BIOSTAT 650 Project

Jaehoon Kim (Group 19)

2024-11-17

```
df = NHANES
```

Initial data exploration of covariates that had a relation to SexAge were difficult to perform a correlation plot due to being factors.

```
covariates = c("SexAge","Gender","HHIncome","Education","PhysActive","SameSex","AlcoholYear","RegularMa
sapply(df[, covariates], is.factor)
```

```
##       SexAge       Gender     HHIncome    Education   PhysActive      SameSex 
##        FALSE         TRUE         TRUE         TRUE         TRUE         TRUE 
##  AlcoholYear RegularMarij    HardDrugs 
##        FALSE         TRUE         TRUE
```

```
#M = cor(df[, covariates])
#corrplot(M, method = 'number')
```

Running different multiple linear regressions, we found two models of interest after some exploratory data analysis with different covariates for which statistical significance persisted even after controlling for some social demographic covariates.

```
model <- lm(SexAge ~ RegularMarij+HardDrugs+RegularMarij*HardDrugs, df)
summary(model)
```

```
## 
## Call:
## lm(formula = SexAge ~ RegularMarij + HardDrugs + RegularMarij * 
##     HardDrugs, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.0399 -2.0399 -0.3123  1.1842 28.9601 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  18.03995    0.06268 287.823  < 2e-16 ***
## RegularMarijYes              -2.22420    0.14750 -15.080  < 2e-16 ***
## HardDrugsYes                 -1.72766    0.20925  -8.256  < 2e-16 ***
## RegularMarijYes:HardDrugsYes  1.44824    0.28116   5.151  2.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.464 on 4712 degrees of freedom
##   (5284 observations deleted due to missingness)
## Multiple R-squared:  0.08977,    Adjusted R-squared:  0.08919 
## F-statistic: 154.9 on 3 and 4712 DF,  p-value: < 2.2e-16
```

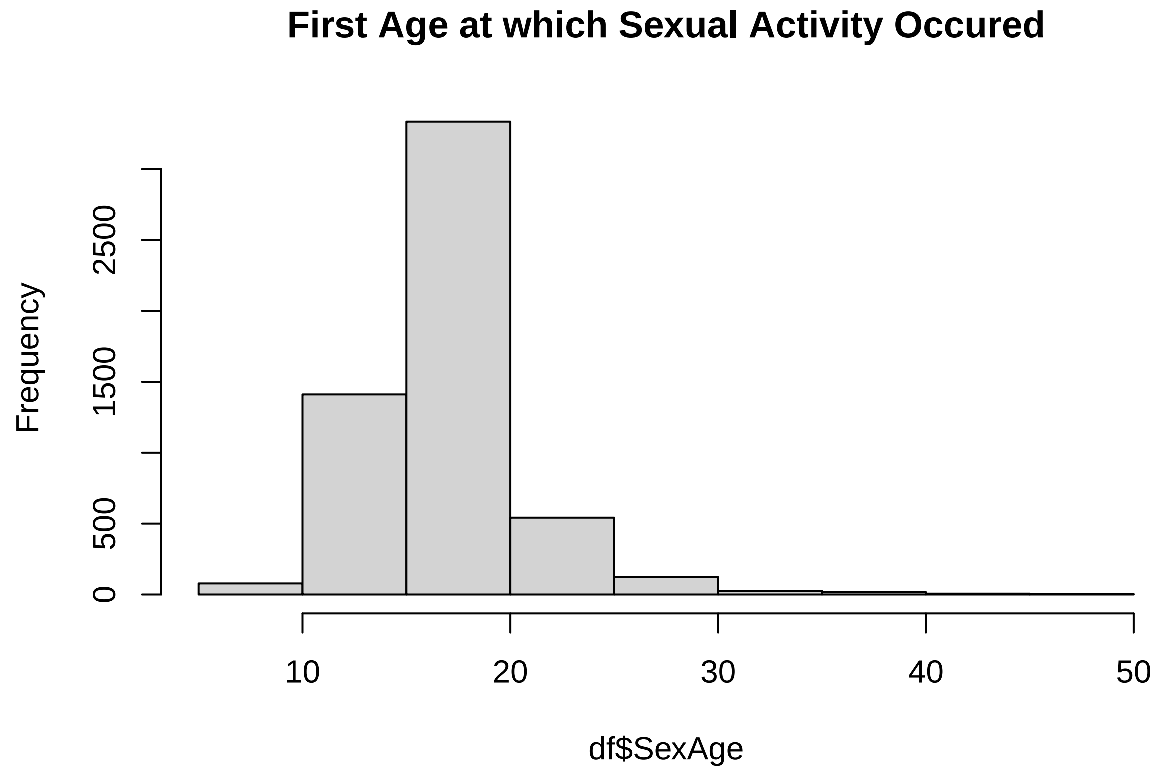
```r
model <- lm(SexNumPartnLife ~ RegularMarij+HardDrugs+RegularMarij*HardDrugs, df)
summary(model)
```

```
##
## Call:
## lm(formula = SexNumPartnLife ~ RegularMarij + HardDrugs + RegularMarij *
##     HardDrugs, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -37.59   -8.41   -5.41   -0.41 1991.59
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   8.4060     1.0513   7.996 1.59e-15 ***
## RegularMarijYes              14.8056     2.5393   5.831 5.88e-09 ***
## HardDrugsYes                 13.5674     3.6078   3.761 0.000171 ***
## RegularMarijYes:HardDrugsYes   0.8151     4.8573   0.168 0.866740
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.88 on 4897 degrees of freedom
##   (5099 observations deleted due to missingness)
## Multiple R-squared:  0.03038,    Adjusted R-squared:  0.02978
## F-statistic: 51.14 on 3 and 4897 DF,  p-value: < 2.2e-16
```

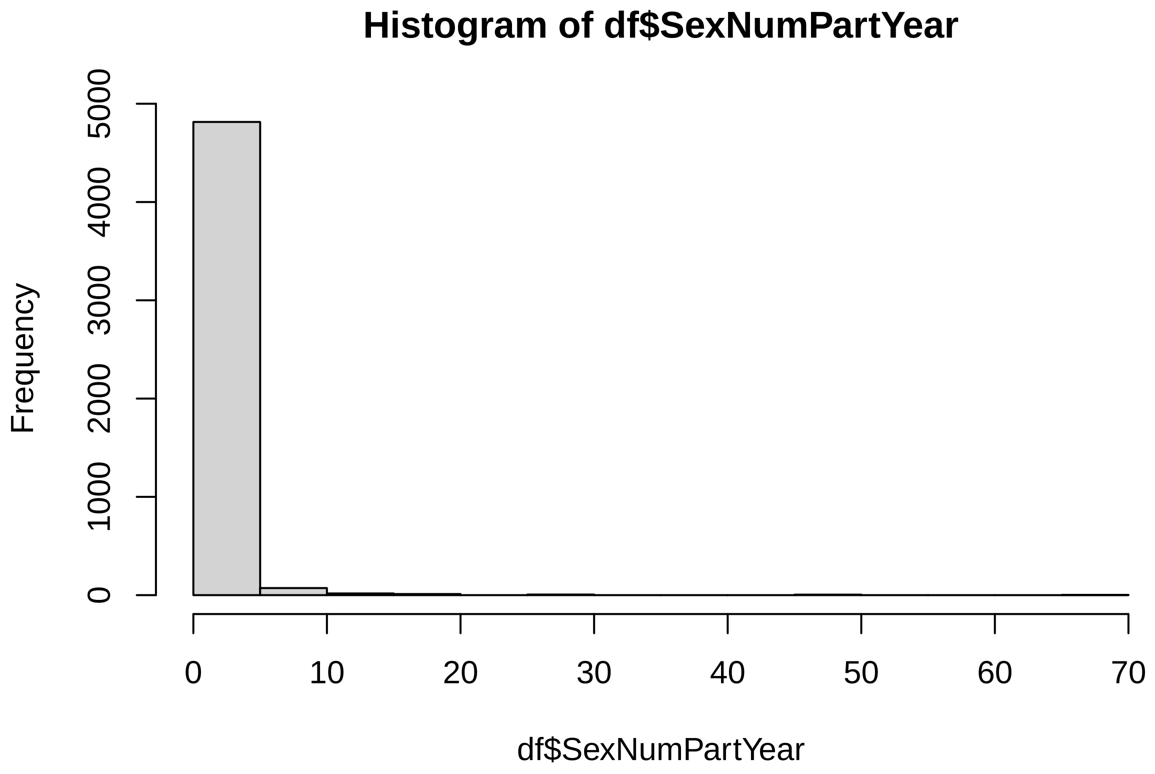
SexAge is has a good distribution but SexNumPartnLife has extreme skenwness and is discrete count data. This requires a Poisson regression which is out side the scopre of this course. Created new variable using the duration, since first sexual activity where (Age - SexAge) since Age >= SexAge, and dividing by the number of sexual partners in life to see frequency of sexual activity. New variable was log transformed due to extreme skewness that violated normality assumption, which could be checked by QQPlot.

