

# Project Title

A STAT 139 Final Project

*Yuyue Wang, Xiangru Shu, Chengye Liu, Chia Chi (Michelle) Ho*

*Due December 13, 2017*

## Abstract

## Introduction

- Obesity is an exerbating problem in the US.
- Explore the association of 21 different factors with bmi, 13 of which are behavior-related factors such as the typical number of hours sleep per night

## Methods

- Data description
- data source is NHANES 2013-2014
- Variables of interest
- Only consider adults of age 20 or above
- Data preprocessing & assumptions
- Merge data by participant sequence number
- Exclude don't know/refused/missing values — discuss implications in limitations
- Perform EDA
- Fit regression models
- Check assumptions

## Results

### Exploratory Data Analysis

### Limitations

### Conclusions

## Appendix

### Appendix I: Data preprocessing

```
## Warning: package 'dplyr' was built under R version 3.4.2
##
## Attaching package: 'dplyr'
```

```

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##      gender          age          race          edu          marriage
##            0              0              0              0                  0
##      famsize        famincome    alcohol12yr    alcoholfrq      grocery
##            0              0              0              0                  0
##      eatout        delivery       milk  meals_nothome meals_fastfood
##            0              0              0              0                  0
##      depressed    sleep_trouble   activity      tv_hrs      sleep_hr
##            0              0              0              0                  0
##      smoke           bmi      bmi_class
##            0              0              0

##      gender          age          race          edu          marriage
##      "factor"      "integer"    "factor"    "factor"    "factor"
##      famsize        famincome    alcohol12yr    alcoholfrq      grocery
##      "integer"      "factor"    "factor"    "integer"    "integer"
##      eatout        delivery       milk  meals_nothome meals_fastfood
##      "integer"      "integer"    "factor"    "integer"    "integer"
##      depressed    sleep_trouble   activity      tv_hrs      sleep_hr
##      "factor"      "factor"    "factor"    "factor"    "integer"
##      smoke           bmi      bmi_class
##      "factor"      "numeric"    "character"

