Panel Data Analysis with Stata Part 1 Fixed Effects and Random Effects Models

Pillai N., Vijayamohanan

2016

Panel Data Analysis with Stata

Part 1

Fixed Effects and Random Effects Models

Vijayamohanan Pillai N.

Centre for Development Studies,

Kerala, India.

e-mail; vijayamohan@cds.ac.in

Panel Data Analysis with Stata

Part 1

Fixed Effects and Random Effects Models

Abstract

The present work is a part of a larger study on panel data. Panel data or longitudinal data (the older terminology) refers to a data set containing observations on multiple phenomena over multiple time periods. Thus it has two dimensions: spatial (cross-sectional) and temporal (time series). The main advantage of panel data comes from its solution to the difficulties involved in interpreting the partial regression coefficients in the framework of a cross-section only or time series only multiple regression. Depending upon the assumptions about the error components of the panel data model, whether they are fixed or random, we have two types of models, fixed effects and random effects. In this paper we explain these models with regression results using a part of a data set from a famous study on investment theory by Yehuda Grunfeld (1958), who tried to analyse the effect of the (previous period) real value of the firm and the (previous period) real capital stock on real gross investment. We consider mainly three types of panel data analytic models: (1) constant coefficients (pooled regression) models, (2) fixed effects models, and (3) random effects models. The fixed effects model is discussed under two assumptions: (1) heterogeneous intercepts and homogeneous slope, and (2) heterogeneous intercepts and slopes. We discuss all the relevant statistical tests in the context of all these models.

Panel Data Analysis with Stata

Part 1

Fixed Effects and Random Effects Models

Panel Data Analysis: A Brief History

According to Marc Nerlove (2002), the fixed effects model of panel data techniques originated from the least squares methods in the astronomical work of Gauss (1809) and Legendre (1805) and the random effects or variance-components models, with an English astronomer George Biddell Airy, who published a monograph in 1861, in which he made explicit use of a variance components model for the analysis of astronomical panel data. The next stage is connected to R. A. Fisher, who coined the terms and developed the methods of variance and analysis of variance (Anova) in 1918; he elaborated both fixed effects and random effects models in Chapter 7: 'Interclass Correlations and the Analysis of Variance' and in Chapter 8: 'Further applications of the Analysis of Variance' of his 1925 work *Statistical Methods for Research Workers*. However, he was not much clear on the distinction between these two models. That had to wait till 1947, when Churchill Eisenhart came out with his 'Survey' that made clear the distinction between fixed effects and random effects models for the analysis of non-experimental versus experimental data. The random effects, mixed, and variance-components models in fact posed considerable computational problems for the statisticians. In 1953, CR Henderson developed the method-of-moments techniques for analysing random effects and mixed models; and in 1967, HO Hartley and JNK Rao devised the maximum likelihood (ML) methods for variance components models. The dynamic panel models started with the famous Balestra-Nerlove (1966) models. Panel data analysis grew into its maturity with the first conference on panel data econometrics in August 1977 in Paris, organized by Pascal Mazodier. Since then, the field has witnessed ever-expanding activities in both methodological and applied research.

Panel data or longitudinal data (the older terminology) refer to a data set containing observations on multiple phenomena over multiple time periods. Thus it has two dimensions: spatial (cross-sectional) and temporal (time series). In general, we can have two panels: micro and macro panels – surveying (usually a large) sample of individuals or households or firms or industries over (usually a short) period of time yields micro panels, whereas macro panels consist of (usually a large) number of countries or regions over (usually a large) number of years.

Nomenclature

A cross sectional variable is denoted by x_i, where i is a given case (household or industry or nation; $i = 1, 2, \ldots, N$), and a time series variable by x_t, where t is a given time point ($t = 1, 2, \ldots, T$). Hence a panel variable can be written as x_{it}, for a given case at a particular time. A typical panel data set is given in Table 1 below, which describes the personal disposable income (PDY) and personal expenditure in three countries, Utopia, Lilliput and Troy over a period of time from 1990 – 2015.

Table 1: A Typical Panel Data Set

Country	Year	PDY	PE
Utopia	1990	6500	5000
Utopia	1991	7000	6000
……	……	……	……
……	……	……	……
Utopia	2015	15000	11000
Lilliput	1990	1500	1300
Lilliput	1991	1700	1600
……	……	……	……
……	……	……	……
Lilliput	2015	5450	5000
Troy	1990	2200	1800
Troy	1991	2400	2000
……	……	……	……
……	……	……	……
Troy	2015	8500	7500

Depending upon the configuration of space and time relative to each other, panels can take two forms: in the first case, time is nested or stacked within the cross-section and in the second, cross-section is nested/stacked within time, as Table 2 below shows:

Table 2: Two Forms of Panel Configuration

Time nested within the cross-section		cross-section nested within time	
Country	Year	Year	Country
Utopia	1990	1990	Utopia
Utopia	1991	1990	Lilliput
……	……	1990	Troy
……	……	1991	Utopia
Utopia	2015	1991	Lilliput
Lilliput	1990	1991	Troy
Lilliput	1991	1992	Utopia
……	……	……	……
……	……	……	……
Lilliput	2015	……	……
Troy	1990	……	……
Troy	1991	……	……
……	……	2015	Utopia
……	……	2015	Lilliput
Troy	2015	2015	Troy

Again, depending upon whether the panels include missing values or not, we can have two varieties: balanced and unbalanced panel. Balanced panel does not have any no missing values, whereas the unbalanced one has, as Table 3 illustrates;

Table 3: Balanced and Unbalanced Panel

Balanced panel					Unbalanced Panel				
Person Sl No	Year	Income	Age	Sex	Person Sl No	Year	Income	Age	Sex
1	2004	800	45	1	1	2005	1750	32	1
1	2005	900	46	1	1	2006	2500	33	1
1	2006	1000	47	1	2	2004	2000	40	2
2	2004	1500	29	2	2	2005	2500	41	2
2	2005	2000	30	2	2	2006	2800	42	2
2	2006	2500	31	2	3	2006	2500	28	2

We have two more models, depending upon the relative size of space and time, short and long panels. In a short panel, the number of time periods (T) is less than the number of cross section units (N), and in a long panel, T > N. Note that Table 1 above gives a long panel.

Advantages of Panel Data

Hsiao (2014) Baltagi (2008) and Andreß *et al*. (2013) list a number of advantages of using panel data, instead of pure cross-section or pure time series data.

The obvious benefit is in terms of obtaining a large sample, giving more degrees of freedom, more variability, more information and less multicollinearity among the variables. A panel has the advantage of having N cross-section and T time series observations, thus contributing a total of NT observations. Another advantage comes with a possibility of controlling for individual or time heterogeneity, which the pure cross-section or pure time series data cannot afford. Panel data also opens up a scope for dynamic analysis.

The main advantage of panel data comes from its solution to the difficulties involved in interpreting the regression coefficients in the framework of a cross-section only or time series only regeression, as we explain below.

Regression Analysis: Some Basics

Let us consider the following cross-sectional multiple regression with two explanatory variables, X_1 and X_2:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \; ; \; i = 1, 2, \ldots, N. \qquad \ldots. (1)$$

Note that X_1 is said to be the covariate with respect to X_2 and vice versa. Covariates act as controlling factors for the variable under consideration. In the presence of the control variables, the regression coefficients βs are partial regression coefficients. Thus, β_1 represents the marginal effect of X_1 on Y, keeping all other variables, here X_2, constant. The latter part, that is, keeping X_2 constant, means the marginal effect of X_1 on Y is obtained after removing the linear effect of X_2 from *both* X_1 and Y. A similar explanation goes for β_2 also. Thus multiple regression facilitates to obtain the *pure* marginal effects by including all the relevant covariates and thus controlling for their heterogeneity.

This we'll discuss in a little detail below. We begin with the concept of partial correlation coefficient. Suppose we have three variables, X_1, X_2 and X_3. The simple correlation coefficient r_{12} gives the degree of correlation between X_1 and X_2. It is possible that X_3 may have an influence on both X_1 and X_2. Hence a question comes up: Is an observed correlation between X_1 and X_2 merely due to the influence of X_3 on both? That is, is the correlation merely due to the common influence of X_3? Or, is there a *net* correlation between X_1 and X_2, over and above the correlation due to the common influence of X_3? It is this *net* correlation between X_1 and X_2 that the partial correlation coefficient captures after removing the influence of X_3 from each, and then estimating the correlation between the *unexplained* residuals that remain. To prove this, we define the following:

Coefficients of correlation between X_1 and X_2, X_1 and X_3, and X_2 and X_3 are given by r_{12}, r_{13}, and r_{23} respectively, defined as

$$r_{12} = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \sum x_2^2}} = \frac{\sum x_1 x_2}{s_1 s_2},\ r_{13} = \frac{\sum x_1 x_3}{\sqrt{\sum x_1^2 \sum x_3^2}} = \frac{\sum x_1 x_3}{s_1 s_3} \text{ and } r_{23} = \frac{\sum x_2 x_3}{\sqrt{\sum x_2^2 \sum x_3^2}} = \frac{\sum x_2 x_3}{s_2 s_3}. \ldots(2')$$

Note that the lower case letters, x_1, x_2, and x_3, denote the respective variables in mean deviation form; thus $(x_1 = X_{1i} - \bar{X}_1)$, etc., and s_1, s_2, and s_3 denote the standard deviations of the three variables.

The common influence of X_3 on both X_1 and X_2 may be modeled in terms of regressions of X_1 on X_3, and X_2 on X_3, with b_{13} as the slope of the regression of X_1 on X_3, given (in deviation form) by $b_{13} = \frac{\sum x_1 x_3}{\sum x_3^2} = r_{13}\frac{s_1}{s_3}$, and b_{23} as that of the regression of X_2 on X_3 given by $b_{23} = \frac{\sum x_2 x_3}{\sum x_3^2} = r_{23}\frac{s_2}{s_3}$.

Given these regressions, we can find the respective unexplained residuals. The residual from the regression of X_1 on X_3 (in deviation form) is $e_{1.3} = x_1 - b_{13}\, x_3$, and that from the regression of X_2 on X_3 is $e_{2.3} = x_2 - b_{23}\, x_3$.

Now the partial correlation between X_1 and X_2, net of the effect of X_3, denoted by $r_{12.3}$, is defined as the correlation between these *unexplained* residuals and is given by $r_{12.3} = \frac{\sum e_{1.3} e_{2.3}}{\sqrt{\sum e_{1.3}^2}\sqrt{\sum e_{2.3}^2}}$. Note that since the least-squares residuals have zero means, we need not write them in mean deviation form. We can directly estimate the two sets of residuals and then find out the correlation coefficient between them. However, the usual practice is to express them in terms of simple correlation coefficients. Using the definitions given above of the residuals and the regression coefficients, we have for the residuals: $e_{1.3} = x_1 - r_{13}\frac{s_1}{s_3}x_3$, and $e_{2.3} = x_2 - r_{23}\frac{s_2}{s_3}x_3$, and hence, upon simplification, we get

$$r_{12.3} = \frac{\sum e_{1.3} e_{2.3}}{\sqrt{\sum e_{1.3}^2}\sqrt{\sum e_{2.3}^2}} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}.$$

"This is the statistical equivalent of the economic theorist's technique of impounding certain variables in a *ceteris paribus* clause." (Johnston, 1972: 58). Thus the partial correlation coefficient between X_1 and X_2 is said to be obtained by keeping X_3 constant. This idea is clear in the above formula for the partial correlation coefficient as a *net* correlation between X_1 and X_2 after removing the influence of X_3 from each.

When this idea is extended to multiple regression coefficients, we have the partial regression coefficients. Consider the regression equation in three variables, X_1, X_2 and X_3:

$$X_{1i} = \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \; ; \; i = 1, 2, \ldots, N. \qquad \ldots. (3)$$

Since the estimated regression coefficients are partial ones, the equation can be written as:

$$X_{1i} = a + b_{12.3} X_{2i} + b_{13.2} X_{3i}, \qquad \ldots. (4)$$

where the lower case letters (a and b) are the OLS estimates of α and β respectively.

The estimate $b_{12.3}$ is given by:

$$b_{12.3} = \frac{\sum x_1 x_2 \sum x_3^2 - \sum x_1 x_3 \sum x_2 x_3}{\sum x_2^2 \sum x_3^2 - (\sum x_2 x_3)^2}.$$

Now using the definitions of simple and partial correlation coefficients in (2) and (2'), we can rewrite the above as:

$$b_{12.3} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \frac{s_1}{s_2}.$$

Why $b_{12.3}$ is called a partial regression coefficient is now clear from the above definition: it is obtained after removing the common influence of X_3 from both X_1 and X_2.

Similarly, we have the estimate $b_{13.2}$ given by:

$$b_{12.3} = \frac{\sum x_1 x_3 \sum x_2^2 - \sum x_1 x_2 \sum x_2 x_3}{\sum x_2^2 \sum x_3^2 - (\sum x_2 x_3)^2} = \frac{r_{13} - r_{12} r_{32}}{1 - r_{23}^2} \frac{s_1}{s_3},$$

obtained after removing the common influence of X_2 from both X_1 and X_3.

Thus the fundamental idea in partial (correlation/regression) coefficient is estimating the *net* correlation between X_1 and X_2 after removing the influence of X_3 from each, by computing the correlation between the *unexplained* residuals that remain (after eliminating the influence of X_3 from both X_1 and X_2). The classical text books describe this procedure as controlling for or accounting for the effect of X_3, or keeping that variable constant; whereas Tukey (in his classic *Exploratory Data Analysis*, 1970, chap. 23) characterizes this as "adjusting for simultaneous linear change in the other predictor", that is, X_3. Above all these seeming semantic differences, let us keep the underlying idea alive, while interpreting the regression coefficients.