Due to extreme skewness, we tried to find some observations that had implausible reported data that could been a typo or non serious answer. For instance, observations 8576 and 3416 reported to have had a first sexual activity at 9 with 360 and 500 sexual partners in life, respectively. Observations 4579 and 4580 reported to have had a first sexual activity at 10 and both reportedly had 700 sexual partners in life. Observations 4579 and 4580 reported to have had a first sexual activity at 10 and both reportedly had 700 sexual partners in life. We removed these outliers.

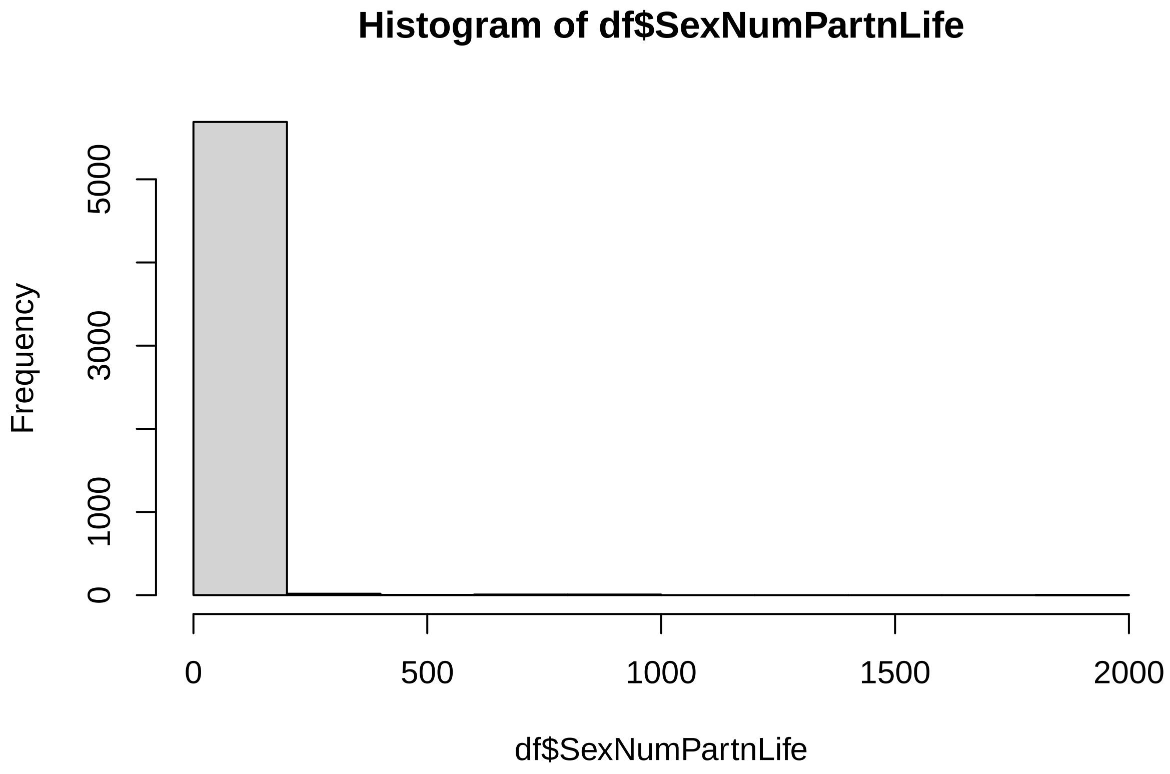
```r
hist(df$SexAge, main= "First Age at which Sexual Activity Occured")
```



```
hist(df$SexNumPartYear, main = )
```



```
hist(df$SexNumPartnLife)
```



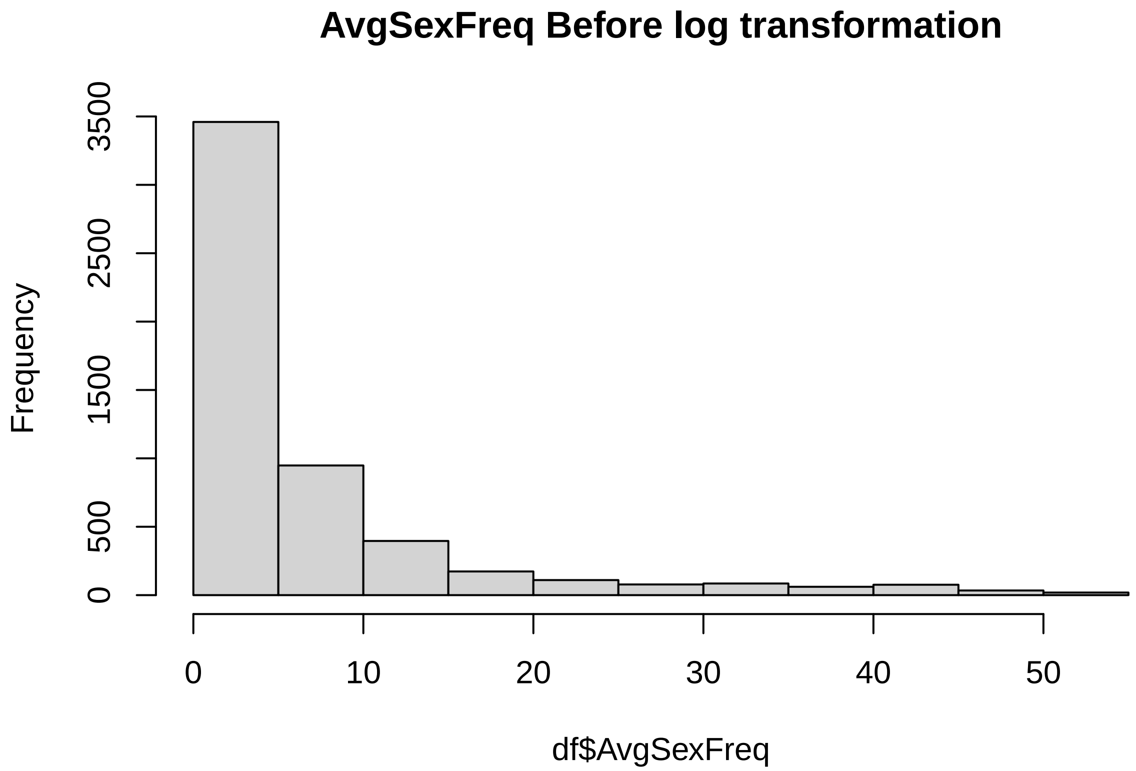
```
#Show observations with more than 300 sexual partners during lifetime
which(df$SexNumPartnLife > 300)
```

```
##  [1] 1353 2764 3416 3724 3795 4579 4580 6964 6965 7953 7954 8122 8123 8124 8428
## [16] 8576 8651 8838 8839 9596 9597 9598 9599 9600 9730
```

