NishantVelagapudi_JoshWilson_DanVolk

Nishant Velagapudi, Josh Wilson, Dan Volk

March 4, 2018

Introduction

The purpose of this analysis is to study the interaction between self-esteem, relationship events, and their impact on an individual's total alcoholic beverage consumption. The data we will use come from a study in which moderate to heavy drinkers recorded how many drinks they had every day over a 30-day period. In addition, participants were asked to fill out a survey covering their current state of mind. Specifically the survey covered how many positive/negative relationship events they experienced, how many positive/negative life events they experienced, and their overall self esteem.

Our analysis will test the following hypotheses layed out by Dehart et al. (2008). "We hypothesized that negative interactions with romantic partners would be associated with alcohol consumption (and an increased desire to drink). We predicted that people with low trait self-esteem would drink more on days they experienced more negative relationship interactions compared with days during which they experienced fewer negative relationship interactions. The relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem."

Our analysis begins with an exploratory data analysis section in which we explore univariate distributions as well as bivariate and multivariate relationships in the data. We then fit a couple of models based on our EDA, evaluate their viability, and follow up with model interpretation and assessment of our hypotheses.

Read in Data

First, we will load in the data, below is a list of the variables in the dataset. We will by default set the variables *id*, *studyday*, *dayweek*, and *gender* to be categorical. The list of variables and their interpretation can be found in the appendix.

```
# Data load
data_c <- read.csv("DeHartSimplified.csv", sep=",")
data_c[,c('id', 'studyday', 'dayweek', 'gender')] <-
  lapply(data_c[,c('id', 'studyday', 'dayweek', 'gender')], as.factor)
```

EDA

Independence assumption

The data was collected over a 30 day period on the same subjects. Assuming each observation is i.i.d. would leave us at risk of violating the independence assumption as behavior for the same subject across multiple study days may be similar and could skew our results. To avoid making this assumption we will examine one single day across all subjects. We chose to analyze Saturday (`dayweek==6`), since many people do not work on Sundays and there is also a complete data set for that day. The rest of our EDA and model analysis will be done with the Saturday subset of data.

```
# Reducing data to first Saturday following the textbook methodology
df_sat <- data_c[data_c$dayweek == 6,]
# Remove the dayweek variable
df_sat <- subset(df_sat, select=-c(dayweek))
```

Data Summary

With our Saturday data, we have 89 individuals between the ages of 24 and 42 and no missing values for any of our variables. The range of age is quite narrow and could bias our results as they may not generalize. Without further information on the experimental design and the sampling process, the analysis below does not claim to be causal. The sample is somewhat skewed in gender, with 56% of the particpants being women.

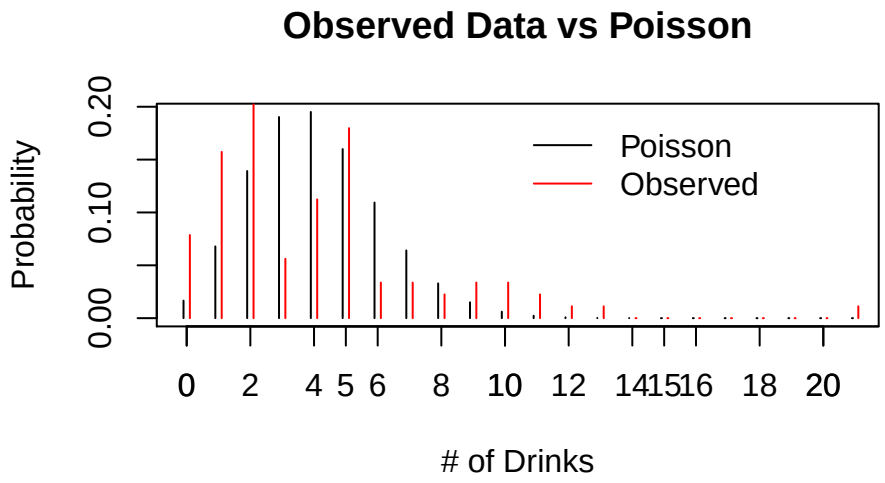
```
summary(df_sat[,c("gender", "age")])
```

```
##  gender      age
##  1:39   Min.   :24.43
##  2:50   1st Qu.:30.53
##         Median :34.57
##         Mean   :34.29
##         3rd Qu.:38.19
##         Max.   :42.28
```

