STA138 Project II

Problem 1: Logistic Regression

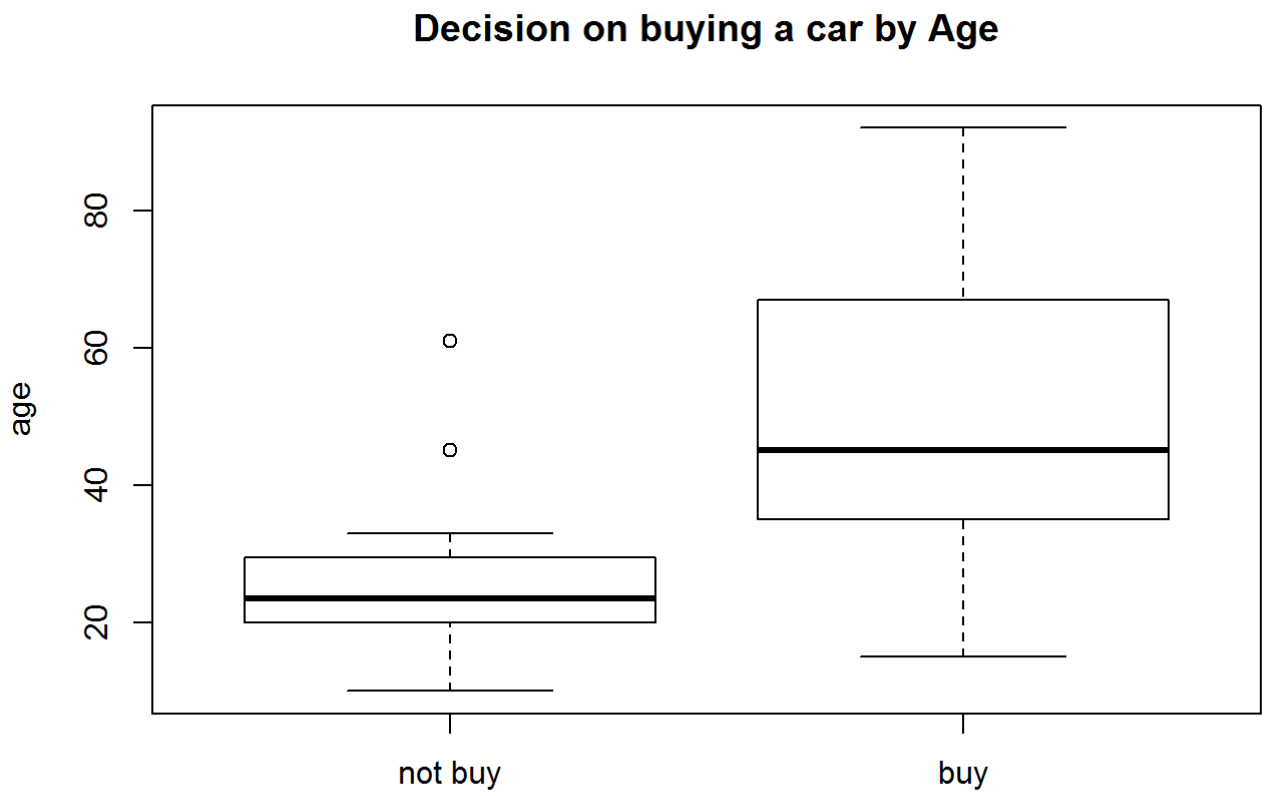
1. Introduction

In this project, we will explore how the age of a consumer affects their decision on buying a car. This analysis can help the dealership to better identify their potential consumers.

2. Summary

The average age for customers who did not buy a car is 26.375 and the average age for customers who did buy a car is 49.4117647.

The standard deviation of age for customers who did not buy a car is 12.2086035 and the standard deviation of age for customers who did buy a car is 22.4445395.

