PHP2511 Homework 2

Kuan-Min Lee

2023-03-02

Regression Analysis:

(a)

In this section, the file kidnet_small.csv is read into the workspace and the last two variables are converted into factor variables.

```
data_test <- read.csv("kidney_small.csv",header=TRUE,sep=",") # reading datasheet
data_test$male <- factor(data_test$male) # convert the variable male into factor type
data_test$black <- factor(data_test$black) # convert the variable black into factor type
```

After the readin, a small data exploratory analysis is conducted.

```
dim_data <- dim(data_test) # retrieve the dimension of the data
print("The dimension of the datasheet is:")
```

```
## [1] "The dimension of the datasheet is:"
```

```
dim_data
```

```
## [1] 1249    8
```

```
portion_data <- head(data_test) # retrieve a portion of the datasheet
print("Glimpse of the data:")
```

```
## [1] "Glimpse of the data:"
```

```
portion_data
```

	gfr	bascre	sbase	dbase	baseu	age	male	black
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<fct>	<fct>
1	12.8	2.9	170	105	1.760	35	1	0
2	20.1	2.6	200	100	4.950	62	1	0
3	10.9	3.8	190	100	0.110	63	1	0
4	11.2	2.0	190	100	3.200	68	0	0

	gfr <dbl>	bascre <dbl>	sbase <dbl>	dbase <dbl>	baseu <dbl>	age <int>	male <fct>	black <fct>
5	31.9	4.2	220	110	4.266	30	1	0
6	10.3	2.2	190	115	0.100	55	0	0

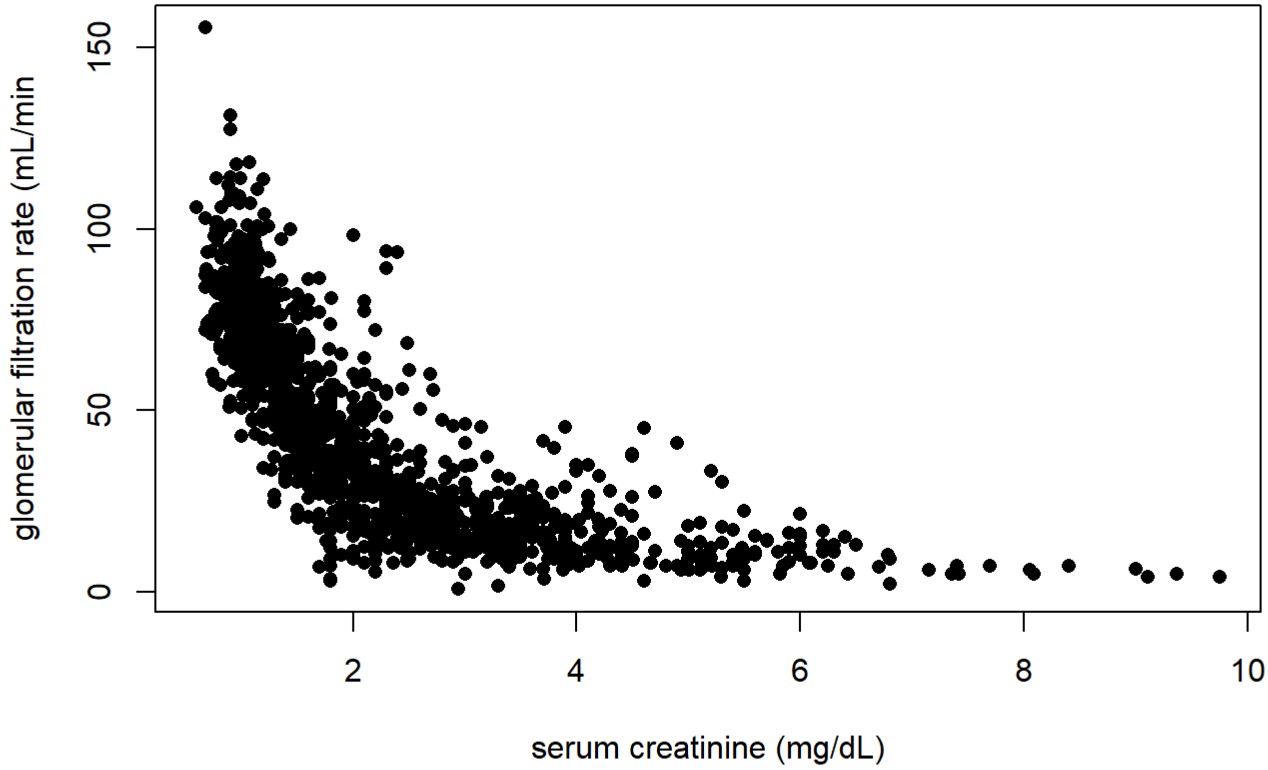
6 rows

Inside this datasheet, there are 8 different types of variables inside the datasheet. From these variables, gfr (glomerular filtration rate) is set as the y variable (or response variables). To view the relationship between gfr and other variables, several dataexplorartory analysis is conducted.

gfr vs bascre

In this portion, a scatter plot of gfr vs bascre variables is created to view the potential relationship between these two variables.

```
plot(data_test$gfr~data_test$bascre,ylab=c("glomerular filtration rate (mL/min"),xlab=c("serum creatinine (mg/dL)"),pch=16) # create a plot of gfr vs bascre
```

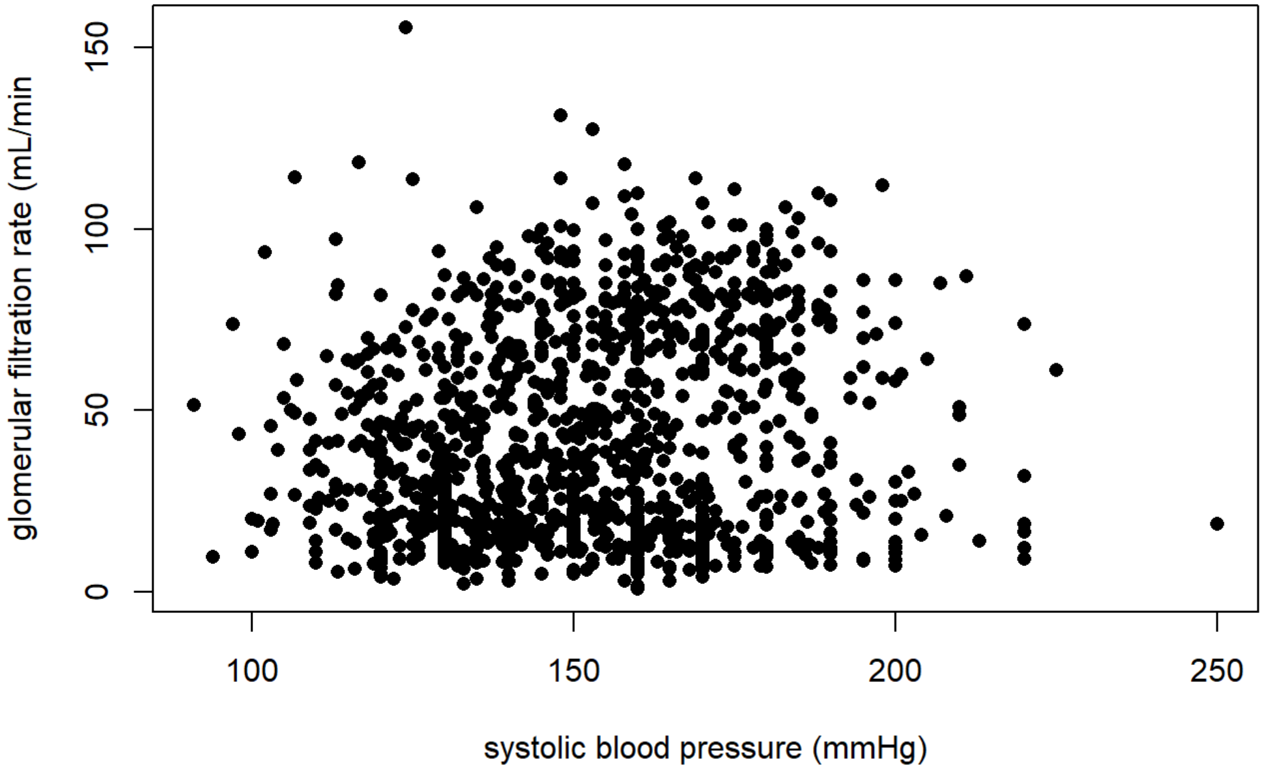


From the above graph, there seem to be an inverse exponential relationship between gfr and bascre variable.

gfr vs sbase:

In this portion, a scatter plot of gfr vs sbase variables is created to view the potential relationship between these two variables.

```
plot(data_test$gfr~data_test$sbase,ylab=c("glomerular filtration rate (mL/min"),xlab=c("systolic
blood pressure (mmHg)"),pch=16) # create a plot of gfr vs sbase
```

