BIOSTAT 650 Project

Jaehoon Kim

2024-11-17

```
df = NHANES
```

```
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')
```

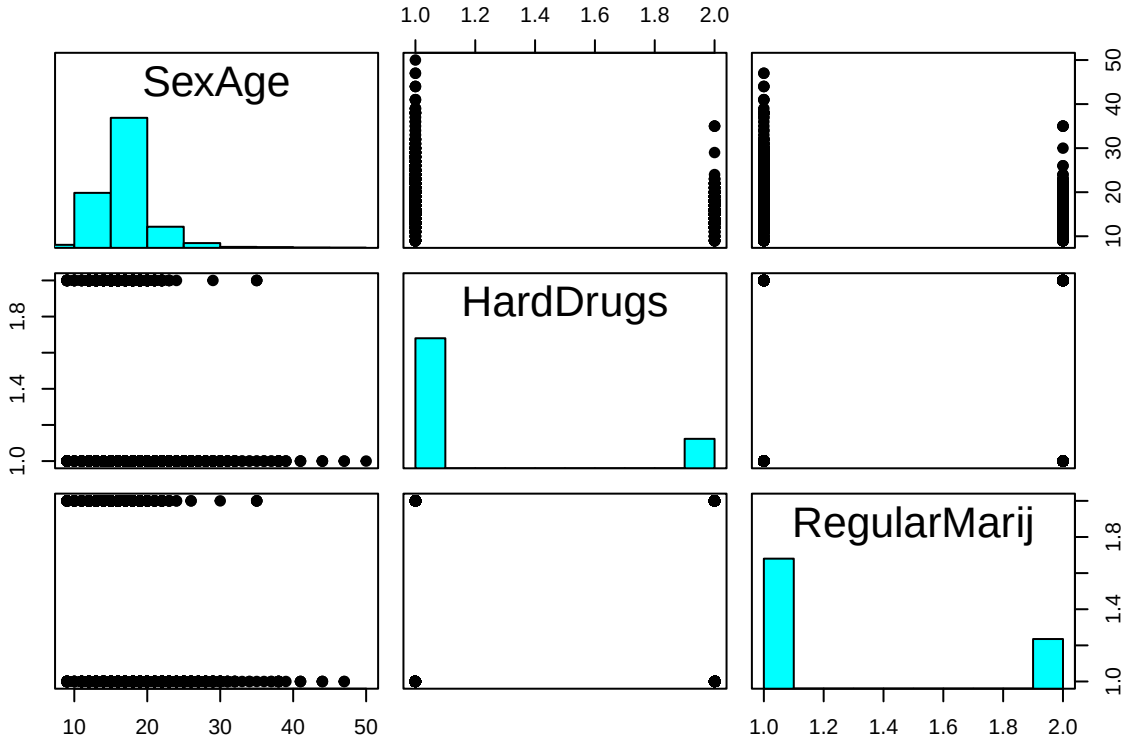
```
df = NHANES
#df = NHANES["DiabetesAge" > 20]
colnames(df)
```

```
##  [1] "ID"               "SurveyYr"         "Gender"           "Age"
##  [5] "AgeDecade"        "AgeMonths"        "Race1"            "Race3"
##  [9] "Education"        "MaritalStatus"    "HHIncome"         "HHIncomeMid"
## [13] "Poverty"          "HomeRooms"        "HomeOwn"          "Work"
## [17] "Weight"           "Length"           "HeadCirc"         "Height"
## [21] "BMI"              "BMICatUnder20yrs" "BMI_WHO"          "Pulse"
## [25] "BPSysAve"         "BPDiaAve"         "BPSys1"           "BPDia1"
## [29] "BPSys2"           "BPDia2"           "BPSys3"           "BPDia3"
## [33] "Testosterone"     "DirectChol"       "TotChol"          "UrineVol1"
## [37] "UrineFlow1"       "UrineVol2"        "UrineFlow2"       "Diabetes"
## [41] "DiabetesAge"      "HealthGen"        "DaysPhysHlthBad"  "DaysMentHlthBad"
## [45] "LittleInterest"   "Depressed"        "nPregnancies"     "nBabies"
## [49] "Age1stBaby"       "SleepHrsNight"    "SleepTrouble"     "PhysActive"
## [53] "PhysActiveDays"   "TVHrsDay"         "CompHrsDay"       "TVHrsDayChild"
## [57] "CompHrsDayChild"  "Alcohol12PlusYr"  "AlcoholDay"       "AlcoholYear"
## [61] "SmokeNow"         "Smoke100"         "Smoke100n"        "SmokeAge"
## [65] "Marijuana"        "AgeFirstMarij"    "RegularMarij"     "AgeRegMarij"
## [69] "HardDrugs"        "SexEver"          "SexAge"           "SexNumPartnLife"
## [73] "SexNumPartYear"   "SameSex"          "SexOrientation"   "PregnantNow"
```

```
scatmatrixData = df[,c("SexAge", "HardDrugs", "RegularMarij")]
panel.hist <- function(x, ...)
{
usr <- par("usr"); on.exit(par(usr))
par(usr = c(usr[1:2], 0, 1.5) )
h <- hist(x, plot = FALSE)
breaks <- h$breaks; nB <- length(breaks)
y <- h$counts; y <- y/max(y)
```

```
rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
pairs(scatmatrixData, pch = 19, diag.panel=panel.hist)
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter
## Warning in par(usr): argument 1 does not name a graphical parameter
## Warning in par(usr): argument 1 does not name a graphical parameter
```




```
model <- lm(DiabetesAge ~ Gender+Poverty+BMI+BPSys1+SleepHrsNight+PhysActiveDays, df)
summary(model)
```

```
##
## Call:
## lm(formula = DiabetesAge ~ Gender + Poverty + BMI + BPSys1 +
##     SleepHrsNight + PhysActiveDays, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.087  -7.907   2.062   8.861  29.318
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    32.96048   10.92836   3.016  0.00287 **
## Gendermale     -2.46465    2.11661  -1.164  0.24553
## Poverty         0.46344    0.62309   0.744  0.45781
## BMI            -0.09236    0.14055  -0.657  0.51180
```

```
## BPSys1          0.13469    0.05758   2.339  0.02024 *
## SleepHrsNight   0.25571    0.73547   0.348  0.72841
## PhysActiveDays -0.19888    0.53308  -0.373  0.70945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 217 degrees of freedom
##   (9776 observations deleted due to missingness)
## Multiple R-squared:  0.04008,    Adjusted R-squared:  0.01354
## F-statistic:  1.51 on 6 and 217 DF,  p-value: 0.176
```

```
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
```

```
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
```

