title: 'HW # 5' author: "Collin Rafferty" date: "10/27/2021" output: pdf_document

```
load("/cloud/project/acs2017_ny_data.RData")
attach(acs2017_ny)
use_varb <- (AGE >= 20) & (AGE <= 60) & (LABFORCE == 2) & (WKSWORK2 > 4) & (U
HRSWORK >= 40) & (AfAm== 1) & (educ_college== 1)
dat_use <- subset(acs2017_ny,use_varb)
detach()
summary(dat_use$AfAm)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1

summary(dat_use$AGE)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.0    30.0    38.0    39.5    49.0    60.0

summary(dat_use$edu_college)

## Length  Class   Mode
##      0   NULL   NULL

attach(dat_use)
lm1 <- lm((INCWAGE ~ AGE + I(AGE^2)+I(AGE^3)+I(AGE^5) +
  AfAm + female + educ_college+veteran+SSMC+NCHILD+in_Brooklyn+in_Manhattan))
summary(lm1)

##
## Call:
## lm(formula = (INCWAGE ~ AGE + I(AGE^2) + I(AGE^3) + I(AGE^5) +
##     AfAm + female + educ_college + veteran + SSMC + NCHILD +
##     in_Brooklyn + in_Manhattan))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -103141  -25407   -6472   16446  563485
##
## Coefficients: (2 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.159e+05  2.724e+05  -2.261   0.0240 *
## AGE           6.166e+04  2.718e+04   2.269   0.0235 *
## I(AGE^2)     -1.978e+03  9.350e+02  -2.116   0.0346 *
## I(AGE^3)      2.371e+01  1.171e+01   2.024   0.0432 *
## I(AGE^5)     -1.308e-03  6.798e-04  -1.924   0.0546 .
## AfAm                 NA         NA      NA       NA
## female       -4.296e+03  2.893e+03  -1.485   0.1379
## educ_college         NA         NA      NA       NA
## veteran      -1.051e+03  9.255e+03  -0.114   0.9096
## SSMC         -6.837e+03  9.293e+03  -0.736   0.4621
## NCHILD        2.115e+03  1.486e+03   1.423   0.1550
```

```
## in_Brooklyn  -2.028e+03  3.077e+03  -0.659   0.5101
## in_Manhattan  2.849e+04  6.182e+03   4.608 4.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48940 on 1185 degrees of freedom
## Multiple R-squared:  0.07176,    Adjusted R-squared:  0.06393
## F-statistic: 9.161 on 10 and 1185 DF,  p-value: 1.043e-14
```

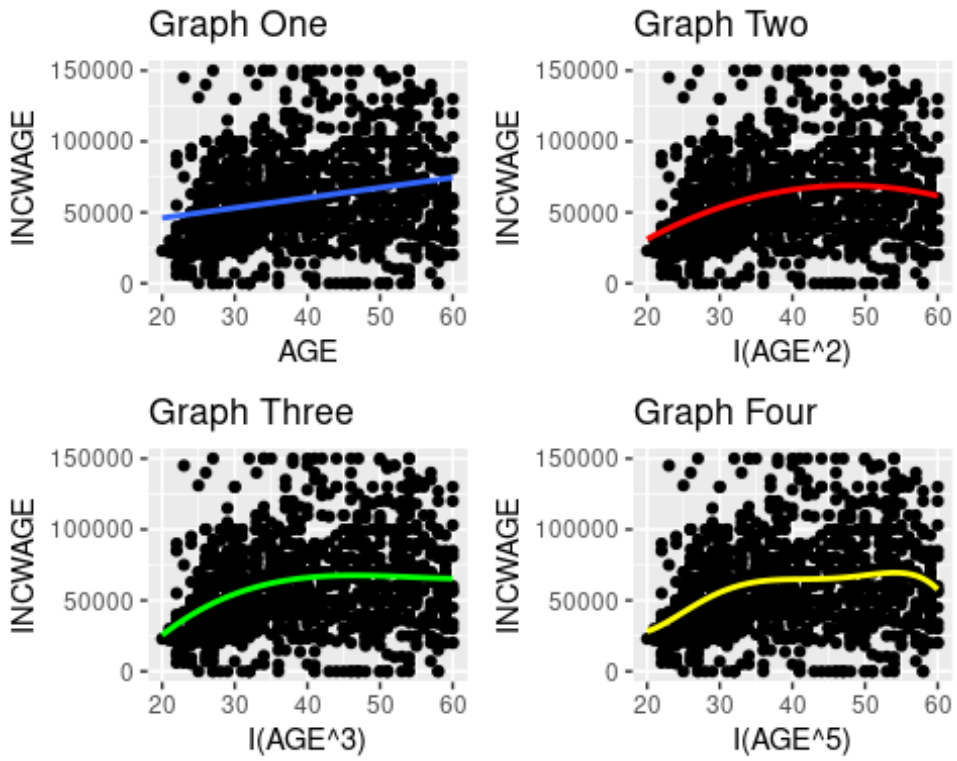
```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
qq1 <- qplot(AGE, INCWAGE,ylim = c(0,150000), data=dat_use)
g1 <-qq1 + geom_smooth(method="lm", formula = y ~ x, se=FALSE)+ ggtitle("Grap
h One")

qq2 <- qplot(AGE, INCWAGE,ylim = c(0,150000), xlab = "I(AGE^2)", data=dat_use
)
g2<- qq2 + geom_smooth(method="lm",
formula = y ~ poly(x, 2, raw=TRUE),color="red", se=FALSE) +ggtitle("Graph Two
")

qq3 <- qplot(AGE, INCWAGE,ylim = c(0,150000), xlab = "I(AGE^3)", data=dat_use
)
g3<- qq3 +geom_smooth(method="lm",
formula = y ~ poly(x, 3, raw=TRUE),color="green", se=FALSE)+ ggtitle("Graph T
hree")

qq4 <- qplot(AGE, INCWAGE,ylim = c(0,150000),xlab = "I(AGE^5)", data=dat_use)
g4<- qq4 +geom_smooth(method="lm",
formula = y ~ poly(x, 5, raw=TRUE),color="yellow", se=FALSE)+ ggtitle("Graph
Four")

grid.arrange(g1,g2,g3,g4)
```



