STA138 Project II

Chad Pickering (913328497), Graham Smith (912355584)

March 8, 2017

Part One: Logistic Regression

1. Introduction:

Nowadays, companies are taking their employee's mental health more seriously; it is desired that we sample a company's employees and "score" their "mental stability" to understand more about the company's workplace environment, but also to potentially make inference to employees at other companies. The evaluation of emotional stability was done with a scoring system where higher scores correspond to more emotional stability. Given an employee's reported stability score, the ability to perform some pre-assigned task was recorded. The columns of the data are as follows:

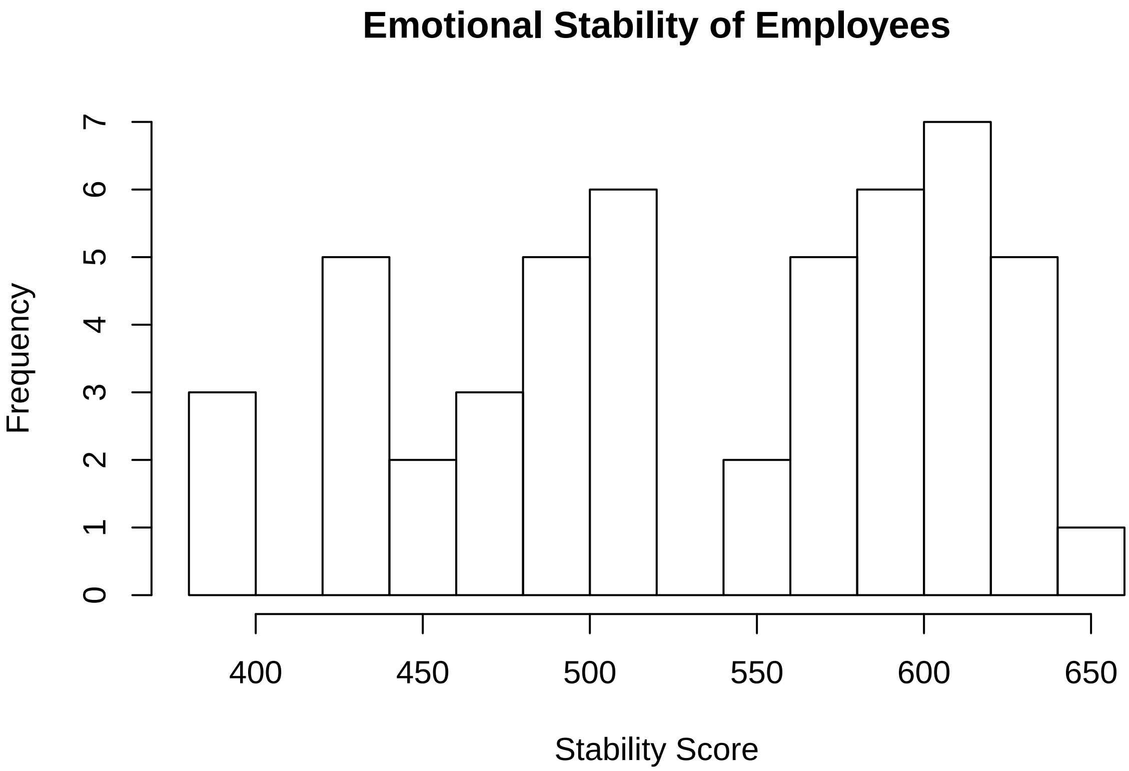
Column 1: Y: 1 if the employee could successfully complete the task, and 0 otherwise.

Column 2: X: The score of the emotional stability test. Range is [300, 700].

We now perform logistic regression in order to predict whether that pre-assigned task can be performed given some mental stability score. This prediction will allow us to know whether we can expect certain tasks to be performed in a given workplace environment; thus, a company can then diagnose possible workplace issues that contribute to mental instability and therefore unproductivity or inability to perform.

2. Summary:

In the sample of 50 employees, we immediately see that the distribution of emotional stability scores is in two semi-distinct groups, a low and a high group (this is more obvious as the number of breaks in the histogram increases). But in general, the distribution is slightly more skewed to the higher scores.



