

# ECON 3040 - Panel Data

Ryan T. Godwin

University of Manitoba

## The problem: Missing variables (unobserved heterogeneity)

- ▶ We return to the same problem that we studied in Instrumental Variables: the dreaded **missing variable** problem. Recall that if a variable is missing, and it's correlated to  $x$ , then the LS estimator for the effect of  $x$  on  $y$  is biased and inconsistent.
- ▶ Correlations in (observational) economics data are almost *never* causal effects. The variables in our data set are endogenous; chosen by people who are trying their best.
- ▶ If we had the right control variable ( $x$ ) we could solve the problem by including it in our model, but the right  $x$  is often *missing*.
- ▶ In panel data, we sometimes call the missing variables problem **unobserved heterogeneity** or **unobserved effects**.
- ▶ If we can observe the same units of observations (e.g. individuals or countries) repeatedly over time, then we can solve this unobserved heterogeneity problem, using the **fixed effects** or **within** estimator.

## The problem: Missing variables (unobserved heterogeneity)

- ▶ This chapter introduces another trick for dealing with the missing variable problem. We've already used DiD and IV, and now we look at *fixed effects* estimation using panel data.
- ▶ In panel data and fixed effects estimation, we usually rephrase the missing variable problem, calling it instead *unobserved heterogeneity*.

- ▶ Unobserved heterogeneity means that people are different, and that some of these differences are unobserved.
- ▶ These differences could be things like *ability* - the missing variable that determines both wages and education, and that we dealt with using the *distance from college* instrument in the IV chapter.
- ▶ Unobserved heterogeneity can also apply to countries, firms, and other observational units.
- ▶ If we can observe the same units of observations (e.g. individuals or countries) repeatedly over time, then we can solve this unobserved heterogeneity problem, using the panel data and the *fixed effects* estimator (also called the *within* estimator).

# What is panel data?

- ▶ “Panel data” is a data set where we observe the same observational units (individuals, countries) repeatedly over time. Each “panel” represents a moment in time, a snapshot of all the individuals in the sample. Then, this panel of information is repeated over time, with some or all of the values of the variables changing in each panel.
- ▶ A panel data set can be arranged by horizontally “stacking” another set of observations for a new time period (think of adding another set of  $n$  observations at the bottom of the data set for each time period, but the units of observation in each stack are the same). See Figure 1. A more common way of arranging panel data, however, is to stack all years of data for an individual together, before moving on to the next observation. See Figure 2.
- ▶ In a panel data set we have  $n \times T$  observations.  $n$  are the number of individuals, and  $T$  are the number of time periods over which these individuals are observed.

**Figure:** Exports ( $y$  variable) from Manitoba to other provinces and territories. The higher the GDP ( $x$  variable) in the destination, the higher the trade. The greater the distance (another  $x$  variable), the lower the trade. Each  $i$  is a different trading partner, and each  $t$  is a different year.

