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
```

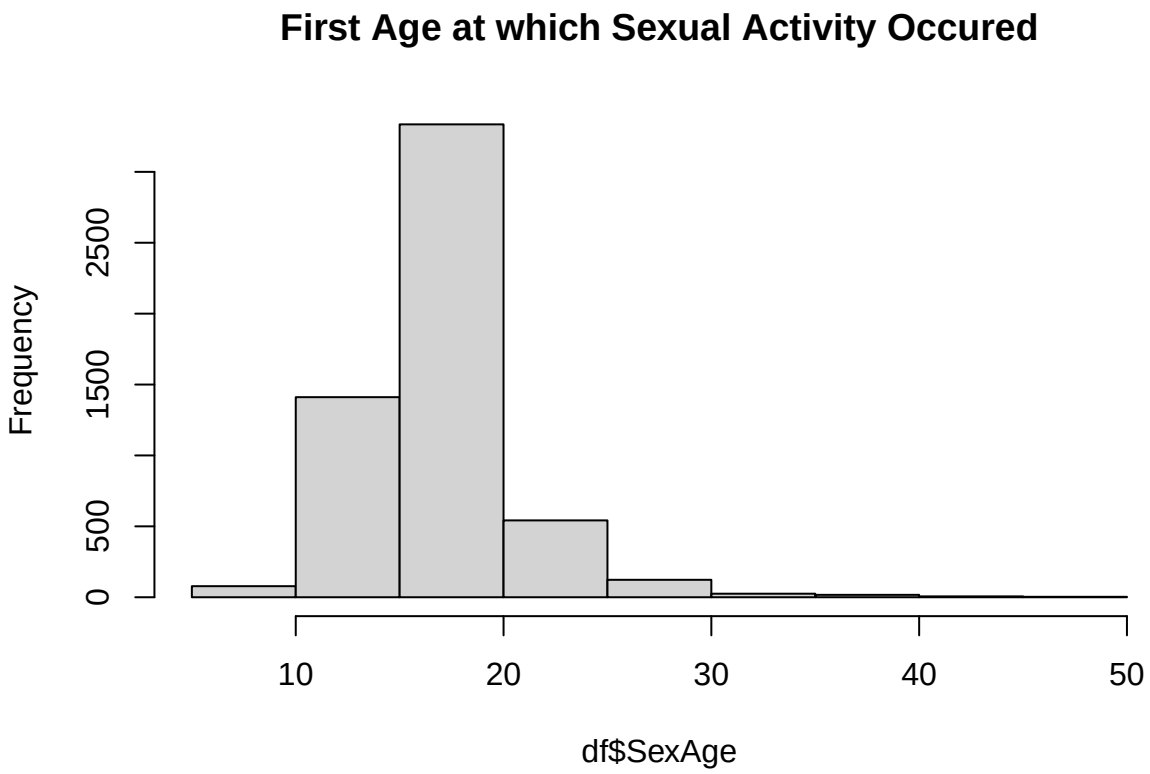
```r
model <- lm(SexNumPartnLife ~ RegularMarij+HardDrugs+RegularMarij*HardDrugs, df)
summary(model)
```

```
##
## 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
```

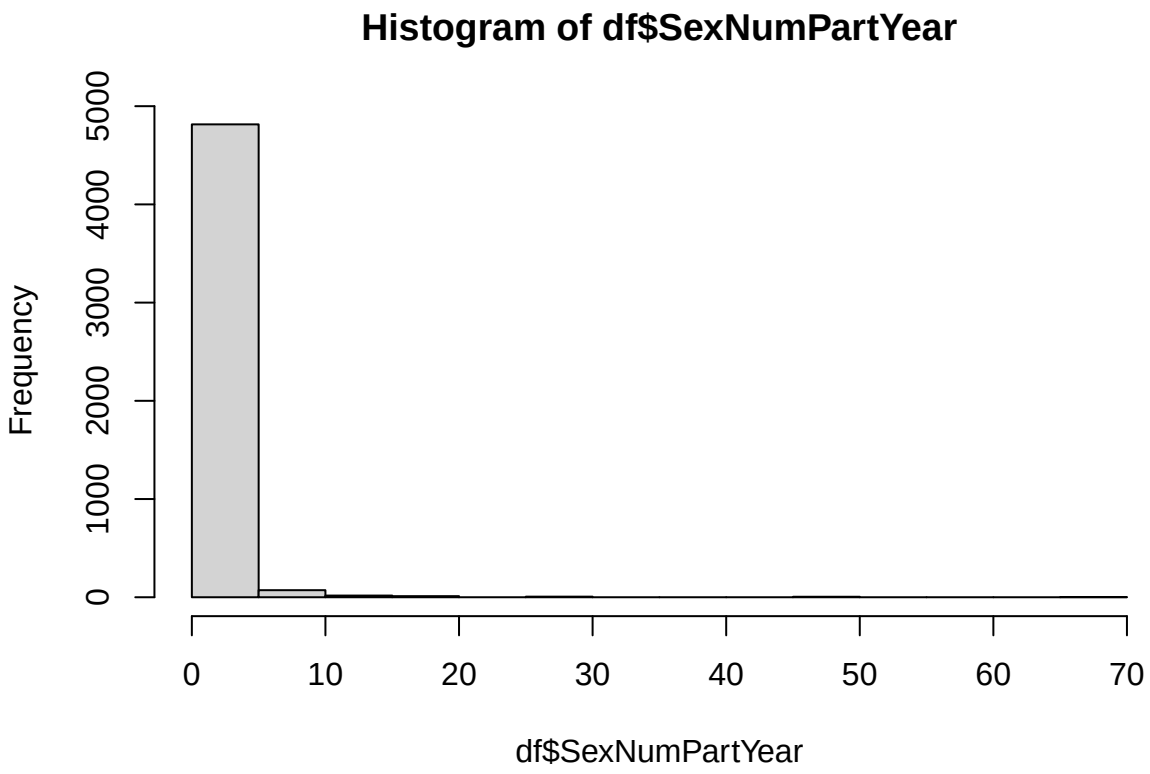
SexAge is has a good distribution but SexNumPartnLife has extreme skenwness and is discrete count data. This requires a Poisson regression which is out side the scopre of this course. Created new variable using the duration, since first sexual activity where (Age - SexAge) since Age >= SexAge, and dividing 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.

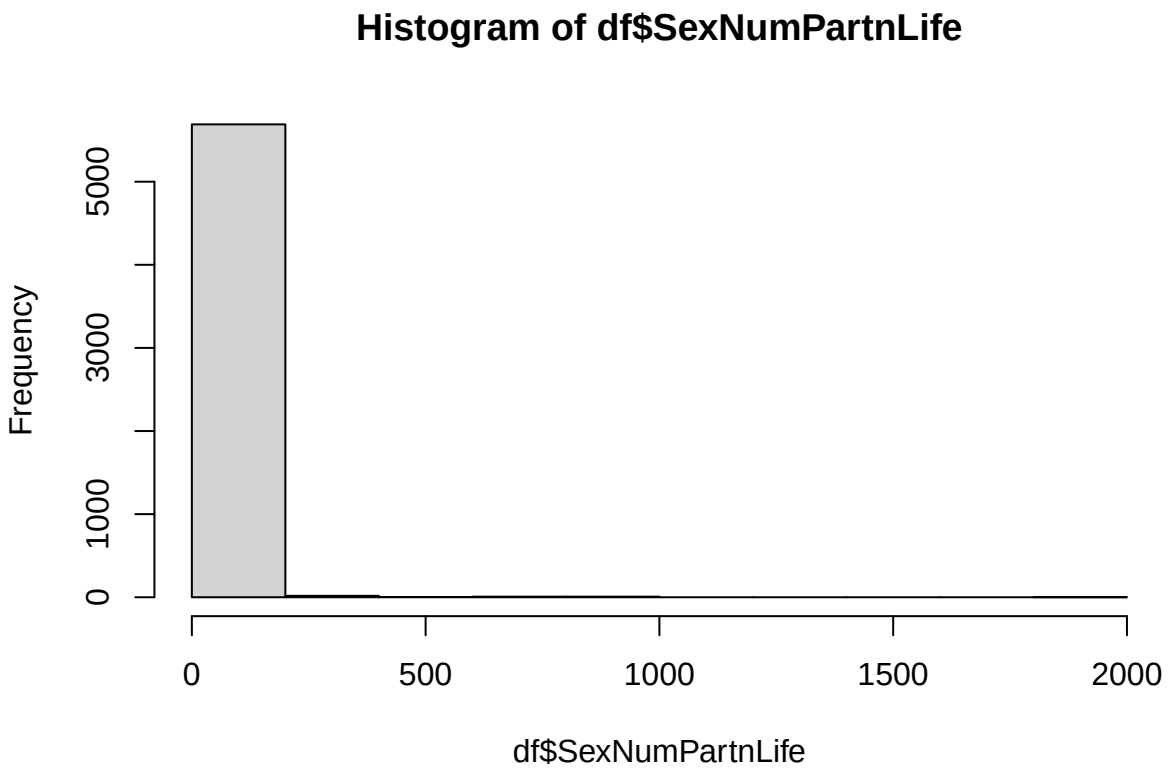
```r
hist(df$SexAge, main= "First Age at which Sexual Activity Occured")
```



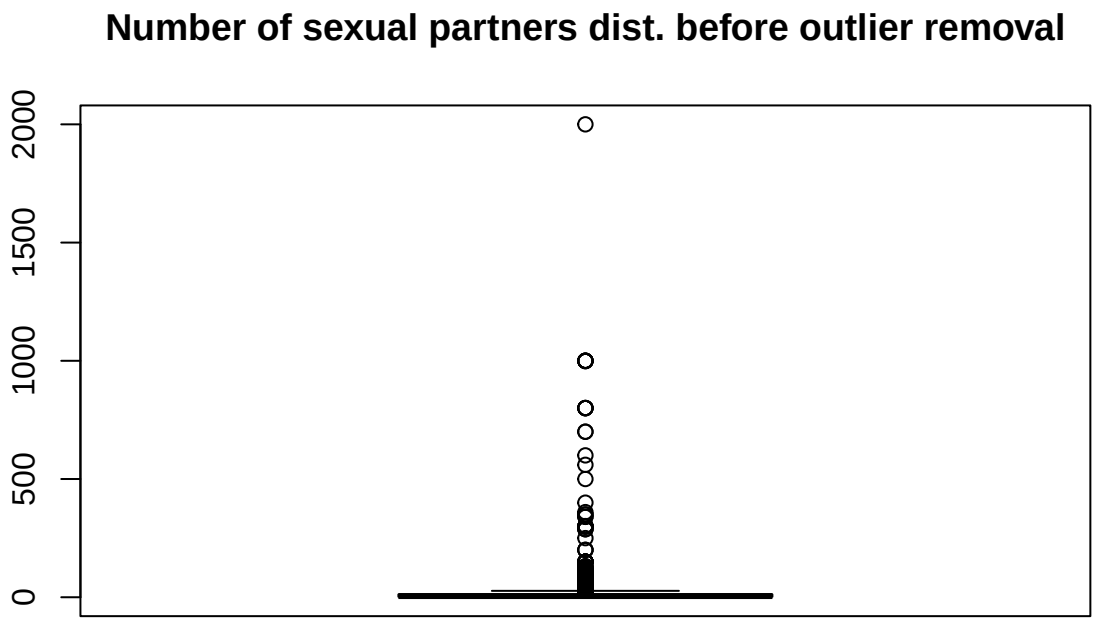

