STA 138 Project 2

PROBLEM 1 - Car.csv

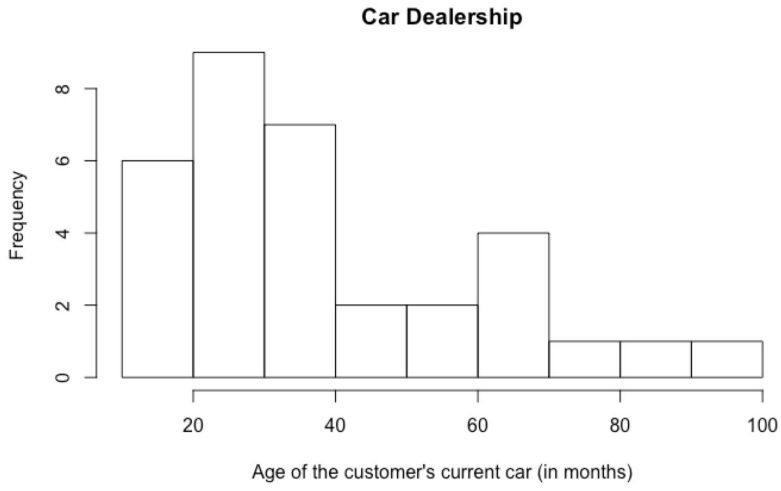
Introduction:

The goal of analyzing the Car data set is to explore the relationship between a variable, X, and the probability that the response is Y=1, a success. We will be using graphs, hypothesis tests, confidence intervals, by fitting the data into a log-linear model. The results will be of interest to car dealerships since it is useful to understand the relationship of when customers tend to buy new cars.

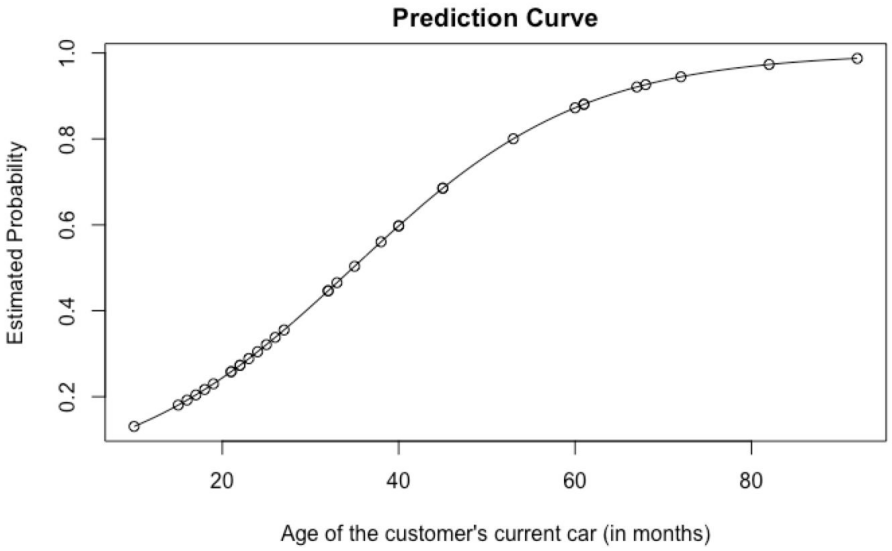
Summary:

The Car.csv is a data frame that consists of two variables, X and Y. In this data, customers who came into a car dealership were polled, and if they bought a car and the age of the current car was recorded.
Column 1 is the Y variable, which takes on the value 1 if the customer bought a new car, and 0 otherwise. Column 2 is the X variable, where the age of the customer's current car in months are recorded.

In order to attain an overview of the data set, we will explore a visual representation of the data set.



This histogram presents the age of the customer's current car in months that visited the car dealership. We can observe that it ranges from around 10 months to a 100 months and that there seems to be more customers with their car age ranging from 10 to 40 months.



This is a prediction curve that fitted the logistic-linear model of the given data set. From this curve, we can see that the relationship between the X and Y variable may be significant since the curve has a relatively steep slope.

Now that we understand the data set better, we will explore the relationship of X and Y variable quantitatively in order to derive accurate conclusions about their relationships.

Analysis:

H_0 : Customer buying a car and the age of their current car is independent.
H_A : Customer buying a car and the age of their current car is dependent.

Wald test statistic is 2.640003 with a p-value of 0.00829053.

Log-likelihood ratio test is 11.79892 with degrees of freedom of 1 and a p-value of 0.00059265.

The confidence interval after exponentiation is (1.0288, 1.1562).

H_0 : Smaller model fits better.
H_A : Saturated model fits better.

The saturated model Y ~ X has an AIC value of 37.92. A smaller model, Y ~ 1 has an AIC value of 47.72.

The logistic-regression model is $\text{logit}(\hat{\pi}) = -2.65826 + 0.07653X$. Therefore, $\hat{\alpha} = -2.65826$ and $\hat{\beta}_1 = 0.07653$.

The $\exp(\hat{\beta}_1) = \exp(0.07653) = 1.07934$.

$X_{50} = \hat{\alpha}/\hat{\beta}_1 = 34.81676$.

