STA 138
MIDTERM 2 PROJECT

998358453

999210972

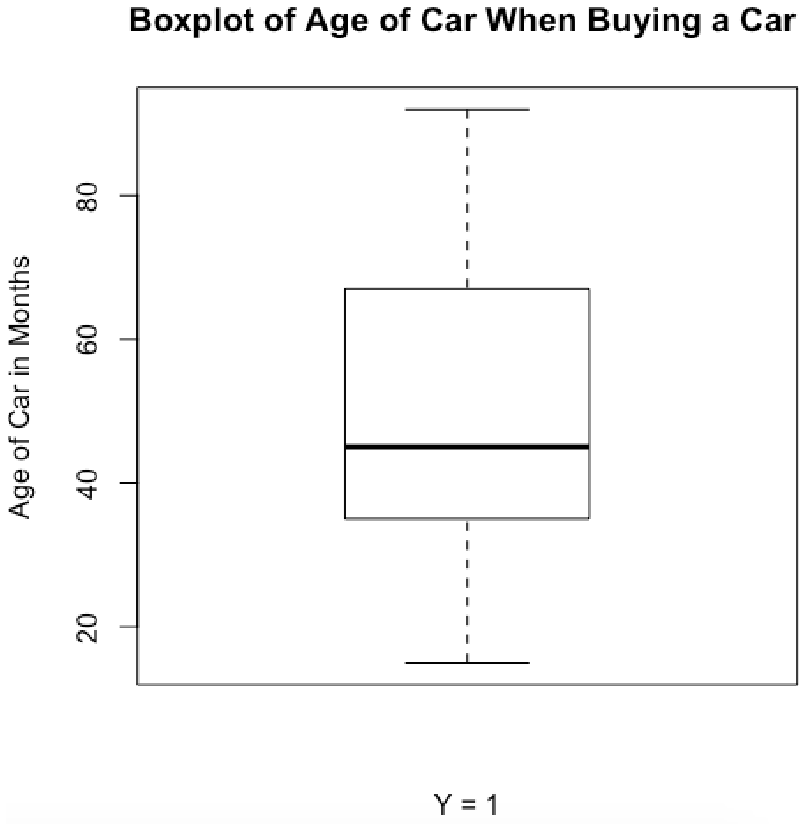
PROBLEM 1

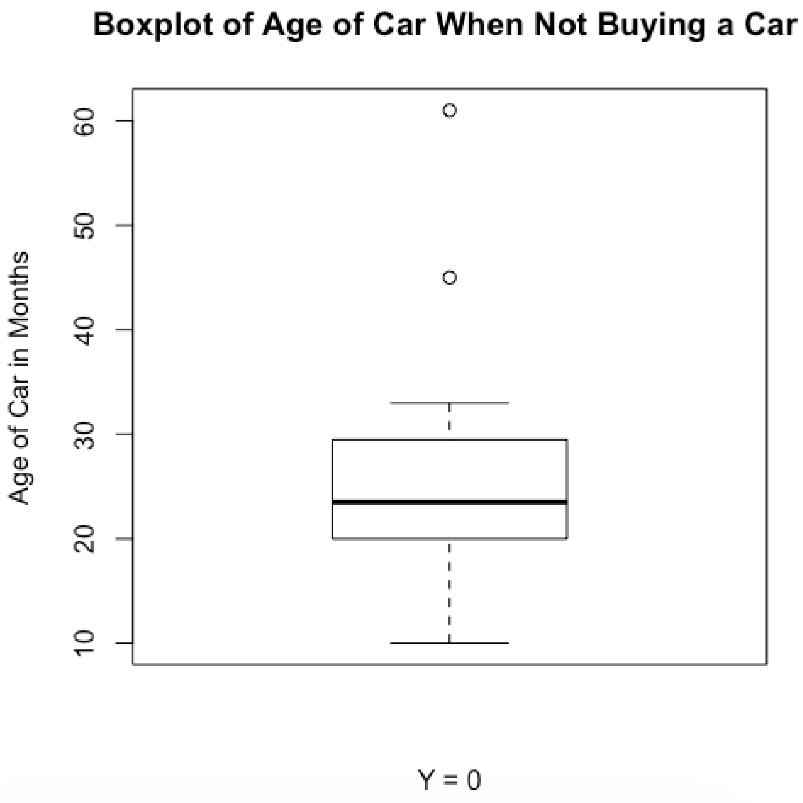
INTRODUCTION

Is there a relationship between buying a car and the age of the currently owned car? Data was gathered on customers who visited car dealerships to see if they bought a car and the age of the current car. All this information was recorded in “Car.csv”. The first column, Y, has the value 1 if the customer bought a new car and 0 if they did not buy a car. The second column X is the age of the current owned car in months. This data set is interesting and important because we can determine if there’s a relationship between buying a car and the age of the currently owned car. This will help dealerships predict the probability of a customer buying a car based on their current owned car’s age in month.

