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. In our analysis, we will use clustered robust standard errors and we 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 on each variable individually. We found that each of the variables and the interaction term added were statistically significant in explaining additional variance in the residuals, and thus will keep them in our final model.

NEED TO DISCUSS: Using cluster robust SE changes significance of prel and interaction, and we might want to exclude them from final model? LRT could be biased b/c it assumes independence of observations. Should we instead build our final model based on an individual day?

```
library("lmtest")
```

```
## Warning: package 'lmtest' was built under R version 3.4.3
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
Anova(pois_int, test="LR")[c("prel","dayweek_f","rosn:prel"),]
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##           LR Chisq Df Pr(>Chisq)
## prel        13.777  1  0.0002059 ***
## dayweek_f  117.624  6  < 2.2e-16 ***
## rosn:prel    9.145  1  0.0024942 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(pois_int)[c("prel","rosn:prel"),]
```

```
##              Estimate Std. Error   z value     Pr(>|z|)
```

```
## prel       0.26660540 0.07519931  3.545317 0.0003921407
## rosn:prel -0.06764523 0.02224098 -3.041468 0.0023542748
```

```
coeftest(pois_int, vcov = vcovHC(pois_int))[c("prel","rosn:prel"),]
```

```
##              Estimate Std. Error   z value   Pr(>|z|)
## prel       0.26660540  0.1285394  2.074115 0.03806864
## rosn:prel -0.06764523  0.0377346 -1.792658 0.07302762
```

```
coeftest(pois_int, vcov = vcovCL(pois_int, cluster=na.omit(dehart)$id))[c("prel","rosn:prel"),]
```

```
##              Estimate Std. Error   z value  Pr(>|z|)
## prel       0.26660540 0.18569989  1.435679 0.1510937
## rosn:prel -0.06764523 0.05137352 -1.316733 0.1879280
```

```
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 full model to see if there were variables that significantly explained additional variation to include in our final model. We found that gender and negative events were statistically significant.

SAME DISCUSSION AS ABOVE: do we include gender and negevent, or not?

```
Anova(pois_full, test="LR")[c("age", "gender_f", "negevent", "posevent", "state"),]
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##          LR Chisq Df Pr(>Chisq)
## age        0.0906  1   0.763475
## gender_f   5.4803  1   0.019232 *
## negevent   7.9020  1   0.004938 **
## posevent   2.2483  1   0.133762
## state      3.2773  1   0.070245 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(pois_full)
```

```
##
## z test of coefficients:
##
##                  Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    -0.0794308  0.4670977 -0.1701 0.8649694
## nrel_f1         1.3207948  0.5489394  2.4061 0.0161245 *
## rosn            0.3289182  0.1085025  3.0314 0.0024340 **
## prel            0.2593329  0.0784006  3.3078 0.0009403 ***
## dayweek_ftue   -0.1254583  0.1107876 -1.1324 0.2574569
## dayweek_fwed   -0.0422304  0.1082113 -0.3903 0.6963451
## dayweek_fthu    0.2198940  0.1014663  2.1672 0.0302225 *
```

```
## dayweek_ffri    0.3862058  0.0976400  3.9554 7.641e-05 ***
## dayweek_fsat    0.6873392  0.0923851  7.4399 1.007e-13 ***
## dayweek_fsun    0.1864523  0.1023063  1.8225 0.0683806 .
## age             0.0017484  0.0058117  0.3008 0.7635419
## gender_ffemale -0.1244068  0.0530979 -2.3430 0.0191310 *
## negevent       -0.2133153  0.0769447 -2.7723 0.0055658 **
## posevent        0.0698705  0.0462575  1.5105 0.1309240
## state          -0.1123743  0.0617790 -1.8190 0.0689158 .
## nrel_f1:rosn   -0.3346419  0.1591678 -2.1024 0.0355141 *
## rosn:prel      -0.0672540  0.0229428 -2.9314 0.0033746 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(pois_full, vcov = vcovCL(pois_full, cluster=na.omit(dehart)$id))
```