```

## Appendix II: Exploratory Data Analysis

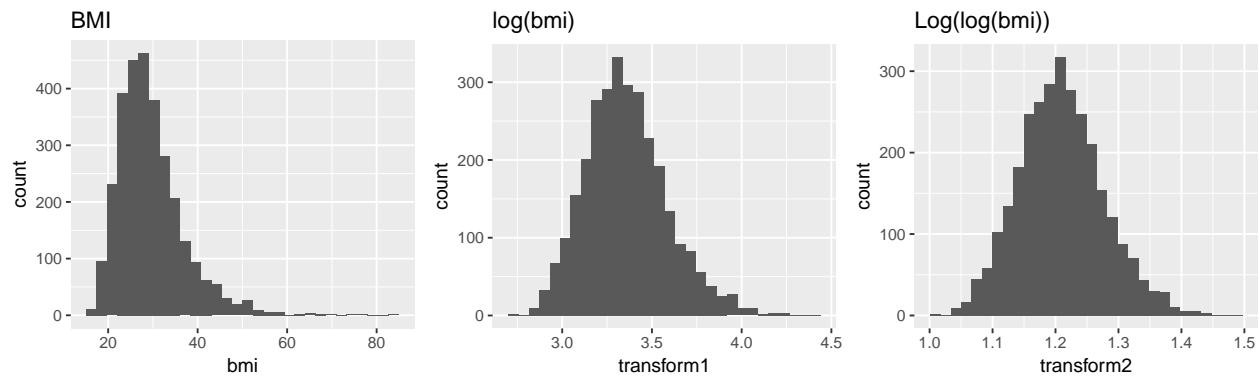
### Response Variable (bmi)

```

## Loading required package: gridExtra
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```

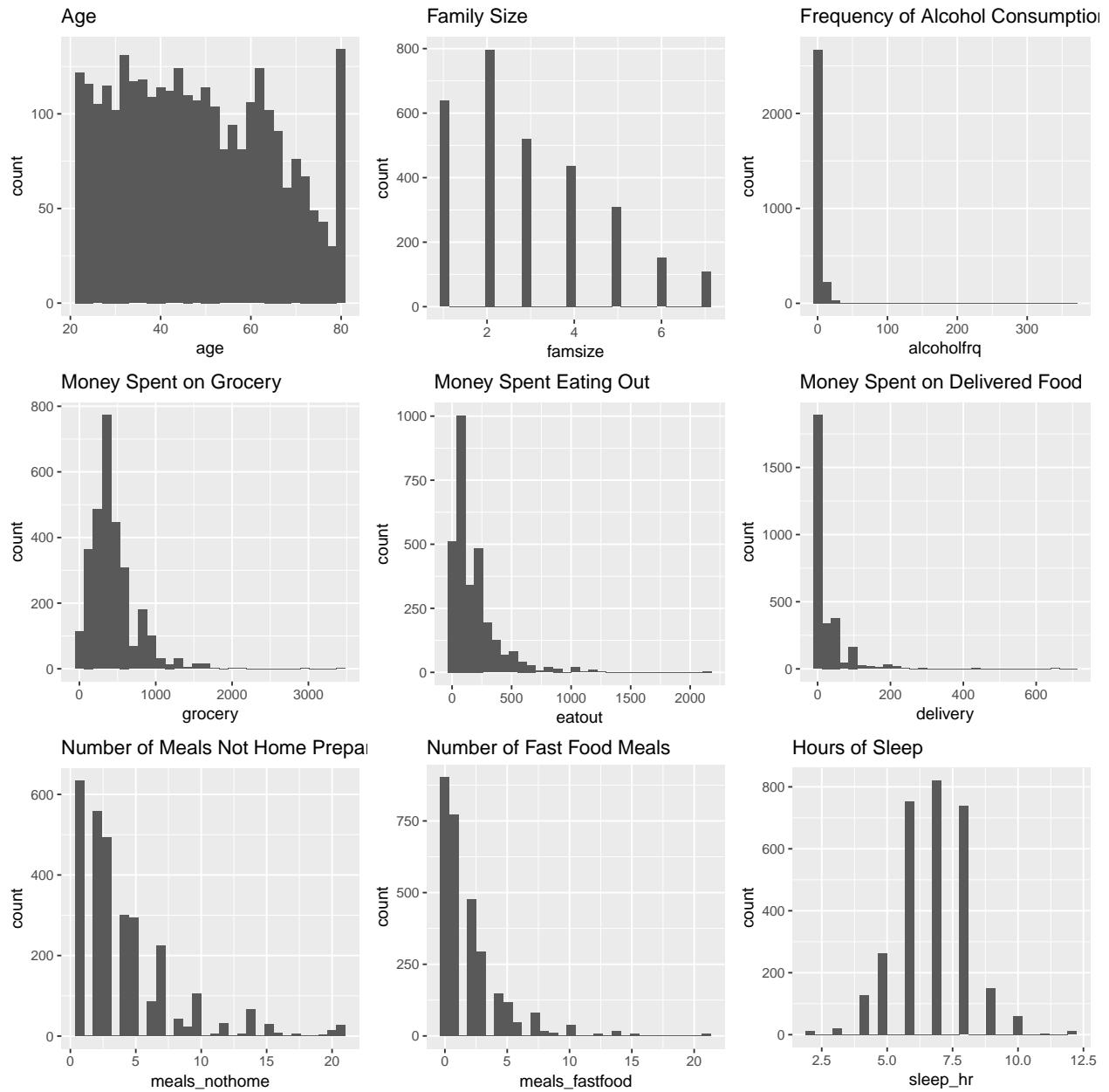


## Predictor Variables

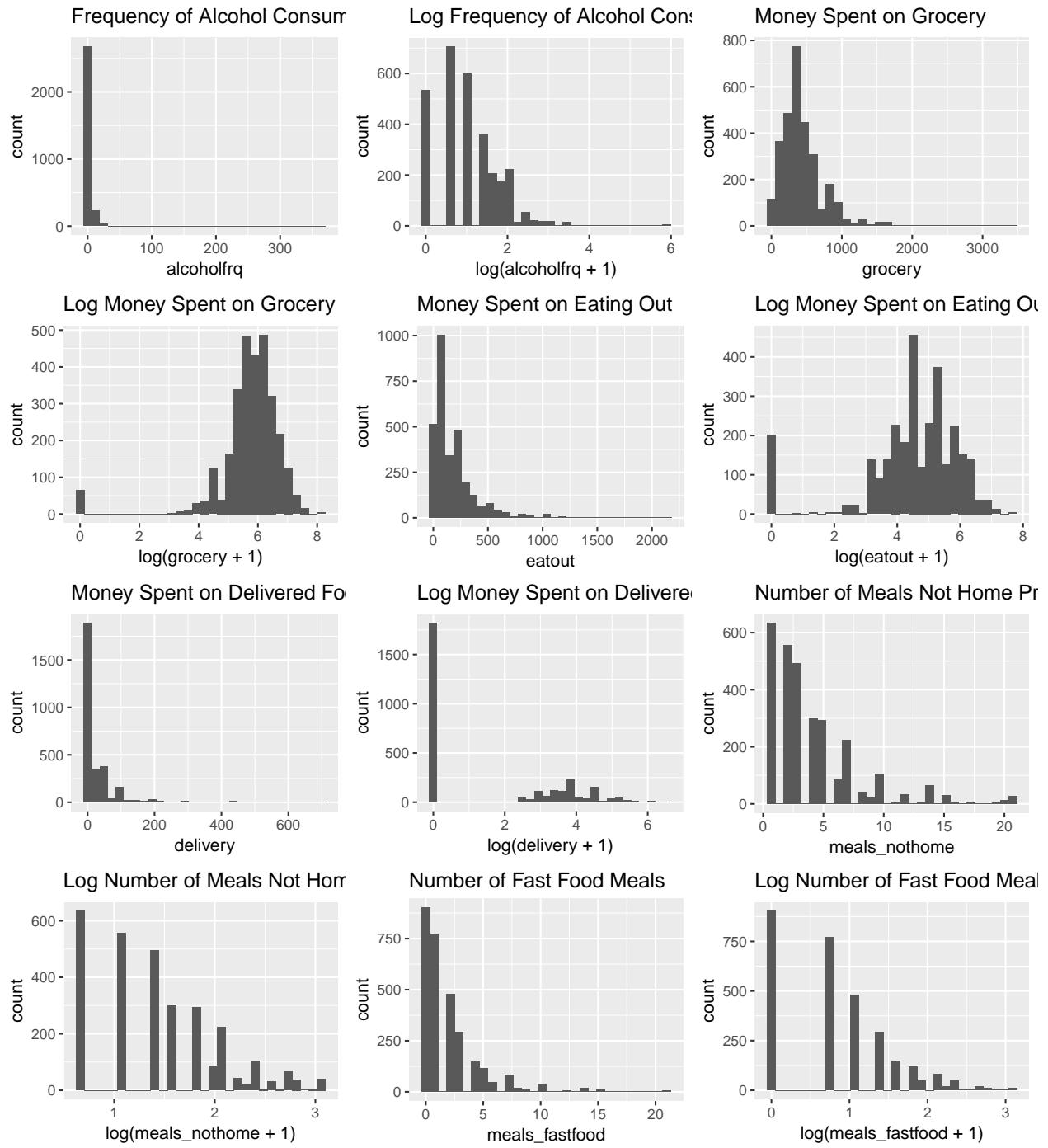
### Numeric Variables

Distribtuion of numeric variables

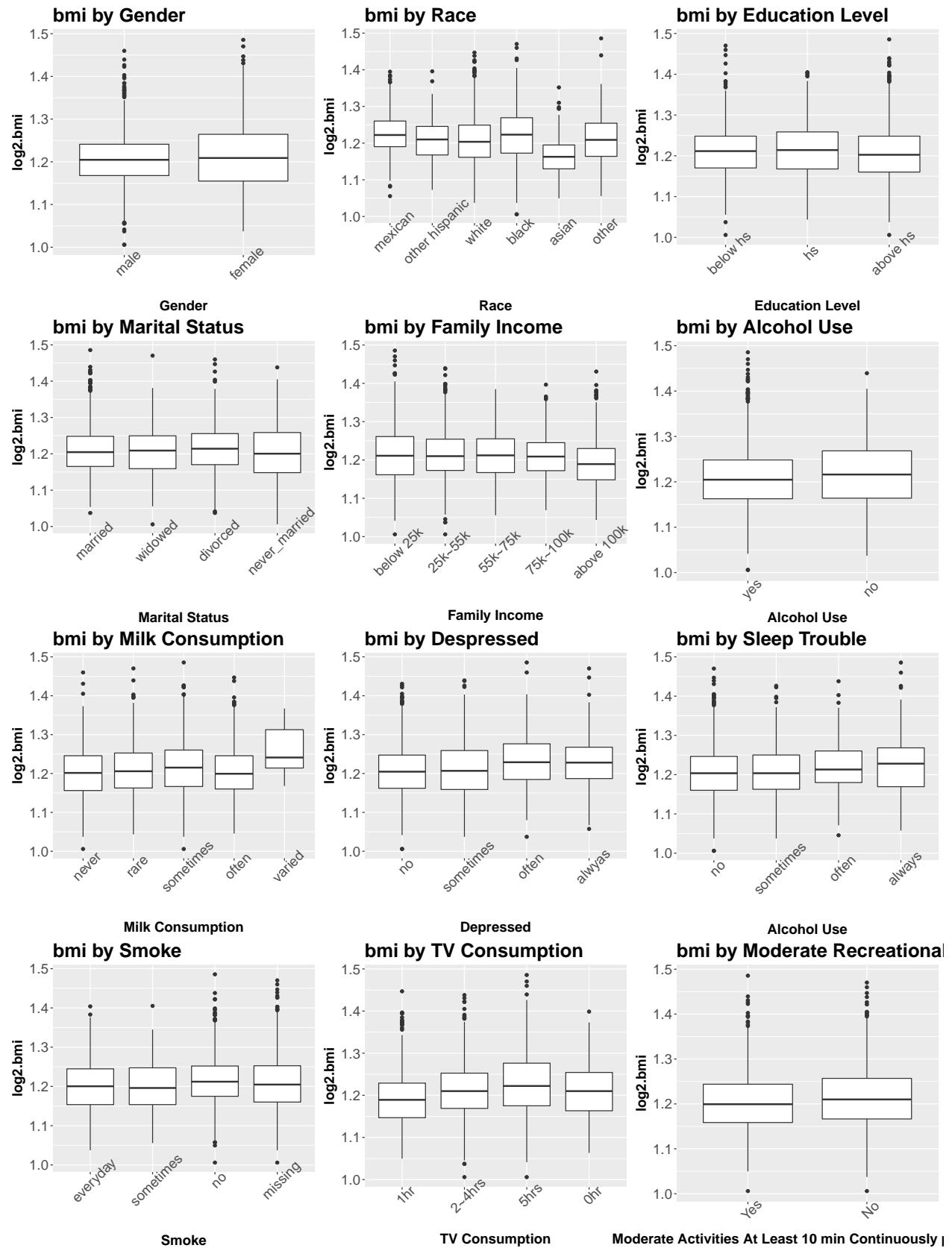
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



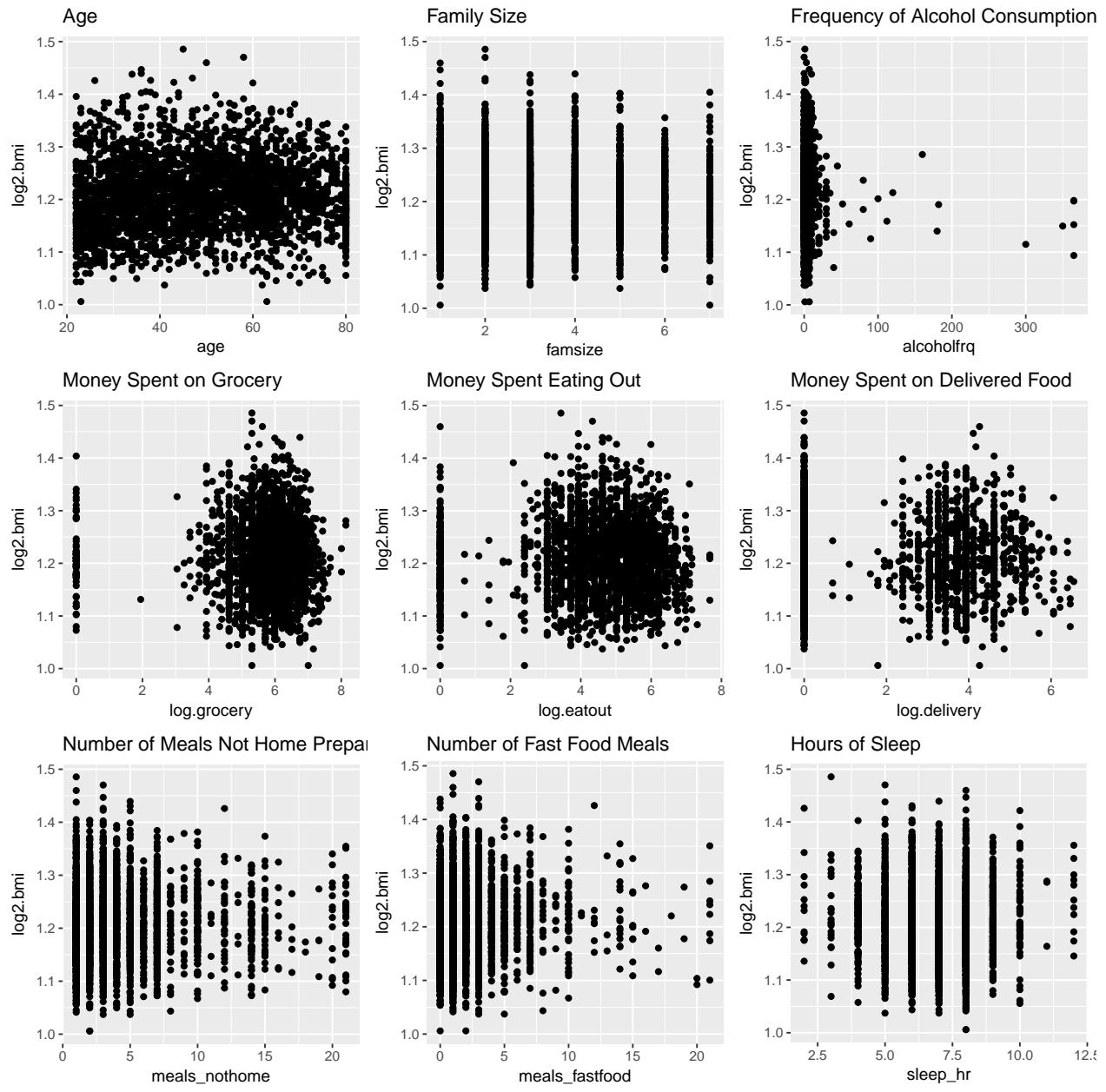
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



## Categorical Variables



Response vs. numeric distribution



## Models

