

Lab 1, Short Questions

Contents

1 Alcohol, self-esteem and negative relationship interactions (5 points)	1
1.1 EDA	2
1.2 Hypothesis One	4
1.3 Hypothesis Two	6

```
library(readr)
library(dplyr)
library(patchwork)
library(forcats)
library(ggplot2)
library(car)
```

1 Alcohol, self-esteem and negative relationship interactions (5 points)

Read the example ‘**Alcohol Consumption**’ in chapter 4.2.2 of the textbook (Bilder and Loughin’s “Analysis of Categorical Data with R”). This is based on a study in which moderate-to-heavy drinkers (defined as at least 12 alcoholic drinks/week for women, 15 for men) were recruited to keep a daily record of each drink that they consumed over a 30-day study period. Participants also completed a variety of rating scales covering daily events in their lives and items related to self-esteem. The data are given in the *DeHartSimplified.csv* data set. Questions 24-26 of chapter 3 of the textbook also relate to this data set and give definitions of its variables: the number of drinks consumed (**numall**), positive romantic-relationship events (**prel**), negative romantic-relationship events (**nrel**), age (**age**), trait (long-term) self-esteem (**rosn**), state (short-term) self-esteem (**state**).

The researchers stated the following hypothesis:

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.

```
drinks <- read_csv('../data/short-questions/DeHartSimplified.csv')
```

