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 ~ SmokeNow, df)
summary(model)
```

```
##
## Call:
## lm(formula = SexAge ~ SmokeNow, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.872 -1.872  0.070  1.128 21.128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.8724     0.0880 191.722  < 2e-16 ***
## SmokeNowYes  -0.9424     0.1241  -7.596 4.35e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.047 on 2411 degrees of freedom
##   (7587 observations deleted due to missingness)
## Multiple R-squared:  0.02337,    Adjusted R-squared:  0.02297
## F-statistic: 57.69 on 1 and 2411 DF,  p-value: 4.352e-14
```

```r
model <- lm(SexAge ~ AlcoholYear, df)
summary(model)
```

```
##
## Call:
## lm(formula = SexAge ~ AlcoholYear, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2924 -2.2326 -0.2855  1.7076 26.7105
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.2923935  0.0603297 286.632   <2e-16 ***
## AlcoholYear -0.0005747  0.0004852  -1.185    0.236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.424 on 5032 degrees of freedom
##   (4966 observations deleted due to missingness)
## Multiple R-squared:  0.0002788,  Adjusted R-squared:  8.014e-05
## F-statistic: 1.403 on 1 and 5032 DF,  p-value: 0.2362
```

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

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

```r
model |>
  tbl_regression(intercept = TRUE, show_single_row = c(RegularMarij, HardDrugs))|>
  as_gt() |>
  gt::tab_header(title = "SexAge MLR")
```

SexAge MLR

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	18	18, 18	<0.001
RegularMarij	-2.2	-2.5, -1.9	<0.001
HardDrugs	-1.7	-2.1, -1.3	<0.001
RegularMarij * HardDrugs			
Yes * Yes	1.4	0.90, 2.0	<0.001

[1] CI = Confidence Interval

```
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
```

```
model |>
  tbl_regression(intercept = TRUE, show_single_row = c(RegularMarij, HardDrugs))|>
  as_gt() |>
  gt::tab_header(title = "SexNumPartnLife MLR")
```

SexNumPartnLife MLR

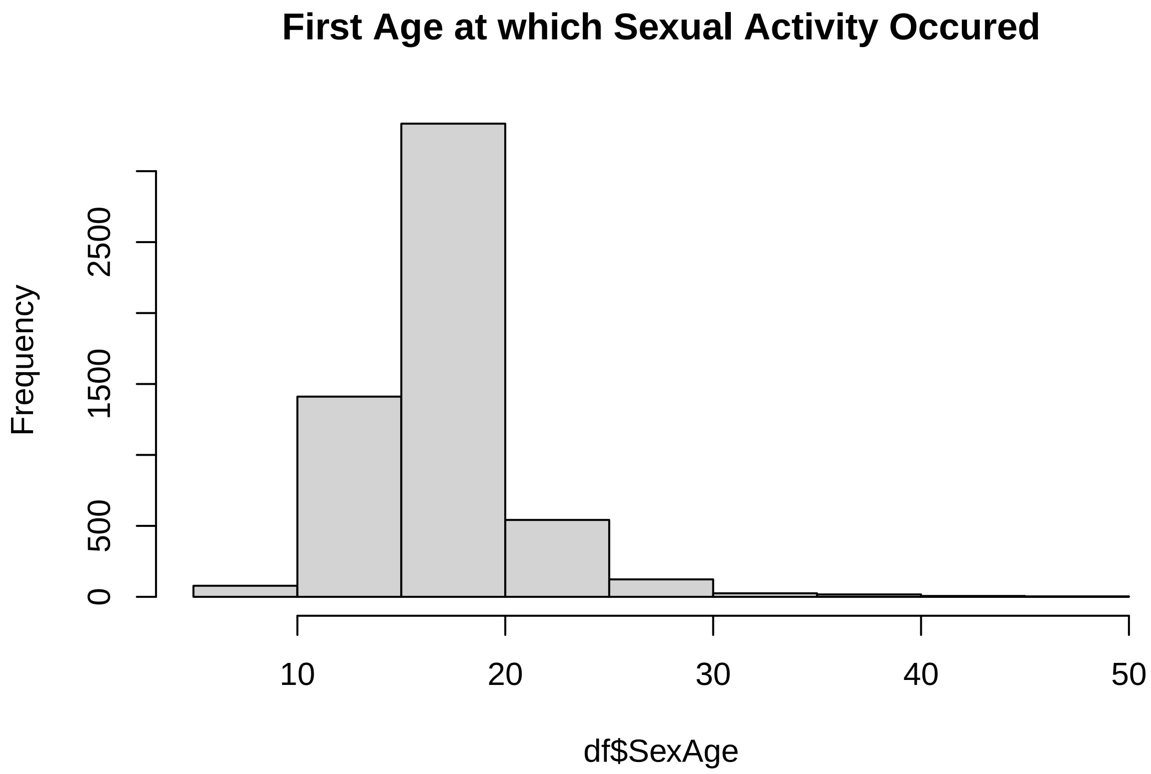
Characteristic	Beta	95% CI[1]	p-value
(Intercept)	8.4	6.3, 10	<0.001
RegularMarij	15	9.8, 20	<0.001
HardDrugs	14	6.5, 21	<0.001
RegularMarij * HardDrugs			
Yes * Yes	0.82	-8.7, 10	0.9

[1] CI = Confidence Interval

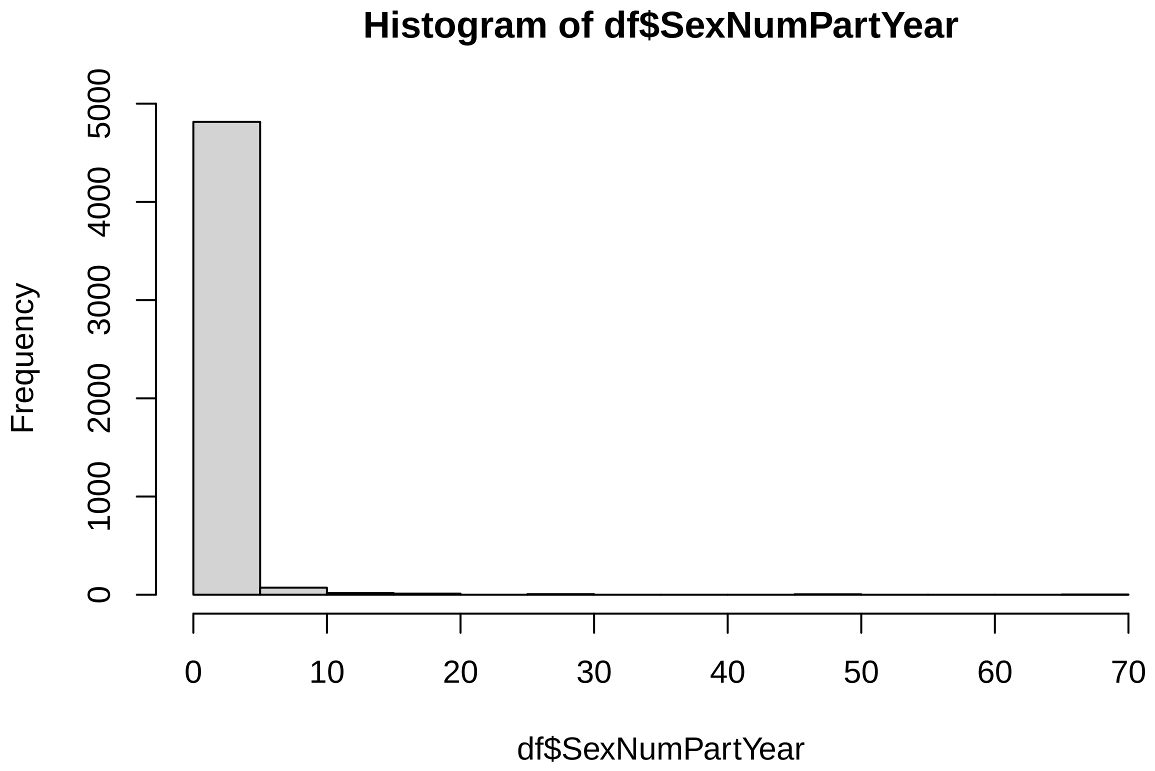
SexAge is has a good distribution but SexNumPartnLife has extreme skenwness and is discrete count data. This requires a Poisson regression which is out side the scopre of this course. Created new variable using the duration, since first sexual activity where (Age - SexAge) since Age >= SexAge, and dividing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

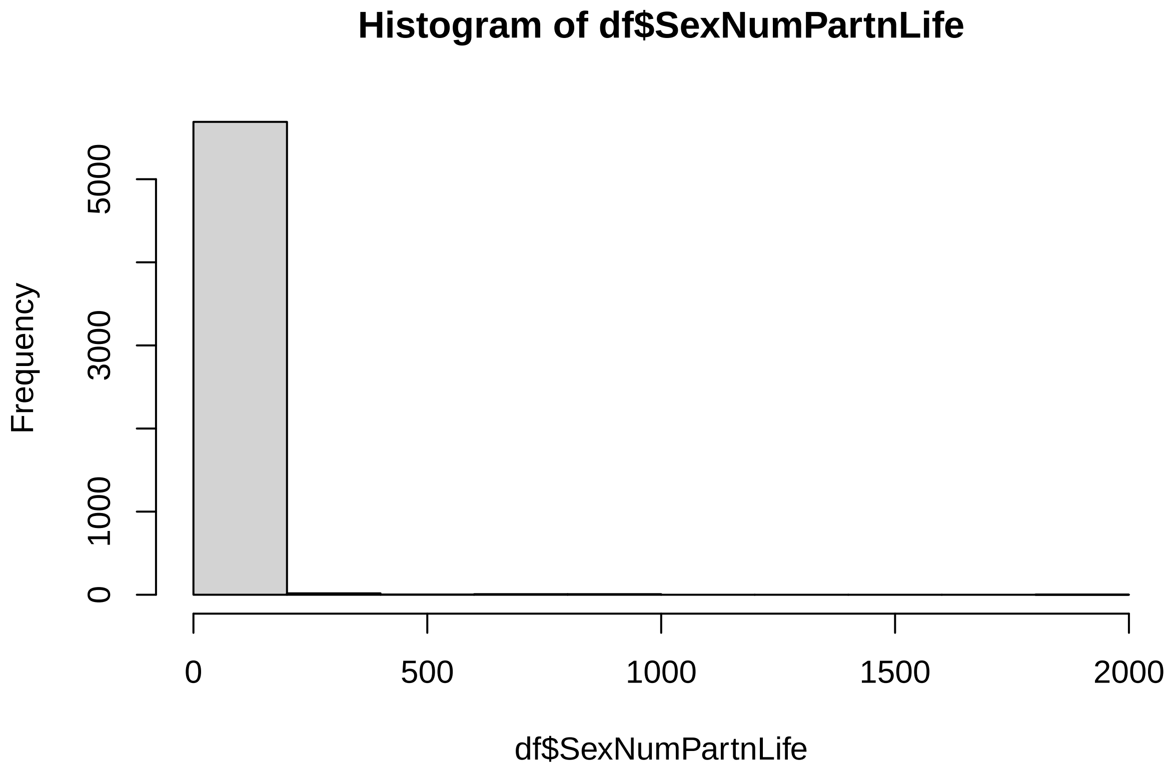
```
hist(df$SexAge, main= "First Age at which Sexual Activity Occured")
```



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



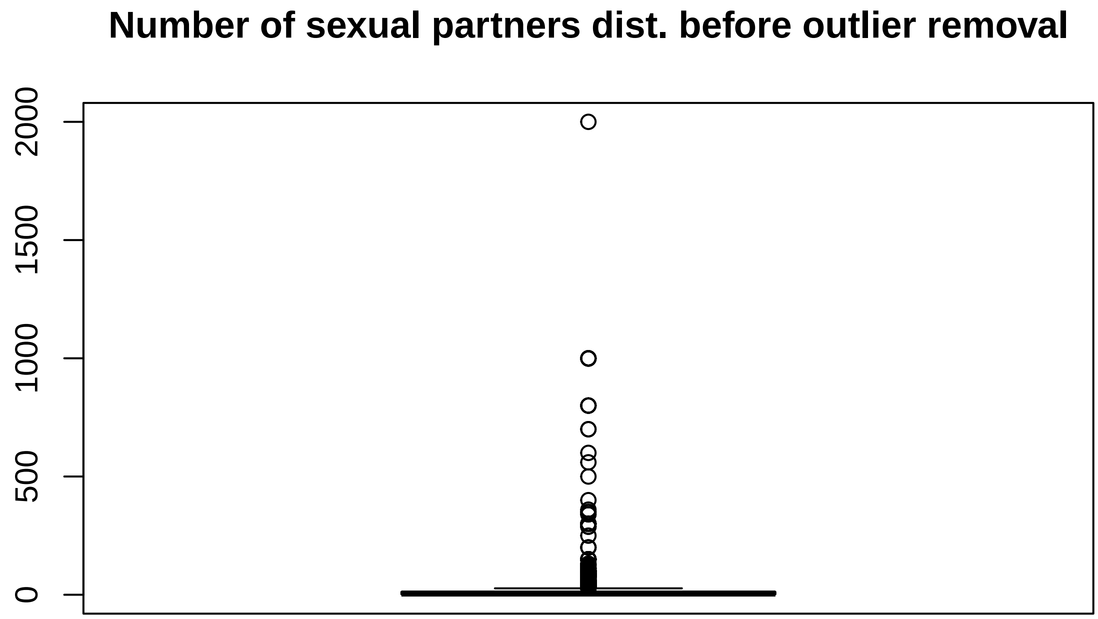
```
hist(df$SexNumPartnLife)
```