```
#df_adult
check_assumptions <- function(model){
  hist(resid(model), col="grey", breaks=15)
  plot(resid(model)~fitted(model), cex=1.5)
  abline(h=0, lwd=2)
}
```

```

library(Matrix)
library(glmnet)

## Warning: package 'glmnet' was built under R version 3.4.2
## Loading required package: foreach
## Loaded glmnet 2.0-13
library(plotmo)

## Loading required package: plotrix
## Warning: package 'plotrix' was built under R version 3.4.3
## Loading required package: TeachingDemos
library(caret)

## Loading required package: lattice
library(lattice)

set.seed(1234)
# define training control
train_control <- trainControl(method="cv", number=5)

#baseline model
model1 = lm(log2.bmi~gender+age+race+edu+marriage+famsize+famincome+alcohol12yr+alcoholfrq+log.grocery+log.eatout+log.delivery+milk+meals_nothome+meals_fastfood+depressed+sleep_trouble+activity+tv_hours+sleep_hr+smoke, data = df_adult)

base_form = formula(model1)

summary(model1)

##
## Call:
## lm(formula = log2.bmi ~ gender + age + race + edu + marriage +
##     famsize + famincome + alcohol12yr + alcoholfrq + log.grocery +
##     log.eatout + log.delivery + milk + meals_nothome + meals_fastfood +
##     depressed + sleep_trouble + activity + tv_hours + sleep_hr +
##     smoke, data = df_adult)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -0.216033 -0.042926 -0.001571  0.037661  0.251263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.174e+00 1.244e-02 94.348 < 2e-16 ***
## genderfemale 6.522e-03 2.542e-03  2.566 0.010334 *
## age          1.238e-05 9.107e-05  0.136 0.891867

```

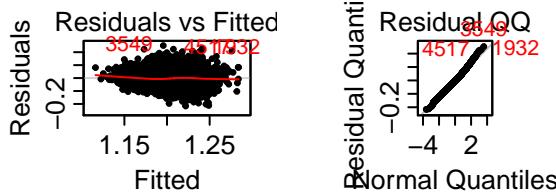
```

## raceother hispanic      -1.977e-02 5.433e-03 -3.639 0.000279 ***
## racewhite               -1.745e-02 4.121e-03 -4.235 2.36e-05 ***
## raceblack                -6.450e-03 4.615e-03 -1.398 0.162343
## raceasian                -5.589e-02 5.625e-03 -9.936 < 2e-16 ***
## raceother                -1.310e-02 7.660e-03 -1.710 0.087362 .
## eduhs                     6.452e-03 4.083e-03 1.580 0.114175
## eduabove hs              8.396e-03 3.834e-03 2.190 0.028624 *
## marriagewidowed          -7.843e-03 5.771e-03 -1.359 0.174211
## marrieddivorced          1.411e-03 3.712e-03 0.380 0.703971
## marriagenever_married   -6.582e-03 3.657e-03 -1.800 0.072007 .
## famsize                   1.048e-03 8.544e-04 1.227 0.220056
## famincome25k~55k          3.582e-03 3.293e-03 1.088 0.276744
## famincome55k~75k          -1.775e-03 4.446e-03 -0.399 0.689739
## famincome75k~100k         -2.414e-03 4.813e-03 -0.502 0.615958
## famincomeabove 100k       -1.285e-02 4.264e-03 -3.015 0.002594 **
## alcohol12yrno             4.590e-03 3.573e-03 1.285 0.199038
## alcoholfrq                -1.576e-04 6.709e-05 -2.349 0.018895 *
## log.grocery                 4.339e-04 1.163e-03 0.373 0.709206
## log.eatout                  2.039e-03 8.723e-04 2.338 0.019463 *
## log.delivery                 -5.234e-04 6.241e-04 -0.839 0.401684
## milkrare                    8.184e-03 3.974e-03 2.060 0.039526 *
## milksometimes                9.926e-03 3.578e-03 2.774 0.005573 **
## milkoften                     3.152e-03 3.527e-03 0.894 0.371559
## milkvaried                   5.545e-02 2.158e-02 2.570 0.010224 *
## meals_nothome                 -4.889e-05 4.116e-04 -0.119 0.905466
## meals_fastfood                 7.922e-04 5.683e-04 1.394 0.163401
## depressedsometimes           -3.224e-04 3.304e-03 -0.098 0.922286
## depressedoften                  2.130e-02 6.621e-03 3.217 0.001310 **
## depresseddalwyas              8.558e-03 6.972e-03 1.227 0.219775
## sleep_troublesometimes        -2.588e-04 2.984e-03 -0.087 0.930885
## sleep_troubleoften            6.435e-03 5.100e-03 1.262 0.207134
## sleep_troublealways            1.046e-02 4.610e-03 2.268 0.023388 *
## activityNo                     5.992e-03 2.445e-03 2.451 0.014309 *
## tv_hrs2~4hrs                  1.630e-02 2.857e-03 5.704 1.29e-08 ***
## tv_hrs5hrs                     2.746e-02 4.107e-03 6.685 2.76e-11 ***
## tv_hrs0hr                      1.430e-02 8.722e-03 1.639 0.101294
## sleep_hr                        -1.509e-03 8.878e-04 -1.699 0.089383 .
## smokesometimes                  6.294e-03 6.565e-03 0.959 0.337729
## smokeno                         2.323e-02 3.941e-03 5.894 4.20e-09 ***
## smokemissing                     1.631e-02 3.667e-03 4.447 9.01e-06 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06387 on 2916 degrees of freedom
## Multiple R-squared: 0.1153, Adjusted R-squared: 0.1025
## F-statistic: 9.047 on 42 and 2916 DF, p-value: < 2.2e-16

```

```
plotres(model1, which=3:4, caption = "Base model assumption check")
```

Base mo...



