BRFSS Health Survey - MLR

Justin Weltz and Andrew Brown

3/4/2018

Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) is conducted by the Centers for Disease Control (CDC) on the United States Population (and is supposed to capture the noninstitutionalized adult population older than 18 years residing in the United States). The public data set contains 486,303 observations. Each of these rows are individuals contacted by telephone (this biases the population they are sampling from and may make inferences taken from this study non-applicable to the general US population). There are 279 accessible variables (a lot of demographic information is omitted in order to preserve anonymity) on demographic characteristics, health-related risk behaviors, chronic health conditions, and use of preventative services. However, we will only be studying a subset of these dimensions.

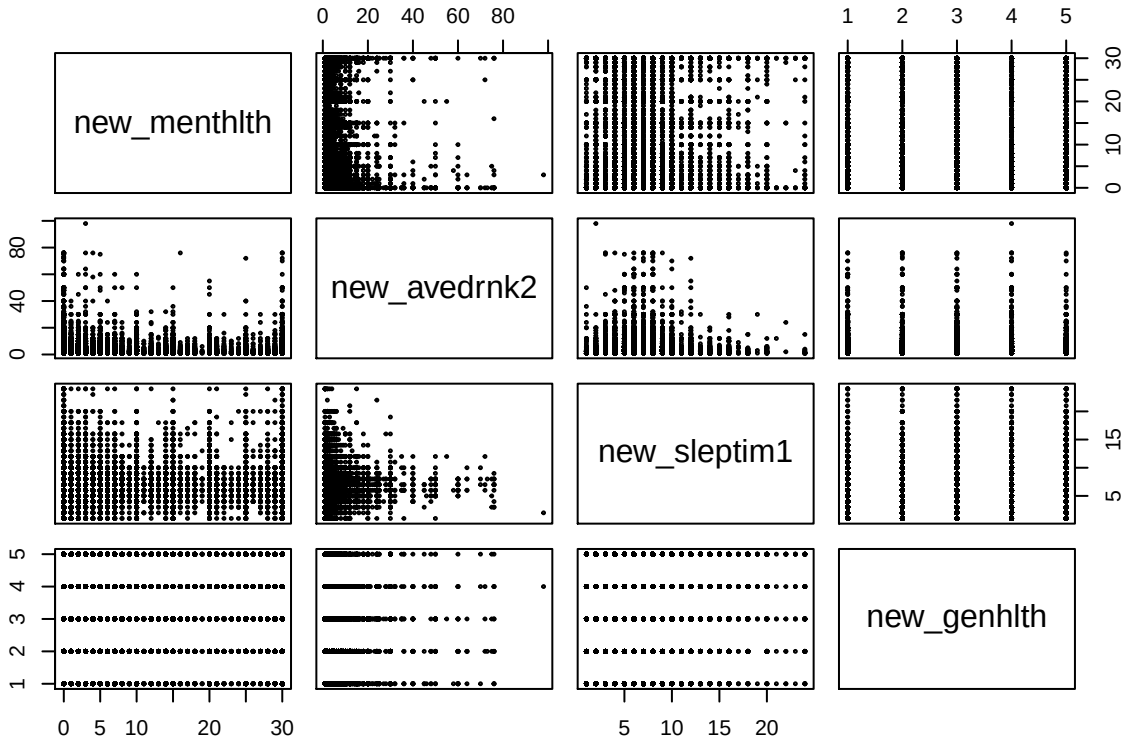
Model Building

The BRFSS Data Set contains a variable called "Number of Days Mental Health Not Good," which is the interviewee's numerical response to the question: "Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?" We decided to regress this continuous variable on five explanatory variables. However, before moving forward, it is important to acknowledge that there is an interesting (and possibly problematic) imbalance to the days of bad mental health variable: 329,500 people report that they have experienced zero days of bad mental health in the past 30 days. This may pull our linear regressions in weird directions later in the project. But now, the explanatory variables:

1. Average Drinks - Continuous (This has distinct outliers; some participants reported that they have as many as 50 drinks a day).
2. General Health - Categorical (We coded this as a factor variable because, while it is on an incrementing scale, the responses are descriptive and therefore hard to quantatify relative to one another).
3. Average Hours Slept - Continuous (This variable is pretty uniformly distributed around a mean of 7.054 hours).
4. Sex - Categorical (The paricipants are pretty evenly distributed between mena and women).
5. Veteran Status - Categorical (The relationship between veteran status and Post Traumatic Stress Disorder may make it an interesting variable to analyze in the context of mental health).