Univariate EDA

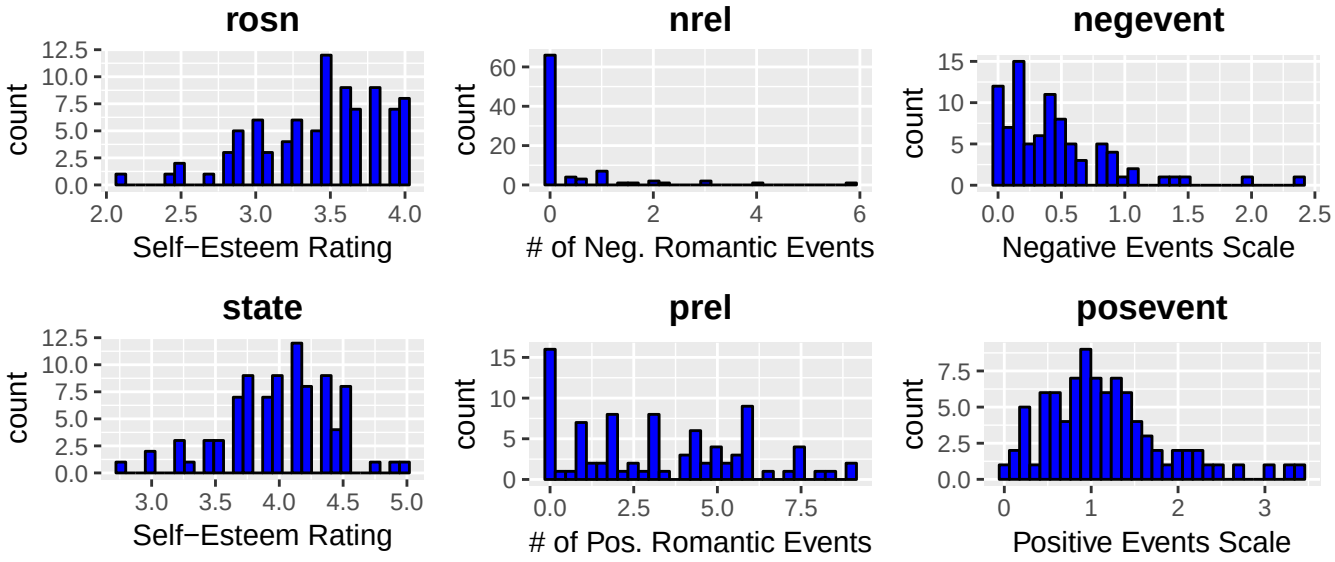
We next investigate and visualize the variables that we use in our model. First, we investigate the *numall* variable which will be the dependent variable in our study. Since it is by nature a count variable, we want to assess whether a Poisson distribution is the correct distribution for this variable. The actual data seems to compare well with the Poisson distribution with notably low values for 3 drinks and a spike at 21 drinks. Several high numbers of drinks will force us to look into possibly influential outliers in our dataset. While, it is not a perfect fit, a Poisson distribution does seem appropriate overall and will be used in our model specification.

```
# Plot of dependent variable vs poisson distribution
rel.freq <- table(factor(df_sat$numall, levels=0:21))/length(df_sat$numall)
y <- 0:21; prob <- round(dpois(y, mean(df_sat$numall)), 4)
plot(y-0.1, prob, type="h", ylab="Probability", xlab="# of Drinks",
     main="Observed Data vs Poisson")
axis(side=1, at=seq(0,22,2)); lines(y+0.1, y=rel.freq, type="h", col='red')
legend(10,0.2, c("Poisson", "Observed"), col=c('black','red'),
       lty=c('solid','solid'), bty='n')
```




```
# Histogram of Rosn
hist_rosn <- ggplot(data = df_sat, aes(x = rosn)) +
  geom_histogram() + stat_bin(colour = "black", fill ="blue") +
  ggtitle("rosn") + xlab('Self-Esteem Rating') +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5))
# Histogram of Nrel
hist_nrel <- ggplot(data = df_sat, aes(x = nrel)) +
  geom_histogram() + stat_bin(colour = "black", fill ="blue") +
  ggtitle("nrel") + xlab('# of Neg. Romantic Events') +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5))
# Histogram  of Negevent
hist_neg <- ggplot(data = df_sat, aes(x = negevent)) +
  geom_histogram() + stat_bin(colour = "black", fill ="blue") +
  ggtitle("negevent") + xlab('Negative Events Scale') +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5))
```

```
# Histogram of State
hist_state <- ggplot(data = df_sat, aes(x = state)) +
  geom_histogram() + stat_bin(colour = "black", fill ="blue") +
  ggtitle("state") +xlab('Self-Esteem Rating') +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5))
# Histogram  of Prel
hist_prel <- ggplot(data = df_sat, aes(x = prel)) +
  geom_histogram() + stat_bin(colour = "black", fill ="blue") +
  ggtitle("prel") + xlab('# of Pos. Romantic Events') +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5))
# Histogram  of Posevent
hist_pos <- ggplot(data = df_sat, aes(x = posevent)) +
  geom_histogram() + stat_bin(colour = "black", fill ="blue") +
  ggtitle("posevent") + xlab('Positive Events Scale') +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5))
ggarrange(hist_rosn, hist_nrel, hist_neg, hist_state, hist_prel, hist_pos,
          ncol = 3, nrow = 2, common.legend = TRUE, legend = "bottom")
```



From the distributions above we can see that the long-term self worth variable *rosn* is more negativley skewed than the day-to-day self esteem rating *state* despite the fact that both ratings seem to be on slightly different scales. This indicates that most respondents have a fairly positive view of themselves.

The *nrel* variable has a large spike at zero with approximatley 74% of respondents reported having zero negative romantic interactions on Saturday. While this number is low, a quick check shows that it is in line with the distribution of negative romantic events for other days. The relative sparsity of this variable indicates that it should probably be binarized when used in our analysis. While *prel* is slightly more distributed, it is still dominated by zero values as 18% of respondents report no positive romantic interactions on saturday. Binarization or binning may also be appropriate for *prel*.

Both *posevent* and *negevent* are more normally distributed with a slight positive skew. Transormations may be appropriate, especially for *negevent* which similar to *nrel* seems to be left consored.

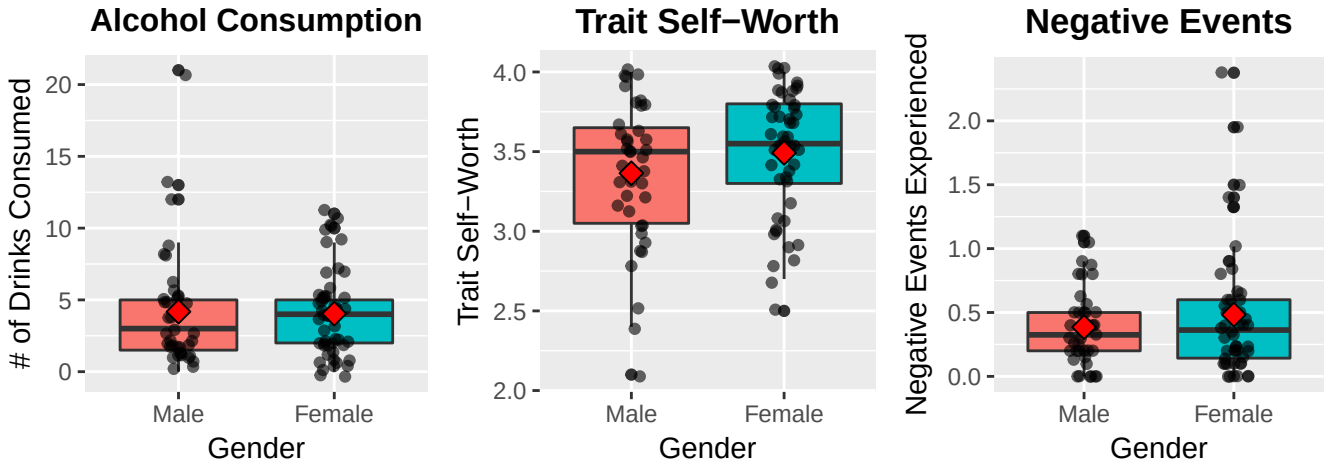
```
# Binarizing Negative Romantic Relationships and Positive Romantic Relationships
df_sat['HasNrel'] <- ifelse(df_sat$nrel > 0, 1, 0)
df_sat['HasPrel'] <- ifelse(df_sat$prel > 0, 1, 0)
```

Bivariate Analysis

We next look at investigating relationships between variables. Below are several boxplots showing the gender divide across several key variables. We notice that female and male distributions are similar for each of the variables. However, females tend to have a higher self esteem and comprise more of the negative relationship outliers. Meanwhile, males are responsible

for several of the high alcohol consumption outliers. These differences, while minor, may make gender worth including in our model specification.