```
##
## z test of coefficients:
##
##                  Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    -0.0794308  0.8624028 -0.0921  0.926615
## nrel_f1         1.3207948  0.6402438  2.0630  0.039117 *
## rosn            0.3289182  0.2009890  1.6365  0.101735
## prel            0.2593329  0.1998124  1.2979  0.194328
## dayweek_ftue   -0.1254583  0.1561470 -0.8035  0.421707
## dayweek_fwed   -0.0422304  0.1302998 -0.3241  0.745861
## dayweek_fthu    0.2198940  0.1161562  1.8931  0.058346 .
## dayweek_ffri    0.3862058  0.1346266  2.8687  0.004121 **
## dayweek_fsat    0.6873392  0.1404493  4.8939 9.888e-07 ***
## dayweek_fsun    0.1864523  0.1622521  1.1492  0.250493
## age             0.0017484  0.0120535  0.1450  0.884672
## gender_ffemale -0.1244068  0.1144995 -1.0865  0.277246
## negevent       -0.2133153  0.1354792 -1.5745  0.115366
## posevent        0.0698705  0.0839534  0.8323  0.405266
## state          -0.1123743  0.1109500 -1.0128  0.311138
## nrel_f1:rosn   -0.3346419  0.1894539 -1.7663  0.077337 .
## rosn:prel      -0.0672540  0.0550824 -1.2210  0.222097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

ONCE FINAL MODEL DETERMINED, come back to AIC analysis commentary

```
# Create final model
pois_final <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn + prel + prel:rosn
                  + dayweek_f + gender_f + negevent, data=na.omit(dehart), family=poisson(link=log))

# 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 2802.964 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 here since there are more social

interactions on Fridays 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 and 6.

NOTE: will need to check final desire to drink model based on decisions about LRT vs cluster-robust SE tests as well

```
# Restrict final model to observations from Fri
pois_final_fri <- glm(formula = numall ~ nrel_f + rosn + nrel_f:rosn + prel
                      + prel:rosn + gender_f + negevent,
                      data=subset(na.omit(dehart), dayweek_f=="fri"), 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 + prel + gender_f, data=na.omit(dehart), method="logistic")
```

With all our models created, here is the output table of all coefficient values and their clustered robust standard errors.

NOTE: maybe try to hide the DOW coeff and include a row for DOW fixed effects Yes/no to save space?

