STA303/1002 H1S - Assignment 3

Last name: Khambaita
First name: Sunil
Student ID: 1000285924
Course section: STA303H1S-L0101/L0201

Out: Mar 16; Due: Apr 5, 23:00

Initializing data and libraries once, before starting the questions.

```
# Loading ggplot2, glmnet, ROCR, psych & aod library
library(ggplot2)
library(glmnet)
library(ROCR)
library(psych)
library(aod)

# Loading in the Abalone data
faba = read.table("abalone.data", sep = ",")

# Creating a new column y which stores 1 if rings variable is greater than
# 9, 0 otherwise.
faba$y = ifelse(faba$V9 > 9, 1, 0)

# Creating a training data set
xtrain = faba[1:3133, 1:8]
ytrain = as.factor(faba[1:3133, 10])

# Creating a testing data set
xtest = faba[-c(1:3133), 1:8]
ytest = as.factor(faba[-c(1:3133), 10])

# Loading in the data for ships for Q2.
ships = read.table("ships.csv", header = T, sep = ",")

# Loading in the data for Q3.
Cholesterol = as.factor(c("H", "H", "L", "L", "H", "H", "L", "L"))
Gender = as.factor(c("M", "M", "M", "M", "F", "F", "F", "F"))
Heart = as.factor(c("Y", "N", "Y", "N", "Y", "N", "Y", "N"))
Results = c(16, 256, 28, 2897, 13, 319, 23, 2565)
data1 = data.frame(Results, Cholesterol, Gender, Heart)
```

Q1. (20 pts) Binary classification of Abalone data.

(1a) (5 pts) We are going to use the first 3133 samples to train the model, and the rest will be used as the test set. Show your R code to get the training data and testing data. Find the mean and standard error of the continous variables (V2-V8). Standardize all the continous predictors (V2-V8) in the training set using formula $(X - \bar{X})/sd(X)$. Use the mean and sd in the training set to standardize the corresponding predictor in the testing data set.

```
#Mean of (V2-V8)
xtrain_mu = apply(xtrain[,2:8],2,mean)
xtrain_mu
```

```
##        V2        V3        V4        V5        V6        V7        V8
## 0.5226317 0.4068353 0.1393345 0.8248355 0.3584299 0.1795303 0.2376143
```

```
#Standard Deviation of (V2-V8)
xtrain_sd = apply(xtrain[,2:8],2,sd)
xtrain_sd
```

```
##         V2         V3         V4         V5         V6         V7
## 0.12089977 0.10000662 0.04310205 0.49357925 0.22421144 0.11002894
##         V8
## 0.14058221
```

```
#Standard Error of (V2-V8)
xtrain_se = xtrain_sd/sqrt(length(ytrain))
xtrain_se
```

```
##           V2           V3           V4           V5           V6
## 0.0021599579 0.0017866873 0.0007700478 0.0088181340 0.0040056922
##           V7           V8
## 0.0019657431 0.0025115983
```