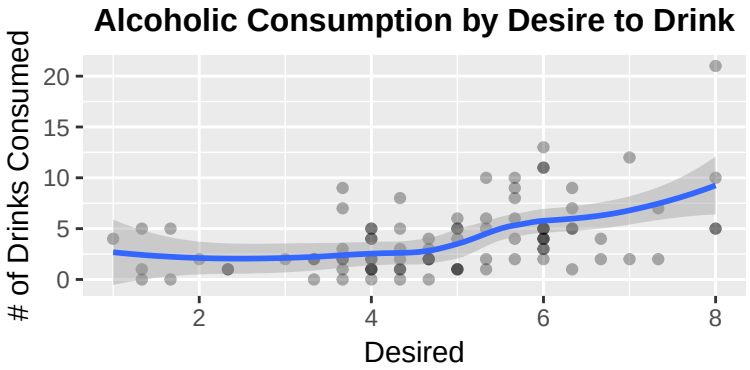
```
#Split out three variables of interest by gender
bp_numall <- ggplot(data = df_sat, aes(factor(gender), numall)) +
         geom_boxplot(aes(fill=factor(gender))) +
         geom_jitter(width = 0.1, alpha = 0.6) + xlab("Gender") +
         ylab("# of Drinks Consumed") + ggtitle("Alcohol Consumption") +
         theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 12)) +
         stat_summary(fun.y = mean, geom = "point", shape= 23, size = 3, fill = "red") +
         scale_x_discrete(labels = c("1" = "Male", "2" = "Female"))
bp_rosn <- ggplot(data = df_sat, aes(factor(gender), rosn)) +
         geom_boxplot(aes(fill=factor(gender))) +
         geom_jitter(width = 0.1, alpha = 0.6) + xlab("Gender") +
         ylab("Trait Self-Worth") + ggtitle("Trait Self-Worth") +
         theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5)) +
         stat_summary(fun.y = mean, geom = "point", shape= 23, size = 3, fill = "red") +
         scale_x_discrete(labels = c("1" = "Male", "2" = "Female"))
bp_negev <- ggplot(data = df_sat, aes(factor(gender), negevent)) +
         geom_boxplot(aes(fill=factor(gender))) +
         geom_jitter(width = 0.1, alpha = 0.6) + xlab("Gender") +
         ylab("Negative Events Experienced") + ggtitle("Negative Events") +
         theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5)) +
         stat_summary(fun.y = mean, geom = "point", shape= 23, size = 3, fill = "red") +
         scale_x_discrete(labels = c("1" = "Male", "2" = "Female"))
ggarrange(bp_numall, bp_rosn, bp_negev, ncol=3, legend = 'none')
```



Not seen in these boxplots, but in others not show, we know that the drinking desire, state self-esteem, and positive events are all relatively similar across gender. Although, it does appear that females experience a higher number of positive events (there are more outliers towards the top end of the scale).

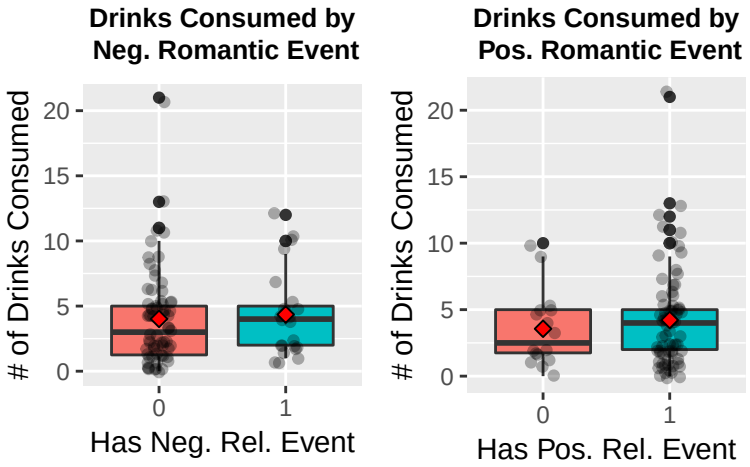
Below we plot desired drinks vs total drinks consumed. There seems to be a positive correlation between the two which makes sense. This correlation is indicative of the fact that *desired* and *numall* are proxies for one another. The problem with this is that using the desire to drink to model the number of drinks could drown out the signal of our other independent variables. To avoid this, we choose only to model *numall* as our dependent variable and we remove *desired* from the model specification.

```
ggplot(data = df_sat, aes(x = desired, y = numall)) + geom_point(alpha = 0.3) +
  geom_smooth(method = 'loess') + xlab("Desired") + ylab("# of Drinks Consumed") +
  ggtitle("Alcoholic Consumption by Desire to Drink") +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 12))
```



Below are boxplots for alcohol consumption vs the binarized variables created above indicating the presence of a positive or negative relationship variable. From the plots below it seems that perhaps the negative relationship events are associated with slightly higher drink consumption as the median is slightly higher. The outliers also stand out and should be considered during our analysis.

