BIOSTAT 653 GROUP PROJECT PROPOSAL

Group: 0ops (Zero-outlier-perfectly-smooth)
Aubrey Annis, Douglas Hannum, Pedro Orozco del Pino, Yichen Si
12/10/18

Introduction

Genes for Good (GfG), a University of Michigan social-media-based study, gathers genetic and longitudinal data from participants through mail-order spit kits and online surveys. The surveys fall into two categories: required surveys that collect participants' demographic information and baseline health/behavior, and optional surveys in which participants can contribute longitudinal data for various behavioral phenotypes like mood, physical activity, sleep patterns, and alcohol use. Since the surveys are done online, there is little measurement error in building the databases. However, the surveys are potentially biased since they rely on voluntary responses and self-reported data.

Our study explores the effect alcohol use has on sleep duration in GfG participants. We consider participants' demographic information, sleep-associated genetics, and longitudinal surveys pertaining to alcohol use and sleep duration. The alcohol use survey asks participants how many drinks they had the previous day, with subcategories asking participants for the number and type of drinks consumed (9 categories). The daily sleep survey asks participants how many hours of sleep they got the previous night. We suspect that low alcohol usage (between 1 and 2 drinks) may have a positive effect on sleep duration, whereas alcohol use greater than 3 drinks may have a negative effect. In addition, we will consider if sleep duration varies based on the type of drink consumed and if participants' genetics have a significant effect.

Scientific Questions

1. Does the consumption of alcohol affect the hours of sleep for a participant?
2. What effects do SNPs associated with sleep have on our study population?
3. Do certain types of alcoholic beverages have an effect on hours of sleep?

Data Cleaning

We subset the GfG dataset to only include participants who had filled out the alcohol use and sleep surveys on the same day, had completed over 10 of these surveys, and had available genetic information. After subsetting, we retained 1,167 individuals with a total of 37,624 observations for same-day alcohol use and sleep duration.

To include genetic information in our model, we created a genetic risk score (GRS) for each study participant using effect sizes obtained for 29 SNPs previously found to be significantly associated with sleep duration (Dashti el at.). Data for 28 of the SNPS were available in the GfG dataset, so these were included in the GRS. The GRS is a weighted sum of each participant's alleles for the associated SNPs and ranges between 25.70 and 45.34. Because the effect sizes for our GRS are drawn from a study with exclusively European ethnicity participants, the GRS is

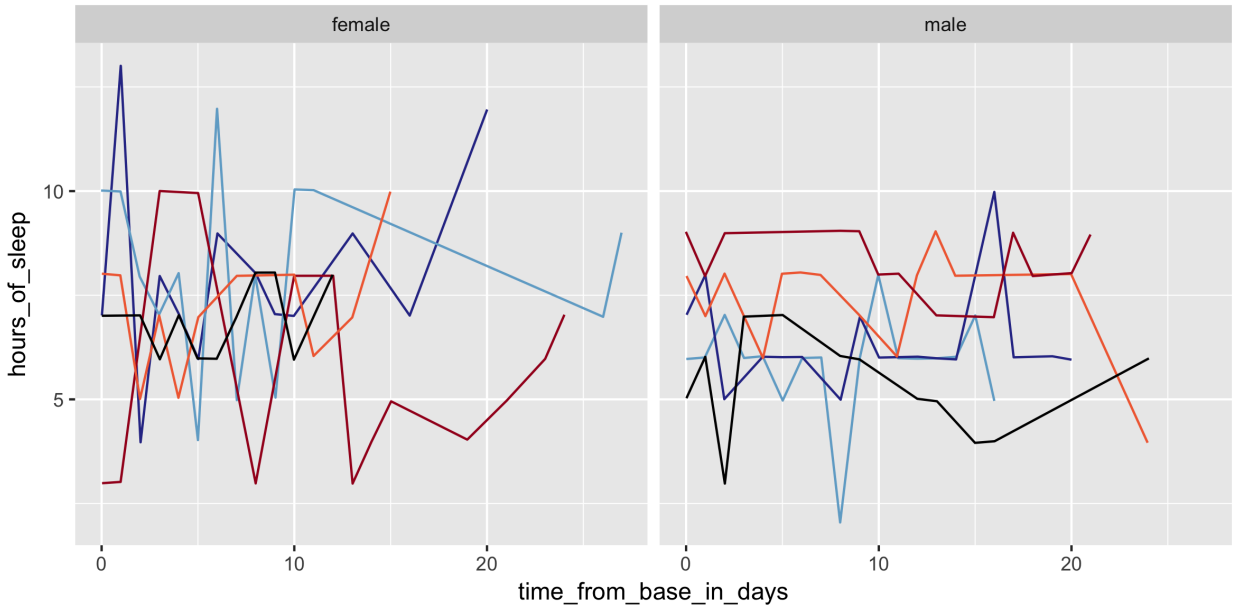
most applicable to the participants with European ancestry (80%) and may not accurately represent association and effect sizes for other participants. In addition, the GRS will not be effective in adjusting for race in our model since it includes only European ancestry genetic information.

To account for alcohol in our model, we chose to bin type of drink into four different categories: beer, wine, liquor, and other. We believe this model choice to be appropriate because we are more interested in determining the effects of hard liquor vs. wine than in determining the effects of bourbon vs. tequila, and binning our data gives us a greater possibility of detecting a significant effect.

EDA

Patterns in longitudinal variables

Below is an example of how the data looks in the sleep and alcohol surveys. In Figure 1, We pick 5 subjects from both genders who have over 10 observations containing alcohol and sleep survey completed on the same day, spanning 20 to 30 days. (The time span restriction is mainly for display purpose; we will include subjects that have different observation time spans in our analysis.) The top row shows the hours of sleep and the bottom shows number of drinks the each individual had on the same day. Female and male are grouped into left and right panel respectively; the same color in the top and bottom figure identifies the same subject.



