FINAL MARK:
46/52 ~ 88%

PHARM609 HOMEWORK2

Yaowen Mei (20470193)

March 11, 2016

Contents

```r
library(stats4)
library(MASS)
library(pander)
library(lattice)
library(survival)
library(Formula)
library(ggplot2)
library(Hmisc)
library(corrplot)
library(matrixStats)
library(varhandle)
# Load the dataset
dat <- read.csv("/Users/Chris/Dropbox/Master Program/lecture note/PHARM 609/Dataset/Carboplatin.csv")
# Change female, male into 0,1
dat$Sex<-factor(dat$Sex,labels = c('0','1'))
# Convert factor into numeric
dat$Sex<- unfactor(dat$Sex)
# Keep only the covariates and response variable
dat=subset(dat, select = -c(Patient,Dose,AUC.Carbo))
head(dat)
```

```
##   Age Sex Weight  BSA GFR CL.Carbo
## 1  62   0   14.0 0.65  88       54
## 2   3   1    4.7 0.25  21       11
## 3  18   0   10.6 0.47  26       20
## 4  47   0   16.4 0.65  55       79
## 5  49   0   21.6 0.86  96       65
## 6  14   1    8.9 0.42  35       36
```

Q1: 8.5/11

Question 1

Perform an exploratory analysis of Carboplatin Clearance as the response variable and potential predictors Sex, Age, Weight, BSA and GFR.

2/3

a. Describe and summarize the distribution of the variables involved in the study

22 patients data was collected, 11 of them were male, and 11 of them were female. The data shown that the age of the patients are varying from 3 to 190 months, and the weights and BSAs for patients are varying from 4.70 to 63.10 kg and 0.25 to 1.63 m2, respectively. The GFRs also have a wide range from 14 to 138 mL/min. Consequently, the carboplatin clearances shown a great variance, and changing from 11mL/min to 137 mL/min

```r
# calculate the mean of each variable
dat.mean=data.frame(colMeans(dat))
colnames(dat.mean) <- c("Mean")
# calculate the quantiles, median, and range
dat.quantiles = colQuantiles(as.matrix(dat),na.rm=TRUE)
colnames(dat.quantiles) <- c("Min","1st Qu.","Median","3rd Qu.","Max")
# calculate the variance
dat.var=as.data.frame(colVars(as.matrix(dat)))
colnames(dat.var) <- c("Variance")
```

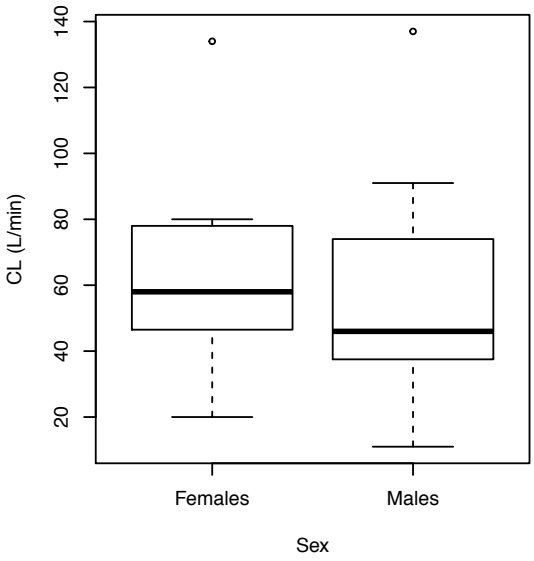
```r
rownames(dat.var) <- names(dat)
# output the summarize the distribution of the variables in this study
sdistri <- cbind(dat.mean,dat.quantiles,dat.var)
round(sdistri,2)
```

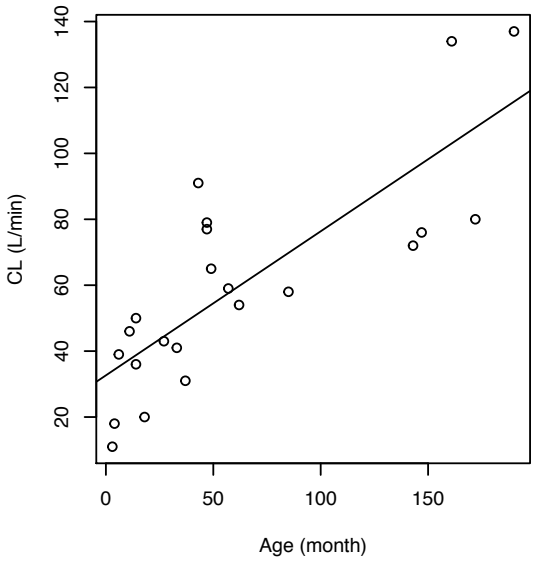
```
##           Mean   Min 1st Qu. Median 3rd Qu.    Max Variance
## Age      62.27  3.00   15.00  45.00   79.25 190.00  3591.45
## Sex       0.50  0.00    0.00   0.50    1.00   1.00     0.26
## Weight   21.21  4.70   10.00  16.00   30.40  63.10   228.08
## BSA       0.79  0.25    0.46   0.66    1.08   1.63     0.18
## GFR      64.41 14.00   30.50  58.00   89.50 138.00  1320.82
## CL.Carbo 59.86 11.00   39.50  56.00   76.75 137.00  1069.08
```

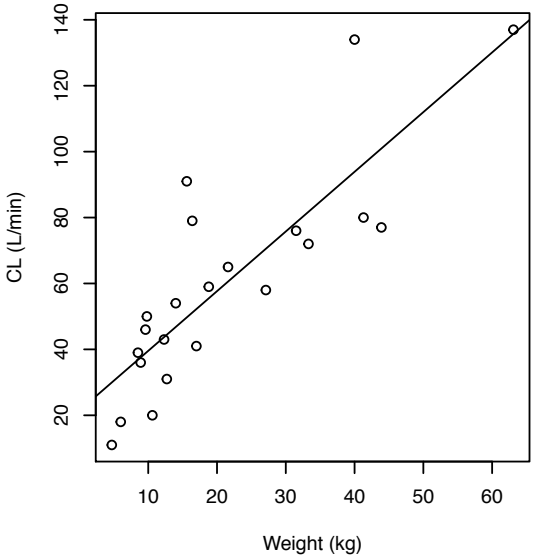
1/2

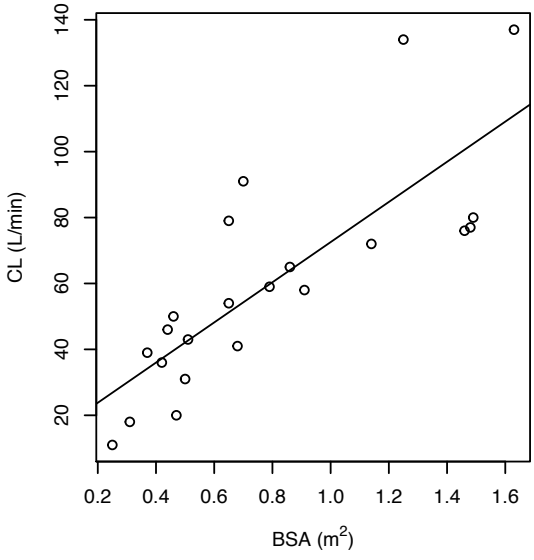
b. Provide a preliminarily assessment of the linear relationships between covariates and response

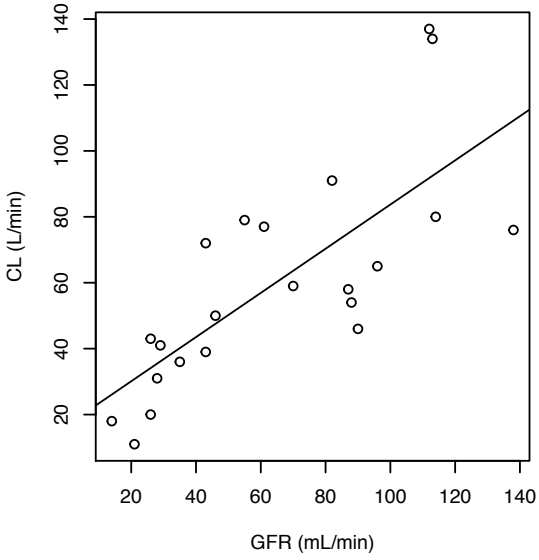
```r
par(mfrow=c(3,2))
par(mar=c(4,4,2,2))
attach(dat)
# Sex
boxplot(CL.Carbo~Sex,xlab="Sex",ylab="CL (L/min)", names=c("Females","Males"),data=dat)
# Age
plot(Age,CL.Carbo,ylim=range(CL.Carbo), xlab="Age (month)",ylab="CL (L/min)")
abline(lm(CL.Carbo~Age,data=dat))
# Weight
plot(Weight,CL.Carbo,ylim=range(CL.Carbo), xlab="Weight (kg)",ylab="CL (L/min)")
abline(lm(CL.Carbo~Weight,data=dat))
# BSA
plot(BSA,CL.Carbo,ylim=range(CL.Carbo), xlab=expression(paste("BSA (",m^2, ")")),ylab="CL (L/min)")
abline(lm(CL.Carbo~BSA,data=dat))
# GFR
plot(GFR,CL.Carbo,ylim=range(CL.Carbo), xlab="GFR (mL/min)",ylab="CL (L/min)")
abline(lm(CL.Carbo~GFR,data=dat))
```


CL (L/min)
20
40
60
80
100
120
140
Females
Males
Sex


CL (L/min)
20
40
60
80
100
120
140
0
50
100
150
Age (month)


CL (L/min)
20
40
60
80
100
120
140
10
20
30
40
50
60
Weight (kg)


CL (L/min)
20
40
60
80
100
120
140
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
BSA (m2)