```
hist(df$SexNumPartYear, main = )
```




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



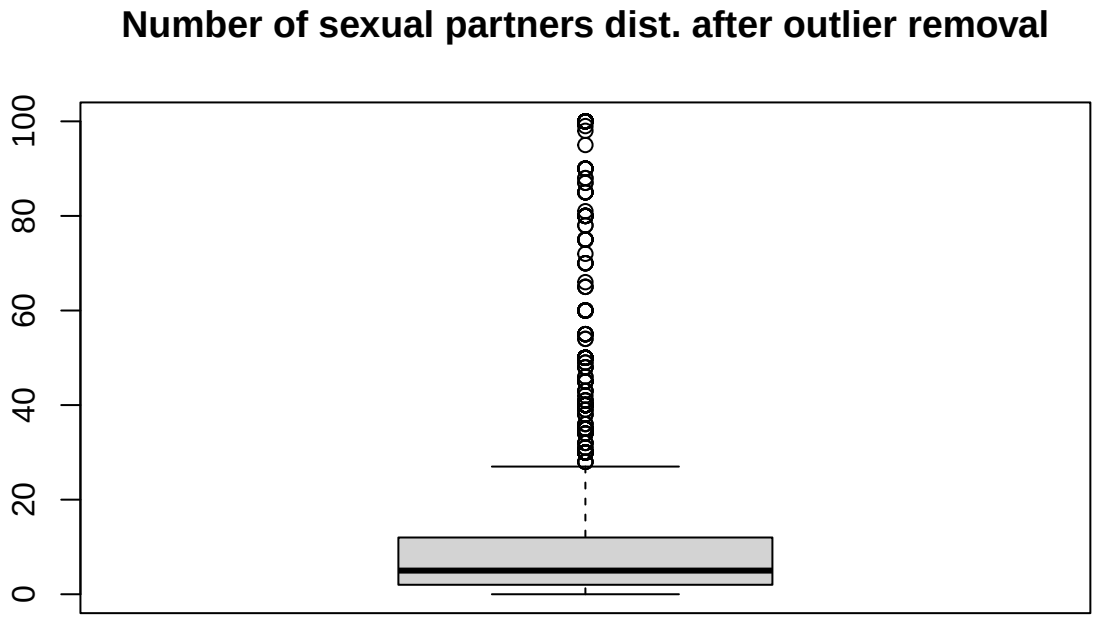

```
#Show observations with more than 300 sexual partners during lifetime
boxplot(df$SexNumPartnLife, main = "Number of sexual partners dist. before outlier removal")
```



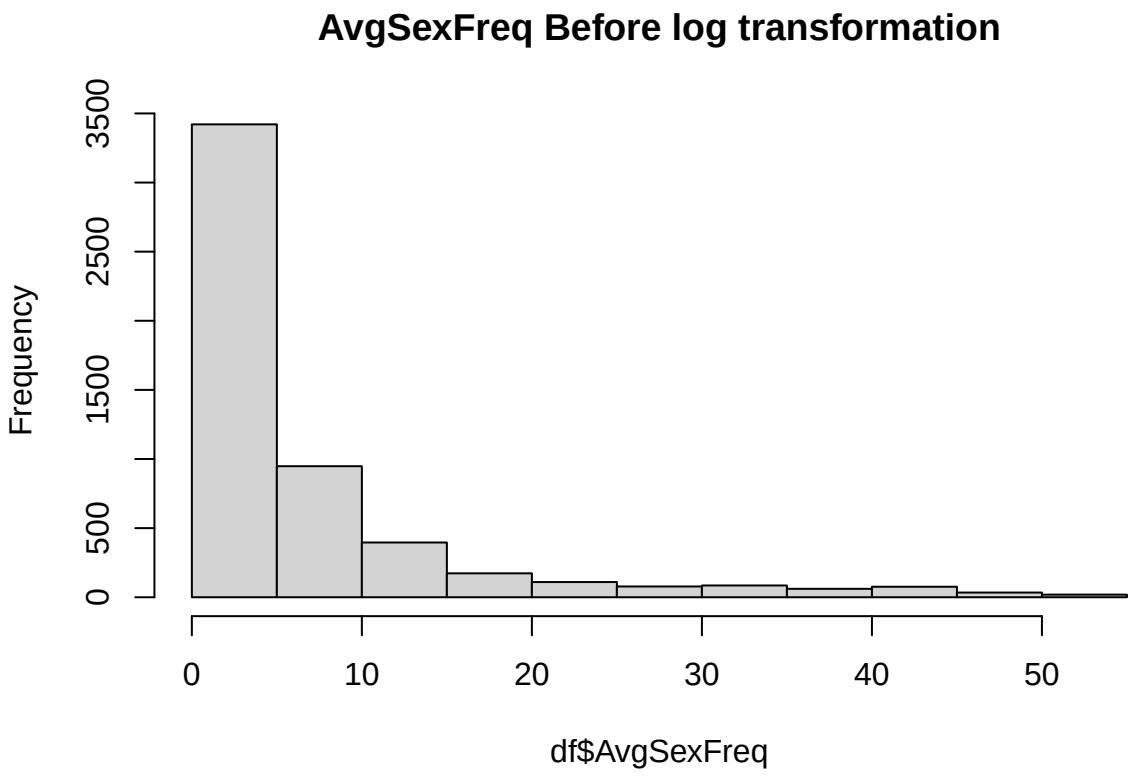

