Econ_A4

Hanpeng Wang

March 31, 2019

```
knitr::opts_chunk$set(echo = TRUE)
options(warn=-1)
library(foreign)
df <- read.dta("C:/Users/wangh/Downloads/utown.dta")
```

A

```
#1
mean_p1 <- mean(df[df[, "pool"] == 1,"price"])
mean_p0 <- mean(df[df[, "pool"] == 0,"price"])
mean_dif <- mean_p1 - mean_p0
se <- sqrt(var(df[df[, "pool"] == 0,"price"])/length(df[df[, "pool"] == 0,"price"]) + var(df[df
[, "pool"] == 1,"price"]) / length(df[df[, "pool"] == 1,"price"]))
t_stat <- (mean_p1 - mean_p0) / se

t_stat
```

```
## [1] 1.668285
```

```
#2

#using t statistics to test difference significance

df$go <- ifelse(df$utown == "0", 1, 0)
df$p0 <- ifelse(df$pool == 0, 1, 0)
df$f0 <- ifelse(df$fplace == 0, 1, 0)

reg1 <- summary(lm(price~age+go, data = df[df[,"go"] == 1, ]))
reg2 <- summary(lm(price~age+utown, data = df[df[,"utown"] == 1, ]))
se1 <-  sqrt(reg1$coefficients['age','Std. Error']**2/ + length(df[df[,"go"] == 1, 1]) +
             reg2$coefficients['age','Std. Error']**2/ + length(df[df[,"utown"] == 1, 1]))
t_stat1 <- (reg1$coefficients['age','Estimate'] - reg2$coefficients['age','Estimate']) / se

#------------------------
reg3 <- summary(lm(price~age+pool, data = df[df[,"pool"] == 1, ]))
reg4 <- summary(lm(price~age+p0, data = df[df[,"p0"] == 1, ]))
se2 <-  sqrt(reg3$coefficients['age','Std. Error']**2/ + length(df[df[,"pool"] == 1, 1]) +
               reg4$coefficients['age','Std. Error']**2/ + length(df[df[,"p0"] == 1, 1]))
t_stat2 <- (reg3$coefficients['age','Estimate'] - reg4$coefficients['age','Estimate']) / se2

#-------------------------
reg5 <- summary(lm(price~age+pool+utown, data = df[df[,"pool"] == 1 & df[,"utown"] == 1, ]))
reg6 <- summary(lm(price~age+pool+go, data = df[df[,"pool"] == 1 & df[,"go"] == 1, ]))
se3 <-  sqrt(reg5$coefficients['age','Std. Error']**2/ + length(df[df[,"pool"] == 1 & df[,"utow
n"] == 1, 1]) +
             reg6$coefficients['age','Std. Error']**2/ + length(df[df[,"pool"] == 1 & df[,"go"]
 == 1, 1]))
t_stat3 <- (reg5$coefficients['age','Estimate'] - reg6$coefficients['age','Estimate']) / se3

#--------------------------
reg7 <- summary(lm(price~age+p0+utown, data = df[df[,"p0"] == 1 & df[,"utown"] == 1, ]))
reg8 <- summary(lm(price~age+pool+go, data = df[df[,"pool"] == 1 & df[,"go"] == 1, ]))
se4 <-  sqrt(reg5$coefficients['age','Std. Error']**2/ + length(df[df[,"p0"] == 1 & df[,"utown"]
== 1, 1]) +
               reg6$coefficients['age','Std. Error']**2/ + length(df[df[,"pool"] == 1 & df[,"go"
] == 1, 1]))
t_stat4 <- (reg5$coefficients['age','Estimate'] - reg6$coefficients['age','Estimate']) / se4

paste("for question a - d, their t stats are", c(t_stat1, t_stat2, t_stat3, t_stat4))
