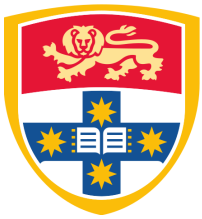

THE UNIVERSITY OF SYDNEY

STAT3022 Applied Linear Models

Assignment 1

Lecturer: Linh Nghiem

Author: Mason Wong

School of Mathematics and Statistics

The University of Sydney

Semester 1, 2022

Question 1

1. The fitted regression equation for the model of the average total SAT score on all other covariates in the data set is:

$$\hat{total} = 1045.97 - 3.62(ratio) + 1.64(salary) + 4.46(expend) - 2.90(perc)$$

2. No they are not contradictory. The reason is because they are testing completely different models!

   - For the t-test for the covariate salary, we see that the test being conducted here is:

$$H_0 : total \sim ratio + expend + perc \quad \text{vs} \quad H_1 : total \sim ratio + salary + expend + perc$$

   - Whereas for the F-test for the covariate salary, we see that the test being conducted here is:

$$H_0 : total \sim ratio \quad \text{vs} \quad H_1 : total \sim ratio + salary$$

3. If we observe the normal quantile-quantile plot see that the standardized residuals are roughly linear so the assumption of normality of the residuals seems to be satisfied.

   Moreover, if we look at the plot of the residuals vs fitted, it seems to be the case that the variance is roughly equal and that there is no obvious pattern. Hence the constant variance assumption seems to be satisfied too.

4. Looking at both the outputs for the cooks distance and the dffit, we see that the states which seem to be influential observations are:

   - Utah because it has high leverage ($\overline{h} = \frac{p}{n} = \frac{5}{50} \implies 2\overline{h} = 0.2$ and the leverage of Utah is 0.29) and it is an outlier (as it is the second when we arrange the data in descending order by magnitude of externally studentized residuals).
   - West Virginia as it is an outlier (being the first when we arrange the data in descending order by magnitude of externally studentized residuals).
   - North Dakota (as it is the third when we arrange the data in descending order by magnitude of externally studentized residuals).
   - New Hampshire (as it is the fourth when we arrange the data in descending order by magnitude of externally studentized residuals).

5. Given that all the covariates are in the model, we see that the correlation between 'salary' and 'ratio' is very small (-0.001). Though coefficients are hard to interpret in the case of multicollinearity, we can see that the correlation between 'ratio' and 'salary' is close to zero, meaning that they are uncorrelated predictors. Hence they will have different effects on the response variable.

6. We calculate the variance inflation factors (VIF's) for the following four covariates:

   - $VIF_{\text{ratio}} = \frac{1}{1-0.589} = 2.433$
   - $VIF_{\text{salary}} = \frac{1}{1-0.892} = 9.26 > 5$
   - $VIF_{\text{expend}} = \frac{1}{1-0.894} = 9.43 > 5$
   - $VIF_{\text{perc}} = \frac{1}{1-0.0.430} = 1.75$

7. We see that the covariates 'salary' and 'expend' have VIF greater than 5. Based on the rule, we see that 'modelA' has serious collinearity. We can also see this through the pairwise correlation plot as as the correlation between 'salary' and 'expend' is 0.870

Question 2

```
library(dplyr)
library(ggplot2)
```

```
mtcars = mtcars %>%
    mutate(vs = factor(vs), am = factor(am))
```

(i) Additive model only

```
additive_model = lm(mpg ~ wt + qsec, data = mtcars)
summary(additive_model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3962 -2.1431 -0.2129  1.4915  5.7486
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.7462     5.2521   3.760 0.000765 ***
## wt           -5.0480     0.4840 -10.430 2.52e-11 ***
## qsec          0.9292     0.2650   3.506 0.001500 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.596 on 29 degrees of freedom
## Multiple R-squared:  0.8264, Adjusted R-squared:  0.8144
## F-statistic: 69.03 on 2 and 29 DF,  p-value: 9.395e-12
```

```
anova(additive_model)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## wt         1 847.73  847.73 125.773 4.603e-12 ***
## qsec       1  82.86   82.86  12.293    0.0015 **
## Residuals 29 195.46    6.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Given `wt` is in the model, does `qsec` depend on `vs`?

```
qsec_dep_vs = lm(mpg ~ wt + qsec * vs, data = mtcars)
anova(qsec_dep_vs)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## wt         1 847.73  847.73 125.2087 1.208e-11 ***
## qsec       1  82.86   82.86  12.2381  0.001641 **
## vs         1   0.07    0.07   0.0110  0.917062
## qsec:vs    1  12.59   12.59   1.8589  0.184017
## Residuals 27 182.80    6.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems like `qsec` doesn't depend on `vs` if we look at the p values for the interaction term. Moreover, It doesn't look like

we should include the vs term at the 5% level of significance

- Given wt is in the model, does qsec depend on am?

```
qsec_dep_am = lm(mpg ~ wt + qsec * am, data = mtcars)
anova(qsec_dep_am)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## wt         1 847.73  847.73 158.9362 7.958e-13 ***
## qsec       1  82.86   82.86  15.5347 0.0005167 ***
## am         1  26.18   26.18   4.9079 0.0353554 *
## qsec:am    1  25.27   25.27   4.7387 0.0384079 *
## Residuals 27 144.01    5.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems like qsec does depend on am and that we should include the interaction term at the 5% level of significance.

- Given wt is in the model, does qsec depend on both vs and am?

```
qsec_dep_both = lm(mpg ~ wt + qsec * am * vs, data = mtcars)
anova(qsec_dep_both)
```

```
## Analysis of Variance Table
##
## Response: mpg
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## wt          1 847.73  847.73 137.6213 3.469e-11 ***
## qsec        1  82.86   82.86  13.4514  0.001279 **
## am          1  26.18   26.18   4.2497  0.050735 .
## vs          1   0.00    0.00   0.0001  0.993137
## qsec:am     1  25.28   25.28   4.1038  0.054540 .
## qsec:vs     1   1.47    1.47   0.2391  0.629500
## am:vs       1   0.00    0.00   0.0001  0.991153
## qsec:am:vs  1   0.86    0.86   0.1391  0.712622
## Residuals  23 141.68    6.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that am and qsec:am are bordering on the 5% level of significance. Given the previous result, all this leads us to conclude that qsec depends on am

```
# check the sum of squared residuals
SSE_1 = sum(qsec_dep_vs$residuals^2)
SSE_2 = sum(qsec_dep_am$residuals^2)
SSE_3 = sum(qsec_dep_both$residuals^2)
c(SSE_1, SSE_2, SSE_3)
```

```
## [1] 182.8034 144.0111 141.6763
```

We see that the sum of squared residuals decreases drastically when we change from vs as being the binary predictor to am being the binary predictor. Moreover, when include both vs and am the decrease in the sum of squared residuals is negligible. Hence, even though the model with both vs and am has the lowest SSE, we have sufficient evidence to not include the coefficient of vs and favour the model with only am and it's interaction term with qsec assuming wt is in the model.

Question 3

Suppose that we have M new observations at x_0^*. That is we have:

$$\begin{cases} y_{01}^* = \beta_0 + \beta_1 x_0^* + \varepsilon_{01} \\ y_{02}^* = \beta_0 + \beta_1 x_0^* + \varepsilon_{02} \\ \vdots \\ y_{0M}^* = \beta_0 + \beta_1 x_0^* + \varepsilon_{0M} \end{cases}$$

Where y_{0i}^* is a random variable for $i = 1, 2, ..., M$ with some random error ε_{0i} for $i = 1, 2, ..., M$

We also denote $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}$ as the least square estimates of β_0, β_1 and σ as given in the lecture. Hence we would have the facts that

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} & -\frac{\overline{x}}{S_{xx}} \\ -\frac{\overline{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{bmatrix} \right)$$

And

$$\hat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

Define the random variable

$$y_0^* = \frac{1}{M} \left[y_{01}^* + y_{02}^* + ... + y_{0M}^* \right]$$

Note: that according to the question, our variable y_0^* corresponds to $\overline{y}_0^*$ and we have omitted the bar for convenience.

We simplify so that we see:

$$\begin{aligned} y_0^* &= \frac{1}{M} \left[(\beta_0 + \beta_1 x_0^* + \varepsilon_{01}) + ... + (\beta_0 + \beta_1 x_0^* + \varepsilon_{0M}) \right] \\ &= \frac{1}{M} \left[M\beta_0 + M\beta_1 x_0^* + (\varepsilon_{01} + ... + \varepsilon_{0M}) \right] \\ &= \beta_0 + \beta_1 x_0^* + \frac{1}{M} (\varepsilon_{01} + ... + \varepsilon_{0M}) \end{aligned}$$

We partition our thinking into 4 steps:

Step 1: Find $\mathbb{E}[y_0^*]$ and $\text{Var}[y_0^*]$

- Since $\mathbb{E}[\varepsilon_{0i}] = 0$ for $i = 1, ..., M$ we have that:

$$\begin{aligned} \mathbb{E}[y_0^*] &= \mathbb{E}[\beta_0 + \beta_1 x_0^* + \frac{1}{M} (\varepsilon_{01} + ...\varepsilon_{0M})] \\ &= \beta_0 + \beta_1 x_0^* + \mathbb{E}[\frac{1}{M} (\varepsilon_{01} + ...\varepsilon_{0M})] \\ &= \beta_0 + \beta_1 x_0^* + \frac{1}{M} \left[\mathbb{E}[\varepsilon_{01}] + ... + \mathbb{E}[\varepsilon_{0M}] \right] \\ &= \beta_0 + \beta_1 x_0^* + \frac{1}{M} \left[\underbrace{0 + ... + 0}_{M \ times} \right] \\ &= \beta_0 + \beta_1 x_0^* \end{aligned}$$

- Since $\text{Cov}(\varepsilon_{0i}, \varepsilon_{0j}) = 0$ for $i \neq j$ (by assumption) we have that $\text{Var}[\varepsilon_{0i} + \varepsilon_{0j}] = \text{Var}[\varepsilon_{0i}] + \text{Var}[\varepsilon_{0j}]$ for $i \neq j$. Furthermore we also have that $\text{Var}[\varepsilon_{0i}] = \sigma^2$ for all $i = 1, ..., M$ By assumption. Hence:

$$\begin{aligned}
\text{Var}[y_0^*] &= \text{Var}[\beta_0 + \beta_1 x_0^* + \frac{1}{M}(\varepsilon_{01} + ... + \varepsilon_{0M})] \\
&= \text{Var}[\frac{1}{M}(\varepsilon_{01} + ... + \varepsilon_{0M})] \\
&= \frac{1}{M^2}\text{Var}[\varepsilon_{01} + ... + \varepsilon_{0M}] \\
&= \frac{1}{M^2}\left[\underbrace{\sigma^2 + ... + \sigma^2}_{m \text{ times}}\right] \\
&= \frac{\sigma^2}{M}
\end{aligned}$$

Step 2: Find the point estimate $\hat{y_0}$

The point estimate $\hat{y_0}$ is obtained by subbing in $\hat{\beta}_0$ and $\hat{\beta}_1$ into y_0^* and removing the error terms (as they zero mean). Hence our point estimate is:

$$\hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0^*$$

Step 3: Find $\mathbb{E}[\hat{y_0}]$ and $\text{Var}[\hat{y_0}]$

- The expectation is obtained by the following

$$\begin{aligned}
\mathbb{E}[\hat{y_0}] &= \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_0^*] \\
&= \mathbb{E}[\hat{\beta}_0] + \mathbb{E}[\hat{\beta}_1] x_0^* \\
&= \beta_0 + \beta_1 x_0^*
\end{aligned}$$

- The variance is obtained by the following:

$$\begin{aligned}
\text{Var}[\hat{y_0}] &= \text{Var}[\hat{\beta}_0 + \hat{\beta}_1 x_0^*] \\
&= \text{Var}[\hat{\beta}_0] + \text{Var}[\hat{\beta}_1 x_0^*] + 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_1 x_0^*) \\
&= \text{Var}[\hat{\beta}_0] + {x_0^*}^2 \text{Var}[\hat{\beta}_1] + 2x_0^* \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} + \frac{{x_0^*}^2}{S_{xx}} - \frac{2x_0^*\overline{x}}{S_{xx}}\right) \\
&= \sigma^2\left(\frac{1}{n} + \frac{(x_0^* - \overline{x})^2}{S_{xx}}\right)
\end{aligned}$$

Step 4: Find the distribution of $y_0^* - \hat{y_0}$

- We obtain the expectation

$$\begin{aligned}
\mathbb{E}[y_0^* - \hat{y_0}] &= \mathbb{E}[y_0^*] - \mathbb{E}[\hat{y_0}] \\
&= 0
\end{aligned}$$

- To find the variance we note that y_0^* and $\hat{y_0}$ are independent due to x_0^* being a new observation. Hence:

$$\begin{aligned}
\text{Var}[y_0^* - \hat{y_0}] &= \text{Var}[y_0^*] + \text{Var}[\hat{y_0}] \\
&= \frac{\sigma^2}{M} + \sigma^2\left(\frac{1}{n} + \frac{(x_0^* - \overline{x})^2}{S_{xx}}\right) \\
&= \sigma^2\left(\frac{1}{M} + \frac{1}{n} + \frac{(x_0^* - \overline{x})^2}{S_{xx}}\right)
\end{aligned}$$

Replacing σ by $\hat{\sigma} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$ we obtain that the standard error is:

$$SE\left(y_0^* - \hat{y_0}\right) = \hat{\sigma}\sqrt{\frac{1}{M} + \frac{1}{n} + \frac{\left(x_0^* - \overline{x}\right)^2}{S_{xx}}}$$

Hence:

$$T = \frac{y_0^* - \hat{y_0} - 0}{SE\left(y_0^* - \hat{y_0}\right)} \sim t_{n-2}$$

Where the degrees of freedom of the student t test come from the $\hat{\sigma}$. Hence we have that a $100(1-\alpha)\%$ prediction interval for y_0^* is:

$$\hat{y_0} \pm t_{n-2,1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{M} + \frac{1}{n} + \frac{\left(x_0^* - \overline{x}\right)^2}{S_{xx}}}$$

Where $t_{n-2,1-\alpha/2}$ is the quantile can be obtained by `qt(p=1-alpha/2, df=n-2, lower.tail=TRUE)` in `R`. Note that if we let $M \to \infty$ then we obtain the confidence interval for the average response (instead of a particular observation) which makes intuitive sense as we are taking an average of the response variables!