Homework ,Lab-5

Thakur Prasad Ghimire

10/21/2021

Study group

Thakur Prasad Ghimire Hassan Fayyaz

knitr::opts\_chunk$set(cache=TRUE)  
  
rm(list=ls(all=TRUE))  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(AER)

## Loading required package: car

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## Loading required package: lmtest

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

library(class)  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:survival':  
##   
## cluster

library(ggplot2)

load("C:/Homework EcoB2000/Homework,Lab-5/acs2017\_ny\_data.RData")  
  
attach(acs2017\_ny)

## The following object is masked from package:survival:  
##   
## veteran

use\_varb <- (AGE >= 25) & (AGE <= 65) & (LABFORCE == 2) & (WKSWORK2 > 4) & (UHRSWORK >= 35) & (Hispanic == 1) & (female == 1) & ((educ\_college == 1) | (educ\_advdeg == 1))  
dat\_use <- subset(acs2017\_ny,use\_varb)

Model 1 shows the linear regression of income wages based on education level, age, and ethnicity.

model1 <- lm(INCWAGE ~ AGE + educ\_hs + educ\_college + educ\_advdeg + white + AfAm + Hispanic+Asian,data=dat\_use)  
summary(model1)

##   
## Call:  
## lm(formula = INCWAGE ~ AGE + educ\_hs + educ\_college + educ\_advdeg +   
## white + AfAm + Hispanic + Asian, data = dat\_use)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -102074 -27393 -8338 14774 566409   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 51266.6 6959.0 7.367 3.15e-13 \*\*\*  
## AGE 597.8 147.6 4.050 5.43e-05 \*\*\*  
## educ\_hs NA NA NA NA   
## educ\_college -21205.6 3263.1 -6.499 1.16e-10 \*\*\*  
## educ\_advdeg NA NA NA NA   
## white 13148.6 3436.7 3.826 0.000137 \*\*\*  
## AfAm 8038.5 5888.9 1.365 0.172488   
## Hispanic NA NA NA NA   
## Asian 22023.1 15219.6 1.447 0.148139   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 56030 on 1260 degrees of freedom  
## Multiple R-squared: 0.06118, Adjusted R-squared: 0.05745   
## F-statistic: 16.42 on 5 and 1260 DF, p-value: 9.994e-16

stargazer(model1,type="text")

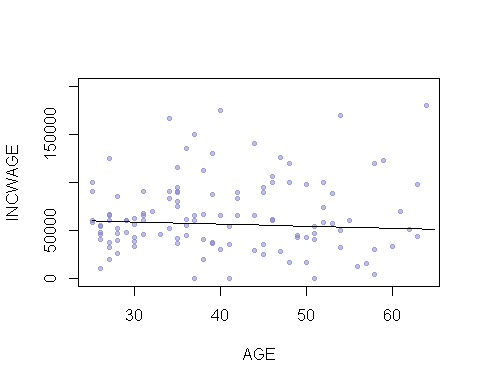
##   
## =================================================  
## Dependent variable:   
## -----------------------------  
## INCWAGE   
## -------------------------------------------------  
## AGE -220.100\*\*\*   
## (8.129)   
##   
## educ\_hs -70.337   
## (381.292)   
##   
## educ\_college 32,983.840\*\*\*   
## (449.419)   
##   
## educ\_advdeg 55,733.840\*\*\*   
## (495.819)   
##   
## white 5,317.250\*\*\*   
## (677.571)   
##   
## AfAm -2,161.436\*\*\*   
## (783.684)   
##   
## Hispanic -3,004.905\*\*\*   
## (542.919)   
##   
## Asian 1,372.544   
## (854.081)   
##   
## Constant 27,046.650\*\*\*   
## (786.352)   
##   
## -------------------------------------------------  
## Observations 163,158   
## R2 0.112   
## Adjusted R2 0.112   
## Residual Std. Error 62,368.800 (df = 163149)   
## F Statistic 2,562.620\*\*\* (df = 8; 163149)  
## =================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

attach(dat\_use)

