

Problem Set 2
Economics 60303
(Due: Friday, February 10, 2012)

Bill Evans
Spring 2012

1. Show that if $\Omega_t = \sigma_\varepsilon^2 I_t + \sigma_u^2 i_t i_t'$ then $\Omega_t^{-1} = \frac{1}{\sigma_\varepsilon^2} [I_t - \frac{\theta}{t} i_t i_t']$ where $\theta = \frac{\sigma_u^2}{T\sigma_u^2 + \sigma_\varepsilon^2}$

2. Consider the two-way fixed-effects model $Y_{it} = \alpha + X_{it}\beta + u_i + v_t + \varepsilon_{it}$ where X is a scalar. In a data set with balanced panels and T=2 observations per group, the author estimates the model as a first difference $\Delta Y_i = \theta + \Delta X_i \beta + \Delta \varepsilon_i$. What does the coefficient on θ represent in this case? The researcher then estimates a first difference model with an entire vector of individual fixed effects $\Delta Y_i = \theta + \Delta X_i \beta + \lambda_i + \Delta \varepsilon_i$ and they estimate this model by adding a dummy for n-1 panels to the first difference model, so the equation they estimate is of the form

$$\Delta Y_i = \theta + \Delta X_i \beta + \sum_{i=1}^{T-1} D_i \pi_i + \Delta \varepsilon_i$$

where D_i is defined as in class. In this model, what is being estimated by the parameter $\hat{\pi}_i$?

3. An author has a balanced panel data set (NT observations) and considers estimating an equation of the form

$$y_{it} = x_{it}\beta + v_t + \varepsilon_{it}$$

where v_t is a random year effects where $E[v_t] = 0$ and $Var(v_t) = \sigma_v^2$.

- a) Suppose the data is sorted by year then group and we can write the equation in matrix notation as

$$Y = X\beta + V$$

Where V is the composite error and $V_{it} = v_t + \varepsilon_{it}$. Using Kroeneker products, what is $E[VV']$?

- b) How does your answer change if the data is sorted by group then year?

4. An author has a balanced panel data set and considers estimating a model of the form

$$y_{it} = x_{it}\beta + v_{it}$$

Where x_{it} is a $(k \times 1)$ row vector. In this case v_{it} has a three-part error structure $v_{it} = u_i + \lambda_t + \varepsilon_{it}$ where u_i is a random effect that varies across individuals but is common over time for person i , and λ_t is a random year effect that is common to all groups but is varying randomly across years. Assume $E[u_i] = E[\lambda_t] = E[\varepsilon_{it}] = 0$, $Var(u_i) = \sigma_u^2$, $Var(\lambda_t) = \sigma_\lambda^2$, $Var(\varepsilon_{it}) = \sigma_\varepsilon^2$ and the errors are random effects such that $cov(u_i, x_{jit}) = cov(\lambda_t, x_{jit}) = 0$.

Write the model in matrix form $Y = X\beta + V$ where Y is an $NT \times 1$ vector.

- a) What is $E[VV']$?
 b) What would be the GLS estimate of $\hat{\beta}$? (Just write the equation).
5. Continuing with problem 5. In order to estimate the model by feasible GLS, we need estimates for σ_u^2 , σ_λ^2 , and σ_ε^2 . Consider estimating $y_{it} = x_{it}\beta + v_{it}$ by OLS and let us use the residuals from that model to obtain estimates of these variances. To make the problem a little easier, suppose we have the actual value of the residual and not an

estimate. So define $v_{it} = y_{it} - x_{it}\beta$. Define $v_{i\cdot} = \frac{1}{T} \sum_{t=1}^T v_{it}$ and $v_{\cdot t} = \frac{1}{N} \sum_{i=1}^N v_{it}$ and

$$v_{\cdot\cdot} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it}$$

What is $Var(v_{i\cdot})$, $Var(v_{\cdot t})$, and $Var(v_{\cdot\cdot})$? With these values, generate a linear combination of v_{it} , $v_{i\cdot}$, $v_{\cdot t}$, and $v_{\cdot\cdot}$ that will have a variance of σ_ε^2 .

6. A fellow classmate has written their own code to estimate within, between, and random effects models with balanced panels. As a check on their work, they estimate between, within and random effects models for the one-way effects models with one covariate. The equation that describes the fixed and random effects models is $Y_{it} = X_{it}\beta + u_i + \varepsilon_{it}$. The results are summarized below. You look at the results and tell the student they have a programming error. You are so smart. What tipped you off?

Variable	Between	Fixed	Random
X_{it}	0.01092 (0.00214)	0.03285 (0.00251)	0.03312 (0.00267)

Your classmate fixes their coding error and generates the following results for a different problem. These are correct. What is the Hausman test statistic for the null hypothesis that u_i and X_{it} are uncorrelated? Can you reject or not reject the null? You can calculate this by hand.

Parameter Estimates and Standard Errors

	Between	Fixed	Random
	0.11012	0.04813	0.0635
	(0.04412)	(0.02429)	(0.02137)

Computer Exercise

All states and the Federal government levy excise taxes on cigarettes. There have been dozens of studies that use changes in taxes and corresponding changes in consumption to estimate the elasticity of demand for cigarettes. In this problem set, we will examine how a large tax hike altered smoking rates in an important and interesting population.

