Homework-3-Q3

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: March 3 2025

Problem 3.

A student is performing a class project to assess the relationship between daily sodium intake and systolic blood pressure. The project requires them to design a study, collect the data, and analyze the results. The student goes door-to-door in their apartment building, and for families that are willing to participate, the student collects data on sodium intake and systolic blood pressure for each member of the family. The data are available in sodium SBP data.csv. The data set contains the following columns:

- **family id**: Which family each observation was collected from
- **sodium**: Self-report of daily sodium consumption (mg)
- **blood pressure**: Measurement of blood pressure (mm Hg).

```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-3/"
data_file <- file.path(data_path, "sodium_SBP_data.csv")
sodium_data <- read.csv(data_file)

head(sodium_data)
```

```
##   family_id   sodium blood_pressure
## 1         1 2712.510       144.4478
## 2         1 3155.709       144.6644
## 3         1 3118.838       151.1809
## 4         1 2653.726       146.1532
## 5         1 4265.315       167.2042
## 6         1 4465.043       161.7727
```

```
str(sodium_data)
```

```
## 'data.frame':    40 obs. of  3 variables:
##  $ family_id     : int  1 1 1 1 1 1 1 2 2 2 ...
##  $ sodium        : num  2713 3156 3119 2654 4265 ...
##  $ blood_pressure: num  144 145 151 146 167 ...
```

predictor variable (x) -> daily sodium consumption (mg) response variable (y) -> blood pressure measurement (mm Hg)

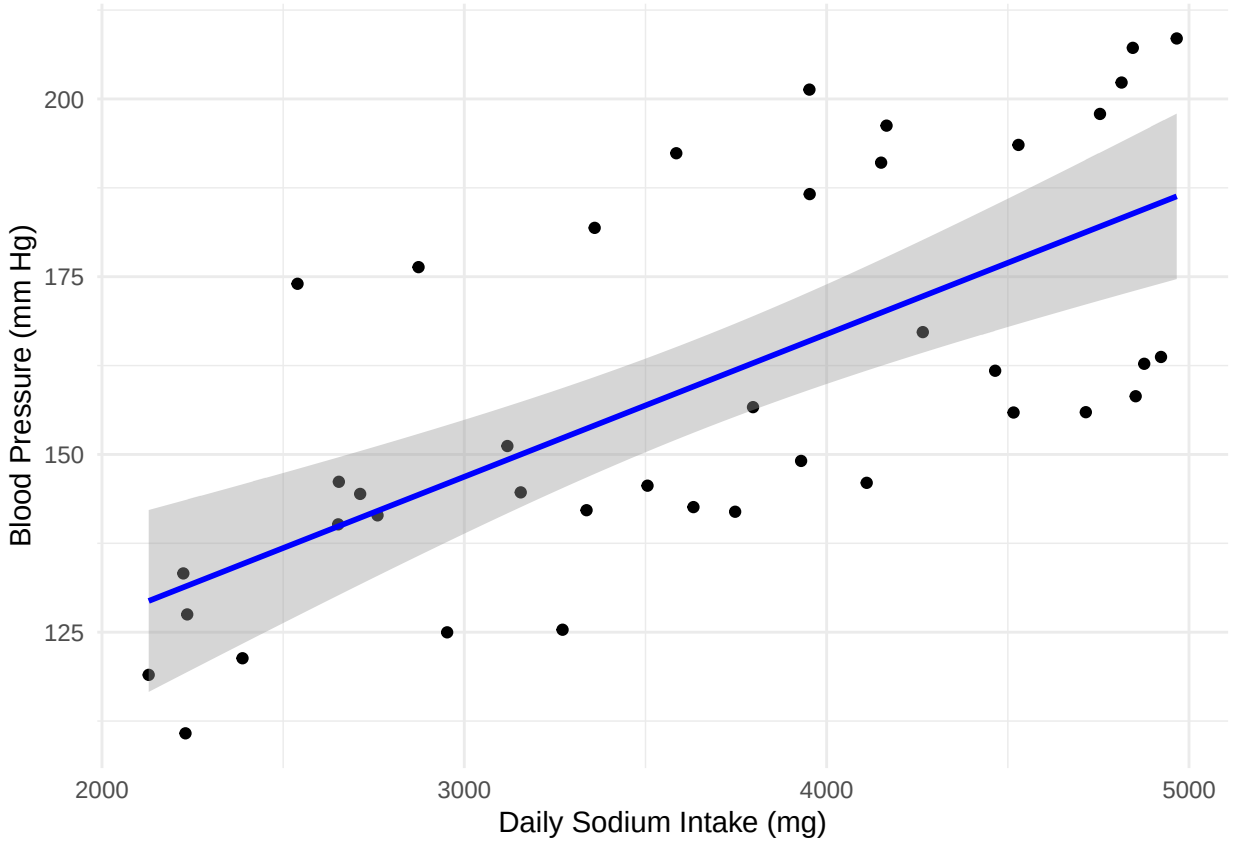
1. Investigate the association between sodium intake and blood pressure using a simple linear regression model. Be sure to carry out all of the usual steps in the analysis.

```
summary(sodium_data)
```

```
##    family_id         sodium     blood_pressure
##  Min.   :1.000   Min.   :2129   Min.   :110.8
##  1st Qu.:2.000   1st Qu.:2845   1st Qu.:142.1
##  Median :4.000   Median :3690   Median :155.9
##  Mean   :3.625   Mean   :3641   Mean   :159.7
##  3rd Qu.:5.000   3rd Qu.:4478   3rd Qu.:183.1
##  Max.   :7.000   Max.   :4966   Max.   :208.5
```

```
sodium_data_scatterplot <- ggplot(
  sodium_data,
  aes(x = sodium, y = blood_pressure)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  theme_minimal() +
  xlab("Daily Sodium Intake (mg)") +
  ylab("Blood Pressure (mm Hg)")
sodium_data_scatterplot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Does a linear relationship appear appropriate? -> Yes

Assumed regression model for Y -> Energy Score = ß0 + ß1(daily sodium intake (mg)) + epsilon

Assumptions about epsilon: - average error should be 0 - errors should follow a normal distribution (bell curve) - there should be no pattern in the errors, we should see homoscedasticity