```

```
## [1] "for question a - d, their t stats are -0.0162143594604482"
## [2] "for question a - d, their t stats are 14.0282710584664"
## [3] "for question a - d, their t stats are 3.45173874478557"
## [4] "for question a - d, their t stats are 4.36264402542886"
```

```
#3

mean_house_ut <- mean(df[df[, "utown"] == 1,"sqft"])
mean_house_go <- mean(df[df[, "go"] == 1,"sqft"])
mean_house_dif <- mean_house_ut - mean_house_go
se_house <- sqrt(var(df[df[, "utown"] == 1,"sqft"])/length(df[df[, "utown"] == 1,"sqft"]) + var
(df[df[, "p0"] == 1,"sqft"]) / length(df[df[, "p0"] == 1,"sqft"]))
t_stat_house <- (mean_house_dif) / se_house
t_stat_house
```

```
## [1] 0.8161393
```

```
print("because size difference between ut and go is not significant, so not most of large houses
in UT")
```

```
## [1] "because size difference between ut and go is not significant, so not most of large house
s in UT"
```

```
mean_house_pool <- mean(df[df[, "pool"] == 1,"sqft"])
mean_house_p0 <- mean(df[df[, "p0"] == 1,"sqft"])
mean_house_dif_pool <- mean_house_pool - mean_house_p0
se_house_pool <- sqrt(var(df[df[, "pool"] == 1,"sqft"])/length(df[df[, "pool"] == 1,"sqft"]) + v
ar(df[df[, "p0"] == 1,"sqft"]) / length(df[df[, "p0"] == 1,"sqft"]))
t_stat_house_pool <- (mean_house_dif_pool) / se_house_pool
t_stat_house_pool
```

```
## [1] -0.1345955
```

```
print("because size difference between pool and without is not significant, so not most of large
houses with or without pool")
```

```
## [1] "because size difference between pool and without is not significant, so not most of larg
e houses with or without pool"
```

```
#4
summary(lm(price~sqft+age+utown+pool+fplace, data = df))
```

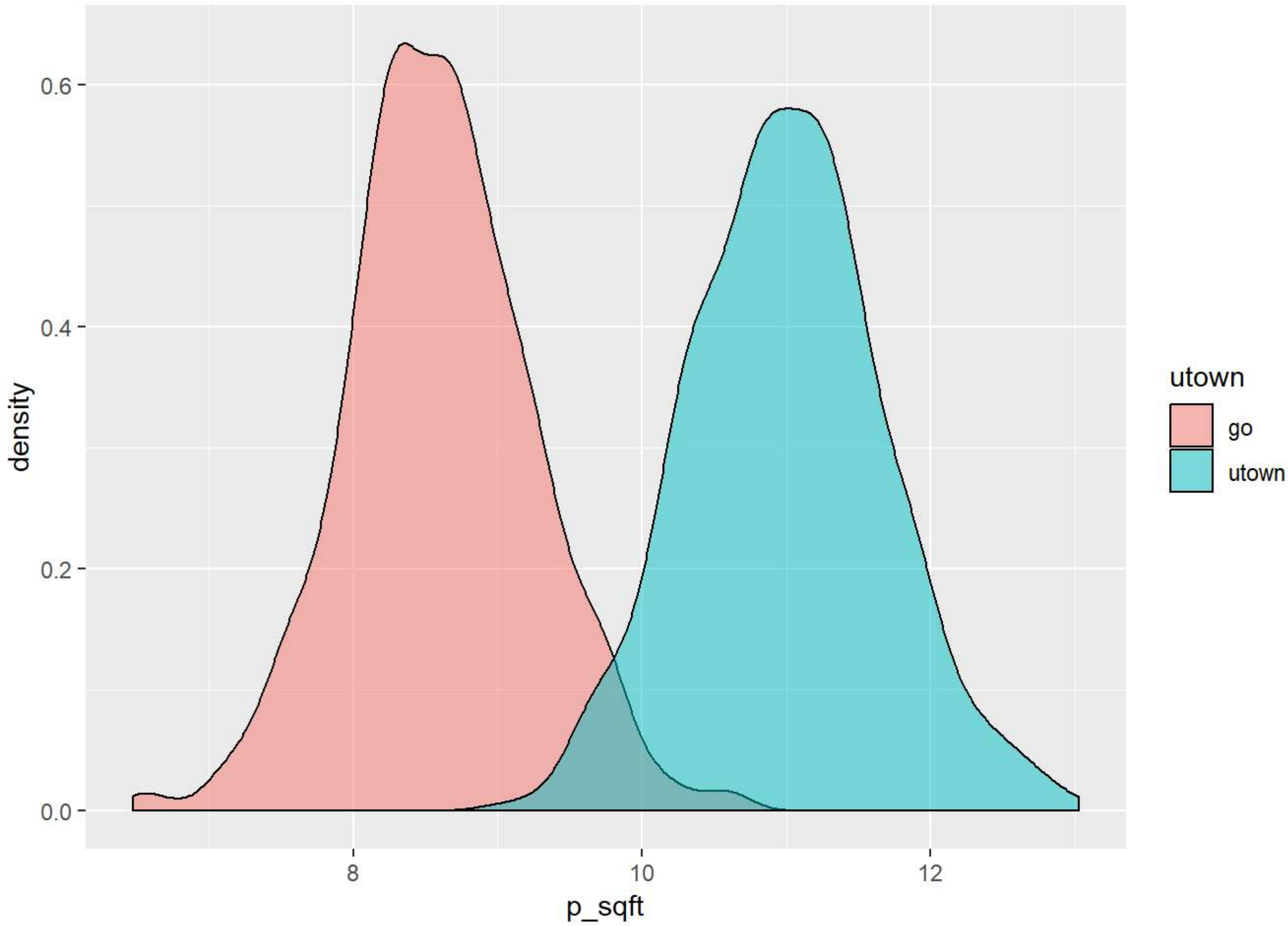
```
##
## Call:
## lm(formula = price ~ sqft + age + utown + pool + fplace, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.971 -10.411   0.198  10.438  44.759
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.91188    4.28937   1.611 0.107410
## sqft         8.31832    0.16717  49.759  < 2e-16 ***
## age         -0.19299    0.05157  -3.743 0.000193 ***
## utown       60.19623    0.97153  61.960  < 2e-16 ***
## pool         4.35257    1.20526   3.611 0.000320 ***
## fplace       1.39881    0.97681   1.432 0.152452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.33 on 994 degrees of freedom
## Multiple R-squared:  0.8686, Adjusted R-squared:  0.8679
## F-statistic:  1314 on 5 and 994 DF,  p-value: < 2.2e-16
```

```
print('the most important factor is location, whether if it is in utown')
```

```
## [1] "the most important factor is location, whether if it is in utown"
```

