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 via a correlation plot due to many covariates 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')
```

Performing several 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. Preliminary analysis suggest that hard drug use and regular marijuana is associated on average 1-2 years earlier first sexual activity. Thus, drug use may be associated with higher frequency of sexual activity.

```
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
```

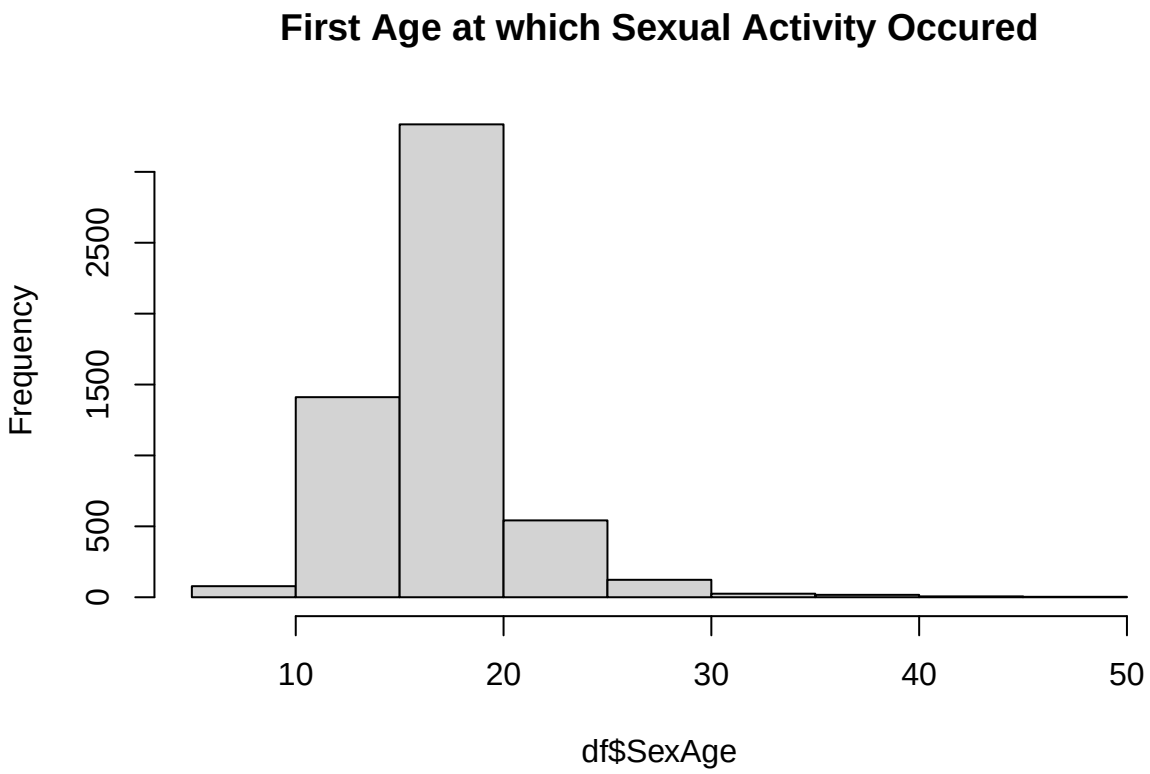
```
model <- lm(SexNumPartnLife ~ RegularMarij+HardDrugs+RegularMarij*HardDrugs, df)
summary(model)
```

```
##
## 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
```

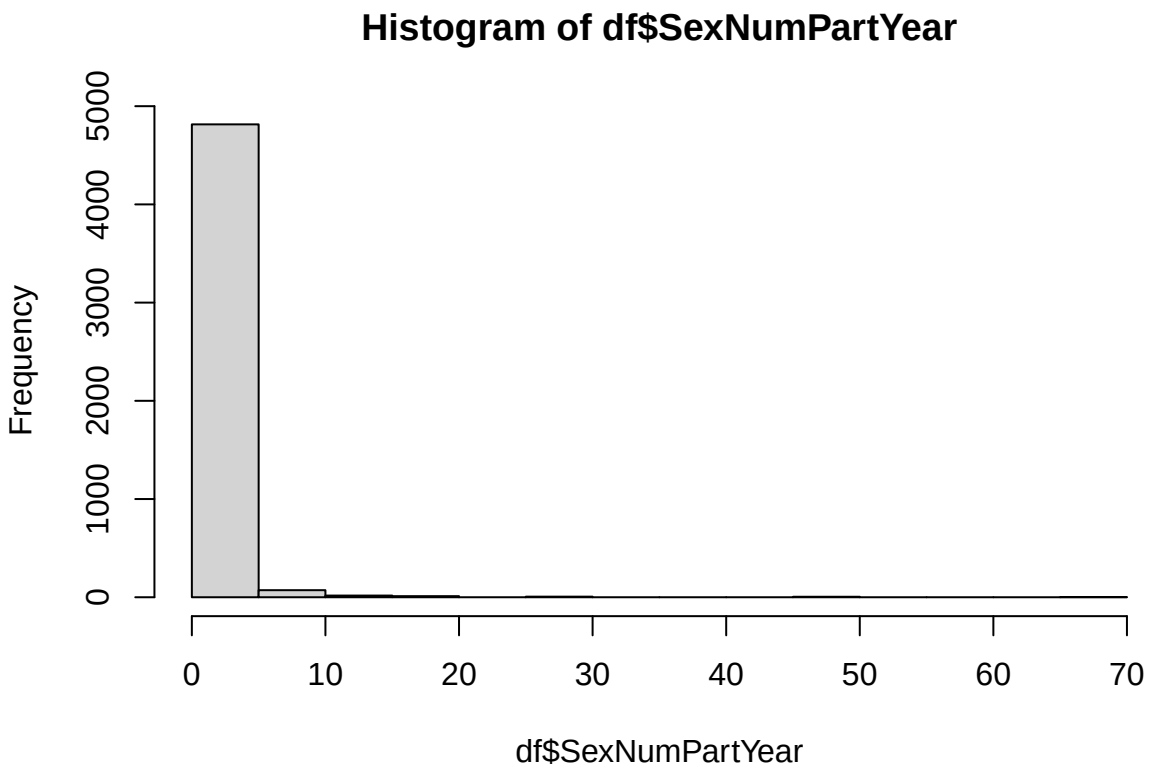
SexAge is has a good distribution but SexNumPartnLife has extreme skenwness and is discrete count data. This requires a Poisson regression which is out side the scopre of this course. Created new variable using the duration, since first sexual activity where (Age - SexAge) since Age >= SexAge, and dividing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

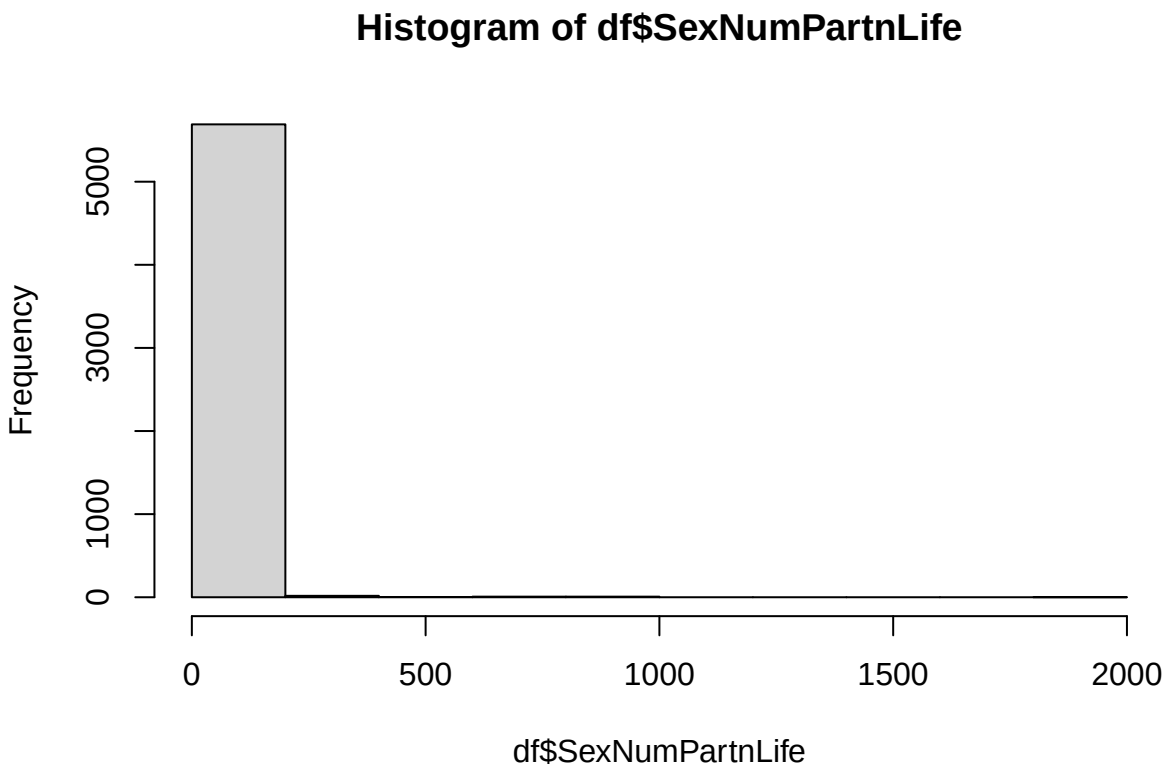
```
hist(df$SexAge, main= "First Age at which Sexual Activity Occured")
```



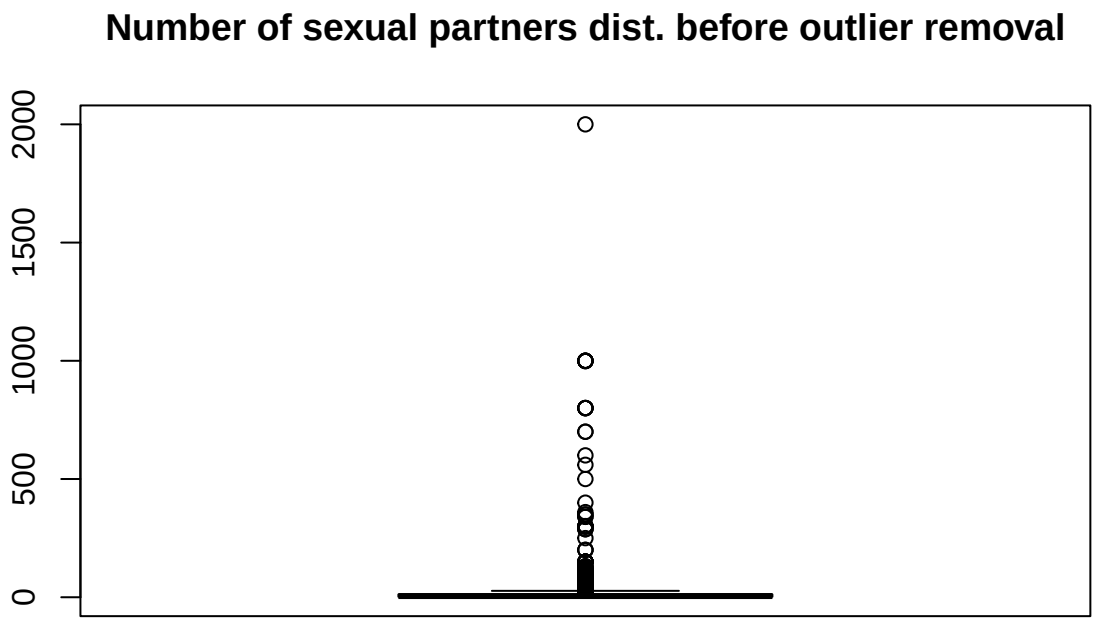