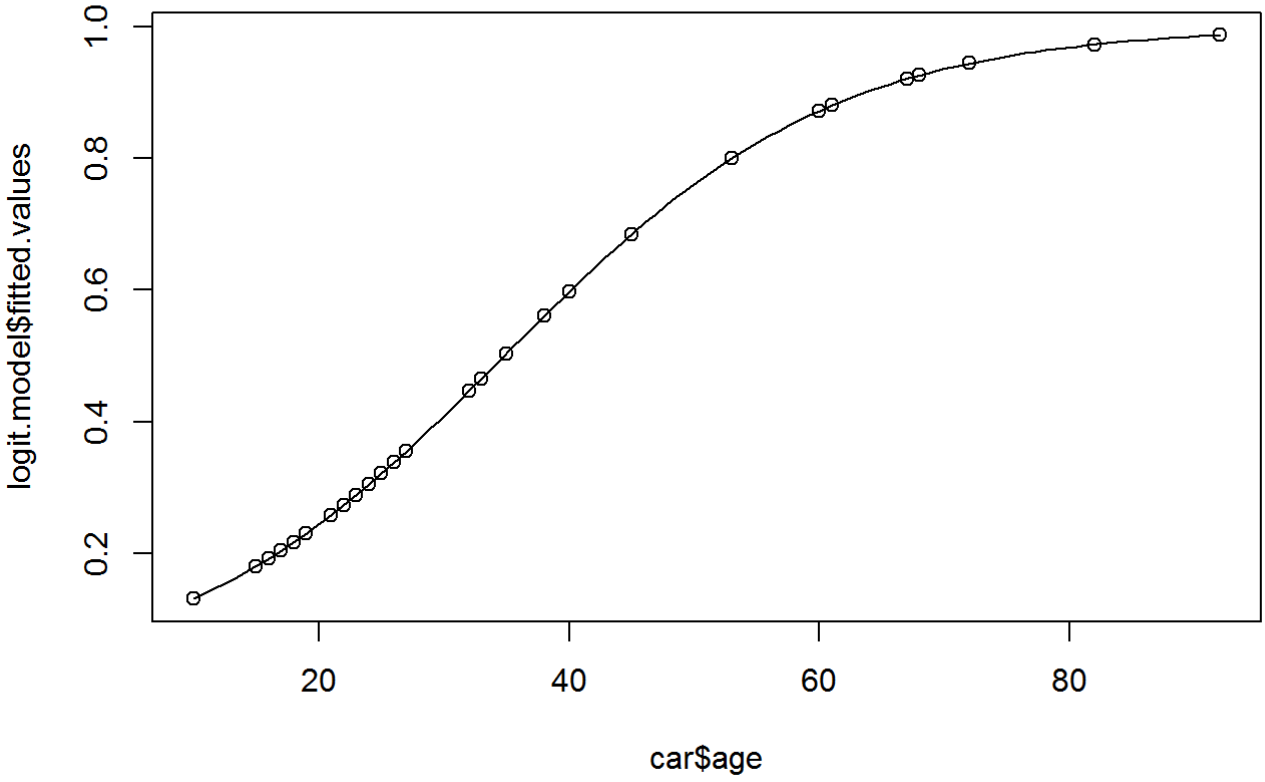


As we can see from the boxplot, the average age for customers who did buy a car is higher than the ones who did not. The boxplot is relatively short for the customers who did not buy the car, which suggests that the overall customers in this group have similar age. There are two outliers for this group. The boxplot is comparatively tall for the customers who did buy a car, which suggests that the customers have relatively different ages.

3. Analysis

```
##                Estimate Std. Error   z value    Pr(>|z|)
## (Intercept) -2.65826472 1.03168026 -2.576636 0.009976689
## age          0.07635294 0.02892154  2.640003 0.008290531
```

This gives estimates $\hat{\alpha}$ = -2.6582647, and $\hat{\beta}$ = 0.0763529. The estimated logistic regression function is logit(π(X)) = -2.6582647 + 0.0763529Xi



Since the curve is quite steep, therefore, we conclude that age has a positive effect (since β is positive) on whether the consumer brought the car.

The exp(β_1) is 1.0793434.

The 99% LR confidence interval for exp(β_1) is (1.0161763, 1.1882431).

We choose the likelihood ratio test because we want a narrower confidence interval.

The LR Test Statistic is 11.7989243

The p value for LR is 5.926495610^{-4}.

4. Interpretation

When the probability of a consumer brought the car is 0.5, the age is 34.8154873 years old.

The odds of a consumer buying a car when age increases by 1 year are 1.0793434 times what they were.

We are 99% confident that the odds of buying a car is within the interval 1.0161763, 1.1882431 when age increases by 1 year. In addition, since the confidence interval for exp(β_1) does not cover one, we believe that age may have an effect on if the consumer brought the car.

If Ho: β_1 = 0 is true, which means that age does not affect the probability of buying a car, the probability of observing our data or more extreme is 5.926495610^{-4}.

Since p-value is less than alpha at any reasonable alpha = 0.1, 0.05 and 0.01. We reject Ho and conclude that there is a positive effect of age on the probability of buying a car.

5. Conclusion

Based on all our analysis above, we can conclude that age has a positive effect on the probability of a consumer buying a car. Therefore, if the consumer is older, he or she will have a higher probability of purchasing the car. This conclusion is not surprising. People are generally more financial stable as they get older. Therefore, their purchasing power also increases.

