2023

Question 1

Part 1

a)

```
plot_residuals <- function(model) {
    par(mfrow=c(2, 2))
    plot(model)
    par(mfrow=c(1, 1))
}

raw_model <- lm(mpg~disp)
raw_model_summary <- summary(raw_model)
log_model <- lm(log.mpg~log.disp)
log_model_summary <- summary(log_model)

plot_residuals(raw_model)
plot_residuals(log_model)

"
The raw model is better from an R^2 perspective, as well as a model
assumptions perspective. The log model suffers from significant
shape in the Residuals vs Fitted plot which indicates a
violation of the linearity assumption.

The log model also significantly deviates from the normal QQ line,
violating the normality assumption.
"
```

Part 2

b)

```
log_model_summary$fstatistic

# H_0: B_{log.disp} = 0
# H_1: B_{log.disp} != 0

# F statistic 31.64008
# Null distribution: F(1, 27)
# p-value: 5.715 * 10^{-6}

"
With a 1% level of significance, we can reject the null hypothesis.
This means that the model with log.disp is better than the
intercept only model.
"
```

c)

```
log_model_summary$coefficients["log.disp", "Estimate"]
# -0.241509
```

d)

```
confint(log_model, level=0.99)
# (-0.3604691, -0.12225489)
```

e)

```
predict(log_model, data.frame(log.disp=log(240)), interval="confidence", level=0.99)
# (2.769225, 3.017715)
```

Part 3

a)

```
carsdat$dummy <- c(rep(0, 5), 1, rep(0, 23))
part_three_model <- lm(log.mpg~log.disp+dummy, data=carsdat)
# a)
summary(part_three_model)
```

b)

```
"
With a t statistic of 0.972, a t distribution of t(26) and
a p-value of 0.34, we cannot reject the null hypothesis, as
0.34 > 0.05.
"
```

Question 2

Part 1

a)

```
# 1 predictor: cyl
# 2 predictors: wt + cyl
# 3 predictors: wt + cyl + disp
```

b)

```
subset_summary$adjr2
# R^2: 0.5996692, 0.6091114, 0.6976197
PRESS <- function(model) {
    p_res <- residuals(model) / (1 - hatvalues(model))
    return(sum(p_res^2))
}
PRESS(lm(mpg~cyl))
PRESS(lm(mpg~wt+cyl))
PRESS(lm(mpg~wt+cyl+disp))

"
The best model is the 3 predictor model, with the highest
adjusted R^2; which indicates that the regression
component explains more of the total variation, and
the lowest PRESS, which indicates the best out of sample
prediction.
"
```

c)

```
"
PRESS is important to consider the out-of-sample prediction.
Using residuals to consider model prediction is a poor
test as the model is fit for the very sample. PRESS residuals
allow for the testing of out-of-sample prediction, and thus
can show relative predictive performance.
"
```

Part 2

d)

```
# Hypotheses:
# H_0 : wt = cyl = disp = hp = 0
# H_1 : Not all betas are 0

# F-statistic: 20.51267
# Null distribution: F(4, 27)
# p-value: 7.294 * 10^{-8}

"
With a p-value < 0.05, we can reject the null hypothesis,
and conclude that the full model is better than the
intercept only model.
"
```

e)

```
e_test <- anova(lm(mpg~wt), full_model)

# Hypotheses:
# H_0: cyl=disp=hp=0
# H_1: Not all betas are 0

# F statistic: 9.6683
# Null distribution: F(3, 27)
# p-value = 0.0001668

"
With a 5% level of significance, we have enough evidence
to reject the null hypothesis. This means that the additional
predictors do have a statistically significant positive
influence on the model.
"
```

f)

```
predict(full_model, data.frame(wt=3.6, cyl=6, disp=220, hp=160),
        interval="prediction", level=0.98)
# (9.123288, 26.53452)
```

QUESTION 3.

$$Y_i = \beta x_i + \varepsilon_i$$

a)

$$S_w(\beta) = \sum_{i=1}^{n} r_i (y_i - \beta x_i)^2$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} r_i x_i (y_i - \beta x_i)$$

$$\sum_{i=1}^{n} r_i x_i (y_i - b_w x_i) = 0.$$

$$= \sum_{i=1}^{n} r_i x_i y_i - b_w \sum_{i=1}^{n} r_i x_i^2 = 0.$$

$$b_w \sum_{i=1}^{n} r_i x_i^2 = \sum_{i=1}^{n} r_i x_i y_i$$

$$b_w = \frac{\sum_{i=1}^{n} r_i x_i y_i}{\sum_{i=1}^{n} r_i x_i^2}$$

b) $$E(b_w) = \frac{\sum_{i=1}^{n} r_i x_i E(y_i)}{\sum_{i=1}^{n} r_i x_i^2}$$

$$= \frac{\sum_{i=1}^{n} r_i x_i \times \beta x_i}{\sum_{i=1}^{n} r_i x_i^2} = \frac{\beta \cancel{\sum_{i=1}^{n} x_i^2 r_i}}{\cancel{\sum_{i=1}^{n} r_i x_i^2}}$$

$$= \beta.$$

$$Var(b_w) = Var\left(\frac{\sum_{i=1}^{n} r_i x_i y_i}{\sum_{i=1}^{n} r_i x_i^2}\right) = \frac{1}{\left(\sum_{i=1}^{n} r_i x_i^2\right)^2} \sum_{i=1}^{n} r_i^2 x_i^2 Var(y_i)$$

