W271 Spring 18: Lab 2

Alyssa Eisenberg, Jeffrey Hsu, Gerard Kelly

Basic setup (confirm all of this is done in EDA section)

```
library(car); require(dplyr); library(Hmisc); library(mcprofile); library(ggplot2); library(gridExtra);
dehart <- read.table(file="DeHartSimplified.csv", header=TRUE, sep=",")
dehart$dayweek_f <- factor(dehart$dayweek); levels(dehart$dayweek_f) = c("mon","tue","wed","thu","fri",
dehart$gender_f <- factor(dehart$gender); levels(dehart$gender_f) = c("male","female")

dehart$nrel_f <- factor(as.numeric(dehart$nrel != 0))
dehart$desired_binS <- cut(dehart$desired, breaks=c(-1, 2, 3, 4, 5, 6, 7, 8))
dehart$desired_binL <- cut(dehart$desired, breaks=c(-1, 3.5, 6, 8))
```

Modeling

Initial Model Specification

First, we decided to run two types of model based on the output variable of interest. The hypothesis under investigation expects a positive association between both desire to drink and actual drinks consumed with negative romantic interaction for those with low trait self-esteem. Thus, we will want to analyze number of drinks as our primary outcome and desire to drink as 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. As part of our EDA, 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, trait self-esteem, and their interaction. We use a dummy for having negative interactions or not because of the high skew (77% of the data have a value of 0), the presence of outliers potentially with high influence (e.g., one points with a value of 8), and the pattern in relationship we saw between this, trait self-esteem, and our output variables in the EDA.

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.

```
# Poisson models for number of drinks:
pois_base <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn,
                 data=na.omit(dehart), family=poisson(link=log))
pois_int <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn + prel + prel:rosn
                + dayweek_f, data=na.omit(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=na.omit(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.

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

```
coeftest(pois_full, vcov = vcovCL(pois_full, cluster=na.omit(dehart)$id))[c("age", "gender_ffemale", "n
```

```
##                   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.

```
# Create final model
pois_final <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn + dayweek_f, data=na.omit(dehart), fami

# 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(na.omit(dehart), dayweek_f=="fri"), family=poisson(link=log))

pois_final_sat <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn,
                      data=subset(na.omit(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=na.omit(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=na.omit(dehart)$id)))
se.pois_int <- sqrt(diag(vcovCL(pois_int, cluster=na.omit(dehart)$id)))
se.pois_final <- sqrt(diag(vcovCL(pois_final, cluster=na.omit(dehart)$id)))
se.pois_full <- sqrt(diag(vcovCL(pois_full, cluster=na.omit(dehart)$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=na.omit(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="dayweek_f",
          #add.lines = list(c("DOW effects?", "No", "Yes", "Yes", "Yes", "No", "No", "Yes")),
          omit.stat = c("aic", "ll"), star.cutoffs = c(.05,.01,.001), header=F, no.space=TRUE)
```