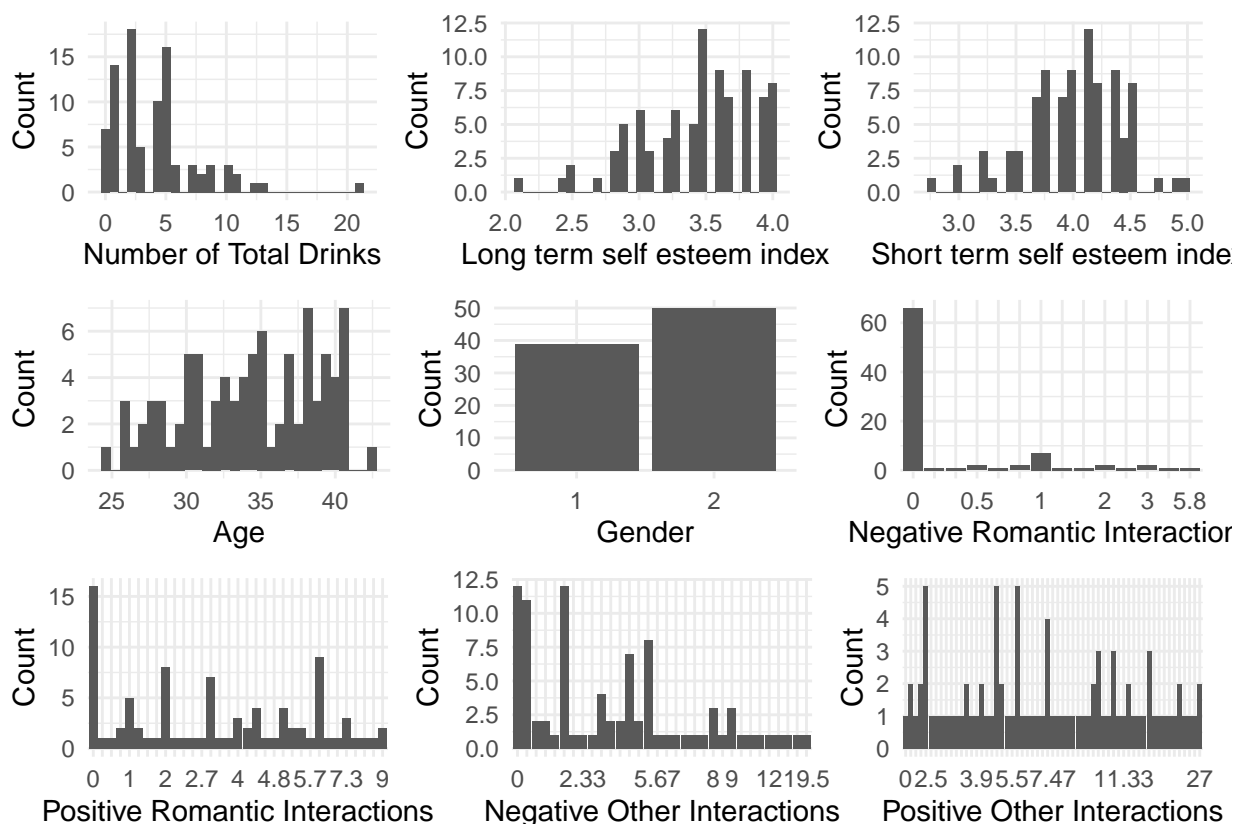
1.1 EDA

(2 points) Conduct a thorough EDA of the data set, giving special attention to the relationships relevant to the researchers' hypotheses. Address the reasons for limiting the study to observations from only one day.

```
drinks_subset <- drinks %>%
  dplyr::filter(dayweek == 6) %>%
  dplyr::select(numall, nrel, prel, rosn,
                negevent, posevent, age, gender, state) %>% # additional controls
  dplyr::mutate(nrel = round(nrel, 1),
               prel = round(prel, 1),
               negother = round(10*negevent - nrel, 2),
               posother = round(10*posevent - prel, 2)) %>%
  dplyr::select(-c("negevent", "posevent"))

drinks_subset$gender <- as_factor(drinks_subset$gender)
## As per the paper, 1 = Male and 2 = Female
```

1.1.1 Starting with some basic histograms



A few key things are revealed and by the histograms and bar charts:

1. Most people drink less than 5 drinks a day.

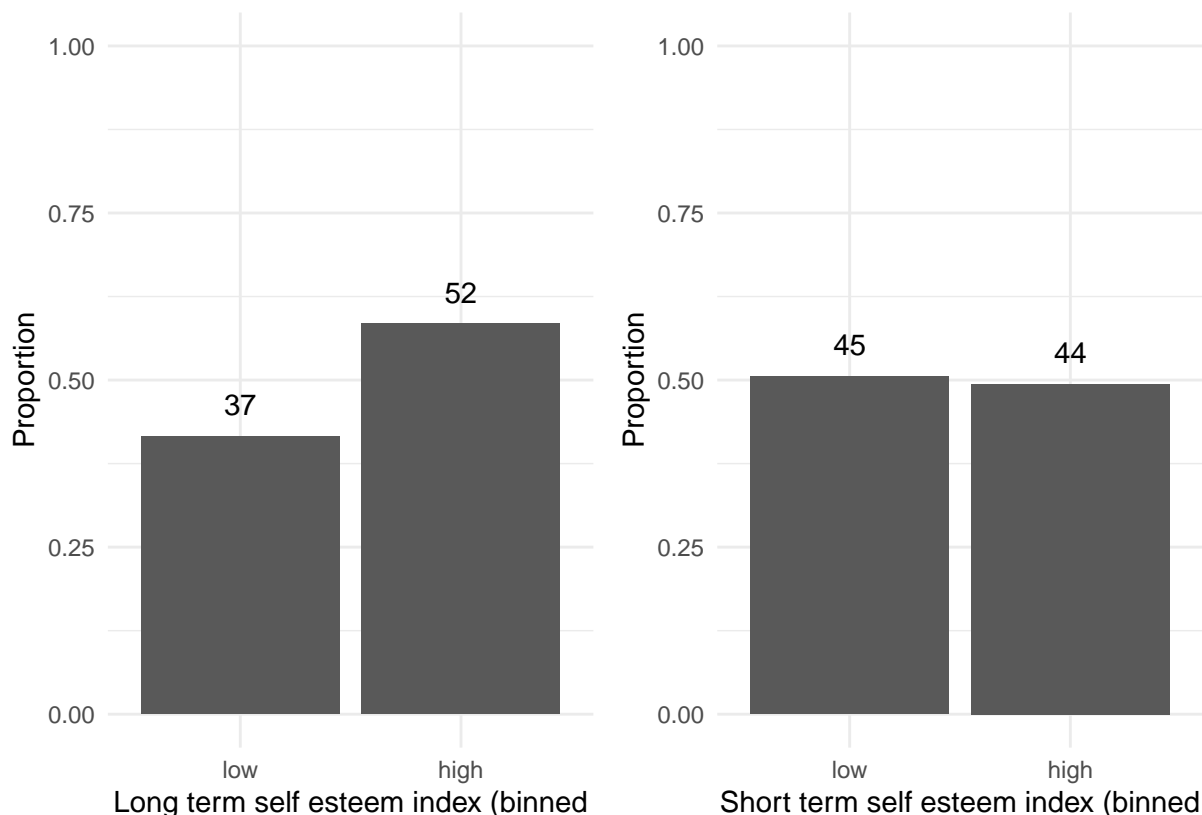
```
table(drinks_subset$numall)
```

```
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 21
##  7 14 18  5 10 16  3  3  2  3  3  2  1  1  1
```