```
hist(df$SexNumPartYear, main = )
```




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



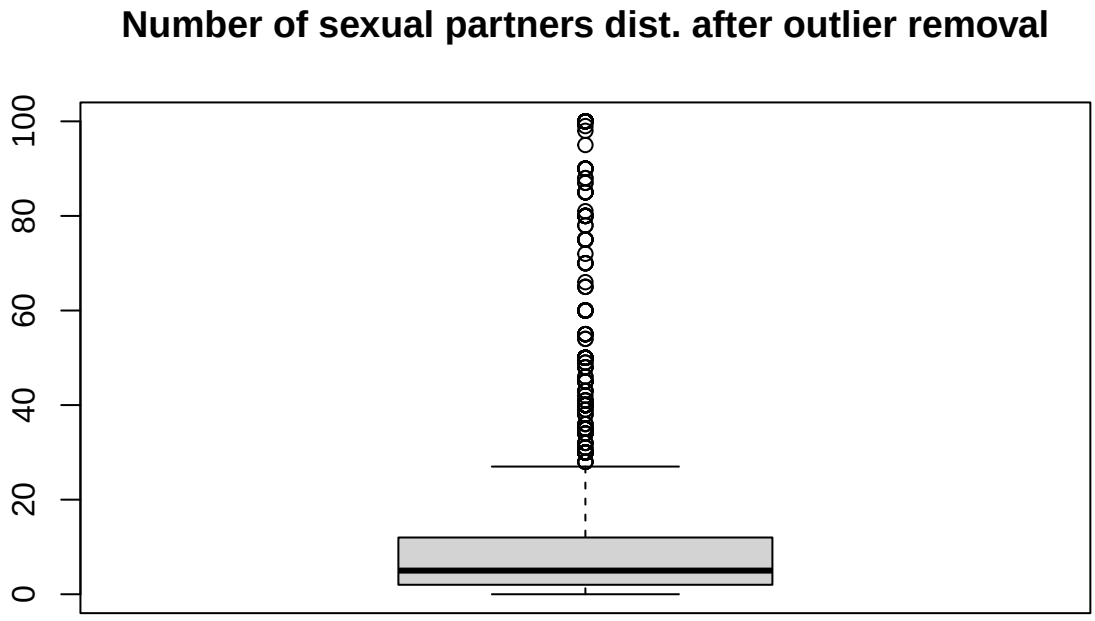

```
#Show observations with more than 300 sexual partners during lifetime
boxplot(df$SexNumPartnLife, main = "Number of sexual partners dist. before outlier removal")
```



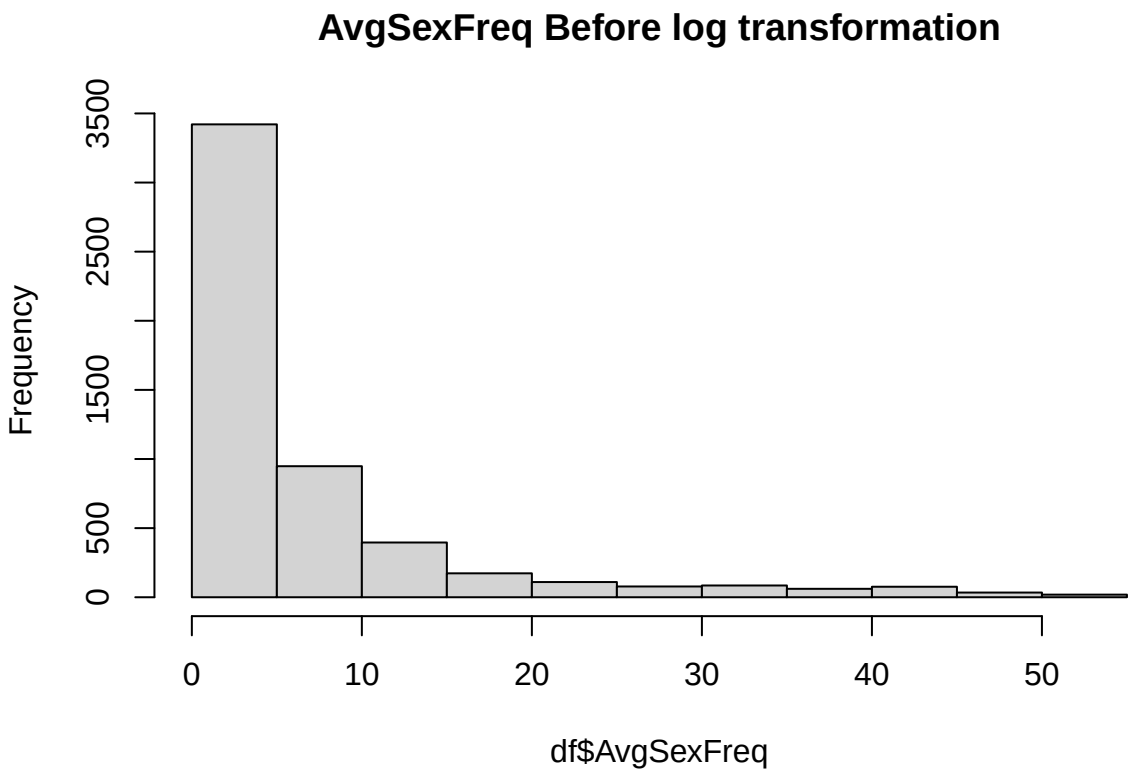

```
df[which(df$SexNumPartnLife > 100), c("Age", "SexAge", "SexNumPartnLife")]
```

