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 test the hypotheses that

- *negative interactions with romantic partners are associated with an increase in alcohol consumption and increased desire to drink*

and that

- *people with low trait self-esteem drink more on days they experienced more negative relationship interactions compared with days during which they experienced fewer negative relationship interactions (with this relation between drinking and negative relationship interactions not evident for individuals with high trait self-esteem).*

Throughout the analysis, the main outcome variables are daily personal alcohol consumption and a 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. We perform an exploratory data analysis on the dataset and propose count-based Poisson regression models and an ordinal response models for modelling 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.

Exploratory Data Analysis

```
library(car); require(dplyr);
library(Hmisc); library(mcprofile);
library(ggplot2); library(gridExtra);
library(reshape2); library(GGally);
library(ggcorrplot); library(package = MASS)
```

```
dehart <- read.table(file="DeHartSimplified.csv",
                     header=TRUE, sep=",")
#describe(dehart)   ## comment out due to page limit
dehart$dayweek_f <- factor(dehart$dayweek);
levels(dehart$dayweek_f) = c("mon","tue","wed",
                             "thu","fri","sat","sun") #categorical variable dayweek
dehart$gender_f <- factor(dehart$gender);
levels(dehart$gender_f) = c("male","female") #categorical variable gender
dehart$nrel_bool = 1*(dehart$nrel > 0) #nrel variable into 2 bins
dehart$rosn_cat <- cut(dehart$rosn,    #rosn variable into 3 bins
                       breaks=c(-1, 2.8, 3.4, Inf),
                       labels = c("low", "mid", "high"))
dehart$trel <- dehart$nrel/sd(dehart$nrel) + dehart$prel/sd(dehart$prel)
#total relationship events; sum of normalized nrel and prel
#dehart_cleaned <- dehart[!is.na(dehart$numall) & !is.na(dehart$desired),] #dunno if this is actually n
```

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. There are 89 unique participants in the dataset, each with 7 data points

studyday: encodes which day of the study it was for the participant. There are 89 observations for each of study days 1 through 7, meaning that we have data for each participant for their first seven days of the study

dayweek: is the day of the week for the observation. There are 89 observations for each day of the week 1 through 7 (Monday is coded as 1). 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 items scored 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 taking 131 distinct values, ranging from 0 to 2.4

posevent: is a combination of several items scored 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 taking 216 distinct values, ranging from 0 to 3.9[1]

gender: is coded as 1 (male) or 2 (female), 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 is a continuous variable taking 22 distinct values ranging from 1 to 8. Note that there are 3 missing values for participant ids 2, 110, 116.

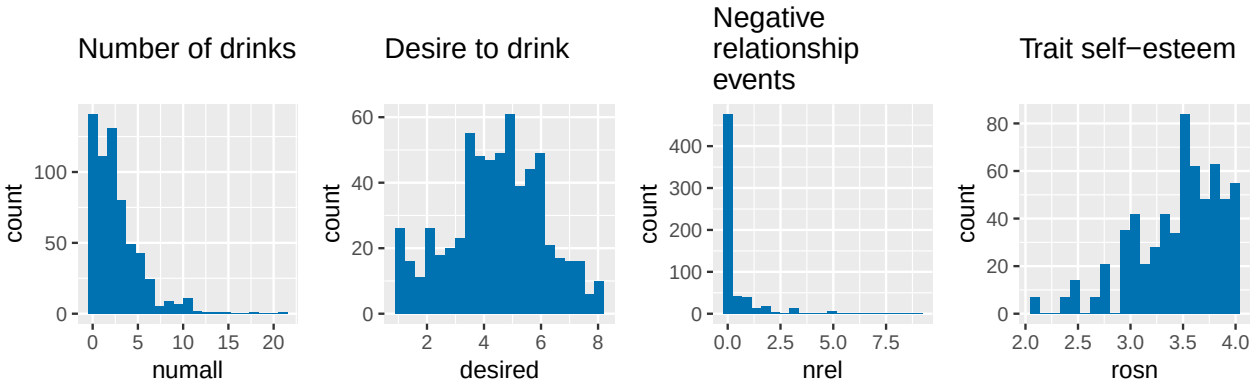
state: is our measure for state self-esteem, which is a short-term view of self-worth. This was measured daily, unlike *rosn* long-term self-esteem. It is a continuous variable taking 25 distinct values ranging from 2.3 to 5. Note that there are 3 missing values for participant ids 2, 4, 110.

