ISEG Business School

Microeconometrics I

Assignment 01

Cheng Amy Qian 59040
Edin Secerkadic 58296
Dhruv Pandit 58802

Descriptive Statistics

Our data set is a balanced panel data, with 10 variables consisting of a total of 630 observations throughout 7 time periods from 1981 to 1987. Out of the 10 variables, 3 are dummy variables specifying geographical characteristics of each of the observation (*Appendix_1*). Additional time dummy variables are added to the data set specifying the year of each of the observations to account for the fixed time effect.

All the variables have observations within –3 standard deviations from the mean, but almost all the variables have observations which are higher than +3 standard deviations from the mean. This finding can indicate the fact that the outliers might be present in the data set and that the distribution of the data is right skewed. Additionally, there are variables with standard deviations relatively small, indicating minimal variation in the data, the accuracy of estimation using the Fixed Effect method may be impacted. Another important note about such spread of the data is the variance may not be constant throughout the data, indicating presence of heteroskedasticity (*Appendix_2*).

Pooled OLS Estimation Method

To use the Pooled OLS estimation method, several assumptions must be made. The unobserved heterogeneity term c_i is independent of the idiosyncratic error term u_{it}. Contemporaneous exogeneity is assumed, implying that the overall error term of the Pooled OLS, v_{it}, is exogenous. No collinearity between the observations is also assumed. And lastly, assumption on homoskedasticity for the error term v_{it} is made which implies homoskedasticity for both the unobserved heterogeneity term c_i as well as the idiosyncratic error term u_{it}.

Under these assumptions, the Pooled OLS method using the non-robust standard error is applied first (*Appendix_3*). To check for violation on homoskedasticity and the absence of serial correlation of the error term, the White Test on heteroskedasticity is applied which shows a p-value of 1.8426e-48 concluding that the residuals are heteroskedastic (*Appendix_4*). To test the autocorrelation between residuals, the ACF of the residuals is plotted (*Appendix_5*). The plot shows clear correlation between the residuals which violates the Gauss Markov assumptions, implying the invalidity of inferences through the standard Pooled OLS method. Instead, the White clustered robust standard error should be used for estimation (*Appendix_6*). The Time Fixed Effect Test is conducted on the adjusted Pooled OLS estimates. With a p-value of 0, the null is rejected, concluding that the model is significant (*Appendix_7*).

Pooled OLS takes out the time dependency effect from the unobserved heterogeneity term c_i and introduces a new error term v_{it} which combines the unobserved term and the idiosyncratic error term. This takes out the need to differentiate between the two terms. And as c_i has a permanent effect on the dependent variable, we can use simple transformation methods to remove it. On top of the convenience in modeling, the Pooled OLS would always be a consistent estimator as long as the model is correctly specified, and all the assumptions are satisfied. Although the Pooled OLS method is always consistent, the underlying assumptions are the most restrictive comparing to other estimating methods. Especially contemporaneous exogeneity which is a strong assumption for panel data.

Random Effect Estimation Method

The Random Effect estimation method, similar to the Pooled OLS method, introduces the error term v_{it} which is a combination of the unobserved term and the idiosyncratic error term. But unlike the Pooled OLS, the heterogeneous unobserved term c_i can now be manipulated. The introduction of time dependency of the unobserved term changes the original assumption of contemporaneous exogeneity to strict exogeneity of both c_{it} and u_{it}. The assumption on no perfect collinearity remains the same. And also, that both c_{it} and u_{it} are homoscedastic with a covariance matrix of Ω.

In presence of heteroskedasticity and serial correlation with our data, referring to the test statistics of the White and Pearson R Tests (*Appendix_8*), as well as the ACF plot of autocorrelation of the residuals (*Appendix_9*), the Feasible Generalized Least Squares method (FGLS) is applied. Instead of the normal covariance matrix, the White clustered robust covariance matrix is used (*Appendix_10*). In fact, the robust covariance matrix is able to provide us with a consistent estimation regardless of the condition of the heteroskedastic and serial correlation properties of the data. The F-test on joint significance of the model shows a p-value of 0 which proves this model to be significant (*Appendix_8*). To formally test for the presence of random effect, the Breusch-Pagan Lagrange Multiplier Test is applied with the null hypothesis of the absence of random effect. The test result shows a p-value of 8.0554e-131, rejecting the null (*Appendix_11*). There exists strong evidence of the presence of random effect in the model.