```
df[which(df$SexNumPartnLife > 100), c("Age", "SexAge", "SexNumPartnLife")]
```

```
## # A tibble: 64 x 3
##      Age SexAge SexNumPartnLife
##    <int>  <int>           <int>
##  1    61     15             288
##  2    61     15             288
##  3    61     15             288
##  4    37     12             126
##  5    37     12             126
##  6    63     18             301
##  7    51     13             131
##  8    51     13             131
##  9    39      9             120
## 10    59     13             150
## # i 54 more rows
```

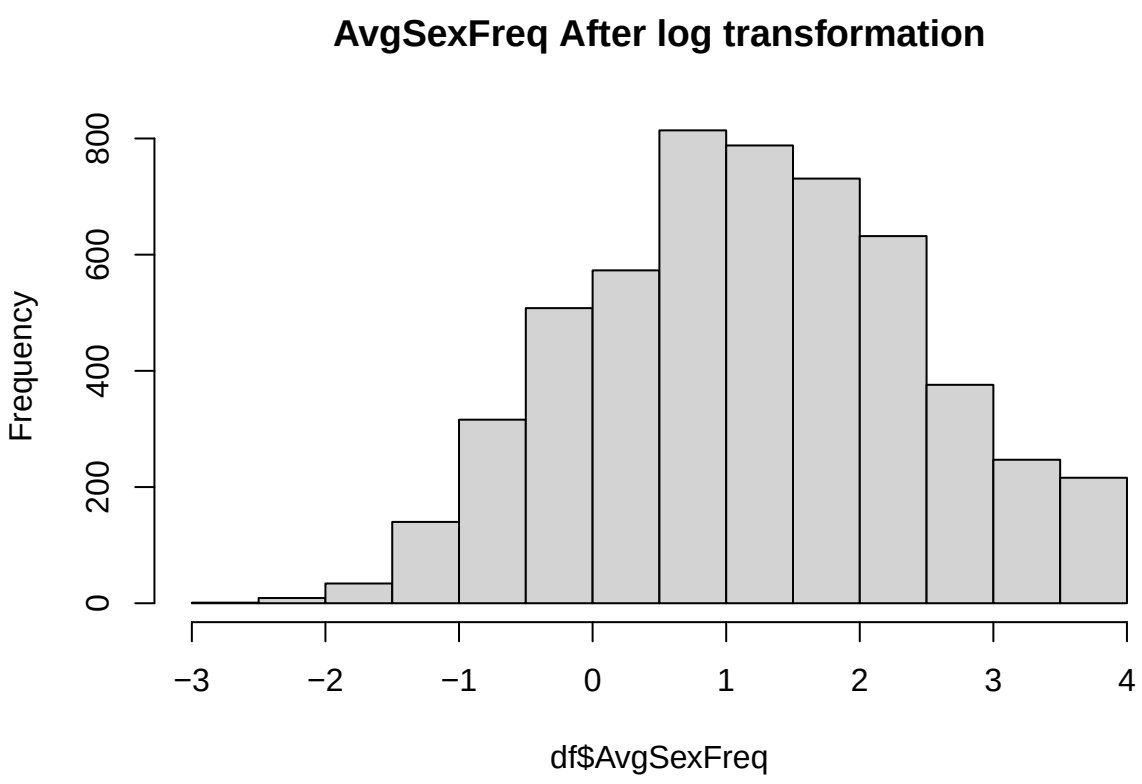
```
df = df[-which(df$SexNumPartnLife > 100),]
boxplot(df$SexNumPartnLife, main = "Number of sexual partners dist. after outlier removal")
```




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




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




```
#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
## -1.4422 -0.2785  0.1172  0.3269  1.9025
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.951383   1.391141  -0.684   0.4973
## SmokeNowYes                  0.289089   0.300845   0.961   0.3413
## AlcoholYear                 -0.001954   0.001615  -1.210   0.2320
## RegularMarijYes              0.713001   0.306404   2.327   0.0241 *
## HardDrugsYes                -1.158128   0.585547  -1.978   0.0536 .
```

```
## Age                          0.055766   0.022981   2.427   0.0190 *
## Gendermale                  -1.295412   0.261920  -4.946 9.31e-06 ***
## HHIncome 5000-9999          -0.866948   0.611280  -1.418   0.1624
## HHIncome10000-14999         -1.272802   0.523385  -2.432   0.0187 *
## HHIncome15000-19999          0.321837   0.868897   0.370   0.7127
## HHIncome20000-24999         -0.486674   0.569341  -0.855   0.3968
## HHIncome25000-34999         -0.473260   0.543180  -0.871   0.3879
## HHIncome35000-44999         -0.010203   0.504876  -0.020   0.9840
## HHIncome45000-54999         -1.915527   0.720635  -2.658   0.0106 *
## HHIncome55000-64999          0.408874   0.591471   0.691   0.4926
## HHIncome65000-74999         -0.788735   0.583832  -1.351   0.1829
## HHIncome75000-99999          0.063837   0.627552   0.102   0.9194
## HHIncomemore 99999          -0.951669   0.505636  -1.882   0.0658 .
## Education9 - 11th Grade     -0.363710   0.471323  -0.772   0.4440
## EducationHigh School        -0.087472   0.550426  -0.159   0.8744
## EducationSome College       -0.013425   0.476881  -0.028   0.9777
## EducationCollege Grad        0.652436   0.600570   1.086   0.2826
## BMI                          0.014850   0.017643   0.842   0.4040
## DiabetesAge                 -0.003065   0.014383  -0.213   0.8321
## DepressedSeveral            -0.373772   0.354654  -1.054   0.2971
## DepressedMost               -0.054524   0.431555  -0.126   0.9000
## LittleInterestSeveral        0.028186   0.323442   0.087   0.9309
## LittleInterestMost           0.638909   0.362941   1.760   0.0846 .
## PhysActiveYes               -0.191463   0.320525  -0.597   0.5530
## SameSexYes                   0.186025   0.470657   0.395   0.6944
## RegularMarijYes:HardDrugsYes 0.693527   0.670479   1.034   0.3060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6988 on 49 degrees of freedom
##   (9856 observations deleted due to missingness)
## Multiple R-squared:  0.7539, Adjusted R-squared:  0.6033
## F-statistic: 5.004 on 30 and 49 DF,  p-value: 3.392e-07
```

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

Characteristic	**Beta**	**95% CI**[1]	**p-value**
(Intercept)	-0.95	-3.7, 1.8	0.5
SmokeNow			
No	—	—	
Yes	0.29	-0.32, 0.89	0.3
AlcoholYear	0.00	-0.01, 0.00	0.2
RegularMarij			
No	—	—	
Yes	0.71	0.10, 1.3	0.024
HardDrugs			
No	—	—	
Yes	-1.2	-2.3, 0.02	0.054
Age	0.06	0.01, 0.10	0.019
Gender			

female	—	—	
male	-1.3	-1.8, -0.77	<0.001
HHIncome			
0-4999	—	—	
5000-9999	-0.87	-2.1, 0.36	0.2
10000-14999	-1.3	-2.3, -0.22	0.019
15000-19999	0.32	-1.4, 2.1	0.7
20000-24999	-0.49	-1.6, 0.66	0.4
25000-34999	-0.47	-1.6, 0.62	0.4
35000-44999	-0.01	-1.0, 1.0	>0.9
45000-54999	-1.9	-3.4, -0.47	0.011
55000-64999	0.41	-0.78, 1.6	0.5
65000-74999	-0.79	-2.0, 0.38	0.2
75000-99999	0.06	-1.2, 1.3	>0.9
more 99999	-0.95	-2.0, 0.06	0.066
Education			
8th Grade	—	—	
9 - 11th Grade	-0.36	-1.3, 0.58	0.4
High School	-0.09	-1.2, 1.0	0.9
Some College	-0.01	-0.97, 0.94	>0.9
College Grad	0.65	-0.55, 1.9	0.3
BMI	0.01	-0.02, 0.05	0.4
DiabetesAge	0.00	-0.03, 0.03	0.8
Depressed			
None	—	—	
Several	-0.37	-1.1, 0.34	0.3
Most	-0.05	-0.92, 0.81	0.9
LittleInterest			
None	—	—	
Several	0.03	-0.62, 0.68	>0.9
Most	0.64	-0.09, 1.4	0.085
PhysActive			
No	—	—	
Yes	-0.19	-0.84, 0.45	0.6
SameSex			
No	—	—	
Yes	0.19	-0.76, 1.1	0.7
RegularMarij * HardDrugs			
Yes * Yes	0.69	-0.65, 2.0	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