## The following objects are masked from acs2017\_ny:  
##   
## AfAm, AGE, Amindian, ANCESTR1, ANCESTR1D, ANCESTR2, ANCESTR2D,  
## Asian, below\_150poverty, below\_200poverty, below\_povertyline, BPL,  
## BPLD, BUILTYR2, CITIZEN, CLASSWKR, CLASSWKRD, Commute\_bus,  
## Commute\_car, Commute\_other, Commute\_rail, Commute\_subway, COSTELEC,  
## COSTFUEL, COSTGAS, COSTWATR, DEGFIELD, DEGFIELD2, DEGFIELD2D,  
## DEGFIELDD, DEPARTS, EDUC, educ\_advdeg, educ\_college, educ\_hs,  
## educ\_nohs, educ\_somecoll, EDUCD, EMPSTAT, EMPSTATD, FAMSIZE,  
## female, foodstamps, FOODSTMP, FTOTINC, FUELHEAT, GQ,  
## has\_AnyHealthIns, has\_PvtHealthIns, HCOVANY, HCOVPRIV, HHINCOME,  
## Hisp\_Cuban, Hisp\_DomR, Hisp\_Mex, Hisp\_PR, HISPAN, HISPAND,  
## Hispanic, in\_Bronx, in\_Brooklyn, in\_Manhattan, in\_Nassau, in\_NYC,  
## in\_Queens, in\_StatenI, in\_Westchester, INCTOT, INCWAGE, IND,  
## LABFORCE, LINGISOL, MARST, MIGCOUNTY1, MIGPLAC1, MIGPUMA1,  
## MIGRATE1, MIGRATE1D, MORTGAGE, NCHILD, NCHLT5, OCC, OWNCOST,  
## OWNERSHP, OWNERSHPD, POVERTY, PUMA, PWPUMA00, RACE, race\_oth,  
## RACED, RELATE, RELATED, RENT, ROOMS, SCHOOL, SEX, SSMC, TRANTIME,  
## TRANWORK, UHRSWORK, UNITSSTR, unmarried, veteran, VETSTAT,  
## VETSTATD, white, WKSWORK2, YRSUSA1

## The following object is masked from package:survival:  
##   
## veteran

NNobs <- length(INCWAGE)  
set.seed(12345)  
graph\_obs <- (runif(NNobs) < 0.1)  
dat\_graph <-subset(dat\_use,graph\_obs)   
plot(INCWAGE ~ AGE, pch = 20, col = rgb(0.5, 0.5, 0.8, alpha = 0.5), ylim = c(0,200000), data = dat\_graph)  
to\_be\_predicted1 <- data.frame(AGE = 25:65, female = 1,white = 1, Hispanic = 0, AfAm =0,Asian=0, educ\_hs = 0, educ\_college = 1, educ\_advdeg = 0)  
to\_be\_predicted1$yhat <- predict(model1, newdata = to\_be\_predicted1)  
  
lines(yhat ~ AGE, data = to\_be\_predicted1)



The linear regression of income earnings depending on education level, age, age squared, and ethnicity is depicted in model 2.

model2 <- lm(INCWAGE ~ AGE + I(AGE^2) + educ\_hs + educ\_college + educ\_advdeg + white + AfAm + Hispanic+Asian,data=dat\_use)  
summary(model2)

##   
## Call:  
## lm(formula = INCWAGE ~ AGE + I(AGE^2) + educ\_hs + educ\_college +   
## educ\_advdeg + white + AfAm + Hispanic + Asian, data = dat\_use)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -97216 -27115 -7640 14724 567415   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -32850.93 25317.52 -1.298 0.19468   
## AGE 4788.68 1222.08 3.918 9.39e-05 \*\*\*  
## I(AGE^2) -49.21 14.25 -3.454 0.00057 \*\*\*  
## educ\_hs NA NA NA NA   
## educ\_college -20303.75 3259.51 -6.229 6.38e-10 \*\*\*  
## educ\_advdeg NA NA NA NA   
## white 13574.75 3424.10 3.964 7.77e-05 \*\*\*  
## AfAm 7762.52 5864.09 1.324 0.18583   
## Hispanic NA NA NA NA   
## Asian 22289.23 15154.20 1.471 0.14159   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 55790 on 1259 degrees of freedom  
## Multiple R-squared: 0.06999, Adjusted R-squared: 0.06556   
## F-statistic: 15.79 on 6 and 1259 DF, p-value: < 2.2e-16