```
#Show observations for which SexAge > Age, None
df[which(df$SexAge > df$Age), ]
```

```
## # A tibble: 0 x 76
## # i 76 variables: ID <int>, SurveyYr <fct>, Gender <fct>, Age <int>,
## #   AgeDecade <fct>, AgeMonths <int>, Race1 <fct>, Race3 <fct>,
## #   Education <fct>, 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>, ...
```

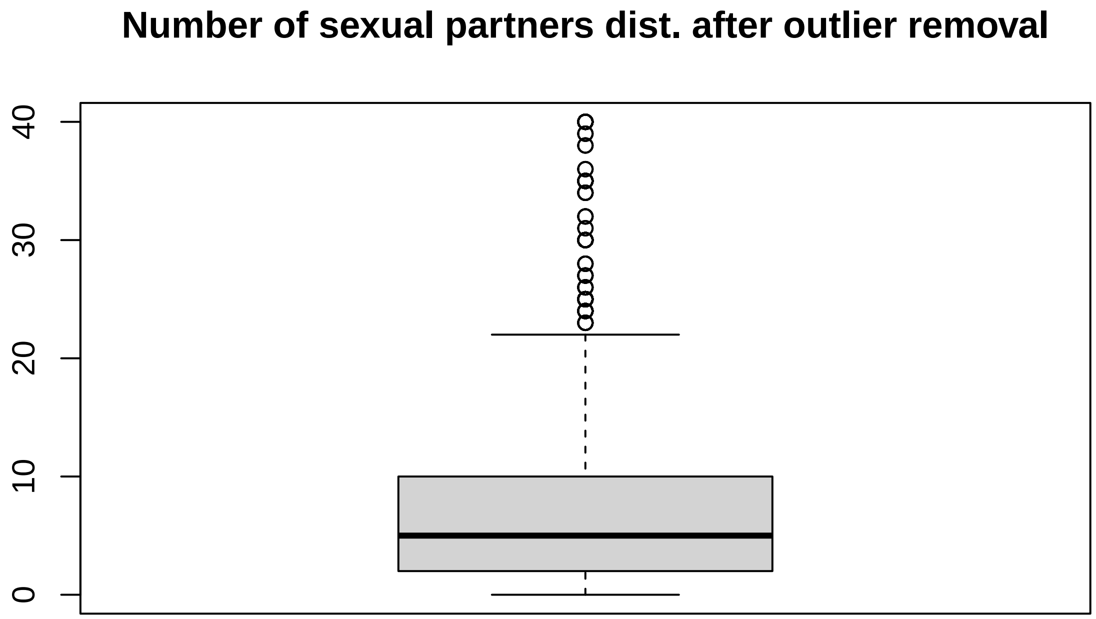
```
#Show observations with more than 40 sexual partners during lifetime
boxplot(df$SexNumPartnLife, main = "Number of sexual partners dist. before outlier removal")
```



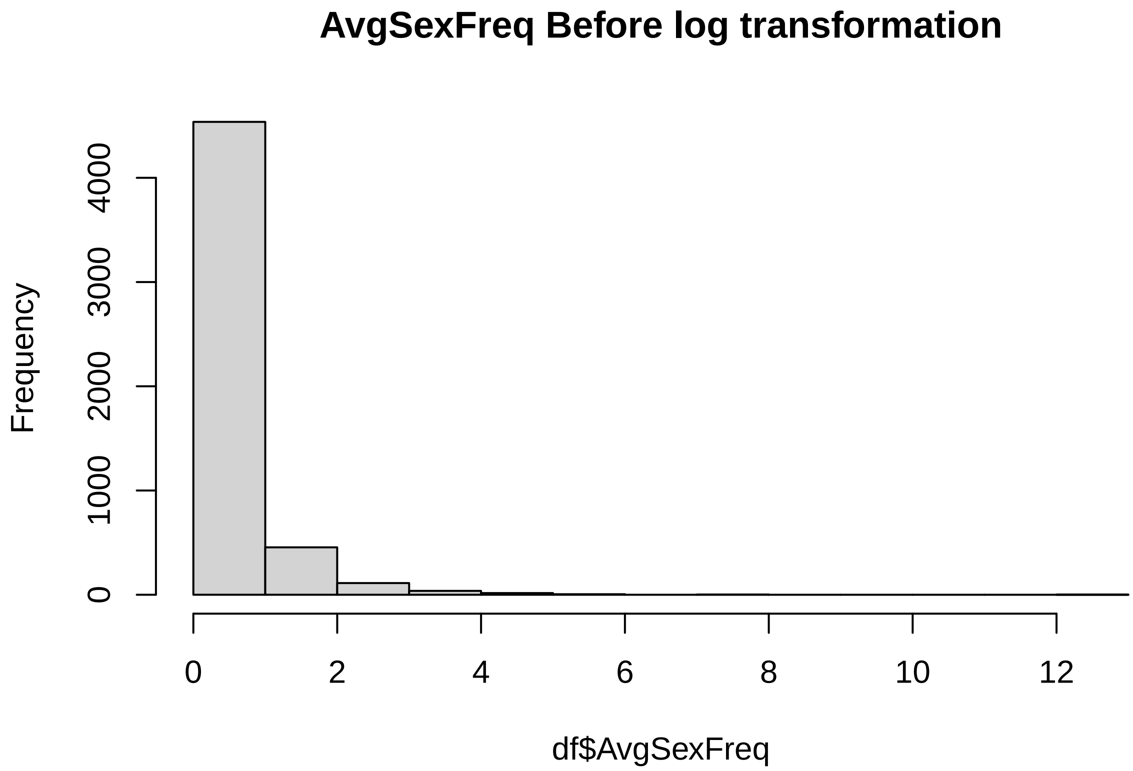
```r
df[which(df$SexNumPartnLife > 40), c("Age", "SexAge", "SexNumPartnLife")]
```

