

ECONOMETRICS ASSIGNMENT 4

- 1) It is a simple linear regression between labour wage and years of education. From the standard error of regression value, it can be said that the estimated model is not a bad fit. The intercept term comes as 5.83. So that means that if we keep the years of education as constant then the labour wage will increase by 5.83 dollars. The slope coefficient is 0.065. That means that with 1 year increase in the years of education, the labour wage will increase by 0.065 dollars. The R² value is 0.155 or the 15.5%. This describes that this regression equation can explain 15.5% variability of the model with certainty. The standard error of the regression is 0.42, which is low. The smaller the value of the standard error of the regression the better the model is. The P-value of the slope coefficient is less than 0.05 so we infer that there is significant evidence that the ED has an impact on the labour wage. If education is increased by one standard deviation, then the labour wage will increase by 0.0023 dollars. After running the coefficient test, the p value is very less than 0.05, so there is heteroskedasticity which means that the systematic change is in the spread of the residuals over the range of measured values. The table below shows the results described above.

```
Call:
lm(formula = LWAGE ~ ED, data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-1.92996 -0.26863  0.00931  0.28453  1.83076

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.838779   0.030997  188.37  <2e-16 ***
ED           0.065204   0.002358   27.65  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4243 on 4163 degrees of freedom
Multiple R-squared:  0.1552,    Adjusted R-squared:  0.155
F-statistic: 764.5 on 1 and 4163 DF,  p-value: < 2.2e-16
```

Table 1: Panel Regression table between LWAGE and ED

- 2) I extend the model by including individual or entity fixed effects and time fixed effects. Years of education, weeks worked, and years of full-time experience are the time fixed effects and the rest are entity fixed effects. I run a panel regression and find that all the variables have p value less than 0.05 which states that all have them have some kind of impact on the labour wage. Also, the R square value of the model is 45%, which is relatively high. I have attached the table below.

```
Call:
p1m(formula = LWAGE ~ EXP + WKS + ED + OCC + IND + SOUTH + SMSA +
      MS + UNION + FEM + BLK, data = df, model = "within",
      index = c("YEAR", "ID"), effects = "twoways")
```

```

Coefficients:
      Estimate Std. Error t-value Pr(>|t|)
EXP    0.00697427 0.00047458 14.6957 < 2.2e-16 ***
WKS    0.00442363 0.00096347  4.5913 4.535e-06 ***
ED     0.05417333 0.00232724 23.2779 < 2.2e-16 ***
OCC    -0.14293191 0.01303951 -10.9614 < 2.2e-16 ***
IND     0.05898222 0.01049250  5.6214 2.019e-08 ***
SOUTH  -0.05848294 0.01115090  -5.2447 1.644e-07 ***
SMSA   0.16054192 0.01074415 14.9423 < 2.2e-16 ***
MS     0.09680071 0.01829639  5.2907 1.281e-07 ***
UNION  0.09208680 0.01138406  8.0891 7.821e-16 ***
FEM    -0.34004867 0.02232483 -15.2319 < 2.2e-16 ***
BLK    -0.16055401 0.01962144  -8.1826 3.659e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    729.43
Residual Sum of Squares: 401.05
R-Squared:                0.45019
Adj. R-Squared:          0.44793
F-statistic: 308.689 on 11 and 4147 DF, p-value: < 2.22e-16

```

Table 2: Panel Regression involving Fixed Effect Variables

- 3) I filtered the original dataset to get the data for the year 1. I then estimate the model as given in the question. The R^2 value of the regression model comes as 28.2% and a low standard error of regression value of 0.33 which tells us that the individual variables and the joint interaction terms explain the model better. We then create two regression equations one for male (i.e., FEM=0) and another for female (i.e., FEM=1). The regression equation for male is $\text{lm}(\text{formula} = \text{LWAGE} \sim \text{BLK} + \text{UNION} + \text{OCC}, \text{data} = \text{year}_1)$ by putting FEM=0 in the main equation shown in the box below. For female we have replace the fem with 1 and rearrange the equation.

```

lm(formula = LWAGE ~ FEM + BLK + UNION + OCC + FEM * BLK + FEM *
    UNION + FEM * OCC, data = year_1)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-1.1108 -0.2095  0.0441  0.2535  0.6676

