BIOSTAT 650 Project

Jaehoon Kim

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')
```

```
model <- lm(BPSys1 ~ Age+Gender+Poverty+BMI+SleepHrsNight+PhysActiveDays+SmokeNow+AlcoholYear+HardDrugs
summary(model)
```

```
##
## Call:
## lm(formula = BPSys1 ~ Age + Gender + Poverty + BMI + SleepHrsNight +
##     PhysActiveDays + SmokeNow + AlcoholYear + HardDrugs, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.397  -8.387  -0.997   7.730  69.906
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    89.959564   3.820975  23.544  < 2e-16 ***
## Age             0.413402   0.035437  11.666  < 2e-16 ***
## Gendermale      5.382522   0.903317   5.959 3.48e-09 ***
## Poverty        -0.843665   0.283924  -2.971  0.00303 **
## BMI             0.345235   0.075337   4.583 5.15e-06 ***
## SleepHrsNight   0.247155   0.331007   0.747  0.45543
## PhysActiveDays -0.021275   0.244823  -0.087  0.93077
## SmokeNowYes     1.325291   0.957252   1.384  0.16651
## AlcoholYear     0.002536   0.004169   0.608  0.54318
## HardDrugsYes    0.141125   0.964282   0.146  0.88367
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.18 on 1038 degrees of freedom
##   (8952 observations deleted due to missingness)
## Multiple R-squared:  0.1709, Adjusted R-squared:  0.1637
## F-statistic: 23.78 on 9 and 1038 DF,  p-value: < 2.2e-16
```

```
model <- lm(SexAge ~ Depressed+LittleInterest+HealthGen+Gender+HHIncome+Education+PhysActive+RegularMar
summary(model)
```

```
##
## Call:
## lm(formula = SexAge ~ Depressed + LittleInterest + HealthGen +
##     Gender + HHIncome + Education + PhysActive + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Depressed * HardDrugs +
##     SmokeAge, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2968 -1.4972 -0.1227  1.1686 20.5223
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 16.342991   0.624806  26.157  < 2e-16 ***
## DepressedSeveral            -0.177236   0.241818  -0.733 0.463700
## DepressedMost               -1.291956   0.374178  -3.453 0.000568 ***
## LittleInterestSeveral       -0.231825   0.191238  -1.212 0.225587
## LittleInterestMost           0.322324   0.277909   1.160 0.246281
## HealthGenVgood               0.200654   0.267130   0.751 0.452665
## HealthGenGood               -0.340287   0.264213  -1.288 0.197942
## HealthGenFair               -0.002334   0.300057  -0.008 0.993793
## HealthGenPoor               -0.184880   0.467620  -0.395 0.692623
## Gendermale                   0.304082   0.129913   2.341 0.019362 *
## HHIncome 5000-9999          -1.348405   0.557167  -2.420 0.015618 *
## HHIncome10000-14999         -1.088389   0.480505  -2.265 0.023629 *
## HHIncome15000-19999         -1.294652   0.483536  -2.677 0.007488 **
## HHIncome20000-24999         -1.369399   0.477907  -2.865 0.004215 **
## HHIncome25000-34999         -0.949078   0.460535  -2.061 0.039469 *
## HHIncome35000-44999         -1.471535   0.469899  -3.132 0.001767 **
## HHIncome45000-54999         -0.426089   0.466347  -0.914 0.361014
## HHIncome55000-64999         -1.784112   0.478566  -3.728 0.000199 ***
## HHIncome65000-74999         -0.933033   0.488515  -1.910 0.056305 .
## HHIncome75000-99999         -1.144292   0.456791  -2.505 0.012333 *
## HHIncomemore 99999          -1.242224   0.442429  -2.808 0.005045 **
## Education9 - 11th Grade     -0.218123   0.341017  -0.640 0.522501
## EducationHigh School        -0.179374   0.332905  -0.539 0.590085
## EducationSome College        0.189442   0.332127   0.570 0.568486
## EducationCollege Grad        1.445331   0.352639   4.099 4.35e-05 ***
## PhysActiveYes               -0.599686   0.133608  -4.488 7.65e-06 ***
## RegularMarijYes             -1.256137   0.167049  -7.520 8.74e-14 ***
## HardDrugsYes                -0.891059   0.248838  -3.581 0.000352 ***
## SmokeAge                     0.100107   0.013415   7.462 1.34e-13 ***
## RegularMarijYes:HardDrugsYes   0.834558   0.290879   2.869 0.004166 **
## DepressedSeveral:HardDrugsYes -0.184463   0.332563  -0.555 0.579190
## DepressedMost:HardDrugsYes     0.565576   0.465395   1.215 0.224432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.624 on 1744 degrees of freedom
##   (8224 observations deleted due to missingness)
## Multiple R-squared:  0.1699, Adjusted R-squared:  0.1551
## F-statistic: 11.51 on 31 and 1744 DF,  p-value: < 2.2e-16
```

```
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
```

```
model <- lm(SexAge ~ Gender+HHIncome+Education+SameSex+PhysActive+RegularMarij+HardDrugs+RegularMarij*Ha
summary(model)
```

```
##
## Call:
## lm(formula = SexAge ~ Gender + HHIncome + Education + SameSex +
##     PhysActive + RegularMarij + HardDrugs + RegularMarij * HardDrugs,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9073 -1.9665 -0.4121  1.2964 27.4144
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 17.54801    0.50328  34.867  < 2e-16 ***
## Gendermale                  -0.07223    0.10749  -0.672   0.5016
## HHIncome 5000-9999          -0.79270    0.54506  -1.454   0.1459
## HHIncome10000-14999         -0.44989    0.46490  -0.968   0.3332
## HHIncome15000-19999         -1.06281    0.46658  -2.278   0.0228 *
## HHIncome20000-24999         -0.44484    0.45888  -0.969   0.3324
## HHIncome25000-34999         -0.38598    0.43784  -0.882   0.3781
## HHIncome35000-44999         -0.18232    0.43789  -0.416   0.6772
## HHIncome45000-54999          0.35222    0.43915   0.802   0.4226
## HHIncome55000-64999         -0.73119    0.44760  -1.634   0.1024
## HHIncome65000-74999           0.32731    0.45372   0.721   0.4707    
## HHIncome75000-99999           0.08799    0.42898   0.205   0.8375    
## HHIncomemore 99999           -0.25391    0.41941  -0.605   0.5449    
## Education9 - 11th Grade       0.16340    0.33500   0.488   0.6257    
## EducationHigh School          0.52625    0.31954   1.647   0.0997 .  
## EducationSome College         0.53590    0.31488   1.702   0.0888 .  
## EducationCollege Grad         1.93066    0.32478   5.945 3.00e-09 ***
## SameSexYes                   -0.49517    0.19924  -2.485   0.0130 *  
## PhysActiveYes                -0.24524    0.11221  -2.186   0.0289 *  
## RegularMarijYes              -2.01369    0.15549 -12.950  < 2e-16 ***
## HardDrugsYes                 -1.54232    0.21857  -7.056 1.99e-12 ***
## RegularMarijYes:HardDrugsYes  1.46429    0.29139   5.025 5.24e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.397 on 4203 degrees of freedom
##   (5775 observations deleted due to missingness)
## Multiple R-squared:  0.1372, Adjusted R-squared:  0.1328 
## F-statistic: 31.81 on 21 and 4203 DF,  p-value: < 2.2e-16
```

```r
model <- lm(SexNumPartnLife ~ Gender+HHIncome+Education+PhysActive+RegularMarij+HardDrugs+RegularMarij*
summary(model)
```

```
## 
## Call:
## lm(formula = SexNumPartnLife ~ Gender + HHIncome + Education + 
##     PhysActive + RegularMarij + HardDrugs + RegularMarij * HardDrugs, 
##     data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -43.88 -11.51  -4.29   2.76 985.61 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  -3.10099    7.13864  -0.434   0.6640    
## Gendermale                    8.77546    1.51990   5.774 8.30e-09 ***
## HHIncome 5000-9999           14.54638    7.76891   1.872   0.0612 .  
## HHIncome10000-14999           3.78538    6.62111   0.572   0.5675    
## HHIncome15000-19999           0.04752    6.67954   0.007   0.9943    
## HHIncome20000-24999           8.46345    6.59501   1.283   0.1995    
## HHIncome25000-34999          11.18533    6.26544   1.785   0.0743 .  
## HHIncome35000-44999           1.12603    6.27352   0.179   0.8576    
## HHIncome45000-54999           1.67325    6.29487   0.266   0.7904    
## HHIncome55000-64999           2.52128    6.40564   0.394   0.6939    
## HHIncome65000-74999           3.25426    6.51323   0.500   0.6174    
## HHIncome75000-99999           4.36560    6.14932   0.710   0.4778    
## HHIncomemore 99999            4.36177    6.01363   0.725   0.4683    
## Education9 - 11th Grade       5.45707    4.69156   1.163   0.2448    
## EducationHigh School          4.54384    4.45914   1.019   0.3083    
## EducationSome College         1.14179    4.38485   0.260   0.7946    
## EducationCollege Grad        -2.03712    4.52072  -0.451   0.6523    
## PhysActiveYes                 3.02096    1.60090   1.887   0.0592 .  
## RegularMarijYes              13.61541    2.23551   6.091 1.22e-09 ***
## HardDrugsYes                 12.66710    3.11864   4.062 4.96e-05 ***
## RegularMarijYes:HardDrugsYes -4.10977    4.21049  -0.976   0.3291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.13 on 4323 degrees of freedom
##   (5656 observations deleted due to missingness)
## Multiple R-squared:  0.05162,    Adjusted R-squared:  0.04723
## F-statistic: 11.77 on 20 and 4323 DF,  p-value: < 2.2e-16
```