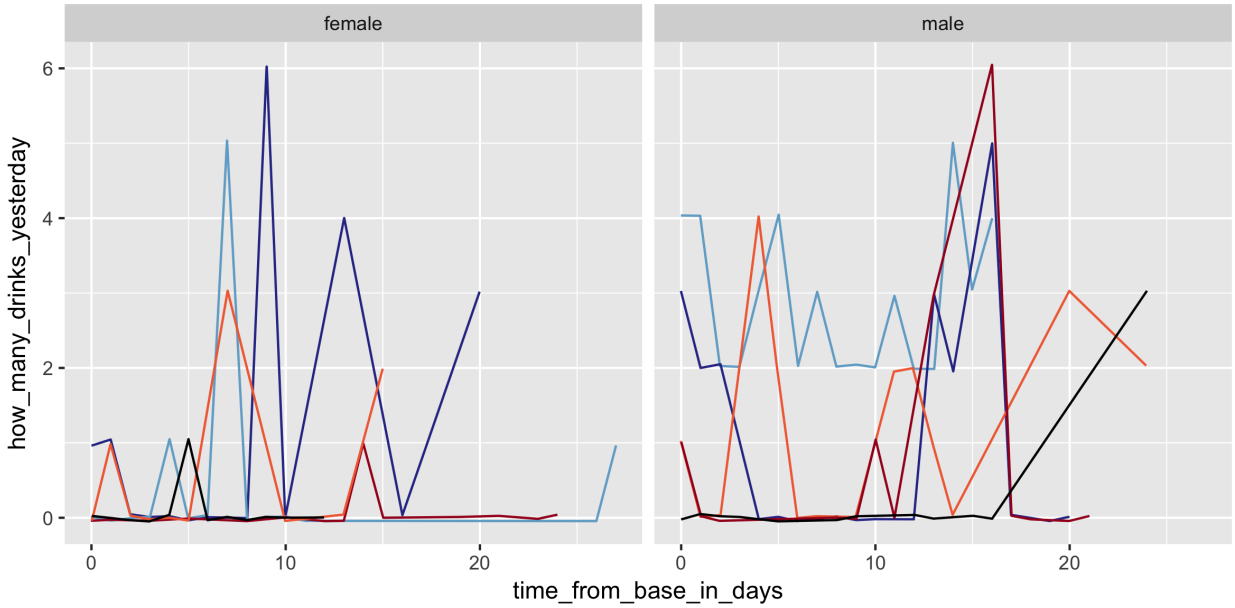


Figure 1. Spaghetti plots for the response variable (hours of sleep the subject got the night before taking the survey) and one explanatory variable (the number of drinks the subject had the day before the survey). X-axis is the time in days, counting from when the first survey is completed.

To further demonstrate the heterogeneity in our data, here are 10 subjects sampled from those having a long span of records and more observations (different from those in Figure 1). The rows (labeled by 1-5) and genders identify the same subjects in the two figures for sleep and drink. Both the time span and intervals between surveys vary greatly across individuals. The records for drinking can be relatively “sparse”, though the reported zero drink is likely to be a true zero. We also included the type of drinks in our model, but here the drinks are not distinguished.

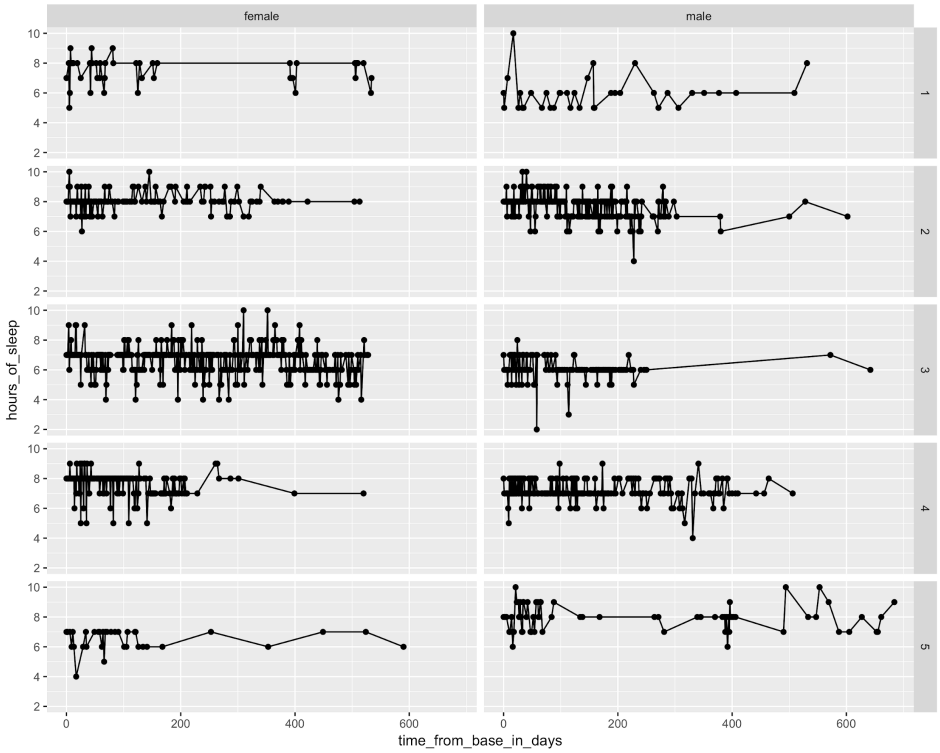


Figure 2.1 Spaghetti plots of the response variable, hours of sleep last night, in subjects having records over a longer time span. Each subplot is one individual.

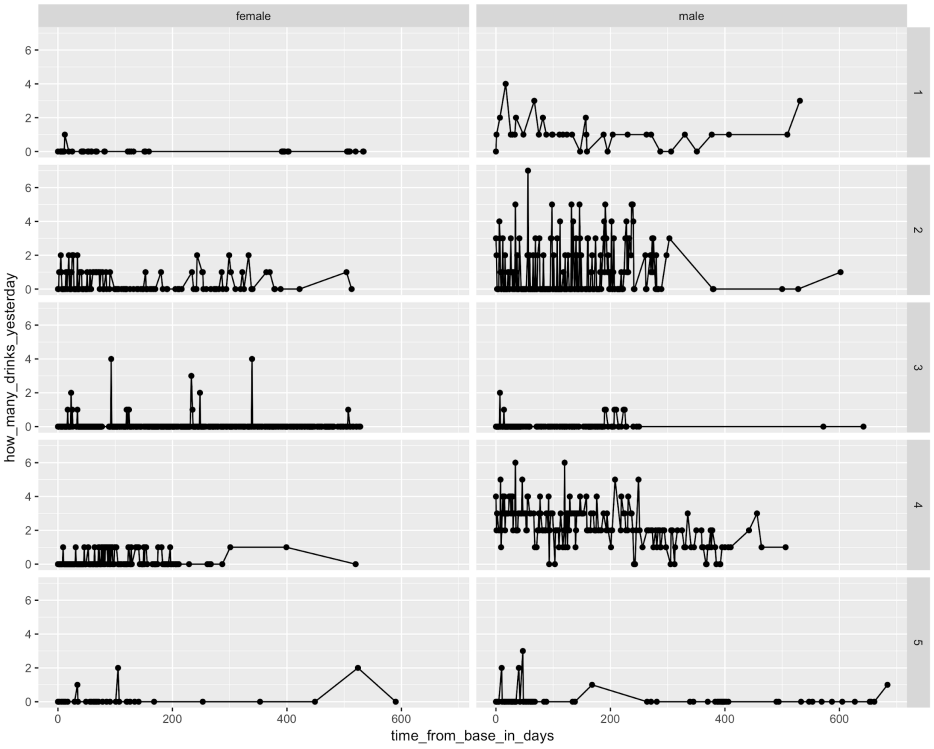


Figure 2.2 Spaghetti plots of the the number of drinks the subjects had last night. Each subplot is one individual, corresponding to the same person in Figure 2.1.

Figure 3 shows the relation between sleeping time and the amount of alcohol consumption in the same day for a few randomly picked subjects. The pattern is not clear, though this is only the marginal relationship.

Considering that the average amount of sleep people need is physiologically limited, we considered the possibility of using the variation of sleeping time during a short period as an alternative response. It also has the potential benefit of reducing the effect of self-reported bias if the bias within each subject is systematic. To model the variation, we would need to define a time window and have enough observations within the window for each subject. Since people take the survey with very different time patterns, we do not have enough data to fully assess this line of thoughts. We put our preliminary analysis of this problem in the supplementary material.