```r
n = 49
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  1.4520  1.4520  2.9738 0.0909273 .
## AlcoholYear               1  4.9797  4.9797 10.1988 0.0024569 **
## RegularMarij              1  0.0737  0.0737  0.1509 0.6993258
## HardDrugs                 1  5.3955  5.3955 11.0503 0.0016842 **
## Age                       1 16.3073 16.3073 33.3982 5.115e-07 ***
## Gender                    1 15.7092 15.7092 32.1735 7.458e-07 ***
## HHIncome                 11 22.2403  2.0218  4.1409 0.0002531 ***
## Education                 4  1.4262  0.3566  0.7302 0.5756717
## BMI                       1  0.5390  0.5390  1.1040 0.2985508
## DiabetesAge               1  0.1886  0.1886  0.3862 0.5371604
## Depressed                 2  2.2372  1.1186  2.2910 0.1119060
## LittleInterest            2  1.9588  0.9794  2.0059 0.1454385
## PhysActive                1  0.2419  0.2419  0.4955 0.4848363
## SameSex                   1  0.0273  0.0273  0.0560 0.8139845
## RegularMarij:HardDrugs    1  0.5224  0.5224  1.0699 0.3060390
## Residuals                49 23.9251  0.4883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## [1] 12.42336
```

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

```
## [1] 5.088753744 0.001445015
```

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

```
## [1] 55.68302
```

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

```
## [1] 2.851052e+01 2.706927e-10
```

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

```
## [1] 5.192894
```

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

```
## [1] 2.12707028 0.08671153
```

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

```
## [1] 0.0273237
```

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

```
## [1] 10.635351411  0.002579227
```

(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	Stop	Not 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] 1761
```

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

```
library(gridExtra)
```

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

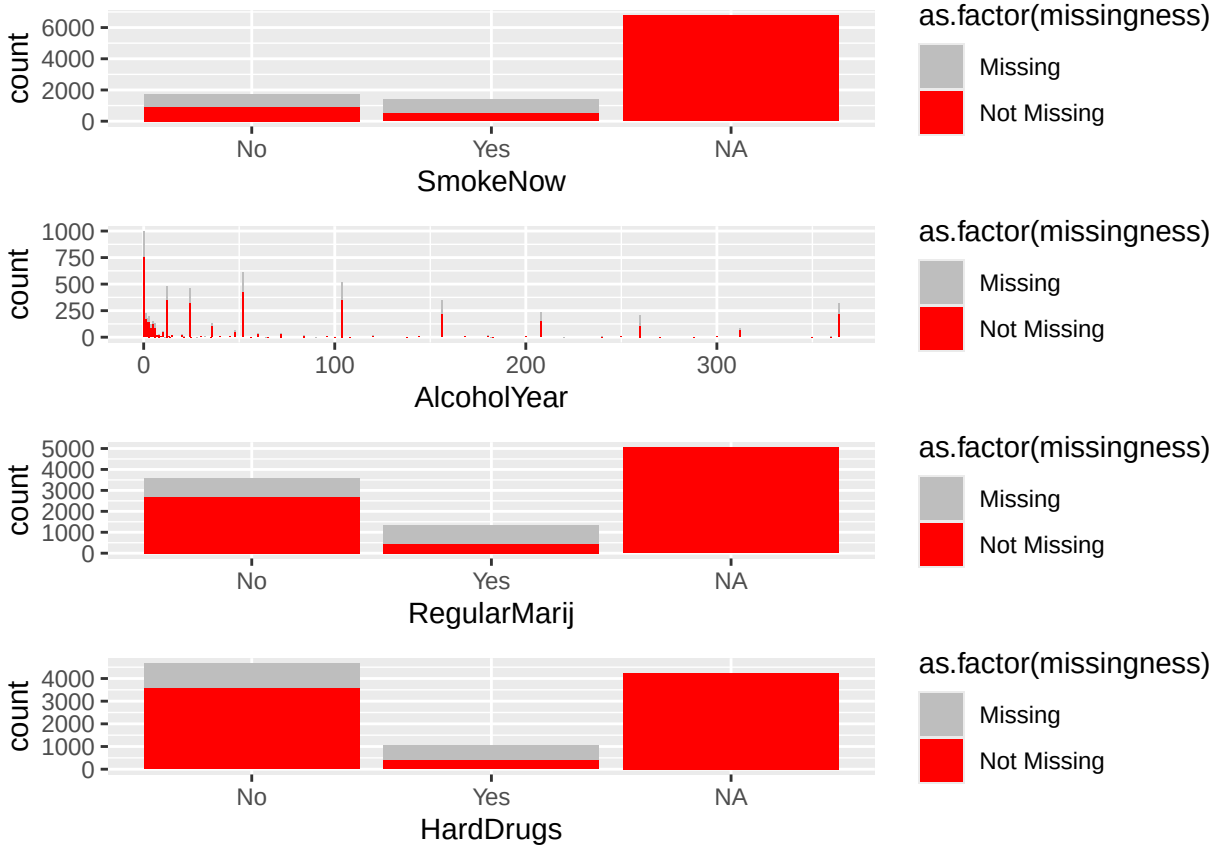
```
##
## Attaching package: 'gridExtra'
```

```
## 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)))+
```