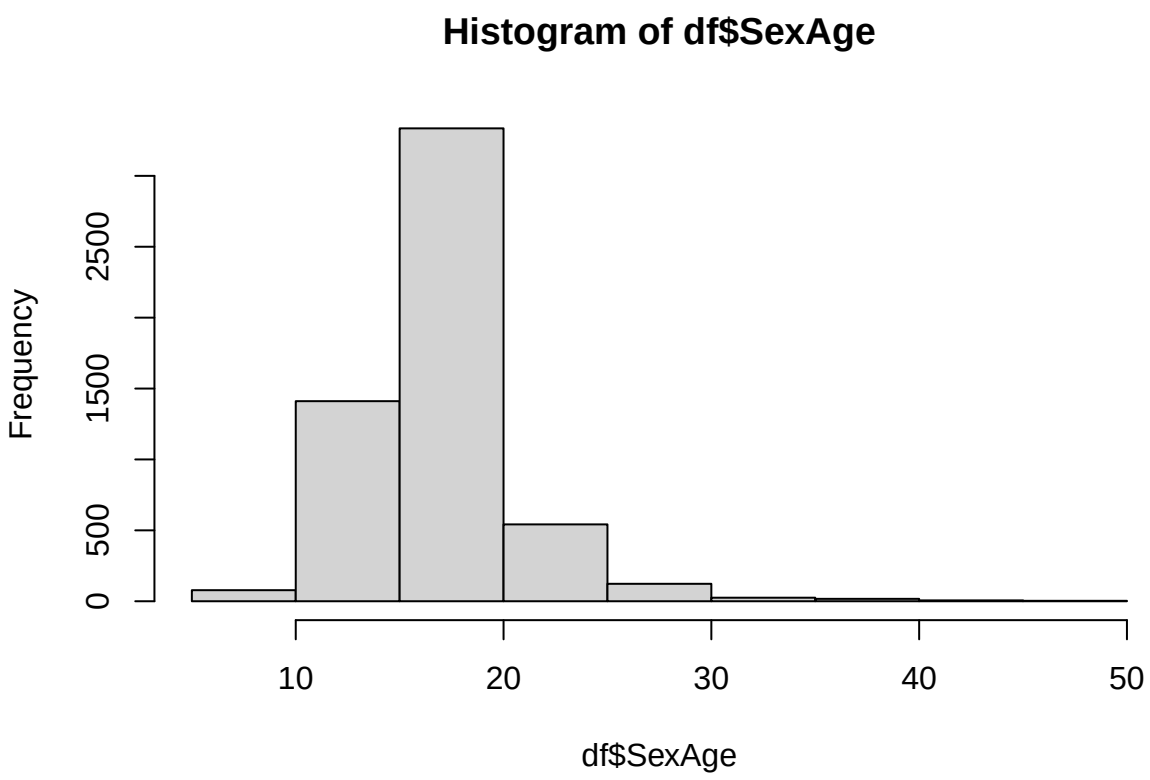
```
model <- lm(SexNumPartnLife ~ Gender+HHIncome+Education+PhysActive+SameSex+RegularMarij+HardDrugs+Regula
summary(model)
```

```
##
## 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
```

Created new variable and log transformed due to extreme skewness

```
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
```

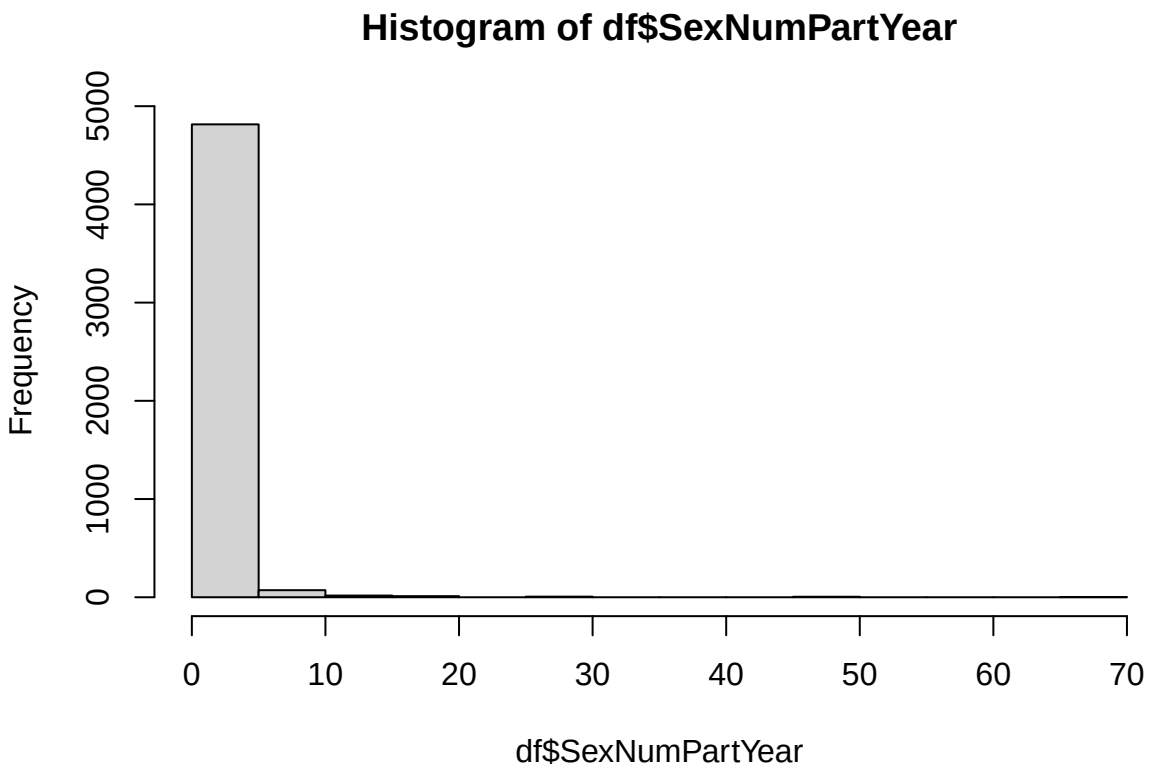
```
typeof(df$SexAge)
```

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

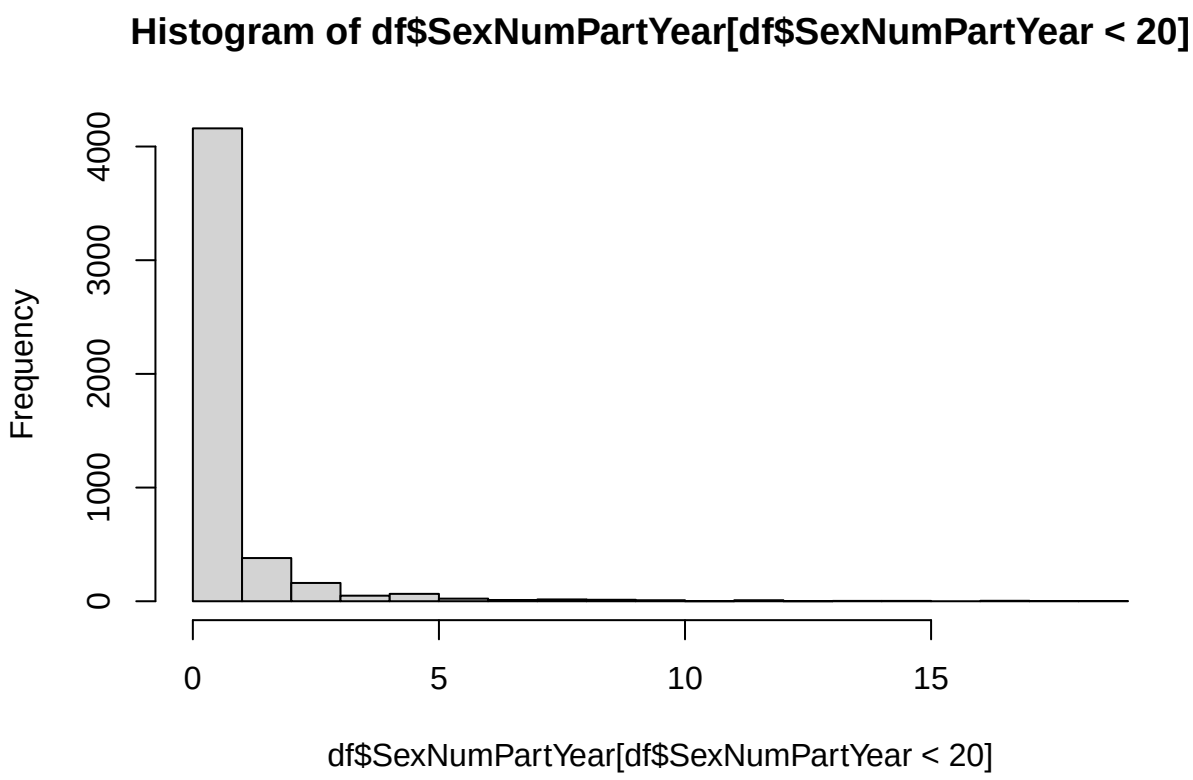
```
subset(df, SexAge == 9 & !is.na(SexAge))$SexNumPartnLife
```

```
##  [1]  30  30  90  90  55  55 120   5   5   5   5  19   3   3   3   5   5   9  88
## [20]  98  27  27  25  30 150 150 150  NA   2  11  85 500 200 200   5   1  23   2
## [39]   8  19  20  20  20   3 100  50  40  40   6 360 150  20  80   3   3   3   5
## [58]  50   7
```