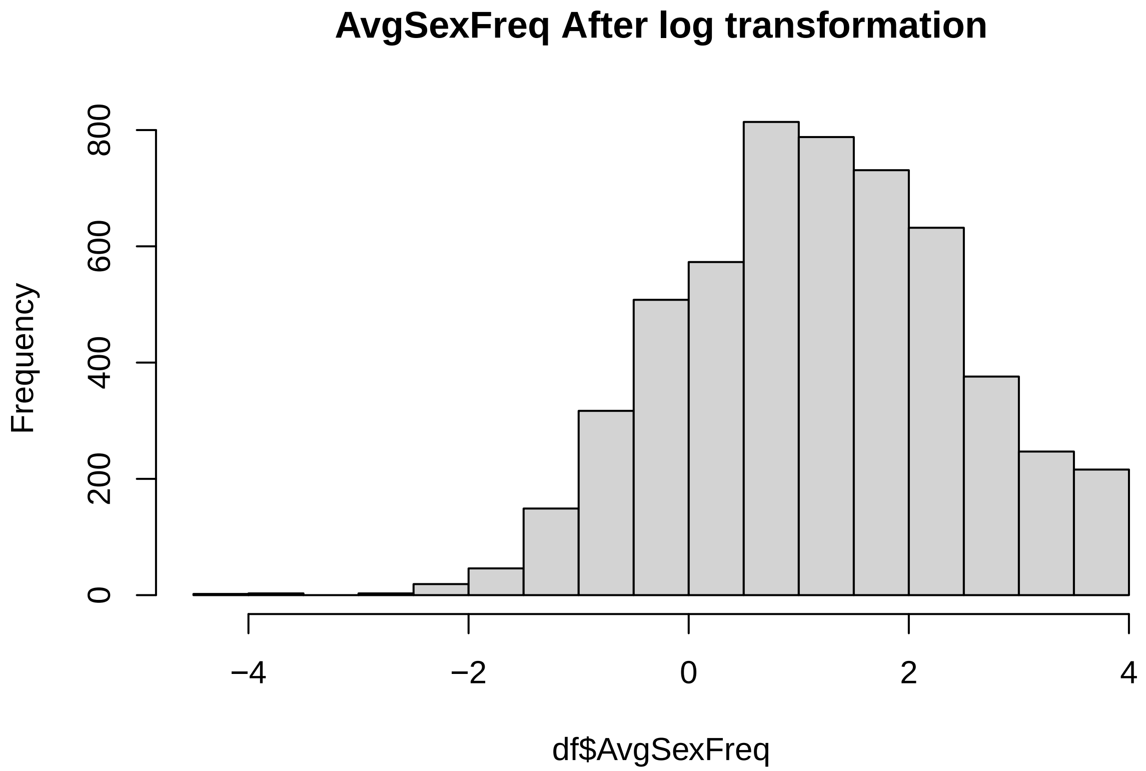
```
df[which(df$SexNumPartnLife > 300), c("Age", "SexAge", "SexNumPartnLife")]
```

```
## # A tibble: 25 x 3
##      Age SexAge SexNumPartnLife
##    <int>  <int>           <int>
##  1    63     18             301
##  2    54     13            1000
##  3    63      9             500
##  4    57     13            1000
##  5    42     14             560
##  6    49     10             700
##  7    49     10             700
##  8    23     11             340
##  9    23     11             340
## 10    50     15            1000
## # i 15 more rows
```

```
df = df[-which(df$SexNumPartnLife > 300),]

#Before log transformation
df = mutate(df, AvgSexFreq = (Age-SexAge)/SexNumPartnLife)
hist(df$AvgSexFreq, main = "AvgSexFreq Before log transformation")
```



```r
#After log transformation
df = mutate(df, AvgSexFreq = log((Age-SexAge)/SexNumPartnLife))
hist(df$AvgSexFreq, main = "AvgSexFreq After log transformation")
```



```r
#Remove negative infinity
df$AvgSexFreq[is.infinite(df$AvgSexFreq)] = NA
#unique(df$AvgSexFreq)

df$nPregnancies = is.factor(df$nPregnancies)
model <- lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+
summary(model)
```

```
##
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + HHIncome +
##     Education + BMI + DiabetesAge + Depressed + LittleInterest +
##     PhysActive + SameSex, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3555 -0.2319  0.1070  0.3372  1.8233
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.638326   1.431952  -1.144   0.2580
## SmokeNowYes                   0.317798   0.315754   1.006   0.3190
## AlcoholYear                  -0.002355   0.001688  -1.395   0.1691
## RegularMarijYes               0.643604   0.320484   2.008   0.0500 .
## HardDrugsYes                 -1.231593   0.614234  -2.005   0.0504 .
## Age                           0.051987   0.024085   2.158   0.0357 *
## Gendermale                   -1.340728   0.274434  -4.885  1.1e-05 ***
## HHIncome 5000-9999           -0.566871   0.629365  -0.901   0.3721
## HHIncome10000-14999          -1.081820   0.543756  -1.990   0.0521 .
## HHIncome15000-19999           0.903343   0.878828   1.028   0.3089
## HHIncome20000-24999          -0.356869   0.595470  -0.599   0.5517
## HHIncome25000-34999          -0.293062   0.565401  -0.518   0.6065
## HHIncome35000-44999           0.156911   0.525551   0.299   0.7665
## HHIncome45000-54999          -1.873535   0.756699  -2.476   0.0167 *
## HHIncome55000-64999           0.636927   0.613700   1.038   0.3043
## HHIncome65000-74999          -0.698542   0.612030  -1.141   0.2592
## HHIncome75000-99999          -0.407544   0.628229  -0.649   0.5195
## HHIncomemore 99999           -0.903659   0.530698  -1.703   0.0948 .
## Education9 - 11th Grade      -0.508748   0.491227  -1.036   0.3053
## EducationHigh School          0.333135   0.550048   0.606   0.5475
## EducationSome College         0.238200   0.489435   0.487   0.6286
## EducationCollege Grad         1.017370   0.611602   1.663   0.1025
## BMI                           0.025369   0.017988   1.410   0.1646
## DiabetesAge                   0.002411   0.014928   0.162   0.8723
## DepressedSeveral             -0.177637   0.363140  -0.489   0.6269
## DepressedMost                 0.236648   0.436207   0.543   0.5899
## LittleInterestSeveral        -0.066404   0.337355  -0.197   0.8448
## LittleInterestMost            0.510451   0.377313   1.353   0.1822
## PhysActiveYes                -0.059868   0.332020  -0.180   0.8576
## SameSexYes                    0.046164   0.490791   0.094   0.9254
## RegularMarijYes:HardDrugsYes  0.675466   0.704185   0.959   0.3421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7339 on 50 degrees of freedom
##   (9894 observations deleted due to missingness)
## Multiple R-squared:  0.7363, Adjusted R-squared:  0.578
## F-statistic: 4.653 on 30 and 50 DF,  p-value: 8.649e-07
```

```
model |>
  tbl_regression(intercept = TRUE)
```

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	-1.6	-4.5, 1.2	0.3
SmokeNow			
No	—	—	
Yes	0.32	-0.32, 0.95	0.3
AlcoholYear	0.00	-0.01, 0.00	0.2
RegularMarij			
No	—	—	
Yes	0.64	0.00, 1.3	0.050
HardDrugs			
No	—	—	
Yes	-1.2	-2.5, 0.00	0.050
Age	0.05	0.00, 0.10	0.036
Gender			

female	—	—	
male	-1.3	-1.9, -0.79	<0.001
HHIncome			
0-4999	—	—	
5000-9999	-0.57	-1.8, 0.70	0.4
10000-14999	-1.1	-2.2, 0.01	0.052
15000-19999	0.90	-0.86, 2.7	0.3
20000-24999	-0.36	-1.6, 0.84	0.6
25000-34999	-0.29	-1.4, 0.84	0.6
35000-44999	0.16	-0.90, 1.2	0.8
45000-54999	-1.9	-3.4, -0.35	0.017
55000-64999	0.64	-0.60, 1.9	0.3
65000-74999	-0.70	-1.9, 0.53	0.3
75000-99999	-0.41	-1.7, 0.85	0.5
more 99999	-0.90	-2.0, 0.16	0.095
Education			
8th Grade	—	—	
9 - 11th Grade	-0.51	-1.5, 0.48	0.3
High School	0.33	-0.77, 1.4	0.5
Some College	0.24	-0.74, 1.2	0.6
College Grad	1.0	-0.21, 2.2	0.10
BMI	0.03	-0.01, 0.06	0.2
DiabetesAge	0.00	-0.03, 0.03	0.9
Depressed			
None	—	—	
Several	-0.18	-0.91, 0.55	0.6
Most	0.24	-0.64, 1.1	0.6
LittleInterest			
None	—	—	
Several	-0.07	-0.74, 0.61	0.8
Most	0.51	-0.25, 1.3	0.2
PhysActive			
No	—	—	
Yes	-0.06	-0.73, 0.61	0.9
SameSex			
No	—	—	
Yes	0.05	-0.94, 1.0	>0.9
RegularMarij * HardDrugs			
Yes * Yes	0.68	-0.74, 2.1	0.3

[1] CI = Confidence Interval

```
#model <- lm(AvgSexFreq ~ #Gender+HHIncome+Education+PhysActive+SameSex+AlcoholYear+RegularMarij+HardDr
#summary(model)
```

Using the sequential sum of squares we tested for each block of covariates at a significance level 0.05

```
n = 50
aov = anova(model <- lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs
aov
```

```
## Analysis of Variance Table
##
## Response: AvgSexFreq
##                        Df  Sum Sq Mean Sq F value    Pr(>F)
## SmokeNow                1  0.7399  0.7399  1.3736 0.2467482
## AlcoholYear             1  6.3185  6.3185 11.7302 0.0012368 **
## RegularMarij            1  0.2515  0.2515  0.4670 0.4975312
## HardDrugs               1  6.0788  6.0788 11.2852 0.0015019 **
## Age                     1 14.9093 14.9093 27.6786 3.000e-06 ***
## Gender                  1 16.2649 16.2649 30.1952 1.318e-06 ***
## HHIncome               11 21.9288  1.9935  3.7009 0.0006885 ***
## Education               4  2.5118  0.6279  1.1658 0.3371471
## BMI                     1  1.5849  1.5849  2.9423 0.0924794 .
## DiabetesAge             1  0.0722  0.0722  0.1340 0.7158242
## Depressed               2  2.3338  1.1669  2.1663 0.1252382
## LittleInterest          2  1.6380  0.8190  1.5205 0.2285478
## PhysActive              1  0.0568  0.0568  0.1054 0.7467409
## SameSex                 1  0.0017  0.0017  0.0032 0.9553125
## RegularMarij:HardDrugs  1  0.4956  0.4956  0.9201 0.3420654
## Residuals              50 26.9329  0.5387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSY = sum(aov$"Sum Sq")
SSQ = aov$"Sum Sq"
MSE = aov$"Mean Sq"[16]
ss1 = sum(SSQ[c(1:4, 15)])
print(ss1)
```

```
## [1] 13.88444
```

```
fstat1 = ss1/5/MSE
pval1 = 1-pf(q = fstat1, df1 = 5, df2 = n-16)
print(c(fstat1, pval1))
```

```
## [1] 5.155204576 0.001262146
```

```
ss2 = sum(SSQ[5:8])
print(ss2)
```

```
## [1] 55.61473
```

```
fstat2 = ss2/4/MSE
pval2 = 1-pf(q = fstat2, df1 = 4, df2 = n-16)
print(c(fstat2, pval2))
```

```
## [1] 2.581174e+01 6.872507e-10
```

```
ss3 = sum(SSQ[9:14])
print(ss3)
```

```
## [1] 5.687399
```

```
fstat3 = ss3/5/MSE
pval3 = 1-pf(q = fstat3, df1 = 5, df2 = n-16)
print(c(fstat3, pval3))
```

```
## [1] 2.11169493 0.08788892
```

```
ss4 = sum(SSQ[14])
print(ss4)
```

```
## [1] 0.001708498
```

```
fstat4 = ss3/1/MSE
pval4 = 1-pf(q = fstat4, df1 = 1, df2 = n-16)
print(c(fstat4, pval4))
```

```
## [1] 10.55847467  0.00260712
```

