GroupProject

Group3

11/23/2021

Loading Packages

Data Cleaning and Variates Selection

```
varinterst = c("Age","Gender","Race1","Depressed","Diabetes","BPSysAve", "BPDiaAve", "TotChol", "Testos
nh1 = nh[,varinterst]
nh1 <- nh1[ which(nh1$Age >=18 & nh1$Age <=59), ] #Due to Marijuana variable.
nh1 <- nh1 %>%
  mutate(HypT = if_else(nh1$BPSysAve >= 130 | nh1$BPDiaAve >=80, 1, 0)) #Add hypertension variable
nh1 <- nh1 %>%
  mutate(Sex = ifelse(nh1$Gender == "female", 1, 0)) #Recode Sex into binary where female=1, male=0
nh1 <- nh1 %>%
  mutate(Diabete = ifelse(nh1$Diabetes == "No", 0, 1)) #Recode Diabetes into binary where No=0, Yes=1
nh1 <- nh1 %>%
  mutate(RegMarij = ifelse(nh1$RegularMarij == "No", 0, 1)) #Recode Regular Marij into binary where No=
nh1 <- nh1 %>%
  mutate(Depress = ifelse(nh1$Depressed == "None", 0, 1)) #Recode Depressed into reports syptoms or doe
nh1$Incomplete = complete.cases(nh1)
nh1$Race1 = as.factor(nh1$Race1) #Race as factor
#Creates Sex Specific Data Sets
drops = c("Sex")
nhF= nh1[ which(nh1$Sex == 1), ]
nhM = nh1[ which(nh1$Sex == 0), ]
nhF = nhF[,!(names(nhF) %in% drops)]
nhM = nhM[,!(names(nhM) %in% drops)]
```

Complete/Incomplete Data Set

```
nh1[,c("Age","Gender","Race1","Depressed","Diabetes","TotChol", "Testosterone", "RegularMarij", "Weight
 tbl_summary(by = Incomplete,
             missing = "no",
             statistic = list(all_continuous() ~ "{mean} ({sd})",
                              all_categorical() ~ "{n} ({p}%)"),
             ) %>%
  add_n %>%
  add_p(test = list(Age ~ "t.test",
                    Race1 ~ "chisq.test",
                    Depressed ~ "chisq.test",
                    Diabetes ~ "chisq.test",
                    HypT ~ "chisq.test",
```

```
                    TotChol ~ "t.test",
                    Testosterone ~ "t.test",
                    RegularMarij ~ "chisq.test",
                    Weight ~ "t.test",
                    Height ~ "t.test"),
       test.args = all_tests("t.test") ~ list(var.equal = TRUE),## Important argument!
       pvalue_fun = function(x) style_pvalue(x, digits = 2)) %>%
  bold_p(t = 0.05) %>%
  bold_labels
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N**	**FALSE**, N = 3,395	**TRUE**, N = 2,247	**p-value**
Age	5,642	38 (12)	39 (12)	0.12
Gender	5,642			**0.001**
female		1,729 (51%)	1,045 (47%)	
male		1,666 (49%)	1,202 (53%)	
Race1	5,642			**0.041**
Black		426 (13%)	246 (11%)	
Hispanic		211 (6.2%)	144 (6.4%)	
Mexican		369 (11%)	208 (9.3%)	
White		2,089 (62%)	1,465 (65%)	
Other		300 (8.8%)	184 (8.2%)	
Depressed	4,984			**0.002**
None		2,072 (76%)	1,790 (80%)	
Several		472 (17%)	310 (14%)	
Most		193 (7.1%)	147 (6.5%)	
Diabetes	5,642	195 (5.7%)	134 (6.0%)	0.77
TotChol	5,349	5.05 (1.05)	5.00 (1.05)	0.074
Testosterone	2,586	172 (222)	233 (230)	**<0.001**
RegularMarij	4,941	734 (27%)	632 (28%)	0.51
Weight	5,607	82 (22)	83 (21)	0.24
Height	5,611	169 (10)	170 (10)	**0.002**
HypT	5,428	887 (28%)	646 (29%)	0.50

Female only Complete/Incomplete Data Set

```
## Shiny app
nhF[,c("Age","Gender","Race1","Depressed","Diabetes","TotChol", "Testosterone", "RegularMarij", "Weight
 tbl_summary(by = Incomplete,
             missing = "no",
             statistic = list(all_continuous() ~ "{mean} ({sd})",
                              all_categorical() ~ "{n} ({p}%)"),
             ) %>%
  add_n %>%
  add_p(test = list(Age ~ "t.test",
                    Race1 ~ "chisq.test",
                    Depressed ~ "chisq.test",
                    Diabetes ~ "chisq.test",
```

```
                    HypT ~ "chisq.test",
                    TotChol ~ "t.test",
                    Testosterone ~ "t.test",
                    RegularMarij ~ "chisq.test",
                    Weight ~ "t.test",
                    Height ~ "t.test"),
       test.args = all_tests("t.test") ~ list(var.equal = TRUE),## Important argument!
       pvalue_fun = function(x) style_pvalue(x, digits = 2)) %>%
  bold_p(t = 0.05) %>%
  bold_labels
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N**	**FALSE**, N = 1,729	**TRUE**, N = 1,045	**p-value**
Age	2,774	38 (12)	39 (12)	0.55
Gender	2,774			>0.99
female		1,729 (100%)	1,045 (100%)	
male		0 (0%)	0 (0%)	
Race1	2,774			0.12
Black		225 (13%)	132 (13%)	
Hispanic		118 (6.8%)	68 (6.5%)	
Mexican		165 (9.5%)	86 (8.2%)	
White		1,050 (61%)	680 (65%)	
Other		171 (9.9%)	79 (7.6%)	
Depressed	2,367			**0.001**
None		943 (71%)	805 (77%)	
Several		269 (20%)	153 (15%)	
Most		110 (8.3%)	87 (8.3%)	
Diabetes	2,774	84 (4.9%)	52 (5.0%)	0.96
TotChol	2,626	5.02 (1.02)	5.08 (1.08)	0.15
Testosterone	1,253	23 (12)	26 (21)	0.067
RegularMarij	2,347	270 (21%)	237 (23%)	0.28
Weight	2,754	76 (22)	76 (20)	0.97
Height	2,754	163 (7)	163 (7)	0.39
HypT	2,644	319 (20%)	229 (22%)	0.24

##Exploratory Stats: Male only Complete/Incomplete Data Set

```
nhM[,c("Age","Gender","Race1","Depressed","Diabetes","TotChol", "Testosterone", "RegularMarij", "Weight
 tbl_summary(by = Incomplete,
             missing = "no",
             statistic = list(all_continuous() ~ "{mean} ({sd})",
                              all_categorical() ~ "{n} ({p}%)"),
             ) %>%
  add_n %>%
  add_p(test = list(Age ~ "t.test",
                    Race1 ~ "chisq.test",
                    Depressed ~ "chisq.test",
                    Diabetes ~ "chisq.test",
                    HypT ~ "chisq.test",
                    TotChol ~ "t.test",
```

```
                    Testosterone ~ "t.test",
                    RegularMarij ~ "chisq.test",
                    Weight ~ "t.test",
                    Height ~ "t.test"),
        test.args = all_tests("t.test") ~ list(var.equal = TRUE),## Important argument!
        pvalue_fun = function(x) style_pvalue(x, digits = 2)) %>%
  bold_p(t = 0.05)
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N**	**FALSE**, N = 1,666	**TRUE**, N = 1,202	**p-value**
Age	2,868	38 (12)	39 (12)	0.13
Gender	2,868			>0.99
female		0 (0%)	0 (0%)	
male		1,666 (100%)	1,202 (100%)	
Race1	2,868			0.052
Black		201 (12%)	114 (9.5%)	
Hispanic		93 (5.6%)	76 (6.3%)	
Mexican		204 (12%)	122 (10%)	
White		1,039 (62%)	785 (65%)	
Other		129 (7.7%)	105 (8.7%)	
Depressed	2,617			0.36
None		1,129 (80%)	985 (82%)	
Several		203 (14%)	157 (13%)	
Most		83 (5.9%)	60 (5.0%)	
Diabetes	2,868	111 (6.7%)	82 (6.8%)	0.93
TotChol	2,723	5.08 (1.09)	4.93 (1.02)	**<0.001**
Testosterone	1,333	409 (189)	413 (168)	0.78
RegularMarij	2,594	464 (33%)	395 (33%)	0.83
Weight	2,853	89 (21)	89 (19)	0.79
Height	2,857	176 (7)	177 (8)	0.27
HypT	2,784	568 (36%)	417 (35%)	0.53

Our first limitation is that complete/incomplete data are significantly different. Even if we adjusted for gender, there is still significant difference between the two data set. Using the complete model can be biased. Not generalizable to US population.

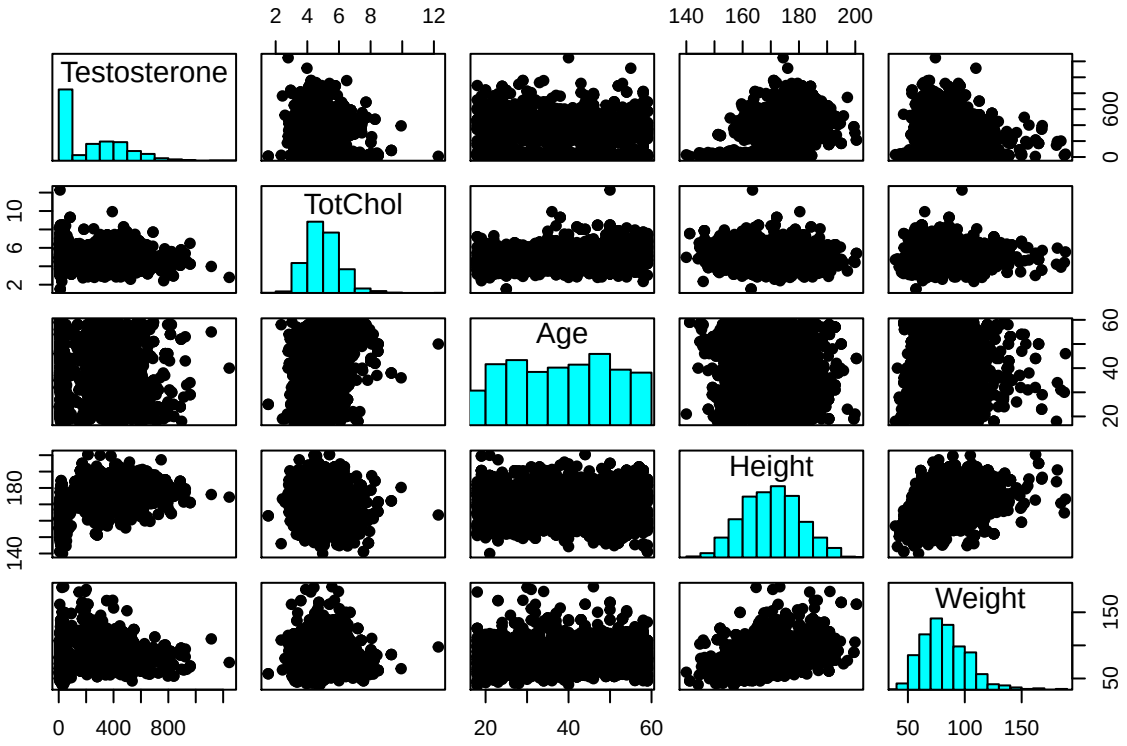
```
nh2 = drop_na(nh1) #complete data set
```

Exploratory Stats: Testosterone (and other continous) by Categorical Variables Tables (assess) (NOT INCLUDE IN PAPER OR PRESENTATION DIRECTLY)

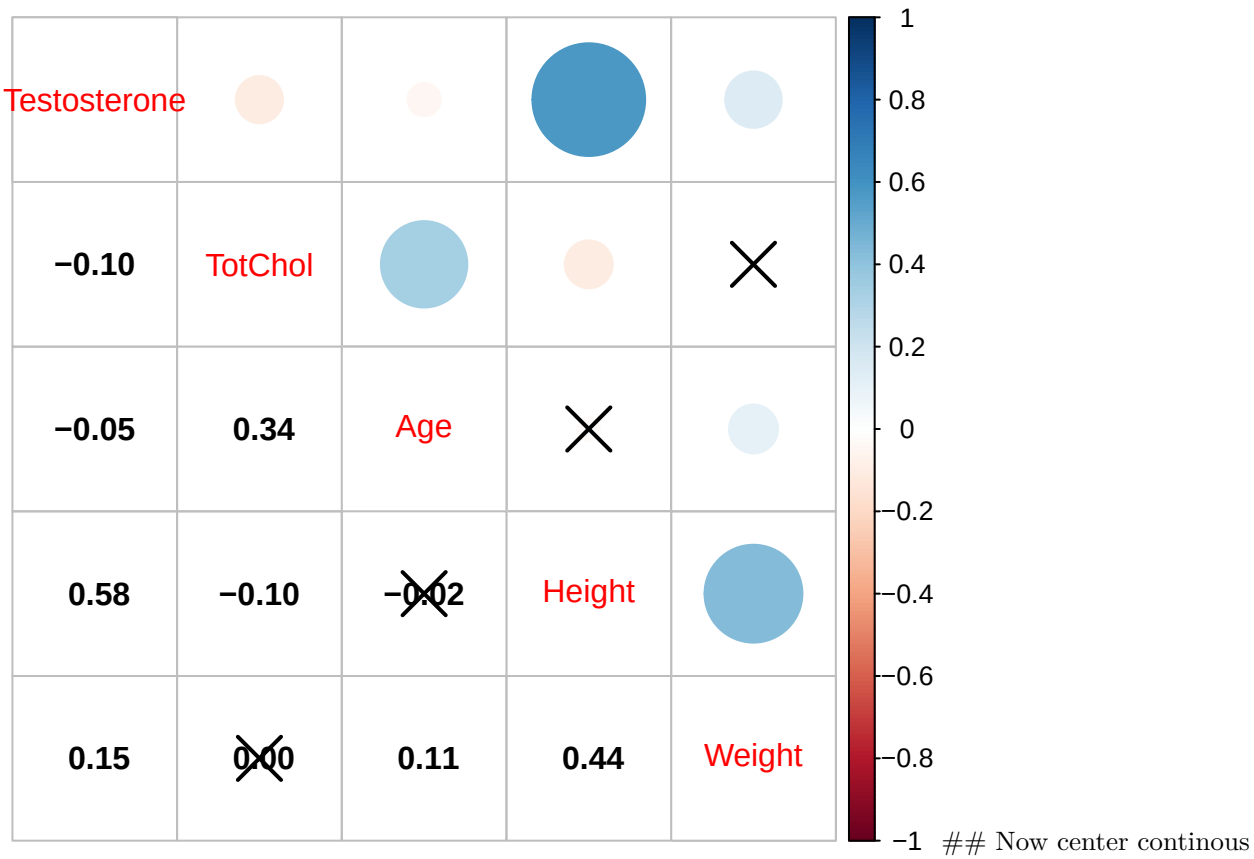
Exploratory Stats: Continous Variables (NOT INCLUDE IN PAPER OR PRESENTATION DIRECTLY)

