FE 541 Midterm Exam

#Problem 1 (20 points) The following source, especially the R-codes associated with Section 8.2 IC: Binary Logistic Regression, would be useful:
http://gattonweb.uky.edu/sheather/book/r_code.php

When you are to use "mmps(…, key=…)," use key=TRUE, instead of key=NULL. Data on 102 male and 100 female athletes were collected at the Australian Institute of Sport. The data are available on Canvas in the file ais.txt. Develop a logistic regression model for gender (y = 1 corresponds to female) or (y = 0 corresponds to male) based on the following predictors (which is a subset of those available): RCC: red cell count, WCC: white cell count, BMI: body mass index

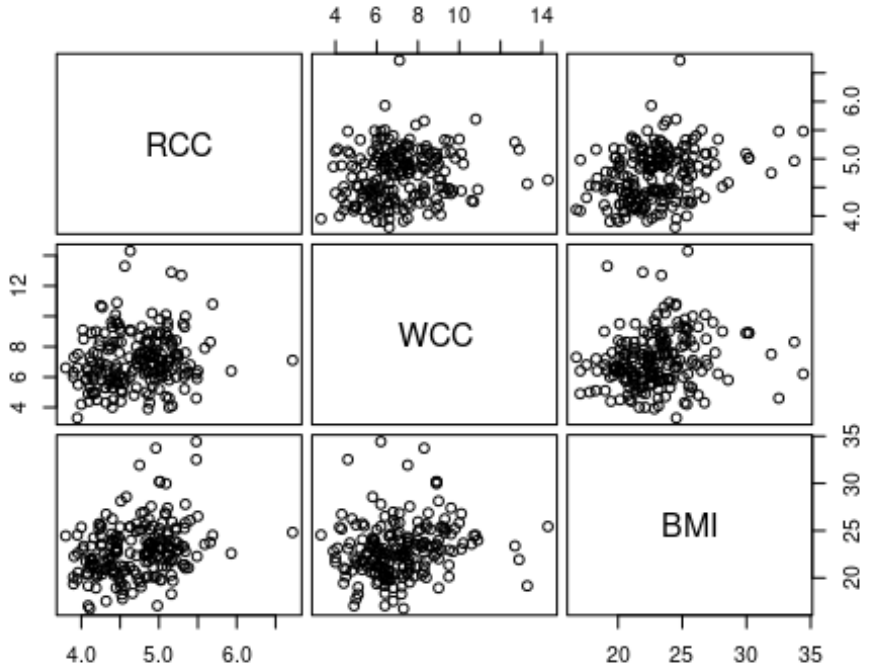
Hint from the lecture note on Chapter 8: When conducting a binary logistic regression with a skewed predictor x, it is often easiest to assess the need for x and log(x) by including them both in the model so that their relative contributions can be assessed directly.

```
Atheletes <- read.table(url("https://gattonweb.uky.edu/sheather/book/docs/datasets/ais.txt"), header=TRUE)
```

y=1 corresponds to female y=0 corresponst to male

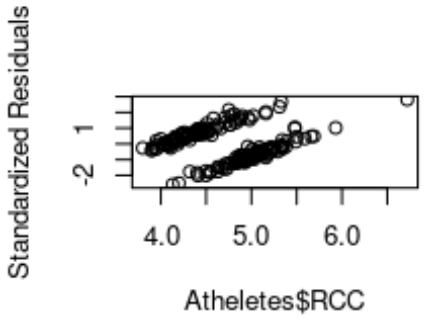
RCC= Red Cell Count (RCC) WCC= White Cell Count (WCC) BMI= Body Mass Index (BMI)

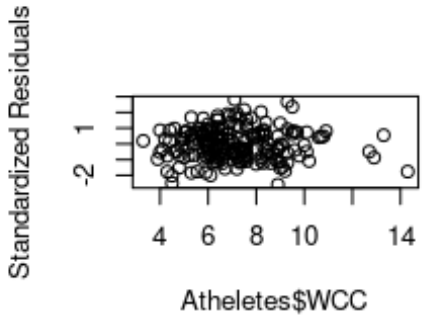
```
pairs(~RCC+WCC+BMI,data=Atheletes,gap=0.4,cex.labels=1.5)
```

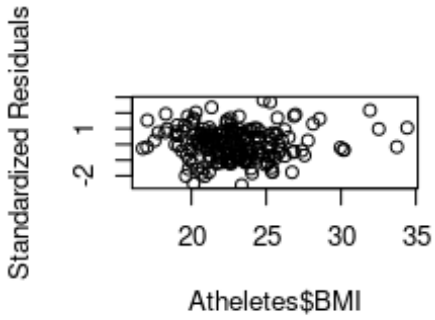



```
m1 <- lm(Sex~RCC+WCC+BMI, Atheletes)
summary(m1)

##
## Call:
## lm(formula = Sex ~ RCC + WCC + BMI, data = Atheletes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92809 -0.26017 -0.01509  0.25649  0.96455
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.316644   0.298901  14.442  < 2e-16 ***
## RCC         -0.704476   0.058758 -11.990  < 2e-16 ***
## WCC          0.016283   0.014488   1.124  0.26241
## BMI         -0.026712   0.009443  -2.829  0.00515 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3622 on 198 degrees of freedom
## Multiple R-squared:  0.4857, Adjusted R-squared:  0.4779
## F-statistic: 62.32 on 3 and 198 DF,  p-value: < 2.2e-16

StanRes1 <- rstandard(m1)
par(mfrow=c(2,2))
plot(Atheletes$RCC,StanRes1, ylab="Standardized Residuals")
```

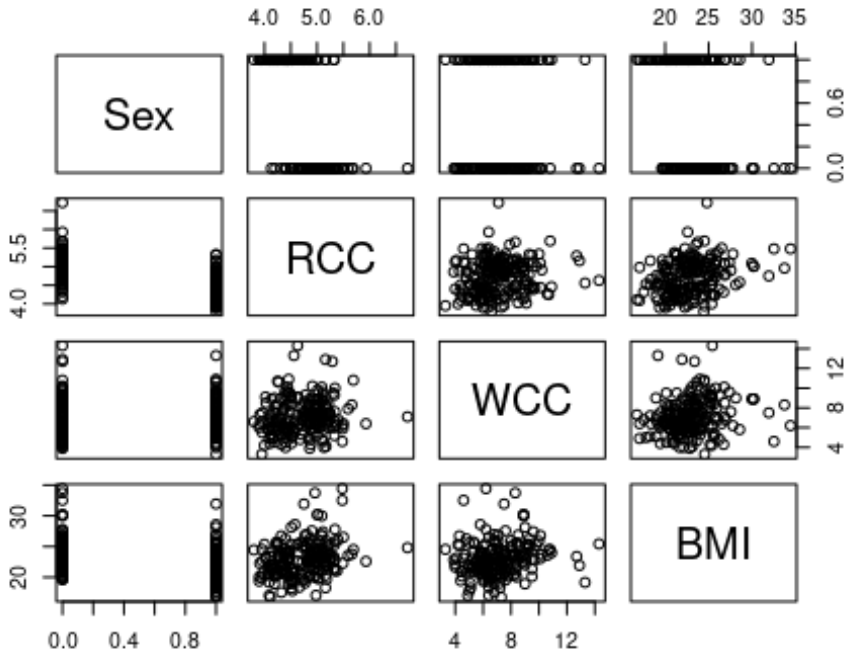
```
plot(Atheletes$WCC,StanRes1, ylab="Standardized Residuals")
plot(Atheletes$BMI,StanRes1, ylab="Standardized Residuals")
```



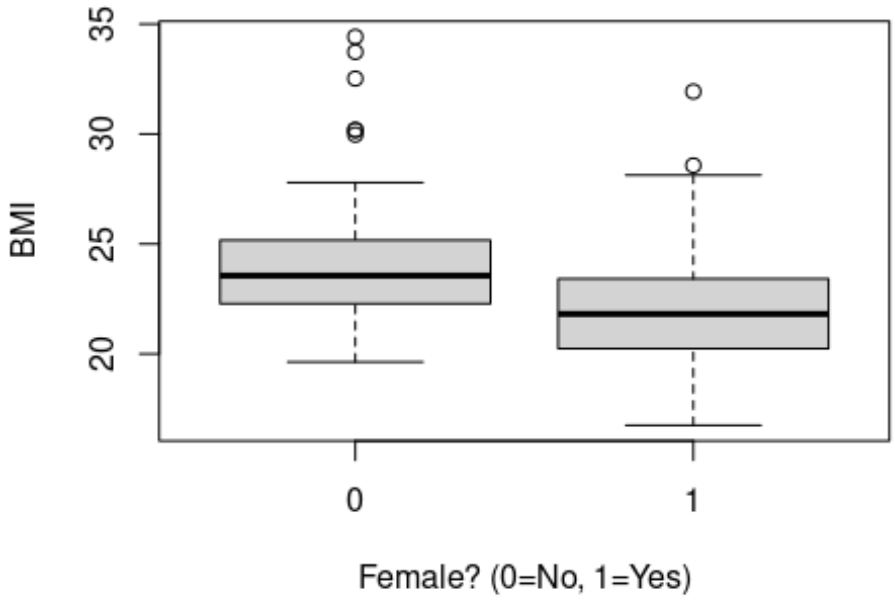




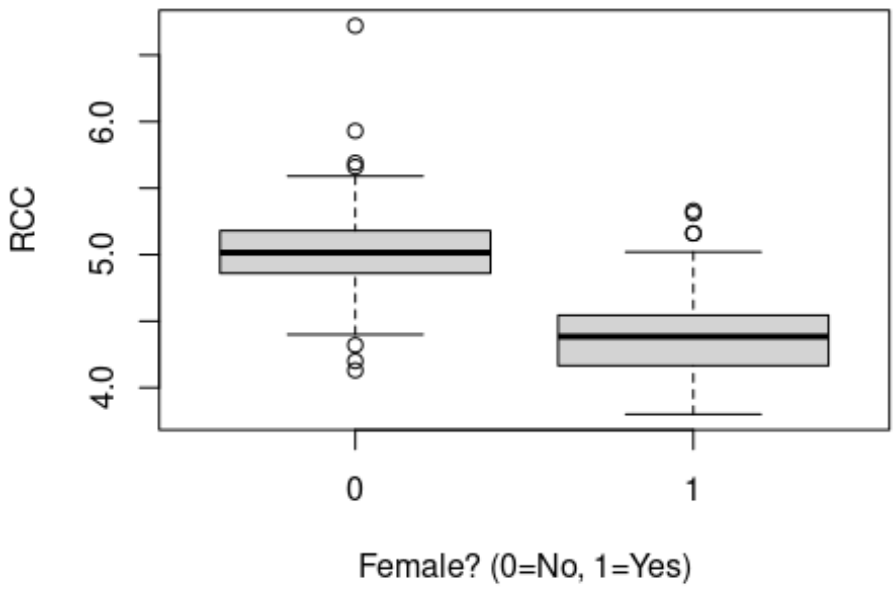

```
library(alr4)

## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

library(car)
library(carData)
data(caution)
attach(caution)
pairs(Sex~RCC+WCC+BMI,Atheletes)
```