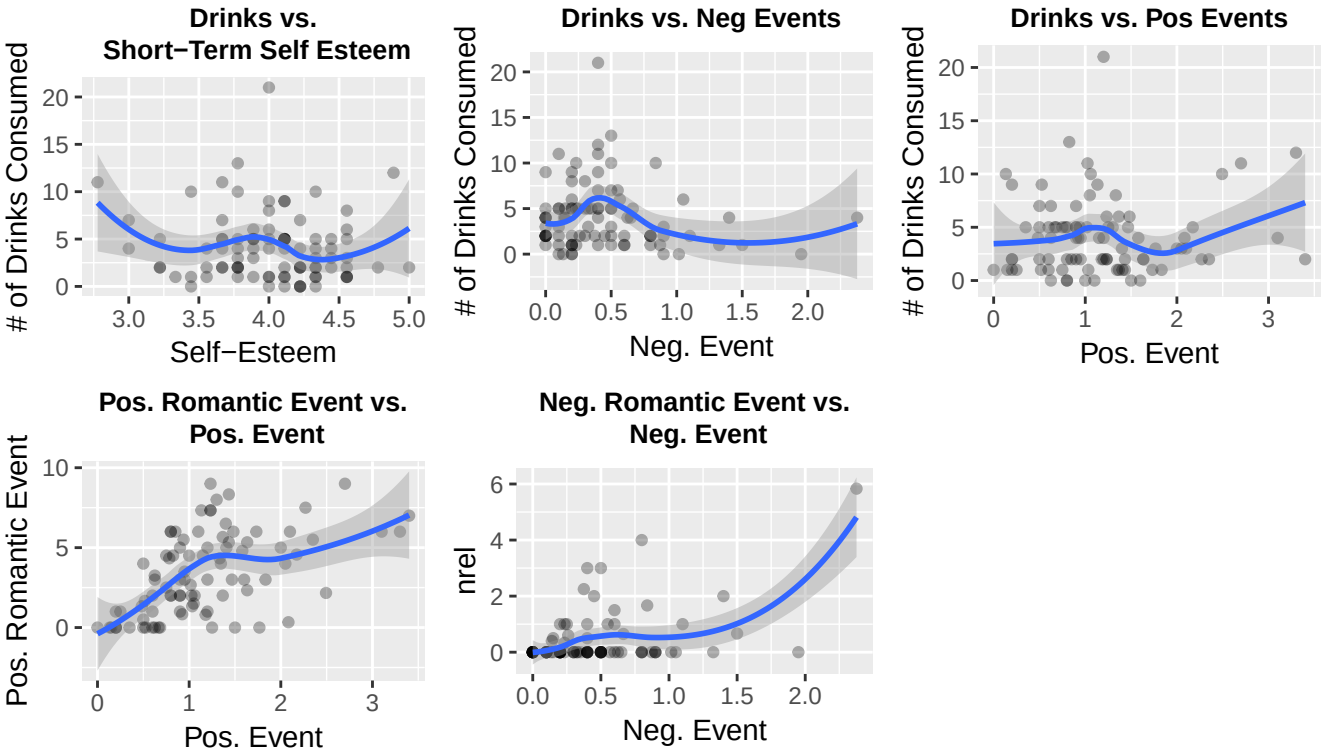
```
#Split out three variables of interest by gender
bp_nrel <- ggplot(data = df_sat, aes(factor(HasNrel), numall)) +
  geom_boxplot(aes(fill=factor(HasNrel))) + geom_jitter(width = 0.1, alpha = 0.3) +
  xlab("Has Neg. Rel. Event") + ggtitle("Drinks Consumed by\n Neg. Romantic Event") +
  ylab("# of Drinks Consumed") +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 10)) +
  stat_summary(fun.y = mean, geom = "point", shape= 23, size = 2, fill = "red")
bp_prel <- ggplot(data = df_sat, aes(factor(HasPrel), numall)) +
  geom_boxplot(aes(fill=factor(HasPrel))) + geom_jitter(width = 0.1, alpha = 0.3) +
  xlab("Has Pos. Rel. Event") + ggtitle("Drinks Consumed by\n Pos. Romantic Event") +
  ylab("# of Drinks Consumed") +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 10)) +
  stat_summary(fun.y = mean, geom = "point", shape= 23, size = 2, fill = "red")
ggarrange(bp_nrel, bp_prel, ncol=2, legend = 'none')
```



In the plots below, we examine the relationships between some of the remaining variables. The *state* variable seems to have an interesting almost quadratic relationship with number of drinks as both tails of the state variable seem to be associated with slightly higher drinking. There also seems to be a potential relationship between *posevent* and number of drinks, while the relationship with *negevent* is less pronounced. Also, *posevent* is strongly correlated with *prel* while *negevent* is less obviously correlated with *nrel*. In each case, the *negevent* and *posevent* variables could be picking up on some variability that our binarized *prel* and *nrel* variables will not. It could be useful to include them in our model. This also makes sense, because in addition to relationship events, outside events could also cause the subjects to drink more.

```
## Self-Esteem vs. Number of Drinks
state_numall <- ggplot(data = df_sat, aes(x = state, y = numall)) +
  geom_point(alpha = 0.3) + geom_smooth() + xlab("Self-Esteem") +
  ylab('# of Drinks Consumed') + ggtitle("Drinks vs.\n Short-Term Self Esteem") +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 10))
```

```
# Negative Event vs. Number of Drinks
nrel_numall <- ggplot(data = df_sat, aes(x = negevent, y = numall)) +
  geom_point(alpha = 0.3) + geom_smooth() + xlab("Neg. Event") +
  ylab('# of Drinks Consumed') + ggtitle("Drinks vs. Neg Events") +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 10))
# Negative Event vs. Number of Drinks
prel_numall <- ggplot(data = df_sat, aes(x = posevent, y = numall)) +
  geom_point(alpha = 0.3) + geom_smooth() + xlab("Pos. Event") +
  ylab('# of Drinks Consumed') + ggtitle("Drinks vs. Pos Events")+
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 10))
# Positive Events vs. Positive Romantic Events
pos_prel <- ggplot(data = df_sat, aes(x = posevent, y = prel)) +
  geom_point(alpha = 0.3) + geom_smooth() + xlab("Pos. Event") +
  ylab("Pos. Romantic Event") + ggtitle("Pos. Romantic Event vs.\n Pos. Event") +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 10))
# Negative Events vs. Negative Romantic Events
neg_nrel <- ggplot(data = df_sat, aes(x = negevent, y = nrel)) +
  geom_point(alpha = 0.3) + geom_smooth() + xlab("Neg. Event") +
  ylab("nrel") + ggtitle("Neg. Romantic Event vs.\n Neg. Event") +
  theme(plot.title=element_text(lineheight=1, face="bold", hjust = 0.5, size = 10))
ggarrange(state_numall, nrel_numall, prel_numall, pos_prel, neg_nrel, nrow=2, ncol=3, legend = 'none')
```