```
# 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.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, 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),
          star.cutoffs = c(.05,.01,.001), header=F
          #,type="text"
          )
```

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. However, the residual variances are clearly different for trait self-esteem, day of week, and negative events. 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. 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.

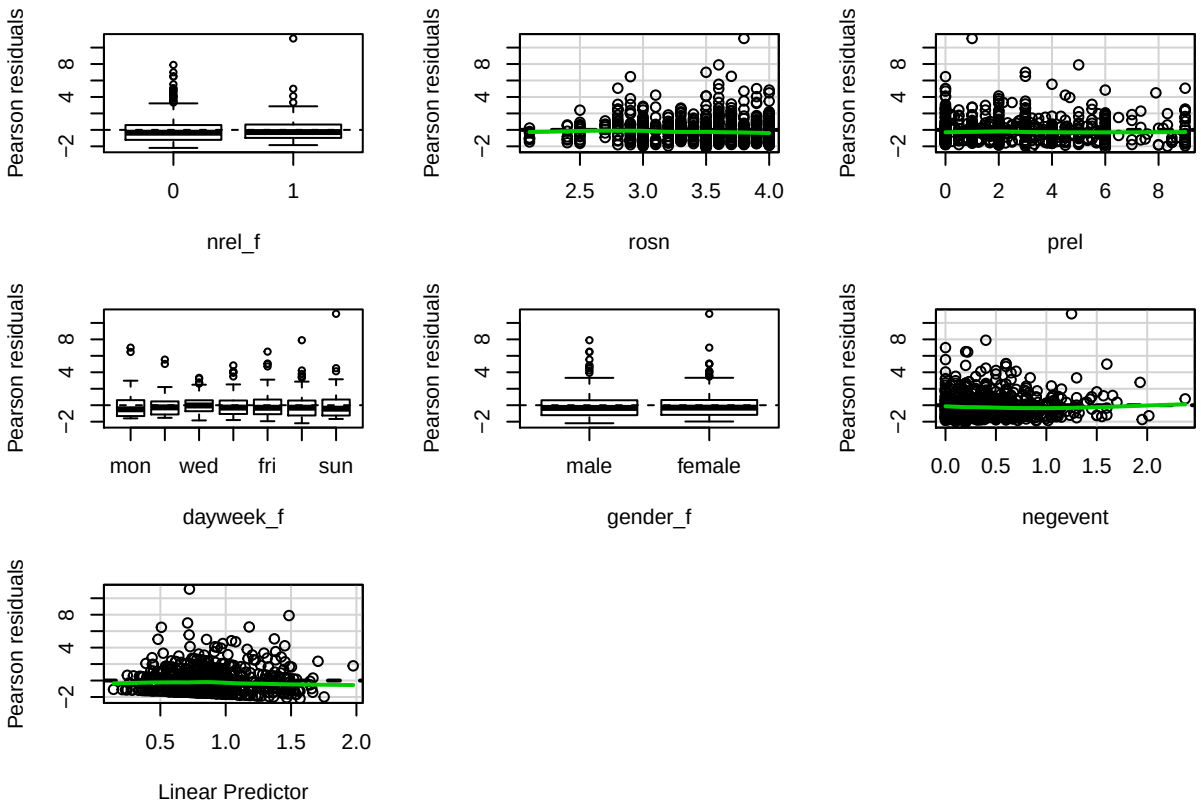
Table 1:

	Dependent variable:					
	numall					desired_binL
	Poisson					*ordered logistic*
	(1)	(2)	(3)	(4)	(5)	(6)
nrel_f1	1.061	1.509*	1.313*	1.321*	0.345	2.617
	(0.771)	(0.654)	(0.635)	(0.640)	(2.415)	(2.082)
rosn	0.050	0.285	0.288	0.329	0.064	−0.588*
	(0.154)	(0.185)	(0.184)	(0.201)	(0.465)	(0.297)
prel		0.267	0.255	0.259	0.292	0.072
		(0.186)	(0.200)	(0.200)	(0.646)	(0.046)
dayweek_ftue		−0.139	−0.134	−0.125		0.511*
		(0.157)	(0.155)	(0.156)		(0.259)
dayweek_fwed		−0.071	−0.057	−0.042		0.574*
		(0.129)	(0.130)	(0.130)		(0.256)
dayweek_fthu		0.204	0.211	0.220		0.578
		(0.117)	(0.117)	(0.116)		(0.298)
dayweek_ffri		0.379**	0.379**	0.386**		0.895**
		(0.132)	(0.133)	(0.135)		(0.311)
dayweek_fsat		0.679***	0.676***	0.687***		0.917***
		(0.142)	(0.140)	(0.140)		(0.254)
dayweek_fsun		0.191	0.179	0.186		−0.107
		(0.160)	(0.162)	(0.162)		(0.273)
age				0.002		
				(0.012)		
gender_ffemale			−0.110	−0.124	−0.416	−0.352
			(0.117)	(0.114)	(0.258)	(0.230)
negevent			−0.182	−0.213	0.558	
			(0.131)	(0.135)	(0.322)	
posevent				0.070		
				(0.084)		
state				−0.112		
				(0.111)		
