W271 Spring 2018: Lab 2

Alyssa Eisenberg, Jeffrey Hsu, Gerard Kelly

Alcohol Consumption, Self-Esteem and Romantic Interactions

Introduction

The following analysis uses the **DeHartSimplified.csv** dataset to examine the main outcome variable of daily personal alcohol consumption and a secondary outcome of daily index of personal desire to drink. We are interested in whether the relationship between these outcome variables and an index level for daily negative romantic interactions differs according to the subject's index level for trait self-esteem because we hypothesize that those with low self-esteem will have higher desire to drink and alcohol consumption on days they experience more negative relationship interactions. We perform an exploratory data analysis on the dataset and propose count-based Poisson regression models and an ordinal response models for modeling alcohol consumption and desire to drink respectively. Confounding factors (such as age, gender and other negative/positive events) are considered on the basis on goodness of fit, parsimony, interpretability and model validity diagnostics, with only day of week fixed effects included in our final model. We find that across our final models, we do not see statistically significant results and cannot conclude that our effect of interest is different from zero.

Exploratory Data Analysis

```r
library(car); require(dplyr); library(Hmisc); library(tidyr); library(ggplot2);
library(gridExtra); library(reshape2); library(GGally); library(ggcorrplot);
library(package = MASS); library(sandwich); library(lmtest); library(stargazer)
```

```r
dehart <- read.table(file="DeHartSimplified.csv", header=TRUE, sep=",")
#describe(dehart)   ## comment out due to page limit
dehart$dayweek_f <- factor(dehart$dayweek); #categorical variable dayweek
levels(dehart$dayweek_f) = c("mon","tue","wed","thu","fri","sat","sun")
dehart$gender_f <- factor(dehart$gender);
levels(dehart$gender_f) = c("male","female") #categorical variable gender
```

This dataset has 623 observations and 13 variables, representing 7 daily entries in records kept by 89 study participants. There are a few missing values, noted at each variable below where they occur. No values appear top or bottom coded in the data.

id: an id number assigned to each unique participant (89 people, with 7 data points each)

studyday: encodes which day of the study it was for the participant. We have data for each participant for their first seven days of the study

dayweek: is the day of the week for the observation. 10 participants begin the study on Monday, 7 on Tuesday, 19 on Wednesday, 15 on Thursday, 16 on Friday, 6 on Saturday, and 16 on Sunday

numall: is the number of alcoholic drinks consumed on that day, an integer variable taking 18 distinct values, ranging from 0 to 21 drinks. Note that there is 1 missing value for participant id 42.

nrel: is a measure of negative romantic relationship interactions experienced on that day. It is a continuous variable taking 33 distinct values, ranging from 0 to 9

prel: is a measure of positive romantic relationship interactions experienced on that day. It is a continuous variable taking 68 distinct values, ranging from 0 to 9

negevent: is a combination of several scores on a 0-3 scale measuring the total number and intensity of negative events on the given day (a higher value indicating a larger number of negative events and/or more extremely negative events). It is a continuous variable with 131 distinct values, ranging from 0 to 2.4

posevent: is a combination of several scores on a 0-3 scale measuring the total number and intensity of positive events on the given day (a higher value indicating a larger number of positive events and/or more extremely positive events). It is a continuous variable with 216 distinct values, ranging from 0 to 3.9[1]

gender: is binary, with a slightly higher proportion of females (39 males, and 50 females)

rosn: is our measure for trait self-esteem, which is a long-term view of self-worth. This value was measured once at the beginning of the study, so the same value carries through all seven observations for each individual. It is a continuous variable taking 17 distinct values ranging from 2.1 to 4

age: is a continuous variable taking 89 distinct values ranging from 24.4 to 42.3

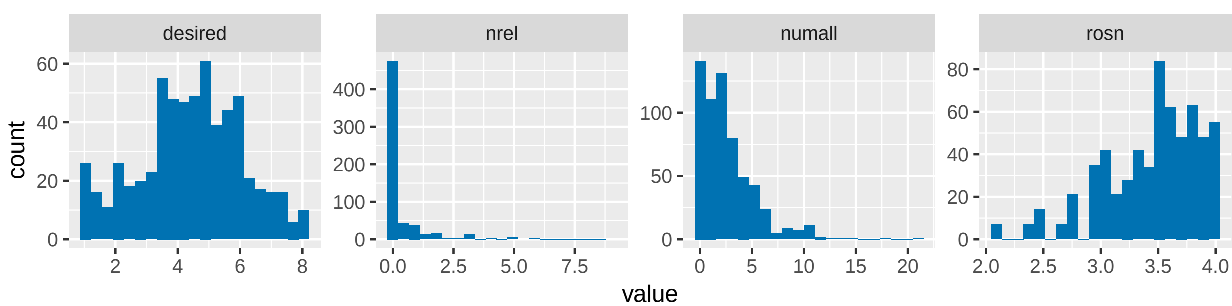
desired: measures the participants' desire to drink, with a higher score meaning a greater desire. It has 22 distinct values ranging from 1 to 8 in $\frac{1}{3}$ steps. Note that there are 3 missing values for participant ids 2, 110, 116.

state: is our measure for state self-esteem, which is a short-term view of self-worth. This was measured daily, unlike **rosn** long-term self-esteem. It is a continuous variable taking 25 distinct values ranging from 2.3 to 5. Note that there are 3 missing values for participant ids 2, 4, 110.