The Random Effect Estimator, also known as the Feasible GLS Estimator, is, in most cases, a more efficient estimator compared to the Pooled OLS estimator. But the higher efficiency of the Random Effect estimator is not necessarily true when there is homoscedasticity and the absence of serial correlation in the idiosyncratic errors. The drawback of using the Random Effect Estimator may include the invalidation of the estimators when the serial correlation is too strong in the error term. Also like the Pooled OLS Estimation method, the estimator would be inconsistent if the correct model is the Fixed Effect Model.

Fixed Effect Estimation Method

Unlike the Pooled OLS estimation method and the Random Effect estimation method, the Fixed Effect estimation method removes the assumption of exogeneity in the unobserved term c_{it}. But the strict exogeneity of the idiosyncratic error term u_{it} still needs to hold for the original model. The assumption of no perfect collinearity remains needed to hold. As well as the assumption on homoskedasticity and the absence of within autocorrelation in the error term.

Once again, to determine the proper covariance matrix used for estimation, the assumption on homoskedasticity and the absence of serial correlation in the error term is tested using the Modified Wald Test (*Appendix_12*) and Woodridge Test (*Appendix_13*). The violation of the assumption of homoskedasticity, based on the Modified Wald Test statistics of p-values of 0, leads to the decision of using the White clustered robust standard error for estimation (*Appendix_14*). By applying the Fixed Effect estimation method, the variables west, central, and urban are absorbed due to their lack of variation in time. To formally test the evidence of fixed effect, the F test with the null hypothesis of no evidence of fixed effect is applied. The p-value of 0 rejects the null hypothesis, suggesting the existence of fixed effect. The overall significance of the model is also tested through a joint significance F test. The p-value of 0 indicates the overall significance of the fixed effect estimation model (*Appendix_15*).

The biggest advantage of the Fixed Effect estimation method is its relaxed assumption on the exogeneity of the unobserved term, which, in practice, occurs in most cases. But one major drawback for the Fixed Effect estimation method is that it loses all variables that are constant over time. And the independent variables with very low variation over time would negatively impact the estimation of the β's by having a low accuracy level. On top of that, by taking out the time factor through transformation, we also lose the number of observations from TN observations to N observations.

Result Comparison

All three estimation methods are significant as the p-value of the F statistics are 0 for all three models. The estimates are relatively close between the Pooled OLS and the Random Effect method. This is due to the similarity in assumptions between the two methods while the Fixed Effect method is more relaxed. The variable *avgsen*, average sentence in days, may require further analysis as Pooled OLS shows a negative estimate while the estimators of Random Effect and Fixed Effect are positive. Another variable that may require further analysis is *density*. While Pooled OLS and Random Effect estimators show significance at 1% level, Fixed Effect estimator shows no statistical significance of the variable (*Appendix_16*).

Choice of Estimator

To determine the estimator of choice, the Robust Hausman Test using bootstrap is done. The choice to use the Robust Hausman Test instead of the Standard Hausman Test is due to the existence of heteroskedasticity in the error term. The test statistics shows a p-value of 0.2079, implying the choice of estimator would be to use the Random Effects estimators, as we fail to reject the null hypothesis at a significance level of 0.05 (5%) while the Bruesch Pagan LM test overwhelmingly rejects the null hypothesis, establishing the significance of random effects (*Appendix_17*). Furthermore, the random effects model is in most cases always preferable to the fixed effects models, due to the latter establishing causation in the absence of context (Bell & Jones, 2015).

Interpretation and Conclusion on Results

Out of the 8 estimators, only 5 variables show statistical significance at a 5% threshold according to the Random Effect estimation model. In the case of the time dummy variables, we observe strong statistical significance for the years 1984 and 1985, with associated negative coefficients of -0.0039 and -0.0036 respectively, indicating a rough decrease of 0.3% in the number of crimes committed if the person was born in the aforementioned years.

A 1% increase in the *average probability of arrest* corresponds to a 0.01% decrease in the observed crime rate. A strong significance is observed for the number of *police per capita* and the population *density*; a 1% increase in the former is associated with a 2.1% increase in the crime rate, while a 1% increase in the latter corresponds to a 0.008% increase in the crime rate.