```
#5

df$p_sqft <- df$price / df$sqft
library(ggplot2)
plot_df <- df[,c('p_sqft','utown','age','fplace')]
plot_df$utown <- ifelse(df$utown == "1", "utown", "go")
ggplot(plot_df, aes(x = p_sqft, fill = utown)) + geom_density(alpha = 0.5)
```




```
mean(df[df[,"fplace"] == 1 & df[,"utown"] == 1 ,"p_sqft"]) - mean(df[df[,"fplace"] == 1 & df[,"u
town"] == 0 ,"p_sqft"])
```

```
## [1] 2.415263
```

```
mean(df[df[,"fplace"] == 0 & df[,"utown"] == 1 ,"p_sqft"]) - mean(df[df[,"fplace"] == 0 & df[,"u
town"] == 0 ,"p_sqft"])
```

```
## [1] 2.391187
```

```
print("fireplace doesn't change price difference much")
```

```
## [1] "fireplace doesn't change price difference much"
```

```
age<- df[,"age"]
age_median <- sort(age)[length(age)/2]
mean(df[df[,"age"] <= age_median & df[,"utown"] == 1 ,"p_sqft"]) - mean(df[df[,"age"] <= age_med
ian & df[,"utown"] == 0 ,"p_sqft"])
```

```
## [1] 2.375417
```

```
mean(df[df[,"age"] > age_median & df[,"utown"] == 1 ,"p_sqft"]) - mean(df[df[,"age"] > age_media
n & df[,"utown"] == 0 ,"p_sqft"])
```

```
## [1] 2.428215
```

```
print("house age doesn't change price difference much")
```

```
## [1] "house age doesn't change price difference much"
```

```
#6

df$old_house <- ifelse(df$age <= mean(df[,"age"]), 0, 1)
df$big_house <- ifelse(df$sqft <= mean(df[,"sqft"]), 0, 1)

paste("difference with new_big and new_small is",
      mean(df[df[,"old_house"] == 0 & df[,"big_house"] == 1 ,"sqft"]) -
      mean(df[df[,"old_house"] == 0 & df[,"big_house"] == 0 ,"sqft"])  )
```

```
## [1] "difference with new_big and new_small is 4.98710199600798"
```

```
paste("difference with old_big and new_small is",
      mean(df[df[,"old_house"] == 1 & df[,"big_house"] == 1 ,"sqft"]) -
        mean(df[df[,"old_house"] == 0 & df[,"big_house"] == 0 ,"sqft"])  )
```

```
## [1] "difference with old_big and new_small is 5.03775579710145"
```

B

```r
#DGM

x1 <-  rep(1, 5000)
x2 <- round(runif(5000,0,100))
x3 <- runif(5000,1,50)
x4 <- rnorm(5000, mean = 5.2 , sd = 1.25)
beta <- c(12, -0.7, 34, -0.17)
equation <- cbind(cbind(x1,x2,x3,x4)%*%beta, x1, x2, x3, x4)

#1

sample_equation <- equation[sample(nrow(equation), 300, replace = T),]
y <- rep(0, 300)
sample_equation <- cbind(sample_equation, y)

beta_hat <- c()
for (i in 1:2000) {
    sample_equation[,6] <- sample_equation[,1] + rnorm(300,0,1)
    coefs <- summary(lm(sample_equation[,6]~sample_equation[,2:5]+0))$coefficients[,1]
    beta_hat<- rbind(beta_hat, coefs)

}

beta_means <- colMeans(beta_hat)
beta_means
```

```
## sample_equation[, 2:5]x1 sample_equation[, 2:5]x2 sample_equation[, 2:5]x3
##               11.9964261               -0.7000155               34.0001462
## sample_equation[, 2:5]x4
##               -0.1698653
```