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



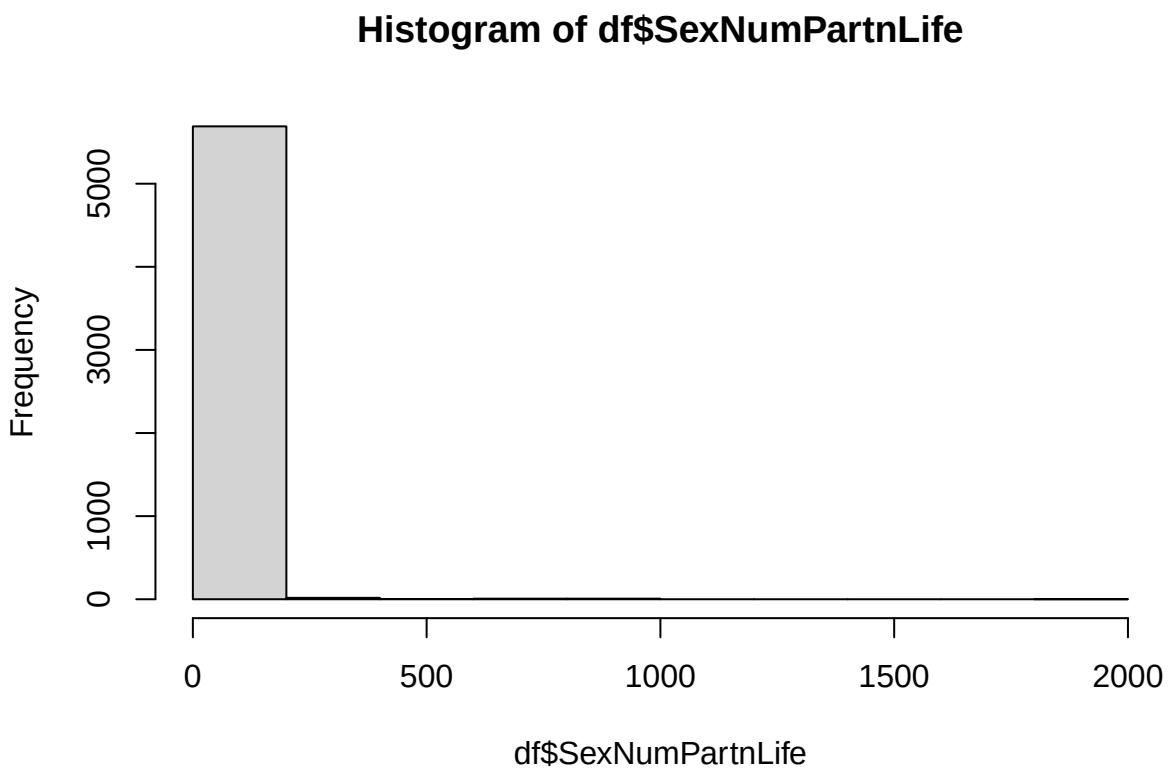

```r
hist(df$SexNumPartYear[df$SexNumPartYear < 20])
```



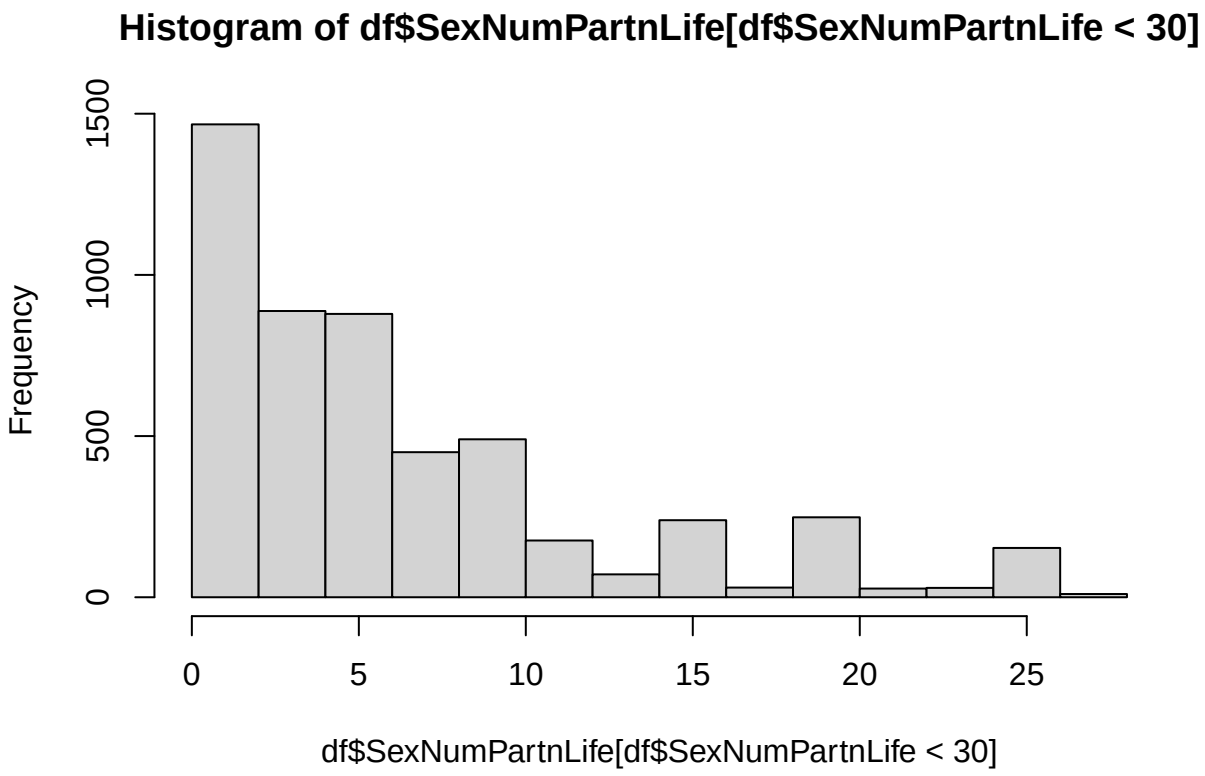

```
sort(unique(df$SexNumPartYear))
```

```
##  [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)
```



