Dan Forbush PA541 Homework 3

Dan Forbush

4/18/2019

## R-Code Setup

#set WD:  
setwd("~/Documents/UIC/PA 541 Data Analysis/Homework 3")  
  
#load libraries:  
library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 2.0.1 ✔ dplyr 0.7.8  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.3.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

options(scipen = 999)  
  
#load data:  
titanic <- read.csv("~/Documents/UIC/PA 541 Data Analysis/Homework 3/TitanicData.csv")

# Part 1

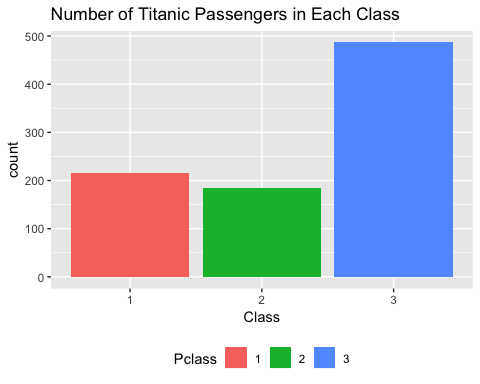
## Question 1

**(i) The variables names regarding siblings/spouses and parents/children are long and not in accordance with good practice (spaces in the names). Update these to the following: ‘num\_sib\_spouse’ and ‘num\_par\_child’. Also, change Pclass to a factor variable.**

titanic$num\_sib\_spouse <- titanic$Siblings.Spouses.Aboard  
titanic$num\_par\_child <- titanic$Parents.Children.Aboard  
titanic$Siblings.Spouses.Aboard <- titanic$Parents.Children.Aboard <- NULL  
  
titanic$Pclass <- factor(titanic$Pclass, levels = c("1", "2", "3"), ordered=FALSE)

**(ii) Create a column chart (hint: use geom\_col()) to show the count of the number of people in each class.**

ggplot(titanic, aes(x=Pclass)) +  
 geom\_bar(aes(fill = Pclass)) +  
 labs(title = "Number of Titanic Passengers in Each Class", x = "Class") +  
 theme(legend.position="bottom")



**(iii) What percentage of females survived? What percentage of males? Write code to find these values and then run a t-test to see if the difference is significant. What do you conclude?**

titanic %>% group\_by(Sex) %>% summarise(mean(Survived)\*100)

## # A tibble: 2 x 2  
## Sex `mean(Survived) \* 100`  
## <fct> <dbl>  
## 1 female 74.2  
## 2 male 19.0

74% of female passengers survived, while 19% of male passengers survived.

t.test(Survived~Sex, data = titanic)

##   
## Welch Two Sample t-test  
##   
## data: Survived by Sex  
## t = 18.592, df = 587.06, p-value < 0.00000000000000022  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.4935209 0.6101018  
## sample estimates:  
## mean in group female mean in group male   
## 0.7420382 0.1902269

These results tell us with a high degree of confidence (because the p value is very small) that there is a statistically significant difference in the mean survival due to sex. Because the mean is higher for females than males, that means that female passengers were significantly more likely to survive the titanic disaster.

**(iv) What was the average price paid for a ticket in each of the three classes?**

titanic %>% group\_by(Pclass) %>% summarise(mean(Fare))

## # A tibble: 3 x 2  
## Pclass `mean(Fare)`  
## <fct> <dbl>  
## 1 1 84.2  
## 2 2 20.7  
## 3 3 13.7

People in 1st class paid an average of $84.15 per ticket (assuming the amounts are in dollars), the people in 2nd class paid an average of $20.66 per ticket, and the people in 3rd class paid an average of $13.71 per ticket.

## Question 2

**In keeping with the sad theme of gender inequality in other datasets we have used, let’s examine how one’s age and gender affect the fare paid to board the Titanic. Run an OLS model predicting Fare by Sex and Age. Interpret the coefficients on Sex and Age.**

mod1 <- lm(Fare ~ Sex + Age, data = titanic)  
summary(mod1)

##   
## Call:  
## lm(formula = Fare ~ Sex + Age, data = titanic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -52.15 -21.08 -13.10 2.05 484.60   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.7700 4.2275 7.515 0.000000000000139 \*\*\*  
## Sexmale -20.0892 3.4259 -5.864 0.000000006384797 \*\*\*  
## Age 0.4585 0.1161 3.950 0.000084360630293 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 48.59 on 884 degrees of freedom  
## Multiple R-squared: 0.04959, Adjusted R-squared: 0.04744   
## F-statistic: 23.06 on 2 and 884 DF, p-value: 0.0000000001728

This model tells us that, holding age constant, men paid $20.09 less on average than women for their fare. Additionally, for each additional year in age, passengers paid $.46 more for their fare. The p-values on both of these coefficients are significant at the .001 level, so it is not likely that these values are due to chance alone. However, we can’t rule out endogeniety or omitted variable bias based on these results.

