2561551

P1

1. (a) (i) $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $\beta = (\alpha)$ 2½

(ii) $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ 2½

(iii) $Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \end{pmatrix}$ $X = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}$ $\beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$ 2½

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

$X^T X = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} n_1 & & \\ & n_2 & \\ & & n_3 \end{pmatrix}$

$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & & \\ & \frac{1}{n_2} & \\ & & \frac{1}{n_3} \end{pmatrix}$

$X^T Y = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n_1} y_{1i} \\ \sum_{i=1}^{n_2} y_{2i} \\ \sum_{i=1}^{n_3} y_{3i} \end{pmatrix}$

$\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{pmatrix} \frac{1}{n_1} & & \\ & \frac{1}{n_2} & \\ & & \frac{1}{n_3} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n_1} y_{1i} \\ \sum_{i=1}^{n_2} y_{2i} \\ \sum_{i=1}^{n_3} y_{3i} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix}$ 5/5

$\Rightarrow \hat{\beta} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix}$

(c) $RSS = Y^T Y - (X^T Y)^T \hat{\beta}$

$= \sum_{i=1}^{n_1} \sum_{j=1}^{n_i} y_{ij}^2 - \begin{pmatrix} \sum_{i=1}^{n_1} y_{1i} \\ \sum_{i=1}^{n_2} y_{2i} \\ \sum_{i=1}^{n_3} y_{3i} \end{pmatrix}^T \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix}$

$= \sum_{i=1}^{n_1} \sum_{j=1}^{n_i} y_{ij}^2 - \left(\sum_{i=1}^{n_1} y_{1i} \bar{y}_1 + \sum_{i=1}^{n_2} y_{2i} \cdot \bar{y}_2 + \sum_{i=1}^{n_3} y_{3i} \cdot \bar{y}_3 \right)$ $\left(\because \sum_{j=1}^{n_i} y_{ij} = n_i \bar{y}_i,\ i = 1,2,3 \right)$

$= \sum_{i=1}^{n_1} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{3} n_i \bar{y}_i^2$ //

4/4

2561551

P2

2 (a) ① good predictor for FSIQ is MRI_Count because there is a strong positive relationship between FSIQ and MRI_Count which is statically significant.

② There is a weak negative relationship between FSIQ and Height, since they are not biologically related in common sense, it is not wierd.

③ There is a weak negative relationship between FSIQ and weight, which is not wierd in common sense.

④ It seems that FSIQ differs in male and female, because male's average FSIQ seems to be higher than female's FSIQ.

3/4

(b) $E(Y) = X\beta$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & \text{Height}_1 & \text{MRI\_Count}_1 \\ \vdots & \vdots & \vdots \\ 1 & \text{Height}_i & \text{MRI\_Count}_i \end{pmatrix}$$

4/2

(c) For Residuals against fitted values plot, as fitted value increase, the amount of variability in residuals decreases, so we can't assume the residuals have constant variance, but has has mean of zero because the blue line fluctuates around 0, there are outliers observation 14, 38.

② Normal Q-Q plot since most of the points are on the dash line, we may assume that the residuals are normally distributed, there are outliers observation 14, 38.

③ For scale location plot, the points are scattering around the blue line, showing that residuals are equally spread around the range of predictors, but there are outliers such as observation 14, 38.

④ For residuals vs Leverage plot, there are some outliers, such as observation 14, 29, 28, which have large residuals, removing the outliers may help improve the model.

3.5/5

(a) For Residuals against fitted value plot, assumption to be checked is constant variance with mean zero.

(b) For Normal Q-Q plot, the assumption to be checked is residuals are normally distributed.

(c) For scale location plot, the assumption is that residuals are equally-spread around constant variance.

(d) For Scale location plot, we are checking outliers which has large residual.

2561551

(d)

95% CI for $b^T\beta$: $b^T\hat{\beta} \pm t(n-p; 0.975)\sqrt{\frac{RSS}{n-p}\left(b^T(X^TX)^{-1}b\right)}$

$b^T = (0\ 1\ 0)$ $\hat{\beta} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$ $b^T\hat{\beta} = \beta$

95% CI for β is:

$$-2.8244 \times 10^0 \pm 2.0262 \times \sqrt{\frac{15805.3}{37} \times 2.585101 \times 10^{-3}}$$

$$(-4.9532, -0.6948)$$

The interval does not include 0 and therefore Height should be retained in model.