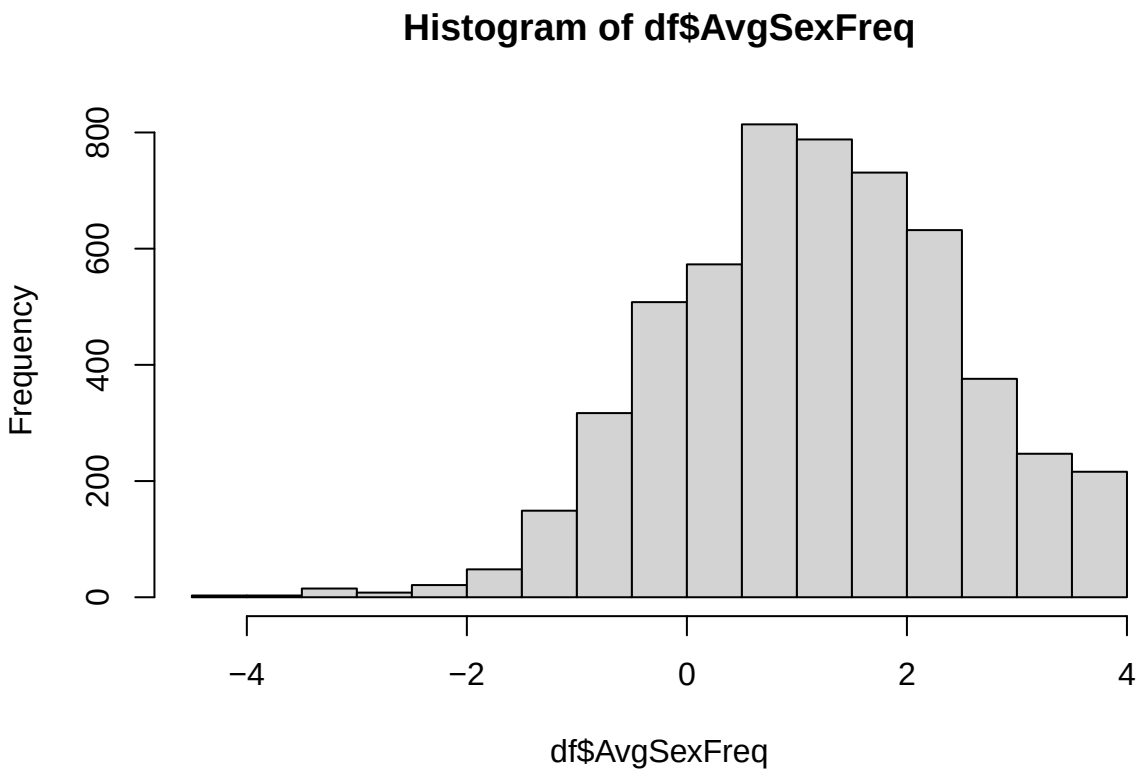

```
hist(df$SexNumPartnLife[df$SexNumPartnLife < 30])
```




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

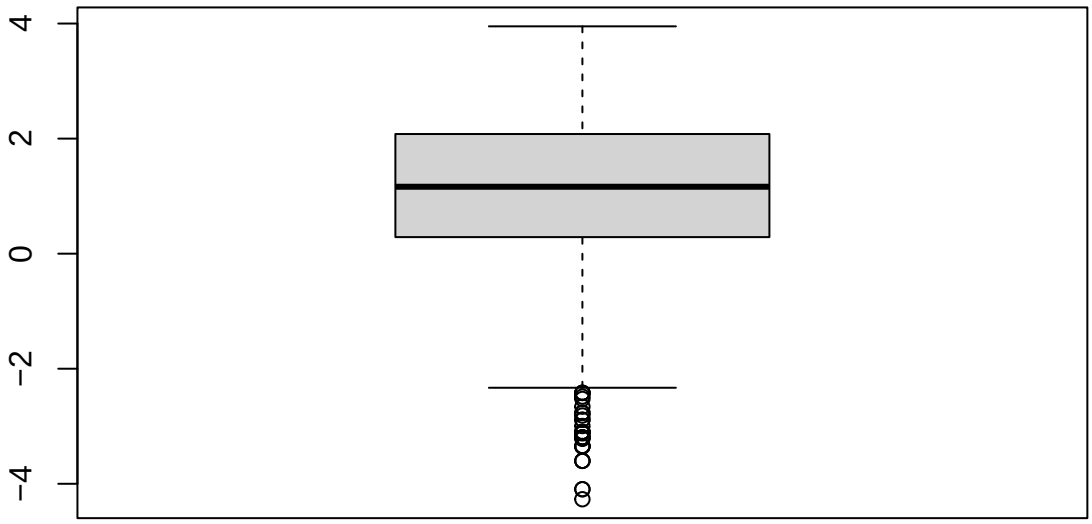
```
##  [1] 16 NA 12 13 17 22 27 20 18 14 23 15 21 24 28 30 19 32 29 26 37 33 35  9 38
## [26] 11 25 10 34 31 50 39 36 44 41 47
```

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




```
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
```



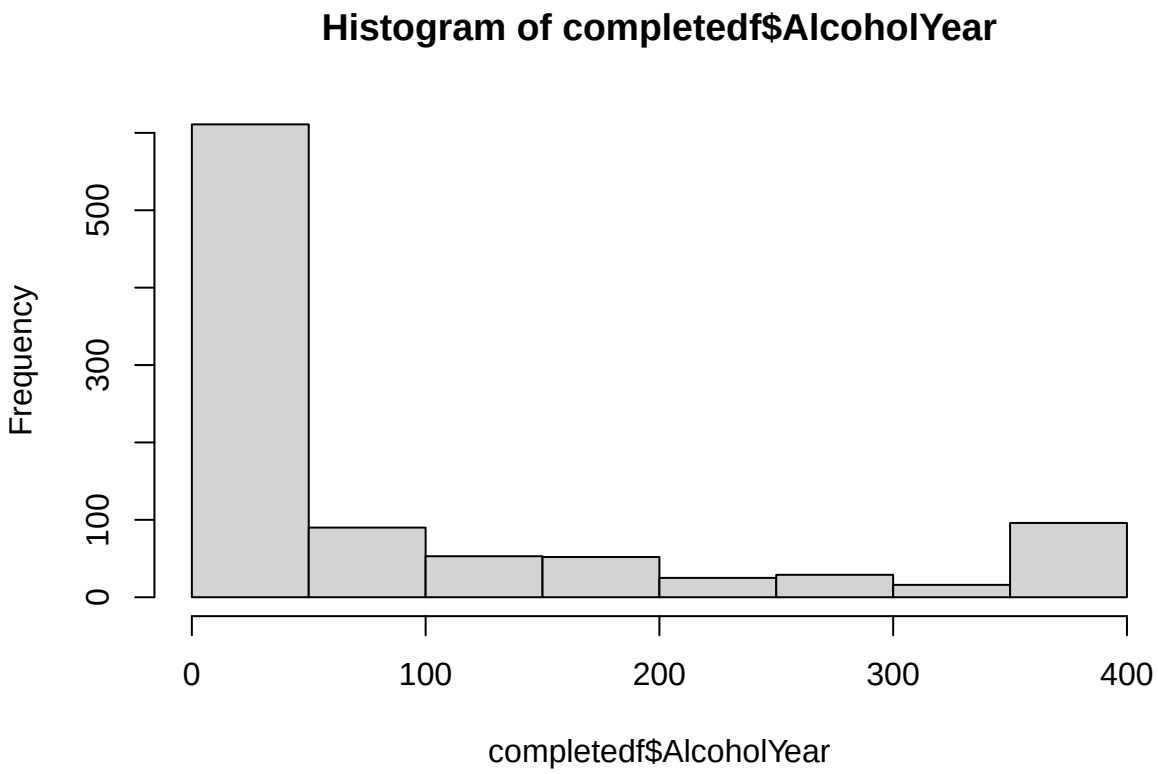

```r
df$AvgSexFreq[is.infinite(df$AvgSexFreq)] = NA
#unique(df$AvgSexFreq)

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 *
```