Thus multiple regression facilitates controlling for the heterogeneity of the covariates.

One major problem with cross section regression is that it fails to control for cross sectional, individual, panel-specific, heterogeneity. Consider a random sample of 50 households; every household is different from one another. This unobserved household heterogeneity can, however, be captured by means of 50 dummy variables in the regression without a constant. But this is just

impossible for this sample, as estimation breaks down because the number of observations is less than the number of parameters to be estimated.

The same problem haunts time series regression also. Consider the following time series multiple regression with two explanatory variables, X_1 and X_2:

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t \; ; t = 1, 2, \ldots, T. \qquad \ldots. (2)$$

We have the same explanation for the marginal effects here also, and we know every time point in this system is different from one another. But we cannot account/control for this time heterogeneity by including time dummies, lest the estimation break down.

It is here panel data regression comes in with a solution. This we explain below.

The Panel Data Regression

Now combining (1) and (2), we get a pooled data set, which forms a panel data with the following panel regression:

$$Y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it} + u_{it} \; ; i = 1, 2, \ldots, N; t = 1, 2, \ldots, T. \qquad \ldots. (3)$$

How do we account for the cross section and time heterogeneity in this model? This is done by using a two-way error component assumption for the disturbances, u_{it}, with

$$u_{it} = \mu_i + \lambda_t + v_{it} \; , \qquad \ldots (4)$$

where μ_i represents the unobservable individual (cross section) heterogeneity, λ_t denotes the unobservable time heterogeneity and v_{it} is the remaining random error term. The first two components (μ_i and λ_t) are also called within component and the last (v_{it}), panel or between component.

Now depending upon the assumptions about these error components, whether they are fixed or random, we have two types of models, fixed effects and random effects. If we assume that the μ_i and λ_t are fixed parameters to be estimated and the random error term, v_{it}, is identically and independently distributed with zero mean and constant variance σ_v^2 (homoscedasticity), that is, $v_{it} \sim \text{IID}(0, \sigma_v^2)$, then equation (3) gives a two-way fixed effects error component model or simply a fixed effects model. On the other hand, if we assume that the μ_i and λ_t are random just like the random error term, that is, μ_i, λ_t and v_{it} are all identically and independently distributed with zero mean and constant variance, or, $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$, $\lambda_t \sim \text{IID}(0, \sigma_\lambda^2)$, and $v_{it} \sim \text{IID}(0, \sigma_v^2)$, with further assumptions that they are all independent of each other and of explanatory variables, then

equation (3) gives a two-way random effects error component model or simply a random effects model.

Instead of both the error components, μ_i and λ_t, if we consider any one component only at a time, then we have a one-way error component model, fixed or random effects. Here the error term u_{it} in (3) will become:

$$u_{it} = \mu_i + v_{it}\,, \quad \text{or,} \qquad \ldots (4')$$

$$u_{it} = \lambda_t + v_{it}\,. \qquad \ldots (4'')$$

We can have one-way error component fixed or random effects model with the appropriate assumptions about the error components, that is, whether μ_i or λ_t is assumed to be fixed or random.

In the following we explain these models with regression results using a part of a data set from a famous study on investment theory by Yehuda Grunfeld (1958), who tried to analyse the effect of the (previous period) real value of the firm (F) and the (previous period) real capital stock (K) on real gross investment (I). For each variable, a positive effect is expected a priori. His original study included 10 US corporations for 20 years during 1935–1954. We consider only four companies – General Electric (GE), General Motor (GM), U.S. Steel (US), and Westinghouse (West) – for the whole period that gives 80 observations.

The investment model of Grunfeld (1958) is given as

> Real gross investment (millions of dollars deflated by implicit price deflator of producers' durable equipment), $I_{it} = f(F_{it\text{-}1}, K_{it\text{-}1})$,

where

> F_{it} = Real value of the firm (share price times number of shares plus total book value of debt; millions of dollars deflated by implicit price deflator of GNP), and
>
> K_{it} = Real capital stock (accumulated sum of net additions to plant and equipment, deflated by depreciation expense deflator – 10 year moving average of WPI of metals and metal products)

The data that we use for the four cross sectional units and 20 time periods are briefly given below:

Table 4: The Panel Data That We Use

Industry	Time	I	F (=F_{it-1})	K (=K_{it-1})	Industry	Time	I	F (=F_{it-1})	K (=K_{it-1})
GE	1935	33.1	1170.6	97.8	US	1935	209.9	1362.4	53.8
GE	1936	45	2015.8	104.4	US	1936	355.3	1807.1	50.5
GE	1937	77.2	2803.3	118	US	1937	469.9	2673.3	118.1
GE	1938	44.6	2039.7	156.2	US	1938	262.3	1801.9	260.2
GE	1939	48.1	2256.2	172.6	US	1939	230.4	1957.3	312.7
GE	1940	74.4	2132.2	186.6	US	1940	361.6	2202.9	254.2
GE	1941	113	1834.1	220.9	US	1941	472.8	2380.5	261.4
GE	1942	91.9	1588	287.8	US	1942	445.6	2168.6	298.7
GE	1943	61.3	1749.4	319.9	US	1943	361.6	1985.1	301.8
GE	1944	56.8	1687.2	321.3	US	1944	288.2	1813.9	279.1
GE	1945	93.6	2007.7	319.6	US	1945	258.7	1850.2	213.8
GE	1946	159.9	2208.3	346	US	1946	420.3	2067.7	232.6
GE	1947	147.2	1656.7	456.4	US	1947	420.5	1796.7	264.8
GE	1948	146.3	1604.4	543.4	US	1948	494.5	1625.8	306.9
GE	1949	98.3	1431.8	618.3	US	1949	405.1	1667	351.1
GE	1950	93.5	1610.5	647.4	US	1950	418.8	1677.4	357.8
GE	1951	135.2	1819.4	671.3	US	1951	588.2	2289.5	341.1
GE	1952	157.3	2079.7	726.1	US	1952	645.2	2159.4	444.2
GE	1953	179.5	2371.6	800.3	US	1953	641	2031.3	623.6
GE	1954	189.6	2759.9	888.9	US	1954	459.3	2115.5	669.7
GM	1935	317.6	3078.5	2.8	WEST	1935	12.93	191.5	1.8
GM	1936	391.8	4661.7	52.6	WEST	1936	25.9	516	0.8
GM	1937	410.6	5387.1	156.9	WEST	1937	35.05	729	7.4
GM	1938	257.7	2792.2	209.2	WEST	1938	22.89	560.4	18.1
GM	1939	330.8	4313.2	203.4	WEST	1939	18.84	519.9	23.5
GM	1940	461.2	4643.9	207.2	WEST	1940	28.57	628.5	26.5
GM	1941	512	4551.2	255.2	WEST	1941	48.51	537.1	36.2
GM	1942	448	3244.1	303.7	WEST	1942	43.34	561.2	60.8
GM	1943	499.6	4053.7	264.1	WEST	1943	37.02	617.2	84.4
GM	1944	547.5	4379.3	201.6	WEST	1944	37.81	626.7	91.2
GM	1945	561.2	4840.9	265	WEST	1945	39.27	737.2	92.4
GM	1946	688.1	4900	402.2	WEST	1946	53.46	760.5	86
GM	1947	568.9	3256.5	761.5	WEST	1947	55.56	581.4	111.1
GM	1948	529.2	3245.7	922.4	WEST	1948	49.56	662.3	130.6
GM	1949	555.1	3700.2	1020.1	WEST	1949	32.04	583.8	141.8
GM	1950	642.9	3755.6	1099	WEST	1950	32.24	635.2	136.7
GM	1951	755.9	4833	1207.7	WEST	1951	54.38	732.8	129.7
GM	1952	891.2	4924.9	1430.5	WEST	1952	71.78	864.1	145.5
GM	1953	1304.4	6241.7	1777.3	WEST	1953	90.08	1193.5	174.8
GM	1954	1486.7	5593.6	2226.3	WEST	1954	68.6	1188.9	213.5

Source: the online complements to Baltagi (2001):
http://www.wiley.com/legacy/wileychi/baltagi/.

Note that we have a balanced long panel (T = 20 > N = 4), where time is nested/stacked within cross section.

The model is generally written in matrix notation as:

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it} ;$$

$$u_{it} \sim \text{IID}(0, \sigma_u^2); \ \text{Cov}(x_{it}, u_{it}) = 0;$$

where y_{it} is the dependent variable, x_{it} is the vector of regressors,

β is the vector of coefficients,

u_{it} is the error term, independently and identically distributed with zero mean and σ_u^2 variance; and

α_i = individual effects: captures effects of the *i*-th individual-specific variables that are constant over time.

Panel Data with Stata

Unlike Gretl and EViews, Stata cannot receive data through dragging and dropping of excel file. We can open only a Stata file through the File → Open command. To enter data saved in Excel format, go to File → Import and select Excel spreadsheet. Next browse your Excel file and import the relevant sheet; mark "Import first row as variable names". Or, in the command space, we can type:

. import excel "C:\Users\CDS 2\Desktop\Panel data Grunfeld.xlsx", sheet("Sheet2") firstrow

Note that we have a string variable "Industry" that Stata cannot identify; we have to generate a corresponding numerical variable by typing:

. encode Industry, gen(ind)

Alternatively, we can also type:

. egen ind = group(Industry)

We can see this new variable "ind" by typing:

. list Industry ind in 1/80, nolabel sepby(Industry)

Next we have to declare the data set to be a panel data. This we do by going to

Statistics → Longitudinal/panel data → Setup and utilities → Declare dataset to be panel data

Now set the panel id variable ("ind") , time variable ("Time") and the time unit (yearly).

Or, w can type:

. xtset ind Time, yearly

When we input this command, Stata will respond with the following:

panel variable: ind (strongly balanced)

time variable: Time, 1935 to 1954

delta: 1 year

Now that we have "xtset" the panel data, we can go for estimation.

Types of Panel Analytic Models:

We consider mainly three types of panel data analytic models: (1) constant coefficients (pooled regression) models, (2) fixed effects models, and (3) random effects models.

1. The Constant Coefficients (Pooled Regression) Model

If there is neither significant cross sectional nor significant temporal effect, we could pool all of the data and run an ordinary least squares (OLS) regression model with an intercept α and slope coefficients βs constant across companies and time:

$$I_{it} = \alpha + \beta_1 F_{it-1} + \beta_2 K_{it-1} + u_{it} \; ; \quad u_{it} \sim \text{IIN}(0, \sigma_u^2); \quad i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20.$$

Note that for OLS regression in Stata, we need not "xtset" panel data; rather we can directly go to OLS regression through

Statistics → Linear models and related → Linear regression

Or, type

. regress I F K

or

. reg I F K

The regression output appears:

```
. regress I F K

      Source |       SS       df       MS              Number of obs =      80
-------------+------------------------------           F(  2,    77) =  119.46
       Model |  4847828.25     2  2423914.13            Prob > F      =  0.0000
    Residual |   1562318.8    77  20289.8545            R-squared     =  0.7563
-------------+------------------------------           Adj R-squared =  0.7499
       Total |  6410147.05    79  81141.1019            Root MSE      =  142.44

------------------------------------------------------------------------------
           I |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           F |   .1101177   .0137482     8.01   0.000     .0827415    .1374938
           K |   .3043034   .0492661     6.18   0.000      .206202    .4024048
       _cons |  -63.30815   29.63878    -2.14   0.036    -122.3265   -4.289797
------------------------------------------------------------------------------
```

The Stata result includes some summary statistics and the estimates of regression coefficients. The upper left part reports an analysis-of-variance (ANOVA) table with sum of squares (SS), degrees of freedom (df), and mean sum of squares (MS). Thus we find the total sum of squares is 6410147.05, of which 4847828.25 is accounted for by the model and 1562318.8 is left unexplained (residual). Note that as the regression includes a constant, the total sum of squares, as well as the sum of squares due to the model, represents the sum of squares after removing the respective means. Also reported are the degrees of freedom, with total degrees of freedom of 79 (that is, 80 observations minus 1 for the mean removal), out of which the model accounts for 2 and the residual for 77. The mean sum of squares is obtained by dividing the sum of squares by the respective degrees of freedom.

The upper right part shows other summary statistics including the F-statistic and the R-squared. The F-statistic is derived from the ANOVA table as the ratio of the MS(Model) to the MS(Residual), that is, $F = \frac{\text{Model SS}/\text{df}_{\text{Model}}}{\text{Residual SS}/\text{df}_{\text{Residual}}}$. Thus F = 2423914.13 / 20289.8545 = 119.46, with 2 numerator degrees of freedom and 77 denominator degrees of freedom. The F-statistic tests the joint null hypothesis that all the coefficients in the model excluding the constant are zero. The p-value associated with this F-statistic is the chance of observing an F-statistic that much large or larger, and is given as 0. Hence we strongly reject the null hypothesis and conclude that the model as a whole is highly significant.

The same test we also obtain by going to

Statistics → Postestimation → Tests → Test parameters

Or by typing

. testparm F K

The result is

```
. testparm F K

 ( 1)  F = 0
 ( 2)  K = 0

       F(  2,    77) =  119.46
            Prob > F =    0.0000
```

This is exactly the same as the above.