For the purposes of the statistical analysis, the small number of observations involving missing values will be removed from the dataset.

The following univariate plots illustrate the sample distributions for variables of key importance in analyzing our current hypothesis; outcomes **numall** and **desired**, as well as the explanatory variables **nrel** and **rosn**.

```
dehart <- na.omit(dehart)
dehart[,c("numall","desired","nrel","rosn")] %>% gather() %>%
  ggplot(aes(value)) + facet_wrap(~key, scales="free", nrow=1) +
  geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2")
```



The number of drinks consumed (**numall**) is highly positively skewed, with around 60% of the data spread fairly evenly across 0, 1, or 2 drinks, and only around 6% of the data at or above 7 drinks. In particular, the five data points with over 12 drinks may be high leverage, an issue to consider in our modeling. Desire to drink (**desired**), does not display this extreme skew. Instead, it looks relatively normal with a higher density area from 3.5 to 6, and tails of fairly uniform lower density out to 1 and 8.

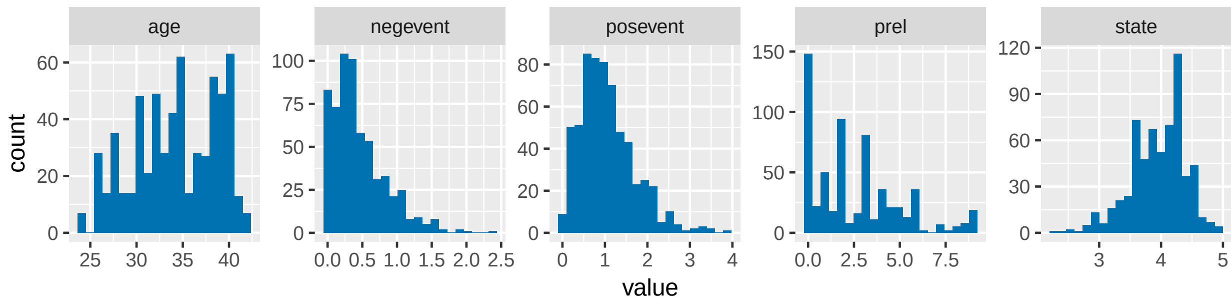
Our main independent variable of interest in analyzing our current hypothesis is the number of negative relationship events (**nrel**). This is highly positively skewed, with 77% of our data points having a value of 0. Among the remaining non-zero data, we still see a positive skew with most of the data having values

[1] Values above 3 may be suspicious since this variable combines several items scored on a 0-3 scale, but we cannot be sure that these are erroneous values since we do not know the combination formula or the individual values. There are only 8 rows with values over 3, and the rest of the data appears normal. Thus, we will assume these are valid data points.

less than 1, and only a couple points with values over 3. There is a particular outlier at a value of 8 which may have high leverage that we should check for during our modeling. This lack of variation in our primary independent variable of interest may make our analysis more challenging. Trait self-esteem (**rosn**) is another primary independent variable. There does not appear to be a clear break-point in the distribution that would easily divide the population into low vs. high self-esteem populations.

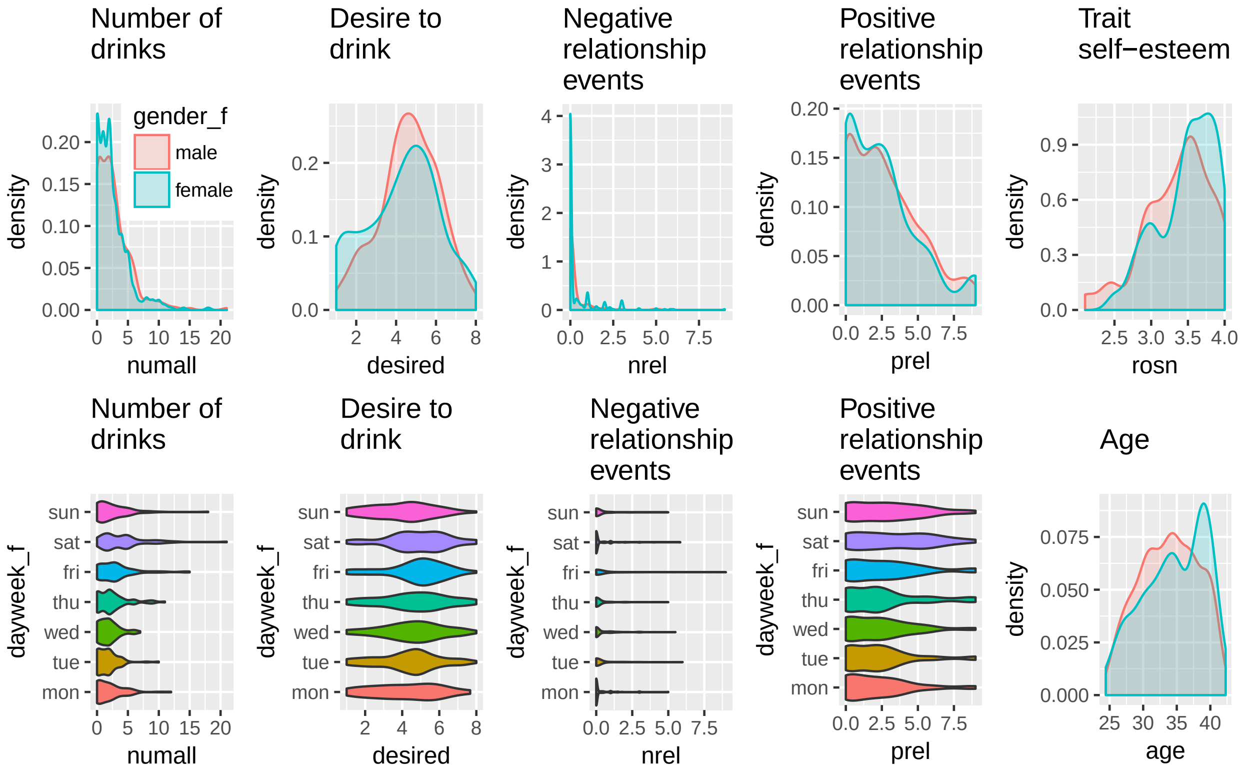
Examining the remaining continuous variables, we see that the other three variables measuring negative and positive event indexes (**negevent**, **posevent** and **prel**) also tend to be positively skewed, albeit less severely than **nrel**. The positive measures have a larger spread than the negative ones. State self-esteem (**state**) is negatively skewed but is much more symmetrically distributed than trait self-esteem (**rosn**), likely due to **state** having a larger number of unique observations (being a daily measure). Age has a relatively uniformly spread from the mid-twenties through to early forties age ranges, with slightly more participants on the older end of the spectrum.

```
dehart[,c("negevent","posevent","prel","age","state")] %>% gather() %>%
  ggplot(aes(value)) + facet_wrap(~key, scales="free", nrow=1) +
  geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2")
```



