W271 Summer 2022 Lecture Video Question Solutions Week 12+13

Contents

Week 12: Analysis of Panel Data: Fixed Effect and Random Effect Models

12.2 An Introduction to Fixed-Effect Models

Q: Does it matter that an intercept is included in the context of a fixed-effect model?

Solution: The specified model is: $y_{i,t} = \beta_0 + \beta_1 x_{i,t} + a_i + \epsilon_{i,t}$

In a fixed effects model, the intercept effectively becomes the reference group. When the intercept is not included, the reference group gets a separate intercept term to represent it. **Therefore, it generally makes no difference whether an intercept is included or not in the context of a fixed effect model.**

As an example, let's say we have two groups $i = 1, 2$.

For $i = 1$ the regression equation with an intercept is $y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \epsilon_{i,t}$

For $i = 1$ the regression equation without an intercept is $y_{i,t} = \beta_1 x_{i,t} + \tilde{a}_1 + \epsilon_{i,t}$

For $i = 2$ the regression equation with an intercept is $y_{i,t} = \beta_0 + \beta_1 x_{i,t} + a_2 + \epsilon_{i,t}$

For $i = 3$ the regression equation without an intercept is $y_{i,t} = \beta_1 x_{i,t} + \tilde{a}_2 + \epsilon_{i,t}$

So we can see that what changes is effectively the coefficients on each fixed effect. For $i = 1$ comparing the equations we have that $\beta_0 = \tilde{a}_1$ and $\beta_0 + a_2 = \tilde{a}_2$. Effectively without an intercept, the reference group indicator can be estimated separately, resulting in essentially the same equation.

The only caveat is that R squared and other diagnostics are invalid without including an intercept. Generally, it is advised to always include an intercept in models.

To see that they are indeed the same, we can compare the model output below with and without an intercept. Note how a coefficient for Group 1 is included in the second model and is equal to the intercept in the first model, but some of the diagnostics are not the same.

```r
dat <- expand.grid("Group" = 1:2, "Time" = 1:4)
dat$X <- rnorm(nrow(dat))
dat$Y <- dat$X + dat$Group + dat$Time + rnorm(nrow(dat))
```

```r
#with intercept
summary(lm(Y ~ X + factor(Group) + factor(Time), data = dat))
```

```
##
## Call:
## lm(formula = Y ~ X + factor(Group) + factor(Time), data = dat)
##
## Residuals:
##       1       2       3       4       5       6       7       8
## -0.4033  0.4033 -0.0625  0.0625  0.2912 -0.2912  0.1746 -0.1746
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.7012     0.4231   4.020   0.0567 .
## X                1.4064     0.2150   6.541   0.0226 *
## factor(Group)2   1.7036     0.3828   4.450   0.0470 *
## factor(Time)2    0.4948     0.7185   0.689   0.5622
## factor(Time)3    0.1427     0.5347   0.267   0.8145
## factor(Time)4    2.1835     0.5331   4.096   0.0548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5309 on 2 degrees of freedom
## Multiple R-squared:  0.9852, Adjusted R-squared:  0.9483
## F-statistic: 26.66 on 5 and 2 DF,  p-value: 0.03655
```

```r
#without intercept; note how the coefficient on Group 1 = the intercept in the previous model
summary(lm(Y ~ X + factor(Group) + factor(Time) - 1, data = dat))
```

```
##
## Call:
## lm(formula = Y ~ X + factor(Group) + factor(Time) - 1, data = dat)
##
## Residuals:
##       1       2       3       4       5       6       7       8
## -0.4033  0.4033 -0.0625  0.0625  0.2912 -0.2912  0.1746 -0.1746
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## X                1.4064     0.2150   6.541   0.0226 *
## factor(Group)1   1.7012     0.4231   4.020   0.0567 .
## factor(Group)2   3.4048     0.4202   8.103   0.0149 *
## factor(Time)2    0.4948     0.7185   0.689   0.5622
## factor(Time)3    0.1427     0.5347   0.267   0.8145
## factor(Time)4    2.1835     0.5331   4.096   0.0548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.5309 on 2 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9863
## F-statistic: 96.67 on 6 and 2 DF,  p-value: 0.01027
```

12.3 An Example: The Effect of Job Training on Firm Scrap Rates

Q: Answer the various questions listed below.

```
library(wooldridge)
```

Solution:

```
## Warning: package 'wooldridge' was built under R version 4.1.1
```

```
data("jtrain")

jtrain.87 <- jtrain[jtrain$year == 1987, ]
```

