**Saad Khan  
100911681**EC3133 problem set 5 (graded homework I)  
  
**Part A**

For there to be internal validity, the statistical inferences made about causal effects must be valid for earnings in the US.

Omitted variable bias is a bias in OLS that occurs when a variable that is correlated to the regressor X (earnings) is omitted. For the regression in question, the only variables included are age, the square of age, a “female” indicator, and a “bachelor” indicator. However, there are a lot more variables than can affect earnings such as the years of work experience; more years worked means higher earning.

The earnings regression function doesn’t suffer from misspecification of the functional form because the non- linear effect on age has been correctly identified by adding the age polynomial, this is evident by the graph plotted in viii for the variable “polagelinfct.male.higsch”.

There could be an error in variable bias that arises from incorrectly assuming that when age increases, “time worked” increases linearly thus equating age to time worked, but this is not the case all the time as people may take long periods of time where they don’t work such as pregnant women, which results in biased OLS estimators.

There isn’t a simultaneous causality bias because gender or earnings cannot possibly have an effect on age, revolutions of the Earth around the sun does.

An inconsistency of the OLS estimator is heteroskedasticity which results in invalid hypothesis tests and wrong confidence intervals. This inconsistency is mitigated by using robust standard errors (which is done on line 83 using the parameters package).

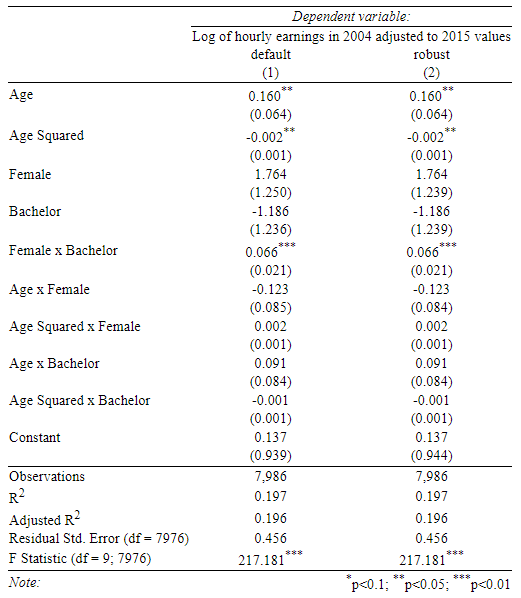
**Part B  
  
Question 1**  
  
1. Report the mean and the standard deviation of the average hourly earnings in CPS96, CPSSW04, and CPS15 measured in 2015 dollars.

|  |  |  |  |
| --- | --- | --- | --- |
| YEAR | 1996 | 2004 | 2015 |
| MEAN | $19.21 | $21.16 | $21.24 |
| STANDARD DEVIATION | $9.63 | $11.05 | $12.12 |
| T-TEST | 155.94 > 1.96 | 171.11 > 1.96 | 147.57 > 1.96 |
| P-VALUE | p < 0.05 | P < 0.05 | P < 0.05 |

What does the mean and standard deviations indicate about real earnings growth and real earnings inequality for young workers in the US over the 1996-2015 period?  
  
The mean has increased from $19.21 to $21.16 to $21.24 for the years 1996, 2004 and 2015 respectively. This indicates that there has been a growth in real earnings.  
  
The standard deviation has also increased from $9.63, $11.05, $12.12 for the years 1996, 2004 and 2015, respectively. This indicates that there has been an increase in real earnings inequality over the period.  
  
Overall, real earnings have increased but real earnings inequality has also increased at a statistically significant level.

**Question 2**

Report the estimated regression in a table.