```

```

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.54886    0.02216 295.519 < 2e-16 ***
FEM          -0.42396    0.05998  -7.068 4.47e-12 ***
BLK          -0.16534    0.06449  -2.564  0.0106 *
UNION         0.13461    0.03236  4.159 3.67e-05 ***
OCC          -0.31069    0.03149  -9.865 < 2e-16 ***
FEM:BLK       0.15469    0.12719  1.216  0.2244
FEM:UNION     0.02828    0.10793  0.262  0.7934
FEM:OCC      -0.09696    0.09849  -0.984  0.3253
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 0.3312 on 587 degrees of freedom
Multiple R-squared:  0.2817,    Adjusted R-squared:  0.2731
F-statistic: 32.88 on 7 and 587 DF, p-value: < 2.2e-16

```

Table 3: Regression Table with Interaction terms

Now I estimate the main equation with FEM=0 and run the regression. I see that the R^2 value has reduced to 16.6% and the residual sum of squares has increased from 0.33 to 0.35. This clearly shows that the female variables had considerable effect on the labour wage in the regression equation. I conduct an Anova test between these two-regression equations and find that p value is less than 0.05 which confirms the fact that female has an impact on the

regression equation. Below are the charts for Regression without female and the Anova table.

```
Call:
lm(formula = LWAGE ~ BLK + UNION + OCC, data = year_1)

Residuals:
    Min       1Q   Median       3Q      Max
-1.17968 -0.23728  0.03433  0.26030  0.71543

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.48776    0.02207  293.915 < 2e-16 ***
BLK          -0.24693    0.05651   -4.370 1.47e-05 ***
UNION         0.16945    0.03261    5.195 2.82e-07 ***
OCC          -0.29744    0.03145   -9.458 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3556 on 591 degrees of freedom
Multiple R-squared:  0.1662,    Adjusted R-squared:  0.162
F-statistic: 39.26 on 3 and 591 DF,  p-value: < 2.2e-16
> anova(model31,model3)
Analysis of Variance Table

Model 1: LWAGE ~ BLK + UNION + OCC
Model 2: LWAGE ~ FEM + BLK + UNION + OCC + FEM * BLK + FEM * UNION + FEM *
OCC
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     591 74.726
2     587 64.377   4    10.349 23.592 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 4: Regression table without Female and the Anova table

Now I try to find out the regression equation of the model without the interaction terms i.e. making them equal to 0. We see that the p-value of the FEM coefficient is less than 0.05. This says that the coefficient of FEM is statistically significant. Now in contrast I try to test the joint significance of the interaction terms. To do this I conduct a anova test with the two models namely the model without any interaction terms and the original model with all the interaction terms. We see that the P value is 0.58 which is greater than 0.05 so fail to reject the null hypothesis and say that the interaction terms does not have any effect on the regression equation. Below are the tables for regression without the interaction terms and the anova table involving the two models

```
Call:
lm(formula = LWAGE ~ FEM + BLK + UNION + OCC, data = year_1)

Residuals:
    Min       1Q   Median       3Q      Max
-1.11119 -0.20787  0.04788  0.25312  0.67468

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.54927    0.02151  304.458 < 2e-16 ***
FEM          -0.42736    0.04441   -9.622 < 2e-16 ***
BLK          -0.13216    0.05391   -2.451  0.0145 *
UNION         0.13957    0.03051    4.575 5.8e-06 ***
OCC          -0.31820    0.02934  -10.845 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3309 on 590 degrees of freedom
Multiple R-squared:  0.2793,    Adjusted R-squared:  0.2744
F-statistic: 57.16 on 4 and 590 DF,  p-value: < 2.2e-16
```

```
> anova(model32,model3)
Analysis of Variance Table

Model 1: LWAGE ~ FEM + BLK + UNION + OCC
Model 2: LWAGE ~ FEM + BLK + UNION + OCC + FEM * BLK + FEM * UNION + FEM *
      OCC
      Res.Df    RSS Df Sum of Sq      F Pr(>F)
1       590 64.590
2       587 64.377   3   0.21331 0.6483 0.5842
```

Table 5: Regression Table without interaction terms and the Anova Table

- 4) Using the subset from Question 3 I create two subsets, one for age greater than 30 and the other less than 30. I run the regression for both the case which are showed in the tables below.