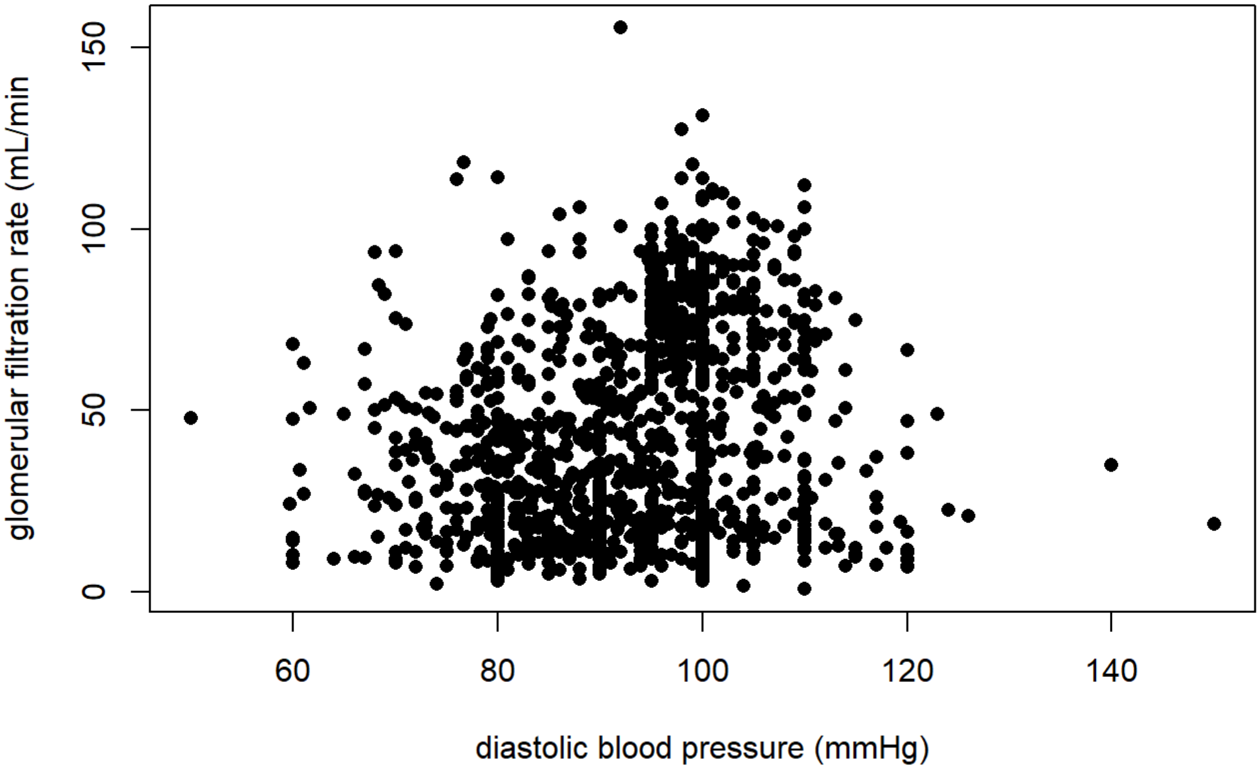


From the above plot, there is not straightforward relationship between gfr and sbase variables. Therefore, a temporarily assumption can be come up that there is no direct relationship between gfr and sbase variables.

gfr vs dbase:

In this portion, a scatter plot of gfr vs dbase variables is created to view the potential relationship between these two variables.

```
plot(data_test$gfr~data_test$dbase,ylab=c("glomerular filtration rate (mL/min"),xlab=c("diastoli
c blood pressure (mmHg)"),pch=16) # create a plot of gfr vs dbase
```

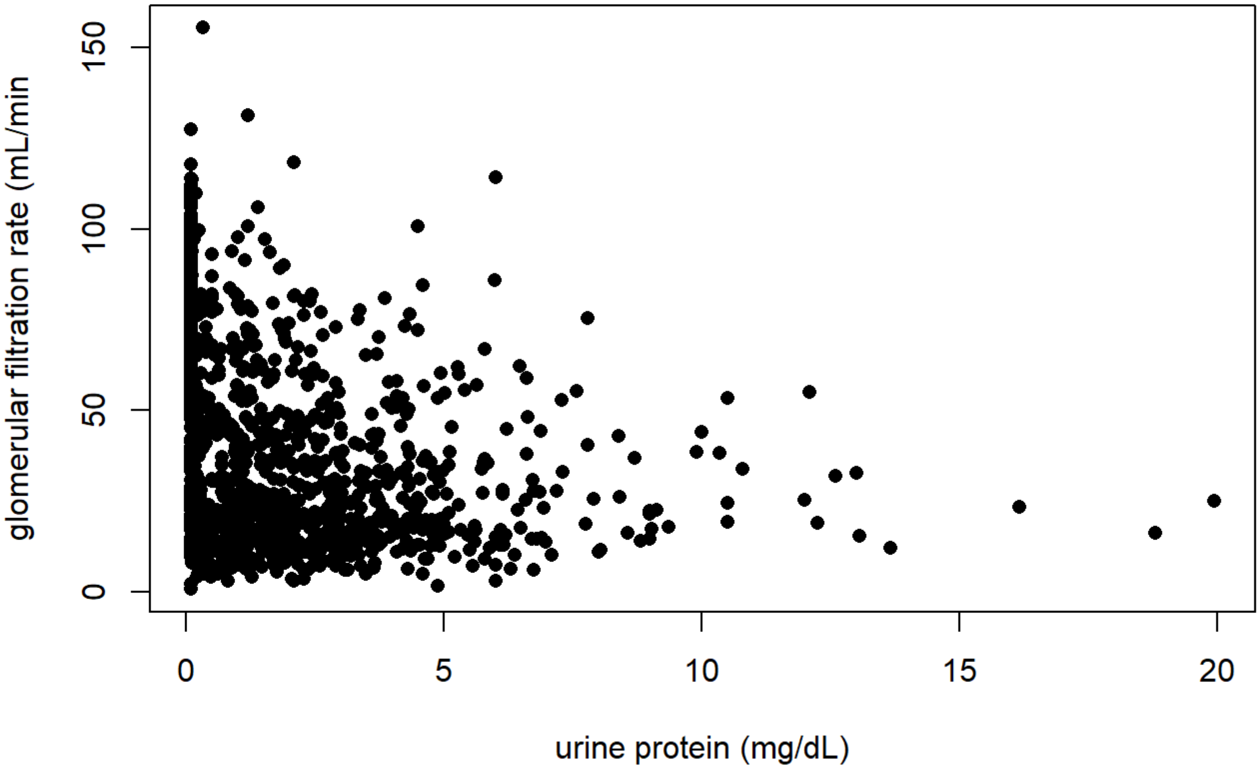


Similar to the previous section, there is not direct relationship between gfr and dbase variables.

gfr vs baseu:

In this portion, a scatter plot of gfr vs baseu variables is created to view the potential relationship between these two variables.

```
plot(data_test$gfr~data_test$baseu,ylab=c("glomerular filtration rate (mL/min"),xlab=c("urine pr
otein (mg/dL)"),pch=16) # create a plot of gfr vs baseu
```

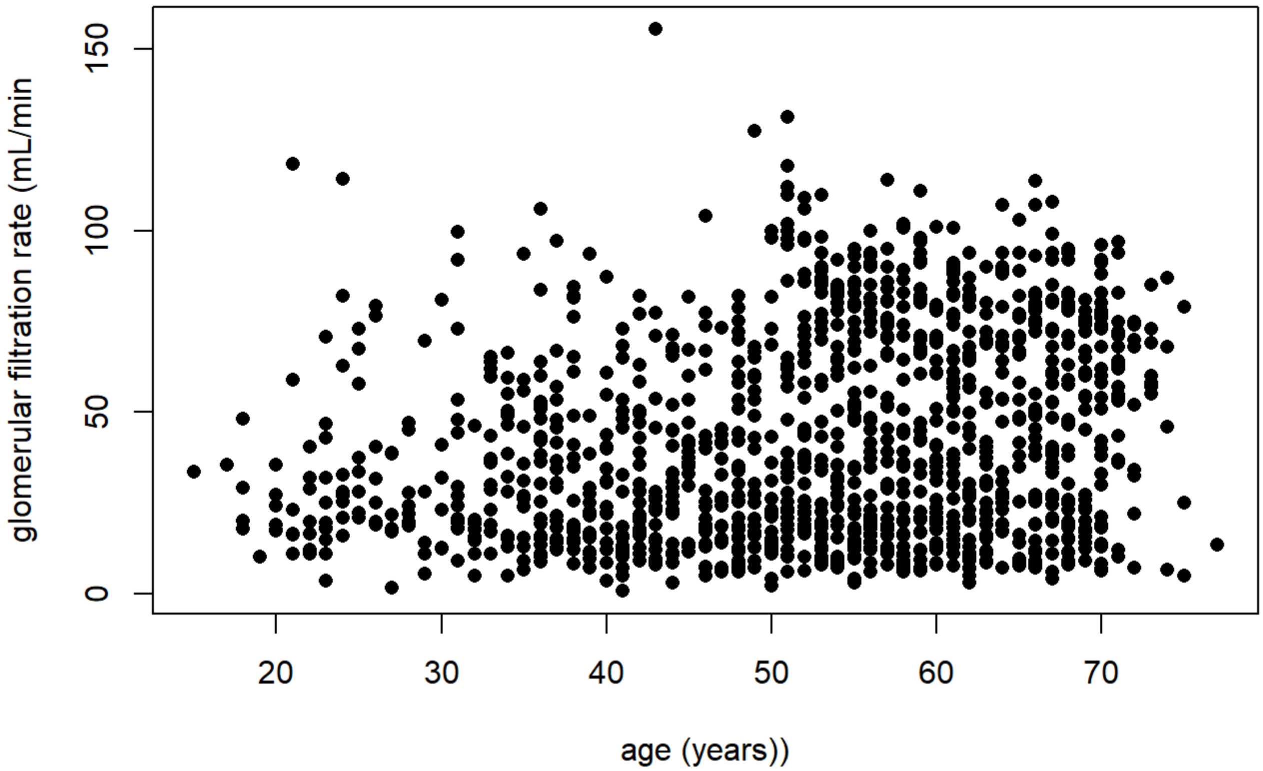


From the above pattern, it roughly follows an inverse exponential relationship between the gfr and baseu variables. Different from the bascre, the slope of inverse exponential relationship is higher.