**Question 3**

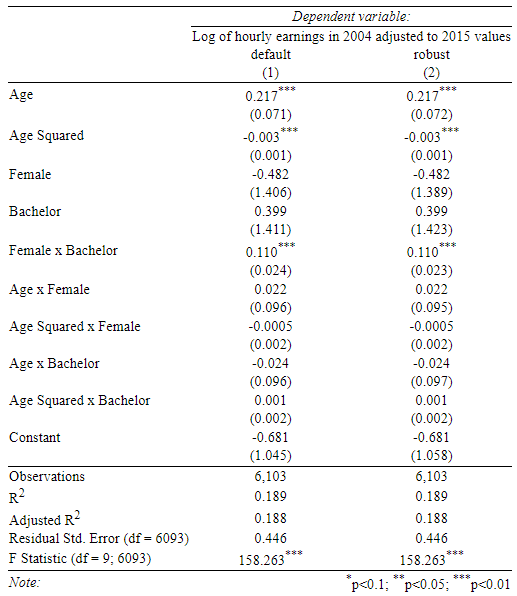
Plot the age-log-earnings profiles (for age 25-34)

Chart, line chart

Description automatically generated

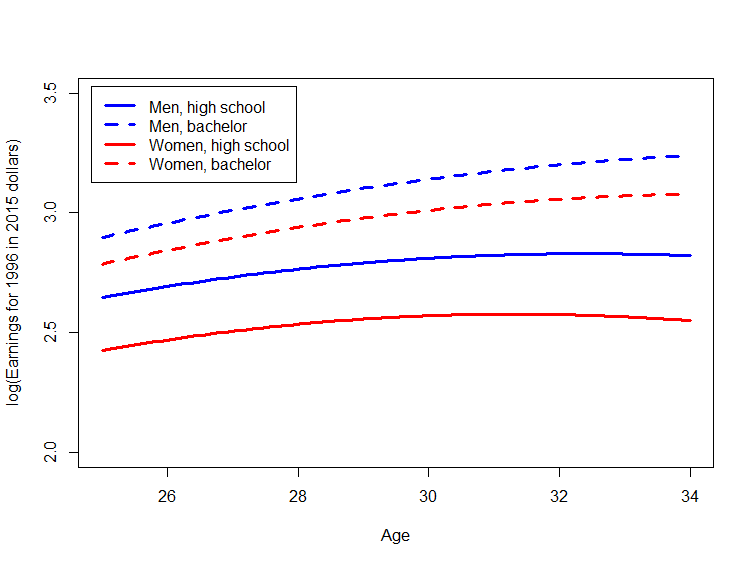
**Question 4**

Run a regression of the logarithm of hourly earnings (2015 dollars) on age

****

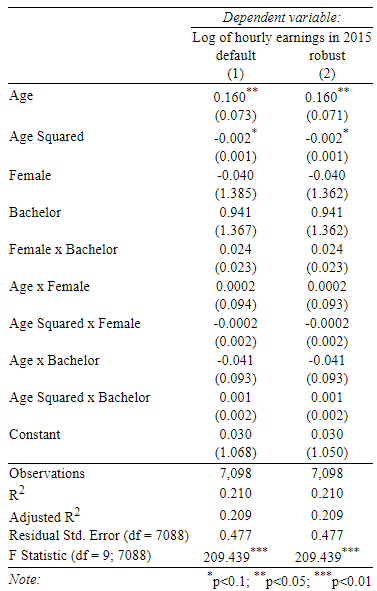
**Question 5**

Plot the age-log-earnings profiles (for age 25-34) implied by the CPS96



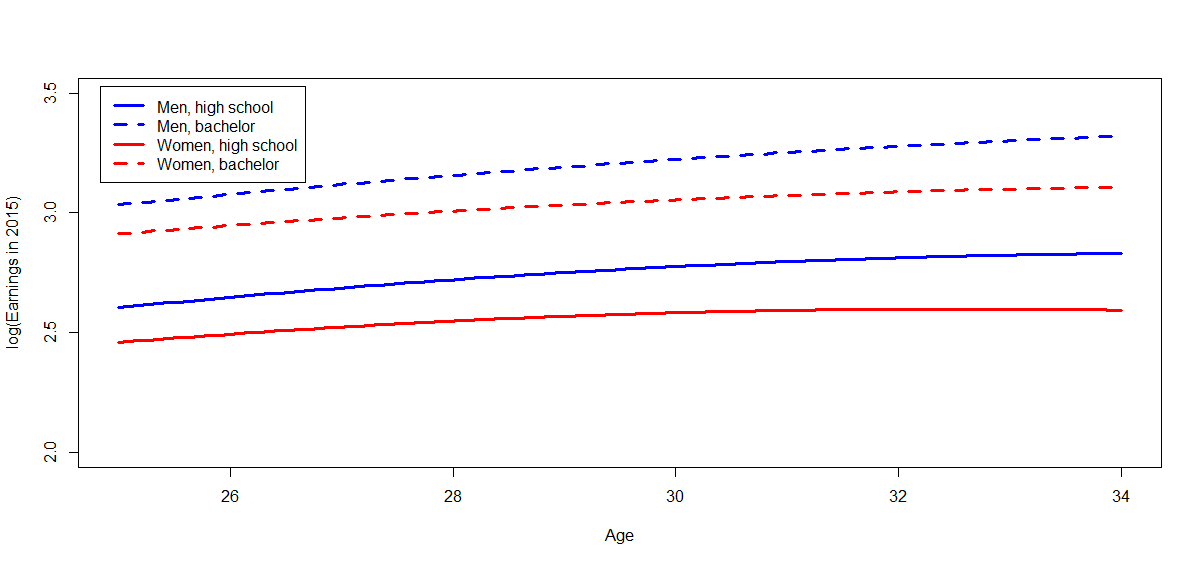
**Question 6**

Using CPS15, run a regression of the logarithm of hourly earnings (2015 dollars) on age