```
## # A tibble: 318 x 3
##      Age SexAge SexNumPartnLife
##    <int>  <int>           <int>
##  1    54     12             100
##  2    56     20              90
##  3    36     16              45
##  4    47     19              45
##  5    61     15             288
##  6    61     15             288
##  7    61     15             288
##  8    42     18              65
##  9    42     18              65
## 10    45     15              50
## # i 308 more rows
```

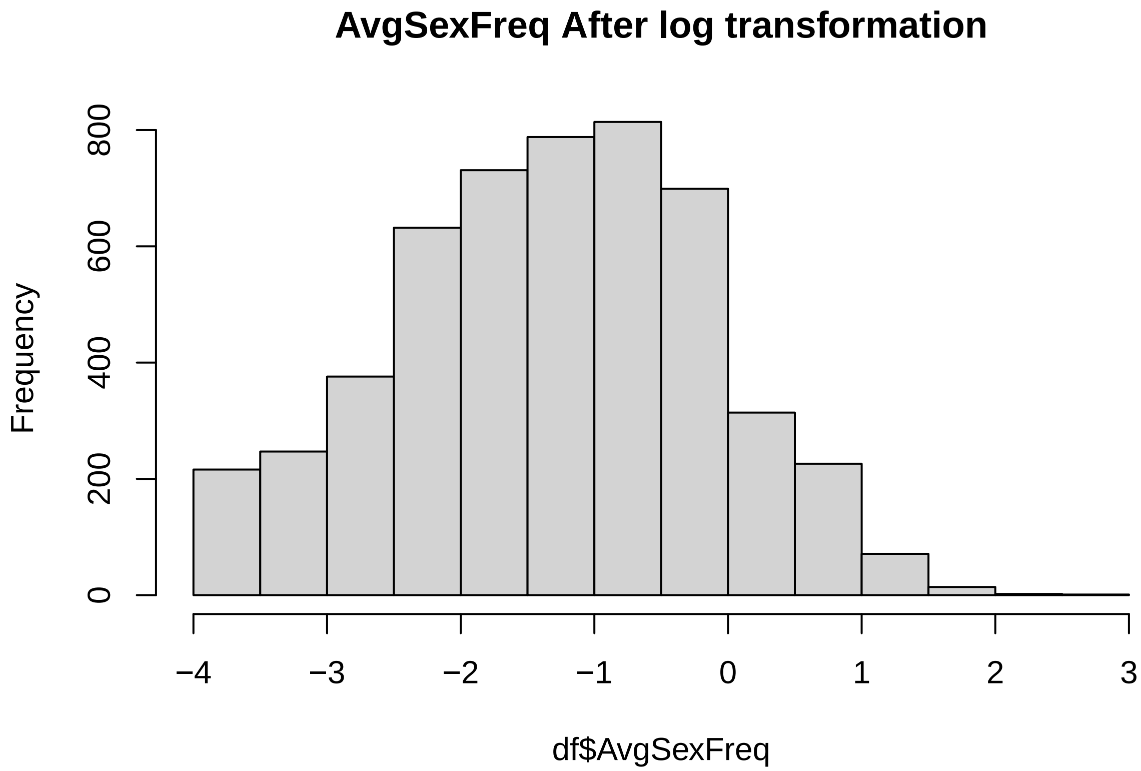
```r
#Remove observations with more than 40 sexual partners during lifetime
df = df[-which(df$SexNumPartnLife > 40),]
boxplot(df$SexNumPartnLife, main = "Number of sexual partners dist. after outlier removal")
```



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



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



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

```
## 4234 missing rows in the "HardDrugs" column have been removed.
```

Characteristic	**No** N = 4,538[1]	**Yes** N = 910[1]
ID	61,879 (5,889)	62,174 (5,956)
SurveyYr		
2009_10	2,286 / 4538 (50%)	442 / 910 (49%)
2011_12	2,252 / 4538 (50%)	468 / 910 (51%)
Gender		
female	2,365 / 4538 (52%)	362 / 910 (40%)
male	2,173 / 4538 (48%)	548 / 910 (60%)
Age	42 (15)	43 (13)
AgeDecade		
0-9	0 / 4538 (0%)	0 / 910 (0%)
10-19	207 / 4538 (4.6%)	22 / 910 (2.4%)
20-29	984 / 4538 (22%)	156 / 910 (17%)
30-39	926 / 4538 (20%)	146 / 910 (16%)
40-49	895 / 4538 (20%)	246 / 910 (27%)

50-59	821 / 4538 (18%)	270 / 910 (30%)
60-69	705 / 4538 (16%)	70 / 910 (7.7%)
70+	0 / 4538 (0%)	0 / 910 (0%)
AgeMonths	507 (175)	497 (144)
Unknown	2,252	468
Race1		
Black	508 / 4538 (11%)	75 / 910 (8.2%)
Hispanic	285 / 4538 (6.3%)	25 / 910 (2.7%)
Mexican	439 / 4538 (9.7%)	66 / 910 (7.3%)
White	2,938 / 4538 (65%)	696 / 910 (76%)
Other	368 / 4538 (8.1%)	48 / 910 (5.3%)
Race3		
Asian	140 / 2252 (6.2%)	9 / 468 (1.9%)
Black	261 / 2252 (12%)	29 / 468 (6.2%)
Hispanic	150 / 2252 (6.7%)	11 / 468 (2.4%)
Mexican	207 / 2252 (9.2%)	28 / 468 (6.0%)
White	1,445 / 2252 (64%)	369 / 468 (79%)
Other	49 / 2252 (2.2%)	22 / 468 (4.7%)
Unknown	2,286	442
Education		
8th Grade	197 / 4322 (4.6%)	17 / 888 (1.9%)
9 - 11th Grade	428 / 4322 (9.9%)	142 / 888 (16%)
High School	866 / 4322 (20%)	199 / 888 (22%)
Some College	1,356 / 4322 (31%)	338 / 888 (38%)
College Grad	1,475 / 4322 (34%)	192 / 888 (22%)
Unknown	216	22
MaritalStatus		
Divorced	373 / 4330 (8.6%)	134 / 887 (15%)
LivePartner	333 / 4330 (7.7%)	130 / 887 (15%)
Married	2,489 / 4330 (57%)	394 / 887 (44%)
NeverMarried	912 / 4330 (21%)	180 / 887 (20%)
Separated	110 / 4330 (2.5%)	32 / 887 (3.6%)
Widowed	113 / 4330 (2.6%)	17 / 887 (1.9%)
Unknown	208	23
HHIncome		
0-4999	64 / 4217 (1.5%)	21 / 850 (2.5%)
5000-9999	84 / 4217 (2.0%)	21 / 850 (2.5%)
10000-14999	225 / 4217 (5.3%)	45 / 850 (5.3%)
15000-19999	190 / 4217 (4.5%)	52 / 850 (6.1%)
20000-24999	231 / 4217 (5.5%)	42 / 850 (4.9%)
25000-34999	384 / 4217 (9.1%)	103 / 850 (12%)
35000-44999	388 / 4217 (9.2%)	66 / 850 (7.8%)
45000-54999	365 / 4217 (8.7%)	80 / 850 (9.4%)
55000-64999	327 / 4217 (7.8%)	51 / 850 (6.0%)
65000-74999	273 / 4217 (6.5%)	59 / 850 (6.9%)

75000-99999	551 / 4217 (13%)	91 / 850 (11%)
more 99999	1,135 / 4217 (27%)	219 / 850 (26%)
Unknown	321	60
HHIncomeMid	61,147 (32,344)	58,129 (33,116)
Unknown	321	60
Poverty	3.06 (1.67)	2.82 (1.69)
Unknown	275	58
HomeRooms	6 (2)	6 (2)
Unknown	25	5
HomeOwn		
Own	3,016 / 4513 (67%)	522 / 905 (58%)
Rent	1,401 / 4513 (31%)	357 / 905 (39%)
Other	96 / 4513 (2.1%)	26 / 905 (2.9%)
Unknown	25	5
Work		
Looking	178 / 4537 (3.9%)	73 / 910 (8.0%)
NotWorking	1,214 / 4537 (27%)	238 / 910 (26%)
Working	3,145 / 4537 (69%)	599 / 910 (66%)
Unknown	1	0
Weight	83 (22)	84 (20)
Unknown	29	1
Length	NA (NA)	NA (NA)
Unknown	4,538	910
HeadCirc	NA (NA)	NA (NA)
Unknown	4,538	910
Height	169 (10)	172 (9)
Unknown	21	1
BMI	29 (7)	28 (6)
Unknown	29	1
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	4,435	903
BMI_WHO		
12.0_18.5	92 / 4492 (2.0%)	9 / 904 (1.0%)
18.5_to_24.9	1,306 / 4492 (29%)	277 / 904 (31%)
25.0_to_29.9	1,444 / 4492 (32%)	311 / 904 (34%)
30.0_plus	1,650 / 4492 (37%)	307 / 904 (34%)
Unknown	46	6
Pulse	73 (12)	72 (11)
Unknown	68	13
BPSysAve	118 (15)	120 (16)
Unknown	73	15

BPDiaAve	70 (12)	72 (11)
Unknown	73	15
BPSys1	119 (15)	120 (16)
Unknown	227	39
BPDia1	71 (12)	73 (11)
Unknown	227	39
BPSys2	119 (15)	120 (17)
Unknown	168	19
BPDia2	70 (12)	73 (11)
Unknown	168	19
BPSys3	118 (15)	120 (16)
Unknown	153	19
BPDia3	70 (12)	72 (12)
Unknown	153	19
Testosterone	215 (228)	245 (250)
Unknown	2,423	470
DirectChol	1.36 (0.41)	1.38 (0.42)
Unknown	188	26
TotChol	5.04 (1.04)	5.25 (1.15)
Unknown	188	26
UrineVol1	126 (94)	133 (94)
Unknown	14	1
UrineFlow1	1.07 (0.98)	1.07 (1.04)
Unknown	240	56
UrineVol2	131 (94)	114 (81)
Unknown	3,802	800
UrineFlow2	1.23 (1.13)	1.10 (1.14)
Unknown	3,804	800
Diabetes	342 / 4536 (7.5%)	75 / 910 (8.2%)
Unknown	2	0
DiabetesAge	46 (13)	43 (13)
Unknown	4,261	852
HealthGen		
Excellent	575 / 4538 (13%)	72 / 904 (8.0%)
Vgood	1,531 / 4538 (34%)	281 / 904 (31%)
Good	1,771 / 4538 (39%)	384 / 904 (42%)
Fair	568 / 4538 (13%)	141 / 904 (16%)
Poor	93 / 4538 (2.0%)	26 / 904 (2.9%)
Unknown	0	6
DaysPhysHlthBad	3 (7)	4 (8)
Unknown	0	6
DaysMentHlthBad	4 (8)	6 (9)
Unknown	1	6
LittleInterest		
None	3,542 / 4536 (78%)	613 / 899 (68%)

Several	741 / 4536 (16%)	186 / 899 (21%)
Most	253 / 4536 (5.6%)	100 / 899 (11%)
Unknown	2	11
Depressed		
None	3,673 / 4538 (81%)	599 / 904 (66%)
Several	626 / 4538 (14%)	208 / 904 (23%)
Most	239 / 4538 (5.3%)	97 / 904 (11%)
Unknown	0	6
nPregnancies	3 (2)	3 (2)
Unknown	2,763	614
nBabies	2 (1)	2 (1)
Unknown	2,885	644
Age1stBaby	23 (5)	23 (5)
Unknown	3,269	729
SleepHrsNight	7 (1)	7 (1)
Unknown	6	5
SleepTrouble	1,028 / 4538 (23%)	365 / 910 (40%)
PhysActive	2,617 / 4538 (58%)	462 / 910 (51%)
PhysActiveDays		
1	279 / 2388 (12%)	43 / 424 (10%)
2	419 / 2388 (18%)	90 / 424 (21%)
3	577 / 2388 (24%)	109 / 424 (26%)
4	296 / 2388 (12%)	62 / 424 (15%)
5	403 / 2388 (17%)	72 / 424 (17%)
6	127 / 2388 (5.3%)	13 / 424 (3.1%)
7	287 / 2388 (12%)	35 / 424 (8.3%)
Unknown	2,150	486
TVHrsDay		
0_hrs	46 / 2251 (2.0%)	16 / 468 (3.4%)
0_to_1_hr	318 / 2251 (14%)	55 / 468 (12%)
1_hr	416 / 2251 (18%)	85 / 468 (18%)
2_hr	582 / 2251 (26%)	128 / 468 (27%)
3_hr	391 / 2251 (17%)	75 / 468 (16%)
4_hr	232 / 2251 (10%)	40 / 468 (8.5%)
More_4_hr	266 / 2251 (12%)	69 / 468 (15%)
Unknown	2,287	442
CompHrsDay		
0_hrs	375 / 2252 (17%)	91 / 468 (19%)
0_to_1_hr	609 / 2252 (27%)	167 / 468 (36%)
1_hr	533 / 2252 (24%)	94 / 468 (20%)
2_hr	306 / 2252 (14%)	48 / 468 (10%)
3_hr	166 / 2252 (7.4%)	27 / 468 (5.8%)
4_hr	109 / 2252 (4.8%)	14 / 468 (3.0%)
More_4_hr	154 / 2252 (6.8%)	27 / 468 (5.8%)
Unknown	2,286	442

TVHrsDayChild	NA (NA)	NA (NA)
Unknown	4,538	910
CompHrsDayChild	NA (NA)	NA (NA)
Unknown	4,538	910
Alcohol12PlusYr	3,454 / 4436 (78%)	847 / 890 (95%)
Unknown	102	20
AlcoholDay	3 (3)	4 (3)
Unknown	1,160	142
AlcoholYear	66 (95)	103 (112)
Unknown	566	26
SmokeNow	722 / 1519 (48%)	374 / 686 (55%)
Unknown	3,019	224
Smoke100	1,519 / 4331 (35%)	686 / 888 (77%)
Unknown	207	22
Smoke100n		
Non-Smoker	2,812 / 4331 (65%)	202 / 888 (23%)
Smoker	1,519 / 4331 (35%)	686 / 888 (77%)
Unknown	207	22
SmokeAge	18 (4)	17 (5)
Unknown	3,077	251
Marijuana	1,847 / 3828 (48%)	809 / 840 (96%)
Unknown	710	70
AgeFirstMarij	18 (4)	16 (4)
Unknown	2,692	101
RegularMarij	617 / 3828 (16%)	569 / 840 (68%)
Unknown	710	70
AgeRegMarij	18 (4)	18 (5)
Unknown	3,921	341
SexEver	4,306 / 4528 (95%)	910 / 910 (100%)
Unknown	10	0
SexAge	18 (4)	16 (3)
Unknown	236	0
SexNumPartnLife	7 (7)	14 (10)
Unknown	44	4
SexNumPartYear	1 (2)	1 (2)
Unknown	724	70
SameSex	204 / 4529 (4.5%)	168 / 910 (18%)
Unknown	9	0
SexOrientation		
Bisexual	70 / 3745 (1.9%)	42 / 827 (5.1%)
Heterosexual	3,625 / 3745 (97%)	758 / 827 (92%)
Homosexual	50 / 3745 (1.3%)	27 / 827 (3.3%)
Unknown	793	83
PregnantNow		
Yes	59 / 1198 (4.9%)	1 / 155 (0.6%)

No	1,114 / 1198 (93%)	154 / 155 (99%)
Unknown	25 / 1198 (2.1%)	0 / 155 (0%)
Unknown	3,340	755
AvgSexFreq	NA (NA)	-Inf (NA)
Unknown	269	4

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

$$AvgSexFreq = \log\left(\frac{SexNumPartnLife}{Age - SexAge}\right)$$