(i) $\boldsymbol{\beta}_{substance} = (\beta_{SmokeNow}, \beta_{AlcoholYear}, \beta_{RegularMarij}, \beta_{HardDrugs}, \beta_{RegularMarij*HardDrugs})^T$
(ii) $\boldsymbol{\beta}_{Demo} = (\beta_{Age}, \beta_{Gender}, \beta_{HHIncome}, \beta_{Education})^T$
(iii) $\boldsymbol{\beta}_{Health} = (\beta_{BMI}, \beta_{DiabetesAges}, \beta_{Depressed}, \beta_{LittleInterest}, \beta_{PhysActive})^T$
(iv) $\boldsymbol{\beta}_{SameSex} = (\beta_{SameSex})^T$

Step	Tested Var.	SS(Num.)	SS(Denom.)	Test Stat.	Dist.	p-value	Decision	Stopping Rule	Decision
I	$\boldsymbol{\beta}_{Substance}$	13.88444	26.9329	5.155204576	$F_{5,34}$	0.001262146	Reject	Do not stop	Collect
II	$\boldsymbol{\beta}_{Demo}$	55.61473	26.9329	25.81174	$F_{4,34}$	6.872507e-10	Reject	Do not stop	Collect
III	$\boldsymbol{\beta}_{Health}$	5.687399	26.9329	2.11169493	$F_{5,34}$	0.08788892	Fail to Reject	Not Collect	Collect
IV	$\boldsymbol{\beta}_{SameSex}$	0.001708498	26.9329	10.55847467	$F_{1,34}$	0.00260712	NA	NA	NA

```
library(ggplot2)
library(tidyr)
#Add new column based on missingness
covariates = c("AvgSexFreq", "SmokeNow","AlcoholYear", "RegularMarij", "HardDrugs", "Age", "Gender","HH
sum(complete.cases(df[, covariates]))
```

```
## [1] 1782
```

```
df$missingness <- ifelse(complete.cases(df[, covariates]), "Missing", "Not Missing")
```

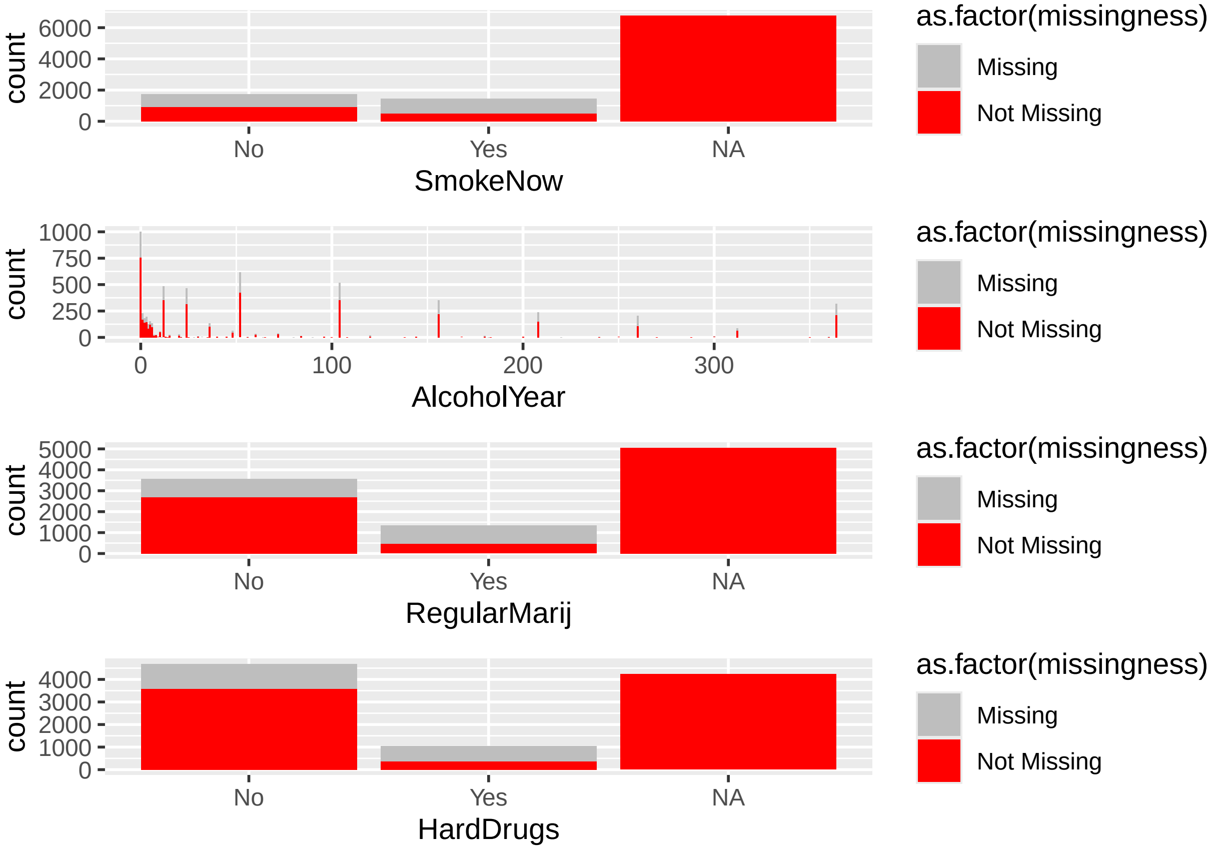
```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.4.2

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
p1 = ggplot(data = df, mapping=aes(x=SmokeNow, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p2 = ggplot(data = df, mapping=aes(x=AlcoholYear, fill=as.factor(missingness)))+
```

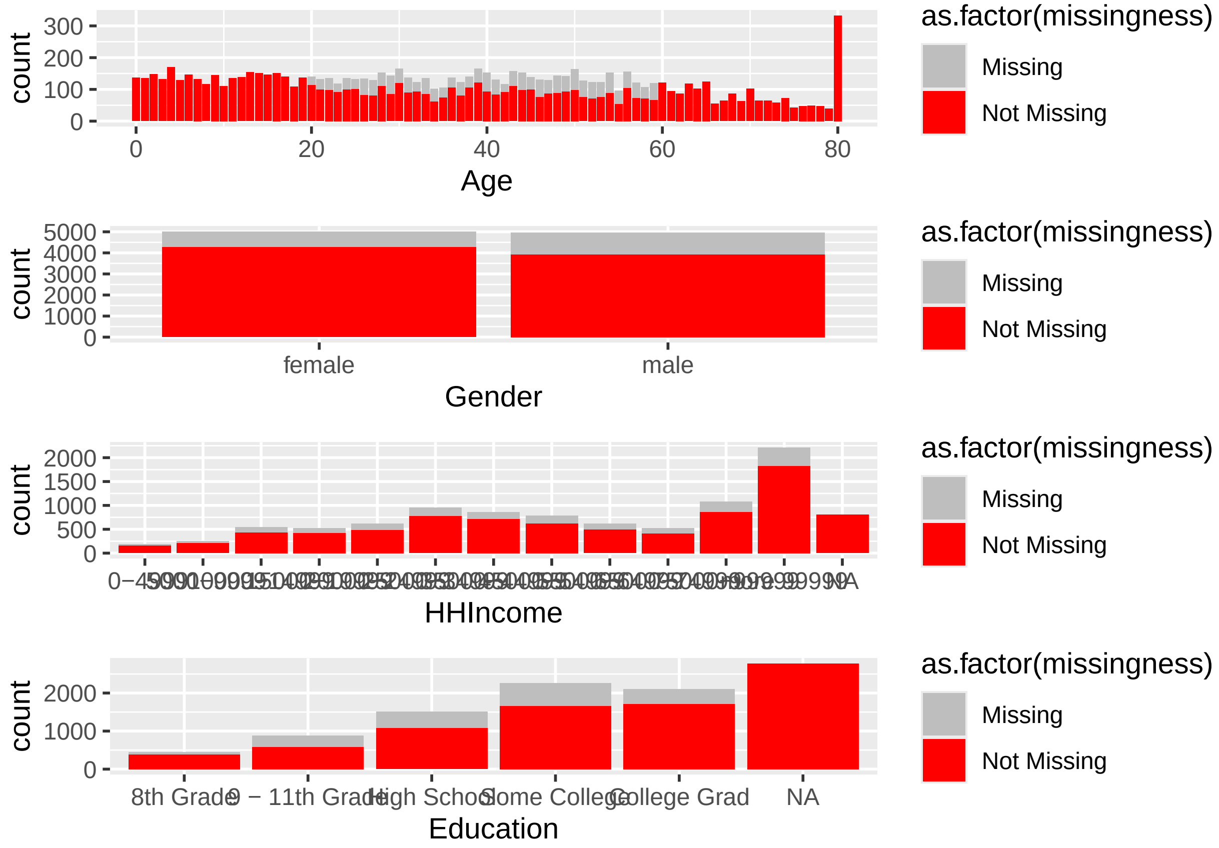
```r
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p3 = ggplot(data = df, mapping=aes(x=RegularMarij, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p4 = ggplot(data = df, mapping=aes(x=HardDrugs, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))

grid.arrange(p1,p2,p3,p4, nrow=4)
```

```
## Warning: Removed 4078 rows containing non-finite outside the scale range
## (`stat_count()`).
```