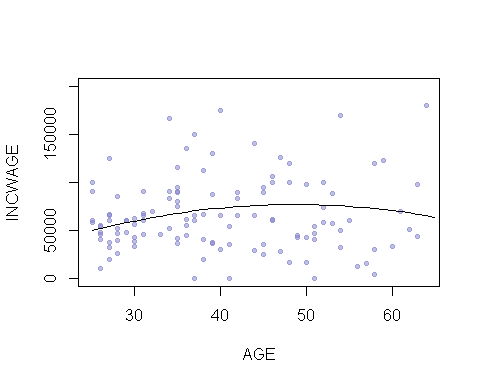
stargazer(model2, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## INCWAGE   
## -----------------------------------------------  
## AGE 4,788.684\*\*\*   
## (1,222.081)   
##   
## I(AGE2) -49.210\*\*\*   
## (14.246)   
##   
## educ\_hs   
##   
##   
## educ\_college -20,303.750\*\*\*   
## (3,259.507)   
##   
## educ\_advdeg   
##   
##   
## white 13,574.750\*\*\*   
## (3,424.096)   
##   
## AfAm 7,762.525   
## (5,864.088)   
##   
## Hispanic   
##   
##   
## Asian 22,289.230   
## (15,154.200)   
##   
## Constant -32,850.930   
## (25,317.530)   
##   
## -----------------------------------------------  
## Observations 1,266   
## R2 0.070   
## Adjusted R2 0.066   
## Residual Std. Error 55,787.020 (df = 1259)   
## F Statistic 15.792\*\*\* (df = 6; 1259)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

NNobs <- length(INCWAGE)  
set.seed(12345)  
graph\_obs <- (runif(NNobs) < 0.1)  
dat\_graph <-subset(dat\_use,graph\_obs)   
  
plot(INCWAGE ~ AGE, pch = 20, col = rgb(0.5, 0.5, 0.8, alpha = 0.5), ylim = c(0,200000), data = dat\_graph)  
  
to\_be\_predicted2 <- data.frame(AGE = 25:65, female = 1, white = 1, Hispanic = 0, AfAm = 0,Asian=0,educ\_hs = 0, educ\_college = 1, educ\_advdeg = 0)  
to\_be\_predicted2$yhat <- predict(model2, newdata = to\_be\_predicted2)

## Warning in predict.lm(model2, newdata = to\_be\_predicted2): prediction from a  
## rank-deficient fit may be misleading

lines(yhat ~ AGE, data = to\_be\_predicted2)



Model 3 depicts the linear regression of income earnings depending on educational attainment as well as other variables such as race, gender, and ethnicity.

model3 <- lm(INCWAGE ~ log(AGE) + I(log(AGE^2)) + educ\_hs + educ\_college + educ\_advdeg + white + AfAm + Hispanic+Asian,data=dat\_use)  
summary(model3)

##   
## Call:  
## lm(formula = INCWAGE ~ log(AGE) + I(log(AGE^2)) + educ\_hs + educ\_college +   
## educ\_advdeg + white + AfAm + Hispanic + Asian, data = dat\_use)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -101248 -27035 -7805 14657 566447   
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -22818 22242 -1.026 0.305142   
## log(AGE) 26762 5967 4.485 7.96e-06 \*\*\*  
## I(log(AGE^2)) NA NA NA NA   
## educ\_hs NA NA NA NA   
## educ\_college -20929 3261 -6.417 1.96e-10 \*\*\*  
## educ\_advdeg NA NA NA NA   
## white 13189 3431 3.844 0.000127 \*\*\*  
## AfAm 7949 5881 1.352 0.176714   
## Hispanic NA NA NA NA   
## Asian 22156 15198 1.458 0.145128   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 55950 on 1260 degrees of freedom  
## Multiple R-squared: 0.0639, Adjusted R-squared: 0.06018   
## F-statistic: 17.2 on 5 and 1260 DF, p-value: < 2.2e-16