****

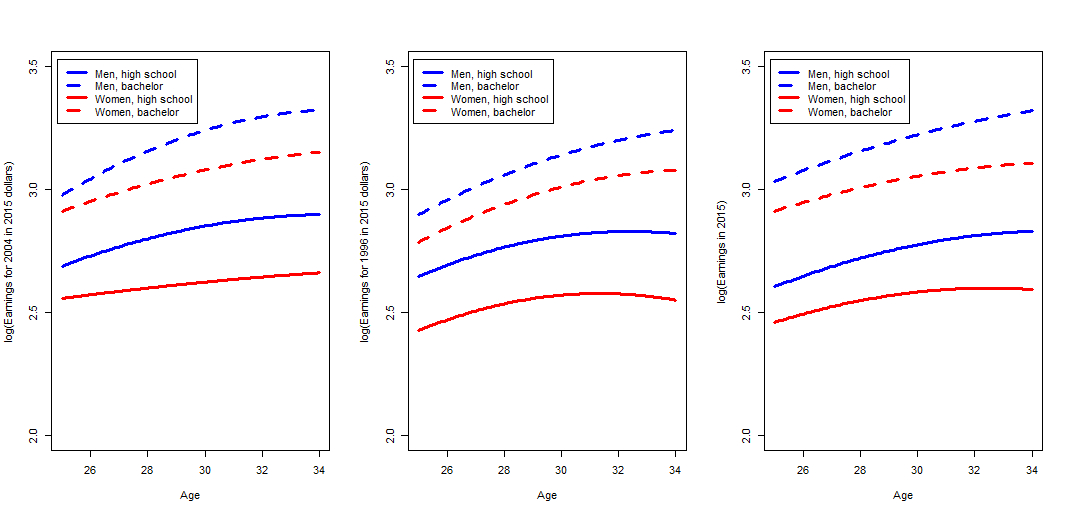
**Question 7**

Plot the age-log-earnings profiles (for age 25-34) implied by the CPS15-regression



**Question 8**

Arrange the age-log-earnings profiles (for age 25-34) implied by the CPS96-, the CPSSW04-, and the CPS15-regressions next to each other, in a single row



Summarize the effect of age on earnings for young workers, paying attention to the role of gender and education, and how these relationships have changed (or not) over time.