```
#baseline model cv
model1_cv = train(base_form, data=df_adult, method="lm", preProcess="scale", trControl=train_control)

print("Base model cv results:")

## [1] "Base model cv results:"
print(paste("RMSE:", model1_cv$results$RMSE, "Rsquared:", model1_cv$results$Rsquared))

## [1] "RMSE: 0.0645052005209794 Rsquared: 0.088067250588867"

#model1a: baseline model step
set.seed(1234)
model1a = step(model1, trace = 0, direction = "backward")

summary(model1a)

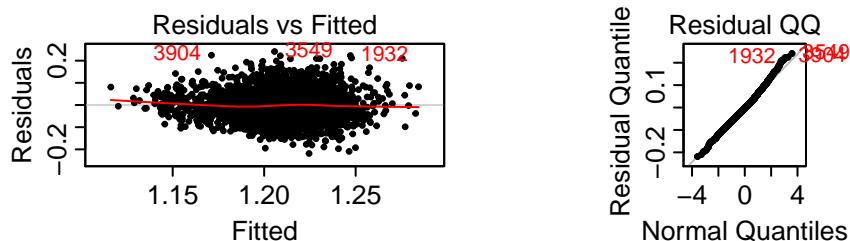
##
## Call:
## lm(formula = log2.bmi ~ gender + race + marriage + famincome +
##     alcoholfrq + log.eatout + milk + meals_fastfood + depressed +
##     sleep_trouble + activity + tv_hours + sleep_hr + smoke, data = df_adult)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.218643 -0.042895 -0.001923  0.038362  0.241206 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.1843326  0.0093716 126.375 < 2e-16 ***
## genderfemale 0.0076379  0.0024725   3.089  0.002026 ** 
## raceother hispanic -0.0191613  0.0053818  -3.560  0.000376 *** 
## racewhite    -0.0165158  0.0039074  -4.227  2.44e-05 *** 
## raceblack    -0.0056150  0.0044921  -1.250  0.211404  
## raceasian    -0.0538556  0.0054758  -9.835 < 2e-16 *** 
## raceother    -0.0114075  0.0075490  -1.511  0.130862  
## marriagewidowed -0.0088122  0.0055040  -1.601  0.109475  
## marriagedivorced  0.0006233  0.0036014   0.173  0.862610  
## marriagenever_married -0.0071471  0.0034005  -2.102  0.035661 *  
## famincome25k~55k  0.0044158  0.0032507   1.358  0.174438 
```

```

## famincome55k~75k      -0.0005382  0.0043425  -0.124  0.901373
## famincome75k~100k    -0.0007125  0.0046866  -0.152  0.879172
## famincomeabove 100k   -0.0109757  0.0040286  -2.724  0.006480 ** 
## alcoholfrq            -0.0001616  0.0000670  -2.412  0.015912 * 
## log.eatout             0.0021437  0.0008550  2.507  0.012221 * 
## milkrare               0.0083961  0.0039669  2.117  0.034383 * 
## milksometimes          0.0102260  0.0035706  2.864  0.004214 ** 
## milkoften               0.0034727  0.0035217  0.986  0.324176
## milkvaried              0.0551352  0.0215682  2.556  0.010629 * 
## meals_fastfood           0.0007148  0.0004502  1.588  0.112457
## depressedsometimes     -0.0002451  0.0032985  -0.074  0.940782
## depressedoften            0.0210983  0.0066160  3.189  0.001443 ** 
## depresseddalwyas        0.0078500  0.0069587  1.128  0.259378
## sleep_troublesometimes  -0.0003573  0.0029785  -0.120  0.904527
## sleep_troubleoften       0.0066860  0.0050872  1.314  0.188851
## sleep_troublealways      0.0100479  0.0046037  2.183  0.029146 * 
## activityNo                0.0057577  0.0024291  2.370  0.017836 * 
## tv_hours2~4hrs            0.0159792  0.0028168  5.673  1.54e-08 *** 
## tv_hours5hrs              0.0267812  0.0040156  6.669  3.06e-11 *** 
## tv_hours0hr                0.0143202  0.0087118  1.644  0.100330
## sleep_hr                  -0.0016220  0.0008811  -1.841  0.065743 .
## smokesometimes             0.0065914  0.0065442  1.007  0.313915
## smokeno                   0.0236999  0.0038542  6.149  8.85e-10 *** 
## smokemissing              0.0174996  0.0035986  4.863  1.22e-06 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.06387 on 2924 degrees of freedom
## Multiple R-squared:  0.1128, Adjusted R-squared:  0.1025
## F-statistic: 10.94 on 34 and 2924 DF,  p-value: < 2.2e-16
plotres(model1a, which=3:4, caption = "model1a assumption check")

```

model1a assu...



```

base_1a_form = formula(model1a)
#baseline model cv
model1a_cv = train(base_1a_form, data=df_adult, method="lm", preProcess="scale", trControl=train_cv)

print("model1a cv results:")

## [1] "model1a cv results:"

```

```

print(paste("RMSE:", model1a_cv$results$RMSE, "Rsquared:", model1a_cv$results$Rsquared))

## [1] "RMSE: 0.0644145278641605 Rsquared: 0.0903970403193655"

#compare model1 with model1a
set.seed(1234)
anova(model1a, model1)

## Analysis of Variance Table
##
## Model 1: log2.bmi ~ gender + race + marriage + famincome + alcoholfrq +
##           log.eatout + milk + meals_fastfood + depressed + sleep_trouble +
##           activity + tv_hrs + sleep_hr + smoke
## Model 2: log2.bmi ~ gender + age + race + edu + marriage + famsize + famincome +
##           alcohol12yr + alcoholfrq + log.grocery + log.eatout + log.delivery +
##           milk + meals_nothome + meals_fastfood + depressed + sleep_trouble +
##           activity + tv_hrs + sleep_hr + smoke
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1   2924 11.928
## 2   2916 11.896  8   0.03278 1.0044 0.4303

#model1b: add sleep^2 to model1a
set.seed(1234)

base_sleep_form = update.formula(base_1a_form, . ~. + I(sleep_hr^2))

model1b = lm(base_sleep_form, data = df_adult)
summary(model1b)

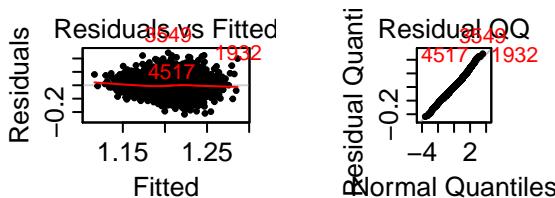
##
## Call:
## lm(formula = base_sleep_form, data = df_adult)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.217409 -0.042988 -0.001923  0.038642  0.243013 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.235e+00  1.891e-02 65.315 < 2e-16 ***
## genderfemale 7.449e-03  2.470e-03  3.016 0.002580 ** 
## raceother hispanic -1.952e-02  5.375e-03 -3.632 0.000286 *** 
## racewhite    -1.674e-02  3.902e-03 -4.289 1.85e-05 *** 
## raceblack    -6.228e-03  4.490e-03 -1.387 0.165540    
## raceasian     -5.392e-02  5.468e-03 -9.862 < 2e-16 *** 
## raceother    -1.194e-02  7.540e-03 -1.584 0.113340    
## mariagewidowed -8.714e-03  5.496e-03 -1.585 0.112965    
## mariagedivorced  5.573e-04  3.596e-03  0.155 0.876858    
## mariagenever_married -7.255e-03  3.396e-03 -2.136 0.032727 *  
## 
```

```

## famincome25k~55k      4.861e-03 3.249e-03 1.496 0.134752
## famincome55k~75k     2.292e-04 4.343e-03 0.053 0.957925
## famincome75k~100k    -7.039e-05 4.684e-03 -0.015 0.988012
## famincomeabove 100k   -1.005e-02 4.034e-03 -2.493 0.012735 *
## alcoholfrq            -1.589e-04 6.691e-05 -2.375 0.017622 *
## log.eatout             2.173e-03 8.538e-04 2.545 0.010964 *
## milkrare               8.423e-03 3.961e-03 2.127 0.033545 *
## milksometimes          1.012e-02 3.566e-03 2.839 0.004557 **
## milkoften               3.346e-03 3.517e-03 0.951 0.341453
## milkvaried              5.522e-02 2.154e-02 2.564 0.010402 *
## meals_fastfood          6.490e-04 4.501e-04 1.442 0.149418
## depressedsometimes     -2.810e-04 3.294e-03 -0.085 0.932007
## depressedoften          1.991e-02 6.618e-03 3.008 0.002651 **
## depresseddalwyas       6.766e-03 6.957e-03 0.972 0.330911
## sleep_troublesometimes -2.703e-04 2.974e-03 -0.091 0.927591
## sleep_troubleoften     6.380e-03 5.081e-03 1.256 0.209333
## sleep_troublealways    8.915e-03 4.611e-03 1.933 0.053309 .
## activityNo              5.632e-03 2.426e-03 2.322 0.020319 *
## tv_hrs2~4hrs            1.603e-02 2.813e-03 5.699 1.32e-08 ***
## tv_hrs5hrs              2.624e-02 4.013e-03 6.539 7.28e-11 ***
## tv_hrs0hr                1.394e-02 8.700e-03 1.602 0.109261
## sleep_hr                 -1.732e-02 5.146e-03 -3.365 0.000775 ***
## smokesometimes           6.946e-03 6.536e-03 1.063 0.287939
## smokeno                  2.396e-02 3.850e-03 6.225 5.52e-10 ***
## smokemissing              1.802e-02 3.597e-03 5.009 5.78e-07 ***
## I(sleep_hr^2)            1.158e-03 3.741e-04 3.096 0.001983 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06378 on 2923 degrees of freedom
## Multiple R-squared: 0.1157, Adjusted R-squared: 0.1052
## F-statistic: 10.93 on 35 and 2923 DF, p-value: < 2.2e-16
plotres(model1b, which=3:4, caption = "model1a plus sleephour^2 assumption check")

```

model1a ...