```
#Remove negative infinity from numerator(NumPartnLife) or Age-SexAge being 0 for and change to zero.
obs = df[is.infinite(df$AvgSexFreq),]
obs[, c("Age","SexAge", "SexNumPartnLife")]
```

```
## # A tibble: 50 x 3
##      Age SexAge SexNumPartnLife
##    <int>  <int>           <int>
##  1    29     29               1
##  2    29     29               1
##  3    29     29               1
##  4    29     29               1
##  5    52     19               0
##  6    28     16               0
##  7    23     23               1
##  8    18     18               1
##  9    26     26               1
## 10    24     14               0
## # i 40 more rows
```

```
df$AvgSexFreq[is.infinite(df$AvgSexFreq)] = 0
#unique(df$AvgSexFreq)

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
## -0.70378 -0.19899 -0.01532  0.11520  0.91187
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2.046433   1.366427  -1.498 0.144672
## SmokeNowYes                  -0.081948   0.366994  -0.223 0.824820
## AlcoholYear                   0.001392   0.001289   1.080 0.288895
## RegularMarijYes              -0.491636   0.255418  -1.925 0.063780 .
## HardDrugsYes                  1.332538   0.429427   3.103 0.004152 **
## Age                           -0.005915   0.021050  -0.281 0.780634
## Gendermale                     0.879161   0.210608   4.174 0.000236 ***
## HHIncome 5000-9999             0.776899   0.486066   1.598 0.120448
## HHIncome10000-14999            0.363545   0.552703   0.658 0.515709
## HHIncome15000-19999            0.402849   0.661642   0.609 0.547199
## HHIncome20000-24999            0.323434   0.485667   0.666 0.510526
## HHIncome25000-34999            0.478661   0.457666   1.046 0.303974
## HHIncome35000-44999            0.535294   0.418417   1.279 0.210587
## HHIncome45000-54999            1.602928   0.747565   2.144 0.040240 *
## HHIncome55000-64999           -0.090747   0.451143  -0.201 0.841940
## HHIncome65000-74999            0.967943   0.411045   2.355 0.025269 *
## HHIncome75000-99999           -0.713722   0.488594  -1.461 0.154475
## HHIncomemore 99999            -0.033470   0.455944  -0.073 0.941968
## Education9 - 11th Grade        0.095578   0.506340   0.189 0.851550
## EducationHigh School           0.554184   0.479747   1.155 0.257144
## EducationSome College          0.343311   0.456860   0.751 0.458232
## EducationCollege Grad         -1.104466   0.534455  -2.067 0.047505 *
## BMI                            0.005673   0.018508   0.306 0.761351
## DiabetesAge                   -0.001027   0.011525  -0.089 0.929550
## DepressedSeveral               0.589864   0.324888   1.816 0.079441 .
## DepressedMost                  0.089268   0.394015   0.227 0.822303
## LittleInterestSeveral         -0.439828   0.289344  -1.520 0.138960
## LittleInterestMost            -0.609118   0.385088  -1.582 0.124191
## PhysActiveYes                  0.040494   0.362289   0.112 0.911748
## SameSexYes                    -0.065757   0.481588  -0.137 0.892306
## RegularMarijYes:HardDrugsYes -1.367908   0.556269  -2.459 0.019916 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4629 on 30 degrees of freedom
##   (9621 observations deleted due to missingness)
## Multiple R-squared:  0.832,  Adjusted R-squared:  0.6641
## F-statistic: 4.954 on 30 and 30 DF,  p-value: 1.706e-05
```

```r
model |>
  tbl_regression(intercept = TRUE,show_single_row = c(RegularMarij, HardDrugs,Gender, PhysActive, SameS
  as_gt() |>
  gt::tab_header(title = "Full model")
```

Full model

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	-2.0	-4.8, 0.74	0.14
SmokeNow			
No	—	—	
Yes	-0.08	-0.83, 0.67	0.8
AlcoholYear	0.00	0.00, 0.00	0.3
RegularMarij	-0.49	-1.0, 0.03	0.064
HardDrugs	1.3	0.46, 2.2	0.004
Age	-0.01	-0.05, 0.04	0.8
Gender	0.88	0.45, 1.3	<0.001
HHIncome			

0-4999	—	—	
5000-9999	0.78	-0.22, 1.8	0.12
10000-14999	0.36	-0.77, 1.5	0.5
15000-19999	0.40	-0.95, 1.8	0.5
20000-24999	0.32	-0.67, 1.3	0.5
25000-34999	0.48	-0.46, 1.4	0.3
35000-44999	0.54	-0.32, 1.4	0.2
45000-54999	1.6	0.08, 3.1	0.040
55000-64999	-0.09	-1.0, 0.83	0.8
65000-74999	0.97	0.13, 1.8	0.025
75000-99999	-0.71	-1.7, 0.28	0.2
more 99999	-0.03	-0.96, 0.90	>0.9
Education			
8th Grade	—	—	
9 - 11th Grade	0.10	-0.94, 1.1	0.9
High School	0.55	-0.43, 1.5	0.3
Some College	0.34	-0.59, 1.3	0.5
College Grad	-1.1	-2.2, -0.01	0.048
BMI	0.01	-0.03, 0.04	0.8
DiabetesAge	0.00	-0.02, 0.02	>0.9
Depressed			
None	—	—	
Several	0.59	-0.07, 1.3	0.079
Most	0.09	-0.72, 0.89	0.8
LittleInterest			
None	—	—	
Several	-0.44	-1.0, 0.15	0.14
Most	-0.61	-1.4, 0.18	0.12
PhysActive	0.04	-0.70, 0.78	>0.9
SameSex	-0.07	-1.0, 0.92	0.9
RegularMarij * HardDrugs			
Yes * Yes	-1.4	-2.5, -0.23	0.020

[1] CI = Confidence Interval

Using the sequential sum of squares we tested for each block of covariates at a significance level 0.001.