```
## # 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
```

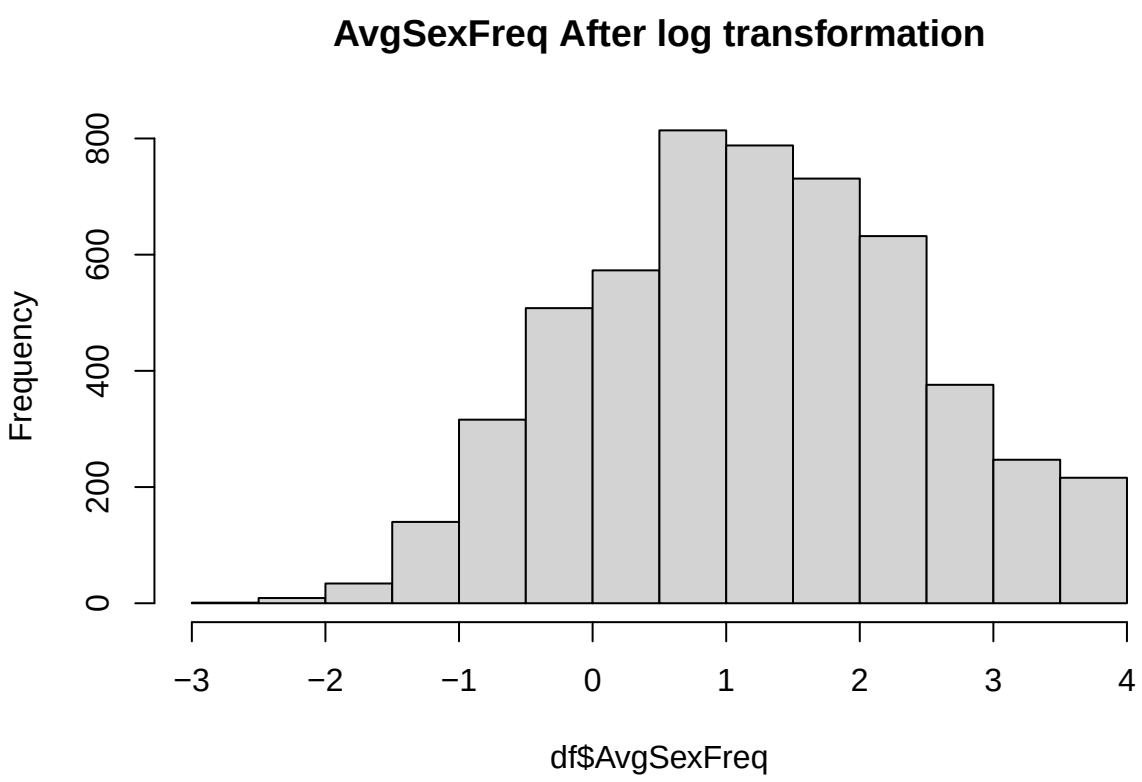
```
df = df[-which(df$SexNumPartnLife > 100),]
boxplot(df$SexNumPartnLife, main = "Number of sexual partners dist. after outlier removal")
```




```
#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")
```




```
tbl_summary(df, by = HardDrugs,
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} / {N} ({p}%)"
    ))
```