Pairs Plot

```
pairs(BRFSS[c(278, 280, 277, 281)], cex=0.3)
```



Considering the number of observations we are analyzing, observations on a scatter plot should be taken with a lot of salt because it is very to distinguish the density of points with the naked eye. For example, the graphs associated with general health contain very little information because all the observations are compressed into 5 lines (We didn't even include the veteran and sex variable in the pairs plot because they had the same problem). Both the daily drinks variable and the sleep time seem to have a roughly parabolic relationship with bad mental health days variable and with eachother. In fact, it is interesting to note that daily drinks and daily sleep time have a very similar relationship to mental health. However, this does not mean that we should fit a parabolic term for each variable necessarily. In fact, these sideways parabolas seem to reflect a non-functional relationship (it fails the vertical line test), but again this hard to eyeball since we don't have a feel for the density of the points at this time. Lastly, it is interesting to note that there seems to be an outlier in the average drink variable that will be interesting to analyze later as an influential point.

Interaction

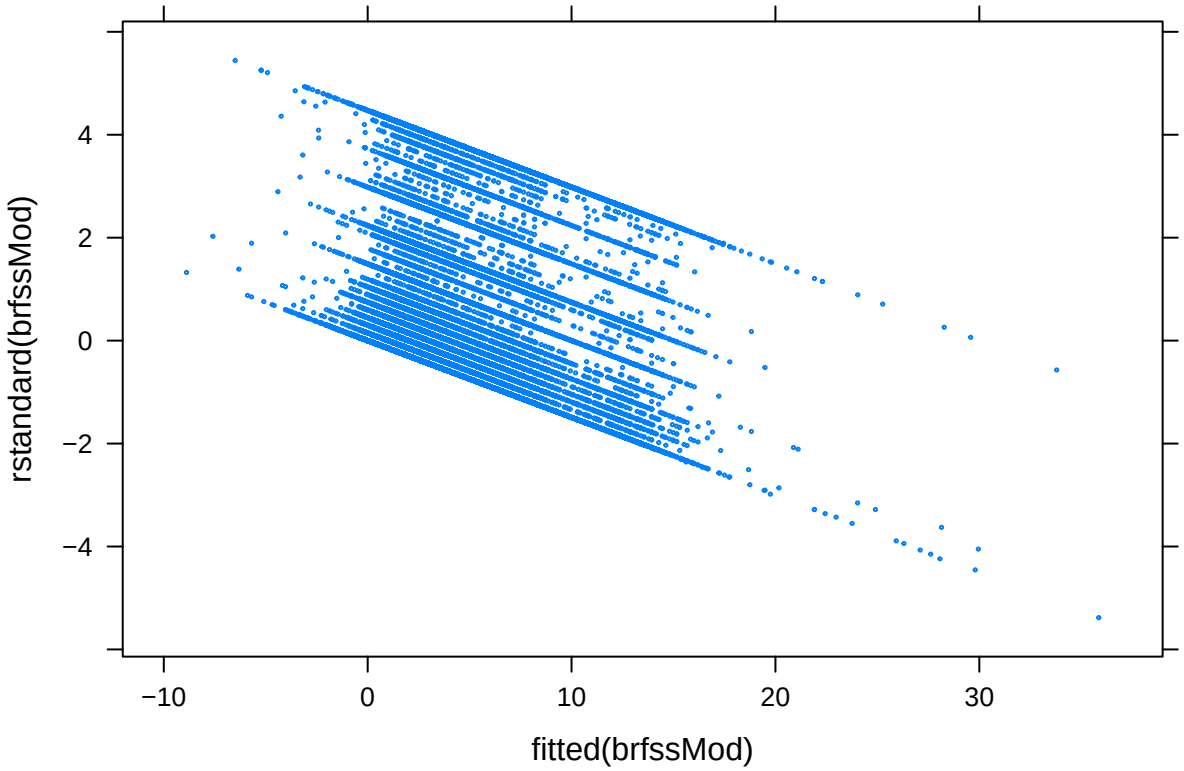
There are a couple interactions that we will include in our model:

1. Veteran and Sex: Male and female soldies may have significantly different experiences (that would have an effect on the prevalence of PTSD or other mental health problems).
2. Veteran and Drink: We hypothesize that being in a generally vulnerable mental state (after war) and then drinking will be significantly different from drinking as a normal citizen.
3. Sex and Drink: Since it seems that society views female and male drinking behavior differently, drinking as a female may be significantly different from drinking as a male (in terms of its correlation with bad mental health days).