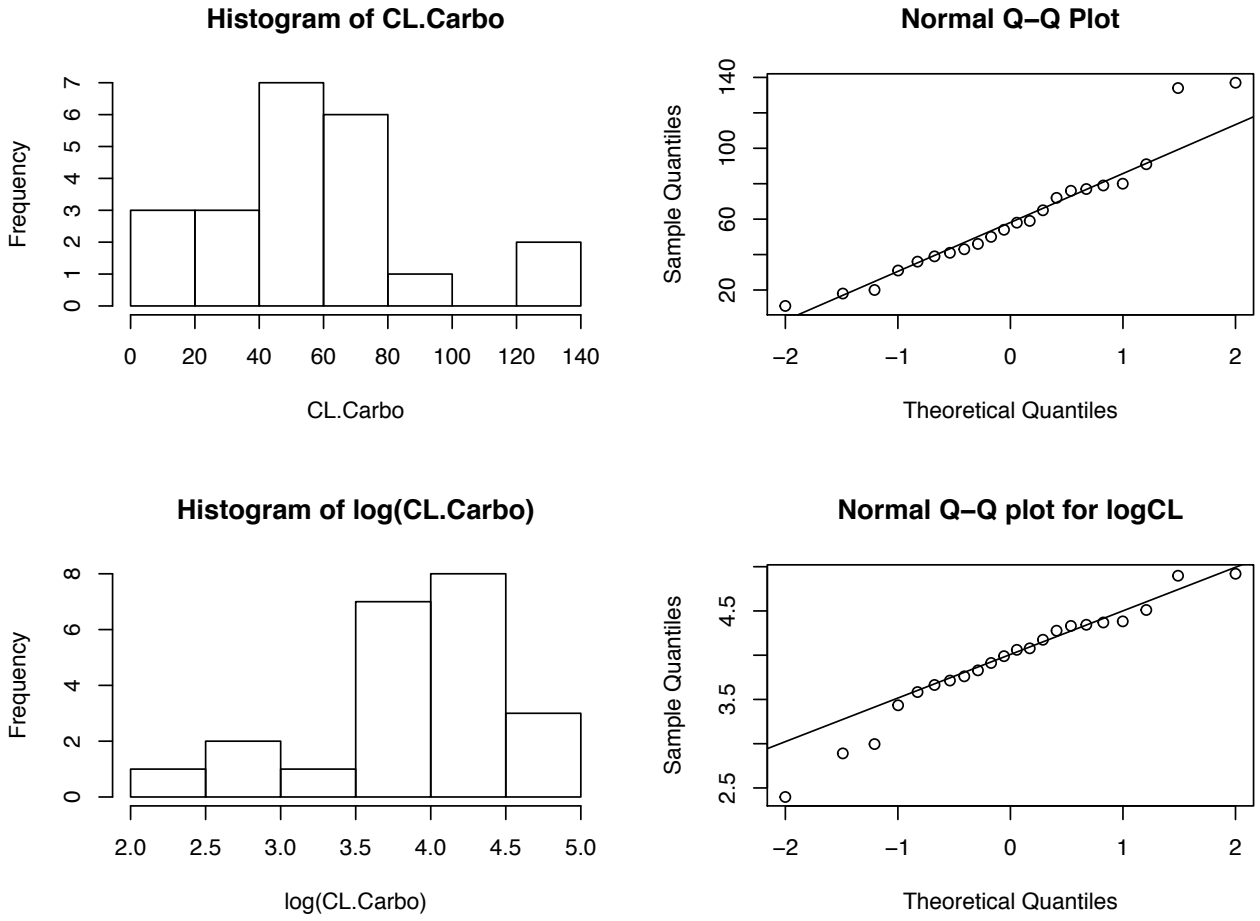

CL (L/min)
20
40
60
80
100
120
140
20
40
60
80
100
120
140
GFR (mL/min)

3/3

c. Assess whether the response variable may require a logarithmic transformation.

As we can see from the Fig below, the response variable, CL.Carbo, is already normally distributed, and the QQ plot behaves exactly as a straight line, so there is no need for us to apply a log transformation on the response variable. Actually, if we make a histogram and Normal Q-Q plot of the logCL.Carbo variable, we can see they are not well shaped (in the hist plot, the data is right skewed, and the median of logCL.Carbo is shifted to the right instead of in the middle; Also, in the QQ plot, the logCL.Carbo data is departure from normality, the head of the distribution seems much lower, and it no longer fit in the qqline).

```
par(mfrow=c(2,2))
hist(CL.Carbo)
qqnorm(CL.Carbo)
qqline(CL.Carbo)
hist(log(CL.Carbo))
qqnorm(log(CL.Carbo), main='Normal Q-Q plot for logCL')
qqline(log(CL.Carbo))
```



2.5/3

d. Assess whether any categorical variables may have an effect in the response.

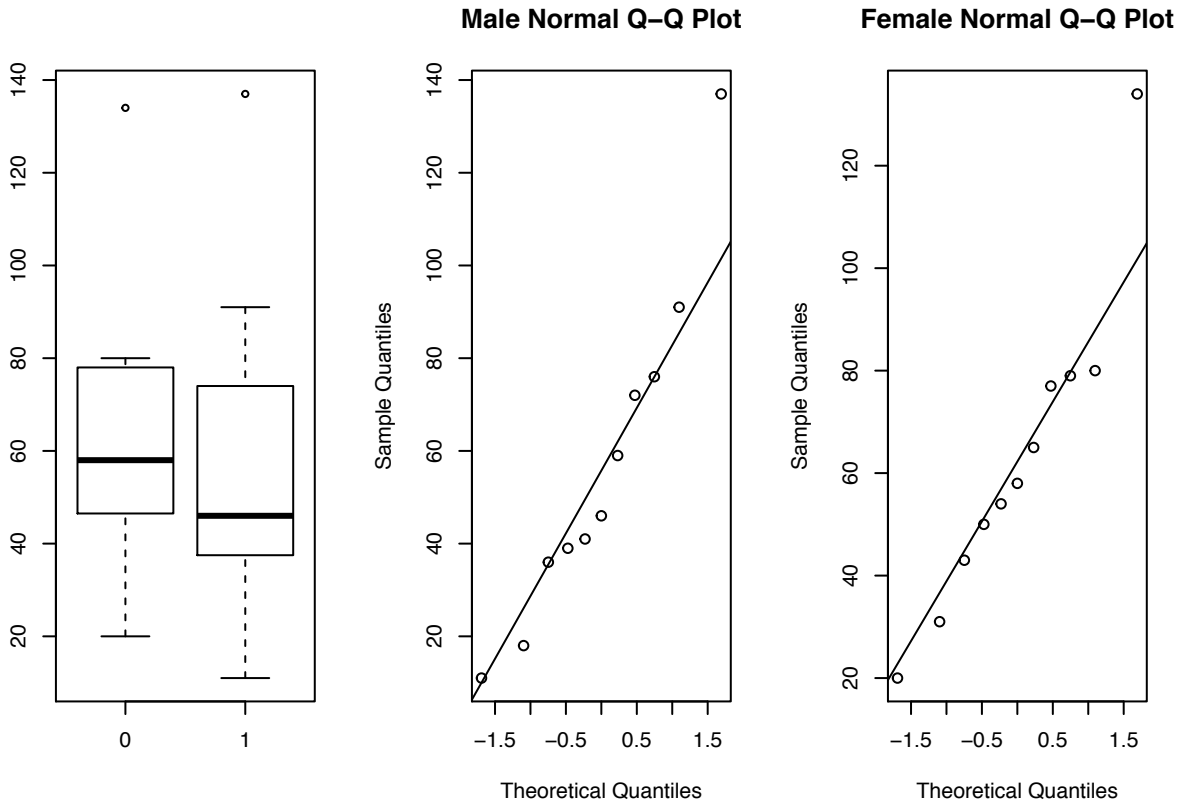
The only catergorical varibale in this dataset is the Sex variable. From the below box plot and the Q-Q plots, one can see that the CL.Carbon data for both male and female are approximately normally distributed, and both the female and male populations roughly have the same variance. To check whether Sex may have an

effect in the response, a hypothesis test should be conducted to see whether the CL.Carbo between female and male is different:

H0: the mean CL.Carbo in the female and male population are the same

H1: the mean CL.Carbo in the female and male population are not the same

```
par(mfrow=c(1,3))
boxplot(CL.Carbo~Sex)
qqnorm(CL.Carbo[Sex==1],main = "Male Normal Q-Q Plot")
qqline(CL.Carbo[Sex==1])
qqnorm(CL.Carbo[Sex==0],main = "Female Normal Q-Q Plot")
qqline(CL.Carbo[Sex==0])
```




```
t.test(CL.Carbo~Sex,var.equal=T,dat)
```

```
##
##  Two Sample t-test
##
## data:  CL.Carbo by Sex
## t = 0.4154, df = 20, p-value = 0.6823
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.76372  35.58191
## sample estimates:
## mean in group 0 mean in group 1
##        62.81818        56.90909
```

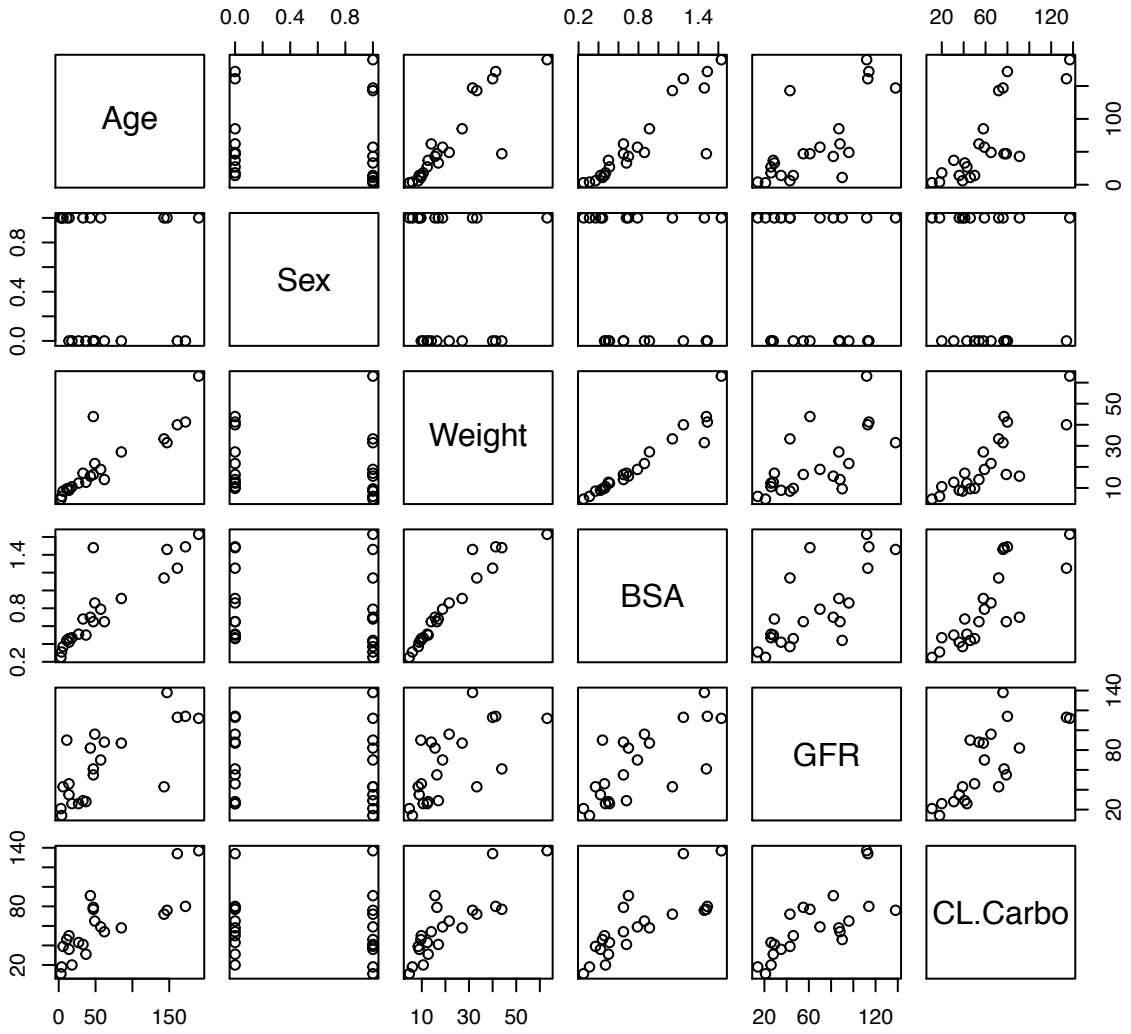
By using a 0.05 significance level, The p-value is 0.6823 » 0.05, therefore providing significant evidence to not reject H0. We then accept H0 and conclude that there is no statistically CL.Carbon difference between genders. Therefore, we conclude that sex has no effect in the response