We can also split the data by whether the task could be performed or not and whether the score reported was greater or less than the mean score of the sample. We see that 15 out of the 50 people in the sample could perform the task, and 13 of those had stability scores above the mean score, 535.62. The employees who could not perform the task had much lower scores, where 22 of those 35 people were below the mean. Specifically, the odds of having a lower than average score for those who cannot perform the task is approximately 11 times the odds of having a lower than average score for those who can perform the task. We expect this relationship to hold, and that as emotional stability increases, the more likely it is for the task to be successfully completed.

```
##
##     FALSE TRUE
##   0    13   22
##   1    13    2
```

```
##          StabScore
## TaskComp  >= mean < mean
##   Failure      13     22
##   Success      13      2
```

3. Analysis/Interpretation:

The estimated logistic regression function is the following:

$$logit(\pi(x)) = -16.8255 + 0.0284x$$

.

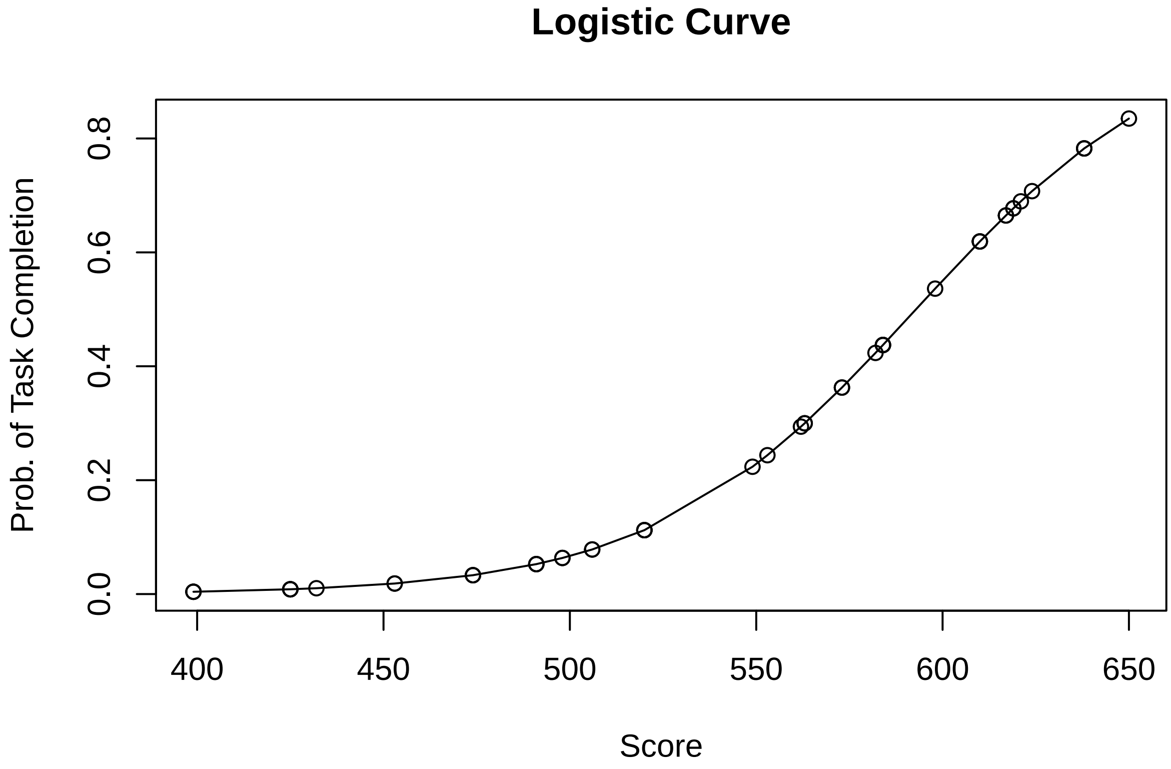
```
##  (Intercept)            X
## -16.82550095   0.02837954
```

In general, as the stability score increases, our probability of task completion also increases ($\hat{\beta} > 0$), and a practical interpretation of α here makes no sense as a score of 0 is never attainable.

We should know if stability score significantly affects the ability to successfully complete a task. We conclude from the model that as $exp(\hat{\beta}) = 1.0288$, for every one unit increase in stability score, the odds of successful task completion are multiplied by approximately 1.0288.

```
## [1] 1.028786
```

The logistic curve shows at least mild visible effect of stability score on successful task completion as the curve is somewhat steep after a score of about 500 - this marked increase points toward significance of score.