```
model1b_form = formula(model1b)
```

```
#model1b cv
```

```
model1b_cv = train(base_sleep_form, data=df_adult, method="lm", preProcess="scale", trControl=train
```

```

print("model1a plus sleephour^2 cv results:")

## [1] "model1a plus sleephour^2 cv results:"
print(paste("RMSE:", model1b_cv$results$RMSE, "Rsquared:", model1b_cv$results$Rsquared))

## [1] "RMSE: 0.064348660409757 Rsquared: 0.0924950285909764"

#compare model1b to model1a
set.seed(1234)
anova(model1a, model1b)

## Analysis of Variance Table

## Model 1: log2.bmi ~ gender + race + marriage + famincome + alcoholfrq +
##           log.eatout + milk + meals_fastfood + depressed + sleep_trouble +
##           activity + tv_hrs + sleep_hr + smoke
## Model 2: log2.bmi ~ gender + race + marriage + famincome + alcoholfrq +
##           log.eatout + milk + meals_fastfood + depressed + sleep_trouble +
##           activity + tv_hrs + sleep_hr + smoke + I(sleep_hr^2)
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1  2924 11.928
## 2  2923 11.889  1  0.038976 9.5821 0.001983 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#model1c: add famsize interactions
set.seed(1234)
model1c = lm(update.formula(model1b_form, .~.+famsize*(famincome+log.grocery+log.eatout+log.delivery))

summary(model1c)

## 
## Call:
## lm(formula = update.formula(model1b_form, . ~ . + famsize * (famincome +
##     log.grocery + log.eatout + log.delivery)), data = df_adult)
## 
## Residuals:
##       Min        1Q        Median        3Q       Max 
## -0.209958 -0.043047 -0.002103  0.038688  0.245642 
## 
## Coefficients:
## (Intercept)      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      1.227e+00  2.340e-02  52.458 < 2e-16 ***
## genderfemale     7.109e-03  2.481e-03   2.865 0.004204 ** 
## raceother hispanic -1.949e-02  5.427e-03  -3.591 0.000334 *** 
## racewhite        -1.645e-02  4.024e-03  -4.088 4.48e-05 *** 
## raceblack        -5.758e-03  4.559e-03  -1.263 0.206664  
## raceasian        -5.371e-02  5.539e-03  -9.698 < 2e-16 *** 
## raceother        -1.238e-02  7.592e-03  -1.631 0.102998 

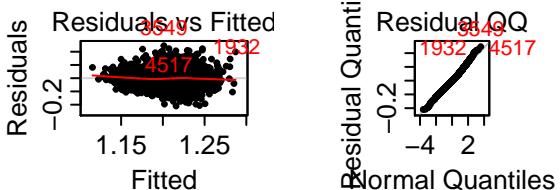
```

```

## marriagewidowed          -8.949e-03  5.657e-03 -1.582 0.113756
## marrieddivorced         1.510e-03  3.733e-03  0.405 0.685774
## marriagenever_married   -6.199e-03  3.483e-03 -1.780 0.075192 .
## famincome25k~55k         1.312e-03  6.033e-03  0.217 0.827835
## famincome55k~75k         -9.220e-03  8.754e-03 -1.053 0.292305
## famincome75k~100k        -3.015e-03  9.913e-03 -0.304 0.761064
## famincomeabove 100k       -1.701e-02  8.343e-03 -2.039 0.041506 *
## alcoholfrq               -1.549e-04  6.701e-05 -2.312 0.020860 *
## log.eatout                2.945e-04  1.649e-03  0.179 0.858299
## milkrare                 8.484e-03  3.970e-03  2.137 0.032684 *
## milksometimes            1.032e-02  3.571e-03  2.890 0.003883 **
## milkoften                3.478e-03  3.520e-03  0.988 0.323243
## milkvaried               5.508e-02  2.156e-02  2.555 0.010666 *
## meals_fastfood            7.499e-04  4.564e-04  1.643 0.100436
## depressedsometimes        -7.110e-05  3.304e-03 -0.022 0.982831
## depressedoften             2.020e-02  6.627e-03  3.048 0.002325 **
## depresseddalwyas          6.918e-03  6.966e-03  0.993 0.320738
## sleep_troublesometimes   -2.440e-04  2.979e-03 -0.082 0.934715
## sleep_troubleoften         6.163e-03  5.091e-03  1.211 0.226127
## sleep_troublealways       8.463e-03  4.619e-03  1.832 0.067003 .
## activityNo                5.238e-03  2.438e-03  2.148 0.031777 *
## tv_hrs2~4hrs              1.636e-02  2.825e-03  5.792 7.71e-09 ***
## tv_hrs5hrs                2.676e-02  4.041e-03  6.621 4.24e-11 ***
## tv_hrs0hr                 1.400e-02  8.712e-03  1.607 0.108083
## sleep_hr                  -1.731e-02  5.152e-03 -3.360 0.000789 ***
## smokesometimes             7.369e-03  6.542e-03  1.126 0.260136
## smokeno                   2.437e-02  3.873e-03  6.290 3.64e-10 ***
## smokemissing               1.833e-02  3.621e-03  5.063 4.38e-07 ***
## I(sleep_hr^2)              1.161e-03  3.744e-04  3.101 0.001950 **
## famsize                    1.386e-03  4.399e-03  0.315 0.752728
## log.grocery                3.224e-03  2.328e-03  1.385 0.166186
## log.delivery               -1.602e-03  1.289e-03 -1.242 0.214193
## famincome25k~55k:famsize   1.296e-03  1.855e-03  0.699 0.484903
## famincome55k~75k:famsize   3.438e-03  2.669e-03  1.288 0.197809
## famincome75k~100k:famsize  9.321e-04  2.828e-03  0.330 0.741759
## famincomeabove 100k:famsize 2.175e-03  2.353e-03  0.924 0.355398
## famsize:log.grocery        -8.860e-04  6.712e-04 -1.320 0.186948
## log.eatout:famsize         6.197e-04  4.639e-04  1.336 0.181685
## famsize:log.delivery       3.407e-04  3.619e-04  0.941 0.346558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06378 on 2913 degrees of freedom
## Multiple R-squared:  0.1186, Adjusted R-squared:  0.105
## F-statistic: 8.709 on 45 and 2913 DF,  p-value: < 2.2e-16
extractAIC(model1c)

```

```

## [1] 46.00 -16242.17
plotres(model1c, which=3:4, caption = "model1c assumption check")
model1c...

#model1c cv
model1c_cv = train(formula(model1c), data=df_adult, method="lm", preProcess="scale", trControl=trainControl(method="cv", number=10))

print("Model1c cv results:")
## [1] "Model1c cv results:"
print(paste("RMSE:", model1c_cv$results$RMSE, "Rsquared:", model1c_cv$results$Rsquared))

## [1] "RMSE: 0.064451866512309 Rsquared: 0.0901468116644371"
#compare model1c to model1b
set.seed(1234)
anova(model1b, model1c)