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 four variables of importance in analyzing our current hypothesis; outcome variables **numall** and **desired**, as well as the key explanatory variables **nrel** and **rosn**.

[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.

```
dehart <- na.omit(dehart)
p1 <- ggplot(dehart, aes(x = numall)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2") +
p2 <- ggplot(dehart, aes(x = desired)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2") +
p3 <- ggplot(dehart, aes(x = nrel)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2") + gg
p4 <- ggplot(dehart, aes(x = rosn)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2") + gg
grid.arrange(p1, p2, p3, p4, ncol = 4)
```

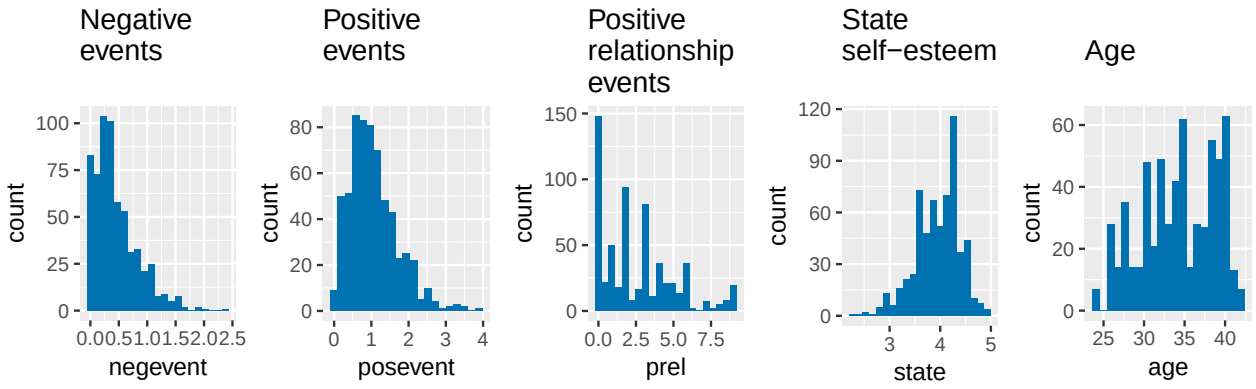


The number of drinks consumed (**numall**) is highly positively skewed, with around 60% of the data spread fairly evenly across 0, 1, or 2 drinks, and only around 6% of the data at or above 7 drinks. In particular, the five data points with over 12 drinks may be high leverage, an issue to consider in our modeling. Our other potential output variable, desire to drink (**desired**), does not display this extreme skew. Instead, it looks relatively normal with a higher density area from 4 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. 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 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 look for during our modeling. This lack of variation in our primary independent variable of interest may make our analysis more challenging. Trait self-esteem is another primary independent variable, as we are investigating differences in behavior between those with high vs low trait self-esteem. 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.

