location plot)

The spread is somewhat equal, so assumption looks reasonable.

Independence: (Check the experimental design)

The IQ of one student should be independent of other students. So assumption may be reasonable.

Sample Question 7

Consider the multiple linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent with $\epsilon_i \sim N(0, \sigma^2)$.

(a) Carefully define what we mean by the least squares estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

(b) State the necessary and sufficient condition on $\boldsymbol{X}$ for $\widehat{\boldsymbol{\beta}}$ to be uniquely identified.

(c) Prove that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}\boldsymbol{y}$.

Sample Question 7 Solution

(a) The least square estimate $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r)^T$ is chosen to minimise the sum of squares error:

$$Q(\beta) = \|y - X\beta\|^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 - \beta_1 x_{i1} + \ldots + \beta_r x_{ir}) \right]^2$$

(b) The columns of X must be linearly independent.

(c) There are a number of ways to prove this.

For example, we could follow Theorem 10 to show that

$$Q(\beta) \geq Q(\hat{\beta})$$

and that equality occurs only if $\beta = \hat{\beta}$.

Or, we could also directly minimise $Q(\beta)$ to find $\hat{\beta}$.

Sample Question 7 Solution

Minimise $Q(\beta) = \|y - X\beta\|^2$ with respect to β

$$= (y - X\beta)^T (y - X\beta)$$

$$= y^T y - y^T X\beta - (X\beta)^T y + (X\beta)^T (X\beta)$$

$$= y^T y - (y^T X)\beta - \beta^T (X^T y) + \beta^T (X^T X)\beta$$

$$\frac{\partial Q(\beta)}{\partial \beta} = -(y^T X)^T - (X^T y) + 2(X^T X)\beta = 0$$

$$-2(X^T y) + 2(X^T X)\beta = 0$$

$$(X^T X)\beta = X^T y$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$