```
* Dr. Steven S. Vickner;
* MSBA 635 - Data Analytics II;

* print data;
```
*Panel or longitudinal data tracks cross-sectional units through time
*Pooled regression models ignore cross-sectional
*heterogeneity in the panel by including only an overall intercept in the model
*One-way fixed effects (FE) models
*control for cross-sectional heterogeneity by including an indicator variable for each cross section except one (to avoid
*the 'dummy variable trap')
*One-way random effects (RE) models control for cross-sectional heterogeneity by adding an error component to the overall error term that only varies by cross section
*Pooled regression models and one-way random effects models can accommodate time-invariant explanatory variable whereas one-way fixed effects models cannot
*Explain variability in wages as a function of education, experience, and demographic variables
```
*5 years of data stacked on top on each other
*degrees of freedom are much larger here. Now you have n*t observations
*when you stack data you get efficiency gains. Your t-statistics go up. 100 observations with 5 cross sections will give you 500 rows of data
*panel data helps break down multicolinearity through the heterogeneity you bring in
*the proc reg ignores the heterogeneity in the data
*one way fixed effect models-put in indicator variable for each cross section in our data set except for one (dummy variable trap)
*one way random effects-adds an error component to a usual error term to control for cross-section heterogeneity
*if you have 100 cross sections, 5years each. And if those 1000 were random samples from 10,000 units then you have to treat them as a cross sectional unit to do the one-way random effect
*your n should be greater than your t (n>t)
*you want more cross sections than you have time periods
*were they from the population? Used fixed effects. If from a sample of the population use random effects.

*see how the data is stacked here.
*id means individual id. Individual 1 has 5 years of data. So does individual 2, 3, 4, 5. You can just run a proc reg on this. You don't have to treat it any other way that a proc reg would treat it
*you are ignoring heterogeneity at the individual level
*__vertically concatenating the data__ (stacking). Individual 2-5 data is under individual 1.
```

```
proc print data=tmp1.nls_panel (obs=25);
run;
```


Obs	id	year	lwage	hours	age	educ	collgrad	msp	mar	nonmsa	ncity	south	black	union	exper	exper2	tenure	tenure2
1	1	82	1.80829	38	30	12	0	1	0	0	1	0	1	1	7.6667	58.778	7.6667	58.778
2	1	83	1.86342	38	31	12	0	1	0	0	1	0	1	1	8.5833	73.674	8.5833	73.674
3	1	85	1.78937	38	33	12	0	0	0	0	1	0	1	1	10.1795	103.622	1.8333	3.361
4	1	87	1.84653	40	35	12	0	0	0	0	1	0	1	1	12.1795	148.340	3.7500	14.063
5	1	88	1.85645	40	37	12	0	0	0	0	1	0	1	1	13.6218	185.553	5.2500	27.563
6	2	82	1.28093	48	36	17	1	1	0	0	0	0	0	0	7.5769	57.410	2.4167	5.840
7	2	83	1.51586	43	37	17	1	1	0	0	0	0	0	0	8.3846	70.302	3.4167	11.674
8	2	85	1.93017	35	39	17	1	1	0	0	0	0	0	0	10.3846	107.840	5.4167	29.340
9	2	87	1.91903	42	41	17	1	1	0	0	0	0	0	1	12.0385	144.925	0.3333	0.111
10	2	88	2.20097	42	43	17	1	1	0	0	0	0	0	1	13.2115	174.545	1.7500	3.063
11	3	82	1.81483	48	35	12	0	1	0	0	0	0	0	0	11.4167	130.340	11.4167	130.340
12	3	83	1.91991	40	36	12	0	1	0	0	0	0	0	1	12.4167	154.174	12.4167	154.174
13	3	85	1.95838	45	38	12	0	1	0	0	0	0	0	0	14.4167	207.840	14.4167	207.840
14	3	87	2.00707	40	40	12	0	1	0	0	0	0	0	0	16.4167	269.507	16.4167	269.507
15	3	88	2.08985	48	42	12	0	1	0	0	0	0	0	0	17.8205	317.571	17.7500	315.063
16	4	82	2.31254	40	30	12	0	0	0	0	1	0	0	1	12.5833	158.340	12.5833	158.340
17	4	83	2.34858	40	31	12	0	0	0	0	1	0	0	1	13.5833	184.507	13.5000	182.250
18	4	85	2.37349	40	33	12	0	0	0	0	1	0	0	1	15.5833	242.840	15.5000	240.250
19	4	87	2.36890	40	35	12	0	0	0	0	1	0	0	1	17.5833	309.174	17.5000	306.250
20	4	88	2.35053	40	37	12	0	0	0	0	1	0	0	1	19.0449	362.707	19.0000	361.000
21	5	82	2.11986	3	34	14	0	1	0	0	0	0	0	0	8.3654	69.980	1.0000	1.000
22	5	83	2.66787	2	35	14	0	1	0	0	0	0	0	0	9.3654	87.710	2.0000	4.000
23	5	85	3.57913	1	37	14	0	1	0	0	0	0	0	0	10.7692	115.976	2.1667	4.694
24	5	87	1.67888	15	39	14	0	1	0	0	0	0	0	0	12.7692	163.053	4.0000	16.000
25	5	88	2.57137	8	40	14	0	1	0	0	0	0	0	0	14.2500	203.063	5.5000	30.250

```
* display data attributes;

proc contents data=tmp1.nls_panel;
run;
```