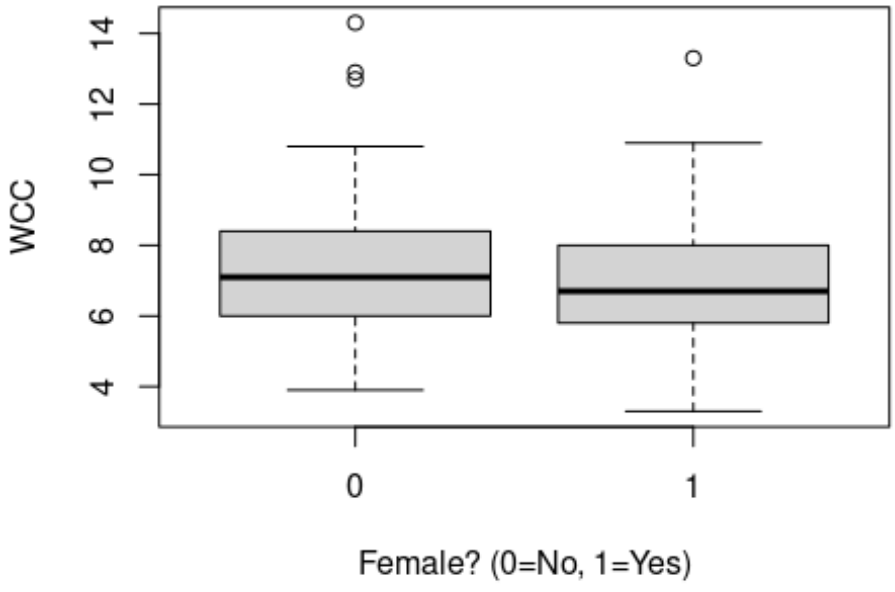

```r
#par(mfrow=c(2,2))
boxplot(Atheletes$BMI~Atheletes$Sex, ylab="BMI",xlab="Female? (0=No, 1=Yes)")
```



```
boxplot(Atheletes$RCC~Atheletes$Sex, ylab="RCC",xlab="Female? (0=No, 1=Yes)")
```




```
boxplot(Atheletes$WCC~Atheletes$Sex, ylab="WCC",xlab="Female? (0=No, 1=Yes)")
```



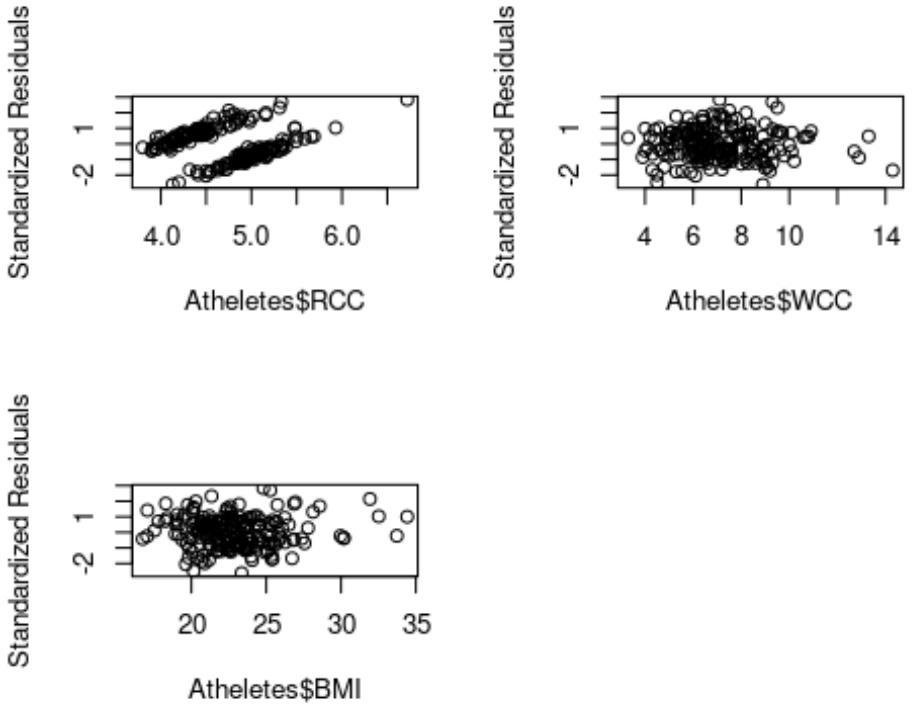
```r
m2 <- glm(Sex~RCC+log(WCC)+log(BMI),family=quasibinomial(),data=Atheletes)

summary(m2)
```

```
##
## Call:
## glm(formula = Sex ~ RCC + log(WCC) + log(BMI), family = quasibinomial(),
##     data = Atheletes)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.61695  -0.50231  -0.02308   0.50879   2.68624
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.2775     7.2256   5.713 4.03e-08 ***
## RCC          -5.4782     0.7475  -7.328 5.78e-12 ***
## log(WCC)      1.6184     0.9289   1.742  0.08302 .
## log(BMI)     -5.9661     2.0329  -2.935  0.00373 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.010601)
##
##     Null deviance: 280.01  on 201  degrees of freedom
## Residual deviance: 144.11  on 198  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

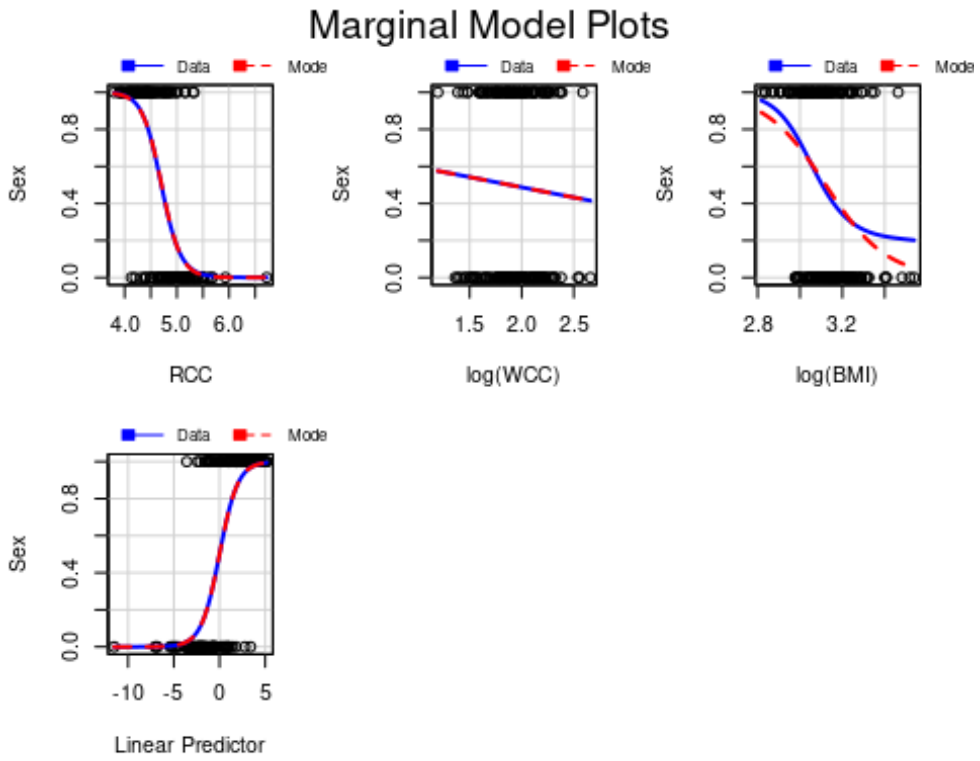
```r
m2.1 <- lm(Sex~RCC+log(WCC)+log(BMI),Atheletes)
summary(m2.1)
```

```
##
## Call:
## lm(formula = Sex ~ RCC + log(WCC) + log(BMI), data = Atheletes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9318 -0.2482 -0.0262  0.2538  0.9671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.70751    0.66487   8.584 2.67e-15 ***
## RCC         -0.70302    0.05851 -12.015  < 2e-16 ***
## log(WCC)     0.15824    0.10416   1.519  0.13030
## log(BMI)    -0.70394    0.22225  -3.167  0.00178 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3598 on 198 degrees of freedom
## Multiple R-squared:  0.4924, Adjusted R-squared:  0.4847
## F-statistic: 64.01 on 3 and 198 DF,  p-value: < 2.2e-16

