Multiple Linear Regresssion

R. Gevorgyan

October 21, 2018

Data

Here, we employ the CPS1988 data frame collected in the March 1988 Current Population Survey (CPS) by the US Census Bureau and analyzed by Bierens and Ginther (2001).

These are cross-section data on males aged 18 to 70 with positive annual income greater than US$ 50 in 1992 who are not self-employed or working without pay.

```
library(AER)
```

```
data("CPS1988")
summary(CPS1988)
```

```
##       wage            education       experience   ethnicity
##  Min.   :   50.05   Min.   : 0.00   Min.   :-4.0   cauc:25923
##  1st Qu.:  308.64   1st Qu.:12.00   1st Qu.: 8.0   afam: 2232
##  Median :  522.32   Median :12.00   Median :16.0
##  Mean   :  603.73   Mean   :13.07   Mean   :18.2
##  3rd Qu.:  783.48   3rd Qu.:15.00   3rd Qu.:27.0
##  Max.   :18777.20   Max.   :18.00   Max.   :63.0
##   smsa             region     parttime
##  no : 7223   northeast:6441   no :25631
##  yes:20932   midwest  :6863   yes: 2524
##              south    :8760
##              west     :6091
##
##
```

```
* wage - the wage in dollars per week,
* education and experience - measured in years
* ethnicity is a factor with levels Caucasian ("cauc") and African-American ("afam").
* smsa - indicating residence in a standard metropolitan statistical area (SMSA)
* region - the region within the United States of America, and
* parttime - whether the individual works part-time.
```

Note that the CPS does not provide actual work experience. It is therefore customary to compute experience as ***age - education - 6***; this may be considered potential experience. This quantity may become negative.

The model of interest is

$$\boldsymbol{log(wage)} = \beta_1 + \beta_2\, \boldsymbol{experience} + \beta_3\, \boldsymbol{experience2} + \beta_4\, \boldsymbol{education} + \beta_5\, \boldsymbol{ethnicity} + \varepsilon$$

This is a semilogarithmic model, which can be fitted in R using

```
cps_lm <- lm(log(wage) ~ experience + I(experience^2) +education + ethnicity, data = CPS1988)
summary(cps_lm)
```

```
##
## Call:
## lm(formula = log(wage) ~ experience + I(experience^2) + education +
##     ethnicity, data = CPS1988)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9428 -0.3162  0.0580  0.3756  4.3830
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.321e+00  1.917e-02  225.38   <2e-16 ***
## experience       7.747e-02  8.800e-04   88.03   <2e-16 ***
## I(experience^2) -1.316e-03  1.899e-05  -69.31   <2e-16 ***
## education        8.567e-02  1.272e-03   67.34   <2e-16 ***
## ethnicityafam   -2.434e-01  1.292e-02  -18.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5839 on 28150 degrees of freedom
## Multiple R-squared:  0.3347, Adjusted R-squared:  0.3346
## F-statistic:  3541 on 4 and 28150 DF,  p-value: < 2.2e-16
```

The summary reveals that all coefficients have the expected sign, and the corresponding variables are highly significant (not surprising in a sample as large as the present one). Specifically, according to this specification, the return on education is 8.57% per year.

To illustrate the general procedure for model comparisons, we explicitly fit the model without ethnicity and then compare both models using anova()

```
cps_noeth <- lm(log(wage) ~ experience + I(experience^2) + education, data = CPS1988)
anova(cps_noeth, cps_lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(wage) ~ experience + I(experience^2) + education
## Model 2: log(wage) ~ experience + I(experience^2) + education + ethnicity
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  28151 9719.6
## 2  28150 9598.6  1    121.02 354.91 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```