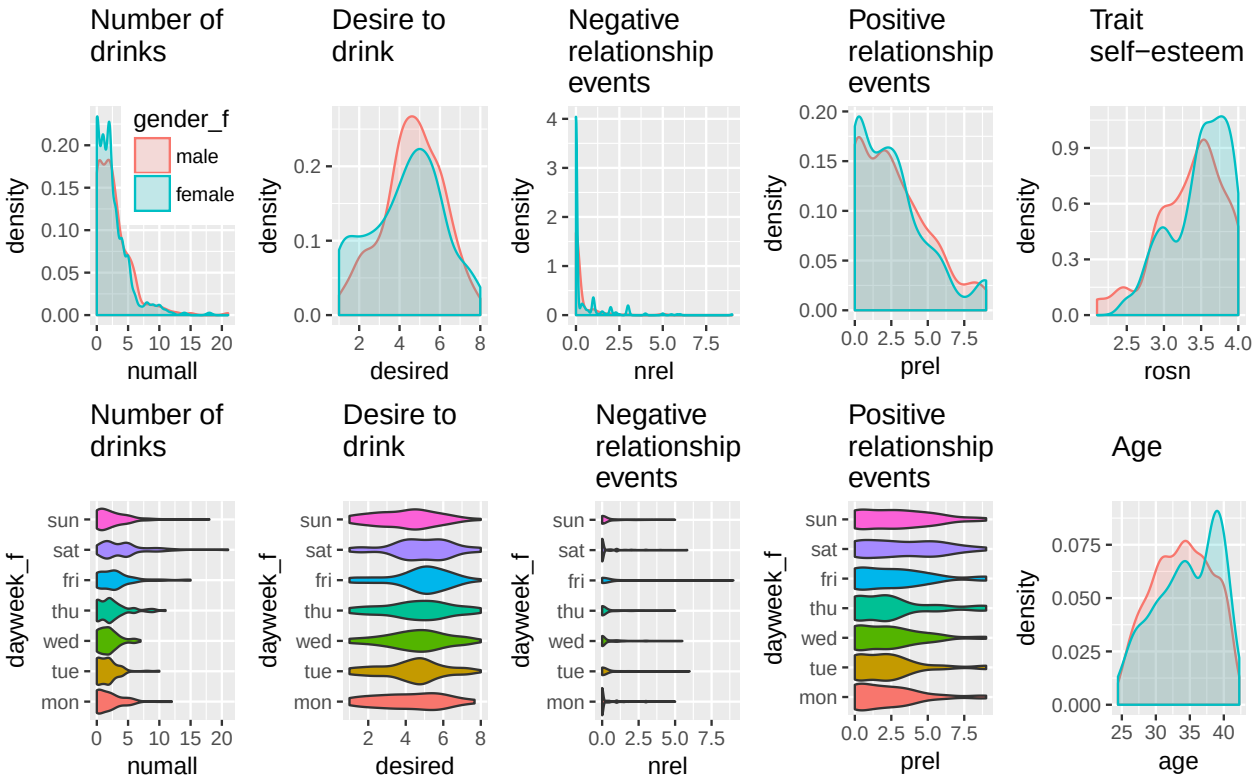
Examining the remaining continuous variables, we see that the other three varaibles measuring negative and positive event indexes (**negevent**, **posevent** and **prel**) also tend to be positively skewed, albeit less severely than **nrel**. The positive measures have a larger spread than the negative 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 much 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.

```
p1 <- ggplot(dehart, aes(x = negevent)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2")
p2 <- ggplot(dehart, aes(x = posevent)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2")
p3 <- ggplot(dehart, aes(x = prel)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2") + gg
p4 <- ggplot(dehart, aes(x = age)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2") + ggt
p5 <- ggplot(dehart, aes(x = state)) + geom_histogram(aes(y = ..count..), bins = 21, fill="#0072B2") + g
grid.arrange(p1, p2, p3, p5, p4, ncol = 5)
```