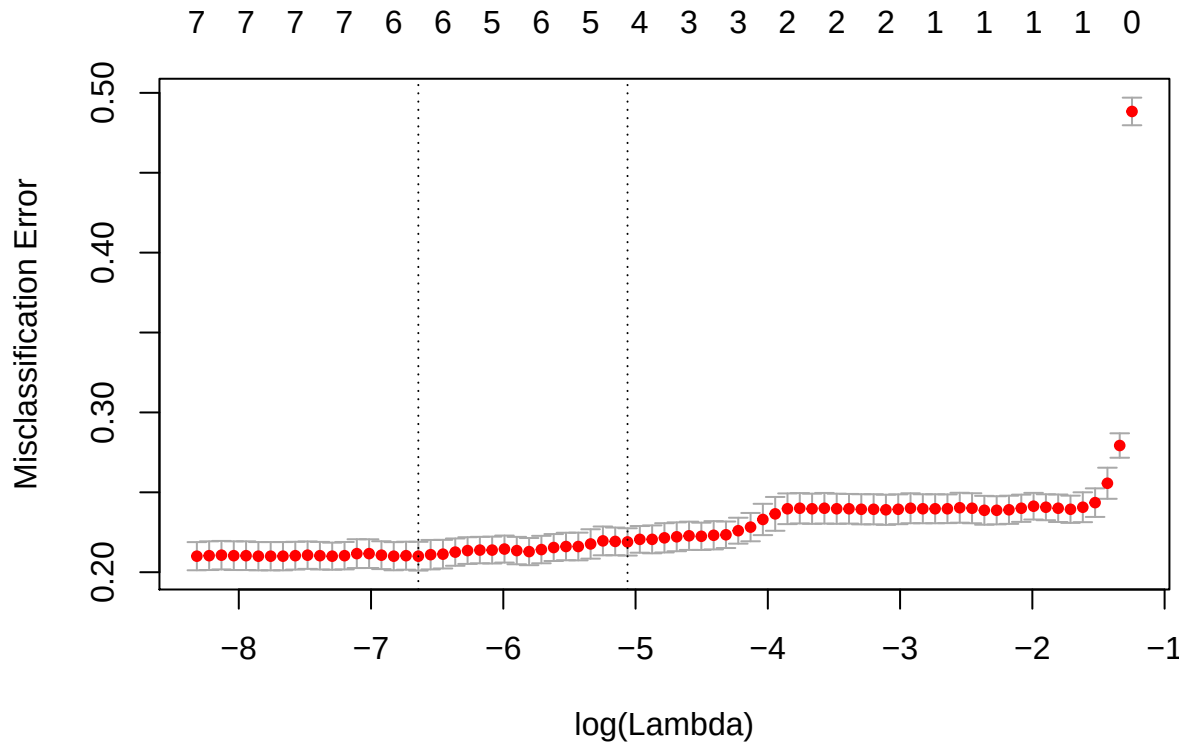
```
#Standardizing (V2-V8) in the training data set
xtrain_std = scale(xtrain[,2:8])
ytrain_std = ytrain

#Standardizing (V2-V8) in the testing data set
xtest_std = scale(xtest[,2:8], center = xtrain_mu, scale = xtrain_sd)
ytest_std = ytest
```

(1b) (5 pts) Fit a LASSO logistic regression (i.e., logistic regression with a LASSO penalty) model using glmnet. Use 10-fold cross-validation to choose the optimal value of the regularizer, show your R code and print the optimal λ obtained from the cross-validation. Predicting with the training and testing data set, print the confusion matrix and report mean error rate (fraction of incorrect labels), respectively.

```
# Training the model on the standardized training set alpha=0 for ridge
# penalty; alpha=1 for the LASSO penalty

# Performing 10-fold cross-validation
cvtrain_std = cv.glmnet(xtrain_std, ytrain_std, type.measure = "class", nfolds = 10,
    family = "binomial", alpha = 1)

# Printing optimal lambda from 10 fold cross validation by lasso penalty
# (alpha = 1)
lambda = cvtrain_std$lambda.min
lambda
```

```
## [1] 0.001303933
```

```
plot(cvtrain_std)
```




```
# Fitting the model alpha=0 for the ridge penalty, alpha=1 for the lasso
# penalty
std_rlogit = glmnet(xtrain_std, ytrain_std, family = "binomial", alpha = 1)

# Predicting with the training data set
srlpred_train = predict(std_rlogit, xtrain_std, type = "class", s = lambda)
```

```r
# Reporting mean error rate training data set (fraction of incorrect labels)
confmatrix_train <- table(ytrain_std, srlpred_train)
confmatrix_train
```

```
##           srlpred_train
## ytrain_std    0    1
##          0 1298  300
##          1  356 1179
```

```r
errorrate_train = (confmatrix_train[1, 2] + confmatrix_train[2, 1])/length(srlpred_train)
errorrate_train
```

```
## [1] 0.209384
```

```r
# Predicting with the testing data set
srlpred_test = predict(std_rlogit, xtest_std, type = "class", s = lambda)

# Reporting mean error rate testing data (fraction of incorrect labels)
confmatrix_test = table(ytest_std, srlpred_test)
confmatrix_test
```