Let's make a numerical conclusion about the effect of score on task completion more formally with a confidence interval. We are 99% confident that the true odds for successful task completion is between 1.0103 and 1.0591 for a one-unit increase in score using a LR confidence interval.

```
##    0.5 %   99.5 %
## 1.010261 1.059147
```

In terms of a tangible p-value, we know given that $\beta = 0$ (score is not an effect), there is a 4.8924×10^{-6} probability of obtaining our data/test statistic (which is reported below) or a more extreme one. We can then reject a null hypothesis that $exp(\beta) = 1$ against any standard α and conclude that stability score has an effect on probability of successful task completion.

```
## $LR_teststat
## [1] 20.87896
##
## $pvalue
## [1] 4.8924e-06
```

Lastly, now that we have a logistic regression model and we know that score has a significant effect on task completion, we can predict whether a task will be completed based on a newly reported score. For example, if an employee scores a 615, they have about a 65.20% chance of being able to complete the task.

```
##         1
## 0.6520173
```

More succinctly, we know that at stability score of about 592.87, the probability of successful task completion is 50%. This score can be the threshold of determining whether a task is more likely to be completed than not given an employee's score, based on this sample.

```
## (Intercept)
##    592.8743
```

Other thresholds can be computed, such as an 80% threshold, if one would like to be more sure that an employee be able to complete the task. With our sample, this threshold is a stability score of about 641.72.

```
## (Intercept)
##    641.7226
```

4. Conclusion:

From our analysis, we can conclude that an employee's emotional stability score has a significant effect on the probability of them being able to successfully complete a pre-assigned task. We know that, through analyzing our sample of 50, that any score greater than around 593 has a greater probability of successful task completion than failure. In general, this study is a fair first step to determining how the workplace environment can be more suited to support employee's mental health at a company by understanding the central tendency and variance of the scores as well as how these scores are associated with productivity, etc. Perhaps in the future, more variables such as daily sleep and average daily exercise duration can be measured in addition to the mental stability score to assess how any or all variables (and combinations thereof) affect task performance. A larger sample would also be helpful in making more general inference to the population of all employees at a company.

Part Two: Log-Linear Models

1. Introduction

In medicine, it is often of interest to understand how different administered drugs affect improvement of a condition, and how other lifestyle choices come into play. Here, we have a random sample of 519, where two types of drugs, one and two, were administered, and the condition status after six months and exercise frequency was recorded. Counts are computed, and the columns of data are as follows:

Column 1: X: Exercise frequency, recorded as Often or Rarely.

Column 2: Y: Condition status, recorded as Imp for improvement and Not otherwise.

Column 3: Z: Drug administered, recorded as One or Two.

Column 4: Freq: The computed count from the sample.

In analyzing this data, we will be able to select a model to better understand significant interactions between variables and derive expected counts for future use. In this way, we can move to make conclusions on how exercise frequency may affect condition status, and potentially determine which of the two drugs has a better chance of improving the condition in general.

2. Summary

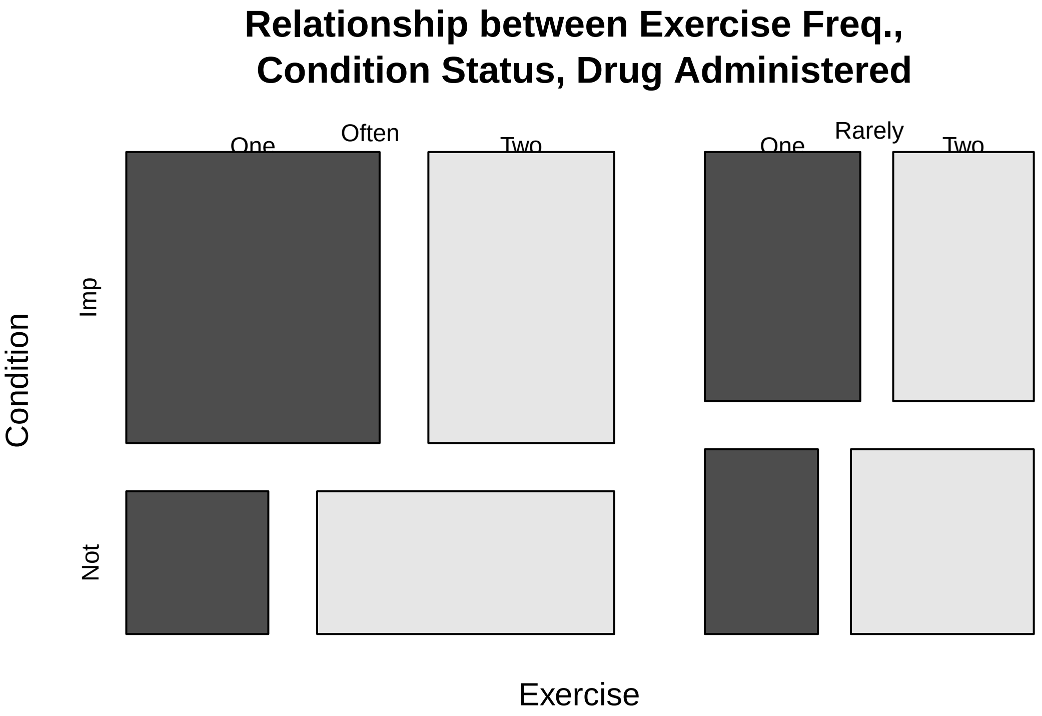
The following is the counts of all of the frequencies, stratified by the three binary variables. We split the data by drug administered so that differences between the two groups would be more apparent.

```
## $DrugOne
##
##          Imp Not
##   Often  120  33
##   Rarely  63  34
##
## $DrugTwo
##
##          Imp Not
##   Often   88  69
##   Rarely  57  55
```