```
#Compare Y to continuous X
scatmatrixData = nh2[,c("Testosterone", "TotChol", "Age", "Height", "Weight")]
```

```
panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
pairs(scatmatrixData, pch = 19, diag.panel=panel.hist)
```




```
cormat = cor(scatmatrixData)
pres <- cor.mtest(scatmatrixData, conf.level = .95)
corrplot.mixed(cormat, lower.col = "black", number.cex = 1,p.mat = pres$p, sig.level = .05)
```



Now center continous variables (Needed)

```
nh2$Weightc = (nh2$Weight-mean(nh2$Weight, na.rm = TRUE))/sd(nh2$Weight, na.rm = TRUE)
nh2$Heightc = (nh2$Height-mean(nh2$Height, na.rm = TRUE))/sd(nh2$Height, na.rm = TRUE)
nh2$Agec = (nh2$Age-mean(nh2$Age, na.rm = TRUE))/sd(nh2$Age, na.rm = TRUE)
nh2$TotCholc = (nh2$TotChol-mean(nh2$TotChol, na.rm = TRUE))/sd(nh2$TotChol, na.rm = TRUE)
```

Function to check assumptions

```
checka <- function (model) {
  car::avPlots(model)
  plot(model$fitted.values, rstudent(model))
  hist(rstudent(model))
  car::qqPlot(rstudent(model))
  shapiro.test(rstudent(model)
               ) ##Should we include shapiro wilk????
}
```

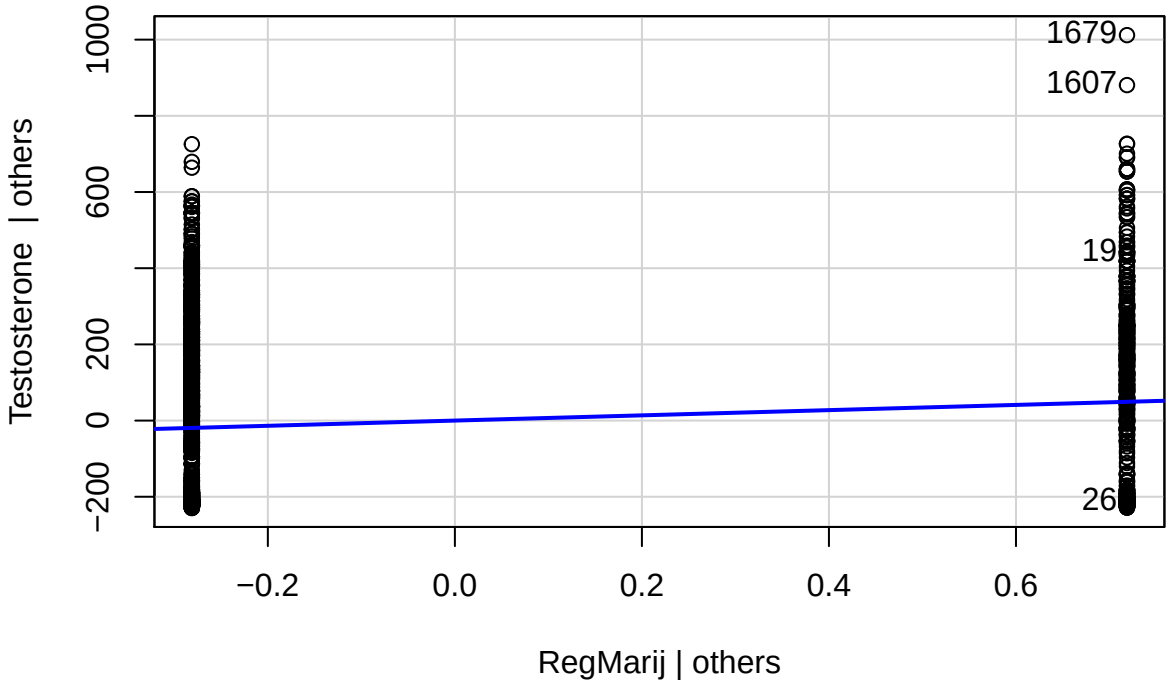
Unrefined Model (what to say about this?)

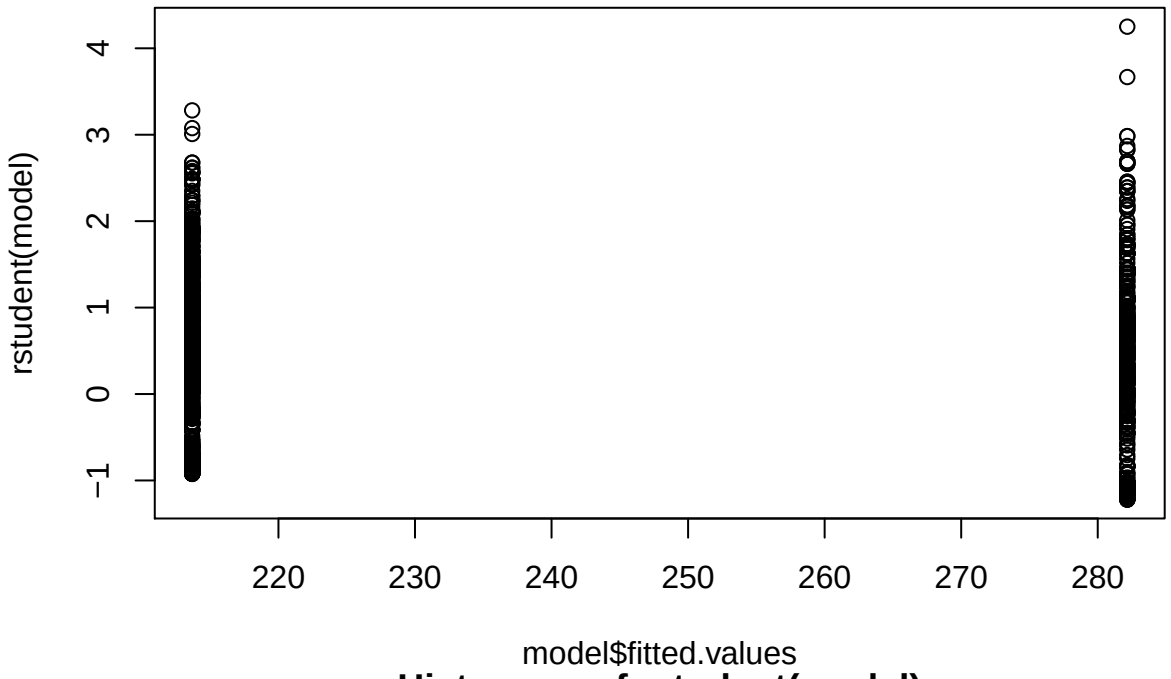
```
m_unrefined <- lm(Testosterone~RegMarij, data=nh2)
summary(m_unrefined)
```

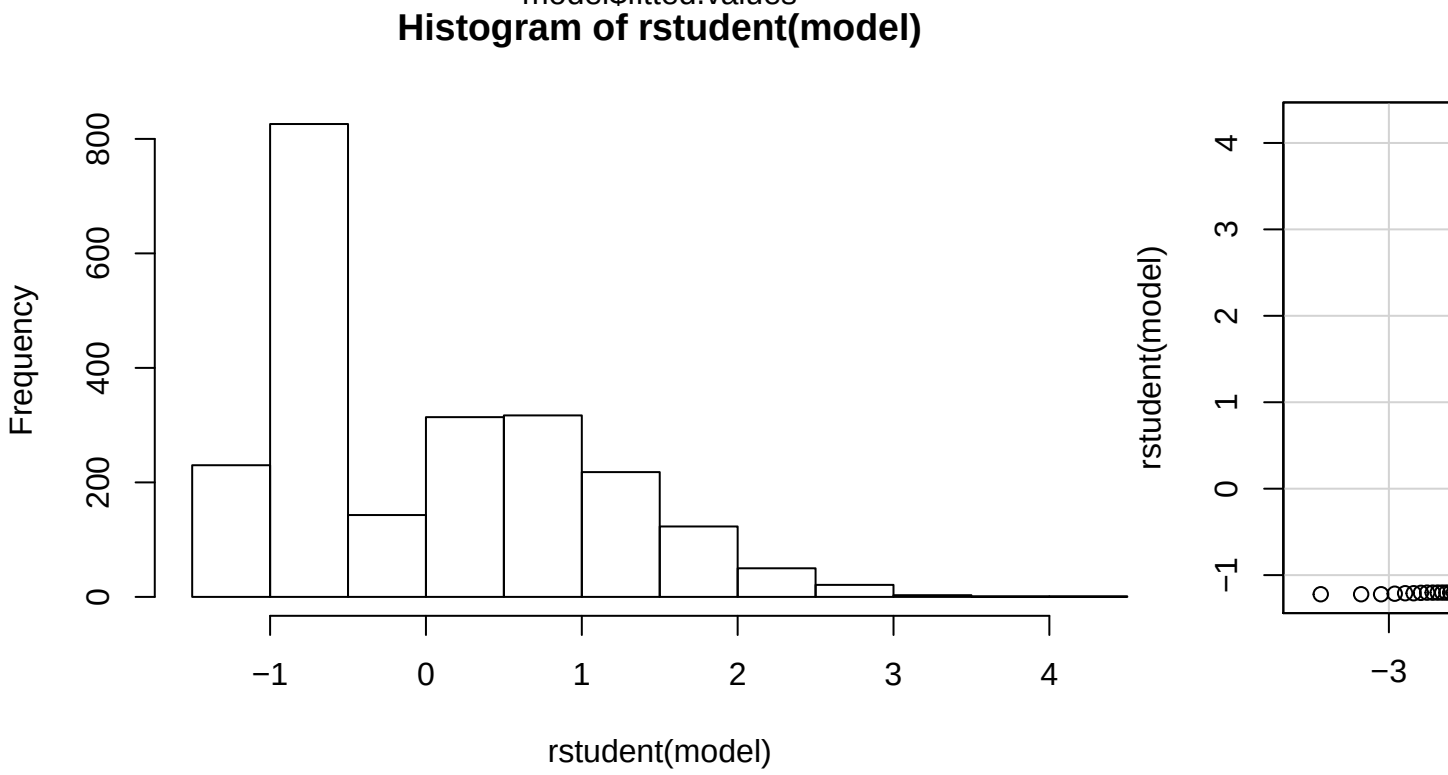
```
##
## Call:
```

```
## lm(formula = Testosterone ~ RegMarij, data = nh2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -277.59 -195.04  -50.66  166.69  962.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  213.682      5.662  37.742  < 2e-16 ***
## RegMarij      68.481     10.676   6.415 1.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 227.5 on 2245 degrees of freedom
## Multiple R-squared:  0.018,  Adjusted R-squared:  0.01756
## F-statistic: 41.15 on 1 and 2245 DF,  p-value: 1.715e-10
```