Provincial trade from Manitoba

Panel 1				
year (t)	trading partner (i)	exports (y)	gdp (x)	distance (x)
2007	Newfoundland and Labrador	143.7	34658	3221
2007	Prince Edward Island	28.6	5748	2537
2007	Nova Scotia	192	39439	2575
2007	New Brunswick	171.5	34342	2297
2007	Quebec	2285	362927	1936
2007	Ontario	5522.5	713004	1515
2007	Saskatchewan	1730.5	65052	536
2007	Alberta	2622.9	274159	1194
2007	British Columbia	1613.4	229376	1904
2007	Yukon	24.2	2098	2635
2007	Northwest Territories	77.6	5556	1746
2007	Nunavut	37.9	1740	2290
Panel 2				
year (t)	trading partner (i)	exports (y)	gdp (x)	distance (x)
2008	Newfoundland and Labrador	149.5	34137	3221
2008	Prince Edward Island	44.5	5807	2537
2008	Nova Scotia	227.3	40219	2575
2008	New Brunswick	240.8	34606	2297
2008	Quebec	2511.8	370086	1936
2008	Ontario	5372.3	712990	1515
2008	Saskatchewan	2040.2	68463	536
2008	Alberta	2737.4	278580	1194
2008	British Columbia	1831.6	231000	1904
2008	Yukon	28.5	2311	2635
2008	Northwest Territories	82.6	5031	1746
2008	Nunavut	49.2	1981	2290
Panel 3				
year (t)	trading partner (i)	exports (y)	gdp (x)	distance (x)
2009	Newfoundland and Labrador	157.7	30714	3221
2009	Prince Edward Island	33.7	5842	2537
2009	Nova Scotia	193.9	40363	2575
2009	New Brunswick	168	34105	2297
2009	Quebec	2095.7	367169	1936
2009	Ontario	6032.1	690852	1515
2009	Saskatchewan	1867.3	64812	536
2009	Alberta	2750.1	263141	1194
2009	British Columbia	1734.5	225497	1904
2009	Yukon	22	2465	2635
2009	Northwest Territories	72.5	4492	1746
2009	Nunavut	27.8	1849	2290

Figure: A more common way of arranging panel data.

### Provincial trade from Manitoba

year (t)	trading partner (i)	exports (y)	gdp (x)	distance (x)
2007	Newfoundland and Labrador	143.7	34658	3221
2008	Newfoundland and Labrador	149.5	34137	3221
2009	Newfoundland and Labrador	157.7	30714	3221
2010	Newfoundland and Labrador	118.6	32386	3221
2011	Newfoundland and Labrador	127.1	33271	3221
2012	Newfoundland and Labrador	155.8	31815	3221
2013	Newfoundland and Labrador	112.8	33486	3221
2014	Newfoundland and Labrador	121.8	33096	3221
2015	Newfoundland and Labrador	108.5	32709	3221
2016	Newfoundland and Labrador	130.2	33201	3221
2017	Newfoundland and Labrador	130.2	33689	3221
2018	Newfoundland and Labrador	124.5	32836	3221
2019	Newfoundland and Labrador	120.3	34157	3221
2020	Newfoundland and Labrador	122.4	32512	3221
2021	Newfoundland and Labrador	122.8	32836	3221
2022	Newfoundland and Labrador	389.1	32488	3221
2007	Prince Edward Island	28.6	5748	2537
2008	Prince Edward Island	44.5	5807	2537

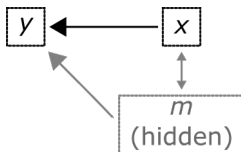
# The model, over time

We need to bring back the *subscripts* in our notation of the population model, and add a  $t$ :

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,i} + m_i + \epsilon_{it},$$

- ▶  $m_i$  is the effect the missing variable has on  $y_{it}$ .
- ▶  $m_i$  has an  $i$  subscript, but no  $t$  subscript. This means that it varies by individual, but not over time.
- ▶  $m_i$  is called *unobserved heterogeneity* because it accounts for differences in each person.
- ▶ Notice that  $y_{it}$  and  $x_{1,it}$  evolve over time, but  $m_i$  and  $x_{2,i}$  do not.  $x_{2,i}$  will end up being useless to us when we do *fixed effects* estimation.

**Figure:** A missing  $m$  variable that is correlated with  $x$  and that determines  $y$  will make estimation of the effect of  $x$  on  $y$  difficult (or impossible).



Let's redraw this and add in the *subscripts*.

If the missing variable  $m_i$  does not vary over time, but if the  $y_{it}$  and  $x_{it}$  variables *do* vary over time, then we can use a trick to solve the missing variable (unobserved heterogeneity) problem.  $y_{it}$  could be an individual's wages, for example, and  $x_{it}$  could be some economic policy or some choice that people are making, that changes over time. Some examples of  $m_i$  (unobserved heterogeneity) that we will be able to deal with, are:

- ▶ gender
- ▶ race
- ▶ ability

These are things that are specific to each individual, but that (usually) do not vary over time. If these things are unobserved, we can solve the problem using the **fixed effects** estimator!