StanRes2 <- rstandard(m2.1)
par(mfrow=c(2,2))
plot(Atheletes$RCC,StanRes2, ylab="Standardized Residuals")
plot(Atheletes$WCC,StanRes2, ylab="Standardized Residuals")
plot(Atheletes$BMI,StanRes2, ylab="Standardized Residuals")
```

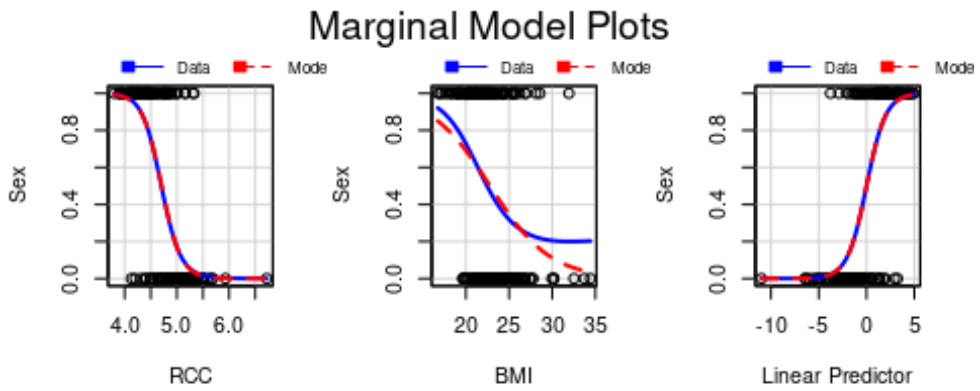



```
library(alr4)
mmps(m2,layout=c(2,3),key=TRUE)
```



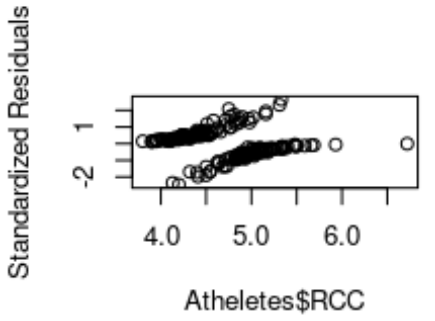

```
m3=glm(formula=Sex~RCC+BMI,family=quasibinomial(), data=Atheletes)
summary(m3)

##
## Call:
## glm(formula = Sex ~ RCC + BMI, family = quasibinomial(), data = Atheletes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5580  -0.5562  -0.0311   0.5088   2.7539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.4902     4.1092   7.177 1.38e-11 ***
## RCC          -5.2635     0.7470  -7.047 2.94e-11 ***
## BMI          -0.2060     0.0878  -2.346     0.02 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.091543)
##
##     Null deviance: 280.01  on 201  degrees of freedom
## Residual deviance: 148.69  on 199  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

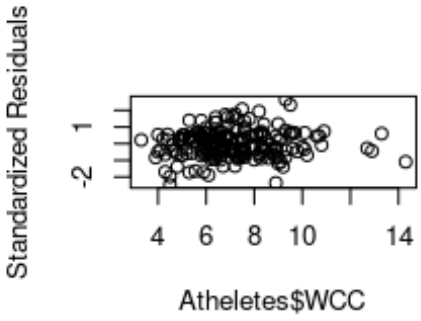
```
mmps(m3,layout=c(2,3),key=TRUE)
```

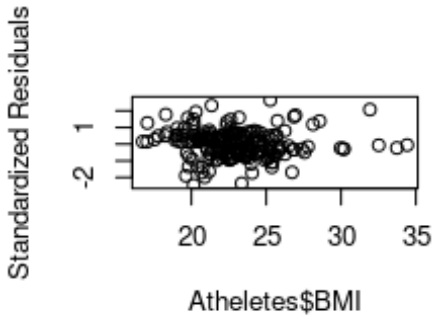



```
m3.1 <- lm(Sex~RCC+BMI,Atheletes)
summary(m3.1)

##
## Call:
## lm(formula = Sex ~ RCC + BMI, data = Atheletes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92772 -0.26680 -0.00676  0.24642  0.98997
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.366847   0.295740  14.766  < 2e-16 ***
## RCC         -0.697860   0.058501 -11.929  < 2e-16 ***
## BMI         -0.025216   0.009355  -2.696  0.00763 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3624 on 199 degrees of freedom
## Multiple R-squared:  0.4824, Adjusted R-squared:  0.4772
## F-statistic: 92.73 on 2 and 199 DF,  p-value: < 2.2e-16

StanRes2 <- rstandard(m3)
par(mfrow=c(2,2))
```

```
plot(Atheletes$RCC,StanRes2, ylab="Standardized Residuals")
plot(Atheletes$WCC,StanRes2, ylab="Standardized Residuals")
plot(Atheletes$BMI,StanRes2, ylab="Standardized Residuals")
```



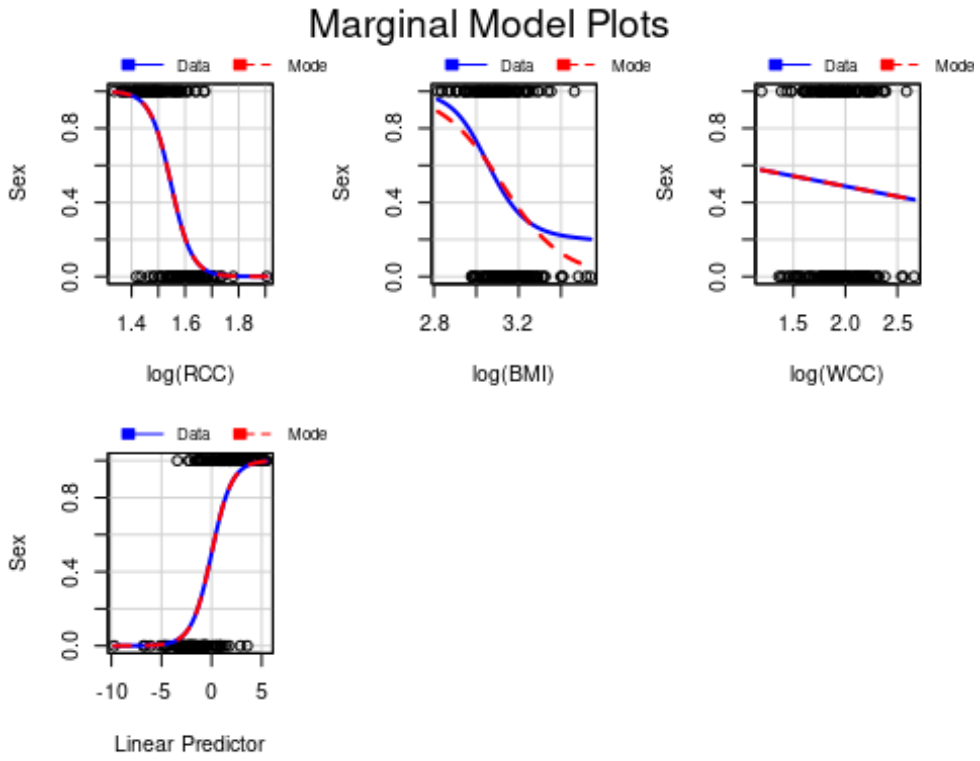





```
m4=glm(formula=Sex~log(RCC)+log(BMI)+log(WCC),family=quasibinomial(), data=At
heletes)
summary(m4)

##
## Call:
## glm(formula = Sex ~ log(RCC) + log(BMI) + log(WCC), family = quasibinomial
(),
##     data = Atheletes)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.67683  -0.50620  -0.02908   0.50527   2.62552
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.3475     8.3984   6.590 3.89e-10 ***
## log(RCC)    -25.7606     3.5220  -7.314 6.29e-12 ***
## log(BMI)     -5.9858     2.0370  -2.939  0.00369 **
## log(WCC)      1.6487     0.9332   1.767  0.07881 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for quasibinomial family taken to be 1.009517)
##
##     Null deviance: 280.01  on 201  degrees of freedom
## Residual deviance: 143.67  on 198  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
library(alr4)
mmps(m4,layout=c(2,3),key=TRUE)
```

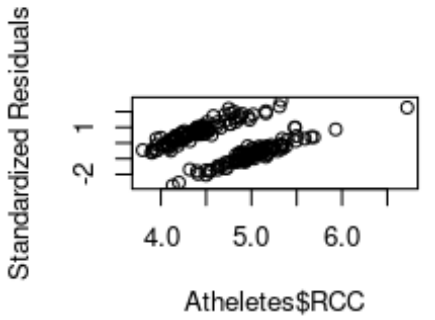


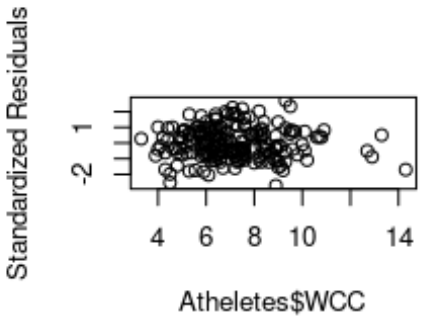

```
m4.1 <- lm(Sex~log(RCC)+log(BMI)+log(WCC), Atheletes)
summary(m4.1)
```

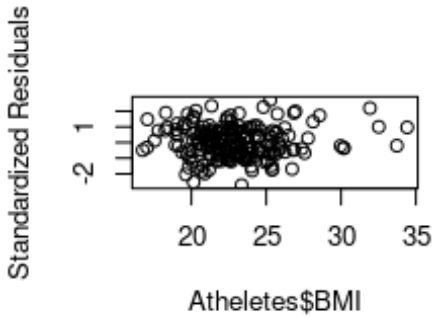
```
##
## Call:
## lm(formula = Sex ~ log(RCC) + log(BMI) + log(WCC), data = Atheletes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9574 -0.2374 -0.0395  0.2561  0.9555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.5861     0.6878  11.029   <2e-16 ***
## log(RCC)     -3.3946     0.2756 -12.316   <2e-16 ***
## log(BMI)     -0.6924     0.2199  -3.148   0.0019 **
```

```
## log(WCC)      0.1682     0.1031   1.630   0.1046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.356 on 198 degrees of freedom
## Multiple R-squared:  0.503,  Adjusted R-squared:  0.4954
## F-statistic: 66.79 on 3 and 198 DF,  p-value: < 2.2e-16
```

```
StanRes4 <- rstandard(m4.1)
par(mfrow=c(2,2))
plot(Atheletes$RCC,StanRes4, ylab="Standardized Residuals")
plot(Atheletes$WCC,StanRes4, ylab="Standardized Residuals")
plot(Atheletes$BMI,StanRes4, ylab="Standardized Residuals")
```