Model Fitting

Appropriateness of Poisson

In our EDA we briefly touched on the appropriateness of Poisson regression for our model of choice given that our *numall* dependent variables is a count variable that seems to generally follow a poisson distribution. We will now formally test our choice of model by comparing fit for a Poisson vs linear model.

```
All_poisson <- glm(numall ~ nrel + rosn + nrel*rosn + desired + age + gender,
                   family=poisson(link="log"), data=df_sat)
All_lm <- lm(numall ~ nrel + rosn + nrel*rosn + desired + age + gender, data=df_sat)
```

```
# AIC for both models
AIC(All_poisson); AIC(All_lm)
```

```
## [1] 448.4493
```

```
## [1] 466.589
```

We can see that the Poisson model has a better AIC: in tandem with the EDA performed comparing a simulated poisson distribution with the observations, this result leads us to use Poisson family models in hypothesis testing.

Forward/Backward Variable Selection

Here, we set up forwards and backwards stepwise selection of variables for inclusion in the model.

```
step_columns <- df_sat[,3:14]
emptyMod <- glm(numall ~ 1, family=poisson(link="log"), data=step_columns)
fullMod <- glm(numall ~ ., family=poisson(link="log"), data=step_columns)

forw.sel <- step(object = emptyMod, scope = list(upper = fullMod),
                 direction = "forward", k = log (nrow (step_columns)), trace = FALSE)

back.sel <- step(object = fullMod, scope = list(lower = emptyMod),
                 direction = "backward", k = log(nrow(step_columns)), trace = FALSE)

Anova(forw.sel)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##          LR Chisq Df Pr(>Chisq)
## desired    69.148  1  < 2.2e-16 ***
## age         7.237  1   0.007143 **
## negevent   15.335  1  9.003e-05 ***
## nrel        8.177  1   0.004243 **
## state       8.072  1   0.004494 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As previously mentioned, we do not feel comfortable using the *desired* variable in predicting number of drinks consumed. We feel that the desire to drink variable is essentially a proxy for what we are trying to study, and including it drowns out the effects of other variables.

We thus manually specify the next model. Based on our EDA and the hypotheses we are testing we include our binary *HasNrel* variable as well as the interaction with *rosn* which indicates our long-term self esteem. We also include *HasPrel*, *posevent*, *negevent*, and *state* as our EDA indicated potential relationships between these variables and the individuals propensity to drink. Our model output below seems to match our hypotheses at first.

```
# Fit Poisson model
Poisson_V2 <- glm(numall ~ HasNrel + rosn + HasNrel:rosn + HasPrel +
                    posevent + negevent + state, family=poisson(link="log"), data=df_sat)
# Coefficients
100*(exp(Poisson_V2$coefficients)-1)
```

```
##  (Intercept)      HasNrel         rosn      HasPrel     posevent
##   735.406248  1222.513288    15.723882    29.815360     9.650475
##     negevent        state HasNrel:rosn
##   -30.249747   -30.040858   -50.782026
```

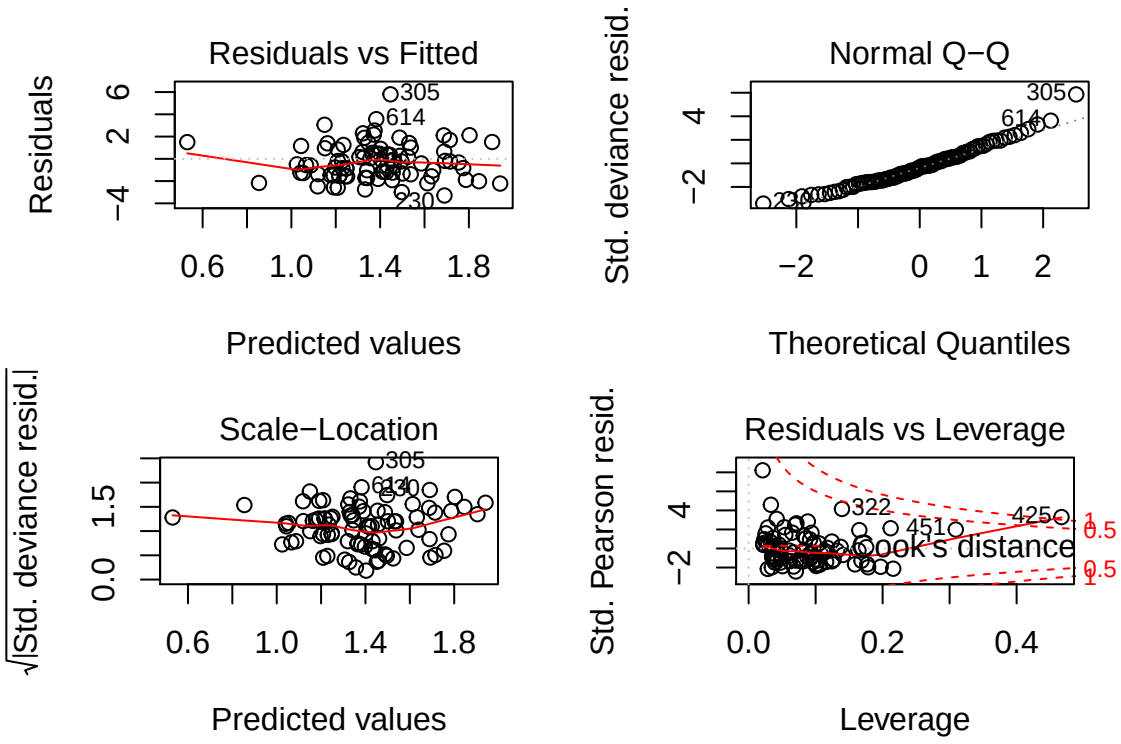
```
#Likelihood ratio test
Anova(Poisson_V2)
```