```
checka(m_unrefined) ##Comment on plots
```







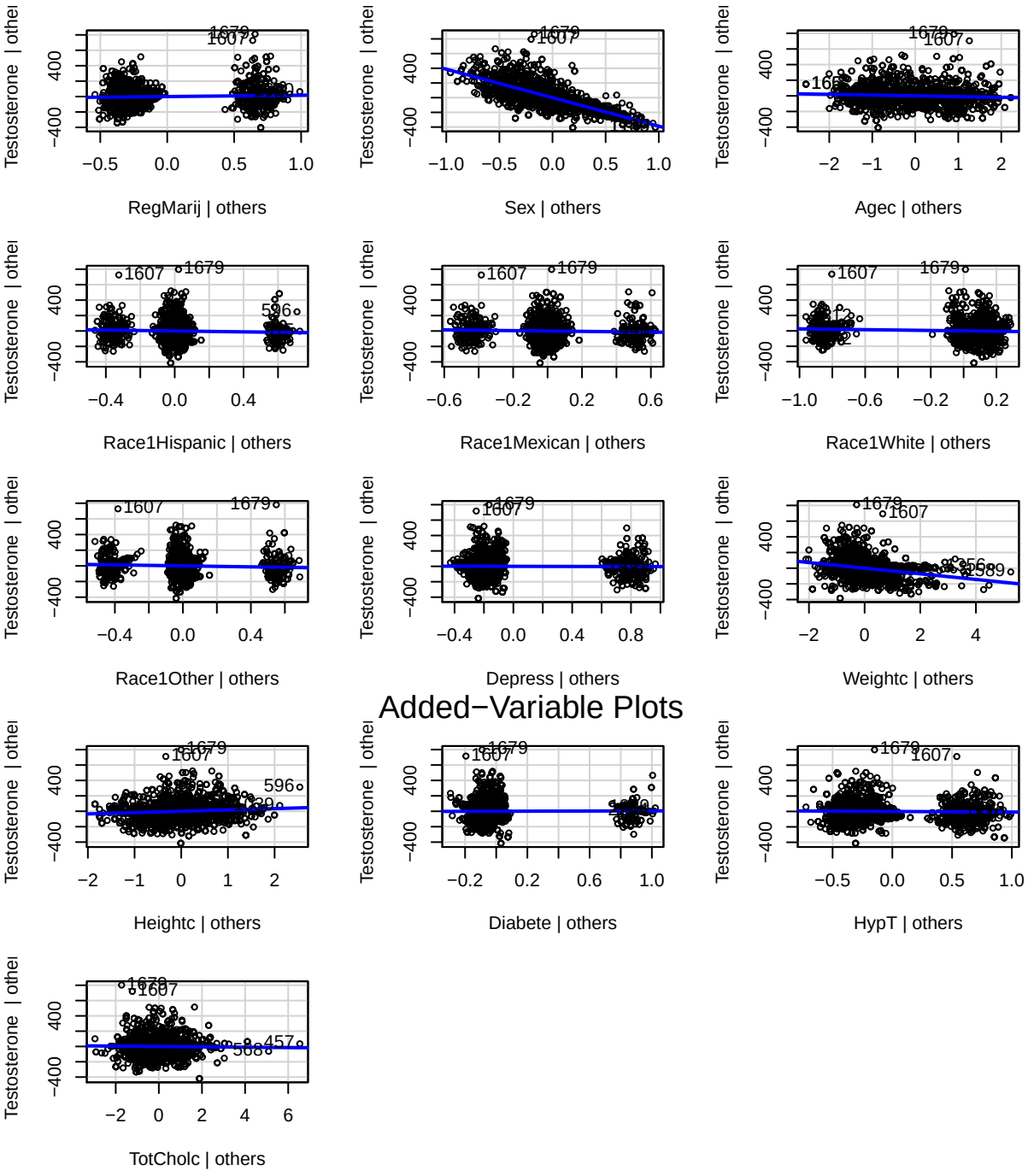

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(model)
## W = 0.8943, p-value < 2.2e-16
```

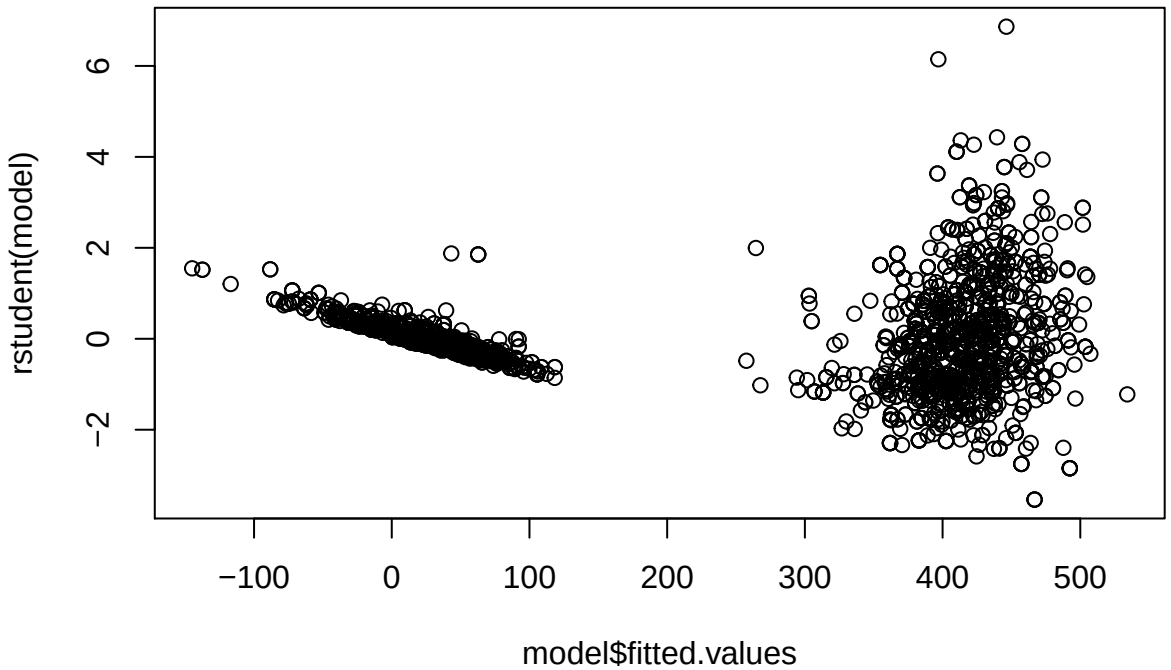
Adjusted Model with no log

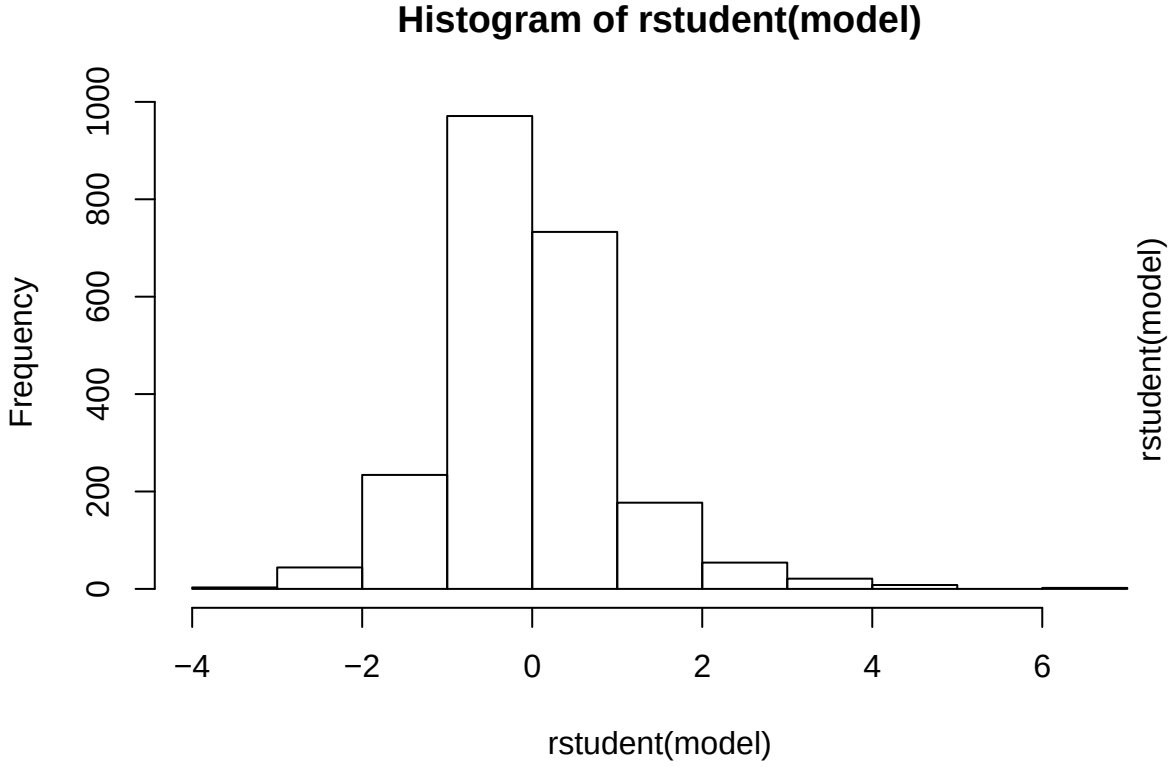
```
m_adjusted_nolog =  lm(Testosterone~ RegMarij + Sex + Agec + Race1  + Depress + Weightc + Heightc +
            Diabete + HypT + TotCholc,
         data = nh2)
summary(m_adjusted_nolog)
```

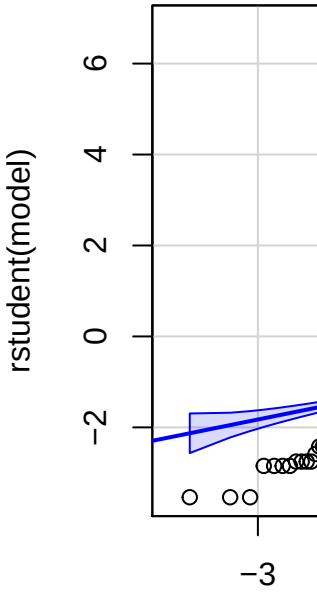
```
##
## Call:
## lm(formula = Testosterone ~ RegMarij + Sex + Agec + Race1 + Depress +
##     Weightc + Heightc + Diabete + HypT + TotCholc, data = nh2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -415.38  -53.31   -6.88   40.59  798.41
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    435.119      9.059  48.032  < 2e-16 ***
## RegMarij        17.009      5.649   3.011 0.002633 **
## Sex           -387.711      7.057 -54.938  < 2e-16 ***
## Agec            -9.661      2.783  -3.471 0.000528 ***
## Race1Hispanic  -29.419     12.657  -2.324 0.020201 *
## Race1Mexican   -24.401     11.423  -2.136 0.032778 *
## Race1White     -26.143      8.230  -3.177 0.001510 **
## Race1Other     -32.226     11.679  -2.759 0.005838 **
## Depress         -3.631      6.275  -0.579 0.562889
## Weightc        -35.945      2.872 -12.515  < 2e-16 ***
## Heightc         17.223      3.809   4.522 6.45e-06 ***
## Diabete          5.109     10.894   0.469 0.639120
## HypT            -8.305      5.830  -1.425 0.154437
## TotCholc        -2.453      2.706  -0.906 0.364847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 118.1 on 2233 degrees of freedom
## Multiple R-squared:  0.737,  Adjusted R-squared:  0.7355
## F-statistic: 481.4 on 13 and 2233 DF,  p-value: < 2.2e-16
```

```
checka(m_adjusted_nolog) ##Constant variance. We think the two clusters are due to sex. Also trend with
```


Added−Variable Plots
Testosterone | others
RegMarij | others
Sex | others
Agec | others
Race1Hispanic | others
Race1Mexican | others
Race1White | others
Race1Other | others
Depress | others
Weightc | others
Heightc | others
Diabete | others
HypT | others
TotCholc | others







