ECO 307: Final Rehearsal of Econometrics Knowledge!

Spring 2019

A researcher is interested in determining whether union employees earn a higher hourly wage on average than non-union employees. The researcher collects data on hourly wages from 545 workers (indexed by $i = 1..545$) over eight years from 1980 through 1987 (indexed by $t = 1..8$). During this time, workers may have changed jobs and moved into or out of unions. The dataset includes the following variables:

- $wage_{i,t}$: wage for individual i at time t, in dollars
- $year_t$: the year including values 1980 through 1987
- $union_{i,t}$: a dummy variable equal to 1 when the worker is a member of the union at the time
- $construc_{i,t}$ a dummy variable equal to 1 when worker i in employed in the construction industry at time t
- $manuf_{i,t}$ a dummy variable equal to 1 when worker i is employed in the manufacturing industry at time t

The researcher estimates the following model:

$$\begin{aligned} log(wage_{i,t}) = \beta_0 + \alpha_i + \beta_1 y1981_t + \beta_2 y1982_t + \beta_3 y1983_t + \beta_4 y1984_t + \beta_5 y1985 + \beta_6 y1986 + \beta_7 y1987 \\ + \beta_8 union_{i,t} + \beta_9 y1981_t \cdot union_{i,t} + \beta_{10} y1982_t \cdot union_{i,t} + \beta_{11} y1983_t \cdot union_{i,t} \\ + \beta_{12} y1984_t \cdot union_{i,t} + \beta_{13} y1985 \cdot union_{i,t} + \beta_{14} y1986 \cdot union_{i,t} + \beta_{15} y1987 \cdot union_{i,t} \\ + construc_{i,t} \cdot union_{i,t} + manuf_{i,t} \cdot union_{i,t} + \epsilon_{i,t} \end{aligned}$$

The estimated model results in the following output:

```
lm1 <- plm(log(wage) ~ manuf + construc + factor(year) + union
           + union:factor(year) + union:construc + union:manuf,
           data=data, index="nr", model="within", effect="individual")
summary(lm1)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(wage) ~ manuf + construc + factor(year) + union +
##     union:factor(year) + union:construc + union:manuf, data = data,
##     effect = "individual", model = "within", index = "nr")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.151549 -0.127544  0.011145  0.162688  1.490286
##
## Coefficients:
##                            Estimate Std. Error t-value  Pr(>|t|)
## manuf                     0.0682902  0.0216623  3.1525  0.001632 **
## construc                 -0.0311133  0.0351518 -0.8851  0.376152
## factor(year)1981          0.1317072  0.0249058  5.2882 1.305e-07 ***
## factor(year)1982          0.1917581  0.0250504  7.6549 2.439e-14 ***
## factor(year)1983          0.2503364  0.0249419 10.0368 < 2.2e-16 ***
## factor(year)1984          0.3089487  0.0251643 12.2773 < 2.2e-16 ***
## factor(year)1985          0.3634849  0.0247911 14.6619 < 2.2e-16 ***
## factor(year)1986          0.4394929  0.0247481 17.7586 < 2.2e-16 ***
## factor(year)1987          0.5042602  0.0251990 20.0111 < 2.2e-16 ***
## union                     0.1549710  0.0404379  3.8323  0.000129 ***
## factor(year)1981:union   -0.0552575  0.0508583 -1.0865  0.277327
## factor(year)1982:union   -0.0721600  0.0510954 -1.4123  0.157955
## factor(year)1983:union   -0.1083754  0.0514555 -2.1062  0.035253 *
## factor(year)1984:union   -0.0819092  0.0515205 -1.5898  0.111955
## factor(year)1985:union   -0.0693145  0.0526037 -1.3177  0.187692
## factor(year)1986:union   -0.1428453  0.0534553 -2.6722  0.007567 **
## factor(year)1987:union   -0.1444939  0.0512731 -2.8181  0.004856 **
## construc:union            0.0601473  0.0652032  0.9225  0.356347
## manuf:union               0.0065812  0.0384778  0.1710  0.864202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    572.05
## Residual Sum of Squares: 473.27
## R-Squared:      0.17268
## Adj. R-Squared: 0.049978
## F-statistic: 41.7007 on 19 and 3796 DF, p-value: < 2.22e-16
```