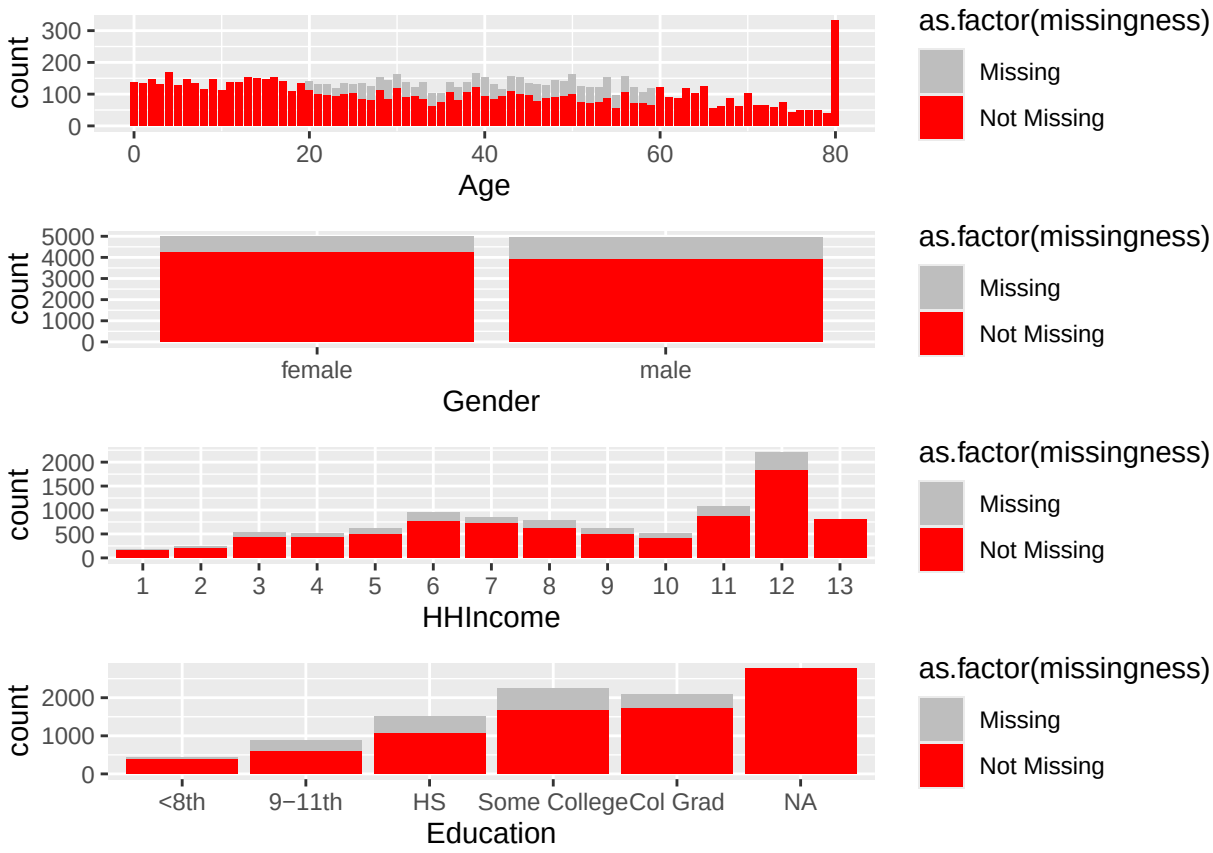
```
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p3 = ggplot(data = df, mapping=aes(x=RegularMarij, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p4 = ggplot(data = df, mapping=aes(x=HardDrugs, fill=as.factor(missingness)))+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

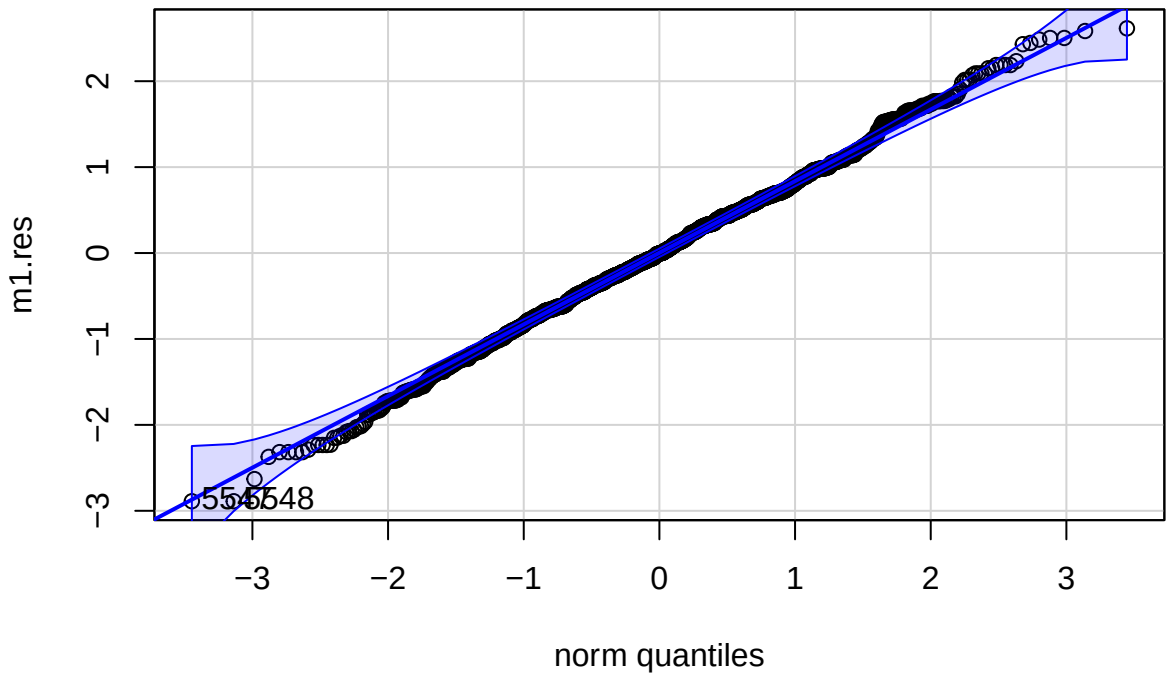



```
p6 = ggplot(data = df, mapping=aes(x=Age, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p7 = ggplot(data = df, mapping=aes(x=Gender, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p8 = ggplot(data = df, mapping=aes(x=HHIncome, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
   scale_x_discrete(labels = c(1,2,3,4,5,6,7,8,9, 10, 11, 12, 13)) +
  scale_fill_manual(values = c("gray", "red"))
p9 = ggplot(data = df, mapping=aes(x=Education, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
```

```
  scale_x_discrete(labels = c("<8th", "9-11th", "HS", "Some College", "Col Grad" )) +
  scale_fill_manual(values = c("gray", "red"))
grid.arrange(p6, p7, p8, p9, nrow = 4)
```

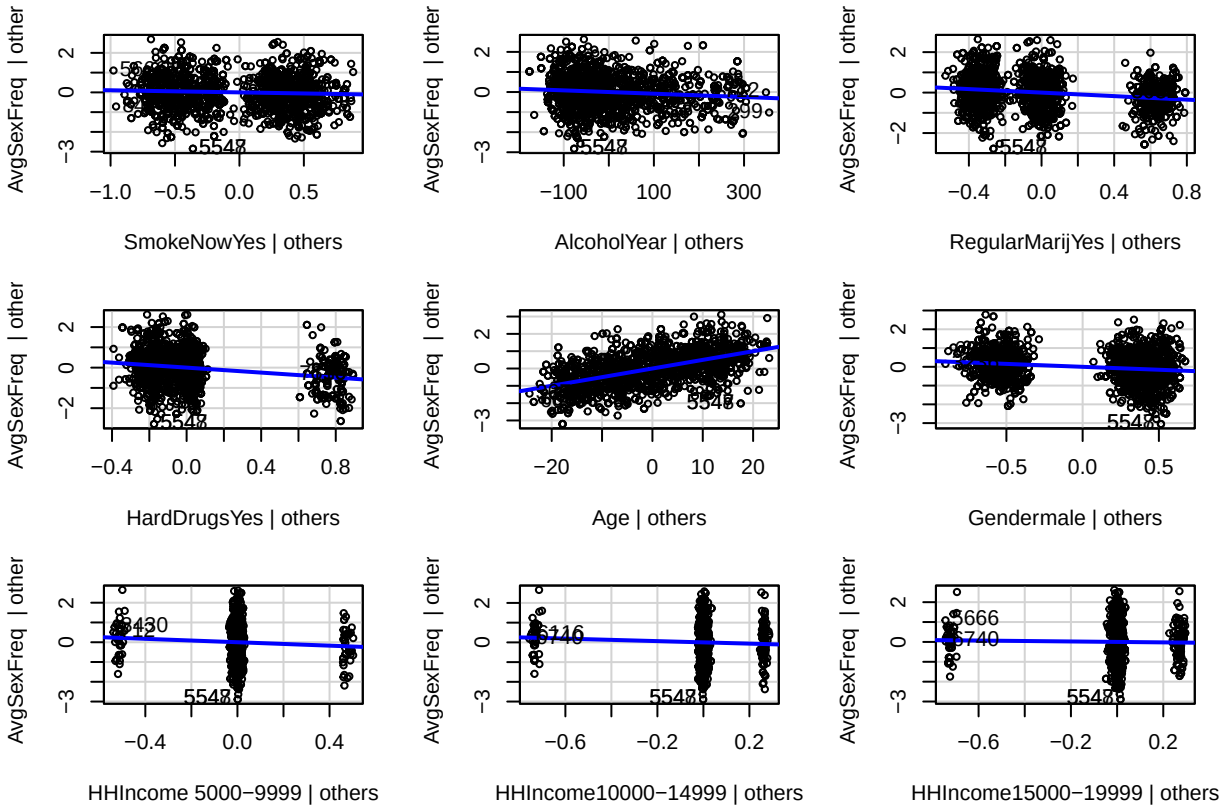


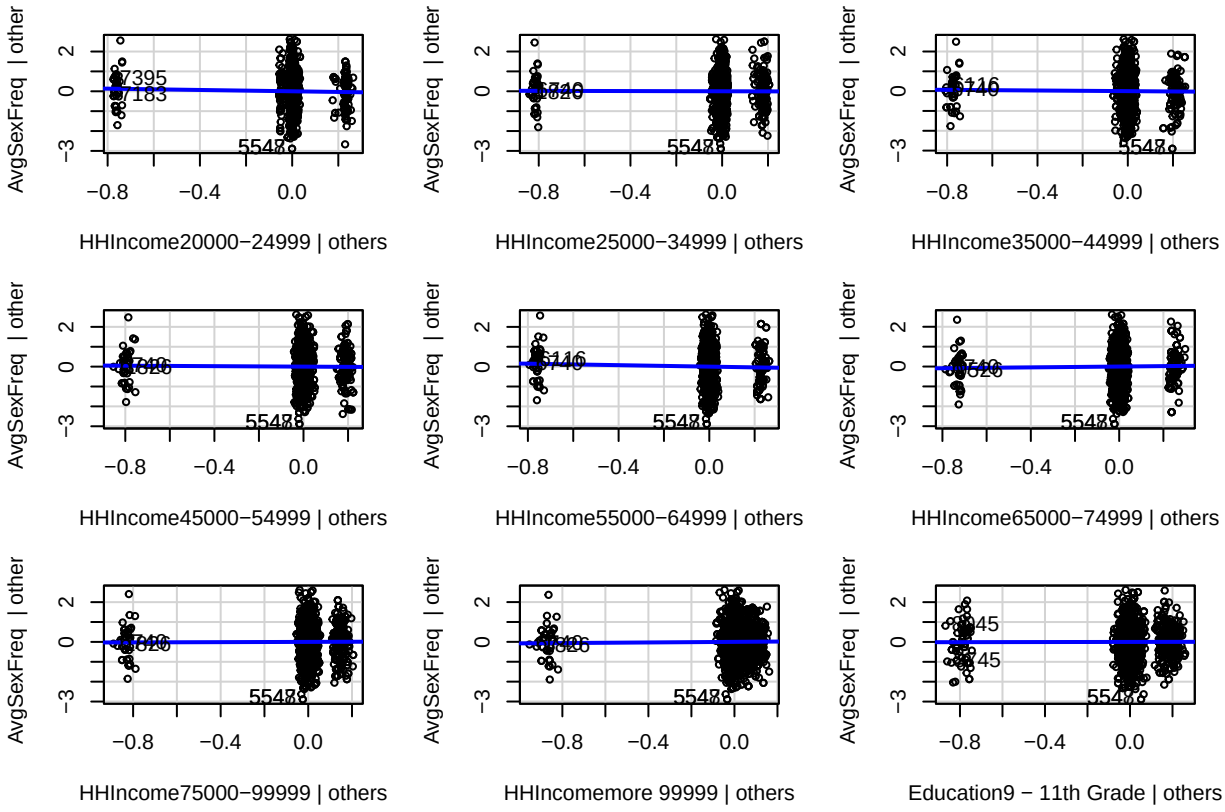