```
model <- lm(SexNumPartnLife ~ Gender+HHIncome+Education+PhysActive+SameSex+RegularMarij+HardDrugs+Regula
summary(model)
```

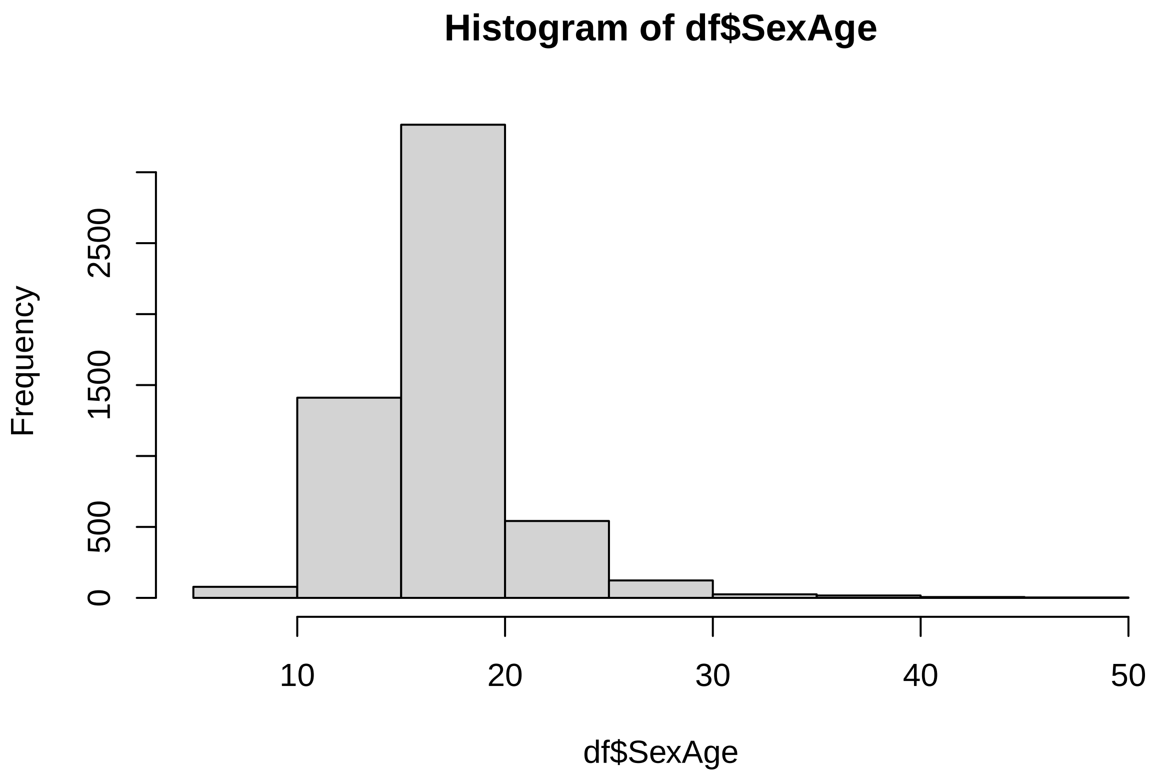
```
##
## Call:
## lm(formula = SexNumPartnLife ~ Gender + HHIncome + Education +
##     PhysActive + SameSex + RegularMarij + HardDrugs + RegularMarij *
##     HardDrugs, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43.99 -11.32  -4.30   2.80 985.80
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2.83227    7.15102  -0.396   0.6921
## Gendermale                    8.62320    1.53271   5.626 1.96e-08 ***
## HHIncome 5000-9999           14.55906    7.77014   1.874   0.0610 .
## HHIncome10000-14999           3.86482    6.62286   0.584   0.5595
## HHIncome15000-19999           0.06679    6.68064   0.010   0.9920
## HHIncome20000-24999           8.50076    6.59625   1.289   0.1976
## HHIncome25000-34999          11.17764    6.26741   1.783   0.0746 .
## HHIncome35000-44999           1.02913    6.27553   0.164   0.8697
## HHIncome45000-54999           1.68879    6.29584   0.268   0.7885
## HHIncome55000-64999           2.53680    6.40663   0.396   0.6922
## HHIncome65000-74999           3.05708    6.51876   0.469   0.6391
## HHIncome75000-99999           4.21680    6.15303   0.685   0.4932
## HHIncomemore 99999            4.27884    6.01544   0.711   0.4769
## Education9 - 11th Grade       5.35105    4.70437   1.137   0.2554
## EducationHigh School          4.45800    4.47243   0.997   0.3189
## EducationSome College         1.10825    4.39882   0.252   0.8011
## EducationCollege Grad        -2.03806    4.53482  -0.449   0.6531
## PhysActiveYes                 3.00891    1.60123   1.879   0.0603 .
## SameSexYes                   -2.32060    2.88395  -0.805   0.4211
## RegularMarijYes              13.77346    2.24501   6.135 9.27e-10 ***
## HardDrugsYes                 13.04387    3.15518   4.134 3.63e-05 ***
## RegularMarijYes:HardDrugsYes -4.26299    4.21578  -1.011   0.3120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.14 on 4321 degrees of freedom
##   (5657 observations deleted due to missingness)
## Multiple R-squared:  0.05177,    Adjusted R-squared:  0.04716
## F-statistic: 11.23 on 21 and 4321 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, given the age 9 was the lowest age at which first sexual activity occured among respondents, 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.

```
hist(df$SexAge)
```



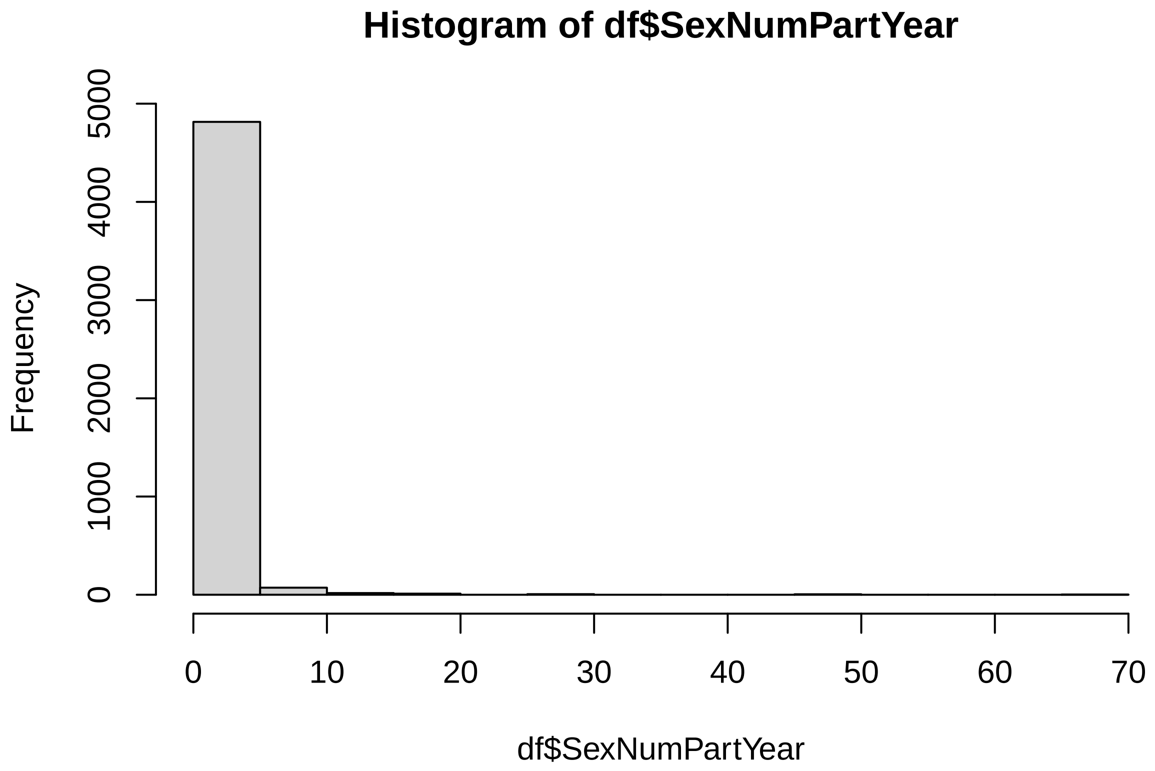
```
sort(unique(df$SexAge))
```

```
##  [1]  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
## [26] 34 35 36 37 38 39 41 44 47 50
```