Having examined univariate distributions, we consider relationships between variables. Firstly, we plot how the two potentially confounding categorical variables **gender** and **dayweek** relate to the outcome variables **numall** and **desired**, as well as the explanatory variable **nrel** and the covariate **prel**. We also plot the interaction between **gender** and both **rosn** and **age**.

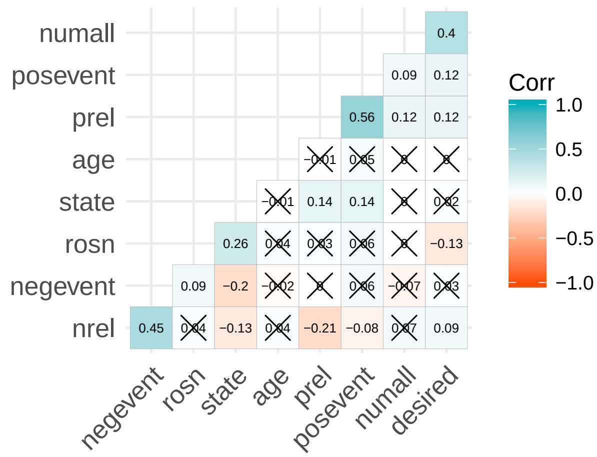
```
p1a<-ggplot(dehart, aes(x = numall, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) +
  ggtitle("Number of\ndrinks\n") + theme(legend.position=c(.65,.75))
p1b<-ggplot(dehart, aes(dayweek_f, numall)) + geom_violin(aes(fill = dayweek_f)) +
  ggtitle("Number of\ndrinks\n") + theme(legend.position="none") + coord_flip()
p2a<-ggplot(dehart, aes(x = desired, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2)+
  ggtitle("Desire to\ndrink\n") + theme(legend.position="none")
p2b<-ggplot(dehart, aes(dayweek_f, desired)) + geom_violin(aes(fill = dayweek_f)) +
  ggtitle("Desire to\ndrink\n") + theme(legend.position="none") + coord_flip()
p3a<-ggplot(dehart, aes(x = nrel, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) +
  ggtitle("Negative\nrelationship\nevents") + theme(legend.position="none")
p3b<-ggplot(dehart, aes(dayweek_f, nrel)) + geom_violin(aes(fill = dayweek_f)) +
  ggtitle("Negative\nrelationship\nevents") + theme(legend.position="none") + coord_flip()
p4a<-ggplot(dehart, aes(x = prel, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) +
  ggtitle("Positive\nrelationship\nevents") + theme(legend.position="none")
p4b<-ggplot(dehart, aes(dayweek_f, prel)) + geom_violin(aes(fill = dayweek_f)) +
  ggtitle("Positive\nrelationship\nevents") + theme(legend.position="none") + coord_flip()
p5<-ggplot(dehart, aes(x = rosn, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) +
  ggtitle("Trait\nself-esteem\n") + theme(legend.position="none")
p6<-ggplot(dehart, aes(x = age, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2)+
  ggtitle("\nAge\n") + theme(legend.position="none")
grid.arrange(p1a, p2a, p3a, p4a, p5, p1b, p2b, p3b, p4b, p6, ncol = 5)
```



Differences by gender are not particularly marked for most variables. A slightly larger proportion of females than males record lower daily numbers of drinks consumed, while a larger proportion of females than males record lower daily desire to drink. There is also a larger proportion of females with higher trait self-esteem and a larger proportion of females in the older age range. Differences by days of the week are more likely to have confounding effects on the relationship between our key variables. There are significant positive skews on the distributions for the number of drinks for Friday, Saturday and Sunday (and higher median numbers of drinks on Friday and Saturday). Mean levels for the desire to drink appear to shift gradually higher throughout the week from Monday through to Saturday before dropping to the lowest level on Sunday (with variance higher on Sunday and Monday than on other days). Differences by weekday in negative and positive relationship events are minor; values for negative relationship events are skewed to a similarly extreme degree on all days (with one outlier on Friday), while values for positive relationship events are higher on average for Saturday and Sunday than for other days. Similar plots for **posevent**, **negevent** and **state** (which we do not present) indicate little interaction between these variables, gender and weekday.