gfr vs age:

In this portion, a scatter plot of gfr vs age variable is created to view the potential relationship.

```
plot(data_test$gfr~data_test$age,ylab=c("glomerular filtration rate (mL/min"),xlab=c("age (year
s))"),pch=16) # create a plot of gfr vs age
```

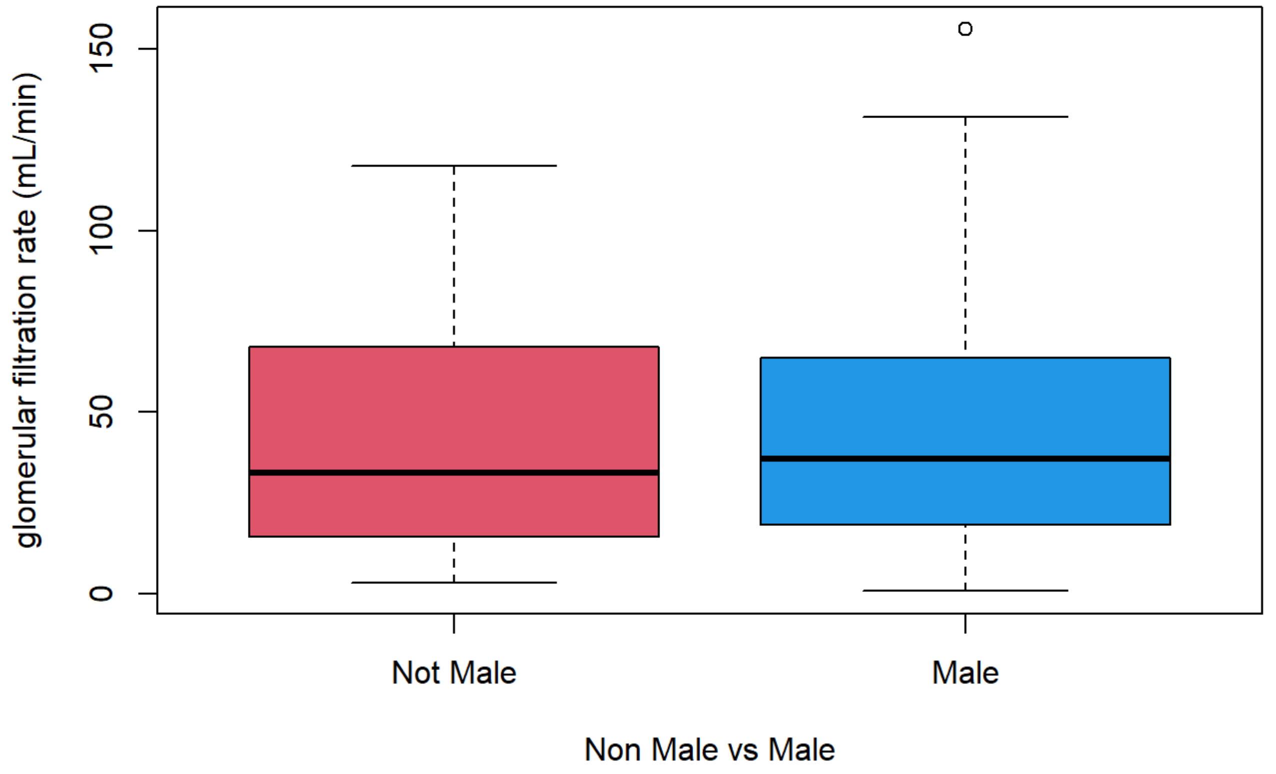


From the above plot, it can be seen that there is again no straightforward relationship between the age variable and gfr variable.

gfr vs male:

Different from the previous where all the variables are continuous variables, male is a categorical variable. Therefore, this portion a boxplot is created to observe the relationship between gfr and male variable.

```
boxplot(data_test$gfr~data_test$male, names=c("Not Male","Male"),
        xlab="Non Male vs Male",ylab="glomerular filtration rate (mL/min)",col=c(2,4))
```



```
t.test(data_test$gfr~data_test$male)
```

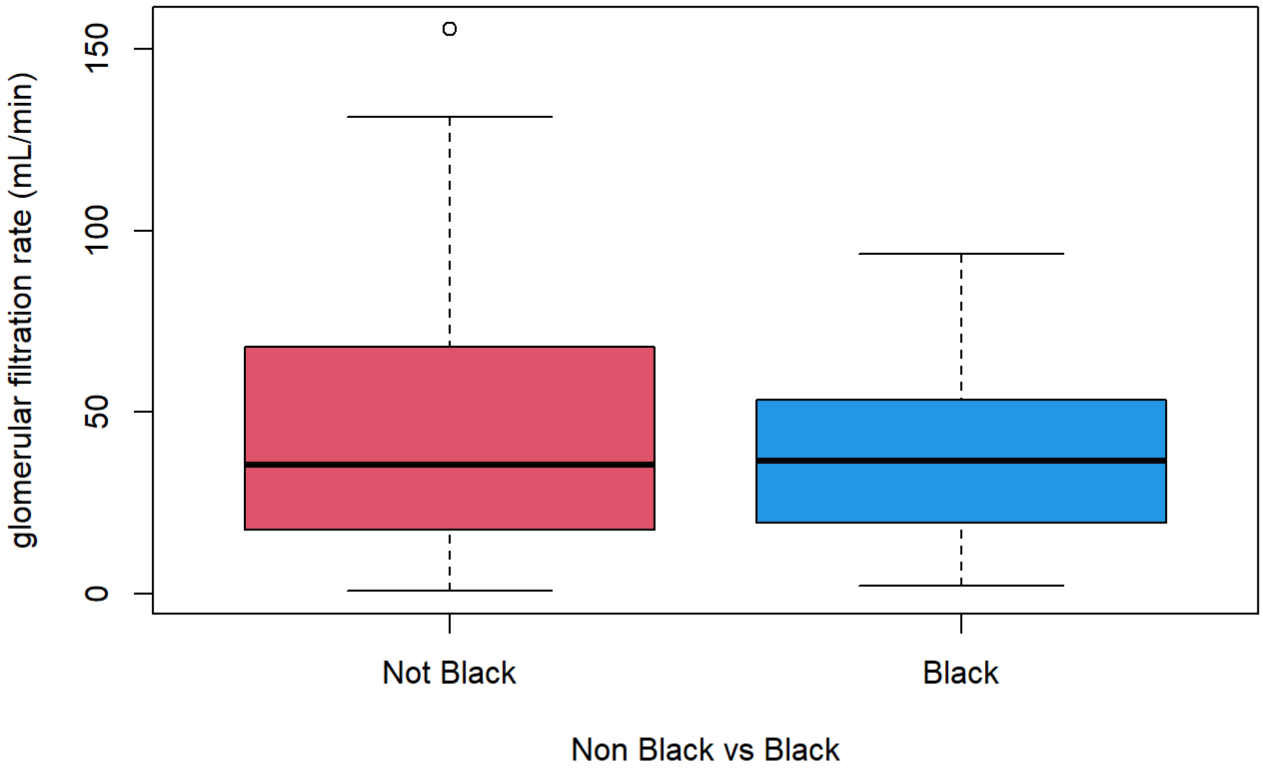
```
##
##  Welch Two Sample t-test
##
## data:  data_test$gfr by data_test$male
## t = -1.1315, df = 944.39, p-value = 0.2581
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
0
## 95 percent confidence interval:
##  -5.192139  1.394473
## sample estimates:
## mean in group 0 mean in group 1
##        41.40081        43.29965
```

From the above plot and the t test that has been implemented, there’s no difference in mean of gfr values for male and non male population.

gfr vs black:

Similarily, this portion a boxplot is created to observe the relationship between gfr and black variable.

```
boxplot(data_test$gfr~data_test$black, names=c("Not Black","Black"),
        xlab="Non Black vs Black",ylab="glomerular filtration rate (mL/min)",col=c(2,4))
```



```
t.test(data_test$gfr~data_test$black)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_test$gfr by data_test$black
## t = 2.4232, df = 162.09, p-value = 0.01649
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
0
## 95 percent confidence interval:
##  0.9458918 9.2759267
## sample estimates:
## mean in group 0 mean in group 1
##        43.05768        37.94677
```

From the above plot, both means and the dispersions of the gfr variables in the two groups are different from each other.

