Dan Forbush PA541 Homework 2

Dan Forbush

3/12/2019

## R-Code Setup

#set WD:  
setwd("~/Documents/UIC/PA 541 Data Analysis/Homework 2")  
  
#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)  
  
options(scipen = 999)  
  
#load data:  
cell <- read.csv("~/Documents/UIC/PA 541 Data Analysis/Homework 2/cell.csv")  
fines <- read.csv("~/Documents/UIC/PA 541 Data Analysis/Homework 2/fines.csv")  
  
#converting cell\_ban and text\_ban to factors is useful for many problems:  
cell$cell\_ban <- as.factor(cell$cell\_ban)  
cell$text\_ban <- as.factor(cell$text\_ban)  
  
#same with converting Female, Hispanic, and Black in fines:   
fines$Female <- as.factor(fines$Female)  
fines$Hispanic <- as.factor(fines$Hispanic)  
fines$Black <- as.factor(fines$Black)

# Part 1

## Question 1

**(a) Of the states with a texting ban, which one has the highest number of traffic deaths?**

cell %>%  
 select(state, cell\_ban, numberofdeaths)%>%  
 filter(cell\_ban==1) %>%  
 arrange((desc(numberofdeaths)))

## state cell\_ban numberofdeaths  
## 1 California 1 2857  
## 2 New York 1 1168  
## 3 New Jersey 1 589  
## 4 Maryland 1 505  
## 5 Washington 1 444  
## 6 Oregon 1 336  
## 7 Nevada 1 258  
## 8 Connecticut 1 236  
## 9 Delaware 1 114  
## 10 New Hampshire 1 108

California has the highest number of deaths among the states that have a cell phone ban.

**(b) How many states have a cell ban, how many have a texting ban, and how many have both?**

table(cell$cell\_ban == 1)

##   
## FALSE TRUE   
## 40 10

table(cell$text\_ban == 1)

##   
## FALSE TRUE   
## 16 34

table(cell$cell\_ban == 1 & cell$text\_ban == 1)

##   
## FALSE TRUE   
## 40 10

10 states have a cell phone ban, 34 states have a text ban, and 10 states have both a cell phone ban and a text ban.

**(c) What is the average number of deaths in states with a texting ban versus those without?**

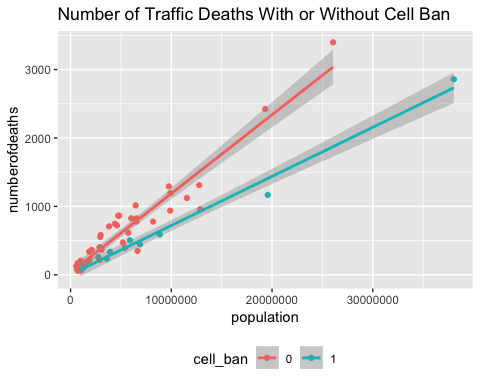
cell %>%  
 group\_by(text\_ban) %>%  
 summarise(mean.deaths = mean(numberofdeaths))

## # A tibble: 2 x 2  
## text\_ban mean.deaths  
## <fct> <dbl>  
## 1 0 892.  
## 2 1 567.

The mean number of traffic deaths is 892.25 for states that don’t have a text ban. The mean number of traffic deaths is 566.76 for states that do have a text ban.

**(d) Plot the relationship between population and number of traffic deaths for those with and without a cell ban on the same plot.**

ggplot(cell, aes(x=population, y=numberofdeaths, group=cell\_ban)) +  
 geom\_point(aes(color=cell\_ban)) +  
 geom\_smooth(method="lm", aes(color=cell\_ban)) +  
 labs(title = "Number of Traffic Deaths With or Without Cell Ban") +  
 theme(legend.position="bottom")



## Question 2

**Although we don’t know how many people are using their phones while driving, we can find the number of cell phone subscriptions in a state (in thousands). Estimate a model with traffic deaths as the dependent variable and number of cell phone subscriptions and population as the independent variables. Interpret the slopes and intercept.**

mod1 <- lm(numberofdeaths~cell\_subscription + population, data = cell)  
summary(mod1)

##   
## Call:  
## lm(formula = numberofdeaths ~ cell\_subscription + population,   
## data = cell)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -811.0 -128.7 -47.8 138.5 882.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 113.93458586 49.91416894 2.283 0.02702 \*   
## cell\_subscription -0.21088577 0.09230331 -2.285 0.02689 \*   
## population 0.00029087 0.00008873 3.278 0.00197 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 261.7 on 47 degrees of freedom  
## Multiple R-squared: 0.8562, Adjusted R-squared: 0.8501   
## F-statistic: 139.9 on 2 and 47 DF, p-value: < 0.00000000000000022

The intercept tells us that, if the number of cellphone subscriptions and population were both equal to 0, we would expect 113.93 traffic deaths. The slope term on cell subscriptions tells us that with each additional 1,000 cell phone subscribers in a state, Model 1 predicts .21 fewer deaths. The slope term on population tells us that with each additional 1,000 people, the model predicts .000290 more traffic deaths.