Question 2

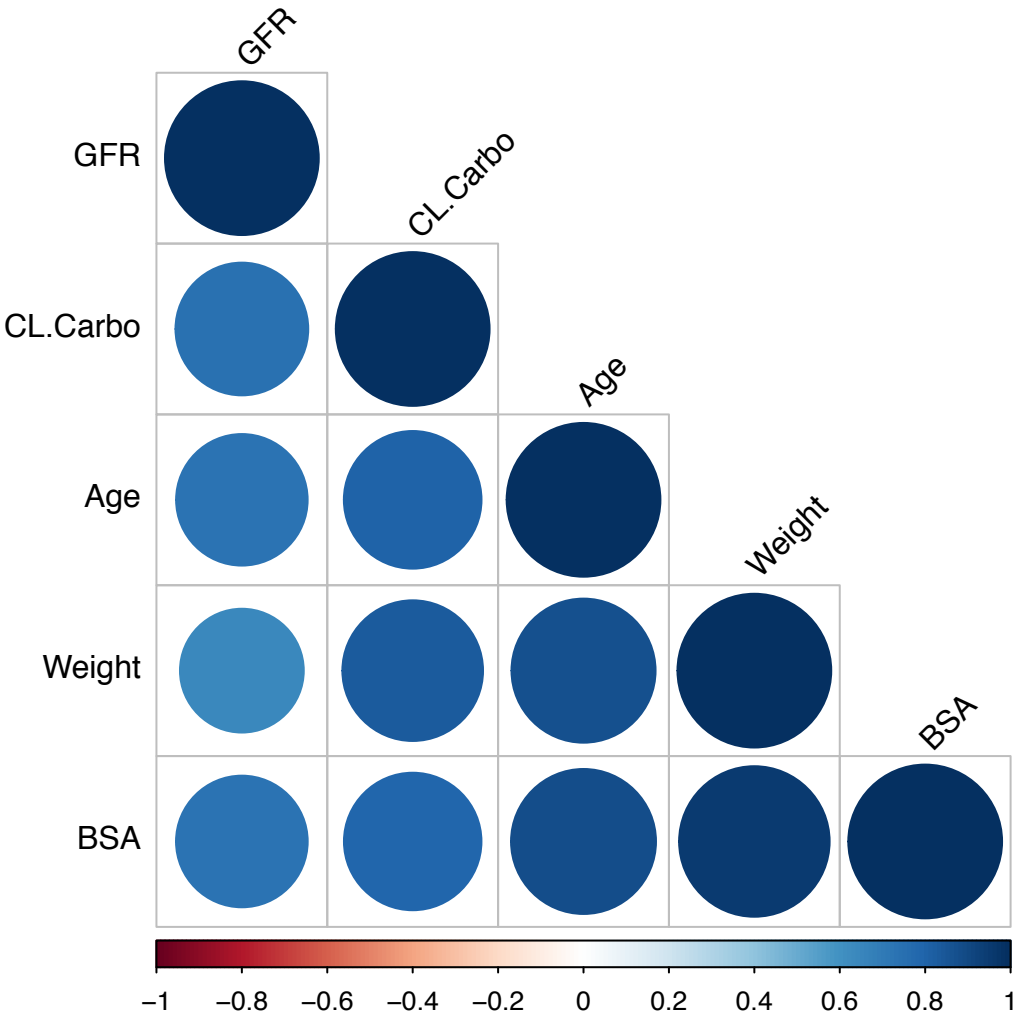
Q2: 1.5/2

Perform a pre-modeling collinearity assessment via correlation table for the continuous variables:

```
# Correlation plot
plot(dat)
```




```
mcor <- cor(subset(dat, select = -c(Sex)))
corrplot(mcor, type="lower", order="hclust", tl.col="black", tl.srt=45)
```




```
# Correlation table
cor_table = rcorr(as.matrix(subset(dat, select = -c(Sex)),type='pearson'))
cor_table
```

```
##            Age Weight  BSA  GFR CL.Carbo
## Age       1.00   0.88 0.89 0.73     0.80
## Weight    0.88   1.00 0.96 0.65     0.84
## BSA       0.89   0.96 1.00 0.73     0.80
## GFR       0.73   0.65 0.73 1.00     0.74
## CL.Carbo 0.80   0.84 0.80 0.74     1.00
##
## n= 22
##
##
## P
##           Age    Weight BSA    GFR    CL.Carbo
## Age              0.0000 0.0000 0.0001 0.0000
## Weight    0.0000        0.0000 0.0011 0.0000
## BSA       0.0000 0.0000        0.0001 0.0000
## GFR       0.0001 0.0011 0.0001        0.0000
## CL.Carbo 0.0000 0.0000 0.0000 0.0000
```

Table 1: The correlation table of covariables

	Age	Weight	BSA	GFR	CL.Carbo
Age	1	0.88	0.89	0.73	0.8
Weight	NA	1	0.96	0.65	0.84
BSA	NA	NA	1	0.73	0.8
GFR	NA	NA	NA	1	0.74
CL.Carbo	NA	NA	NA	NA	1

b. Identify the set of variables in (a) that may contribute to collinearity (if any)

By conducting Pearson type correlation and significance test in part (a), we set significance level to be 0.05, if p-val < 0.05 => we reject H0 => reject X,Y are not correlated => X,Y contribute to collinearity.

Therefore, as indicated in the correlation plot and correlation talbe in part (a), all the pairs of continuous covariables contribute to collinearity. Finally, the paris that contribute to collinearity are namely to be:

- Age & Weight, Age & BSA, Age & GFR
- Weight & BSA, Weight & GFR
- BSA & GFR

Q3: 6/8

Question 3

a. Perform individual simple linear fits for each covariate vs. Carboplatin Clearance.

1/1

```
# Sex (individual simple linear fits)
fit.sex <- lm(CL.Carbo ~ Sex, data=dat)
# Age (individual simple linear fits)
fit.age <- lm(CL.Carbo ~ Age, data=dat)
# Weight (individual simple linear fits)
fit.wt <- lm(CL.Carbo ~ Weight, data=dat)
# BSA (individual simple linear fits)
fit.bsa <- lm(CL.Carbo ~ BSA, data=dat)
# GFR (individual simple linear fits)
fit.gfr <- lm(CL.Carbo ~ GFR, data=dat)

# define my summary function that produce a list of summary for each induvidual simple linear fits
mysummaryfunction <- function(covariate){
  a<- summary(lm(CL.Carbo ~ covariate, data=dat))
  b<- a$coef[,c(1,2,4)]
  c<- a$r.squared
  d<- cbind(b,R2=c)
  # get the variable name as a string
  #vname <- deparse(substitute(covariate))[1]
  rownames(d)<- c("Intercept","Slope")
  return(d)
}
mylist <- as.list(subset(dat, select = -c(CL.Carbo)))
```

```
table_lm = lapply(mylist,"mysummaryfunction")
pander(table_lm,caption = 'The Tabel of simple linear regressions CL.Carbon vs. potential covariates')
```

- **Age**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	32.65	6.243	4.062e-05	0.6416
Slope	0.437	0.07304	7.513e-06	0.6416

- **Sex**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	62.82	10.06	4.244e-06	0.008554
Slope	-5.909	14.22	0.6823	0.008554

- **Weight**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	21.45	6.857	0.005295	0.6996
Slope	1.811	0.2654	1.239e-06	0.6996

- **BSA**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	11.63	9.175	0.2197	0.6389
Slope	60.92	10.24	8.104e-06	0.6389

- **GFR**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	16.7	9.87	0.1063	0.555
Slope	0.6702	0.1342	6.966e-05	0.555

1/2

b. Give the interpretation of the estimated slope when using the models with sex and age variables in (a).

1) For each 1 unit increase in Age(month), there is statistically significant 0.43701 increase in the mean Carboplatin Clearance (p-val = 7.51E-6 « 0.05)