## Question 3

**Test whether the assumption of homoskedasticy has been met. Discuss results. Calculate the VIF for each variable. Should we be concerned with multicollinearity. (Note, the vif() command is in the ‘car’ package. You can also calculate the VIF yourself as we did in class)**

#I am assuming we are looking only at homoskedasticity for Age.   
#For that, I'll perform a Breusch-Pagan test for mod1.  
#Step 1 - add residiuals and squared residuals to the dataset  
titanic$mod1res <- resid(mod1)  
titanic$mod1ressq <- (titanic$mod1res ^ 2)  
  
#Step 2 - run a regression of the squared residuals on Age:  
mod1BPtest <- lm(mod1ressq~Age, data = titanic)  
summary(mod1BPtest)

##   
## Call:  
## lm(formula = mod1ressq ~ Age, data = titanic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3908 -2218 -1862 -1273 232278   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1248.81 1120.22 1.115 0.265  
## Age 37.46 34.28 1.093 0.275  
##   
## Residual standard error: 14410 on 885 degrees of freedom  
## Multiple R-squared: 0.001347, Adjusted R-squared: 0.0002187   
## F-statistic: 1.194 on 1 and 885 DF, p-value: 0.2749

Based on the Breusch-Pagan test run above, we should not be concerned with heteroskedasticity. The value on the Age coefficient in the mod1BPtest model is not statistically significant at any level. That means we cannot reject the null hypothesis that the true coefficient on Age is 0 in this regression that predicts the squared residuals of mod1 by Age. Therefore, the assumption of homoskedasticity has been met. We don’t have evidence that the size of the squared residuals changes as Age increases.

#Next, calculate the VIF of mod1 to examine possible multicollinearity  
vif(mod1)

## Sex Age   
## 1.008513 1.008513

The VIF is the is the variance inflation factor, which is calculated as , where is the r-squared of one predictor in the model on the other, i.e. how much one predictor predicts the value of the other. The higher the VIF value, the greater the collinearity. Here, the VIF value is very close to 1, indicating that a very small amount of variance in one variable is predicted by the other. Therefore, multicollinearity is not an issue and we have met the corresponding linear regression assumption.

## Question 4

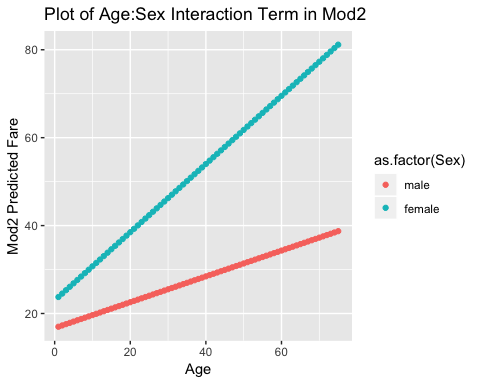
**Create an interaction between Sex and Age and rerun the model predicting Fare. Assume that Sex moderates the effect of Age in your interpretation of the interaction. (i) Interpret the results on each coefficient. (ii) Create a plot to visualize the interaction**

#First, create the model  
mod2 <- lm(Fare ~ Sex + Age + Sex:Age, data = titanic)  
summary(mod2)

##   
## Call:  
## lm(formula = Fare ~ Sex + Age + Sex:Age, data = titanic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -63.02 -20.17 -14.23 2.23 485.35   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.9781 6.1379 3.744 0.000193 \*\*\*  
## Sexmale -6.2821 7.7897 -0.806 0.420191   
## Age 0.7757 0.1982 3.914 0.0000978 \*\*\*  
## Sexmale:Age -0.4820 0.2443 -1.973 0.048827 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 48.51 on 883 degrees of freedom  
## Multiple R-squared: 0.05376, Adjusted R-squared: 0.05054   
## F-statistic: 16.72 on 3 and 883 DF, p-value: 0.0000000001426

The coefficient on Sex (that’s not part of the interaction term) tells us the simple main effect of sex. It says that fare will be $6.28 less for men than women, but only at age 0. The coefficient on Age tells us that, for women, with each additional year in age, fare will increase by $.78. The coefficient on the interaction term tells us that, when the Sex is male, the coefficient on Age will be $.48 less than the coefficient when Sex is Female. In other words, for males only, with each additional year in Age, Fare will increase by $.30 (= $.78 - $.48).