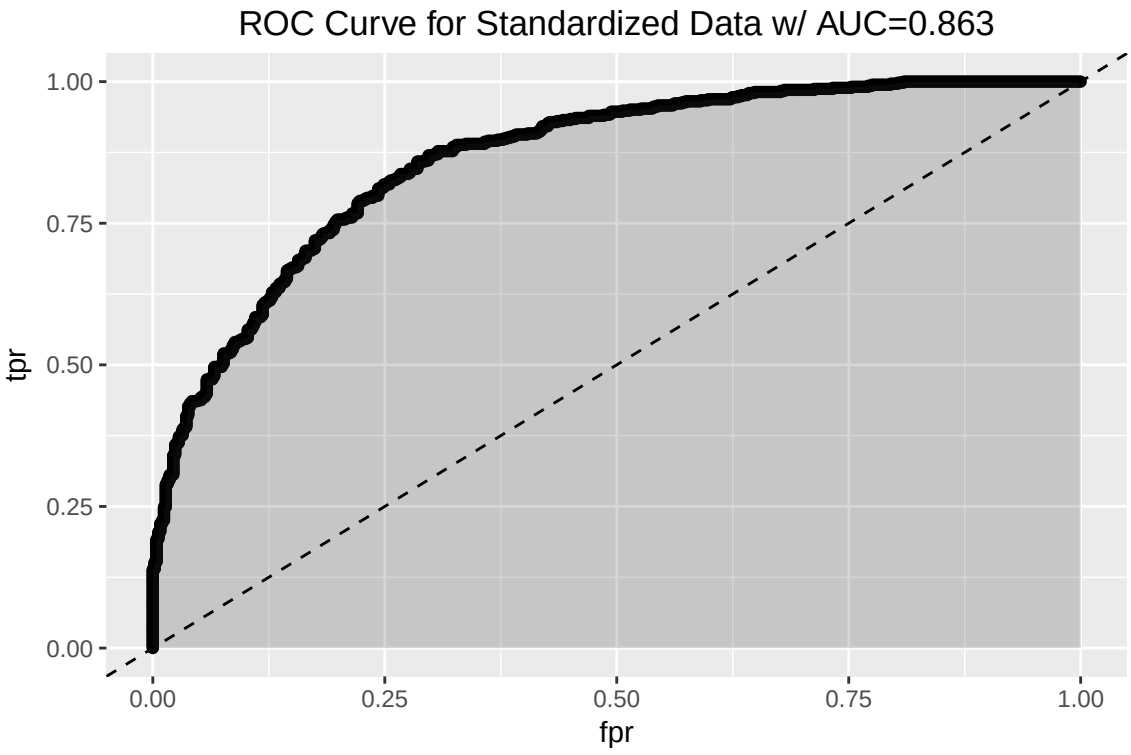
```
##          srlpred_test
## ytest_std   0   1
##         0 391 107
##         1 129 417
```

```r
errorrate_test = (confmatrix_test[1, 2] + confmatrix_test[2, 1])/length(srlpred_test)
errorrate_test
```

```
## [1] 0.2260536
```