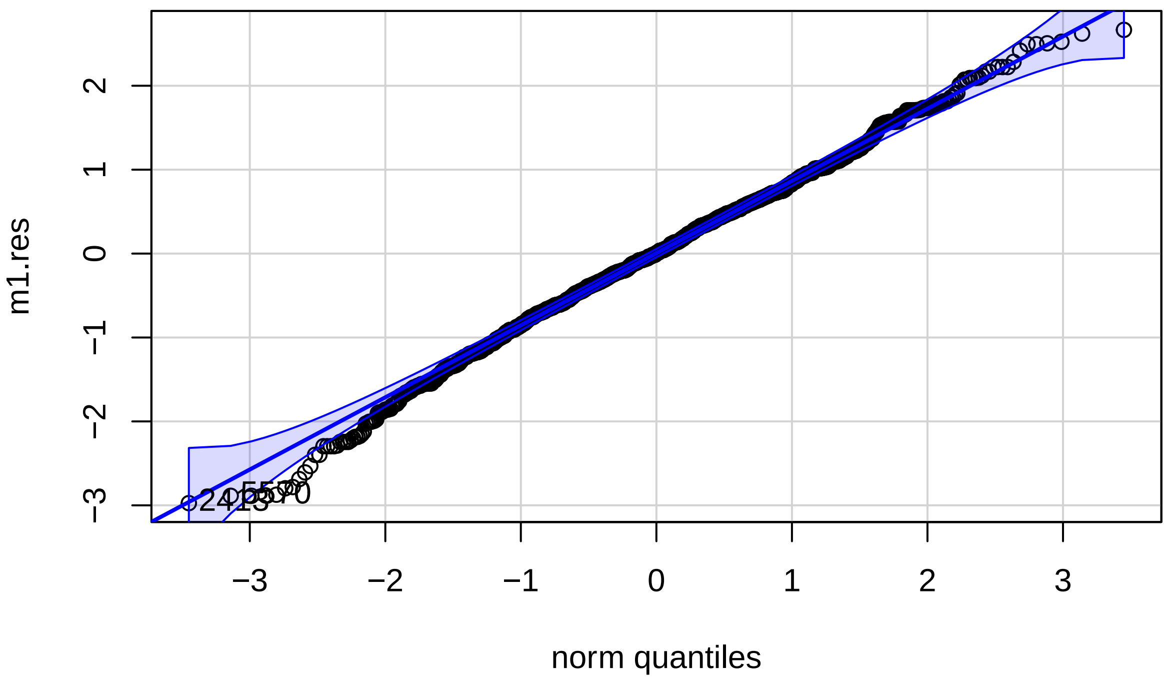


```r
p6 = ggplot(data = df, mapping=aes(x=Age, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p7 = ggplot(data = df, mapping=aes(x=Gender, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p8 = ggplot(data = df, mapping=aes(x=HHIncome, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p9 = ggplot(data = df, mapping=aes(x=Education, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
```

```r
grid.arrange(p6, p7, p8, p9, nrow = 4)
```

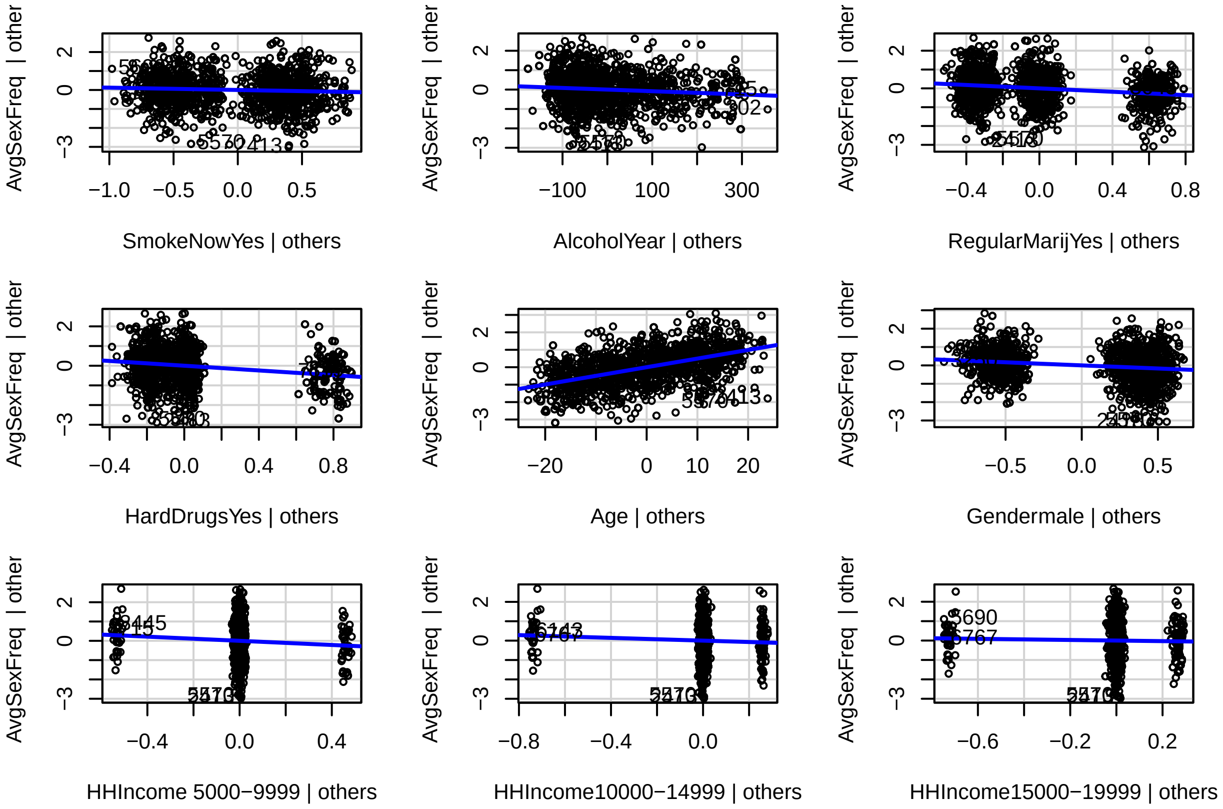


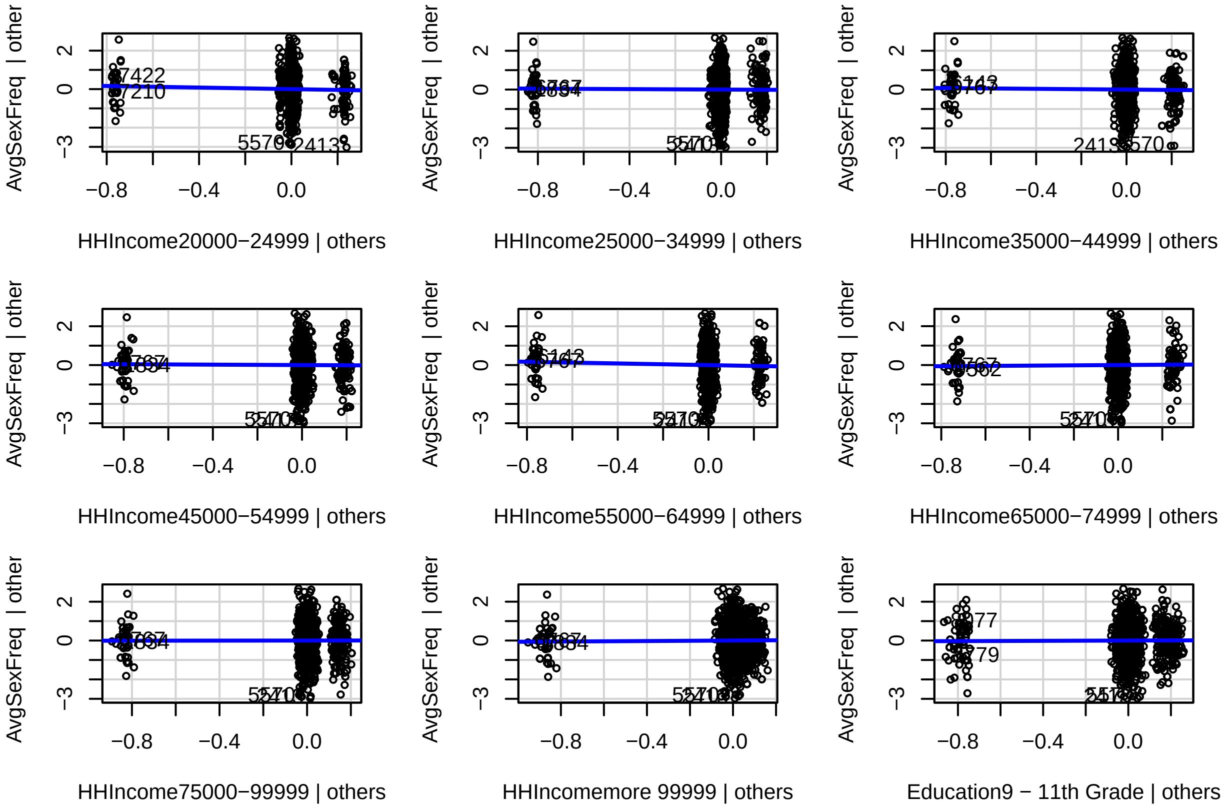
```r
m1 = lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+HHIn
m1.res = m1$residuals

car::qqPlot(m1.res)
```