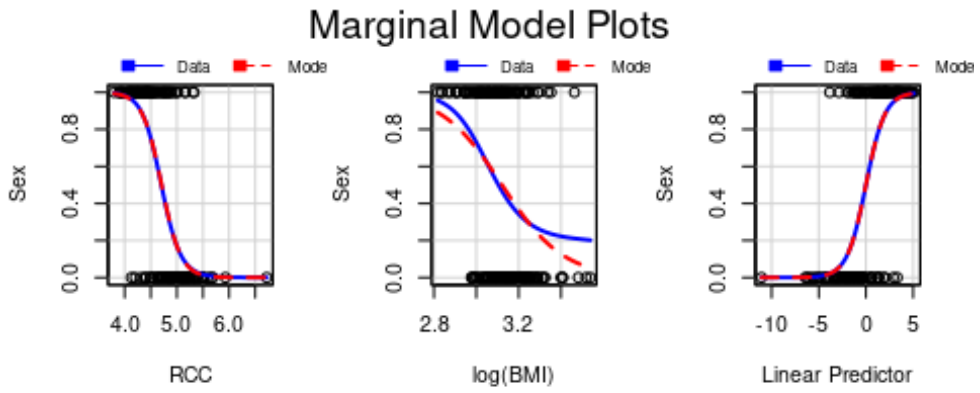



```
m5=glm(formula=Sex~RCC+log(BMI),family=quasibinomial(), data=Atheletes)
summary(m5)
```

```
##
## Call:
## glm(formula = Sex ~ RCC + log(BMI), family = quasibinomial(),
##     data = Atheletes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5867  -0.5640  -0.0314   0.4971   2.7775
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.1871     7.4747   5.510 1.10e-07 ***
## RCC          -5.2745     0.7547  -6.989 4.08e-11 ***
## log(BMI)     -5.2360     2.0428  -2.563   0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.104853)
##
##     Null deviance: 280.01  on 201  degrees of freedom
## Residual deviance: 147.34  on 199  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
mmps(m5,layout=c(2,3),key=TRUE)
```

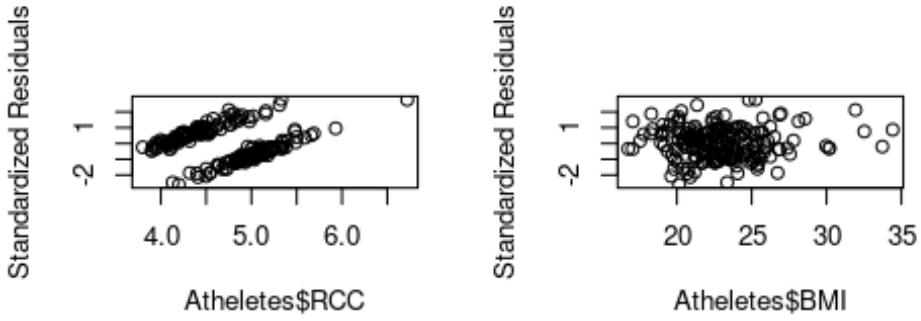



```
m5.1 <- lm(Sex~RCC+log(BMI), Atheletes)
summary(m5.1)
```

```
##
## Call:
## lm(formula = Sex ~ RCC + log(BMI), data = Atheletes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93533 -0.27030 -0.01137  0.24617  0.99661
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.82358    0.66263   8.789  7.1e-16 ***
## RCC         -0.69270    0.05831 -11.880  < 2e-16 ***
## log(BMI)    -0.65894    0.22099  -2.982  0.00322 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.361 on 199 degrees of freedom
## Multiple R-squared:  0.4864, Adjusted R-squared:  0.4813
## F-statistic: 94.24 on 2 and 199 DF,  p-value: < 2.2e-16
```

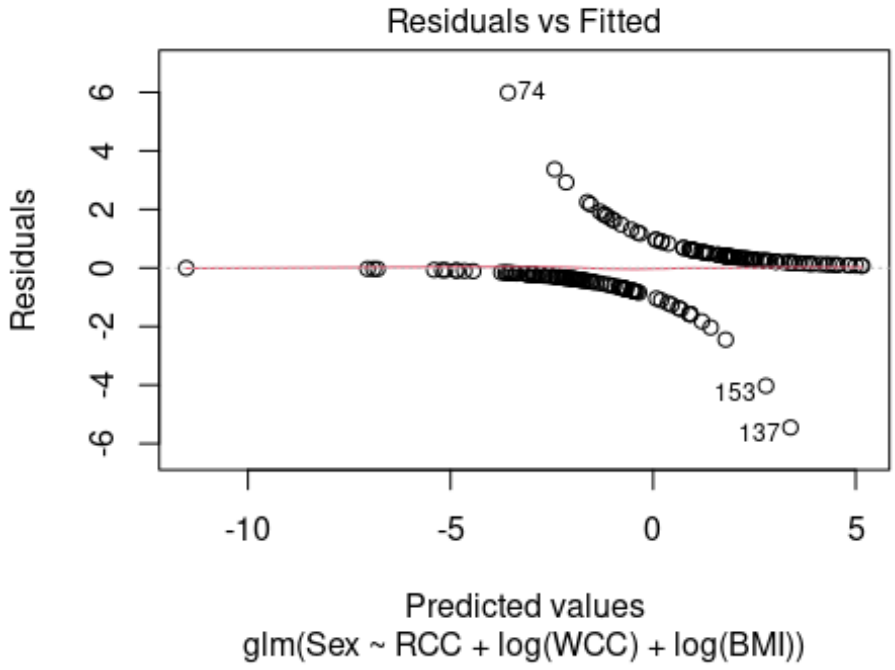
```
StanRes5 <- rstandard(m5.1)
par(mfrow=c(2,2))
plot(Atheletes$RCC,StanRes5, ylab="Standardized Residuals")
plot(Atheletes$BMI,StanRes5, ylab="Standardized Residuals")
```