Problem 2: Log-Linear Model

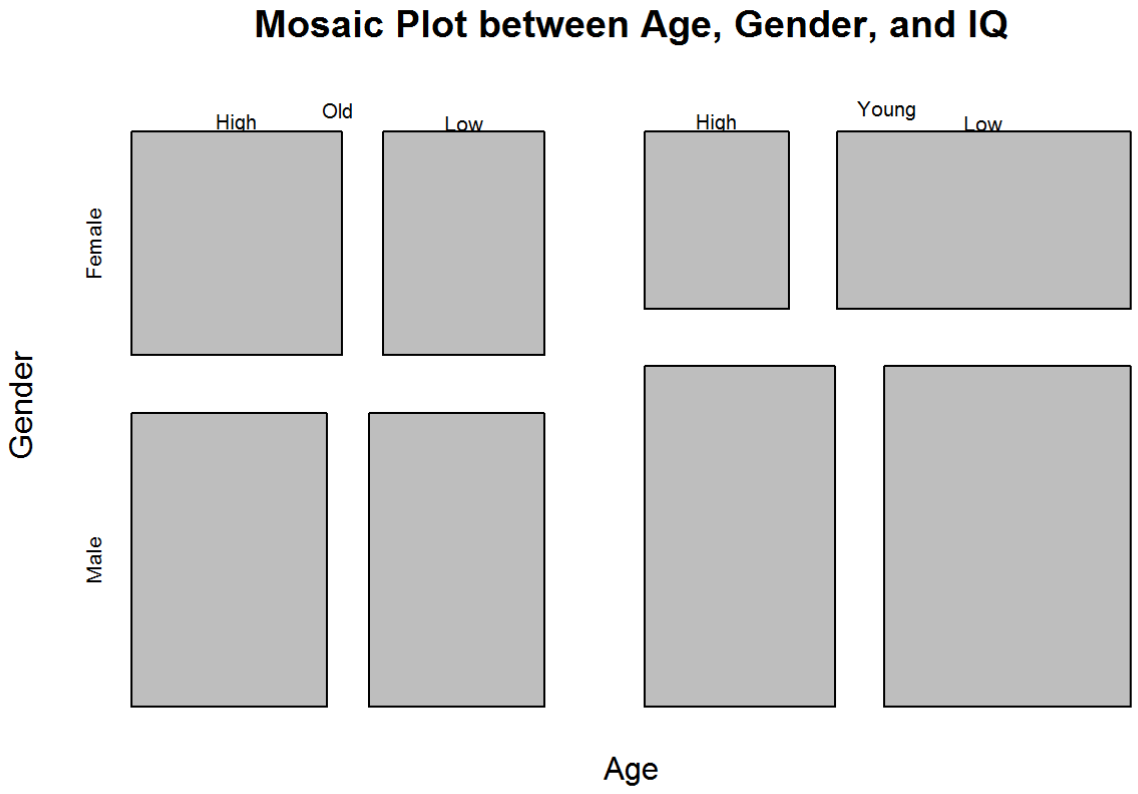
1. Introduction:

In this project, we will examine the dependence between age and gender, between age and IQ, and between gender and IQ.

2. Summary

The frequencies are shown below.

```
## , , IQ = High
##
##        Gender
## Age     Female Male
##   Old       57   70
##   Young     31   79
##
## , , IQ = Low
##
##        Gender
## Age     Female Male
##   Old       44   63
##   Young     63  102
```



Regardless of age or IQ, there is a higher proportion of males than females. Regardless of gender, old people tend to have a higher proportion of high IQ than low IQ while young people tend to have a higher proportion of low IQ than high IQ. Old females have a higher proportion of having a high IQ compared to old males. Young males have a higher proportion of having high IQ compared to young females.

3. Analysis

Below are all the AICs for each model:

```
##    (X,Y,Z)     (X,YZ)     (Y,XZ)     (Z,XY)    (XY,XZ)    (XY,YZ)
##    73.5115    75.2502    65.1341    71.1999    62.8224    72.9386
##    (XZ,YZ) (XY,XZ,YZ)
##    66.8728    64.1510
```

The best model is the (XY, XZ) model because it has the lowest AIC. We used AIC because we want to penalize the model for having too many parameters.

Our model looks like $ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$

Our model is conditionally independent. Looking at the interaction terms, we have λ_{ij}^{XY} so age and gender are dependent and we have λ_{ik}^{XZ} so age and IQ are dependent.

The coefficients for the model is shown below.

```
##  (Intercept)       XYoung        YMale         ZLow XYoung:YMale
##    4.0039865   -0.3769824    0.2752286   -0.1713583    0.3799736
##  XYoung:ZLow
##    0.5768234
```

Since we cannot construct a confidence interval using the Pearson's test statistic, we will use the Likelihood Ratio test statistics.

The LR test statistic is 3.3055.

The p-value for LR is 0.1915.