## Analysis of Variance Table
##
## Model 1: log2.bmi ~ gender + race + marriage + famincome + alcoholfrq +
##           log.eatout + milk + meals_fastfood + depressed + sleep_trouble +
##           activity + tv_hours + sleep_hr + smoke + I(sleep_hr^2)
## Model 2: log2.bmi ~ gender + race + marriage + famincome + alcoholfrq +
##           log.eatout + milk + meals_fastfood + depressed + sleep_trouble +
##           activity + tv_hours + sleep_hr + smoke + I(sleep_hr^2) + famsize +
##           log.grocery + log.delivery + famincome:famsize + famsize:log.grocery +
##           log.eatout:famsize + famsize:log.delivery
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1   2923 11.889
## 2   2913 11.851 10  0.038099 0.9365 0.4981
##
## Call:
## lm(formula = log2.bmi ~ race + tv_hours + smoke + depressed + milk +
##     famincome + gender + log.eatout + activity + alcoholfrq +
##     marriage + meals_fastfood + sleep_trouble + tv_hours:depressed +
##     famincome:gender + race:gender + smoke:gender + race:activity +
##     smoke:marriage + race:meals_fastfood + gender:meals_fastfood +
##     depressed:milk + tv_hours:log.eatout, data = df_adult)
##
## Residuals:

```

```

##      Min       1Q    Median       3Q      Max
## -0.200795 -0.040473 -0.002891  0.037567  0.219548
##
## Coefficients: (2 not defined because of singularities)
##                               Estimate Std. Error t value
## (Intercept)                1.148e+00  1.215e-02 94.550
## raceother hispanic        -6.377e-03  1.051e-02 -0.607
## racewhite                  -1.149e-02  7.842e-03 -1.465
## raceblack                  1.110e-02  8.875e-03  1.251
## raceasian                 -2.578e-02  1.012e-02 -2.548
## raceother                  2.100e-02  1.562e-02  1.344
## tv_hrs2~4hrs               2.176e-02  9.357e-03  2.326
## tv_hrs5hrs                 7.280e-03  1.153e-02  0.631
## tv_hrs0hr                  1.184e-02  2.673e-02  0.443
## smokesometimes             -5.033e-03  1.018e-02 -0.494
## smokeno                     3.052e-02  5.810e-03  5.254
## smokemissing                3.314e-02  5.636e-03  5.880
## depressedsometimes          -1.944e-02  9.465e-03 -2.054
## depressedoften              -1.178e-02  1.770e-02 -0.666
## depresseddalwyas           7.267e-03  2.083e-02  0.349
## milkrare                    5.320e-03  4.487e-03  1.186
## milksometimes              8.312e-03  4.003e-03  2.076
## milkoften                   1.192e-04  3.950e-03  0.030
## milkvaried                  5.257e-02  2.242e-02  2.344
## famincome25k~55k            1.041e-02  4.465e-03  2.332
## famincome55k~75k            6.884e-03  5.673e-03  1.213
## famincome75k~100k           7.329e-03  6.120e-03  1.198
## famincomeabove 100k          2.493e-03  5.163e-03  0.483
## genderfemale                3.552e-02  9.897e-03  3.589
## log.eatout                  3.181e-03  1.595e-03  1.994
## activityNo                  1.815e-02  6.950e-03  2.611
## alcoholfrq                  -1.759e-04  6.576e-05 -2.674
## marriagewidowed             -2.478e-02  1.886e-02 -1.314
## marrieddivorced              9.006e-03  7.573e-03  1.189
## marriagenever_married       -1.730e-03  7.144e-03 -0.242
## meals_fastfood                8.205e-04  1.207e-03  0.680
## sleep_troublesometimes       8.224e-05  2.936e-03  0.028
## sleep_troubleoften           8.354e-03  4.983e-03  1.677
## sleep_troublealways          9.883e-03  4.490e-03  2.201
## tv_hrs2~4hrs:depressedsometimes 8.323e-03  7.516e-03  1.107
## tv_hrs5hrs:depressedsometimes 1.577e-02  9.635e-03  1.636
## tv_hrs0hr:depressedsometimes 3.996e-02  2.244e-02  1.781
## tv_hrs2~4hrs:depressedoften   1.555e-02  1.575e-02  0.987
## tv_hrs5hrs:depressedoften     7.323e-02  2.011e-02  3.640
## tv_hrs0hr:depressedoften      2.634e-03  2.919e-02  0.090
## tv_hrs2~4hrs:depresseddalwyas -1.618e-02  1.735e-02 -0.933
## tv_hrs5hrs:depresseddalwyas -1.862e-02  1.884e-02 -0.988
## tv_hrs0hr:depresseddalwyas -6.512e-02  6.578e-02 -0.990

```

## famincome25k~55k:genderfemale	-1.037e-02	6.296e-03	-1.648
## famincome55k~75k:genderfemale	-1.300e-02	8.367e-03	-1.554
## famincome75k~100k:genderfemale	-1.710e-02	8.864e-03	-1.929
## famincomeabove 100k:genderfemale	-2.853e-02	7.044e-03	-4.050
## raceother hispanic:genderfemale	-1.402e-02	1.084e-02	-1.293
## racewhite:genderfemale	-8.346e-03	7.754e-03	-1.076
## raceblack:genderfemale	8.942e-03	8.740e-03	1.023
## raceasian:genderfemale	-2.434e-02	1.083e-02	-2.248
## raceother:genderfemale	-2.078e-02	1.569e-02	-1.324
## smokesometimes:genderfemale	-2.045e-03	1.334e-02	-0.153
## smokeno:genderfemale	-4.893e-03	7.585e-03	-0.645
## smokemissing:genderfemale	-2.318e-02	7.008e-03	-3.308
## raceother hispanic:activityNo	-1.373e-02	1.077e-02	-1.275
## racewhite:activityNo	-7.821e-03	7.703e-03	-1.015
## raceblack:activityNo	-2.208e-02	8.838e-03	-2.498
## raceasian:activityNo	-3.017e-02	1.062e-02	-2.842
## raceother:activityNo	-1.405e-02	1.512e-02	-0.929
## smokesometimes:marriagewidowed	2.820e-02	3.554e-02	0.793
## smokeno:marriagewidowed	-3.129e-03	2.063e-02	-0.152
## smokemissing:marriagewidowed	3.382e-02	2.023e-02	1.672
## smokesometimes:marriagedivorced	2.225e-02	1.785e-02	1.247
## smokeno:marriagedivorced	-1.047e-02	9.743e-03	-1.075
## smokemissing:marriagedivorced	-1.556e-02	9.094e-03	-1.711
## smokesometimes:marriagenever_married	2.818e-02	1.527e-02	1.845
## smokeno:marriagenever_married	4.362e-03	1.032e-02	0.423
## smokemissing:marriagenever_married	-1.494e-02	8.293e-03	-1.802
## raceother hispanic:meals_fastfood	9.099e-04	1.922e-03	0.473
## racewhite:meals_fastfood	8.098e-04	1.332e-03	0.608
## raceblack:meals_fastfood	-3.007e-03	1.432e-03	-2.101
## raceasian:meals_fastfood	-1.690e-03	2.511e-03	-0.673
## raceother:meals_fastfood	-4.747e-03	2.427e-03	-1.956
## genderfemale:meals_fastfood	2.647e-03	9.281e-04	2.852
## depressedsometimes:milkRare	1.830e-02	1.047e-02	1.749
## depressedoften:milkRare	-6.829e-03	2.010e-02	-0.340
## depressedalways:milkRare	3.656e-02	2.106e-02	1.736
## depressedsometimes:milkSometimes	7.195e-03	9.558e-03	0.753
## depressedoften:milkSometimes	3.048e-02	1.904e-02	1.601
## depressedalways:milkSometimes	-1.515e-02	1.937e-02	-0.782
## depressedsometimes:milkOften	1.525e-02	9.384e-03	1.625
## depressedoften:milkOften	2.109e-02	1.767e-02	1.193
## depressedalways:milkOften	3.418e-02	1.905e-02	1.794
## depressedsometimes:milkVaried	1.003e-01	6.727e-02	1.491
## depressedoften:milkVaried	NA	NA	NA
## depressedalways:milkVaried	NA	NA	NA
## tv_hrs2~4hrs:log.eatout	-1.934e-03	1.863e-03	-1.038
## tv_hrs5hrs:log.eatout	3.814e-03	2.370e-03	1.609
## tv_hrs0hr:log.eatout	-3.707e-04	5.890e-03	-0.063
##	Pr(> t )		