# Provincial trade

Let's rewrite the population model in terms of a Manitoba trade model. We have data on the amount of exports from Manitoba to other provinces and territories, the GDP of the trading partner, and the distance of the trading partner.

$$\log(\text{exports}_{it}) = \beta_0 + \beta_1 \log(\text{GDP}_{it}) + \beta_2 \log(\text{distance}_i) + \epsilon_{it},$$

What variables might be missing from this model that could be correlated with both GDP and exports?

- ▶ Terrain difficulties and quality of roads / rails
- ▶ Historical linkages and path dependence
- ▶ Shared languages, cultures, and borders

Some of these characteristics we might be able to include in our model and control for, but not all. There will still be some *unobserved heterogeneity*.

# Pooled LS

All we really know to do so far is to ignore the panel structure, throw all the data together, and estimate *pooled* least squares (pool the data together without taking into account the time component). In R, this looks like:

```
1 trade <- read.csv("https://rtgodwin.com/data/trade2.csv")
2 mod <- lm(log(exports) ~ log(gdp) + log(distance), data =
   trade)
3 summary(mod)
```

```

1 Coefficients:
2           Estimate Std. Error t value Pr(>|t|)
3 (Intercept)   5.74166   0.41817   13.73  <2e-16 ***
4 log(gdp)       0.83199   0.01115   74.63  <2e-16 ***
5 log(distance) -1.15155   0.04733  -24.33  <2e-16 ***
6 ---
7 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
8
9 Residual standard error: 0.2753 on 189 degrees of freedom
10 Multiple R-squared:  0.9794, Adjusted R-squared:  0.9792
11 F-statistic: 4499 on 2 and 189 DF, p-value: < 2.2e-16

```

The interpretation is that an increase in the GDP of a trading partner of 1% is associated with a 0.83% increase in trade. But this estimate is wrong if there is unobserved heterogeneity!

# Marriage and wages

As another example, let's look at the long standing observation that married individuals make higher wages than single individuals. That is, we observe a positive correlation between marriage and wages. Does marriage **cause** wages? What could be some possible explanations for this association?

$$wage_{it} = \beta_0 + \beta_1 marriage_{it} + \beta_2 x_{it} + \beta_3 x_i + ability_i + \epsilon_{it},$$

- ▶  $wage_{it}$  and  $marriage_{it}$  are the workers' wages and marital status over time.
- ▶ We also have some time-varying controls ( $x_{it}$  - things like the amount of time worked that year) and some time invariant controls ( $x_i$  - like gender or race). It turns out that the  $x_i$  variable can't be used; it will disappear when we make the missing  $ability_i$  variable cancel out.
- ▶  $ability_i$  is a missing variable that determines wage.
- ▶ We can't estimate  $\beta_1$  and call it the causal effect of marriage on wage. Why?

It might be the case that *ability* is an attractive quality in a partner. People with higher ability are more likely to get married, and to have higher wages. Marriage just might be an indicator for a higher ability person, which is the true cause for higher wages (much like how fireplaces indicated a bigger house). Without some trick, we can't get an unbiased and consistent estimator for the effect of marriage on wages using LS.

Let's try pooled LS with some marriage data. The data comes from the Panel Study of Income Dynamics, a survey run in the U.S. that tracks the same workers yearly. See Figure 4.