The 95% LR confidence intervals are shown below.

```
##                    2.5 %      97.5 %
## (Intercept)   3.76935921  4.22514933
## XYoung       -0.71544116 -0.04029104
## YMale         0.01771007  0.53580743
## ZLow         -0.42985555  0.08524838
## XYoung:YMale  0.02130284  0.74021576
## XYoung:ZLow   0.22536306  0.93110843
```

The 95% LR exp(CI) are shown below.

```
##                    2.5 %     97.5 %
## (Intercept)  43.3522761 68.3847149
## XYoung        0.4889763  0.9605099
## YMale         1.0178678  1.7088274
## ZLow          0.6506031  1.0889875
## XYoung:YMale  1.0215314  2.0963878
## XYoung:ZLow   1.2527775  2.5373200
```

4. Interpretation

Ho: Current model fits well compared to the saturated model

Ha: Current model doesn't fit well compared to the saturated model

Since our p-value > α for all α = 0.01, 0.05, 0.10, we fail to reject Ho and conclude the (XY, XZ) model fits well.

We are 95% confident that young age has a negative effect on the log frequency because it is between -0.7154412 and -0.040291.

We are 95% confident that males have a positive effect on the log frequency because it is between 0.0177101 and 0.5358074.

We are 95% confident that low IQ has no effect on the log frequency because it is between -0.4298556 and 0.0852484.

We are 95% confident that the estimate for old females with high IQ is between 43.3522761 and 68.3847149.

We are 95% confident that the odds ratio for young people for males vs females regardless of IQ is between 1.0215314 and 2.0963878.

We are 95% confident that the odds ratio for young people for low vs high IQ regardless of gender is between 1.2527775 and 2.53732.

5. Conclusion

We conclude that age and gender are dependent as well as age and IQ. The probability of being a young male is higher than being a young female. The probability of being young and having a low IQ is more likely than being young and having a high IQ.

Code Appendix

```r
library(readr)
car <- read_csv("C:/Users/pichu/Desktop/STA138/Project II/Car.csv")
library(plyr)
car = rename(car, c("Y" = "buy", "X" = "age"))
age_no = with(car, mean(age[buy == 0]))
age_buy = with(car, mean(age[buy == 1]))
age0 = with(car, sd(age[buy == 0]))
age1 = with(car, sd(age[buy == 1]))
boxplot(age~buy, data = car, main = "Decision on buying a car by Age", ylab = "age", names =
c("not buy", "buy"))
#Step 1: Fit the model
logit.model = glm(formula = buy ~ age, family = binomial(logit), data = car)
summary(logit.model)$coefficients
#Step 2: Plot the logistic curve and get the covariance between $\alpha$ and $\beta$
plot(car$age, logit.model$fitted.values)
curve(predict(logit.model, data.frame(age=x), type = "response"), add = TRUE)
exp.beta = exp(logit.model$coefficients[2])
#Step 4:
alpha = 0.01
CI = confint(logit.model,level = (1-alpha))[2, ]
exp = exp(CI)
smaller.model = glm(buy ~ 1, family = binomial(logit), data = car)
LR = as.numeric(-2*(logLik(smaller.model) - logLik(logit.model)))
d.f = length(logit.model$coefficients) - length(smaller.model$coefficients)
p.val.LR = pchisq(LR, d.f, lower.tail = FALSE)
#Step 3: Find the value of age when the probability of a consumer brought the car is 0.5
pi_x = -logit.model$coefficients[1]/ logit.model$coefficients[2]
iqs = read_csv("C:/Users/pichu/Desktop/STA138/Project II/IQshort.csv")
iql = read_csv("C:/Users/pichu/Desktop/STA138/Project II/IQlong.csv")
colnames(iql) = c("Age", "Gender", "IQ")

tab = table(iql)
tab

mosaicplot(tab, main = "Mosaic Plot between Age, Gender, and IQ")
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)
}

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")
```

```
all.model.fits = lapply(all.model.formulas,function(the.model){
  glm(the.model,data = iqs, family = poisson)
})

all.GOF = sapply(all.model.fits,function(the.model){
  good.fit.LL(the.model)
}) #It is the wrong orientation so I flip it
all.GOF = t(all.GOF)
#I also add the model formulas for reference
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,Y
Z)")
rownames(all.GOF) = book.notation
round(all.GOF[,7],digits = 4)

model = glm(F~X+Y+Z+X*Y+X*Z, data = iqs, family = poisson)

model1 = glm(F~X+Y+Z+X*Y+X*Z+Y*Z, data = iqs, family = poisson)
model$coefficients
stuff = round(good.fit.LL(model),4)
alpha = 0.05
ci = confint(model,level = 1-alpha)
ci
ci2 = exp(ci)
ci2
```