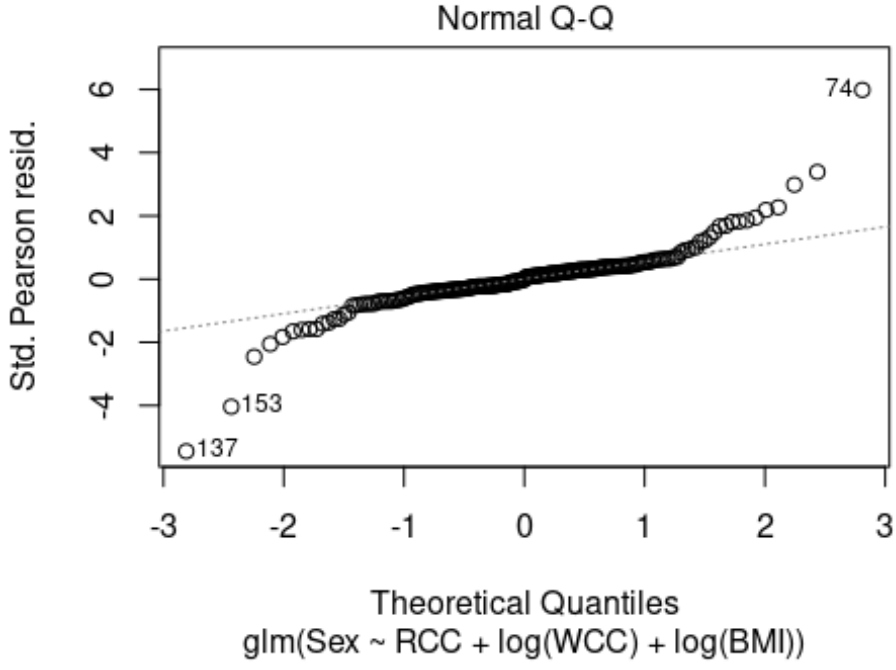


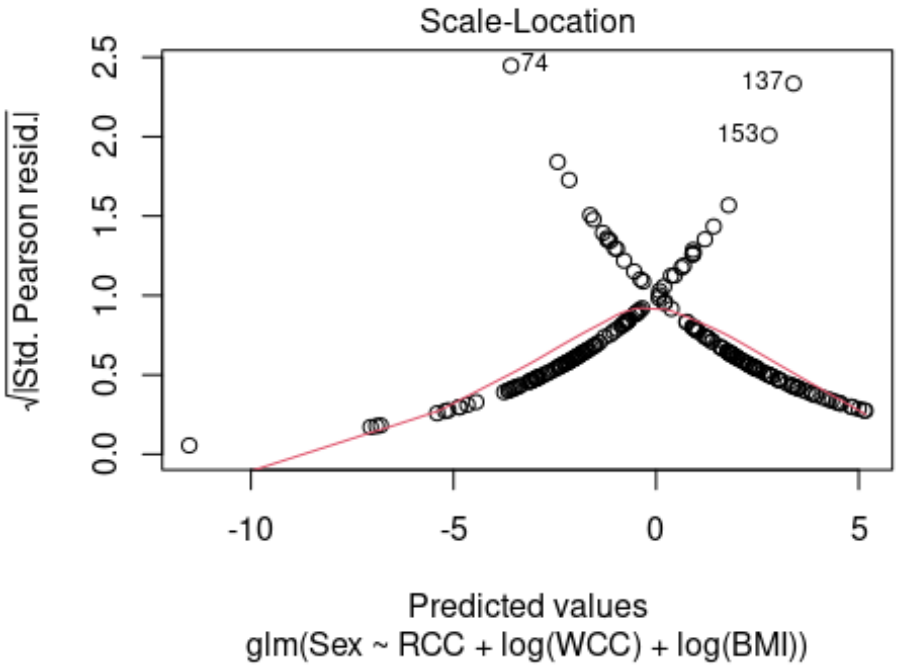
Including Plots

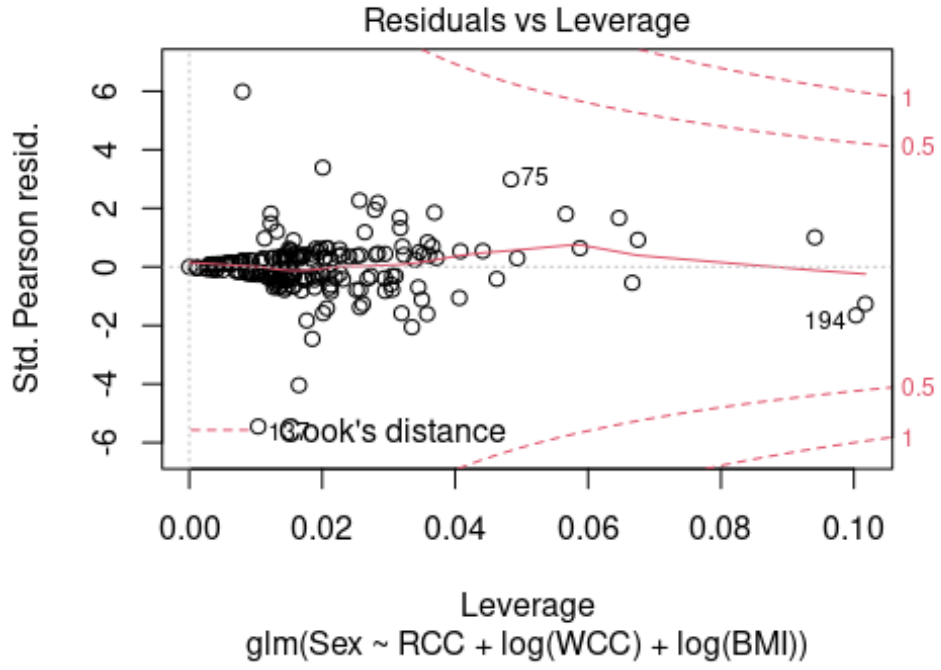
You can also embed plots, for example:

```
plot(m2)
```

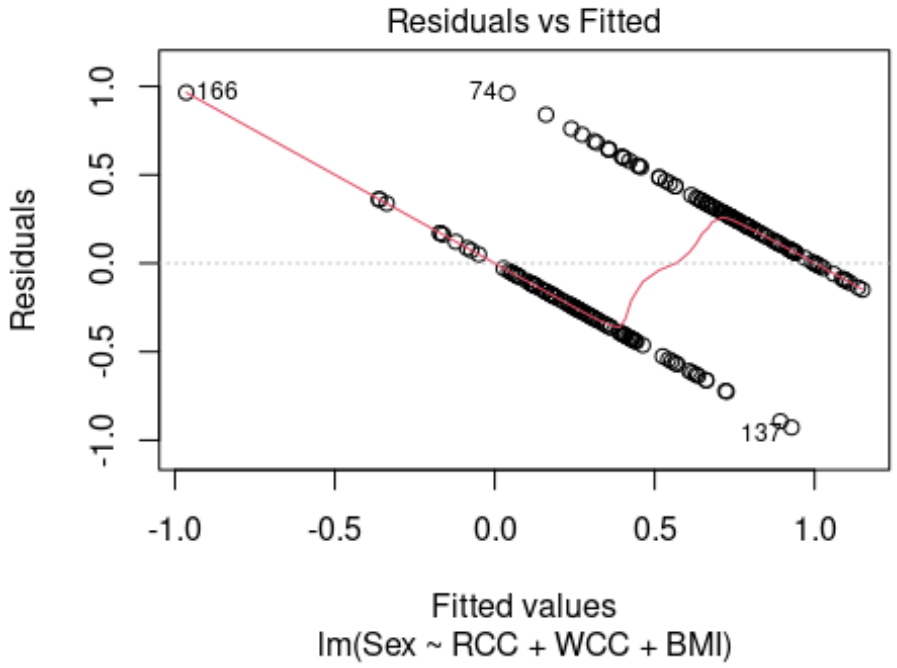

Residuals vs Fitted
Residuals
74
153
137
-10
-5
0
5
Predicted values
glm(Sex ~ RCC + log(WCC) + log(BMI))

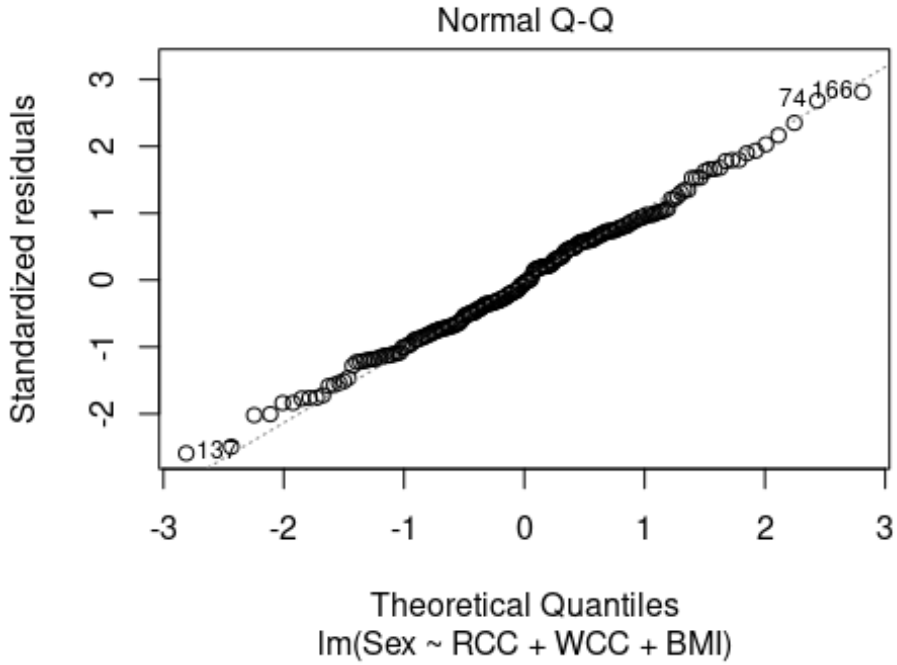

Normal Q-Q
Std. Pearson resid.
74
153
137
-3
-2
-1
0
1
2
3
Theoretical Quantiles
glm(Sex ~ RCC + log(WCC) + log(BMI))

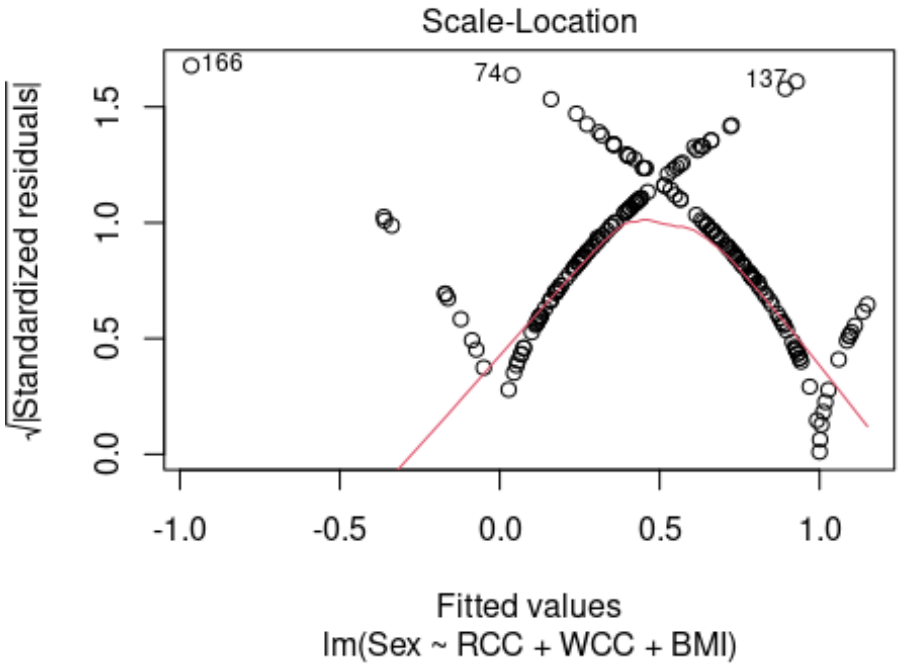

Scale-Location
74
137
153
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
-10
-5
0
5
Predicted values
glm(Sex ~ RCC + log(WCC) + log(BMI))

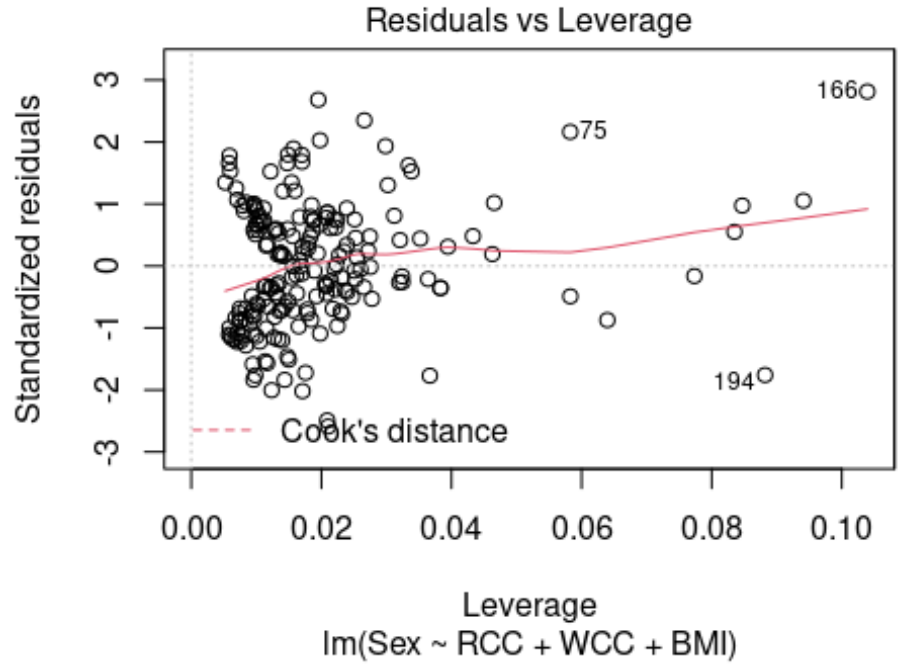

Residuals vs Leverage
Std. Pearson resid.
6
4
2
0
-2
-4
-6
75
194
137
Cook's distance
1
0.5
0.5
1
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
glm(Sex ~ RCC + log(WCC) + log(BMI))

```
plot(m1)
```