```
m1 = lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+HHIn
m1.res = m1$residuals

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

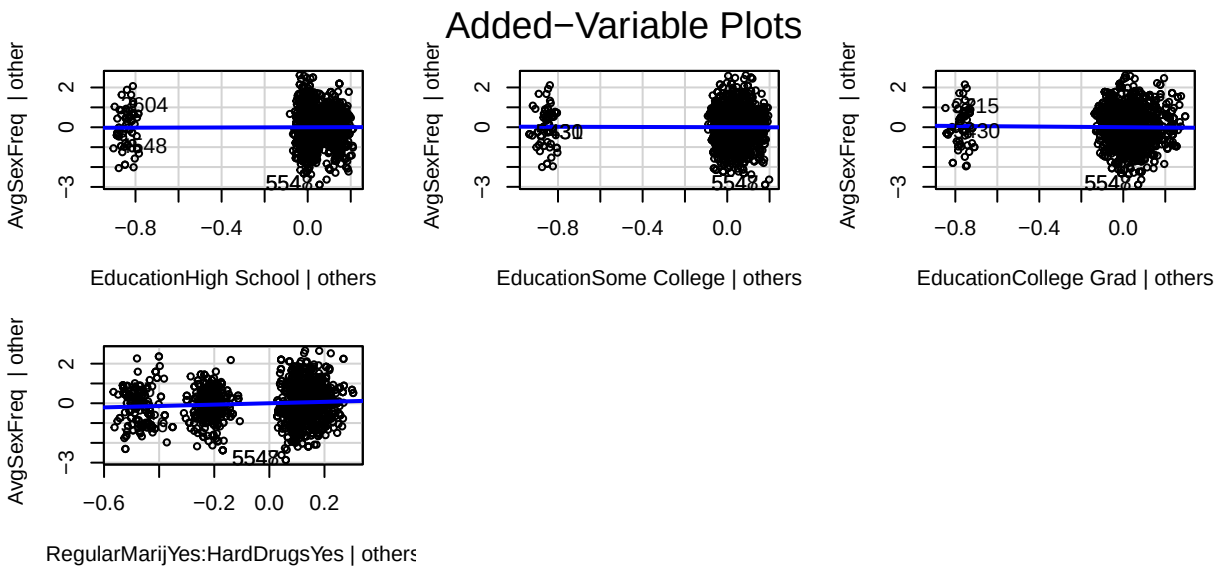



```
## 5547 5548
## 1013 1014
```

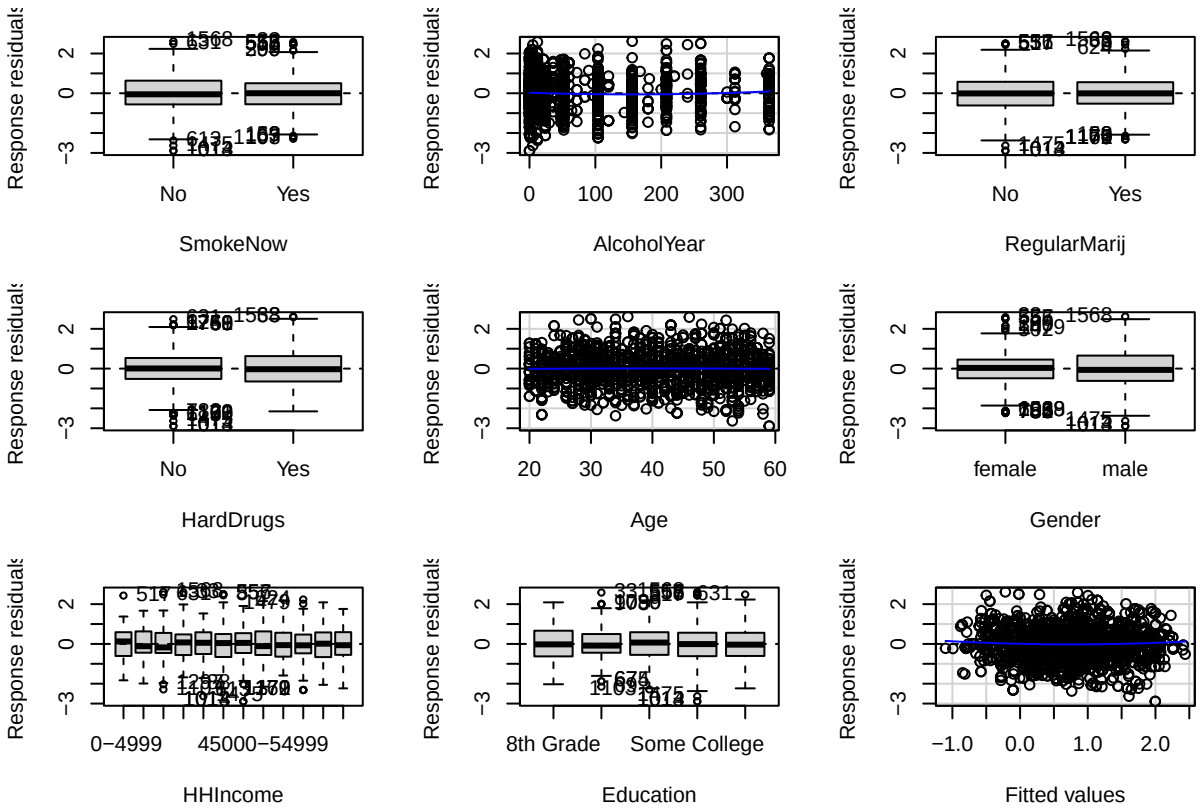
```
car::avPlots(m1)
```


AvgSexFreq | other
SmokeNowYes | others
AlcoholYear | others
RegularMarijYes | others
HardDrugsYes | others
Age | others
Gendermale | others
HHIncome 5000−9999 | others
HHIncome10000−14999 | others
HHIncome15000−19999 | others


AvgSexFreq | other
HHIncome20000−24999 | others
HHIncome25000−34999 | others
HHIncome35000−44999 | others
HHIncome45000−54999 | others
HHIncome55000−64999 | others
HHIncome65000−74999 | others
HHIncome75000−99999 | others
HHIncomemore 99999 | others
Education9 − 11th Grade | others




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




```
##              Test stat Pr(>|Test stat|)
## SmokeNow
## AlcoholYear     1.9664          0.04941 *
## RegularMarij
## HardDrugs
## Age            -0.2929          0.76966
## Gender
## HHIncome
## Education
## Tukey test      1.3957          0.16279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Use a non interactive model to check for collinearity
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     1.171232  1        1.082235           --
## AlcoholYear  1.119563  1        1.058094           --
## RegularMarij 1.034122  1        1.016918           --
## Age          1.092913  1        1.045425           --
## Gender       1.045458  1        1.022477           --
## HHIncome     1.431548 11        1.016441           --
## Education    1.412827  4        1.044146           --
```

```
##                                                          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
```