Model Fitting

We will first attempt to fit the model normally with the explanatory variables and interaction terms and examine the residual plot.

```
xyplot(rstandard(brfssMod) ~ fitted(brfssMod), cex = 0.2)
```



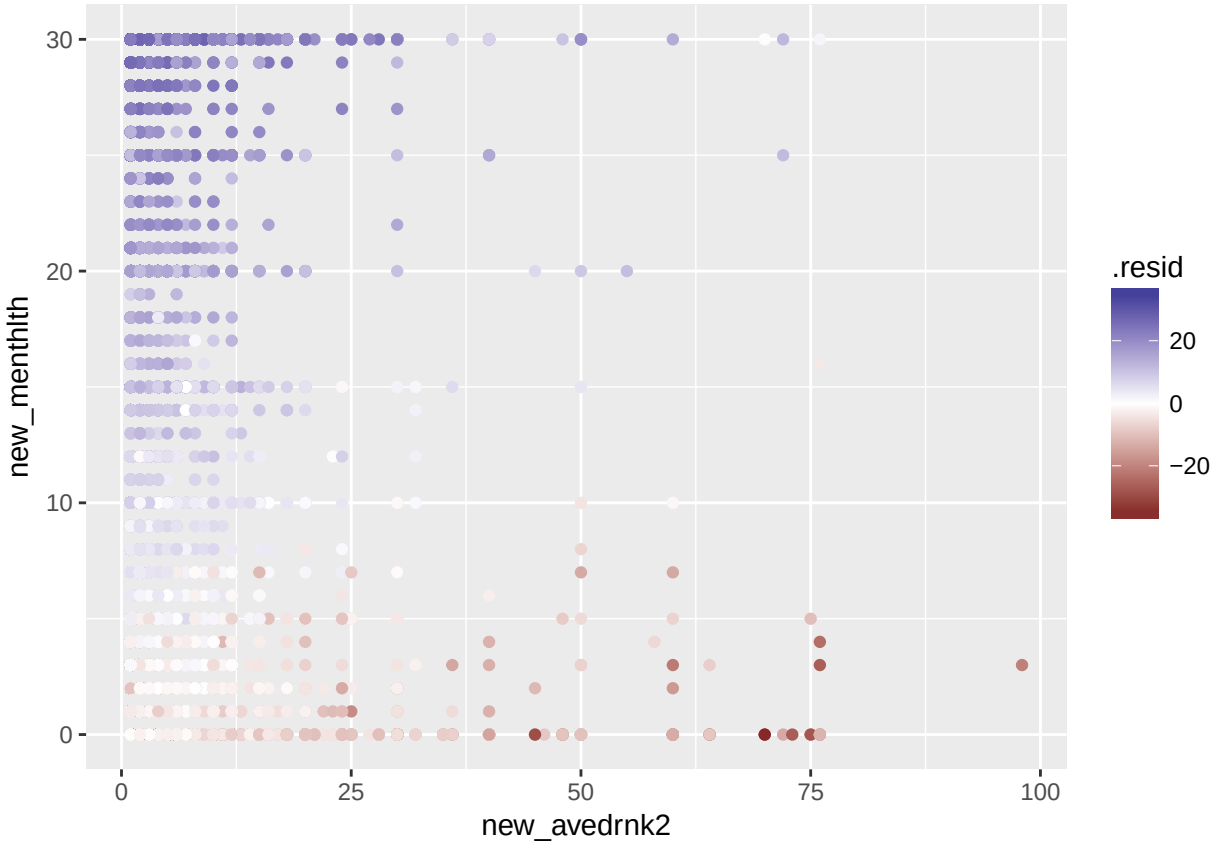
It is clear that there are some problems here. First and foremost, the errors are far from normally distributed. This seems to be an issue with the functional relationship between X and Y - a problem that we would fix by doing some tranformations on X. However, after enumerable attempts (that I won't list here), we were not able to change this residual plot with transformations.

We decided to look more closely at the relationship between our explanatory variables and bad mental health days:

First, we looked more closely at the residuals graphed on the scatterplot in order to get a better sense for how the linear model was interpreting the relationship between the explanatory variables and the response.

The most interesting plot is below:

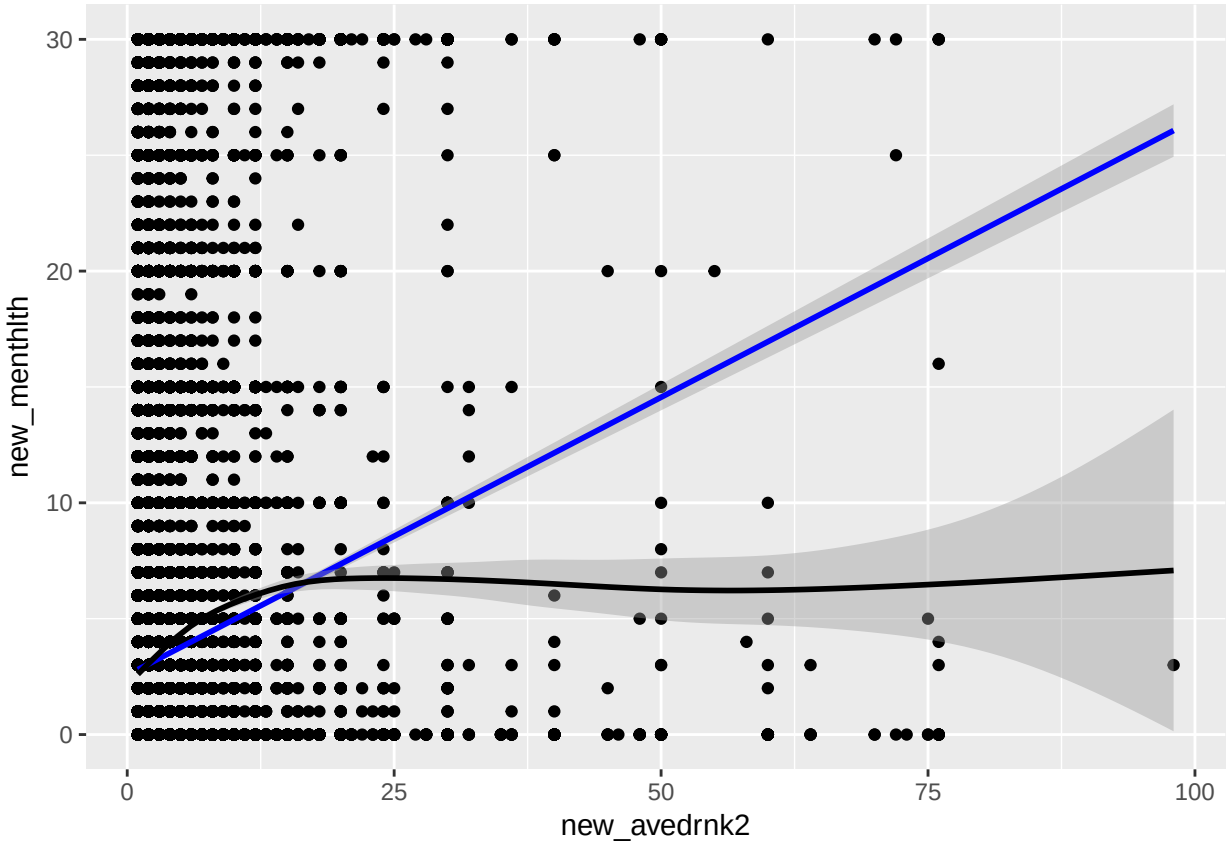
```
ggplot(data = residual_table, aes(x = new_avedrnk2, y = new_menthlth)) +
  geom_point(aes(color = .resid)) + scale_color_gradient2()
```



It is clear from this graph that the linear model believes daily drinks and bad mental health days to be generally positively correlated. This contradicts our initial obeservation that there doesn't seem to be a functional relationship between the two variables. Is this causing the problem? Should there be another variable in the model? Or, is this the true mean relationship between the two variables. It is still not possible to truly answer these questions with the plot above.

Consequently, we try another analysis of the relationship between daily drinks and bad mental health days. If the average of bad mental health days conditional on daily drinks is in fact a positive linear relationship, then a moving weighted average (or spline?), should capture this trend as well. With this in mind, we create the graph below:

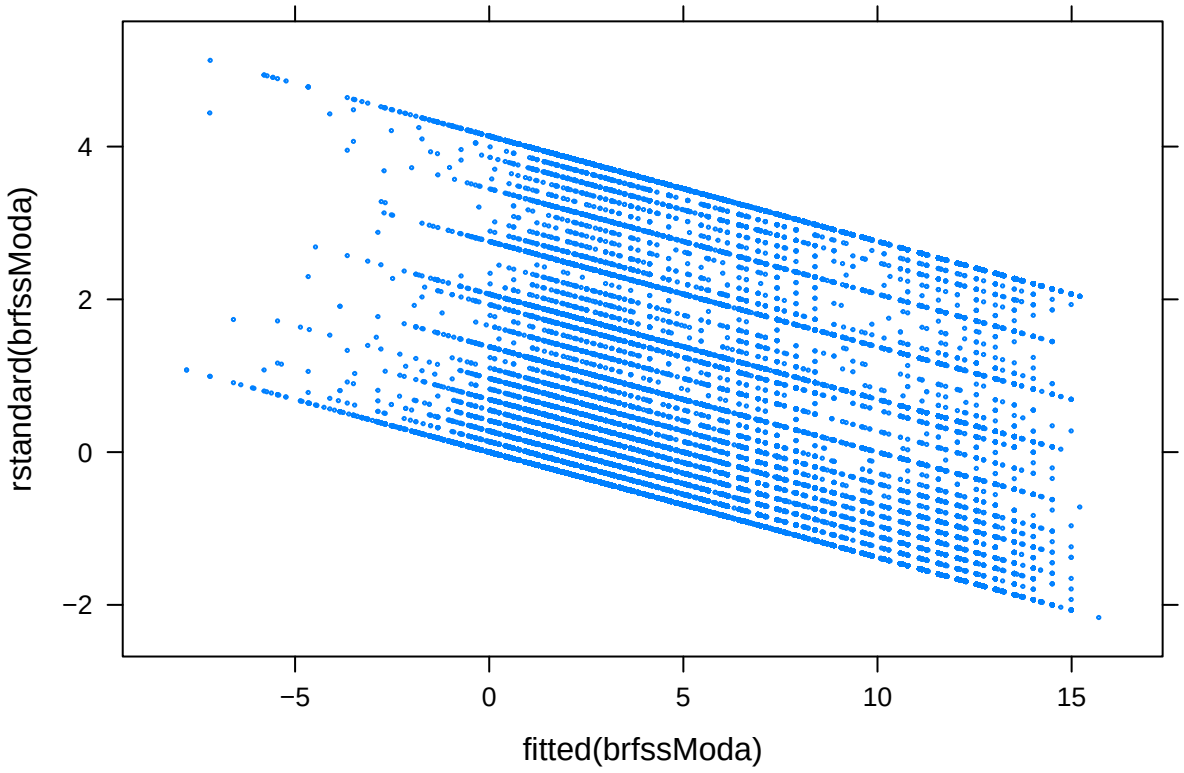
```
ggplot(data=BRFSS, aes(x= new_avedrnk2, y = new_menthlth )) + geom_point() + geom_smooth(method = "lm",
```