Firstly, commenting on the general trend, for both men and women the earning overall increased from age 25-34 for all years 1996, 2004, 2015. Moreover, it is clear to see that having a bachelor’s degree increases earnings significantly more than not only having high school education for both men and women. Men always earn more than women for the same level of education. Women earn more than men when compared to women who have degree level education and men who only have high school education.

The rate of increase slows down for the groups: men with bachelor’s degrees, women with bachelor’s degrees and men with high school education. However, the more the age increases the larger the wage gap increases between men and women. This trend does not occur for women with high school education in the data set for 1996 and 2015 as their earnings increases at first then seems to decrease. This is not the case for 2004 as the earnings for women with high school education looks constant for the age group.

What do you conclude regarding the temporal external validity of your analysis in EC3133 problem set 4?

Temporal external validity means if these results can be used to generalise for another point in time. In the US, I believe these results have little accuracy when comparing to a point in the past, for example, in the early 1900s, there was a culture where women would normally become house wives, so even if the women have a degree or high school education, it was likely that their husband would be the sole source of income in a household. This could indicate that while the earnings for men obtained from these results could be accurate in explaining the effect of age on earnings during an earlier time, the results for women do not have temporal external validity. If we were to look at temporal external validity somewhere in the near future (suppose 2020), then these results obtained can be significant because of gender discrimination and lack of gender discrimination legislation. In a further future, there may be better legislation protecting against gender wage discrimination which could result in less of a difference between earnings to age between men and women.

To conclude, the graph of the results for 1996, 2004 and 2015 (across 19 years of data) do have similarities indicating that future results could also be similar but a sufficiently far enough future can have very different results.

# R script EC3133 Problem set 5

# Author: Saad Khan

# Date: 09/11/2020

library(AER)

library(parameters)

library(readxl)

CPS96\_15 <- read\_excel("./Problem set 5/CPS96\_15.xlsx")

CPS96 <- data.frame(CPS96\_15[(CPS96\_15$year == 1996),])

CPS15 <- data.frame(CPS96\_15[(CPS96\_15$year == 2015),])

data("CPSSW04")

### Question 1

# creating variables with base year 2015

CPS96Base2015 <- (CPS96$ahe\*(233.7/154.4))

CPSSW04Base2015 <- (CPSSW04$earnings\*(233.7/185.2))

# mean

mean(CPS96Base2015)

mean(CPSSW04Base2015)

mean(CPS15$ahe)

# Standard deviation

sd(CPS96Base2015)

sd(CPSSW04Base2015)

sd(CPS15$ahe)

# T test for mean

t.test(CPS96Base2015)

t.test(CPSSW04Base2015)

t.test(CPS15$ahe)

# At the 5% level, all means are statistically significant

### Question 2

# Variable for adjusted to 2015 values

CPSSW04$CPSSW04Base2015 <- CPSSW04Base2015

# Age squared

CPSSW04$agesquared <- CPSSW04$age^2

# log of earnings 2004, adjusted to 2015

CPSSW04$logearnings\_base2015 <- log(CPSSW04$CPSSW04Base2015)

# interaction variable between female and bachelor

CPSSW04$female.x.bachelor <- (CPSSW04$gender == "female")\*(CPSSW04$degree == "bachelor")

# interaction variable between age and female

CPSSW04$age.x.female <- (CPSSW04$gender == "female")\*(CPSSW04$age)

# interaction variable between age and bachelor

CPSSW04$age.x.bachelor <- (CPSSW04$degree == "bachelor")\*(CPSSW04$age)

# interaction variable between age squared and female

CPSSW04$agesquared.x.female <- (CPSSW04$gender == "female")\*(CPSSW04$agesquared)

# interaction variable between age squared and bachelor

CPSSW04$agesquared.x.bachelor <- (CPSSW04$degree == "bachelor")\*(CPSSW04$agesquared)

# Regression

lm1 <- lm(logearnings\_base2015 ~

age +

agesquared +

gender +

degree +

female.x.bachelor +

age.x.female +

agesquared.x.female +

age.x.bachelor +

agesquared.x.bachelor,

data= CPSSW04)