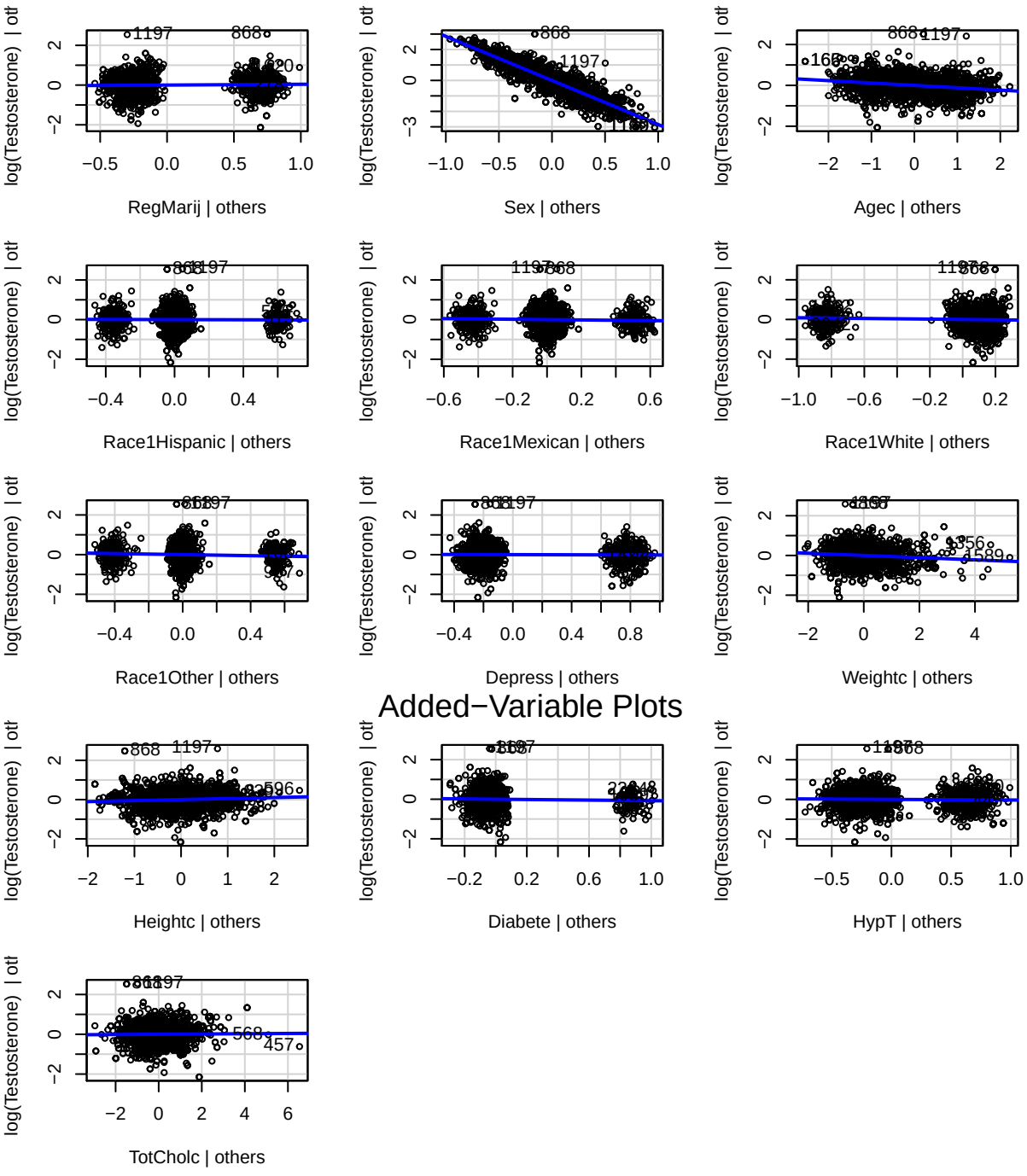

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(model)
## W = 0.92958, p-value < 2.2e-16
```

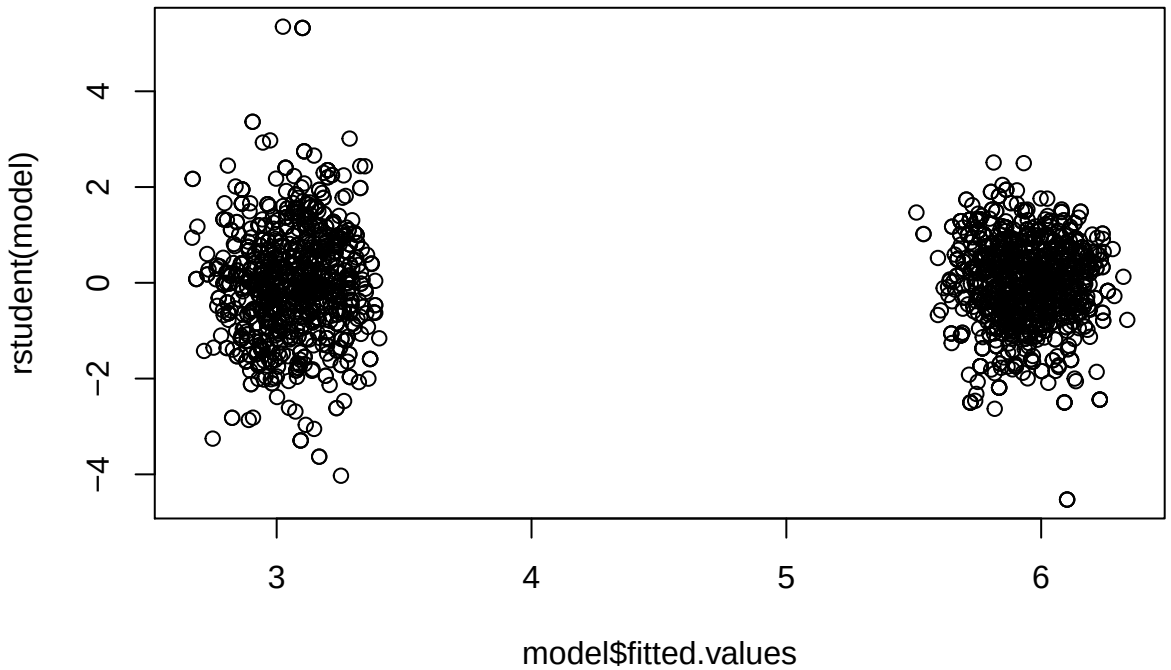
Both models above had linear trend in plots, so we adjusted model log(Y)-THIS IS FINAL MAIN EFFECTS MODEL

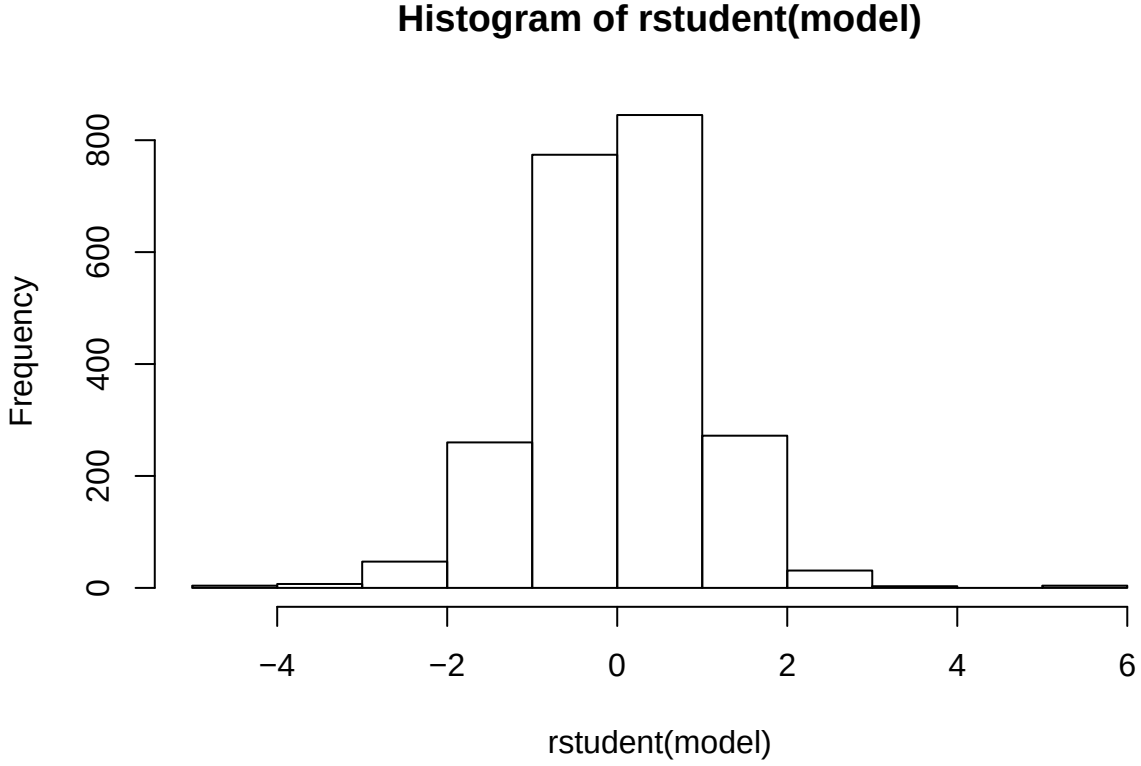
```
m_adjust <- lm(log(Testosterone)~ RegMarij + Sex + Agec + Race1  + Depress + Weightc + Heightc +
             Diabete + HypT + TotCholc, data = nh2)
summary(m_adjust)
```

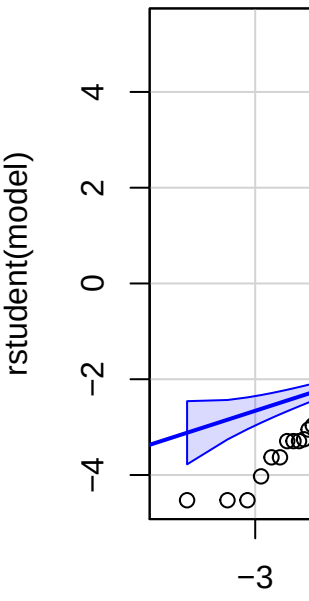
```
##
## Call:
## lm(formula = log(Testosterone) ~ RegMarij + Sex + Agec + Race1 +
##     Depress + Weightc + Heightc + Diabete + HypT + TotCholc,
##     data = nh2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16195 -0.27606  0.01975  0.30368  2.55213
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.008396   0.036897 162.844  < 2e-16 ***
## RegMarij       0.040059   0.023008   1.741 0.081804 .
## Sex           -2.849736   0.028744 -99.143  < 2e-16 ***
## Agec          -0.115816   0.011337 -10.216  < 2e-16 ***
## Race1Hispanic -0.022528   0.051553  -0.437 0.662164
## Race1Mexican  -0.085709   0.046525  -1.842 0.065577 .
## Race1White    -0.086492   0.033519  -2.580 0.009932 **
## Race1Other    -0.135437   0.047566  -2.847 0.004449 **
## Depress       -0.018457   0.025556  -0.722 0.470249
## Weightc       -0.055488   0.011698  -4.743 2.23e-06 ***
## Heightc        0.051861   0.015513   3.343 0.000843 ***
## Diabete       -0.075805   0.044369  -1.709 0.087681 .
## HypT          -0.032210   0.023745  -1.357 0.175073
## TotCholc       0.006758   0.011023   0.613 0.539866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4808 on 2233 degrees of freedom
## Multiple R-squared:  0.9004, Adjusted R-squared:  0.8998
## F-statistic:  1552 on 13 and 2233 DF,  p-value: < 2.2e-16
```

```
checka(m_adjust)
```


Added−Variable Plots
log(Testosterone) | others
RegMarij | others
Sex | others
Agec | others
Race1Hispanic | others
Race1Mexican | others
Race1White | others
Race1Other | others
Depress | others
Weightc | others
Heightc | others
Diabete | others
HypT | others
TotCholc | others








```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(model)
## W = 0.98214, p-value = 3.535e-16
```

Due to constant variance plot having clusters and literature, we think that TT distribution by sex underlies the trend. Test this by stratifying model. See that in stratified models, this is elimnated

```
nhFemale = nh2[ which(nh2$Sex == 1), ]
nhMale = nh2[ which(nh2$Sex == 0), ]
m_F = lm(log(Testosterone)~ RegMarij + Agec + Race1  + Depress + Weightc + Heightc +
             Diabete + HypT + TotCholc,
         data = nhFemale)
summary(m_F)
```

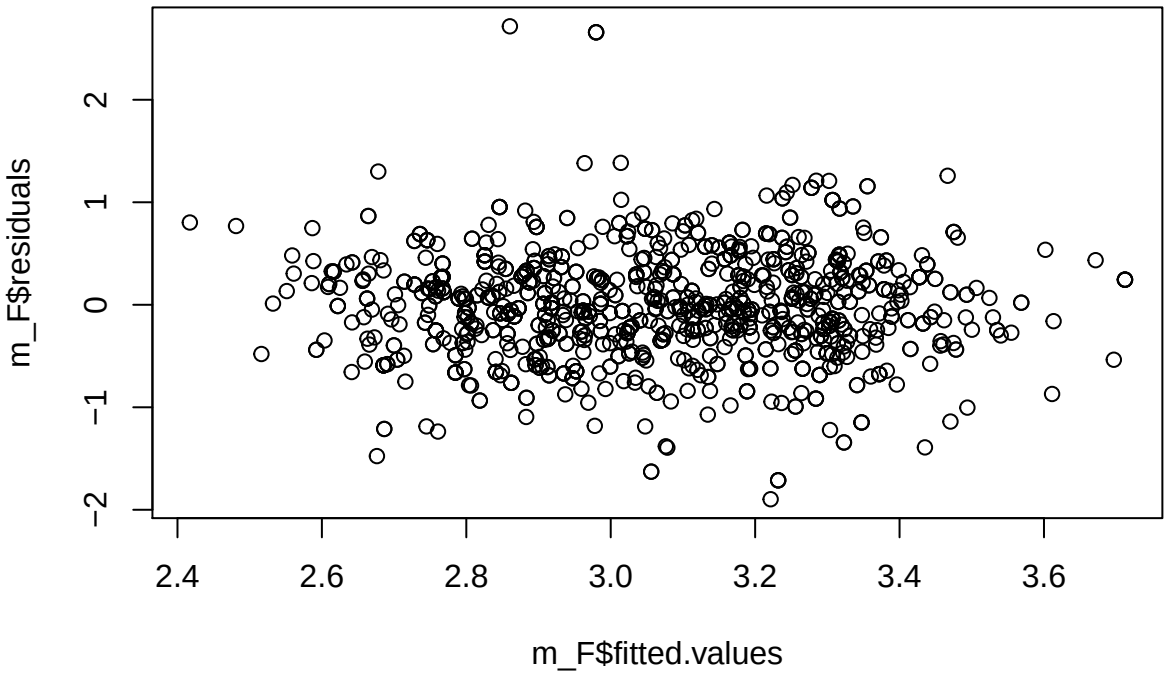
```
##
## Call:
## lm(formula = log(Testosterone) ~ RegMarij + Agec + Race1 + Depress +
##     Weightc + Heightc + Diabete + HypT + TotCholc, data = nhFemale)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89704 -0.29545 -0.00775  0.30405  2.71637
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.1132300  0.0512814  60.709  < 2e-16 ***
## RegMarij       0.0195351  0.0392456   0.498 0.618755
## Agec          -0.2075484  0.0184817 -11.230  < 2e-16 ***
## Race1Hispanic  0.1780629  0.0797391   2.233 0.025758 *
## Race1Mexican  -0.0154863  0.0738236  -0.210 0.833885
## Race1White    -0.0076547  0.0506123  -0.151 0.879814
## Race1Other    -0.0596064  0.0765196  -0.779 0.436177
## Depress       -0.0531885  0.0388554  -1.369 0.171334
## Weightc        0.0864876  0.0183900   4.703 2.91e-06 ***
## Heightc       -0.0009671  0.0253610  -0.038 0.969590
## Diabete       -0.2989161  0.0784031  -3.813 0.000146 ***
## HypT          -0.0323136  0.0422939  -0.764 0.445027
## TotCholc       0.0357328  0.0175339   2.038 0.041811 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5206 on 1032 degrees of freedom
## Multiple R-squared:  0.169,  Adjusted R-squared:  0.1594
## F-statistic: 17.49 on 12 and 1032 DF,  p-value: < 2.2e-16
```