## Question 3

**Now add total miles driven to the model. What happens to the coefficient on cell phone subscriptions? Why?**

mod2 <- lm(numberofdeaths~cell\_subscription + population + total\_miles\_driven, data = cell)  
summary(mod2)

##   
## Call:  
## lm(formula = numberofdeaths ~ cell\_subscription + population +   
## total\_miles\_driven, data = cell)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -555.98 -92.75 -12.18 60.67 788.37   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.34594956 41.07541922 0.106 0.916   
## cell\_subscription 0.00246485 0.07670690 0.032 0.975   
## population -0.00007422 0.00008821 -0.841 0.404   
## total\_miles\_driven 0.01883208 0.00301818 6.240 0.000000127 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 194.7 on 46 degrees of freedom  
## Multiple R-squared: 0.9221, Adjusted R-squared: 0.917   
## F-statistic: 181.5 on 3 and 46 DF, p-value: < 0.00000000000000022

When total miles driven was added in as predictor, the cell subscription variable went from -.21 to +.002. This is an example of omitted variable bias. There was a downward bias because the true coefficient (.002) is higher than the observed coefficient without total miles driven (-.21).

## Question 4

**Add the dummy variables for cell phone bans and texting bans to the model from question 3. Interpret the coefficients on these dummy variables.**

#converting cell\_ban and   
mod3 <- lm(numberofdeaths~cell\_subscription + population + total\_miles\_driven + cell\_ban + text\_ban, data = cell)  
summary(mod3)

##   
## Call:  
## lm(formula = numberofdeaths ~ cell\_subscription + population +   
## total\_miles\_driven + cell\_ban + text\_ban, data = cell)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -434.62 -91.65 -5.73 91.54 681.74   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 150.09798799 54.70507784 2.744 0.00876 \*\*   
## cell\_subscription 0.02733338 0.06877471 0.397 0.69297   
## population -0.00007226 0.00007952 -0.909 0.36845   
## total\_miles\_driven 0.01588496 0.00293226 5.417 0.0000024 \*\*\*  
## cell\_ban1 -100.78962448 73.32850317 -1.374 0.17625   
## text\_ban1 -165.23438232 56.94955795 -2.901 0.00578 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 173.5 on 44 degrees of freedom  
## Multiple R-squared: 0.9409, Adjusted R-squared: 0.9342   
## F-statistic: 140 on 5 and 44 DF, p-value: < 0.00000000000000022

The cell phone ban and text ban are both categorial variables, meaning they are either present or absent. For each state that has a text ban, approximately 165 fewer traffic deaths are predicted. The model tells us that for each state that has a cell phone ban, approximately 101 fewer traffic deaths are predicted on average. However, it should be noted that the p-value on the cell ban is .176 so we cannot reject the null hypothesis that the true effect of a cell phone ban on traffic deaths is zero.

## Question 5

**Based on the results from question 4, how many lives would be saved if California banned cell phone use while driving? How many lives would be saved if Wyoming had such a ban? Discuss implications for the proper specification of the model.**

There are no interaction effects in Model 3 in question 4, so it is not specific to situations where a cell phone ban may be more or less effective. Therefore, our model predicts that both Wyoming and California would see the same benefit from a cell phone ban: approximately 101 fewer traffic deaths. Because cell phone ban is a categorical variable, it shifts the linear model by a constant amount regardless of the values of any of the continuous variables.

