Homework-4-Q3

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: March 17 2025

Problem 3.

Estimated glomerular filtration rate (GFR) measures the level of kidney function. It can be used to determine the stage of kidney disease. GFR is calculated primarily from the results of blood creatinine test. An investigator is interested in the association between GFR (response) and age, sex, race, and BMI in patients with coronary artery disease. She plans to look at the association in a study cohort of 366 patients aged 19-90 years. This dataset is GFR.txt. The following variables are relevant to this analysis:

- ID: The patient ID (x366)
- BL_GFR: The estimated glomerular filtration rate
- Age: the patient age in years
- Male: male sex, Male = 1 if male and 0 if female
- Black: race, Black = 1 if black and 0 if non-black
- BMI: body mass index (kg/m2)
- BMIcat: BMI categories based on BMI cut points 25 and 30 (3 unique categories)
- **Y (response variable)**: 'BL_GFR', estimated GFR
- **X1**: 'Age', patient age in years (19-90 years)
- **X2**: 'Male', sex
- **X3**: 'Black', race
- **X4**: 'BMI', BMI in patients with coronary artery disease

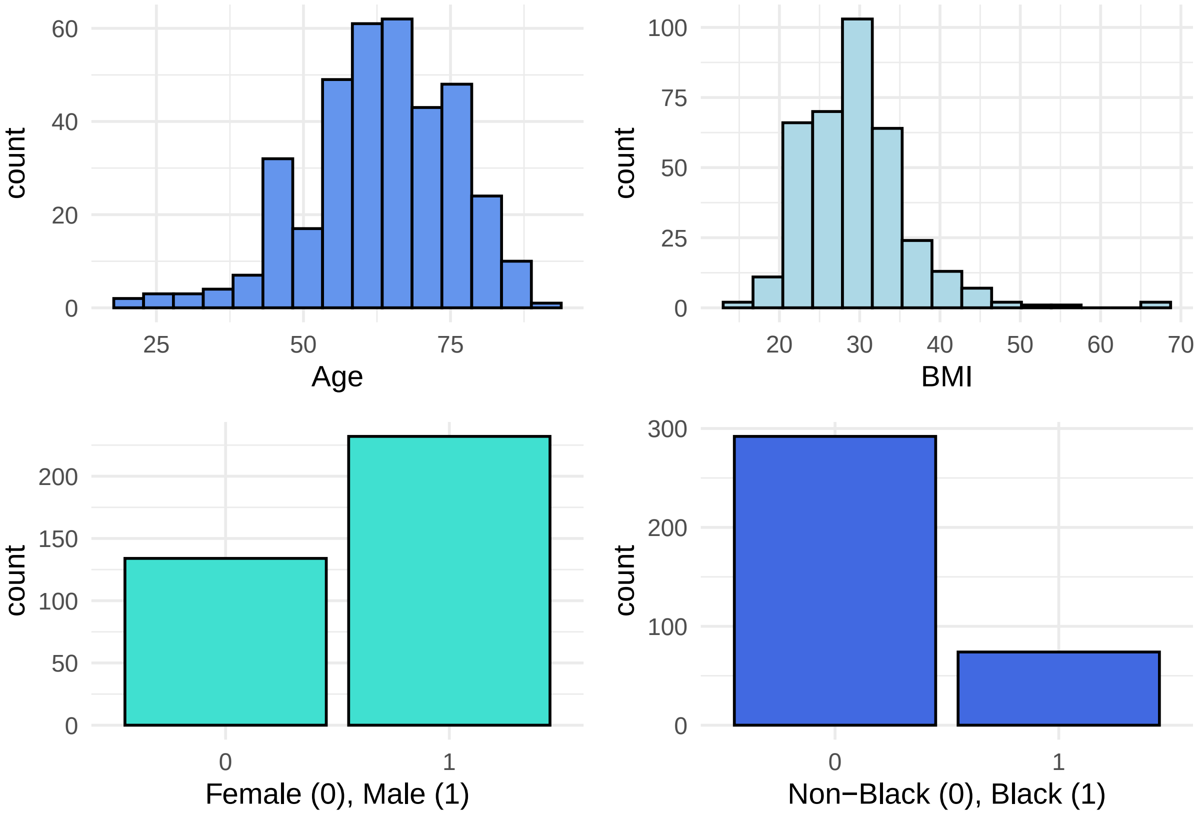
```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-4/"
data_file <- file.path(data_path, "GFR.txt")
GFR_data <- read.delim(data_file, header = TRUE)

head(GFR_data)
#str(GFR_data)
```

1. Conduct exploratory data analyses for the variables that the researcher is interested in (outcome and predictors). Provide any relevant plots or tables.