stargazer(model3, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## INCWAGE   
## -----------------------------------------------  
## log(AGE) 26,761.560\*\*\*   
## (5,967.058)   
##   
## I(log(AGE2))   
##   
##   
## educ\_hs   
##   
##   
## educ\_college -20,929.280\*\*\*   
## (3,261.427)   
##   
## educ\_advdeg   
##   
##   
## white 13,189.300\*\*\*   
## (3,431.284)   
##   
## AfAm 7,948.740   
## (5,880.557)   
##   
## Hispanic   
##   
##   
## Asian 22,156.130   
## (15,197.680)   
##   
## Constant -22,817.610   
## (22,241.830)   
##   
## -----------------------------------------------  
## Observations 1,266   
## R2 0.064   
## Adjusted R2 0.060   
## Residual Std. Error 55,947.290 (df = 1260)   
## F Statistic 17.202\*\*\* (df = 5; 1260)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

attach(dat\_use)

## The following objects are masked from dat\_use (pos = 3):  
##   
## AfAm, AGE, Amindian, ANCESTR1, ANCESTR1D, ANCESTR2, ANCESTR2D,  
## Asian, below\_150poverty, below\_200poverty, below\_povertyline, BPL,  
## BPLD, BUILTYR2, CITIZEN, CLASSWKR, CLASSWKRD, Commute\_bus,  
## Commute\_car, Commute\_other, Commute\_rail, Commute\_subway, COSTELEC,  
## COSTFUEL, COSTGAS, COSTWATR, DEGFIELD, DEGFIELD2, DEGFIELD2D,  
## DEGFIELDD, DEPARTS, EDUC, educ\_advdeg, educ\_college, educ\_hs,  
## educ\_nohs, educ\_somecoll, EDUCD, EMPSTAT, EMPSTATD, FAMSIZE,  
## female, foodstamps, FOODSTMP, FTOTINC, FUELHEAT, GQ,  
## has\_AnyHealthIns, has\_PvtHealthIns, HCOVANY, HCOVPRIV, HHINCOME,  
## Hisp\_Cuban, Hisp\_DomR, Hisp\_Mex, Hisp\_PR, HISPAN, HISPAND,  
## Hispanic, in\_Bronx, in\_Brooklyn, in\_Manhattan, in\_Nassau, in\_NYC,  
## in\_Queens, in\_StatenI, in\_Westchester, INCTOT, INCWAGE, IND,  
## LABFORCE, LINGISOL, MARST, MIGCOUNTY1, MIGPLAC1, MIGPUMA1,  
## MIGRATE1, MIGRATE1D, MORTGAGE, NCHILD, NCHLT5, OCC, OWNCOST,  
## OWNERSHP, OWNERSHPD, POVERTY, PUMA, PWPUMA00, RACE, race\_oth,  
## RACED, RELATE, RELATED, RENT, ROOMS, SCHOOL, SEX, SSMC, TRANTIME,  
## TRANWORK, UHRSWORK, UNITSSTR, unmarried, veteran, VETSTAT,  
## VETSTATD, white, WKSWORK2, YRSUSA1