## Question 6

**Estimate a model in which total miles is interacted with both the cell phone ban and the prohibition of texting variables. What is the estimated effect of a cell phone ban for California? For Wyoming? What is the effect of a texting ban for California? For Wyoming? What is the effect of total miles?**

mod4 <- lm(numberofdeaths~cell\_subscription + population + total\_miles\_driven + cell\_ban + text\_ban + cell\_ban\*total\_miles\_driven + text\_ban\*total\_miles\_driven, data = cell)  
summary(mod4)

##   
## Call:  
## lm(formula = numberofdeaths ~ cell\_subscription + population +   
## total\_miles\_driven + cell\_ban + text\_ban + cell\_ban \* total\_miles\_driven +   
## text\_ban \* total\_miles\_driven, data = cell)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -369.40 -48.11 -9.84 44.28 247.11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 8.84526915 47.04605654 0.188 0.85177  
## cell\_subscription -0.02199976 0.05476504 -0.402 0.68993  
## population 0.00001217 0.00006627 0.184 0.85519  
## total\_miles\_driven 0.01416978 0.00245091 5.781 0.000000819  
## cell\_ban1 -20.08987974 72.03140219 -0.279 0.78169  
## text\_ban1 21.83219016 65.71372430 0.332 0.74137  
## total\_miles\_driven:cell\_ban1 -0.00146219 0.00105064 -1.392 0.17134  
## total\_miles\_driven:text\_ban1 -0.00292451 0.00097184 -3.009 0.00441  
##   
## (Intercept)   
## cell\_subscription   
## population   
## total\_miles\_driven \*\*\*  
## cell\_ban1   
## text\_ban1   
## total\_miles\_driven:cell\_ban1   
## total\_miles\_driven:text\_ban1 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 127.7 on 42 degrees of freedom  
## Multiple R-squared: 0.9694, Adjusted R-squared: 0.9643   
## F-statistic: 190 on 7 and 42 DF, p-value: < 0.00000000000000022

Model 4 tells us that the coefficient on total miles driven is moderated by the presence of a cell phone ban and/or text ban. The coefficient on total miles driven says that with each million miles driven, .0142 traffic deaths will result. However, if a cell phone ban is present, then that coefficient is reduced by about 0.0015 to about .0127 traffic deaths per million total miles. Likewise with the text ban, if it is present, the coefficient on total miles driven will decrease by 0.0029 to .0113 deaths per million miles driven.

newdata1 <- cell %>% filter(state == "California")  
mod4$coefficients[5] + (mod4$coefficients[7]\*newdata1$total\_miles\_driven)

## cell\_ban1   
## -497.1601

mod4$coefficients[6] + (mod4$coefficients[8]\*newdata1$total\_miles\_driven)

## text\_ban1   
## -932.351

newdata1$total\_miles\_driven

## [1] 326271.7

Our model predicts that a cell phone ban in California will save approximately 497 traffic deaths per year. It does this in part because of the cell phone ban alone which, according to the model, saves about 20 lives per year. The other part is because of the interaction between the cell phone ban and total miles driven, which is calculated as the total miles driven in California (326,271 million) times the interaction coefficient (.00146) and equals about 477 lives saved per year.

In California, the text ban saves approximately 932 lives per year. This is in part because of the text ban alone, the presence of which *adds* 21.83 deaths per year, and in part because of the interaction between the text ban and total miles driven in California, which saves 954 lives per year.

newdata2 <- cell %>% filter(state == "Wyoming")  
mod4$coefficients[5] + (mod4$coefficients[7]\*newdata2$total\_miles\_driven)

## cell\_ban1   
## -33.64581

mod4$coefficients[6] + (mod4$coefficients[8]\*newdata2$total\_miles\_driven)

## text\_ban1   
## -5.280879

newdata2$total\_miles\_driven

## [1] 9270.994

In Wyoming, by contrast, the cell phone ban is predicted to save 33.65 lives per year and the text ban is predicted to save 5.28 lives per year. The calculation is the same as for California, but the lives saved is smaller because the total miles driven in Wyoming (9,271 million) is much smaller than that of California.

# Part 2

## Question 7

**Write out a model (in notation similar to that which we use in class or the Wooldridge text; in other words write out the regression model) that predicts the size of the fine based on Age, MPHover, Female, Hispanic, and Black. You can use the Microsoft word equation editor or simply enter the model using regular text in word. Given the model and how the variables are defined in the dataset, what is the base group? Write out the condition expectation for a Female Hispanic. Write out the conditional expectation for a Black Male.**

The linear model equation, with Fine Amount represented by , would be:

This linear model contains two continuous variables: Age and MPHover, which are represented with Beta coefficients. The dummy variables include Female, Hispanic, and Black, coded with alpha coefficients. The data set defines 0 = male, non-hispanic, and non-black (respectively), so those three characteristics would be the base group.