```
n = 30
aov = anova(model <- lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs
aov
```

```
## Analysis of Variance Table
##
## Response: AvgSexFreq
##                         Df Sum Sq Mean Sq F value    Pr(>F)
## SmokeNow                 1 0.6174  0.6174  2.8811  0.099975 .
## AlcoholYear              1 0.8397  0.8397  3.9187  0.056992 .
## RegularMarij             1 1.2551  1.2551  5.8574  0.021777 *
## HardDrugs                1 1.6838  1.6838  7.8579  0.008784 **
## Age                      1 8.4943  8.4943 39.6406 6.113e-07 ***
## Gender                   1 3.1811  3.1811 14.8454  0.000571 ***
## HHIncome                11 7.4549  0.6777  3.1628  0.005990 **
## Education                4 3.1691  0.7923  3.6973  0.014586 *
## BMI                      1 0.0581  0.0581  0.2713  0.606307
## DiabetesAge              1 0.0287  0.0287  0.1341  0.716830
## Depressed                2 2.8144  1.4072  6.5671  0.004310 **
## LittleInterest           2 0.8567  0.4283  1.9990  0.153115
## PhysActive               1 0.0878  0.0878  0.4097  0.526967
## SameSex                  1 0.0067  0.0067  0.0312  0.860945
## RegularMarij:HardDrugs  1 1.2958  1.2958  6.0471  0.019916 *
## Residuals               30 6.4285  0.2143
## ---
## 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] 5.691796
```

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

```
## [1] 5.312440703 0.006065241
```

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

```
## [1] 22.29934
```

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

```
## [1] 2.601638e+01 2.363474e-06
```

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

```
## [1] 3.852443
```

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

```
## [1] 3.59568007 0.02665494
```

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

```
## [1] 0.006689209
```

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

```
print(c(fstat4, pval4))
```

```
## [1] 1.797840e+01 8.238241e-04
```

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

Step	Tested Var.	SS(Num.)	SS(Denom.)	Test Stat.	Dist.	p-value	Decision	Stopping Rule	Decision
I	$\boldsymbol{\beta}_{Substance}$	13.88444	26.9329	5.155204576	$F_{5,14}$	0.001262146	Reject	Do not stop	Collect
II	$\boldsymbol{\beta}_{Demo}$	55.61473	26.9329	25.81174	$F_{4,14}$	6.872507e-10	Reject	Do not stop	Collect
III	$\boldsymbol{\beta}_{Health}$	5.687399	26.9329	2.11169493	$F_{5,14}$	0.08788892	Fail to Reject	Stop	Not Collect
IV	$\boldsymbol{\beta}_{SameSex}$	0.00170849	26.9329	10.55847467	$F_{1,14}$	0.00260712	NA	NA	NA

Final model

$$AvgSexFreq = X_{Substance}\boldsymbol{\beta}_{Substance} + X_{Demo}\boldsymbol{\beta}_{Demo} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$$

```
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] 1639
```

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

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

Characteristic	**Missing** N = 8,043[1]	**Not Missing** N = 1,639[1]
Age	36 (24)	41 (11)
Gender		
female	4,235 / 8043 (53%)	730 / 1639 (45%)
male	3,808 / 8043 (47%)	909 / 1639 (55%)
HHIncome		
0-4999	152 / 7266 (2.1%)	36 / 1639 (2.2%)
5000-9999	204 / 7266 (2.8%)	36 / 1639 (2.2%)
10000-14999	425 / 7266 (5.8%)	99 / 1639 (6.0%)
15000-19999	410 / 7266 (5.6%)	98 / 1639 (6.0%)
20000-24999	469 / 7266 (6.5%)	117 / 1639 (7.1%)
25000-34999	757 / 7266 (10%)	165 / 1639 (10%)

35000-44999	709 / 7266 (9.8%)	133 / 1639 (8.1%)
45000-54999	611 / 7266 (8.4%)	154 / 1639 (9.4%)
55000-64999	483 / 7266 (6.6%)	125 / 1639 (7.6%)
65000-74999	407 / 7266 (5.6%)	107 / 1639 (6.5%)
75000-99999	841 / 7266 (12%)	204 / 1639 (12%)
more 99999	1,798 / 7266 (25%)	365 / 1639 (22%)
Unknown	777	0
Education		
8th Grade	372 / 5267 (7.1%)	64 / 1639 (3.9%)
9 - 11th Grade	554 / 5267 (11%)	277 / 1639 (17%)
High School	1,044 / 5267 (20%)	399 / 1639 (24%)
Some College	1,612 / 5267 (31%)	541 / 1639 (33%)
College Grad	1,685 / 5267 (32%)	358 / 1639 (22%)
Unknown	2,776	0
MaritalStatus		
Divorced	453 / 5279 (8.6%)	191 / 1637 (12%)
LivePartner	273 / 5279 (5.2%)	247 / 1637 (15%)
Married	3,022 / 5279 (57%)	798 / 1637 (49%)
NeverMarried	970 / 5279 (18%)	335 / 1637 (20%)
Separated	134 / 5279 (2.5%)	45 / 1637 (2.7%)
Widowed	427 / 5279 (8.1%)	21 / 1637 (1.3%)
Unknown	2,764	2

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

```r
missingness_comparison = glm(as.factor(missingness)~Age+Gender+HHIncome+Education+MaritalStatus, family
missingness_comparison |>
  tbl_regression(intercept = TRUE)|>
  as_gt() |>
  gt::tab_header(title = "Missingness Comparison")
```

Missingness Comparison

Characteristic	**log(OR)**[1]	**95% CI**[1]	**p-value**
(Intercept)	0.50	-0.09, 1.1	0.092
Age	-0.03	-0.04, -0.03	<0.001
Gender			
female	—	—	
male	0.38	0.26, 0.50	<0.001
HHIncome			
0-4999	—	—	
5000-9999	-0.30	-0.89, 0.29	0.3
10000-14999	-0.17	-0.66, 0.32	0.5
15000-19999	-0.11	-0.60, 0.39	0.7
20000-24999	0.09	-0.39, 0.58	0.7
25000-34999	-0.29	-0.75, 0.18	0.2

35000-44999	-0.43	-0.89, 0.05	0.077
45000-54999	-0.04	-0.51, 0.44	0.9
55000-64999	0.03	-0.44, 0.52	0.9
65000-74999	0.07	-0.41, 0.57	0.8
75000-99999	0.07	-0.39, 0.54	0.8
more 99999	-0.02	-0.46, 0.44	>0.9
Education			
8th Grade	—	—	
9 - 11th Grade	0.89	0.57, 1.2	<0.001
High School	0.53	0.22, 0.85	0.001
Some College	0.29	-0.02, 0.61	0.072
College Grad	-0.17	-0.49, 0.17	0.3
MaritalStatus			
Divorced	—	—	
LivePartner	0.43	0.16, 0.70	0.002
Married	-0.61	-0.81, -0.40	<0.001
NeverMarried	-0.86	-1.1, -0.62	<0.001
Separated	-0.42	-0.83, -0.02	0.043
Widowed	-1.6	-2.1, -1.1	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

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

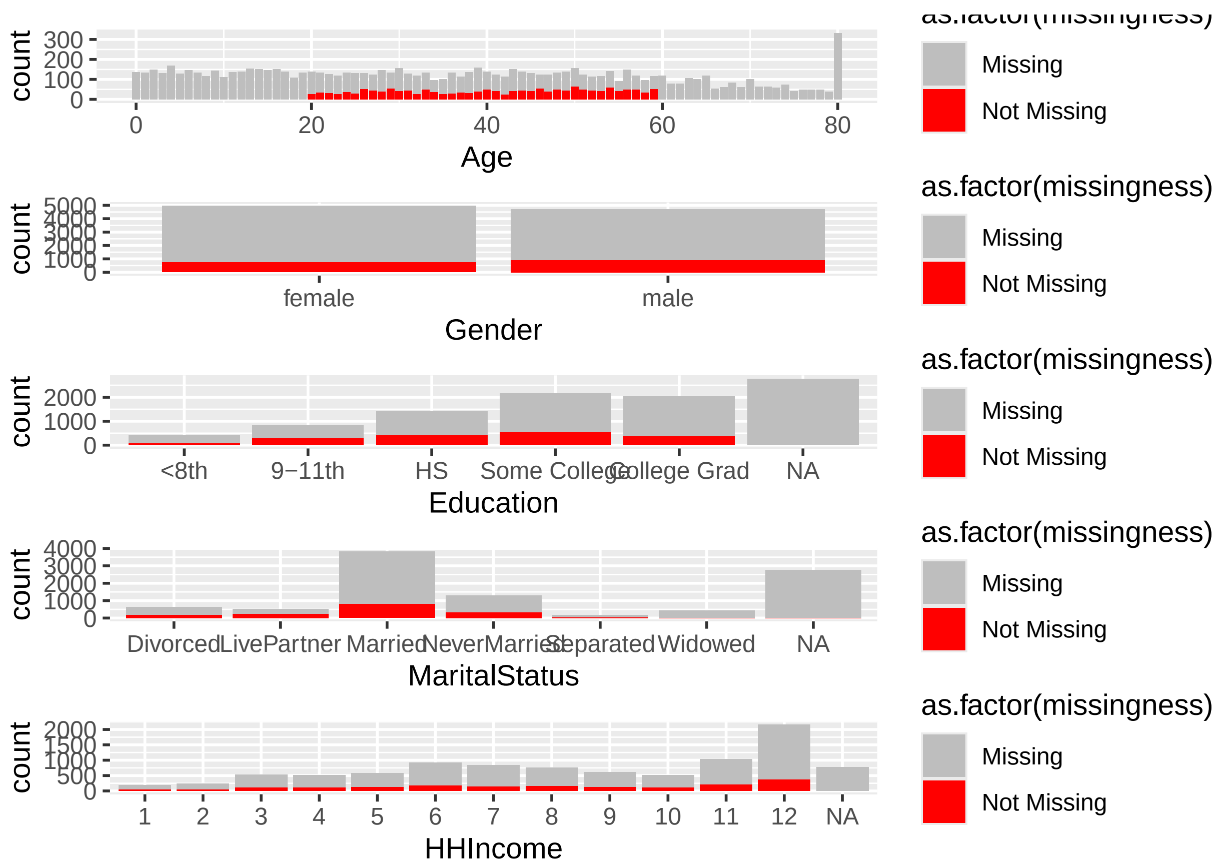
Missingness for occurs for those aged below 20 because they are not recorded for some covariates. Why missingness for those aged above 60 occurs is unclear.