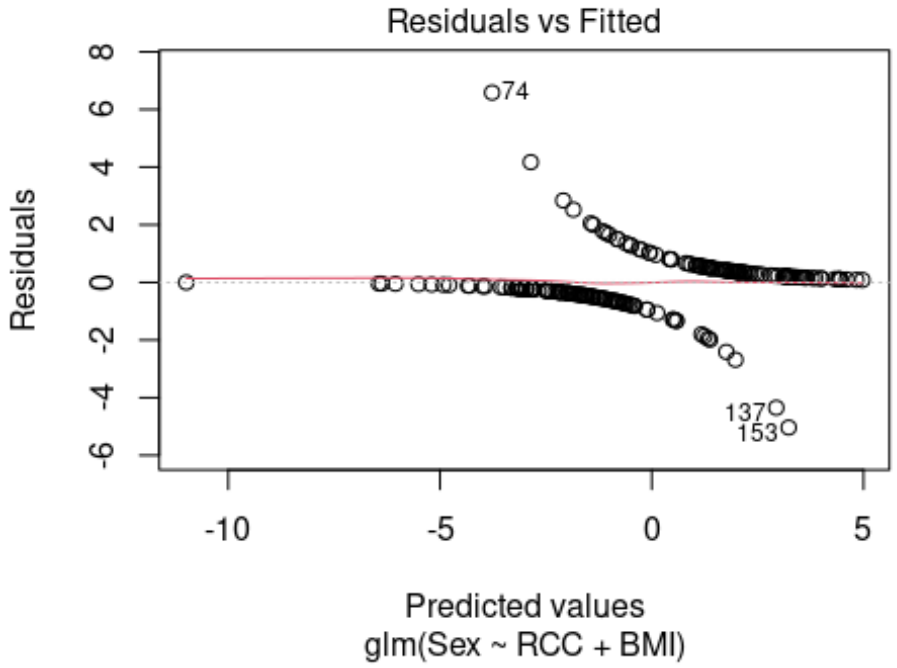

Residuals vs Fitted
Residuals
Fitted values
lm(Sex ~ RCC + WCC + BMI)
166
74
137
1.0
0.5
0.0
-0.5
-1.0

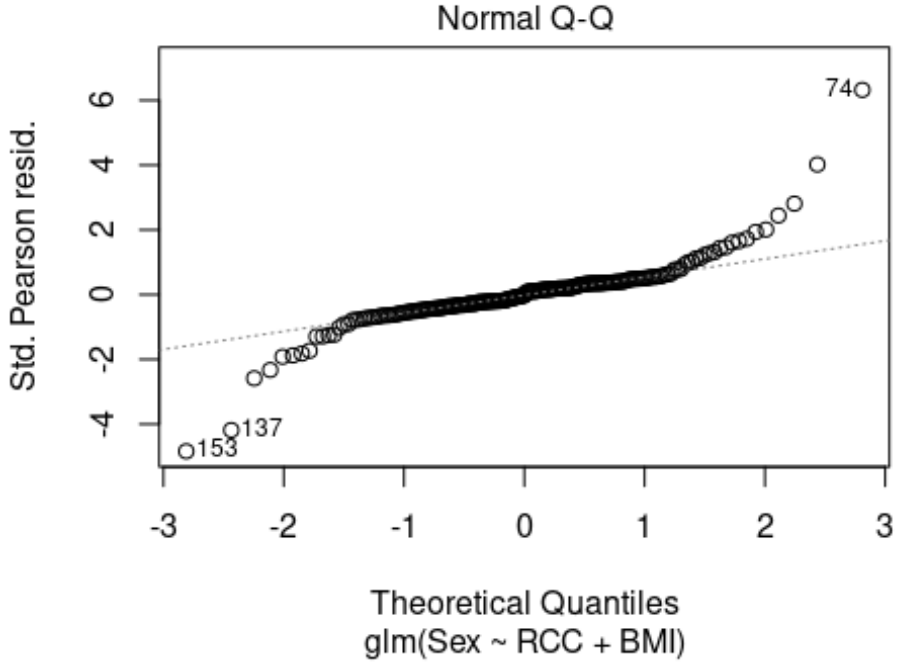

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Sex ~ RCC + WCC + BMI)
74
166
137
3
2
1
0
-1
-2
-3

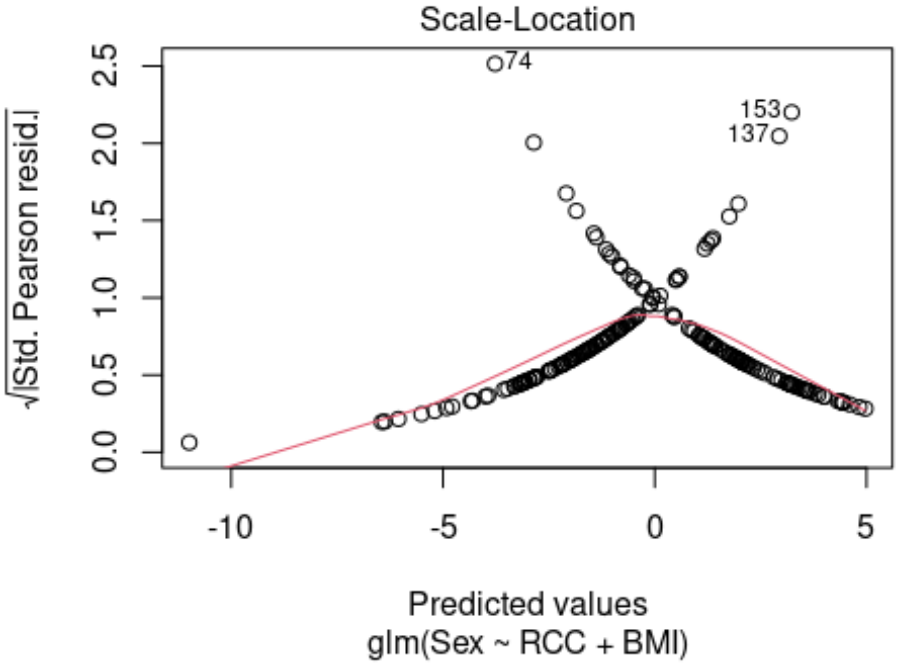

Scale-Location
166
74
137
√|Standardized residuals|
0.0
0.5
1.0
1.5
-1.0
-0.5
0.0
0.5
1.0
Fitted values
lm(Sex ~ RCC + WCC + BMI)

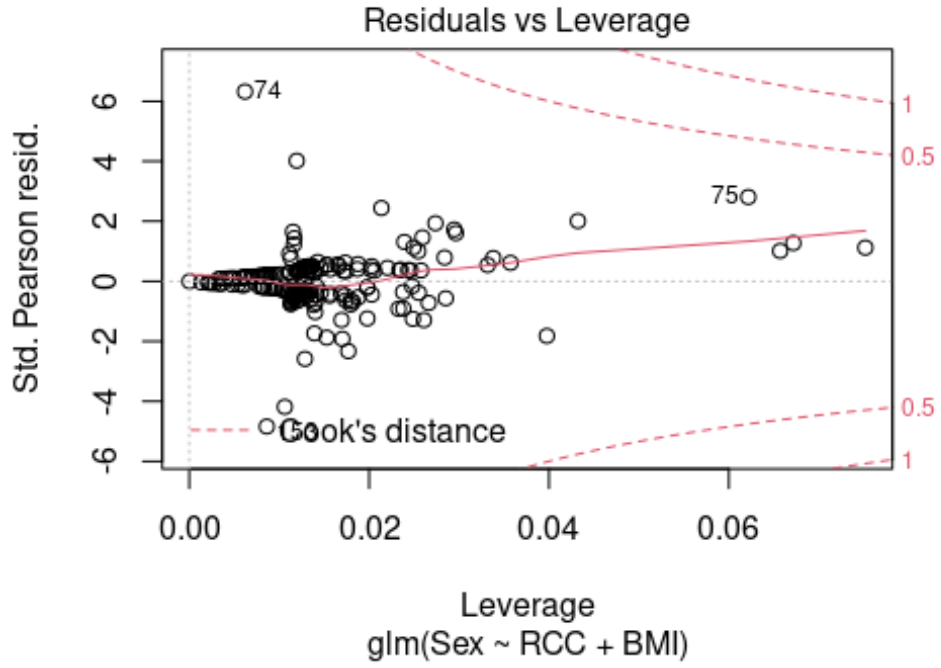

Residuals vs Leverage
166
75
194
Cook's distance
Standardized residuals
-3
-2
-1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
lm(Sex ~ RCC + WCC + BMI)

```
plot(m3)
```