This is fascinating. While the linear regression is a positive, straight line (by construct), the spline reflects a different relationship. In the range of about 0-15 daily drinks, the two lines seem to approximate eachother. However, after the initial period, the line continues on its predetermined path while the spline curves into a horizontal line (suggesting very little relationship between the two variables in this higher range). While the spline is not necessarily the truth (this method has its own problems), the large discrepancy between these two methods' results may reflect a larger problem with including this variable in the model.

In this vein, we decide to fit a model without the daily drinks variable and look at the residual plot.

```
brfssModa <- lm(new_menthlth ~ new_genhlth + new_sex + new_veteran3 + new_sleptim1 + new_sex*new_veteran
xyplot(rstandard(brfssModa) ~ fitted(brfssModa), cex = 0.2)
```



The residual plot looks the same! At this point, for fear of banging our heads against too many walls, we move on and hope to find the key to this mysteriously mishappen residual plot later on in the project.

Coefficients and Inferences

```
summary(brfssMod)
```

```
##
## Call:
## lm(formula = new_menthlth ~ new_genhlth + new_sex + new_veteran3 +
##     new_sleptim1 + new_avedrnk2 + new_sex * new_avedrnk2 + new_sex *
##     new_veteran3 + new_veteran3 * new_avedrnk2, data = BRFSS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.860  -2.881  -1.759  -0.253  36.498
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.65894    0.09798  37.343  < 2e-16 ***
## new_genhlth2                 0.65915    0.03821  17.251  < 2e-16 ***
## new_genhlth3                 1.93458    0.04040  47.889  < 2e-16 ***
## new_genhlth4                 5.24423    0.05573  94.108  < 2e-16 ***
## new_genhlth5                10.46393    0.09343 111.999  < 2e-16 ***
## new_sex2                     1.82252    0.13655  13.347  < 2e-16 ***
```

```
## new_veteran32               0.68491    0.05996  11.422  < 2e-16 ***
## new_sleptim1               -0.53117    0.01074 -49.440  < 2e-16 ***
## new_avedrnk2                0.20725    0.01456  14.236  < 2e-16 ***
## new_sex2:new_avedrnk2       0.19039    0.01348  14.125  < 2e-16 ***
## new_sex2:new_veteran32     -0.92371    0.13779  -6.704 2.04e-11 ***
## new_veteran32:new_avedrnk2 -0.04903    0.01629  -3.010  0.00261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.707 on 229175 degrees of freedom
##   (257116 observations deleted due to missingness)
## Multiple R-squared:  0.1099, Adjusted R-squared:  0.1098
## F-statistic:  2571 on 11 and 229175 DF,  p-value: < 2.2e-16
```

Now that we know our residual plot is far from ideal, it is important to take the inferences above with a grain of salt. That being said, this is a very impressive looking coefficient table. Every p-value is essentialy 0, except for two of the interaction variables. However, in order to be conservative, let's just look at the sign of our coefficients so we are less likely to make false conclusions based on magnitudes that we can't properly contextualize with good estimates of standard error.

Sex - Men report more bad mental health days than women. Interesting.