The CONTENTS Procedure

Data Set Name	TMP1.NLS_PANEL	Observations	3580
Member Type	DATA	Variables	18
Engine	V9	Indexes	0
Created	11/13/2010 16:36:43	Observation Length	144
Last Modified	11/13/2010 16:36:43	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	12288
Number of Data Set Pages	43
First Data Page	1
Max Obs per Page	85
Obs in First Data Page	63
Number of Data Set Repairs	0
Filename	C:\Users\nxnguy01\Desktop\nls_panel.sas7bdat
Release Created	9.0202M3
Host Created	W32_VSPRO

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
5	age	Num	8	age in current year
13	black	Num	8	= 1 if black; 0 if white
11	c_city	Num	8	= 1 if central city
7	collgrad	Num	8	= 1 if college graduate
6	educ	Num	8	current grade completed
15	exper	Num	8	= total work experience
16	exper2	Num	8	= exper^2
4	hours	Num	8	usual hours worked
1	id	Num	8	identifier for panel individual; 716 total
3	lwage	Num	8	ln(wage/GNP deflator)
8	msp	Num	8	= 1 if married, spouse present
9	nev_mar	Num	8	= 1 if never yet married
10	not_smsa	Num	8	= 1 if not SMSA
12	south	Num	8	= 1 if south
17	tenure	Num	8	= job tenure, in years
18	tenure2	Num	8	= tenure^2
14	union	Num	8	= 1 if union member
2	year	Num	8	year interviewed (1982, 1983, 1985, 1987, 1988)

```
* keep selected variables;
*the "keep" statement will let you keep a small number of columns for your model so that
you don't have to see them all. It will create a temporary sas data set
*brings you down to 11 variables instead of the 18 variables.
*you have the other variables in your permanent sas data set but not in your work data
set

data nlsdata;
set tmp1.nls_panel;
keep id year lwage educ exper exper2 tenure tenure2 black south union;
run;


* estimate model using proc reg;
*the proc reg is run on stacked data
*one year increase in education leads to a 7.145% increase in wages

options nolabel;
proc reg data=work.nlsdata;
model lwage = educ exper exper2 tenure tenure2 black south union;
run;
quit;
```


The REG Procedure
Model: MODEL1
Dependent Variable: lwage

Number of Observations Read	3580
Number of Observations Used	3580

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	251.53504	31.44188	215.50	<.0001
Error	3571	521.02619	0.14590		
Corrected Total	3579	772.56123			