```
Call:
lm(formula = LWAGE ~ EXP, data = exp_greater_30)

Residuals:
    Min       1Q   Median       3Q      Max
-0.98016 -0.20765  0.00225  0.24349  0.73136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.38597    0.32567  22.680 < 2e-16 ***
EXP         -0.02754    0.00957  -2.877  0.00489 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3357 on 101 degrees of freedom
Multiple R-squared:  0.07576,    Adjusted R-squared:  0.06661
F-statistic: 8.279 on 1 and 101 DF,  p-value: 0.004894
```

Table : Regression table for experience greater than 30

```
Call:
lm(formula = LWAGE ~ EXP, data = exp_lesser_30)

Residuals:
    Min       1Q   Median       3Q      Max
-1.29568 -0.24904  0.04975  0.29724  0.67896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.147261    0.032517 189.045 < 2e-16 ***
EXP          0.015906    0.002092   7.604 1.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3738 on 490 degrees of freedom
Multiple R-squared:  0.1055,    Adjusted R-squared:  0.1037
F-statistic: 57.81 on 1 and 490 DF,  p-value: 1.481e-13
```

Table : Regression table for experience less than 30

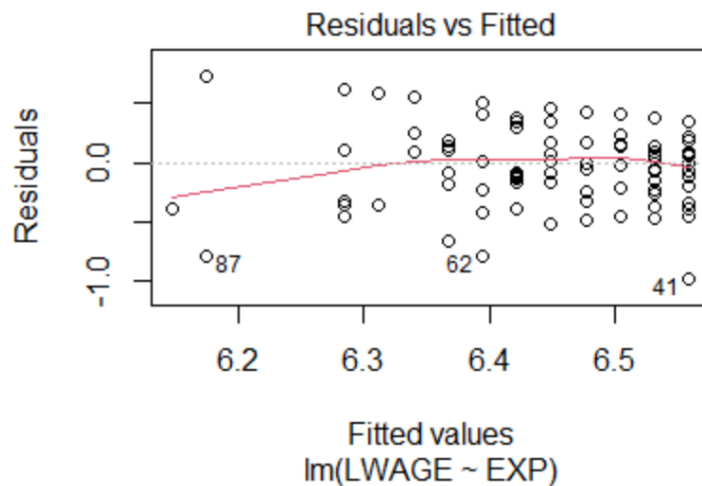


Chart1: Scatter Plot for age less than 30

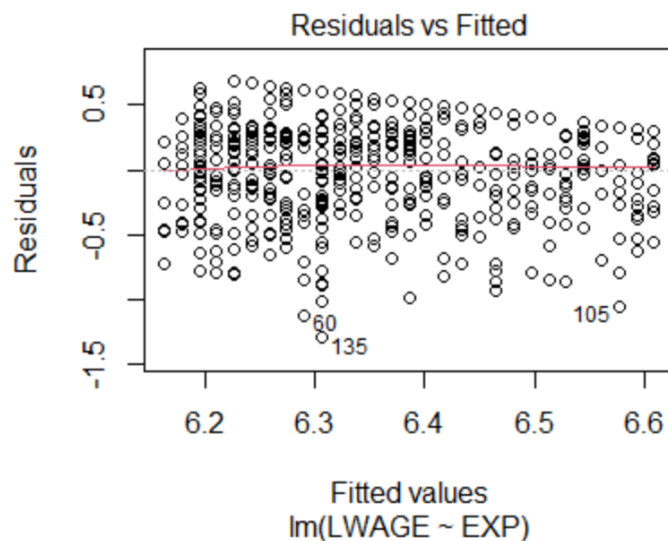


Chart 2: Scatter Plot for age greater than 30

So, from the above chart 1 we can infer that people having age less than 30, experience and labour wage are perfectly related but as the age goes beyond 30 experience does not only effects the wage. There are other factors that come in the way and hence the regression line is not perfectly linear as in our case in chart 2. Here regression lines are denoted by the red line.

- 5) Using the original panel data, I estimate the regression equation that experience has both time and entity fixed effect on labour wage. The p value of the slope coefficient is less than 0.05. But the R^2 value is very low, 2.2%. So, experience weakly describes labour wage. Now we add another fixed effect variable, ED to this model. Now the R^2 value goes up to 25% and the residual sum of squares goes down to 546 from 713. This shows that addition of one more fixed effect drastically impacts the regression model and helps define the model more efficiently. Wages increase with the work experience and this is true when we compare the data to the real world. Without adding the experience as a predictor for wages the model's predictions have increased error.