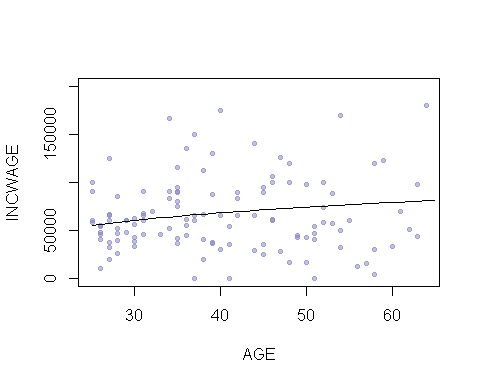
## The following objects are masked from acs2017\_ny:  
##   
## AfAm, AGE, Amindian, ANCESTR1, ANCESTR1D, ANCESTR2, ANCESTR2D,  
## Asian, below\_150poverty, below\_200poverty, below\_povertyline, BPL,  
## BPLD, BUILTYR2, CITIZEN, CLASSWKR, CLASSWKRD, Commute\_bus,  
## Commute\_car, Commute\_other, Commute\_rail, Commute\_subway, COSTELEC,  
## COSTFUEL, COSTGAS, COSTWATR, DEGFIELD, DEGFIELD2, DEGFIELD2D,  
## DEGFIELDD, DEPARTS, EDUC, educ\_advdeg, educ\_college, educ\_hs,  
## educ\_nohs, educ\_somecoll, EDUCD, EMPSTAT, EMPSTATD, FAMSIZE,  
## female, foodstamps, FOODSTMP, FTOTINC, FUELHEAT, GQ,  
## has\_AnyHealthIns, has\_PvtHealthIns, HCOVANY, HCOVPRIV, HHINCOME,  
## Hisp\_Cuban, Hisp\_DomR, Hisp\_Mex, Hisp\_PR, HISPAN, HISPAND,  
## Hispanic, in\_Bronx, in\_Brooklyn, in\_Manhattan, in\_Nassau, in\_NYC,  
## in\_Queens, in\_StatenI, in\_Westchester, INCTOT, INCWAGE, IND,  
## LABFORCE, LINGISOL, MARST, MIGCOUNTY1, MIGPLAC1, MIGPUMA1,  
## MIGRATE1, MIGRATE1D, MORTGAGE, NCHILD, NCHLT5, OCC, OWNCOST,  
## OWNERSHP, OWNERSHPD, POVERTY, PUMA, PWPUMA00, RACE, race\_oth,  
## RACED, RELATE, RELATED, RENT, ROOMS, SCHOOL, SEX, SSMC, TRANTIME,  
## TRANWORK, UHRSWORK, UNITSSTR, unmarried, veteran, VETSTAT,  
## VETSTATD, white, WKSWORK2, YRSUSA1

## The following object is masked from package:survival:  
##   
## veteran

NNobs <- length(INCWAGE)  
set.seed(12345)  
graph\_obs <- (runif(NNobs) < 0.1)  
dat\_graph <-subset(dat\_use,graph\_obs)   
  
plot(INCWAGE ~ (AGE), pch = 20, col = rgb(0.5, 0.5, 0.8, alpha = 0.5), ylim = c(0,200000), data = dat\_graph)  
  
to\_be\_predicted3 <- data.frame(AGE = 25:65, female = 1, white = 1, Hispanic =0, AfAm = 0,Asian=0,educ\_hs = 0, educ\_college = 1, educ\_advdeg = 0)  
to\_be\_predicted3$yhat <- predict(model3, newdata = to\_be\_predicted3)

## Warning in predict.lm(model3, newdata = to\_be\_predicted3): prediction from a  
## rank-deficient fit may be misleading

lines(yhat ~ AGE, data = to\_be\_predicted3)



Model 4 depicts a linear regression of income salaries depending on educational attainment, age, age polynomials (Ageexp 2,3,4), and ethnicity.

model4 <- lm(INCWAGE ~ AGE + I(AGE^2) + I(AGE^3) + I(AGE^4) +I(AGE^5) + I(AGE^6) + educ\_hs + educ\_college+ educ\_advdeg + white + AfAm + Hispanic+Asian,data=dat\_use)  
  
summary(model4)