```r
beta
```

```
## [1] 12.00 -0.70 34.00 -0.17
```

```r
print('yes, they are very close')
```

```
## [1] "yes, they are very close"
```

```
#2

sample_equation2 <- equation[sample(nrow(equation), 300, replace = T), - 5]
y2 <- rep(0, 300)
sample_equation2 <- cbind(sample_equation2, y2)

beta_hat2 <- c()
for (i in 1:2000) {
  sample_equation2[,5] <- sample_equation2[,1] + rnorm(300,0,1)
  coefs <- summary(lm(sample_equation2[,5]~sample_equation2[,2:4]+0))$coefficients[,1]
  beta_hat2<- rbind(beta_hat2, coefs)

}

beta2_means <- colMeans(beta_hat2)
beta2_means
```

```
## sample_equation2[, 2:4]x1 sample_equation2[, 2:4]x2
##                11.1149007                -0.6998734
## sample_equation2[, 2:4]x3
##                34.0003150
```

```
beta
```

```
## [1] 12.00 -0.70 34.00 -0.17
```

```
print('they are close, but estimation error for x1 is a bit bigger when x4 is ommited')
```

```
## [1] "they are close, but estimation error for x1 is a bit bigger when x4 is ommited"
```

```
#3

sample_equation3 <- equation[sample(nrow(equation), 300, replace = T), ]


y3 <- rep(0, 300)
sample_equation3 <- cbind(sample_equation3, y3)
x4_me <- rep(0, 300)
sample_equation3 <- cbind(sample_equation3, x4_me)

beta_hat3 <- c()
for (i in 1:2000) {
  sample_equation3[,6] <- sample_equation3[,1] + rnorm(300,0,1)
  sample_equation3[,7] <- sample_equation3[,5] + rnorm(300,0,1) # x4_me = x4+e
  coefs <- summary(lm(sample_equation3[,1]~sample_equation3[,c(2:4,7)]+0))$coefficients[,1]
  beta_hat3<- rbind(beta_hat3, coefs)

}

beta3_means <- colMeans(beta_hat3)
beta3_means
```

```
##    sample_equation3[, c(2:4, 7)]x1    sample_equation3[, c(2:4, 7)]x2
##                         11.6384818                         -0.6997947
##    sample_equation3[, c(2:4, 7)]x3 sample_equation3[, c(2:4, 7)]x4_me
##                         33.9998042                         -0.1015480
```

```
beta
```

```
## [1] 12.00 -0.70 34.00 -0.17
```

```
print('they are close, but still have some estimation errors')
```

```
## [1] "they are close, but still have some estimation errors"
```

```
#4
sample_equation4 <- equation[sample(nrow(equation), 300, replace = T), ]
y4 <- rep(0, 300)
sample_equation4 <- cbind(sample_equation4, y4)


beta_hat4 <- c()
for (i in 1:2000) {
  sample_equation4[,6] <- sample_equation4[,1] + rnorm(300,0,1)
  coefs <- summary(lm(sample_equation4[,6] + rnorm(300,0,1) # measurement error to y
                      ~sample_equation4[,2:5]+0))$coefficients[,1]
  beta_hat4<- rbind(beta_hat4, coefs)

}

beta4_means <- colMeans(beta_hat4)