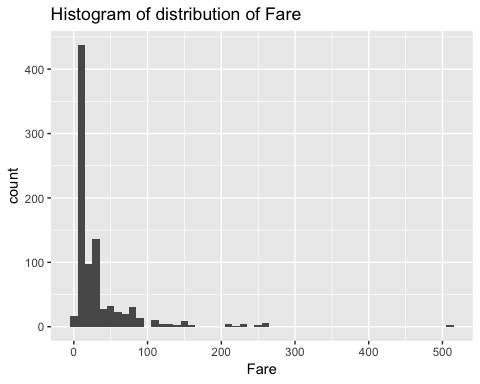
#Below is the code to create a plot of this effect:  
newdata = expand.grid(Sex = c("male", "female"), Age = c(1:75))  
newdata$preds <- predict(mod2, new = newdata, type = "response")  
ggplot(newdata, aes(x=Age, y = preds, group = Sex)) +   
 geom\_point(aes(colour=as.factor(Sex))) +  
 labs(title = "Plot of Age:Sex Interaction Term in Mod2", y = "Mod2 Predicted Fare")



## Question 5

**Create a histogram of the variable Fare. Notice the long right tail and the bunching of the data on the left.**

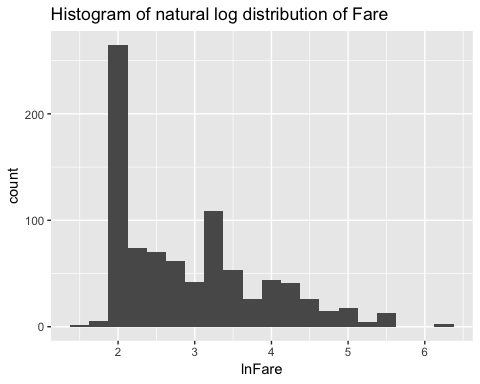
ggplot(titanic, aes(Fare)) + geom\_histogram(binwidth = 10) +  
 labs(title = "Histogram of distribution of Fare")



**Let’s log transform (using the natural log) the variable Fare. Note, there are a few fares with the value of zero. The log of zero is not defined. A common approach to deal with this issue when wanting to log transform a variable is to add a small number to those observations before taking the log. For example, add the value of 1 to each fare. However, in this case, there are very few observations with a fare of zero. Because of this, let’s just drop them from the analysis. (If you were doing this work for real, you would want to try to understand why they were zero. Were they all babies? And then see how sensitive your results are to including or not includign those observations in your analysis. For now, let’s just drop those observations)**

**Create a new datset that only includes non-zero fares. Once you remove the observations with a fare of zero, log transform Fare and a create a histogram of the new variable. Notice the change in the distribution.**

titanic2 <- titanic[titanic$Fare != 0,]  
titanic2$lnFare <- log(titanic2$Fare)  
  
ggplot(titanic2, aes(lnFare)) + geom\_histogram(binwidth = .25) +  
 labs(title = "Histogram of natural log distribution of Fare")



**Create a model predicting the log transformed version of Fare by Age. Interpret the coefficient on Age.**

mod3 <- lm(lnFare ~ Age, data = titanic2)  
summary(mod3)

##   
## Call:  
## lm(formula = lnFare ~ Age, data = titanic2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.4637 -0.8123 -0.2251 0.5595 3.2371   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.654891 0.072345 36.698 < 0.0000000000000002 \*\*\*  
## Age 0.009913 0.002214 4.477 0.00000859 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9271 on 870 degrees of freedom  
## Multiple R-squared: 0.02252, Adjusted R-squared: 0.02139   
## F-statistic: 20.04 on 1 and 870 DF, p-value: 0.000008589

Using natural logs in regression functions acts as an approximation of percentage change in a variable. In mod3, the dependent variable has been log-transformed but the indepdendent variable has not been log transformed. Thus, this is a log-lin model. The coefficient on this model tells us that with each additional year of age, there is an associated 1% increase in Fare. This coefficient is statistically significant at the .001 level.