Further relationships between our numeric variables are summarized in the below table of correllation coefficients.

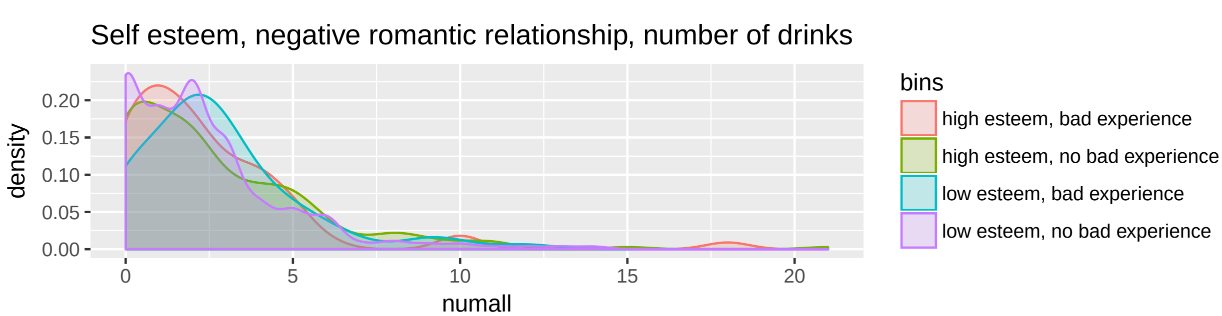
```
corr <- round(cor(dehart[,c(4,5,6,7,8,10,11,12,13)]), 2)
ggcorrplot(corr, p.mat = cor_pmat(dehart[,c(4,5,6,7,8,10,11,12,13)]),hc.order = TRUE,
           type="lower", color=c("#FC4E07","white","#00AFBB"), lab=TRUE,  lab_size=2)
```



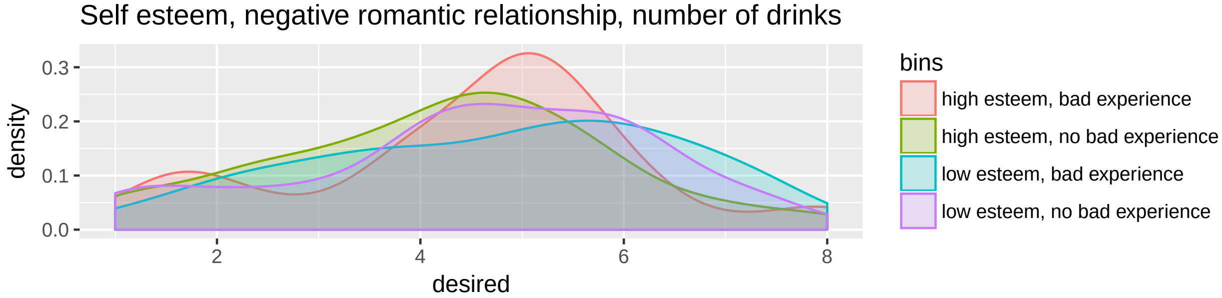
There are fairly strong, positive relationships between the number of drinks and the desire to drink, between positive romantic events and all postive events, and between negative romantic events and all negative events. There is a mild negative correlation between positive and negative romantic relationship events. We also see a moderately positive correlation between trait self-esteem and state self-esteem. In general, these relationships are fairly weak (as can be observed using bivariate scatterplots, which we do not include here). Of particular relevance to the hypothesis are the very weak positive correlations between negative romantic relationship events and both the number of drinks consumed and the desire to drink. There is a marginally negative relationship between trait self-esteem and the desire to drink but none between trait self-esteem and the number of drinks consumed.

Considering the question of whether trait self-esteem affects the relationship between negative romantic relaitonship events and either drinks consumed or the desire to drink, we examine distributions of the variables **numall** and **desired** for four subsets of the data representing combinations of above- and below-median values for negative romantic events and trait self-esteem (**nrel** and **rosn**). Since the median value of **nrel** is zero, this split corresponds to the presence or absence of any negative romantic relationship event. We consider the transformation of **nrel** from a continuous index to a two-level categorical variable to be useful for statistical modeling as well as this visualization, due to the variable's extreme skew (with much of its variation accounted for by a small share of observations). Furthermore, the index **nrel** measures not only the number but also the intensity of these events. Due to the subjectivity involved in this combined measure, the binary version of the variable may be more relevant.

```
dehart$bins[dehart$rosn <= median(dehart$rosn) & dehart$nrel <= median(dehart$nrel)] = "low esteem, no b
dehart$bins[dehart$rosn <= median(dehart$rosn) & dehart$nrel > median(dehart$nrel)] = "low esteem, bad e
dehart$bins[dehart$rosn > median(dehart$rosn) & dehart$nrel <= median(dehart$nrel)] = "high esteem, no b
dehart$bins[dehart$rosn > median(dehart$rosn) & dehart$nrel > median(dehart$nrel)] = "high esteem, bad e
ggplot(dehart, aes(numall, fill = bins, colour = bins)) + geom_density(alpha=0.2) +
  ggtitle("Self esteem, negative romantic relationship, number of drinks")
```