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

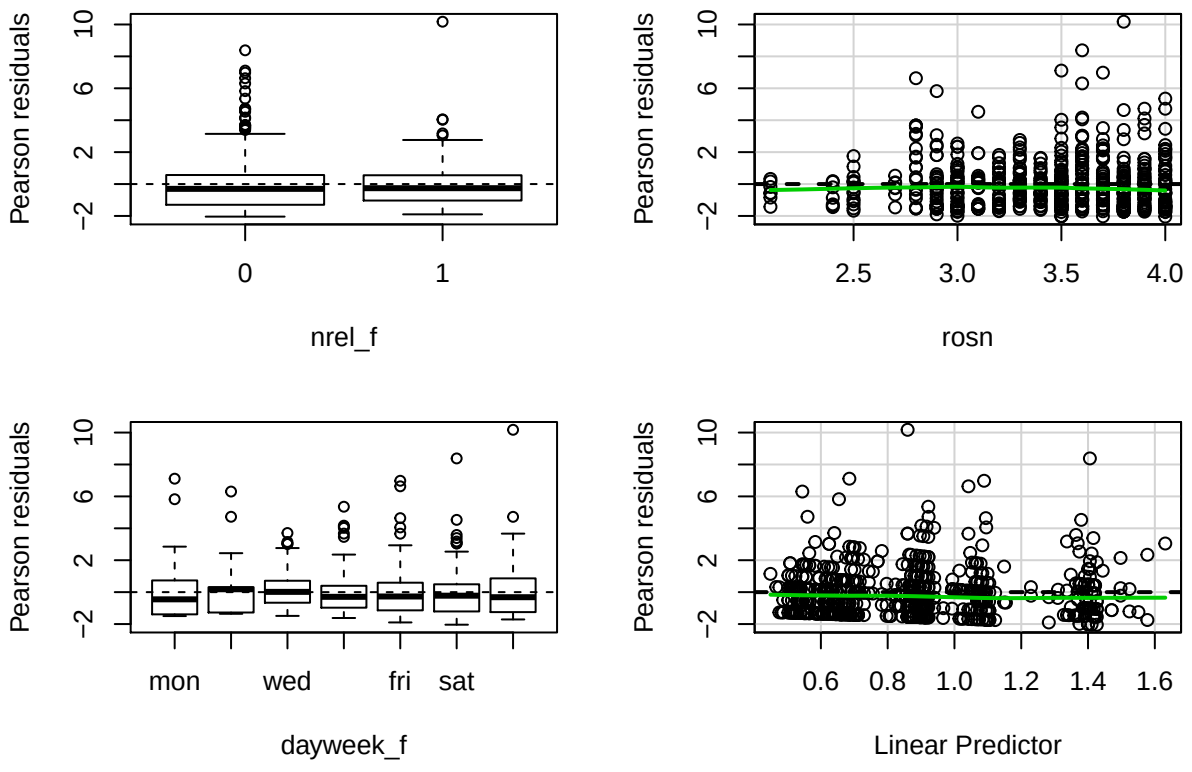
Residual Diagnostics

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.

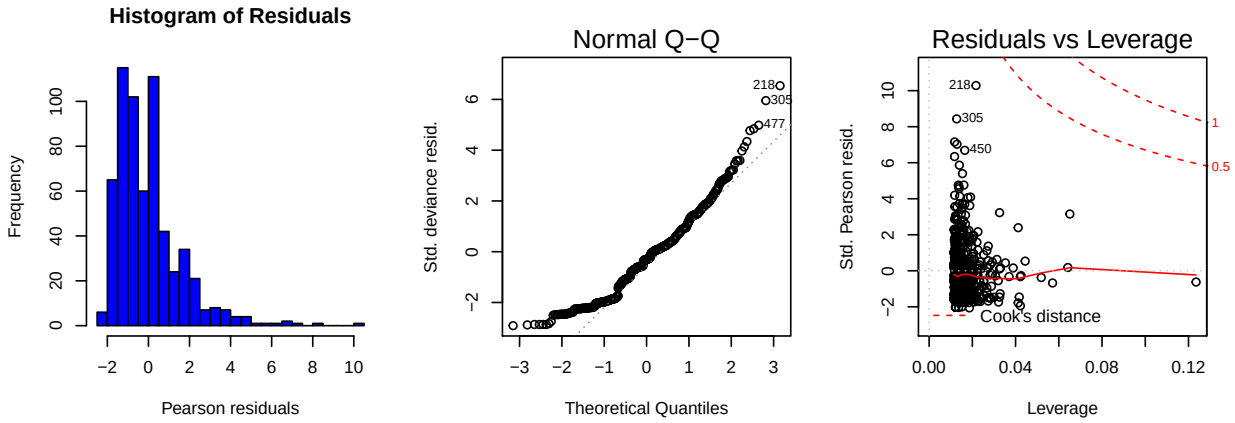
In the plots of Pearson residuals vs. each explanatory variable and the linear predictor, we see that the residual means are pretty constant across the explanatory variable range (for the individual day models, there was a bit more curvature to the means). However, the residual variances are clearly 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 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.

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




```
# 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.

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

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

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

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