The R-squared (R^2) for the regression model represents the measure of goodness of fit or the coefficient of determination, obtained as the proportion of the model SS in total SS, that is, 4847828.25/ 6410147.05 = 0.7563, indicating that our model with two explanatory variables, *F* and *K*, accounts for (or explain) about 76% of the variation in investment, leaving 24% unexplained. The adjusted R^2 (or R-bar-squared, $\overline{R}^2$) is the R-squared adjusted for degrees of freedom, obtained as

$$\overline{R}^2 = 1 - \frac{MS_{Residual}}{MS_{Total}} = 1 - \frac{\text{Residual SS}/df_{Residual}}{\text{Total SS}/df_{Total}} = 1 - (1 - R^2)\left(\frac{df_{Total}}{df_{Residual}}\right).$$

Thus $\overline{R}^2 = 1 - (1 - 0.7563)\left(\frac{79}{77}\right) = 0.7499$. The root mean squared error, reported below the adjusted R-squared as Root MSE. is the square root of the MS(Residual) in the ANOVA table, and equals $\sqrt{20289.8545} = 142.44$. Note that this is the standard error (SE) of the residual.

Below the summary statistics, we have the table of the estimated coefficients. The first term (I) on the first line of the table gives the dependent variable. The estimates of the marginal effects of F and K and the intercept are given as coefficients (coef) along with the standard error (Std. Err.) and the corresponding t-values (t) and the two-sided significance level (p-value, P > |t|). Note that the t-value is estimated as the ratio of the coefficient value to the corresponding standard error; thus for the coefficient of F, the t-value is 0.1101177/ 0.0137482 = 8.01, which is much greater than 2, as a rule of thumb, and hence the coefficient is highly significant. The zero p-value corresponds to this. To the right of the p-value is reported the 95% confidence interval for the coefficients.

The marginal effects of F and K are positive as expected, and highly significant, with K registering an effect nearly three times higher than that of F. The constant intercept also is significant.

Statistics – Postestimation – Reports and statistics

Unfortunately, we cannot have DW statistic for multiple panels:

```
. estat dwatson
sample may not include multiple panels
r(459);

. estat durbinalt
sample may not include multiple panels
r(459);
```

2. The Fixed Effects Model

We have two models here: (i) Least Squares Dummy Variable model and (ii) Within-groups regression model.

2.1 The Fixed Effects (Least Squares Dummy Variable) Model:

If there is significant cross sectional or significant temporal effect, we cannot assume a constant intercept α for all the companies and years; rather we have to consider the one-way or two-way error components models; if the errors are assumed to be fixed, we have fixed effects model.

$$I_{it} = \beta_1 F_{it\text{-}1} + \beta_2 K_{it\text{-}1} + u_{it}\ ; \quad i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20. \qquad \ldots (5)$$

$u_{it} = \mu_i + \lambda_t + v_{it}$, or $u_{it} = \mu_i + v_{it}$, or, $u_{it} = \lambda_t + v_{it}$.

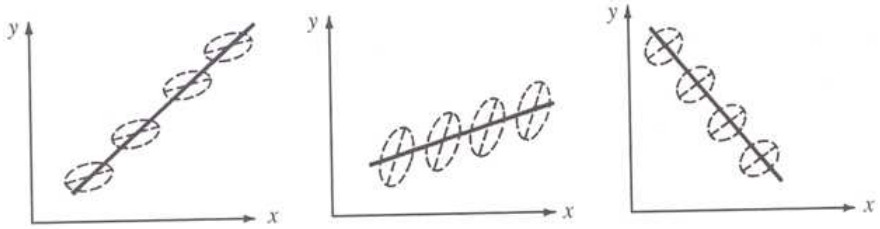
$v_{it} \sim \text{IID}(0, \sigma_v^2)$;

Note that we have not explicitly included the fixed intercept α; it is subsumed under the error components, as will be clear later on.

The model (5) is also called an analysis of covariance (ANCOVA) model (Hsiao 2014:35). The usual regression model assumes that the expected value of investment, *I*, is a function of the exogenous variables, F and K, whereas the traditional Anova gives a general linear model that

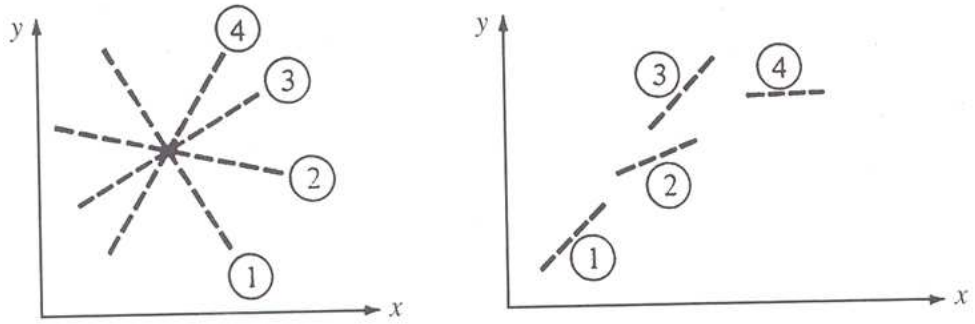
describes every single dependent variable as an equation; for example, $I_{it} = \mu_i + v_{it}$, when we consider only the company heterogeneity. This enables us to test, in the Anova framework, for the mean differences among the companies. One major problem with this (Anova) model is that it is not controlled for the relevant factors, for example, for the differences in F and K, such that the within-group (company) sum of squares will be an overestimate of the stochastic component in *I*, and the differences between company means will reflect not only the company effects but also the effects of differences in the values of the uncontrolled variables in different companies. When we include the covariates (F and K) to the Anova model to account for their effects, we get the Ancova model. Note that this interpretation is obtained when we consider Ancova as a regression within an Anova framework; on the other hand, when we consider Ancova as an Anova in a regression framework, the interpretation is in terms of assessing the marginal effects of the covariates after controlling for the effects of company differences. And this is precisely what we do in model (5). Note that the regression model gives us the marginal effects of quantitative variables, while the Anova model, those of qualitative factors; the Ancova model includes both quantitative and qualitative factors in a framework of controlling their effects.

Now the fixed effects model (5) can be discussed under two assumptions: (1) heterogeneous intercepts ($\mu_i \neq \mu_j$, $\lambda_t \neq \lambda_s$) and homogeneous slope ($\beta_i = \beta_j$; $\beta_t = \beta_s$), and (2) heterogeneous intercepts and slopes ($\mu_i \neq \mu_j$, $\lambda_t \neq \lambda_s$); ($\beta_i \neq \beta_j$; $\beta_t \neq \beta_s$). (Judge *et al*., 1985: Chapter 11, and Hsiao, 1986: Chapter 1). In the former case, cross section and/or time heterogeneity applies only to intercepts, not to slopes; that is, we will have separate intercept for each company and/or for each year, but for all the companies and/or years, the slope will be common; for example, see the following figures, where we consider only the cross-section (company) heterogeneity:



The broken line ellipses in the above graphs represent the scatter plot of data points of each company over time, and the broken straight line in each scatter plot represent individual regression for each company. Note that the company intercepts vary, but the slopes are the same for all the companies ($\mu_i \neq \mu_j$; $\beta_i = \beta_j$). Now if we pool the entire NT data points, the resultant pooled regression is represented by the solid line, with altogether different intercept and slope that highlights the obvious consequence of pooling with biased estimates.

The second case of heterogeneous intercepts and slopes is illustrated below for the four companies, where each company has its own intercept and slope ($\mu_i \neq \mu_j$; $\beta_i \neq \beta_j$).



In general, we can consider the fixed effects panel data models with the following possible assumptions:

1. Slope coefficients constant but intercept varies over companies.
2. Slope coefficients constant but intercept varies over time.
3. Slope coefficients constant but intercept varies over companies and time.
4. All coefficients (intercept and slope) vary over companies.
5. All coefficients (intercept and slope) vary over time.
6. All coefficients (intercept and slope) vary over companies and time.

Last one = Random coefficients model. A random-coefficients model is a panel-data model in which group specific heterogeneity is introduced by assuming that each group has its own parameter vector, which is drawn from a population common to all panels.

Now consider our model:

$$I_{it} = \beta_1 F_{it-1} + \beta_2 K_{it-1} + u_{it} ; \quad i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20. \qquad \ldots (5)$$

$u_{it} = \mu_i + \lambda_t + v_{it}$, or $u_{it} = \mu_i + v_{it}$, or, $u_{it} = \lambda_t + v_{it}$.

$v_{it} \sim \text{IID}(0, \sigma_v^2)$;

(i) Slope coefficients constant but intercept varies over companies.

Our first assumption is: no significant temporal effects, but significant differences among companies. That is, a linear regression model in which the intercept terms vary over individual companies; so our model can be written as a one-way error component model:

$$I_{it} = \beta_1 F_{it-1} + \beta_2 K_{it-1} + u_{it} ;$$

$$u_{it} = \mu_i + v_{it} , \quad v_{it} \sim \text{IID}(0, \sigma_v^2); \quad i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20.$$

Or,

$$I_{it} = \mu_i + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it} ; \quad v_{it} \sim \text{IID}(0, \sigma_v^2); \; i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20 \ldots (6)$$

We also assume that the explanatory variables are independent of the error term.

In regression equation (6), we have for all the four companies separate intercepts, μ_i, which can be estimated by including a dummy variable for each unit i in the model. A dummy variable or an indicator variable is a variable that takes on the values 1 and 0, where 1 means something is true (such as Industry is GE, sex is male, etc.). Thus our model may be written as

$$I_{it} = \Sigma\mu_i D_i + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it} ; \quad v_{it} \sim \text{IID}(0, \sigma_v^2); \; i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20 \ldots (6)$$

Or,

$$I_{it} = \mu_1 D_1 + \mu_2 D_2 + \mu_3 D_3 + \mu_4 D_4 + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it};$$

where $D_1 = 1$ for GE; and zero otherwise.

$D_2 = 1$ for GM; and zero otherwise.

$D_3 = 1$ for US; and zero otherwise.

$D_4 = 1$ for WEST; and zero otherwise.

Note that the model we have started with does not have a constant intercept, and that is why we have included four dummies for the four companies. If the model does have a constant intercept, we need to include only three dummies, lest the model should fall in the 'dummy variable trap' of perfect multicollinearilty. In this case, our model will be

$$I_{it} = \mu + \mu_2 D_2 + \mu_3 D_3 + \mu_4 D_4 + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it};$$

When $D_2 = D_3 = D_4 = 0$, the model becomes

$$I_{it} = \mu + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it}.$$

This is the model for the remaining company, GE. Hence, GE is said to be the 'base company', and μ , the constant intercept, serves as the intercept for GE.

If all the μs are statistically significant, we have differential intercepts, and our model thus accounts for cross section heterogeneity. For example, if μ and μ_2 are significant, the intercept for GM = $\mu + \mu_2$.

An advantage of this model is that all the parameters can be estimated by OLS. Hence this fixed effects model is also called least squares dummy variable (LSDV) Model. It is also known as covariance model, since the explanatory variables are covariates.

Now let us estimate this model in Stata by OLS. Note that we need not xtset our data for OLS estimation. But the LSDV estimation requires dummy variables for the four companies. In Stata this estimation we can do in two ways: one way is to create dummy variables in Stata using the **tabulate** command and the **generate( )** option, and use them directly in the regression command. Remember, we have already created a variable "ind" from the string variable "Industry". Now typing

. tabulate ind, generate(D)

will generate four dummy variables, D1, D2, D3, and D4, corresponding to the four groups in "ind", GE, GM, US and WEST. We can see these dummy variables by typing the command **list** or going to Data → Data Editor → Data Editor (Edit).

Now we can have our OLS result with a constant and the last three dummy variables by typing:

. regress I F K D2 D3 D4

And the result is:

```
. regress I F K D2 D3 D4
```

Source	SS	df	MS	Number of obs =	80
				F(5, 74) =	210.70
Model	5989428.28	5	1197885.66	Prob > F =	0.0000
Residual	420718.767	74	5685.38874	R-squared =	0.9344
				Adj R-squared =	0.9299
Total	6410147.05	79	81141.1019	Root MSE =	75.402

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	.1060964	.0172848	6.14	0.000	.0716557	.140537
K	.347562	.0266309	13.05	0.000	.2944988	.4006252
D2	167.0862	45.86362	3.64	0.000	75.7009	258.4714
D3	339.8296	24.02047	14.15	0.000	291.9677	387.6914
D4	184.6554	31.39971	5.88	0.000	122.0901	247.2207
_cons	-242.758	35.52478	-6.83	0.000	-313.5426	-171.9733

The same result we can have without using the dummy variables directly; this second method is to use what Stata calls "factor variables", a kind of "virtual variables". With reference to the variable ind, the notation i.ind tells Stata that ind is a categorical variable rather than continuous and Stata, in effect, creates dummy variables coded 0/1 from this categorical variable. Note that

our ind variables is coded as GE = 1, GM = 2, US = 3 and WEST = 4. Then i.ind would cause Stata to create three 0/1 dummies. By default, the first category (in this case GE) is the reference (base) category, but we can change that, e.g. ib2.ind would make GM the reference category, or ib(last).ind would make the last category, WEST, as the base.