```
## 4235 missing rows in the "HardDrugs" column have been removed.
```

Characteristic	**No** N = 4,663[1]	**Yes** N = 1,038[1]
ID	61,874 (5,893)	62,143 (5,938)
SurveyYr		
2009_10	2,353 / 4663 (50%)	501 / 1038 (48%)
2011_12	2,310 / 4663 (50%)	537 / 1038 (52%)
Gender		
female	2,384 / 4663 (51%)	383 / 1038 (37%)
male	2,279 / 4663 (49%)	655 / 1038 (63%)
Age	42 (15)	43 (12)
AgeDecade		
0-9	0 / 4663 (0%)	0 / 1038 (0%)
10-19	207 / 4663 (4.4%)	22 / 1038 (2.1%)
20-29	1,004 / 4663 (22%)	165 / 1038 (16%)
30-39	942 / 4663 (20%)	182 / 1038 (18%)
40-49	914 / 4663 (20%)	299 / 1038 (29%)

50-59	856 / 4663 (18%)	294 / 1038 (28%)
60-69	740 / 4663 (16%)	76 / 1038 (7.3%)
70+	0 / 4663 (0%)	0 / 1038 (0%)
AgeMonths	509 (175)	499 (140)
Unknown	2,310	537
Race1		
Black	547 / 4663 (12%)	87 / 1038 (8.4%)
Hispanic	289 / 4663 (6.2%)	29 / 1038 (2.8%)
Mexican	443 / 4663 (9.5%)	74 / 1038 (7.1%)
White	3,011 / 4663 (65%)	789 / 1038 (76%)
Other	373 / 4663 (8.0%)	59 / 1038 (5.7%)
Race3		
Asian	143 / 2310 (6.2%)	10 / 537 (1.9%)
Black	281 / 2310 (12%)	36 / 537 (6.7%)
Hispanic	152 / 2310 (6.6%)	15 / 537 (2.8%)
Mexican	207 / 2310 (9.0%)	31 / 537 (5.8%)
White	1,477 / 2310 (64%)	416 / 537 (77%)
Other	50 / 2310 (2.2%)	29 / 537 (5.4%)
Unknown	2,353	501
Education		
8th Grade	204 / 4447 (4.6%)	19 / 1016 (1.9%)
9 - 11th Grade	451 / 4447 (10%)	161 / 1016 (16%)
High School	898 / 4447 (20%)	221 / 1016 (22%)
Some College	1,390 / 4447 (31%)	406 / 1016 (40%)
College Grad	1,504 / 4447 (34%)	209 / 1016 (21%)
Unknown	216	22
MaritalStatus		
Divorced	407 / 4455 (9.1%)	155 / 1015 (15%)
LivePartner	345 / 4455 (7.7%)	145 / 1015 (14%)
Married	2,536 / 4455 (57%)	452 / 1015 (45%)
NeverMarried	940 / 4455 (21%)	209 / 1015 (21%)
Separated	111 / 4455 (2.5%)	34 / 1015 (3.3%)
Widowed	116 / 4455 (2.6%)	20 / 1015 (2.0%)
Unknown	208	23
HHIncome		
0-4999	66 / 4338 (1.5%)	23 / 959 (2.4%)
5000-9999	92 / 4338 (2.1%)	22 / 959 (2.3%)
10000-14999	234 / 4338 (5.4%)	51 / 959 (5.3%)
15000-19999	205 / 4338 (4.7%)	54 / 959 (5.6%)
20000-24999	240 / 4338 (5.5%)	57 / 959 (5.9%)
25000-34999	396 / 4338 (9.1%)	118 / 959 (12%)
35000-44999	400 / 4338 (9.2%)	72 / 959 (7.5%)
45000-54999	370 / 4338 (8.5%)	92 / 959 (9.6%)
55000-64999	330 / 4338 (7.6%)	58 / 959 (6.0%)
65000-74999	279 / 4338 (6.4%)	62 / 959 (6.5%)

75000-99999	572 / 4338 (13%)	105 / 959 (11%)
more 99999	1,154 / 4338 (27%)	245 / 959 (26%)
Unknown	325	79
HHIncomeMid	60,841 (32,442)	57,993 (33,065)
Unknown	325	79
Poverty	3.05 (1.67)	2.80 (1.69)
Unknown	279	74
HomeRooms	6 (2)	6 (2)
Unknown	25	5
HomeOwn		
Own	3,096 / 4638 (67%)	584 / 1033 (57%)
Rent	1,443 / 4638 (31%)	420 / 1033 (41%)
Other	99 / 4638 (2.1%)	29 / 1033 (2.8%)
Unknown	25	5
Work		
Looking	179 / 4662 (3.8%)	77 / 1038 (7.4%)
NotWorking	1,260 / 4662 (27%)	266 / 1038 (26%)
Working	3,223 / 4662 (69%)	695 / 1038 (67%)
Unknown	1	0
Weight	83 (22)	85 (20)
Unknown	29	2
Length	NA (NA)	NA (NA)
Unknown	4,663	1,038
HeadCirc	NA (NA)	NA (NA)
Unknown	4,663	1,038
Height	169 (10)	172 (9)
Unknown	21	2
BMI	29 (7)	28 (6)
Unknown	29	2
BMICatUnder20yrs		
UnderWeight	15 / 103 (15%)	0 / 7 (0%)
NormWeight	54 / 103 (52%)	7 / 7 (100%)
OverWeight	10 / 103 (9.7%)	0 / 7 (0%)
Obese	24 / 103 (23%)	0 / 7 (0%)
Unknown	4,560	1,031
BMI_WHO		
12.0_18.5	95 / 4615 (2.1%)	9 / 1029 (0.9%)
18.5_to_24.9	1,343 / 4615 (29%)	310 / 1029 (30%)
25.0_to_29.9	1,480 / 4615 (32%)	354 / 1029 (34%)
30.0_plus	1,697 / 4615 (37%)	356 / 1029 (35%)
Unknown	48	9
Pulse	73 (12)	72 (11)
Unknown	73	15
BPSysAve	118 (15)	120 (16)
Unknown	78	17

BPDiaAve	70 (12)	73 (11)
Unknown	78	17
BPSys1	120 (16)	121 (16)
Unknown	237	49
BPDia1	71 (12)	73 (11)
Unknown	237	49
BPSys2	119 (15)	120 (16)
Unknown	174	28
BPDia2	71 (12)	73 (11)
Unknown	174	28
BPSys3	118 (15)	120 (16)
Unknown	159	25
BPDia3	70 (12)	72 (12)
Unknown	159	25
Testosterone	219 (228)	268 (258)
Unknown	2,492	529
DirectChol	1.36 (0.42)	1.36 (0.42)
Unknown	193	26
TotChol	5.04 (1.05)	5.23 (1.14)
Unknown	193	26
UrineVol1	127 (94)	132 (93)
Unknown	16	1
UrineFlow1	1.07 (0.98)	1.07 (1.01)
Unknown	256	61
UrineVol2	130 (93)	123 (87)
Unknown	3,910	912
UrineFlow2	1.23 (1.12)	1.22 (1.20)
Unknown	3,912	912
Diabetes	368 / 4661 (7.9%)	95 / 1038 (9.2%)
Unknown	2	0
DiabetesAge	46 (13)	42 (12)
Unknown	4,362	963
HealthGen		
Excellent	586 / 4663 (13%)	82 / 1032 (7.9%)
Vgood	1,561 / 4663 (33%)	327 / 1032 (32%)
Good	1,824 / 4663 (39%)	420 / 1032 (41%)
Fair	593 / 4663 (13%)	169 / 1032 (16%)
Poor	99 / 4663 (2.1%)	34 / 1032 (3.3%)
Unknown	0	6
DaysPhysHlthBad	3 (7)	4 (8)
Unknown	0	6
DaysMentHlthBad	4 (8)	6 (9)
Unknown	1	6
LittleInterest		
None	3,639 / 4661 (78%)	704 / 1027 (69%)

Several	762 / 4661 (16%)	215 / 1027 (21%)
Most	260 / 4661 (5.6%)	108 / 1027 (11%)
Unknown	2	11
Depressed		
None	3,762 / 4663 (81%)	694 / 1032 (67%)
Several	652 / 4663 (14%)	231 / 1032 (22%)
Most	249 / 4663 (5.3%)	107 / 1032 (10%)
Unknown	0	6
nPregnancies	3 (2)	3 (2)
Unknown	2,872	721
nBabies	2 (1)	2 (1)
Unknown	2,998	760
Age1stBaby	23 (5)	22 (5)
Unknown	3,389	846
SleepHrsNight	7 (1)	7 (1)
Unknown	6	5
SleepTrouble	1,062 / 4663 (23%)	399 / 1038 (38%)
PhysActive	2,684 / 4663 (58%)	540 / 1038 (52%)
PhysActiveDays		
1	287 / 2446 (12%)	52 / 500 (10%)
2	427 / 2446 (17%)	106 / 500 (21%)
3	584 / 2446 (24%)	137 / 500 (27%)
4	306 / 2446 (13%)	68 / 500 (14%)
5	411 / 2446 (17%)	81 / 500 (16%)
6	130 / 2446 (5.3%)	14 / 500 (2.8%)
7	301 / 2446 (12%)	42 / 500 (8.4%)
Unknown	2,217	538
TVHrsDay		
0_hrs	48 / 2309 (2.1%)	19 / 537 (3.5%)
0_to_1_hr	326 / 2309 (14%)	66 / 537 (12%)
1_hr	420 / 2309 (18%)	89 / 537 (17%)
2_hr	593 / 2309 (26%)	147 / 537 (27%)
3_hr	403 / 2309 (17%)	90 / 537 (17%)
4_hr	238 / 2309 (10%)	52 / 537 (9.7%)
More_4_hr	281 / 2309 (12%)	74 / 537 (14%)
Unknown	2,354	501
CompHrsDay		
0_hrs	391 / 2310 (17%)	103 / 537 (19%)
0_to_1_hr	622 / 2310 (27%)	190 / 537 (35%)
1_hr	540 / 2310 (23%)	111 / 537 (21%)
2_hr	318 / 2310 (14%)	57 / 537 (11%)
3_hr	170 / 2310 (7.4%)	34 / 537 (6.3%)
4_hr	111 / 2310 (4.8%)	14 / 537 (2.6%)
More_4_hr	158 / 2310 (6.8%)	28 / 537 (5.2%)
Unknown	2,353	501

TVHrsDayChild	NA (NA)	NA (NA)
Unknown	4,663	1,038
CompHrsDayChild	NA (NA)	NA (NA)
Unknown	4,663	1,038
Alcohol12PlusYr	3,572 / 4561 (78%)	968 / 1018 (95%)
Unknown	102	20
AlcoholDay	3 (3)	4 (3)
Unknown	1,174	172
AlcoholYear	67 (95)	100 (109)
Unknown	570	26
SmokeNow	759 / 1598 (47%)	434 / 791 (55%)
Unknown	3,065	247
Smoke100	1,598 / 4456 (36%)	791 / 1016 (78%)
Unknown	207	22
Smoke100n		
Non-Smoker	2,858 / 4456 (64%)	225 / 1016 (22%)
Smoker	1,598 / 4456 (36%)	791 / 1016 (78%)
Unknown	207	22
SmokeAge	18 (4)	17 (6)
Unknown	3,124	281
Marijuana	1,913 / 3916 (49%)	927 / 962 (96%)
Unknown	747	76
AgeFirstMarij	18 (4)	16 (4)
Unknown	2,751	111
RegularMarij	659 / 3916 (17%)	662 / 962 (69%)
Unknown	747	76
AgeRegMarij	18 (4)	17 (5)
Unknown	4,004	376
SexEver	4,431 / 4653 (95%)	1,038 / 1038 (100%)
Unknown	10	0
SexAge	18 (4)	16 (3)
Unknown	236	0
SexNumPartnLife	8 (13)	21 (22)
Unknown	44	4
SexNumPartYear	1 (2)	1 (2)
Unknown	759	76
SameSex	210 / 4654 (4.5%)	190 / 1038 (18%)
Unknown	9	0
SexOrientation		
Bisexual	70 / 3833 (1.8%)	45 / 949 (4.7%)
Heterosexual	3,713 / 3833 (97%)	877 / 949 (92%)
Homosexual	50 / 3833 (1.3%)	27 / 949 (2.8%)
Unknown	830	89
PregnantNow		
Yes	60 / 1209 (5.0%)	1 / 170 (0.6%)

No	1,124 / 1209 (93%)	169 / 170 (99%)
Unknown	25 / 1209 (2.1%)	0 / 170 (0%)
Unknown	3,454	868
AvgSexFreq	NA (NA)	Inf (NA)
Unknown	269	4

[1] Mean (SD); n / N (%)

```
#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

```
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
```

```
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
```

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

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

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

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

```
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
```

```
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	$\beta_{Substance}$	13.88444	26.9329	5.155204576	$F_{5,34}$	0.001262146	Reject	Do not stop	Collect
II	β_{Demo}	55.61473	26.9329	25.81174	$F_{4,34}$	6.872507e-10	Reject	Do not stop	Collect
III	β_{Health}	5.687399	26.9329	2.11169493	$F_{5,34}$	0.08788892	Fail to Reject	Stop	Not Collect
IV	$\beta_{SameSex}$	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]), "Not Missing", "Missing")

tbl_summary(df[,c("Age", "Gender", "HHIncome", "Education", "MaritalStatus", "missingness")], by = missi
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} / {N} ({p}%)"
    ))
```

Characteristic	**Missing** N = 8,175[1]	**Not Missing** N = 1,761[1]
Age	36 (24)	41 (11)
Gender		
female	4,254 / 8175 (52%)	751 / 1761 (43%)
male	3,921 / 8175 (48%)	1,010 / 1761 (57%)
HHIncome		
0-4999	153 / 7374 (2.1%)	39 / 1761 (2.2%)
5000-9999	207 / 7374 (2.8%)	42 / 1761 (2.4%)
10000-14999	431 / 7374 (5.8%)	108 / 1761 (6.1%)
15000-19999	421 / 7374 (5.7%)	104 / 1761 (5.9%)
20000-24999	481 / 7374 (6.5%)	129 / 1761 (7.3%)
25000-34999	770 / 7374 (10%)	179 / 1761 (10%)
35000-44999	717 / 7374 (9.7%)	143 / 1761 (8.1%)
45000-54999	617 / 7374 (8.4%)	165 / 1761 (9.4%)
55000-64999	489 / 7374 (6.6%)	129 / 1761 (7.3%)
65000-74999	410 / 7374 (5.6%)	113 / 1761 (6.4%)
75000-99999	860 / 7374 (12%)	220 / 1761 (12%)
more 99999	1,818 / 7374 (25%)	390 / 1761 (22%)
Unknown	801	0
Education		
8th Grade	378 / 5399 (7.0%)	68 / 1761 (3.9%)
9 - 11th Grade	581 / 5399 (11%)	292 / 1761 (17%)