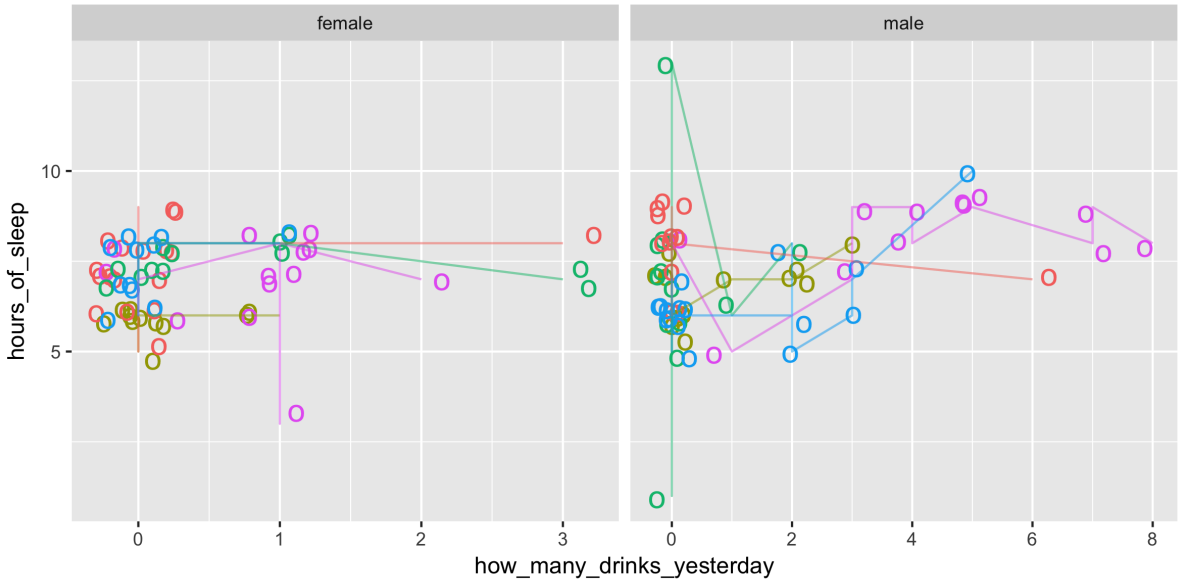


Figure 3. Scatter plot of hours of sleep vs. the number of drinks for 10 random subjects, 5 in each gender. Color identifies individual.

Splines

To assess if a spline in age is necessary, we looked at partial regression plots using number of drinks, beer, wine, liquor, other types of drink, GRS, gender, and weekend as covariates (model described in more detail below). From the scatter plot of the residuals in Figure 4 we cannot distinguish any non-linear pattern in the relation of age and hours of sleep, adjusting for the list of covariates. We also assessed the linear relationship of number of drinks and hours of sleep with using partial regression plots in Figure 5. Again we found no evidence of a non-linear pattern.

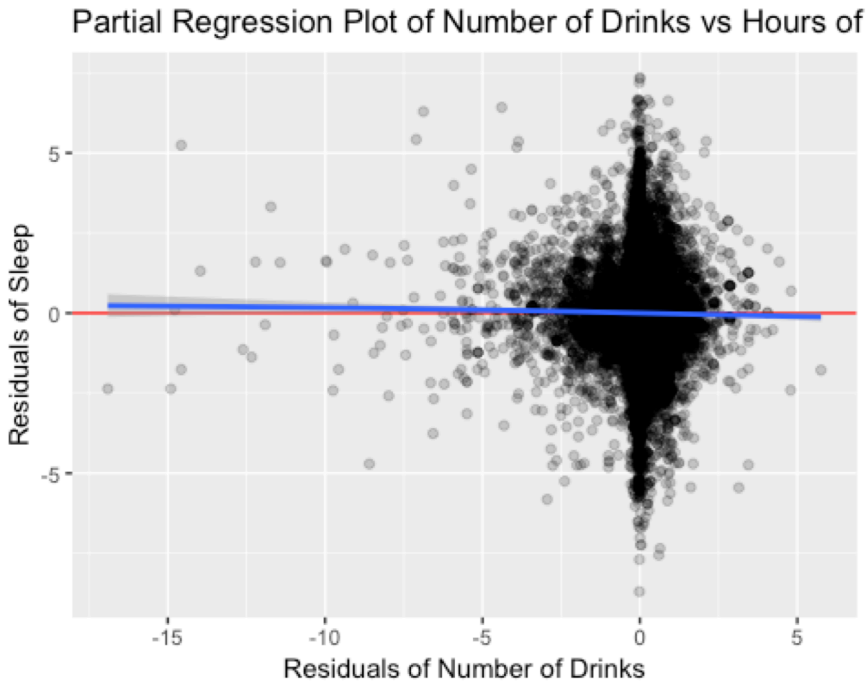


Figure 4. Partial regression plot of Number of drinks vs Hours of Sleep. The red line is a horizontal line around zero and the blue line is the geom_smooth() with linear trend.

One of our first thoughts when fitting the model is that the relationship between age and sleep may not be completely linear. As a first way to see if this is the case we will plot the partial regression plot using as covariates number of drinks, beers, wines, liquors, other type of drink, PRS, gender and weekend. From the scatter plot of the residuals we cannot distinguish any non-linear pattern on the relation of age and hours of sleep adjusting for the list of covariates. However, we wanted to test furthermore the relationship and to do so we fit the model with a 3 knots spline and without it.

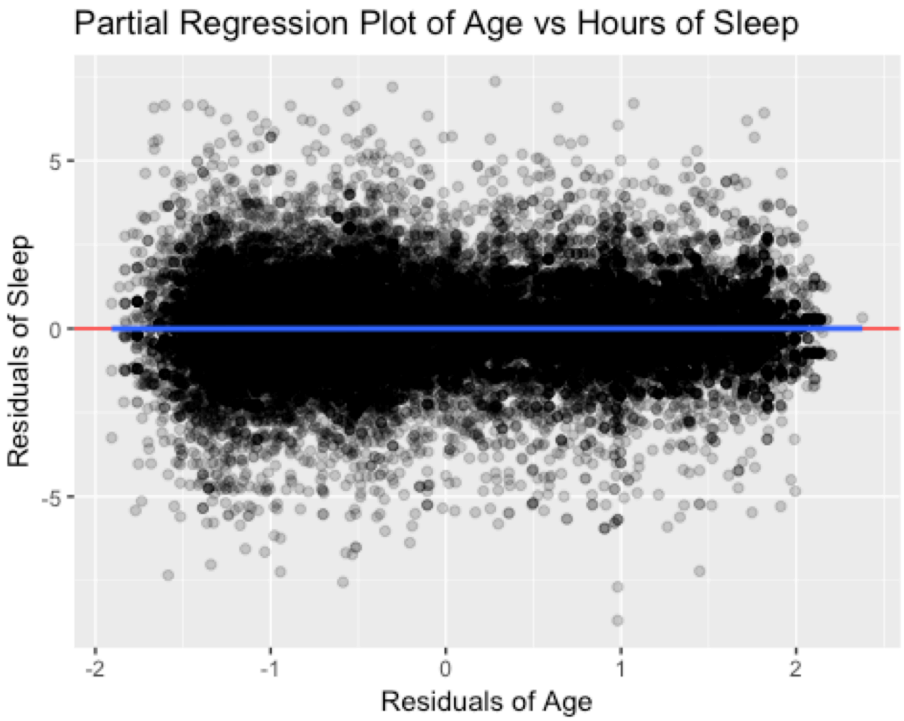


Figure 5. Partial regression plot of Age vs Hours of Sleep. The red line is a horizontal line around zero and the blue line is the geom_smooth() with linear trend.