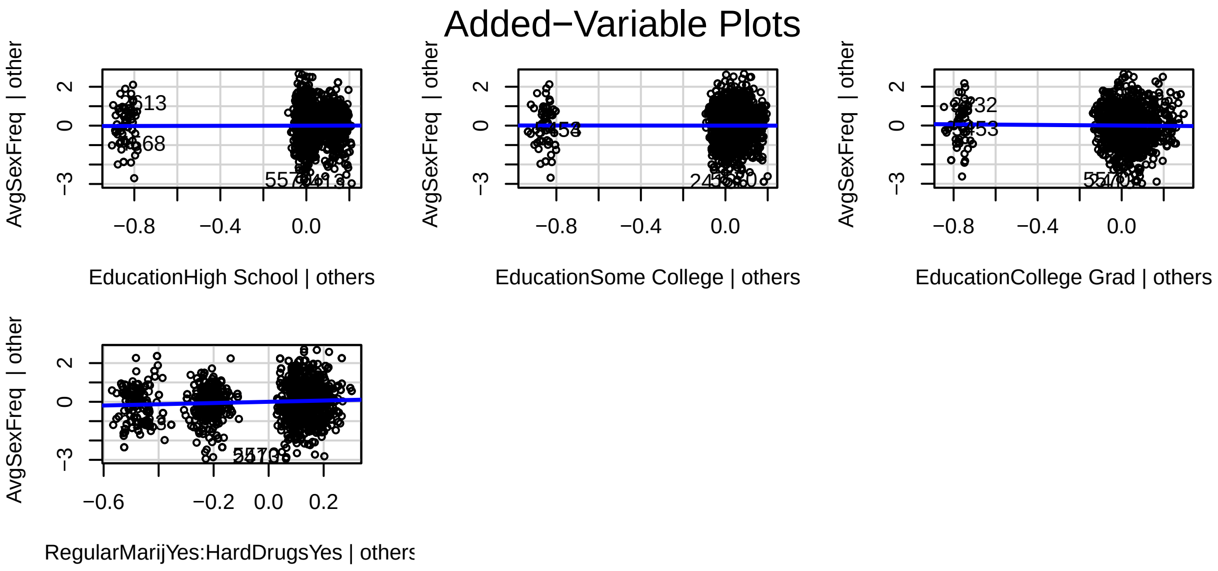


```
## 2413 5570
##  458 1030
```

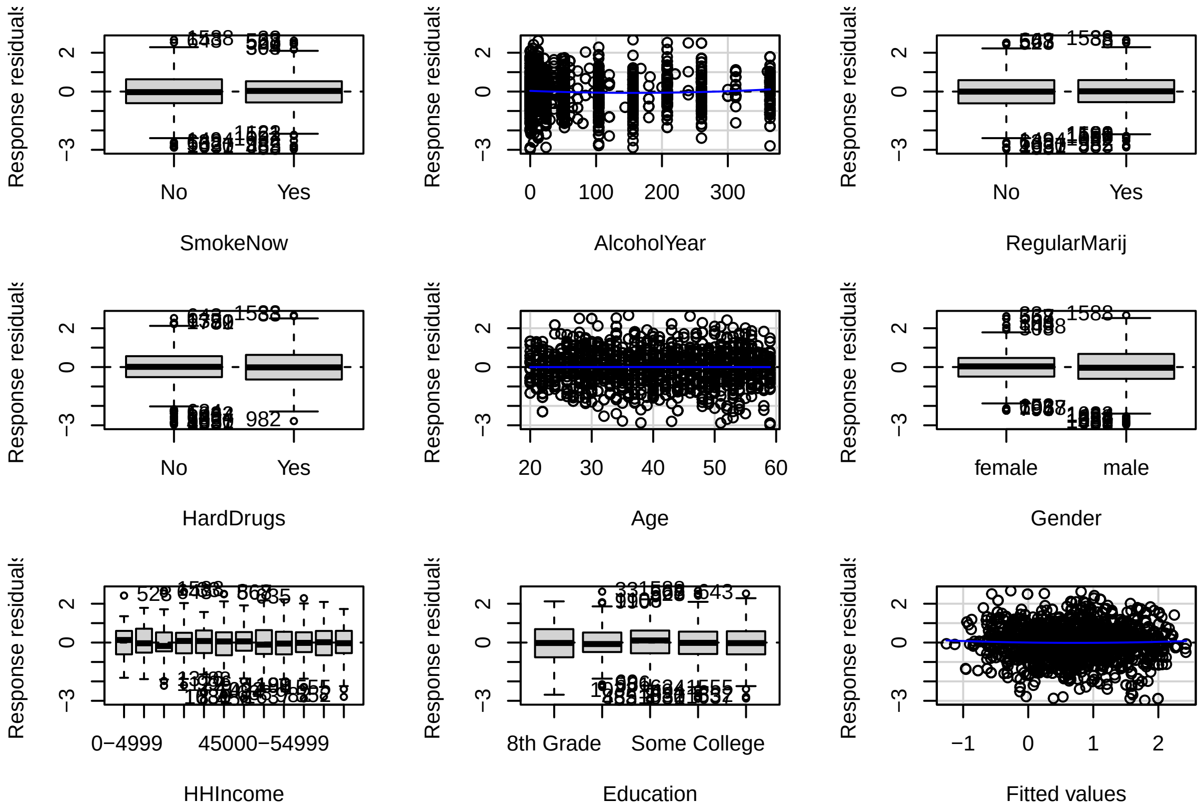
```
car::avPlots(m1)
```


AvgSexFreq | other
2
0
−3
−1.0
−0.5
0.0
0.5
SmokeNowYes | others
AvgSexFreq | other
−100
100
300
AlcoholYear | others
502
AvgSexFreq | other
−0.4
0.0
0.4
0.8
RegularMarijYes | others
AvgSexFreq | other
−0.4
0.0
0.4
0.8
HardDrugsYes | others
AvgSexFreq | other
−20
0
10
20
Age | others
AvgSexFreq | other
−0.5
0.0
0.5
Gendermale | others
AvgSexFreq | other
−0.4
0.0
0.4
HHIncome 5000−9999 | others
3445
AvgSexFreq | other
−0.8
−0.4
0.0
HHIncome10000−14999 | others
AvgSexFreq | other
−0.6
−0.2
0.2
HHIncome15000−19999 | others
3690
5767