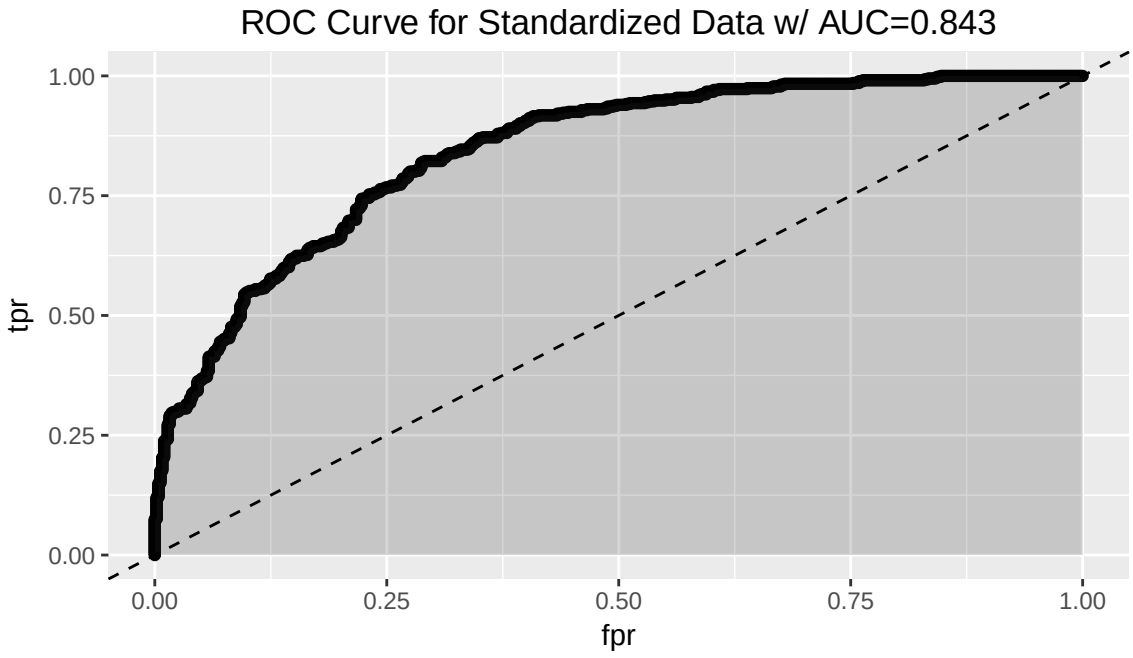
(1c) (5 pts) Plot the receiver operating characteristic (ROC) curve on the test data. Use package ROCR to get the ROC curve and use ggplot2 to plot the ROC curves. Report the area under the ROC curve (AUC).

```
# Using lasso penalty

# ROC curve for standardized data
std_prob = predict(std_rlogit, xtest_std, type="response", s=lambda)
std_pred = prediction(std_prob, ytest_std)
std_perf = performance(std_pred, measure = "tpr", x.measure = "fpr")

# True positive rate
tpr.ptstd = attr(std_perf, "y.values")[[1]]

# False positive rate
fpr.ptstd = attr(std_perf,"x.values")[[1]]

# Area under the curve
auc.std = attr(performance(std_pred, "auc"), "y.values")[[1]]
auc1 = signif(auc.std, digits=3)

roc.std = data.frame(fpr = fpr.ptstd, tpr = tpr.ptstd, model = "GLM")

# Plotting the ROC curve under test data
ggplot(roc.std, aes(x = fpr, y = tpr, ymin = 0, ymax = tpr)) + geom_point() +
  geom_ribbon(alpha = 0.2) + geom_abline(intercept = 0, slope = 1,lty = 2) +
  ggtitle(paste0("ROC Curve for Standardized Data w/ AUC=", auc1)) +
  theme(plot.title = element_text(hjust = 0.5))
```



(1d) (5 pts) Plot the receiver operating characteristic (ROC) curve on the test data using ridge penalty. Also, report the area under the ROC curve (AUC).