High School	1,071 / 5399 (20%)	426 / 1761 (24%)
Some College	1,660 / 5399 (31%)	595 / 1761 (34%)
College Grad	1,709 / 5399 (32%)	380 / 1761 (22%)
Unknown	2,776	0
MaritalStatus		
Divorced	485 / 5411 (9.0%)	214 / 1759 (12%)
LivePartner	282 / 5411 (5.2%)	266 / 1759 (15%)
Married	3,080 / 5411 (57%)	845 / 1759 (48%)
NeverMarried	997 / 5411 (18%)	365 / 1759 (21%)
Separated	134 / 5411 (2.5%)	48 / 1759 (2.7%)
Widowed	433 / 5411 (8.0%)	21 / 1759 (1.2%)
Unknown	2,764	2

[1] Mean (SD); n / N (%)

```
#logit(missingness~Age+Gender+HHIncome+Education+MaritalStatus,df)
t.test(df[df$missingness == "Missing",]$Age, df[df$missingness == "Not Missing",]$Age)
```

```
##
##  Welch Two Sample t-test
##
## data:  df[df$missingness == "Missing", ]$Age and df[df$missingness == "Not Missing", ]$Age
## t = -13.182, df = 5677.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.747404 -4.259286
## sample estimates:
## mean of x mean of y
##  35.81040  40.81374
```

```
#for{}
#pdf export
```

Missingness for occurs for those aged below 20 because they are not recorded for some covariates. Why missingness for those aged above 60 occurs is unclear.

```
library(gridExtra)
```

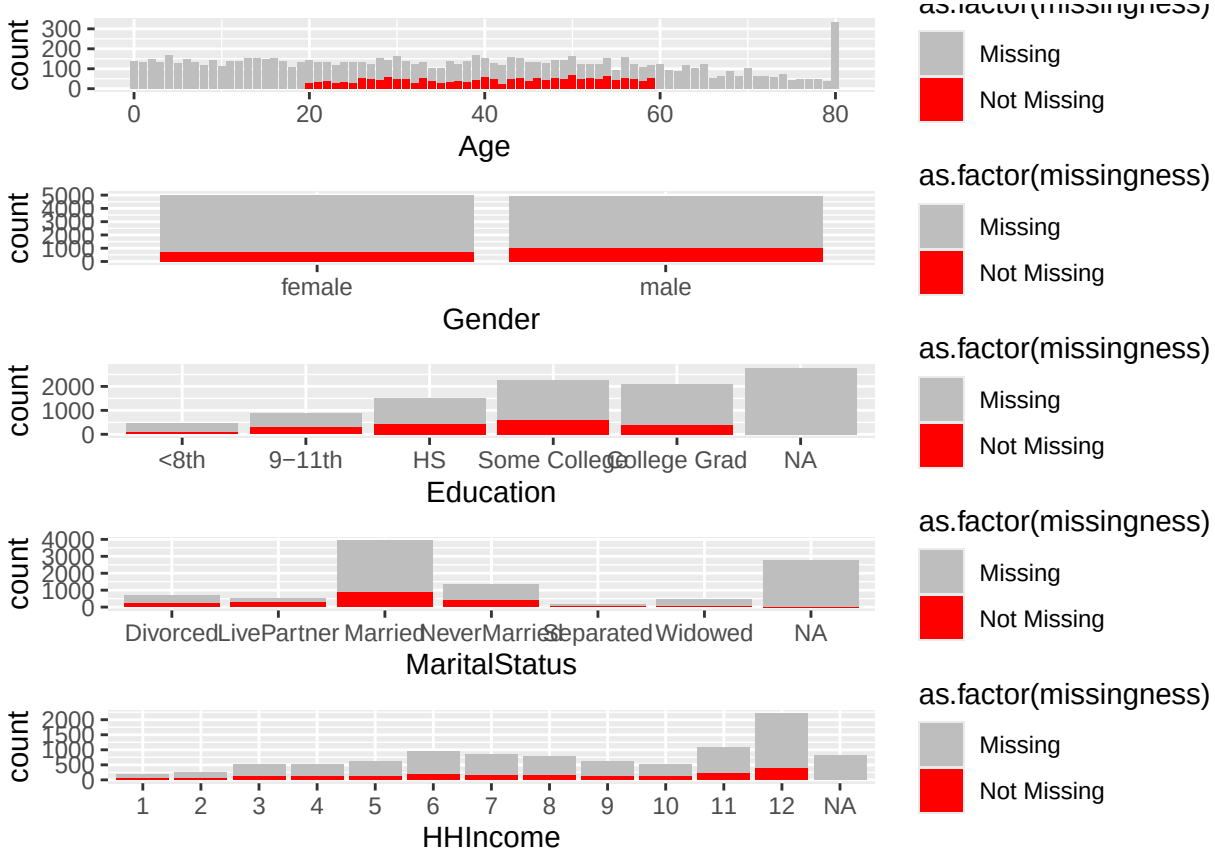
```
## Warning: package 'gridExtra' was built under R version 4.4.2
```

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

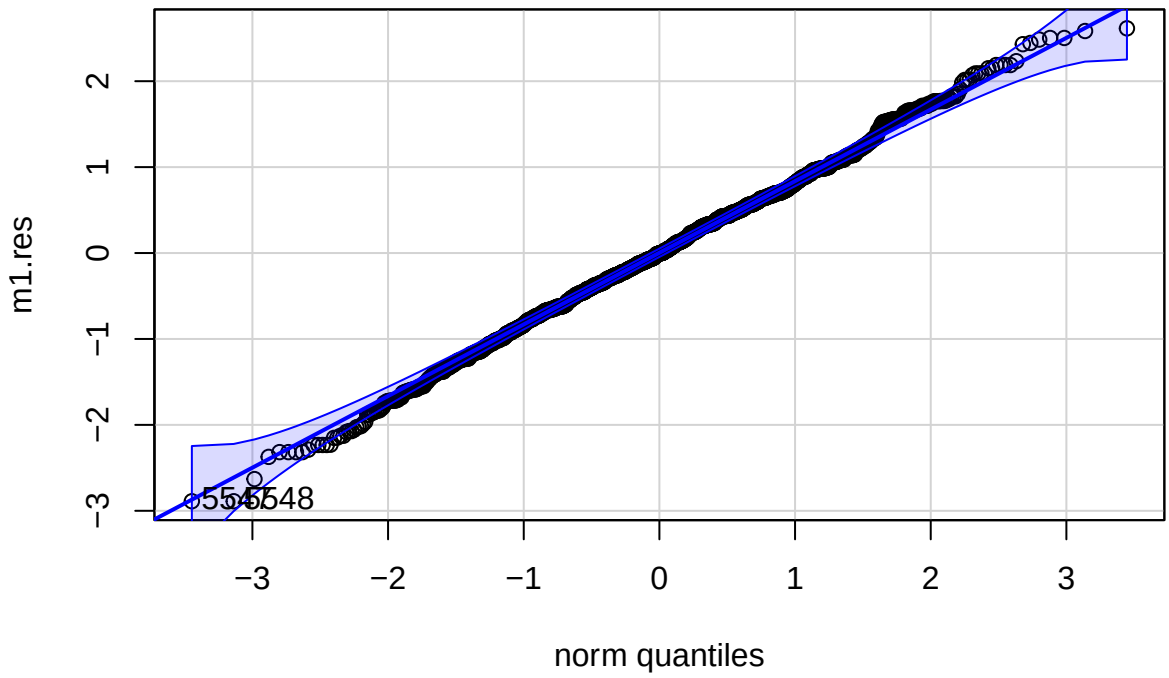
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
p1 = ggplot(data = df, mapping=aes(x=Age, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p2 = ggplot(data = df, mapping=aes(x=Gender, 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_x_discrete(labels = c("<8th", "9-11th", "HS", "Some College", "College Grad" ))+
  scale_fill_manual(values = c("gray", "red"))
p4 = ggplot(data = df, mapping=aes(x=MaritalStatus, fill=as.factor(missingness)))+
  geom_bar(stat="count")+
  scale_fill_manual(values = c("gray", "red"))
p5 = 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, "NA")) +
  scale_fill_manual(values = c("gray", "red"))