# Part 2

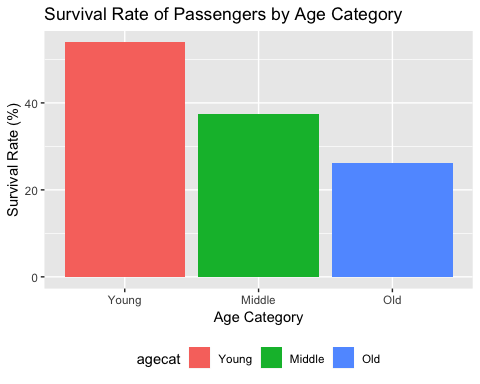
## Question 6

**We are going to try to understand more about who survived on the Titanic and the factors associated with survival. Let’s look at the survival rate based on age. (i) Create a new variable that breaks age into three categories: young, middle, and old. Create the breaks as follows: 0 - 14, 15 - 55, and 56+. (ii) Once you have created this variable, create a column chart for the survival rate for each age group (hint: use geom\_col() ).**

titanic3 = titanic %>%   
 mutate(agecat = case\_when(Age > 55 ~ 'Old',  
 Age >= 15 & Age <= 55 ~ 'Middle',  
 Age < 15 ~ 'Young'))  
titanic3$agecat <- factor(titanic3$agecat, levels = c("Young", "Middle", "Old"))  
  
q6df <- titanic3 %>% group\_by(agecat) %>% summarise(AgeCatSurvival = mean(Survived)\*100)  
q6df

## # A tibble: 3 x 2  
## agecat AgeCatSurvival  
## <fct> <dbl>  
## 1 Young 53.9  
## 2 Middle 37.5  
## 3 Old 26.1

ggplot(q6df, aes(x=agecat, y = AgeCatSurvival)) +  
 geom\_col(aes(fill = agecat)) +  
 labs(title = "Survival Rate of Passengers by Age Category", x = "Age Category", y = "Survival Rate (%)") +  
 theme(legend.position="bottom")



## Question 7

**Run a linear probability model to predict whether a passenger survives based on Age, Sex, and Pclass. Interpret the coefficients along with the intercept.**

mod4 <- lm(Survived ~ Age + Sex + Pclass, data = titanic3)  
summary(mod4)

##   
## Call:  
## lm(formula = Survived ~ Age + Sex + Pclass, data = titanic3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.09191 -0.23675 -0.07911 0.22250 1.00040   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.101854 0.047486 23.204 < 0.0000000000000002 \*\*\*  
## Age -0.004970 0.001005 -4.944 0.000000915 \*\*\*  
## Sexmale -0.494781 0.027494 -17.996 < 0.0000000000000002 \*\*\*  
## Pclass2 -0.190178 0.039514 -4.813 0.000001749 \*\*\*  
## Pclass3 -0.383843 0.034638 -11.082 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3833 on 882 degrees of freedom  
## Multiple R-squared: 0.3832, Adjusted R-squared: 0.3804   
## F-statistic: 137 on 4 and 882 DF, p-value: < 0.00000000000000022

The intercept tells us that the base category survival - a female, age 0, who is in 1st class - has a 110% chance of surviving the titanic. This is out of the bounds of possibility because it is greater than 100%. But all the other variables decrease the chance of survival.

The coefficient on Age tells us that, for each additional year of Age, the chance of survival decreases by .5%, holding all other variables constant. The coefficient on Sexmale tells us that, if the passenger is male, the chance of survival decreases nearly 50%, holding all other variables constant. If the passenger is in 2nd class, the chance of survival decreases by 19%, holding all other variables constant. And if the passenger is in 3rd class, the chance of survival decreases 38%, holding all other variables constant.

These coefficients are all statistically significant at the .001 level, as shown by the p-values.

## Question 8

**Now run a logistic regression using the same three predictors as you did in the LPM above. Interpret the coefficients on Age and Sex.**

mod5 <- glm(Survived ~ Age + Sex + Pclass, data = titanic3, family=binomial(link = "logit"))  
summary(mod5)

##   
## Call:  
## glm(formula = Survived ~ Age + Sex + Pclass, family = binomial(link = "logit"),   
## data = titanic3)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6811 -0.6653 -0.4137 0.6367 2.4505   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.63492 0.37045 9.812 < 0.0000000000000002 \*\*\*  
## Age -0.03427 0.00716 -4.787 0.00000169 \*\*\*  
## Sexmale -2.58872 0.18701 -13.843 < 0.0000000000000002 \*\*\*  
## Pclass2 -1.19911 0.26158 -4.584 0.00000456 \*\*\*  
## Pclass3 -2.45544 0.25322 -9.697 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1182.77 on 886 degrees of freedom  
## Residual deviance: 801.59 on 882 degrees of freedom  
## AIC: 811.59  
##   
## Number of Fisher Scoring iterations: 5

