Simple Regression Analysis

Yingying Ji

2019/03/25

Introduction

The goal of this report is making empirical analysis of real data in R language, using different regression functions we have studied in class: **linear regression**, **logarithms in regression** and **polynomial regression**.

Data Description

The data file contains the data for 2008 (from the March 2009 "Current Population Survey" by the Bureau of Labor Statistics in the U.S. Department of Labor). These data are for full-time workers, defined as workers employed more than 35 hours per week for at least 48 weeks in the previous year, age 25-34, with a high school diploma or a bachelor's degree as their highest degree. [1]

Series Name in Data Set	Interpretion
FEMALE	1 if female; 0 if male
YEAR	Year
AHE	Average Hourly Earnings
BACHELOR	1 if worker has a bachelor's degree; 0 if worker has a high school degree

Empirical Analysis

In this part, I will make simple analysis about the relationship between a worker's age and earnings, using different regression functions.

Simple Linear Regression

- Regression Function 1

$$AHE = \alpha + \beta_1 age + \beta_2 female + \beta_3 bachelor + e$$

```r
rm(list=ls())
#import data
data<-read.csv("C:/Users/15068/Desktop/Microeconometrics/HW2/dataforhw2.csv",header = T)
reg1=lm(ahe~age+female+bachelor,data=data)
summary(reg1)
```

```
##
## Call:
## lm(formula = ahe ~ age + female + bachelor, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.139  -5.773  -1.509   4.112  57.414
```

[1] The data are from ***Student Resources of Introductin To Econometrics***.

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6357     1.0854  -0.586    0.558
## age           0.5852     0.0362  16.165   <2e-16 ***
## female       -3.6640     0.2107 -17.391   <2e-16 ***
## bachelor      8.0830     0.2088  38.709   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.072 on 7707 degrees of freedom
## Multiple R-squared:  0.1998, Adjusted R-squared:  0.1995
## F-statistic: 641.5 on 3 and 7707 DF,  p-value: < 2.2e-16
```

- Result: As the regression results show, if Age increases 1 unit, average hourly earnings are predicted to increase by \$0.5852.

Log-Linear Regression

- Regression Function 2

$$\ln(AHE) = \alpha + \beta_1 age + \beta_2 female + \beta_3 bachelor + e$$

```r
reg2=lm(log(ahe)~age+female+bachelor,data=data)
summary(reg2)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34755 -0.27810  0.01842  0.30954  1.66410
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.876340   0.056160   33.41   <2e-16 ***
## age          0.027327   0.001873   14.59   <2e-16 ***
## female      -0.185924   0.010901  -17.06   <2e-16 ***
## bachelor     0.428127   0.010804   39.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7707 degrees of freedom
## Multiple R-squared:  0.2007, Adjusted R-squared:  0.2003
## F-statistic: 644.9 on 3 and 7707 DF,  p-value: < 2.2e-16
```

- Result: As the regression results show, if Age increases 1 unit, $\ln(AHE)$ is predicted to increase by 0.0273, which means that average hourly earnings are predicted to increase by 2.73%.

Log-Log Regression

- Regression Function 3

$$\ln(AHE) = \alpha + \beta_1 \ln(age) + \beta_2 female + \beta_3 bachelor + e$$

```r
reg3=lm(log(ahe)~log(age)+female+bachelor,data=data)
summary(reg3)
```

```
##
## Call:
## lm(formula = log(ahe) ~ log(age) + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34852 -0.27913  0.02117  0.30921  1.66325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03453    0.18622  -0.185    0.853
## log(age)     0.80391    0.05496  14.626   <2e-16 ***
## female      -0.18589    0.01090 -17.054   <2e-16 ***
## bachelor     0.42825    0.01080  39.641   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7707 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.2005
## F-statistic: 645.3 on 3 and 7707 DF,  p-value: < 2.2e-16
```

- Result: Since it's log-log function, case becomes different. If Age increases 1 unit, average hourly earnings are predicted to increase by different amount. If Age increases from x to $x+1$, then $\ln(AHE)$ increases by $\ln(x+1) - \ln(x)$. The predicted increase in $\ln(AHE)$ is $0.804 * (\ln(x+1) - \ln(x))$. This means that average hourly earnings are predicted to increase by $0.804 * (\ln(x+1) - \ln(x)) * 100\%$.

Polynomial Regression

- Regression Function 4

$$\ln(AHE) = \alpha + \beta_1 age + \beta_2 female + \beta_3 bachelor + \beta_4 age^2 + e$$

```r
reg4=lm(log(ahe)~age+female+bachelor+I(age^2),data=data)
summary(reg4)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + female + bachelor + I(age^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34922 -0.27960  0.02046  0.30927  1.66268
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0854298  0.6382725   1.701   0.0891 .
## age          0.0813725  0.0434864   1.871   0.0614 .
## female      -0.1858687  0.0109006 -17.051   <2e-16 ***
## bachelor     0.4283780  0.0108057  39.644   <2e-16 ***
## I(age^2)    -0.0009148  0.0007354  -1.244   0.2135
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7706 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.2004
## F-statistic: 484.1 on 4 and 7706 DF,  p-value: < 2.2e-16
```

- Result: After adding the quadratic term in the function, case is also different. If Age increases from x to $x+1$, then predicted increase in $\ln(AHE)$ is $(0.0814 + 0.00091 * (x^2 - (x+1)^2)) = 0.08049 - 0.00182 * x$. This means that average hourly earnings are predicted to increase by $(0.08049 - 0.00182 * x) * 100\%$.

Comparison

After using different regression functions, we can compare which is better.

- Compare $reg2$ with $reg3$, the regressions differ in their choice of one of the regressors. They can be compared on the basis of $\bar{R}^2$: $reg3$ has a higher $\bar{R}^2$, so it's better.
- Compare $reg2$ with $reg4$, $reg4$ adds another variable Age^2. The coefficient on Age^2 is not statistically significant and the estimated coefficient is very close to zero. This suggests that $reg2$ is better.
- Compare $reg3$ with $reg4$, the regressions differ in their choice of the regressors: $\ln(Age)$ in $reg3$ and Age and Age^2 in $reg4$. They can also be compared on the basis of $\bar{R}^2$: $reg3$ has a higher $\bar{R}^2$, so it's better.

Conclusion

In this empirical case, after comparison, it seems that $reg3$:the log-log regression function is the best one among them. The result of it can be more reliable. If Age increases from x to $x+1$, the average hourly earnings are predicted to increase by $0.804 * (\ln(x+1) - \ln(x)) * 100\%$.