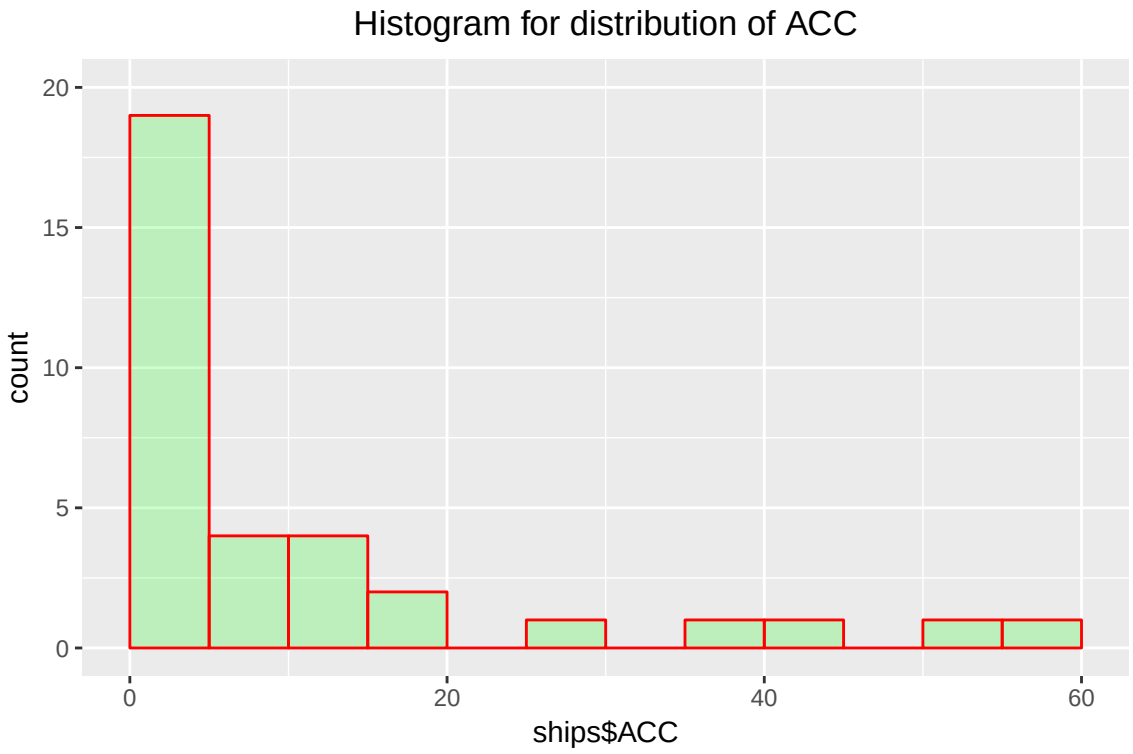
```
# Using Ridge penalty

# Fitting the model, alpha = 0 for the ridge penalty, alpha=1 for the lasso penalty
std_rlogit2 = glmnet(xtrain_std, ytrain_std, family="binomial", alpha=0)

# ROC curve for standardized data
std_prob2 = predict(std_rlogit2, xtest_std, type = "response", s = lambda)
std_pred2 = prediction(std_prob2, ytest_std)
std_pref2 = performance(std_pred2, measure = "tpr", x.measure = "fpr")

# True positive rate
tpr2.ptstd = attr(std_pref2, "y.values")[[1]]

# False positive rate
fpr2.ptstd2 = attr(std_pref2,"x.values")[[1]]

# Area under the curve
auc2.std = attr(performance(std_pred2, "auc"), "y.values")[[1]]
auc2 = signif(auc2.std, digits=3)

roc2.std = data.frame(fpr = fpr2.ptstd2, tpr = tpr2.ptstd, model = "GLM")

# Plotting the ROC curve under test data
ggplot(roc2.std, aes(x = fpr, y = tpr, ymin = 0, ymax = tpr)) + geom_point() +
  geom_ribbon(alpha = 0.2) + geom_abline(intercept = 0, slope = 1,lty = 2) +
  ggtitle(paste0("ROC Curve for Standardized Data w/ AUC=", auc2)) +
  theme(plot.title = element_text(hjust = 0.5))
```



Q2. (15 pts) Analysis of ships data.