Figure: Panel data on wages and marital status.

	experience	weeks	occupation	industry	south	smsa	married	gender	union	education	ethnicity	wage	year	id
1	3	32	white	no	yes	no	yes	male	no	9	other	260	1976	1
2	4	43	white	no	yes	no	yes	male	no	9	other	305	1977	1
3	5	40	white	no	yes	no	yes	male	no	9	other	402	1978	1
4	6	39	white	no	yes	no	yes	male	no	9	other	402	1979	1
5	7	42	white	yes	yes	no	yes	male	no	9	other	429	1980	1
6	8	35	white	yes	yes	no	yes	male	no	9	other	480	1981	1
7	9	32	white	yes	yes	no	yes	male	no	9	other	515	1982	1
8	30	34	blue	no	no	no	yes	male	no	11	other	475	1976	2
9	31	27	blue	no	no	no	yes	male	no	11	other	500	1977	2
10	32	33	blue	yes	no	no	yes	male	yes	11	other	525	1978	2
11	33	30	blue	yes	no	no	yes	male	no	11	other	695	1979	2
12	34	30	blue	yes	no	no	yes	male	no	11	other	810	1980	2
13	35	37	blue	yes	no	no	yes	male	no	11	other	890	1981	2
14	36	30	blue	yes	no	no	yes	male	no	11	other	912	1982	2
15	6	50	blue	yes	no	no	yes	male	yes	12	other	285	1976	3
16	7	51	blue	yes	no	no	yes	male	yes	12	other	624	1977	3

```

1 marriage <- read.csv("https://rtgodwin.com/data/marriage.csv
  ")
2 pooled.ls <- lm(log(wage) ~ experience + weeks + occupation
  + industry + south + smsa + married
3   + gender + union + education + ethnicity, data = marriage)
4 summary(pooled.ls)

```

```

1 Coefficients:
2           Estimate Std. Error t value Pr(>|t|)
3 (Intercept)   5.4411580   0.0716880   75.901 < 2e-16 ***
4 experience     0.0103709   0.0005354   19.372 < 2e-16 ***
5 weeks         0.0049445   0.0011059    4.471 7.99e-06 ***
6 occupationblue -0.1486343   0.0149933  -9.913 < 2e-16 ***
7 industryyes   0.0530577   0.0120663    4.397 1.12e-05 ***
8 southyes     -0.0532102   0.0128247  -4.149 3.41e-05 ***
9 smsayes      0.1453038   0.0123480   11.767 < 2e-16 ***
10 marriedyes   0.0660797   0.0210208    3.144 0.00168 **
11 genderfemale -0.3533020   0.0256743  -13.761 < 2e-16 ***
12 unionyes     0.1020757   0.0130870    7.800 7.79e-15 ***
13 education    0.0571540   0.0026749   21.366 < 2e-16 ***
14 ethnicityafam -0.1671226   0.0225679  -7.405 1.58e-13 ***

```

So, marriage increases wage by 6.6%, and the effect is significant at 1%. But if marriage is positively correlated with ability, this effect is over-stated!

# The within estimator

We can use panel data to solve the missing variable problem (and other problems of endogeneity), by using the *within*, also called the *fixed effects*, estimator. Let's rewrite a simple population model with the missing variable:

$$y_{it} = \beta_0 + \beta_1 x_{it} + m_i + \epsilon_{it},$$

- ▶ What happens if something doesn't change over time, and you take it's average over time?
- ▶ How can we get rid of (subtract) the missing variable (the unobserved heterogeneity)?

The trick involves taking the averages of the variables *over time*. That is, we're going to take each person's wage, and average it over the time in which they are observed. Let's define these time-averages:

$$\bar{y}_i \equiv \frac{1}{T} \sum_{t=1}^T y_{it} \quad ; \quad \bar{x}_i \equiv \frac{1}{T} \sum_{t=1}^T x_{it}$$

# The trick

- ▶ If we could see the missing variable (we can't), what would be its average over time? Does it change?
- ▶ What about  $\beta_0$  and  $\beta_1$ ?
- ▶ What about the error term?

$$\frac{1}{T} \sum_{t=1}^T m_i = m_i \quad ; \quad \frac{1}{T} \sum_{t=1}^T \beta_0 = \beta_0 \quad ; \quad \frac{1}{T} \sum_{t=1}^T \beta_1 = \beta_1 \quad ; \quad \frac{1}{T} \sum_{t=1}^T \epsilon_{it} = \bar{\epsilon}_i$$

So anything that is constant over time, is the same as its average over time.