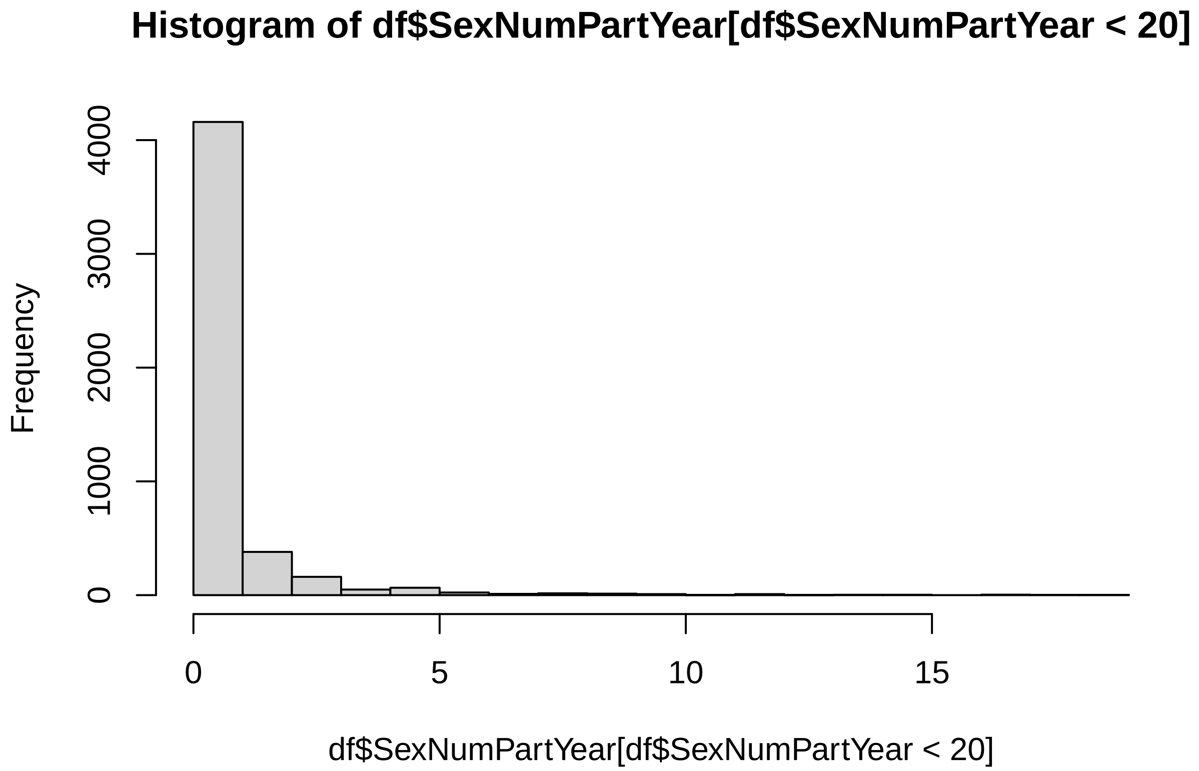
```
typeof(df$SexAge)
```

```
## [1] "integer"
```

```
hist(df$SexNumPartYear)
```



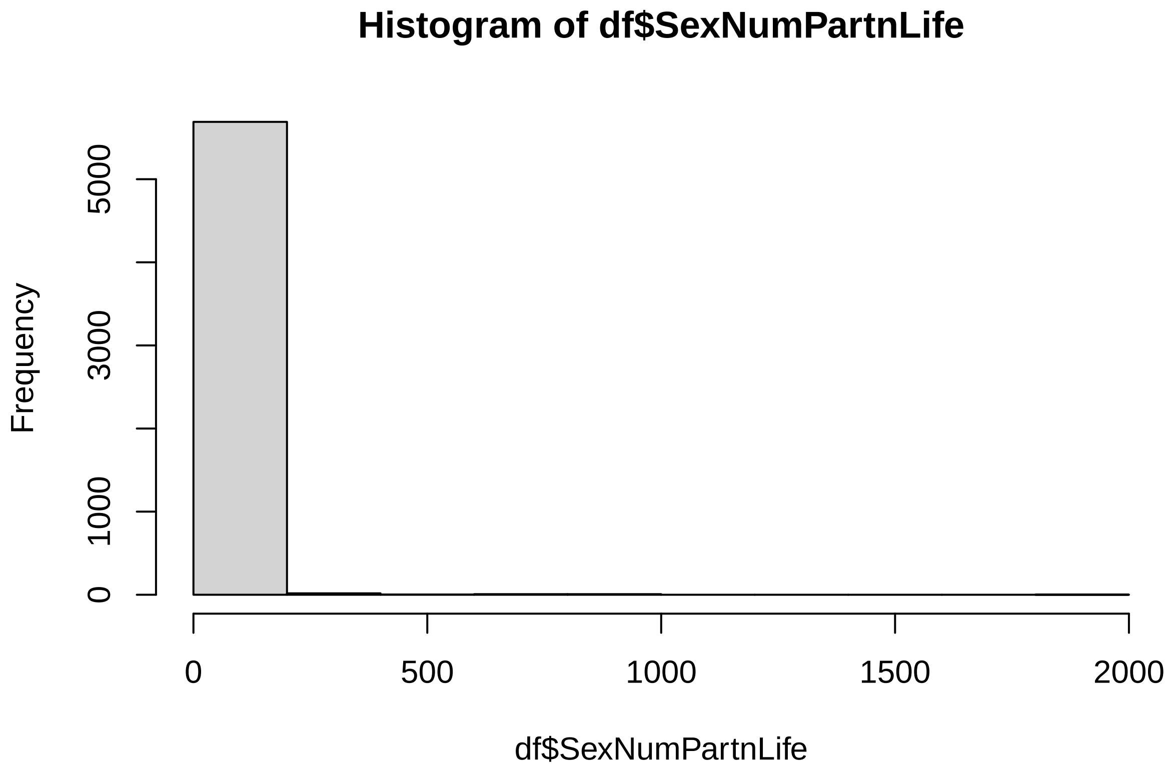
```
hist(df$SexNumPartYear[df$SexNumPartYear < 20])
```



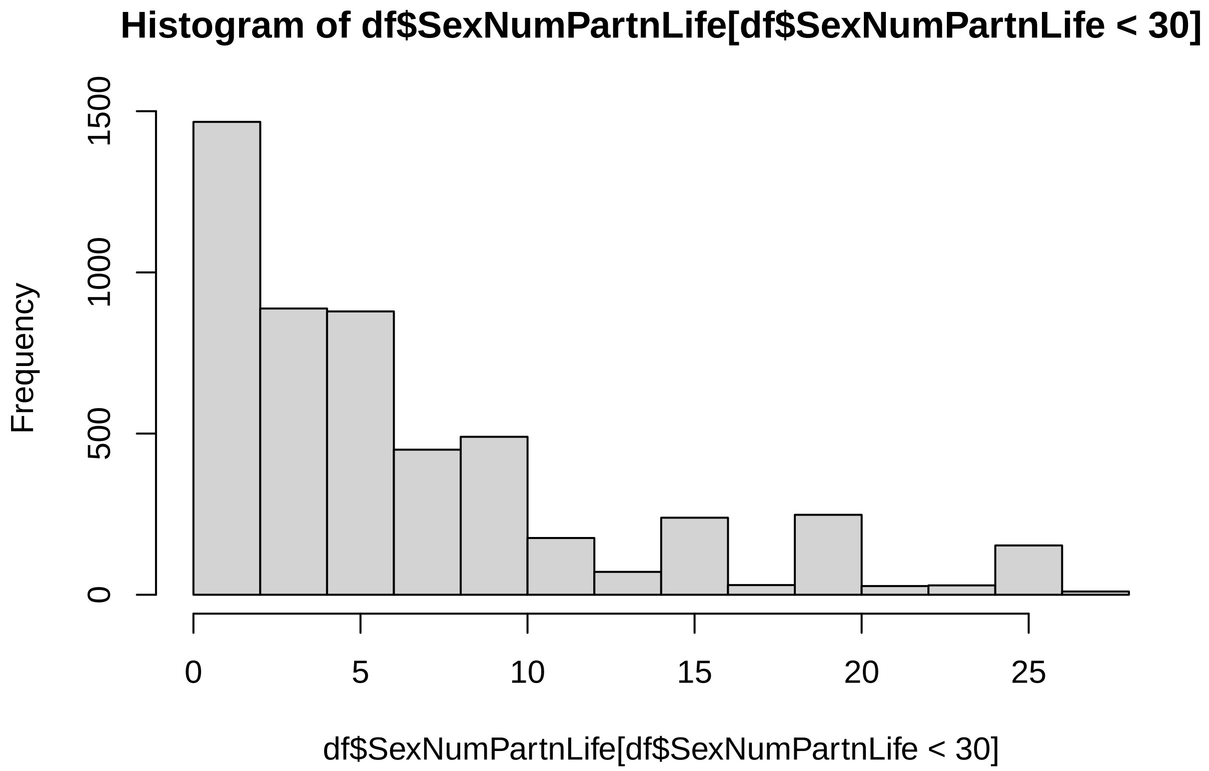
```
sort(unique(df$SexNumPartYear))
```

```
##  [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 17 18 19 20 30 50 69
```

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



```
hist(df$SexNumPartnLife[df$SexNumPartnLife < 30])
```



```
#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
```