```
plot(m_F$fitted.values, m_F$residuals)
m_M = lm(log(Testosterone)~ RegMarij + Agec + Race1  + Depress + Weightc + Heightc +
             Diabete + HypT + TotCholc,
         data = nhMale)
summary(m_M)
```

```
##
## Call:
## lm(formula = log(Testosterone) ~ RegMarij + Agec + Race1 + Depress +
##     Weightc + Heightc + Diabete + HypT + TotCholc, data = nhMale)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.23100 -0.24223  0.03502  0.27506  1.20166
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.01093    0.04194 143.316  < 2e-16 ***
## RegMarij       0.04066    0.02481   1.639   0.1015
## Agec          -0.02981    0.01274  -2.340   0.0194 *
## Race1Hispanic -0.10475    0.06042  -1.734   0.0832 .
## Race1Mexican  -0.06914    0.05363  -1.289   0.1976
## Race1White    -0.09479    0.04040  -2.346   0.0191 *
## Race1Other    -0.09599    0.05439  -1.765   0.0778 .
## Depress       -0.02035    0.03039  -0.670   0.5033
## Weightc       -0.19147    0.01367 -14.009  < 2e-16 ***
## Heightc        0.10140    0.01754   5.780 9.52e-09 ***
## Diabete        0.01353    0.04746   0.285   0.7756
## HypT          -0.02100    0.02503  -0.839   0.4017
## TotCholc      -0.01173    0.01258  -0.932   0.3514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3985 on 1189 degrees of freedom
## Multiple R-squared:  0.1723, Adjusted R-squared:  0.1639
## F-statistic: 20.62 on 12 and 1189 DF,  p-value: < 2.2e-16
```

```
plot(m_F$fitted.values, m_F$residuals)
```



Multicolinearity check

```
car::vif(m_adjusted_nolog)
```

```
##              GVIF Df GVIF^(1/(2*Df))
```

```
## RegMarij 1.039990  1        1.019799
## Sex      1.997513  1        1.413334
## Agec     1.248448  1        1.117340
## Race1    1.214469  4        1.024586
## Depress  1.028377  1        1.014089
## Weightc  1.329264  1        1.152937
## Heightc  2.337791  1        1.528983
## Diabete  1.072891  1        1.035805
## HypT     1.122416  1        1.059441
## TotCholc 1.180220  1        1.086379
```

##Adding interactions. Based on analysis and stratified models, we assume Sex has effect modifcation on variables.

```
m_int <- lm(lm(log(Testosterone)~ RegMarij + Sex + Agec + Race1  + Depress + Weightc + Heightc +
             Diabete + HypT + TotCholc + Sex*(RegMarij + Agec + Race1  + Depress + Weightc + Heightc +
             Diabete + HypT + TotCholc),  data = nh2))
summary(m_int)
```

```
##
## Call:
## lm(formula = lm(log(Testosterone) ~ RegMarij + Sex + Agec + Race1 +
##     Depress + Weightc + Heightc + Diabete + HypT + TotCholc +
##     Sex * (RegMarij + Agec + Race1 + Depress + Weightc + Heightc +
##         Diabete + HypT + TotCholc), data = nh2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23100 -0.25470  0.02054  0.29052  2.71637
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.01093    0.04834 124.342  < 2e-16 ***
## RegMarij           0.04066    0.02859   1.422 0.155140
## Sex               -2.89770    0.06621 -43.766  < 2e-16 ***
## Agec              -0.02981    0.01468  -2.030 0.042431 *
## Race1Hispanic     -0.10475    0.06964  -1.504 0.132688
## Race1Mexican      -0.06914    0.06182  -1.118 0.263496
## Race1White        -0.09479    0.04657  -2.035 0.041930 *
## Race1Other        -0.09599    0.06269  -1.531 0.125846
## Depress           -0.02035    0.03503  -0.581 0.561384
## Weightc           -0.19147    0.01575 -12.154  < 2e-16 ***
## Heightc            0.10140    0.02022   5.015 5.72e-07 ***
## Diabete            0.01353    0.05470   0.247 0.804658
## HypT              -0.02100    0.02885  -0.728 0.466837
## TotCholc          -0.01173    0.01450  -0.809 0.418701
## RegMarij:Sex      -0.02113    0.04490  -0.470 0.638056
## Sex:Agec          -0.17774    0.02194  -8.102 8.85e-16 ***
## Sex:Race1Hispanic  0.28281    0.09899   2.857 0.004315 **
## Sex:Race1Mexican   0.05365    0.08979   0.598 0.550231
## Sex:Race1White     0.08713    0.06452   1.351 0.176975
## Sex:Race1Other     0.03638    0.09212   0.395 0.692922
## Sex:Depress       -0.03284    0.04901  -0.670 0.502892
## Sex:Weightc        0.27796    0.02261  12.291  < 2e-16 ***
## Sex:Heightc       -0.10237    0.03016  -3.395 0.000699 ***
```

```
## Sex:Diabete       -0.31245    0.08818  -3.543 0.000404 ***
## Sex:HypT          -0.01132    0.04716  -0.240 0.810389
## Sex:TotCholc       0.04746    0.02120   2.239 0.025282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4593 on 2221 degrees of freedom
## Multiple R-squared:  0.9096, Adjusted R-squared:  0.9086
## F-statistic:   894 on 25 and 2221 DF,  p-value: < 2.2e-16
```

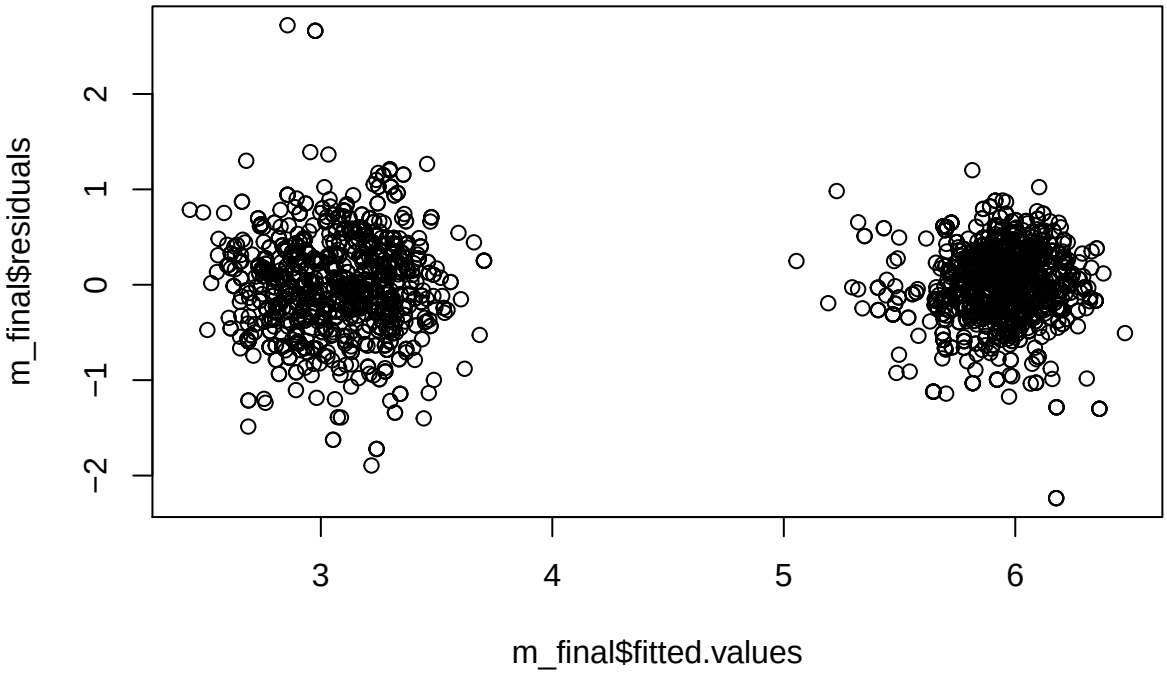
Remove all insig interactions, and checks assumptions (not linearity, as this is checked only for main effects (see above))

```
m_final <- lm(lm(log(Testosterone)~ RegMarij + Sex + Agec + Race1  + Depress + Weightc + Heightc +
            Diabete + HypT + TotCholc + Sex*(Agec + Race1  + RegMarij +  Weightc + Heightc +
            Diabete + TotCholc),  data = nh2))
summary(m_final)
```

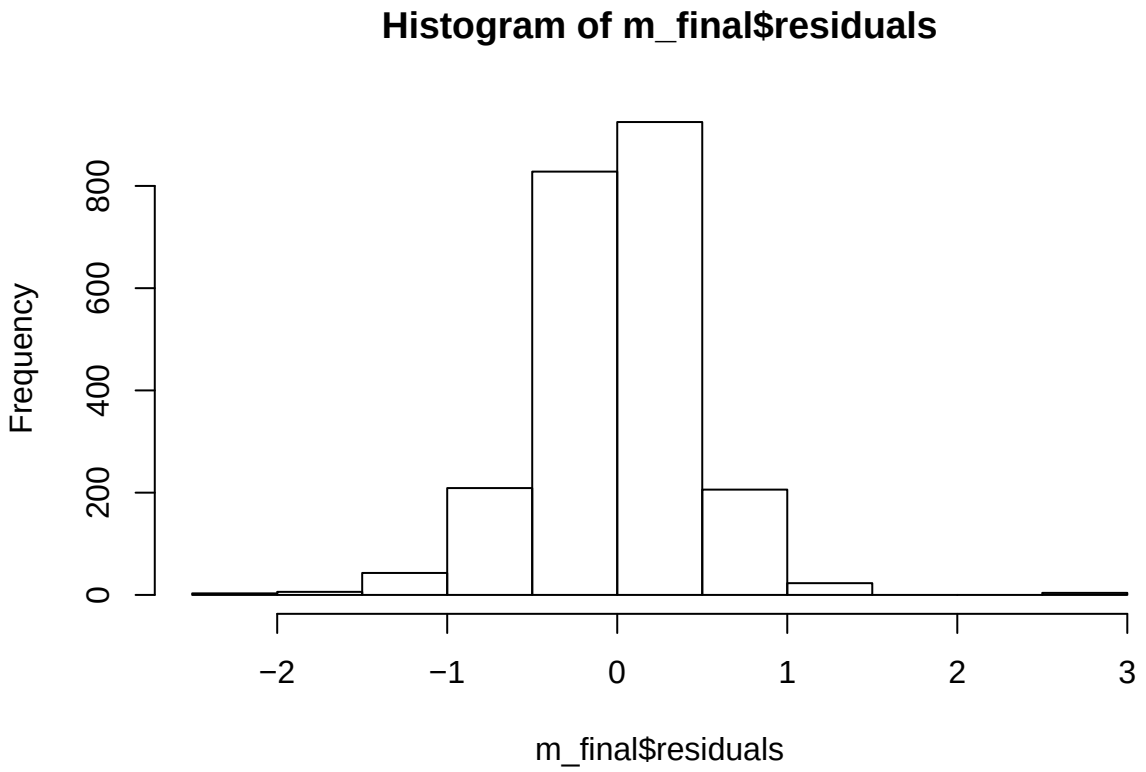
```
##
## Call:
## lm(formula = lm(log(Testosterone) ~ RegMarij + Sex + Agec + Race1 +
##     Depress + Weightc + Heightc + Diabete + HypT + TotCholc +
##     Sex * (Agec + Race1 + RegMarij + Weightc + Heightc + Diabete +
##         TotCholc), data = nh2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23678 -0.25606  0.01645  0.29577  2.72053
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.01414    0.04781 125.788  < 2e-16 ***
## RegMarij           0.04187    0.02852   1.468 0.142192
## Sex               -2.90672    0.06391 -45.482  < 2e-16 ***
## Agec              -0.02924    0.01464  -1.998 0.045867 *
## Race1Hispanic     -0.10328    0.06959  -1.484 0.137889
## Race1Mexican      -0.06812    0.06169  -1.104 0.269563
## Race1White        -0.09392    0.04652  -2.019 0.043613 *
## Race1Other        -0.09510    0.06264  -1.518 0.129087
## Depress           -0.03696    0.02448  -1.510 0.131303
## Weightc           -0.19136    0.01572 -12.173  < 2e-16 ***
## Heightc            0.10096    0.02020   4.997 6.28e-07 ***
## Diabete            0.01495    0.05461   0.274 0.784314
## HypT              -0.02464    0.02280  -1.081 0.279842
## TotCholc          -0.01103    0.01445  -0.763 0.445414
## Sex:Agec          -0.17916    0.02152  -8.324  < 2e-16 ***
## Sex:Race1Hispanic  0.27959    0.09884   2.829 0.004715 **
## Sex:Race1Mexican   0.05260    0.08966   0.587 0.557475
## Sex:Race1White     0.08729    0.06446   1.354 0.175789
## Sex:Race1Other     0.03682    0.09202   0.400 0.689061
## RegMarij:Sex      -0.02366    0.04462  -0.530 0.595962
## Sex:Weightc        0.27701    0.02248  12.323  < 2e-16 ***
## Sex:Heightc       -0.10166    0.03013  -3.374 0.000753 ***
```