AvgSexFreq | other
2
0
−3
−0.8
−0.4
0.0
7422
7210
5570
2413
HHIncome20000−24999 | others
AvgSexFreq | other
5570
HHIncome25000−34999 | others
AvgSexFreq | other
2413
5570
HHIncome35000−44999 | others
AvgSexFreq | other
5570
HHIncome45000−54999 | others
AvgSexFreq | other
5570
HHIncome55000−64999 | others
AvgSexFreq | other
5570
HHIncome65000−74999 | others
AvgSexFreq | other
5570
HHIncome75000−99999 | others
AvgSexFreq | other
5570
HHIncomemore 99999 | others
AvgSexFreq | other
77
79
Education9 − 11th Grade | others



```
car::residualPlots(m1, type="response")
```



```
##              Test stat Pr(>|Test stat|)
## SmokeNow
## AlcoholYear     2.3041          0.02134 *
## RegularMarij
## HardDrugs
## Age            -0.0525          0.95818
## Gender
## HHIncome
## Education
## Tukey test      0.9173          0.35898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#interactions(???)
nonintmodel <- lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+Age+Gender+HHIncome+Education, df)
car::vif(nonintmodel,type = 'predictor')
```

```
## GVIFs computed for predictors
```

```
##                  GVIF Df GVIF^(1/(2*Df)) Interacts With
## SmokeNow     1.176162  1        1.084510           --
## AlcoholYear  1.121568  1        1.059041           --
## RegularMarij 1.036768  1        1.018218           --
## Age          1.093531  1        1.045720           --
## Gender       1.046471  1        1.022972           --
## HHIncome     1.437208 11        1.016623           --
## Education    1.418796  4        1.044696           --
##                                                               Other Predictors
## SmokeNow          AlcoholYear, RegularMarij, Age, Gender, HHIncome, Education
## AlcoholYear          SmokeNow, RegularMarij, Age, Gender, HHIncome, Education
## RegularMarij          SmokeNow, AlcoholYear, Age, Gender, HHIncome, Education
## Age          SmokeNow, AlcoholYear, RegularMarij, Gender, HHIncome, Education
## Gender          SmokeNow, AlcoholYear, RegularMarij, Age, HHIncome, Education
## HHIncome          SmokeNow, AlcoholYear, RegularMarij, Age, Gender, Education
## Education          SmokeNow, AlcoholYear, RegularMarij, Age, Gender, HHIncome
```

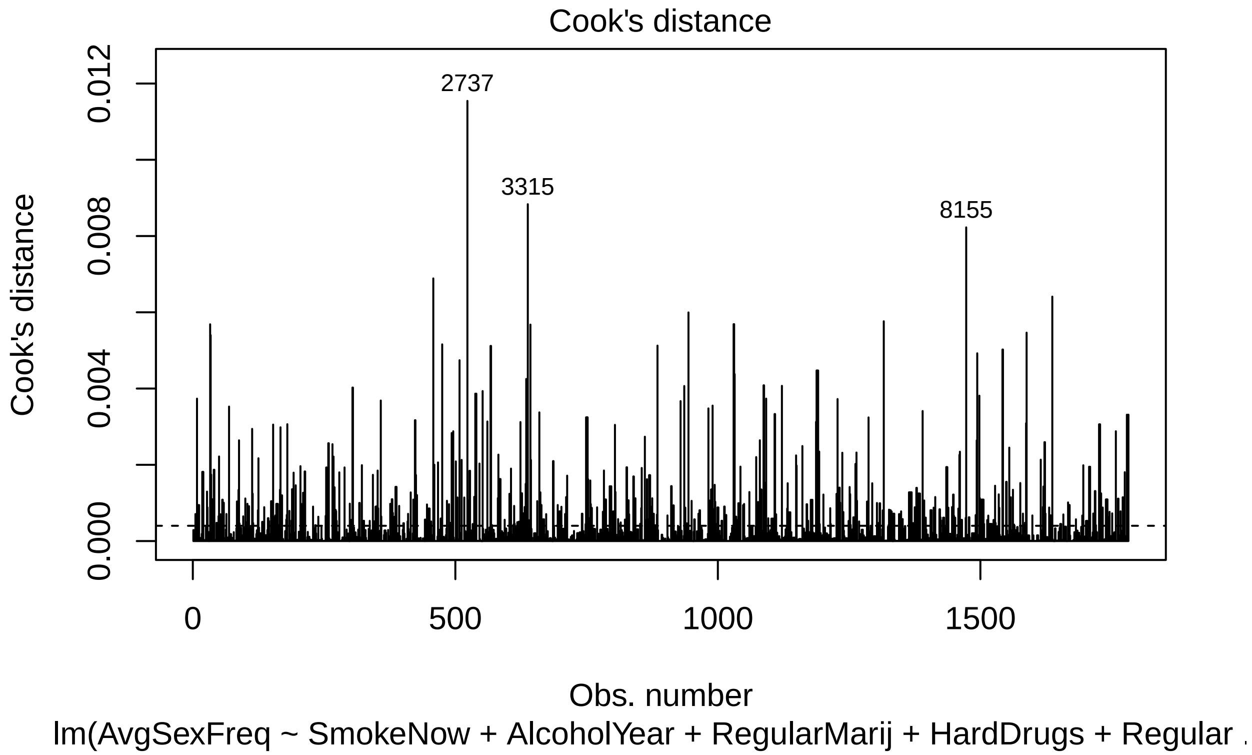
```
model.deffits=dffits(m1)
model.CD = cooks.distance(m1)
model.deffits[which.max(model.deffits)]
```

```
##      2737
## 0.5162887
```

```
model.CD[which.max(model.CD)]
```

```
##       2737
## 0.01154526
```

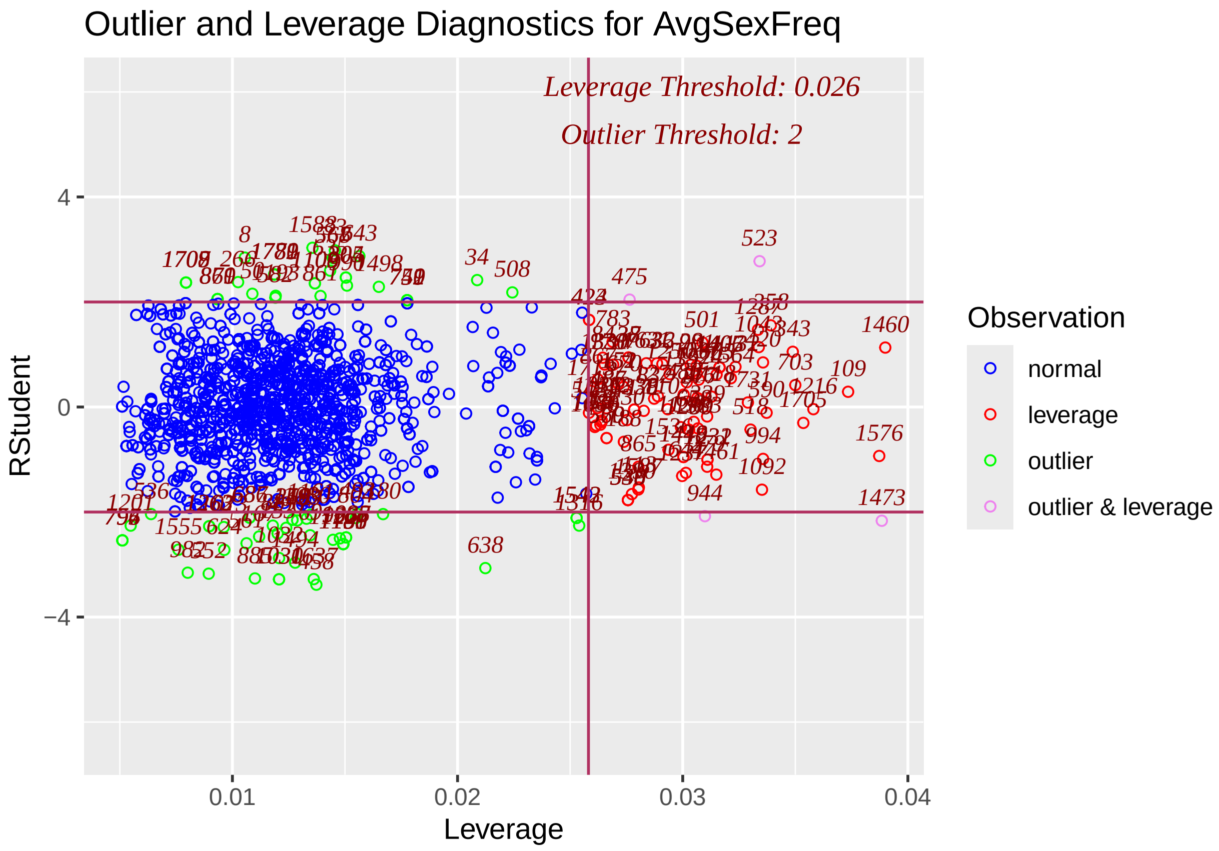
```
n = nrow(df)
p = m1$rank
plot(m1, which = 4)
abline(h=4/n,lty=2)
```



```
df[c(2737, 3315, 8155),]
```

```
## # A tibble: 3 x 78
##      ID SurveyYr Gender   Age AgeDecade AgeMonths Race1    Race3    Education
##   <int> <fct>    <fct>  <int> <fct>         <int> <fct>    <fct>    <fct>
## 1 57411 2009_10  male      52 " 50-59"        633 White    <NA>     Some College
## 2 58645 2009_10  male      52 " 50-59"        629 Mexican <NA>     8th Grade
## 3 68401 2011_12  male      43 " 40-49"         NA Mexican Mexican 8th Grade
## # i 69 more variables: MaritalStatus <fct>, HHIncome <fct>, HHIncomeMid <int>,
## #   Poverty <dbl>, HomeRooms <int>, HomeOwn <fct>, Work <fct>, Weight <dbl>,
## #   Length <dbl>, HeadCirc <dbl>, Height <dbl>, BMI <dbl>,
## #   BMICatUnder20yrs <fct>, BMI_WHO <fct>, Pulse <int>, BPSysAve <int>,
## #   BPDiaAve <int>, BPSys1 <int>, BPDia1 <int>, BPSys2 <int>, BPDia2 <int>,
## #   BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>, DirectChol <dbl>,
## #   TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>, UrineVol2 <int>, ...
```

```
ols_plot_resid_lev(m1)
```



```
df[c(475, 523, 944, 1473),]
```

```
## # A tibble: 4 x 78
##      ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education
##   <int> <fct>    <fct>  <int> <fct>         <int> <fct> <fct> <fct>
## 1 52577 2009_10  male      78 " 70+"          944 White <NA>  College Grad
## 2 52689 2009_10  female    44 " 40-49"        530 White <NA>  College Grad
## 3 53532 2009_10  male      51 " 50-59"        615 White <NA>  High School
## 4 54672 2009_10  female     1 " 0-9"           12 White <NA>  <NA>
## # i 69 more variables: MaritalStatus <fct>, HHIncome <fct>, HHIncomeMid <int>,
## #   Poverty <dbl>, HomeRooms <int>, HomeOwn <fct>, Work <fct>, Weight <dbl>,
## #   Length <dbl>, HeadCirc <dbl>, Height <dbl>, BMI <dbl>,
## #   BMICatUnder20yrs <fct>, BMI_WHO <fct>, Pulse <int>, BPSysAve <int>,
## #   BPDiaAve <int>, BPSys1 <int>, BPDia1 <int>, BPSys2 <int>, BPDia2 <int>,
## #   BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>, DirectChol <dbl>,
## #   TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>, UrineVol2 <int>, ...
```

```
df2 = df[-c(3315),]
m2 = lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+HHInc
summary(m1)
```