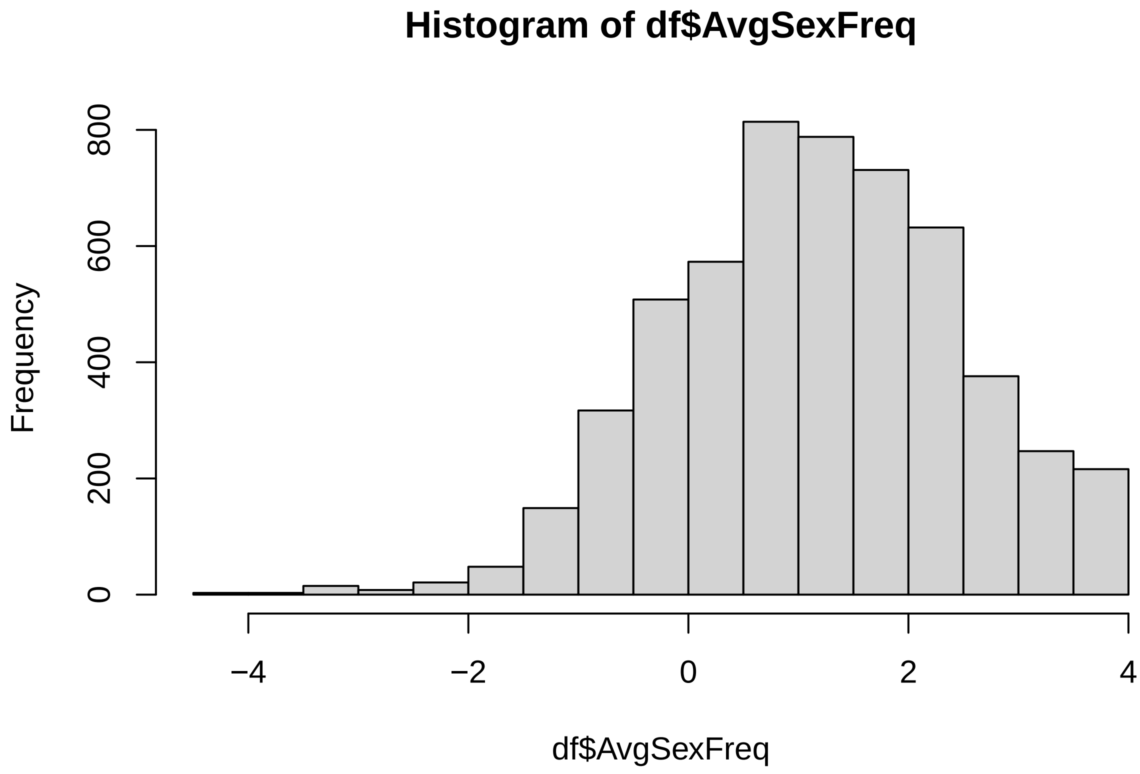
```
df[which(df$SexNumPartnLife > 300), c("Age", "SexAge", "SexNumPartnLife")]
```

```
## # 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
```

```
unique(df$SexAge)
```

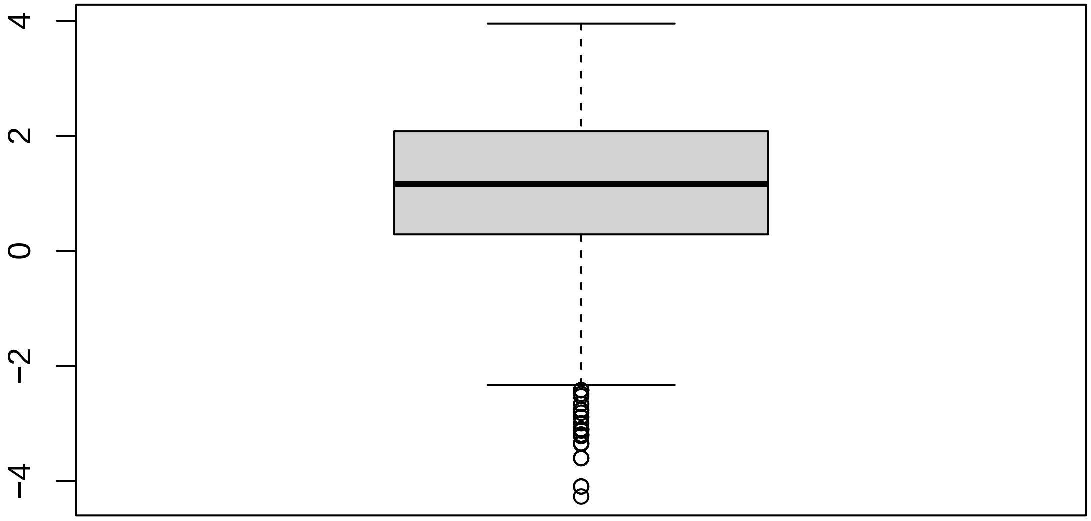
```
##  [1] 16 NA 12 13 17 22 27 20 18 14 23 15 21 24 28 30 19 32 29 26 37 33 35  9 38
## [26] 11 25 10 34 31 50 39 36 44 41 47
```

```r
df = mutate(df, AvgSexFreq = log((Age-SexAge)/SexNumPartnLife))
hist(df$AvgSexFreq)
```



```r
boxplot(df$AvgSexFreq)
```

```
## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
## z$out[z$group == : Outliers (-Inf, Inf) in boxplot 1 are not drawn
```



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

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

```
##
## Call:
## lm(formula = AvgSexFreq ~ Gender + HHIncome + Education + PhysActive +
##     SameSex + AlcoholYear + RegularMarij + HardDrugs + RegularMarij *
##     HardDrugs, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6281 -0.7327  0.0013  0.7379  3.3856
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.4312253  0.1707177   8.384  < 2e-16 ***
## Gendermale                     -0.3447323  0.0358114  -9.626  < 2e-16 ***
## HHIncome 5000-9999             -0.1849921  0.1848121  -1.001 0.316902
## HHIncome10000-14999            -0.0200325  0.1551540  -0.129 0.897274
## HHIncome15000-19999             0.1206346  0.1555663   0.775 0.438119
## HHIncome20000-24999            -0.0224108  0.1527872  -0.147 0.883392
## HHIncome25000-34999             0.1347785  0.1460209   0.923 0.356060
## HHIncome35000-44999           0.3210884  0.1461312   2.197 0.028061 *
## HHIncome45000-54999           0.2533943  0.1460033   1.736 0.082725 .
## HHIncome55000-64999           0.3779310  0.1488853   2.538 0.011175 *
## HHIncome65000-74999           0.5014736  0.1506628   3.328 0.000881 ***
## HHIncome75000-99999           0.3277854  0.1424487   2.301 0.021440 *
## HHIncomemore 99999            0.5833739  0.1398801   4.171 3.11e-05 ***
## Education9 - 11th Grade      -0.1721010  0.1155834  -1.489 0.136575
## EducationHigh School         -0.1078566  0.1108868  -0.973 0.330777
## EducationSome College        -0.1964943  0.1091879  -1.800 0.072002 .
## EducationCollege Grad        -0.0008423  0.1123243  -0.007 0.994017
## PhysActiveYes                -0.2863758  0.0370669  -7.726 1.41e-14 ***
## SameSexYes                   -0.2472752  0.0658888  -3.753 0.000177 ***
## AlcoholYear                  -0.0003505  0.0001879  -1.865 0.062193 .
## RegularMarijYes              -0.7544967  0.0504630 -14.951  < 2e-16 ***
## HardDrugsYes                 -0.5926214  0.0707266  -8.379  < 2e-16 ***
## RegularMarijYes:HardDrugsYes  0.6119568  0.0935046   6.545 6.75e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 3858 degrees of freedom
##   (6119 observations deleted due to missingness)
## Multiple R-squared:  0.189,  Adjusted R-squared:  0.1844
## F-statistic: 40.88 on 22 and 3858 DF,  p-value: < 2.2e-16
```

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

```
library(ggplot2)
library(tidyr)
#Add new column based on missingness
covariates = c("Gender","HHIncome","Education","PhysActive","SameSex","AlcoholYear","RegularMarij","Har
df$missingness <- ifelse(complete.cases(df[, covariates]), "Missing", "Not Missing")
```

```
library(gridExtra)
```