nrel_f1:rosn	−0.292	−0.407*	−0.329	−0.335	−0.149	−0.650
	(0.226)	(0.194)	(0.188)	(0.189)	(0.697)	(0.600)
rosn:prel		−0.068	−0.063	−0.067	−0.092	
		(0.051)	(0.055)	(0.055)	(0.187)	
Constant	0.742	−0.389	−0.286	−0.079	0.910	
	(0.544)	(0.667)	(0.668)	(0.862)	(1.530)	

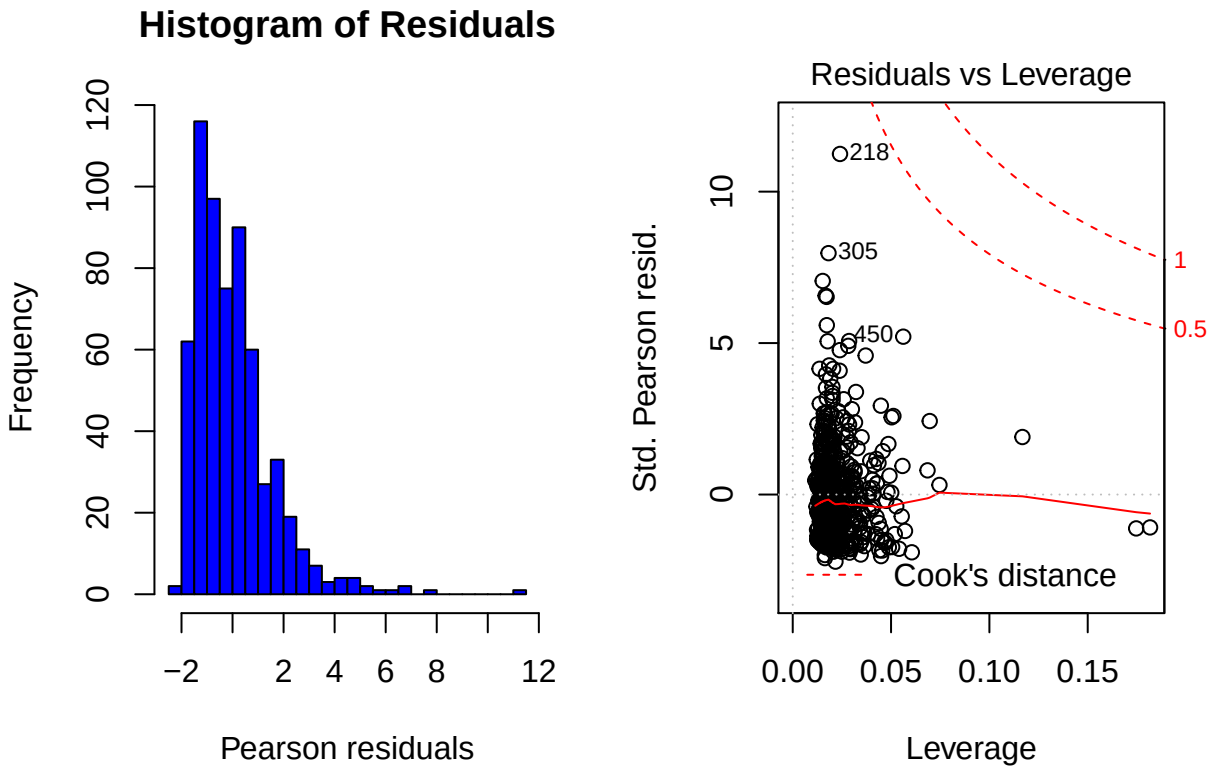
Finally, we checked the residuals vs. leverage plot, but did not see any points with high influence (Cook's distance>1) that we would need to be concerned about and check robustness without those outliers.

NOTE: still need to check how to do diagnostic plots for ordinal regression model NOTE: check that individual day Poisson models show the same thing Do we want to talk about OVB for the interaction term? How would we think about this?

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




```
# Additional plots - histogram of residuals and leverage/influence plot
par(mfrow=c(1,2))
hist(residuals(pois_final, "pearson"), breaks = 20, col = "blue",
     xlab="Pearson residuals", main="Histogram of Residuals")
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 0.89 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 1.02 times for every 0.4 decrease in trait self-esteem. Across all our models in Table 1, 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.

NOTE: double check significance with final models Should we explain coefficient in ordinal regression model as well? Do we want a visual - fitted vs actual values? trait self esteem vs. num of drinks with scatter of actual data and lines for model output with and without nrel?

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

```
## [1] 0.89
```

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

```
## [1] 1.02
```