$$= \frac{1}{\left(\sum_{i=1}^{n} r_i x_i^2\right)^2} \sum_{i=1}^{n} r_i^2 x_i^2 \times \frac{\sigma^2}{r_i}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} r_i x_i^2}$$

c)

$$\varepsilon_i \sim N\left(0, \frac{\sigma^2}{r_i}\right)$$

$\therefore$ y is also normally distributed

$$Y_i \sim N\left(\beta x_i, \frac{\sigma^2}{r_i}\right)$$

$$f_Y(y_i) = \frac{\sqrt{r_i}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right)$$

$$L(\beta, \sigma^2; y) = \frac{\sqrt{\prod r_i}}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{\cancel{n}}{2\sigma^2} \sum_{i=1}^{n} r_i (y_i - \beta x_i)^2\right)$$

$$l(\beta, \sigma^2; y) = \ln\left(\sqrt{\prod r_i}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} r_i (y_i - \beta x_i)^2 - \frac{n}{2}\ln(2\pi\sigma^2).$$

$$\frac{\partial l(\beta, \sigma^2; y)}{\partial \beta} = -\frac{1}{2\sigma^2} \frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n} r_i (y_i - \beta x_i)^2\right).$$

This is the same as least squares.

d)

$$\sqrt{r}\,\mathrm{Cov}(\sqrt{r_i}\, e_i, \sqrt{r_j}\, e_j) = \sqrt{r_i r_j}\,\mathrm{Cov}(e_i, e_j)$$

$=$ e_i & e_j are given as uncorrelated, $\therefore$

$\mathrm{Cov}(e_i, e_j) = 0$. $\therefore = 0$.

ii) $E(\sqrt{r_i}\, e_i) = \sqrt{r_i}\, E(e_i)$

$= \sqrt{r_i} \times 0$

$= 0.$

iii) $\mathrm{Var}(\sqrt{r_i}\, e_i) = r_i \mathrm{Var}(e_i)$

$= r_i \times \frac{\sigma^2}{r_i}$

$= \sigma^2.$

e)

$$S(\beta) = \sum_{i=1}^{n} (\sqrt{r_i}\, y_i - \sqrt{r_i}\, \beta x_i)^2 = \sum_{i=1}^{n} r_i (y_i - \beta x_i)^2$$

= same objective function.

f) $r_i = \frac{1}{x_i^2}$.

$\therefore$ Consider:

$$Var(b_w) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 \times \frac{1}{x_i^2}}$$

$$= \frac{\sigma^2}{n} \quad \therefore \text{CONSTANT.}$$

If the variance ~~is~~ has the relationship $Var(b_w) = x_i^2 \sigma^2$, the transformation seen in model (2) stabilises the variance.

QUESTION 4.

PART 1.

a) $b - b_{-i}$ is the vector of differences between the fitted β parameters with all observations - without i-th observation. Cook's Distance attempts to consider the total magnitude of distance change induced by the removal of the i-th observation.

b) First, recall that:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

$$D_i = \frac{\left(\frac{(X^TX)^{-1} x_i e_i}{1-h_{ii}}\right)^T (X^TX)^{-1} \left(\frac{(X^TX)^{-1} x_i e_i}{1-h_{ii}}\right)}{p\,\hat{\sigma}^2}.$$

$$= \frac{e_i^2 x_i^T (X^TX)^{-1}(X^TX)(X^TX)^{-1} x_i}{\hat{\sigma}^2 p (1-h_{ii})^2}$$

$$= \frac{e_i^2 x_i^T (X^TX)^{-1} x_i}{\hat{\sigma}^2 p (1-h_{ii})^2}$$

$$= \left(\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\right)^2 \times \frac{h_{ii}}{p(1-h_{ii})}$$

$$= \frac{r_i^2}{10} \times \frac{h_{ii}}{1-h_{ii}}$$

PART 2.

c) $\hat{Y}_{i,-i}$ is the value fitted at the i-th observation with parameters fitted WITHOUT the i-th observation

d) We assume $\hat{y}_i$ is a linear combination of the predictors and their betas. Thereby, define $x_i = (x_{i1}, x_{i2} \dots x_{in})^T$. We can then define:

$$\hat{Y}_i = x_i^T b \quad \& \quad \hat{Y}_{i,-i} = x_i^T b_{-i}$$

where b_{-i} is fitted without the i-th observation

e) $$\hat{Y}_i - \hat{Y}_{i,-i} = x_i^T b - x_i^T b_{-i}$$

$$= x_i^T b - x_i^T \left(b - \frac{(X^TX)^{-1} x_i e_i}{1-h_{ii}}\right)$$

$$= \frac{x_i^T (X^TX)^{-1} x_i e_i}{1-h_{ii}}$$

$x_i^T (X^TX)^{-1} x_i = h_{ii}$, as $H = X(X^TX)^{-1}X^T$.

$$= \frac{h_{ii} e_i}{1-h_{ii}}$$

f)

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i,-i}}{\hat{\sigma}_{-i}\sqrt{h_{ii}}} \qquad t_i = \frac{e_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}}$$

$$= \frac{\frac{h_{ii}e_i}{1-h_{ii}}}{\hat{\sigma}_{-i}\sqrt{h_{ii}}}$$

$$= \frac{e_i\sqrt{h_{ii}}}{\hat{\sigma}_{-i}(1-h_{ii})}$$

$$= \frac{e_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}} \times \frac{\sqrt{h_{ii}}}{\sqrt{1-h_{ii}}}$$

$$= t_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$