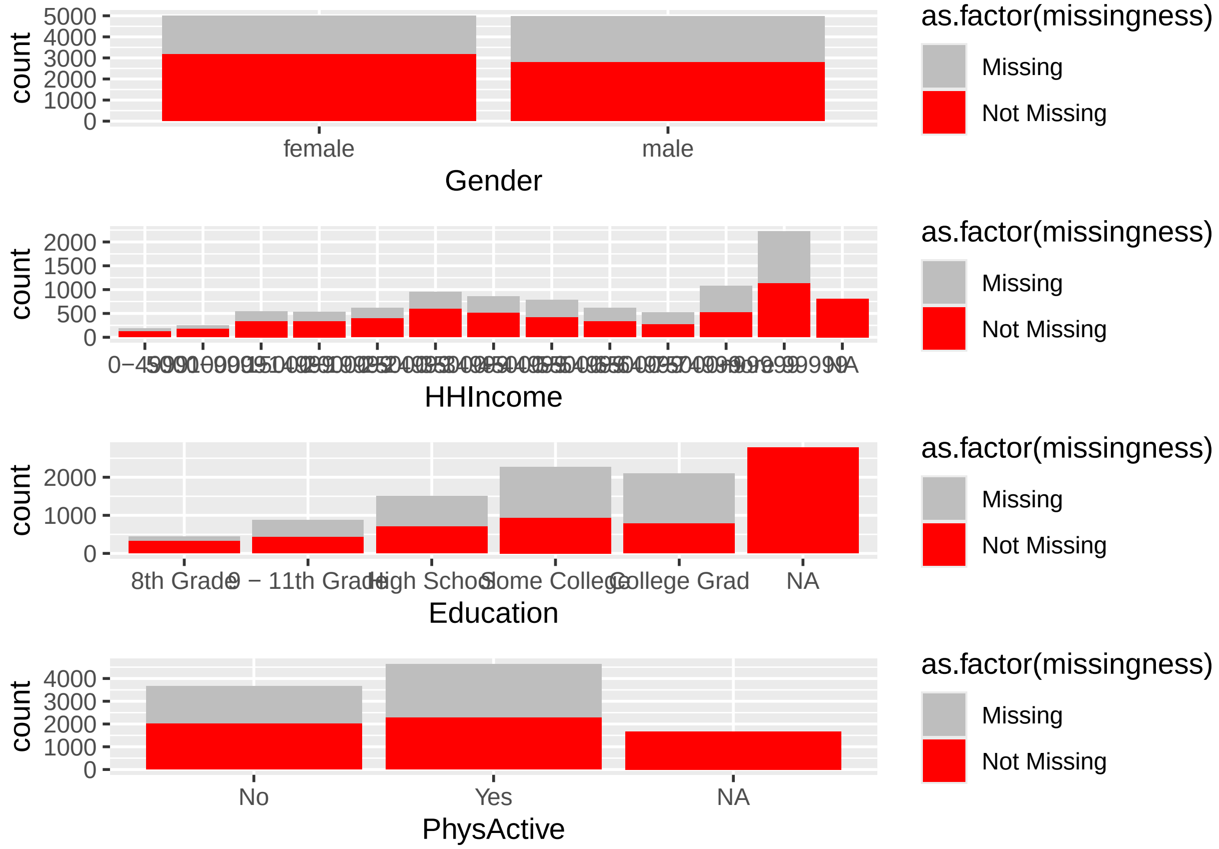
```
## Warning: package 'gridExtra' was built under R version 4.4.2
```

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

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

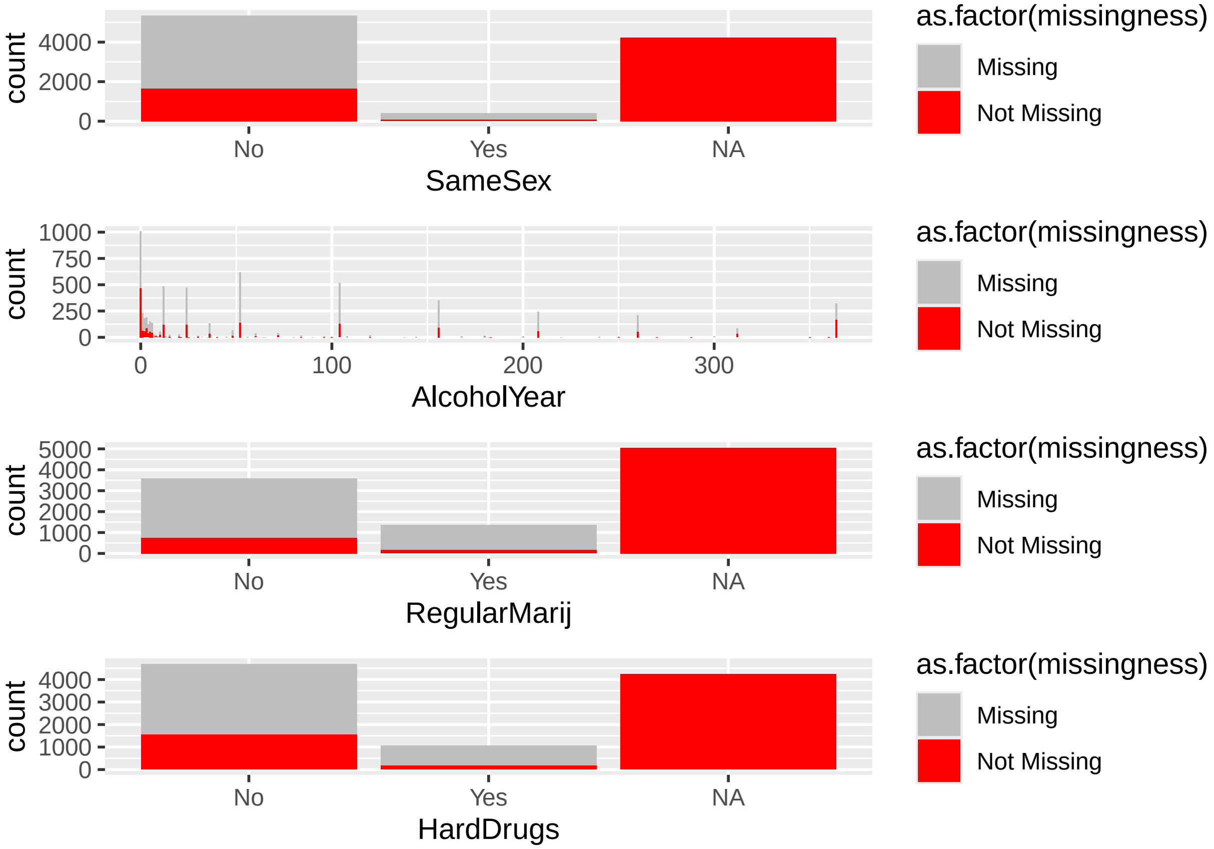
```
p1 = ggplot(data = df, mapping=aes(x=Gender, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p2 = ggplot(data = df, mapping=aes(x=HHIncome, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p3 = ggplot(data = df, mapping=aes(x=Education, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p4 = ggplot(data = df, mapping=aes(x=PhysActive, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
```

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



```r
p6 = ggplot(data = df, mapping=aes(x=SameSex, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p7 = ggplot(data = df, mapping=aes(x=AlcoholYear, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p8 = ggplot(data = df, mapping=aes(x=RegularMarij, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p9 = ggplot(data = df, mapping=aes(x=HardDrugs, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
grid.arrange(p6, p7, p8, p9, nrow = 4)
```

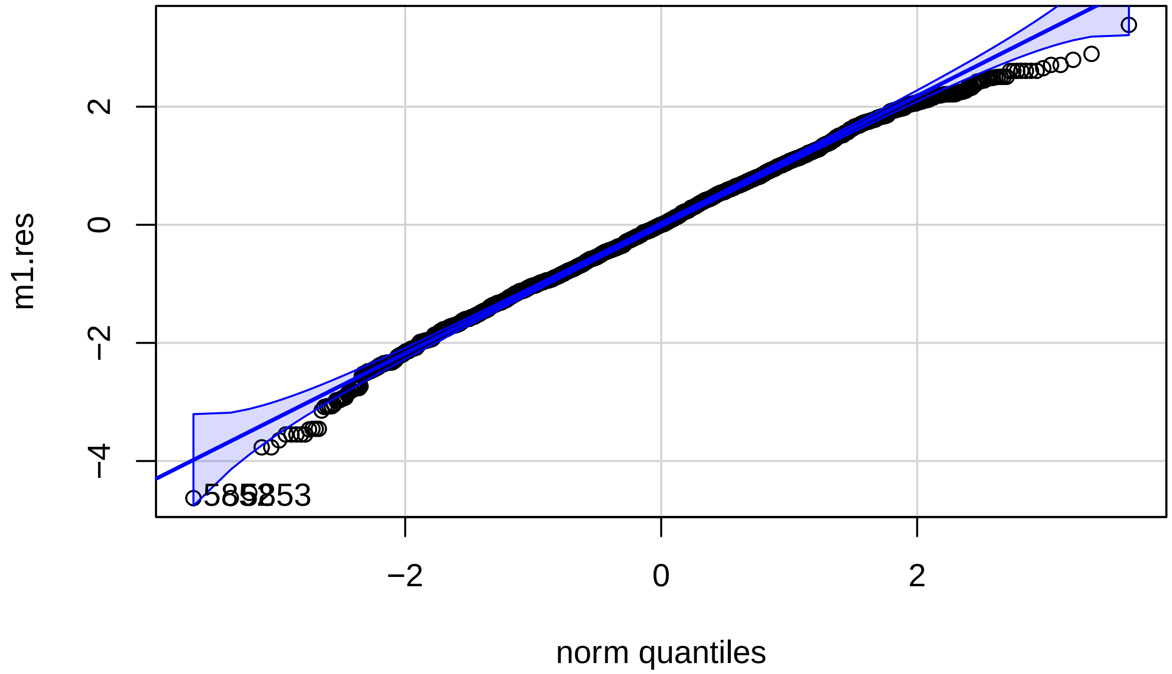
```
## Warning: Removed 4078 rows containing non-finite outside the scale range
## (`stat_count()`).
```



```
library(car)
car::Anova(lm(AvgSexFreq ~ Gender+HHIncome+Education+PhysActive+SameSex+AlcoholYear+RegularMarij+HardDru
```