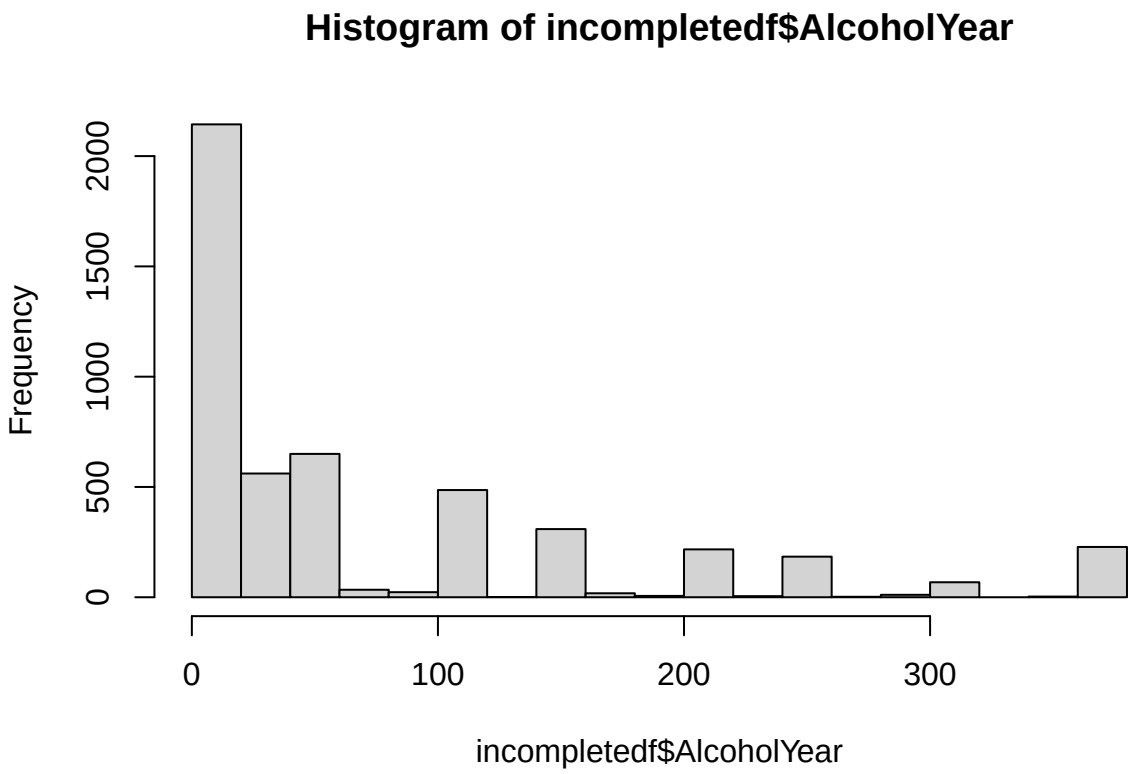
```
## HHIncome45000-54999           0.2533943  0.1460033   1.736 0.082725 .
## HHIncome55000-64999           0.3779310  0.1488853   2.538 0.011175 *
## HHIncome65000-74999           0.5014736  0.1506628   3.328 0.000881 ***
## HHIncome75000-99999           0.3277854  0.1424487   2.301 0.021440 *
## 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)
```

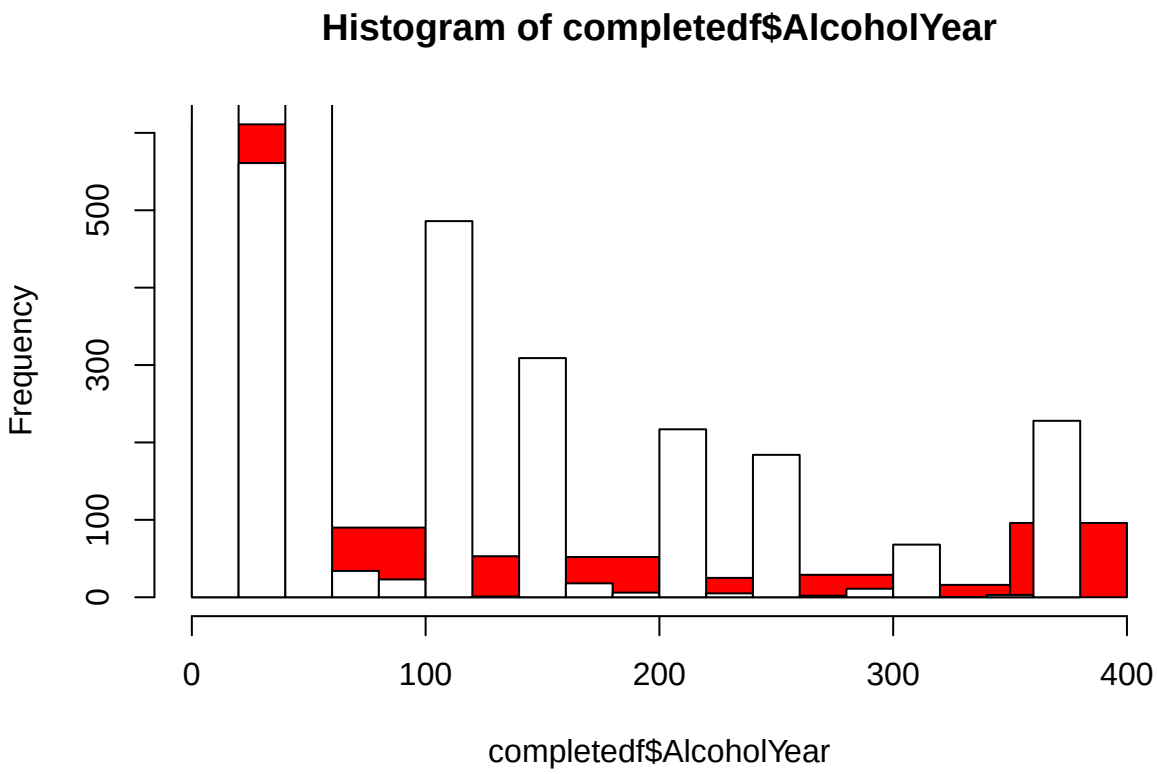
```
library(ggplot2)
library(tidyr)
completedf = df[is.na(df$AvgSexFreq),]
incompletedf = df[!is.na(df$AvgSexFreq),]
#Add new column based on missingness
df$missingness <- ifelse(is.na(df$AvgSexFreq), "Missing", "Not Missing")
covariates = c("Gender","HHIncome","Education","PhysActive","SameSex","AlcoholYear","RegularMarij","Har

A = hist(completedf$AlcoholYear)
```