And the number of people drinking more than six or more drinks a day is drastically lower (3 and lower) than the number drinking five or less drinks a day.

2. Since the short and long term self esteem indices are discrete, we can bin them into low and high using a middle point. We used the mean value of the variables to determine the cut-off point. (We compared mean and median values for both variables and found them to be very close for each variable)

```
drinks_subset$roasn_binned <- ifelse(drinks_subset$roasn < mean(drinks_subset$roasn), "low", "high")
drinks_subset$state_binned <- ifelse(drinks_subset$state < mean(drinks_subset$state), "low", "high")
```



A higher proportion of people in the study have high long-term self esteem and people with lower long-term self esteem. The short-term self esteem is approximately equally distributed in the dataset.

3. Most study participants are between the ages of 25 and 41.
4. There are more females in the study than males.
5. Of the 89 people in the sample, 66 have had 0 negative relationship interactions (`nrel`). This skewed distribution will have implications for the models constructed with this variable.

6. A lot more people have had positive romantic interactions, with only 15 with no positive romantic interactions.
7. The two variables `negother` and `posother` are created by subtracting the `nrel` and `prel` from `negevent` and `posevent` respectively. This is done to isolate the effects of romantic and all other interactions. A higher count of people have had minimal negative other interactions.

Why was the study limited to observations from only one day?

By limiting observations to one day, we reduce the scope of the problem from one involving the time component and can therefore assume constant conditions across the subjects. If we were to include data from all days, we would need to model the impact of time series component.

1.2 Hypothesis One

(2 points) The researchers hypothesize that negative interactions with romantic partners would be associated with alcohol consumption and an increased desire to drink. Using appropriate models, evaluate the evidence that negative relationship interactions are associated with higher alcohol consumption and an increased desire to drink.

Answer: As shown above, only 23 people in the sample had non-zero negative romantic relationships on the Saturday of the study. Given the large skew in the data, it will be difficult to detect a statistically significant relationship between negative interactions, `nrel`, and alcohol consumption, `numall`. Below we use multiple models to try and detect the relationship:

```
simple_model <- glm(formula = numall ~ nrel,
                   family = poisson(link = "log"), data = drinks_subset)

summary(simple_model)
```

```
##
## Call:
## glm(formula = numall ~ nrel, family = poisson(link = "log"),
##      data = drinks_subset)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.8330  -1.3203  -0.5295   0.4744   5.9612
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.38951    0.05719  24.298  <2e-16 ***
## nrel         0.05087    0.05083   1.001    0.317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 249.39  on 87  degrees of freedom
## AIC: 508.79
```

```
##
```

```
## Number of Fisher Scoring iterations: 5
```

The simplest model, regressing the number of negative romantic interactions on the number of overall drinks consumed shows a positive relationship between the variables. However, this estimate is not statistically significant. It is important to note that the `nrel` variable does not have much variability in its values since a majority of the respondents answered 0 negative romantic interactions on the day of the study.