First, we will plot the variables of interest. We have 2 continuous variables (age, BMI) for which we use a histogram, and we have 2 categorical variables (race, sex) for which we rather use a bar chart.

```
age_histogram <- GFR_data %>%
  ggplot(aes(x = Age)) +
  geom_histogram(bins = 15, color = "black",fill = "cornflowerblue") +
  theme_minimal()

bmi_histogram <- GFR_data %>%
  ggplot(aes(x = BMI)) +
  geom_histogram(bins = 15, color = "black", fill = "lightblue") +
```

```r
  theme_minimal()

sex_barchart <- GFR_data %>%
  ggplot(aes(x = factor(Male))) +
  geom_bar(color = "black", fill = "turquoise") +
  theme_minimal() +
  labs(x = "Female (0), Male (1)")

race_barchart <- GFR_data %>%
  ggplot(aes(x = factor(Black))) +
  geom_bar(color = "black", fill = "royalblue") +
  theme_minimal() +
  labs(x = "Non-Black (0), Black (1)")

age_histogram + bmi_histogram + sex_barchart + race_barchart
```

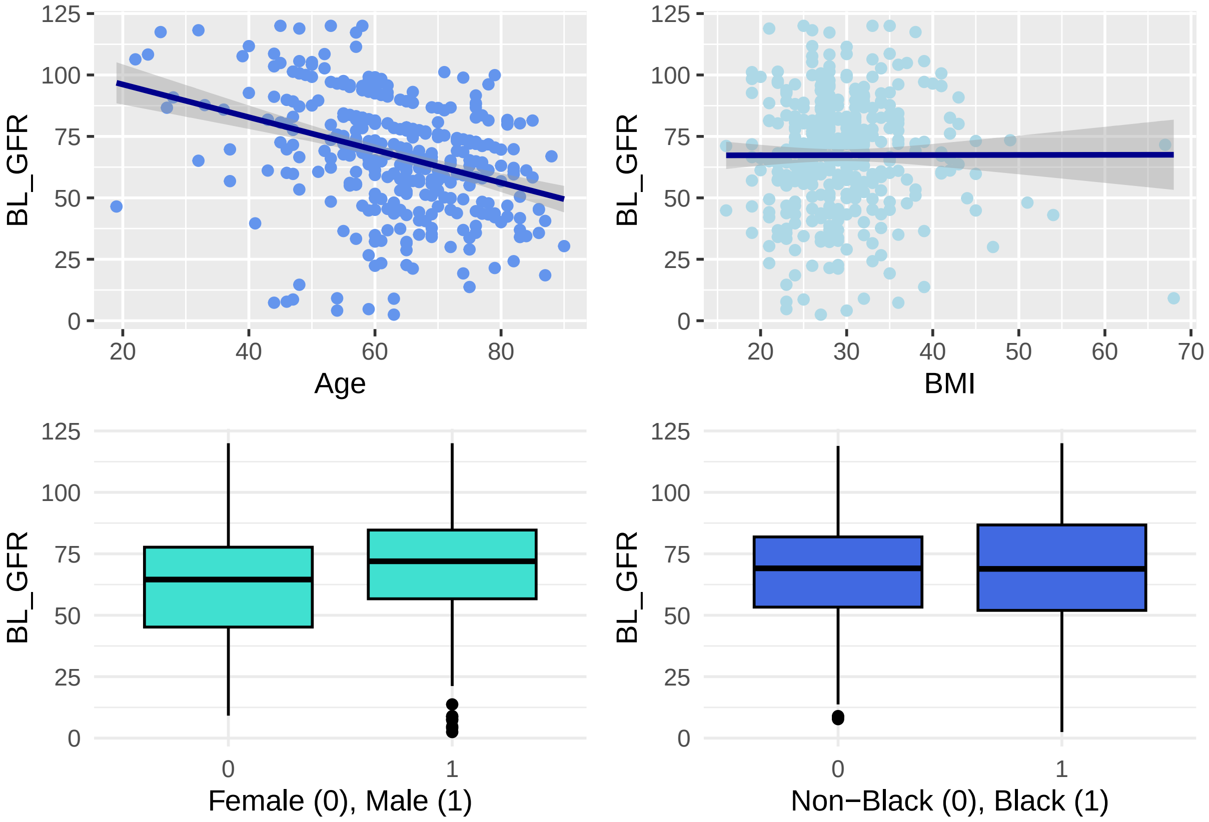


At first glance, the data appears to not have any outliers. To be noted as a potential bias down the line, there is some under-representation in the cohort, with fewer females and fewer black participants included in the study. This could be for various reasons e.g. more men getting diagnose with CAD, or the study taking place in a region with a less diverse racial profile in the local population. There is a skew towards people over 50 participating in the study, however this is not unusual, as all the patients included have been diagnosed with coronary heart disease which manifests later in life. For the BMI variable, if we look up the BMI band ranges (checked, NHS inform site), then under 18.5 is underweight, 18.5-24.9 is healthy, 25-29.9 is overweight, 30-39.9 is obese, and 40+ is severe obesity. Considering this study is American, with participants living in the U.S. and the U.S. has an obesity epidemic, it is less surprising that the majority of the population's BMI is past the "healthy range". Additionally, if we look at a study e.g. Ades & Savage 2018, they state "Over 80% of patients with CHD are overweight or obese", and CAD is the most common type of CHD, therefore, we may naturally have a more obese population as only participants with CAD were selected.

Next, we will generate scatterplots of our continuous predictors against the outcome of interest to indicate whether or not a linear relationship is appropriate. For our cateogiral preedictors, we will use boxplots.

```
GFR_age_scatter <- GFR_data %>%
  ggplot(aes(x = Age, y = BL_GFR)) +
  geom_point(color = "cornflowerblue") +
  geom_smooth(method = "lm", color = "darkblue")
  theme_minimal()

GFR_bmi_scatter <- GFR_data %>%
  ggplot(aes(x = BMI, y = BL_GFR)) +
  geom_point(color = "lightblue") +
  geom_smooth(method = "lm", color = "darkblue")
  theme_minimal()

GFR_sex_box <- GFR_data %>%
  ggplot(aes(x = factor(Male), y = BL_GFR)) +
  geom_boxplot(color = "black", fill = "turquoise") +
  theme_minimal() +
  labs(x = "Female (0), Male (1)", y = "BL_GFR")

GFR_race_box <- GFR_data %>%
  ggplot(aes(x = factor(Black), y = BL_GFR)) +
  geom_boxplot(color = "black", fill = "royalblue") +
  theme_minimal() +
  labs(x = "Non-Black (0), Black (1)", y = "BL_GFR")

GFR_age_scatter + GFR_bmi_scatter + GFR_sex_box + GFR_race_box
```

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



Age and BMI appear to roughly show a linear relationship with GFR, where GFR decreases as age increases. The relationship between GFR and BMI is less clear at this stage. When looking at race and sex predictors, it now becomes clear that there are some outliers for individuals with particularly low GFR that were not visible with just a histogram. With such a low rate, those individuals may be participants at a more severe case of kidney disease.There's a small difference between sexes, where females tend to show slightly lower GFR. The GFR across the two racial groups is approximately the same.

Finally, we can check the summary statistics for each variable to complete our EDA. For our categorical variables it would be more informative to use table rather than summary.

```
print("Age Summary:"); summary(GFR_data$Age)
print("BMI summary:"); summary(GFR_data$BMI)
print("Sex summary:"); table(GFR_data$Male)
print("Race summary:"); table(GFR_data$Black)
```

2. Test whether the variable BMIcat is contributing significantly to GFR given age, sex, and race. Report the terms in the full and reduced models, the degrees of freedom of the test, the test statistic, p-value, and your conclusions.

H0: BMIcat does not contribute significantly to GFR given age, sex, and race.
HA: BMIcat contributes significantly to GFR given age, sex, and race.

Since we want to check whether BMIcat contributes significantly to GFR, so we leave it out in the reduced model. To handle BMIcat's 3 categories we do k-1, where k=3 for 3 categories. Therfore, 3-1, leaves us now with 2 binary coded variables for ß4 and ß5.

Full model: E[BL_GFR] = ß0 + ß1(Age) + ß2(Male) + ß3(Black) + ß4(BMIcat2) + ß5(BMIcat3)
Reduced model: E[BL_GFR] = ß0 + ß1(Age) + ß2(Male) + ß3(Black)

```
# make BMIcat a factor since it has 3 categories
GFR_data$BMIcat <- factor(GFR_data$BMIcat)
```

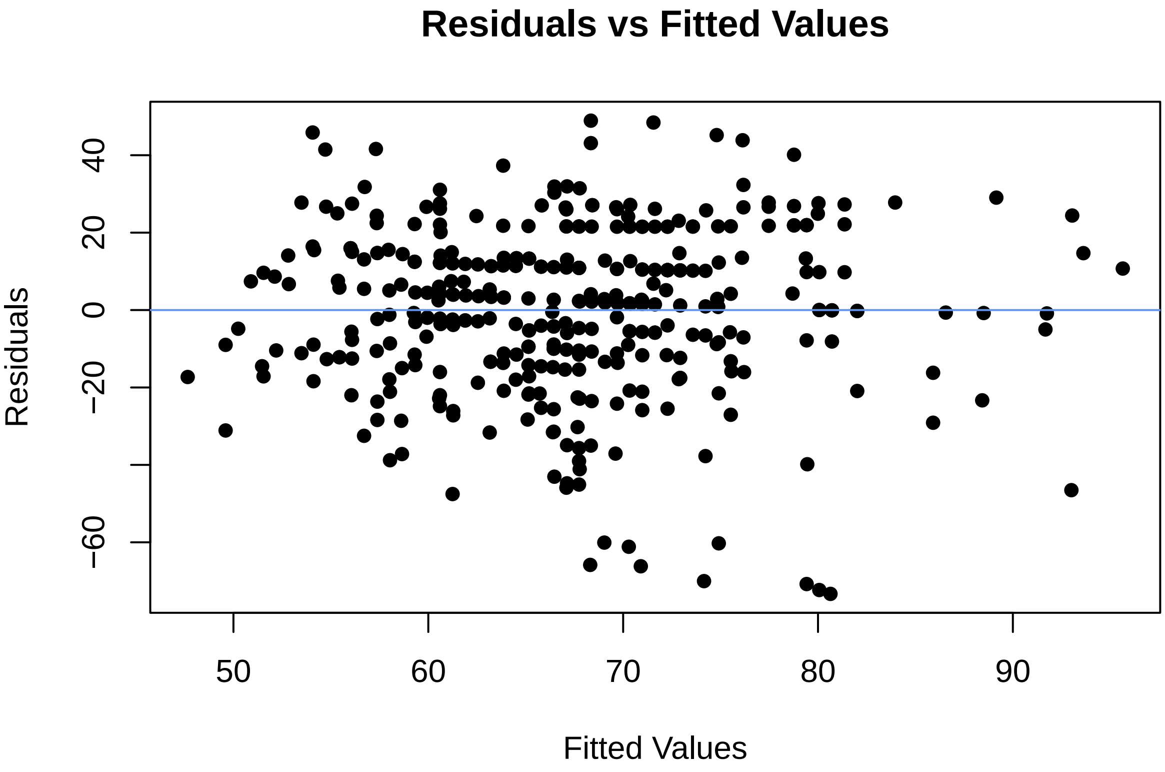
```
# first we fit the models
full_model <- lm(BL_GFR ~ Age + Male + Black + BMIcat, data = GFR_data)
reduced_model <- lm(BL_GFR ~ Age + Male + Black, data = GFR_data)

# then we compare with anova
anova_test <- anova(reduced_model, full_model)
print(anova_test)
```

The f-statistic is 0.1624. Since p = 0.8502 > 0.05, there if no evidence BMIcat contributes significantly to GFR given age, sex, and race. Therefore, we fail to reject the H0 at a(alpha) = 0.05. The numerator df (df1) is 2, calculated as the difference between the residual df of the reduced model (362) and the full model (360). The df1 of 2 to the fact that the full model has two additional parameters (the categories of BMIcat) compared to the reduced model. The f-statistic is 0.1624, therefore including the BMIcat parameter does not provide the model with a better fit, and it therefore not a significant predictor once we account for age, sex, and race.

3. Conduct residual analyses for whichever model you decide to use based on part (2). Report any relevant plots and comment on the validity of your assumptions based on the plots.

To continue the analysis, we should use the reduced model as BMIcat did not contribute significantly to GFR given age, sex, and race.

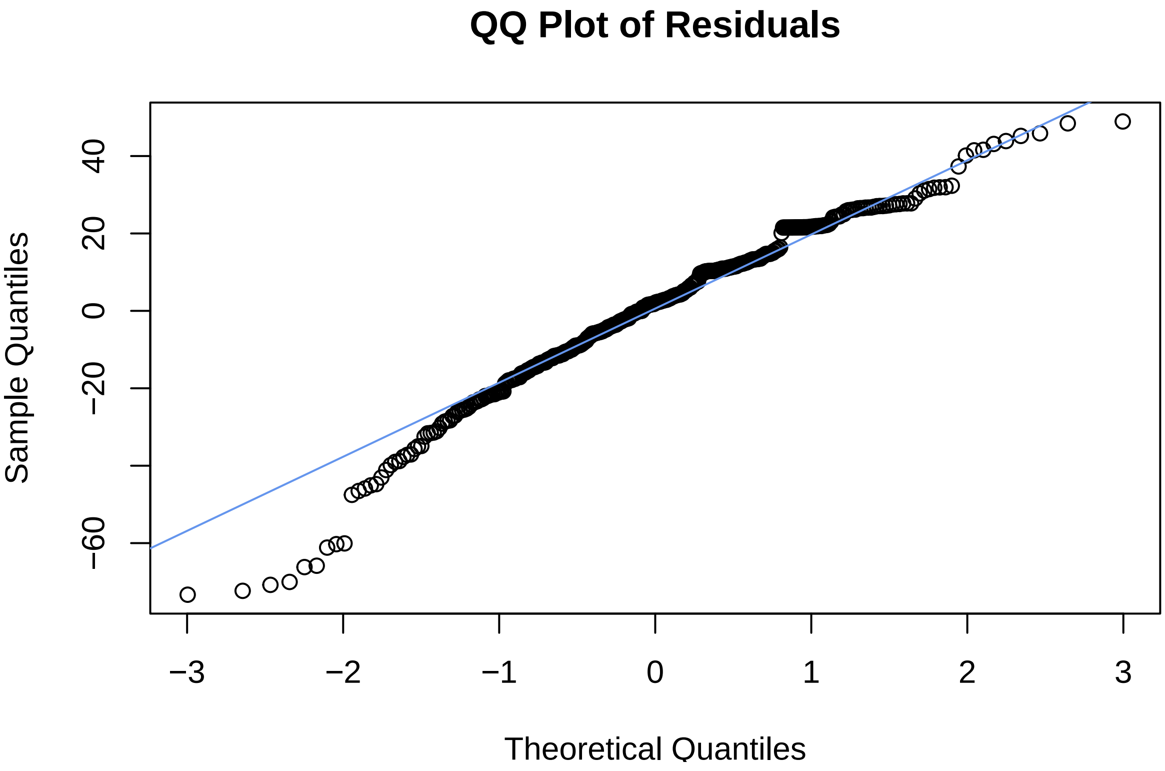
```
reduced_model <- lm(BL_GFR ~ Age + Male + Black, data = GFR_data)

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



Looking at the fitted values vs residuals, most of the residuals are roughly center around 0. There are some outliers around -60 or 40 to be noted as exceptions. There is no obvious fanning or curvature, suggesting that the variance of the errors is constant (homoscedasticity). This means the model's prediction accuracy doesn't depend on the level of GFR, instead it makes predictions consistently.

Next, we'll check the QQ plot.

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



While most of the residuals lie close to the diagonal, the left tail deviates away, aka a 'heavy tail', indicating that the assumption of normality is violated. Specifically, there are more extreme negative residuals than expected under a normal distribution, meaning that for some patients the observed GFR is much lower than the model predicts. This pattern might indicate that for patients with severe kidney disease (very low GFR), the model overestimates GFR. This may be because the model is lacking key predictors that could better capture the complexities of severe kidney dysfunction.

4. The investigator wants to know whether the association between GFR and age is being modified by race, that is, whether the association is different in blacks and non-blacks. Use appropriate modeling and conduct the corresponding tests. Show all models that you fit and clearly state your hypotheses. Report your findings.

H0: The effect of Age on GFR is the same for both racial groups (blacks vs non-blacks). ßAge:Black = 0

HA: The effect of Age on GFR is not the same for both racial groups (blacks vs non-blacks). ßAge:Black != 0

Since earlier we found BMI was not an explanatory predictor, we'll continue without it. We're looking for interaction between Age and Race, therefore we'll include an interaction term in our full model, but not our reduced model.

Full model: E[BL_GFR] = ß0 + ß1(Age) + ß2(Male) + ß3(Black) + ß4(Age)(Black)
Reduced model: E[BL_GFR] = ß0 + ß1(Age) + ß2(Male) + ß3(Black)

```r
# time to fit the models and run anova
race_full_model <- lm(BL_GFR ~ Age + Male + Black + Age:Black, data = GFR_data)
race_reduced_model <- lm(BL_GFR ~ Age + Male + Black, data = GFR_data)

race_anova <- anova(race_reduced_model, race_full_model)
print(race_anova)
```

Since p = 0.03136 < 0.05 there is evidence that race (between blacks and non black groups) significantly modifies the relationship between Age and GFR. Therefore, we reject H0 at a(alpha) = 0.05. The numerator df (df1) is 1, calculated as the difference between the residual df of the reduced model (362) and the full model (361). The df1 of 1 to the fact that the full model has one additional parameter (the interaction term between Age and Black) compared to the reduced model. The f-statistic is 4.6694, therefore including this interaction term in the model provides much improvement in the model's fit, relative to the leftover unexplained variance. Therefore, the effect of age on GFR does differ by racial group.