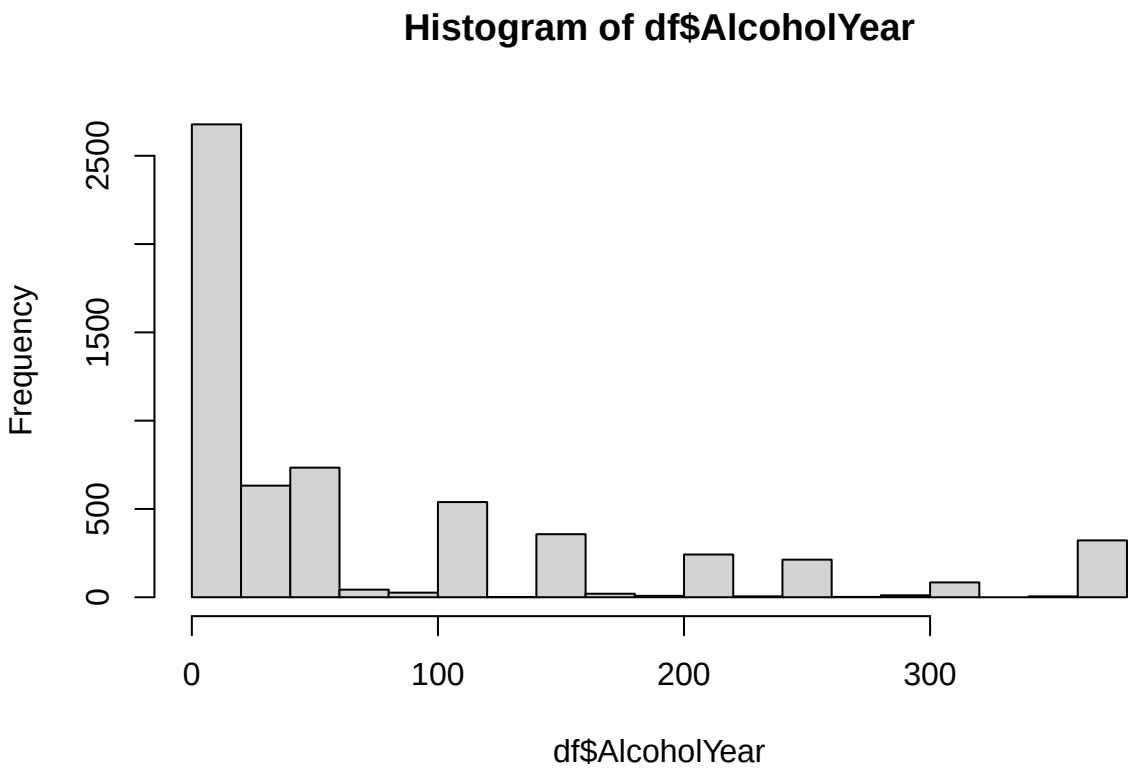

```
B = hist(incompletedf$AlcoholYear)
```




```
plot(A, col = "red")
plot(B, col = "white", add = TRUE)
```




```
hist(df$AlcoholYear)
```




```
library(gridExtra)
```

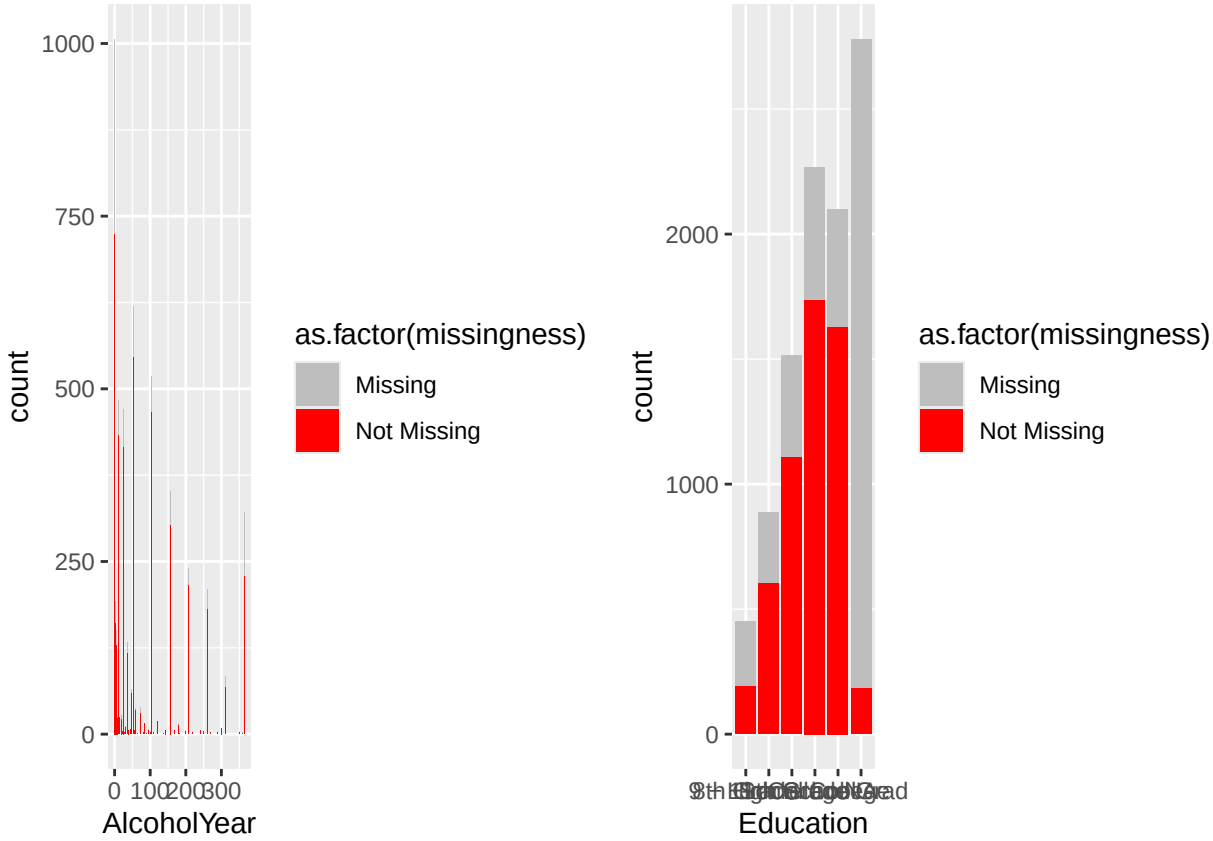
```
## Warning: package 'gridExtra' was built under R version 4.4.2
```