Some relevant cases of predicted wage from model:

$\widehat{Wage}_1 = ß_0 + ß_1$SSMC + $ß_2$ In_Brooklyn + $ß_3$veteran

$\widehat{Wage}_1$= 0 + 1(-6,836) +1(-2,027)+1(-1,051)=$ -9,914

A college-educated African American male or female who served in the military lives in Brooklyn and is part of a same-sex couple makes on average $9,914 less than a college-educated African American male or female who did not serve in the military, who doesn't live in Brooklyn and is not part of a same-sex couple.

$\widehat{Wage}_2 = ß_0 + ß_1$NChild + $ß_2$ In_Manhattan + $ß_3$female

$\widehat{Wage}_2$= 0 + 1(2,115)+1(28,485)+0(-4,295)=$30,600

A college-educated African American male who lives in Manhattan and has one of her own children in the household makes on average $30,600 more than a college-educated African American female who doesn't live in Manhattan and does not have one of her own children in the household.

```
lm2 <- lm(INCWAGE ~ I(AGE^2)+I(AGE^3)+I(AGE^4)+I(AGE^5))
anova(lm1,lm2)

## Analysis of Variance Table
##
## Model 1: INCWAGE ~ AGE + I(AGE^2) + I(AGE^3) + I(AGE^5) + AfAm + female +
##     educ_college + veteran + SSMC + NCHILD + in_Brooklyn + in_Manhattan
```

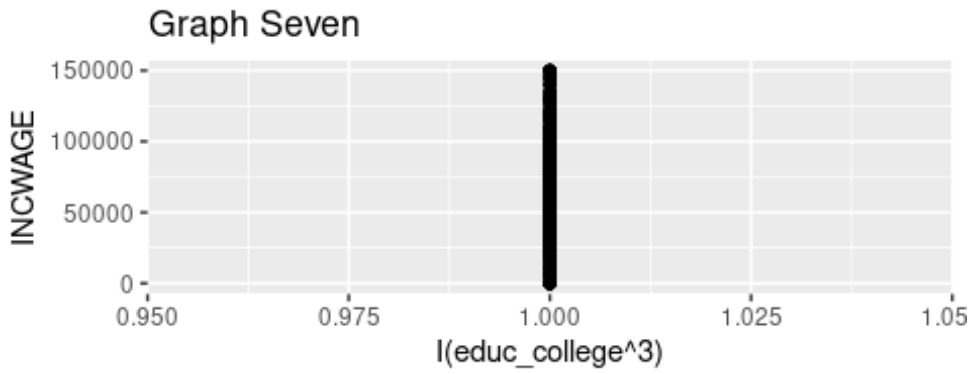
```
## Model 2: INCWAGE ~ I(AGE^2) + I(AGE^3) + I(AGE^4) + I(AGE^5)
##   Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
## 1   1185 2.8382e+12
## 2   1191 2.9077e+12 -6 -6.9494e+10 4.8358 6.891e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

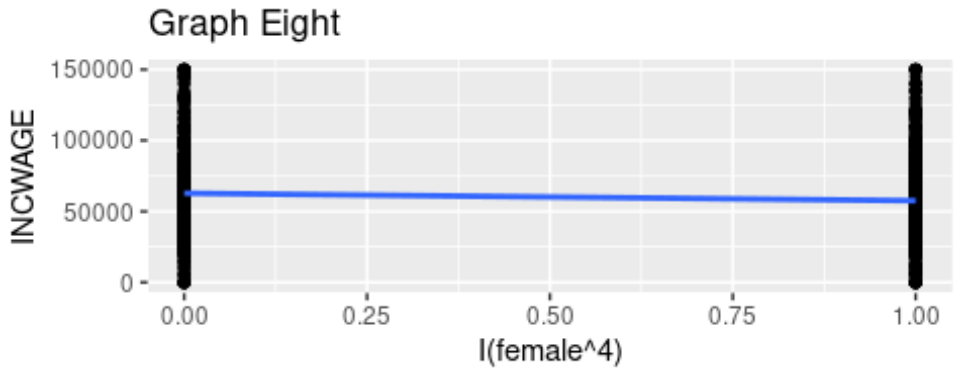
Looking at the p-value, we can see these higher-order polynomials terms are jointly highly significant, so we can reject the null. These terms do affect the regression.