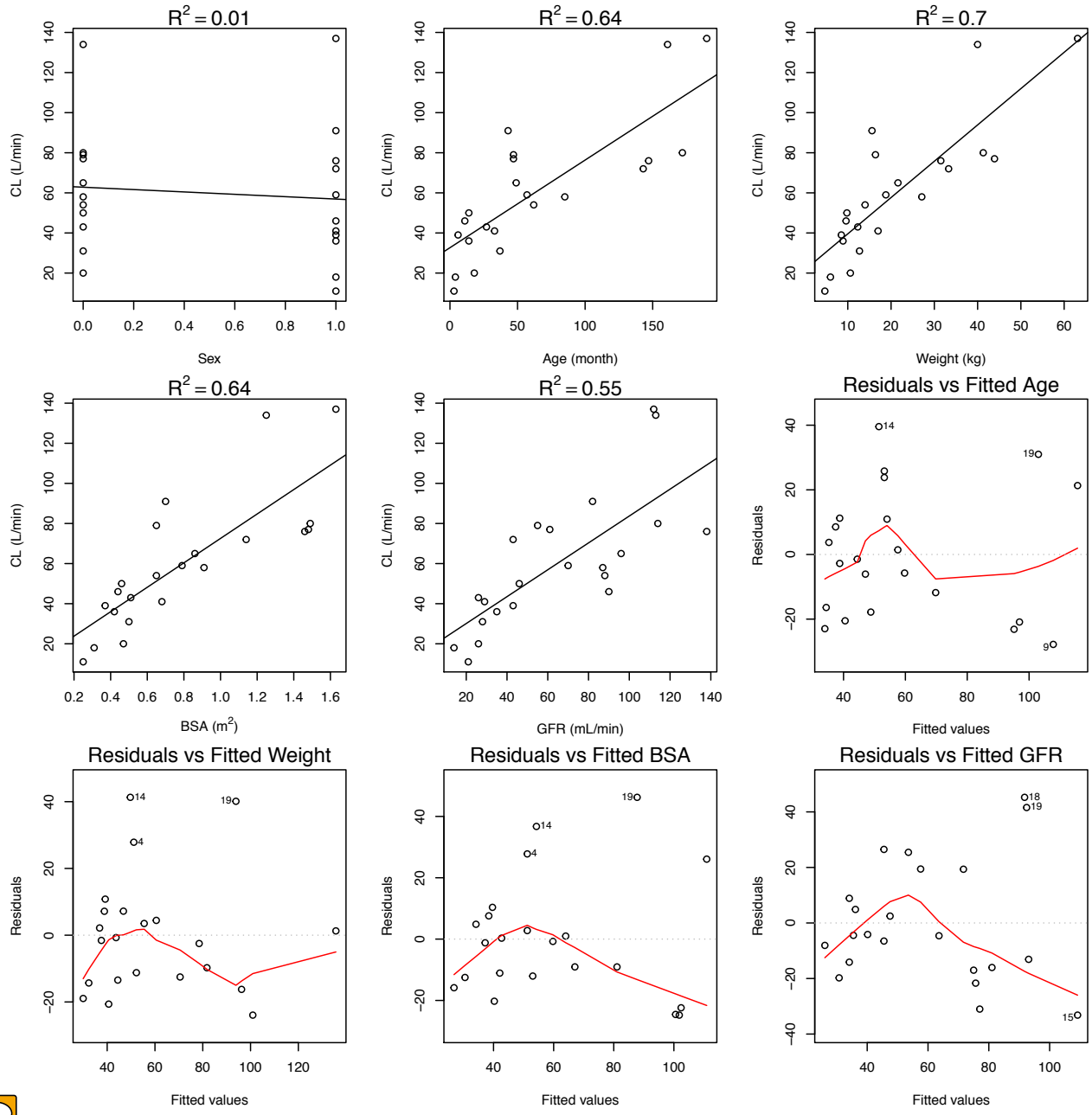
2) For the sex variables, the slope reported by R is negative (-5.9), actually, there is no slope (as we can see from Question 3 part (a)). The reason for this is that sex is a categorical variable and it has no effect in the Carboplatin Clearance as we have already proofed in Quetion 1 part (d). Therefore, the slope itself has no meaning.

2/3

c. Produce scatter plots for each variable in (a) adding the fitted line, the value of the coefficient of determination R2 and report any evidence of a relationship other than linear (e.g. curvilinear, no slope). Provide an interpretation of R2.

Hint: use the mtext() function to add text to the plots.

```
par(mfrow=c(4,3))
par(mar=c(4,4,2,2))
attach(dat)
plot(Sex,CL.Carbo,ylim=range(CL.Carbo), xlab="Sex",ylab="CL (L/min)")
abline(lm(CL.Carbo~Sex,data=dat))
mtext(expression(R^2==0.01))
plot(Age,CL.Carbo,ylim=range(CL.Carbo), xlab="Age (month)",ylab="CL (L/min)")
abline(lm(CL.Carbo~Age,data=dat))
mtext(expression(R^2==0.64))
plot(Weight,CL.Carbo,ylim=range(CL.Carbo), xlab="Weight (kg)",ylab="CL (L/min)")
abline(lm(CL.Carbo~Weight,data=dat))
mtext(expression(R^2==0.70))
plot(BSA,CL.Carbo,ylim=range(CL.Carbo), xlab=expression(paste("BSA (",m^2, ")")),ylab="CL (L/min)")
abline(lm(CL.Carbo~BSA,data=dat))
mtext(expression(R^2==0.64))
plot(GFR,CL.Carbo,ylim=range(CL.Carbo), xlab="GFR (mL/min)",ylab="CL (L/min)")
abline(lm(CL.Carbo~GFR,data=dat))
mtext(expression(R^2==0.55))
plot(fit.age,1,caption="Residuals vs Fitted Age")
plot(fit.wt,1,caption="Residuals vs Fitted Weight")
plot(fit.bsa,1,caption="Residuals vs Fitted BSA")
plot(fit.gfr,1,caption="Residuals vs Fitted GFR")
```



As we can see from the plot above, Sex has no linear realtionship with CL.Carbo (no slope).

Interpretation for R2: R2, coefficient of determination, is a measure of how close the regression line fits the raw data. R2 is ranging from 0 to 1, with 1 represent a perfect fit of the raw data. Generally speaking, the higher the R2, the better the model fits the data. However, R2 itself does not provide us the entire relationship between the model and the raw data, sometimes, high R2 values could even be obtained from a bad model.Therefore,R2 should be evaluated in conjunction with residual plots.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{\sum \left(Y_i - \bar{Y}_i\right)^2}$$

As we can see from the residual plots above, even though, the residuals shown a little bit un-linear trend (blow downward), but this is not very significant; therefore, we can still roughly consider the residuals are

almost normal distributed. The R2 value telling us that Weight is the best predictor with the strongest linear relationship with Cl.Carbon (R2=0.70), even though, other variables, Age (R2 = 0.64), BSA (R2=0.64), GFR(R2=0.55) also have a significant linear relationship with Cl.Carbon. The R2 value for Sex is 0.01, which indicates that Sex has no linear relationship with Cl.Carbon at all.

2/2

d. Fit a linear model for Carboplatin Clearance vs. Age, Weight and BSA. Does the conclusion regarding the significance of these variables agree with the corresponding individual simple linear fits performed in (a)? Support your answer.

Hint: look at the correlations between Age, Weight and BSA calculated in (2).

```
full.mod <- lm(CL.Carbo ~ Age + Weight + BSA, data=dat)
summary(full.mod)
```

```
##
## Call:
## lm(formula = CL.Carbo ~ Age + Weight + BSA, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.008 -14.269  -0.896   7.071  42.275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.1841    10.8385   2.508   0.0219 *
## Age           0.1839     0.1506   1.221   0.2378
## Weight        1.6274     0.9578   1.699   0.1065
## BSA         -16.7927    35.2151  -0.477   0.6392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.6 on 18 degrees of freedom
## Multiple R-squared:  0.7226, Adjusted R-squared:  0.6764
## F-statistic: 15.63 on 3 and 18 DF,  p-value: 2.969e-05
```

The p-val for these variables (Age, Weight, BSA) in this multiple regression mode are 0.2378, 0.1065, 0.6392, respectively. p-val $>>$ 0.05, these facts indicate that there are no statistical significant linear relationship between these two variables with Carbonplation clearance. This conclusion is contrasting with the conclusion we drawn from the simple linear fits in (a). Because in (a), each individual continues coverable shows a statistical significant linear relationship with Carbonplation clearance; however, variables that are expected to be important ($p < 0.0001$) in (a) are not found statistically significant ($P > 0.05$) in (d). Looking at Tabel below, one may noticed that this contrasting conclusion was caused by collinearity between Age, Weight, and BSA. A collinearity diagnostics should be conducted in order to get a better prediction for these data.

Table 7: The correlation talbe of continues coverables

Covariables	Weight	BSA	Age	GFR
Weight	1.00	0.96	0.88	0.67
BSA		1.00	0.89	0.73
Age			1.00	0.73
GFR				1.00

Q4: 2/2

Question 4

Find a regression equation that best fits Carboplatin Clearance by implementing the two stepwise selection methods given in class. Establish a level of significance to use throughout the process.

Whenever deemed applicable, please present tables with relevant statistical summaries to back up your decisions.

The significant level is set to be 0.05

Backward elimination

```
# The full model of backward elimination
backfull.mod <- lm(CL.Carbo~Weight + BSA + Age + GFR + Sex, data=dat)
summary(backfull.mod)
```

```
##
## Call:
## lm(formula = CL.Carbo ~ Weight + BSA + Age + GFR + Sex, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.975  -7.980  -2.340   5.211  34.400
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.80162   11.18626   1.949   0.0691 .
## Weight        2.42747    0.91689   2.648   0.0176 *
## BSA         -54.67498   35.04350  -1.560   0.1383
## Age           0.06633    0.14413   0.460   0.6516
## GFR           0.40811    0.16271   2.508   0.0233 *
## Sex          -1.11475    7.20676  -0.155   0.8790
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.7 on 16 degrees of freedom
## Multiple R-squared:  0.8013, Adjusted R-squared:  0.7392
## F-statistic: 12.91 on 5 and 16 DF,  p-value: 3.839e-05
```

```
# Eliminate Sex first (P-val = 0.87)
back.mod1 <- update(backfull.mod,.~. -Sex)
summary(back.mod1)
```

```
##
## Call:
## lm(formula = CL.Carbo ~ Weight + BSA + Age + GFR, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.428  -7.932  -2.353   4.696  33.825
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.03621    9.74002   2.160   0.0454 *
## Weight        2.42802    0.89017   2.728   0.0143 *
## BSA         -54.29605   33.93937  -1.600   0.1281
## Age           0.06402    0.13918   0.460   0.6513
## GFR           0.40873    0.15792   2.588   0.0191 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.21 on 17 degrees of freedom
## Multiple R-squared:  0.801,  Adjusted R-squared:  0.7542
## F-statistic: 17.11 on 4 and 17 DF,  p-value: 8.564e-06
```

```
# Eliminate Age then (P-val = 0.6513)
back.mod2 <- update(back.mod1,.~. -Age)
summary(back.mod2)
```

```
##
## Call:
## lm(formula = CL.Carbo ~ Weight + BSA + GFR, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.508  -7.113  -4.267   6.580  33.060
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.0223     8.5079   2.236  0.03827 *
## Weight        2.5480     0.8323   3.062  0.00672 **
## BSA         -51.8985    32.7941  -1.583  0.13093
## GFR           0.4329     0.1456   2.973  0.00815 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.85 on 18 degrees of freedom
## Multiple R-squared:  0.7985, Adjusted R-squared:  0.765
## F-statistic: 23.78 on 3 and 18 DF,  p-value: 1.743e-06
```

```
# Eliminate BSA finally (P-val =0.13093)
back.mod3 <- update(back.mod2,.~. -BSA)
summary(back.mod3)
```

```
##
## Call:
## lm(formula = CL.Carbo ~ Weight + GFR, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.982 -10.449  -2.962   5.884  34.049
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.5947     7.3718   1.573 0.132259
```

```
## Weight        1.3200     0.3125   4.224 0.000459 ***
## GFR           0.3147     0.1299   2.423 0.025535 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.47 on 19 degrees of freedom
## Multiple R-squared:  0.7705, Adjusted R-squared:  0.7463
## F-statistic: 31.89 on 2 and 19 DF,  p-value: 8.461e-07
```

$$CL.Carbon_i = \beta_0 + \beta_1 Weight_i + \beta_2 GFR_i + \varepsilon_i$$
$$CL.Carbon_i = 11.59 + 1.32 \times Weight_i + 0.31 \times GFR_i + \varepsilon_i$$