##   
## Call:  
## lm(formula = INCWAGE ~ AGE + I(AGE^2) + I(AGE^3) + I(AGE^4) +   
## I(AGE^5) + I(AGE^6) + educ\_hs + educ\_college + educ\_advdeg +   
## white + AfAm + Hispanic + Asian, data = dat\_use)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -96297 -26906 -6714 14512 562394   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.552e+06 7.173e+06 0.216 0.828680   
## AGE -1.311e+05 1.068e+06 -0.123 0.902277   
## I(AGE^2) 1.589e+03 6.491e+04 0.024 0.980474   
## I(AGE^3) 1.690e+02 2.064e+03 0.082 0.934734   
## I(AGE^4) -6.900e+00 3.622e+01 -0.191 0.848937   
## I(AGE^5) 9.894e-02 3.329e-01 0.297 0.766350   
## I(AGE^6) -5.002e-04 1.253e-03 -0.399 0.689771   
## educ\_hs NA NA NA NA   
## educ\_college -2.011e+04 3.293e+03 -6.109 1.33e-09 \*\*\*  
## educ\_advdeg NA NA NA NA   
## white 1.329e+04 3.421e+03 3.885 0.000108 \*\*\*  
## AfAm 8.031e+03 5.846e+03 1.374 0.169796   
## Hispanic NA NA NA NA   
## Asian 2.056e+04 1.512e+04 1.359 0.174285   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 55600 on 1255 degrees of freedom  
## Multiple R-squared: 0.07908, Adjusted R-squared: 0.07174   
## F-statistic: 10.78 on 10 and 1255 DF, p-value: < 2.2e-16

stargazer(model4, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## INCWAGE   
## -----------------------------------------------  
## AGE -131,119.300   
## (1,067,653.000)   
##   
## I(AGE2) 1,588.934   
## (64,908.870)   
##   
## I(AGE3) 169.043   
## (2,063.867)   
##   
## I(AGE4) -6.900   
## (36.219)   
##   
## I(AGE5) 0.099   
## (0.333)   
##   
## I(AGE6) -0.001   
## (0.001)   
##   
## educ\_hs   
##   
##   
## educ\_college -20,114.790\*\*\*   
## (3,292.581)   
##   
## educ\_advdeg   
##   
##   
## white 13,292.300\*\*\*   
## (3,421.430)   
##   
## AfAm 8,030.551   
## (5,846.146)   
##   
## Hispanic   
##   
##   
## Asian 20,558.540   
## (15,123.970)   
##   
## Constant 1,552,499.000   
## (7,172,901.000)   
##   
## -----------------------------------------------  
## Observations 1,266   
## R2 0.079   
## Adjusted R2 0.072   
## Residual Std. Error 55,602.290 (df = 1255)   
## F Statistic 10.776\*\*\* (df = 10; 1255)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

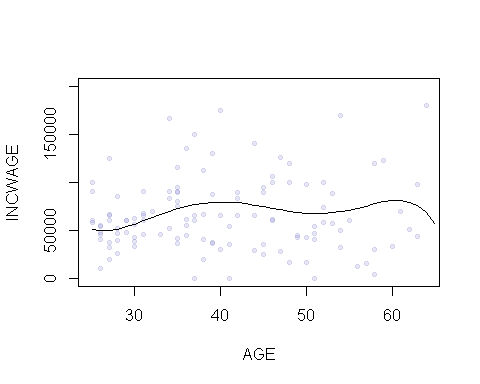
attach(dat\_use)

## The following object is masked from package:survival:  
##   
## veteran

NNobs <- length(INCWAGE)  
set.seed(12345)  
graph\_obs <- (runif(NNobs) < 0.1)  
dat\_graph <-subset(dat\_use,graph\_obs)   
  
plot(INCWAGE ~ (AGE), pch = 20, col = rgb(0.5, 0.5, 0.8, alpha = 0.2), ylim = c(0,200000), data = dat\_graph)  
to\_be\_predicted4 <- data.frame(AGE = 25:65, female = 1, white = 1, Hispanic =0, AfAm = 0,Asian=0,educ\_hs = 0, educ\_college = 1, educ\_advdeg = 0)  
to\_be\_predicted4$yhat <- predict(model4, newdata = to\_be\_predicted4)

## Warning in predict.lm(model4, newdata = to\_be\_predicted4): prediction from a  
## rank-deficient fit may be misleading

lines(yhat ~ AGE, data = to\_be\_predicted4)

 Model 4 builds on the conditions of Model 2, but this time the regression model incorporates age as a polynomial component with varying exponents (2,3,4). Using age as polynomials in the projected analysis reveals that as one gets older AGE to AGE^2 earnings also rise after then start to decline.