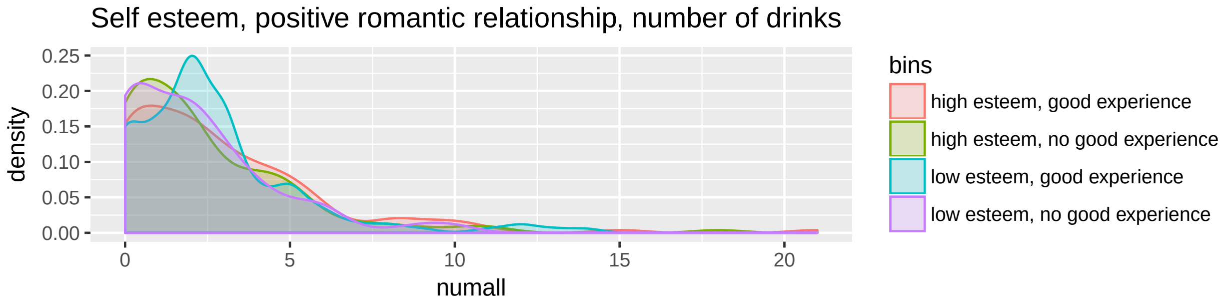


```
ggplot(dehart, aes(desired, fill = bins, colour = bins)) + geom_density(alpha=0.2) +
  ggtitle("Self esteem, negative romantic relationship, number of drinks")
```

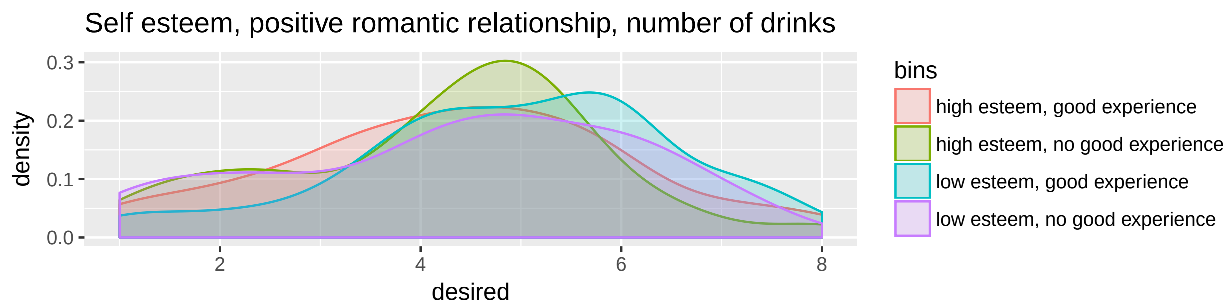


These comparisons show a difference in the number of drinks consumed by respondents who have had a bad negative relationship experience, according to whether or not they have high or low self-esteem, with the lower self-esteem individuals drinking notably more. Distributions in the high self-esteem category are quite similar regardless of whether a bad experience occurred on not; in fact, the presence of bad romantic experiences has a less flat distribution with a lower density at higher number of drinks. Distributions for low self-esteem individuals differs much more noticably betwen those who have had a bad relationship experience that day and those who haven't. Bad romantic relationship experiences are associated with higher average desire to drink for both high and low self-esteem individuals, however the presence of bad romantic relationship experiences seems to have sligthly less effect on the desire to drink for low esteem individuals than for high esteem individuals. Interestingly it appears that with regard to drinks consumed, low self-esteem individuals have a more significant difference from high self-esteem individuals in their response to *good* romantic relationship expeirences, as illustrated:

```
dehart$bins[dehart$rosn <= median(dehart$rosn) & dehart$prel <= median(dehart$prel)] = "low esteem, no g
dehart$bins[dehart$rosn <= median(dehart$rosn) & dehart$prel > median(dehart$prel)] = "low esteem, good
dehart$bins[dehart$rosn > median(dehart$rosn) & dehart$prel <= median(dehart$prel)] = "high esteem, no g
dehart$bins[dehart$rosn > median(dehart$rosn) & dehart$prel > median(dehart$prel)] = "high esteem, good
ggplot(dehart, aes(numall, fill = bins, colour = bins)) + geom_density(alpha=0.2) +
  ggtitle("Self esteem, positive romantic relationship, number of drinks")
```




```
ggplot(dehart, aes(desired, fill = bins, colour = bins)) + geom_density(alpha=0.2) +
  ggtitle("Self esteem, positive romantic relationship, number of drinks")
```



Modeling

Initial Model Specification

First, we decided to run two types of model based on the output variable of interest. As already stated, number of drinks is our primary outcome and desire to drink is a secondary outcome variable. For the number of drinks model, we chose a Poisson model because this is a count variable and its distribution approximately followed a Poisson model. While we also considered binning the variable and using a proportional odds model, we did not see clear break points in the distribution to support that choice. For the desire to drink model, we chose a proportional log odds model because it is a ranking on a scale of 1 to 8. Thus, a Poisson model is not appropriate because the output is not a count variable, and a linear regression is not appropriate because it is a ranked output instead of a continuous one. Based on its univariate distribution, we decided to bin the variable into low (1-3.5), mid (3.5-6), and high (6-8) desire to drink.