The researcher saves the residuals ($e_{i,t}$) and predicted values (let $\hat{l}_{i,t}$ denote the predicted value for $log(wage_{i,t})$) and estimates the following regression:

$$e_{i,t}^2 = g_0 + g_1\hat{l}_{i,t} + g_2\hat{l}_{i,t}^2 + u_{i,t}$$

The results are here:

```
fitted.values <- log(data$wage) - lm1$residuals
lm2 <- lm(lm1$residuals ~ fitted.values + I(fitted.values^2))
summary(lm2)
```

```
##
## Call:
## lm(formula = lm1$residuals ~ fitted.values + I(fitted.values^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1709 -0.1273  0.0117  0.1622  1.4937
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.05051    0.05326   0.948    0.343
## fitted.values      -0.06595    0.06539  -1.008    0.313
## I(fitted.values^2)  0.02012    0.01962   1.026    0.305
##
## Residual standard error: 0.3295 on 4357 degrees of freedom
## Multiple R-squared:  0.0002414,  Adjusted R-squared:  -0.0002175
## F-statistic: 0.526 on 2 and 4357 DF,  p-value: 0.591
```

The researcher decides to ***remove the interaction terms between `union` and the year dummy variables***, resulting in the following output:

```
lm3 <- plm(log(wage) ~ manuf + construc + factor(year) + union
           + union:construc + union:manuf,
           data=data,index="nr", model="within", effect="individual")
```

```
anova(lm3,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: log(wage) ~ factor(nr) + manuf + construc + factor(year) + union +
##     union:construc + union:manuf
## Model 2: log(wage) ~ factor(nr) + manuf + construc + factor(year) + union +
##     union:factor(year) + union:construc + union:manuf
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   3803 474.71
## 2   3796 473.27  7    1.4437 1.6542 0.1156
```

The researcher restores the original model, then decides to ***remove the interaction terms between `union` and the manufacturing and construction industry variables variables***, resulting in the following output:

```
lm4 <- plm(log(wage) ~ manuf + construc + factor(year) + union
           + union:factor(year),
           data=data,index="nr", model="within", effect="individual")
```

```
anova(lm4,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: log(wage) ~ factor(nr) + manuf + construc + factor(year) + union +
##     union:factor(year)
## Model 2: log(wage) ~ factor(nr) + manuf + construc + factor(year) + union +
##     union:factor(year) + union:construc + union:manuf
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   3798 473.38
## 2   3796 473.27  2    0.1061 0.4255 0.6535
```

The researcher restores the original model, then decides to ***remove the `union` variable from the regression all together***, also removing all interaction terms involving union, resulting in the following output:

```
lm5 <- plm(log(wage) ~ manuf + construc + factor(year),
           data=data, index="nr", model="within", effect="individual")
```

```
anova(lm5,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: log(wage) ~ factor(nr) + manuf + construc + factor(year)
## Model 2: log(wage) ~ factor(nr) + manuf + construc + factor(year) + union +
##     union:factor(year) + union:construc + union:manuf
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   3806 476.94
## 2   3796 473.27 10    3.6703 2.9439 0.001091 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The researcher restores the original model, then decides to ***remove all the year dummies in the regression***, also removing all interaction terms involving union, resulting in the following output:

```
lm6 <- plm(log(wage) ~ manuf + construc + union + union:construc + union:manuf,
           data=data, index="nr", model="within", effect="individual")
```

```
anova(lm6,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: log(wage) ~ factor(nr) + manuf + construc + union + union:construc +
##     union:manuf
## Model 2: log(wage) ~ factor(nr) + manuf + construc + factor(year) + union +
##     union:factor(year) + union:construc + union:manuf
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3810 565.68
## 2   3796 473.27 14    92.413 52.944 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer all of the following questions on a separate sheet of paper.

1. Is there statistical evidence that any of the terms in the original equation help explain the outcome variable? Test the appropriate hypothesis.
2. Is there evidence for heteroskedasticity in the original model? Test the appropriate hypothesis.
3. Interpret the meaning of the α_i term in the model. How does it limit omitted variable bias? What types of variables could this be accounting for?
4. Accounting for the other variables in the model, is there evidence that being a member of a union affects average hourly wage? Test the appropriate hypothesis.
5. Describe the meaning of the regression coefficient on `union`.
6. Describe the meaning of the regression coefficient on `year1981`.
7. Describe the meaning of the regression coefficient on $year1987 \cdot union_{i,t}$
8. Accounting for all the other variables in the model, how much more or less did unionized construction workers earn in 1987 than non-union construction workers at the same time period?
9. Describe how the year dummy variables help reduce omitted variable bias. What types of variables could these be accounting for?
10. Another researcher points out that gender is not included in the model, that gender is positively related to the probability that a person is in a union and positively related to people's average hourly wages, and therefore you have omitted variable bias. Do you agree? Explain.
11. Does the effect that union has on wages differ by industry? Test the appropriate hypothesis.
12. Labor market experts have argued that union power fluctuated during this sample period, sometimes diminishing the benefit that union membership has on wages. Do you find evidence for differences in union influence on wages across time? Test the appropriate hypothesis(es).
13. How much more or less did a union construction worker earn in 1987 versus 1986?
14. How well does the full model explain wages? Provide a measure of the strength of explanatory power of the explanatory variables. Is this number high or low? Is this good or is this problematic?