(2a) (2 pts) Make a histogram of the variable ACC. Comment on its form.

```
ggplot(data = ships, aes(ships$ACC)) +
  geom_histogram(breaks = seq(0, 60, by = 5),
                 col="red", fill="green", alpha = .2) +
  labs(title = "Histogram for distribution of ACC") +
  xlim(c(0,60)) + ylim(c(0,20)) +
  theme(plot.title = element_text(hjust = 0.5))
```



Comments:

From initial visual inspection we note that the ACC is highly asymmetric on the right side, which is a sign that the chances of having none or one accident per month for ships is quite high. All in all, the histogram shows us the unconditional distribution of ACC.

(2b) (5 pts) Estimate the Poisson regression model including all explicative variables and a constant term.Show your R code and summary output, comment on the coefficient for the variables MONTHS, is it significant?

Be careful on fitting the Poisson model. Note that if you include all the Type (TA-TE) and years (T6569-T7579) dummy variables, an error message would be generated, and no estimation would be performed. To avoid it, TA was chosen to be the reference category for type, and T6064 was chosen to be the reference category for construction year.

```
options(scipen=5)

new_ships = subset(ships, , -c(TYPE, TA, T6064, O6074, O7579))
mod = glm(new_ships$ACC ~ ., family = poisson, data = new_ships)
summary(mod)
```

```
##
## Call:
## glm(formula = new_ships$ACC ~ ., family = poisson, data = new_ships)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9343  -1.5022  -0.5824   0.8669   3.5497
##
## Coefficients:
##                 Estimate   Std. Error z value      Pr(>|z|)
## (Intercept)  0.867529561  0.241285007   3.595      0.000324 ***
## TB           0.989657607  0.212251763   4.663 0.00000312150 ***
## TC          -1.219121872  0.327417344  -3.723      0.000197 ***
## TD          -0.858781207  0.287597135  -2.986      0.002826 **
## TE          -0.242659429  0.236351387  -1.027      0.304567
## T6569        0.950926701  0.176265225   5.395 0.00000006858 ***
## T7074        1.266905841  0.227427122   5.571 0.00000002539 ***
## T7579        0.719229970  0.277311555   2.594      0.009498 **
## MONTHS       0.000044822  0.000007418   6.042 0.00000000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 614.54  on 33  degrees of freedom
## Residual deviance: 106.36  on 25  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 222.22
##
## Number of Fisher Scoring iterations: 6
```

Comments on the coefficient for the variable MONTH: The coefficient of the variable MONTH is very significant which means that we reject the null hypothesis stating that the MONTHS coefficient is equal to zero and we take on the alternative hypothesis stating that it's not equal to zero instead.

We also notice that the value we have for the coefficient is positive which implies that the marginal effect on the MONTHS variable is a positive one. This also basically implies that an increase in the amount of aggregate months in which a ship has been functioning, results in a positive impact on the number of accidents, which intuitevely makes sense.

(2c) (3 pts) Perform a Wald test for the joint significance of all the type dummy variables. Specify the H_0 and H_a, and your conclusion.

```
# Performing Wald Test
wald_mod = glm(new_ships$ACC ~ new_ships$MONTHS + new_ships$TB + new_ships$TC +
                 new_ships$TD + new_ships$TE + new_ships$T6569 + new_ships$T7074 +
                 new_ships$T7579, family = poisson, data = new_ships)

wald.test(b = coef(wald_mod), Sigma = vcov(wald_mod), Terms = c(3:6))
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 76.3, df = 4, P(> X2) = 1e-15
```

$$\begin{cases} H_0 : \beta_{TB} = \beta_{TC} = \beta_{TD} = \beta_{TE} = 0 \\ H_a : \exists i, s.t. \beta_i \neq 0 \end{cases}$$

After performing our Wald Test we received a value 76.35 for the χ^2 statistic, we also received a smaller p-value, indicating that the type dummy variables are jointly significant.