```

## (Intercept) < 2e-16 ***
## raceother hispanic 0.543892
## racewhite 0.142916
## raceblack 0.210948
## raceasian 0.010872 *
## raceother 0.178922
## tv_hrs2~4hrs 0.020088 *
## tv_hrs5hrs 0.527863
## tv_hrs0hr 0.657737
## smokesometimes 0.621020
## smokeno 1.60e-07 ***
## smokemissing 4.58e-09 ***
## depressedsometimes 0.040052 *
## depressedoften 0.505640
## depresseddalwyas 0.727259
## milkrare 0.235859
## milksometimes 0.037947 *
## milkoften 0.975925
## milkvaried 0.019130 *
## famincome25k~55k 0.019794 *
## famincome55k~75k 0.225075
## famincome75k~100k 0.231208
## famincomeabove 100k 0.629275
## genderfemale 0.000338 ***
## log.eatout 0.046259 *
## activityNo 0.009063 **
## alcoholfrq 0.007533 **
## marriagewidowed 0.189031
## marrieddivorced 0.234445
## marriagenever_married 0.808701
## meals_fastfood 0.496494
## sleep_troublesometimes 0.977659
## sleep_troubleoften 0.093744 .
## sleep_troublealways 0.027823 *
## tv_hrs2~4hrs:depressedsometimes 0.268203
## tv_hrs5hrs:depressedsometimes 0.101873
## tv_hrs0hr:depressedsometimes 0.075022 .
## tv_hrs2~4hrs:depressedoften 0.323580
## tv_hrs5hrs:depressedoften 0.000277 ***
## tv_hrs0hr:depressedoften 0.928104
## tv_hrs2~4hrs:depresseddalwyas 0.351029
## tv_hrs5hrs:depresseddalwyas 0.323085
## tv_hrs0hr:depresseddalwyas 0.322296
## famincome25k~55k:genderfemale 0.099523 .
## famincome55k~75k:genderfemale 0.120323
## famincome75k~100k:genderfemale 0.053776 .
## famincomeabove 100k:genderfemale 5.25e-05 ***
## raceother hispanic:genderfemale 0.195981

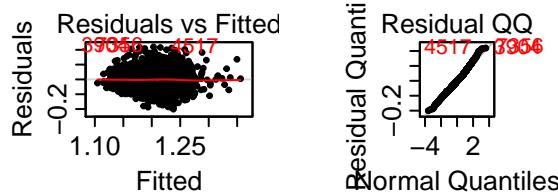
```

```

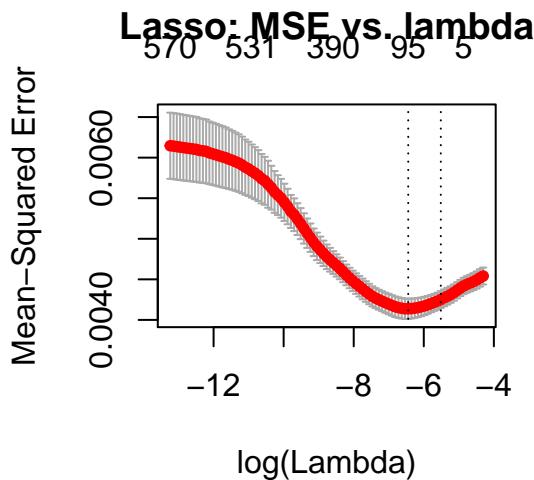
## racewhite:genderfemale          0.281832
## raceblack:genderfemale         0.306336
## raceasian:genderfemale         0.024639 *
## raceother:genderfemale         0.185608
## smokesometimes:genderfemale   0.878136
## smokeno:genderfemale          0.518889
## smokemissing:genderfemale    0.000952 ***
## raceother hispanic:activityNo 0.202484
## racewhite:activityNo          0.310063
## raceblack:activityNo          0.012535 *
## raceasian:activityNo          0.004516 **
## raceother:activityNo          0.352884
## smokesometimes:marriagewidowed 0.427572
## smokeno:marriagewidowed      0.879435
## smokemissing:marriagewidowed 0.094677 .
## smokesometimes:marriagedivorced 0.212546
## smokeno:marriagedivorced     0.282530
## smokemissing:marriagedivorced 0.087206 .
## smokesometimes:marriagenever_married 0.065179 .
## smokeno:marriagenever_married 0.672569
## smokemissing:marriagenever_married 0.071681 .
## raceother hispanic:meals_fastfood 0.635912
## racewhite:meals_fastfood       0.543311
## raceblack:meals_fastfood      0.035759 *
## raceasian:meals_fastfood      0.500897
## raceother:meals_fastfood      0.050541 .
## genderfemale:meals_fastfood   0.004371 **
## depressedsometimes:milkrare    0.080461 .
## depressedoften:milkrare        0.734057
## depresseddalwyas:milkrare     0.082692 .
## depressedsometimes:milksometimes 0.451671
## depressedoften:milksometimes   0.109455
## depresseddalwyas:milksometimes 0.434429
## depressedsometimes:milkoften    0.104281
## depressedoften:milkoften        0.232896
## depresseddalwyas:milkoften     0.072965 .
## depressedsometimes:milkvaried   0.136021
## depressedoften:milkvaried       NA
## depresseddalwyas:milkvaried    NA
## tv_hours2~4hrs:log.eatout     0.299346
## tv_hours5hrs:log.eatout       0.107636
## tv_hours0hr:log.eatout       0.949828
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06233 on 2871 degrees of freedom
## Multiple R-squared: 0.1705, Adjusted R-squared: 0.1453
## F-statistic: 6.782 on 87 and 2871 DF, p-value: < 2.2e-16

```

model1d...



```
## [1] "model1d cv results:"  
## [1] "RMSE: 0.0635994339787107 Rsquared: 0.116642266562635"  
  
#compare model1d to model1b  
set.seed(1234)  
anova(model1b, model1d)  
  
## Analysis of Variance Table  
##  
## Model 1: log2.bmi ~ gender + race + marriage + famincome + alcoholfrq +  
##           log.eatout + milk + meals_fastfood + depressed + sleep_trouble +  
##           activity + tv_hrs + sleep_hr + smoke + I(sleep_hr^2)  
## Model 2: log2.bmi ~ race + tv_hrs + smoke + depressed + milk + famincome +  
##           gender + log.eatout + activity + alcoholfrq + marriage +  
##           meals_fastfood + sleep_trouble + tv_hrs:depressed + famincome:gender +  
##           race:gender + smoke:gender + race:activity + smoke:marriage +  
##           race:meals_fastfood + gender:meals_fastfood + depressed:milk +  
##           tv_hrs:log.eatout  
##   Res.Df     RSS Df Sum of Sq    F    Pr(>F)  
## 1  2923 11.889  
## 2  2871 11.153 52  0.73587 3.6427 < 2.2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
#lasso cv  
set.seed(1234)  
x.lasso = model.matrix(update(full_intr_form, .~.+0), data=df_adult)  
cv_lasso = cv.glmnet(x.lasso, df_adult$log2.bmi, nfolds = 5)  
  
lasso.lambda.min = cv_lasso$lambda.min  
lasso.mse.min = min(cv_lasso$cvm)  
  
coef.lasso = coef(cv_lasso, s = "lambda.min")  
nonzero_coef_lasso = cbind(rownames(coef.lasso)[which(coef.lasso != 0)], coef.lasso[which(coef.lasso != 0)])  
  
#lasso  
model_lasso = glmnet(x.lasso, df_adult$log2.bmi, lambda = lasso.lambda.min)  
  
#plot  
plot(cv_lasso, main="Lasso: MSE vs. lambda")
```

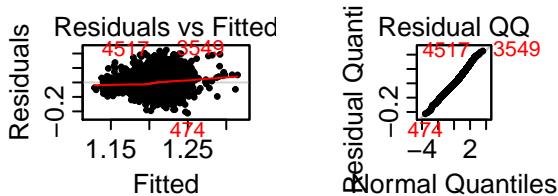


```
summary(model_lasso)
```

```
##          Length Class      Mode
## a0            1   -none- numeric
## beta         594   dgCMatrix S4
## df            1   -none- numeric
## dim           2   -none- numeric
## lambda        1   -none- numeric
## dev.ratio     1   -none- numeric
## nulldev       1   -none- numeric
## npasses       1   -none- numeric
## jerr           1   -none- numeric
## offset         1   -none- logical
## call           4   -none- call
## nobs           1   -none- numeric
```

```
plotres(model_lasso, which=3:4, caption = "Full interaction Lasso model assumption check")
```

Full int...



