Homework_5

Marjan Rezvani

10/29/2020

```
load('/Users/marjanrezvani/Documents/Fall2020/eco_stat/data/acs2017_ny/acs2017_ny_dat
a.RData')
attach(acs2017_ny)
use_varb <- (AGE >= 25) & (AGE <= 55) & (LABFORCE == 2) & (WKSWORK2 > 4) & (UHRSWORK
>= 35) & (Hispanic == 1) & (female == 1) & ((educ_college == 1) | (educ_advdeg == 1))
dat_use <- subset(acs2017_ny,use_varb)
detach()
attach(dat_use)
```

In this homework, we'll explore how to generate the Wage dataset models we saw in class. I first fit the polynomial regression model using the following command:

```
fit <- lm(INCWAGE ~ poly(AGE, 4), data = dat_use)
coef(summary(fit))
```

```
##                   Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)      69660.30   1566.007 44.482755 1.182036e-250
## poly(AGE, 4)1  257068.67  52781.633  4.870419  1.271877e-06
## poly(AGE, 4)2 -253566.44  52781.633 -4.804066  1.763702e-06
## poly(AGE, 4)3   60144.07  52781.633  1.139489  2.547407e-01
## poly(AGE, 4)4  116704.68  52781.633  2.211085  2.722942e-02
```

This syntax fits a linear model, using the lm() function, in order to predict wage using a fourth-degree polynomial in age: poly(age,4). The poly() command allows us to avoid having to write out a long formula with powers of age. The function returns a matrix whose columns are a basis of orthogonal polynomials, which essentially means that each column is a linear combination of the variables age, age^2, age^3 and age^4.

In performing a polynomial regression we must decide on the degree of the polynomial to use. One way to do this is by using hypothesis tests. I now fit models ranging from linear to a degree-5 polynomial and seek to determine the simplest model which is sufficient to explain the relationship between wage and age. We can do this using the anova() function, which performs an analysis of variance (ANOVA, using an F-test) in order to test the null hypothesis that a model M1 is sufficient to explain the data against the alternative hypothesis that a more complex model M2 is required. In order to use the anova() function, M1 and M2 must be nested models: the predictors in M1 must be a subset of the predictors in M2. In this case, I fit five different models and sequentially compare the simpler model to the more complex model:

```
fit_1 = lm(INCWAGE~AGE, data = dat_use)
fit_2 = lm(INCWAGE~poly(AGE,2), data = dat_use)
fit_3 = lm(INCWAGE~poly(AGE,3), data = dat_use)
fit_4 = lm(INCWAGE~poly(AGE,4), data = dat_use)
fit_5 = lm(INCWAGE~poly(AGE,5), data = dat_use)
print(anova(fit_1,fit_2,fit_3,fit_4,fit_5))
```

```
## Analysis of Variance Table
##
## Model 1: INCWAGE ~ AGE
## Model 2: INCWAGE ~ poly(AGE, 2)
## Model 3: INCWAGE ~ poly(AGE, 3)
## Model 4: INCWAGE ~ poly(AGE, 4)
## Model 5: INCWAGE ~ poly(AGE, 5)
##   Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
## 1   1134 3.2324e+12
## 2   1133 3.1681e+12  1 6.4296e+10 23.0620 1.779e-06 ***
## 3   1132 3.1645e+12  1 3.6173e+09  1.2975   0.25492
## 4   1131 3.1509e+12  1 1.3620e+10  4.8853   0.02729 *
## 5   1130 3.1504e+12  1 4.5349e+08  0.1627   0.68680
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value comparing the linear Model 1 to the quadratic Model 2 is essentially zero (<10–15), indicating that a linear fit is not sufficient. Similarly the p-value comparing the quadratic Model 2 to the cubic Model 3 is very low (0.0017), so the quadratic fit is also insufficient. The p-value comparing the cubic and degree-4 polynomials, Model 3 and Model 4, is approximately 0.05 while the degree-5 polynomial Model 5 seems unnecessary because its p-value is 0.37. Hence, either a cubic or a quartic polynomial appear to provide a reasonable fit to the data, but lower- or higher-order models are not justified.

we see the straight relationship between age and wage, but in the other models, it’s not going that way we expected.

Taylor series approximations tell us that pretty much any smooth function can be approximated by a polynomial, so including terms like x2or x3(where x is age) let us estimate the coefficients for the approximation for a known or unknown non-linear function of x, or age in our case. Testing these coefficients is also a simple way to test if the relationship is reasonably linear or if non-linear terms will give a better fit.

adding the square of the variable allows us to model more accurately the effect of age, which may have a non-linear relationship with the independent variable. For instance, the effect of age could be positive up until, say, the age of 55, and then negative thereafter.

we can also use anova() to compare the other models using different subset and variables:

```r
attach(acs2017_ny)
```

```
## The following objects are masked from dat_use:
##
##     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
```

```r
use_varb_1 <- (AGE >= 25) & (AGE <= 55) & (LABFORCE == 2) & (WKSWORK2 > 4) & (UHRSWOR
K >= 35) & (in_Westchester == 1) & (Commute_car == 1) & (female == 1) & ((educ_colleg
e == 1) | (educ_advdeg == 1))
dat_use_1 <- subset(acs2017_ny,use_varb_1)
detach()
attach(dat_use_1)
```

```
## The following objects are masked from dat_use:
##
##     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
```

```
fit_6 = lm(INCWAGE~EDUC+AGE, data = dat_use_1)
fit_7 = lm(INCWAGE~EDUC+poly(AGE,2), data = dat_use_1)
fit_8 = lm(INCWAGE~EDUC+poly(AGE,3), data = dat_use_1)
print(anova(fit_1,fit_2,fit_3))
```

```
## Analysis of Variance Table
##
## Model 1: INCWAGE ~ AGE
## Model 2: INCWAGE ~ poly(AGE, 2)
## Model 3: INCWAGE ~ poly(AGE, 3)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1134 3.2324e+12
## 2   1133 3.1681e+12  1 6.4296e+10 23.000 1.836e-06 ***
## 3   1132 3.1645e+12  1 3.6173e+09  1.294    0.2556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

there are clear relationships between wage and age power 2, and education. Now we will find out if they are nonlinear or not.

```
fit_9 <- lm(INCWAGE ~ EDUC + poly(AGE, 2) + poly(FAMSIZE, 4),data = dat_use_1)
summary(fit_9)
```

```
##
## Call:
## lm(formula = INCWAGE ~ EDUC + poly(AGE, 2) + poly(FAMSIZE, 4),
##     data = dat_use_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -131017  -36317  -12132   19113  531061
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 85077       5555  15.316  < 2e-16 ***
## EDUC5+ years of college     24524       7609   3.223  0.00137 **
## poly(AGE, 2)1              252168      77602   3.249  0.00125 **
## poly(AGE, 2)2             -104437      79842  -1.308  0.19159
## poly(FAMSIZE, 4)1          154551      78823   1.961  0.05059 .
## poly(FAMSIZE, 4)2           23171      76734   0.302  0.76283
## poly(FAMSIZE, 4)3          -19668      76765  -0.256  0.79791
## poly(FAMSIZE, 4)4          215544      77101   2.796  0.00543 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76180 on 407 degrees of freedom
## Multiple R-squared:  0.09705,    Adjusted R-squared:  0.08152
## F-statistic: 6.249 on 7 and 407 DF,  p-value: 5.674e-07
```

i am going to perform polynomial regression to predict wage using age . Use cross-validation to select the optimal degree d for the polynomial. What degree was chosen, and how does this compare to the results of hypothesis testing using ANOVA? Make a plot of the resulting polynomial fit to the data.

```
require(stargazer)
```

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

```
stargazer(fit_8, type = "text")
```

```
##
## =====================================================
##                               Dependent variable:
##                           ---------------------------
##                                     INCWAGE
## -----------------------------------------------------
## EDUC5+ years of college        23,946.890***
##                                 (7,660.066)
##
## poly(AGE, 3)1                 283,739.800***
##                                (77,396.880)
##
## poly(AGE, 3)2                 -160,893.300**
##                                (77,284.880)
##
## poly(AGE, 3)3                   -1,998.247
##                                (77,047.200)
##
## Constant                       85,388.430***
##                                 (5,601.098)
##
## -----------------------------------------------------
## Observations                         415
## R2                                  0.071
## Adjusted R2                         0.062
## Residual Std. Error        76,974.980 (df = 410)
## F Statistic              7.880*** (df = 4; 410)
## =====================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

There's strong quadratic relation between wage and age. There's strong linear relation between age and education more than 4 years of college. and as we can see there is a relevant statistical relationship between the variables. As we add more polynomials like AGE^3 the p-value increases,and also there is age was more statistically significant than age^2 age^3 age^4.