Now typing the following

. reg I F K i.ind

We get the same result as above.

```
. reg I F K i.ind
```

Source	SS	df	MS		
Model	5989428.28	5	1197885.66	Number of obs	= 80
Residual	420718.767	74	5685.38874	F(5, 74)	= 210.70
				Prob > F	= 0.0000
				R-squared	= 0.9344
				Adj R-squared	= 0.9299
Total	6410147.05	79	81141.1019	Root MSE	= 75.402

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	.1060964	.0172848	6.14	0.000	.0716557	.140537
K	.347562	.0266309	13.05	0.000	.2944988	.4006252
ind						
GM	167.0862	45.86362	3.64	0.000	75.7009	258.4714
US	339.8296	24.02047	14.15	0.000	291.9677	387.6914
WEST	184.6554	31.39971	5.88	0.000	122.0901	247.2207
_cons	-242.758	35.52478	-6.83	0.000	-313.5426	-171.9733

The results on the marginal effects of F and K are similar to those from the pooled regression above; both the coefficients are positive and highly significant, with a very marginal fall in respect of F and a marginal increase in respect of K; now K has an effect a little more than three times higher than that of F.

The cross section (company) heterogeneity also is highly significant. Thus every company has its own significant intercept. The intercept for the base company, GE, is given by the constant intercept of the model, that is, – 242.758. And the intercepts of other companies are:

For GM = – 75.672 (= – 242.758 + 167.0862)

For US = 97.072 (= – 242.758 + 339.8296), and

For WEST = – 58.103 (= – 242.758 + 184.6554)

Poolability Test (between Pooled Regression and FE Model)

Compared with our old pooled regression model, the new LSDV fixed effects model has a higher R^2 value. Hence the question comes up: Which model is better? The pooled regression with constant slope and constant intercept or the LSDV fixed effects model with constant slope and variable intercept for companies? The question can be reframed also as: Can we assume that there is neither significant cross sectional nor significant temporal effect, and pool the data and run an OLS regression model with an intercept α and slope coefficients βs constant across companies and time? This is the poolability test.

Note that compared with the second (FE) model, the first one (pooled regression) is a restricted model; it imposes a common intercept on all companies: $\mu_2 = \mu_3 = \mu_4 = \mu$. Hence we have to do the restricted F test given by

$$F = \frac{(R^2_{UR} - R^2_R)/J}{(1-R^2_{UR})/(n-k)}$$

where $R_{UR}^2 = R^2$ of the unrestricted regression (second model) = 0.9344;

$R_R^2 = R^2$ of the restricted regression (first model) = 0.7563;

J = number of linear restrictions on the first model = 3;

k = number of parameters in the unrestricted regression = 6; and

n = NT = number of observations = 80.

Hence $F = \frac{(0.9344-0.7563)/3}{(1-0.9344)/74} = 66.968$, with a p-value equal to zero. Comparing this with $F_{3,74}$ = 4.05787 at 1% right tail significance level, we find that the difference in the explanatory powers of the two models is highly significant and so conclude that the restricted regression (pooled regression) is invalid.

This poolability test we can do in Stata after the regression with the factor variable i.ind, by typing

. testparm i.ind

And the result is:

```
. testparm i.ind

 ( 1)  2.ind = 0
 ( 2)  3.ind = 0
 ( 3)  4.ind = 0

       F(  3,    74) =   66.93
            Prob > F =    0.0000
```

With this p-vale, we strongly reject the three null hypotheses of zero company effect.

Random Coefficient models (Another Poolability test)

In random-coefficients models, we wish to treat the parameter vector as a realization (in each panel) of a stochastic process. The Stata command xtrc fits the Swamy (1970) random-coefficients model, which is suitable for linear regression of panel data.

To take a first look at the assumption of parameter constancy, we go to

Statistics > Longitudinal/panel data > Random-coefficients regression by GLS

Or typing

. xtrc I F K

```
. xtrc I F K

Random-coefficients regression                  Number of obs      =        80
Group variable: Time                            Number of groups   =        20

                                                Obs per group: min =         4
                                                               avg =       4.0
                                                               max =         4

                                                Wald chi2(2)       =      2.52
                                                Prob > chi2        =    0.2837

------------------------------------------------------------------------------
           I |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           F |   .1551084    .156705     0.99   0.322    -.1520278    .4622446
           K |   -.147248   .4339527    -0.34   0.734    -.9977797    .7032838
       _cons |   10.63852   99.10235     0.11   0.915    -183.5985    204.8756
------------------------------------------------------------------------------
Test of parameter constancy:    chi2(57) =   165.54       Prob > chi2 = 0.0000
```

The test included with the random-coefficients model also indicates that the assumption of parameter constancy is not valid for these data.

(ii) Slope coefficients constant but intercept varies over time.

Our second assumption is: no significant cross section differences, but significant temporal effects. That is, a linear regression model in which the intercept terms vary over time; so our model can be written as a one-way error component model:

$$I_{it} = \beta_1 F_{it-1} + \beta_2 K_{it-1} + u_{it} ;$$

$$u_{it} = \lambda_t + v_{it} , \quad v_{it} \sim \text{IID}(0, \sigma_v^2); \quad i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20.$$

Or,

$$I_{it} = \lambda_t + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it} ; \quad v_{it} \sim \text{IID}(0, \sigma_v^2); i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20 \ldots (7)$$

We also assume that the explanatory variables are independent of the error term.

In regression equation (7), we have for all the 20 years separate intercepts, λ_t, which can be estimated by including a dummy variable for each year t in the model. Thus our model may be written as

$$I_{it} = \Sigma\lambda_t d_t + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it} ; \quad v_{it} \sim \text{IID}(0, \sigma_v^2); i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20 \ldots (6)$$

Or,

$$I_{it} = \lambda_1 d_1 + \lambda_2 d_2 + \ldots. + \lambda_{19} d_{19} + \lambda_{20} d_{20} + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it};$$

where $d_1 = 1$ for year 1935; and zero otherwise, etc. up to $d_{20} = 1$ for year 1954; and zero otherwise.

If the model is assumed to have a constant intercept, we need to include 19 time dummies, and our model will be

$$I_{it} = \lambda + \lambda_2 d_2 + \ldots. + \lambda_{19} d_{19} + \lambda_{20} d_{20} + \beta_1 F_{it-1} + \beta_2 K_{it-1} + v_{it};$$

Here the 'base year' is 1935, and λ , the constant intercept, serves as the intercept for that year.

If all the λs are statistically significant, we have differential intercepts, and our model thus accounts for temporal heterogeneity. For example, if λ and λ_2 are significant, the intercept for 1936 = $\lambda + \lambda_2$.

An advantage of this model is that all the parameters can be estimated by OLS. Hence this fixed effects model is also called least squares dummy variable (LSDV) Model. It is also known as covariance model, since the explanatory variables are covariates.

Now we turn to estimating this model in Stata by OLS. First we create time dummy variables in Stata using the **tabulate** command and the **generate( )** option, as before:

. tabulate Time, generate(d)

This will generate 20 dummy variables, *d*1, *d*2, …, *d*19, and *d*20, corresponding to the 20 years from 1935 to 1954. We can see these dummy variables by typing the command **list** or going to Data → Data Editor → Data Editor (Edit).

Now we can have our OLS result with a constant and the last 19 dummy variables by typing:

. regress I F K d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20

And the result is:

```
. regress I F K d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20
```

Source	SS	df	MS		
Model	4938989.32	21	235189.968	Number of obs	= 80
Residual	1471157.73	58	25364.7885	F(21, 58)	= 9.27
				Prob > F	= 0.0000
Total	6410147.05	79	81141.1019	R-squared	= 0.7705
				Adj R-squared	= 0.6874
				Root MSE	= 159.26

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	.1160908	.0181922	6.38	0.000	.0796752	.1525065
K	.2707146	.0832103	3.25	0.002	.104151	.4372782
d2	-35.21155	113.46	-0.31	0.757	-262.3264	191.9033
d3	-79.75487	114.96	-0.69	0.491	-309.8722	150.3625
d4	-69.87972	112.8546	-0.62	0.538	-295.7827	156.0233
d5	-118.1149	113.1231	-1.04	0.301	-344.5553	108.3256
d6	-57.43471	113.3091	-0.51	0.614	-284.2474	169.378
d7	-.1731213	113.2113	-0.00	0.999	-226.7902	226.444
d8	8.988479	113.2964	0.08	0.937	-217.7988	235.7758
d9	-34.1216	113.2734	-0.30	0.764	-260.863	192.6198
d10	-39.16445	113.1673	-0.35	0.731	-265.6934	187.3645
d11	-60.34621	113.306	-0.53	0.596	-287.1529	166.4605
d12	5.463983	113.5785	0.05	0.962	-221.888	232.816
d13	14.16819	115.5051	0.12	0.903	-217.0404	245.3768
d14	4.51502	117.1907	0.04	0.969	-230.0676	239.0977
d15	-50.26966	118.3731	-0.42	0.673	-287.2191	186.6798
d16	-42.05006	118.8364	-0.35	0.725	-279.9269	195.8268
d17	-20.78958	118.1475	-0.18	0.861	-257.2876	215.7084
d18	.0692122	120.4859	0.00	1.000	-241.1095	241.2479
d19	17.29588	123.9711	0.14	0.890	-230.8593	265.4511
d20	-22.29242	129.7075	-0.17	0.864	-281.9302	237.3453
_cons	-35.60765	83.24187	-0.43	0.670	-202.2344	131.0191

Now the same result we get using the factor variable i.Time, by typing the following:

. reg I F K i.Time

```
. reg I F K i.Time
```

Source	SS	df	MS	Number of obs =	80
				F(21, 58) =	9.27
Model	4938989.32	21	235189.968	Prob > F =	0.0000
Residual	1471157.73	58	25364.7885	R-squared =	0.7705
				Adj R-squared =	0.6874
Total	6410147.05	79	81141.1019	Root MSE =	159.26

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	.1160908	.0181922	6.38	0.000	.0796752	.1525065
K	.2707146	.0832103	3.25	0.002	.104151	.4372782
Time						
1936	-35.21155	113.46	-0.31	0.757	-262.3264	191.9033
1937	-79.75487	114.96	-0.69	0.491	-309.8722	150.3625
1938	-69.87972	112.8546	-0.62	0.538	-295.7827	156.0233
1939	-118.1149	113.1231	-1.04	0.301	-344.5553	108.3256
1940	-57.43471	113.3091	-0.51	0.614	-284.2474	169.378
1941	-.1731213	113.2113	-0.00	0.999	-226.7902	226.444
1942	8.988479	113.2964	0.08	0.937	-217.7988	235.7758
1943	-34.1216	113.2734	-0.30	0.764	-260.863	192.6198
1944	-39.16445	113.1673	-0.35	0.731	-265.6934	187.3645
1945	-60.34621	113.306	-0.53	0.596	-287.1529	166.4605
1946	5.463983	113.5785	0.05	0.962	-221.888	232.816
1947	14.16819	115.5051	0.12	0.903	-217.0404	245.3768
1948	4.51502	117.1907	0.04	0.969	-230.0676	239.0977
1949	-50.26966	118.3731	-0.42	0.673	-287.2191	186.6798
1950	-42.05006	118.8364	-0.35	0.725	-279.9269	195.8268
1951	-20.78958	118.1475	-0.18	0.861	-257.2876	215.7084
1952	.0692122	120.4859	0.00	1.000	-241.1095	241.2479
1953	17.29588	123.9711	0.14	0.890	-230.8593	265.4511
1954	-22.29242	129.7075	-0.17	0.864	-281.9302	237.3453
_cons	-35.60765	83.24187	-0.43	0.670	-202.2344	131.0191

The marginal effects are positive and significant, with marginal differences compared with the other models. We also have an interesting result here; all the time dummies are insignificant, indicating that the investment function has not changed much over time, and the R^2 is only 0.7705, irrespective of a large number of variables. Now comparing this LSDV time effect model with the pooled regression with $R^2 = 0.7563$, which one is better? With the increment in R^2 equal to only 0.0142, the F test does not reject; we had better pool the data and run an OLS model with constant intercept.

Now Stata gives the poolability test result after the regression with the factor variable i.Time:

```
. testparm i.Time

 ( 1)  1936.Time = 0
 ( 2)  1937.Time = 0
 ( 3)  1938.Time = 0
 ( 4)  1939.Time = 0
 ( 5)  1940.Time = 0
 ( 6)  1941.Time = 0
 ( 7)  1942.Time = 0
 ( 8)  1943.Time = 0
 ( 9)  1944.Time = 0
 (10)  1945.Time = 0
 (11)  1946.Time = 0
 (12)  1947.Time = 0
 (13)  1948.Time = 0
 (14)  1949.Time = 0
 (15)  1950.Time = 0
 (16)  1951.Time = 0
 (17)  1952.Time = 0
 (18)  1953.Time = 0
 (19)  1954.Time = 0

       F( 19,    58) =    0.19
            Prob > F =    0.9999
```

With this p-value, we cannot reject the F-test null of zero time effect.

Thus we have found that the company effects are statistically significant, but the time effects not. Does that mean our model is somehow misspecified? Let us now consider both company and time effects together.

(iii) Slope coefficients constant but intercept varies over companies and time.

This gives our two-way error components model:

$$I_{it} = \beta_1 F_{it\text{-}1} + \beta_2 K_{it\text{-}1} + u_{it}\ ; \quad i = 1, 2, 3, 4; \quad t = 1, 2, \ldots, 20. \qquad \ldots (5)$$

$$u_{it} = \mu_i + \lambda_t + v_{it}\ ,$$

$$v_{it} \sim \text{IID}(0,\ \sigma_v^2).$$

We also assume that the explanatory variables are independent of the error term.