GRS

The GRS variable is the score obtained from 28 loci previously associated with sleep duration from (Dashti *et al.*). The GWAS study is based on UK Biobank data, which means that it will not necessarily be useful for non-white individuals. To see how worrisome this is, we plot the distribution of the GRS divided by white and non-white. As expected, the distributions are different, but not so much as to invalidate our use of GRS in the model .

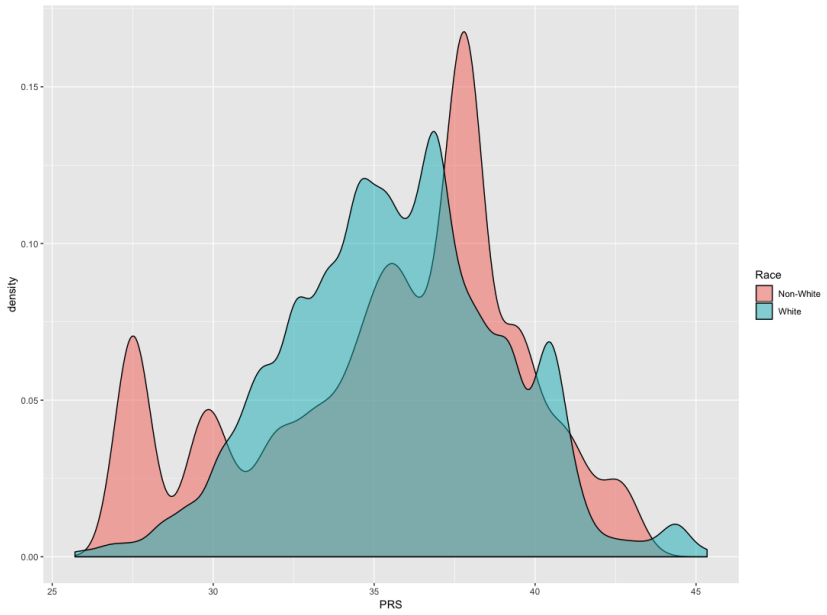


Figure 6. Estimated densities of the GRS by white and non-white participants. The distribution of white participants is in blue, and the non-white participants is in red.

We now wanted to see if the score is useful, so we plotted a boxplot of hours of sleep by quantiles of GRS. This means that the left most boxplot is the boxplot of hours of sleep with individuals with a GRS less than 75% of the population. We can see that there is a small negative trend, which means that the genetic risk score seems to be having a small effect on duration of sleep in GfG participants.

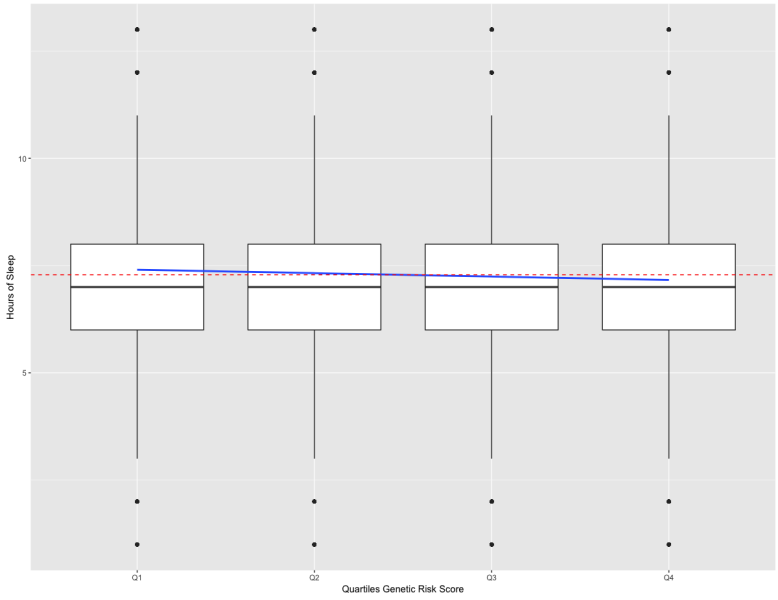


Figure 7. Distribution of hours of sleep by quartiles of the GRS. The dashed red line is the mean hours of sleep and the blue line is the linear trend of the quartile classification.

Model Methods

Null Model

By looking at the results of the EDA we propose a model that has all the covariates we think could be meaningful. We will test some of these to come up with a final model. The null model is as follows:

$$sleep = \alpha_0 + (\alpha_1, \ldots, \alpha_{23})race + (\alpha_{24}, \alpha_{25})spline(age) + (\alpha_{26}, \alpha_{27})spline(no.drinks) + \alpha_{28}gender + \alpha_{29}GRS + \alpha_{30}wine + \alpha_{31}beer + \alpha_{32}liquor + \alpha_{33}other.drink + (\alpha_{34}, \ldots, \alpha_{37})(no.drinks) * (wine, beer, liquor, other.drink) + b_0 + b_1 weekend + \epsilon$$

Since we have 24 different races, 4 types of different beverages, 2 splines (age and number of beverages) with 2 knots each, gender, GRS, number of beverages interacted with type of beverage, individual, and weekend, we have a total of 40 parameters in the null model.

Race

We explored two methods for including race in our model. The first was to include an indicator for each of the 24 unique categories for race in our data. The second was to bin participants as either white or non-white, which seemed appropriate since 80% of study participants described themselves as white. Although the white/non-white binary variable was significant, we found that including all 24 categories of race was more significant (p-value = 0.044 vs. p-value = 4.60e-4). In addition, including all race categories resulted in a lower AIC than including the white/non-white binary variable. It is possible that the white/non-white binary variable was not as significant because the “non-white” categorization included several races that were mixtures of white and non-white ethnicities.