```
qq5 <- qplot(log(AGE^3), INCWAGE,ylim = c(0,150000), data=dat_use)
g5 <-qq5 + geom_smooth(method="lm", formula = y ~ x, se=FALSE) +ggtitle("Grap
h Five")

qq6 <- qplot(log(AGE), INCWAGE,ylim = c(0,150000), xlab = "I(AGE^3)", data=da
t_use)
g6<- qq6 +geom_smooth(method="lm",
formula = y ~ poly(x, 3, raw=TRUE),color="green", se=FALSE) +ggtitle("Graph S
ix")
```

If you put educ_college into a regression, it splits the data set into 2 groups: those with college degrees and those individuals without college degrees. However, division creates two very homogeneous groups because it does not distinguish between different degrees. If you put both educ_college and educ_advdeg into a regression, it would create 3 groups. As you subset the data further, it gets less and less homogeneous.

```
qq7 <- qplot(educ_college, INCWAGE,ylim = c(0,150000), xlab = "I(educ_college
^3)", data=dat_use)
g7<- qq7 +geom_smooth(method="lm",
formula = y ~ poly(x, 3, raw=TRUE), se=FALSE) +ggtitle("Graph Seven")

qq8 <- qplot(female, INCWAGE,ylim = c(0,150000), xlab = "I(female^4)", data=d
at_use)
g8<- qq8 +geom_smooth(method="lm",
formula = y ~ poly(x, 4, raw=TRUE), se=FALSE) +ggtitle("Graph Eight")

grid.arrange(g7,g8)
```





The only values a dummy variable can take is 0 or 1, so by cubing the variable, you change nothing; 0^3 is still 0, and 1^3 is still 1. That is why the graphs make above are not that helpful.

```
detach()
dat_noZeroWage <- subset(dat_use,(INCWAGE > 0))


lm3 <- lm((log(INCWAGE) ~ AGE + I(AGE^2)+I(AGE^3)+I(AGE^5) +
  AfAm + female + educ_college+ +veteran+SSMC+NCHILD+in_Brooklyn+ in_Manhatta
n))

summary(lm3)

##
## Call:
## lm(formula = (log(INCWAGE) ~ AGE + I(AGE^2) + I(AGE^3) + I(AGE^5) +
##     AfAm + female + educ_college + +veteran + SSMC + NCHILD +
##     in_Brooklyn + in_Manhattan))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76831 -0.33622  0.03559  0.36141  2.30425
##
## Coefficients: (2 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.685e+00  3.305e+00  -0.812 0.416707
```