With a constant intercept, our LSDV model is

$$I_{it} = \alpha + \mu_2 D_2 + \mu_3 D_3 + \mu_4 D_4 + \lambda_2 d_2 + \ldots. + \lambda_{20} d_{20} + \beta_1 F_{it\text{-}1} + \beta_2 K_{it\text{-}1} + v_{it}, \qquad \ldots(7)$$

with the same definitions for the dummy variables as above.

The constant intercept α, if significant, denotes the base company, GE, for the base year, 1935; if α and λ_2 are significant, then $\alpha + \lambda_2$ gives the intercept for GE for the year 1936, and so on. The Stata output for this model is:

```
. regress I F K D2 D3 D4 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20
```

Source	SS	df	MS		
Model	6082815.02	24	253450.626	Number of obs	= 80
Residual	327332.029	55	5951.49144	F(24, 55)	= 42.59
				Prob > F	= 0.0000
				R-squared	= 0.9489
				Adj R-squared	= 0.9267
Total	6410147.05	79	81141.1019	Root MSE	= 77.146

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	.1267775	.0268777	4.72	0.000	.0729134	.1806416
K	.3691081	.0415325	8.89	0.000	.2858752	.452341
D2	112.5462	66.24191	1.70	0.095	-20.20559	245.2979
D3	341.3641	24.80427	13.76	0.000	291.6553	391.073
D4	217.6964	40.74364	5.34	0.000	136.0443	299.3485
d2	-45.03609	58.57156	-0.77	0.445	-162.4161	72.34394
d3	-101.23	66.6013	-1.52	0.134	-234.702	32.24197
d4	-85.58827	55.31736	-1.55	0.128	-196.4467	25.27019
d5	-140.4574	58.37524	-2.41	0.020	-257.444	-23.47082
d6	-80.34846	59.78844	-1.34	0.185	-200.1672	39.47025
d7	-24.71323	58.97732	-0.42	0.677	-142.9064	93.47996
d8	-15.26153	55.89531	-0.27	0.786	-127.2782	96.75518
d9	-61.09746	57.12428	-1.07	0.289	-175.5771	53.38216
d10	-64.51794	57.2712	-1.13	0.265	-179.292	50.25611
d11	-88.12239	59.30216	-1.49	0.143	-206.9666	30.72179
d12	-27.97868	60.63872	-0.46	0.646	-149.5014	93.54404
d13	-25.17069	56.63821	-0.44	0.658	-138.6762	88.33481
d14	-42.02802	57.32754	-0.73	0.467	-156.915	72.85893
d15	-103.0746	58.10986	-1.77	0.082	-219.5294	13.38014
d16	-98.34156	58.62943	-1.68	0.099	-215.8376	19.15445
d17	-85.09251	61.73637	-1.38	0.174	-208.815	38.62993
d18	-74.93115	63.57047	-1.18	0.244	-202.3292	52.46693
d19	-78.02982	70.10255	-1.11	0.271	-218.5185	62.45882
d20	-132.4467	71.84455	-1.84	0.071	-276.4264	11.53299
_cons	-222.8553	51.3875	-4.34	0.000	-325.8382	-119.8725

The same output we get by typing

. reg I F K i.ind i.Time

```
. reg I F K i.ind i.Time
```

Source	SS	df	MS	Number of obs =	80
				F(24, 55) =	42.59
Model	6082815.02	24	253450.626	Prob > F =	0.0000
Residual	327332.029	55	5951.49144	R-squared =	0.9489
				Adj R-squared =	0.9267
Total	6410147.05	79	81141.1019	Root MSE =	77.146

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	.1267775	.0268777	4.72	0.000	.0729134	.1806416
K	.3691081	.0415325	8.89	0.000	.2858752	.452341
ind						
GM	112.5462	66.24191	1.70	0.095	-20.20559	245.2979
US	341.3641	24.80427	13.76	0.000	291.6553	391.073
WEST	217.6964	40.74364	5.34	0.000	136.0443	299.3485
Time						
1936	-45.03609	58.57156	-0.77	0.445	-162.4161	72.34394
1937	-101.23	66.6013	-1.52	0.134	-234.702	32.24197
1938	-85.58827	55.31736	-1.55	0.128	-196.4467	25.27019
1939	-140.4574	58.37524	-2.41	0.020	-257.444	-23.47082
1940	-80.34846	59.78844	-1.34	0.185	-200.1672	39.47025
1941	-24.71323	58.97732	-0.42	0.677	-142.9064	93.47996
1942	-15.26153	55.89531	-0.27	0.786	-127.2782	96.75518
1943	-61.09746	57.12428	-1.07	0.289	-175.5771	53.38216
1944	-64.51794	57.2712	-1.13	0.265	-179.292	50.25611
1945	-88.12239	59.30216	-1.49	0.143	-206.9666	30.72179
1946	-27.97868	60.63872	-0.46	0.646	-149.5014	93.54404
1947	-25.17069	56.63821	-0.44	0.658	-138.6762	88.33481
1948	-42.02802	57.32754	-0.73	0.467	-156.915	72.85893
1949	-103.0746	58.10986	-1.77	0.082	-219.5294	13.38014
1950	-98.34156	58.62943	-1.68	0.099	-215.8376	19.15445
1951	-85.09251	61.73637	-1.38	0.174	-208.815	38.62993
1952	-74.93115	63.57047	-1.18	0.244	-202.3292	52.46693
1953	-78.02982	70.10255	-1.11	0.271	-218.5185	62.45882
1954	-132.4467	71.84455	-1.84	0.071	-276.4264	11.53299
_cons	-222.8553	51.3875	-4.34	0.000	-325.8382	-119.8725

Time						
1936	-45.03609	58.57156	-0.77	0.445	-162.4161	72.34394
1937	-101.23	66.6013	-1.52	0.134	-234.702	32.24197
1938	-85.58827	55.31736	-1.55	0.128	-196.4467	25.27019
1939	-140.4574	58.37524	-2.41	0.020	-257.444	-23.47082
1940	-80.34846	59.78844	-1.34	0.185	-200.1672	39.47025
1941	-24.71323	58.97732	-0.42	0.677	-142.9064	93.47996
1942	-15.26153	55.89531	-0.27	0.786	-127.2782	96.75518
1943	-61.09746	57.12428	-1.07	0.289	-175.5771	53.38216
1944	-64.51794	57.2712	-1.13	0.265	-179.292	50.25611
1945	-88.12239	59.30216	-1.49	0.143	-206.9666	30.72179
1946	-27.97868	60.63872	-0.46	0.646	-149.5014	93.54404
1947	-25.17069	56.63821	-0.44	0.658	-138.6762	88.33481
1948	-42.02802	57.32754	-0.73	0.467	-156.915	72.85893
1949	-103.0746	58.10986	-1.77	0.082	-219.5294	13.38014
1950	-98.34156	58.62943	-1.68	0.099	-215.8376	19.15445
1951	-85.09251	61.73637	-1.38	0.174	-208.815	38.62993
1952	-74.93115	63.57047	-1.18	0.244	-202.3292	52.46693
1953	-78.02982	70.10255	-1.11	0.271	-218.5185	62.45882
1954	-132.4467	71.84455	-1.84	0.071	-276.4264	11.53299
_cons	-222.8553	51.3875	-4.34	0.000	-325.8382	-119.8725

We have a little mixed results here; the dummy variable D_2 associated with GM is significant only at 10% level, and a few time dummies are significant at 5% or 10% level. The covariates and other two company dummies are highly significant. And the R^2 value is higher at 0.9489. Compared with the pooled regression (with $R^2 = 0.7563$), the F-test rejects in favour of our new LSDV model, but against out first LSDV model (with differential intercepts for companies, having $R^2 = 0.9344$), this model fails the test with an increment of only 0.0145, indicating that the time effect is insignificant in general. We conclude that the investment function has not changed much over time, but changed over companies.

In this case in Stata, we can do the poolability test in three ways. First we test the null of zero cross section *and* temporal effects:

```
. testparm i.ind i.Time

 ( 1)  2.ind = 0
 ( 2)  3.ind = 0
 ( 3)  4.ind = 0
 ( 4)  1936.Time = 0
 ( 5)  1937.Time = 0
 ( 6)  1938.Time = 0
 ( 7)  1939.Time = 0
 ( 8)  1940.Time = 0
 ( 9)  1941.Time = 0
 (10)  1942.Time = 0
 (11)  1943.Time = 0
 (12)  1944.Time = 0
 (13)  1945.Time = 0
 (14)  1946.Time = 0
 (15)  1947.Time = 0
 (16)  1948.Time = 0
 (17)  1949.Time = 0
 (18)  1950.Time = 0
 (19)  1951.Time = 0
 (20)  1952.Time = 0
 (21)  1953.Time = 0
 (22)  1954.Time = 0

       F( 22,     55) =    9.43
            Prob > F =    0.0000
```

We reject the null: the intercepts are different across the companies *and* time in general.

Next we do the F-test only for the temporal effects:

```
. testparm i.Time

 ( 1)  1936.Time = 0
 ( 2)  1937.Time = 0
 ( 3)  1938.Time = 0
 ( 4)  1939.Time = 0
 ( 5)  1940.Time = 0
 ( 6)  1941.Time = 0
 ( 7)  1942.Time = 0
 ( 8)  1943.Time = 0
 ( 9)  1944.Time = 0
 (10)  1945.Time = 0
 (11)  1946.Time = 0
 (12)  1947.Time = 0
 (13)  1948.Time = 0
 (14)  1949.Time = 0
 (15)  1950.Time = 0
 (16)  1951.Time = 0
 (17)  1952.Time = 0
 (18)  1953.Time = 0
 (19)  1954.Time = 0

       F( 19,    55) =    0.83
            Prob > F =    0.6683
```

Here we cannot the reject the null of zero time effects!

Then we do the F-test only for the company effects:

```
. testparm i.ind

 ( 1)  2.ind = 0
 ( 2)  3.ind = 0
 ( 3)  4.ind = 0

       F(  3,    55) =   64.06
            Prob > F =    0.0000
```

We do reject the null: the company effects are significant.

(iv) All coefficients (intercept and slope) vary over companies.

Our next model assumes that all the slope coefficients as well as the intercept are variable over companies; this means that all the four companies, GE, GM, US and WEST, have altogether different investment functions. This assumption can be incorporated in our LSDV model by assigning one more role to the company dummies. In the earlier LSDV models, these three company dummies were included along with the constant intercept in an additive way to account for intercept differences. Now to account for slope differences, these three company dummies

have to be included in the LSDV model in an interactive/multiplicative way, by multiplying each of the company dummies by each of the explanatory variables. Thus our extended LSDV model is:

$$I_{it} = \mu + \mu_2 D_2 + \mu_3 D_3 + \mu_4 D_4 + \beta_1 F_{it-1} + \beta_2 K_{it-1} + \gamma_1 (D_2 F_{it-1}) + \gamma_2 (D_2 K_{it-1})$$

$$+ \gamma_3 (D_3 F_{it-1}) + \gamma_4 (D_3 K_{it-1}) + \gamma_5 (D_4 F_{it-1}) + \gamma_6 (D_4 K_{it-1}) + v_{it}, \qquad \ldots(8)$$

where the μs represent differential intercepts, and the βs and γs together give differential slope coefficients. The base company, as before, is GE, with a differential intercept of μ; β_1 is the slope coefficient of F_{it-1} of the base company GE. If β_1 and γ_1 are statistically significant, then the slope coefficient of F_{it-1} of GM is given by $(\beta_1 + \gamma_1)$, which is different from that of GE.

It is very difficult to specify the regression equation command using so many dummy variables in additive and multiplicative ways. Stata has certain easy ways to deal with this problem, using the factor variables and the cross operator #; the latter is used for interactions and product terms. However, note that when we use the cross operator along with the i. prefix with variables, Stata by default assumes that the variables on both the sides of the # operator are categorical and computes interaction terms accordingly. Hence we must use the i. prefix only with categorical variables. When we have a categorical variable (ind) along with a continuous variable (F or K), we must use the i. prefix with the categorical variable (i.ind) and c. prefix with the continuous variable (c.F or c.K). Thus the simple command i.ind#c.F or i.ind#c.K will give us an indication of the slope differential over the companies. Also note that c.F#c.F tells Stata to include the squared term of F (F^2) in the model; we need not compute the variable separately.