SUMMARY

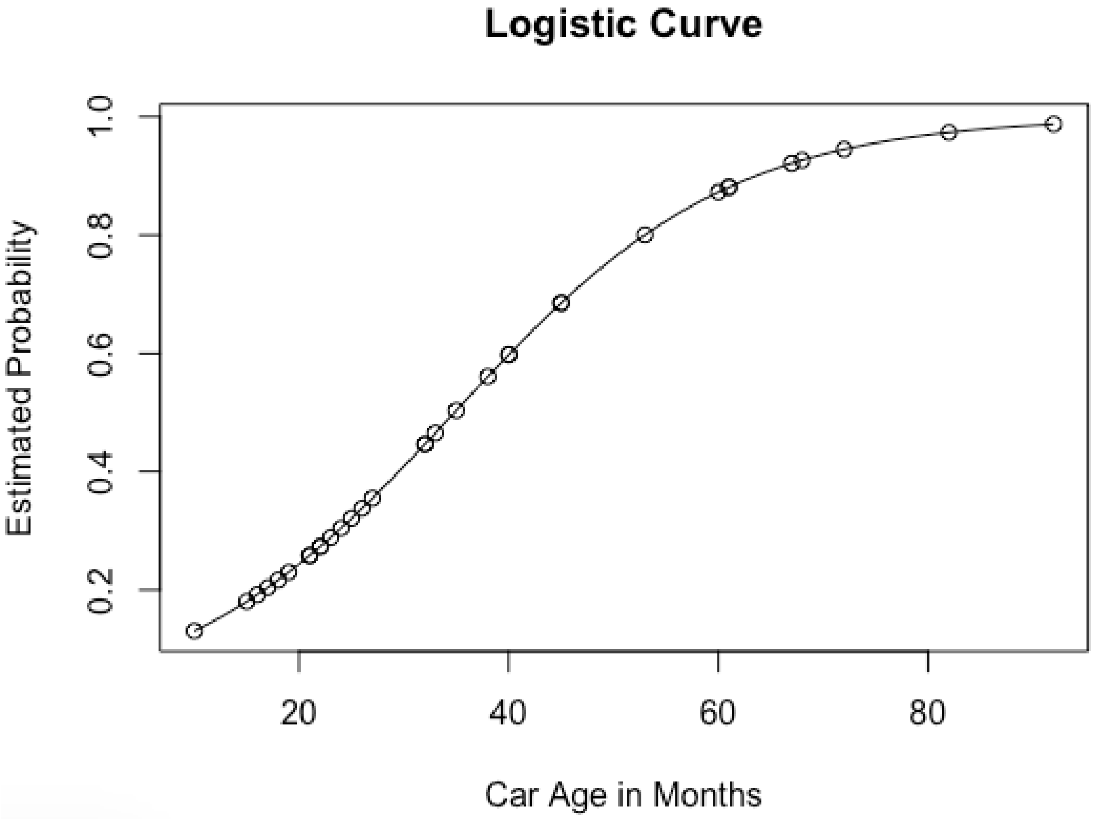
There are a total of 33 observations and 17 of those bought a car. The range of a car’s age in month is [10, 92].





This first box plot shows the range of the age of the currently owned car in months for customers who successfully purchased a car. It seems like the customer who bought a car had the youngest car age at 15 months and the oldest car age at 92 months. The median age of car for a customer to buy a new is car is around 45 months.

The second boxplot shows the range of age of cars in months for customers who did not buy a car when going into a car dealership. The youngest currently owned car is 10 months old whereas the oldest car age that is not the outliers is around 32 months old. However, this plot has 2 outliers at 45 months and 61 months old.



In the above logistic curve plot, there isn't a steep slope so it does not appear that the age of a car between [10, 20] and [60, 92] has a significant effect on whether a customer buys a car or not. However, the slope is steeper when the car age is between [20, 60] where there may be an effect on whether a customer buys a car or not.