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

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

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

grid.arrange(p1,p2,nrow=1)
```

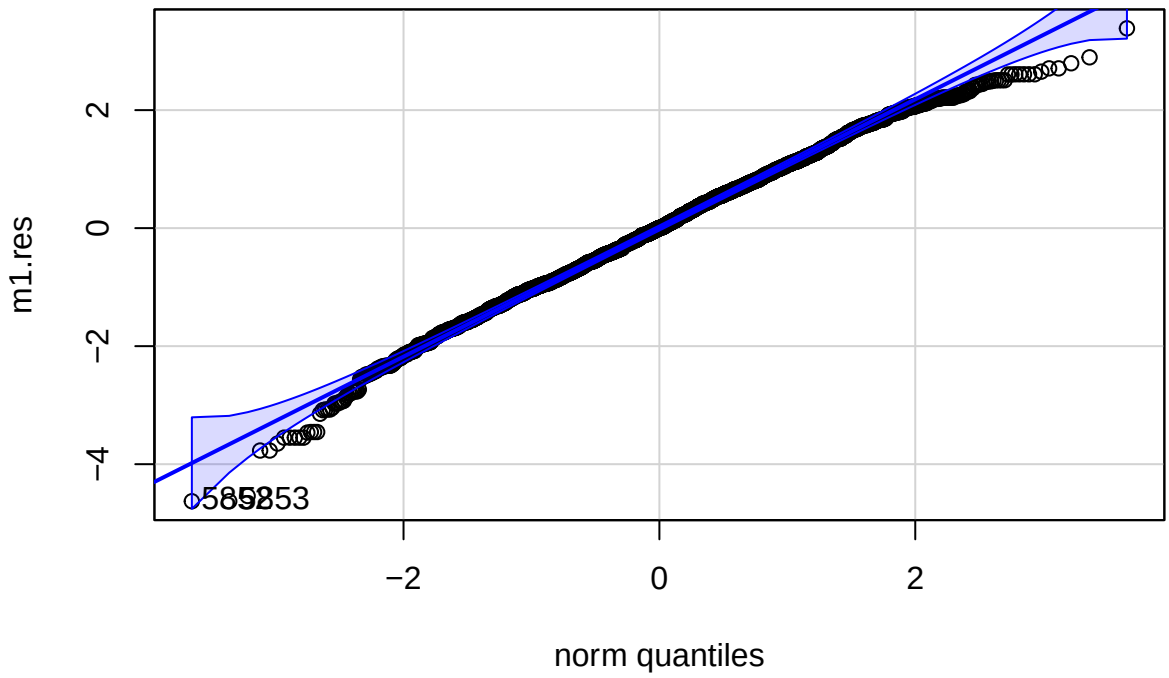
```
## Warning: Removed 4078 rows containing non-finite outside the scale range
## (`stat_count()`).
```




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

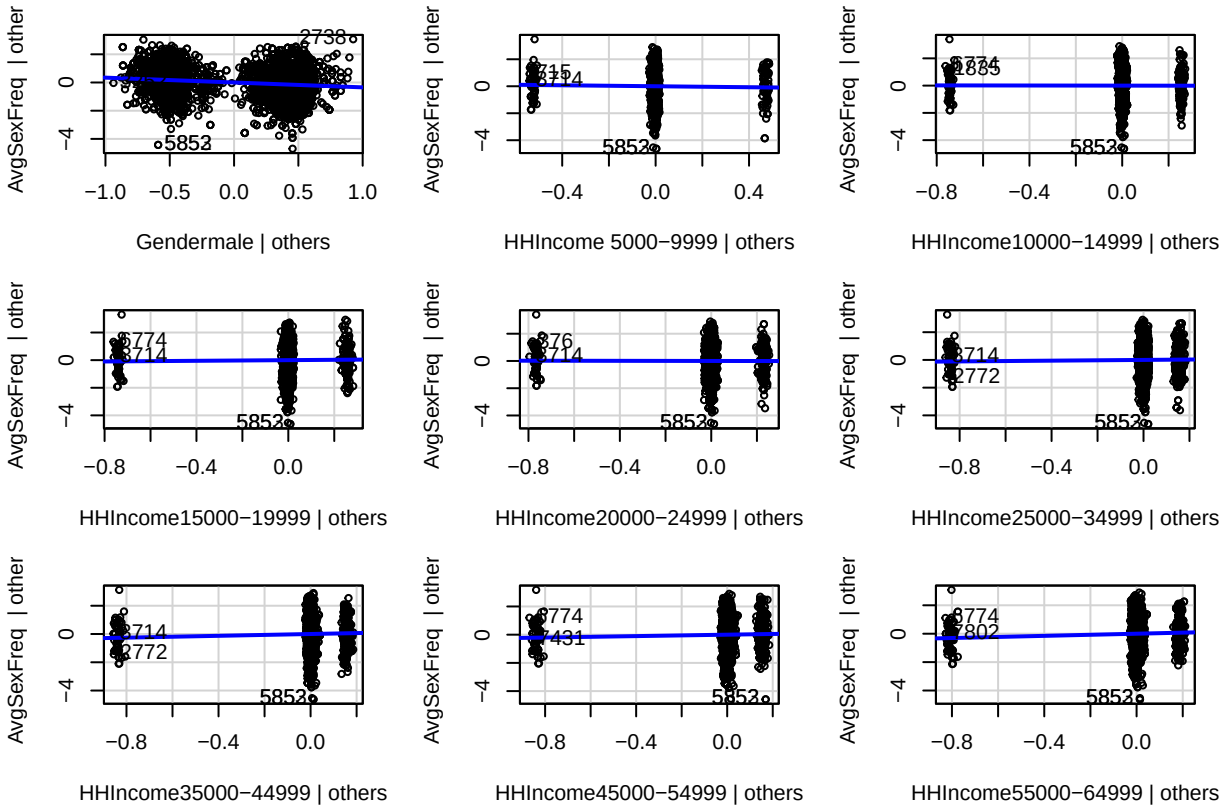
```
## Anova Table (Type III tests)
##
## Response: AvgSexFreq
##                          Sum Sq   Df  F value    Pr(>F)
## (Intercept)               80.4    1  70.2844 < 2.2e-16 ***
## Gender                   106.0    1  92.6660 < 2.2e-16 ***
## HHIncome                 142.6   11  11.3271 < 2.2e-16 ***
## 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
```

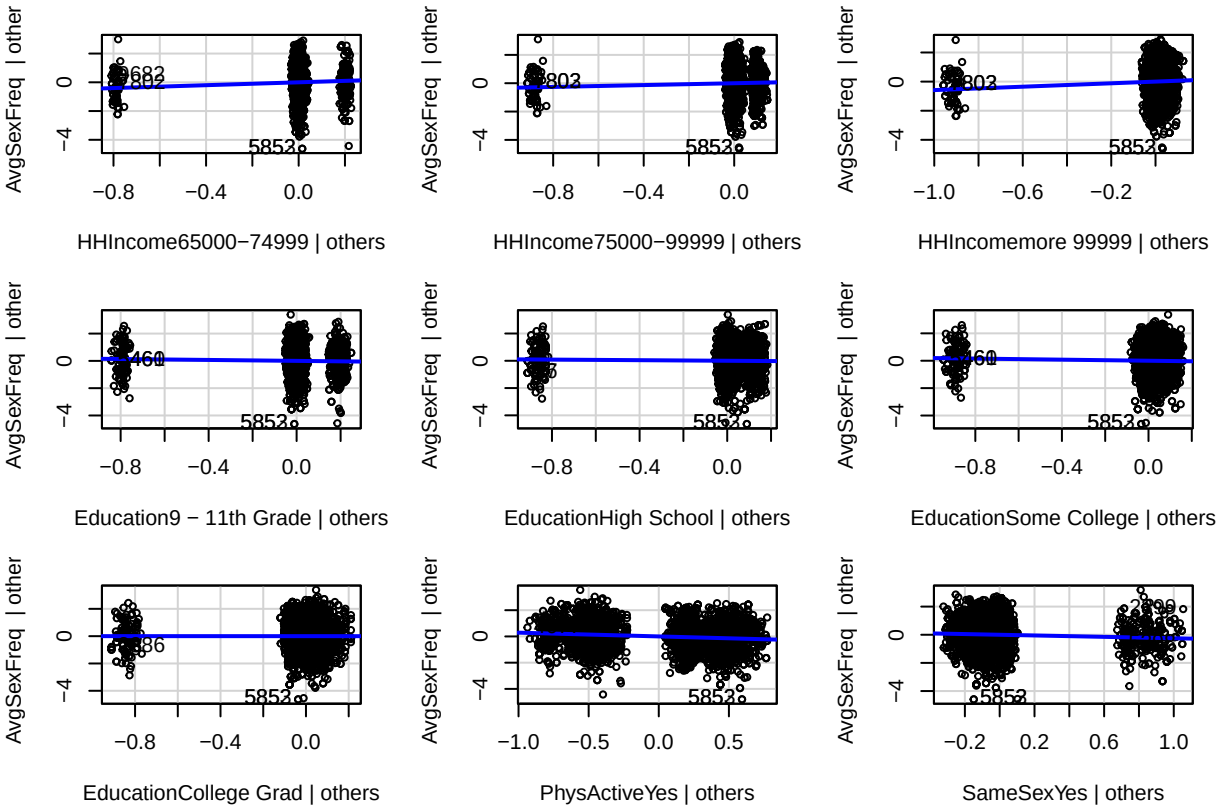
```
m1 = lm(AvgSexFreq ~ Gender+HHIncome+Education+PhysActive+SameSex+AlcoholYear+RegularMarij+HardDrugs+Re
m1.res = m1$residuals