The dummy variables *west* and *central* indicating the location of the county are statistically significant, with negative coefficients. If the county is located in the west, the observed crime rate is roughly 1% less, while a 0.7% decrease in the crime rate is observed in central counties.

Appendix

Appendix_1 – Balanced Panel Data

```
Panel variable: county (strongly balanced)
 Time variable: year, 81 to 87
         Delta: 1 year
```

Appendix_2 – Descriptive Statistics

	county	year	crmrte	prbconv	prbarr	avgsen	polpc	density	taxpc	west	central	urban
count	630.00000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000
mean	100.60000	1984.000000	0.031588	0.688618	0.307368	8.954540	0.001917	1.386062	30.239194	0.233333	0.377778	0.088889
std	58.03627	2.001589	0.018121	1.690345	0.171205	2.658082	0.002735	1.439703	11.454694	0.423289	0.485217	0.284809
min	1.00000	1981.000000	0.001812	0.068376	0.058823	4.220000	0.000459	0.197719	14.302565	0.000000	0.000000	0.000000
25%	51.00000	1982.000000	0.018352	0.347692	0.217902	7.160000	0.001191	0.532944	23.425596	0.000000	0.000000	0.000000
50%	103.00000	1984.000000	0.028441	0.474375	0.278240	8.495000	0.001451	0.952595	27.792328	0.000000	0.000000	0.000000
75%	151.00000	1986.000000	0.038406	0.635597	0.352518	10.197500	0.001803	1.507818	33.271218	0.000000	1.000000	0.000000
max	197.00000	1987.000000	0.163835	37.000000	2.750000	25.830000	0.035578	8.827652	119.761452	1.000000	1.000000	1.000000

Appendix_3 – Pooled OLS, Standard

PooledOLS Estimation Summary

Dep. Variable:	crmrte	**R-squared:**	0.7131
Estimator:	PooledOLS	**R-squared (Between):**	0.7948
No. Observations:	630	**R-squared (Within):**	0.1633
Date:	Tue, Mar 14 2023	**R-squared (Overall):**	0.7131
Time:	17:13:05	**Log-likelihood**	2026.6
Cov. Estimator:	Unadjusted		
		F-statistic:	101.75
Entities:	90	**P-value**	0.0000
Avg Obs:	7.0000	**Distribution:**	F(15,614)
Min Obs:	7.0000		
Max Obs:	7.0000	**F-statistic (robust):**	101.75
		P-value	0.0000
Time periods:	7	**Distribution:**	F(15,614)
Avg Obs:	90.000		
Min Obs:	90.000		
Max Obs:	90.000		

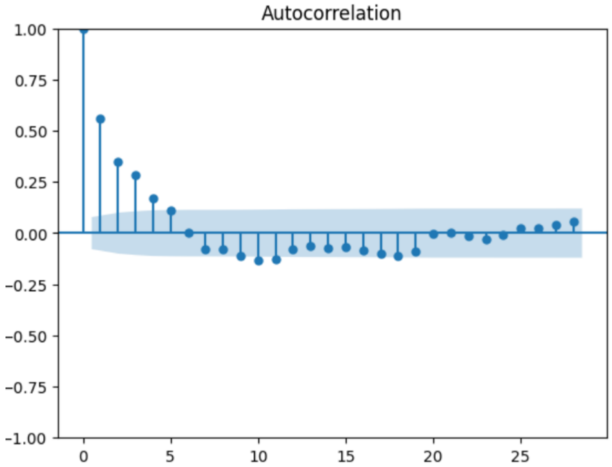
Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0334	0.0026	12.698	0.0000	0.0282	0.0385
prbconv	-0.0021	0.0003	-7.7803	0.0000	-0.0026	-0.0016
prbarr	-0.0294	0.0025	-11.597	0.0000	-0.0343	-0.0244
avgsen	-0.0002	0.0002	-1.1773	0.2395	-0.0005	0.0001
polpc	2.6118	0.1714	15.239	0.0000	2.2752	2.9484
density	0.0082	0.0005	15.340	0.0000	0.0071	0.0092
taxpc	3.842e-05	4.073e-05	0.9432	0.3460	-4.157e-05	0.0001
west	-0.0131	0.0011	-12.417	0.0000	-0.0151	-0.0110
central	-0.0071	0.0010	-7.2581	0.0000	-0.0090	-0.0051
urban	-0.0023	0.0025	-0.9073	0.3646	-0.0073	0.0027
1981	0.0012	0.0016	0.7778	0.4370	-0.0019	0.0043
1982	0.0006	0.0016	0.3727	0.7095	-0.0025	0.0037
1983	-0.0013	0.0016	-0.8354	0.4038	-0.0043	0.0018
1984	-0.0039	0.0015	-2.5633	0.0106	-0.0069	-0.0009
1985	-0.0035	0.0015	-2.3083	0.0213	-0.0064	-0.0005
1986	-0.0018	0.0015	-1.1980	0.2314	-0.0047	0.0011