grid.arrange(p1,p2,p3,p4,p5, nrow=5)
```

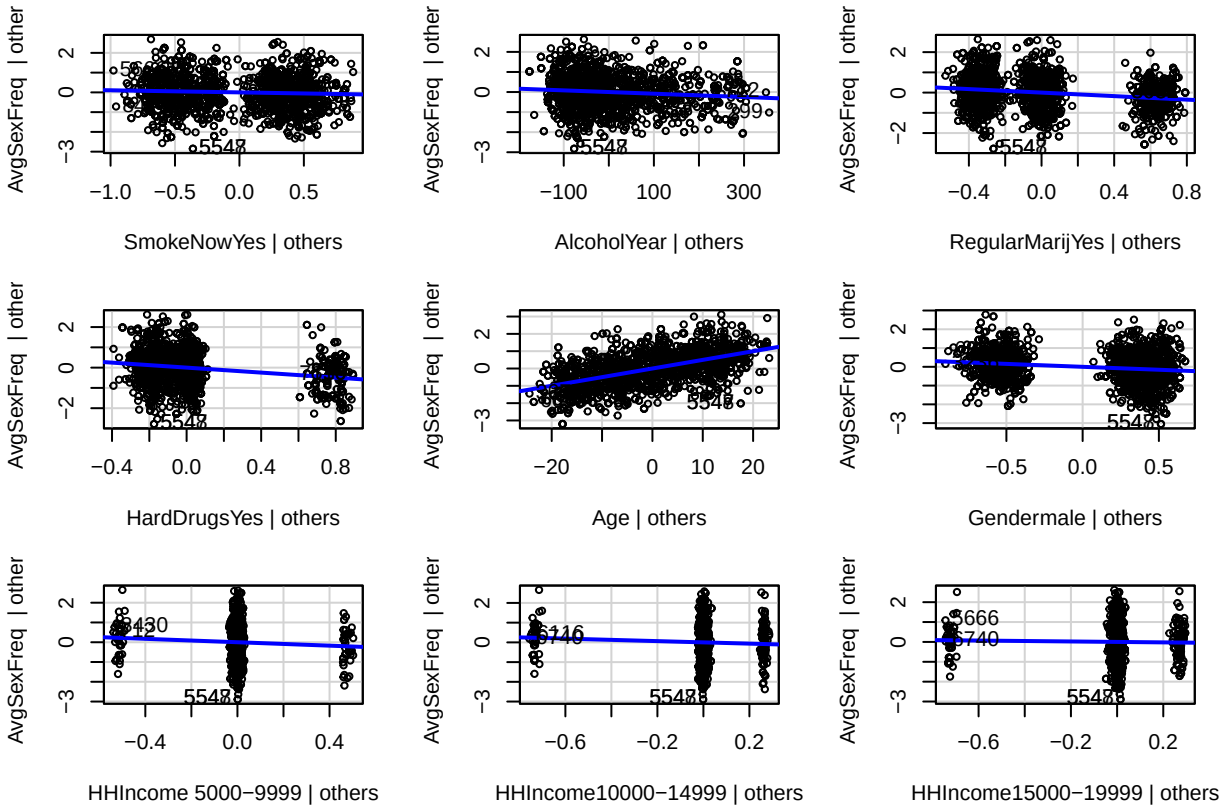


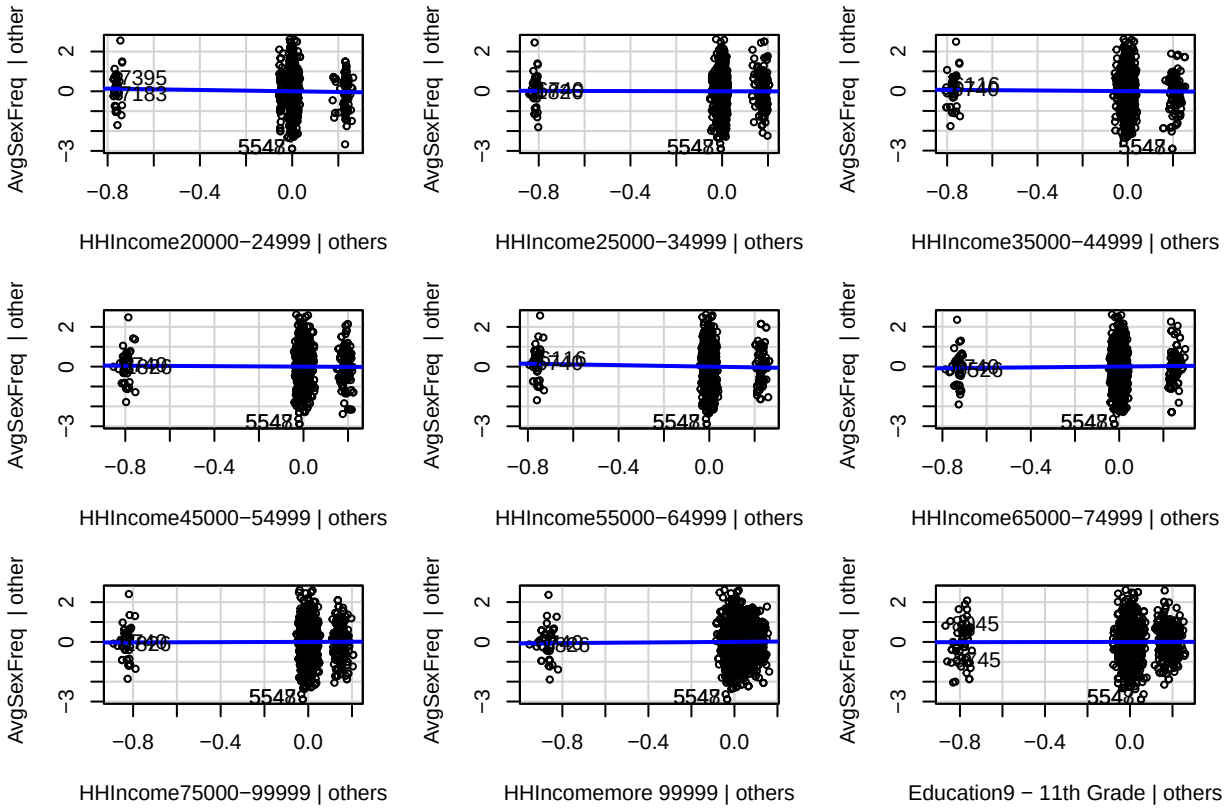

```
m1 = lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+HHIn
m1.res = m1$residuals