The odds of improvement within six months for those who exercise regularly is 1.9625 times the odds of improvement within six months for those who exercise rarely, conditioned on drug one. On the other hand, conditioned on drug two, the odds of improvement within six months for those who exercise regularly is 1.2306 times the odds of improvement within six months for those who exercise rarely. This difference in odds suggests that improvement is dependent on drug administered to some degree.

```
## [1] 1.962482 1.230613
```

Finally, as for a visualization of the data, a mosaic plot shows some interesting initial insights. It is more likely that improvement of the condition in six months happens when a person exercises often. In addition, it appears it is more likely that a subject had taken drug 1 when condition had improved than taken drug 1 when the condition had not improved. Both of these aforementioned associations seem visually striking enough to be numerically significant, and we will test these using standard analytic procedures now.



3. Analysis/Interpretation

In order to determine which variables are dependent on one another (if any) we can use model selection with AIC/BIC as criteria to see which interaction terms are significant. It should be noted that this method tells us nothing about how good the models are in absolute terms, rather only how well they fit in relation to one another. However, since we're only trying to check for dependence and not actually predicting or making inferences, simple model selection will suffice.

Using the glm() and Dr. Melcon's good.fit.LL() functions, we are able to find the AIC and BIC for every possible model, as seen in the table below. We are choosing to use AIC rather than the other criteria because it penalizes larger models and avoids choosing a model that is overfit. In fact, the model chosen below minimizes both AIC and BIC (BIC penalizes larger models more than AIC).

```
##               Log-Li      LR Pearson df p-val:LR p-val:Pear     AIC     BIC
## (X,Y,Z)     -37.4907 27.5388 27.1055  4   0.0000     0.0000 82.9814 83.2992
## (X,YZ)      -26.9962  6.5498  6.5710  3   0.0877     0.0869 63.9924 64.3896
## (Y,XZ)      -37.2740 27.1055 26.1329  3   0.0000     0.0000 84.5481 84.9453
## (Z,XY)      -34.9878 22.5329 22.2182  3   0.0001     0.0001 79.9755 80.3727
## (XY,XZ)     -34.7711 22.0996 21.7804  2   0.0000     0.0000 81.5422 82.0188
## (XY,YZ)     -24.4933  1.5439  1.5459  2   0.4621     0.4616 60.9865 61.4632
## (XZ,YZ)     -26.7795  6.1164  6.2010  2   0.0470     0.0450 65.5590 66.0357
## (XY,XZ,YZ) -24.4701  1.4976  1.4995  1   0.2210     0.2207 62.9402 63.4963
## (XYZ)       -23.7213  0.0000  0.0000  0   1.0000     1.0000 63.4426 64.0781
```

Based on the lowest AIC and BIC, we determine that the correct model to choose is clearly (XY, YZ), or

$$ln(\hat{\mu_{ijk}}) = \hat{\lambda} + \hat{\lambda}_i^X + \hat{\lambda}_j^Y + \hat{\lambda}_k^Z + \hat{\lambda}_{ij}^{XY} + \hat{\lambda}_{jk}^{YZ}$$

. This indicates that there are interaction terms between XY and YZ; in other words, these interaction terms imply that there is likely dependence between frequency of exercise and status of condition, as well as status of condition and type of drug administered. This is an example of conditional independence - X (amount of exercise) and Z (drug administered) are independent. We should note that because we lack at least one interaction term in this model, it follows that the interaction term because all three variables common in the saturated model will be insignificant.