```
sodium_model <- lm(blood_pressure ~ sodium, data = sodium_data)
summary(sodium_model)
```

```
##
## Call:
## lm(formula = blood_pressure ~ sodium, data = sodium_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.957 -17.695  -4.534  19.453  36.372
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 86.69256   13.49451   6.424 1.49e-07 ***
## sodium       0.02006    0.00360   5.572 2.20e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.31 on 38 degrees of freedom
## Multiple R-squared:  0.4497, Adjusted R-squared:  0.4352
## F-statistic: 31.05 on 1 and 38 DF,  p-value: 2.2e-06
```

ß0 (intercept) = 86.69256 (aka Y when daily sodium intake is 0) ß1hat (slope) = 0.02006 (aka for each unit increase in sodium (mg), blood pressure (mm Hg) increases by 0.02006)

Estimated model -> E[blood_pressure] = (86.69256) + (0.02006) x (sodium)

```
anova(sodium_model)
```

```
## Analysis of Variance Table
##
## Response: blood_pressure
##           Df Sum Sq Mean Sq F value  Pr(>F)
## sodium     1  12808 12807.9  31.047 2.2e-06 ***
## Residuals 38  15676   412.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(sodium_model)$r.squared
```

```
## [1] 0.4496534
```

R squared being 0.4496534 means 44.97% of the variation in Blood Pressure (mm Hg) can be explained by daily sodium intake (mg) and it's therefore it is not a great predictor by itself, more than half of the variation would have to be explained by other predictors beyond daily sodium intake (mg).

H0: ß1 = 0 -> daily sodium intake (mg) does not affect Blood Pressure (mm Hg) HA: ß1!= 0 -> daily sodium intake (mg) does affect Blood Pressure (mm Hg) = 0.05

Two tailed test, as we're testing if the slope is different from 0 in either direction, not just one. Therefore, (alpha)/2 = 0.025, and the test statistic is the t-statistic.

Residuals' DF for t test = 38

```
# t-value via manual t-test
ß1hat <- 0.02006
SE_ß1hat <- 0.00360
t = ß1hat/SE_ß1hat
t
```

```
## [1] 5.572222
```

```
summary(sodium_model)$coefficients
```

```
##                Estimate   Std. Error  t value     Pr(>|t|)
## (Intercept) 86.69255758 13.494506381 6.424285 1.493111e-07
## sodium       0.02005751  0.003599684 5.572019 2.199896e-06
```

```
# 5.572 matches our t value from the summary table
```

```r
# critical t-value
qt(0.025, df = 38, lower.tail = TRUE)  # get -ve val
```

```
## [1] -2.024394
```

```r
qt(0.025, df = 38, lower.tail = FALSE)  # get +ve val
```

```
## [1] 2.024394
```

```r
# = ±2.024394
```

To decide on whether to reject the null hypothesis, we need to check if |t| > critical value.

|5.572| > 2.024394, therefore we reject the null hypothesis.

Now, to get our 95% CI:

```r
confint(sodium_model)
```

```
##                   2.5 %       97.5 %
## (Intercept) 59.37435762 114.01075755
## sodium       0.01277033   0.02734469
```

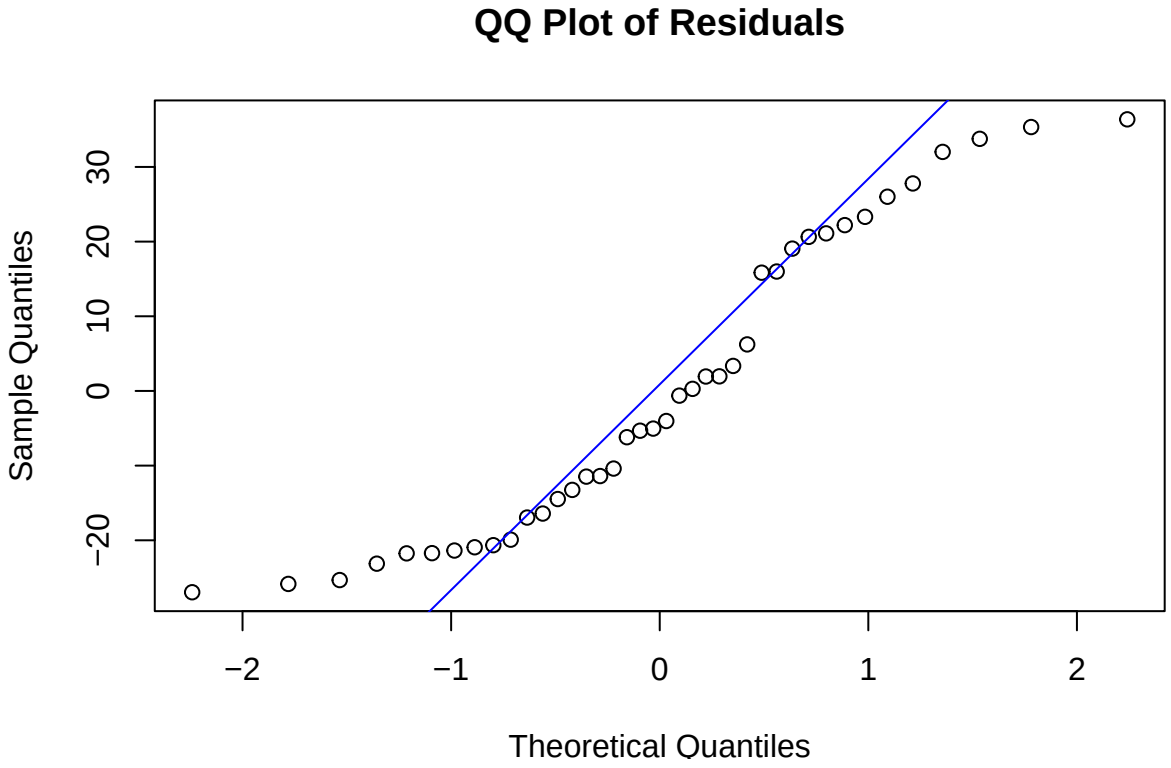
In the context of the nutrition study, since the p-value (2.199896e-06) is much smaller than = 0.05, we have strong evidence to reject the H0, and that daily sodium intake (mm) influences blood pressure (mm hg). The R squared value being 0.45, indicates that while daily sodium intake has influence on the participants' blood pressure, it explains less than half the variance in blood pressure, and we should look for other predictors. We are 95% confident that the interval 0.013-0.027 units contains the true slope (ß1), whereby for each additional mm of sodium, we see an increase of about 0.02 mm hg in participants' blood pressure.

2. What is a potential issue with this study? How would it affect the quantities that you reported in Part 1? (HINT: when I created this dataset, I set the true association between sodium intake and blood pressure to be ß1 = 0.01.)

If the true slope has been set to ß1 = 0.01, but our estimated slope ß1hat = 0.02, then our model overestimates the effect of sodium on blood pressure. This is likely due to the datasets' omission of key variables generally factored into analyses of patient data, i.e. age, sex, weight, height, relation to other participants (e.g. whether participants are family or entirely independent), prior disease history (e.g. preexisting issues with sodium intake like hypertension). Without controlling for these variables that are likely correlated with blood pressure, our model ends up estimating that sodium is more influential than it truly is.

3. Generate a qqplot and a histogram of the residuals. What potential problem do you see? Why might this problem NOT be your first priority to fix.

```r
qqnorm(residuals(sodium_model), main = "QQ Plot of Residuals")
qqline(residuals(sodium_model), col = "blue")
```



Looking at the QQ plot, there are clear "heavy tails", where the data points do not align with the diagonal line and stray far from the normal distribution. This indicates that an assumption of normality has been violated. One possible reason is that participants are not independent—since the data collector measured multiple family members per household, their blood pressure and sodium intake may be more similar than if all participants were independent. Nonetheless, addressing this normality violation may not be our first priority, because ß1hat is overestimated. Before considering normality, we must first account for the omitted variables causing bias in the model, such as age, sex, height, or genetic predisposition, which may confound the relationship between sodium intake and blood pressure. Obtaining an unbiased estimator is the more important step before addressing normality.