Root MSE	0.38197	R-Square	0.3256
Dependent Mean	1.91824	Adj R-Sq	0.3241
Coeff Var	19.91280		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.47660	0.05616	8.49	<.0001
educ	1	0.07145	0.00269	26.57	<.0001
exper	1	0.05569	0.00861	6.47	<.0001
exper2	1	-0.00115	0.00036129	-3.18	0.0015
tenure	1	0.01496	0.00441	3.39	0.0007
tenure2	1	-0.00048604	0.00025770	-1.89	0.0594
black	1	-0.11671	0.01572	-7.43	<.0001
south	1	-0.10600	0.01420	-7.46	<.0001
union	1	0.13224	0.01496	8.84	<.0001

```
* estimate model using proc panel;
*new procedure for panel data
*the id statement asks if you have a cross sectional identifier "id" and a time-series
point (aka time periods)"year"
*id in our data should have been named individual
*pooled does a whites correction from heteroscedasticity  (hccme=4) you want the model to
go cross section by cross section and within those 5 time periods for individual lets
make sure we control for heteroscedasticity for that level. Make sure the variance for
those 5 error terms for each id is constant.
*the "=4" tells it what option to use
*what's different is the standard errors here on proc panel the standard errors for this
test are bigger than the previous proc reg. the whites test generally make the standard
errors larger
*whenever something is statistically significant close to zero(<.0001) its standard error
will be close to zero
*t-values go down
*p-values go up
*this is trying to help you decide what goes on the right hand side
*you would throw out tenure2 at 0.2353 p-value
*5*716=3580. 3580-9=3571
*tells us the number of individual here and the cross section numbers
*R-square is 32.56%
*Get rid of the heteroscedasticity as much as you can

options nolabel;
proc panel data=work.nlsdata;
id id year;
model lwage = educ exper exper2 tenure tenure2 black south union / pooled hccme=4;
run;
quit;
```


The PANEL Procedure
Pooled (OLS) Estimates

Dependent Variable: lwage

Model Description

Estimation Method	Pooled
Number of Cross Sections	716
Time Series Length	5
Hetero. Corr. Cov. Matrix Estimator	4

Fit Statistics

SSE	521.0262	DFE	3571
MSE	0.1459	Root MSE	0.3820
R-Square	0.3256		

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	0.4766	0.0844	5.65	<.0001	Intercept
educ	1	0.071449	0.00549	13.02	<.0001	
exper	1	0.055685	0.0113	4.93	<.0001	
exper2	1	-0.00115	0.000492	-2.33	0.0196	
tenure	1	0.01496	0.00711	2.10	0.0354	

tenure2	1	-0.00049	0.000409	-1.19	0.2353
black	1	-0.11671	0.0281	-4.16	<.0001
south	1	-0.106	0.0270	-3.92	<.0001
union	1	0.132243	0.0270	4.89	<.0001

```
* estimate model using proc panel;
*one way fixed effect model
*in the first model demographic variable for education and for ethnicity are included
*in the second model both of the terms are dropped out
*this looks odd, check out your log, it “does not have full rank”
*you have 2 columns that are exactly the same, so they can’t be inverted
*there is perfect multicolinearity here
*set up a dummy variable for each cross section except for one
*set up 715 dummy variables, the 716th one is tossed out
*find a linear combo of your dummies to be perfectly parallel with those 2 variables
*only works in a fixed effect model
*education level and ethnicity level did not change in these models
*do you want to drop this model where you drop education and black or do you want your
one-way random effects model
*715-1=714. Set up 715 indicator/dummy variables, let their coefficients equal zero
*reject the null. The coefficients are statistically different from zero
*715 is here again
*reduced set of explanatory variables are here
*the R-square jumped up by like 50 percentage points
*ON THE FINAL
```

```
options nolabel;
proc panel data=work.nlsdata;
id id year;
model lwage = educ exper exper2 tenure tenure2 black south union / fixone;
model lwage = exper exper2 tenure tenure2 south union / fixone;
run;
quit;
```