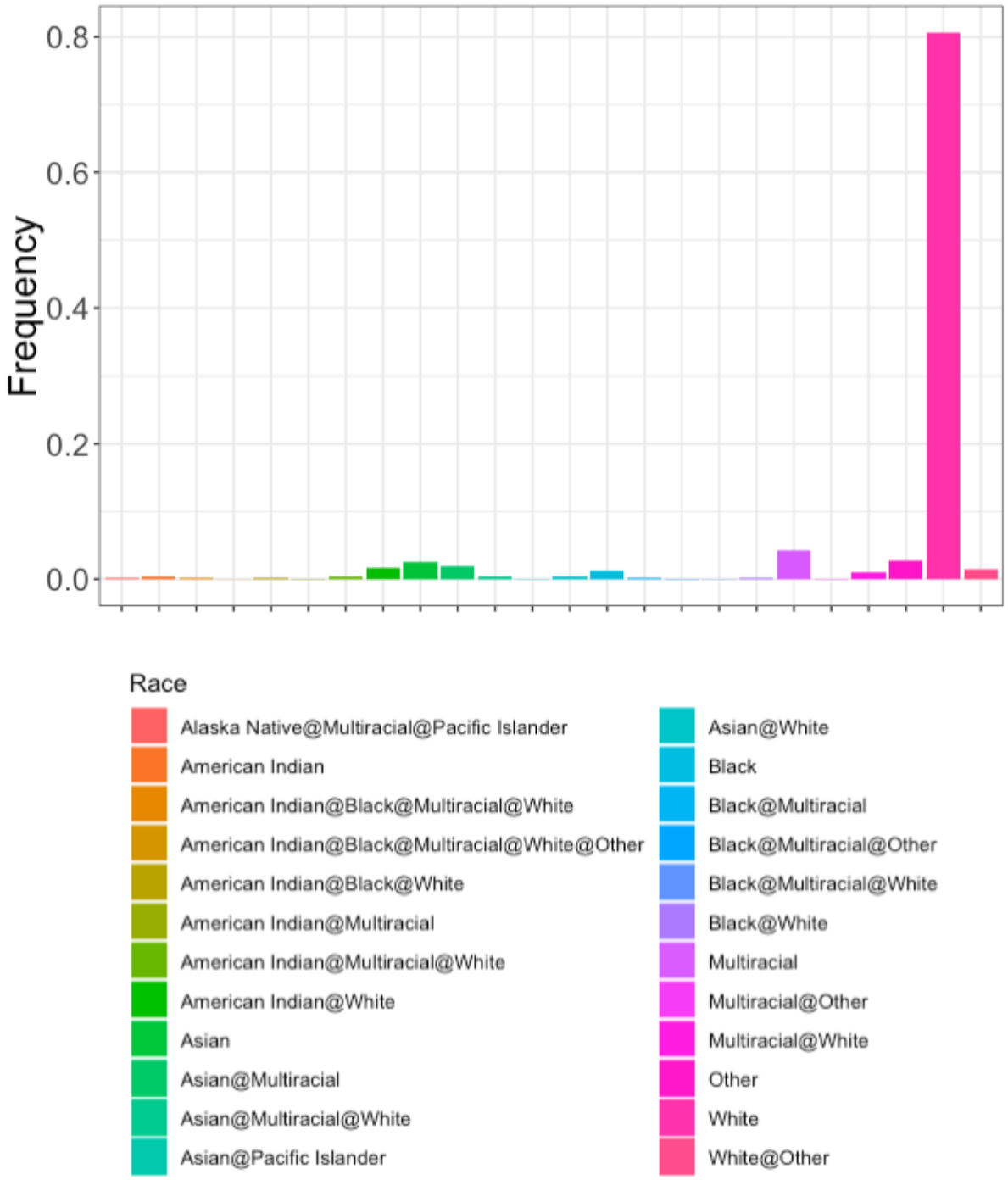


Figure 8. Distribution of races.

Model	df	AIC
No race	15	129460.5
White vs. non-white	16	129461.5
24 race categories	38	129448.4

Figure 9. AIC values for a model without any race term, a model with a binary white/non-white covariate, and a model including all 24 categorizations of race.

Splines

Despite seeing no evidence for splines in the EDA partial regression plots, we decided to fit our model with and without splines to see if their inclusion produced any model improvement. When we did so, we found that the spline term for age was significant (p-value = 2.79e-4) and slightly reduced AIC. The spline term for number of drinks also slightly reduced AIC, but it was not significant (p-value = 0.89). Given these results, we chose to include a spline for age in the model but not include a spline for number of drinks.

Model	df	AIC
No splines	36	129448.7
Age spline	38	129448.4
Number of drinks spline	38	129445.6

Figure 10. AIC values for models with and without splines.

Random Effect: Individual

We expected there to be some differences between individuals when it came to their sleeping patterns. Figure 11 was generated by taking a random sample of individuals from our subsetted data. On the x-axis is the twenty random individuals, while on the y-axis we plotted box-plots for the hours of sleep. The variability between different individuals is evident in the plot. We then calculated the intra-class correlation (ICC), and the score was equal to 0.3. While this is not generally considered a very high score we are dealing with discrete data, and so we thought this score along with Figure 11 was enough evidence for us to consider using a random effect for the intercept of each subject. Using this random effect greatly improved the variability explained by our model.

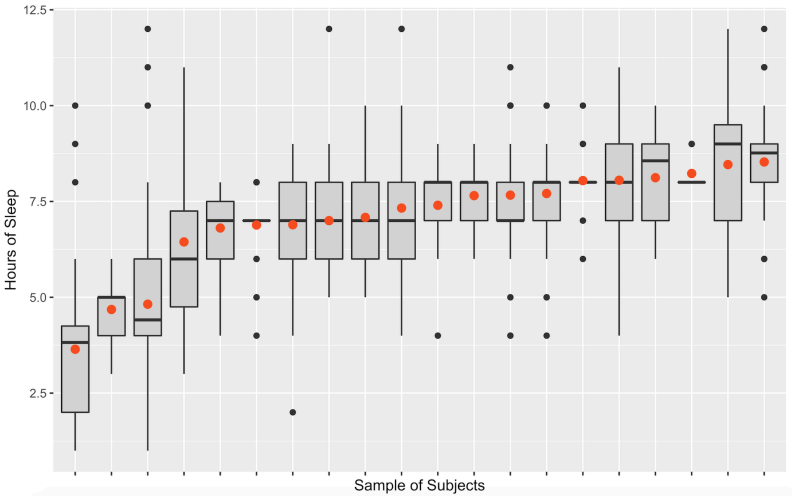


Figure 11. Sample of individuals' sleep duration.

Random Effect: Weekend

It is natural to consider that the day on which a subject completes a survey can introduce extra structure in the data. Figure 12 shows the distribution of hours of sleep among all subjects, stratified by the day in a week. Each color represents an amount of sleep (reported as integer in hours). Those with more than 11 hours of sleep are grouped into the purple bar while those below 3 hours are grouped into the light blue bar. We observed that weekends (Friday and Saturday) do have a different pattern than weekdays, and we therefore decided to include weekends vs. weekdays, rather than every day individually, as a random effect in our model to account for this relation.