To ensure that we have chosen the best model and that no other interaction terms, particularly for (XZ) (amount of exercise and drug administered), are significant, we perform the following hypothesis test.

H_o: The reduced/smaller model is better. $\lambda_{ik}^{XZ} = 0$.

H_a: The larger model is better. $\lambda_{ik}^{XZ} \neq 0$.

Our focus is to see if our 95% Wald CI for $exp(\lambda_{ik}^{\hat{XZ}})$ in the larger model contains 1. As indicated by the exponentiated CIs for $\lambda_{ik}^{\hat{XZ}}$, the bounds are 0.7256706 and 1.4915131. So, because the interval contains 1, we can fail to reject the null hypothesis - we have evidence that the smaller model that we had originally chosen is the superior model, and we can drop the XZ term.

```
## lower.bound upper.bound
##   0.7256706   1.4915131
```

Besides the confidence interval constructed above, we can derive a LR test statistic and p-value in order to see if this conclusion agrees with what was indicated above. First, we derive a p-value from the Likelihood Ratio test statistic, which compares the maximum log-likelihood of the smaller model and the larger model.

Likelihood Ratio test statistic = -2(L_0 - L_1) = 0.0463392 . p-value = 0.8295604 - this is larger than any significance level α reasonably chosen. Thus, we thoroughly fail to reject the null hypothesis that the larger model is superior. As before, we can drop the (XZ) term from the model.

As an additional test, we can look at the confidence intervals for the odds ratio point estimates for the interaction terms that exist in the smaller model to confirm that they are indeed significant (we expect both to be given the model we have chosen).

The odds ratios for the terms we think are significant, XY and YZ, are 1.0526919, 2.1729141 and 1.6158039, 3.3765352 respectively. Neither of these confidence intervals contain 1, which implies that both interaction terms cannot be dropped from the model. However, the confidence interval for the XZ term, 0.7256706, 1.4915131, DOES contain 1 which agrees with our earlier results.

4. Conclusion:

The above analysis indicates that there is a dependence between amount of exercise and condition status (XY), and condition status and drug administered (YZ), but not between amount of exercise and drug administered (XZ) terms. We can confirm that our initial analysis, both graphically with the mosaic plot and mathematically with the model selection were correct in deciding the dependency relationships.