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

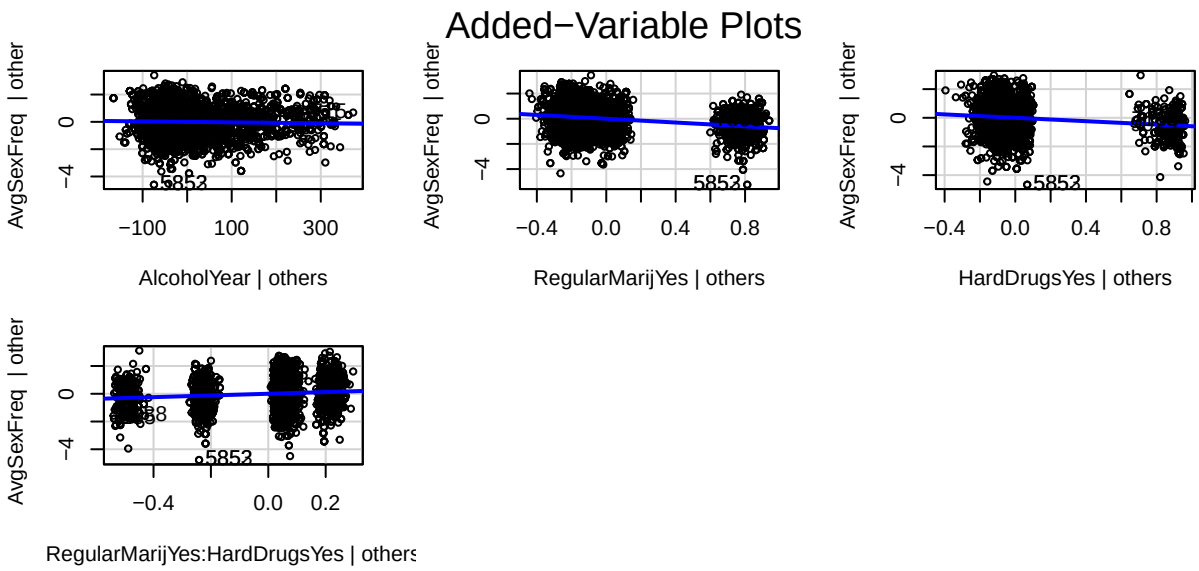



```
## 5852 5853
## 2288 2289
```

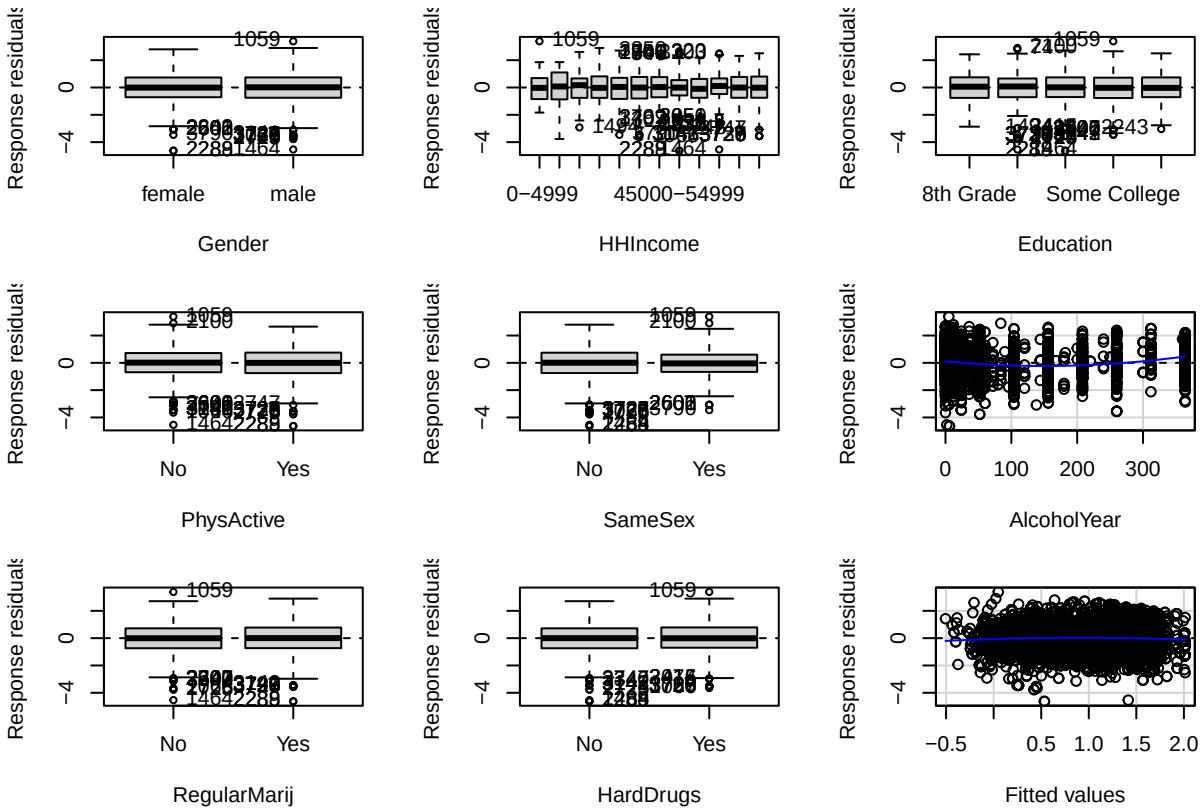
```
car::avPlots(m1)
```


AvgSexFreq | other
Gendermale | others
HHIncome 5000−9999 | others
HHIncome10000−14999 | others
HHIncome15000−19999 | others
HHIncome20000−24999 | others
HHIncome25000−34999 | others
HHIncome35000−44999 | others
HHIncome45000−54999 | others
HHIncome55000−64999 | others


AvgSexFreq | other
HHIncome65000−74999 | others
HHIncome75000−99999 | others
HHIncomemore 99999 | others
Education9 − 11th Grade | others
EducationHigh School | others
EducationSome College | others
EducationCollege Grad | others
PhysActiveYes | others
SameSexYes | others




```
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(???)
car::vif(m1, type = 'predictor')
```

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

```
##                  GVIF Df GVIF^(1/(2*Df)) Interacts With
## Gender       1.081224  1        1.039819           --
## HHIncome     1.350283 11        1.013744           --
## Education    1.431171  4        1.045831           --
## PhysActive   1.133111  1        1.064477           --
## SameSex      1.104540  1        1.050971           --
## AlcoholYear  1.112480  1        1.054741           --
## RegularMarij 1.188909  3        1.029259     HardDrugs
## HardDrugs    1.188909  3        1.029259   RegularMarij
```

```
##                                                                                 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                    Gender, HHIncome, Education, PhysActive, SameSex, AlcoholYear
```