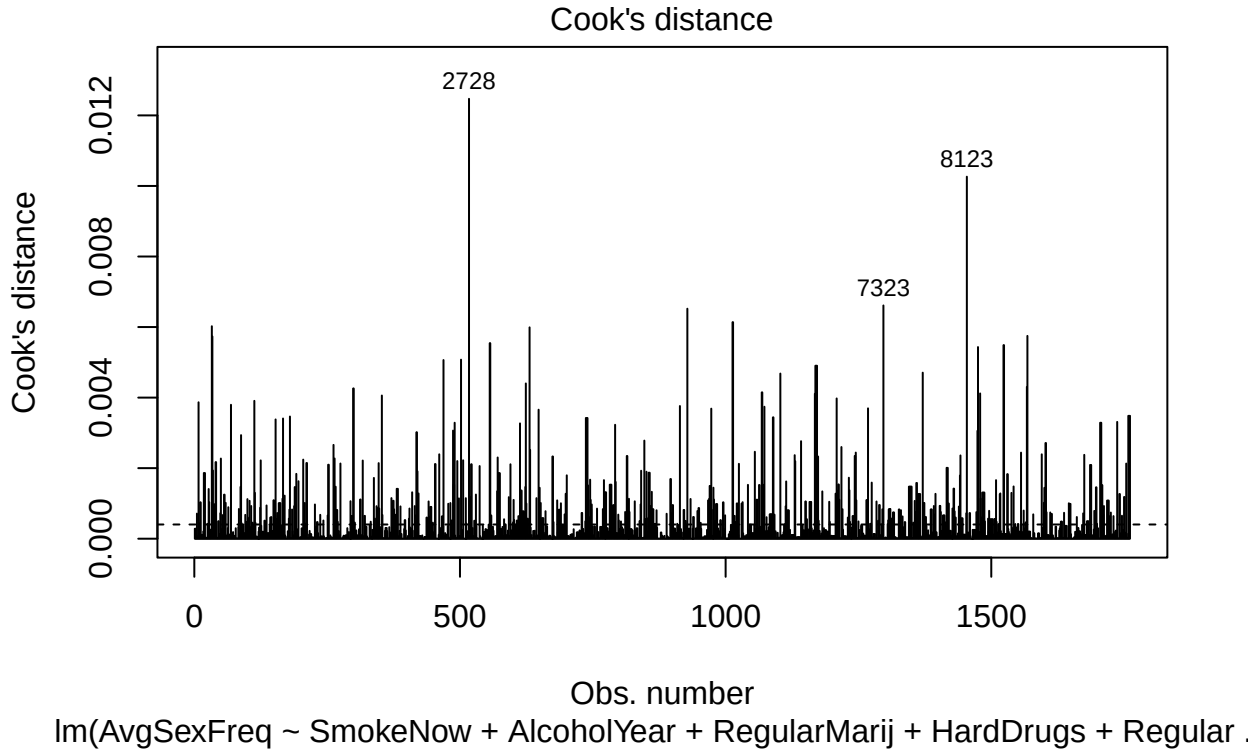
```
model.deffits=dffits(m1)
model.CD = cooks.distance(m1)
model.deffits[which.max(model.deffits)]
```

```
##      2728
## 0.5366936
```

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

```
##       2728
## 0.01247095
```

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

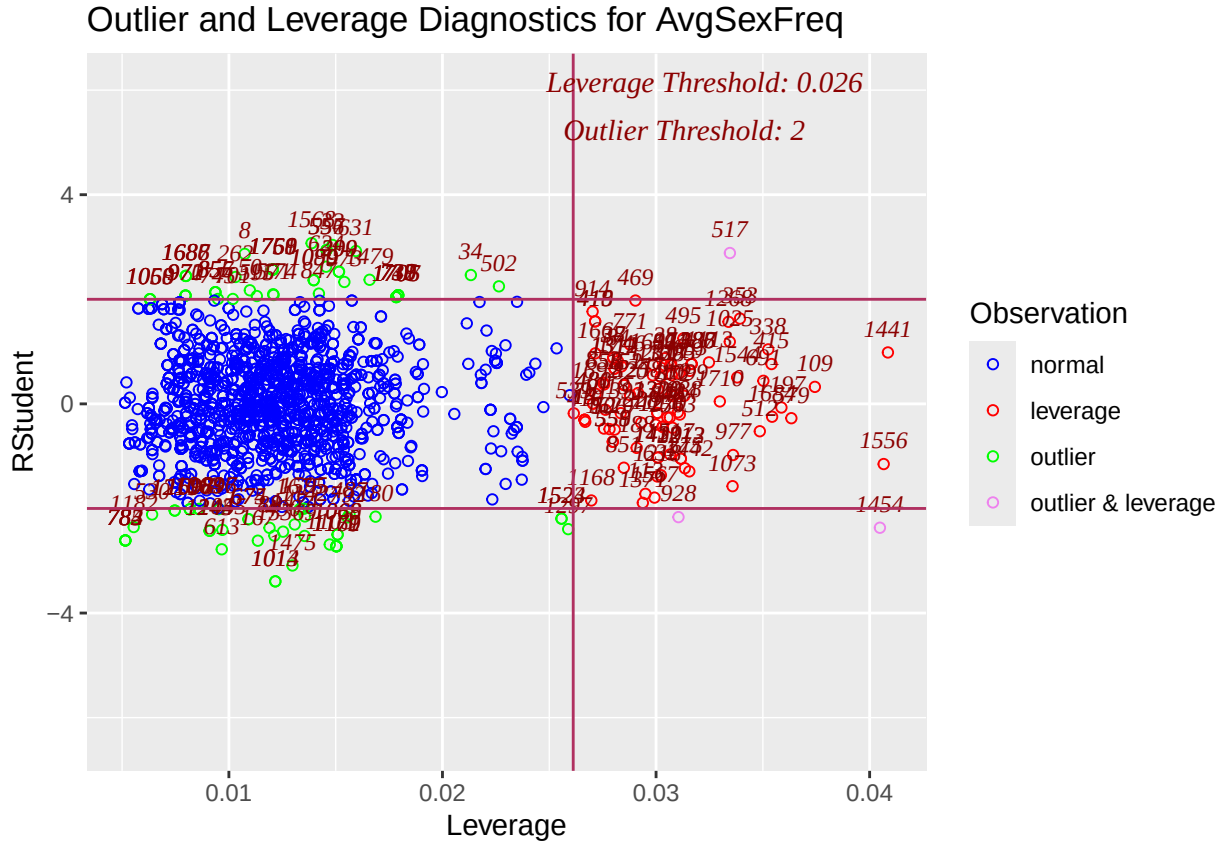



```
df[c(2737, 3315, 8155),]
```

```
## # A tibble: 3 x 78
```

```
##      ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education
##   <int> <fct>    <fct>  <int> <fct>         <int> <fct> <fct> <fct>
## 1 57426 2009_10  male      12 " 10-19"        152 Black <NA>  <NA>
## 2 58668 2009_10  female    65 " 60-69"        783 White <NA>  High School
## 3 68447 2011_12  female    68 " 60-69"         NA White White Some College
## # 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(517, 928, 1454),]
```

```
## # A tibble: 3 x 78
##      ID SurveyYr Gender   Age AgeDecade AgeMonths Race1    Race3 Education
##   <int> <fct>    <fct>  <int> <fct>         <int> <fct>    <fct> <fct>
## 1 52676 2009_10  female     8 " 0-9"           99 White    <NA>  <NA>
## 2 53515 2009_10  male      41 " 40-49"        503 White    <NA>  High School
## 3 54659 2009_10  female    45 " 40-49"        544 Hispanic <NA>  High School
## # 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>, ...
```

```
df = mutate(df, nlAvgSexFreq = (Age-SexAge)/SexNumPartnLife)
df$nlAvgSexFreq[is.infinite(df$nlAvgSexFreq)] = NA
m1 = lm(nlAvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+HH
df2 = df[-c(517, 928, 1454),]
m2 = lm(nlAvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+HH
summary(m1)
```

```
##
## Call:
## lm(formula = nlAvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + HHIncome +
##     Education, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.063 -2.382 -0.641  0.979 32.539
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.354594   1.034953  -0.343 0.731927
## SmokeNowYes                  -0.538056   0.233876  -2.301 0.021532 *
## AlcoholYear                  -0.003828   0.001067  -3.586 0.000345 ***
## RegularMarijYes              -1.919860   0.289391  -6.634 4.35e-11 ***
## HardDrugsYes                 -2.489531   0.387660  -6.422 1.73e-10 ***
## Age                           0.169582   0.010007  16.946  < 2e-16 ***
## Gendermale                   -0.639674   0.223670  -2.860 0.004288 **
## HHIncome 5000-9999           -1.070635   1.010039  -1.060 0.289295
## HHIncome10000-14999          -1.111943   0.845926  -1.314 0.188862
## HHIncome15000-19999           0.023072   0.851923   0.027 0.978397
## HHIncome20000-24999          -0.763861   0.829435  -0.921 0.357208
## HHIncome25000-34999           0.127388   0.801685   0.159 0.873766
## HHIncome35000-44999          -0.601915   0.821812  -0.732 0.464009
## HHIncome45000-54999          -0.596442   0.810263  -0.736 0.461764
## HHIncome55000-64999          -0.621963   0.829876  -0.749 0.453678
## HHIncome65000-74999           1.016601   0.847394   1.200 0.230428
## HHIncome75000-99999          -0.023887   0.793685  -0.030 0.975994
## HHIncomemore 99999            0.800385   0.774784   1.033 0.301728
## Education9 - 11th Grade      -0.171983   0.616568  -0.279 0.780328
## EducationHigh School          0.076190   0.597914   0.127 0.898617
## EducationSome College         0.063907   0.592290   0.108 0.914089
## EducationCollege Grad        -0.552754   0.625781  -0.883 0.377194
## RegularMarijYes:HardDrugsYes  1.426457   0.497623   2.867 0.004200 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.517 on 1738 degrees of freedom
##   (8175 observations deleted due to missingness)
## Multiple R-squared:  0.2381, Adjusted R-squared:  0.2284
## F-statistic: 24.68 on 22 and 1738 DF,  p-value: < 2.2e-16
```

```r
summary(m2)
```