# Robust standard errors

parameters(lm1, robust = TRUE, vcov\_type = "HC1")

library(sandwich)

cov <- vcovHC(lm1, type = "HC")

robust.se <- sqrt(diag(cov))

summary(lm1)

install.packages("stargazer")

library(stargazer)

# Report the estimated regression in a table (with robust standard errors)

stargazer(lm1, lm1,

se= list(NULL, robust.se),

column.labels=c("default","robust"), align=TRUE,

type = "html",

dep.var.labels=c("Log of hourly earnings in 2004 adjusted to 2015 values"),

covariate.labels = c("Age",

"Age Squared",

"Female",

"Bachelor",

"Female x Bachelor",

"Age x Female",

"Age Squared x Female",

"Age x Bachelor",

"Age Squared x Bachelor"),

out = "lm1.html")

remove(cov)

remove(robust.se)

### Question 3

regfct1.agelogearnings.male.bachelor <- function(x){lm1$coefficients["(Intercept)"] +

lm1$coefficients["age"]\*x +

lm1$coefficients["agesquared"]\*x^2 +

lm1$coefficients["degreebachelor"] +

lm1$coefficients["age.x.bachelor"]\*x +

lm1$coefficients["agesquared.x.bachelor"]\*x^2}

regfct1.agelogearnings.male.highsch <- function(x){lm1$coefficients["(Intercept)"] +

lm1$coefficients["age"]\*x +

lm1$coefficients["agesquared"]\*x^2}

regfct1.agelogearnings.female.bachelor <- function(x){lm1$coefficients["(Intercept)"] +

lm1$coefficients["age"]\*x +

lm1$coefficients["agesquared"]\*x^2 +

lm1$coefficients["genderfemale"] +

lm1$coefficients["degreebachelor"] +

lm1$coefficients["female.x.bachelor"] +

lm1$coefficients["age.x.female"]\*x +

lm1$coefficients["age.x.bachelor"]\*x +

lm1$coefficients["agesquared.x.female"]\*x^2 +

lm1$coefficients["agesquared.x.bachelor"]\*x^2}

regfct1.agelogearnings.female.highsch <- function(x){lm1$coefficients["(Intercept)"] +

lm1$coefficients["age"]\*x +

lm1$coefficients["agesquared"]\*x^2 +

lm1$coefficients["genderfemale"] +

lm1$coefficients["age.x.female"]\*x +

lm1$coefficients["agesquared.x.female"]\*x^2}

curve(regfct1.agelogearnings.male.bachelor,

from = 25, to = 34, # Evaluate in x values from 25 to 34

xlab = "Age", # x- axis label

ylab = "log(Earnings for 2004 in 2015 dollars)", # y- axis label

xlim = c(25, 34),

ylim = c(2, 3.5),

lwd = 3, # Set linewidth to 3

lty = 2,

col = "blue")

curve(regfct1.agelogearnings.male.highsch,

from = 25, to = 34,

lwd = 3,

col = "blue", # Make the plotted line blue

add = TRUE) # Add second curve to current plot

curve(regfct1.agelogearnings.female.bachelor,

from = 25, to = 34,

lwd = 3,

lty = 2,

col = "red",

add = TRUE)

curve(regfct1.agelogearnings.female.highsch,

from = 25, to = 34,

lwd = 3,

col = "red",

add = TRUE)

legend("topleft", inset = 0.02,

legend=c("Men, high school", "Men, bachelor", "Women, high school", "Women, bachelor"),

col=c("blue", "blue", "red", "red"),

lwd=c(3,3,3,3),

lty=c(1,2,1,2))

### Question 4

CPS96$CPS96Base2015 <- CPS96Base2015

# Creating variables

CPS96$logEarnings\_Base2015 <- log(CPS96$CPS96Base2015)