Appendix_4 – White Test Statistics, POLS

```
(359.7724127439798,
 1.84256960596813e-48,
 15.417243597812835,
 3.087510937963384e-77)
```

Appendix_5 – ACF of Residuals Plot, POLS



Appendix_6 – Pooled OLS, White Clustered Robust Standard Error

PooledOLS Estimation Summary

Dep. Variable:	crmrte	**R-squared:**	0.7131
Estimator:	PooledOLS	**R-squared (Between):**	0.7948
No. Observations:	630	**R-squared (Within):**	0.1633
Date:	Tue, Mar 14 2023	**R-squared (Overall):**	0.7131
Time:	17:17:30	**Log-likelihood**	2026.6
Cov. Estimator:	Robust		
		F-statistic:	101.75
Entities:	90	**P-value**	0.0000
Avg Obs:	7.0000	**Distribution:**	F(15,614)
Min Obs:	7.0000		
Max Obs:	7.0000	**F-statistic (robust):**	140.34
		P-value	0.0000
Time periods:	7	**Distribution:**	F(15,614)
Avg Obs:	90.000		
Min Obs:	90.000		
Max Obs:	90.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0334	0.0045	7.4450	0.0000	0.0246	0.0422
prbconv	-0.0021	0.0008	-2.6242	0.0089	-0.0036	-0.0005
prbarr	-0.0294	0.0063	-4.6621	0.0000	-0.0417	-0.0170
avgsen	-0.0002	0.0002	-1.0363	0.3005	-0.0005	0.0002
polpc	2.6118	0.6336	4.1220	0.0000	1.3675	3.8562
density	0.0082	0.0005	16.947	0.0000	0.0072	0.0091
taxpc	3.842e-05	8.241e-05	0.4662	0.6413	-0.0001	0.0002
west	-0.0131	0.0010	-13.554	0.0000	-0.0149	-0.0112
central	-0.0071	0.0010	-7.3959	0.0000	-0.0089	-0.0052
urban	-0.0023	0.0022	-1.0597	0.2897	-0.0066	0.0020
1981	0.0012	0.0017	0.7201	0.4717	-0.0021	0.0046
1982	0.0006	0.0017	0.3483	0.7278	-0.0027	0.0039
1983	-0.0013	0.0017	-0.7767	0.4377	-0.0046	0.0020
1984	-0.0039	0.0016	-2.4888	0.0131	-0.0070	-0.0008
1985	-0.0035	0.0015	-2.2623	0.0240	-0.0065	-0.0005
1986	-0.0018	0.0016	-1.0962	0.2734	-0.0049	0.0014

Appendix_7 – Time Time Fixed Effect Test

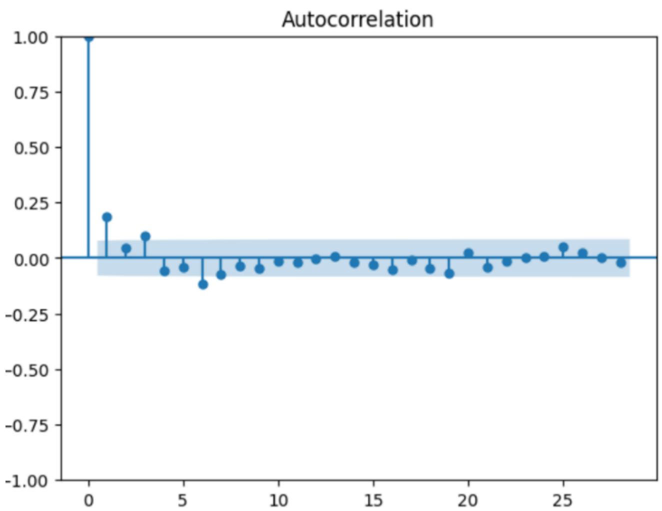
```
. . test (y1 y2 y3 y4 y5 y6)

 ( 1)  y1 = 0
 ( 2)  y2 = 0
 ( 3)  y3 = 0
 ( 4)  y4 = 0
 ( 5)  y5 = 0
 ( 6)  y6 = 0

       F(  6,    89) =   12.95
            Prob > F =    0.0000
```