(b):

For this portion, a linear regression model of formula:

gfr~bascre+age+male

is implemented to fit the data trend of the gfr.

```
# setup variables for model comstruction
gfr <- data_test$gfr
bascre <- data_test$bascre
age <- data_test$age
male <- data_test$male
# construct the model
mod_gfr_basc_ag_mal <- lm(gfr~bascre+age+male,data=data_test)
summary(mod_gfr_basc_ag_mal)
```

```
##
## Call:
## lm(formula = gfr ~ bascre + age + male, data = data_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.039 -14.108  -3.093  13.060  89.607
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.92764    2.75456  23.208  < 2e-16 ***
## bascre      -14.37248    0.39794 -36.117  < 2e-16 ***
## age           0.18848    0.04265   4.419 1.08e-05 ***
## male1         3.61560    1.13475   3.186  0.00148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.35 on 1245 degrees of freedom
## Multiple R-squared:  0.536,  Adjusted R-squared:  0.5348
## F-statistic: 479.3 on 3 and 1245 DF,  p-value: < 2.2e-16
```

(c):

For this part, the R^2 value of the above model will be observed. From the definition of R value of a regression model, R can be shown as:

R^2 = 1 - var(e_i)/var(y_i)

where e_i is the residual values for each dataset and y_i is the true y_i values.

To calculate this, the residual and true values of the above model can be etraploated and its variance can be calculated.

```
test_residuals <- mod_gfr_basc_ag_mal$residuals # extract residuals
var_res <- var(test_residuals) # calculate the variance of residuals
test_y <- gfr # extract the true values
var_y <- var(test_y) #calculate the variance of y

# calculate R^2 value
test_R2 <- 1-var_res/var_y

print("The R^2 value of the model is:")
```

```
## [1] "The R^2 value of the model is:"
```

```
test_R2
```

```
## [1] 0.5359523
```

From the above outcome, it can be seen that the R2 value is roughly 0.54, which represents the proportion of the variance for a dependent variable explained by the predictor variables in the regression model. It roughly passes half, which shows theh model didn’t perform well in predicting the trend of the response variable.

(d):

For this portion of the problem, the same procedure is performed except for this part, the model is changed into:

log(gfr) ~ log(bascre) + age + male

```
# construct the model
log_mod_gfr_basc_ag_mal <- lm(log(gfr)~log(bascre)+age+male,data=data_test)
summary(log_mod_gfr_basc_ag_mal)
```

```
## 
## Call:
## lm(formula = log(gfr) ~ log(bascre) + age + male, data = data_test)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4466 -0.1929  0.0339  0.2162  1.5180 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.2583535  0.0579339  73.504   <2e-16 ***
## log(bascre) -1.2452406  0.0220023 -56.596   <2e-16 ***
## age         -0.0012383  0.0009292  -1.333    0.183    
## male1        0.2257503  0.0244811   9.221   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4164 on 1245 degrees of freedom
## Multiple R-squared:  0.7311, Adjusted R-squared:  0.7304 
## F-statistic:  1128 on 3 and 1245 DF,  p-value: < 2.2e-16
```

```
log_test_residuals <- log_mod_gfr_basc_ag_mal$residuals # extract residuals
log_var_res <- var(log_test_residuals) # calculate the variance of residuals
test_y <- gfr # extract the true values
var_y <- var(test_y) #calculate the variance of y

# calculate R^2 value
test_R2 <- 1-log_var_res/var_y

print("The R^2 value of the model is:")
```

```
## [1] "The R^2 value of the model is:"
```

```
test_R2
```

```
## [1] 0.9997852
```

From the above R2 value, it can be seen that the portion of variance of the response variables that can be explained by the predictor variables in the model is enlarged, which indicates an improvement due to the modification of the model.

Hypothesis Tests:

(a):

For this portion, different hypothesis tests will be shown for each coefficient.

p-value for each coefficient is determined by the data distribution of the test statistic that is used to judge whether the null hypothesis holds.

From the above model built in the previous section, there are three variables: log(bascre), age, and male variables.

First is log(bascre) variable and the below is its null hypothesis and alternative hypothesis:

null hypothesis (H0): there is no relationship (slope=0) between log(gfr) and log(bascre) variable. alternative hypothesis (Ha): there exists a relationship between the variable log(gfr) and log(bascre) variables.

```r
coe_tbl <- tidy(log_mod_gfr_basc_ag_mal) # create a table for coefficients extraction
coe_tbl
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	4.258353523	0.0579339462	73.503598	0.000000e+00
log(bascre)	-1.245240633	0.0220023290	-56.595856	0.000000e+00
age	-0.001238335	0.0009291812	-1.332716	1.828689e-01
male1	0.225750266	0.0244811115	9.221406	1.215464e-19

4 rows

From above, it can be observed that the p_value for log(bascre) variable is less than 0.05, which indicates that the null hypothesis of log(bascre) is rejected.

As for the second variable, age, its null hypothesis and alternative hypothesis is shown below:

null hypothesis (H0): there is no relationship (slope=0) between log(gfr) and age variable. alternative hypothesis (Ha): there exists a relationship between the variable log(gfr) and age variables

From above, the p-value of age variable is larger than 0.05, which indicates that we should accpet the null hypothesis of age variable.

Last is the male variable and its hypothesises can be shown as below:

null hypothesis (H0): there is no relationship (slope=0) between log(gfr) and the fact that the tested sample is a male or not. alternative hypothesis (Ha): there exists a relationship between the variable log(gfr) and male variables

From aboe table, the p-value for male variable is less than 0.05, which indicates that we should not accept null hypothesis of the variable.

(b):

For this part, the 95% confidence interval will be constructed for log(bascre) variable.

```
# get the t statistic
t_val_log_basc_upp <- qt(0.975,1245)
t_val_log_basc_low <- qt(0.025,1245)

mean_log_basc <- mean(log(bascre)) # receive the mean of log(bascre)
std_log_basc <- sd(log(bascre)) # receive the standard deviation of log(bascre)
n_log_basc <- length(log(bascre)) # receive the sample size of log(bascre)

t_val_log_basc_upp_bound <- mean_log_basc+t_val_log_basc_upp*std_log_basc/sqrt(n_log_basc) # calculate the upper bound
t_val_log_basc_low_bound <- mean_log_basc+t_val_log_basc_low*std_log_basc/sqrt(n_log_basc) # calculate the lower bound

print("The 95% confidence interval for log(bascre) is:")
```

```
## [1] "The 95% confidence interval for log(bascre) is:"
```

```
c(t_val_log_basc_low_bound,t_val_log_basc_upp_bound)
```

```
## [1] 0.6570687 0.7186359
```

(c):

For this part, the F-statistics are viewed and degree of freedom for F-statistics are observed:

```
summary(mod_gfr_basc_ag_mal) # view F statistics
```

```
##
## Call:
## lm(formula = gfr ~ bascre + age + male, data = data_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.039 -14.108  -3.093  13.060  89.607
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.92764    2.75456  23.208  < 2e-16 ***
## bascre      -14.37248    0.39794 -36.117  < 2e-16 ***
## age           0.18848    0.04265   4.419 1.08e-05 ***
## male1         3.61560    1.13475   3.186  0.00148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.35 on 1245 degrees of freedom
## Multiple R-squared:  0.536,  Adjusted R-squared:  0.5348
## F-statistic: 479.3 on 3 and 1245 DF,  p-value: < 2.2e-16
```

The above is the result of the F test. From above, it can be seen taht the degree of freedom is 1245, which is defined as:

degree of freedom = (sample number) - (predictor) - 1

For this part, we have 1249 samples, and 3 predictors, which contirbutes to the degree of freedom of 1245.

(d):

For this part, the hypothesis for F statistics is implemented for the model:

null hypothesis: all coefficients are zero in the model (no relationship is found between the predictors and the response variables) alternative hypothesis: at least one coefficient is found nonzero (there exists at least one relationship between one of the predictor variable and the response variable)

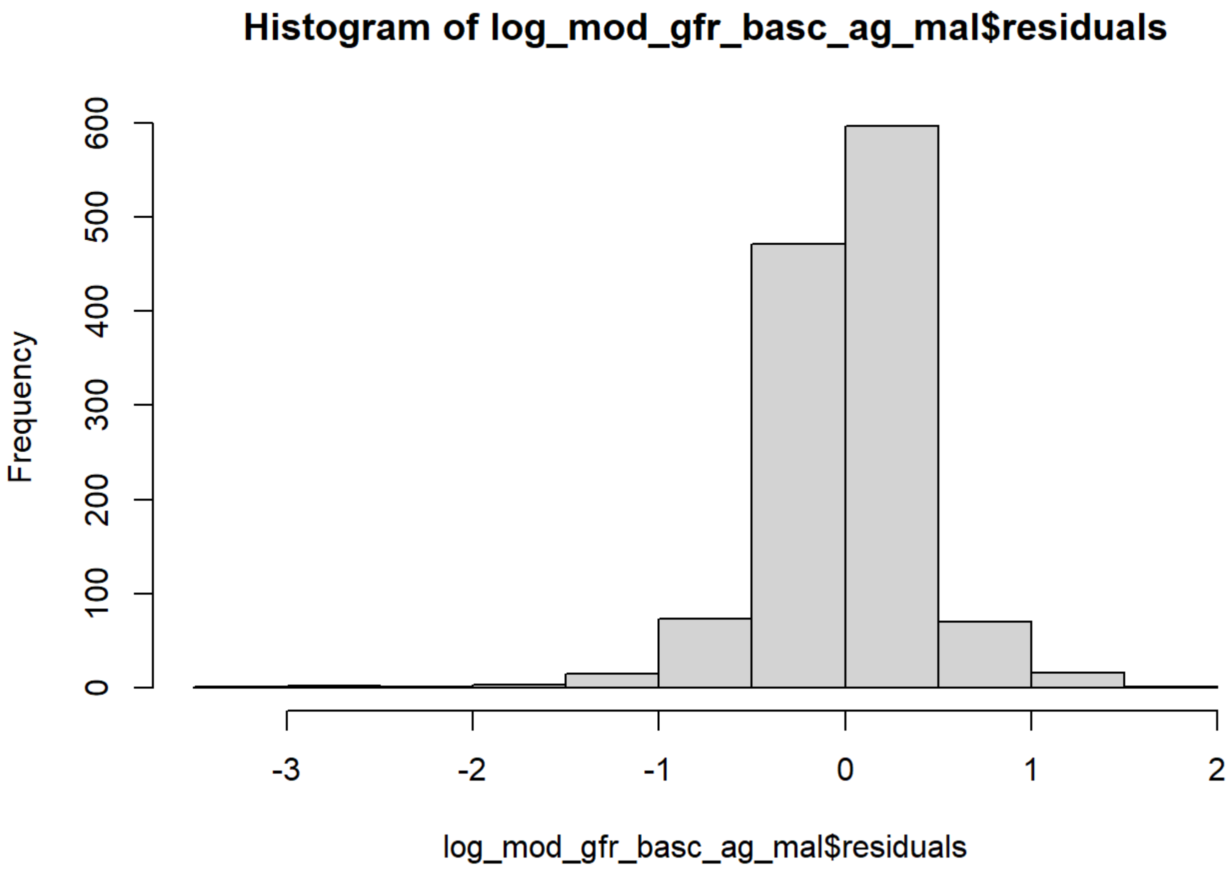
From the reported p value for F statistics, it shows that the null hypothesis is rejected.

Diagnostic Plots:

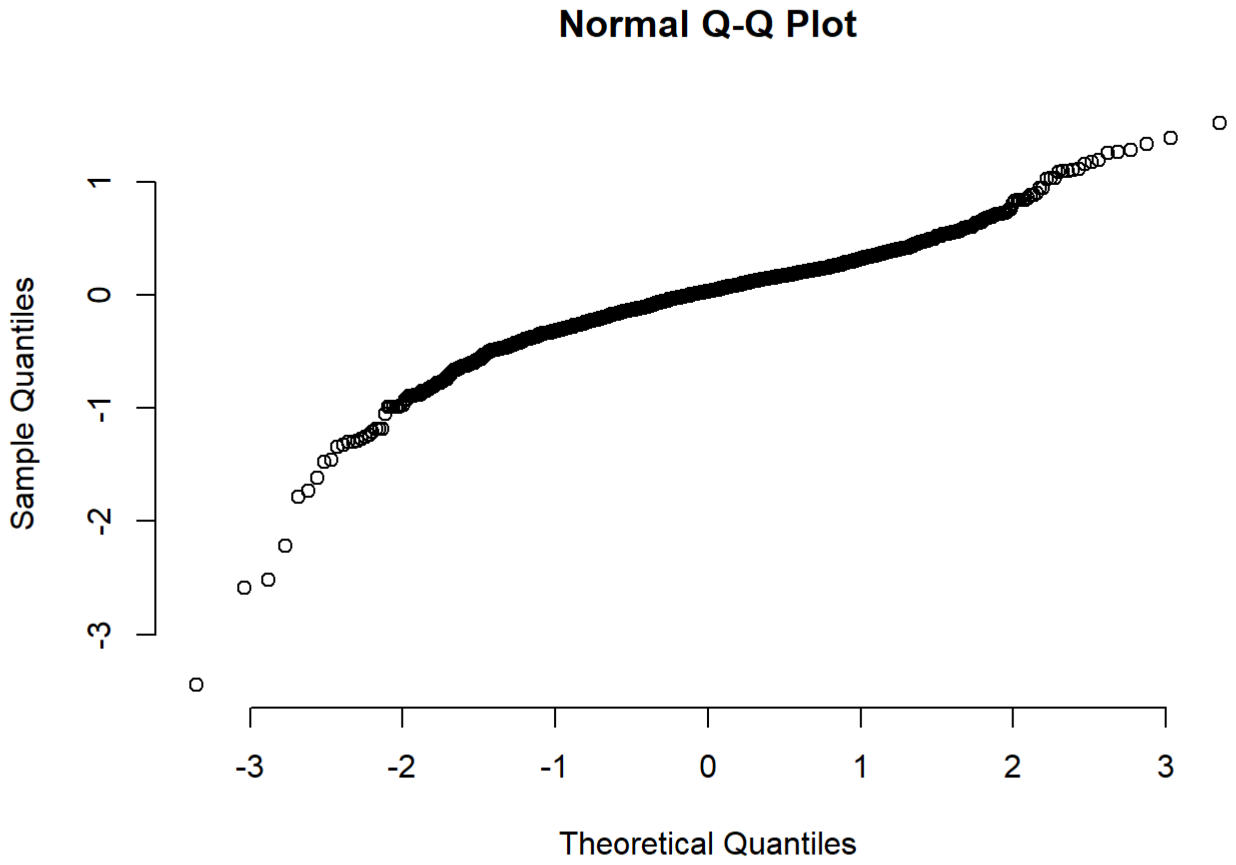
(a):

In this part, a histogram of residuals and a QQ plot of that will be shown.

```
hist(log_mod_gfr_basc_ag_mal$residuals) # create a histogram for residuals
```