CPS96$agesquared <- CPS96$age^2

CPS96$female.x.bachelor <- CPS96$female\*CPS96$bachelor

CPS96$age.x.female <- CPS96$age\*CPS96$female

CPS96$age.x.bachelor <- CPS96$age\*CPS96$bachelor

CPS96$agesquared.x.female <- (CPS96$age^2)\*CPS96$female

CPS96$agesquared.x.bachelor <- (CPS96$age^2)\*CPS96$bachelor

# Regression using CPS1996 (adjusted to 2015)

lm2 <- lm(logEarnings\_Base2015 ~

age +

agesquared +

female +

bachelor +

female.x.bachelor +

age.x.female +

agesquared.x.female +

age.x.bachelor +

agesquared.x.bachelor,

data = CPS96)

summary(lm2)

# Robust standard errors

parameters(lm2, robust = TRUE, vcov\_type = "HC1")

cov <- vcovHC(lm2, type = "HC")

robust.se <- sqrt(diag(cov))

# Report the estimated regression in a table (with robust standard errors)

stargazer(lm2, lm2,

se= list(NULL, robust.se),

column.labels=c("default","robust"), align=TRUE,

type = "html",

dep.var.labels=c("Log of hourly earnings in 2004 adjusted to 2015 values"),

covariate.labels = c("Age",

"Age Squared",

"Female",

"Bachelor",

"Female x Bachelor",

"Age x Female",

"Age Squared x Female",

"Age x Bachelor",

"Age Squared x Bachelor"),

out = "lm2.html")

### Question 5

regfct2.agelogearnings.male.bachelor <- function(x){lm2$coefficients["(Intercept)"] +

lm2$coefficients["age"]\*x +

lm2$coefficients["agesquared"]\*x^2 +

lm2$coefficients["bachelor"] +

lm2$coefficients["age.x.bachelor"]\*x +

lm2$coefficients["agesquared.x.bachelor"]\*x^2}

regfct2.agelogearnings.male.highsch <- function(x){lm2$coefficients["(Intercept)"] +

lm2$coefficients["age"]\*x +

lm2$coefficients["agesquared"]\*x^2}

regfct2.agelogearnings.female.bachelor <- function(x){lm2$coefficients["(Intercept)"] +

lm2$coefficients["age"]\*x +

lm2$coefficients["agesquared"]\*x^2 +

lm2$coefficients["female"] +

lm2$coefficients["bachelor"] +

lm2$coefficients["female.x.bachelor"] +

lm2$coefficients["age.x.female"]\*x +

lm2$coefficients["age.x.bachelor"]\*x +

lm2$coefficients["agesquared.x.female"]\*x^2 +

lm2$coefficients["agesquared.x.bachelor"]\*x^2}

regfct2.agelogearnings.female.highsch <- function(x){lm2$coefficients["(Intercept)"] +

lm2$coefficients["age"]\*x +

lm2$coefficients["agesquared"]\*x^2 +

lm2$coefficients["female"] +

lm2$coefficients["age.x.female"]\*x +

lm2$coefficients["agesquared.x.female"]\*x^2}

curve(regfct2.agelogearnings.male.bachelor,

from = 25, to = 34, # Evaluate in x values from 25 to 34

xlab = "Age", # x- axis label

ylab = "log(Earnings for 1996 in 2015 dollars)", # y- axis label

xlim = c(25, 34),

ylim = c(2, 3.5),

lwd = 3, # Set linewidth to 3

lty = 2,

col = "blue")

curve(regfct2.agelogearnings.male.highsch,

from = 25, to = 34,

lwd = 3,

col = "blue", # Make the plotted line blue

add = TRUE) # Add second curve to current plot

curve(regfct2.agelogearnings.female.bachelor,

from = 25, to = 34,

lwd = 3,

lty = 2,

col = "red",

add = TRUE)

curve(regfct2.agelogearnings.female.highsch,

from = 25, to = 34,

lwd = 3,

col = "red",

add = TRUE)

legend("topleft", inset = 0.02,

legend=c("Men, high school", "Men, bachelor", "Women, high school", "Women, bachelor"),

col=c("blue", "blue", "red", "red"),

lwd=c(3,3,3,3),

lty=c(1,2,1,2))