(2d) (5 pts) Given a ship of category TA, constructed in the year period 65-69, with MONTHS=1000. Predict the number of accidents per month. ALso, estimate (1) The probability that no accidents will occur for this ship. (2) the probability that at least two accidents will occur.

```
# Finding the estimation on the number of accidents per month
```

$E[Y|x] = \lambda = e^{(\beta_0 + \beta_{T6569} + (1000(MONTHS)))}$

$e^{(0.867529561 + 0.950926701 + (1000(0.000044822)))} = \mathbf{6.4448300193}$

```
# (1) The probability that no accidents will occur for this ship.
```

$e^{-6.4448300193} = \mathbf{0.001588714597}$

```
# (2) The probability that at least two accidents will occur.
# Equivalent to: 1 - Pr(at most 2 will occur).

# Finding Probability at most two will occur.
```

$P(ACC \leq 1|x) = P(ACC = 0|x) + P(ACC = 1|x)$

$P(ACC \leq 1|x) = e^{-\lambda(x)} + e^{-\lambda(x)}\lambda(x)$

$P(ACC \leq 1|x) = 0.001588932 + 0.010240178 = 0.01182911$

```
# (2) The probability that at least two accidents will occur.
```

$P(ACC \geq 2|x) = 1 - P(ACC \leq 1|x) = 1 - 0.01182911 = \mathbf{0.98817089}$

Q3. (15 pts) Analysis of 3-way contingency table

You investigate the relationship between serum cholesterol (C), gender (G) and heart disease (H) using the given data.

(3a) (5 pts) State the loglinear model that only expresses the main effects of the three characteristics on the expected counts. Interpret the assumption of the model, and compute the fitted values in the top left count of the table, i.e. (male, high cholesterol, with the disease) according to the model.

Solutions:

We have Y_{ijk} = count of Cholesterol i, Gender j and Heart disease k.

Conditional on $n = \sum_{i,j,k} n_{ijk}$

Assume the count $Y_{ijk}|n \sim Multinom(\pi_{ijk})$ where:

i, j and k are mutually independent if $\pi_{ijk} = \pi_i \times \pi_j \times \pi_k$ which is our model assumption.

$$(1)\ E(Y_{ijk}) = n \times \pi_{ijk} = n \times \pi_i \times \pi_j \times \pi_k$$

$$(2)\ \log E(Y_{ijk}) = \log n + \log \pi_i + \log \pi_j + \log \pi_k$$

$$\log E(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$$

which shows that the three variables are independent.

If we let n_{ijk} to be the observed count for i, j, k = 1,2.

We can let $n_{i++} = \sum_{j=1}^{2} \sum_{k=1}^{2} n_{ijk}$,

$n_{+j+} = \sum_{i=1}^{2} \sum_{k=2}^{2} n_{ijk}$,

$n_{++k} = \sum_{i=1}^{2} \sum_{j=1}^{k} n_{ijk}$ and

$n_{+++} = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} n_{ijk}$

What we have is then is an independence model and the fitted value can be calculated as follows:

$$\hat{n}_{ijk} = \frac{n_{i++} n_{+j+} n_{++k}}{n_{+++}^2}$$

```
# Creating independence model
C.G.H.glm = glm(Results ~ Cholesterol + Gender + Heart, data = data1, family =
                  poisson)
# Predict male, high cholesterol, with the disease
predict(C.G.H.glm,type="response")[1]
```

```
##        1
## 4.128503
```

Giving us the result of Male, high cholesterol, with the disease = **4.13**.

(3b) (5 pts) State the loglinear model that expresses all the main effects, and also an interaction between **Cholesterol** and **Gender**, and an interaction between **Cholesterol** and **Heart disease**. Interpret the assumption of the model, and compute the fitted values in the top left count of the table, i.e. (male, high cholesterol, with the disease) according to the model.