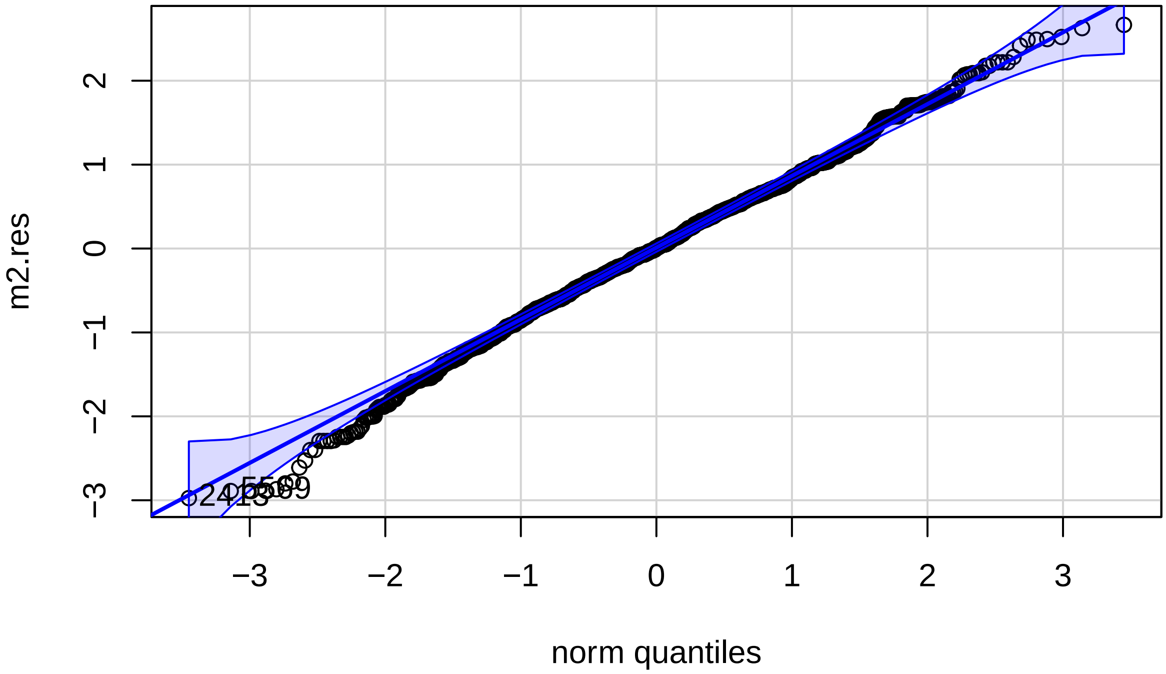
```
##
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + HHIncome +
##     Education, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97351 -0.57280  0.00155  0.58754  2.66593
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.4866093  0.2026882  -2.401  0.01646 *
## SmokeNowYes                  -0.1226982  0.0457879  -2.680  0.00744 **
## AlcoholYear                  -0.0008466  0.0002090  -4.051 5.33e-05 ***
## RegularMarijYes              -0.4499456  0.0565859  -7.952 3.26e-15 ***
## HardDrugsYes                 -0.6026914  0.0761115  -7.919 4.22e-15 ***
## Age                           0.0495315  0.0019548  25.338  < 2e-16 ***
## Gendermale                   -0.3373153  0.0437338  -7.713 2.04e-14 ***
## HHIncome 5000-9999           -0.5386945  0.1950816  -2.761  0.00582 **
## HHIncome10000-14999          -0.3560340  0.1656085  -2.150  0.03170 *
## HHIncome15000-19999          -0.1488586  0.1670012  -0.891  0.37286
## HHIncome20000-24999          -0.2047610  0.1626641  -1.259  0.20827
## HHIncome25000-34999          -0.0578691  0.1573318  -0.368  0.71305
## HHIncome35000-44999          -0.0974220  0.1613428  -0.604  0.54604
## HHIncome45000-54999          -0.0548363  0.1591968  -0.344  0.73054
## HHIncome55000-64999          -0.2137773  0.1627365  -1.314  0.18914
## HHIncome65000-74999           0.0757010  0.1663099   0.455  0.64904
## HHIncome75000-99999           0.0086152  0.1558027   0.055  0.95591
## HHIncomemore 99999            0.0654338  0.1522073   0.430  0.66732
## Education9 - 11th Grade       0.0351823  0.1203788   0.292  0.77012
## EducationHigh School          0.0205410  0.1166196   0.176  0.86021
## EducationSome College        -0.0062633  0.1156260  -0.054  0.95681
## EducationCollege Grad        -0.0796581  0.1221844  -0.652  0.51452
## RegularMarijYes:HardDrugsYes  0.3197590  0.0973576   3.284  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8875 on 1759 degrees of freedom
##   (8193 observations deleted due to missingness)
## Multiple R-squared:  0.3868, Adjusted R-squared:  0.3791
## F-statistic: 50.43 on 22 and 1759 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
```

```
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + HHIncome +
##     Education, data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97442 -0.56586  0.00393  0.58777  2.66542
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.4504301  0.2025488  -2.224 0.026288 *
## SmokeNowYes                  -0.1262317  0.0456933  -2.763 0.005794 **
## AlcoholYear                  -0.0008394  0.0002085  -4.026 5.92e-05 ***
## RegularMarijYes              -0.4526468  0.0564579  -8.017 1.95e-15 ***
## HardDrugsYes                 -0.6072452  0.0759447  -7.996 2.31e-15 ***
## Age                           0.0496820  0.0019508  25.467  < 2e-16 ***
## Gendermale                   -0.3353620  0.0436343  -7.686 2.51e-14 ***
## HHIncome 5000-9999           -0.5415626  0.1946190  -2.783 0.005449 **
## HHIncome10000-14999          -0.3577768  0.1652149  -2.166 0.030482 *
## HHIncome15000-19999          -0.1516692  0.1666058  -0.910 0.362764
## HHIncome20000-24999          -0.2093735  0.1622834  -1.290 0.197161
## HHIncome25000-34999          -0.0462766  0.1570024  -0.295 0.768219
## HHIncome35000-44999          -0.1024180  0.1609666  -0.636 0.524684
## HHIncome45000-54999          -0.0578562  0.1588205  -0.364 0.715688
## HHIncome55000-64999          -0.2168239  0.1623518  -1.336 0.181880
## HHIncome65000-74999           0.0736871  0.1659149   0.444 0.657006
## HHIncome75000-99999           0.0059989  0.1554339   0.039 0.969218
## HHIncomemore 99999            0.0627278  0.1518472   0.413 0.679585
## Education9 - 11th Grade      -0.0029744  0.1207343  -0.025 0.980348
## EducationHigh School         -0.0174155  0.1169977  -0.149 0.881686
## EducationSome College        -0.0448210  0.1160331  -0.386 0.699338
## EducationCollege Grad        -0.1192098  0.1225732  -0.973 0.330905
## RegularMarijYes:HardDrugsYes  0.3231571  0.0971319   3.327 0.000896 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8853 on 1758 degrees of freedom
##   (8193 observations deleted due to missingness)
## Multiple R-squared:  0.3891, Adjusted R-squared:  0.3815
## F-statistic: 50.91 on 22 and 1758 DF,  p-value: < 2.2e-16
```

```
m2.res = m2$residuals

car::qqPlot(m2.res)
```



```
## 2413 5569
##  458 1029
```