Is there anything wrong with this estimated regression (in terms of understanding impact of training on scrap rate)?

```
model <- lm(lscrap ~ hrsemp + lsales + lemploy, data = jtrain.87)
summary(model)
```

```
##
## Call:
## lm(formula = lscrap ~ hrsemp + lsales + lemploy, data = jtrain.87)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81878 -0.91530  0.03304  0.87052  2.68042
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.74426    4.57470   2.567  0.01420 *
## hrsemp      -0.04218    0.01868  -2.259  0.02957 *
## lsales      -0.95064    0.36984  -2.570  0.01409 *
## lemploy      0.99213    0.35692   2.780  0.00833 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.3 on 39 degrees of freedom
##   (114 observations deleted due to missingness)
## Multiple R-squared:  0.3099, Adjusted R-squared:  0.2568
## F-statistic: 5.838 on 3 and 39 DF,  p-value: 0.002148
```

Technically speaking, there is nothing wrong with this regression as is except for the fact that it does not leverage the full data set and structure of the panel data. It is really estimating the impact of training on scrap rate in 1987.

Interpret the coefficient associated with the variable hrsemp.

The model results imply that there is a significant, negative relationship between training and scrap rate, meaning that as we increase training we decrease the scrap rate. This is the intuitive relationship we might expect, given that training should make workers more effective at their jobs and help reduce errors.

Is the effect large? Is there any other information (perhaps including that not included in the regression) you would need in order to answer this question?

The coefficient is -0.04, meaning that for each unit increase (hour) in hrsemp, we decrease lscrap by -0.04. However, it is hard to know the impact without taking into account the typical variations in hrsemp and lscrap. We would want to understand the standard deviation of hrsemp and mean of lscrap to contextualize this coefficient results.

To better understand if this coefficient is a large impact, we can multiply it by the standard deviation of hrsemp and divide that by the average lscrap rate. This is effectively comparing the marginal impact of a one standard deviation increase in hrsemp to the average lscrap to see if the typical change is relatively large.

As the number below shows, the coefficient on hrsemp implies a large impact to lscrap.

```
#take logs first given high skew in hrsemp to avoid outliers
hrsemp.sd <- exp(sd(log(1 + jtrain.87$hrsemp), na.rm = T))

#average lscrap
avg.lscrap <- mean(jtrain.87$lscrap, na.rm = T)

#1 sd increase in hrsemp leads to a relative decrease in lscrap
hrsemp.sd * coef(model)["hrsemp"]/ avg.lscrap
```

```
##     hrsemp
## -0.2973252
```

How would you estimate a cross-sectional model differently, if at all?

To estimate a cross sectional model, we can use the same formula on the larger data set except add fixed effects for year to control for differences across time in lscrap.

When we utilize the full data set and include fixed effects for year, the hrsemp coefficient is no longer significant. Obviously in a real analysis, we would want to likely include additional controls and do more analysis in the model. But this highlights how sometimes accounting for the panel data can change the results seen in a more basic model or how results in a subset of the data do not always match the results in a larger sample.

```
model <- lm(lscrap ~ hrsemp + lsales + lemploy + factor(year), data = jtrain)
summary(model)
```

```
##
## Call:
## lm(formula = lscrap ~ hrsemp + lsales + lemploy + factor(year),
##     data = jtrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8545 -0.8664 -0.1430  0.9743  3.0391
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.648462   2.725628   3.173 0.001886 **
## hrsemp            -0.001664   0.004728  -0.352 0.725540
## lsales            -0.703702   0.218048  -3.227 0.001584 **
## lemploy            0.734902   0.209594   3.506 0.000626 ***
## factor(year)1988 -0.143537   0.301746  -0.476 0.635102
## factor(year)1989 -0.382020   0.306539  -1.246 0.214935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 129 degrees of freedom
##   (336 observations deleted due to missingness)
## Multiple R-squared:  0.1074, Adjusted R-squared:  0.07282
## F-statistic: 3.105 on 5 and 129 DF,  p-value: 0.01113
```

12.4 A Digression: Differencing When There Are More Than Two Time Periods

Q: Why in the fixed-effect models is the transformation of more than two time periods naturally handled?

Solution: In a fixed effects model, we can take successive differences between periods to eliminate any unobserved differences in groups that are time invariant i.e. do not change with time. This just collapses the number of observations and reduces the number of time periods we are analyzing, but the model itself can still be run after removing the fixed effects that are differenced out.

12.5 Remarks on Fixed-Effect Models

Q: What does homoskedasticity and serially uncorrelated errors across t mean?

Solution: Homoskedasticity across time means that the $Var(\epsilon_t|t) = \sigma^2$ i.e. that the variance of the residuals does not depend on time and is constant over time. If there is a relationship with the variance over time, then the regression coefficients will not be consistent, leading to incorrect statistical inferences.

Serial correlation means that the residuals are correlated over time i.e. $Cov(\epsilon_t, \epsilon_{t'}) \neq 0$. This means that the residuals are not longer independent and that the regression model is no longer consistent, again leading to incorrect statistical inferences.