```
## Analysis of Deviance Table (Type II tests)
```

```
##
## Response: numall
##              LR Chisq Df Pr(>Chisq)
## HasNrel        1.5756  1    0.20939
## rosn           0.0333  1    0.85513
## HasPrel        2.7549  1    0.09696 .
## posevent       1.3528  1    0.24479
## negevent       5.7832  1    0.01618 *
## state          7.7912  1    0.00525 **
## HasNrel:rosn   4.1711  1    0.04112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *HasNrel* variable is positively correlated and significant indicating that subjects with negative relationship events drink more than those without. Also the interaction term *HasNrel:rosn* is negative and significant indicating that as self esteem improves, the drinking rate will decrease by about 50% for each 1-unit change in self esteem. However, there is some cause for concern as our significance levels do not hold up when looking at the likelihood ratio test which is more appropriate for Poisson regression models. Also, the estimated impact of our *HasNrel* variable is over 1200% increase in the rate of drinking for those individuals with negative relationship events vs those without. This does not seem reasonable considering the less prominent relationship between *HasNrel* and *numall* observed in the EDA.

```
# Residual Plots
par(mfrow=c(1,2)); plot(Poisson_V2)
```



The model above does not seem to display prominent trends in the residuals based on the plots above. However, there is at least one highly influential point as shown in the leverage plot above. It is important to remove the point from our dataset as it could be significantly skewing our model results. This could also explain the unusually high coefficient on the *HasNrel* variable. Below we fit the same model, but with the influential point removed.

```
# Fit model removing outliers
Poisson_V2_prune <- glm(numall ~ HasNrel + rosn + HasNrel:rosn + HasPrel +
                          posevent + negevent + state,
                        family=poisson(link="log"),
                        data=df_sat[!(rownames(df_sat) %in% c(425)),])
# Coefficient Analysis
100*(exp(Poisson_V2_prune$coefficients)-1)
```

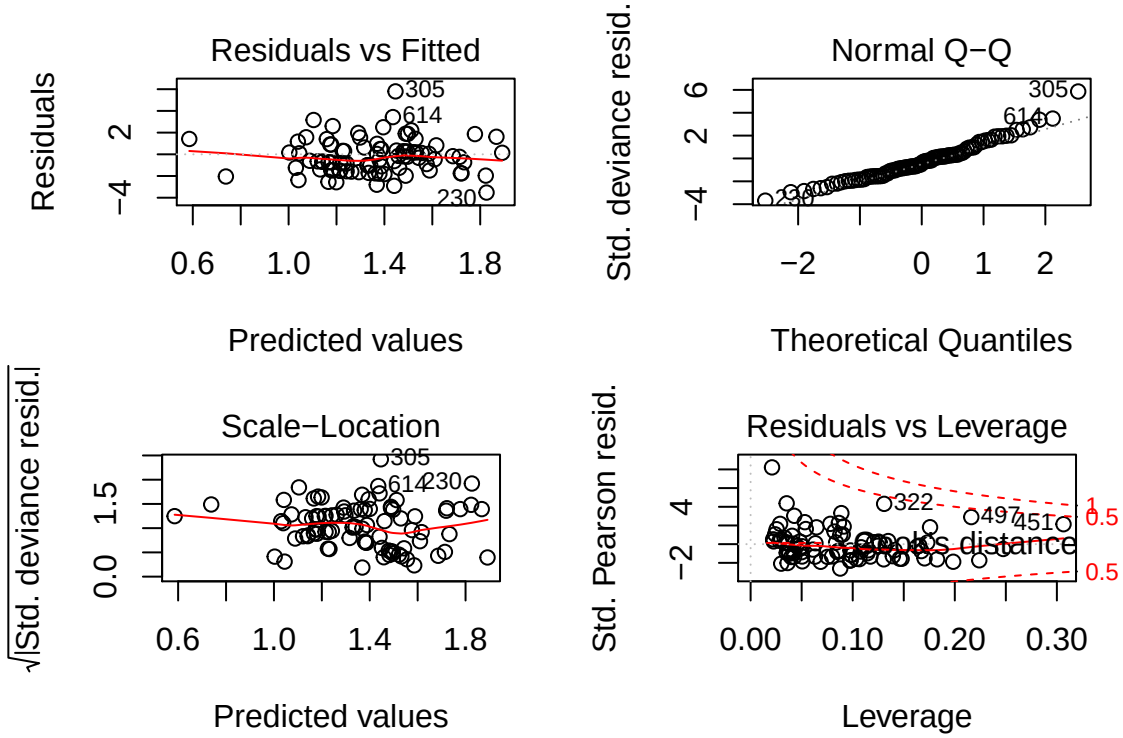
```
##  (Intercept)      HasNrel         rosn      HasPrel     posevent
##  1410.562893    38.550609    18.620429    31.413011    -2.530268
##     negevent        state HasNrel:rosn
##   -35.242533   -38.617365    -8.164136
```

```
# Likelihood Ratio
Anova(Poisson_V2_prune)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##              LR Chisq Df Pr(>Chisq)
## HasNrel        0.0519  1  0.8197976
## rosn           1.3971  1  0.2372097
## HasPrel        3.0351  1  0.0814828 .
## posevent       0.0842  1  0.7717309
## negevent       8.2407  1  0.0040961 **
## state         13.0418  1  0.0003046 ***
## HasNrel:rosn   0.0459  1  0.8304368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our model coefficients make more sense now. However, the variables of interest are no longer significant. According to the model above, an individual starts out with a high propensity to drink which makes sense as our subjects are noted to be heavy drinkers. Having a negative relationship event on Saturday has a 38% increase on an individual's drinking rate, but this is not significant. Also, the interaction term is also insignificant, despite also being directionally in line with the hypothesis. Instead, negative events and short term self esteem seem to dominate this model. For *negevent* a 1-unit increase results in a 35% decrease in drinking rate. For state, a 1-unit increase is associated with a 38% decrease in drinking rate.

```
# Plot residuals
par(mfrow=c(1,2)); plot(Poisson_V2_prune)
```



The model above seems to be more appropriately specified. We do not see non-normality of our residuals or any real patterns to the residual plots to indicate heteroskedasticity. No points are overly influentail, but some are close to the Cook's distance lines.

Below we show side by side comparisons of the two specified models.