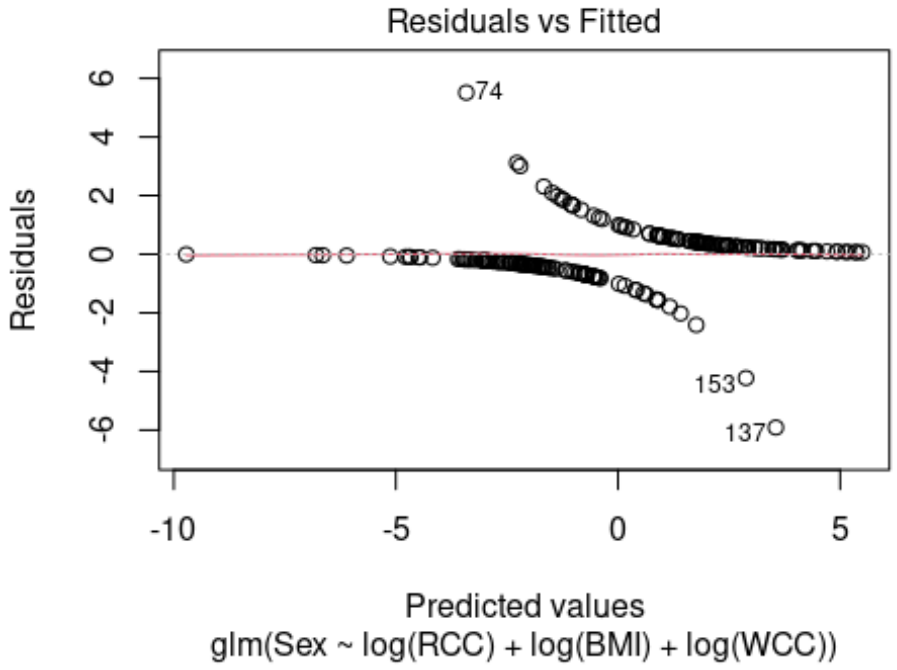

Residuals vs Fitted
Residuals
74
137
153
-10
-5
0
5
Predicted values
glm(Sex ~ RCC + BMI)

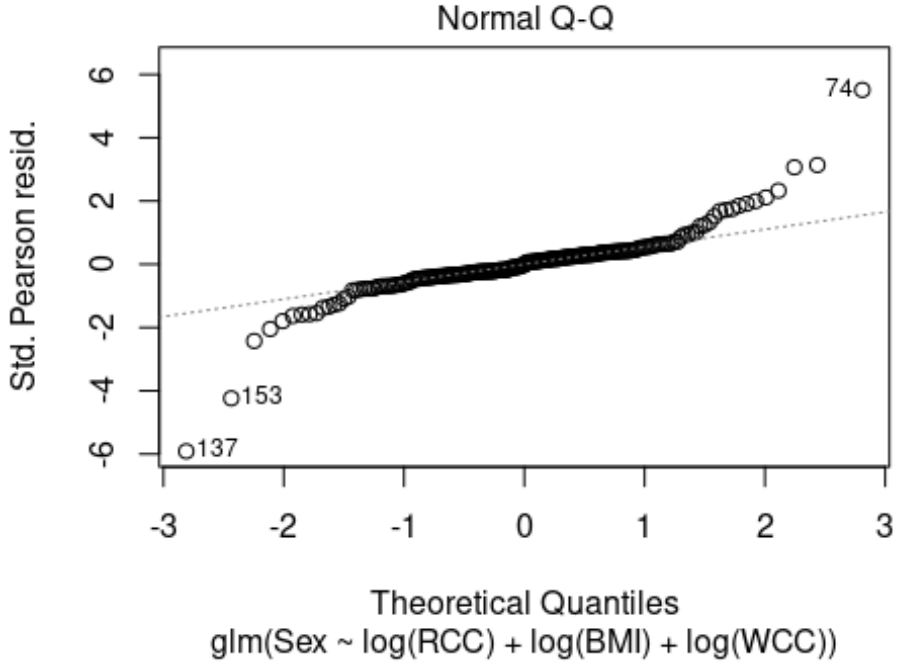

Normal Q-Q
Std. Pearson resid.
74
137
153
-3
-2
-1
0
1
2
3
Theoretical Quantiles
glm(Sex ~ RCC + BMI)

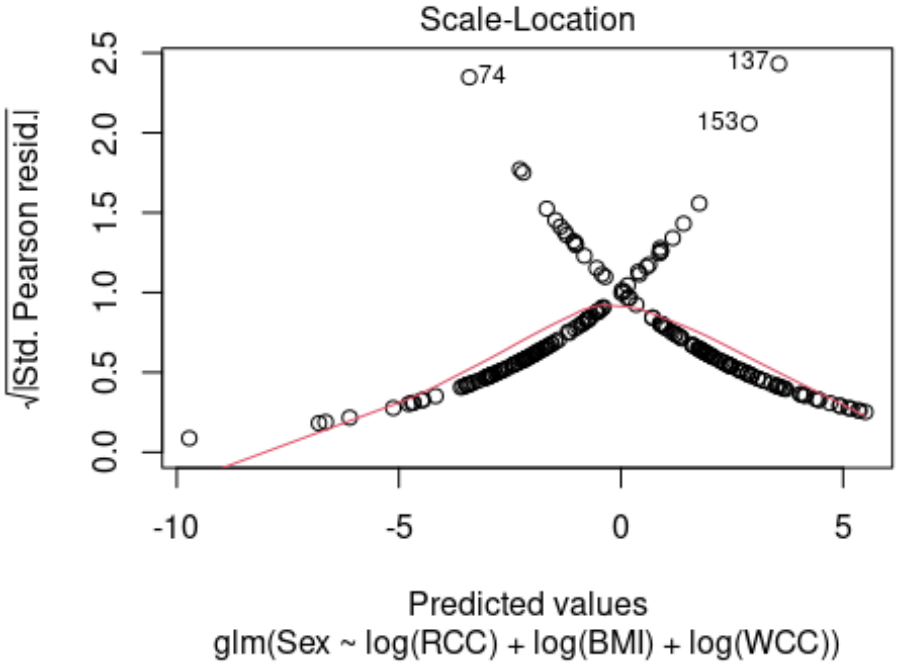

Scale-Location
√|Std. Pearson resid.|
74
153
137
0.0
0.5
1.0
1.5
2.0
2.5
-10
-5
0
5
Predicted values
glm(Sex ~ RCC + BMI)

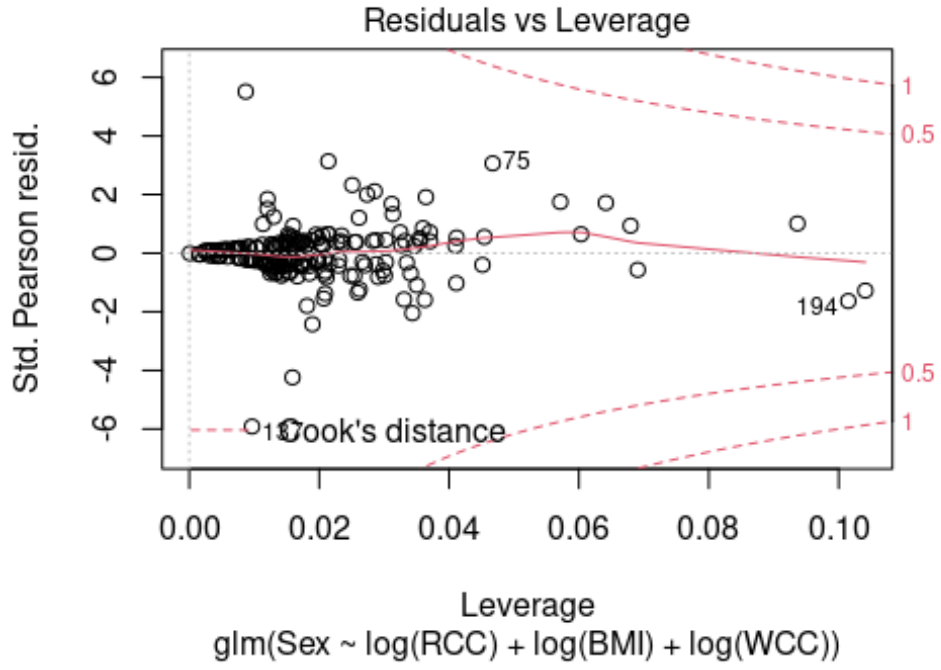

Residuals vs Leverage
Std. Pearson resid.
74
75
50
Cook's distance
1
0.5
0.5
1
6
4
2
0
-2
-4
-6
0.00
0.02
0.04
0.06
Leverage
glm(Sex ~ RCC + BMI)

```
plot(m4)
```


Residuals vs Fitted
Residuals
74
153
137
-10
-5
0
5
-6
-4
-2
0
2
4
6
Predicted values
glm(Sex ~ log(RCC) + log(BMI) + log(WCC))

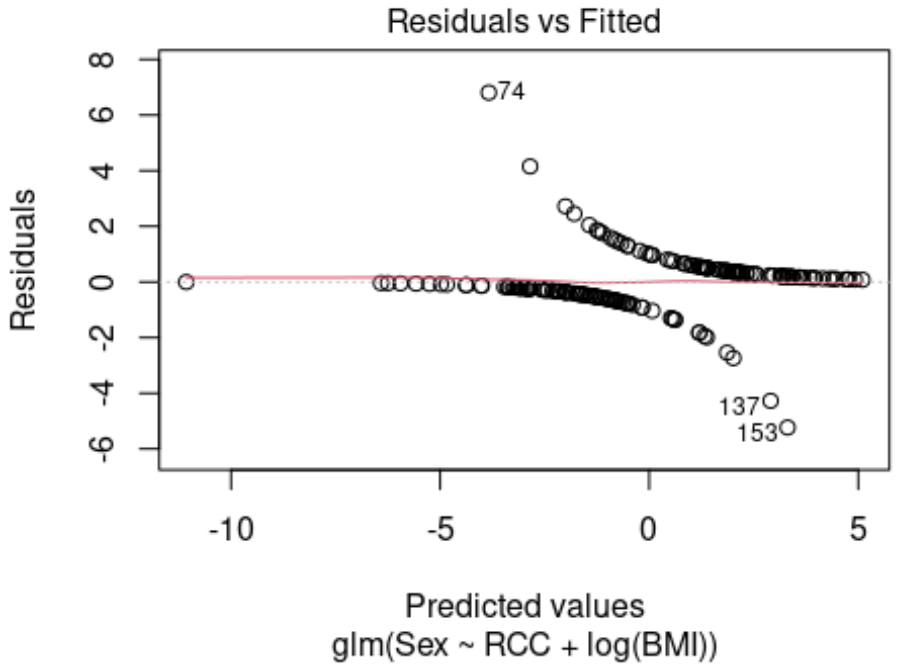

Normal Q-Q
Std. Pearson resid.
74
153
137
-3
-2
-1
0
1
2
3
-6
-4
-2
0
2
4
6
Theoretical Quantiles
glm(Sex ~ log(RCC) + log(BMI) + log(WCC))