### Question 6

CPS15$logEarnings <- log(CPS15$ahe)

CPS15$agesquared <- CPS15$age^2

CPS15$female.x.bachelor <- CPS15$female\*CPS15$bachelor

CPS15$age.x.female <- CPS15$age\*CPS15$female

CPS15$age.x.bachelor <- CPS15$age\*CPS15$bachelor

CPS15$agesquared.x.female <- (CPS15$age^2)\*CPS15$female

CPS15$agesquared.x.bachelor <- (CPS15$age^2)\*CPS15$bachelor

lm3 <- lm(logEarnings ~

age +

agesquared +

female +

bachelor +

female.x.bachelor +

age.x.female +

agesquared.x.female +

age.x.bachelor +

agesquared.x.bachelor,

data = CPS15)

summary(lm3)

# Robust standard errors

parameters(lm3, robust = TRUE, vcov\_type = "HC1")

cov <- vcovHC(lm3, type = "HC")

robust.se <- sqrt(diag(cov))

# Report the estimated regression in a table (with robust standard errors)

stargazer(lm3, lm3,

se= list(NULL, robust.se),

column.labels=c("default","robust"), align=TRUE,

type = "html",

dep.var.labels=c("Log of hourly earnings in 2015"),

covariate.labels = c("Age",

"Age Squared",

"Female",

"Bachelor",

"Female x Bachelor",

"Age x Female",

"Age Squared x Female",

"Age x Bachelor",

"Age Squared x Bachelor"),

out = "lm3.html")

### Question 7

regfct3.agelogearnings.male.bachelor <- function(x){lm3$coefficients["(Intercept)"] +

lm3$coefficients["age"]\*x +

lm3$coefficients["agesquared"]\*x^2 +

lm3$coefficients["bachelor"] +

lm3$coefficients["age.x.bachelor"]\*x +

lm3$coefficients["agesquared.x.bachelor"]\*x^2}

regfct3.agelogearnings.male.highsch <- function(x){lm3$coefficients["(Intercept)"] +

lm3$coefficients["age"]\*x +

lm3$coefficients["agesquared"]\*x^2}

regfct3.agelogearnings.female.bachelor <- function(x){lm3$coefficients["(Intercept)"] +

lm3$coefficients["age"]\*x +

lm3$coefficients["agesquared"]\*x^2 +

lm3$coefficients["female"] +

lm3$coefficients["bachelor"] +

lm3$coefficients["female.x.bachelor"] +

lm3$coefficients["age.x.female"]\*x +

lm3$coefficients["age.x.bachelor"]\*x +

lm3$coefficients["agesquared.x.female"]\*x^2 +

lm3$coefficients["agesquared.x.bachelor"]\*x^2}

regfct3.agelogearnings.female.highsch <- function(x){lm3$coefficients["(Intercept)"] +

lm3$coefficients["age"]\*x +

lm3$coefficients["agesquared"]\*x^2 +

lm3$coefficients["female"] +

lm3$coefficients["age.x.female"]\*x +

lm3$coefficients["agesquared.x.female"]\*x^2}

curve(regfct3.agelogearnings.male.bachelor,

from = 25, to = 34, # Evaluate in x values from 25 to 34

xlab = "Age", # x- axis label

ylab = "log(Earnings in 2015)", # y- axis label

xlim = c(25, 34),

ylim = c(2, 3.5),

lwd = 3, # Set linewidth to 3

lty = 2,

col = "blue")

curve(regfct3.agelogearnings.male.highsch,

from = 25, to = 34,

lwd = 3,

col = "blue", # Make the plotted line blue

add = TRUE) # Add second curve to current plot

curve(regfct3.agelogearnings.female.bachelor,

from = 25, to = 34,

lwd = 3,

lty = 2,

col = "red",

add = TRUE)

curve(regfct3.agelogearnings.female.highsch,

from = 25, to = 34,

lwd = 3,

col = "red",

add = TRUE)

legend("topleft", inset = 0.02,

legend=c("Men, high school", "Men, bachelor", "Women, high school", "Women, bachelor"),

col=c("blue", "blue", "red", "red"),

lwd=c(3,3,3,3),

lty=c(1,2,1,2))