The conditional expectation for a female hispanic person is:

The conditional expectation for a black male is:

## Question 8

**Run the model discussed above. Interpret the coefficients on Female, Hispanic, and Black. Look at standard errors on coefficients for the Female, Hispanic, and Black variables. Why they are different?**

mod5 <- lm(Amount ~ Age + MPHover + Female + Hispanic + Black, data = fines)  
summary(mod5)

##   
## Call:  
## lm(formula = Amount ~ Age + MPHover + Female + Hispanic + Black,   
## data = fines)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -309.477 -20.231 6.834 25.457 225.559   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.84512 0.97546 4.967 0.0000006834222092 \*\*\*  
## Age 0.02682 0.01761 1.523 0.1278   
## MPHover 6.87850 0.03874 177.547 < 0.0000000000000002 \*\*\*  
## Female1 -3.54327 0.47457 -7.466 0.0000000000000847 \*\*\*  
## Hispanic1 1.93201 1.05784 1.826 0.0678 .   
## Black1 -2.02930 1.01789 -1.994 0.0462 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 39.63 on 31668 degrees of freedom  
## (36683 observations deleted due to missingness)  
## Multiple R-squared: 0.5036, Adjusted R-squared: 0.5035   
## F-statistic: 6426 on 5 and 31668 DF, p-value: < 0.00000000000000022

The coefficients from Model 5 tell us that, on average, the fine amount will decrease by $3.54 if the driver is female; will increase by $1.93 if the driver is hispanic; and will decrease by $2.03 if the driver is black.

The standard error for each coefficient is an indicator of the variability within the sample. It is calculated based on the variance within the sample and can be used to determine if the coefficient value is significantly different from 0. Comparatively, the standard error on the Female variable is smaller than the standard errors on the Hispanic or Black variables (both of which are about equal). This means the model is more confident that the coefficient on female is statistically significant not likely to be 0 if the dataset was resampled.

On the other hand, the standard errors for the Hispanic and Black variables are on the border of statistical significance at the 0.05 level: Black has a p-value just under 0.05, while Hispanic has a p-value just over 0.05. This indicates that we can be somewhat confident that the coefficients on these variables are significantly different from 0, but not as confident as we are for the coefficient on the Female variable.

## Question 9

**The model above implicitly assumes the effect of race is the same for both men and women. Test whether this assumption is true and briefly discuss your results (i.e., tell my whether the assumption is true or not).**

In order to test the assumption that the effect of race is the same for both women and men, we can add two variables to Model 5 where the Female variable interacts with the Hispanic and Black variables:

mod6 <- lm(Amount ~ Age + MPHover + Female + Hispanic + Black + Female\*Hispanic + Female\*Black, data = fines)  
summary(mod6)

##   
## Call:  
## lm(formula = Amount ~ Age + MPHover + Female + Hispanic + Black +   
## Female \* Hispanic + Female \* Black, data = fines)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -309.531 -20.423 6.803 25.452 225.504   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.90573 0.97683 5.022 0.0000005138533725 \*\*\*  
## Age 0.02679 0.01761 1.521 0.128   
## MPHover 6.87841 0.03874 177.545 < 0.0000000000000002 \*\*\*  
## Female1 -3.71327 0.49495 -7.502 0.0000000000000643 \*\*\*  
## Hispanic1 1.84582 1.20206 1.536 0.125   
## Black1 -2.97691 1.18506 -2.512 0.012 \*   
## Female1:Hispanic1 0.29703 2.52501 0.118 0.906   
## Female1:Black1 3.60666 2.31052 1.561 0.119   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 39.63 on 31666 degrees of freedom  
## (36683 observations deleted due to missingness)  
## Multiple R-squared: 0.5037, Adjusted R-squared: 0.5035   
## F-statistic: 4590 on 7 and 31666 DF, p-value: < 0.00000000000000022

The addition of the moderating variables where Female interacts with Hispanic and Female interacts with Black indicates that our assumption is likely true that the effect of race is the same for both men and women.

If there were no difference between how men and women are treated based on race, than the coefficient would be 0 on the interaction terms for both of those variables. Our model does return a non-zero value for both coefficients, so we look to the p-values to see if that value is due to chance or if it is significantly different from 0 at the 0.05 level.

The p-value for the coefficeint where Female interacts with Hispanic is .906, meaning that the value of this coefficient is not significantly different from 0. The same is true for the interaction between Female and Black as the p-value on this coefficient is .119.