Now the above model we estimate in Stata by typing

```
. reg I F K i.ind i.ind#c.F i.ind#c.K
```

And the result is:

```
. reg I F K i.ind i.ind#c.F i.ind#c.K
```

Source	SS	df	MS	Number of obs =	80
Model	6095228.45	11	554111.677	F(11, 68) =	119.65
Residual	314918.6	68	4631.15588	Prob > F =	0.0000
				R-squared =	0.9509
				Adj R-squared =	0.9429
Total	6410147.05	79	81141.1019	Root MSE =	68.053

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	.0265512	.0379918	0.70	0.487	-.0492602	.1023626
K	.1516939	.0627352	2.42	0.018	.0265076	.2768801
ind						
GM	-126.1637	108.3931	-1.16	0.249	-342.4589	90.1314
US	-40.12174	129.6108	-0.31	0.758	-298.7561	218.5127
WEST	9.375904	93.3885	0.10	0.920	-176.9779	195.7298
ind#c.F						
GM	.0895841	.0423627	2.11	0.038	.0050508	.1741175
US	.1448793	.0648388	2.23	0.029	.0154955	.274263
WEST	.0265042	.111464	0.24	0.813	-.1959187	.2489271
ind#c.K						
GM	.2222108	.068435	3.25	0.002	.0856508	.3587707
US	.2570148	.1208285	2.13	0.037	.0159054	.4981243
WEST	-.0600001	.3797005	-0.16	0.875	-.8176807	.6976805
_cons	-9.956306	76.57426	-0.13	0.897	-162.7579	142.8453

We have the following results of F-tests:

```
. testparm i.ind

 ( 1)  2.ind = 0
 ( 2)  3.ind = 0
 ( 3)  4.ind = 0

       F(  3,    68) =    0.76
            Prob > F =    0.5217
```

```
. testparm c.F c.K

 ( 1)  F = 0
 ( 2)  K = 0

       F(  2,    68) =    3.42
            Prob > F =    0.0386
```

```
. testparm i.ind#c.F

 ( 1)  2.ind#c.F = 0
 ( 2)  3.ind#c.F = 0
 ( 3)  4.ind#c.F = 0

       F(  3,    68) =    2.15
            Prob > F =    0.1016
```

```
. testparm i.ind#c.K

 ( 1)  2.ind#c.K = 0
 ( 2)  3.ind#c.K = 0
 ( 3)  4.ind#c.K = 0

       F(  3,    68) =    3.81
            Prob > F =    0.0139
```

We have some mixed results here. Investment is significantly related only to K (by the individual t-test), even though the F-test rejects the null of zero coefficients for both F and K in general. Thus the slope coefficient of the base company (GE) is significant only in respect of K. The slope differentials (γs) in respect of both F and K are significant only for GM and US, not for WEST, even though the F-test rejects the joint null (for all the three companies) strongly in respect of K and at a little more than 10% for F. Also note that all the company intercepts, including the constant, are insignificant.

(v) All coefficients (intercept and slope) vary over time.

This model assumes that all the slope coefficients as well as the intercept are variable over time; this means that all 20 years, from 1935 to 1954, have altogether different investment functions. This assumption is incorporated in our LSDV model by including the time dummies in both additive and interactive/multiplicative way. Thus our extended LSDV model is:

$$I_{it} = \lambda + \lambda_2 d_2 + \lambda_3 d_3 + \ldots + \lambda_{20} d_{20} + \beta_1 F_{it-1} + \beta_2 K_{it-1} + \gamma_1 (d_2 F_{it-1}) + \gamma_2 (d_2 K_{it-1})$$

$$+ \gamma_3 (d_3 F_{it-1}) + \gamma_4 (d_3 K_{it-1}) + \ldots + \gamma_{37} (d_{20} F_{it-1}) + \gamma_{38} (d_{20} K_{it-1}) + v_{it}, \qquad \ldots(9)$$

where the μs represent differential intercepts, and the βs and γs together give differential slope coefficients. The base year, as before, is 1935, with a differential intercept of λ; β_1 is the slope coefficient of F_{it-1} of the base year, 1935. If β_1 and γ_1 are statistically significant, then the slope coefficient of F_{it-1} of 1936 is given by $(\beta_1 + \gamma_1)$, which is different from that of the base year.

The Stata results using the indicator variables and cross operator are obtained by typing

. reg I F K i.Time i.Time#c.F i.Time#c.K

And the result is given below and is left for your own interpretation:

```
. reg I F K i.Time i.Time#c.F i.Time#c.K
```

Source	SS	df	MS	Number of obs =	80
				F(59, 20) =	3.36
Model	5821984.52	59	98677.7036	Prob > F =	0.0020
Residual	588162.534	20	29408.1267	R-squared =	0.9082
				Adj R-squared =	0.6376
Total	6410147.05	79	81141.1019	Root MSE =	171.49

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	.1052894	.0838614	1.26	0.224	-.0696424	.2802212
K	-.5980396	2.184176	-0.27	0.787	-5.15415	3.958071
Time						
1936	48.39522	252.1546	0.19	0.850	-477.59	574.3805
1937	-8.622768	254.4624	-0.03	0.973	-539.4221	522.1765
1938	-39.12789	278.3113	-0.14	0.890	-619.6751	541.4193
1939	-75.21356	262.1928	-0.29	0.777	-622.1383	471.7111
1940	-94.21622	266.5826	-0.35	0.727	-650.2977	461.8653
1941	-64.74314	270.1124	-0.24	0.813	-628.1877	498.7014
1942	-68.3191	283.2693	-0.24	0.812	-659.2085	522.5703
1943	-49.67035	301.8949	-0.16	0.871	-679.4121	580.0714
1944	-26.89948	311.656	-0.09	0.932	-677.0025	623.2036
1945	9.898586	303.6664	0.03	0.974	-623.5385	643.3356
1946	24.89088	279.8911	0.09	0.930	-558.9518	608.7336
1947	-115.8048	263.1106	-0.44	0.665	-664.644	433.0343
1948	-123.5759	269.1321	-0.46	0.651	-684.9757	437.8239
1949	-22.00557	245.9603	-0.09	0.930	-535.0698	491.0587
1950	-106.7305	248.2333	-0.43	0.672	-624.536	411.075
1951	-63.44114	244.111	-0.26	0.798	-572.6478	445.7655
1952	-201.0213	258.3286	-0.78	0.446	-739.8853	337.8427
1953	-135.7339	241.5208	-0.56	0.580	-639.5374	368.0696
1954	1763.298	1085.33	1.62	0.120	-500.66	4027.256

Time#c.F						
1936	-.0141468	.1036556	-0.14	0.893	-.2303687	.202075
1937	-.0919587	.1510385	-0.61	0.549	-.4070195	.223102
1938	-.1063311	.1879293	-0.57	0.578	-.4983447	.2856826
1939	-.0438613	.1112838	-0.39	0.698	-.2759952	.1882725
1940	-.0241785	.1143537	-0.21	0.835	-.2627161	.2143591
1941	-.0162411	.1245705	-0.13	0.898	-.2760907	.2436085
1942	.0799395	.1761323	0.45	0.655	-.2874661	.4473452
1943	.0382249	.1166844	0.33	0.747	-.2051746	.2816244
1944	.0370799	.1050209	0.35	0.728	-.1819899	.2561497
1945	.0395751	.1083095	0.37	0.719	-.1863545	.2655047
1946	.102726	.1419185	0.72	0.478	-.1933108	.3987628
1947	.3471329	.2824935	1.23	0.233	-.2421382	.9364041
1948	.5026666	.3293273	1.53	0.143	-.1842982	1.189631
1949	.2439426	.2186174	1.12	0.278	-.2120853	.6999705
1950	.3676762	.2549471	1.44	0.165	-.1641341	.8994866
1951	.268405	.1677148	1.60	0.125	-.081442	.6182521
1952	.5355867	.2624107	2.04	0.055	-.0117925	1.082966
1953	.164237	.263211	0.62	0.540	-.3848115	.7132854
1954	-2.792771	1.472943	-1.90	0.072	-5.865276	.2797343
Time#c.K						
1936	-.611121	3.323083	-0.18	0.856	-7.542951	6.320709
1937	2.637883	4.318102	0.61	0.548	-6.369521	11.64529
1938	1.678583	2.647134	0.63	0.533	-3.843242	7.200409
1939	1.043602	2.385403	0.44	0.666	-3.932261	6.019466
1940	1.291015	2.545085	0.51	0.618	-4.01794	6.599971
1941	1.271331	2.621369	0.48	0.633	-4.196748	6.73941
1942	.4355624	2.632311	0.17	0.870	-5.055342	5.926466
1943	.4908004	2.434874	0.20	0.842	-4.588257	5.569858
1944	.3414439	2.400829	0.14	0.888	-4.666598	5.349486
1945	.0258404	2.509344	0.01	0.992	-5.20856	5.26024
1946	-.2462594	2.607919	-0.09	0.926	-5.686282	5.193763
1947	-.4681653	2.426983	-0.19	0.849	-5.530764	4.594434
1948	-.8109853	2.401271	-0.34	0.739	-5.819948	4.197978
1949	-.0849694	2.294793	-0.04	0.971	-4.871824	4.701885
1950	-.3271907	2.312885	-0.14	0.889	-5.151785	4.497404
1951	-.2036824	2.249207	-0.09	0.929	-4.895446	4.488081
1952	-.8268425	2.31822	-0.36	0.725	-5.662566	4.008881
1953	.4532783	2.336016	0.19	0.848	-4.419565	5.326122
1954	7.207045	3.898233	1.85	0.079	-.9245257	15.33862
_cons	13.9873	182.1079	0.08	0.940	-365.8832	393.8577

(vi) All coefficients (intercept and slope) vary over companies and time.

Our final model assumes that all the (intercept and slope) coefficients vary over both companies and time. The model with the following specification really looks very formidable:

$$I_{it} = \alpha + \mu_2 D_2 + \mu_3 D_3 + \mu_4 D_4 + \beta_1 F_{it-1} + \beta_2 K_{it-1} + \gamma_1 (D_2 F_{it-1}) + \gamma_2 (D_2 K_{it-1}) +$$

$$\gamma_3 (D_3 F_{it-1}) + \gamma_4 (D_3 K_{it-1}) + \gamma_5 (D_4 F_{it-1}) + \gamma_6 (D_4 K_{it-1}) +$$

$$\lambda_2 d_2 + \lambda_3 d_3 + \ldots + \lambda_{20} d_{20} + \beta_1 F_{it-1} + \beta_2 K_{it-1} + \delta_1 (d_2 F_{it-1}) + \delta_2 (d_2 K_{it-1}) +$$

$$\delta_3 (d_3 F_{it-1}) + \delta_4 (d_3 K_{it-1}) + \ldots + \delta_{37} (d_{20} F_{it-1}) + \delta_{38} (d_{20} K_{it-1}) + v_{it}, \qquad \ldots(10)$$

This model is estimated by typing

. reg I F K i.ind i.Time i.ind#c.F i.ind#c.K i.Time#c.F i.Time#c.K

And the results are:

```
. reg I F K i.ind i.Time i.ind#c.F i.ind#c.K i.Time#c.F i.Time#c.K
```

Source	SS	df	MS		
Model	6391633.52	68	93994.6106	Number of obs	= 80
Residual	18513.5286	11	1683.04805	F(68, 11)	= 55.85
				Prob > F	= 0.0000
Total	6410147.05	79	81141.1019	R-squared	= 0.9971
				Adj R-squared	= 0.9793
				Root MSE	= 41.025

I	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F	-.3705667	.1036748	-3.57	0.004	-.5987533	-.1423801
K	-3.845659	.8712557	-4.41	0.001	-5.76328	-1.928038
ind						
GM	114.871	230.2471	0.50	0.628	-391.8995	621.6416
US	-306.8151	119.5875	-2.57	0.026	-570.0254	-43.6048
WEST	-855.8763	172.4571	-4.96	0.000	-1235.452	-476.3008
Time						
1936	-82.43587	105.1064	-0.78	0.449	-313.7735	148.9018
1937	-170.1636	162.1791	-1.05	0.317	-527.1174	186.7902
1938	-182.5426	101.2965	-1.80	0.099	-405.4948	40.40954
1939	-216.2866	102.9246	-2.10	0.059	-442.8221	10.24898
1940	-282.7636	123.386	-2.29	0.043	-554.3343	-11.19284
1941	-299.3507	111.0047	-2.70	0.021	-543.6703	-55.03097
1942	-461.6698	145.9172	-3.16	0.009	-782.8314	-140.5082
1943	-639.4738	201.1655	-3.18	0.009	-1082.236	-196.7115
1944	-691.2451	220.8914	-3.13	0.010	-1177.424	-205.0664
1945	-796.4654	223.8735	-3.56	0.004	-1289.208	-303.7232
1946	-755.3998	209.9574	-3.60	0.004	-1217.513	-293.2867
1947	-916.7204	281.921	-3.25	0.008	-1537.224	-296.2164
1948	-1078.229	340.9232	-3.16	0.009	-1828.596	-327.8624
1949	-1103.296	346.7723	-3.18	0.009	-1866.537	-340.0559
1950	-1163.5	351.3863	-3.31	0.007	-1936.896	-390.1044
1951	-1095.021	333.477	-3.28	0.007	-1828.999	-361.0436
1952	-1301.748	393.7757	-3.31	0.007	-2168.443	-435.0539
1953	-1582.253	455.4728	-3.47	0.005	-2584.742	-579.764
1954	-2425.422	588.0215	-4.12	0.002	-3719.649	-1131.195