# The trick

$$y_{it} = \beta_0 + \beta_1 x_{it} + m_i + \epsilon_{it} \quad (1)$$

Let's instead write the population model in terms of the over-time-averages (note that the  $\beta$  are the same in both equations):

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + m_i + \bar{\epsilon}_i \quad (2)$$

Can you see a way to delete the missing variable? Subtract (2) from (1)!

$$(y_{it} - \bar{y}_i) = (\beta_0 - \beta_0) + (\beta_1 x_{it} - \beta_1 \bar{x}_i) + (m_i - m_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$

$$(y_{it} - \bar{y}_i) = (\beta_1 x_{it} - \beta_1 \bar{x}_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{\epsilon}_{it}$$

The  $\ddot{y}_{it}$  and  $\ddot{x}_{it}$  are called *time demeaned variables*. They only use variation *within* each individual. Estimating the model using time demeaned variables solves the unobserved heterogeneity problem.

# Time-demeaning deletes the unobserved heterogeneity

- ▶ When we subtract the time-averages (time demeaned variables) from both sides of the equation, we are deleting the unobserved heterogeneity (missing variable).
- ▶ Note that any individual characteristics that are time invariant, such as race, gender, etc., will also be deleted from the regression. That is why  $x_i$  is not needed as a control, only  $x_{it}$ .
- ▶ We can estimate this model in R by manually de-meaning the variables, and estimating the model by LS. This is called the **within** estimator, which is also equivalent to the fixed effects estimator.

## Within estimator for marriage

Let's focus on a simple model for the purposes of the example (we can add the extra controls but they do not change the estimates on marriage much):

$$\log(\ddot{wage})_{it} = \beta_1 \ddot{marriage}_{it} + \beta_2 \ddot{experience}_{it} + \ddot{\epsilon}_{it},$$

Here is the R code to manually create the time demeaned variables for the marriage data set:

```
1 marriage$log.wage.bar <- rep(tapply(log(marriage$wage),
2   marriage$id, mean), each = 7)
3 marriage$experience.bar <- rep(tapply(marriage$experience,
4   marriage$id, mean), each = 7)
5 marriage$married.bar <- rep(tapply(marriage$married,
6   marriage$id, mean), each = 7)
7 marriage$log.wage.within <- log(marriage$wage) - marriage$log.wage.bar
8 marriage$experience.within <- marriage$experience - marriage$experience.bar
9 marriage$married.within <- marriage$married - marriage$married.bar
```

The above code is hard to read. Run it, and see what happens to the data set. Now, estimate the population model using the time demeaned variables (the `-1` tells R to not include the intercept):

```
1 within <- lm(log.wage.within ~ married.within + experience.  
  within -1, data = marriage)  
2 summary(within)
```

```
1 Coefficients:  
2      Estimate Std. Error t value Pr(>|t|)  
3 married.within -0.032350  0.017692  -1.829  0.0675 .  
4 experience.within  0.096866  0.001101  87.947 <2e-16 ***  
5 ---  
6 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
7  
8 Residual standard error: 0.1421 on 4163 degrees of freedom  
9 Multiple R-squared:  0.6507, Adjusted R-squared:  0.6506  
10 F-statistic: 3878 on 2 and 4163 DF, p-value: < 2.2e-16
```

The effect of marriage is now  $-3.2\%$  and insignificant at the  $5\%$  level.

# Fixed effects estimation

There is another way to think about the problem. There is unobserved heterogeneity amongst the individuals in the data set. Each person is unique. Why can't we control for this? We can! Give each person their own dummy variable. The population model when we control for unobserved heterogeneity looks like:

$$y_{it} = \alpha_i D_i + \beta x_{it} + \epsilon_{it}$$

$D_i$  is a dummy variable equal to 1 if the individual is person  $i$ , and 0 otherwise. For example,  $D_4 = 1$  for person 4, and 0 for everyone else. This means we are adding  $n$  dummy variables to the model (that's a lot). We need to have reasonably large  $T$  for this to work. When each person gets their own dummy variable, this is called fixed effects estimation.

