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)
```