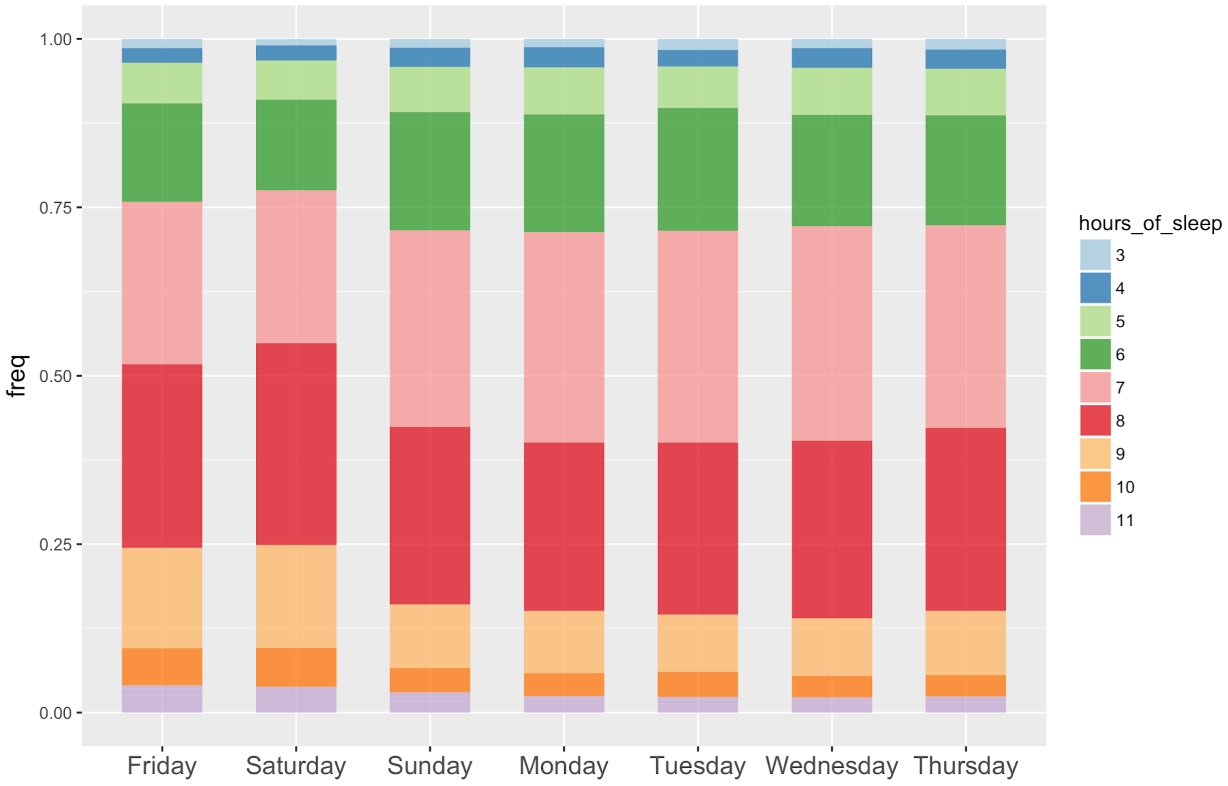


Figure 12. Distribution of hours of sleep among all subjects stratified by the day in a week.

Interactions

We were interested in looking at the interactions with the type of drink and the number of drinks. Figure 13 shows the amount of drinks consumed the night before on the x-axis and the hours of sleep gotten on the y-axis. It is divided in four panels, each of which plots a trend for a particular type of drink. In the bottom left panel we can see that the slope of hours of sleep vs. amount of drinks is less when liquor is consumed and that the confidence bands don't overlap. This demonstrates that drinking liquor, as opposed to other drinks, significantly affects the association between hours of sleep and number of drinks. It is consequently evidence that we should keep the interaction. For the case of beer and wine (top left and top right respectively), the plots show almost no evidence of an interaction. Finally, the bottom right plots shows a very different slope for consumption of "other" drinks; however, the uncertainty in the slope, demonstrated by the overlapping confidence bands, is so high that we would not feel justified including the interaction in the model.

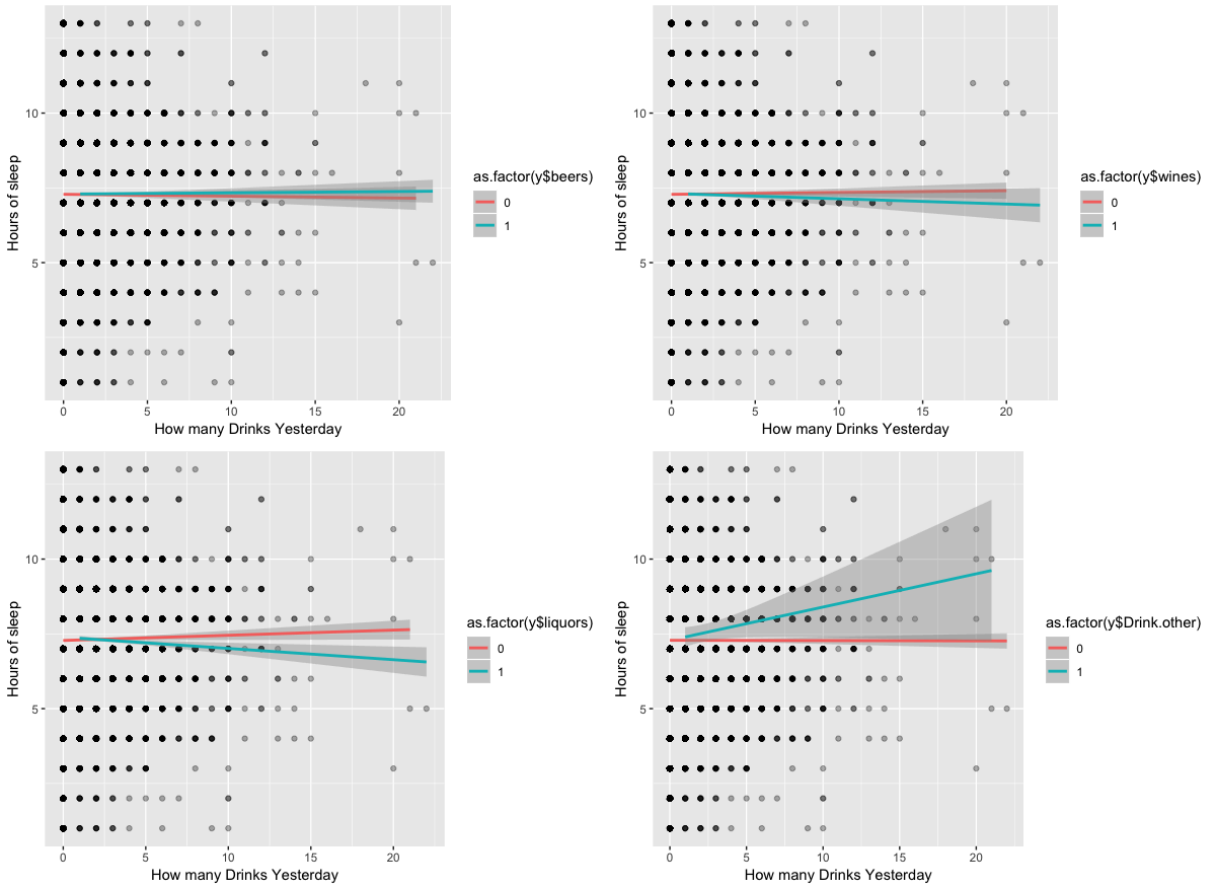


Figure 13. Scatter plot of how many drinks vs. hours of sleep divided in four panels(top left = beer, top right = wine, bottom left = liquor, bottom right = other drink). The blue line in each panel is the linear trend of number of drinks consumed vs. hours of sleep of participants that had at least one drink of the type indicated in that panel.