Note that neither of these violations makes the regression equation necessarily biased. The coefficient estimate is still unbiased if we assume the residuals are centered at zero, but it is no longer BLUE or consistent.

Q: In a general fixed-effect model, we have $N\ddot{O}T$ observations and k independent variables. As such, we should have $NT - k$ degrees of freedom. Is that correct?

Solution: Generally speaking, this is not correct if we include fixed effects for each group and time period in the regression model. If we omit these fixed effects though, then this is correct.

In a true fixed effect model, we have k independent variables but also $(N-1)$ variables for each group and $(T-1)$ variables for each time period (we subtract one for the reference group and reference time period).

This means we really have $NT - k - (N-1) - (T-1) - 1$ degrees of freedom in the model.

12.6 Random-Effect Models

Q: Estimate the model using pooled OLS and fixed effects.

Solution: We use the plm package because it makes estimating pooled OLS, fixed effects, and random effects models easy on panel data. We just have to specify the input dataset and the indices that describe the panel struture.

```
#install.packages("wooldridge")
#install.packages("plm")
library(wooldridge)
library(plm)
```

```
## Warning: package 'plm' was built under R version 4.1.1
```

```
data("wagepan")

wagepan.pl <- pdata.frame(wagepan, index = c("nr", "year"))
panel.model.pool <- plm(lwage ~ educ + black + hisp + exper + I(exper^2) + married + union, wagepan.pl,
panel.model.fe <- plm(lwage ~ educ + black + hisp + exper + I(exper^2) + married + union, wagepan.pl, m
```

The pooled regression results, which ignores any grouping strucure is similar to the random effects results seen in the lecture.

```
#pooled regression
summary(panel.model.pool)
```

```
## Pooling Model
##
## Call:
## plm(formula = lwage ~ educ + black + hisp + exper + I(exper^2) +
##     married + union, data = wagepan.pl, model = "pooling")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -5.268937 -0.248691  0.033205  0.296163  2.560777
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept) -0.03470569  0.06456900 -0.5375    0.5910
## educ         0.09938779  0.00467760 21.2476 < 2.2e-16 ***
## black       -0.14384171  0.02355950 -6.1055 1.114e-09 ***
## hisp         0.01569798  0.02081119  0.7543    0.4507
## exper        0.08917907  0.01011105  8.8200 < 2.2e-16 ***
## I(exper^2)  -0.00284866  0.00070736 -4.0272 5.742e-05 ***
## married      0.10766558  0.01569647  6.8592 7.897e-12 ***
## union        0.18007257  0.01712053 10.5179 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Total Sum of Squares:    1236.5
## Residual Sum of Squares: 1005.8
## R-Squared:      0.18659
## Adj. R-Squared: 0.18528
## F-statistic: 142.613 on 7 and 4352 DF, p-value: < 2.22e-16
```

The fixed effects model results drop the educ, black, and hisp terms because when we include fixed effects for each person, that absorbs these variables, which are at the person level. This is one benefit of random effects models which allow us to include these in addition to person level random intercepts.

```
#fixed effects regression
summary(panel.model.fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lwage ~ educ + black + hisp + exper + I(exper^2) +
##     married + union, data = wagepan.pl, model = "within")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -4.1726214 -0.1257010  0.0092527  0.1595770  1.4701690
##
## Coefficients:
##               Estimate  Std. Error t-value  Pr(>|t|)
## exper       0.11684669  0.00841968 13.8778 < 2.2e-16 ***
## I(exper^2) -0.00430089  0.00060527 -7.1057 1.422e-12 ***
## married     0.04530332  0.01830968  2.4743   0.01339 *
## union       0.08208713  0.01929073  4.2553 2.138e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    572.05
## Residual Sum of Squares: 470.2
## R-Squared:      0.17804
## Adj. R-Squared: 0.059852
## F-statistic: 206.375 on 4 and 3811 DF, p-value: < 2.22e-16
```

Week 13: Analysis of Panel Data: Linear Mixed Effect Models

Q: Test the specified models using a likelihood ratio test and explain your result.

```
#note the data no longer appears to be available at the url in the lecture video
#we can obtain a similar data set here:
#https://raw.githubusercontent.com/michael-franke/intro-data-analysis/master/data_sets/politeness_data.
#frequency = pitch and attitude = context
dat <- read.csv("https://raw.githubusercontent.com/michael-franke/intro-data-analysis/master/data_sets/p

head(dat)
```