```
## Sex:Diabete       -0.31372    0.08813  -3.560 0.000379 ***
## Sex:TotCholc       0.04590    0.02097   2.189 0.028727 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4591 on 2223 degrees of freedom
## Multiple R-squared:  0.9096, Adjusted R-squared:  0.9086
## F-statistic: 972.3 on 23 and 2223 DF,  p-value: < 2.2e-16
```

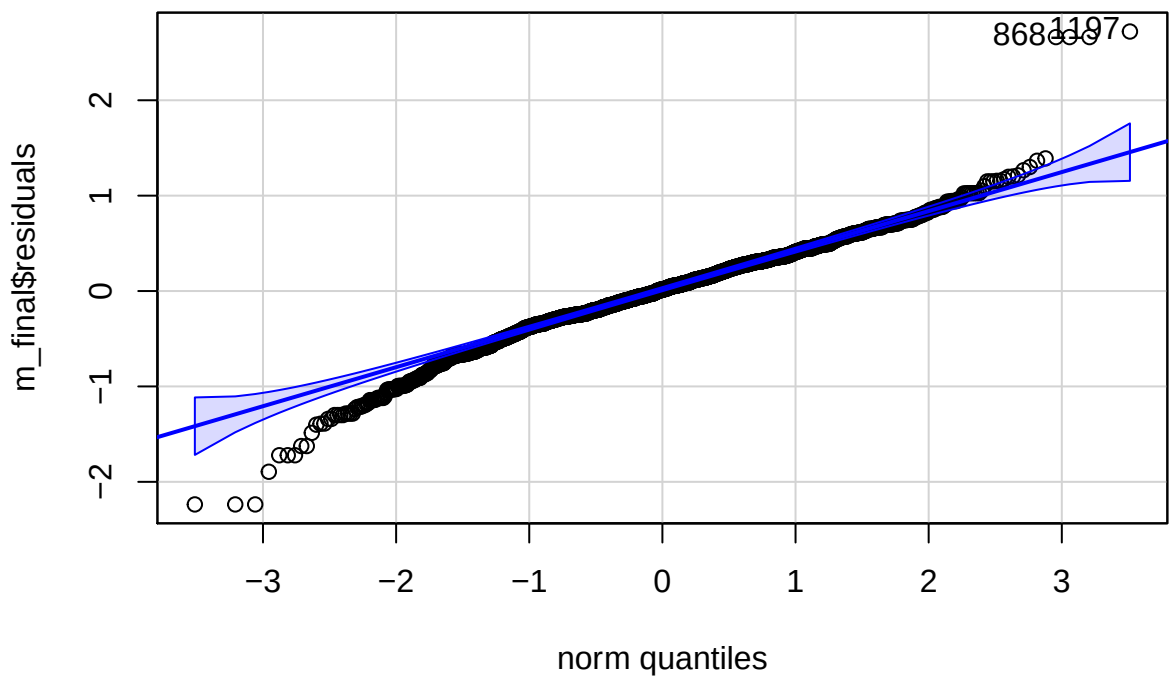
```
plot(m_final$fitted.values, m_final$residuals)
```




```
hist(m_final$residuals)
```




```
car::qqPlot(m_final$residuals)
```



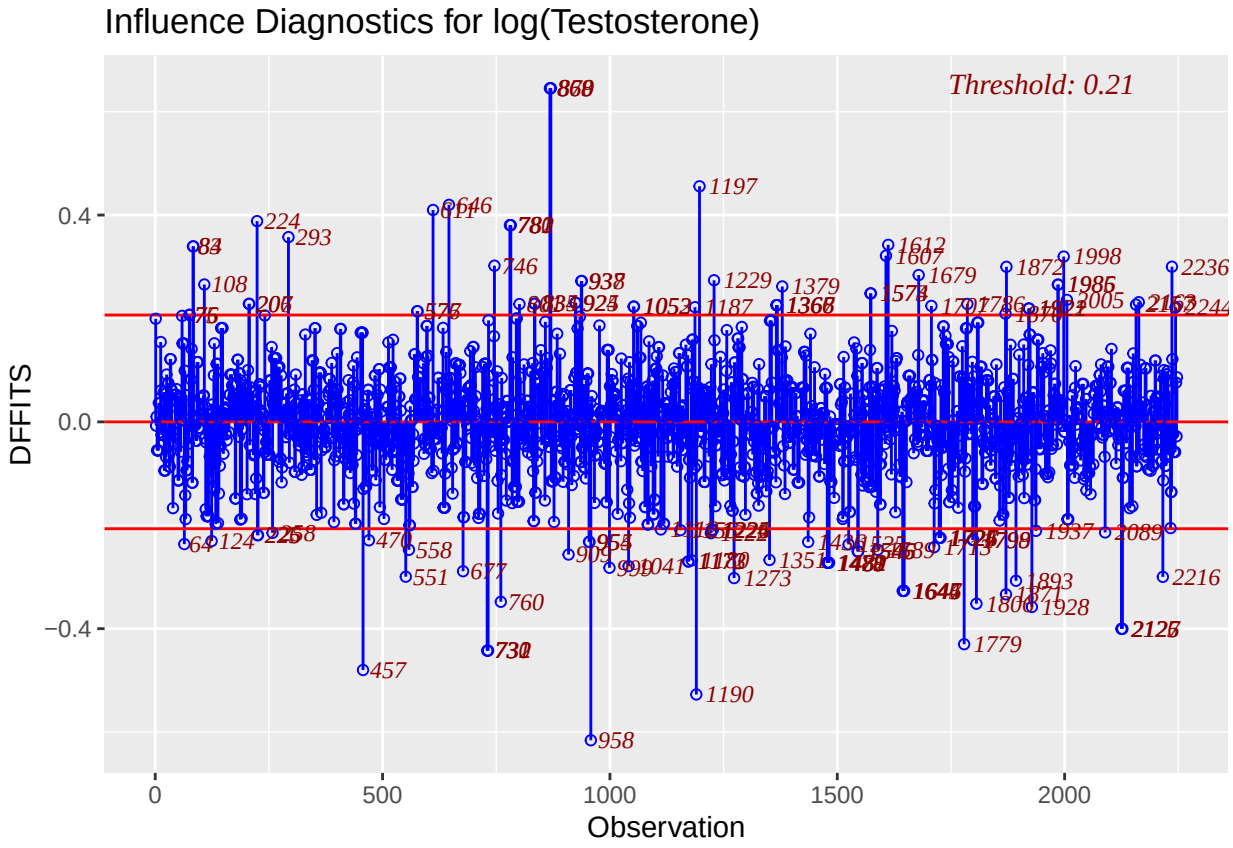

```
## [1] 1197  868
```

```
shapiro.test(m_final$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
```

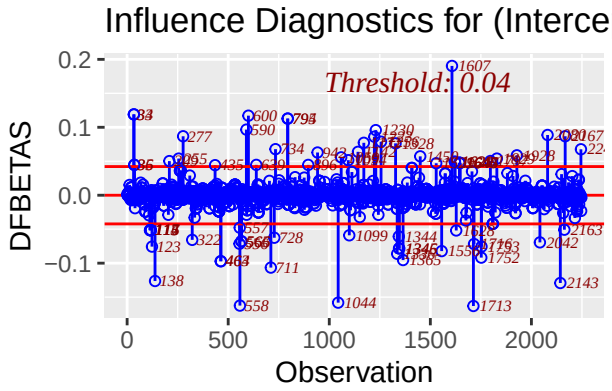
```
## data:  m_final$residuals
## W = 0.97064, p-value < 2.2e-16
```

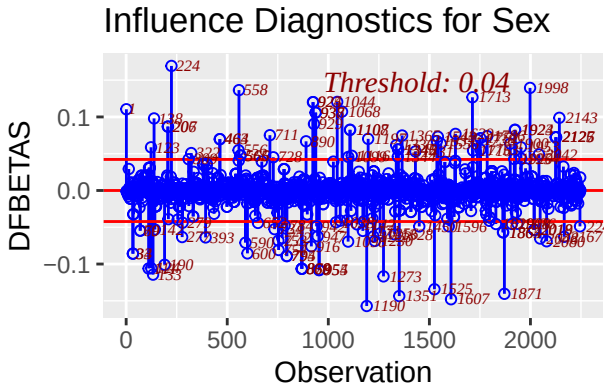
```
olsrr::ols_plot_dffits(m_final)
```

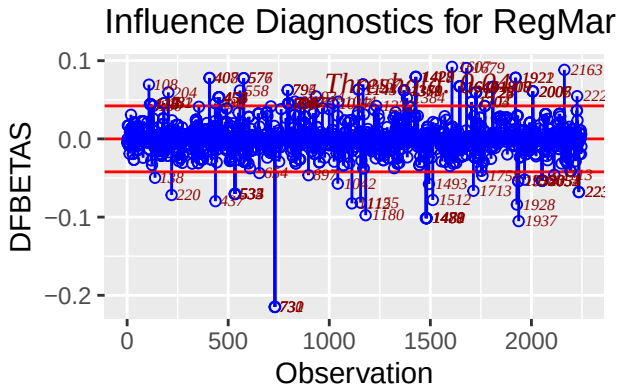


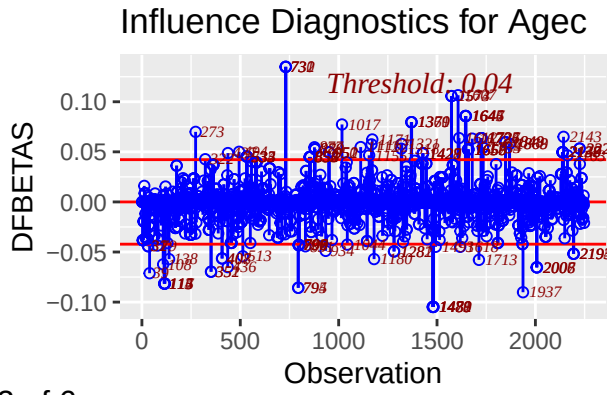