Appendix_8 – White, Pearson R, and F Test Statistics, RE

```
(436.96603142142015, 3.293170535058554e-63, 26.21334825740892, 1.619160857772303e-117)
PearsonRResult(statistic=0.4810353510979873, pvalue=8.494831762945844e-38)
```

```
Model F-statistic (homoskedastic)
H0: All parameters ex. constant are zero
Statistic: 33.9077
P-value: 0.0000
Distributed: F(15,614)
```

Appendix_9 – ACF of Residuals Plot, RE



Appendix_10 – Random Effect, White Clustered Robust Standard Error

RandomEffects Estimation Summary

Dep. Variable:	crmrte	**R-squared:**	0.4531
Estimator:	RandomEffects	**R-squared (Between):**	0.7403
No. Observations:	630	**R-squared (Within):**	0.3007
Date:	Tue, Mar 14 2023	**R-squared (Overall):**	0.6834
Time:	16:59:40	**Log-likelihood**	2333.5
Cov. Estimator:	Clustered		
		F-statistic:	33.908
Entities:	90	**P-value**	0.0000
Avg Obs:	7.0000	**Distribution:**	F(15,614)
Min Obs:	7.0000		
Max Obs:	7.0000	**F-statistic (robust):**	67.393
		P-value	0.0000
Time periods:	7	**Distribution:**	F(15,614)
Avg Obs:	90.000		
Min Obs:	90.000		
Max Obs:	90.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0257	0.0044	5.8582	0.0000	0.0171	0.0343
prbconv	-0.0011	0.0006	-1.8337	0.0672	-0.0023	7.925e-05
prbarr	-0.0111	0.0046	-2.4167	0.0160	-0.0202	-0.0021
avgsen	5.137e-05	8.439e-05	0.6087	0.5429	-0.0001	0.0002
polpc	2.1610	0.4329	4.9915	0.0000	1.3108	3.0112
density	0.0088	0.0010	8.9455	0.0000	0.0069	0.0108
taxpc	2.945e-05	0.0001	0.2710	0.7865	-0.0002	0.0002
west	-0.0136	0.0021	-6.3768	0.0000	-0.0178	-0.0094
central	-0.0075	0.0023	-3.2734	0.0011	-0.0120	-0.0030
urban	-0.0028	0.0044	-0.6194	0.5359	-0.0115	0.0060
1981	0.0007	0.0015	0.4853	0.6276	-0.0022	0.0036
1982	0.0005	0.0014	0.3881	0.6981	-0.0022	0.0032
1983	-0.0017	0.0014	-1.2168	0.2242	-0.0044	0.0010
1984	-0.0039	0.0011	-3.5657	0.0004	-0.0061	-0.0018
1985	-0.0036	0.0009	-4.2322	0.0000	-0.0053	-0.0019
1986	-0.0015	0.0009	-1.5797	0.1147	-0.0034	0.0004

Appendix_11 – Breusch-Pagan Lagrange Multiplier Test

```
LM Test: chisq = 592.2657107270627, df = 1, p-value = 8.055389883186393e-131
'Random Effects are significant at a level of 0.05'
```

Appendix_12 – Modified Wald Test, FE

```
Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (90)  =    40453.85
Prob>chi2 =       0.0000
```

Appendix_13 – Wooldridge Test, FE

```
Wooldridge test for autocorrelation in panel data
H0: no first order autocorrelation
    F(  1,      89) =      0.381
           Prob > F =      0.5387
```