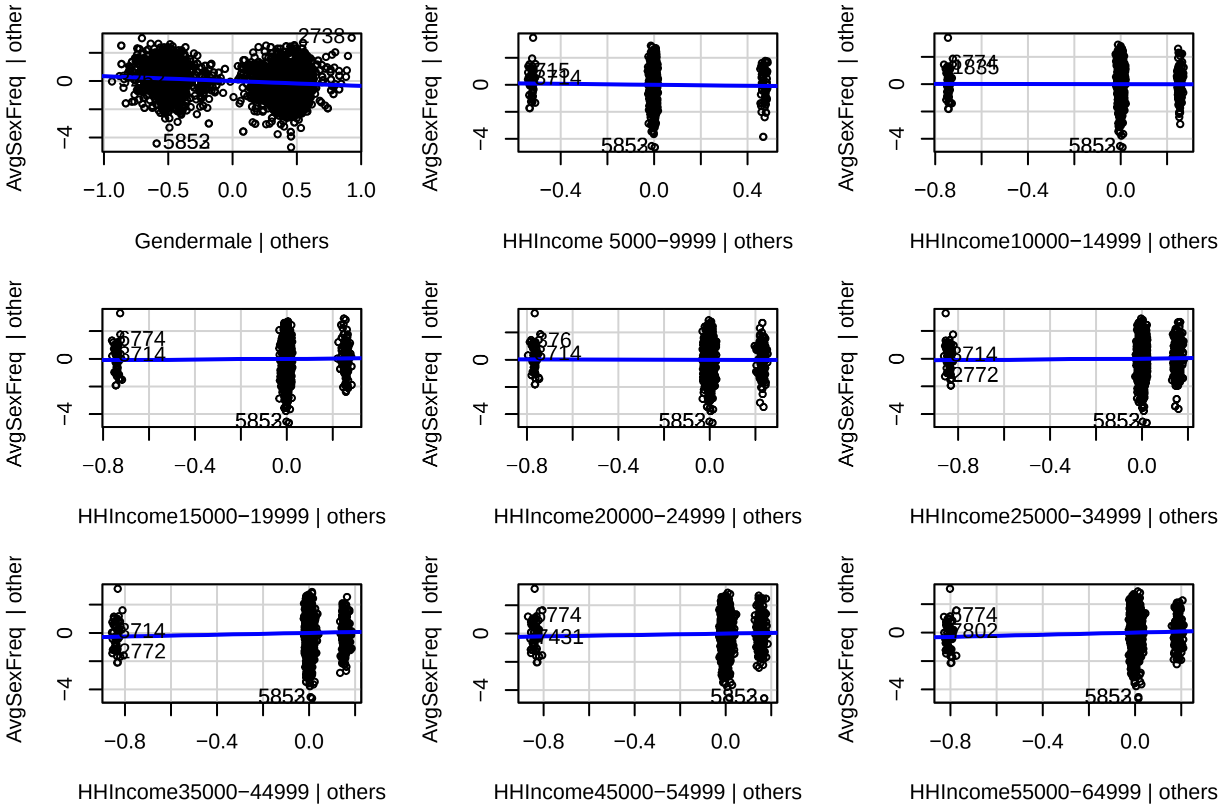
```
## Anova Table (Type III tests)
##
## Response: AvgSexFreq
##                        Sum Sq   Df  F value    Pr(>F)
## (Intercept)              80.4    1  70.2844 < 2.2e-16 ***
## Gender                  106.0    1  92.6660 < 2.2e-16 ***
## HHIncome                142.6   11  11.3271 < 2.2e-16 ***
## Education                23.9    4   5.2216 0.0003413 ***
## PhysActive               68.3    1  59.6899 1.406e-14 ***
## SameSex                  16.1    1  14.0844 0.0001774 ***
## AlcoholYear               4.0    1   3.4799 0.0621926 .
## RegularMarij            255.8    1 223.5470 < 2.2e-16 ***
## HardDrugs                80.3    1  70.2085 < 2.2e-16 ***
## RegularMarij:HardDrugs   49.0    1  42.8328 6.746e-11 ***
## Residuals              4414.3 3858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

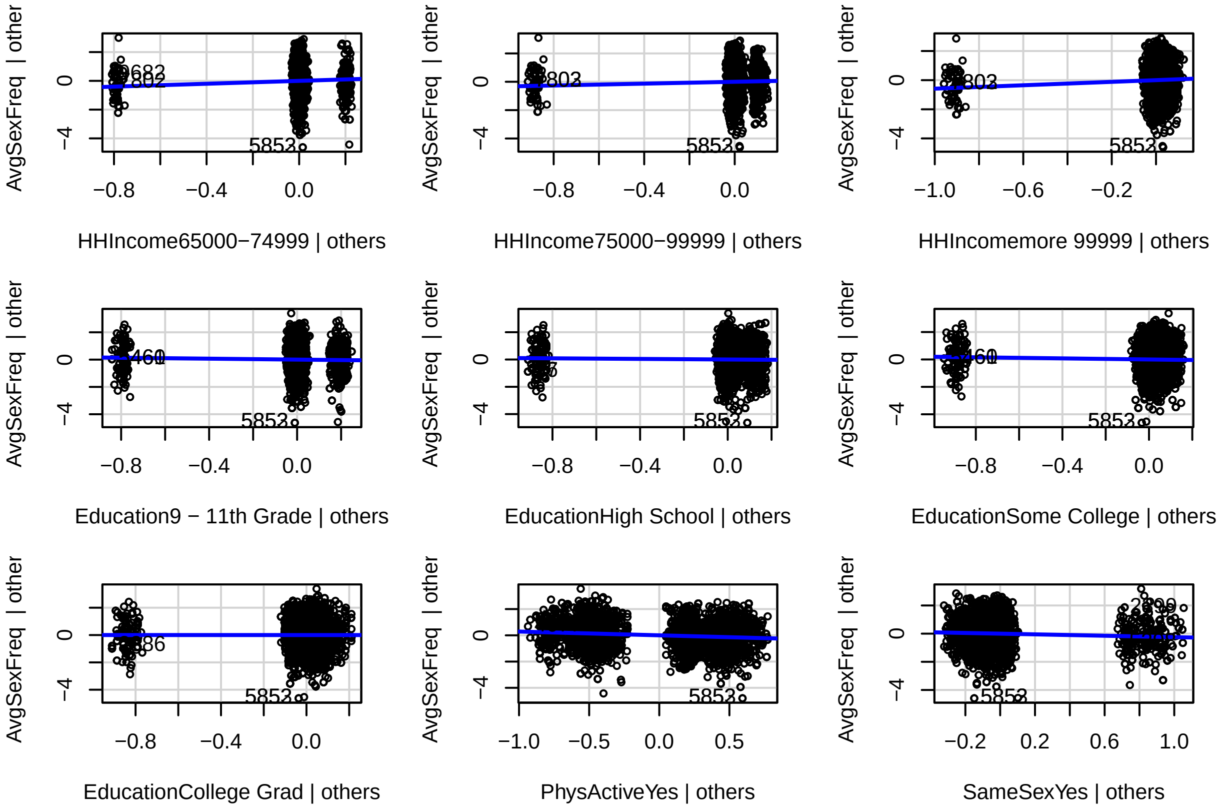
```
m1 = lm(AvgSexFreq ~ Gender+HHIncome+Education+PhysActive+SameSex+AlcoholYear+RegularMarij+HardDrugs+Re
m1.res = m1$residuals