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) + gg
p1b<-ggplot(dehart, aes(dayweek_f, numall)) + geom_violin(aes(fill = dayweek_f)) + ggtitle("Number of\n
p2a<-ggplot(dehart, aes(x = desired, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) + gg
p2b<-ggplot(dehart, aes(dayweek_f, desired)) + geom_violin(aes(fill = dayweek_f)) + ggtitle("Desire to\n
p3a<-ggplot(dehart, aes(x = nrel, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) + ggti
p3b<-ggplot(dehart, aes(dayweek_f, nrel)) + geom_violin(aes(fill = dayweek_f)) + ggtitle("Negative\nrela
p4a<-ggplot(dehart, aes(x = prel, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) + ggti
p4b<-ggplot(dehart, aes(dayweek_f, prel)) + geom_violin(aes(fill = dayweek_f)) + ggtitle("Positive\nrela
p5<-ggplot(dehart, aes(x = rosn, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) + ggtitl
p6<-ggplot(dehart, aes(x = age, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2) + ggtitle
grid.arrange(p1a, p2a, p3a, p4a, p5, p1b, p2b, p3b, p4b, p6, ncol = 5)
```

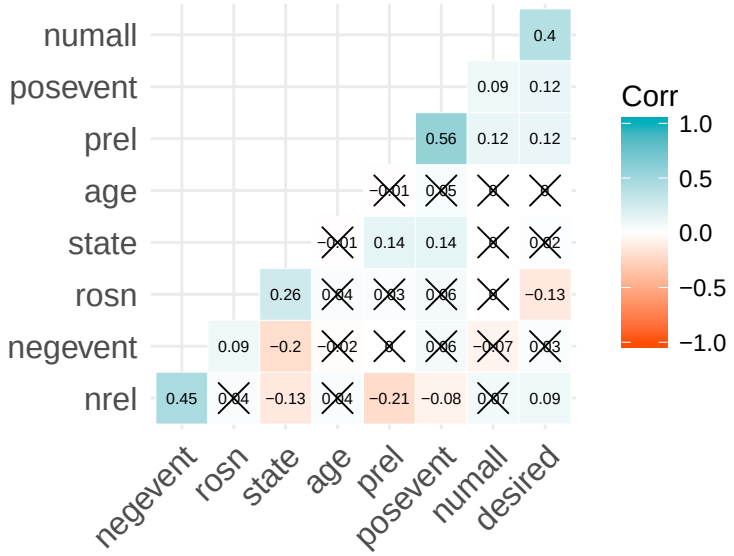


Differences by gender are not particularly marked for most variables. A slighly larger proportion of females

than males record lower daily numbers of drinks consumed, while a more significantly 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 significant 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.

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

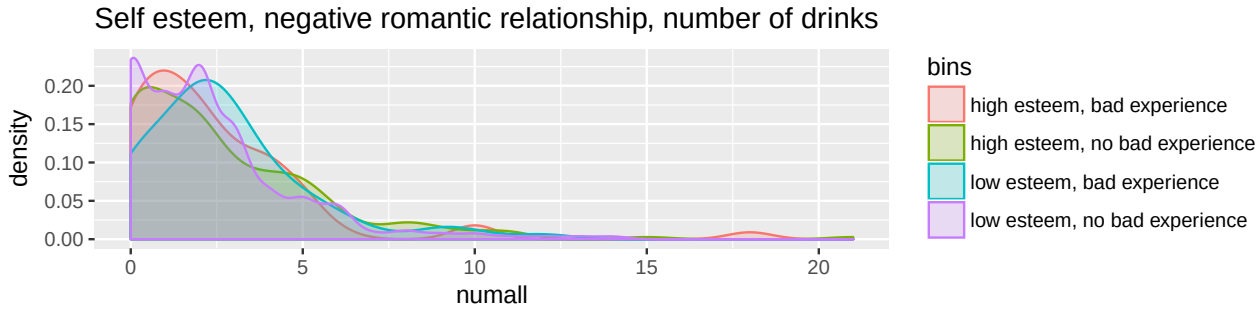


There are fairly strong, positive relationship between the number of drinks and the desire to drink, between positive romantic events and all postive events, and between negative romantic events and all negative events. There is a mild negative correlation between positive and negative romantic relationship events. We also see a moderately positive correlation between trait self-esteem and state self-esteem. In general however relationships between these variables 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 significant negative relationship between trait self-esteem and the desire to drink but none between trait self-estee and the number of drinks consumed.

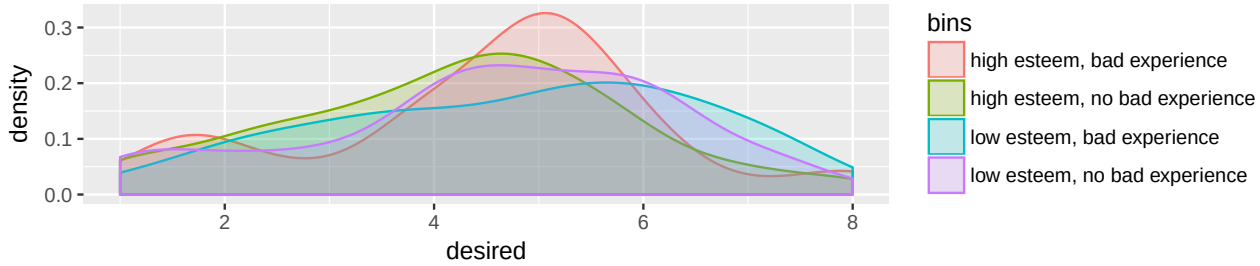
Considering the question of whether trait self-esteem affects the relationship between negative romantic relaitonship events and either drinks consumed or the desire to drink, we examine distributions of the variables **numall** and **desired** for four subsets of the data representing combinations of 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. Although these transformations are made for visualisation of differences, we consider the transformation of **nrel** from a continuous index to a two-level categorical variable to be useful for statistical modeling

as well, due to the variable's extreme skew (with much of its variation accounted for by a small share of observations). Furthermore, although the hypothesis considers the difference between "more" and "fewer" romantic relationship events, 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(na.omit(dehart), aes(numall, fill = bins, colour = bins)) +
  geom_density(alpha=0.2) + ggtitle("Self esteem, negative romantic relationship, number of drinks") + g
```