In 1993, Michigan voters passed a referendum eliminating local property taxes, which are the main source of revenues for schools. To make up for lost revenue, the Michigan legislature raised the cigarette tax from 25 to 75 cents per pack. The higher tax rate went into effect on May 1, 1994. The Surgeon General of the US estimates that smoking during pregnancy doubles the chance a baby will be born with a low birth weight (<2500 grams). Although smoking rates among pregnant women have fallen considerably over the past 20 years, roughly 17 percent of births are to women who smoked during their pregnancy. In recent years, a number of public health officials have suggested that higher cigarette taxes can be used as way to improve birth outcomes. We will use the data from the Michigan “experiment” to evaluate this conjecture.

The data for this project are taken from the Natality Detail File, which is an annual census of births in the US. I have taken a 5% random sample of births for the state of Michigan for the 32 months prior and 24 months after the tax hike. I have also include a 5% random sample of data over the same period for two Midwestern states that had no nominal change in their state cigarette tax rates over this period: Iowa and Pennsylvania.

The data for this project is in the STATA data file michigan.dta. The data set has a little more than 76,000 observations. Detailed variable definitions are listed below.

Variable	Definition
MONTH	This is an index that equals 1 in the first month (September 1991) 2 in the second (October 1991), through month 56. Month 33 is the month the new tax went into effect (May of 1994).
STATE	2-digit state FIPS code. Michigan is state 26.
SMOKED	Dummy variable, =1 if a mother self-reported that she smoked during her pregnancy, =0 otherwise.

MRACE3	3 level variable, =1 if mother wife, =2 if Black, =3 if other race.
MEDUC6	6-level variable for mother's education: =1 if <9 years, =2 if 9-11 years, =3 if 12 years, =4 if 13-15 years, =5 if 16+ years, =6 if education was not reported.
PARITY	4-level variable for mother's parity of birth. =1 if this is the first birth, =2 if the second birth, =3 if third birth, =4 if fourth or higher birth.
HISPANIC	Dummy variable, =1 if mother is Hispanic, =0 otherwise.
MARRIED	Dummy variable, =1 if mother is married, =0 otherwise.

This is a fairly large data set so to read this file into STATA, you need to set the memory to about 50 Meg. Another option you must change is the MATSIZE which is the maximum number of variables that can be added to a regression model. For this program, set the MATSIZE to 100.

For the first four problems, treat the data from Iowa and Pennsylvania as one control group.

1. Construct two variables: A dummy variable for Michigan (name it MI) and another for the period after the tax rate is increased (HIKE). Sort the data by these two variables and calculate the mean smoking rate before and after the tax hike for Michigan and the control group. Using these means, calculate the difference in difference estimate of the impact of higher taxes on smoking in Michigan.
 2. Using the two variables from problem 1 and any other necessary variables, calculate a "different in difference" estimate in a regression framework. For guidance, see the Meyer, Viscusi and Durbin *AER* paper on the reading list. How does this estimate compare to the estimate from problem 1? Did the tax hike reduce smoking rates by a statistically significant amount
 3. Using the XI: command, re-run the model from problem 2 but add month effects, Does the estimate of the treatment effect variable change when this set of variables are added to the model? How do you interpret this result?
1. Again using the XI: command, re-run the model from problem 3 but add race effects, age effects, education effects, parity effects and dummies for Hispanic and Married. How does the estimate of the treatment effect variable change when these variables are added to the model? How do you interpret this change?
 2. Calculate the average smoking rate in Michigan in the 12 months prior to the tax hike. Note that the average retail price of cigarettes in Michigan in the year before the tax hike was 177 cents. Assume that retail prices increase one cent for every cent increase in the excise tax, and using the means calculated above, what is the implied elasticity of smoking participation from the treatment effect coefficient you estimated in problem 4? (for a discussion about how to calculate the elasticity, see the Cook and Tauchen paper in your reading packet).

3. Restrict your sample to the 32 months prior to the tax hike. Estimate a model with a Michigan dummy, month effects, race effects, age effects, education effects, parity effects and dummies for Hispanic and Married. Next, estimate separate monthly dummy variables for Michigan and the control states (You can do this using the XI command. Check the online help for instructions). Using the printout, do an F-test that the monthly dummy variables for Michigan differ from the time-series in the control states. What are the degrees of freedom of the f-test and what is a 95% critical value for this test. For the pre-treatment period, can you accept or reject the null hypothesis that the monthly dummies are the same in the treatment and control states?
4. Redo question 6, but estimate two separate models, one with MI and Iowa in the sample, another with MI and Pennsylvania.
5. This data set is a only a 5% random sample of the populations in each state. What do you think would happen to the value of the f-test statistic in problem 5 if you had a census of births in each state? It is easier to answer this question by re-writing the f-test as a function of R^2 's rather than SSE's (sums of squared errors).
6. Generate a new variable that equals 50 in control states and 1 in Michigan and call this WGT1. Next, re-run the model in problem 4) using the fweight command – Fweight is a frequency weight for the model. For example, in the STATA code

```
reg model y x [fweight=wgt1]
```

the regression will treat each observation as if there are WGT1 observations per line. In this case, this will increase the sample size of the control group considerably. Looking at the regression printout, what is the sample size for this new model? What has happened to the estimate and precision of the treatment effect when the sample size of the control group increases by a factor of 50? Construct a new variable WGT2 that equals 1 for MI and 200 for the control group and add this to the FWEIGHT command. Does the precision of the estimate change much? What does this about the importance of the size of the comparison sample necessary for these types of analyses?