For each of these models, we decided to run them on all observations without missing values (fewer than 5 had missing values). While we are aware that this violates the model assumption of independence of the observations since each individual has 7 data points collected over time, we wanted to take advantage of the increased number of observations due to the low variation in negative relationship interactions. Due to this, in our analysis we will use the more conservative clustered robust standard errors. We also plan on validating the results on the full data set by also running our final model seven more times on data for each individual day of the week.

For each output, we will want to run a base model, an intermediate model, and a full model to check robustness of any results to model specification. Beyond these initial models, we will check for statistical significance to see if we should change any variable inclusion from the intermediate model to a final model.

In our base model, we will include only the variables of interest: dummy for negative relationship interactions (see EDA for reasoning behind this transformation), trait self-esteem, and their interaction.

In the intermediate model, we will add day-of-week fixed-effects, positive relationship interactions, and the interaction of positive interactions with trait self-esteem. We saw a relationship among each of these variables in our EDA, and they also theoretically make sense to have an impact on desire to drink and number of drinks consumed.

Finally, the full model will include all of our other variables: age, gender, negative events, positive events, and state self-esteem. These are variables where we did not see a strong relationship with our outcome variables in the EDA.

We now create all the Poisson models for number of drinks discussed above on all observations. While we also examined the desire to drink model, due to space constraints we will only show and discuss the final model version for this secondary outcome variable. Please see Table 1 columns 1, 2, and 4 for the model results using their clustered robust standard errors.

```
# Additional variables:
dehart$nrel_f <- factor(as.numeric(dehart$nrel != 0))
dehart$desired_binL <- cut(dehart$desired, breaks=c(-1, 3.5, 6, 8))
```

```r
# Poisson models for number of drinks:
pois_base <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn,
                 data=dehart, family=poisson(link=log))
pois_int <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn + prel + prel:rosn
                + dayweek_f, data=dehart, family=poisson(link=log))
pois_full <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn + prel + prel:rosn
                 + dayweek_f + age + gender_f + negevent + posevent + state,
                 data=dehart, family=poisson(link=log))
```

Final Model Choice

From our intermediate model, we checked whether the additional variables included beyond the base model were significant using a likelihood ratio test and Wald tests using clustered standard errors. The LRT assumes independence, which we know does not hold, and the Wald test assumes a normal distribution, which can be reasonable due to having a large enough sample. We found that the significance depends on the type of test and standard errors used, and thus followed the more conservative results from the Wald test with clustered standard errors in determining statistical significance. We found that prel and its interaction with rosn were jointly not significant, but that the day of week fixed effects were significant with a p-value of practically 0 (omitted full test results for brevity). Thus, we choose to remove prel and its interaction term from our final model.

```r
pois_int_test <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn
                + dayweek_f, data=dehart, family=poisson(link=log))
waldtest(pois_int,pois_int_test,vcov=vcovCL(pois_int, cluster=dehart$id))
```

```
## Wald test
##
## Model 1: numall ~ nrel_f + rosn + nrel_f:rosn + prel + prel:rosn + dayweek_f
## Model 2: numall ~ nrel_f + rosn + nrel_f:rosn + dayweek_f
##   Res.Df Df      F Pr(>F)
## 1    606
## 2    608 -2 1.4493 0.2356
```

We also checked the additional variables added in the full model to see if any significantly explained additional variation to include in our final model. We found that none of them were individually statistically significant.

```r
coeftest(pois_full, vcov = vcovCL(pois_full, cluster=dehart$id))[c("age", "gender_ffemale", "negevent",
```

```
##                    Estimate Std. Error    z value  Pr(>|z|)
## age             0.001748351 0.01205351  0.1450491 0.8846721
## gender_ffemale -0.124406787 0.11449946 -1.0865273 0.2772458
## negevent       -0.213315313 0.13547925 -1.5745239 0.1153664
## posevent        0.069870545 0.08395339  0.8322541 0.4052656
## state          -0.112374307 0.11095001 -1.0128373 0.3111379
```

Based on this analysis, we created a final Poisson model (see results in Table 1 column 3). We then compared the AIC values for a measure of in-sample fit including a penalty for more parameters between our various model specifications. While the final model did not have the lowest AIC, it was fairly similar to that of the intermediate and full models.

```r
# Create final model
pois_final <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn + dayweek_f, data=dehart, family=poiss

# Obtain AIC values
cbind(base=AIC(pois_base),int=AIC(pois_int), final=AIC(pois_final),full=AIC(pois_full))
```

```
##          base      int    final     full
## [1,] 2949.037 2810.925 2829.847 2803.752
```

Finally, for robustness we ran the final model for number of drinks consumed seven times, once for each day of the week. Due to space constraints, we only include the version for Friday and Saturday here since there are more social interactions than other days of the week. Note that we also went through a similar process to this for the desire to drink model. We create the final model here, but omit the details leading up to it due to space constraints. These model outputs can be found in Table 1, columns 5-7.

```
# Restrict final model to observations from Fri and Sat
pois_final_fri <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn,
                      data=subset(dehart, dayweek_f=="fri"), family=poisson(link=log))

pois_final_sat <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn,
                      data=subset(dehart, dayweek_f=="sat"), family=poisson(link=log))

# Create final model for desire to drink
prop_odds_final <- polr(formula = desired_binL ~ nrel_f + rosn + nrel_f:rosn
                        + dayweek_f, data=dehart, method="logistic")
```

