Homework-6-Q3

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: April 7 2025

Problem 3.

In a previous homework, you analyzed a random sample of data on infants' birth weights (Y , lbs) and gestation period (X1, weeks). I've uploaded a new version of this data to canvas called birth weight expanded.csv. There are 4 variables in this data set:

- SNAP Score: for Neonatal Acute Physiology, a continuous measure of illness severity. Higher scores indicate more severe illness
- Birthweight: Birth weight in pounds
- Gestation period: gestational age at birth, in weeks
- Sex: whether the infant is male or female

The researcher is interested in understanding whether birth weight and/or gestation period are useful for predicting the SNAP score. The researcher also wants to be sure to control for sex. Conduct regression analyses to answer the researcher's question. Be sure to list and check your assumptions, describe any hypothesis testing you perform (including full and reduced models), and describe any remedial measures used.

- **Y (response)**: SNAP (continuous)
- **X1 (predictor)**: birthweight (lbs) (continuous)
- **X2 (predictor)**: gestation_period (weeks) (continuous)
- **X3 (predictor)**: sex (male/female) (categorical)

```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-6/"
data_file <- file.path(data_path, "birth_weight_expanded.csv")
BIRTHWEIGHT_data <- read.csv(data_file, header = TRUE)

str(BIRTHWEIGHT_data)
```

```
## 'data.frame':    30 obs. of  4 variables:
##  $ SNAP            : num  20.28 9.04 14.84 22.51 21.74 ...
##  $ birthweight     : num  6.75 8 7.5 6.5 7.25 7 5.5 7.5 8 6.75 ...
##  $ gestation_period: num  36 39 38 36 37 37 35 38 39 36 ...
##  $ sex             : chr  "M" "M" "F" "M" ...
```

```
head(BIRTHWEIGHT_data)
```

```
##        SNAP birthweight gestation_period sex
## 1 20.275989        6.75              36   M
## 2  9.044135        8.00              39   M
## 3 14.843212        7.50              38   F
## 4 22.512288        6.50              36   M
## 5 21.738022        7.25              37   M
## 6 21.934176        7.00              37   M
```

```r
# let's make sure that sex is categorical
# sexM = 0 is female (also reference group)
# sexM = 1 is male
BIRTHWEIGHT_data <- BIRTHWEIGHT_data %>%
  mutate(sex = factor(sex))
```

To start, we'll fit the full MLR model: E[SNAP] = ß0 + ß1(birthweight) + ß2(gestation_period) + ß3(sex).

```r
BIRTHWEIGHT_model_full <- lm(SNAP ~ birthweight + gestation_period + sex, data = BIRTHWEIGHT_data)
summary(BIRTHWEIGHT_model_full)
```

```
##
## Call:
## lm(formula = SNAP ~ birthweight + gestation_period + sex, data = BIRTHWEIGHT_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9653 -3.6636  0.3764  3.4426 19.0265
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       60.6024    22.9765   2.638   0.0139 *
## birthweight       -2.7536     2.2478  -1.225   0.2315
## gestation_period  -0.4965     0.7658  -0.648   0.5224
## sexM              -2.6789     2.5239  -1.061   0.2983
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.622 on 26 degrees of freedom
## Multiple R-squared:  0.1968, Adjusted R-squared:  0.1041
## F-statistic: 2.124 on 3 and 26 DF,  p-value: 0.1215
```

Looking at the summary() output, the estimated intercept is 60.6024, therefore when birth weight = 0, gestation_period = 0, and sex = female (reference category), then SNAP score is 60.60. This doesn't hold any clinical meaning as gestation weeks and birth weights can't be 0 to measure SNAP score. ß1 for birthweight and ß2 for gestation period are both negative, indicating that as these predictors decrease, SNAP score (illness severity) increases, this makes clinical sense. ß3 for sexM is -2.68, indicating that males have lower SNAP scores by 2.68 than females typically.

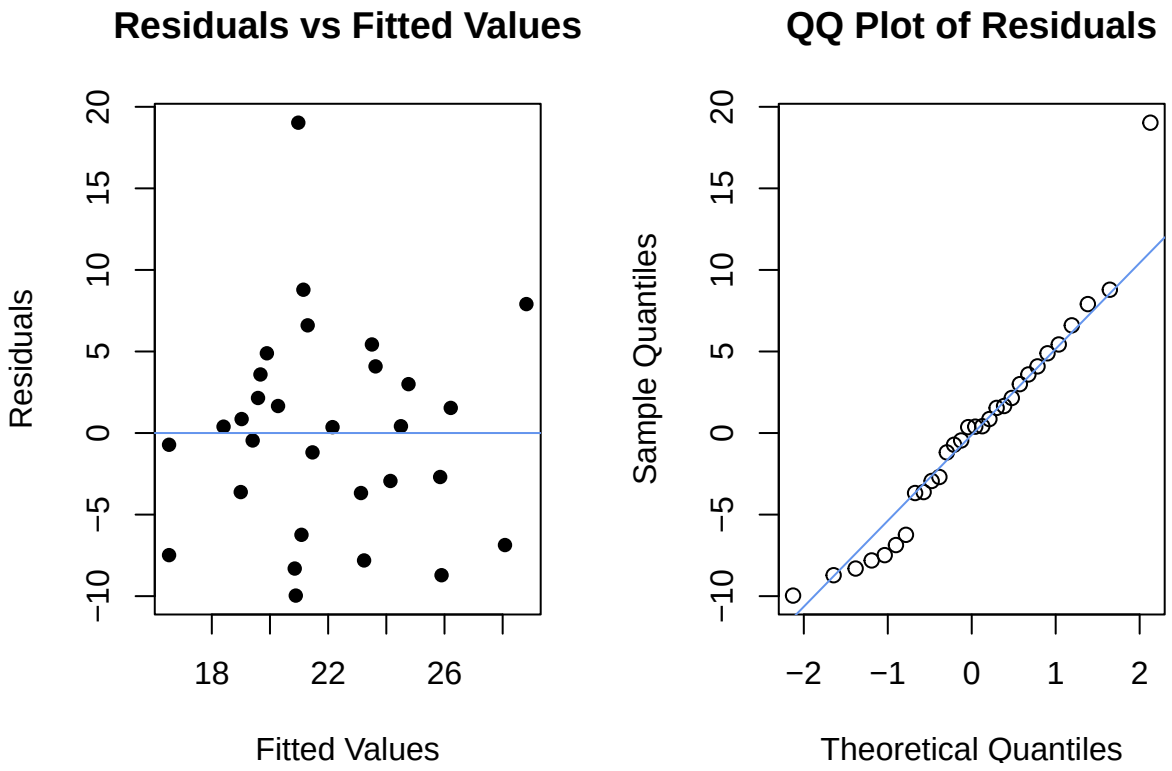
Assumptions for the model:
- relationship between each of the predictors and response is linear
- there is no multicollinearity
- residuals are independent
- residuals have constant variance (homoscedasticity)
- residuals are normally distributed

```r
par(mfrow = c(1, 2))

plot(BIRTHWEIGHT_model_full$fitted.values, residuals(BIRTHWEIGHT_model_full),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values",
     pch = 16)
abline(h = 0, col = "cornflowerblue")

qqnorm(residuals(BIRTHWEIGHT_model_full), main = "QQ Plot of Residuals")
```

```
qqline(residuals(BIRTHWEIGHT_model_full), col = "cornflowerblue")
```