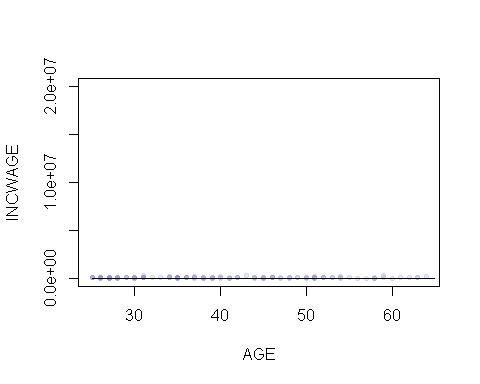
model5 <- lm(log1p(INCWAGE) ~ AGE + I(AGE^2) + I(AGE^3) + I(AGE^4) + educ\_hs + educ\_college + educ\_advdeg + white + AfAm + Hispanic+Asian,data=dat\_use)  
summary(model5)

##   
## Call:  
## lm(formula = log1p(INCWAGE) ~ AGE + I(AGE^2) + I(AGE^3) + I(AGE^4) +   
## educ\_hs + educ\_college + educ\_advdeg + white + AfAm + Hispanic +   
## Asian, data = dat\_use)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.1266 -0.1613 0.2277 0.5770 2.7908   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.660e+00 1.243e+01 0.777 0.4373   
## AGE -1.014e-01 1.231e+00 -0.082 0.9344   
## I(AGE^2) 1.142e-02 4.441e-02 0.257 0.7971   
## I(AGE^3) -2.939e-04 6.920e-04 -0.425 0.6711   
## I(AGE^4) 2.270e-06 3.938e-06 0.576 0.5644   
## educ\_hs NA NA NA NA   
## educ\_college -2.393e-01 9.711e-02 -2.464 0.0139 \*  
## educ\_advdeg NA NA NA NA   
## white 2.206e-01 1.012e-01 2.181 0.0294 \*  
## AfAm 4.213e-01 1.732e-01 2.433 0.0151 \*  
## Hispanic NA NA NA NA   
## Asian 6.784e-01 4.477e-01 1.515 0.1299   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.647 on 1257 degrees of freedom  
## Multiple R-squared: 0.02076, Adjusted R-squared: 0.01452   
## F-statistic: 3.331 on 8 and 1257 DF, p-value: 0.0008869

stargazer(model5, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## log1p(INCWAGE)   
## -----------------------------------------------  
## AGE -0.101   
## (1.231)   
##   
## I(AGE2) 0.011   
## (0.044)   
##   
## I(AGE3) -0.0003   
## (0.001)   
##   
## I(AGE4) 0.00000   
## (0.00000)   
##   
## educ\_hs   
##   
##   
## educ\_college -0.239\*\*   
## (0.097)   
##   
## educ\_advdeg   
##   
##   
## white 0.221\*\*   
## (0.101)   
##   
## AfAm 0.421\*\*   
## (0.173)   
##   
## Hispanic   
##   
##   
## Asian 0.678   
## (0.448)   
##   
## Constant 9.660   
## (12.434)   
##   
## -----------------------------------------------  
## Observations 1,266   
## R2 0.021   
## Adjusted R2 0.015   
## Residual Std. Error 1.647 (df = 1257)   
## F Statistic 3.331\*\*\* (df = 8; 1257)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