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

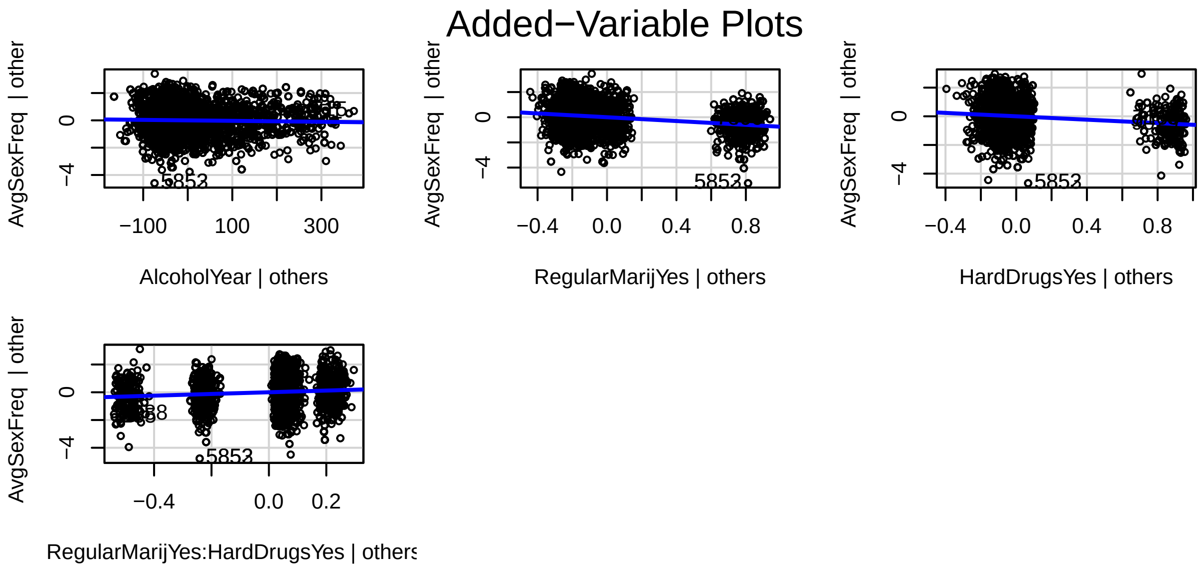


```
## 5852 5853
## 2288 2289
```

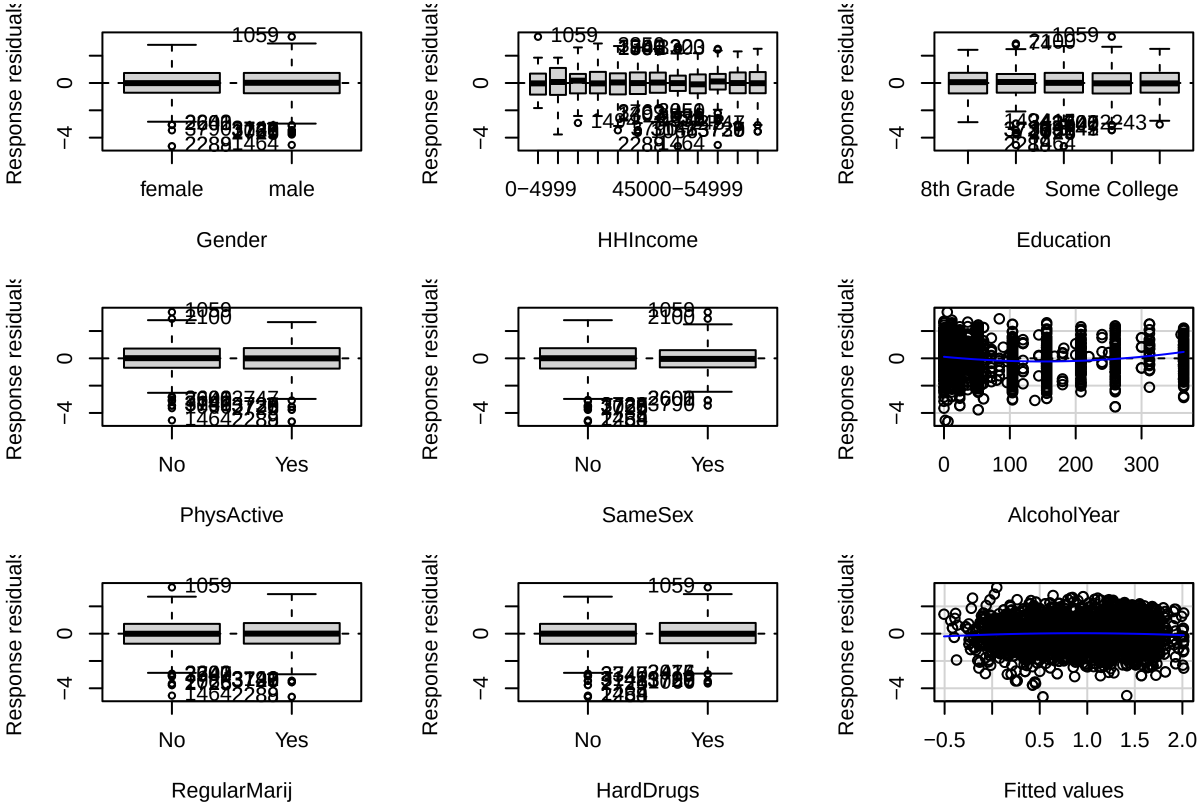
```
car::avPlots(m1)
```


AvgSexFreq | other
0
−4
2738
5852
−1.0
−0.5
0.0
0.5
1.0
Gendermale | others
AvgSexFreq | other
715
3714
5852
−0.4
0.0
0.4
HHIncome 5000−9999 | others
AvgSexFreq | other
6774
1835
5853
−0.8
−0.4
0.0
HHIncome10000−14999 | others
AvgSexFreq | other
6774
3714
5852
−0.8
−0.4
0.0
HHIncome15000−19999 | others
AvgSexFreq | other
876
3714
5852
−0.8
−0.4
0.0
HHIncome20000−24999 | others
AvgSexFreq | other
3714
2772
5852
−0.8
−0.4
0.0
HHIncome25000−34999 | others
AvgSexFreq | other
3714
2772
5852
−0.8
−0.4
0.0
HHIncome35000−44999 | others
AvgSexFreq | other
6774
7431
5852
−0.8
−0.4
0.0
HHIncome45000−54999 | others
AvgSexFreq | other
6774
7802
5853
−0.8
−0.4
0.0
HHIncome55000−64999 | others


AvgSexFreq | other
0
−4
−0.8
−0.4
0.0
9682
802
5852
HHIncome65000−74999 | others
AvgSexFreq | other
0
−4
−0.8
−0.4
0.0
803
5852
HHIncome75000−99999 | others
AvgSexFreq | other
0
−4
−1.0
−0.6
−0.2
803
5852
HHIncomemore 99999 | others
AvgSexFreq | other
0
−4
−0.8
−0.4
0.0
461
5852
Education9 − 11th Grade | others
AvgSexFreq | other
0
−4
−0.8
−0.4
0.0
7
5852
EducationHigh School | others
AvgSexFreq | other
0
−4
−0.8
−0.4
0.0
461
5852
EducationSome College | others
AvgSexFreq | other
0
−4
−0.8
−0.4
0.0
3086
5852
EducationCollege Grad | others
AvgSexFreq | other
0
−4
−1.0
−0.5
0.0
0.5
5852
PhysActiveYes | others
AvgSexFreq | other
0
−4
−0.2
0.2
0.6
1.0
5852
SameSexYes | others



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



```
##                Test stat Pr(>|Test stat|)
## Gender
## HHIncome
## Education
## PhysActive
## SameSex
## AlcoholYear      9.1617         < 2e-16 ***
## RegularMarij
## HardDrugs
## Tukey test      -2.0236         0.04301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

```
##                  GVIF Df GVIF^(1/(2*Df)) Interacts With
## Gender       1.079422  1        1.038952           --
## HHIncome     1.340755 11        1.013418           --
## Education    1.429918  4        1.045716           --
## PhysActive   1.133017  1        1.064433           --
## SameSex      1.103131  1        1.050300           --
## AlcoholYear  1.109762  1        1.053452           --
## RegularMarij 1.316648  1        1.147453           --
## HardDrugs    1.350139  1        1.161955          --
##                                                                               Other Predictors
## Gender       HHIncome, Education, PhysActive, SameSex, AlcoholYear, RegularMarij, HardDrugs
## HHIncome       Gender, Education, PhysActive, SameSex, AlcoholYear, RegularMarij, HardDrugs
## Education       Gender, HHIncome, PhysActive, SameSex, AlcoholYear, RegularMarij, HardDrugs
## PhysActive       Gender, HHIncome, Education, SameSex, AlcoholYear, RegularMarij, HardDrugs
## SameSex      Gender, HHIncome, Education, PhysActive, AlcoholYear, RegularMarij, HardDrugs
## AlcoholYear       Gender, HHIncome, Education, PhysActive, SameSex, RegularMarij, HardDrugs
## RegularMarij       Gender, HHIncome, Education, PhysActive, SameSex, AlcoholYear, HardDrugs
## HardDrugs       Gender, HHIncome, Education, PhysActive, SameSex, AlcoholYear, RegularMarij
```

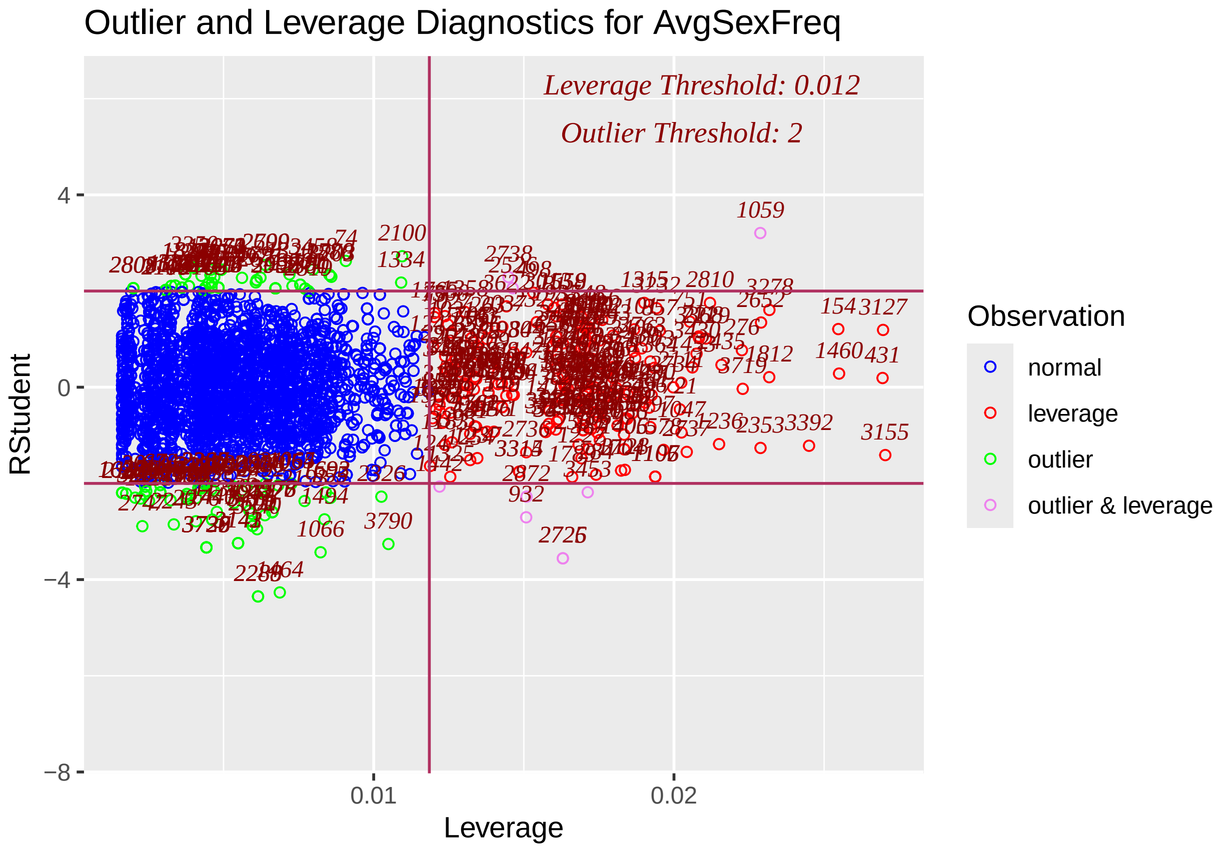
```
model.deffits=dffits(m1)
model.CD = cooks.distance(m1)
model.deffits[which.max(model.deffits)]
```