```
olsrr::ols_plot_dfbetas(m_final)
```

page 1 of 6

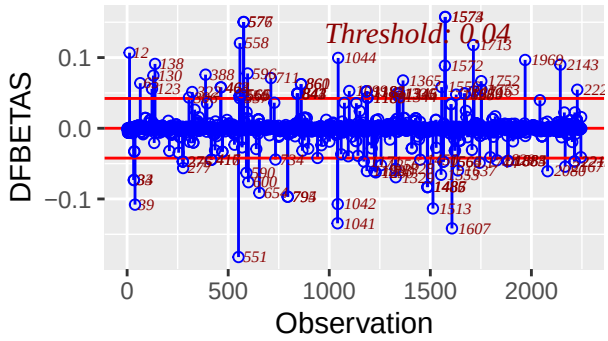


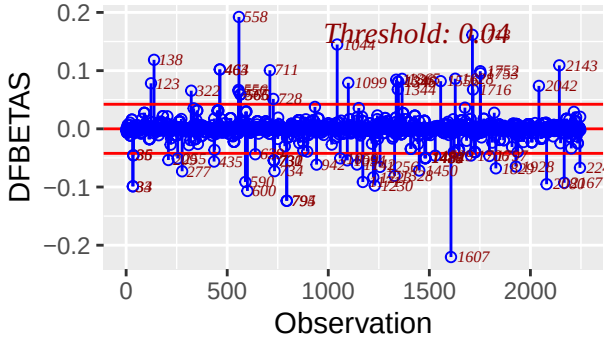


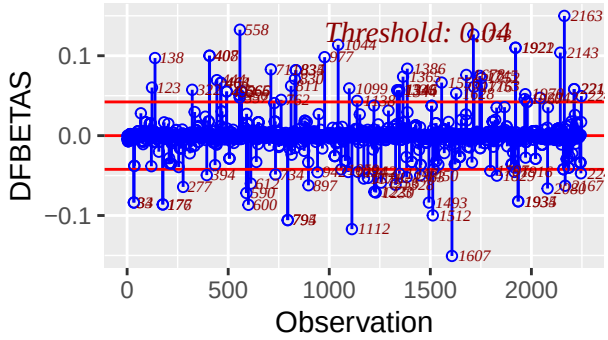


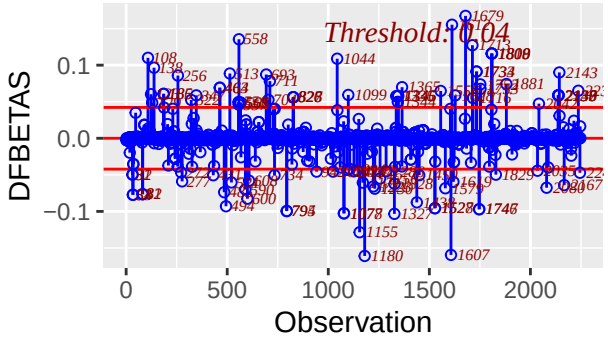


page 2 of 6

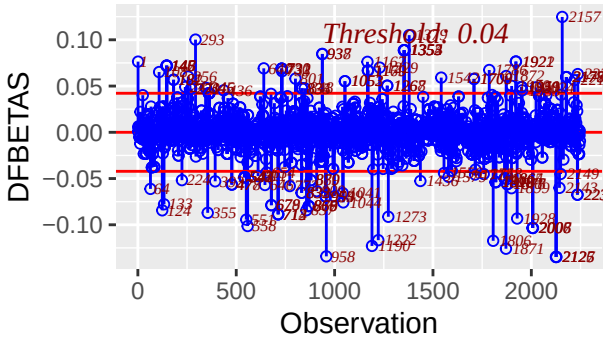




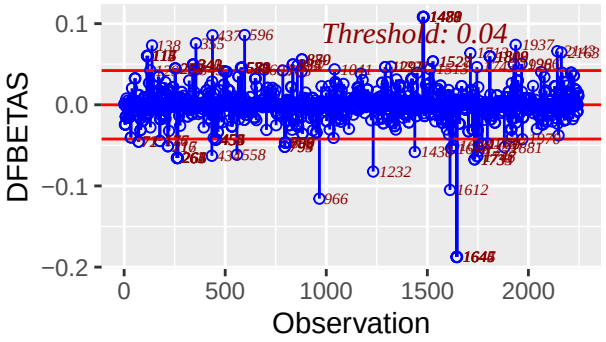




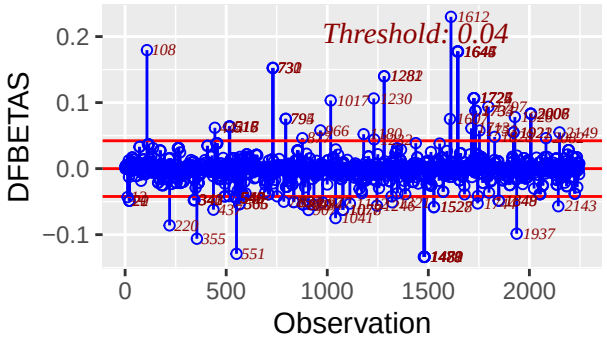
Influence Diagnostics for Depres



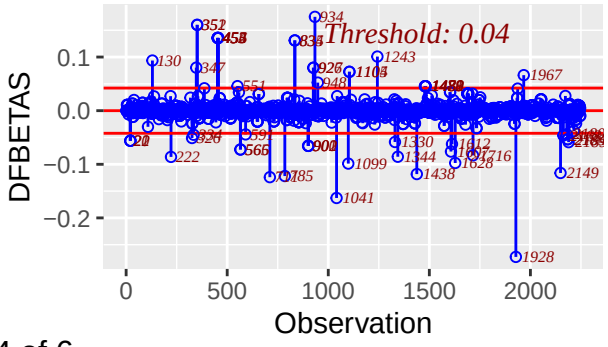
Influence Diagnostics for Heightc



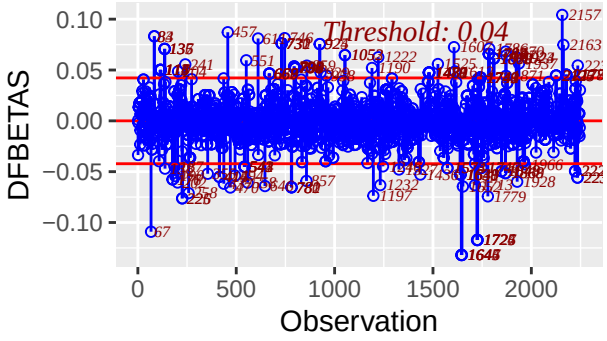
Influence Diagnostics for Weightc



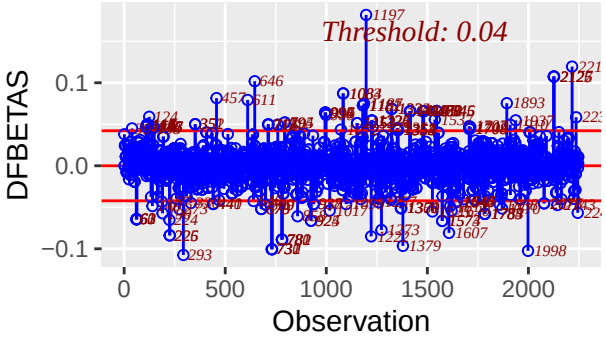
Influence Diagnostics for Diabete



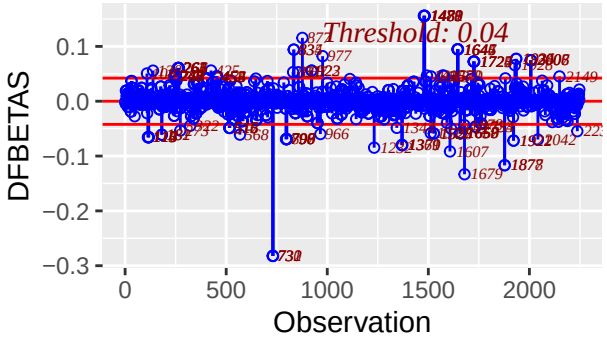
Influence Diagnostics for HypT



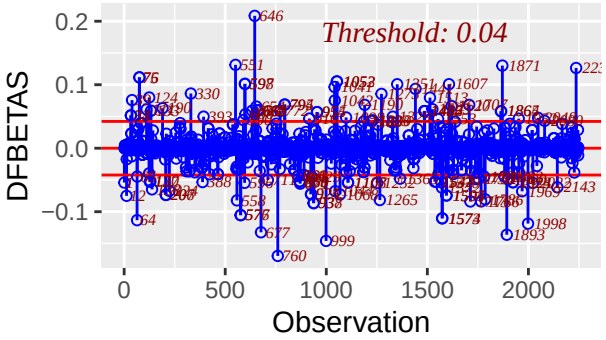
Influence Diagnostics for Sex:Age

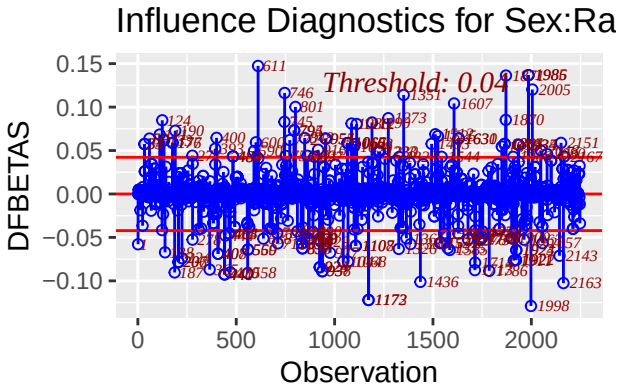


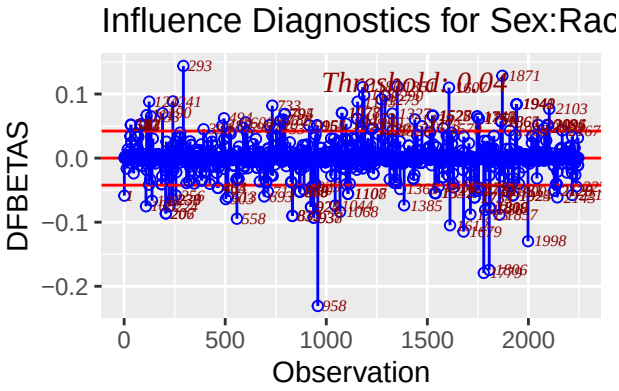
Influence Diagnostics for TotChol

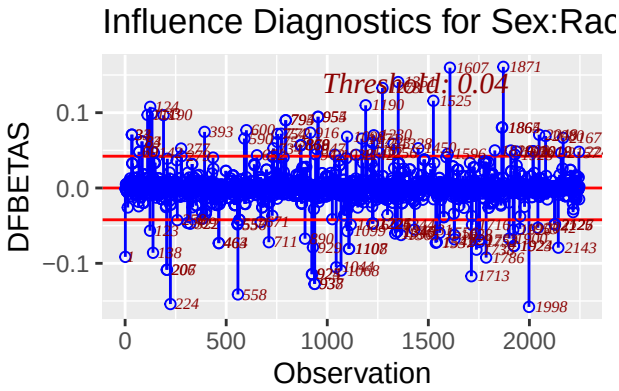


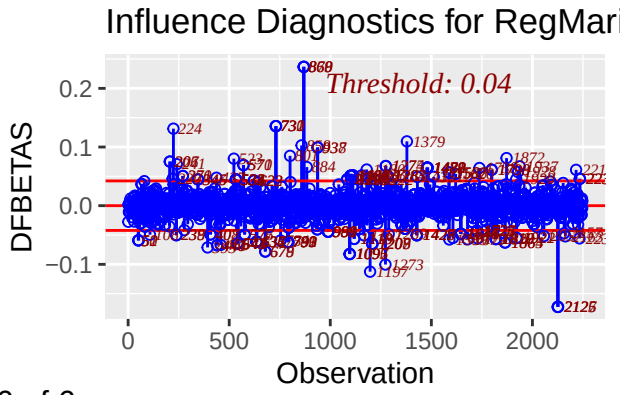
Influence Diagnostics for Sex:Rac

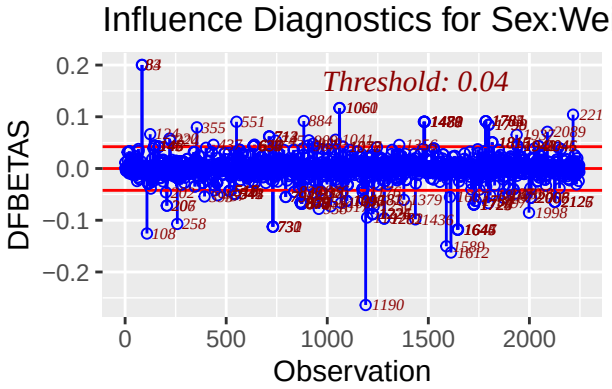


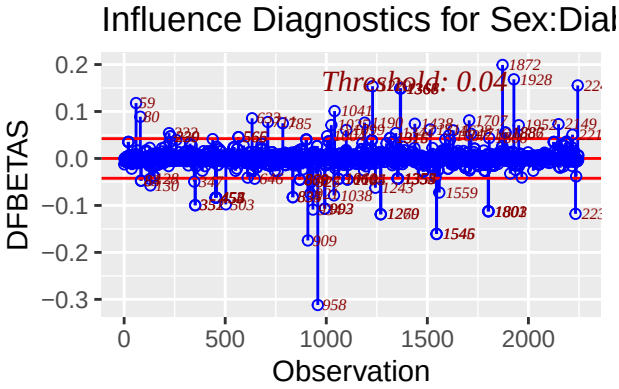

Influence Diagnostics for Sex:Ra
Threshold: 0.04
DFBETAS
Observation

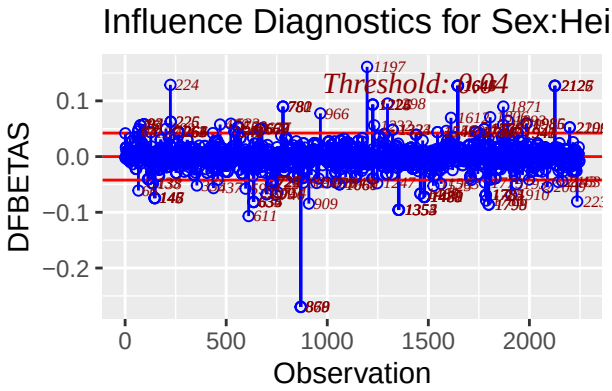

Influence Diagnostics for Sex:Rac
Threshold: 0.04
DFBETAS
Observation

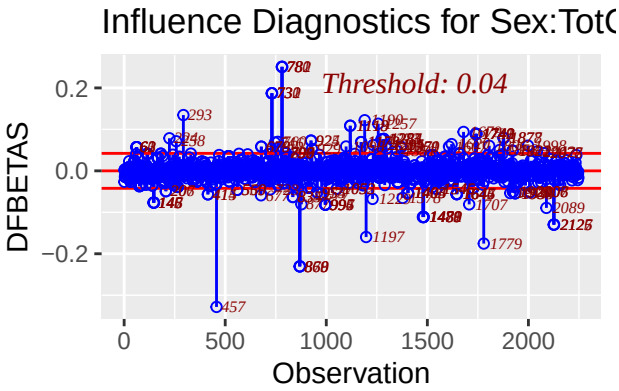

Influence Diagnostics for Sex:Rac
Threshold: 0.04
DFBETAS
Observation


Influence Diagnostics for RegMar
Threshold: 0.04
DFBETAS
Observation

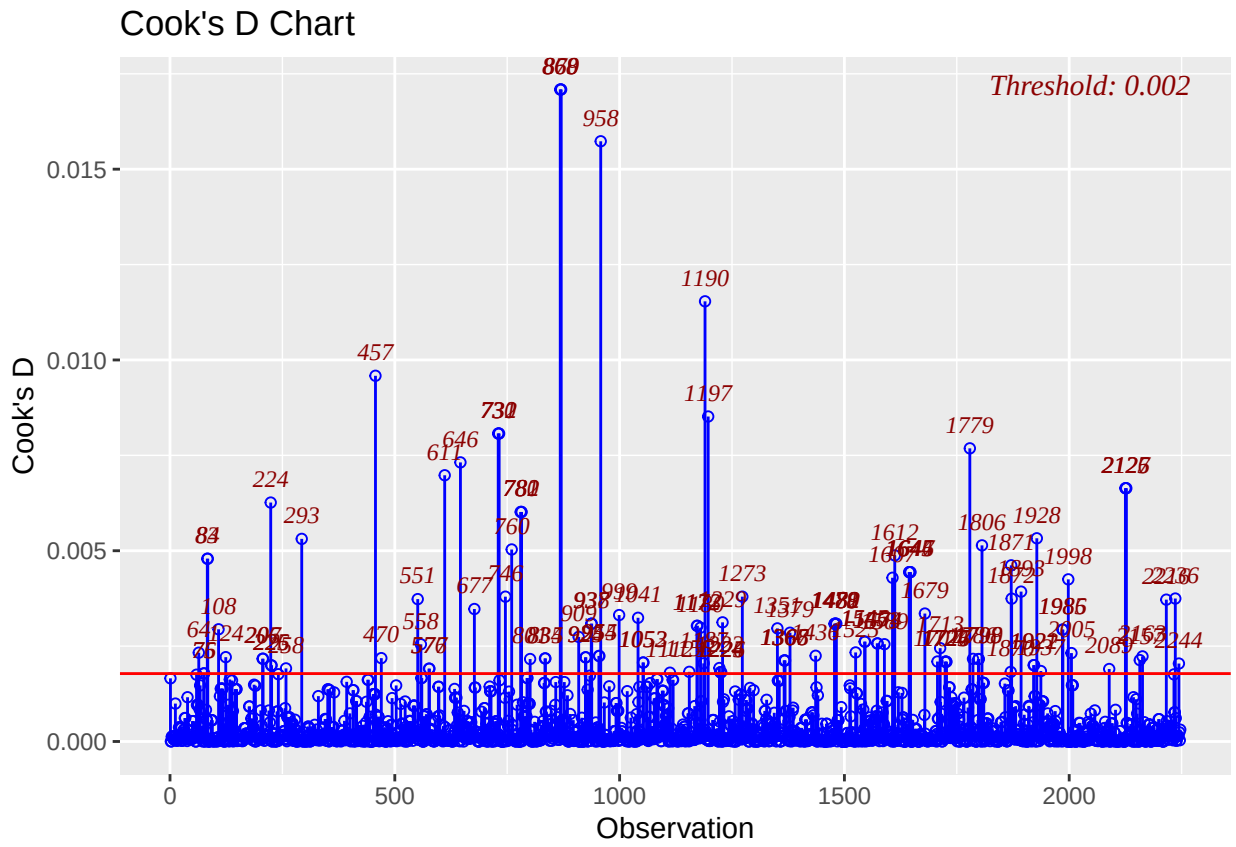

Influence Diagnostics for Sex:Wei
Threshold: 0.04
DFBETAS
Observation


Influence Diagnostics for Sex:Dial
Threshold: 0.04
DFBETAS
Observation

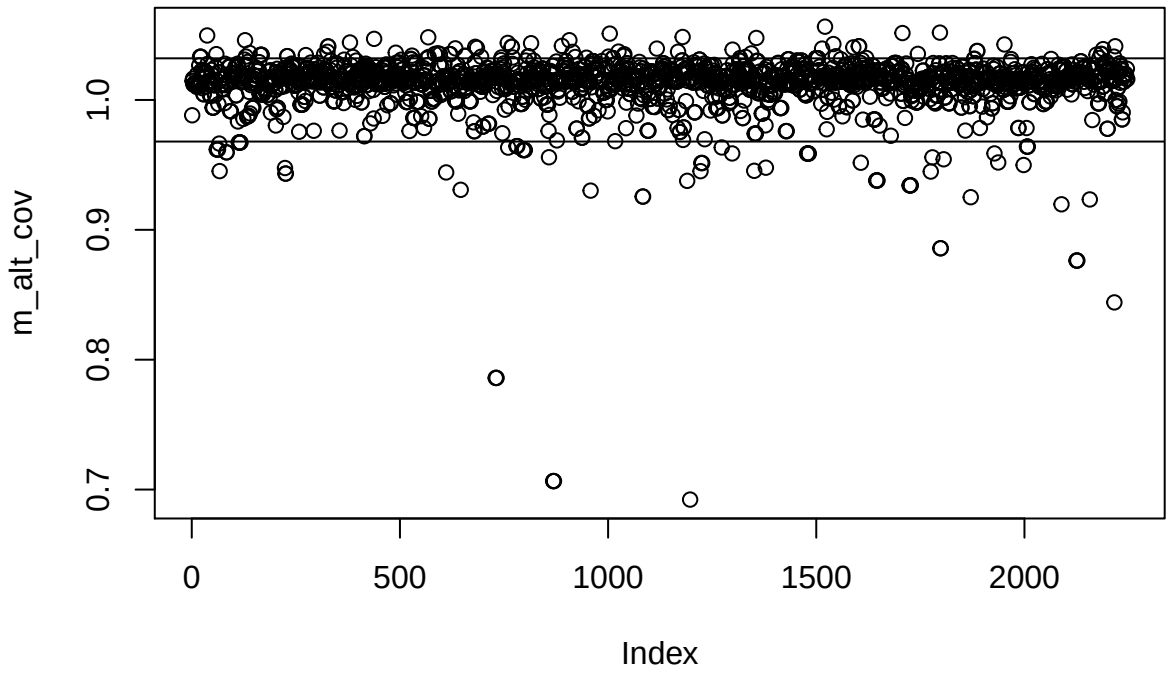

Influence Diagnostics for Sex:Hei
Threshold: 0.04
DFBETAS
Observation


Influence Diagnostics for Sex:Tot
Threshold: 0.04
DFBETAS
Observation