NNobs <- length(INCWAGE)  
set.seed(12345)  
graph\_obs <- (runif(NNobs) < 0.1)  
dat\_graph <-subset(dat\_use,graph\_obs)   
plot(INCWAGE ~ (AGE), pch = 20, col = rgb(0.5, 0.5, 0.8, alpha = 0.2), ylim = c(0,20000000), data = dat\_graph)  
to\_be\_predicted5 <- data.frame(AGE = 25:65, female = 1, white = 1, Hispanic =0, AfAm = 0,Asian=0,educ\_hs = 0, educ\_college = 1, educ\_advdeg = 0)  
to\_be\_predicted5$yhat <- predict(model5, newdata = to\_be\_predicted5)  
  
lines(yhat ~ AGE, data = to\_be\_predicted5)

 Model 5 expands on the conditions of model 2 but this time, the regression model is based on log of income wages which is the dependent variable. The table below demonstrates the estimates for the coefficient variables for all the models that were run above (Models 1-5).

stargazer(model1, model2, model3, model4, model5, type = "text")

##   
## =====================================================================================================================================================  
## Dependent variable:   
## ---------------------------------------------------------------------------------------------------------------------------------  
## INCWAGE log1p(INCWAGE)   
## (1) (2) (3) (4) (5)   
## -----------------------------------------------------------------------------------------------------------------------------------------------------  
## AGE -220.100\*\*\* 4,788.684\*\*\* -131,119.300 -0.101   
## (8.129) (1,222.081) (1,067,653.000) (1.231)   
##   
## I(AGE2) -49.210\*\*\* 1,588.934 0.011   
## (14.246) (64,908.870) (0.044)   
##   
## log(AGE) 26,761.560\*\*\*   
## (5,967.058)   
##   
## I(log(AGE2))   
##   
##   
## I(AGE3) 169.043 -0.0003   
## (2,063.867) (0.001)   
##   
## I(AGE4) -6.900 0.00000   
## (36.219) (0.00000)   
##   
## I(AGE5) 0.099   
## (0.333)   
##   
## I(AGE6) -0.001   
## (0.001)   
##   
## educ\_hs -70.337   
## (381.292)   
##   
## educ\_college 32,983.840\*\*\* -20,303.750\*\*\* -20,929.280\*\*\* -20,114.790\*\*\* -0.239\*\*   
## (449.419) (3,259.507) (3,261.427) (3,292.581) (0.097)   
##   
## educ\_advdeg 55,733.840\*\*\*   
## (495.819)   
##   
## white 5,317.250\*\*\* 13,574.750\*\*\* 13,189.300\*\*\* 13,292.300\*\*\* 0.221\*\*   
## (677.571) (3,424.096) (3,431.284) (3,421.430) (0.101)   
##   
## AfAm -2,161.436\*\*\* 7,762.525 7,948.740 8,030.551 0.421\*\*   
## (783.684) (5,864.088) (5,880.557) (5,846.146) (0.173)   
##   
## Hispanic -3,004.905\*\*\*   
## (542.919)   
##   
## Asian 1,372.544 22,289.230 22,156.130 20,558.540 0.678   
## (854.081) (15,154.200) (15,197.680) (15,123.970) (0.448)   
##   
## Constant 27,046.650\*\*\* -32,850.930 -22,817.610 1,552,499.000 9.660   
## (786.352) (25,317.530) (22,241.830) (7,172,901.000) (12.434)   
##   
## -----------------------------------------------------------------------------------------------------------------------------------------------------  
## Observations 163,158 1,266 1,266 1,266 1,266   
## R2 0.112 0.070 0.064 0.079 0.021   
## Adjusted R2 0.112 0.066 0.060 0.072 0.015   
## Residual Std. Error 62,368.800 (df = 163149) 55,787.020 (df = 1259) 55,947.290 (df = 1260) 55,602.290 (df = 1255) 1.647 (df = 1257)   
## F Statistic 2,562.620\*\*\* (df = 8; 163149) 15.792\*\*\* (df = 6; 1259) 17.202\*\*\* (df = 5; 1260) 10.776\*\*\* (df = 10; 1255) 3.331\*\*\* (df = 8; 1257)  
## =====================================================================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01