ind#c.F						
GM	.1631697	.079258	2.06	0.064	-.0112759	.3376153
US	.2129046	.0584225	3.64	0.004	.0843176	.3414916
WEST	.4368447	.2157061	2.03	0.068	-.0379213	.9116107
ind#c.K						
GM	-.645235	.3513944	-1.84	0.093	-1.418649	.1281789
US	.799231	.1810003	4.42	0.001	.400852	1.19761
WEST	5.434994	2.083356	2.61	0.024	.8495575	10.02043
Time#c.F						
1936	.1349252	.0572089	2.36	0.038	.0090092	.2608413
1937	.1522291	.0946924	1.61	0.136	-.0561875	.3606457
1938	.1513868	.06562	2.31	0.042	.0069581	.2958154
1939	.2076969	.0673806	3.08	0.010	.0593932	.3560006
1940	.2372419	.0736022	3.22	0.008	.0752445	.3992393
1941	.2670278	.0803637	3.32	0.007	.0901484	.4439071
1942	.3231763	.0829668	3.90	0.002	.1405676	.5057851
1943	.3006946	.0788039	3.82	0.003	.1272483	.4741409
1944	.2992464	.0790301	3.79	0.003	.1253022	.4731906
1945	.2905874	.082037	3.54	0.005	.1100251	.4711496
1946	.3129331	.0915145	3.42	0.006	.1115109	.5143552
1947	.4560514	.1509522	3.02	0.012	.1238078	.7882949
1948	.6594446	.1897628	3.48	0.005	.2417795	1.07711
1949	.5090536	.1573973	3.23	0.008	.1626245	.8554828
1950	.5978037	.178968	3.34	0.007	.2038977	.9917096
1951	.5024197	.1481896	3.39	0.006	.1762565	.8285829
1952	.5965298	.1765534	3.38	0.006	.2079385	.9851211
1953	.3727521	.1569314	2.38	0.037	.0273483	.7181559
1954	1.089454	.5670916	1.92	0.081	-.1587061	2.337614
Time#c.K						
1936	1.394761	.9254125	1.51	0.160	-.6420584	3.43158
1937	3.862934	1.673926	2.31	0.041	.1786471	7.547221
1938	2.667286	.9001629	2.96	0.013	.6860405	4.648531
1939	2.406562	.7755587	3.10	0.010	.6995691	4.113556
1940	2.697141	.819789	3.29	0.007	.892798	4.501485
1941	2.791305	.8183275	3.41	0.006	.9901782	4.592432
1942	3.032671	.8657278	3.50	0.005	1.127217	4.938126
1943	3.664144	.939462	3.90	0.002	1.596402	5.731886
1944	3.779855	.97795	3.87	0.003	1.627401	5.932309
1945	4.393232	1.049381	4.19	0.002	2.083559	6.702905
1946	4.363707	1.018202	4.29	0.001	2.122659	6.604755
1947	4.062844	.9408668	4.32	0.001	1.99201	6.133678
1948	3.608075	.9331779	3.87	0.003	1.554165	5.661986
1949	4.069421	.9413913	4.32	0.001	1.997433	6.141409
1950	3.90444	.9267221	4.21	0.001	1.864738	5.944141
1951	4.035491	.9333531	4.32	0.001	1.981194	6.089787
1952	4.007791	.9710069	4.13	0.002	1.870619	6.144963
1953	4.985038	1.084637	4.60	0.001	2.597768	7.372309
1954	3.592048	1.630307	2.20	0.050	.0037673	7.180328
_cons	861.3578	184.5213	4.67	0.001	455.2292	1267.486

2.2 The Fixed Effects (Within-groups Regression) Model

The main problem with the above fixed effects (LSDV) model, as is clear from the above, is that it hosts too many regressors; this makes the model numerically unattractive and infects it with the problems of multicollinearity. Moreover, as the number of regressors increases, the degrees of freedom fall, and the error variance rises, leading to Type 2 error in inference (not rejecting a false null hypothesis). Another problem is that this model is unable to identify the impact of time-invariant variables (such as sex, colour, ethnicity, education, which are invariant over time). Again, the assumption that the error term follows classical rules [that $u_{it} \sim N(0, \sigma^2)$] can go wrong. For example, for a given period, it is possible that the error term for GM is correlated with the error term for, say, US or both US and WEST. If it so happens, we have to deal with it in terms of the seemingly unrelated regression (SURE) modelling aka Arnold Zellner (see Jan Kmenta 1986 *Elements of Econometrics*).

However, there is a simple way to estimate the fixed effects model without using dummy variables. Below we describe this.

Let us consider a simple one-way error components panel data model (with differential intercepts across individuals, which necessitate including dummy variables in estimation equation):

$$Y_{it} = \alpha_i + \beta X_{it} + v_{it}\ ; \quad i = 1, 2, \ldots, N; \quad t = 1, 2, \ldots, T. \qquad \ldots(2.2.1)$$

$$v_{it} \sim \text{IID}(0, \sigma^2);\ \text{Cov}(X_{it}, v_{is}) = 0;\ \forall\ t \text{ and } s.$$

Averaging the regression equation over time gives

$$\bar{Y}_{i.} = \alpha_i + \beta \bar{X}_{i.} + \bar{v}_{i.}, \qquad \ldots(2.2.2)$$

where $\bar{Y}_{i.} = \sum_t Y_{it}/T$, $\bar{X}_{i.} = \sum_t X_{it}/T$, and $\bar{v}_{i.} = \sum_t v_{it}/T$.

Now subtracting the first (2.2.1) equation from the second (2.2.2), we get

$$(Y_{it} - \bar{Y}_{i.}) = \beta(X_{it} - \bar{X}_{i.}) + (v_{it} - \bar{v}_{i.}). \qquad \ldots(2.2.3)$$

This deviations from means transformation is called Q transformation (Baltagi, 2008:15), which wipes out the differential intercepts. The OLS estimator for β from this transformed model is called within-groups FE estimator, or simply within estimator, as this estimator is based only on the variation within each company; this is exactly identical to the LSDV estimator. Since our panel model (2.2.1) is an Ancova model, the within estimator is also called covariance (CV) estimator. The individual-specific intercepts are estimated unbiasedly as:

$$\hat{\alpha}_i = \bar{Y}_{i.} - \hat{\beta}\bar{X}_{i.}, \quad i = 1, \ldots, N. \qquad \ldots(2.2.4)$$

We can also have an OLS estimator for β from the mean equation (2.2.2); this estimate is known as the between-group FE estimator, or simply between estimator.

All the statistical packages report the within estimator for the FE model. Stata has an additional option to give the between estimator also.

Note that we have estimated all the LSDV fixed effects models by OLS, without setting our data to panel data mode, that is, without invoking the xtset command. Now once we have xtset our data (as we did earlier), we can have the Stata within-groups fixed effects estimation by going to

Statistics → Longitudinal/panel data → Linear models → Linear regression (FE, RE, PA, BE)

When the xtreg window appears, enter the dependent (I) and independent (F, K) variables and mark the model type as fixed effects. We can also type the command

. xtreg I F K, fe

The output is:

```
. xtreg I F K, fe

Fixed-effects (within) regression               Number of obs      =        80
Group variable: ind                             Number of groups   =         4

R-sq:  within  = 0.8062                         Obs per group: min =        20
       between = 0.7294                                        avg =      20.0
       overall = 0.7548                                        max =        20

                                                F(2,74)            =    153.96
corr(u_i, Xb)  = -0.0822                        Prob > F           =    0.0000

------------------------------------------------------------------------------
           I |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           F |   .1060964   .0172848     6.14   0.000     .0716557     .140537
           K |    .347562   .0266309    13.05   0.000     .2944988    .4006252
       _cons |  -69.86518   37.06412    -1.88   0.063     -143.717    3.986688
-------------+----------------------------------------------------------------
     sigma_u |  138.96268
     sigma_e |  75.401517
         rho |  .77254819   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(3, 74) =    66.93                 Prob > F = 0.0000
```

Note that the results here on the marginal effects are identical with those of the FE LSDV model, as already indicated. The estimate of the constant intercept however is different and it is significant at 10% only here.

The xtreg command in Stata reports three R-squares, within, between and overall. Note that these reported R-squares do not share many of the properties of the OLS R^2. The common properties of OLS R^2 include the following:

(i) R^2 is equal to the squared correlation between the dependent variable (Y) and its estimate ($\hat{Y}$); and

(ii) R^2 is equal to the proportion of the variation in Y explained by its estimate ($\hat{Y}$); formally defined as R^2 = Var($\hat{Y}$) / Var(Y), and lying in the range of 0 and 1. These variances are reported in the text books in terms of the sum of the squared deviations, or simply, sum of squares (SS): R^2 = Explained ($\hat{Y}$) SS / Total (Y) SS.

It is important to note that this identity of the definitions is a special property of the OLS estimates (see Johnston, 1972:34-35); in general, the squared correlation between a variable (Y) and its estimate ($\hat{Y}$) need not be equal to the ratio of the variances, and the ratio of the variances need not be less than 1.

As already noted, the command xtreg, fe estimates (2.2.3) and (2.2.4) by OLS; hence its reported R^2 within has all the properties of the usual R^2. Other two R^2s are correlations squared, corresponding to the between estimator equation and an overall equation with a constant intercept. Thus the usual R^2 for our FE model is 0.8062, less than those for our earlier LSDV models. The overall R^2 is similar to that of the pooled regression. The Stata reports a poolability test at the bottom of the results; Stata uses u_i for our μ_i; the F-test rejects the null of zero company heterogeneity. Hence, between the pooled regression and FE model, we select the latter.

Stata also reports sigma_u, sigma_e, and rho; note that Stata's u is our μ (intercept heterogeneity) and e stands for the random error term v in our one-way error component model. The FE model assumes that the μ_i (or Stata's u_i) are formally fixed, having no distribution. Hence, we need not bother about this estimate. However, in the random effects model, this estimate does natter.

Estimating Panel Effects

We can have the estimates of the individual (cross section or panel) effects in Stata; first estimate the FE model using the command

. xtreg I F K, fe

Then type in the command area

. predict IE, u

This will generate the individual effects (IE) that can be viewed in the Data Editor (Edit). Note that we can give any name , instead of IE, for example, the command

. predict pe, u

will give the same series with name pe (panel effects).

The Random Effects Model

Let us consider a simple one-way error components model;

$$Y_{it} = \alpha + \beta X_{it} + u_{it} \; ; \; i = 1, 2, \ldots, N; \; t = 1, 2, \ldots, T. \qquad \ldots. (3)$$

$$u_{it} = \mu_i + v_{it} \; , \qquad \ldots (4)$$

In a FE model, the μ_is are assumed to be fixed. However, the main problem with the FE model is its specification with too many parameters, resulting in heavy loss of degrees of freedom. This problem can be averted if the μ_is are assumed to be random; this gives us a random effects (RE) model with

$$v_{it} \sim \text{IIN}(0, \sigma_v^2); \quad \mu_i \sim \text{IID}(0, \sigma_\mu^2);$$

$$\text{Cov}(v_{it}, \mu_i) = 0 \qquad \text{Cov}(v_{it}, X_{it}) = 0 \quad \text{Cov}(X_{it}, \mu_i) = 0. \quad \ldots.(5)$$

Individual error components are not correlated with each other, and not autocorrelated across both cross-section and time series units.

The presence of α and μ_i in the equation means that the sample of our four companies are drawn from the same population and have a common mean value for the intercept (α); the individual differences in the intercept values of each company are reflected in the error term μ_i.

Now let us consider the statistical properties of the composite error term $u_{it} = \mu_i + v_{it}$:

Evidently, $\text{E}(u_{it}) = 0$; and

$\text{Var}(u_{it}) = \sigma_\mu^2 + \sigma_v^2$ (sum of within and between component variances).

Here we have a significant result. Note that if $\sigma_\mu^2 = 0$, then $\text{Var}(u_{it}) = \sigma_v^2$; and there is no difference between the pooled regression model and the RE model; we can pool the data and run OLS. Hence the test on the null $\sigma_\mu^2 = 0$ can be taken as a poolability test in the context of pooled regression vs. RE model. Such a test is available in Breusch-Pagan test, discussed below.

It is also evident that the variance of the composite error term [$\text{Var}(u_{it}) = \sigma_\mu^2 + \sigma_v^2$] is constant and hence, the composite error term is homoscedastic for all i and t; but serially correlated over time only between the errors of the same company (unless $\sigma_\mu^2 = 0$). That is, under the assumptions in (5),

$$\text{Cov}(u_{it}, u_{js}) = \text{E}[(\mu_i + v_{it})(\mu_j + v_{js})] = \sigma_\mu^2 + \sigma_v^2, \text{ for } i = j, t = s \ [=\text{Var}(u_{it})]$$

$$= \text{E}(\mu_i^2) = \sigma_\mu^2, \text{ for } i = j, t \neq s \text{ (same company, over time)}$$

$$= 0, \text{ otherwise.}$$

And the correlation coefficient of u_{it} and u_{js} is given by

$$\rho(u_{it}, u_{is}) = 1, \text{ for } i = j, t = s \ [=\text{Var}(u_{it})/\ \text{Var}(u_{it})]$$

$$= \sigma_\mu^2 /(\sigma_\mu^2 + \sigma_v^2), \text{ for } i = j, t \neq s \text{ (same company, over time)}$$

$$= 0, \text{ otherwise.}$$

Thus the errors of each company are correlated over time; hence we call this correlation equi-correlation. The presence of such serial correlation makes the composite error term nonspherical, and the OLS estimation, inefficient.

In matrix notation, the OLS estimate of β is given by $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$, However, in the panel context, it is often the case that the OLS assumptions about the spherical error will not be accurate, as shown above. If we knew the shape of the errors (that is, their variance-covariance matrix) we could simply use it to modify our data and then apply OLS to the transformed data; this would give the generalized least squares (GLS) estimates. If the shape of the errors is Ω (an NT x NT variance-covariance matrix of the errors), the estimate of β is given by $\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$. In reality, often we might not know the shape of the errors and we could only use an estimate of Ω; this would give the feasible generalized least squares (FGLS) estimates: $\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y$.

In the presence of the (serially correlated) non-spherical error in our RE model, we need to modify our data using the information on the shape of the non-spherical error, and then apply OLS to the transformed data. This GLS estimator of the RE model is obtained by applying OLS to the data after the following transformation into quasi deviations:

$$(y_{it} - \theta\bar{Y}_i) = (1-\theta)\alpha + \beta(X_{it} - \theta\bar{X}_i) + \{(1-\theta)\mu_i + (v_{it} - \theta\bar{v}_i)\},$$

where

$$\theta = 1 - \sqrt{\sigma_v^2/(\sigma_v^2 + T\sigma_\mu^2)}.$$

This is given without proof in Hausman (1978:1262); also see Johnston (1984:402).