## Fixed effects and within estimation is the same

Way back in Section 5.3.3 we saw that including a dummy for gender is the same as dividing the sample into two parts, and taking the average difference. So it shouldn't be too much of a surprise that we can affect a difference in sample means by using dummy variables. That is, including the fixed effects in the model is equivalent to time demeaning the variables.

Although I started with time demeaning, economists usually prefer to think in terms of fixed effects. A fixed effects model is typically written something like:

$$y_{it} = \beta x_{it} + \alpha_i + \epsilon_{it}$$

where the dummy variable has been omitted and the  $\alpha_i$  represents the fixed effect for each individual. While this model is conceptually preferred, it is actually estimated using the time demeaned variables, which is computationally easier.

## Fixed effects for marriage data

We should be able to get the same -3.2% effect of marriage by giving everyone a dummy variable:

```
1 fe.dummies <- lm(log(wage) ~ married + experience + as.  
  factor(id), data = marriage)  
2 summary(fe.dummies)
```

That's a lot of dummy variables! (There were 595). We do get the same effect for marriage though.

```

1 Coefficients:
2           Estimate Std. Error t value Pr(>|t|)
3 (Intercept)    5.415910   0.061565   87.971 < 2e-16 ***
4 married       -0.032350   0.019110   -1.693 0.090571 .
5 experience     0.096866   0.001190   81.419 < 2e-16 ***
6 as.factor(id)2 -2.081706   0.088103  -23.628 < 2e-16 ***
7 as.factor(id)3  0.246070   0.082289   2.990 0.002806 **
8 as.factor(id)4 -2.322404   0.090352  -25.704 < 2e-16 ***
9 as.factor(id)5  0.261531   0.082460   3.172 0.001529 **
10 as.factor(id)6 -1.099762   0.086481  -12.717 < 2e-16 ***
11 as.factor(id)7 -0.816274   0.083271   -9.803 < 2e-16 ***
12 as.factor(id)8 -1.310293   0.085419  -15.340 < 2e-16 ***
13 as.factor(id)9  0.932437   0.082038   11.366 < 2e-16 ***
14 ...
15 as.factor(id)196 0.166094   0.082176   2.021 0.043334 *
16 as.factor(id)197 -1.697629   0.086482  -19.630 < 2e-16 ***
17 as.factor(id)198 -1.592690   0.087676  -18.166 < 2e-16 ***
18 [ reached 'max' / getOption("max.print") -- omitted 397
    rows ]
19 ---
20 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
21
22 Residual standard error: 0.1535 on 3568 degrees of freedom
23 Multiple R-squared: 0.9052, Adjusted R-squared: 0.8894
24 F-statistic: 57.19 on 596 and 3568 DF, p-value: < 2.2e-16

```

# feols()

Of course R can make this much easier for us.

```
1 library(fixest)
2 fe.mod <- feols(log(wage) ~ experience + married | id, data
  = marriage)
3 summary(fe.mod)
```

```
1 OLS estimation, Dep. Var.: log(wage)
2 Observations: 4,165
3 Fixed-effects: id: 595
4 Standard-errors: Clustered (id)
5           Estimate Std. Error  t value  Pr(>|t|)
6 experience  0.096866   0.001770  54.72401 < 2.2e-16 ***
7 married    -0.032350   0.026273  -1.23133  0.21869
8 ---
9 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
10 RMSE: 0.142054      Adj. R2: 0.889406
11                   Within R2: 0.65075
```

A few things to notice:

- ▶ `feols()` reports the **Within R2**. This is the one we should use. The dummies for each person should not be counted.
- ▶ While the estimated effect of  $-3.2\%$  is the same across all methods, `feols()` now reports it as insignificant at the 10% level. This is because it has accounted for *heteroskedasticity*. More specifically, it has clustered the standard errors at the individual level. Each individual is allowed to have different variance, but the variance does not change over time.