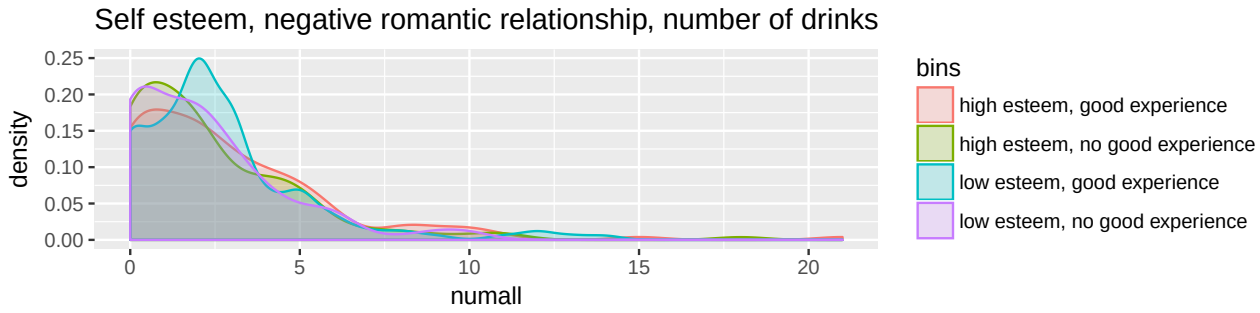



```
ggplot(na.omit(dehart), aes(desired, fill = bins, colour = bins)) +
  geom_density(alpha=0.2) + ggtitle("Self esteem, negative romantic relationship, number of drinks") + g
```



These comparisons show a difference in the number of drinks consumed by respondants who have had a bad negative relationship experience, according to whether or not they have high or low self-esteem, wth 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, at higher numbers of drinks the presence of bad romantic experiences is here associated with *less* drinks consumed. 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(na.omit(dehart), aes(numall, fill = bins, colour = bins)) +
  geom_density(alpha=0.2) + ggtitle("Self esteem, positive romantic relationship, number of drinks") + g
```




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

Modeling

Summary