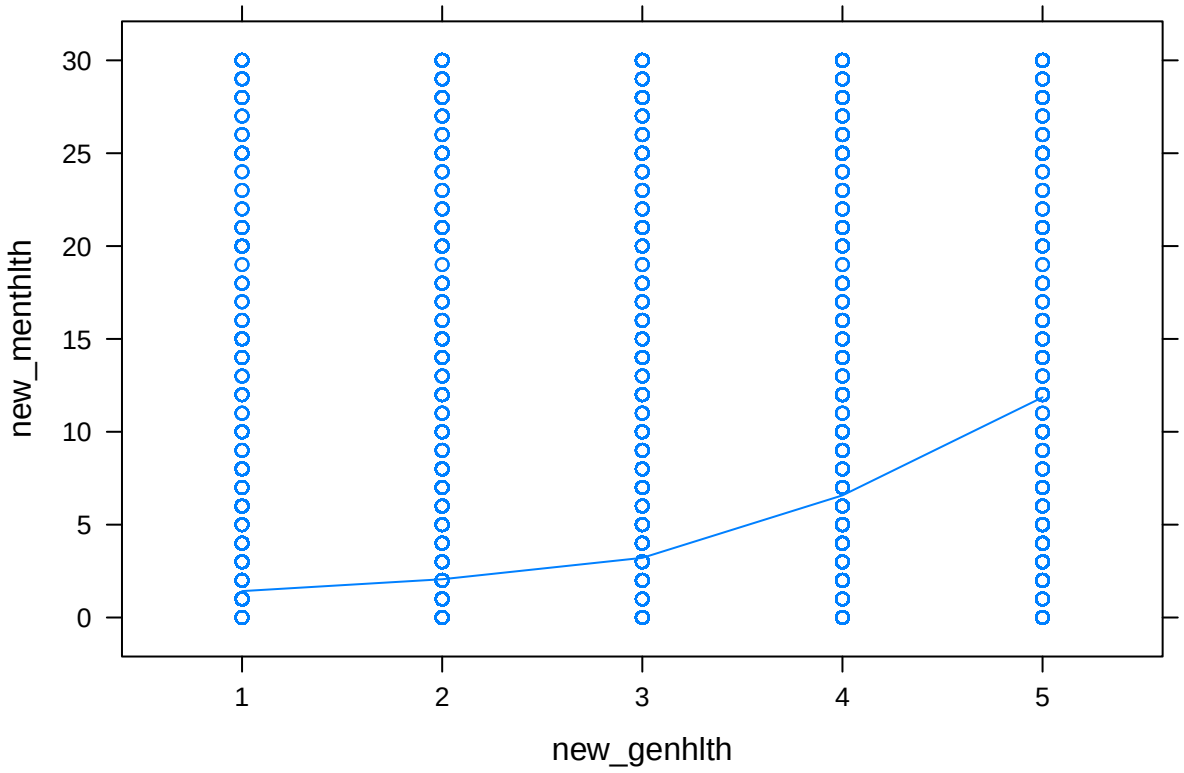
Veterans - Counter to our initial hypothesis, veterans actually report less bad mental health days than non-veterans.

Daily Drinks - As we saw above, bad mental health days and daily drinks are positively correlated.

Sleep Time - Sleep and bad mental health days are negatively correlated, which seems natural given that sleep seems to be good for just about everything.

General Health - This is by far the weirdest finding. It seems that general health is positively correlated with bad mental health days (and with large magnitudes too!). This would mean that reporting you are in a better general state of health is positively correlated with reporting more mental health days. In order to check the validity of this relationship, we connect the mean bad mental health days of each general health category.

```
require(mosaic)
plotModel(lm(new_menthlth ~ new_genhlth, data = BRFSS))
```



The relationship holds! We will comment more on this oddity in our conclusion.

F-test

```
brfssModNoInt <- lm(new_menthlth ~ new_genhlth + new_sex + new_veteran3 + new_sleptim1 + new_avedrnk2, d
anova(brfssMod, brfssModNoInt)
```

```
## Analysis of Variance Table
##
## Model 1: new_menthlth ~ new_genhlth + new_sex + new_veteran3 + new_sleptim1 +
##     new_avedrnk2 + new_sex * new_avedrnk2 + new_sex * new_veteran3 +
##     new_veteran3 * new_avedrnk2
## Model 2: new_menthlth ~ new_genhlth + new_sex + new_veteran3 + new_sleptim1 +
##     new_avedrnk2
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1 229175 10310669
## 2 229178 10321513 -3    -10844 80.343 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this F-test, we can see that at least one of the interaction terms is nonzero, indicating that ……

R^2 Values