Interpretation:

If the relationship between customers buying the car and the age of their current car is independent, the probability of observing our data or more extreme is 0.00829053 or 0.00059265. Since the p-value is very small, (p-value $< \alpha = 0.05$), we reject the null hypothesis and conclude that there is a dependent relationship between customers buying the car and the age of their current car. The exponentiated confidence interval of (1.0288, 1.1562) indicates that we are 99% confident that when age of the customer's current car increases by 1 month, the odds of purchasing a new car are between 1.0288 and 1.1562 times what they were.

Looking at the logistic-regression model the AIC of the saturated model is smaller. Therefore, we reject the null hypothesis and conclude that the saturated model fits better. The saturated model is $\text{logit}(\hat{\pi}) = -2.65826 + 0.07653X$. Since the sign of $\hat{\beta}_1$ is positive, it suggests the probability increases as age increases. When the age of the customer's current car increases by 1 month, the odds of buying a new car is 1.07934 times what they were. The 50% probability of buying a new car is when the customer's current car is around 34.81676 months old.

Conclusion:

The first set of hypothesis test is with null hypothesis stating that the customer buying a car and the age of their current car is independent. Since the p-value for the test statistic is very small, (p-value $< \alpha = 0.05$), we reject the null hypothesis and conclude that there is a dependent relationship between customers buying the car and the age of their current car.
For the second set of hypothesis test where the null hypothesis states that smaller model fits better, we can conclude that the bigger model fits better and that there is a dependency between a customer buying a car and the age of their current car.

PROBLEM 2 - TrialsShort.csv

Introduction:

We have a data set that shows the frequency of exercise, the type of drug taken, and improvement on a medical condition after 6 months. We would like to use log-linear models to see if there is a relationship between or among any of the three variables. This dependency might be of interest to researchers or medical professionals so that they can give helpful advice to patients who want to improve their medical condition.

Summary

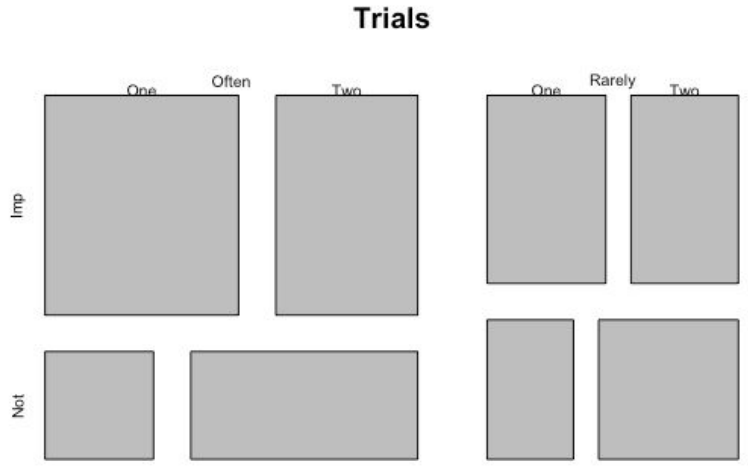
In this dataset, we have three categorical variables (X=exercise, Y=Condition Status, and Z=Drug Type). Each variable has two categories contained within it (X: Often/Rarely; Y: Imp/Not; Z: One/Two).

Our goal is to use log-linear models in order to estimate expected counts and log-odds, so that we can determine if there is dependence among the variables. We want to know if an aspect of a variable will affect the outcome of another variable. More specifically, we want to know if exercise and/or drug use can improve a medical condition over a span of time.

The table below shows the data we are given, as well as the estimated counts for each category. From the numbers, we can tell that X=Exercise Often, Y=Improvement, and Z=Group 1 is a large category, as well as X=Exercise Rarely, Y=No Improvement, and Z=Group 2.

X	Y	Z	F	Estimated Counts
Often	Imp	One	120	116.04878
Often	Imp	Two	88	91.95122
Often	Not	One	33	35.7801
Often	Not	Two	69	66.2199
Rarely	Imp	One	63	66.95122
Rarely	Imp	Two	57	53.04878
Rarely	Not	One	34	31.2199
Rarely	Not	Two	55	57.7801

From our mosaic plot below, it seems that those who exercise often and show improvement are greater in group 1 than group 2. Those who exercise often and don't show improvement have greater numbers in group 2 than group 1. Those who rarely exercise and show improvement are roughly the same in both groups. Those who rarely exercise and don't show improvement have greater occurrence in group 2 than group 1.



Analysis and Interpretation: Choose estimates/ CI / Hypothesis tests. Report numerical results.

Since we have 3 variables in our data, R fits nine models, which include mutually independent, jointly independent, conditionally independent, and fully dependent models. We choose the model F~X+Y+Z+XY+YZ as our model of best fit because it has the lowest AIC =60.9865, (-2(maximized log-likelihood-# parameters)). This model suggests dependence between X and Y, and Y and Z.