Forward selection

recall the tabel of simple linear regressions CL.Carbon vs. potential covariates, Weight should be choosen as the first predictor (p-val = 1.239e-06)

- **Age**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	32.65	6.243	4.062e-05	0.6416
Slope	0.437	0.07304	**7.513e-06**	0.6416

- **Sex**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	62.82	10.06	4.244e-06	0.008554
Slope	-5.909	14.22	0.6823	0.008554

- **Weight**:

	Estimate	Std. Error	Pr(>|t|)	R2adj
Intercept	21.45	6.857	0.005295	0.6996
Slope	1.811	0.2654	**1.239e-06**	0.6996

- **BSA**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	11.63	9.175	0.2197	0.6389
Slope	60.92	10.24	**8.104e-06**	0.6389

- **GFR**:

	Estimate	Std. Error	Pr(>|t|)	R2
Intercept	16.7	9.87	0.1063	0.555
Slope	0.6702	0.1342	**6.966e-05**	0.555

	Estimate	Std. Error	Pr(>|t|)	R2

Weight + remaining variable:

```
# Weight + remaining variable:
fit.sex1 <- lm(CL.Carbo ~ Weight + Sex, data=dat)
fit.age1 <- lm(CL.Carbo ~ Weight + Age, data=dat)
fit.bsa1 <-lm(CL.Carbo ~ Weight + BSA, data=dat)
fit.gfr1 <- lm(CL.Carbo ~ Weight + GFR, data=dat)

ests.w.weight <- rbind(
summary(fit.sex1)$coef[3,],
summary(fit.age1)$coef[3,],
summary(fit.bsa1)$coef[3,],
summary(fit.gfr1)$coef[3,])

R2.1 <- c(
summary(fit.sex1)$r.squared,
summary(fit.age1)$r.squared,
summary(fit.bsa1)$r.squared,
summary(fit.gfr1)$r.squared)

mat.ests.w.weight <- cbind(ests.w.weight,R2.1)
dimnames(mat.ests.w.weight)[[1]] <- c("Sex","Age","BSA","GFR")
mat.ests.w.weight <- round(mat.ests.w.weight,3)
mat.ests.w.weight
```

```
##     Estimate Std. Error t value Pr(>|t|)  R2.1
## Sex   -0.531      8.074  -0.066    0.948 0.700
## Age    0.159      0.138   1.149    0.265 0.719
## BSA   -1.883     33.455  -0.056    0.956 0.700
## GFR    0.315      0.130   2.423    0.026 0.770
```

Table 13: Weight + each remaining variable

	Estimate	Std. Error	t value	Pr(>|t|)	R2
Sex	-0.531	8.074	-0.066	0.948	0.7
Age	0.159	0.138	1.149	0.265	0.719
BSA	-1.883	33.45	-0.056	0.956	0.7
GFR	0.315	0.13	2.423	**0.026**	0.77

```
# Weight + GFR + remaining variable:
fit.sex2 <- lm(CL.Carbo ~ Weight + GFR + Sex, data=dat)
fit.age2 <- lm(CL.Carbo ~ Weight + GFR + Age, data=dat)
fit.bsa2 <-lm(CL.Carbo ~ Weight + GFR + BSA, data=dat)

ests.w.weight2 <- rbind(
summary(fit.sex2)$coef[4,],
summary(fit.age2)$coef[4,],
summary(fit.bsa2)$coef[4,])
```

```r
R2.2 <- c(
summary(fit.sex2)$r.squared,
summary(fit.age2)$r.squared,
summary(fit.bsa2)$r.squared)

mat.ests.w.weight2 <- cbind(ests.w.weight2,R2.2)
dimnames(mat.ests.w.weight2)[[1]] <- c("Sex","Age","BSA")
mat.ests.w.weight2 <- round(mat.ests.w.weight2,3)
mat.ests.w.weight2
```

```
##     Estimate Std. Error t value Pr(>|t|)  R2.2
## Sex   -0.185      7.252  -0.026    0.980 0.771
## Age    0.030      0.143   0.208    0.838 0.771
## BSA  -51.899     32.794  -1.583    0.131 0.799
```

Table 14: Weight + GFR + each remaining variable

	Estimate	Std. Error	t value	Pr(>|t|)	R2
Sex	-0.185	7.252	-0.026	0.98	0.771
Age	0.03	0.143	0.208	0.838	0.771
BSA	-51.9	32.79	-1.583	0.131	0.799

At a 5 % level of significance, there are no other variables that are significant

$$CL.Carbon_i = \beta_0 + \beta_1 Weight_i + \beta_2 GFR_i + \varepsilon_i$$

$$CL.Carbon_i = 11.59 + 1.32 \times Weight_i + 0.31 \times GFR_i + \varepsilon_i$$

Q5: 4/4

Question 5

1/1

a. If the models that resulted from stepwise selection methods in 3 are not the same, please select the most appropriate and explain why?

The model got from forward selection is the same as the one got from backwards elimination

2/2

b. Write the equation of the model selected in 5(a) with all of its components, including the underlying statistical assumptions.

$$CL.Carbon_i = 11.59 + 1.32 \times Weight_i + 0.31 \times GFR_i + \varepsilon_i$$

Underlying statistical assumptions are: 1) ε_i is normally distributed, uncorrelated with each other and have mean zero with variance σ^2; 2) the covariates are measured without error and they have a linear relationship with CL.Carbon

1/1

c. Write the matrix representation of the model and give the dimensions of the matrices and vectors.

$$\begin{bmatrix} CL.Carbon_1 \\ CL.Carbon_2 \\ \vdots \\ CL.Carbon_{22} \end{bmatrix}_{22\times 1} = \begin{bmatrix} 1 & Weight_1 & GFR_1 \\ 1 & Weight_2 & GFR_2 \\ \vdots & \vdots & \vdots \\ 1 & Weight_{22} & GFR_{22} \end{bmatrix}_{22\times 3} \begin{bmatrix} 11.59 \\ 1.32 \\ 0.31 \end{bmatrix}_{3\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{22} \end{bmatrix}_{22\times 1}$$

Q6: 13/14

Question 6

1/1

a. Perform a post-modeling collinearity assessment through the Variance Inflation Factor and the Condition Number. Comment on the results.

Variance inflation factor

$$Weight = \alpha_0 + \alpha_1 GFR + +\varepsilon$$

```
# Calculating Variance inflation factor
w.vif <- lm(Weight~GFR, data=dat)
vif.r2 <- summary(w.vif)$r.squared
myvif <- 1/(1-vif.r2)
myvif
```

```
## [1] 1.724696
```

If VIF > 5, there could be collinearity between covariables, if VIF > 10, there are almost sure to have collinearity between covariables. In this case, VIF $= 1.72$, which is far less than 5. So there is less likely to be collinearity in our model now

Condition Number

```
# Condition number (K) defined as the ratio of the largest to the
# smallest eigenvalues of the correlation matrix
mod.mat <- model.matrix(CL.Carbo~Weight+GFR,data=dat)
kappa(mod.mat)
```

```
## [1] 141.789
```

K $= 141.798$, by using Bonates guideline: $K < 10^4$ indicates no collinearity. Therefore, there is no collinearity in our model now

7/8

b. Perform a residual analysis and state the model assumptions that are to be checked in each plot, as well as whether the assumptions are being met.

Major Assumptions

1. The relationship between the response and the regressors is linear, at least approximately.
2. The error term has zero mean
3. The error term has constant variance (homoscedastic)

4. The errors are uncorrelated
5. The errors are normally distributed

The Quantile-Quantile Normal Residual Plot checked whether the errors are normally distributed, as all the points fitted in the straight line except for three potential outliers, and the hist plot shown a well shaped bell structure, assumption (5) is met.

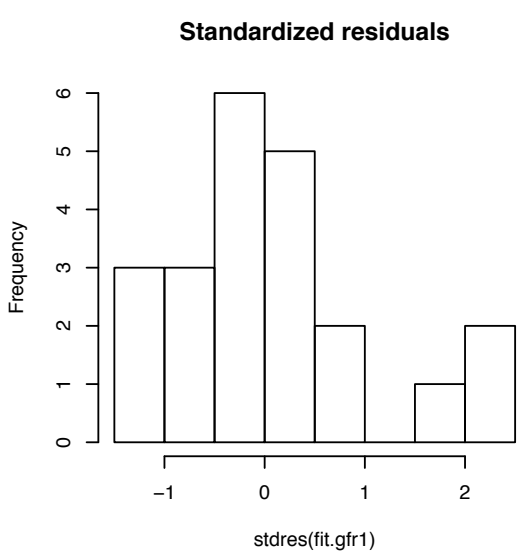
The plot of Residuals vs.fitted values checked whether the error term has a constant variance with a zero mean. As the residuals are contained in a horizontal band around zero, without showing specials patterns. Assumptions (2) and Assumptions (3) are met

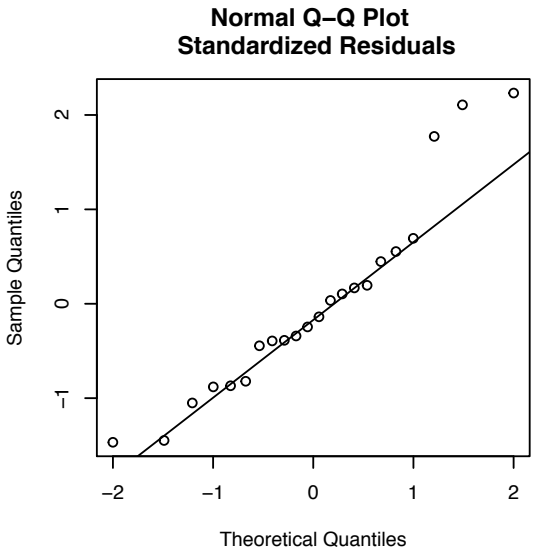
The plots of Residuals vs. Covariates are used to check for linearity and constant variance of CL.Carbo with Weight and GFR respectively. As the residuals are contained in a horizontal band around zero, without showing specials patterns, which indicateds that the residuals are not correlated, and the variables each has a linear realtionship with CL.Carbo. Assumptions (1), (2), (3) and (4) are met