```
lasso.mse.min
```

```
## [1] 0.004137175
#ridge cv
set.seed(1234)
x.ridge = model.matrix(update(full_intr_form, .~.+0), data=df_adult)
cv_ridge = cv.glmnet(x.ridge, df_adult$log2.bmi, alpha = 0, nfolds = 5)

ridge.lambda.min = cv_ridge$lambda.min
```

```

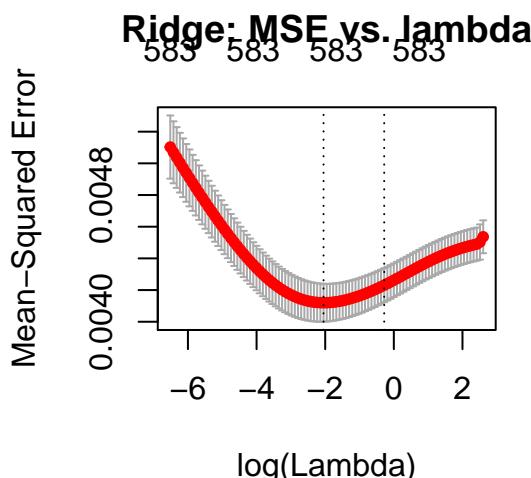
ridge.mse.min = min(cv_ridge$cvm)

coef.ridge = coef(cv_ridge, s = "lambda.min")
nonzero_coef_ridge = cbind(rownames(coef.ridge)[which(coef.ridge != 0)], coef.ridge[which(coef.ridge != 0),])

#ridge
model_ridge = glmmnet(x.ridge, df_adult$log2.bmi, lambda = ridge.lambda.min, alpha = 0)

#plot
plot(cv_ridge, main="Ridge: MSE vs. lambda")

```

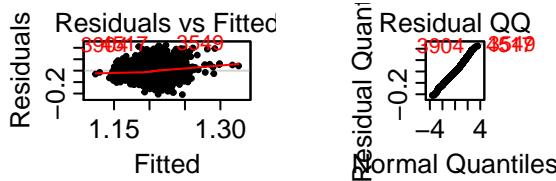


```
summary(model_ridge)
```

	Length	Class	Mode
## a0	1	-none-	numeric
## beta	594	dgCMatrix	S4
## df	1	-none-	numeric
## dim	2	-none-	numeric
## lambda	1	-none-	numeric
## dev.ratio	1	-none-	numeric
## nulldev	1	-none-	numeric
## npasses	1	-none-	numeric
## jerr	1	-none-	numeric
## offset	1	-none-	logical
## call	5	-none-	call
## nobs	1	-none-	numeric

```
plotres(model_ridge, which=3:4, caption = "Full interaction Lasso model \n assumption check")
```

Full in...  
assump...



```
ridge.mse.min
```

```
## [1] 0.004119922

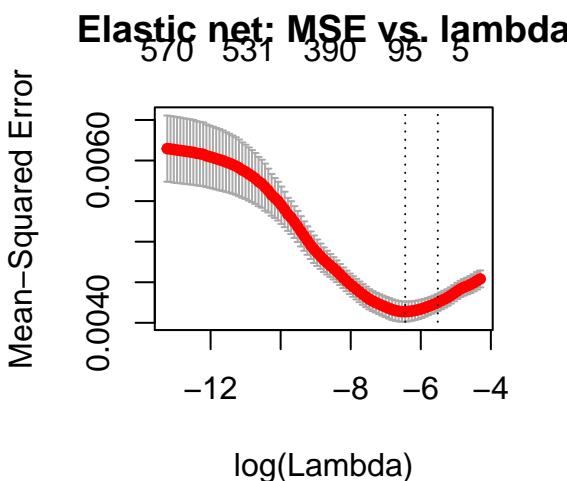
#elastic net cv
set.seed(1234)
x.elastic = model.matrix(update(full_intr_form, .~.+0), data=df_adult)
cv_elastic = cv.glmnet(x.elastic, df_adult$log2.bmi, nfolds = 5)

elastic.lambda.min = cv_elastic$lambda.min
elastic.mse.min = min(cv_elastic$cvm)

coef.elastic = coef(cv_elastic, s = "lambda.min")
nonzero_coef_elastic = cbind(rownames(coef.elastic)[which(coef.elastic != 0)], coef.elastic[which(coef.elastic != 0),])

#elastic net
model_elastic = glmnet(x.elastic, df_adult$log2.bmi, lambda = elastic.lambda.min, alpha=0.5)

#plot
plot(cv_elastic, main="Elastic net: MSE vs. lambda")
```



```
summary(model_elastic)
```

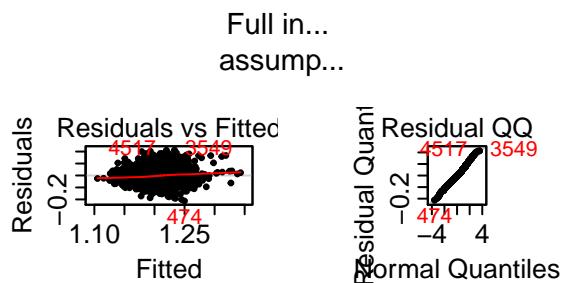
```
##          Length Class      Mode
## a0            1   -none-   numeric
## beta         594    dgCMatrix S4
## df            1   -none-   numeric
```

```

## dim      2 -none- numeric
## lambda   1 -none- numeric
## dev.ratio 1 -none- numeric
## nulldev   1 -none- numeric
## npasses   1 -none- numeric
## jerr      1 -none- numeric
## offset    1 -none- logical
## call      5 -none- call
## nobs      1 -none- numeric

plotres(model_elastic, which=3:4, caption = "Full interaction Elastic model \n assumption check")

```



```
elastic.mse.min
```

```

## [1] 0.004137175

# best model (model1d) bootstrapping

# Bootstrap 95% CI for regression coefficients
set.seed(1234)
library(boot)

## 
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
## 
##     melanoma

# function to obtain regression weights
bs <- function(formula, data, indices) {
  d <- data[indices,] # allows boot to select sample
  fit <- lm(formula, data=d)
  return(coef(fit))
}
# bootstrapping with 1000 replications
nsim = 500
results <- boot(data=df_adult, statistic=bs, R=nsim, formula=formula(model1d))
names_ = names(coef(lm(formula(model1d), data=df_adult)))

ci = data.frame(matrix(ncol = 3, nrow = length(names_)))

```

```

print("the following coef is statistically significant from 0:")

## [1] "the following coef is statistically significant from 0:"

for (i in c(1:85, 88:90)){
  ci[i, 1:2] = boot.ci/boot.out=results, conf = 0.95, type = "perc", index= i)$perc[,4:5]
  ci[i, 3] = (0 < ci[i, 2]) & (0 > ci[i, 1])
  if (ci[i, 3] == FALSE) {print(names_[i])}
}

## [1] "(Intercept)"
## [1] "raceasian"
## [1] "tv_hrs2~4hrs"
## [1] "smokeno"
## [1] "smokemissing"
## [1] "milksometimes"
## [1] "milkvaried"
## [1] "famincome25k~55k"
## [1] "genderfemale"
## [1] "log.eatout"
## [1] "activityNo"
## [1] "alcoholfrq"
## [1] "tv_hrs5hrs:depressedoften"
## [1] "tv_hrs0hr:depressedalwyas"
## [1] "famincomeabove 100k:genderfemale"
## [1] "raceasian:genderfemale"
## [1] "smokemissing:genderfemale"
## [1] "raceblack:activityNo"
## [1] "raceasian:activityNo"
## [1] "raceblack:meals_fastfood"
## [1] "raceother:meals_fastfood"
## [1] "genderfemale:meals_fastfood"
## [1] "depressedsometimes:milkvaried"

## [1] "Multinomial classification base model cv accuracy: 0.490360456372346"
## [1] "Multinomial classification base sleep model cv accuracy: 0.483256732927276"
## [1] "Multinomial classification model1d cv results: 0.468077371933699"
## [1] "Ordinal classification base model cv accuracy: 0.480552885021367"

## Fitting ordinalNet on full training data
## Fitting ordinalNet on fold 1 of 5
## Fitting ordinalNet on fold 2 of 5
## Fitting ordinalNet on fold 3 of 5
## Fitting ordinalNet on fold 4 of 5
## Fitting ordinalNet on fold 5 of 5
## Done

## [1] "Ordinal classification base model cv accuracy: 0.412638907943476"

```

```
write.csv(data.frame(summary(model1)$coefficients), file="table1.csv")
write.csv(data.frame(summary(model1a)$coefficients), file="table1a.csv")
write.csv(data.frame(summary(model1b)$coefficients), file="table1b.csv")
write.csv(data.frame(summary(model1c)$coefficients), file="table1c.csv")

write.csv(data.frame(nonzero_coef_lasso), file="table_lasso.csv")
write.csv(data.frame(nonzero_coef_ridge), file="table_ridge.csv")
write.csv(data.frame(nonzero_coef_elastic), file="table_elastic.csv")
```