Appendix_14 – Fixed Effect, White Clustered Robust Standard Error

PanelOLS Estimation Summary

Dep. Variable:	crmrte	**R-squared:**	0.3075
Estimator:	PanelOLS	**R-squared (Between):**	0.3588
No. Observations:	630	**R-squared (Within):**	0.3075
Date:	Tue, Mar 14 2023	**R-squared (Overall):**	0.3522
Time:	17:18:08	**Log-likelihood**	2393.2
Cov. Estimator:	Clustered		
		F-statistic:	19.537
Entities:	90	**P-value**	0.0000
Avg Obs:	7.0000	**Distribution:**	F(12,528)
Min Obs:	7.0000		
Max Obs:	7.0000	**F-statistic (robust):**	44.295
		P-value	0.0000
Time periods:	7	**Distribution:**	F(12,528)
Avg Obs:	90.000		
Min Obs:	90.000		
Max Obs:	90.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0265	0.0083	3.1884	0.0015	0.0102	0.0428
prbconv	-0.0010	0.0006	-1.6287	0.1040	-0.0021	0.0002
prbarr	-0.0071	0.0044	-1.5991	0.1104	-0.0157	0.0016
avgsen	8.775e-05	0.0001	0.8280	0.4081	-0.0001	0.0003
polpc	2.1206	0.3796	5.5870	0.0000	1.3750	2.8662
density	0.0024	0.0061	0.4018	0.6880	-0.0095	0.0144
taxpc	4.375e-05	9.678e-05	0.4521	0.6514	-0.0001	0.0002
1981	0.0002	0.0013	0.1540	0.8776	-0.0023	0.0027
1982	0.0001	0.0012	0.1243	0.9011	-0.0022	0.0025
1983	-0.0021	0.0012	-1.6961	0.0905	-0.0045	0.0003
1984	-0.0042	0.0010	-4.2833	0.0000	-0.0061	-0.0023
1985	-0.0038	0.0008	-4.8799	0.0000	-0.0053	-0.0023
1986	-0.0015	0.0009	-1.6216	0.1055	-0.0034	0.0003

Appendix_15 –Evidence of Fixed Effect Test, Joint Significance Test, FE

```
Pooled F-statistic
H0: Effects are zero
Statistic: 18.0961
P-value: 0.0000
Distributed: F(89,528)


Model F-statistic (homoskedastic)
H0: All parameters ex. constant are zero
Statistic: 19.5367
P-value: 0.0000
Distributed: F(12,528)
```

Appendix_16 – Result Comparison

Variable	beta_POLS	beta_RE	beta_FE
prbconv	-.00207375***	-.00112743***	-.00095477***
prbarr	-.02936678***	-.01138751***	-.00705592***
avgsen	-.0001851	.00004853	.00008775
polpc	2.6118221***	2.1639995***	2.1206051***
density	.00816042***	.00883316***	.00244864
taxpc	.00003842	.00002959	.00004375
west	-.0130544***	-.01359322***	(omitted)
central	-.00705607***	-.00752424***	(omitted)
urban	-.00230481	-.00282236	(omitted)
y1	.00123103	.00071608	.00019593
y2	.00058333	.0005353	.00014625
y3	-.00129636	-.00168561	-.00208813**
y4	-.00391078**	-.00391906***	-.00417305***
y5	-.00347369***	-.0036166***	-.003814***
y6	-.00177261*	-.0014982	-.00153686*
_cons	.03337131***	.02576131***	.02645621***

Legend: * p<.1; ** p<.05; *** p<.01

Appendix_17 – Robust Hausman Test

```
%%stata
. quietly xtreg $yxvars y1-y6, fe vce(cluster county)
. estimates store b_rob_FE
. quietly xtreg $yxvars y1-y6, re vce(cluster county)
. estimates store b_rob_RE
. *set the seed of the random generator for boostrap
. set seed 5271982
. rhausman b_rob_FE b_rob_RE, reps(500) cluster

✓ 41.2s
```

```
Output exceeds the size limit. Open the full output data in a text editor

. . quietly xtreg $yxvars y1-y6, fe vce(cluster county)

. . estimates store b_rob_FE

. . quietly xtreg $yxvars y1-y6, re vce(cluster county)

. . estimates store b_rob_RE

. . *set the seed of the random generator for boostrap

. . set seed 5271982

. . rhausman b_rob_FE b_rob_RE, reps(500) cluster
bootstrap in progress
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................. 50
(This bootstrap will approximately take another 0h. 0min. 36sec.)
.................................................. 100
.................................................. 150
.................................................. 200
.................................................. 250
.................................................. 300
.................................................. 350
.................................................. 400
...
                       =       16.81
            Prob>chi2 =       0.2079
```