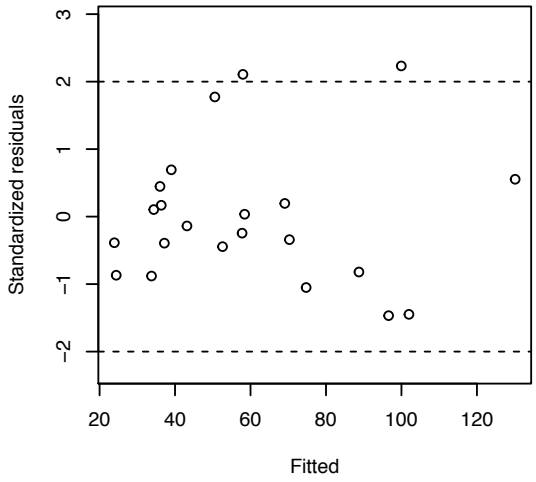
```
par(mfrow=c(3,2))
# Quantile-Quantile Normal Residual Plot
hist(stdres(fit.gfr1),main="Standardized residuals")
qqnorm(stdres(fit.gfr1),main="Normal Q-Q Plot \n Standardized Residuals")
qqline(stdres(fit.gfr1))

# Residuals vs. Fitted Values
plot(fit.gfr1$fitted,stdres(fit.gfr1),
    ylim=range(stdres(fit.gfr1)+c(-.8,.8)),
    xlab="Fitted",ylab="Standardized residuals")
abline(h=2,lty=2); abline(h=-2,lty=2)
stdres(fit.gfr1)[stdres(fit.gfr1)>2]
```

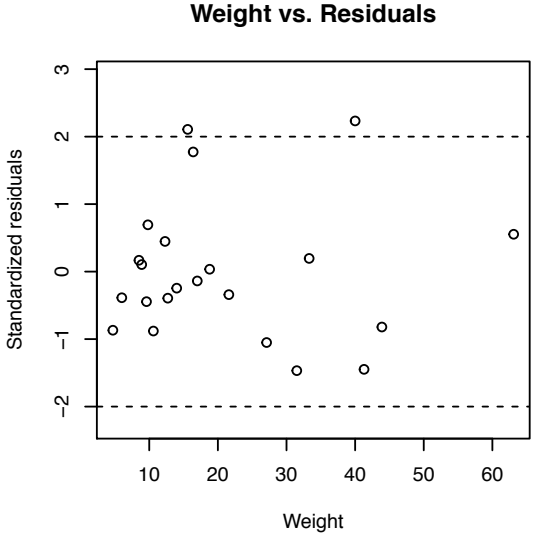
```
##       14       19
## 2.107417 2.232533
```

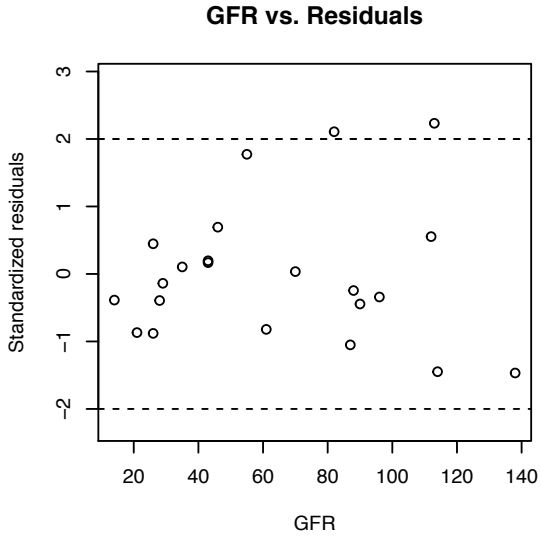
```
# Residuals vs. Covariates

plot(Weight,stdres(fit.gfr1),xlab="Weight",
               ylim=range(stdres(fit.gfr1)+c(-.8,.8)),
               ylab="Standardized residuals",main="Weight vs. Residuals")
abline(h=c(-2,2),lty=2)
plot(GFR,stdres(fit.gfr1),xlab="GFR",
                ylim=range(stdres(fit.gfr1)+c(-.8,.8)),
                ylab="Standardized residuals",main="GFR vs. Residuals")
abline(h=2,lty=2)
abline(h=c(-2,2),lty=2)
```


Standardized residuals
Frequency
0
1
2
3
4
5
6
−1
0
1
2
stdres(fit.gfr1)


Normal Q−Q Plot
Standardized Residuals
Sample Quantiles
−1
0
1
2
−2
−1
0
1
2
Theoretical Quantiles


Standardized residuals
−2
−1
0
1
2
3
20
40
60
80
100
120
Fitted


Weight vs. Residuals
Standardized residuals
−2
−1
0
1
2
3
10
20
30
40
50
60
Weight

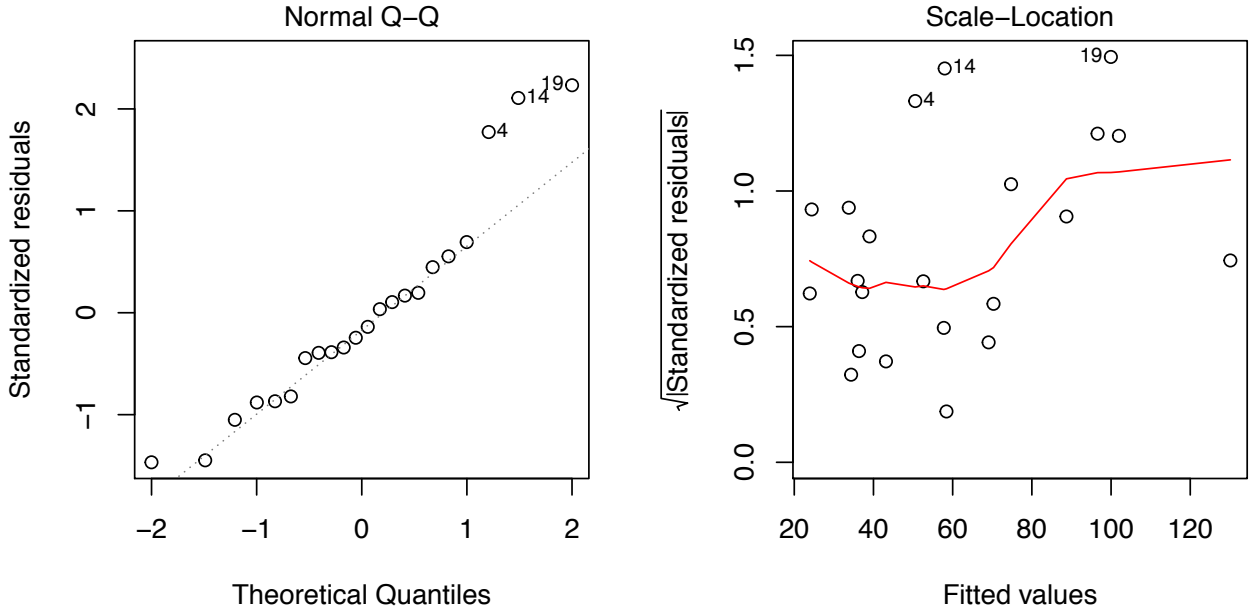

GFR vs. Residuals
Standardized residuals
−2
−1
0
1
2
3
20
40
60
80
100
120
140
GFR

5/5

c. Identify outliers through the analysis on the residuals in (b) and assess their influence. Are there any influential outliers? Are there any influential observations that are not outliers? If so, please give the patients number.

Detection of outliers

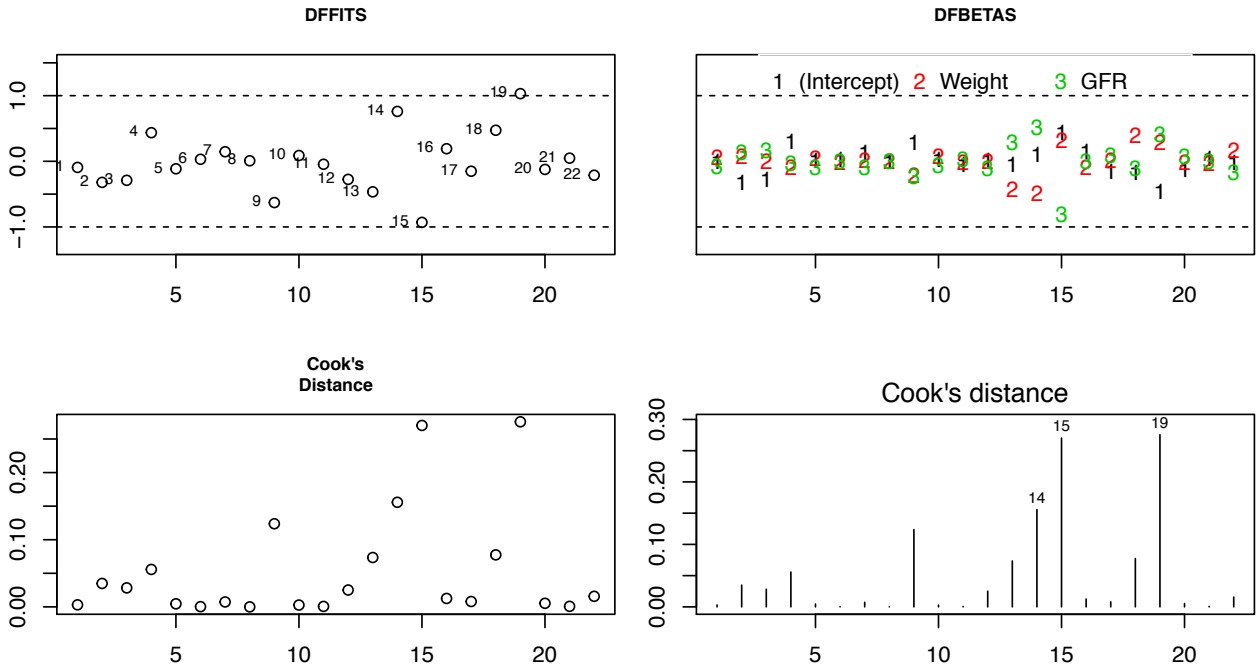
```
# Identify the outliers
# fit.gfr1 <- lm(CL.Carbo ~ Weight + GFR, data=dat)
par(mfrow=c(1,2))
plot(fit.gfr1,2)
plot(fit.gfr1,3)
```