```
##      2738
## 0.4905041
```

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

```
##       2738
## 0.01043553
```

```
ols_plot_resid_lev(m1)
```



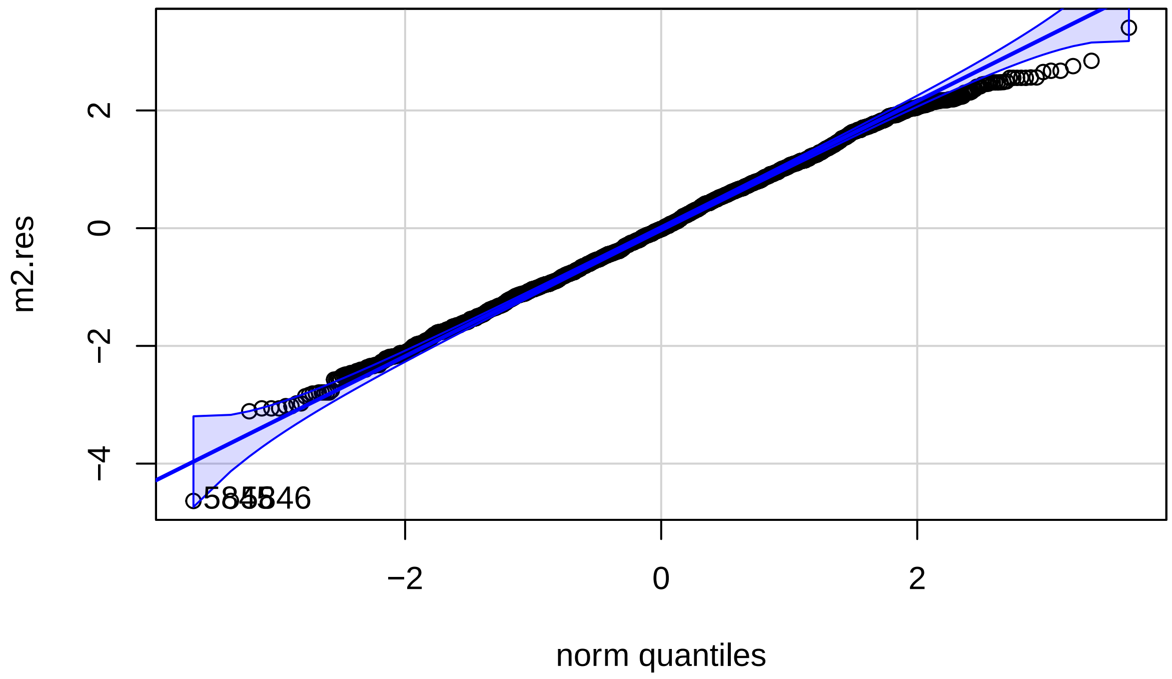
```
df2 = df[-which(df$SexNumPartnLife > 300),]
m2 = lm(AvgSexFreq ~ Gender+HHIncome+Education+PhysActive+SameSex+AlcoholYear+RegularMarij+HardDrugs+Reg

summary(m2)
```

```
## 
## Call:
## lm(formula = AvgSexFreq ~ Gender + HHIncome + Education + PhysActive + 
##     SameSex + AlcoholYear + RegularMarij + HardDrugs + RegularMarij * 
##     HardDrugs, data = df2)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.6339 -0.7391 -0.0148  0.7223  3.4059 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  1.3898750  0.1669358   8.326  < 2e-16 ***
## Gendermale                  -0.3278360  0.0350528  -9.353  < 2e-16 ***
## HHIncome 5000-9999          -0.0830288  0.1818996  -0.456 0.648089    
## HHIncome10000-14999         -0.0074734  0.1517960  -0.049 0.960736    
## HHIncome15000-19999          0.1157245  0.1520939   0.761 0.446778    
## HHIncome20000-24999          0.0088823  0.1495366   0.059 0.952638    
## HHIncome25000-34999          0.1761905  0.1428793   1.233 0.217598    
## HHIncome35000-44999          0.3242550  0.1428706   2.270 0.023289 *  
## HHIncome45000-54999          0.2580116  0.1427465   1.807 0.070765 .  
## HHIncome55000-64999          0.3841671  0.1455639   2.639 0.008345 ** 
## HHIncome65000-74999          0.5244299  0.1473705   3.559 0.000377 ***
## HHIncome75000-99999          0.3480667  0.1393122   2.498 0.012515 *  
## HHIncomemore 99999           0.6116639  0.1367873   4.472 7.99e-06 ***
## Education9 - 11th Grade     -0.1190802  0.1131697  -1.052 0.292762    
## EducationHigh School        -0.0816215  0.1084438  -0.753 0.451699    
## EducationSome College       -0.1877321  0.1068029  -1.758 0.078871 .  
## EducationCollege Grad       -0.0020736  0.1098626  -0.019 0.984942    
## PhysActiveYes               -0.2766208  0.0363267  -7.615 3.30e-14 ***
## SameSexYes                  -0.2550131  0.0645098  -3.953 7.85e-05 ***
## AlcoholYear                 -0.0003276  0.0001840  -1.781 0.075060 .  
## RegularMarijYes             -0.7481164  0.0494597 -15.126  < 2e-16 ***
## HardDrugsYes                -0.5895750  0.0692630  -8.512  < 2e-16 ***
## RegularMarijYes:HardDrugsYes  0.6357582  0.0916886   6.934 4.78e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.046 on 3841 degrees of freedom
##   (6111 observations deleted due to missingness)
## Multiple R-squared:  0.1863, Adjusted R-squared:  0.1816 
## F-statistic: 39.96 on 22 and 3841 DF,  p-value: < 2.2e-16
```

```
#detect beta change by more than 10 percent?
100*(coef(m1)-coef(m2))/coef(m1)
```

```
##                  (Intercept)                   Gendermale 
##                    2.8891543                    4.9012724 
##           HHIncome 5000-9999          HHIncome10000-14999 
##                   55.1176451                   62.6934739 
##          HHIncome15000-19999          HHIncome20000-24999 
##                    4.0702250                  139.6338248 
##          HHIncome25000-34999          HHIncome35000-44999 
##                  -30.7259720                   -0.9862154
##          HHIncome45000-54999          HHIncome55000-64999
##                   -1.8221936                   -1.6500616
##          HHIncome65000-74999          HHIncome75000-99999
##                   -4.5777582                   -6.1873887
##           HHIncomemore 99999      Education9 - 11th Grade
##                   -4.8493849                   30.8079538
##         EducationHigh School        EducationSome College
##                   24.3240246                    4.4592503
##        EducationCollege Grad                PhysActiveYes
##                 -146.1844900                    3.4063594
##                   SameSexYes                  AlcoholYear
##                   -3.1292504                    6.5439861
##              RegularMarijYes                 HardDrugsYes
##                    0.8456350                    0.5140547
## RegularMarijYes:HardDrugsYes
##                   -3.8893858
```

```
m2.res = m2$residuals

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



```
## 5845 5846
## 2285 2286
```