```
qqnorm(log_mod_gfr_basc_ag_mal$residuals, pch=1, frame=FALSE) # create a qq plot for residuals
```

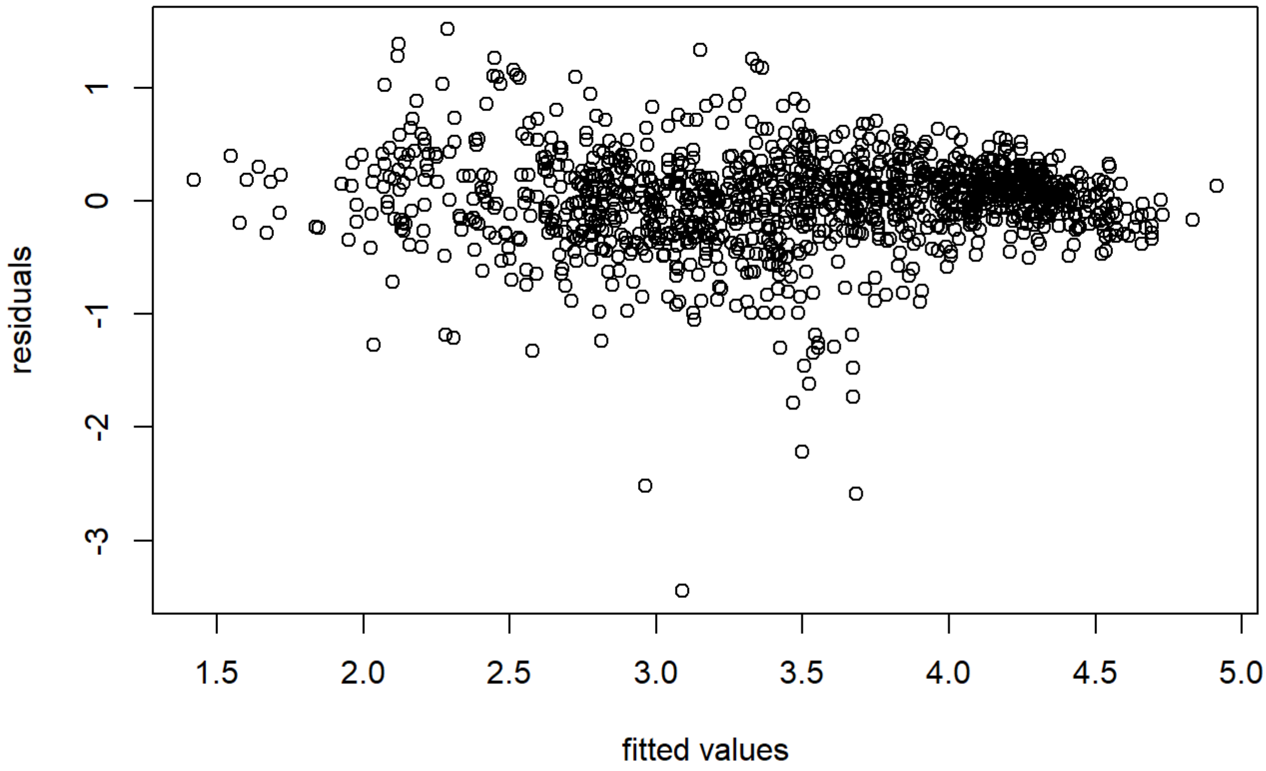


Based on both the histogram and qq plot for the residual, it can be seen that the distrubution of residuals mostly follow normal distribution. The shape of the histogram looks similar to the shape of normal distribution and the qq plot mostly follows a linear relation.

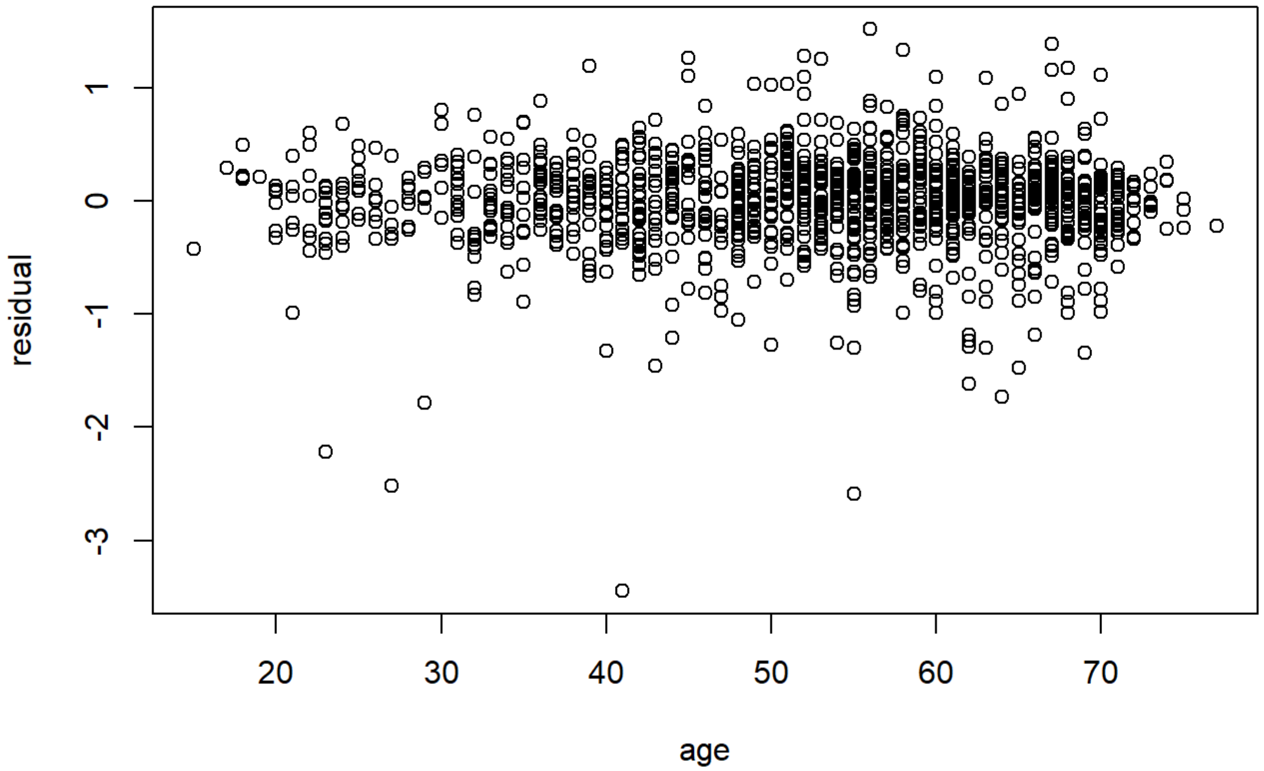
(b):

In this part, there are two graphs will be created to make observation on the fitted model: residuals vs fitted values plot and resituals vs covariates plot.

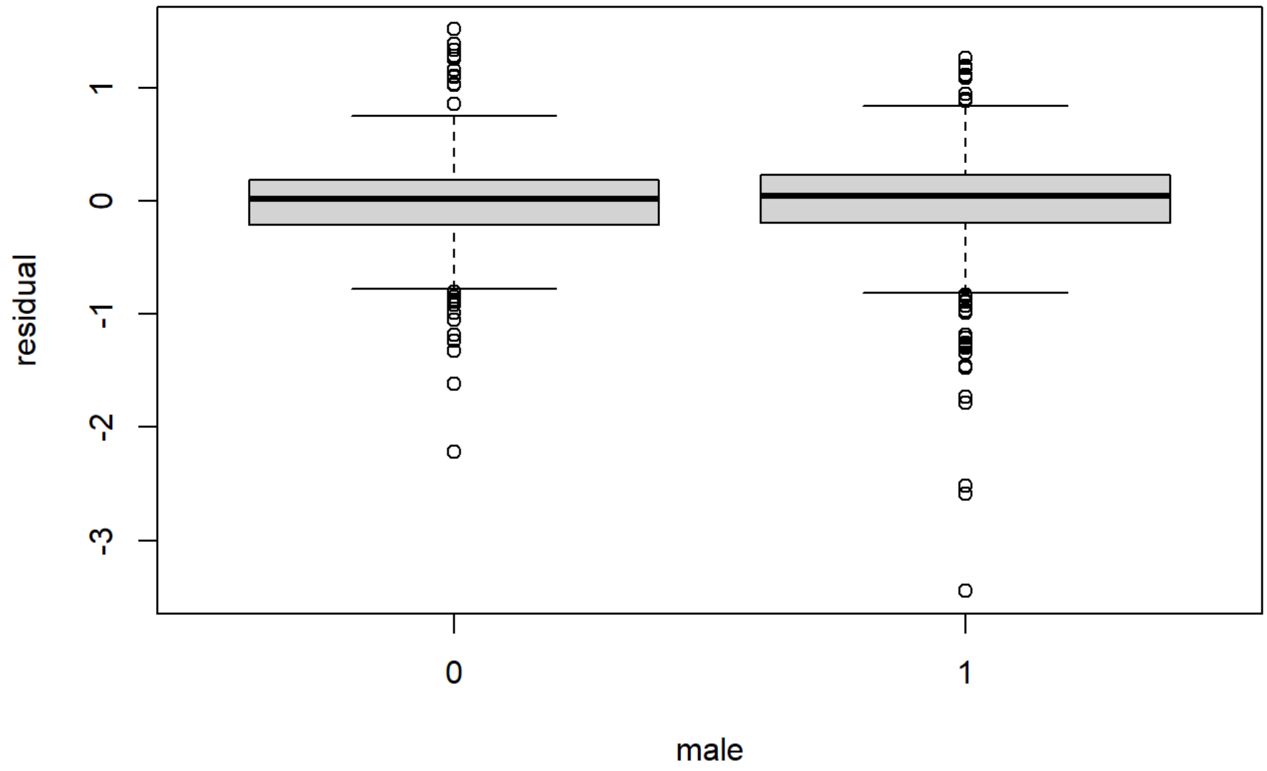
```
plot(log_mod_gfr_basc_ag_mal$fitted.values,log_mod_gfr_basc_ag_mal$residuals,xlab=c("fitted valu
es"),ylab=c("residuals")) # create the plot for residuals vs fitted values
```



```
plot(age,log_mod_gfr_basc_ag_mal$residuals,xlab=c("age"),ylab=c("residual")) # create the plot f
or residuals vs covariate age
```



```
boxplot(log_mod_gfr_basc_ag_mal$residuals~male,xlab=c("male"),ylab=c("residual")) # create the p
lot for residuals vs covariate age
```



From the results shown above, it can be seen that the residuals of the fitted model is roughly homescedasticity, since the variance of the residuals does not vary a lots for different fitted value and covariates.

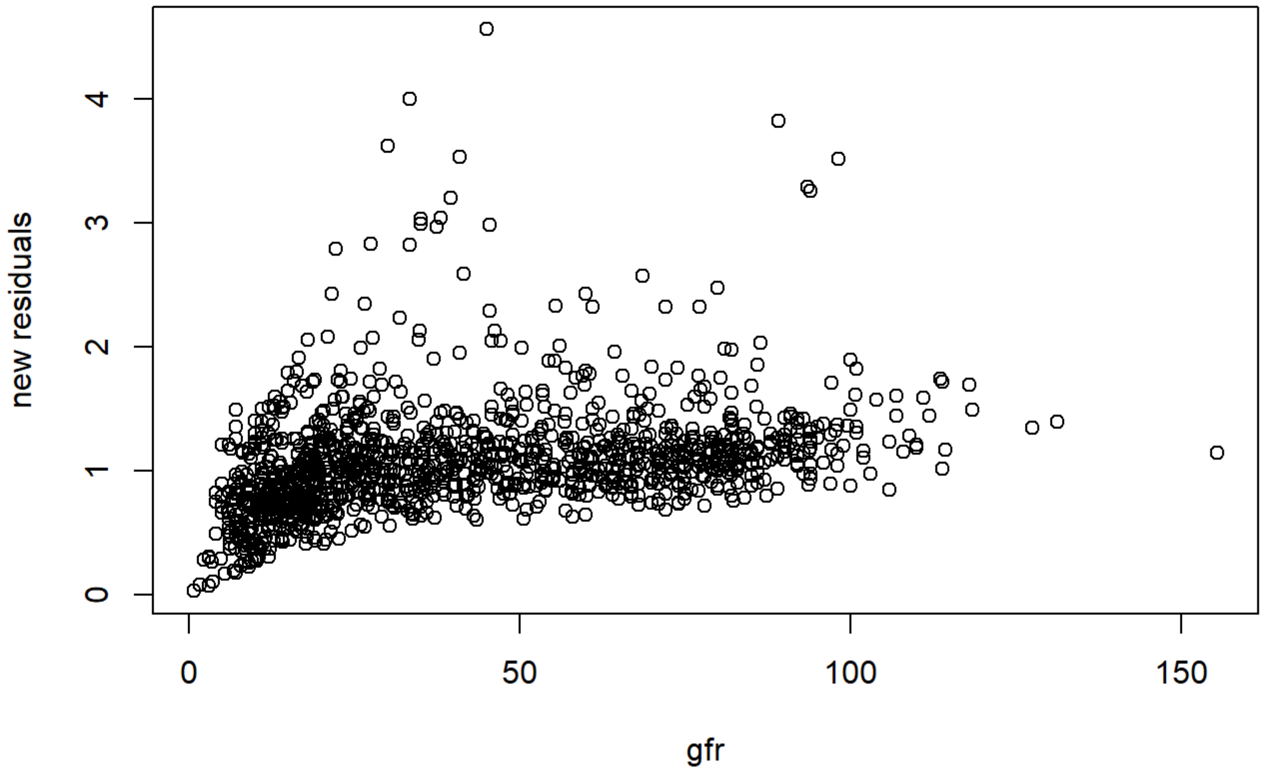
Outliers, Leverage, and Influence:

(a):

In this part of the problem, the predicted values of the predicted value will be exponentiated and then plotted into the plot of residuals vs gfr