Detection of Influential points

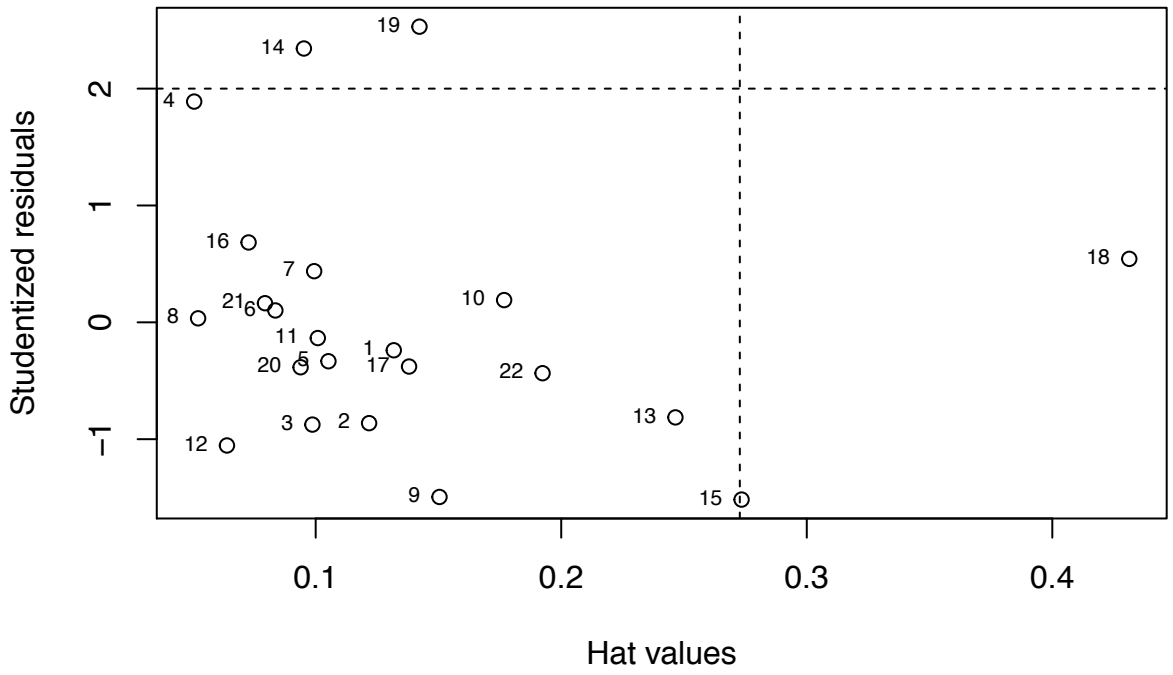
```
# Influence in the y-direction
par(mfrow=c(2,2), mai = c(.5, .5, 0.5, 0.01))

# plotting dffits +/-1
plot(dffits(fit.gfr1),
       main="DFFITS",ylab="",
       ylim=range(dfbetas(fit.gfr1))+c(-.5,1),cex.main=.8)
abline(h=-1,lty=2)
abline(h=1,lty=2)
text(dffits(fit.gfr1), labels=names(dffits(fit.gfr1)), cex= 0.7,pos=2)
#dffits(fit.gfr1)[dffits(fit.gfr1) >1]


# plotting dfbetas +/-1
matplot(dfbetas(fit.gfr1),
    main="DFBETAS",ylab="",
    ylim=range(dfbetas(fit.gfr1))+c(-.5,1),
    yaxt = "n",cex.main=.8)
legend("top",dimnames(dfbetas(fit.gfr1))[[2]],
      box.col="white",pch=c("1","2","3","4"),
```

```
      col=1:4,horiz=T)
abline(h=-1,lty=2)
abline(h=1,lty=2)
# cook's distance
plot(cooks.distance(fit.gfr1),main="Cook's
Distance",ylab="Cook's Distance",cex.main=.8)

# cook's distance
plot(fit.gfr1,4)
```



As the above plots shown, from the cook distance, points # 14, 15, 19 are possibly influential points. However, when comparing DFBETAS and DFFITS plots together, one might notice that points 14, 15, 19 are not significant influential points in DFBETAS plot, and point 19 and point 15 are the only significant influential point in DFFITS plot. By cosidering all these plots together, we conclude that point # 15 and 19 are the influential points in y-direction.

```
# Influence measure in the x-direction
hs <- hatvalues(fit.gfr1)

p <- 3
rule.thumb <- 2*p/nrow(dat)

# graph
plot(hs,studres(fit.gfr1),ylab="Studentized residuals",
                xlab="Hat values")
abline(h=-2,lty=2)
abline(h=2,lty=2)
abline(v=rule.thumb,lty=2)
text(hs, studres(fit.gfr1), labels=names(hs), cex= 0.7,pos=2)
```




```
# identify observations greater than 2p/n
#eval.hatvalues <- (hs>rule.thumb)*1
#hs[eval.hatvalues==1]

# points are possibly not influential
#studres(fit.gfr1)[studres(fit.gfr1)>2]
```

The hat value plot shows that point # 15 and point 18 are the influential points in x-direction, but points # 4, 14, 19 are possibly not influential at all.

```
new.dat <- dat[-c(19, 15), ]
fit.gfr2 <- lm(CL.Carbo ~ Weight + GFR, data=new.dat)
pander(summary(fit.gfr2)$coef , caption = 'The coefficient talbe for dataset deleting patient #15, 19')
```

Table 15: The coefficient talbe for dataset deleting patient #15, 19

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.98	6.878	1.742	0.09953
Weight	1.175	0.2792	4.209	0.0005901
GFR	0.3456	0.1319	2.62	0.01793

```
pander(summary(fit.gfr1)$coef , caption ='The coefficient talbe for the raw dataset')
```

Table 16: The coefficient talbe for the raw dataset

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.59	7.372	1.573	0.1323
Weight	1.32	0.3125	4.224	0.0004594

	Estimate	Std. Error	t value	Pr(>|t|)
GFR	0.3147	0.1299	2.423	0.02554

In summary, plots show that the outliers are patients # 14, 19. In y-direction, patients # 15, 19 are possibly influential points. In x-direction, patients # 15 and 18 are possibly influential points.

The patients # 19 is the influential outlier that have influence both on x and y-direction.

The patients # 15, and 18 are influential observations that are not outliers, and # 15 has influence both on x and y-directon

As the two tables indicated above, by deleting patients # 15 and 19, we improved the p-val for intercept from 0.132 to 0.099, the p-val for GFR coefficient from 0.025 to 0.017.

O7: 4/4

Question 7

1/1

a. Fit a model for the logarithm of Carboplatin Clearance with Weight and BSA as covariates.

```
# Append the dataframe with an additional column 'logCL'
logCL <- log(dat$CL.Carbo)
dat$logCL <- logCL
# Fit logCL with Weight and BSA
fit.wtbsa <- lm(logCL~Weight + BSA, data=dat)
pander(summary(fit.wtbsa), caption='Fit logCL with Weight and BSA')
```

	Estimate	Std. Error	t value	Pr(>|t|)
Weight	0.007074	0.02078	0.3405	0.7372
BSA	0.8742	0.7315	1.195	0.2467
(Intercept)	3.09	0.2197	14.07	1.696e-11

Table 18: Fit logCL with Weight and BSA

Observations	Residual Std. Error	R^2	Adjusted R^2
22	0.4119	0.5985	0.5562

2/2

b. Provide the interpretation of the estimated coefficients for Weight and BSA on the response under this model.

$$\ln(CL) = 3.09 + 0.007 \times Weight + 0.874 \times BSA$$

$$CL = 21.977 \times e^{0.007 Weight} e^{0.874 BSA}$$

ON THE EFFECT OF WEIGHT

The estimated coefficient of Weight is $\beta = 0.007$ so we would say that:On average and while holding BSA constant, a one unit increase in Weight would result in a increase of $e^{0.007} = 1.007$ times the value in CL (p-val=0.7372).

On average and while holding BSA constant, a one unit increase in Weight would result in a 0.7 % percent increase in CL (p-val=0.7372).

$$\left(e^{0.007} - 1\right) \times 100\% = 0.7\%$$

ON THE EFFECT OF BSA

On average and while holding Weight constant, a one unit increase in BSA would result in a increase of $e^{0.847}$ = 2.333 times the value of CL (p-val=0.2467).

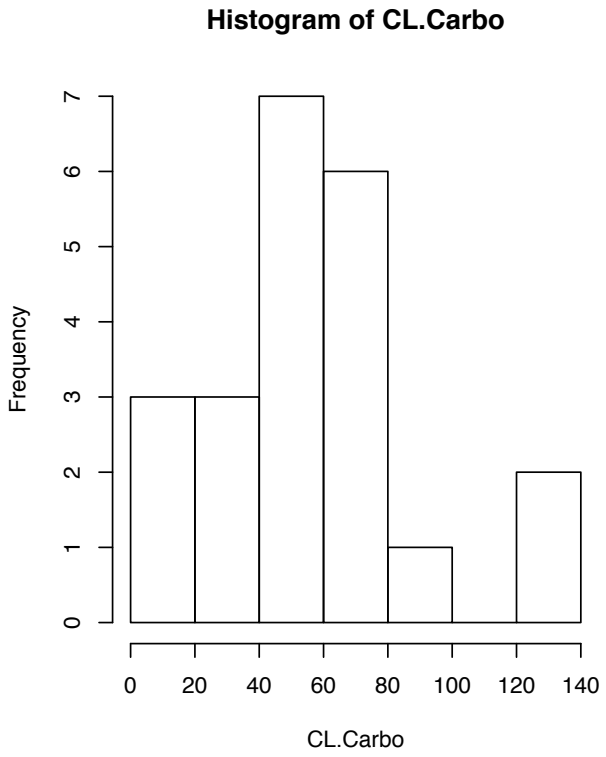
On average and while holding Weight constant, a one unit increase in BSA would result in a 133.26% increase in CL (p-val=0.2467), since:

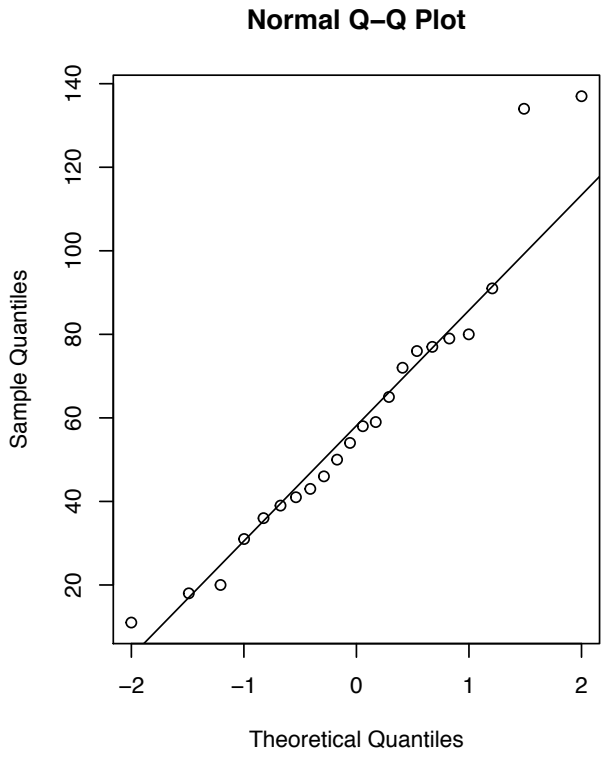
$$\left(e^{0.874} - 1\right) \times 100\% = 133.26\%$$