```
summary(brfssMod)
```

```
##
## Call:
## lm(formula = new_menthlth ~ new_genhlth + new_sex + new_veteran3 +
##     new_sleptim1 + new_avedrnk2 + new_sex * new_avedrnk2 + new_sex *
##     new_veteran3 + new_veteran3 * new_avedrnk2, data = BRFSS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.860  -2.881  -1.759  -0.253  36.498
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.65894    0.09798  37.343  < 2e-16 ***
## new_genhlth2                  0.65915    0.03821  17.251  < 2e-16 ***
## new_genhlth3                  1.93458    0.04040  47.889  < 2e-16 ***
## new_genhlth4                  5.24423    0.05573  94.108  < 2e-16 ***
## new_genhlth5                 10.46393    0.09343 111.999  < 2e-16 ***
## new_sex2                      1.82252    0.13655  13.347  < 2e-16 ***
## new_veteran32                 0.68491    0.05996  11.422  < 2e-16 ***
## new_sleptim1                 -0.53117    0.01074 -49.440  < 2e-16 ***
## new_avedrnk2                  0.20725    0.01456  14.236  < 2e-16 ***
## new_sex2:new_avedrnk2         0.19039    0.01348  14.125  < 2e-16 ***
## new_sex2:new_veteran32       -0.92371    0.13779  -6.704 2.04e-11 ***
## new_veteran32:new_avedrnk2 -0.04903    0.01629  -3.010  0.00261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.707 on 229175 degrees of freedom
##   (257116 observations deleted due to missingness)
## Multiple R-squared:  0.1099, Adjusted R-squared:  0.1098
## F-statistic:  2571 on 11 and 229175 DF,  p-value: < 2.2e-16
```

While all variables are significant, our overall model is not that precise, as we have a R^2 Value of 0.1099, and an Adjusted-R^2 Value of 0.1098. .....

Residuals

From the residual plot that was given earlier, it was evident

Variable Selection

ASK ABOUT THIS SECTION

```
require(leaps)
```

```
## Loading required package: leaps
```

```
col.best <- regsubsets(new_menthlth ~ new_genhlth + new_sex + new_veteran3 + new_avedrnk2 + new_sleptim
col.best.sum <- summary(col.best)
which.min(col.best.sum$cp)
```

```
## [1] 5
```

```
which.max(col.best.sum$adjr2)
```

```
## [1] 5
```

```
which.min(col.best.sum$bic)
```

```
## [1] 5
```

```
add1(lm(new_menthlth~1, data=BRFSS), new_menthlth ~ new_genhlth + new_sex + new_veteran3 + new_sleptim1
```

```
## Warning in add1.lm(lm(new_menthlth ~ 1, data = BRFSS), new_menthlth ~
## new_genhlth + : using the 229187/478348 rows from a combined fit
```

```
## Single term additions
##
## Model:
## new_menthlth ~ 1
##              Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>                    11583248 899051
## new_genhlth   4    986124 10597124 878667 11128.1 < 2.2e-16 ***
## new_sex       1     64803 11518445 897768  2691.2 < 2.2e-16 ***
## new_veteran3  1     17237 11566011 898712   712.9 < 2.2e-16 ***
## new_sleptim1  1    163217 11420031 895801  6836.6 < 2.2e-16 ***
## new_avedrnk2  1     76699 11506550 897531  3188.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(lm(new_menthlth~new_genhlth, data=BRFSS), new_menthlth ~ new_genhlth + new_sex + new_veteran3 + new
```

```
## Warning in add1.lm(lm(new_menthlth ~ new_genhlth, data = BRFSS),
## new_menthlth ~ : using the 229187/477204 rows from a combined fit
```

```
## Single term additions
##
## Model:
## new_menthlth ~ new_genhlth
##              Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>                    10597124 878667
## new_sex       1     88080 10509045 876756  3999.6 < 2.2e-16 ***
## new_veteran3  1     34844 10562280 877914  1574.3 < 2.2e-16 ***
## new_sleptim1  1    113557 10483567 876200  5169.0 < 2.2e-16 ***
## new_avedrnk2  1     45220 10551904 877689  2045.0 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(lm(new_menthlth~new_genhlth + new_sleptim1, data=BRFSS), new_menthlth ~ new_genhlth + new_sex + new
```

```
## Warning in add1.lm(lm(new_menthlth ~ new_genhlth + new_sleptim1, data =
## BRFSS), : using the 229187/472253 rows from a combined fit
```

```
## Single term additions
##
## Model:
## new_menthlth ~ new_genhlth + new_sleptim1
##              Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>                    10483567 876200
## new_sex       1     92214 10391353 874177  4190.8 < 2.2e-16 ***
```

```
## new_veteran3  1     33412 10450155 875470  1509.9 < 2.2e-16 ***
## new_avedrnk2  1     40274 10443294 875320  1821.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(lm(new_menthlth~new_genhlth + new_sleptim1 + new_sex, data=BRFSS), new_menthlth ~ new_genhlth + new
```

```
## Warning in add1.lm(lm(new_menthlth ~ new_genhlth + new_sleptim1 +
## new_sex, : using the 229187/472198 rows from a combined fit
```

```
## Single term additions
##
## Model:
## new_menthlth ~ new_genhlth + new_sleptim1 + new_sex
##              Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>                    10391353 874177
## new_veteran3  1      7054 10384299 874023  320.74 < 2.2e-16 ***
## new_avedrnk2  1     64924 10326429 872743 2968.76 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(lm(new_menthlth~new_genhlth + new_sleptim1 + new_sex + new_avedrnk2, data=BRFSS), new_menthlth ~ n
```

```
## Warning in add1.lm(lm(new_menthlth ~ new_genhlth + new_sleptim1 + new_sex
## + : using the 229187/229385 rows from a combined fit
```

