Fall 2022 STAT 706 HW 4

Reina Li

Chapter 4: 4.2.1, 4.2.3; Chapter 6: 6.7.4

4.2.1

```
# Create data frame-----
df1 <- data.frame(yearsExperience = c(0, 2, 4, 6, 8, 12, 17, 22, 28, 34),
                  n = c(17, 33, 19, 25, 18, 60, 58, 31, 34, 19),
                  thirdQuantile = c(101300, 111303, 98000, 124000, 128475, 117410,
                                    115825, 134300, 128066, 164700))

# Fit WLS model-----
model1b_weights <- 1 / df1$n
model1b <- lm(thirdQuantile ~ yearsExperience,
              weights = model1b_weights,
              data = df1)
summary(model1b)$coefficient
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     103925.587  5232.1030 19.863062 4.299901e-08
## yearsExperience   1517.902   305.8019  4.963678 1.101801e-03
```

```
# Estimate third quantile for salary for full professors with 6 years of experience
# thirdQuantile = 103925.587 + 1517.902*yearsExperience
yearsExperience <- 6
103925.587 + 1517.902 * yearsExperience
```

```
## [1] 113033
```

The weighted least squares model is $ThirdQuantile = 1.0392559 \times 10^5 + 1517.9022738 * YearsExperience$. Using weighted least squares, the estimate of the 2005-2006 third quantile for salary of full professors with 6 years of experience is $\$1.13033 \times 10^5$.

4.2.3 (a)

It is necessary to use weighted least squares to fit model (4.6) because the assumption of homoscedasticity (constant variance in the errors) is violated. $w_i = n_i$ is the appropriate choice for the weights because Y_i is the median of n_i observations.

4.2.3 (b)

Model (4.6) is not a valid regression model because:

1. Looking at the standardized residuals plot, there are many outliers, of which many are bad leverage points that drag the least squares line away from the bulk of the points. –> There may be potential problems with the model. Maybe consider fitting another model.
2. Additionally, the standardized residuals plot does not produce points close to a straight line. –> Evidence of non-normality.
3. Looking at the standardized residuals plot and the square root of the absolute value of the standardized residuals plot, the points group together as fitted values increase. The variability in the standardized residuals also tends to increase as the fitted values increase. –> Violated the assumption of homoscedasticity.

4.2.3 (c)

The steps I would take to obtain a valid regression model are:

1. Output the standardized residuals plot, the square root of the absolute value of the standardized residuals plot, and the Q-Q plot.
2. Look at the plots to determine the validity of the model.
3. Figure out which assumptions are being violated.
4. Options: (a) Try fitting another (better) model; (b) Use transformations; (c) Fit the line based on weighted least squares; (d) Select a subset of predictor variables; (e) Regularization

6.7.4 (a)

Model (6.38) is a valid model because looking at the scatter plot matrix first, we can see that some of the plots are linear and some of the plots have a curve. Looking at the residuals vs fitted plot next, we can see a random scatter of points around the horizontal axis. It also shows constant variability.

6.7.4 (b)

Since the plots of standardized residuals against RA and VTINV produce curved patterns, that means that there is a non-linear relationship between RA and VTINV. So based on the residual plots, we cannot say anything about what part of the model is misspecified.

6.7.4 (c)

The correlation coefficient measures the strength and direction of a linear relationship between two variables (-1 to +1). However, using the correlation coefficient to choose between competing regression models only works for linear regression models and will not be good for non-linear regression models. So I think that the correlation coefficient should not be used to choose between competing regression models.

The standard deviation measures how the data spreads and can be used to approximate the variation in the data and give us information about the normality. But it is not enough to help us choose between competing regression models because standard deviation measures how close each observation is to the mean and thus tell us how precise the measurements are. Instead, I think that the standard error would be more useful because standard error tells us if the sample data accurately represents the population data and how well the model fits the data.

The F value is a result of the F-test to determine if the all the regression coefficients are equal to zero (versus the null hypothesis that there is at least one regression coefficient that is not equal to zero). So I think that the F value is not so useful when choosing between competing regression models.