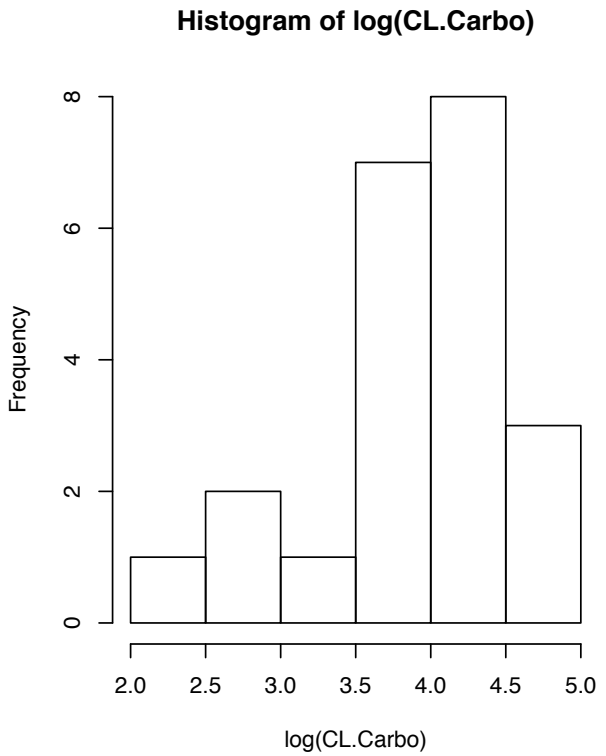
1/1

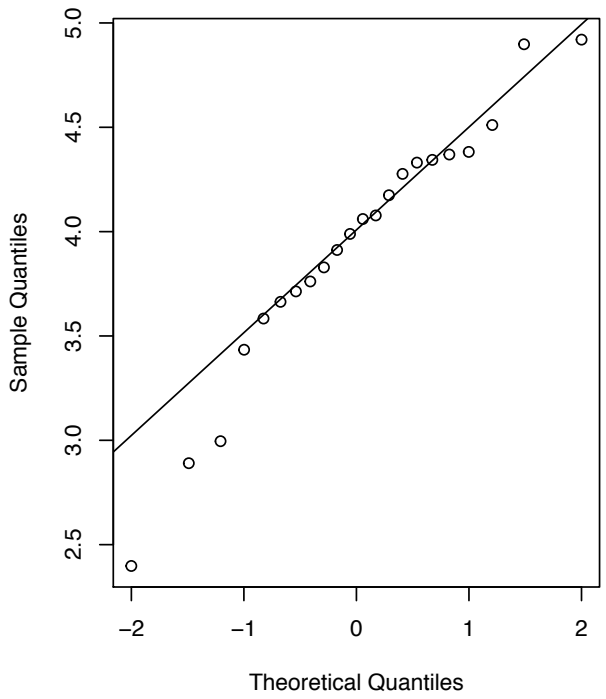
c. Construct a Normal Quantile plot (add the straight line as visual aid) and histogram for the transformed and untransformed variables. Compare the plots for the two and comment on the effect of transforming the data. Do you think the transformation works for these data?

```
par(mfrow=c(2,2))
hist(CL.Carbo)
qqnorm(CL.Carbo)
qqline(CL.Carbo)
#text(qqnorm(CL.Carbo), labels=names(CL.Carbo), cex= 0.7,pos=2)
hist(log(CL.Carbo))
qqnorm(log(CL.Carbo), main='Normal Q-Q plot for logCL')
qqline(log(CL.Carbo))
```









I think transformation does not work for these data. As we can see from the Fig above: CL.Carbo, is already normally distributed, and the Q-Q plot behaves exactly as a straight line, except for the two points in the heavy tail (these two points are probably outliers and should be delete from the dataset). Actually, if we make a histogram and Normal Q-Q plot of the logCL.Carbo, we can see the hist plot is not well shaped (the

data is right skewed, and the median of logCL.Carbo is shifted to the right instead of in the middle), and also, in the QQ plot, the logCL.Carbo data is departure from normality, the outliers in the tail are now in the line, but the head of the distribution seems much lower).

Q8: 3/3

Question 8

1/1

a. Fit a model for the Carboplatin Clearance with Weight centered to its mean and BSA as covariates.

```
m.Weight <- mean(Weight)
dat$cent.Weight <- Weight-m.Weight
fit.cwtbsa <- lm(CL.Carbo~cent.Weight + BSA, data=dat)
pander(summary(fit.cwtbsa),caption = 'Table of coefficients of CL with centered Weight and BSA')
```

	Estimate	Std. Error	t value	Pr(>|t|)
cent.Weight	1.862	0.9503	1.959	0.06491
BSA	-1.883	33.46	-0.05628	0.9557
(Intercept)	61.35	26.79	2.29	0.03364

Table 20: Table of coefficients of CL with centered Weight and BSA

Observations	Residual Std. Error	R^2	Adjusted R^2
22	18.84	0.6996	0.668

```
pander(summary(lm(CL.Carbo~ Weight + BSA, data=dat)),caption = 'Table of coefficients of CL with Weight
```

	Estimate	Std. Error	t value	Pr(>|t|)
Weight	1.862	0.9503	1.959	0.06491
BSA	-1.883	33.46	-0.05628	0.9557
(Intercept)	21.85	10.05	2.175	0.04247

Table 22: Table of coefficients of CL with Weight and BSA

Observations	Residual Std. Error	R^2	Adjusted R^2
22	18.84	0.6996	0.668

2/2

b. Interpret the value of the estimated slope.

$$CL = 61.35 + 1.862 \times (Weight - 21.21) - 1.883 \times BSA$$

ON THE EFFECT OF WEIGHT

On average and while holding BSA constant, a one unit increase in Weight would result in a 1.862 increase of the value in CL (p-val=0.06491).

ON THE EFFFECT OF BSA

On average and while holding Weight constant, a one unit increase in BSA would result in a 1.833 decrease of the value of CL (p-val=0.955).

Q9:4 /4

Question 9

1/1

a. Explain in your own words the Maximum Likelihood Estimation method.

Generally speaking, there are two methods for estimating statistical parameters, one is the Least Square estimation method, and the other is the Maximum Likelihood Estimation method. Maximum Likelihood Estimation method, so-called the MLE method, is a approach to get estimated statistical parameters based on a set of data that we already have. In another word, by approaching the Maximum Likelihood Estimation method, we are trying to get a unknown distribution from analysing a random sample from it.

3/3

b. Estimate the coefficients β_0, β_1 and σ^2 of the linear model of Carboplatin Clearance vs. Weight and BSA using the Maximum Likelihood method through the mle() function (library(stats4)). Provide 2 sets of starting values. Please comment.

```
# Define my Maximum Likelihood method functon
LL.reg.fun <- function (b0,b1,b2,sigma){
  #this function calculates the ???log Likelihood for
  # the normal residuals of a linear regression
  # CL.Carbon vs. Weight and BSA
  # it takes as argument a list with four components:
  # list=(b0,b1,b2,sigma)
  var <- dat$CL.Carbo - b0 - b1 * dat$Weight - b2 * dat$BSA
  pdf <- dnorm(var, 0, sigma)
  -sum(log(pdf))
}
# First Starting value (21,1,-1,18):
fit.mle1 <- mle(LL.reg.fun , start=list(b0=21,b1=1,b2=-1,sigma=18),
                method="L-BFGS-B",lower=c(-Inf,-Inf,-Inf,0),upper=c(Inf,Inf,Inf,Inf))
summary(fit.mle1)
```

```
## Maximum likelihood estimation
##
## Call:
## mle(minuslogl = LL.reg.fun, start = list(b0 = 21, b1 = 1, b2 = -1,
##     sigma = 18), method = "L-BFGS-B", lower = c(-Inf, -Inf, -Inf,
##     0), upper = c(Inf, Inf, Inf, Inf))
##
## Coefficients:
##        Estimate Std. Error
## b0    21.857662  9.3379577
## b1     1.862502  0.8831919
```

```
## b2    -1.899023 31.0911590
## sigma 17.508548  2.6396093
##
## -2 log L: 188.3906
```

```
# Second Starting value (1,-1,18,21)
fit.mle2 <- mle(LL.reg.fun , start=list(b0=-100,b1=-100,b2=100,sigma=1000))
```

```
## Warning in dnorm(var, 0, sigma): NaNs produced

## Warning in dnorm(var, 0, sigma): NaNs produced

## Warning in dnorm(var, 0, sigma): NaNs produced

## Warning in dnorm(var, 0, sigma): NaNs produced

## Warning in dnorm(var, 0, sigma): NaNs produced

## Warning in dnorm(var, 0, sigma): NaNs produced
```

```
summary(fit.mle2)
```

```
## Maximum likelihood estimation
##
## Call:
## mle(minuslogl = LL.reg.fun, start = list(b0 = -100, b1 = -100,
##     b2 = 100, sigma = 1000))
##
## Coefficients:
##         Estimate Std. Error
## b0     7.7934195  13.030884
## b1     0.1120244   1.335433
## b2    60.6722476  47.266531
## sigma 21.3306124   4.820357
##
## -2 log L: 192.7844
```

```
# Comparing with the coefficients provided by lm() method
summary(lm(CL.Carbo~ Weight + BSA, data=dat))$coef
```

```
##              Estimate Std. Error   t value   Pr(>|t|)
## (Intercept) 21.853297  10.047923  2.174907 0.04247198
## Weight       1.862070   0.950341  1.959370 0.06490770
## BSA         -1.882916  33.455028 -0.056282 0.95570482
```

```
summary(lm(CL.Carbo~ Weight + BSA, data=dat))$sigma
```

```
## [1] 18.83971
```

As we can see from the previous code, if we give the mle() function a good starting value, the coefficients calculated by mle() is pretty much as same as the coefficients provided by lm() function. However, if a really

weird staring value is passed to mle(), the output can be relatively off from what it should be. These facts indicate that the estimates we got from the Maximum Likelihood Estimation method are not very stable, and the outputs are really sensitive to the initial value. In order to get the best result from mle(), careful consideration should be given when choosing the staring value.