```
Call:
plm(formula = LWAGE ~ EXP, data = df, model = "within",
     index = c("YEAR", "ID"), effects = "twoways")

Balanced Panel: n = 7, T = 595, N = 4165

Residuals:
    Min.    1st Qu.    Median    3rd Qu.    Max.
-2.148385 -0.254883  0.037788  0.268494  1.906929

Coefficients:
            Estimate Std. Error t-value Pr(>|t|)
EXP 0.00577751 0.00059534  9.7045 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    729.43
Residual Sum of Squares: 713.27
R-Squared:    0.022153
Adj. R-Squared: 0.020507
F-statistic: 94.1778 on 1 and 4157 DF, p-value: < 2.22e-16
```

Table : Panel Regression table with Exp

```
plm(formula = LWAGE ~ EXP + ED, data = df, model = "within",
     index = c("YEAR", "ID"), effects = "twoways")

Balanced Panel: n = 7, T = 595, N = 4165

Residuals:
    Min.    1st Qu.    Median    3rd Qu.    Max.
-2.03768 -0.22758  0.03844  0.24221  1.95842

Coefficients:
            Estimate Std. Error t-value Pr(>|t|)
EXP 0.01001214 0.00053411  18.745 < 2.2e-16 ***
ED 0.07379691 0.00206563  35.726 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    729.43
Residual Sum of Squares: 545.69
R-Squared:    0.2519
Adj. R-Squared: 0.25046
F-statistic: 699.712 on 2 and 4156 DF, p-value: < 2.22e-16
```

Table : Panel Regression table with Exp and ED

R Codes

setwd("C:\\Users\\soham\\Desktop\\REcotrics1")
df <- read.csv('cornwellrupert1988_rev.csv')
install.packages("AER")
library(AER)
install.packages("ggplot2")
library(ggplot2)
install.packages("plm")
library(plm)
install.packages("sandwich")
library(sandwich)
install.packages("dplyr")
library(dplyr)
##Q1

estimating LWAGE with Education
model1 <- lm(LWAGE ~ED,data=df)
Summary of the model
summary(model1)
checking for heteroskedasticity
coeftest(model1, vcov = vcovHC, type = "HC1")
##Q2
model2 <- plm(LWAGE~EXP+WKS+ED+OCC+IND+SOUTH+SMSA+MS+UNION+FEM+BLK, data=df,index = c("YEAR","ID"), model="within", effects="twoways")
summary(model2)
##Q3
#Filter for year 1
year_1 <- df %>% filter(YEAR == 1)
model3<-lm(LWAGE~FEM+BLK+UNION+OCC+FEM*BLK+FEM*UNION +FEM*OCC,data=year_1)
summary(model3)
model31<-lm(LWAGE~BLK+UNION+OCC,data=year_1)
summary(model31)
anova(model31,model3)
model32<-lm(LWAGE~FEM+BLK+UNION+OCC,data=year_1)
summary(model32)
anova(model32,model3)
##Q4
Filtering out Experiance
exp_greater_30 <- year_1 %>% filter(EXP >=30)
exp_lesser_30 <- year_1 %>% filter(EXP <30)
exp_greater_30_model <- lm(LWAGE ~ EXP,data = exp_greater_30)
plot(exp_greater_30_model)
summary(exp_greater_30_model)
exp_lesser_30_model <- lm(LWAGE ~ EXP,data = exp_lesser_30)
plot(exp_lesser_30_model)
summary(exp_lesser_30_model)
##Q5
#fitting with EXP
model5 <- plm(LWAGE ~EXP,data=df,index=c("YEAR","ID"),model="within",effects="twoways")
#fitting with experince and also with education
model5_with_ED <- plm(LWAGE ~EXP+ED,data=df,index=c("YEAR","ID"),model="within",effects="twoways")
summary(model5)
summary(model5_with_ED)