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

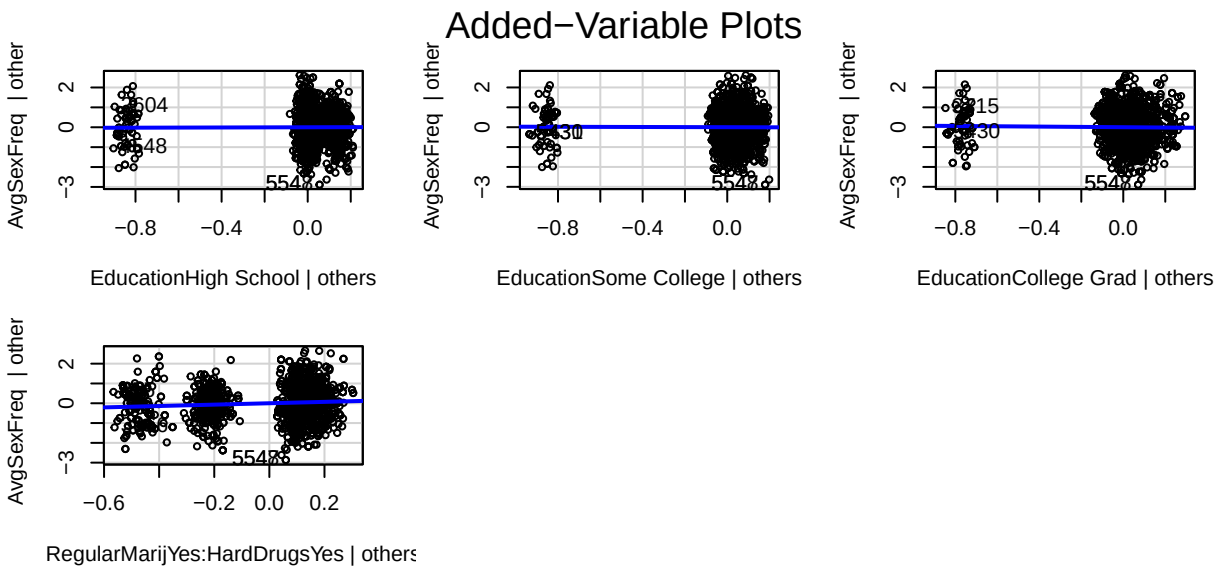



```
## 5547 5548
## 1013 1014
```

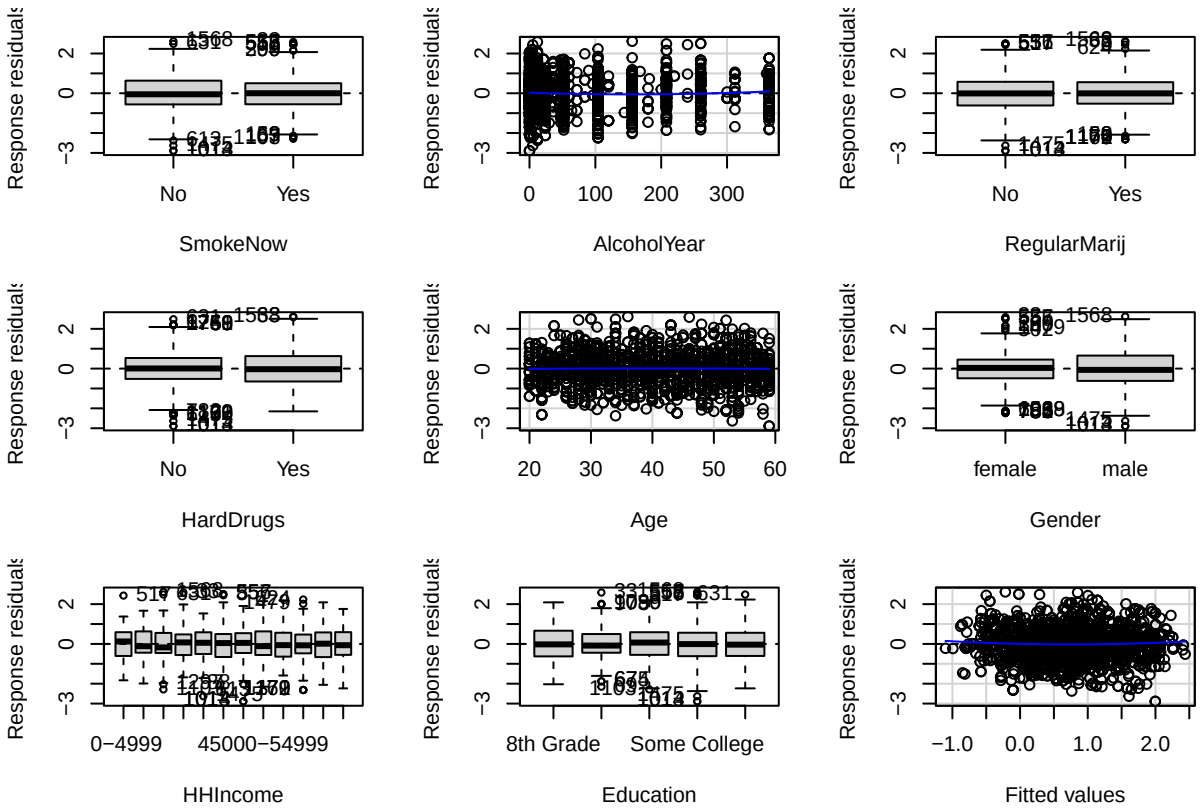
```
car::avPlots(m1)
```


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
```

```
car::durbinWatsonTest(m1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3904876      1.215041       0
##  Alternative hypothesis: rho != 0
```

```
#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
```

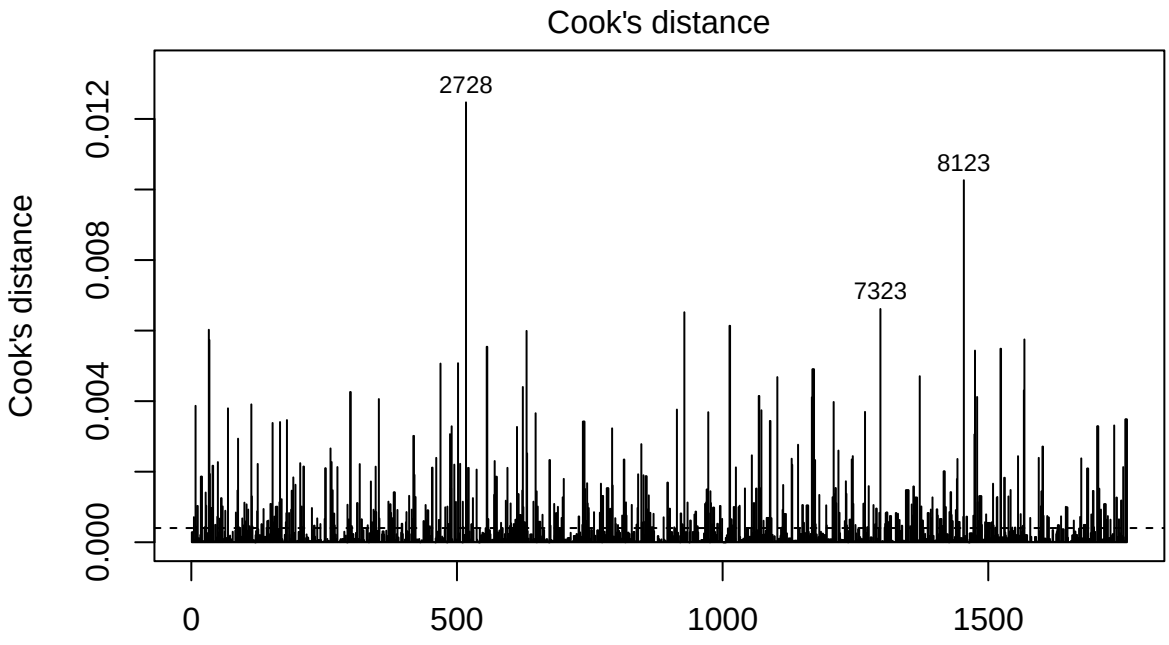
```
model.deffits=dffits(m1)
model.CD = cooks.distance(m1)
model.deffits[which.max(model.deffits)]
```

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

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

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

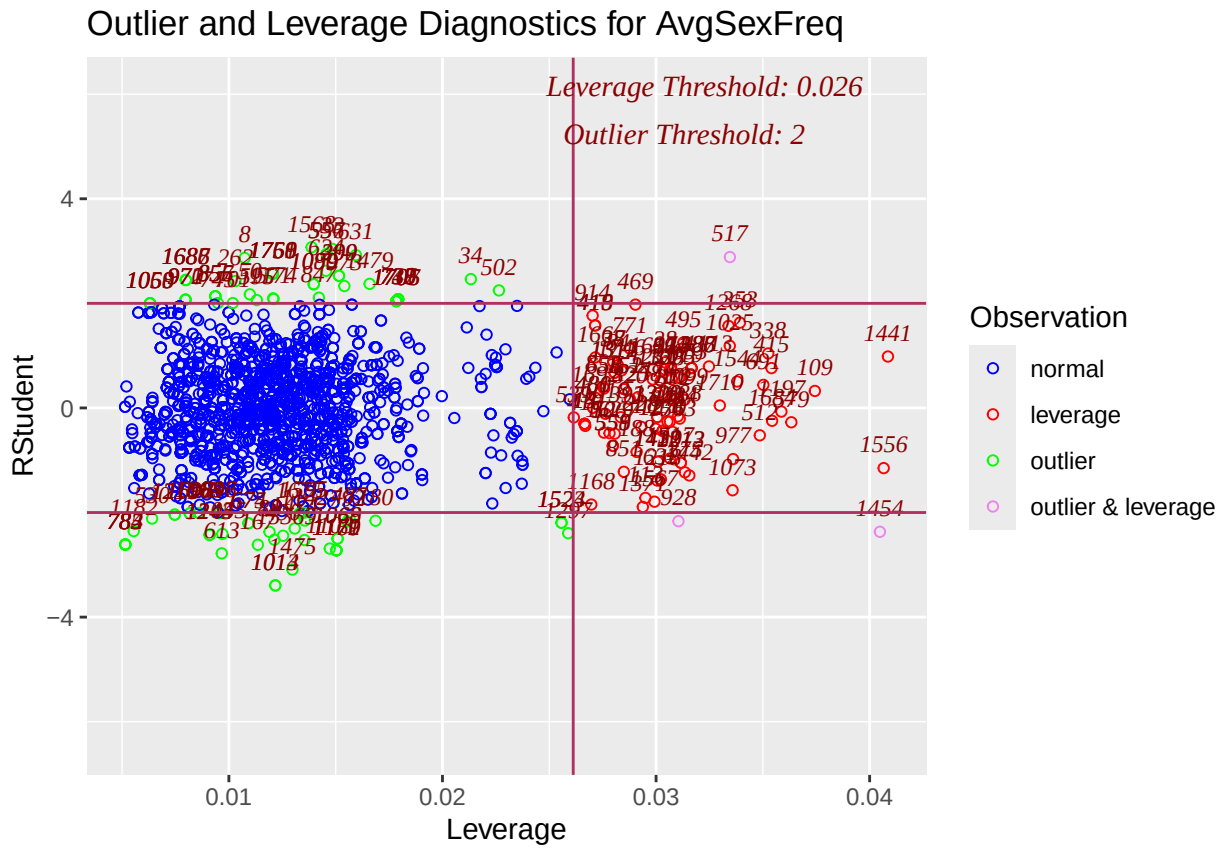
```
n = nrow(df)
p = m1$rank
plot(m1, which = 4)
abline(h=4/n,lty=2)
```




