hw7

Zeren Li

10/23/2019

This problem set is due at 8:05 am on 10/31

Please upload both Rmd and PDF files on Sakai

Do not show the code in the pdf, show outputs and write-up only

Total points: 10

Goodness of Fit

Is there a marriage premium for professional athletes? Korenman and Neumark (1991) found a significant wage premium for married men after using a variety of econometric methods, but their analysis is limited because they cannot directly observe productivity. Professional athletes provide an interesting group in which to study the marriage premium because we can easily collect data on various productivity measures, in addition to salary. The data set NBASAL.RAW, on players in the National Basketball Association (NBA), is one example. For each player, we have information on points scored, rebounds, assists, playing time, and demographics.

```
data("nbasal", package="wooldridge")  #from wooldridge package
```

1. Provide summary statistics and data visualization of the following variables: `marr`, `points`, `coll`, `exper`, `age`, `black`,`children`.

- Hint: You can use `stargazer()` and `ggpairs()`.

2. Split the data into a training set (75%) and test set (25%). Create a new dummy variable of college basketball experience using `coll`. Use the training data, estimate a linear regression model relating points per game to NBA experience and college basketball experience dummy. Include experience in quadratic form. Interpret your results.

3. Holding NBA experience fixed, does a player with college basketball experience score more than his peers without such experience? How much more or less? Is the difference statistically significant?

4. Now, add marital status(`marr`) to the equation. Holding college basketball experience and NBA experience fixed, are married players more productive (based on points per game)?

5. Compute the r-squared, adjusted r-squared, RMSE by hand. Double-check it with the result from built-in function.

6. Conduct a F-test, testing the null hypothesis that `marr` has no effect on points per game against the alternative that `marr` has a positive effect. Compare `(y ~ exper + exper^2 + coll + marr)` vs that model without `marr`. Based on the test, will you include `marr` in a final model explaining the points scored by NBA players?

7. Add the variables, `age`, `black`,`children` to the model you develop in question 6. Which of these factors are individually significant? Are these factors jointly significant?

8. Try adding or dropping variables, using the F-test or the RMSE, until you find a model you're most happy with. Then run this final model on both the training dataset and the test dataset. Is the RMSE larger or smaller in the test data compared to the training set?

Measurement Error

Hamermesh and Biddle (1994) used measures of physical attractiveness in a wage equation.

```
data("beauty", package="wooldridge")  #from wooldridge package
summary(beauty)
```

```
##       wage            lwage            belavg          abvavg
##  Min.   : 1.020   Min.   :0.0198   Min.   :0.000   Min.   :0.000
##  1st Qu.: 3.708   1st Qu.:1.3104   1st Qu.:0.000   1st Qu.:0.000
##  Median : 5.300   Median :1.6677   Median :0.000   Median :0.000
##  Mean   : 6.307   Mean   :1.6588   Mean   :0.123   Mean   :0.304
##  3rd Qu.: 7.695   3rd Qu.:2.0406   3rd Qu.:0.000   3rd Qu.:1.000
##  Max.   :77.720   Max.   :4.3531   Max.   :1.000   Max.   :1.000
##      exper           looks           union           goodhlth
##  Min.   : 0.00   Min.   :1.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 8.00   1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median :15.00   Median :3.000   Median :0.0000   Median :1.0000
##  Mean   :18.21   Mean   :3.186   Mean   :0.2722   Mean   :0.9333
##  3rd Qu.:27.00   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :48.00   Max.   :5.000   Max.   :1.0000   Max.   :1.0000
##      black            female         married          south
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.000   Median :1.0000   Median :0.0000
##  Mean   :0.07381   Mean   :0.346   Mean   :0.6913   Mean   :0.1746
##  3rd Qu.:0.00000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##     bigcity         smllcity         service          expersq
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :   0.0
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  64.0
##  Median :0.000   Median :0.0000   Median :0.0000   Median : 225.0
##  Mean   :0.219   Mean   :0.4667   Mean   :0.2738   Mean   : 474.5
##  3rd Qu.:0.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 729.0
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :2304.0
##       educ
##  Min.   : 5.00
##  1st Qu.:12.00
##  Median :12.00
##  Mean   :12.56
##  3rd Qu.:13.00
##  Max.   :17.00
```

9. Regress `lwage` on `looks`, controlling for a set of control variables using the following equation. Report heteroskedasticity-robust standard errors below coefficients. Interpret the results.

$$lwage = \beta_0 + \beta_1 looks + \beta_2 black + \beta_3 female + \beta_4 educ + \beta_5 exper + \beta_6 exper2 + u$$

10. Does this model suffer from measurement error in dependent variable or independent variables? If you think there are measurement errors, state the type of errors and explain how these errors bias our results?