Looking at the residuals vs. fitted plot, there doesn't appear to be a clear fan shape or strong increase in spread as fitted values increase. While there are a few outliers near the edges of the plot, the overall pattern supports the assumption of constant variance. The QQ plot of residuals also supports the assumption of normality, showing only a slight deviation in the left tail and one outlier far from the diagonal line. These plots suggest that the model assumptions of linearity and normality are reasonably met.

Next, we will check for multicollinearity, using the variance inflation factor (VIF) to assess whether the predictors are highly correlated with each other.

```
vif(BIRTHWEIGHT_model_full)
```

```
##      birthweight gestation_period              sex
##         1.526195         1.485359         1.045763
```

The VIF values are: birthweight = 1.52, gestation_period = 1.49, sex = 1.05. The values are quite close to 1, indicating that there is very little to no multicollinearity among the predictors.

Now that we have checked the assumptions, and none are violated we can continue to our hypotheses and testing the predictors.

First, we'll check the full model vs the sex only model.

- H0: ß1 = ß2 = 0 (birthweight and gestation period don't improve the model controlling for sex)
- HA: ß1 or ß2 != 0 (birthweight and gestation period do improve the model controlling for sex)

```
SEX_model_reduced <- lm(SNAP ~ sex, data = BIRTHWEIGHT_data)
anova(SEX_model_reduced, BIRTHWEIGHT_model_full)
```

```
## Analysis of Variance Table
##
## Model 1: SNAP ~ sex
```

```
## Model 2: SNAP ~ birthweight + gestation_period + sex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 1320.4
## 2     26 1140.3  2    180.07 2.0529 0.1487
```

Based on the F-test comparing the full model and a reduced model including only sex.

The F-statistic is 2.0529 at 2 df, and the p value is 0.1487, where $p > 0.05$. Therefore, we fail to reject the null hypothesis. Adding birth weight and gestation both doesn't significantly improve the model compared to using sex alone. Now to test the individual predictors.

To test just birthweight:

- H0: ß1 = 0 (birthweight doesn't improve the model controlling for sex)
- HA: ß1 != 0 (birthweight does improve the model controlling for sex)

```
BIRTHWEIGHT_model_reduced <- lm(SNAP ~ sex + birthweight, data = BIRTHWEIGHT_data)
anova(SEX_model_reduced, BIRTHWEIGHT_model_reduced)
```

```
## Analysis of Variance Table
##
## Model 1: SNAP ~ sex
## Model 2: SNAP ~ sex + birthweight
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 1320.4
## 2     27 1158.7  1    161.63 3.7663 0.0628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is 3.7663 at 1 df, and the p value is 0.0628, where $p > 0.05$. Therefore, we fail to reject the null hypothesis. Adding birth weight doesn't significantly improve the model compared to once sex is accounted for.

To test just gestation period:

- H0: ß2 = 0 (gestation period doesn't improve the model controlling for sex)
- HA: ß2 != 0 (gestation period does improve the model controlling for sex)

```
GESTATION_model_reduced <- lm(SNAP ~ sex + gestation_period, data = BIRTHWEIGHT_data)
anova(SEX_model_reduced, GESTATION_model_reduced)
```

```
## Analysis of Variance Table
##
## Model 1: SNAP ~ sex
## Model 2: SNAP ~ sex + gestation_period
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 1320.4
## 2     27 1206.1  1    114.25 2.5577 0.1214
```

The F-statistic is 2.56 at 1 df, and the p value is 0.1214, where $p > 0.05$. Again, since $p > 0.05$, we fail to reject the null hypothesis. Adding gestation period alone also doesn't significantly improve the model compared to using sex alone.

In conclusion, neither birthweight nor the gestation period significantly improved the prediction of the SNAP score when controlling for sex, based on the F-tests using nested model comparisons. Out of the predictors, gestation period came "closest" to 0.05, but since it did not fall under the threshold, we cannot consider it as a better predictor. Therefore, the reduced model including only sex is the most favourable due to its simplicity and sufficient explanatory power.