```
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>, ...
```

```
df2 = df[-c(517, 928, 1454),]
m2 = lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+HHIn
summary(m1)
```

```
##
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + HHIncome +
##     Education, data = df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.88889 -0.55938 -0.00535  0.56525  2.61514
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.5234173  0.1968264  -2.659 0.007903 **
## SmokeNowYes                 -0.1070640  0.0444784  -2.407 0.016183 *
## AlcoholYear                 -0.0008302  0.0002030  -4.090 4.51e-05 ***
## RegularMarijYes             -0.4340106  0.0550361  -7.886 5.47e-15 ***
## HardDrugsYes                -0.6147064  0.0737248  -8.338  < 2e-16 ***
## Age                          0.0498163  0.0019031  26.176  < 2e-16 ***
## Gendermale                  -0.3120928  0.0425374  -7.337 3.34e-13 ***
## HHIncome 5000-9999          -0.4319851  0.1920884  -2.249 0.024644 *
## HHIncome10000-14999         -0.3191854  0.1608774  -1.984 0.047410 *
## HHIncome15000-19999         -0.1152378  0.1620180  -0.711 0.477015
## HHIncome20000-24999         -0.1653589  0.1577412  -1.048 0.294649
## HHIncome25000-34999         -0.0259102  0.1524637  -0.170 0.865074
## HHIncome35000-44999         -0.0804727  0.1562915  -0.515 0.606696
## HHIncome45000-54999         -0.0592029  0.1540952  -0.384 0.700879
## HHIncome55000-64999         -0.1858836  0.1578251  -1.178 0.239045
## HHIncome65000-74999          0.1062340  0.1611567   0.659 0.509856
## HHIncome75000-99999          0.0317870  0.1509424   0.211 0.833232
## HHIncomemore 99999           0.0765356  0.1473477   0.519 0.603533
## Education9 - 11th Grade      0.0144927  0.1172584   0.124 0.901649
## EducationHigh School         0.0267950  0.1137107   0.236 0.813739
## EducationSome College       -0.0258883  0.1126412  -0.230 0.818251
## EducationCollege Grad       -0.0785033  0.1190104  -0.660 0.509576
## RegularMarijYes:HardDrugsYes  0.3481896  0.0946375   3.679 0.000241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8589 on 1738 degrees of freedom
##   (8175 observations deleted due to missingness)
## Multiple R-squared:  0.3928, Adjusted R-squared:  0.3851
## F-statistic: 51.11 on 22 and 1738 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + HHIncome +
##     Education, data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.88971 -0.55959 -0.00742  0.56539  2.61480
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.5229637  0.1967912  -2.657 0.007945 **
## SmokeNowYes                 -0.1055055  0.0444871  -2.372 0.017820 *
## AlcoholYear                 -0.0008325  0.0002030  -4.102 4.29e-05 ***
## RegularMarijYes             -0.4340014  0.0550261  -7.887 5.42e-15 ***
## HardDrugsYes                -0.6215872  0.0739085  -8.410  < 2e-16 ***
```

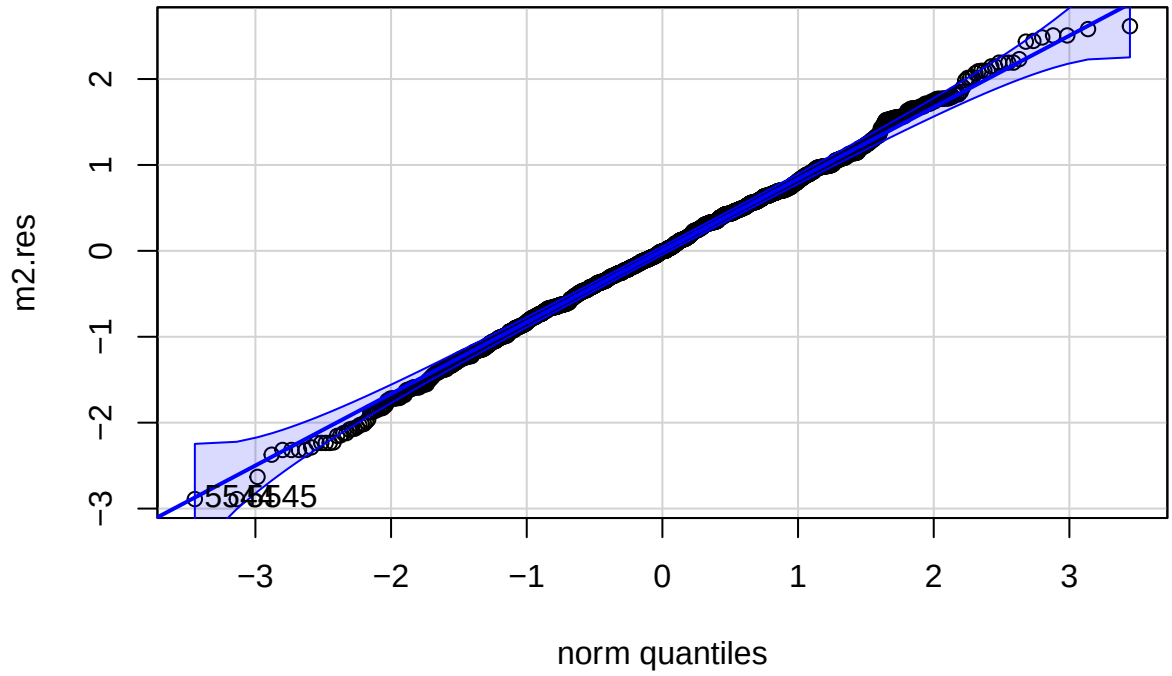
```
## Age                           0.0498322  0.0019028  26.188  < 2e-16 ***
## Gendermale                   -0.3132620  0.0425396  -7.364 2.74e-13 ***
## HHIncome 5000-9999           -0.4329887  0.1920553  -2.255 0.024289 *
## HHIncome10000-14999          -0.3193753  0.1608484  -1.986 0.047238 *
## HHIncome15000-19999          -0.1153468  0.1619888  -0.712 0.476519
## HHIncome20000-24999          -0.1660616  0.1577136  -1.053 0.292519
## HHIncome25000-34999          -0.0258725  0.1524362  -0.170 0.865245
## HHIncome35000-44999          -0.0806444  0.1562633  -0.516 0.605864
## HHIncome45000-54999          -0.0597495  0.1540679  -0.388 0.698202
## HHIncome55000-64999          -0.1863950  0.1577970  -1.181 0.237672
## HHIncome65000-74999           0.1065264  0.1611277   0.661 0.508616
## HHIncome75000-99999           0.0260743  0.1509815   0.173 0.862909
## HHIncomemore 99999            0.0757158  0.1473225   0.514 0.607355
## Education9 - 11th Grade       0.0144719  0.1172373   0.123 0.901772
## EducationHigh School          0.0243232  0.1137067   0.214 0.830641
## EducationSome College        -0.0251184  0.1126224  -0.223 0.823537
## EducationCollege Grad        -0.0769181  0.1189954  -0.646 0.518109
## RegularMarijYes:HardDrugsYes  0.3553097  0.0947848   3.749 0.000184 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8588 on 1737 degrees of freedom
##   (8173 observations deleted due to missingness)
## Multiple R-squared:  0.3932, Adjusted R-squared:  0.3855
## F-statistic: 51.16 on 22 and 1737 DF,  p-value: < 2.2e-16
```

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

```
##                   (Intercept)                  SmokeNowYes
##                  -0.086645412                -1.455691842
##                   AlcoholYear              RegularMarijYes
##                  -0.270904233                -0.002101475
##                  HardDrugsYes                         Age
##                  -1.119365199                 0.031929309
##                    Gendermale           HHIncome 5000-9999
##                  -0.374620377                -0.232307029
##           HHIncome10000-14999          HHIncome15000-19999
##                  -0.059495684                -0.094581094
##           HHIncome20000-24999          HHIncome25000-34999
##                  -0.424901976                -0.145609079
##           HHIncome35000-44999          HHIncome45000-54999
##                  -0.213287919                -0.923212655
##           HHIncome55000-64999          HHIncome65000-74999
##                  -0.275094065                 0.275258406
##           HHIncome75000-99999           HHIncomemore 99999
##                  17.972045624                 1.071034164
##       Education9 - 11th Grade         EducationHigh School
##                   0.143389165                 9.224812852
##         EducationSome College        EducationCollege Grad
##                  -2.974114203                -2.019306283
## RegularMarijYes:HardDrugsYes
##                   2.044908730
```

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

```
m2.res = m2$residuals

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




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