Answer:

We have Y_{ijk} = count of Cholesterol i, Gender j and Heart disease k.

Conditional on $n = \sum_{i,j,k} n_{ijk}$.

Assume the count $Y_{ijk}|n \sim Multinom(\pi_{ijk})$ where:

j, k are independent given i if $\pi_{jk|i} = \pi_{j|i} \times \pi_{k|i}$ which is our model assumption.

$$(1)\ E(Y_{ijk}) = n \times \pi_{ijk} = n \times \pi_{jk|i} \times \pi_i = n \times \pi_{j|i} \times \pi_{k|i} \times \pi_i$$

$$(1)\ E(Y_{ijk}) = \frac{(n \times \pi_{j|i} \times \pi_i) \times (\pi_{k|i} \times \pi_i)}{\pi_i} = \frac{n \times \pi_{ij} \times \pi_{ik}}{\pi_i}$$

$$(2)\ \log E(Y_{ijk}) = \log n + \log \pi_{ij} + \log \pi_{ik} - \log \pi_i$$

$$\log E(Y_{ijk}) = \mu + \alpha_j + \beta_k + \gamma_i + (\alpha\gamma)_{ji} + (\beta\gamma)_{ki}$$

This is the conditional model where we have Gender and Heart disease are independent given Cholesterol. We can also calculate the fitted value as follows:

$$\hat{n}_{ijk} = \frac{n_{ij+} n_{i+k}}{n_{i++}}$$

```
# Creating a conditional model comparison
CG.CH.glm = glm(Results ~ Cholesterol * Heart + Cholesterol * Gender, data = data1,
                family = poisson)

# Predict male, high cholesterol, with the disease
predict(CG.CH.glm, type = "response")[1]
```

```
##       1
## 13.0596
```

Giving us the result of Male, with high cholesterol, with the disease = **13.06**.

(3c) (5 pts) For model in (a) and (b), which one is better? Make your conclusion based on AIC and likelihood ratio test.

```
# Model A
# Mutual Independence.
C.G.H.glm = glm(Results ~ Cholesterol + Gender + Heart, data = data1, family =
                  poisson)

# Printing AIC for Model A
round(c(AIC(C.G.H.glm)), digits = 4)
```

```
## [1] 117.9922
```

```
# Model B
# Conditional independence of Gender(G) and Heart(H) given Cholesterol(C).
CG.CH.glm = glm(Results ~ Cholesterol * Heart + Cholesterol * Gender, data = data1,
                family = poisson)

# Printing AIC for Model B
round(c(AIC(CG.CH.glm)), digits = 4)
```

```
## [1] 67.0053
```

```
# Performing the likelihood ratio test on the two models
anova(C.G.H.glm, CG.CH.glm, test = "Chis")
```

```
## Analysis of Deviance Table
##
## Model 1: Results ~ Cholesterol + Gender + Heart
## Model 2: Results ~ Cholesterol * Heart + Cholesterol * Gender
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         4     56.315
## 2         2      1.328  2   54.987 1.147e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solutions:

Looking at the AIC we received a value of 117.9922 for Model A and 67.0053 for Model B. For AIC, the smaller the value the better the fit of the model is (from a statistical perspective) as they reflect a trade-off between the lack of fit and the number of parameters in the model. In our case using the AIC we prefer Model B (Conditional Model) which has the lower AIC.

Looking at the Likelihood ratio test, which is a test we use for comparing the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model). In our scenario, our null model is Model A (Mutual Independence) and our alternative model is Model B (Conditional Independence). The test helps us expresses how many times more likely the data are under one model than the other. After this test was performed we received a p-value of 1.147e-12 which means we would reject the null model and take on the alternative model which is Model B.

In both cases the Conditional Model (Model B) seems to be the better one.