Height is a significant predictor of FS2Q, the coefficient is highly likely to lie between -4.9532 and -0.6948

95% CI for γ is: $b^T = (0\ 0\ 1)$ $\hat{\beta} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$ $b^T\hat{\beta} = \gamma$

$$2.096 \times 10^{-4} \pm 2.0262 \times \sqrt{\frac{15805.3}{37} \times 7.703296 \times 10^{-12}}$$

$$(9.337 \times 10^{-5}, 3.2583 \times 10^{-4})$$

The interval does not include 0, and therefore NR2_cont should be retained in model. NR2_cont is a significant predictor of FS2Q. The coefficient is highly likely to lie between 9.337×10^{-5} to 3.2583×10^{-4}

6/6

(e) According to part d, both Height & and NR2-count are significant predictors of FS2Q, the coefficient of Height is highly likely to be between -4.9532 and -0.6948, the coefficient of NR2-cont is highly likely to lie between 9.337×10^{-5} to 3.2583×10^{-4}.

0/4

2. (f) 95% PI: $b^T\hat{\beta} \pm t(n-p; 0.975)\sqrt{\frac{RSS}{n-p}(1 + b^T(X^TX)^{-1}b)}$

$b^T = (1 \quad 70 \quad 860000)$ $\hat{\beta} = (1.172 \times 10^2, -2.824, 2.096 \times 10^{-4})^T$

$$\sqrt{\frac{RSS}{n-p}(1+b^T(X^TX)^{-1}b)} = \sqrt{\frac{1505.3}{37}\left[1+(1 \quad 70 \quad 860000)(X^TX^{-1})\begin{pmatrix}1\\70\\860000\end{pmatrix}\right]}$$

$$= 21.3014$$

$b^T\beta \pm t(n-p, 0.975)\sqrt{\frac{RSS}{n-p}(1+b^T(X^TX)^{-1}b)}$

$= 99.776 \pm 2.0262 \times 21.3014$

$= (56.616, 142.936)$

Interpretation: The FSZQ for a future person with height 70 inches and MRI count equal to 860000 is highly likely to lie between 56.616 to 142.936

4/4

256155

3. (a) There seems to be a moderate positive relationship between study time and final grade for both male and female. Also, it seems that the relationship differ by gender, i.e. male's slope is slightly larger than female's slope.

2/2

(b)

According to part A, my model is as follows.

Model: $E(Y_{ij}) = \alpha_i + \beta_i x_{ij}$, $i=1,2$, $j=1 \cdots n_i$, where y_{ij} is the final grade of person j of gender i ($i=1$, female, $i=2$, male)

x_{ij} the study time of person j of gender i.

β_i: β_1 is slope of female, β_2 is slope of male

α_i: α_1 is slope of female, α_2 is slope of male

3/4

(c)

Model 1: model 1 describes a different relationship between final grade and study time depending on gender. For each gender, there is two regression lines with different slope and different intercept terms.

model 2 describe systematic difference between final grade and study time depending on gender. For each gender, there are two regression lines with same slope, and different intercept terms.

Model 3 For each gender, there are one regression line with same slope and same intercept terms.

model 4 For each gender, there is only and intercept term. i.e. Final grade is not related to study time in this model.

4/4

25 of 54

3.(d)

For model 1:

E(final grade | male) = 29.0337 + 0.3017 + 7.656 × studytime + 0.1605 × study time

⇒ E(final grade | male) = 29.0337 + 0.3017 + (7.656 + 0.1605) × 5

= 68.4178

E(final grade | female) = 29.0337 + studytime × 7.656

= 29.0337 + 5 × 7.656

= 67.3137

2/2

2) the expected final grade for a male student spend average of 5 hour study per week is 68.4178, for female studying 5 hours per week is 67.3137.

(e) For model 3: multiple R-squared = 0.7274 = R^2

2/2

⇒ $\hat{r} = \sqrt{R^2} = 0.853$

(f) By table, $n = 100$, $\alpha = 0.05$, $r_{n,\alpha} = 0.1966$

Since $\hat{r} = 0.853 > 0.1966$

⇒ we reject the null hypothesis $H_0: \rho = 0$ and conclude that there is a statically significant linear relationship between final grade and study time. //

3/3

25/155

3. (9)

For lowest AIC criterion, priority: 3 > 2 > 1 > 4 (the lower AIC the better)

For lowest BIC criterion, model choose priority: 3 > 2 > 1 > 4 (the lower BIC the better)

For R^2 adj criterion, the higher R^2 adj the better: model priority: 2 > 3 > 1 > 4

R^2 is not a good criterion compared with R^2 adj because it increase when we add more explanatory varichles.

Overall, I would choose model 3 because it minimize AIC and BIC, also it has a high R^2 (adj) and R^2. //

3/3