The term θ gives a measure of the relative sizes of the within and between component variances. We have the following results on the transformed quasi-deviation form model:

1. If $\theta = 1$, the RE-estimator is identical with the FE-within estimator; this is possible when $\sigma_v^2 = 0$, which means that every v_{it} is zero, given $E(v_{it}) = 0$; in this case the FE regression will have an R^2 of 1.
2. If $\theta = 0$, the RE-estimator is identical with the pooled OLS-estimator; this is because, $\sigma_\mu^2 = 0$, which means that μ_i is always zero, given $E(\mu_i) = 0$.

Normally, θ will lie between 0 and 1.

If $Cov(X_{it}, \mu_i) \neq 0$, the RE-estimator will be biased. The degree of the bias will depend on the size of θ. If σ_μ^2 is much larger than σ_v^2, then θ will be close to 1, and the bias of the RE-estimator will be low.

One major difficulty with RE estimator is that its small sample properties are unknown; it has only asymptotic properties.

Now let us turn to estimating the RE model for our data. Once we have xtset our data (as we did earlier), we can have the Stata random effects estimation by going to

Statistics → Longitudinal/panel data → Linear models → Linear regression (FE, RE, PA, BE)

When the xtreg window appears, enter the dependent (I) and independent (F, K) variables and mark the model type as GLS random-effects. We can also type the command

. xtreg I F K, re

The output is:

```
. xtreg I F K, re

Random-effects GLS regression                   Number of obs      =        80
Group variable: ind                             Number of groups   =         4

R-sq:  within  = 0.8062                         Obs per group: min =        20
       between = 0.7294                                        avg =      20.0
       overall = 0.7548                                        max =        20

                                                Wald chi2(2)       =    316.59
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

I	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
F	.1059662	.0166155	6.38	0.000	.0734005	.1385319
K	.3470571	.0265224	13.09	0.000	.2950742	.39904
_cons	-69.39439	83.17499	-0.83	0.404	-232.4144	93.62561
sigma_u	150.78857					
sigma_e	75.401517					
rho	.79996932	(fraction of variance due to u_i)				

Note that the marginal effects and intercept are almost equal to those of the FE-within model reported above; however, the intercept here is not at all significant. Also note that all the R^2s are equal to those of the FE-within model. Since the RE estimator has only asymptotic properties, the F statistic for overall model significance is not reported here; rather, we have the results from a Wald chi-square test that indicates that the model as a whole is (all the coefficients taken jointly are) significant.

Stata obtains the result by assuming that the correlation of μ_i and the explanatory variables is zero, or $\text{Cov}(X_{it}, \mu_i) = 0$. This is reported as corr(u_i, X) = 0 (assumed).

Stata also reports sigma_u (our μ), sigma_e (our v), and rho. We have σ_μ = 150.78857 and σ_v = 75.401517 and rho = 0.79996932. Stata reports rho as "fraction of variance due to u_i"; remember our definition of this correlation: $\rho = \sigma_\mu^2 / (\sigma_\mu^2 + \sigma_v^2)$, the proportion of the variance of μ_i in the total variance of the error components. Note that we do not have the estimate of theta, used in the quasi-deviation in the results above, because we have not explicitly specified for it; we can estimate it, using the formula $\theta = 1 - \sqrt{\sigma_v^2 / (\sigma_v^2 + T\sigma_\mu^2)}$, and the values of σ_μ = 150.78857 and σ_v = 75.401517 and T = 20 as theta = 0.8889, somewhat close to unity. If we want the estimate of theta to be reported in the results, then we have to type

. xtreg I F K, re theta

And the result is

```
. xtreg I F K, re theta

Random-effects GLS regression                   Number of obs      =        80
Group variable: ind                             Number of groups   =         4

R-sq:  within  = 0.8062                         Obs per group: min =        20
       between = 0.7294                                        avg =      20.0
       overall = 0.7548                                        max =        20

                                                Wald chi2(2)       =    316.59
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
theta          = .88887837
```

I	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
F	.1059662	.0166155	6.38	0.000	.0734005	.1385319
K	.3470571	.0265224	13.09	0.000	.2950742	.39904
_cons	-69.39439	83.17499	-0.83	0.404	-232.4144	93.62561
sigma_u	150.78857					
sigma_e	75.401517					
rho	.79996932	(fraction of variance due to u_i)				

We have seen that if $\sigma_\mu^2 = 0$, then the variance of the composite error term reduces to Var(u_{it}) = σ_v^2; and there is no difference between the pooled regression model and the RE model; we can pool the data and run OLS. Now given that σ_μ = 150.78857 here, we cannot do this. However, we can have a formal test in terms of Breusch-Pagan poolability test in the context of pooled regression vs. RE model by going to

Statistics → Longitudinal/panel data → Linear models → Lagrange multiplier test for random effects

or, by typing

. xttest0

The result is:

```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        I[ind,t] = Xb + u[ind] + e[ind,t]

        Estimated results:
                     |       Var     sd = sqrt(Var)
             ---------+------------------------------
                   I |     81141.1       284.8528
                   e |    5685.389       75.40152
                   u |    22737.19       150.7886

        Test:   Var(u) = 0
                             chibar2(01) =   378.65
                          Prob > chibar2 =   0.0000
```

We reject the null of $\sigma_\mu^2 = 0$; we cannot pool the data, but select the RE model.

We have earlier seen that in the context of pooled regression vs. FE model, we have favoured the FE model, and now in the context of pooled regression vs. RE model, we have selected the RE model. Now the question is: Which one is better, FE or RE?

FE- or RE-Modelling?

For most of the research problems, there is room to suspect that Cov(X_{it}, μ_i) ≠ 0. That means the RE-estimator will be biased. Hence, it would be wiser to use the FE-estimator to get unbiased estimates. The RE-estimator, however, provides estimates for time-invariant covariates. Many studies would attempt to analyse the marginal effects of certain variables after accounting for the effects of sex, race, etc. This is possible only with the RE modeling. Suppose the cross section units are individual workers, and we want to study the workers' earnings (Y_i), including a categorical variable for race (Z_i) in the model:

$$Y_{it} = \beta_1 X_{it} + \beta_2 Z_i + u_{it} ;$$

$$u_{it} = \mu_i + v_{it} , \quad v_{it} \sim \text{IID}(0, \sigma_v^2); \quad i = 1, 2, \ldots, \text{N}; \quad t = 1, 2, \ldots, \text{T}.$$

In the FE model, it would be impossible to estimate β_2, because it would not be possible to distinguish between the worker-specific constant term (μ_i) and the effect of the time-invariant variable, race (Z_i); the two would be perfectly multicollinear. Since RE accepts μ_i as a random variable, it easily allows for the estimation of β_2, by averting this multicollinearity.

Judge, et al. (1988) propose the following simple rules:

1. If T is large and N small, there is little difference in the parameter estimates of FE and RE models. Hence computational convenience prefers FE model.

2. If N is large and T small, the two methods differ. If cross-sectional units in the sample are random drawings from a larger sample, RE model is appropriate; otherwise, FE model.

3. If the individual error component, μ_i, and one or more regressors are correlated, RE estimators are biased and FE estimators unbiased

4. If N is large and T small, and if the assumptions of RE modeling hold, RE estimators are more efficient.

In most applications the assumption that Cov(x_{it}, μ_i) = 0 may be wrong, and the RE-estimator will be biased. "This is risking to throw away the big advantage of panel data only to be able to write a paper on "The determinants of Y"". (Josef Brüderl, 2005: *Panel Data Analysis*).

However, we can have a test for RE vs. FE, in terms of the null hypothesis H_0: E(u_{it}| X_{it}) = 0. Note that this null implies H_0: E(μ_i | X_i) = 0. Hausman (1978) proposes to compare $\hat{\beta}_{RE}$ and $\hat{\beta}_{FE}$, both of which are consistent under the null H_0: E(u_{it}| X_{it}) = 0. In fact, $\hat{\beta}_{FE}$ is consistent whether the null is true or not, whereas $\hat{\beta}_{RE}$ is best linear unbiased estimator (BLUE), consistent and asymptotically efficient under the null, but is inconsistent when the null is false. Note that any test statistic for a mean difference comparison consists in the ratio of the difference between the statistics to its standard error, or the squared ratio in asymptotic cases. Thus a test statistic in our case can be based on the mean difference $\hat{q} = \hat{\beta}_{RE} - \hat{\beta}_{FE}$; under the null, the probability limit of this value is: plim $\hat{q}$ = 0, and Cov($\hat{\beta}_{RE}$, $\hat{q}$) = 0. The variance of this mean difference is Var($\hat{q}$) = Var($\hat{\beta}_{RE}$) – Var($\hat{\beta}_{FE}$). Thus the test statistic for Hausman's specification test is $h = \hat{q}'[Var(\hat{q})]^{-1}\hat{q}$, where $\hat{q} = \hat{\beta}_{RE} - \hat{\beta}_{FE}$ and Var($\hat{q}$) = Var($\hat{\beta}_{RE}$) – Var($\hat{\beta}_{FE}$), to test the null H_0: E(μ_i | X_i) = 0 against the alternative H_a: E(μ_i | X_i) ≠ 0. Under the null hypothesis, this statistic is distributed asymptotically as central chi-squared, with k (= number of parameters) degrees of freedom.

Now we turn to conducting the Hausman test to see whether a fixed-effects or random effects model is more appropriate for the Grunfeld data that we consider. The procedure in Stata is as follows:

Estimate the FE model by typing

. xtreg I F K, fe

And store the result as fe by typing

. estimates store fe

Then estimate the RE model by typing

. xtreg I F K, re

And do the Hausman test by typing

. hausman fe

The output is

```
. hausman fe
```

	Coefficients			
	(b) fe	(B) .	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
F	.1060964	.1059662	.0001302	.0047634
K	.347562	.3470571	.0005049	.0024016

```
                b = consistent under Ho and Ha; obtained from xtreg
 B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        0.07
                Prob>chi2 =      0.9660
```

The test fails to reject the null, as the p-value (Prob>chi2) is greater than 5%. Note that the Hausman test is a test of

H_0: random effects would be consistent and efficient, versus

H_1: random effects would be inconsistent.

Hence we select the RE model. The tests imply that the company effects though present in the data set are not correlated with the explanatory variables, and can very well be taken as random; the RE estimators will be consistent and efficient.

The statistical tests in the context of panel data analysis in a nutshell

FE vs. OLS $H_0 = \mu_1 = \mu_2 = \ldots = \mu$ F or Wald Test	RE vs. OLS $H_0 = \text{Var}(\mu_i) = 0$ Breusch-Pagan Test	Your Model
H_0 not rejected $\Rightarrow$ No FE	H_0 not rejected $\Rightarrow$ No RE	Pooled OLS
H_0 rejected $\Rightarrow$ FE	H_0 not rejected $\Rightarrow$ No RE	FE Model
H_0 not rejected $\Rightarrow$ No FE	H_0 rejected $\Rightarrow$ RE	RE Model
H_0 rejected $\Rightarrow$ FE	H_0 rejected $\Rightarrow$ RE	Choose one based on Hausman test.

"There is no simple rule to help the researcher navigate past the Scylla of fixed effects and the Charybdis of measurement error and dynamic selection. Although they are an improvement over cross-section data, panel data do not provide a cure-all for all of an econometrician's problems." (Johnston and DiNardo 1997: 403).

References

Airy (1861) *On the algebraical and numerical theory of errors of observations and the combination of observations* Macmillan, Cambridge and London

Andreß, Hans-Jürgen; Katrin Golsch and Alexander W. Schmidt (2013) Applied Panel Data Analysis for Economic and Social Surveys. Springer-Verlag, Berlin Heidelberg

Arellano, Manuel (2003) *Panel Data Econometrics* Oxford University Press

Baltagi, Badi H. (2001) *Econometric Analysis of Panel Data*, 2nd ed., John Wiley and Sons.

Baltagi, Badi H. (2005) *Econometric Analysis of Panel Data*, 3rd Edition, John Wiley and Sons.

Eisenhart (1947) 'The assumptions underlying the Analysis of Variance', *Biometrics* Vol. 3: pp. 1-21.

Fisher R. A. (1925) *Statistical Methods for Research Workers* Oliver and Boyd, London.Reprinted in *Statistical Methods, Experimental Design and Scientific Inference* OUP, Oxford

Frees, Edward W. (2004) *Longitudinal and Panel Data Analysis and Applications in the Social Sciences* Cambridge University Press.

Grunfeld, Y. (1958) "The Determinants of Corporate Investment," unpublished Ph.D. thesis, Department of Economics, University of Chicago,.

Hausman, J. A. (1978). *Specification Tests in Econometrics*, Econometrica, Vol. 46, No. 6. (Nov., 1978), pp. 1251-1271.

Hsiao, Cheng (2003) *Analysis of Panel Data*, 2nd Edition, Cambridge University Press.

Hsiao, Cheng (2014) *Analysis of Panel Data*. Third edition. Cambridge University Press, New York.

Johnston, J. (1972) *Econometric Methods*. Second edition. McGraw-Hill.

Judge, GG, Hill, RC, Griffiths, WE, Lutkepohl, H and Lee, TS (1988) *Introduction to the Theory and Practice of Econometrics*

Lillard, L.A. and R.J. Willis, 1978, Dynamic aspects of earning mobility, *Econometrica* **46**, 985–1012.

Nerlove, Marc (2002) *Essays in Panel Data Econometrics* Cambridge University Press

Wooldridge, Jeffrey M (2001) *Econometric Analysis of Cross Section and Panel Data* The MIT Press.