```
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
summary(m1)
```

```
##
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + HHIncome +
##     Education, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5442 -0.4806  0.0178  0.5349  2.2544
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.4989789  0.1776065   2.809  0.00502 **
## SmokeNowYes                 0.1232020  0.0403154   3.056  0.00228 **
## AlcoholYear                 0.0009830  0.0001831   5.367 9.15e-08 ***
## RegularMarijYes             0.3926936  0.0490751   8.002 2.32e-15 ***
## HardDrugsYes                   0.5239528  0.0675210   7.760 1.50e-14 ***
## Age                           -0.0498245  0.0017127 -29.091  < 2e-16 ***
## Gendermale                     0.1939702  0.0382140   5.076 4.30e-07 ***
## HHIncome 5000-9999             0.3091670  0.1775244   1.742  0.08178 .
## HHIncome10000-14999            0.2968935  0.1462576   2.030  0.04253 *
## HHIncome15000-19999            0.1435646  0.1467499   0.978  0.32808
## HHIncome20000-24999            0.1097613  0.1435960   0.764  0.44475
## HHIncome25000-34999            0.0373360  0.1384265   0.270  0.78741
## HHIncome35000-44999            0.0618094  0.1416846   0.436  0.66272
## HHIncome45000-54999            0.0783545  0.1399578   0.560  0.57566
## HHIncome55000-64999            0.2258547  0.1425091   1.585  0.11320
## HHIncome65000-74999           -0.0910084  0.1457700  -0.624  0.53250
## HHIncome75000-99999           -0.0161207  0.1369970  -0.118  0.90634
## HHIncomemore 99999            -0.0498470  0.1334491  -0.374  0.70880
## Education9 - 11th Grade        0.0460313  0.1053077   0.437  0.66209
## EducationHigh School          -0.0380138  0.1023653  -0.371  0.71042
## EducationSome College         -0.0188031  0.1014977  -0.185  0.85305
## EducationCollege Grad          0.0585822  0.1069348   0.548  0.58388
## RegularMarijYes:HardDrugsYes -0.3380416  0.0863001  -3.917 9.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7493 on 1616 degrees of freedom
##   (8043 observations deleted due to missingness)
## Multiple R-squared:  0.4388, Adjusted R-squared:  0.4312
## F-statistic: 57.44 on 22 and 1616 DF,  p-value: < 2.2e-16
```

```
#Perform GLH to collapse the income categories
Contrast.T = matrix(c(0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
                      0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,0,0,
                      0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,0,
                      0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,
                      0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,
                      0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,
                      0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,
                      0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,
                      0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,
                      1,0,0,0,0,0,0,0,0,-1,0,0,0,0,0,0,0,0,0,0,0,0,0), byrow=T, nrow=10)
car::linearHypothesis(model=m1,hypothesis.matrix=Contrast.T)
```

```
##
## Linear hypothesis test:
##
##
## Model 1: restricted model
## Model 2: AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij + HardDrugs +
##     RegularMarij * HardDrugs + Age + Gender + HHIncome + Education
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   1626 919.65
## 2   1616 907.29 10    12.361 2.2017 0.01551 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
df = df |> mutate(segmentincome = ifelse(HHIncome == "5000-9999" | HHIncome == "10000-14999", "Low", "Hi
m1 = lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+segme
summary(m1)
```

```
##
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + segmentincome +
##     Education, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.47836 -0.49235  0.01859  0.53738  2.48544
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.6042138  0.1228635   4.918 9.64e-07 ***
## SmokeNowYes                   0.1529443  0.0397066   3.852 0.000122 ***
## AlcoholYear                   0.0009325  0.0001815   5.137 3.12e-07 ***
## RegularMarijYes               0.3838283  0.0491256   7.813 9.93e-15 ***
## HardDrugsYes                  0.5086340  0.0668837   7.605 4.80e-14 ***
## Age                          -0.0502942  0.0016930 -29.707  < 2e-16 ***
## Gendermale                    0.1875971  0.0382049   4.910 1.00e-06 ***
## segmentincomeLow              0.2417697  0.0790067   3.060 0.002249 **
## Education9 - 11th Grade       0.0136409  0.1047108   0.130 0.896367
## EducationHigh School         -0.0710857  0.1018240  -0.698 0.485200
## EducationSome College        -0.0797806  0.1000099  -0.798 0.425145
## EducationCollege Grad        -0.0107724  0.1041574  -0.103 0.917639
## RegularMarijYes:HardDrugsYes -0.3165513  0.0860656  -3.678 0.000243 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7523 on 1626 degrees of freedom
##   (8043 observations deleted due to missingness)
## Multiple R-squared:  0.4308, Adjusted R-squared:  0.4266
## F-statistic: 102.6 on 12 and 1626 DF,  p-value: < 2.2e-16
```

```r
m1|>
  tbl_regression(intercept = TRUE,show_single_row = c(SmokeNow, RegularMarij, HardDrugs, Gender, segmen
  as_gt() |>
  gt::tab_header(title = "AvgSexFreq MLR")
```

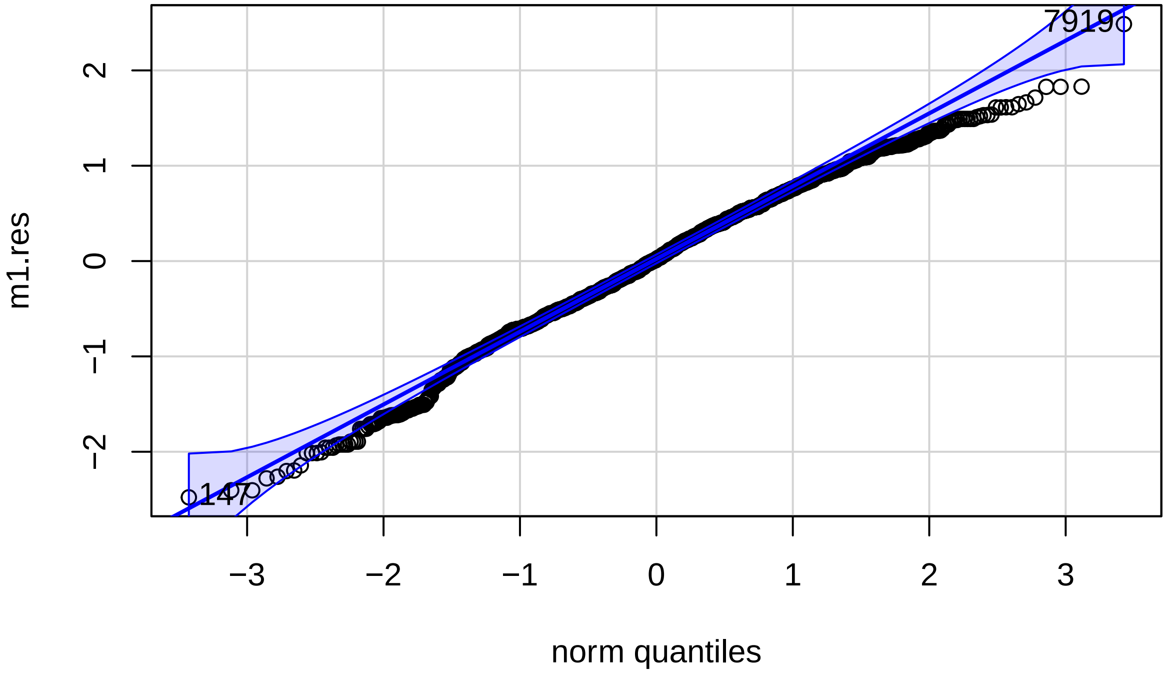
AvgSexFreq MLR

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	0.60	0.36, 0.85	<0.001
SmokeNow	0.15	0.08, 0.23	<0.001
AlcoholYear	0.00	0.00, 0.00	<0.001
RegularMarij	0.38	0.29, 0.48	<0.001
HardDrugs	0.51	0.38, 0.64	<0.001
Age	-0.05	-0.05, -0.05	<0.001
Gender	0.19	0.11, 0.26	<0.001

segmentincome	0.24	0.09, 0.40	0.002
Education			
8th Grade	—	—	
9 - 11th Grade	0.01	-0.19, 0.22	0.9
High School	-0.07	-0.27, 0.13	0.5
Some College	-0.08	-0.28, 0.12	0.4
College Grad	-0.01	-0.22, 0.19	>0.9
RegularMarij * HardDrugs			
Yes * Yes	-0.32	-0.49, -0.15	<0.001

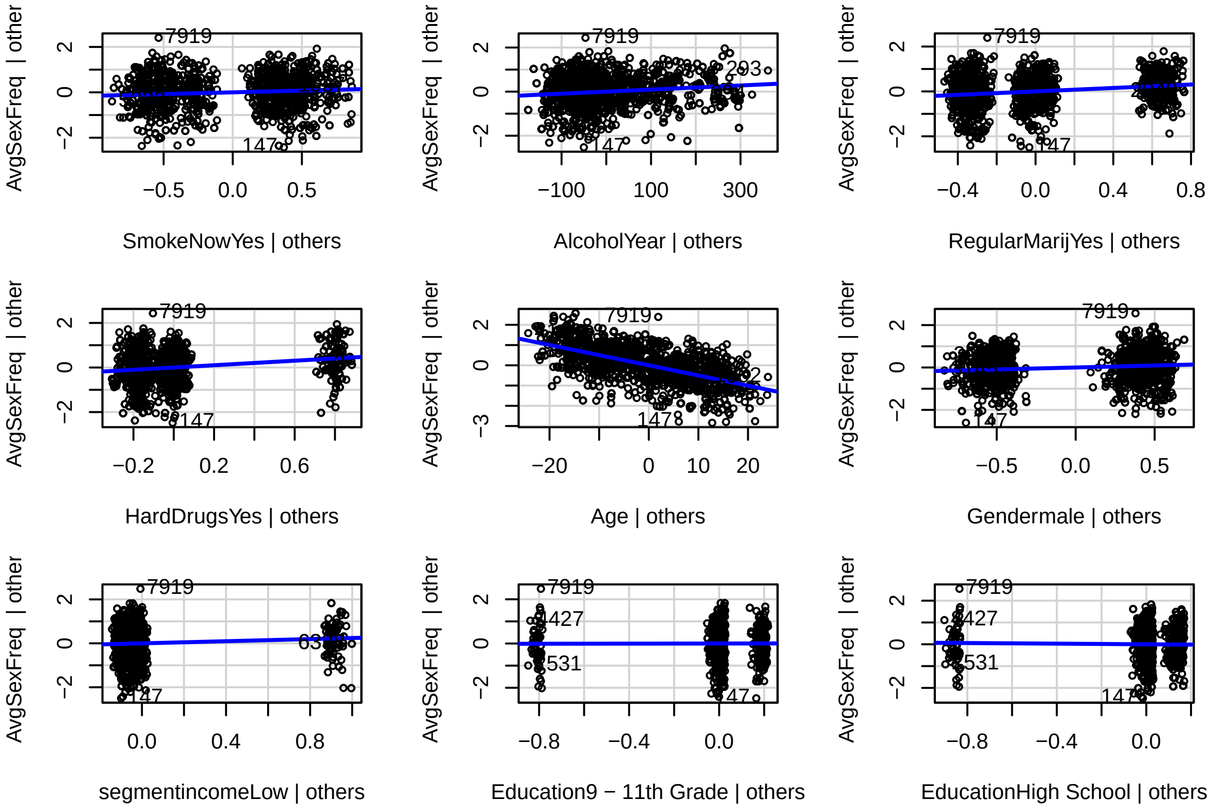
[1] CI = Confidence Interval

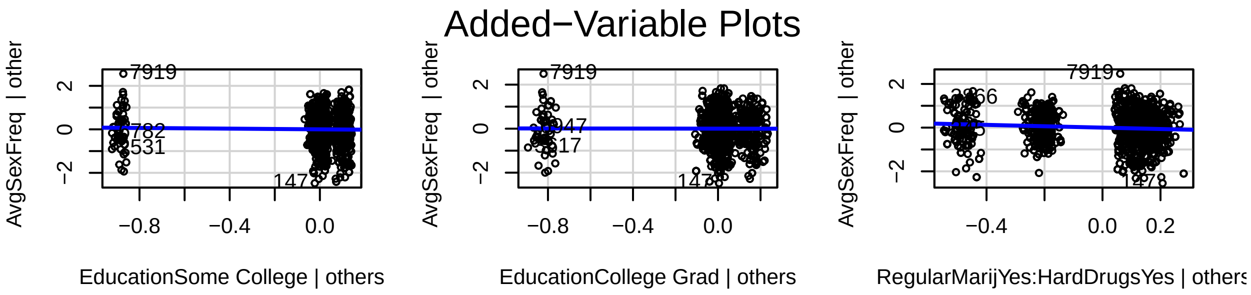
```r
m1.res = m1$residuals