Final Model

$$sleep = \alpha_0 + (\alpha_1, \ldots, \alpha_{23})race + (\alpha_{24}, \alpha_{25})spline(age) + \alpha_{26}(no.\,drinks) + \alpha_{27}gender + \alpha_{28}GRS + \alpha_{29}wine + \alpha_{30}beer + \alpha_{31}liquor + \alpha_{32}other.\,drink + \alpha_{33}(no.\,drinks) * liquor + b_0 + b_1 weekend + \epsilon$$

We fit this model with the lmer function (lme4 package), using bs() (splines package) to do the splines for age.

Results

```
fmod <- lmer(data = y, hours_of_sleep ~ bs(sd_age,3) +
                  how_many_drinks_yesterday + gender.x +
                  (1|masked_id) + (1|weekend) + sd_PRS +
                  wines + beers + liquors + Drink.other + race + how_many_drinks_yesterday:liquors)
```

```
Type III Analysis of Variance Table with Satterthwaite's method
                                  Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
bs(sd_age, 3)                     32.358 10.7860     3  1279  6.3646 0.0002786 ***
how_many_drinks_yesterday          0.035  0.0354     1 37432  0.0209 0.8850491
gender.x                           0.183  0.1828     1  1122  0.1079 0.7426337
sd_PRS                             7.769  7.7691     1  1135  4.5844 0.0324766 *
wines                              0.105  0.1051     1 37567  0.0620 0.8033524
beers                              0.084  0.0841     1 37494  0.0496 0.8237068
liquors                            0.980  0.9799     1 37402  0.5782 0.4470140
Drink.other                        1.242  1.2423     1 37021  0.7330 0.3919071
race                              89.823  3.9053    23  1118  2.3045 0.0004601 ***
how_many_drinks_yesterday:liquors 13.289 13.2895     1 37432  7.8419 0.0051074 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']

REML criterion at convergence: 129372.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.4137 -0.5089  0.0069  0.5234  6.8754

Random effects:
 Groups    Name        Variance Std.Dev.
 masked_id (Intercept) 0.69429  0.8332
 weekend   (Intercept) 0.05381  0.2320
 Residual              1.69467  1.3018
Number of obs: 37624, groups:  masked_id, 1167; weekend, 2


Fixed effects:
                                                  Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                                       8.188e+00  8.805e-01  4.655e+02   9.299  < 2e-16 ***
bs(sd_age, 3)1                                   -2.446e-01  3.771e-01  1.445e+03  -0.649  0.51663
bs(sd_age, 3)2                                   -5.764e-01  2.141e-01  1.326e+03  -2.692  0.00720 **
bs(sd_age, 3)3                                   -3.254e-01  2.389e-01  1.443e+03  -1.362  0.17341
how_many_drinks_yesterday                        -1.535e-03  1.062e-02  3.743e+04  -0.145  0.88505
gender.xmale                                     -1.948e-02  5.931e-02  1.122e+03  -0.328  0.74263
sd_PRS                                           -5.784e-02  2.701e-02  1.135e+03  -2.141  0.03248 *
wines                                            -8.239e-03  3.309e-02  3.757e+04  -0.249  0.80335
beers                                            -7.080e-03  3.178e-02  3.749e+04  -0.223  0.82371
liquors                                           3.358e-02  4.415e-02  3.740e+04   0.760  0.44701
Drink.other                                       9.511e-02  1.111e-01  3.702e+04   0.856  0.39191
raceAmerican Indian                              -7.441e-01  9.286e-01  1.008e+03  -0.801  0.42312
raceAmerican Indian@Black@Multiracial@White      -1.058e+00  1.001e+00  1.028e+03  -1.058  0.29039
raceAmerican Indian@Black@Multiracial@White@Other -6.006e-02  1.256e+00  1.136e+03  -0.048  0.96187
raceAmerican Indian@Black@White                  -3.541e+00  1.202e+00  9.659e+02  -2.948  0.00328 **
raceAmerican Indian@Multiracial                  -1.062e+00  1.222e+00  1.028e+03  -0.869  0.38523
raceAmerican Indian@Multiracial@White            -6.149e-01  9.414e-01  1.004e+03  -0.653  0.51377
raceAmerican Indian@White                        -5.767e-01  8.712e-01  9.908e+02  -0.662  0.50815
raceAsian                                        -5.968e-01  8.761e-01  9.918e+02  -0.681  0.49588
raceAsian@Multiracial                             1.484e+00  1.057e+00  1.007e+03   1.404  0.16070
raceAsian@Multiracial@White                      -7.103e-01  9.066e-01  1.004e+03  -0.784  0.43351
raceAsian@Pacific Islander                        1.047e-01  1.236e+00  1.069e+03   0.085  0.93250
raceAsian@White                                   4.213e-01  9.612e-01  1.006e+03   0.438  0.66122
raceBlack                                        -7.242e-01  8.743e-01  9.936e+02  -0.828  0.40769
raceBlack@Multiracial                            -4.155e-01  1.064e+00  1.048e+03  -0.390  0.69631
raceBlack@Multiracial@Other                       1.081e+00  1.225e+00  1.035e+03   0.882  0.37790
raceBlack@Multiracial@White                       5.778e-01  1.242e+00  1.085e+03   0.465  0.64185
raceBlack@White                                  -3.110e-02  1.070e+00  1.068e+03  -0.029  0.97682
raceMultiracial                                  -1.016e+00  8.690e-01  9.903e+02  -1.169  0.24261
raceMultiracial@Other                            -1.637e+00  1.224e+00  1.028e+03  -1.338  0.18115
raceMultiracial@White                            -8.506e-01  8.875e-01  9.986e+02  -0.958  0.33808
raceOther                                        -8.578e-01  8.702e-01  9.903e+02  -0.986  0.32447
raceWhite                                        -5.793e-01  8.561e-01  9.866e+02  -0.677  0.49877
raceWhite@Other                                  -9.968e-01  9.279e-01  9.992e+02  -1.074  0.28299
how_many_drinks_yesterday:liquors                -4.122e-02  1.472e-02  3.743e+04  -2.800  0.00511 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14. Results of the final model.

Discussion

Genes for Good (GfG) is a fantastic resource for longitudinal analysis and has never been analyzed longitudinally prior to this study. Our project started with a simple desire to analyze this dataset, but it took us a while to determine what kind of question to ask that would relate to longitudinal analysis.

Limitations