We can add other variables to the model to evaluate whether the relationship between these two variables changes. We are including interaction terms below because the impact of a negative romantic interaction on the number of total drinks consumed may be affected by one or more of these factors.

```
model2 <- glm(formula = numall ~ nrel + negother + posother + prel +
              state_binned + rosn_binned +
              nrel:state_binned + nrel:rosn_binned,
              family = poisson(link = "log"), data = drinks_subset)

summary(model2)
```

```
##
```

```
## Call:
```

```
## glm(formula = numall ~ nrel + negother + posother + prel + state_binned +
##      rosn_binned + nrel:state_binned + nrel:rosn_binned, family = poisson(link = "log"),
##      data = drinks_subset)
```

```
##
```

```
## Deviance Residuals:
```

```
##      Min        1Q      Median        3Q        Max
## -3.4578  -1.3647  -0.3666    0.8181    5.1407
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.681367   0.144348  11.648 < 2e-16 ***
## nrel             -0.080902   0.107581  -0.752  0.452045
## negother         -0.055894   0.017118  -3.265  0.001094 **
## posother          0.008572   0.009237   0.928  0.353428
## prel             0.014080   0.022207   0.634  0.526055
## state_binnedhigh -0.565932   0.121513  -4.657  3.2e-06 ***
## rosn_binnedhigh   0.079482   0.117776   0.675  0.499769
## nrel:state_binnedhigh 0.409081  0.122613   3.336  0.000849 ***
## nrel:rosn_binnedhigh -0.105759  0.120389  -0.878  0.379687
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
```

```
##      Null deviance: 250.34  on 88  degrees of freedom
```

```
## Residual deviance: 214.75  on 80  degrees of freedom
```

```
## AIC: 488.15
##
## Number of Fisher Scoring iterations: 5
```

In the more involved model, `nrel` is negatively correlated with the number of total drinks consumed and the estimate remains statistically insignificant. The changing sign of the estimate is of concern and points to its unreliability in explaining the number of total drinks consumed.

Other variables are show statistical significance in the model: `negother` (other negative interactions), `state_binnedhigh` (the short term self-esteem index), and `nrel:statebinnedhigh` (the interaction of negative romantic relationships and high long-term self-esteem).

```
Anova(model2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##
##          LR Chisq Df Pr(>Chisq)
## nrel          2.6206  1  0.1054875
## negother      11.5175  1  0.0006894 ***
## posother       0.8496  1  0.3566656
## prel          0.4005  1  0.5268174
## state_binned  13.2051  1  0.0002792 ***
## rosn_binned   0.1436  1  0.7047543
## nrel:state_binned 11.9792  1  0.0005380 ***
## nrel:rosn_binned  0.7681  1  0.3808172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio tests show the same variables to be statistically significant as those shown by the `summary()` function. Interesting points to note are:

1. Negative Romantic Interactions are not correlated with the total number of drinks
2. Negative Other Interactions are negatively correlated with the total number of drinks, counter to the researchers' intuition.
3. The short term self-esteem index, individually as well as interacted with the number of negative romantic interactions, is correlated with the total number of drinks and is statistically significant.

1.3 Hypothesis Two

(1 point) The researchers hypothesize that the relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem. Conduct an analysis to address this hypothesis.

The model shown above includes the `rosn_binned` term to account for long term high self-esteem. We see that the interaction between `nrel:rosn_binned` is statistically insignificant via both the `summary()` function and the LR tests' `Anova()` function. The researchers' hypothesis that, "the relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem" can mean one of two things:

1. Individuals with high self-esteem do not show high volumes of drinking, or

2. The relationship between `nrel:rosn_binned` and `numall` is statistically insignificant

The output of the `Anova()` function validates point #2 above. To test point #1, we would need to collect more data to reduce the amount of skew towards 0 within `nrel` in the dataset.