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



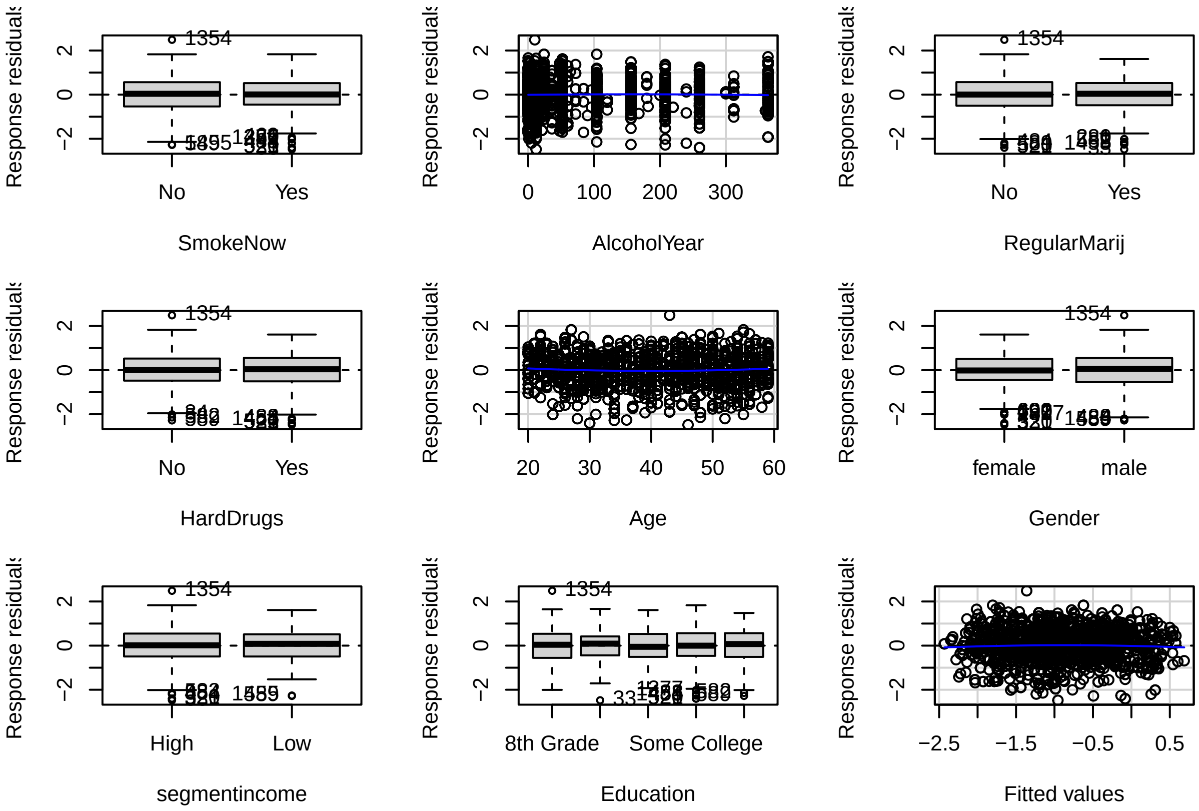
```
## 7919  147
## 1354   33
```

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


AvgSexFreq | other
SmokeNowYes | others
AlcoholYear | others
RegularMarijYes | others
HardDrugsYes | others
Age | others
Gendermale | others
segmentincomeLow | others
Education9 − 11th Grade | others
EducationHigh School | others
7919
147
293
1478
1427
531
−0.5
0.0
0.5
−100
100
300
−0.4
0.0
0.4
0.8
−0.2
0.2
0.6
−20
0
10
20
0.0
0.4
0.8
−0.8
−0.4
0.0
2
0
−2
−3



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



```
##                Test stat Pr(>|Test stat|)
## SmokeNow
## AlcoholYear      -0.5797          0.56219
## RegularMarij
## HardDrugs
## Age               1.7910          0.07347 .
## Gender
## segmentincome
## Education
## Tukey test       -1.0680          0.28553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.4029405      1.190102       0
##  Alternative hypothesis: rho != 0
```

```r
#Use a non interactive model to check for collinearity
nonintmodel <- lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+Age+Gender+segmentincome+Education, df
car::vif(nonintmodel,type = 'predictor')
```

```
## GVIFs computed for predictors

##                   GVIF Df GVIF^(1/(2*Df)) Interacts With
## SmokeNow      1.129942  1        1.062987             --
## AlcoholYear   1.103500  1        1.050476             --
## RegularMarij  1.024763  1        1.012306           --
## Age           1.079130  1        1.038812           --
## Gender        1.035586  1        1.017638           --
## segmentincome 1.025706  1        1.012772           --
## Education     1.195547  4        1.022577           --
##                                                                     Other Predictors
## SmokeNow           AlcoholYear, RegularMarij, Age, Gender, segmentincome, Education
## AlcoholYear           SmokeNow, RegularMarij, Age, Gender, segmentincome, Education
## RegularMarij           SmokeNow, AlcoholYear, Age, Gender, segmentincome, Education
## Age           SmokeNow, AlcoholYear, RegularMarij, Gender, segmentincome, Education
## Gender           SmokeNow, AlcoholYear, RegularMarij, Age, segmentincome, Education
## segmentincome           SmokeNow, AlcoholYear, RegularMarij, Age, Gender, Education
## Education           SmokeNow, AlcoholYear, RegularMarij, Age, Gender, segmentincome
```

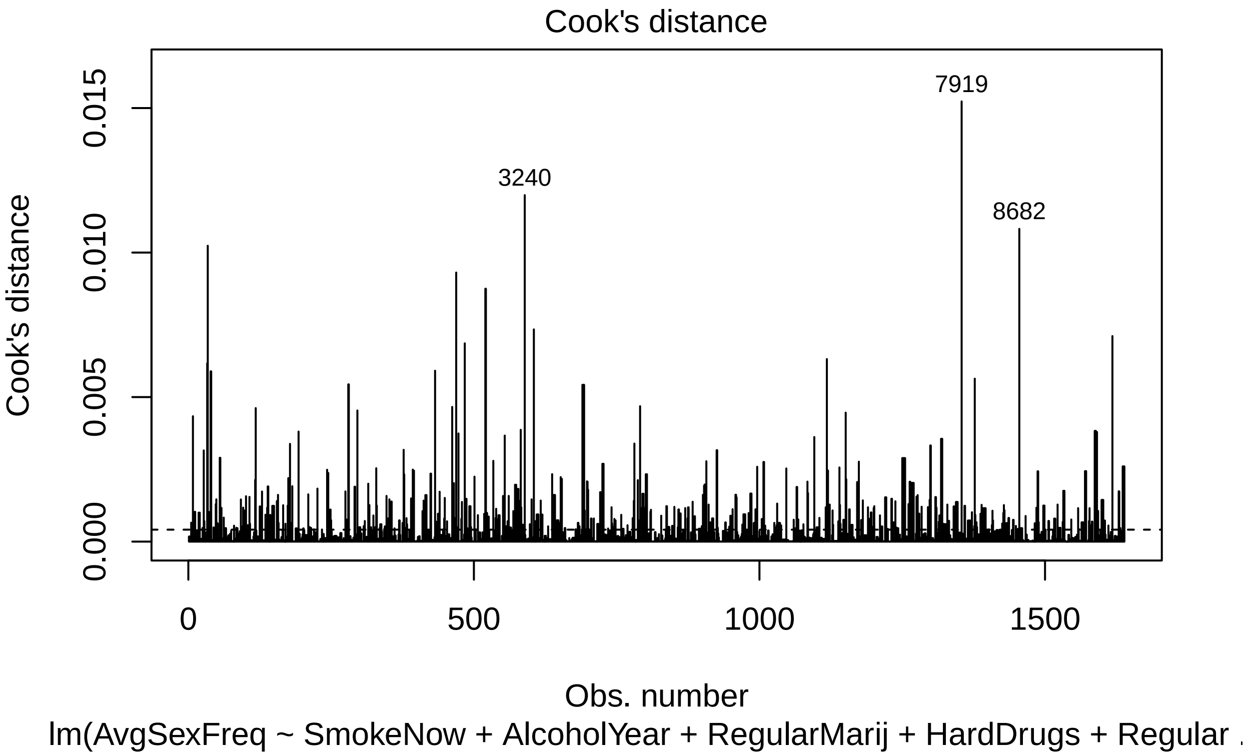
```
model.deffits=dffits(m1)
model.CD = cooks.distance(m1)
model.deffits[which.max(model.deffits)]
```

```
##      7919
## 0.4463536
```

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

```
##       7919
## 0.01523016
```