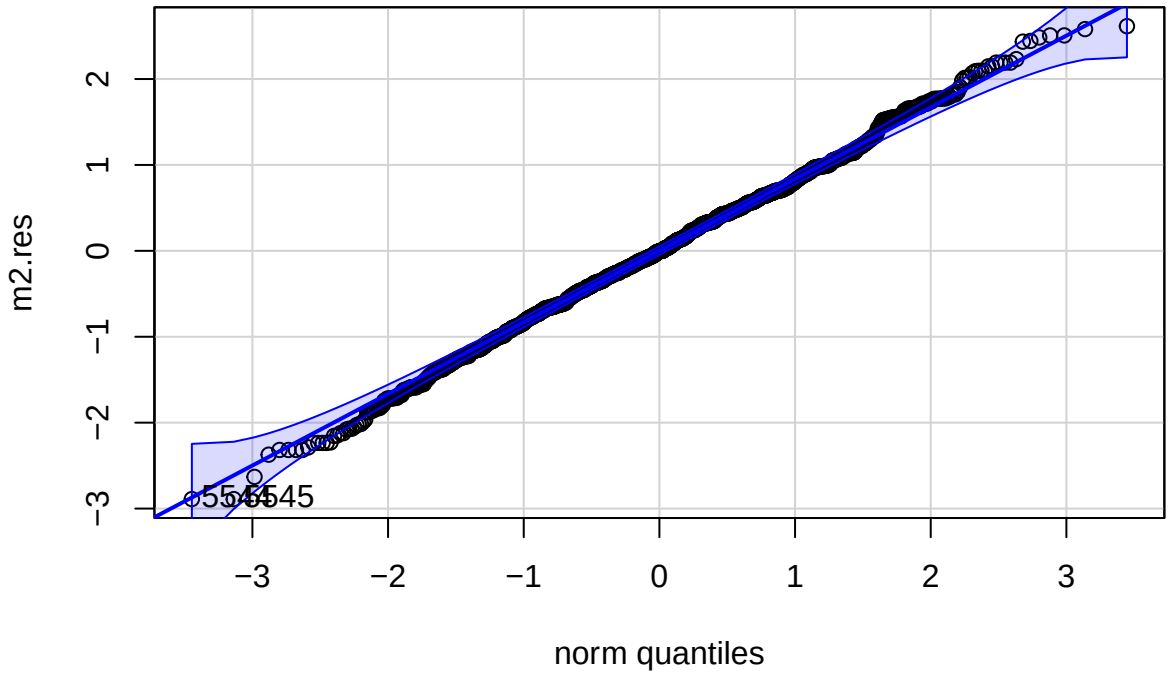
```
##
## Call:
## lm(formula = nlAvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + HHIncome +
##     Education, data = df2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -2.384 -0.634  0.980 32.540
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.353729   1.035189  -0.342 0.732614
## SmokeNowYes                  -0.535083   0.234018  -2.287 0.022344 *
## AlcoholYear                  -0.003832   0.001068  -3.589 0.000341 ***
## RegularMarijYes              -1.919842   0.289456  -6.633 4.39e-11 ***
## HardDrugsYes                 -2.502657   0.388784  -6.437 1.57e-10 ***
## Age                           0.169613   0.010010  16.945  < 2e-16 ***
## Gendermale                   -0.641904   0.223773  -2.869 0.004173 **
## HHIncome 5000-9999           -1.072550   1.010276  -1.062 0.288547
## HHIncome10000-14999          -1.112305   0.846117  -1.315 0.188818
## HHIncome15000-19999           0.022864   0.852116   0.027 0.978597
## HHIncome20000-24999          -0.765202   0.829627  -0.922 0.356477
## HHIncome25000-34999           0.127460   0.801866   0.159 0.873724
## HHIncome35000-44999          -0.602242   0.821998  -0.733 0.463867
## HHIncome45000-54999          -0.597484   0.810450  -0.737 0.461085
## HHIncome55000-64999          -0.622939   0.830066  -0.750 0.453074
## HHIncome65000-74999           1.017159   0.847587   1.200 0.230278
## HHIncome75000-99999          -0.034785   0.794214  -0.044 0.965071
## HHIncomemore 99999            0.798821   0.774966   1.031 0.302787
## Education9 - 11th Grade      -0.172022   0.616708  -0.279 0.780327
## EducationHigh School          0.071475   0.598136   0.119 0.904896
## EducationSome College         0.065376   0.592432   0.110 0.912143
## EducationCollege Grad        -0.549730   0.625956  -0.878 0.379943
## RegularMarijYes:HardDrugsYes  1.440040   0.498600   2.888 0.003923 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.518 on 1737 degrees of freedom
##   (8173 observations deleted due to missingness)
## Multiple R-squared:  0.2381, Adjusted R-squared:  0.2285
## F-statistic: 24.68 on 22 and 1737 DF,  p-value: < 2.2e-16
```

```r
100*(abs(coef(m1)-coef(m2)))/coef(m1)
```

```
##                  (Intercept)                  SmokeNowYes
##                -2.439842e-01                -5.525667e-01
##                  AlcoholYear              RegularMarijYes
##                -1.120890e-01                -9.062635e-04
##                 HardDrugsYes                          Age
##                -5.272558e-01                 1.789284e-02
##                   Gendermale           HHIncome 5000-9999
```

```
##                -3.486711e-01              -1.788089e-01
##          HHIncome10000-14999          HHIncome15000-19999
##                -3.257957e-02               9.011864e-01
##          HHIncome20000-24999          HHIncome25000-34999
##                -1.754694e-01               5.649775e-02
##          HHIncome35000-44999          HHIncome45000-54999
##                -5.439758e-02              -1.748140e-01
##          HHIncome55000-64999          HHIncome65000-74999
##                -1.568400e-01               5.487230e-02
##          HHIncome75000-99999           HHIncomemore 99999
##                -4.562366e+01               1.953742e-01
##      Education9 - 11th Grade         EducationHigh School
##                -2.305047e-02               6.188849e+00
##        EducationSome College        EducationCollege Grad
##                 2.298324e+00              -5.470884e-01
## RegularMarijYes:HardDrugsYes
##                 9.522050e-01
```

```
m2 = lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+HHIn

m2.res = m2$residuals

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




```
## 5544 5545
## 1012 1013
```