Code Appendix:

```
# Summary:
psych <- read.csv("Psych.csv")
hist(psych$X, breaks = 10,
     main = "Emotional Stability of Employees", xlab = "Stability Score")
table(psych$Y, psych$X<535.62)
cont_tbl <- c(13, 22, 13, 2)
as.table(matrix(cont_tbl, nrow = 2, byrow = TRUE, dimnames = list(TaskComp=c('Failure', 'Success'), Sta
logit.model = glm(formula = Y ~ X, family = binomial(logit), data = psych)
logit.model$coefficients
exp_beta1 <- exp(0.02837954)
exp_beta1
plot(psych$X, logit.model$fitted.values, main = "Logistic Curve", xlab = "Score", ylab = "Prob. of Task
lines(psych$X[order(psych$X)], logit.model$fitted.values[order(psych$X)], xlim=range(psych$X), ylim=ran
suppressMessages(exp(confint(logit.model, level = 0.99)[2,]))
smaller.model = glm(Y ~ 1, family = binomial(logit), data = psych)
lrt <- as.numeric(-2*(logLik(smaller.model) - logLik(logit.model)))
lrt_df <- length(logit.model$coefficients) - length(smaller.model$coefficients)
list_stats <- list(lrt, pchisq(lrt, lrt_df, lower.tail = FALSE))
names(list_stats) <- c("LR_teststat", "pvalue")
list_stats
predict(logit.model, newdata = data.frame(X = 615), type = "response")
-logit.model$coefficients[1]/logit.model$coefficients[2]
(log(0.8/(1-0.8))-logit.model$coefficients[1])/logit.model$coefficients[2]
# Summary:
trials <- read.csv("TrialsShort.csv")
trials_long <- read.csv("TrialsLong.csv")
cont_tbl_two <- table(trials_long$X, trials_long$Y, trials_long$Z)
data_split <- split(trials_long, trials_long$Z)
tbl_1 <- table(data_split$One$X, data_split$One$Y)
tbl_2 <- table(data_split$Two$X, data_split$Two$Y)
list_z <- list(tbl_1, tbl_2)
names(list_z) <- c("DrugOne", "DrugTwo")
list_z
or_1 <- (120*34)/(63*33)
or_2 <- (88*55)/(57*69)
c(or_1, or_2)
mosaicplot(cont_tbl_two, main = "Relationship between Exercise Freq., \n Condition Status, Drug Adminis
good.fit.LL = function(the.model){
  K = length(the.model$coefficients)
  df.model = length(the.model$residuals) - K
  Pearson.TS = round(sum(residuals(the.model,type = "pearson")^2),4)
  LL = as.numeric(logLik(the.model))
  Dev = round(the.model$deviance,4)
  the.AIC = AIC(the.model)
  the.BIC = BIC(the.model)
  pval.Pear = round(pchisq(Pearson.TS,df.model,lower.tail = F),digits =8)
  pval.LR = round(pchisq(Dev,df.model,lower.tail = F),digits =8)
  All.GOF = c(LL,Dev,Pearson.TS,df.model,pval.LR,pval.Pear,the.AIC,the.BIC)
  names(All.GOF) = c("Log-Li","LR","Pearson","df", "p-val:LR","p-val:Pear","AIC", "BIC")
  return(All.GOF)
}
```

```r
all.model.formulas = c("F~X+Y+Z","F~X+Y+Z+Y*Z","F~X+Y+Z+X*Z","F~X+Y+Z+X*Y",
                       "F~X+Y+Z+X*Y+X*Z","F~X+Y+Z+X*Y+Y*Z","F~X+Y+Z+X*Z+Y*Z",
                       "F~X+Y+Z+X*Y+X*Z+Y*Z",
                       "F~X+Y+Z+X*Y+X*Z+Y*Z+X*Y*Z")
all.model.fits = lapply(all.model.formulas,
  function(the.model){
    glm(the.model, data = trials, family = poisson)
})

#goodness of fit testing
all.GOF = sapply(all.model.fits,function(the.model){
  good.fit.LL(the.model)
})
all.GOF = t(all.GOF)
rownames(all.GOF) = all.model.formulas
book.notation = c("(X,Y,Z)","(X,YZ)","(Y,XZ)","(Z,XY)","(XY,XZ)","(XY,YZ)","(XZ,YZ)","(XY,XZ,YZ)","(XYZ
rownames(all.GOF) = book.notation

round(all.GOF, digits = 4)
# Build the model
model = glm(F ~ X + Y + Z + X*Y + Y*Z, data = trials, family = poisson)

alpha = 0.05
za = qnorm(1-alpha/2)
lower.bound = summary(model)$coefficients[,1] -za*summary(model)$coefficients[,2]
upper.bound = summary(model)$coefficients[,1] +za*summary(model)$coefficients[,2]
CIs = cbind(lower.bound,upper.bound)
#round(exp(model2$coefficients), 8)[6] #interaction terms are now odds ratios point estimates
#round(exp(model2$coefficients), 8)[5]

# let's test a larger model to be sure that the lambda for the XZ interaction term is actually 0
bigmodel = glm(F ~ X + Y + Z + X*Y + X*Z + Y*Z, data = trials, family = poisson)
lower.bound = summary(bigmodel)$coefficients[,1] -za*summary(bigmodel)$coefficients[,2]
upper.bound = summary(bigmodel)$coefficients[,1] +za*summary(bigmodel)$coefficients[,2]
CIb = cbind(lower.bound,upper.bound)
XZ.CI = exp(CIb)[6,] # 1 is in this CI which means we fail to reject - no effect
XZ.CI
L0 = logLik(model)
L1 = logLik(bigmodel)
LR.test = as.numeric(-2*(L0 - L1))
LR.pval = pchisq(LR.test, df = 1,lower.tail = F )
# the Likelihood Ratio test statistic yields a p-value that's larger than any reasonable value of alpha
#CI for XY term
XY.CI = exp(CIs)[5,]

#CI for YZ term
YZ.CI = exp(CIs)[6,]
```