```r
# Compare to model with outliers
stargazer(Poisson_V2_prune, Poisson_V2, type="text", digits = 2)
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                numall
##                           (1)            (2)
## ------------------------------------------------
## HasNrel                   0.33          2.58**
##                          (1.39)         (1.19)
##
## rosn                      0.17           0.15
##                          (0.15)         (0.14)
##
## HasPrel                  0.27*           0.26
##                          (0.16)         (0.16)
##
## posevent                 -0.03           0.09
##                          (0.09)         (0.08)
##
## negevent                -0.43***        -0.36**
##                          (0.16)         (0.16)
##
## state                   -0.49***       -0.36***
##                          (0.13)         (0.13)
##
## HasNrel:rosn             -0.09          -0.71**
##                          (0.40)         (0.35)
##
## Constant                2.72***         2.12***
##                          (0.68)         (0.65)
##
## ------------------------------------------------
## Observations               88             89
## Log Likelihood          -236.02        -243.05
## Akaike Inf. Crit.        488.05         502.11
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Conclusion

After careful EDA and model specification, we fail to reject the null hypothesis that there is no relationship between negative romantic events and drinking rate. Similarly, we found no significant indication that negative relationship events would have a stronger impact on individuals with a lower self esteem. Independence assumption violations forced us to dramatically reduce the size of our sample to ensure that our observations were iid. This smaller sample size paired with the lack of variance in the *nrel* variable led to a few highly impactful observations having a strong impact on the outcome of our model. There were in reality only 13 individuals who had negative relationship events on Saturday and each of those had a strong impact on the direction and magnitude of the *HasNrel* variable.

Appendix

0. Variable List

- *id*: identifier of survey respondant
- *studyday*: day of study
- *dayweek*: day of week (saturday == 6)
- *numall*: number of alcoholic beverages consumed on that day
- *nrel*: negative romantic relationship event
- *prel*: positive romantic relatinship event
- *negevent*: number & intensity of negative events in a day (0-3)
- *posevent*: number & intensity of positive events in a day (0-3)
- *gender*: male (1) female (2)
- *rosn*: trait self-esteem
- *age*: age of study subjects
- *desired*: desire to drink each day, higher number corresponds to more desire
- *state*: short-term self-esteem state

1. Boxplots of various variables by gender

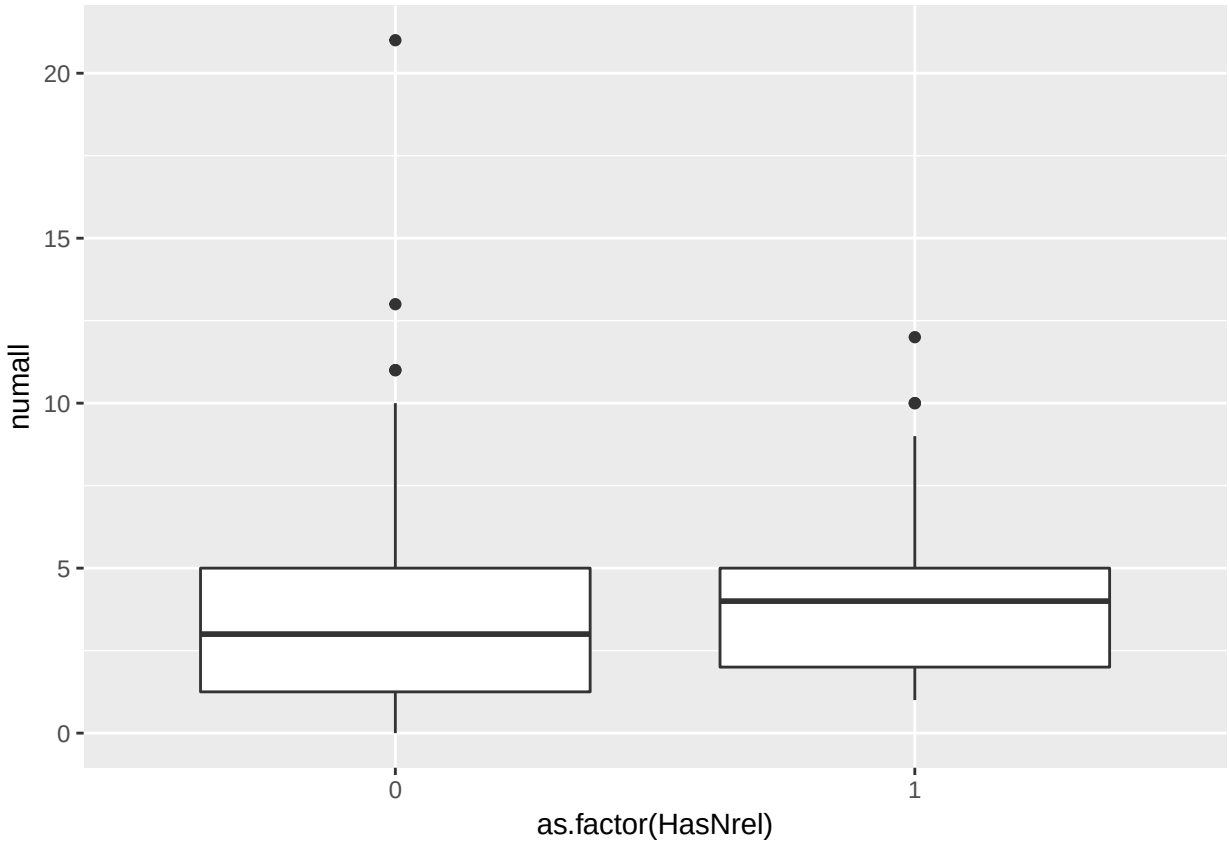
```
biv_1 <- ggplot(data = df_sat, aes(factor(gender), posevent)) + geom_boxplot(aes(fill=factor(gender))) +
         geom_jitter(width=0.1) + xlab("Gender")       + ggtitle("Positive Events")      +
         theme(plot.title=element_text(lineheight=1, face="bold"))

biv_2 <- ggplot(data = df_sat, aes(factor(gender), state)) + geom_boxplot(aes(fill=factor(gender))) +
         geom_jitter(width=0.1) + xlab("Gender")       + ggtitle("Self-Esteem State")      +
         theme(plot.title=element_text(lineheight=1, face="bold"))

biv_3 <- ggplot(data = df_sat, aes(factor(gender), desired)) + geom_boxplot(aes(fill=factor(gender))) +
         geom_jitter(width=0.1) + xlab("desired")       + ggtitle("Drinking Desire")      +
         theme(plot.title=element_text(lineheight=1, face="bold"))


ggarrange(biv_1, biv_2, biv_3, ncol=3, common.legend = TRUE, legend = "bottom")
```

```
drk_x_hasnrel <- ggplot(data = df_sat, aes(x = as.factor(HasNrel), y = numall)) + geom_boxplot()
drk_x_hasnrel
```