#Exponentiate the coefficients to make it easier to interpret  
exp(mod5$coefficients)

## (Intercept) Age Sexmale Pclass2 Pclass3   
## 37.8988400 0.9663058 0.0751161 0.3014609 0.0858252

The coefficients that we get from a logistic regression are the log odds of survival for a one unit change in that variable. This is difficult to interpret, so I’ve exponentiated the beta coefficients from the logit to make it easier to interpret.

The exponentiated intercept tells us that, for the base category - a person age 0 who is female in 1st class, the odds of survival are nearly 38:1.

The exponentiated coefficient on age tell us that the odds change by .9663058 for each one unit increase in age. Thus, with each additional year in age, you multiply the exponentiated intercept by .966. For instance, someone who differs from the base category by one year would have a 37.89884 x .9663058 = 36.6:1 chance of survival; a person who differs from the base category by two years would have a 37.89884 x (.9663058)^2 = 35:1 chance of survival.

For sex, if a person is male, you multiply the exponentiated intercept times .0751161 to get the odds ratio of that person surviving. So, a person age 0 in first class who is male has a 37.8988400 \* .0751161 = 2.8:1 chance of survival.

## Question 9

**Using the output you produced in question 8, calculate the predicted probability of survival for (i) a female in first class who is 35 years old, (ii) a male in 3rd class who is 10 years old, (iii) a male in 1st class who is ten years old. Please show your work. Hint: you will want to calculate the total logit value first (just as you calculate any predicted value for any other regression equation) and then using that value calculate the probability associated with it. See class slides and the excel file for an example.**

The general formula for finding the odds of survival is as follows:

For part i:

#save the coefficients  
B0 <- mod5[["coefficients"]][["(Intercept)"]]  
Bage <- mod5[["coefficients"]][["Age"]]  
Bsexmale <- mod5[["coefficients"]][["Sexmale"]]  
BPclass2 <- mod5[["coefficients"]][["Pclass2"]]  
BPclass3 <- mod5[["coefficients"]][["Pclass3"]]  
  
#calculate the odds ratio for a female in 1st class who is 35 years old:  
q9iodds <- exp(B0 + (Bage \* 35))  
q9iodds

## [1] 11.4192

#calculate the probability  
q9iodds/(1+q9iodds)

## [1] 0.9194795

A female age 35 in 1st class has a 92% chance of survival. To calculate the probability, first I saved the coefficients from the model. Then, I calculated the new logged odds ratio by adding the relevant coefficients - here, that is just . Then, I exponentiated the logged odds ratio to get the odds ratio. Finally, I got the probability by following the formula of .

For part ii:

#calculate the odds ratio for a female in 1st class who is 35 years old:  
q9iiodds <- exp(B0 + (Bage \* 10) + Bsexmale + BPclass3)  
q9iiodds

## [1] 0.1734281

#calculate the probability  
q9iiodds/(1+q9iiodds)

## [1] 0.1477961

A 10 year old male in 3rd class has a 14% chance or survival. Here, to get the logged odds ratio, the relevant coefficients are . The rest of the calculation was the same as in part i.

For part iii:

#calculate the odds ratio for a female in 1st class who is 35 years old:  
q9iiiodds <- exp(B0 + (Bage \* 10) + Bsexmale)  
q9iiiodds

## [1] 2.020713

#calculate the probability  
q9iiiodds/(1+q9iiiodds)

## [1] 0.6689523

A 10 year old male in 1st class has a 67% chance or survival. Here, to get the logged odds ratio, the relevant coefficeints are . The rest of the calculation was the same as in part i.

## Bonus

**BONUS - Create a plot/visualization that shows the effect of Age on the probability of Survival holding all other predictors constant.**

#make sample data  
newdata2 = expand.grid(Age = seq(from = 1, to = 90, by = 1), Sex="female", Pclass = "1")  
  
#predict the results  
prednew <- predict(mod5, newdata = newdata2, type = "response")  
  
ggplot(data=newdata2, aes(x=Age, y=prednew\*100)) + geom\_point() +  
 labs(title = "Probability of Female in 1st Class Surviving by Age",   
 x = "Age", y = "Survival Probability (%)")