There were some issues with how variables were coded in the dataset that made the outcome and covariate choices difficult. For example, the age variable is measured in two different ways: as a categorical variable with intervals, and by year of birth. Since we wanted to use splines for the age, it was better to use the year of birth to create a discrete numeric variable for age. When trying to create this variable, we noticed that sometimes the year of birth was coded as '< 1950' (~2% of the observations). In contrast, around 6% of the data had a year of birth coded as a numeric variable before 1950. Our first thought was to not use the individuals with year of birth coded as '< 1950', but we thought this might bias our results since we were only using a portion of the people born before 1950. The final decision was made that we wouldn't use any subjects that had a birth year before 1950.

The self-reported nature of the survey also proved to be a challenge. Self-reported data may suffer from low accuracy or even bias in some situations. This is especially the case with our outcome variable: hours of sleep gotten the previous night. The accuracy of the reported outcomes is not something we were able to address in our model. One way this could be resolved in further studies would be to link the survey results with activity trackers. While activity trackers have their own issues with tracking sleep time accurately, they would be more accurate and consistent than self-reported outcomes.

It is important to consider how answering survey questions may influence participants' behaviors (i.e. alter sleep or drinking patterns), and perhaps model the change over time as the outcome. We did see a slight negative correlation between the variation of sleep time and the time the subject fills out the survey (more details in supplementary). Unfortunately, there are not yet enough participants who have consistently filled out the survey to model the change in sleep and drinking variation over time. An interesting hypothesis is that the variability in sleep might decrease the longer someone took the surveys. The thinking behind this hypothesis is that the longer the person participated in the survey, the more aware they would become of their sleeping patterns, and in turn try to maintain a more consistent sleeping pattern.

Self-selection bias was also something we were not able to control for in this analysis. Self-selection bias (similar to non-response bias) arises from only analyzing the subset of the population who decided to participate in the study and, in our case, decided to fill out multiple surveys over a length of time. This subset may not be a good sample of the population as a whole, and it can be hard to extrapolate the results to a larger population. With more time and resources, we could have tested this by comparing our results to population data to see how they differ. Even within one subject, his/her choice of filling the survey on a particular day is also

subject to a selection bias. The bias can be in different directions across individuals, so our hope that the bias mostly averages out across participants.

Findings

While there were some limitations to our model, we were nonetheless able to analyze our data in an informative way and answer our scientific questions. Specifically, we found that the number of drinks consumed alone did not significantly affect sleep duration, but that the interaction between number of drinks consumed and liquor had a significant negative effect on sleep duration. In addition, we found that the GRS was also a significant factor in sleep duration, suggesting that genetics plays a meaningful role the sleep patterns of GfG participants. This is an exciting discovery since genetics generally have a very small effect on complex phenotypes like sleep duration.

Other significant covariates include age and race. It was not surprising to find that age was significant since we saw some variation in drinking habits among older participants and GfG also has participants below the legal drinking age. The significance of race, however, is more surprising since 80% of GfG participants classify themselves as white. We suspect there may be a few individuals with very influential data that have non-white ethnicity, and these may account for the significance of race in our model.

References

Dashti, Hassan *et al*. **GWAS in 446,118 European adults identifies 78 genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates.** April 19, 2018. https://www.biorxiv.org/content/early/2018/04/19/274977.

Supplementary Material

We explored the possibility of using variation of the amount of sleep over a time period as a response, since it may be more heterogeneous in the population than the actual hours the subjects sleep.

We define overlapping sliding windows of 10 records (not restricting the time span of the 10 records, with the assumption that the days people fill the survey are a random sample from the whole time period) with 5 records overlapped between two adjacent windows. We calculate the variance of the hours of sleep and the mean of the numbers of drinks in 10 records for each subject. We then calculate the correlation between the two quantities (across an array of such windows) within each subject. Figure S1 shows the distribution of this correlation, stratified by gender. While we do see that female and male display different distribution and the correlation may vary greatly across individuals, we do not have much evidence supporting an overall pattern.

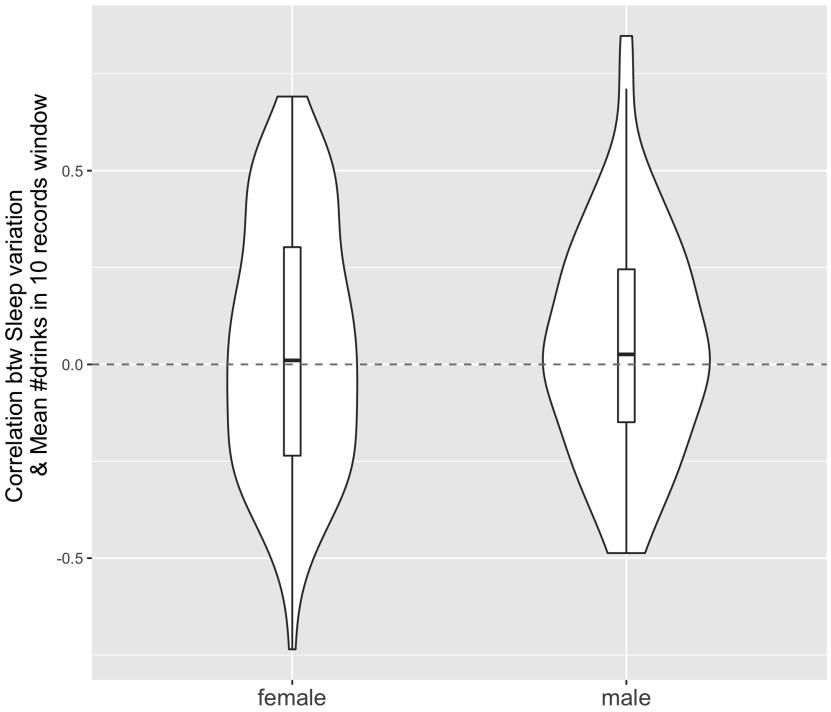


Figure S1. Correlation between the variation of sleep and the amount of drink.

To explore our hypothesis that filling the survey potentially changes people's behavior, we looked at the change of the variation of sleep over time. Similar as above, we calculated the variance of the hours of sleep in sliding windows for each subject, then calculate its correlation with the time starting from his/her first survey. Figure S2 suggests a slightly negative correlation, in the same direction with this hypothesis. It is possible that we can use the variation as a response and do a more careful modeling to describe the pattern, but this is out of the scope of this project.

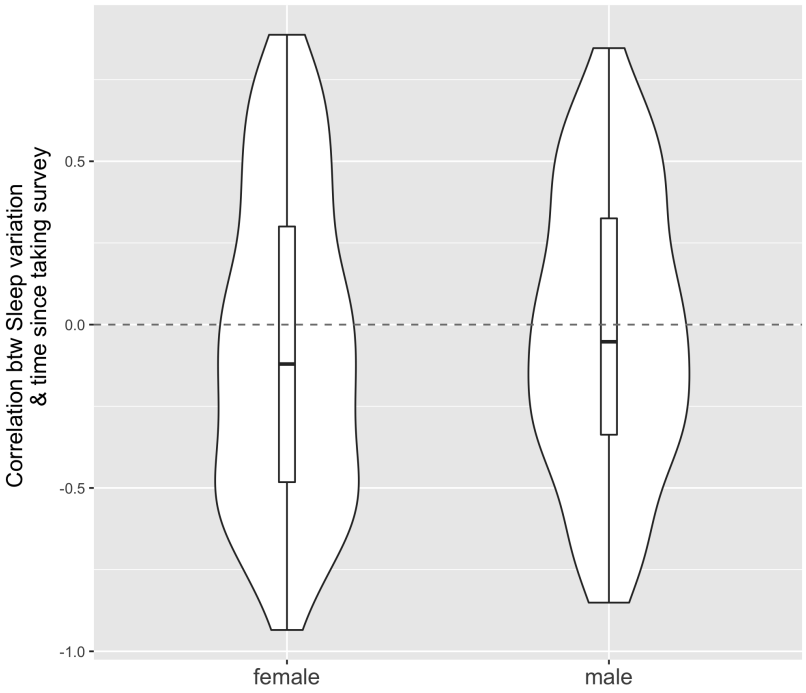


Figure S2. Correlation between the variation of sleep and the time since the subject first took the survey.