```
new_predicted_vals <- exp(log_mod_gfr_basc_ag_mal$fitted.value) # exponentiate the original fitted values
new_residual_vals <- exp(log_mod_gfr_basc_ag_mal$residuals) # exponentiate the original fitted values

plot(gfr,new_residual_vals,xlab=c("gfr"),ylab=c("new residuals")) # plot the predicted values vs gfr plot
```

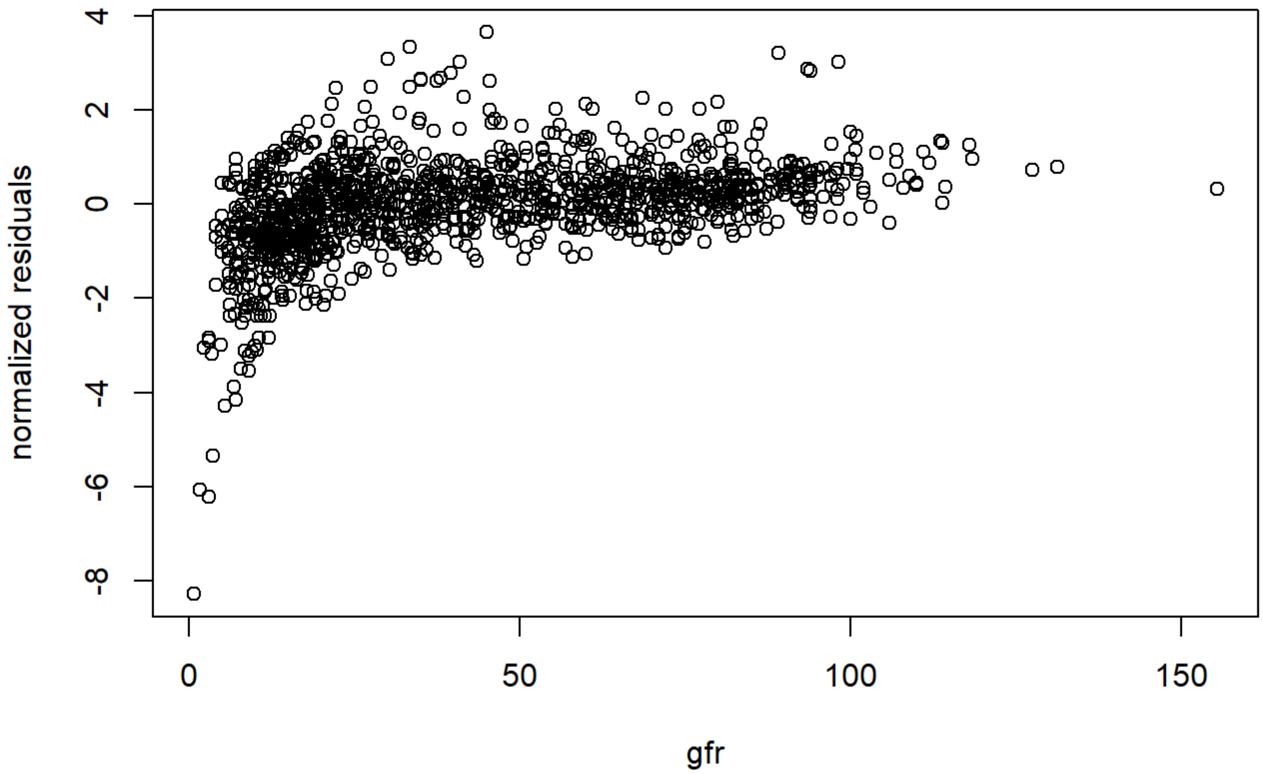


From the above observation, we can make the conclusion that it has the trend that as gfr increases, the residual increases as well.

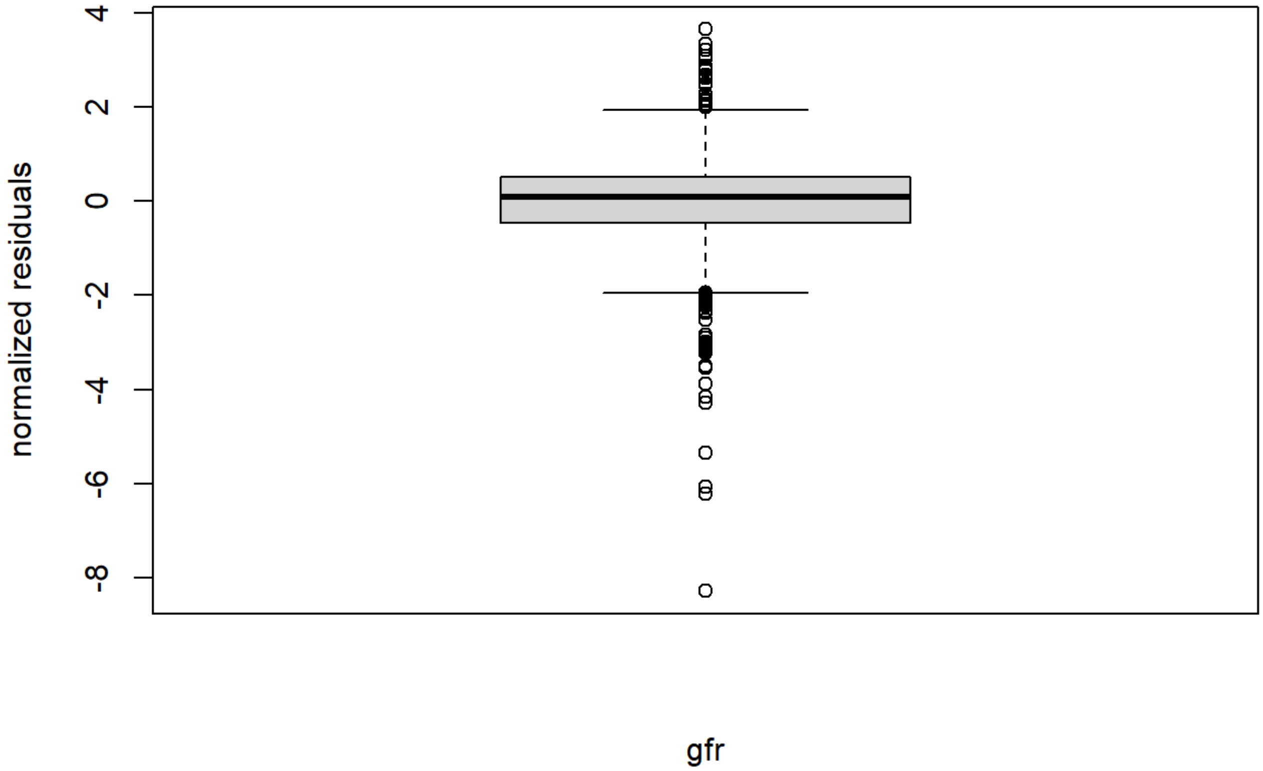
(b):

In this part of the problem, a new funciton, rstandard, will be inserted to normalize the residuals of the above model and use the new normalized residuals as a criterion for judging the outliers:

```
normal_log_res<-rstandard(log_mod_gfr_basc_ag_mal) # standardized the residuals
plot(gfr,normal_log_res,xlab=c("gfr"),ylab=c("normalized residuals"))
```

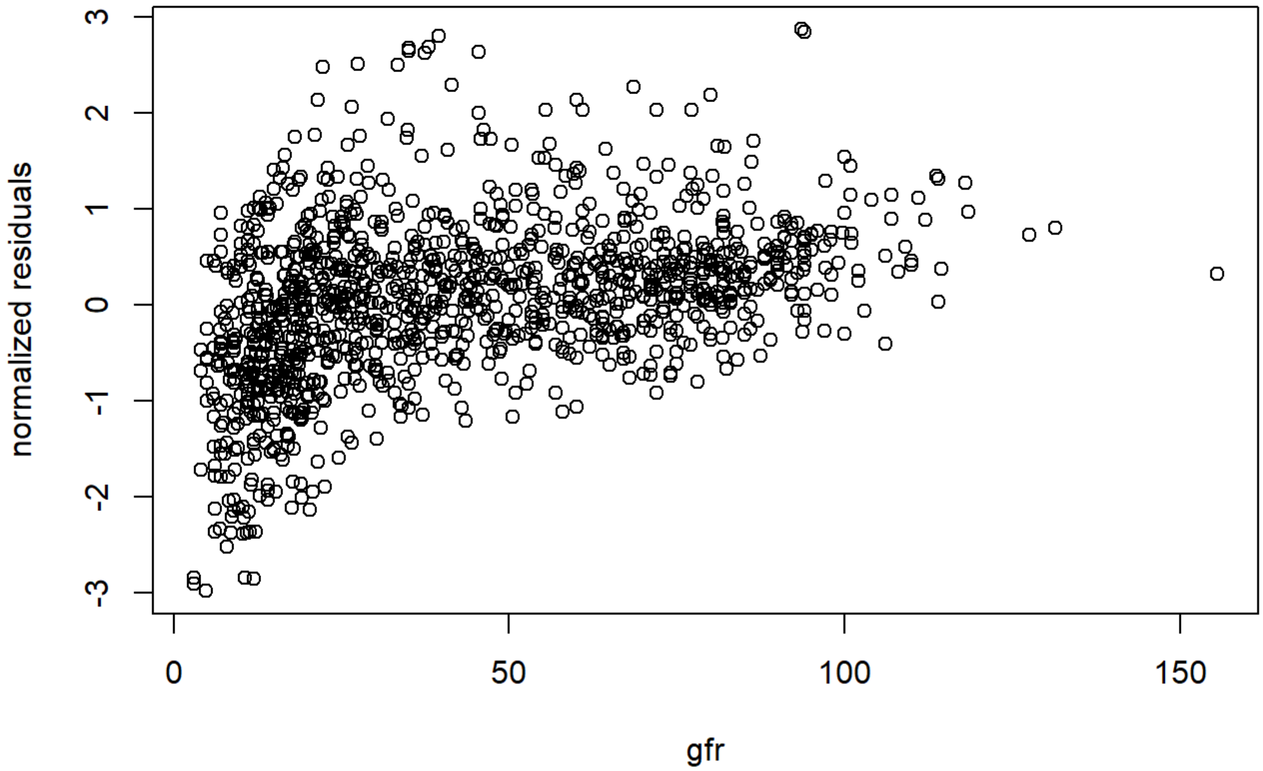


```
boxplot(normal_log_res,xlab=c("gfr"),ylab=c("normalized residuals"))
```

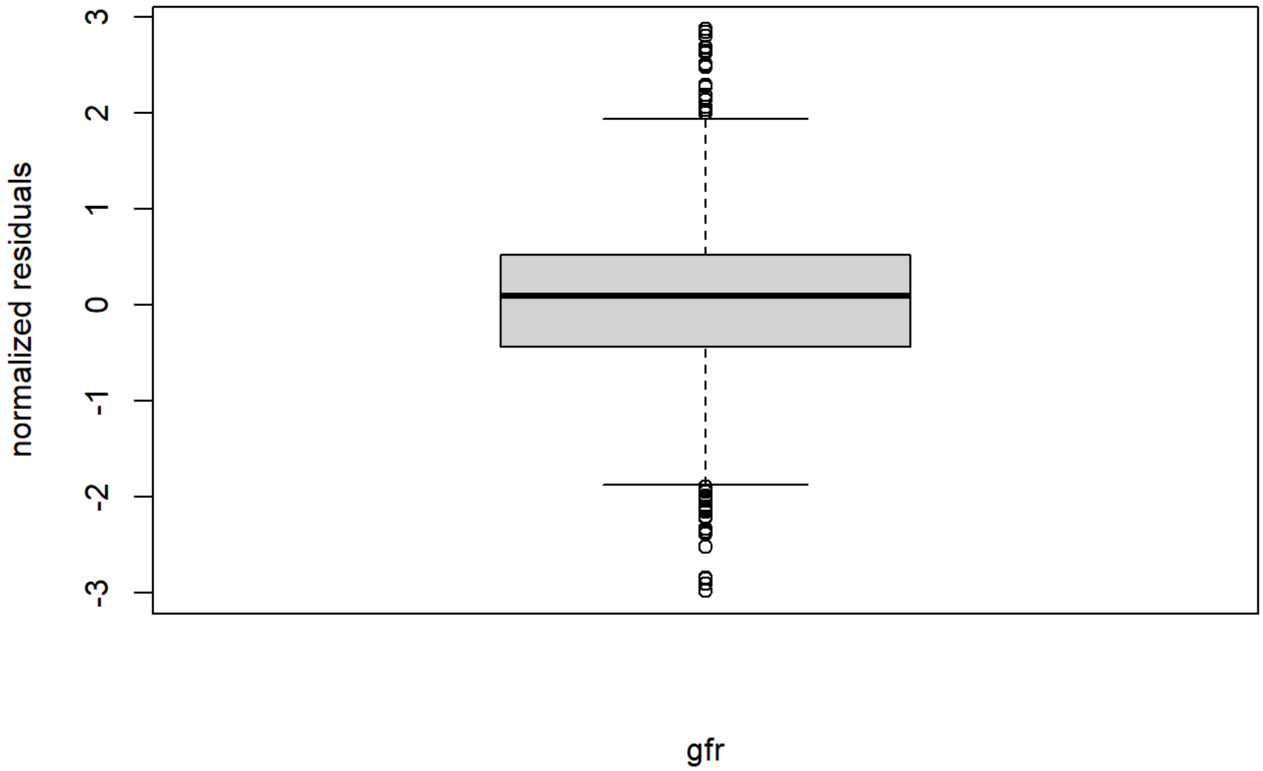


The above plot is the normlized residual of the log model vs gfr plot. To define the outliers, the concept of the value of normalized error is larger than 3 is taken as outliers.

```
filtered_normal_log_res <- normal_log_res[abs(normal_log_res)<3] # get rid off outliers
filtered_gfr <- gfr[abs(normal_log_res)<3] # get the gfr values
plot(filtered_gfr,filtered_normal_log_res,xlab=c("gfr"),ylab=c("normalized residuals")) # plot n
ew graph
```



```
boxplot(filtered_normal_log_res,xlab=c("gfr"),ylab=c("normalized residuals"))
```



##

(c):

For this part, the leverage for the observations will be calculated using the hatvalues function in r.

```
lever_log_mod <- hatvalues(log_mod_gfr_basc_ag_mal) # calculate the leverage of the model
max_lever_ind <- which.max(lever_log_mod) # get the index of the maximum leverage
pre_vals <- log_mod_gfr_basc_ag_mal$fitted.value # extract predicted values
max_pre_val <- pre_vals[max_lever_ind] # get the predicted values with the largest leverage
mean_pre_val <- mean(pre_vals) # get the mean of the predicted value
residuals <- log_mod_gfr_basc_ag_mal$residuals # retrieve the residuals
max_res_val <- residuals[max_lever_ind] # obtain the residuals

print("The predicted values with the highest leverege is:")
```

```
## [1] "The predicted values with the highest leverege is:"
```

```
max_pre_val
```

```
##     1185
## 3.942145
```

```
print("The mean of the predicted value is:")
```

```
## [1] "The mean of the predicted value is:"
```

```
mean_pre_val
```

```
## [1] 3.478452
```

```
print("The residual of the point is:")
```

```
## [1] "The residual of the point is:"
```

```
max_res_val
```

```
##       1185
## -0.4255377
```

Leverage measures how far the datapoint is to the mean of the dataset. Therefore, the point with the largest leverage is the point with the furthest distance to the mean of the data population. For this definition, it only states that the point is pretty far away from the mean of the entire population. However, it does not necessary mean that the different between the fitted values and the leverage point is big.

(d):

For this part, the influence of each observation of the model will be observed:

```
influences_log_mod <- cooks.distance(log_mod_gfr_basc_ag_mal) # calculate the influence of each observation
max_inf_ind <- which.max(influences_log_mod) # get the index of the maximum value of influence
new_data_test <- data_test[-c(max_inf_ind), ]
new_log_mod<-lm(log(gfr)~log(bascre)+age+male,data=new_data_test)
summary(new_log_mod)
```

```
##
## Call:
## lm(formula = log(gfr) ~ log(bascre) + age + male, data = new_data_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4495 -0.1910  0.0328  0.2152  1.5160
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.282712   0.057465  74.528   <2e-16 ***
## log(bascre) -1.247413   0.021761 -57.324   <2e-16 ***
## age         -0.001579   0.000921  -1.715   0.0867 .
## male1        0.220616   0.024227   9.106   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4118 on 1244 degrees of freedom
## Multiple R-squared:  0.7356, Adjusted R-squared:  0.735
## F-statistic:  1154 on 3 and 1244 DF,  p-value: < 2.2e-16
```

From removing the observation for the largest influence, the coefficient log(bascre) is changed but its sign remains the same. Therefore, although removing the observation cause some changes in model coefficients, it would not change the sign of the coefficients.