beta4_means
```

```
## sample_equation4[, 2:5]x1 sample_equation4[, 2:5]x2
##                11.9967020                -0.7000034
## sample_equation4[, 2:5]x3 sample_equation4[, 2:5]x4
##                34.0001482                -0.1693946
```

```
beta
```

```
## [1] 12.00 -0.70 34.00 -0.17
```

```
print('they are close, but still have some estimation errors')
```

```
## [1] "they are close, but still have some estimation errors"
```

```
#5
#E(beta_hat) = beta + E(cov(x4_me, u)/var(x4_me) - beta * cov(x4_me, v#error to x4_me)/var(x4_m
e))
beta_hat5<- c()
bias_collector <- c() # collect E(cov(x4_me, u)/var(x4_me) - beta * cov(x4_me, v#error to x4_m
e)/var(x4_me))
for (i in 1:2000) {
  u <- rnorm(300,0,1)
  v <- rnorm(300,0,1)
  sample_equation3[,6] <- sample_equation3[,1] + u
  sample_equation3[,7] <- sample_equation3[,5] + v
  bias_collector <- append(bias_collector,
                           cov(sample_equation3[,7], u)/ var(sample_equation3[,7]) - beta[4] * (
cov(sample_equation3[,7], v)
                                                                        / var(sample_equation3[,
7])))
  coefs <- summary(lm(sample_equation3[,1]~sample_equation3[,c(2:4,7)]+0))$coefficients[,1]
  beta_hat5<- rbind(beta_hat5, coefs)

}

beta5_means <- colMeans(beta_hat5)

beta[4] + mean(bias_collector)
```

```
## [1] -0.1009942
```

```
beta5_means[4]
```

```
## sample_equation3[, c(2:4, 7)]x4_me
##                         -0.1015519
```

```
print('yes, we confirm bias function is true')
```

```
## [1] "yes, we confirm bias function is true"
```

C

```
df <- read.dta("C:/Users/wangh/Downloads/utown.dta")
price_list <- df$price
top_25 <- sort(price_list)[length(price_list)*0.25]
df$high <- ifelse(df$price >= top_25, 1, 0)


#1

reg_lpm <- summary(lm(high~ age+sqft+fplace+utown, data = df))

reg_lpm
```

```
##
## Call:
## lm(formula = high ~ age + sqft + fplace + utown, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76579 -0.23891  0.00007  0.26160  0.79240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.026872   0.086686 -11.846   <2e-16 ***
## age         -0.001523   0.001043  -1.460    0.145
## sqft         0.061006   0.003384  18.029   <2e-16 ***
## fplace       0.013381   0.019753   0.677    0.498
## utown        0.476482   0.019661  24.235   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3104 on 995 degrees of freedom
## Multiple R-squared:  0.4874, Adjusted R-squared:  0.4853
## F-statistic: 236.5 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
print("the meaning of this lpm is that estimate is margin effect to high price
      e.g. if house is in utown then high price will have 0.47 more probability")