```
olsrr::ols_plot_cooksd_chart(m_final)
```




```
m_alt_cov = covratio(m_final)
n = nrow(nh2)
p = m_final$rank
plot(m_alt_cov); abline(1+3*p/n,0); abline(1-3*p/n,0)
```



For discussion, in main effects model, test if mean test is same for race groups (not something we need to include, but idk, we can get rid of, just thought it was a way to utilize GLH learned in class)

```
m_wo_its =  lm(log(Testosterone)~ -1 + RegMarij + Sex + Agec + Race1  + Depress + Weightc + Heightc +
             Diabete + HypT + TotCholc, data = nh2)
summary(m_wo_its)
```

```
##
## Call:
## lm(formula = log(Testosterone) ~ -1 + RegMarij + Sex + Agec +
##     Race1 + Depress + Weightc + Heightc + Diabete + HypT + TotCholc,
##     data = nh2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16195 -0.27606  0.01975  0.30368  2.55213
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## RegMarij       0.040059   0.023008   1.741 0.081804 .
## Sex           -2.849736   0.028744 -99.143  < 2e-16 ***
## Agec          -0.115816   0.011337 -10.216  < 2e-16 ***
## Race1Black     6.008396   0.036897 162.844  < 2e-16 ***
## Race1Hispanic  5.985868   0.043338 138.120  < 2e-16 ***
## Race1Mexican   5.922687   0.036291 163.201  < 2e-16 ***
## Race1White     5.921905   0.022836 259.326  < 2e-16 ***
## Race1Other     5.872959   0.038662 151.906  < 2e-16 ***
## Depress       -0.018457   0.025556  -0.722 0.470249
## Weightc       -0.055488   0.011698  -4.743 2.23e-06 ***
## Heightc        0.051861   0.015513   3.343 0.000843 ***
```

```
## Diabete       -0.075805   0.044369  -1.709 0.087681 .
## HypT          -0.032210   0.023745  -1.357 0.175073
## TotCholc       0.006758   0.011023   0.613 0.539866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4808 on 2233 degrees of freedom
## Multiple R-squared:  0.9902, Adjusted R-squared:  0.9901
## F-statistic: 1.613e+04 on 14 and 2233 DF,  p-value: < 2.2e-16
```

```
#Test if all five are same
Contrast.T = matrix(c(0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0
car::linearHypothesis(model=m_wo_its,hypothesis.matrix=Contrast.T, rsh=c(0,0,0,0))
```

```
## Linear hypothesis test
##
## Hypothesis:
## Race1Black - Race1Hispanic = 0
## Race1Hispanic - Race1Mexican = 0
## Race1Mexican - Race1White = 0
## Race1White - Race1Other = 0
##
## Model 1: restricted model
## Model 2: log(Testosterone) ~ -1 + RegMarij + Sex + Agec + Race1 + Depress +
##     Weightc + Heightc + Diabete + HypT + TotCholc
##
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   2237 518.9
## 2   2233 516.3  4     2.608 2.8199 0.02383 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Test if (4) are same
Contrast.T = matrix(c(0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0
car::linearHypothesis(model=m_wo_its,hypothesis.matrix=Contrast.T, rsh=c(0,0,0))
```

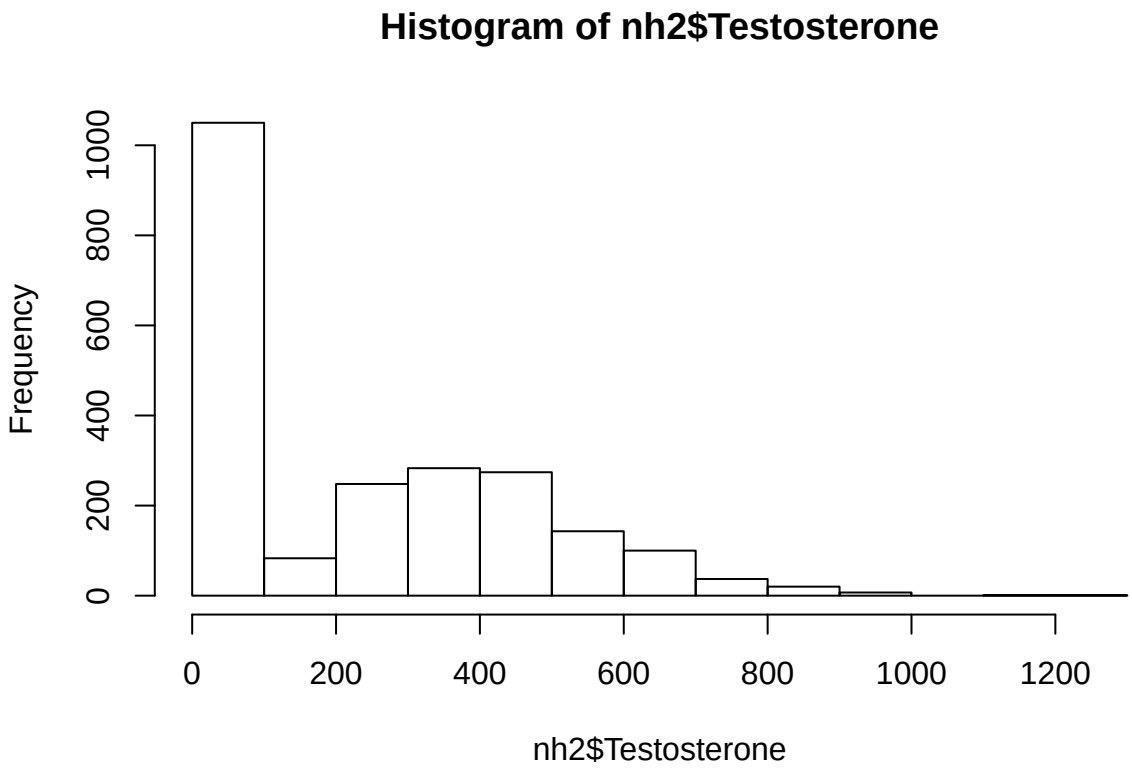
```
## Linear hypothesis test
##
## Hypothesis:
## Race1Hispanic - Race1Mexican = 0
## Race1Mexican - Race1White = 0
## Race1White - Race1Other = 0
##
## Model 1: restricted model
## Model 2: log(Testosterone) ~ -1 + RegMarij + Sex + Agec + Race1 + Depress +
##     Weightc + Heightc + Diabete + HypT + TotCholc
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   2236 517.31
## 2   2233 516.30  3    1.0171 1.4664 0.2218
```

Additional Analysis (Leah)

```
summary(nh2$Testosterone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.32   22.25  188.02  232.94  405.88 1244.73
```

```
hist(nh2$Testosterone)
```




```
m_int <- lm(lm(log(Testosterone)~ RegMarij + Sex + Agec + Race1
               + Depress + Weightc + Heightc + Diabete + HypT +
                 TotCholc + Sex*RegMarij,  data = nh2))
summary(m_int)
```

```
##
## Call:
## lm(formula = lm(log(Testosterone) ~ RegMarij + Sex + Agec + Race1 +
##     Depress + Weightc + Heightc + Diabete + HypT + TotCholc +
##     Sex * RegMarij, data = nh2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17857 -0.28028  0.01832  0.30216  2.56299
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.99897    0.03760 159.554  < 2e-16 ***
## RegMarij       0.06454    0.02975   2.170 0.030141 *
## Sex           -2.83372    0.03128 -90.603  < 2e-16 ***
## Agec          -0.11597    0.01134 -10.230  < 2e-16 ***
```

```
## Race1Hispanic -0.02056    0.05157  -0.399 0.690125
## Race1Mexican  -0.08471    0.04652  -1.821 0.068782 .
## Race1White    -0.08435    0.03355  -2.514 0.012009 *
## Race1Other    -0.13461    0.04756  -2.830 0.004694 **
## Depress       -0.01783    0.02556  -0.698 0.485405
## Weightc       -0.05528    0.01170  -4.726 2.43e-06 ***
## Heightc        0.05176    0.01551   3.337 0.000861 ***
## Diabete       -0.07809    0.04440  -1.759 0.078727 .
## HypT          -0.03304    0.02375  -1.391 0.164293
## TotCholc       0.00649    0.01102   0.589 0.556081
## RegMarij:Sex  -0.06016    0.04634  -1.298 0.194403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4808 on 2232 degrees of freedom
## Multiple R-squared:  0.9005, Adjusted R-squared:  0.8998
## F-statistic:  1442 on 14 and 2232 DF,  p-value: < 2.2e-16
```