ANALYSIS

Null Hypothesis: $B_1 = 0$
Alternative Hypothesis: $B_1 \neq 0$

X_1 = age of car in months
logit($\boldsymbol{\pi}$(x)) = $-2.65826 + 0.07635X_1$

The probability of 0.5 for a customer buying a car is when the age of the currently owned car is 34.815485 months old.

The value of exp(B_1) is 1.079343.

The Wald test-statistic for the null hypothesis is 2.64. Therefore, the Wald p-value is 0.008290531.

The Wald 99% confidence interval after exponential is (1.001858, 1.162822).

INTERPRETATION

We are testing to see if there's an effect on the age of car and a customer buying a new car.

exp(α) is an estimate of the odds of a customer buying a car when the age of the currently owned car is 0. This has no practical interpretation because when the age is 0, it exceeds our data's range. The minimum of our data's range for the age of car is 10 months. Also, a car of age 0 means the car is newly bought and that means the customer will not need to buy a new car again just yet.

The odds of buying a new car when the age of car increases by 1 month are exp(B_1) = 1.079343 times what they were.

A customer has a 50% chance of a buying a new car if their currently owned car is 34.815485 months old.

Since alpha = 0.01 and our p-value is 0.008290531 which is less than alpha, we reject the null hypothesis and conclude that the age of car has some (positive) effect on a customer buying a new car. A p-value of 0.008290531 means that if in reality, there is no effect on the age of car on a customer buying a new car, the probability of observing our data or more extreme is 0.8290531%.

When the age of car increases by 1 month, the odds a customer buying a new car are between 1.001858 and 1.162822 times that of what they were with 99% confidence.

CONCLUSION

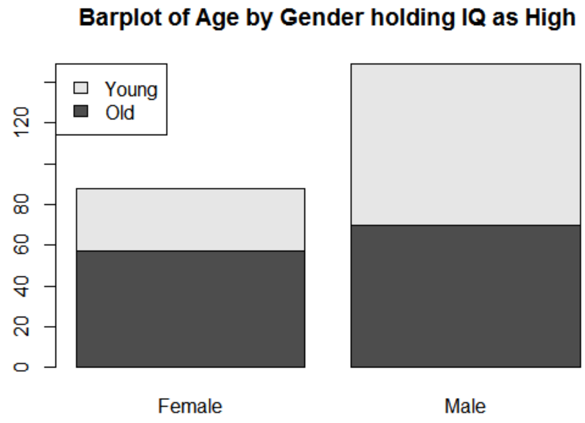
We tested to see if there was an effect on age of currently owned car in months on buying a new car when a customer goes to the car dealership. Since our 99% confidence interval for exp(B_1) did not contain 1, it suggests an influence on age of currently owned car on the odds of buying a car ($Y = 1$). Overall, the older the car, the higher the probability of buying a car. This makes sense since an older car means it is time for change and an upgrade by buying a new car.

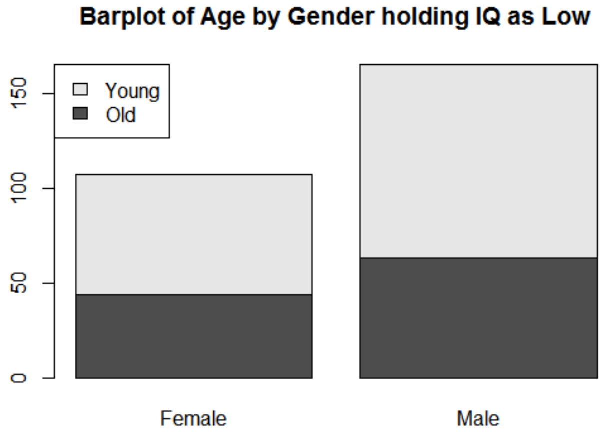
PROBLEM 2

Introduction

For the dataset "IQshort.csv," we examine the dependence between age (young and old), gender (male and female), and IQ (high and low). The data of 509 observations were obtained and the age (X), gender(Y), and IQ levels(Z) were collected. For this data, we selected the best model to fit these three variables, and then tested the dependence of each pair of variables determined to be dependent.