```

```
## [1] "the meaning of this lpm is that estimate is margin effect to high price\n      e.g. if h
ouse is in utown then high price will have 0.47 more probability"
```

```
#2a

reg_lr <- summary(glm(formula =high~ age+sqft+fplace+utown, data = df, family = binomial ))

reg_lr
```

```
##
## Call:
## glm(formula = high ~ age + sqft + fplace + utown, family = binomial,
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.58580   0.00667   0.05174   0.28960   2.85693
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -21.46083    1.80547 -11.887   <2e-16 ***
## age          -0.02487    0.01270  -1.959   0.0501 .
## sqft          0.86024    0.07138  12.052   <2e-16 ***
## fplace       -0.03611    0.25290  -0.143   0.8865
## utown         6.43714    0.49389  13.033   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1122.47  on 999  degrees of freedom
## Residual deviance:  418.58  on 995  degrees of freedom
## AIC: 428.58
##
## Number of Fisher Scoring iterations: 7
```

```
print("these coeffcients are parameters for logistic regression,defined
      as 1/1+e(-z). They have non linear effects on
      classfication of housing price ")
```

```
## [1] "these coeffcients are parameters for logistic regression,defined\n      as 1/1+e(-z). Th
ey have non linear effects on \n      classfication of housing price "
```

```
#2b
df$intercept <-  rep(1, length(df[,1]))

or <- exp(as.matrix(df[,c('intercept', 'age', 'sqft','fplace','utown')]) %*% reg_lr$coefficients
[,1])
print("odd ratios are the ratio of probability of expensive house over inexpensive one, generall
y this
      ratio the bigger the better, because higher ratio gives more confidence on prediction accu
racy")
```

```
## [1] "odd ratios are the ratio of probability of expensive house over inexpensive one, general
ly this\n      ratio the bigger the better, because higher ratio gives more confidence on predic
tion accuracy"
```

```
#3b
#applying partial derivative, we know &xi/&y is not constant, example for effect by age

marginal_effect_age <- reg_lr$coefficients['age',1] * or

#3d

probability_highprice <- 1 / (1 + 1/exp(as.matrix(df[,c('intercept', 'age', 'sqft','fplace','uto
wn')])
                                        %*% reg_lr$coefficients[,1]) )
df$prediction <- probability_highprice

#3e
# we use 0.5 as threshold of high price and low price

df$evaluation <- ifelse(df$prediction >= 0.5, 1, 0)

false_positive <- length(df[df$high == 0 & df$evaluation == 1, ]) / length(df[df$evaluation == 1
, 1 ])

paste('false positive rate is', false_positive)
```

```
## [1] "false positive rate is 0.0130548302872063"
```