```
## AGE           1.201e+00  3.298e-01   3.641 0.000283 ***
## I(AGE^2)     -3.732e-02  1.135e-02  -3.289 0.001035 **
## I(AGE^3)      4.330e-04  1.422e-04   3.045 0.002378 **
## I(AGE^5)     -2.262e-08  8.253e-09  -2.741 0.006226 **
## AfAm                NA         NA      NA       NA
## female       -6.442e-02  3.533e-02  -1.824 0.068477 .
## educ_college         NA         NA      NA       NA
## veteran       3.394e-02  1.116e-01   0.304 0.761093
## SSMC         -5.898e-02  1.119e-01  -0.527 0.598244
## NCHILD        2.098e-02  1.816e-02   1.155 0.248319
## in_Brooklyn  -2.068e-03  3.747e-02  -0.055 0.956001
## in_Manhattan  2.418e-01  7.602e-02   3.182 0.001504 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5893 on 1154 degrees of freedom
## Multiple R-squared:  0.128,  Adjusted R-squared:  0.1205
## F-statistic: 16.94 on 10 and 1154 DF,  p-value: < 2.2e-16
```

Predicted values using log wage:

$ln(\widehat{Wage}_3) = ß_0 + ß_1 \text{SSMC} + ß_2 \text{ In\_Brooklyn} + ß_3 \text{female}$

$ln(\widehat{Wage}_3) = 0 + 1(-.06) + 1(-.003) + 1(-.06) =. -123$

A college-educated African American female who lives in Brooklyn and is part of a same-sex couple makes on average 12.3% less than a college-educated African American male who doesn't live in Brooklyn and is part of a heterosexual couple.

```
lm4 <-lm(INCWAGE ~ AGE + I(AGE^2) + female + I(female*AGE)+ I(female*(AGE^2)+
veteran+I(veteran*AGE)))
summary(lm4)

##
## Call:
## lm(formula = INCWAGE ~ AGE + I(AGE^2) + female + I(female * AGE) +
##     I(female * (AGE^2) + veteran + I(veteran * AGE)))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73504 -25469  -8815  13996 576178
##
## Coefficients:
##                                                  Estimate Std. Error t va
lue
## (Intercept)                                    -55375.06   34313.73  -1.
614
## AGE                                              5871.70    1770.22   3.
317
## I(AGE^2)                                          -63.12      21.71  -2.
908
```

```
## female                                                206.29   44399.02   0.
005
## I(female * AGE)                                      -624.67    2296.37  -0.
272
## I(female * (AGE^2) + veteran + I(veteran * AGE))     11.15      28.16   0.
396
##                                                    Pr(>|t|)
## (Intercept)                                        0.106846
## AGE                                                0.000938 ***
## I(AGE^2)                                           0.003709 **
## female                                             0.996294
## I(female * AGE)                                    0.785650
## I(female * (AGE^2) + veteran + I(veteran * AGE)) 0.692187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48740 on 1159 degrees of freedom
## Multiple R-squared:  0.05694,    Adjusted R-squared:  0.05287
## F-statistic:    14 on 5 and 1159 DF,  p-value: 2.627e-13
```

```
lm5 <-lm(INCWAGE ~ AGE + female + I(female*AGE))
summary(lm4)
```

```
##
## Call:
## lm(formula = INCWAGE ~ AGE + I(AGE^2) + female + I(female * AGE) +
##     I(female * (AGE^2) + veteran + I(veteran * AGE)))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73504 -25469  -8815  13996 576178
##
## Coefficients:
##                                                     Estimate Std. Error t va
lue
## (Intercept)                                        -55375.06   34313.73  -1.
614
## AGE                                                  5871.70    1770.22   3.
317
## I(AGE^2)                                              -63.12      21.71  -2.
908
## female                                               206.29   44399.02   0.
005
## I(female * AGE)                                      -624.67    2296.37  -0.
272
## I(female * (AGE^2) + veteran + I(veteran * AGE))     11.15      28.16   0.
396
##                                                    Pr(>|t|)
## (Intercept)                                        0.106846
## AGE                                                0.000938 ***
```

```
## I(AGE^2)                                          0.003709 **
## female                                            0.996294
## I(female * AGE)                                   0.785650
## I(female * (AGE^2) + veteran + I(veteran * AGE)) 0.692187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48740 on 1159 degrees of freedom
## Multiple R-squared:  0.05694,    Adjusted R-squared:  0.05287
## F-statistic:    14 on 5 and 1159 DF,  p-value: 2.627e-13
```

$\hat{U}_f$ = 41,481.5 +700.5Age-16,393.6female+262.1AGE*female

$\hat{U}_f$= 25,087+9,962.6AGEhat

$\hat{U}_m$= 41,481.5 +700.5AGE-16,393.6female+262.1AGE*female

$\hat{U}_m$=41,481.5 +700.5AGE

```
 plot(AGE[female==1],INCWAGE[female==1], col="red", ylim = c(0,150000), xlab=
"Gender", ylab = "INCWage",main="Wage vs Gender")
points(AGE[female==0],INCWAGE[female==0], col="blue",pch=16)
legend("topleft",legend = c("female", "male"), col = c("red","blue"), pch = c
(1, 16), bty = "n")
abline(a=25087.9,b=962.6, col="red", lwd=3)
abline(a=41481.5,b=700.5, col="blue", lwd=3)
```