Summary





Model Selection:

Model	*AIC*	*Pearson T-stat*	*Pearson p-value*	*LR*	*LR p-value*
F~X+Y+Z	73.5115	17.4660	0.0016	17.9946	0.0012
F~X+Y+Z+Y*Z	75.2502	17.6293	0.0005	17.7333	0.0005
F~X+Y+Z+X*Z	65.1341	7.4544	0.0587	7.6171	0.0546
F~X+Y+Z+X*Y	71.1999	13.5152	0.0036	13.6829	0.0034

F~X+Y+Z+X*Y+X*Z	**62.8224**	**3.2683**	**0.1951**	**3.3055**	**0.1915**
F~X+Y+Z+X*Y+Y*Z	72.6386	13.3005	0.0013	13.4216	0.0012
F~X+Y+Z+X*Z+Y*Z	66.8728	7.2773	0.0263	7.3558	0.0253
F~X+Y+Z+X*Y+X*Z+Y*Z	64.1510	2.6204	0.1055	2.6341	0.1046
F~X+Y+Z+X*Y+X*Z+Y*Z+X*Y*Z	63.5169	0	1.0000	0	1

Model selection using the AIC criteria gives us the "best model" as $\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$.

Analysis

To select the best model, chose to use the AIC criteria because AIC penalizes for large models, but we still want to minimize AIC. We fit all possible models and chose the model with the lowest AIC because a lower AIC tells us that the quality of the model is better relative to each of the other models fitted. The resulting "best model" is $\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$.

- The model suggests there may be dependence of age on gender because the interaction term λ_{ij}^{XY} is in the model
- The model suggests there may be dependence of age on IQ level because the interaction term λ_{ik}^{XZ} is in the model

To test whether the current model fits well, we used Pearson's test to test the fit by testing $H_0\colon \textit{the current model fits well}$ against $H_a\colon \textit{the current model does not fit well.}$

- The resulting Pearson's test statistics is $X^2 = 3.2683$ and the p-value is 0.1951

To test whether we can drop age and gender interaction term (implies that age and gender are independent), we can use the Likelihood Ratio test to test $H_0\colon \lambda_{ij}^{XY} = 0$ against $H_a\colon \lambda_{ij}^{XY} \neq 0$.

- The test statistic is $G^2 = G^2[(Y, XZ)] - G^2[(XY, XZ)] = 7.6171 - 3.3055 = 4.3116$
- The p-value is 0.0379

To test whether we can drop age and IQ interaction term (implies that age and IQ are independent), we can use the Likelihood Ratio test to test $H_0\colon \lambda_{ik}^{XZ} = 0$ against $H_a\colon \lambda_{ik}^{XZ} \neq 0$

- The test statistic is $G^2 = G^2[(Z, XY)] - G^2[(XY, XZ)] = 13.6829 - 3.3055 = 10.3774$
- The p-value is 0.0013

Since age and gender as well as gender and IQ were found to be dependent, we found the Wald confidence intervals for the exponential of the interaction terms λ_{ij}^{XY} and λ_{ik}^{XZ} to determine if there is an effect of interaction between age and gender as well as age and IQ.
To have an overall 95% CI, each CI was corrected to be a 1-(α/g) = 1-(0.05/2) = 0.975 = 97.5% confidence interval.

- For the interaction term between age and gender, the confidence interval was between -0.0308 and 0.7907
  - For the exponentiated interaction term between age and gender, the confidence interval was between 1.0313 and 2.2049

- For the interaction term between age and IQ, the confidence interval was between -0.1736 and 0.9801
  - For the exponentiated interaction term between age and IQ, the confidence interval was between 1.1896 and 2.6647

Interpretation

For Pearson's test for determining if the current model fits well, since the p-value is 0.1951, at any reasonable level of α, we fail to reject the null hypothesis that model fits well. We conclude that the current model of $\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$ fits well.

For the Likelihood Ratio test for whether we can drop the age and gender interaction term, since the p-value is 0.0379 > 0.01, we fail to reject H_0: $\lambda_{ij}^{XY} = 0$ and conclude that age and gender are independent.

For the Likelihood Ratio test for whether we can drop the age and IQ interaction term, since the p-value is 0.0013 < 0.01, we reject H_0: $\lambda_{ik}^{XZ} = 0$ and conclude that age and gender are dependent.

Using the Wald confidence interval to test whether there is an effect for interaction between age and gender as well as age and IQ, we determine if the interaction term is significant.