The PANEL Procedure
Fixed One-Way Estimates

Dependent Variable: lwage

Model Description

Estimation Method	FixOne
Number of Cross Sections	716
Time Series Length	5

Fit Statistics

SSE	108.7985	DFE	2858
MSE	0.0381	Root MSE	0.1951
R-Square	0.8592		

F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
715	2858	15.15	<.0001

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	0.899436	0.1019	8.82	<.0001	Intercept
educ	0	0	.	.	.	
exper	1	0.041083	0.00662	6.21	<.0001	
exper2	1	-0.00041	0.000273	-1.50	0.1346	
tenure	1	0.013909	0.00328	4.24	<.0001	
tenure2	1	-0.0009	0.000206	-4.35	<.0001	
black	0	0	.	.	.	
south	1	-0.01632	0.0361	-0.45	0.6516	
union	1	0.063697	0.0143	4.47	<.0001	

The PANEL Procedure
Fixed One-Way Estimates

Dependent Variable: lwage

Model Description

Estimation Method	FixOne
Number of Cross Sections	716
Time Series Length	5

Fit Statistics

SSE	108.7985	DFE	2858
MSE	0.0381	Root MSE	0.1951
R-Square	0.8592		

F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
715	2858	19.66	<.0001

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	0.899436	0.1019	8.82	<.0001	Intercept
exper	1	0.041083	0.00662	6.21	<.0001	
exper2	1	-0.00041	0.000273	-1.50	0.1346	
tenure	1	0.013909	0.00328	4.24	<.0001	
tenure2	1	-0.0009	0.000206	-4.35	<.0001	
south	1	-0.01632	0.0361	-0.45	0.6516	
union	1	0.063697	0.0143	4.47	<.0001	

```
* estimate model using proc panel;
*educ and black are back.
*this is a random effects model which doesn't care about time or variance
*the Breusch Pagan test says that there is no random effects and you reject the null.
Conclude that there are random effects
*the variance of these random effects are equal to zero
*we are talking about spread in the bell shaped curve. Is it larger or smaller? Smaller
means smaller r-square
*a vertical line instead of a bell shaped curve means that there is no variance on the
random effects. It's a degenerate distribution. Meaning that the mass all piles up at one
point
*the null that we are testing is that variance for our individual variance is equal to
zero. We are rejecting this
*parameter estimates: tenure2 and black. Tenure2 has a smaller p-value. Tenure2 and black
are estimable here

options nolabel;
proc panel data=work.nlsdata;
id id year;
model lwage = educ exper exper2 tenure tenure2 black south union / ranone bp;
run;
quit;
```


The PANEL Procedure
Fuller and Battese Variance Components (RanOne)

Dependent Variable: lwage

Model Description

Estimation Method	RanOne
Number of Cross Sections	716
Time Series Length	5

Fit Statistics

SSE	136.4066	DFE	3571
MSE	0.0382	Root MSE	0.1954
R-Square	0.1940		

Variance Component Estimates

Variance Component for Cross Sections	0.10863
Variance Component for Error	0.038068

Hausman Test for Random Effects

Coefficients	DF	m Value	Pr > m
6	6	20.64	0.0021

Breusch Pagan Test for Random Effects (One Way)

DF	m Value	Pr > m
1	3859.28	<.0001

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.533933	0.0800	6.68	<.0001
educ	1	0.073256	0.00534	13.73	<.0001
exper	1	0.04361	0.00636	6.86	<.0001
exper2	1	-0.00056	0.000263	-2.14	0.0328
tenure	1	0.014154	0.00317	4.47	<.0001
tenure2	1	-0.00076	0.000195	-3.88	0.0001
black	1	-0.11676	0.0302	-3.86	0.0001
south	1	-0.08174	0.0224	-3.65	0.0003
union	1	0.080196	0.0132	6.07	<.0001