To test if this model of conditional independence fits well, we will state our null hypothesis as: "The model F~X+Y+Z+XY+YZ fits our data well", and our alternative hypothesis as: "The model F~X+Y+Z+XY+YZ does not fit our data well". The value of our Likelihood Ratio test statistic is **1.5439**, with a p-value of **.4621.** The value of Pearson's test statistic is **1.5459** with a p-value of **.4616.** Since both of these p-values are greater than alpha (at any significance level .1, .05, and .01), we fail to reject our null hypothesis and conclude that this particular model fits our data well, and that there are dependencies among X and Y, and Y and Z.

Now that we have concluded that X and Y are dependent on each other, and Y and Z are dependent on each other, we would like to understand how and by how much these variables are dependent on each other. In other words, we want to know how exercise affects condition status, as well as how drug type affects condition status. We will estimate odds ratios by exponentiating the interaction terms. For X and Y, we will get e ^(λ_{11}^{XY}) = e^(.4137099) = 1.512. Since X=1 corresponds to Rarely, and Y=1 corresponds to No Improvement, we will interpret this as "the odds of rarely exercising and not showing improvement are 1.512 times the odds of exercising rarely and showing improvement on condition". For every 10 people who rarely exercise and show improvement, we expect roughly 15 people who rarely exercise to show <u>no</u> improvement. For Y and Z, we estimate the odds ratio as e^(λ_{11}^{yz}) = e^(.8483414) = 2.336. The odds of showing no improvement on condition and being in group 2 are 2.336 times the odds of showing no improvement and not being in group 2. For every 10 people in group 1 who show no improvement, we expect roughly 24 people in group 2 to show no improvement.

The 95% Wald confidence interval for λ_{11}^{XY} is: (e^.0514, e^.7761) = (1.0527, 2.173). We are 95% confident that the odds of rarely exercising and not showing improvement are between 1.0527 and 2.173 times the odds of exercising rarely and showing improvement on condition. The 95% Wald CI for λ_{11}^{yz} is: (e^.4798, e^1.2169) = (1.6158, 3.3767). We are 95% confident that the odds of showing no improvement on condition and being in group 2 are between 1.6158 and 3.3767 times the odds of showing no improvement and not being in group 2.

The 95% Likelihood Ratio confidence interval for λ_{11}^{XY} (e^.0513, e^.7767) = (1.0526, 2.1743). The LR confidence interval for λ_{11}^{yz} is (e^.4827, e^1.2204) = (1.6204, 3.3885). We would interpret this interval the same way we interpreted the Wald CI above.

<u>Conclusion:</u>

We used hypothesis testing, confidence intervals, and model selection in order to see which categorical variables (X, Y, and Z) are dependent on each other. To see how much these variables depended on each other, we calculate odds ratios. We found that X and Y, and Y and Z, are dependent on each other. Furthermore, we found that the odds of X=rarely and Y=no improvement are greater than the odds of X=rarely and Y=improvement. We also found that the odds of Y=no improvement and Z=group 2 are greater than the odds of Y=no improvement and Z=group 1.

Code Appendix:

Part1:

```
car = read.csv("/Users/rurikoimai/Downloads/Car.csv")
hist(car$X, xlab = "Age of the customer's current car (in months)", main = "Car Dealership")
table(car$Y) #16 0s, 17 0s
logit.model =  glm(formula = Y ~ X, family = binomial(logit), data = car)
summary(logit.model)
logit.model
exp(0.07635)
alpha = -2.65826
beta = 0.07635
-alpha/beta
plot(car$X, logit.model$fitted.values, xlab = "Age of the customer's current car (in months)",
ylab = "Estimated Probability", main = "Prediction Curve")
curve(predict(logit.model, data.frame(X=x), type = "response"), add = TRUE)
wald.ts = summary(logit.model)$coefficients[2,3]
wald.ts
wald.pval = round(summary(logit.model)$coefficients[2,4],8)
wald.pval
smaller.model = glm(Y ~ 1, family = binomial(logit), data = car)
smaller.model
LR.ts = as.numeric(-2*(logLik(smaller.model) - logLik(logit.model)))
LR.ts
d.f. = length(logit.model$coefficients) - length(smaller.model$coefficients)
d.f.
LR.pval = round(pchisq(LR.ts, d.f., lower.tail = FALSE), 8)
LR.pval
CI = round(exp(as.numeric(confint(logit.model)[2,], level = 0.99)), 4)
CI
```

Part 2:

```
trial=read.csv("TrialsShort.csv",header=TRUE)
trial
library(MASS)

# Summary

trial
```

```
summary(trial)
head(trial)

trialslong=read.csv("TrialsLong.csv",header=TRUE)
trialslong
head(trialslong)

Trials = table(trialslong$X,trialslong$Y,trialslong$Z)
Trials

mosaicplot(Trials)

# Analysis

names(trial) = c("X","Y","Z","F")
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 = trial, family = poisson)
})
all.model.fits

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)
}
round(good.fit.LL(Model),4)
```

```
Model = glm(F ~ X+Y+Z, data=trial,family = poisson)
names(Model)

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)
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

all.model.fits[[6]]$coefficients
all.model.fits[[6]]$fitted.values

# Model with lowest AIC: F~ X + Y + Z + XY + YZ
```