With all our models created, we create Table 1 with the output of all coefficient values and their clustered robust standard errors.

```
# Get clustered robust SE for models with all observations, robust SE for Fri model
se.pois_base <- sqrt(diag(vcovCL(pois_base, cluster=dehart$id)))
se.pois_int <- sqrt(diag(vcovCL(pois_int, cluster=dehart$id)))
se.pois_final <- sqrt(diag(vcovCL(pois_final, cluster=dehart$id)))
se.pois_full <- sqrt(diag(vcovCL(pois_full, cluster=dehart$id)))
se.pois_final_fri <- sqrt(diag(vcovHC(pois_final_fri)))
se.pois_final_sat <- sqrt(diag(vcovHC(pois_final_sat)))
se.prop_odds_final <- sqrt(diag(vcovCL(prop_odds_final, cluster=dehart$id)))
```

```
##
## Re-fitting to get Hessian
```

```
# Output stargazer table
stargazer(pois_base,pois_int, pois_final, pois_full, pois_final_fri, pois_final_sat, prop_odds_final,
          se = list(se.pois_base, se.pois_int, se.pois_final, se.pois_full,
                    se.pois_final_fri, se.prop_odds_final),
          column.labels=c("Base", "Int", "Final", "Full", "Fri", "Sat", "Final"),
          omit.stat = c("aic", "ll"), star.cutoffs = c(.05,.01,.001), header=F, no.space=TRUE)
```

Residual Diagnostics

Due to space constraints, we only show residual diagnostics for the final Poisson model on all observations here. However, we saw similar patterns for the individual day Poisson models.

In the plots of Pearson residuals vs. each explanatory variable and the linear predictor, we see that the residual means are pretty constant across the explanatory variable range (for the individual day models, there was a bit more curvature to the means). However, the residual variances are clearly different for trait self-esteem and day of week. This heteroskedasticity indicates we should be using robust standard errors. The other thing to note in these plots is that there are numerous extreme values of residuals greater than 2. This is confirmed in the histogram of the residuals showing a positive skew and the q-q plot showing curvature away from the normal line. This indicates overdispersion in our model, potentially indicating that we have omitted variables which would help reduce the additional variability to the counts that the model

Table 1:

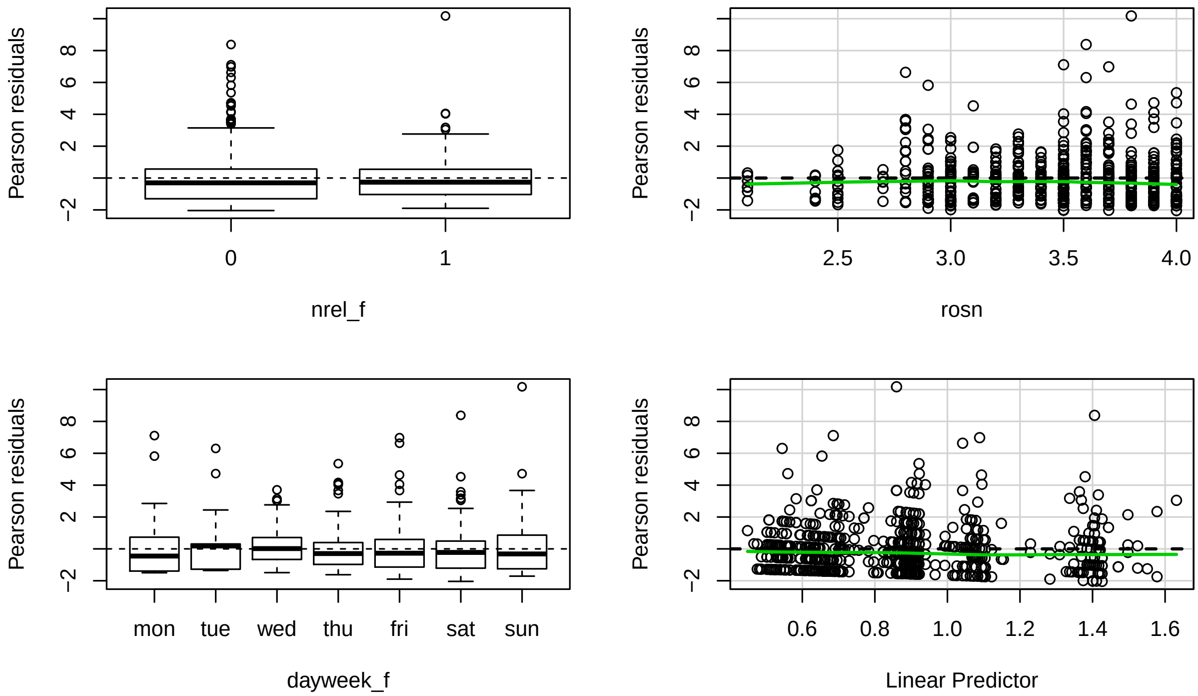
	Dependent variable:						
	numall						desired_binL
	Poisson						*ordered logistic*
	Base	Int	Final	Full	Fri	Sat	Final
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
nrel_f1	1.061	1.509*	1.162	1.321*	−1.006	2.416	2.565
	(0.771)	(0.654)	(0.729)	(0.640)	(1.917)	(2.016)	(1.762)
rosn	0.050	0.285	0.052	0.329	−0.365	0.044	−0.623**
	(0.154)	(0.185)	(0.155)	(0.201)	(0.272)	(0.301)	(0.215)
prel		0.267		0.259			
		(0.186)		(0.200)			
dayweek_ftue		−0.139	−0.146	−0.125			0.492
		(0.157)	(0.155)	(0.156)			(0.301)
dayweek_fwed		−0.071	−0.066	−0.042			0.568
		(0.129)	(0.128)	(0.130)			(0.303)
dayweek_fthu		0.204	0.210	0.220			0.575
		(0.117)	(0.117)	(0.116)			(0.304)
dayweek_ffri		0.379**	0.393**	0.386**			0.904**
		(0.132)	(0.134)	(0.135)			(0.303)
dayweek_fsat		0.679***	0.715***	0.687***			0.979**
		(0.142)	(0.146)	(0.140)			(0.303)
dayweek_fsun		0.191	0.211	0.186			−0.080
		(0.160)	(0.161)	(0.162)			(0.302)
age				0.002			
				(0.012)			
gender_ffemale				−0.124			
				(0.114)			
negevent				−0.213			
				(0.135)			
posevent				0.070			
				(0.084)			
state				−0.112			
				(0.111)			
nrel_f1:rosn	−0.292	−0.407*	−0.320	−0.335	0.270	−0.678	−0.674
	(0.226)	(0.194)	(0.214)	(0.189)	(0.573)	(0.578)	(0.503)
rosn:prel		−0.068		−0.067			
		(0.051)		(0.055)			
Constant	0.742	−0.389	0.503	−0.079	2.342*	1.240	
	(0.544)	(0.667)	(0.558)	(0.862)	(0.916)		
Observations	618	618	618	618	88	89	618