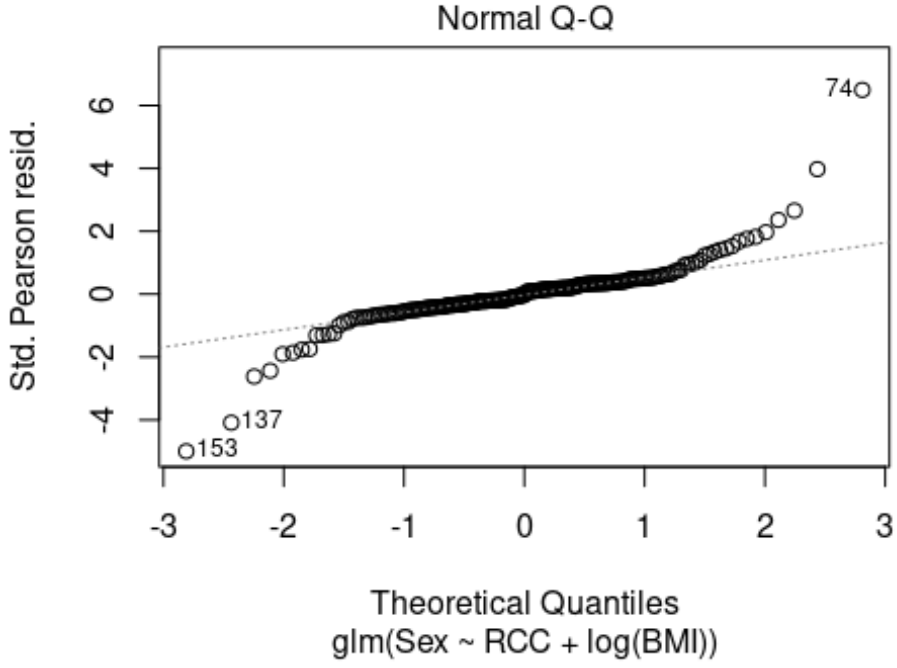


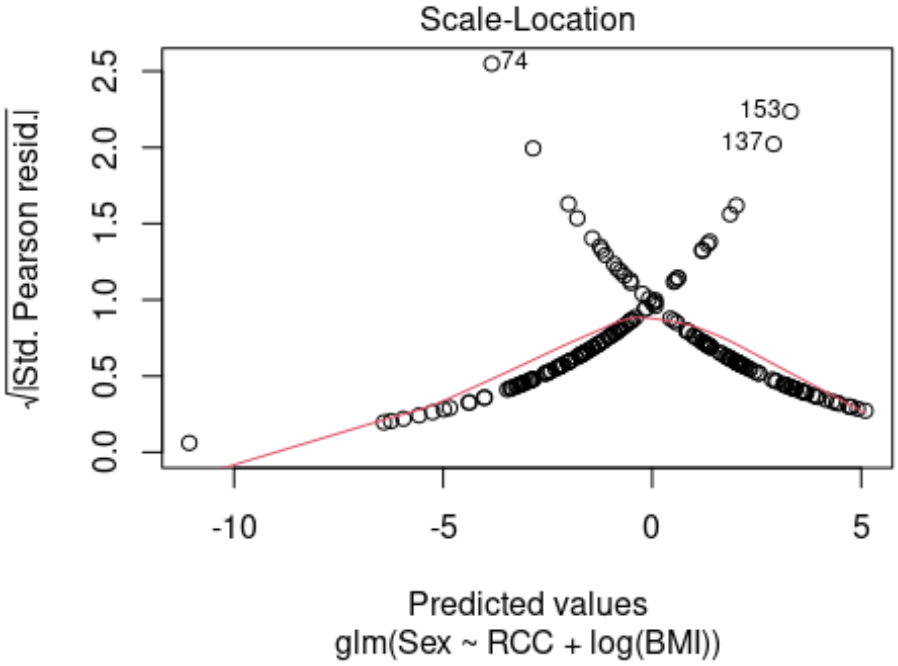


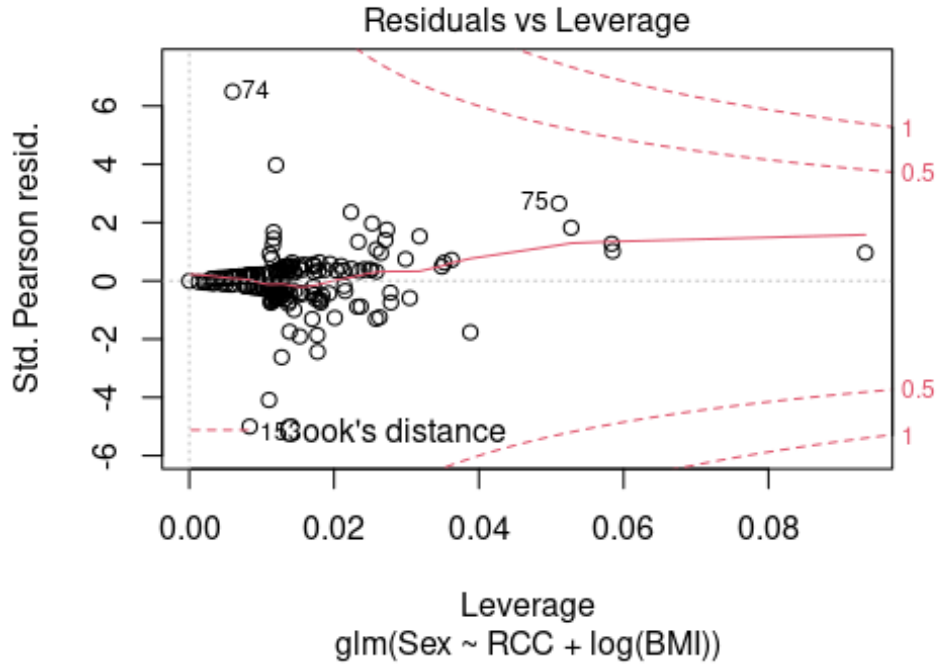
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
plot(m5)
```


Residuals vs Fitted
Residuals
74
137
153
-10
-5
0
5
Predicted values
glm(Sex ~ RCC + log(BMI))


Normal Q-Q
Std. Pearson resid.
74
137
153
-3
-2
-1
0
1
2
3
Theoretical Quantiles
glm(Sex ~ RCC + log(BMI))





Explaination of Problem 1:

From the above boxplots, it shows that all three predictor x variables, Red Cell Count (RCC), White Cell Count (WCC), and Body Mass Index (BMI) are all skewed. It is apparent that the

data for Males is much more skewed than the data for females. When looking at all predictors variables, we can see WCC, BMI, and RCC all have many outliers that need further investigation. This means that the models may not be as accurate as they could be.

I assessed 5 models (Model 1 (m1), Model 2 (m2), Model 3 (m3 and m3.1), Model 4 (m4 and m4.1), and Model 5 (m5 and m 5.1)). When looking at the modes and their summary output, it appears that they are all very similar in accuracy.

When looking at m1, it is also clear that WCC is insignificant in comparison to the BMI and RCC. This is because the p-value for WCC is 0.20338>0.05. This may mean that the WCC is not significant in predicting the Gender and WCC may possiblly be removed from the model. However, even if we have some terms that are "significant", this does not mean that the model accurately describes the data.

This told me that I needed to run diagnostics of the data. So, one way is to look at the usual plots of residuals versus fitted value and versus predictors. However, the residuals are very hard to use in ungrouped logistic regression because they are essentially either 1− fitted value for Y = 1 or − fitted value for Y = 0. A useful alternative is called a marginal model plot (mmp) which I will use to help assess the accuracy of each model. When using mmp, it is easy to assess the accuracy of each model. If the two lines are similar, the model is reproducing the data in that direction. If they differ, there is something wrong with the model.

When assessing each model using mmp, they are all very similar, where all models are approachign the data in the direction of the data. When looking at the R-squared infomraiton of each model, Model 4 (m4 and m 4.1) is slightly better with an R-squared value of 0.503. Additionally, the mmp for Model 4 (m4 and m 4.1) appears to be a better fit. The two lines are similar and the model is reproducting the data in the direction of the data. There are still some points that are outliers that need further investigation (74, 153, 137). However, Model 4 (m4 and m 4.1) is the best model of the 5 models assessed.

#Problem 2 Based on observed characteristics of an email message we want to build a classification rule for assigning the message either as spam (marked with a "1") or not spam messages, we obtain predictions from each selected model and flag a message ("0"). To build our filter we have a data of 4601 emails, and for each message we have a human-ed P[spam= 1|X] > 0.5 (just like part (f)). For each of the three models, we assigned label spam (1 for spam, 0 for not spam) and hold out 1000 emails for classification testing. We have considered choosing variables using forward stepwise selection based on AIC, BIC, and a selection using the LASSO. For each of the three models, we then computed the we then computed the

	#ERROR		
	-1	0	1
#AIC	#40	#940	#20
#BIC	#50	#900	#50
#LASSO	#20	#920	#60

Based on these results, which method of model selection do you prefer and why?

Problem 2 Explination:

A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis. This means that your report that your findings are significant when in fact they have occurred by chance. In Problem 2, a type 1 error is when messged is flagged as SPAM when it is not SPAM.

A type II error is also known as a false negative and occurs when a researcher fails to reject a null hypothesis which is really false. Here a researcher concludes there is not a significant effect, when actually there really is. In Problem 2, a type 2 error is when messged is not flagged when it is SPAM

For this problem, the type 1 and type 2 errors are defined as the following:

Type 1 Error: If a messged is flagged as SPAM when it is not SPAM

Type 2 Error: A message is not flagged when it is SPAM

From the data output, BIC has more errors with both type 1 and type 2 errors and is the worst choice of the three.

When it comes to type 1 errors, I prefer LASSO wich has the least number of type 1 errors.

When it comes to type 2 errors, I perfer AIC because it has the least number of type 2 errors.

However, if I had to pick one model, I would pick AIC because has the least number of errors.