2. Adjusting desire to drink by mean desire

```
#we can see that the average desire to drink is also a normal distribution. This suggests that when we are lo

#we can do this by taking a desire to drink as a distance from average desire to drink

# df_sat = merge(df_sat, avgDrink, by="id")
# df_sat = rename(df_sat, c('desired.y' = 'Id_MeanDrink', 'desired.x' = 'desired'))

# df_sat['desired_adj'] = df_sat$desired - df_sat$Id_MeanDrink
# hist(df_sat$desired_adj)

#this gives a normal distribution of desired
```

3. Testing significance of mean drink consumed delta between those with negative relationship events and those without

Performing a wilcox rank-sum test informs us of whether or not there is a statistically significant average difference in desire to drink when the survey respondent has either no negative relationship events or some negative relationship events.

```
data.frame(aggregate(desired ~ HasNrel, df_sat, mean))
```

```
##   HasNrel  desired
## 1       0 4.828283
## 2       1 4.898551
```

```
#little difference between 0 and nonzero here.
#test for significance of difference using wilcoxon rank sum test
#unaltered desired to drink - nonsignificant, closer to significance
wilcox.test(df_sat[df_sat$nrel > 0, "numall"], df_sat[df_sat$nrel == 0, "numall"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  df_sat[df_sat$nrel > 0, "numall"] and df_sat[df_sat$nrel == 0, "numall"]
## W = 821, p-value = 0.5604
## alternative hypothesis: true location shift is not equal to 0
```

```
#adjusted desire to drink - nonsignificant
wilcox.test(df_sat[df_sat$nrel > 0, "numall"], df_sat[df_sat$nrel == 0, "numall"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df_sat[df_sat$nrel > 0, "numall"] and df_sat[df_sat$nrel == 0, "numall"]
## W = 821, p-value = 0.5604
## alternative hypothesis: true location shift is not equal to 0
```

We find that the difference is not significant: either when we test for desire to drink or desire to drink adjusted by the average desire to drink by respondant.

4. Forwards and backwards selection
5. Independence Assumption validation

```
# Test independence assumption
mod <- lm(desired ~ as.factor(id), data=data_c)
mean(summary(mod)$coefficients[,4]<0.05)
```

```
## [1] 0.2022472
```

6. Bucketing

```
#bucket data
df_sat['nrelBucket'] <- transform(df_sat, group=cut(as.numeric(sub('[%]', '', nrel+.001)),
    breaks=c(0,1, 2, 3, 4,5,6,7,8.1),
    labels=c('0-1', '1-2', '2-3', '3-4', '4-5', '5-6','6-7','7-8')))$group

df_sat['rosnBucket'] <- transform(df_sat, group=cut(as.numeric(sub('[%]', '', rosn)),
    breaks=c(2,2.5, 3, 3.5, 4),
    labels=c('2-2.5', '2.5-3', '3-3.5', '3.5-4')))$group

df_sat['desiredBucket'] <- transform(df_sat, group=cut(as.numeric(sub('[%]', '', desired+.001)),
    breaks=c(0,1, 2, 3, 4,5,6,7,8.1),
    labels=c('0-1', '1-2', '2-3', '3-4', '4-5', '5-6','6-7','7-8')))$group
```

We cross-tabulate our data to see how many observations fall into each bucket of negative relationship events, self worth values, and desire to drink values.

```
# volume and mean desire to drink in each bucket
# dplyr::count(df_sat, nrelBucket, rosnBucket, desiredBucket)

 xtabs(desired ~ nrelBucket, aggregate(desired ~ nrelBucket, df_sat, mean))
```

```
## nrelBucket
##      0-1      1-2      2-3      3-4      4-5      5-6      6-7      7-8
## 4.799087 5.555556 4.444444 4.333333 3.333333 5.666667 0.000000 0.000000
```

```
 xtabs(desired ~ rosnBucket, aggregate(desired ~ rosnBucket, df_sat, mean))
```

```
## rosnBucket
##    2-2.5    2.5-3    3-3.5    3.5-4
## 4.000000 4.933333 4.922222 4.841667
```

```
 xtabs(desired ~ nrelBucket + rosnBucket, aggregate(desired ~ nrelBucket + rosnBucket, df_sat, mean))
```

```
##           rosnBucket
## nrelBucket    2-2.5    2.5-3    3-3.5    3.5-4
##        0-1 4.000000 4.916667 4.819444 4.838384
```

```
##        1-2 0.000000 4.000000 5.733333 6.666667
##        2-3 0.000000 0.000000 0.000000 4.444444
##        3-4 0.000000 7.000000 0.000000 1.666667
##        4-5 0.000000 0.000000 3.333333 0.000000
##        5-6 0.000000 0.000000 0.000000 5.666667
##        6-7 0.000000 0.000000 0.000000 0.000000
##        7-8 0.000000 0.000000 0.000000 0.000000
```

```
table(df_sat$nrelBucket, df_sat$rosnBucket)
```

```
##
##       2-2.5 2.5-3 3-3.5 3.5-4
##   0-1     4    12    24    33
##   1-2     0     2     5     2
##   2-3     0     0     0     3
##   3-4     0     1     0     1
##   4-5     0     0     1     0
##   5-6     0     0     0     1
##   6-7     0     0     0     0
##   7-8     0     0     0     0
```

We observe that desire to drink does appear to climb with negative relationship events, but there are so few observations with multiple negative relationship events that this finding cannot be taken at face value.

9. Additional Loess plots