```
## Single term additions
##
## Model:
## new_menthlth ~ new_genhlth + new_sleptim1 + new_sex + new_avedrnk2
##                      Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>                            10326429 872743
## new_veteran3          1    4915.3 10321513 872635  109.23 < 2.2e-16 ***
## new_sex:new_avedrnk2  1    8076.4 10318352 872565  179.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(lm(new_menthlth~new_genhlth + new_sleptim1 + new_sex + new_avedrnk2 + new_sex*new_avedrnk2, data=B
```

```
## Warning in add1.lm(lm(new_menthlth ~ new_genhlth + new_sleptim1 + new_sex
## + : using the 229187/229385 rows from a combined fit
```

```
## Single term additions
##
## Model:
## new_menthlth ~ new_genhlth + new_sleptim1 + new_sex + new_avedrnk2 +
##     new_sex * new_avedrnk2
##              Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>                    10318352 872565
## new_veteran3  1    5358.7 10312994 872448  119.19 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(lm(new_menthlth~new_genhlth + new_sleptim1 + new_sex + new_avedrnk2 + new_sex*new_avedrnk2 + new_v
```

```
## Single term additions
##
## Model:
```

```
## new_menthlth ~ new_genhlth + new_sleptim1 + new_sex + new_avedrnk2 +
##     new_sex * new_avedrnk2 + new_veteran3
##                            Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>                                   10312994 872448
## new_sex:new_veteran3        1   1916.51 10311077 872408 42.5967 6.741e-11
## new_veteran3:new_avedrnk2   1    302.44 10312691 872443  6.7211  0.009529
##
## <none>
## new_sex:new_veteran3      ***
## new_veteran3:new_avedrnk2 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(lm(new_menthlth~new_genhlth + new_sleptim1 + new_sex + new_avedrnk2 + new_sex*new_avedrnk2 + new_v
```

```
## Single term additions
##
## Model:
## new_menthlth ~ new_genhlth + new_sleptim1 + new_sex + new_avedrnk2 +
##     new_sex * new_avedrnk2 + new_veteran3 + new_sex * new_veteran3
##                            Df Sum of Sq      RSS    AIC F value   Pr(>F)
## <none>                                   10311077 872408
## new_veteran3:new_avedrnk2   1    407.75 10310669 872401  9.0631 0.002609 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#add1(lm(new_menthlth~new_genhlth + new_sleptim1 + new_sex + new_avedrnk2 + new_sex*new_avedrnk2 + new_
```

Through forward selection, we end up with the same variables in the model as we had before. But does this make sense? And why did different variables show up??

Model Interpretation

Confidence Intervals

```
newdata <- data.frame(new_genhlth = 4,
                      new_veteran3 = 1,
                      new_sex = 1,
                      new_avedrnk2 = 10,
                      new_sleptim1 = 5)

newdata$new_genhlth <- as.factor(newdata$new_genhlth)
newdata$new_sex <- as.factor(newdata$new_sex)
newdata$new_veteran3 <- as.factor(newdata$new_veteran3)

predict.lm(brfssMod, newdata, interval = "confidence")
```

```
##       fit     lwr      upr
## 1 8.31983 8.07081 8.568849
```

We are 95% confident that the true mean value for days of poor mental health for a male veteran who had 10 drinks in the past 30 days, gets an average of 5 hours of sleep every 24 hours, and reports fair general health is between 8.07081 and 8.568849 days in the past 30 days.

```r
predict.lm(brfssMod, newdata, interval = "predict")
```

```
##       fit      lwr      upr
## 1 8.31983 -4.82903 21.46869
```

We are 95% confident that a male veteran who had 10 drinks in the past 30 days, gets an average of 5 hours of sleep every 24 hours, and reports fair general health will have had between -4.82903 (0) and 21.46869 days of poor mental health in the past 30 days.

Coefficient of Partial Determination

```r
anova(brfssMod)
```

```
## Analysis of Variance Table
##
## Response: new_menthlth
##                               Df   Sum Sq Mean Sq   F value    Pr(>F)
## new_genhlth                    4   986124  246531 5479.6375 < 2.2e-16 ***
## new_sex                        1    88080   88080 1957.7479 < 2.2e-16 ***
## new_veteran3                   1     8228    8228  182.8897 < 2.2e-16 ***
## new_sleptim1                   1   116517  116517 2589.8183 < 2.2e-16 ***
## new_avedrnk2                   1    62786   62786 1395.5439 < 2.2e-16 ***
## new_sex:new_avedrnk2           1     8520    8520  189.3685 < 2.2e-16 ***
## new_sex:new_veteran3           1     1917    1917   42.5982 6.736e-11 ***
## new_veteran3:new_avedrnk2      1      408     408    9.0631  0.002609 **
## Residuals                 229175 10310669      45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary