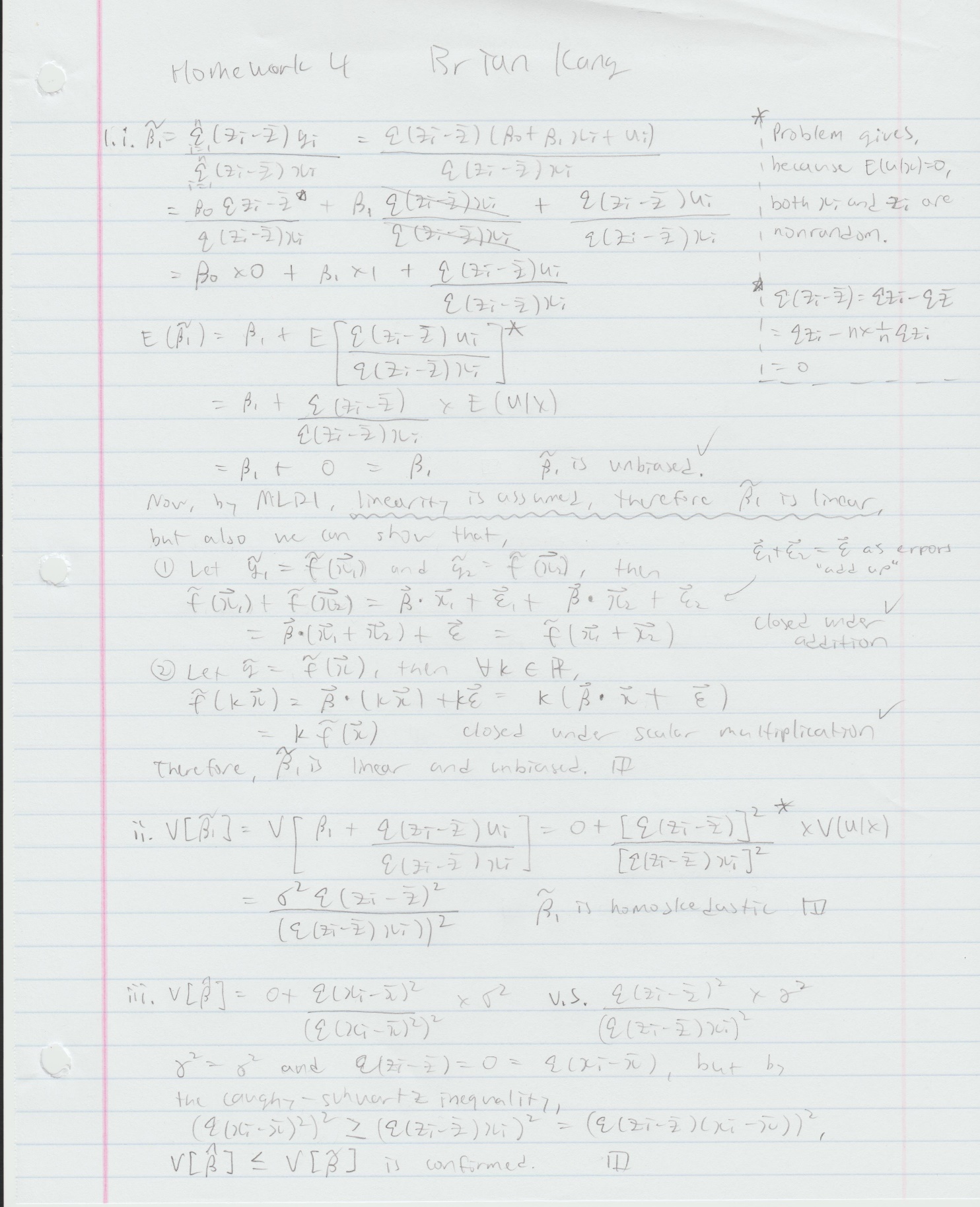
hw04

Brian Kang

Tue Feb 12 21:37:41 2019



#rm(list = ls())  
  
# C1)  
# i)  
# The sign for Beta2 is most likely positive. It is conceived  
# that infants born in wealthier families will be well fed,  
# therefore, the birth weight will likely increase as income  
# increases.  
  
# ii)  
# I think cigs and faminc are likely to be inversely correlated.  
# The wealthier families might be more concerned with personal  
# development and health and may be more informed of the harms  
# of smoking, so they might smoke less. On the other hand, those  
# with lower income, people may be less educated and are more  
# prone to smoke as an habit.   
# To simplify, I think cigarettes are an inferior good, so the  
# correlation between cigs and faminc might be negative.  
  
# iii)  
library(wooldridge)

## Warning: package 'wooldridge' was built under R version 3.5.2

data(bwght)  
lm.1 <- lm(bwght ~ cigs, data = bwght)  
# sample size  
length(bwght$cigs) # or

## [1] 1388

# R-squared value  
summary(lm.1)$r.squared

## [1] 0.02272912

# equation form  
# bwght = 119.77190 - 0.51377\*cigs  
lm.2 <- lm(bwght ~ cigs+faminc, data = bwght) # overfit?  
# sample size  
length(lm.2$fitted)

## [1] 1388

# R-squared value  
summary(lm.2)$r.squared

## [1] 0.02980484

# equation form  
# bwght = 116.97413 - 0.46341\*cigs + 0.09276\*faminc  
summary(lm.1)

##   
## Call:  
## lm(formula = bwght ~ cigs, data = bwght)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -96.772 -11.772 0.297 13.228 151.228   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 119.77190 0.57234 209.267 < 2e-16 \*\*\*  
## cigs -0.51377 0.09049 -5.678 1.66e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.13 on 1386 degrees of freedom  
## Multiple R-squared: 0.02273, Adjusted R-squared: 0.02202   
## F-statistic: 32.24 on 1 and 1386 DF, p-value: 1.662e-08

summary(lm.2)

##   
## Call:  
## lm(formula = bwght ~ cigs + faminc, data = bwght)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -96.061 -11.543 0.638 13.126 150.083   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 116.97413 1.04898 111.512 < 2e-16 \*\*\*  
## cigs -0.46341 0.09158 -5.060 4.75e-07 \*\*\*  
## faminc 0.09276 0.02919 3.178 0.00151 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.06 on 1385 degrees of freedom  
## Multiple R-squared: 0.0298, Adjusted R-squared: 0.0284   
## F-statistic: 21.27 on 2 and 1385 DF, p-value: 7.942e-10

summary(lm(cigs~faminc, data = bwght))

##   
## Call:  
## lm(formula = cigs ~ faminc, data = bwght)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.661 -2.668 -1.896 -0.103 47.829   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.688107 0.291297 12.661 < 2e-16 \*\*\*  
## faminc -0.055154 0.008432 -6.541 8.58e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.885 on 1386 degrees of freedom  
## Multiple R-squared: 0.02994, Adjusted R-squared: 0.02924   
## F-statistic: 42.78 on 1 and 1386 DF, p-value: 8.575e-11

# From the summary of the results of the linear model, we can  
# conclude that adding faminc does not substantially change the  
# estimated effect of cigs on bwght. The R-squared values both  
# with and without faminc are low, close to 0.02, and the std.  
# error values don't change significantly. The std. error for  
# the intercept estimate even increases from 0.57 to 1.04. When  
# there is barely any benefit from adding another predictor,  
# might as well not add the term. But also, we don't want to fit  
# too many terms everytime just because more variation can be  
# explained because overfitting can be misleading.  
# From summary() you can see the Beta coefficients. They do not  
# change much from estimaiting with faminc. From the lm()  
# between cigs and faminc we see that they do not strongly  
# correlate either.  
  
# C2)  
# i)  
data("hprice1")  
lm.3 <- lm(price ~ sqrft+bdrms, data = hprice1)  
# equation form  
# price = -19.31500 + 0.12844\*sqrft + 15.19819\*bdrms  
  
# ii)  
# estimated increase in price with 1 more bdrm, holding sqrft  
summary(lm.3)$coefficient[3,1]

## [1] 15.19819

# iii)  
# estimated increase in price with 1 bdrm ie 140 sqrft  
summary(lm.3)$coefficient[3,1]+140\*summary(lm.3)$coefficient[2,1]

## [1] 33.17926

# (ii) shows the increase in price when we get 1 more bdrm with  
# 0 sqrt, which does not make sense. (iii) shows the more  
# realistic idea of how much price increases with an extra bdrm  
# with a physical area of 140 sqrft.  
  
# iv)  
summary(lm.3)$r.squared

## [1] 0.6319184

# About 63.2% of the variation in price is explained by the   
# square footage and the number of bedrooms.  
  
# v)  
# predicted selling price of the first house  
lm.3$fitted[1]

## 1   
## 354.6052

# vi)  
# residual for the first house  
lm.3$residuals[1]

## 1   
## -54.60525

# This suggests that the buyer underpaid for this house by $54k  
  
# C8)  
# i)  
data("discrim")  
# to check if AT LEAST 1 NA, [anyNA()] or [sum(is.na(data))>0]  
# deal with NA values  
mean(discrim$prpblck, na.rm = T)

## [1] 0.1134864

sd(discrim$prpblck, na.rm = T)

## [1] 0.1824165

# unit of measurement is proporiton of blacks within a zipcode  
mean(discrim$income, na.rm = T)

## [1] 47053.78

sd(discrim$income, na.rm = T)

## [1] 13179.29

# unit of measurement is median income ($s) within a zipcode  
  
# ii)  
lm.4 <- lm(psoda ~ prpblck+income, data = discrim,   
 na.action = na.omit)  
# sample size  
length(lm.4$fitted)

## [1] 401

# R-squared value  
summary(lm.4)$r.squared

## [1] 0.06422039

# equation form  
# psoda = 9.563e-01 + 1.150e-01\*prpblck + 1.603e-06\*income  
# Interpretation of 1.150e-01:  
# We estimate that with each increase of prpblck by 1, holding  
# all other variables constant, psoda will increase by 1.150e-01.  
# I think this is a insignificant value. It is saying that with  
# each 100 percentage point increase in the proportion of blacks,  
# The price of medium soda increases by 10 cents.  
  
# iii)  
lm.5 <- lm(psoda ~ prpblck, data = discrim,   
 na.action = na.omit)  
summary(lm.5)

##   
## Call:  
## lm(formula = psoda ~ prpblck, data = discrim, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.30884 -0.05963 0.01135 0.03206 0.44840   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.03740 0.00519 199.87 < 2e-16 \*\*\*  
## prpblck 0.06493 0.02396 2.71 0.00702 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0881 on 399 degrees of freedom  
## (9 observations deleted due to missingness)  
## Multiple R-squared: 0.01808, Adjusted R-squared: 0.01561   
## F-statistic: 7.345 on 1 and 399 DF, p-value: 0.007015

# equation form  
# psoda = 1.03740 + 0.06493\*prpblck  
# The discrimination effect smaller without the income.  
# But the estimated intercept has increased.  
  
# iv)  
lm.6 <- lm(log(psoda) ~ prpblck+log(income), data = discrim,   
 na.action = na.omit)  
summary(lm.6)

##   
## Call:  
## lm(formula = log(psoda) ~ prpblck + log(income), data = discrim,   
## na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33563 -0.04695 0.00658 0.04334 0.35413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.79377 0.17943 -4.424 1.25e-05 \*\*\*  
## prpblck 0.12158 0.02575 4.722 3.24e-06 \*\*\*  
## log(income) 0.07651 0.01660 4.610 5.43e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0821 on 398 degrees of freedom  
## (9 observations deleted due to missingness)  
## Multiple R-squared: 0.06809, Adjusted R-squared: 0.06341   
## F-statistic: 14.54 on 2 and 398 DF, p-value: 8.039e-07

# equation form  
# log(psoda) = -0.79377 + 0.12158\*prpblck + 0.07651\*log(income)  
# According to Appendix A, in a semi-elastic relationship,  
# % change in psoda = 100\*slope\*change in prpblack. Therefore,  
# % change in psoda = 100\*0.12158\*0.2 = 2.4316  
  
# v)  
lm.7 <- lm(log(psoda) ~ prpblck+log(income)+prppov, data = discrim,   
 na.action = na.omit)  
lm.7$coefficients["prpblck"]

## prpblck   
## 0.07280726

# The slope decreased by about 0.05 after addint the new term.  
  
# vi)  
# solve correlation coeffecicient AND deal with NA values  
cor(log(discrim$income),discrim$prppov, use = "complete.obs")

## [1] -0.838467

# No, the sign is roughly what I expected, but I expected a  
# much smaller value. I was expecting a value of 0.2 maximum.  
  
# vii)  
summary(lm(log(psoda) ~ prpblck+prppov, data = discrim,   
 na.action = na.omit))

##   
## Call:  
## lm(formula = log(psoda) ~ prpblck + prppov, data = discrim, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.35032 -0.05212 0.01336 0.03965 0.35834   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.039803 0.006123 6.501 2.4e-10 \*\*\*  
## prpblck 0.101558 0.031097 3.266 0.00119 \*\*   
## prppov -0.155739 0.084182 -1.850 0.06505 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0839 on 398 degrees of freedom  
## (9 observations deleted due to missingness)  
## Multiple R-squared: 0.0267, Adjusted R-squared: 0.02181   
## F-statistic: 5.459 on 2 and 398 DF, p-value: 0.004581

# equation form:  
# log(psoda) = 0.039803 + 0.101558\*prpblck - 0.155739\*prppov  
# The statement is not true at all. High correlation does not mean  
# they will produce equal results. Creating a lm with prppov in  
# place for log(income), we can see that the signs of our estimates  
# changed. R^2 decreased from 0.0821 to 0.02. Also, our t values  
# and their Pr(>|t|) value also drastically changed. P-value  
# increased from 8.039e-07 to 0.004581 which is not good. Therefore,  
# we can conclude that the statement is not true.