- We are 95% confident that the log odds of people who are young and male is between -0.0308 and 0.7907 that of people who are young and female.
- We are 95% confident that the estimated odds of people who are young and male is between 0.9697 and 2.2049 that of those who are young and female. Since the confidence interval contains 1, age and gender are independent.
- We are 95% confident that the log odds of people who are young and have low IQ is between 0.1736 and 0.9801 that of those who are young and female.
- We are 95% confident that the estimated odds of people who are young and have low IQs is between 1.1896 and 2.6647 that of those who are young and have high IQ. Since this confidence interval does not contain 1, age and IQ are dependent.

Conclusion

Using the AIC criteria for model selection, the "best fit" model was determined to be $\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$. This also suggested that age and gender is dependent, as well as age and IQ levels. We used Pearson's test to test if the current model is a good fit and determined that it is and that age and gender are dependent as well as age and IQ. We then tested whether age and gender was dependent using the Likelihood ratio test and concluded that they are independent. Testing age and IQ using the Likelihood ratio test also concluded that they are dependent. Further supporting this idea, is the overall 95% confidence intervals for the odds ratio for age and gender was found to be independent while age and IQ are dependent.

APPENDIX PROBLEM 1

```
#import data
Car <- read.csv("~/Desktop/Car.csv")

zero = Car[Car$Y == 0, ]
one = Car[Car$Y == 1, ]

boxplot(one$X, xlab = "Y = 1", ylab = "Age of Car in Months", main = "Boxplot of Age of Car When Buying a Car ")
boxplot(zero$X, xlab = "Y = 0", ylab = "Age of Car in Months", main = "Boxplot of Age of Car When Not Buying a Car ")

#logistic model
logit.model =  glm(formula = Y ~ X, family = binomial(logit), data = Car)
summary(logit.model)

#basic statistics on data
tot = sum(Car$Y)
mini = min(Car$X)
maxi = max(Car$X)

#plot curve
plot(Car$X,logit.model$fitted.values, xlab = "Car Age in Months",ylab = "Estimated Probability", main = "Logistic Curve")
curve(predict(logit.model, data.frame(X=x), type="response"), add=TRUE)

estimates =  summary(logit.model)$coefficients[,1] # A vector of only the estimates
SE =  summary(logit.model)$coefficients[,2] #A vector of only the Wald SE's

bestimate = exp(estimates[2])

#calculate B CI
alpha = 0.01
z.a.2 = qnorm(1-alpha/2)
upper.bounds = estimates +z.a.2*SE
lower.bounds = estimates -z.a.2*SE
Wald.CI = cbind(lower.bounds,upper.bounds)
Wald.CI
```

```
#exp(B) CI
lower = exp(Wald.CI[2,1])
upper = exp(Wald.CI[2,2])

(-coef(logit.model)[1]) / coef(logit.model)[2]

predict(logit.model,newdata = data.frame(X = 34.815485),type = "response")
```

APPENDIX PROBLEM 2

```
IQshort = read.csv("C:\\Users\\Shirley\\Desktop\\IQshort.csv", header = T)

#Goodness of Fit Function
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)
}

#fit all models
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 = IQshort, family = poisson)
})

#goodness of fit statistics
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
round(all.GOF,digits =4) #Rounding components for readability

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) #Rounding components for readability

#overall 95% CI (1-alpha/g) where g = 2 (two CI)
alpha = 0.05/2
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)
CIs

IQlong = read.csv("C:\\Users\\Shirley\\Desktop\\IQlong.csv", header = T)
IQlong_table = table(IQlong)

#Bar plot of age vs gender holding IQ as high
IQlong_low = matrix(c(44,63,63,102),ncol=2,byrow=TRUE)
```

```
colnames(IQlong_low) = c("Female","Male")
rownames(IQlong_low) = c("Old","Young")
IQlong_low = as.table(IQlong_low)
barplot(IQlong_low, main = "Barplot of Age by Gender holding IQ as Low", legend =
rownames(IQlong_low), args.legend = list(x="topleft"))

#Bar plot of age vs gender holding IQ as high
IQlong_high = matrix(c(57,70,31,79),ncol=2,byrow=TRUE)
colnames(IQlong_high) = c("Female","Male")
rownames(IQlong_high) = c("Old","Young")
IQlong_high = as.table(IQlong_high)
barplot(IQlong_high, main = "Barplot of Age by Gender holding IQ as High", legend =
rownames(IQlong_high), args.legend = list(x="topleft"))
```