Note: *p<0.05; **p<0.01; ***p<0.001

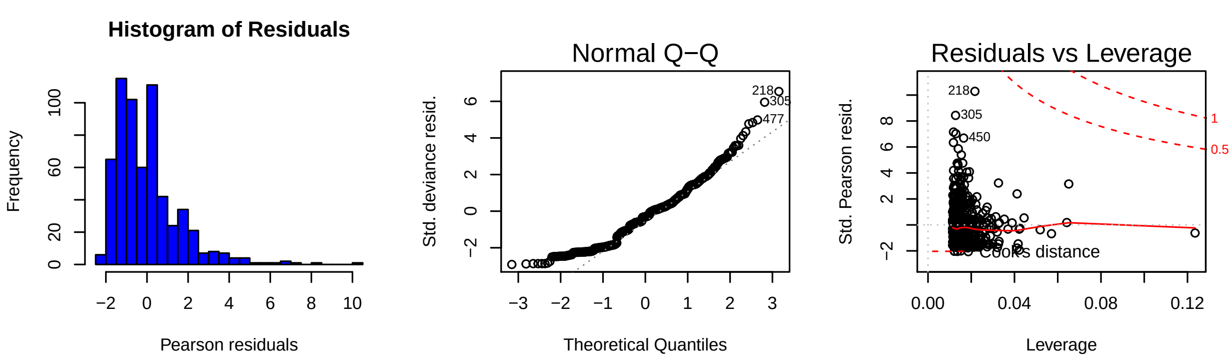
cannot currently explain. A couple examples of potential missing explanatory variables are socioeconomic status, attractiveness, and quality of friendships.

Finally, we checked the residuals vs. leverage plot, but did not see any points with high influence (Cook's distance>1) that we would need to be concerned about and check robustness without those outliers.

```r
# Pearson residual plots
suppressWarnings(residualPlots(pois_final, test=FALSE))
```




```r
# Additional plots - histogram of residuals and leverage/influence plot
par(mfrow=c(1,3))
hist(residuals(pois_final, "pearson"), breaks = 20, col = "blue",
     xlab="Pearson residuals", main="Histogram of Residuals")
plot(pois_final, which=2)
plot(pois_final, which=5)
```



Interpret Model Results

Returning to our hypothesis, the coefficient of interest to us is the interaction between negative relationship interactions and trait self-esteem. In our final model for number of drinks, we see that the expected mean number of drinks decreases by 0.98 times for every 0.4 decrease (one standard deviation) in trait self-esteem when negative relationship interactions are not present. When negative relationship interactions are present, the expected mean number of drinks instead increases by 1.12 times for every 0.4 decrease in trait self-esteem. Across all our models in Table 1 except the Friday specific model, this coefficient remains negative with a moderate magnitude. This is in the correct direction to support the hypothesis that negative relationship interactions are positively associated with drinking for those with low trait self-esteem. However, this coefficient is only statistically significant in one specification (our intermediate specification of the number of drinks).

Due to this lack of robustness in the day-specific models and in the model for desire to drink as well as the potential for omitted variables in the model, we have to conclude that we fail to reject the null that this interaction is 0 and there is no difference in effect of negative relationship interactions based on level of trait self-esteem.

```r
# Interpretation of coefficient when nrel=0
unname(round(1/exp(sd(dehart$rosn)*pois_final$coefficients["rosn"]),2))
```

```
## [1] 0.98
```

```r
# Interpretation of coefficient when nrel=1
unname(round(1/exp(sd(dehart$rosn)*(pois_final$coefficients["rosn"]+pois_final$coefficients["nrel_f1:ro
```

```
## [1] 1.12
```

Summary