Solution:

```
##   subject gender sentence context pitch
## 1      F1      F       S1     pol 213.3
## 2      F1      F       S1     inf 204.5
## 3      F1      F       S2     pol 285.1
## 4      F1      F       S2     inf 259.7
## 5      F1      F       S3     pol 203.9
## 6      F1      F       S3     inf 286.9
```

```
base.model <- lm(pitch ~ gender + context, data = dat)
full.model <- lm(pitch ~ gender + context + gender:context, data = dat)

anova(base.model, full.model, test = "LRT")
```

```
## Analysis of Variance Table
##
## Model 1: pitch ~ gender + context
## Model 2: pitch ~ gender + context + gender:context
##   Res.Df    RSS Df Sum of Sq Pr(>Chi)
## 1     80 101820
## 2     79 100511  1    1309.1   0.3104
```

Comparing the models with a LRT has a p-value pf 0.31, meaning that the full model with the interaction term is not significantly better than the smaller model without the interaction term. There is no evidence of a different politness effect on pitch by gender in this data set.

Q: Compare the results of your model to the final model with both fixed effects, random intercepts, and random slopes from lecture.

```
#note the data no longer appears to be available at the url in the lecture video
#we can obtain a similar data set here:
#https://raw.githubusercontent.com/michael-franke/intro-data-analysis/master/data_sets/politeness_data.
#frequency = pitch and attitude = context and scenario = sentence
dat <- read.csv("https://raw.githubusercontent.com/michael-franke/intro-data-analysis/master/data_sets/p

head(dat)
```

Solution:

```
##   subject gender sentence context pitch
## 1      F1      F       S1     pol 213.3
## 2      F1      F       S1     inf 204.5
## 3      F1      F       S2     pol 285.1
## 4      F1      F       S2     inf 259.7
## 5      F1      F       S3     pol 203.9
## 6      F1      F       S3     inf 286.9
```

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
#note convergence issue is likely due to differences in data set
re.model <- lmer(pitch ~ gender + context + (1 + context | subject) + (1 + context | sentence), data = d
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(re.model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: pitch ~ gender + context + (1 + context | subject) + (1 + context |
##     sentence)
##    Data: dat
##
##      AIC      BIC   logLik deviance df.resid
##    814.9    839.1   -397.4    794.9       73
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1946 -0.6690 -0.0789  0.5256  3.4251
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  sentence (Intercept) 182.083  13.494
##           contextpol   31.244   5.590   0.22
##  subject  (Intercept) 392.344  19.808
##           contextpol    1.714   1.309   1.00
##  Residual             627.890  25.058
## Number of obs: 83, groups:  sentence, 7; subject, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  257.991     13.528  19.071
## genderM     -110.806     17.510  -6.328
## contextpol   -19.747      5.922  -3.335
##
## Correlation of Fixed Effects:
##            (Intr) gendrM
## genderM    -0.647
## contextpol -0.105  0.003
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

We will fit a model with fixed effects and an interaction between sentence and context.

```
alt.model <- lm(pitch ~ gender + context + sentence + sentence * context, data = dat)

summary(alt.model)
```

```
##
## Call:
## lm(formula = pitch ~ gender + context + sentence + sentence *
##     context, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.022 -20.667   0.422  20.858  79.778
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            235.178     14.238  16.517   <2e-16 ***
## genderM               -108.823      7.402 -14.701   <2e-16 ***
## contextpol              -9.717     19.444  -0.500   0.6189
## sentenceS2              17.450     19.444   0.897   0.3726
## sentenceS3              46.667     19.444   2.400   0.0191 *
## sentenceS4              44.833     19.444   2.306   0.0242 *
## sentenceS5              16.800     19.444   0.864   0.3906
## sentenceS6               8.867     19.444   0.456   0.6498
## sentenceS7              18.133     19.444   0.933   0.3543
## contextpol:sentenceS2   15.133     27.498   0.550   0.5839
## contextpol:sentenceS3  -31.283     27.498  -1.138   0.2593
## contextpol:sentenceS4   -4.650     27.498  -0.169   0.8662
## contextpol:sentenceS5   -4.783     27.498  -0.174   0.8624
## contextpol:sentenceS6  -16.559     28.187  -0.587   0.5588
## contextpol:sentenceS7  -30.033     27.498  -1.092   0.2786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.68 on 68 degrees of freedom
## Multiple R-squared:  0.781,  Adjusted R-squared:  0.736
## F-statistic: 17.33 on 14 and 68 DF,  p-value: < 2.2e-16
```

```
anova(re.model, alt.model)
```

```
## Data: dat
## Models:
## re.model: pitch ~ gender + context + (1 + context | subject) + (1 + context | sentence)
## alt.model: pitch ~ gender + context + sentence + sentence * context
##           npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## re.model    10 814.90 839.09 -397.45   794.90
## alt.model   16 834.79 873.50 -401.40   802.79     0  6          1
```

Comparing the models, our alternative model is not significantly better compared to the random effects model.