### Question 8

# Displaying the graphs next to each other

par(mfrow=c(1,3))

# Graph 1

curve(regfct1.agelogearnings.male.bachelor,

from = 25, to = 34, # Evaluate in x values from 25 to 34

xlab = "Age", # x- axis label

ylab = "log(Earnings for 2004 in 2015 dollars)", # y- axis label

xlim = c(25, 34),

ylim = c(2, 3.5),

lwd = 3, # Set linewidth to 3

lty = 2,

col = "blue")

curve(regfct1.agelogearnings.male.highsch,

from = 25, to = 34,

lwd = 3,

col = "blue", # Make the plotted line blue

add = TRUE) # Add second curve to current plot

curve(regfct1.agelogearnings.female.bachelor,

from = 25, to = 34,

lwd = 3,

lty = 2,

col = "red",

add = TRUE)

curve(regfct1.agelogearnings.female.highsch,

from = 25, to = 34,

lwd = 3,

col = "red",

add = TRUE)

legend("topleft", inset = 0.02,

legend=c("Men, high school", "Men, bachelor", "Women, high school", "Women, bachelor"),

col=c("blue", "blue", "red", "red"),

lwd=c(3,3,3,3),

lty=c(1,2,1,2))

# Graph 2

curve(regfct2.agelogearnings.male.bachelor,

from = 25, to = 34, # Evaluate in x values from 25 to 34

xlab = "Age", # x- axis label

ylab = "log(Earnings for 1996 in 2015 dollars)", # y- axis label

xlim = c(25, 34),

ylim = c(2, 3.5),

lwd = 3, # Set linewidth to 3

lty = 2,

col = "blue")

curve(regfct2.agelogearnings.male.highsch,

from = 25, to = 34,

lwd = 3,

col = "blue", # Make the plotted line blue

add = TRUE) # Add second curve to current plot

curve(regfct2.agelogearnings.female.bachelor,

from = 25, to = 34,

lwd = 3,

lty = 2,

col = "red",

add = TRUE)

curve(regfct2.agelogearnings.female.highsch,

from = 25, to = 34,

lwd = 3,

col = "red",

add = TRUE)

legend("topleft", inset = 0.02,

legend=c("Men, high school", "Men, bachelor", "Women, high school", "Women, bachelor"),

col=c("blue", "blue", "red", "red"),

lwd=c(3,3,3,3),

lty=c(1,2,1,2))

# Graph 3

curve(regfct3.agelogearnings.male.bachelor,

from = 25, to = 34, # Evaluate in x values from 25 to 34

xlab = "Age", # x- axis label

ylab = "log(Earnings in 2015)", # y- axis label

xlim = c(25, 34),

ylim = c(2, 3.5),

lwd = 3, # Set linewidth to 3

lty = 2,

col = "blue")

curve(regfct3.agelogearnings.male.highsch,

from = 25, to = 34,

lwd = 3,

col = "blue", # Make the plotted line blue

add = TRUE) # Add second curve to current plot

curve(regfct3.agelogearnings.female.bachelor,

from = 25, to = 34,

lwd = 3,

lty = 2,

col = "red",

add = TRUE)

curve(regfct3.agelogearnings.female.highsch,

from = 25, to = 34,

lwd = 3,

col = "red",

add = TRUE)

legend("topleft", inset = 0.02,

legend=c("Men, high school", "Men, bachelor", "Women, high school", "Women, bachelor"),

col=c("blue", "blue", "red", "red"),

lwd=c(3,3,3,3),

lty=c(1,2,1,2))