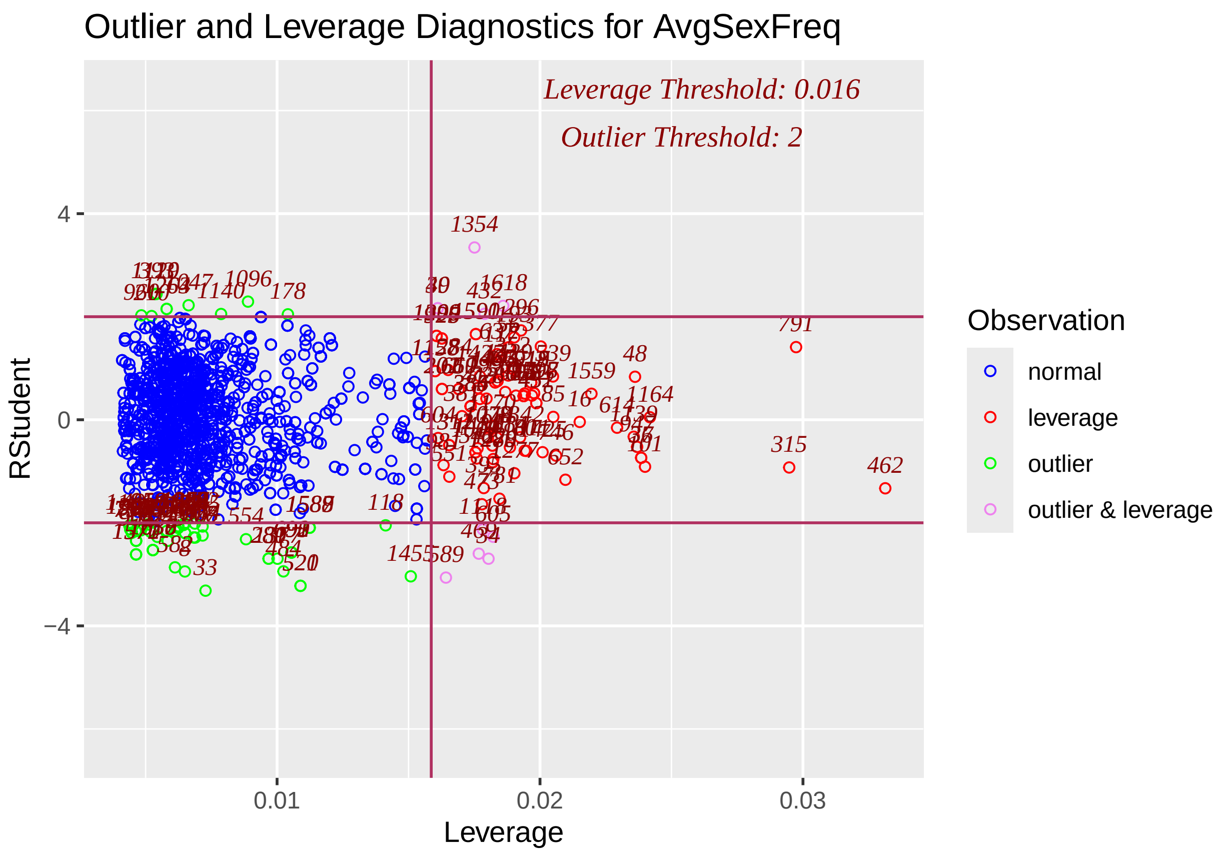
```
n = nrow(df)
p = m1$rank
plot(m1, which = 4)
abline(h=4/n,lty=2)
```



```
df[c(3240, 7919, 8682),]
```

```
## # A tibble: 3 x 79
##      ID SurveyYr Gender   Age AgeDecade AgeMonths Race1   Race3   Education
##   <int> <fct>    <fct>  <int> <fct>         <int> <fct>   <fct>   <fct>
## 1 58706 2009_10  male      34 " 30-39"        415 Mexican <NA>    College Grad
## 2 68401 2011_12  male      43 " 40-49"         NA Mexican Mexican 8th Grade
## 3 69888 2011_12  male      36 " 30-39"         NA White   White   Some College
## # i 70 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)
```



```r
df[c(1354, 1618),]
```

```
## # A tibble: 2 x 79
##      ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education
##   <int> <fct>    <fct>  <int> <fct>         <int> <fct> <fct> <fct>
## 1 54520 2009_10  male      27 " 20-29"        326 White <NA>  High School
## 2 55096 2009_10  female    17 " 10-19"        210 White <NA>  <NA>
## # i 70 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>, ...
```

Coefficient Interpretation is as follows:

$$\frac{f(x+1)}{f(x)} - 1 = (e^{\beta_1} - 1) * 100$$

```r
df2 = df[-c(7919),]
m2 = lm(AvgSexFreq ~ SmokeNow+AlcoholYear+RegularMarij+HardDrugs+RegularMarij*HardDrugs+Age+Gender+segme
summary(m1)
```

```
##
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + segmentincome +
##     Education, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.47836 -0.49235  0.01859  0.53738  2.48544
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.6042138  0.1228635   4.918 9.64e-07 ***
## SmokeNowYes                  0.1529443  0.0397066   3.852 0.000122 ***
## AlcoholYear                  0.0009325  0.0001815   5.137 3.12e-07 ***
## RegularMarijYes              0.3838283  0.0491256   7.813 9.93e-15 ***
## HardDrugsYes                 0.5086340  0.0668837   7.605 4.80e-14 ***
## Age                         -0.0502942  0.0016930 -29.707  < 2e-16 ***
## Gendermale                   0.1875971  0.0382049   4.910 1.00e-06 ***
## segmentincomeLow             0.2417697  0.0790067   3.060 0.002249 **
## Education9 - 11th Grade      0.0136409  0.1047108   0.130 0.896367
## EducationHigh School        -0.0710857  0.1018240  -0.698 0.485200
## EducationSome College       -0.0797806  0.1000099  -0.798 0.425145
## EducationCollege Grad       -0.0107724  0.1041574  -0.103 0.917639
## RegularMarijYes:HardDrugsYes -0.3165513  0.0860656  -3.678 0.000243 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7523 on 1626 degrees of freedom
##   (8043 observations deleted due to missingness)
## Multiple R-squared:  0.4308, Adjusted R-squared:  0.4266
## F-statistic: 102.6 on 12 and 1626 DF,  p-value: < 2.2e-16
```

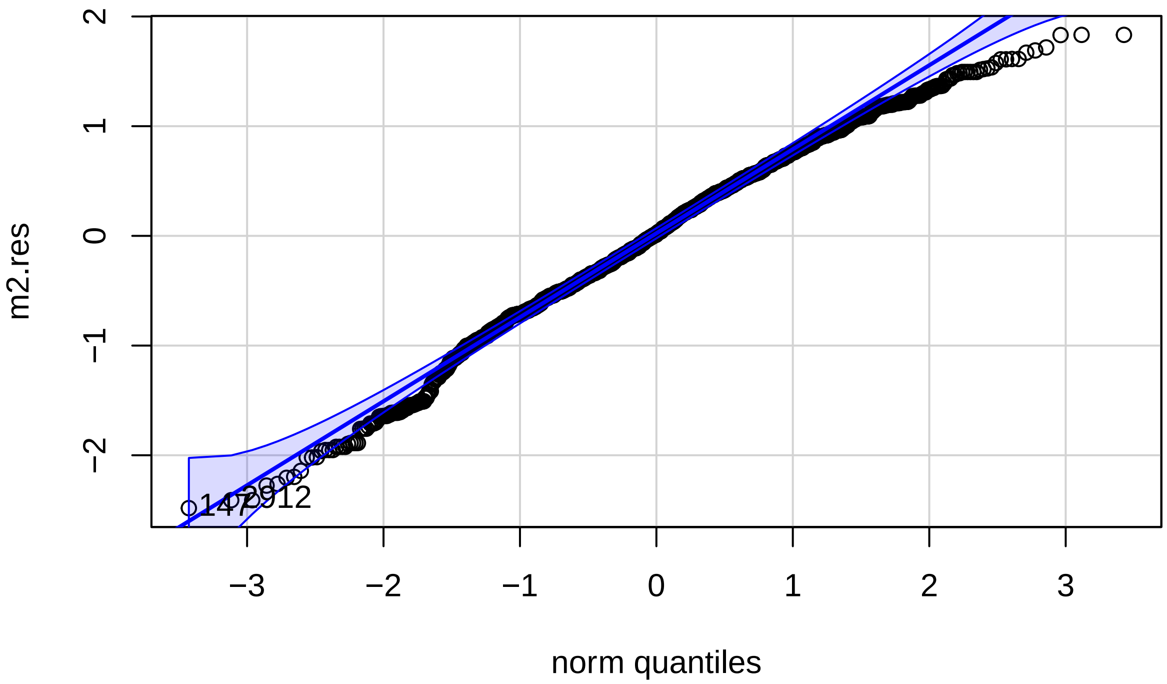
```
summary(m2)
```

```
##
## Call:
## lm(formula = AvgSexFreq ~ SmokeNow + AlcoholYear + RegularMarij +
##     HardDrugs + RegularMarij * HardDrugs + Age + Gender + segmentincome +
##     Education, data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48179 -0.49274  0.01562  0.54035  1.83234
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.5633778  0.1230881   4.577 5.07e-06 ***
## SmokeNowYes                  0.1567232  0.0395990   3.958 7.89e-05 ***
## AlcoholYear                  0.0009393  0.0001810   5.191 2.36e-07 ***
## RegularMarijYes              0.3865065  0.0489791   7.891 5.45e-15 ***
## HardDrugsYes                 0.5106993  0.0666782   7.659 3.19e-14 ***
## Age                         -0.0503187  0.0016877 -29.814  < 2e-16 ***
## Gendermale                   0.1851264  0.0380930   4.860 1.29e-06 ***
## segmentincomeLow             0.2419673  0.0787606   3.072 0.002160 **
## Education9 - 11th Grade      0.0524570  0.1050282   0.499 0.617525
## EducationHigh School        -0.0323726  0.1021649  -0.317 0.751386
## EducationSome College       -0.0408125  0.1003772  -0.407 0.684361
## EducationCollege Grad        0.0290340  0.1045132   0.278 0.781200
## RegularMarijYes:HardDrugsYes -0.3185743  0.0857996  -3.713 0.000212 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7499 on 1625 degrees of freedom
##   (8043 observations deleted due to missingness)
## Multiple R-squared:  0.4332, Adjusted R-squared:  0.4291
## F-statistic: 103.5 on 12 and 1625 DF,  p-value: < 2.2e-16
```

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

```
##                  (Intercept)                   SmokeNowYes
##                   6.75853945                    2.47078537
##                  AlcoholYear               RegularMarijYes
##                   0.73530672                    0.69774687
##                 HardDrugsYes                           Age
##                   0.40605571                   -0.04875762
##                   Gendermale             segmentincomeLow
##                   1.31700599                    0.08173055
##      Education9 - 11th Grade          EducationHigh School
##                 284.55599155                  -54.45968206
##       EducationSome College         EducationCollege Grad
##                 -48.84401287                 -369.52142736
## RegularMarijYes:HardDrugsYes
##                  -0.63905527
```

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

m2.res = m2$residuals

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



```
##  147 2912
##   33  520
```