Luke Ehrenstrom Problem Set 4

Question 1

1. χ2 = 3.7912
2. SuperTroopers = matrix(c(14, 6, 7, 7, 7, 1), ncol = 3, byrow=TRUE)
3. colnames(SuperTroopers) = c("Not Stopped", "Bribe requested", "Stopped/given warning")
4. rownames(SuperTroopers) = c("Upper class", "Lower class")
5. SuperTroopers
6. SuperTroopersExpected = matrix(c((27/42)\*21, (27/42)\*13, (27/42)\*8, (15/42)\*21, (15/42)\*13, (15/42)\*8), ncol = 3, byrow = TRUE)
7. SuperTroopersExpected
8. ResidualST = SuperTroopers - SuperTroopersExpected
9. ChiSqST = sum((ResidualST^2)/SuperTroopersExpected)
10. # x2 = 3.791
11. p = .1502. You fail to reject the null hypothesis that there is no difference between the frequency of traffic stops for lower-class and upper-class people, providing some evidence that there is no difference between the frequency of traffic stops for lower-class and upper-class people.
12. chisq.test(SuperTroopers)
13. # p = .1502
14. 
15. StandardizedResiduals = (SuperTroopers - SuperTroopersExpected)/sqrt(SuperTroopersExpected)
16. colnames(StandardizedResiduals) = c("Not Stopped", "Bribe requested", "Stopped/given warning")
17. rownames(StandardizedResiduals) = c("Upper class", "Lower class")
18. StandardizedResiduals
19. The standardized residuals provide information on how many standard deviations away from the expected values the observed values were. Considering that all observed values were within roughly one standard deviation, we can confirm (only somewhat due to the low level of expected values) that it is probable for the study to have yielded the observed values without rejecting the null hypothesis.

Sources on making tables, chi-square

https://www.youtube.com/watch?v=53kYOOr5Yhk

https://www.cyclismo.org/tutorial/R/types.html#tables

https://www.statisticshowto.datasciencecentral.com/what-is-a-standardized-residuals/

Question 2

1. *H*0: There is no correlation between the presence of a yard sign in a precinct and vote share.

*H*a : There is a correlation between the presence of a yard sign in a precinct and vote share.

p = .016 is less than our level of significance .05, so we can reject the null hypothesis that there is no correlation between the presence of a yard sign and vote share, providing some evidence that there is a correlation between the presence of a yard sign and vote share.

1. # Question 2 a
2. r2 = .094
3. k = 2
4. n = 30
5. df = n - 2
6. stderror = .016
7. coeff = .042
8. t = .042/.016
9. p = dt(t, df)
10. # p-value .016
11. *H*0: There is no correlation between the presence of a yard sign in an adjacent precinct and vote share.

*H*a : There is a correlation between the presence of a yard sign in an adjacent precinct and vote share.

p = .0028 is less than our level of significance .05, so we can reject the null hypothesis that there is no correlation between the presence of a yard sign in an adjacent precinct and vote share, providing some evidence that there is a correlation between the presence of a yard sign in an adjacent precinct and vote share.

1. # Question 2 b
2. n = 76
3. df = n - 2
4. stderror = .013
5. coeff = .042
6. t = .042/.013
7. p = dt(t, df)
8. # p-value: .0028
9. With no presence of a yard sign either in the precinct or adjacent, the vote share percentage is .302.
10. The model fit is only .094, indicating that there is much unaccounted for variance, which indicates that there are likely much more significant factors that influence vote share.

Sources: <https://stattrek.com/regression/slope-test.aspx>

https://www.researchgate.net/publication/288056551\_The\_Effects\_of\_Lawn\_Signs\_on\_Vote\_Outcomes\_Results\_from\_Four\_Randomized\_Field\_Experiments

Question 3

* 1. *H*0: There is no correlation between the number of new or repaired drinking-water facilities within a Gram Panchayat and whether the leadership was reserved for a woman or not.

*H*a : There is a correlation between the number of new or repaired drinking-water facilities within a Gram Panchayat and whether the leadership was reserved for a woman or not.

* 1. With a p-value of 0.0197, we can reject *H*0 and there is some evidence to suggest that there is a correlation between the reservation policy and the number of new or repaired drinking-water facilities within a Gram Panchayat.

1. WomenLeaders = read.csv(url("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv"))
2. reserved = WomenLeaders$reserved
3. water = WomenLeaders$water
4. RegressionWomenLeaders = lm(water ~ reserved)
5. summary(RegressionWomenLeaders)
6. plot(WomenLeaders$reserved, WomenLeaders$water)
7. abline(RegressionWomenLeaders, col="red")
8. # p = 0.0197, coefficient: 9.252
   1. When the reservation policy is in place in a GP, there are approximately 9.252 more new or repaired drinking-water facilities than when the reservation policy is not in place in a GP.

Question 4

1. install.packages(car)
2. library(car)
3. data(Prestige)
4. help(Prestige)
5. View(Prestige)
7. Prestige$professional = NA # make empty column
8. Prestige$professional = Prestige$type # make the column
9. Prestige$professional = as.character(Prestige$professional) # First part of recoding, essentially telling R you must
10. # read the code like characters rather than a factor (just makes R not freak out)
11. Prestige$professional[ Prestige$professional=="prof" ] <- "1" # recoding
12. Prestige$professional[ Prestige$professional=="bc" ] <- "0" # recoding
13. Prestige$professional[ Prestige$professional=="wc" ] <- "0" # recoding
14. PrestigeLM = lm(prestige ~ income + professional + professional:income, data=Prestige)
15. summary(PrestigeLM)
16. ŷ = 21.142 + .0032x1 + 37.78x2 + -0.0023x3

where ŷ is prestige

where x1 is income (dollars)

where x2 is whether a job is considered professional or not

where x3 is the interaction between whether a job is professional and income

1. For every $1 increase in income, there is a .0032 increase in the measure of prestige.
2. If a career is considered professional vs. not, there is a 37.78 increase in the measure of prestige.
3. answer.e1 = 21.142 + .0032\*(0) + 37.78\*(1) - 0.0023\*(0)\*1
4. answer.e2 = 21.142 + .0032\*(1000) + 37.78 - 0.0023\*(1000)\*1
5. answer.e = answer.e2 - answer.e1
6. answer.e # 0.9
7. An increase in $1000 of salary leads to an estimated 0.9 increase in the measure of prestige.
8. answer.f1 = 21.142 + .0032\*(6000) + 37.78\*(0) - 0.0023\*(6000)\*0
9. answer.f2 = 21.142 + .0032\*(6000) + 37.78\*(1) - 0.0023\*(6000)\*1
10. answer.f = answer.f2 - answer.f1
11. answer.f # 23.98
12. Switching to a professional career with $6000 income leads to a 23.98 increase in the measure of prestige.

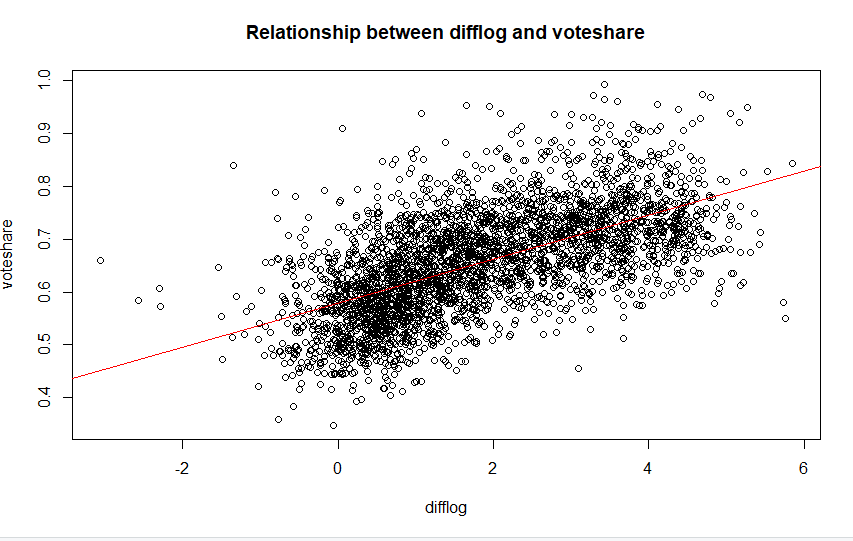
Question 5

1. Code for each:
2. data("newhamp")
3. colnames(newhamp)
4. View(NewHamp)
5. NewHamp = newhamp
6. NewHamp = as.data.frame(NewHamp)
7. NewHamp$votesys = as.character(NewHamp$votesys)
8. NewHamp$votetsys[NewHamp$votesys=="D"] = 1
9. NewHamp$votetsys[NewHamp$votesys=="H"] = 0
10. votesysonly = lm(pObama ~ votesys, data=NewHamp)
11. summary(votesysonly)
12. VSPR = lm(pObama ~ votesys + povrate, data=NewHamp)
13. summary(VSPR)
14. VSPRPCI = lm(pObama ~ votesys + povrate + pci, data=NewHamp)
15. summary(VSPRPCI)
16. VSPRPCID = lm(pObama ~ votesys + povrate + pci + Dean, data=NewHamp)
17. summary(VSPRPCID)
18. VSPRCIDW = lm(pObama ~ votesys + povrate + pci + Dean + white, data=NewHamp)
19. summary(VSPRCIDW)
20. Deanonly = lm(pObama ~ Dean, data=NewHamp)
21. summary(Deanonly)

|  |  |  |
| --- | --- | --- |
|  | R^2 | Adjusted R^2 |
| Model 1 | 0.0834 | 0.08006 |
| Model 2 | 0.08971 | 0.08304 |
| Model 3 | 0.2441 | 0.2358 |
| Model 4 | 0.5091 | 0.5018 |
| Model 5 | 0.5093 | 0.5002 |
| Model 6 | 0.4181 | 0.416 |

1. The model with the highest R^2 is Model 5, the one with the most dummy variables. As the number of variables increases, R^2 also increases.

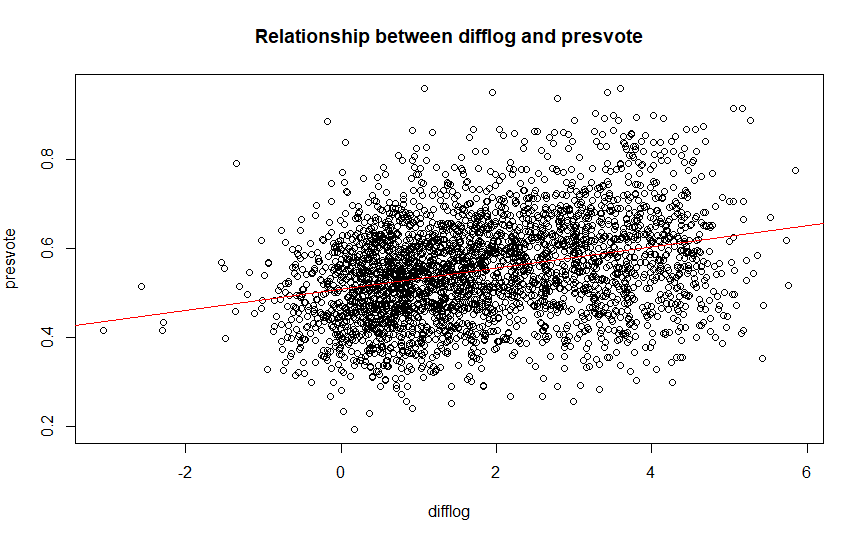
Question 6

1. 

ŷ = 0.579031 + 0.041666x1

where ŷ is voteshare

where x1 is difflog

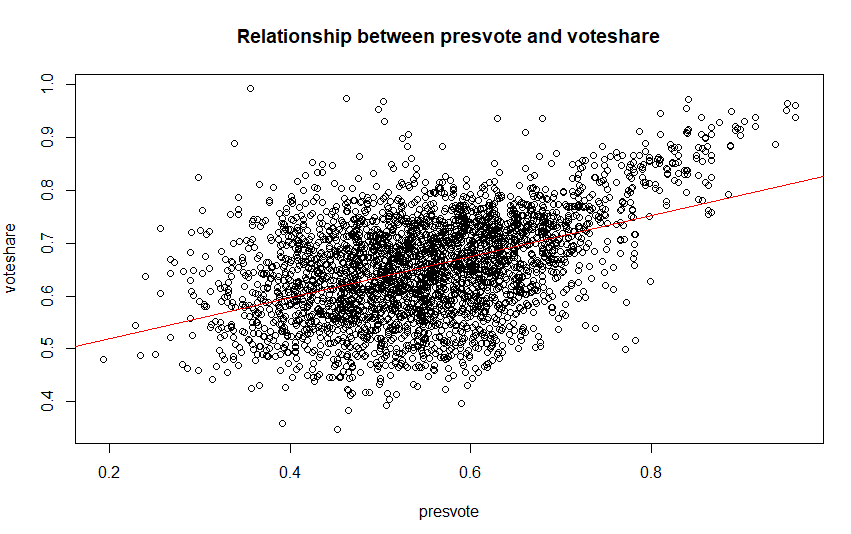
1. setwd("C:/Users/lehre/Desktop/QPM/PS4")
2. incumbents\_subset = read.csv("incumbents\_subset.csv")
3. View(incumbents\_subset)
4. VSdifflog.lm = lm(voteshare ~ difflog, data=incumbents\_subset)
5. summary(VSdifflog.lm)
6. VSdifflog.rs = summary(VSdifflog.lm)$residuals
7. VSdifflog.fv = VSdifflog.lm$fitted.values
8. pdf("Question6a.pdf")
9. plot(incumbents\_subset$difflog, incumbents\_subset$voteshare, xlab = "difflog", ylab = "voteshare",main = "Relationship between difflog and voteshare")
10. abline(VSdifflog.lm, col = "red")
11. dev.off()
12. 

ŷ = 0.507583 + .023837x1

where ŷ is presvote

where x1 is difflog

1. # Question 6 b
2. PVdifflog.lm = lm(presvote ~ difflog, data=incumbents\_subset)
3. summary(PVdifflog.lm)
4. PVdifflog.rs = summary(PVdifflog.lm)$residuals
5. pdf("Question6b.pdf")
6. plot(incumbents\_subset$difflog, incumbents\_subset$presvote, xlab = "difflog", ylab = "presvote",main = "Relationship between difflog and presvote")
7. abline(PVdifflog.lm, col = "red")
8. dev.off()

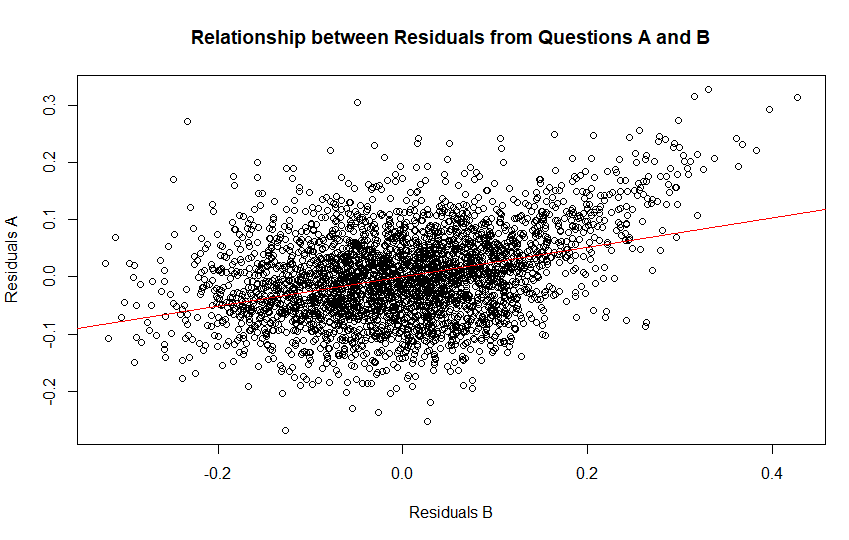


1. ŷ = 0.441330 + .388018x1

where ŷ is voteshare

where x1 is presvote

1. # Question 6 c
2. VSpresvote.lm = lm(voteshare ~ presvote, data=incumbents\_subset)
3. summary(VSpresvote.lm)
4. pdf("Question6c.pdf")
5. plot(incumbents\_subset$presvote, incumbents\_subset$voteshare, xlab = "presvote", ylab = "voteshare",main = "Relationship between presvote and voteshare")
6. abline(VSpresvote.lm, col = "red")
7. dev.off()



1. ŷ = -4.860e-18 + 2.569e-01x1

where ŷ is Residuals A

where x1 is Residuals B

1. # Question 6 d
2. Residual.lm= lm(VSdifflog.rs ~ PVdifflog.rs, data=incumbents\_subset)
3. summary(Residual.lm)
4. pdf("Question6d.pdf")
5. plot(PVdifflog.rs, VSdifflog.rs, xlab = "Residuals B", ylab = "Residuals A",main = "Relationship between Residuals from Questions A and B")
6. abline(Residual.lm, col = "red")
7. dev.off()
8. ŷ = 0.4486442 + .00355431x1 + 0.2568770x2

where ŷ is voteshare

where x1 is difflog

where x2 is presvote

1. # Question 6 e
2. VSdifflogpresvote.lm = lm(voteshare ~ difflog + presvote, data=incumbents\_subset)
3. summary(VSdifflogpresvote.lm)

c. Residual standard error is nearly identical.

Collaborated with: Sinclair Bowman, Michael Rimmey