Problem set #4

1)

(a) Calculate the χ2 test statistic by hand (even better if you can do ”by hand” in R).

F1e = (27/42)\*21 = 13.5

F2e = (27/42)\*13 = 8.35

F3e = (27/42)\*8 = 5.14

F4e = (15/42)\*21 = 7.5

F5e = (15/42)\*13 = 4.64

F6e = (15/42)\*8 = 2.85

X2 = (14-13.5)2/13.5 + (6-8.35)2/8.35 + (7-5.14)2/5.14 + (7-7.5)2/7.5 + (7-4.64)2/4.64 + (1-2.85)2/2.85 = 3.78

DF= (3-1)\*(2-1)= 2

pchisq(3.78, 2, lower.tail = FALSE) = .151

the variables are statistically independent

Z1 = (14-13.5)/(sqrt(13.5\*(1-27/42)\*(1-21/42))) = .322

Z2 = (6-8.35)/(sqrt(8.35\*(1-27/42)\*(1-13/42))) = -1.637

Z3 = (7-5.14)/(sqrt(5.14\*(1-27/42)\*(1-8/42))) = 1.525

Z4 = (7-7.5)/(sqrt(7.5\*(1-15/42)\*(1-21/42))) = -.322

Z5 = (7-4.64)/(sqrt(4.64\*(1-15/42)\*(1-13/42))) = 1.644

Z6 = (1-2.85)/(sqrt(2.85\*(1-15/42)\*(1-8/42))) -1.51

|  |  |  |  |
| --- | --- | --- | --- |
|  | Not stopped | Bribe requested | Stopped/given warning |
| Upper class | .322 | -1.637 | 1.525 |
| Lower class | -.322 | 1.637 | -1.525 |

the standardized residuals show us how far away each observed value is from the expectation

#So, they can show that the variables are independent because values follow no trend between upper class and lower class.

2)

#Ho: pi=.302

#Ha: pi doesnt equal .302

(.042-.302)/(sqrt(.302\*(1-.302)/30))

# -3.10

2\*pnorm(3.10, lower.tail=FALSE)

# .0019

# meaning the signs affect vote share because we can reject the null.

#2b

#Ho: pi=.302

#Ha: pi doesnt equal .302

(.042-.302)/(sqrt(.302\*(1-.302)/76))

#-4.93

2\*pnorm(4.93, lower.tail = FALSE)

#8.222e-07

#reject the null. Again Signs have an effect.

#2c

#the constant term is the value at which the regression line crosses the y-axis

#so, the constant is the proportion of votes that went to Cuccinelli when there are no signs at all.

#2d

#R^2 = .094, so yard signs impact voters very little compared to other factors.

#E.g. how you voted last election, how your parents vote, etc.

#3a

#Ho: #of repaired drinking water facilities under male leaders = # repaired drinking water facilities under female leaders

#Ha: #of repaired drinking water facilities under male leaders =/ # repaired drinking water facilities under female leaders

#3b

x <- read.csv(url("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv"))

model3b <- lm(water ~ reserved, data=x)

summary(model3b)

#3c

# On average, keeping all else the same, if the seat is reserved for a woman, they are 9.25 times more likely to repair drinking water facilities.

#4

install.packages("car")

library("car")

data(Prestige)

help(Prestige)

Prestige$professional <- ifelse(Prestige$type =="prof", "1", "0")

#b

model4b <- lm(prestige ~ income+professional+ income:professional, data=Prestige)

summary(model4b)

#c

#Y=21.1422589 + .0031709\*Income + 37.7812800\*professional1 + -.0023257\*income:professional1

#d

#on average, keeping all else the same, a one unit increase in income (one dollar increase in income) leads to a .0031709 increase in prestige score controlling for the independent effects of X2 and X3

#e

#professional is an indicator variable for having a professional job.

# on average, keeping all else the same, if you have a professional job, it leads to a 37.7812800 unit increase in your prestige score.

#f

.0031709\*1000

#3.1709

.0031709\*2000

#6.3418

6.3418-3.1709

# on average, keeping all else the same , there is a 3.1709 unit increase in prestige score when you have a professional job and your income increases by 1000 dollars.

#g

Y=21.1422589 + .0031709\*6000 + 37.7812800

# Y=77.94894, when job is professional

Y=21.1422589 + .0031709\*6000

# Y= 40.16766

77.94894-40.16766

# on average, keeping all else the same, the marginal effect of having a professional job when your income in $6000 is a 37.78128 unit increase in prestige score.

#5

library("faraway")

data("newhamp")

colnames(newhamp)

#a

model5aa <- lm(pObama ~ votesys, data=newhamp)

model5ab <- lm(pObama ~ votesys + povrate, data=newhamp)

model5ac <- lm(pObama ~ votesys + povrate + pci, data=newhamp)

model5ad <- lm(pObama ~ votesys + povrate + pci + Dean, data=newhamp)

model5ae <- lm(pObama ~ votesys + povrate + pci + Dean + white, data=newhamp)

model5af <- lm(pObama ~ Dean, data=newhamp)

summary(model5aa)

summary(model5ab)

summary(model5ac)

summary(model5ad)

summary(model5ae)

summary(model5af)

|  |  |  |
| --- | --- | --- |
|  | R^2 | Adjusted R^2 |
| Model1 | .0834 | .08006 |
| Model2 | .08971 | .08304 |
| Model3 | .2441 | .2358 |
| Model4 | .5091 | .5018 |
| Model5 | .5093 | .5002 |
| Model6 | .4181 | .416 |

#c

Model 5.

R^2 increases when you add a lot of variables because you are accounting for more of the variance.

#6

#a

read.csv("incumbents\_subset.csv")

y <- read.csv("incumbents\_subset.csv")

model6a <- lm(voteshare ~ difflog, data = y)

summary(model6a)

plot(y$voteshare, y$difflog)

abline(model6a, col= "red")

residualmodel6a <- predict(model6a)

segments(y$difflog, y$voteshare, y$difflog, residualmodel6a)



#b

model6b <- lm(presvote ~ difflog, data = y)

plot(y$difflog, y$presvote)

abline(model6b, col="red")

summary(model6b)

residuals6b <- predict(model6b)

segments(y$difflog, y$presvote, y$difflog, residuals6b)



#c

model6c <- lm(voteshare ~ presvote, data=y)

plot(y$presvote,y$voteshare)

abline(model6c, col="red")

summary(model6c)

residuals6c <- predict(model6c)

segments(y$presvote, y$voteshare, y$presvote, residuals6c)



#d

part1 <- resid(model6a)

part2 <- resid(model6b)

model6d <- lm(part1 ~ part2)

summary(model6d)

plot(part1, part2)

abline(model6d, col="red")

prediction <- predict(model6d)



#e

model6e <- lm(voteshare ~ difflog + presvote, data=y)

summary(model6e)

predict6e <- predict(model6e)

The residuals are exactly the same as in part 6d.