~~Homework~~ assignment 4

Group -10

Shweta Siddha, Manisha Gupta, Achint Khanijo, Kartikay Nigam

19/20

Q1. I have given you a panel data on wages (Wage data) in which N=334, T=3 years (1984-1986). For each ID, the data is sorted by year. You need to create the ID and year variables.

Columns	Variable name	Description
C1	Edu	Education in years
C2	Hr	Work hours per year
C3	Wage	Dollar wage per hour
C4	Famearn	Family earnings in dollars per year
C5	Self	Dummy for self-employed
C6	Sal	Dummy for salaried
C7	Mar	Dummy for married
C8	Numkid	Number of children
C9	Age	
C10	unemp	Local unemployment percentage

We need to do a regression to understand the determinants of "natural log (wages)" that is {ln(wage)}.
We need to understand the effect of the following variables: age, edu, numkid, hr, mar, sal, self, unemp.

1. Find the best linear regression model. Check for multicollinearity and take appropriate actions.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	1.35635	0.18086	7.50	<.0001	0	0
age	1	0.01225	0.00213	5.75	<.0001	0.17327	1.20133
edu	1	0.06793	0.00723	9.40	<.0001	0.28153	1.18581
numkid	1	0.02640	0.01957	1.35	0.1775	0.04148	1.24890
hr	1	-0.00014238	0.00002526	-5.64	<.0001	-0.16716	1.16278
married	1	0.13837	0.07512	1.84	0.0658	0.05318	1.10220
salaried	1	0.29291	0.04436	6.60	<.0001	0.20436	1.26600
selfempl	1	-0.35071	0.05203	-6.74	<.0001	-0.19653	1.12365
locunemp	1	-0.01499	0.01134	-1.32	0.1867	-0.03688	1.02955

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	125.74484	15.71811	41.24	<.0001
Error	992	378.13268	0.38118		
Corrected Total	1000	503.87752			

Root MSE	0.61740	R-Square	0.2496
Dependent Mean	2.59390	Adj R-Sq	0.2435
Coeff Var	23.80201		

Above regression is done with
Null hypothesis: Coefficient of age, education, Number of children, Work hours per year, Dummy for married, Dummy for salaried, Dummy for self-employed and Local unemployment percentage is equal to zero.
Alternate hyposthesis: Atleast one of the coefficient is not zero.
As the p-value (0.0001) is less than alpha (0.05), therefore we have sufficient proof to reject the null hypothesis, hence the coefficient of atleast one independent variables is not zero.
For the above resultset, the R-squared is 0.2496 and ajusted R-squared is 0.2435, this means that the given 8 independent variables could explain only 24% of variation in dependent variable.

Variable significance:
Variables age, education, Work hours per year, Dummy for salaried and Dummy for self-employed are significant variables at 95% confidence level. But variables Number of children, Dummy for married and Local unemployment percentage are not significant.

Test for Multicollinearity:
As the VIF of all the variable coefficient is less than 10, and Collin index is less than 100, we can say that there is no multicollinearity in the dataset.

Re-running the regression model with significant variables:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	124.37957	20.72993	54.30	<.0001
Error	994	379.49795	0.38179		
Corrected Total	1000	503.87752			

Root MSE	0.61789	R-Square	0.2468
Dependent Mean	2.59390	Adj R-Sq	0.2423
Coeff Var	23.82094		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	1.28890	0.14926	8.64	<.0001	0	0
age	1	0.01128	0.00198	5.68	<.0001	0.15951	1.03994
edu	1	0.06806	0.00718	9.48	<.0001	0.28204	1.16764
hr	1	-0.00014082	0.00002515	-5.60	<.0001	-0.16533	1.15085
married	1	0.16489	0.07288	2.26	0.0239	0.06338	1.03575
salaried	1	0.29524	0.04415	6.69	<.0001	0.20599	1.25214
selfempl	1	-0.34800	0.05202	-6.69	<.0001	-0.19501	1.12141

2

In the first analysis, we found that Dummy for married variable was not significant at 95% confidence level, but after removing number of children and Local unemployment percentage from the model; Dummy for married variable became significant.
For the above resultset, the R-squared is 0.2468 and adjusted R-squared is 0.2423, that means the given 6 significant independent variables are able to again explain only 24% of variation in dependent variable (wage).

2. Develop a model to test if there are nonlinear effects for some variables. Which variables have non-linear effect on ln(wages).

NON-LINEARITY TEST:

1. Wage (log term) as dependent variable and Age, Age squared term as independent variable:
   ln(wage) = b0 + b1*age + b2*(age*age)   ln(wage) = b0 + b1*age + b2*(age*age) +b3*edu+b4*hr+b5*mar+b6*sal+b7*self

Nonlinear OLS Summary of Residual Errors							
Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq
wage_lm	3	998	484.6	0.4856	0.6968	0.0383	0.0363

Nonlinear OLS Parameter Estimates				
Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
b0	0.858198	0.4574	1.88	0.0609
b1	0.071203	0.0219	3.26	0.0012
b2	-0.00067	0.000250	-2.70	0.0070

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	0.53590	0.41338	1.30	0.1951	0	0
age	1	0.04924	0.01954	2.52	0.0119	0.69652	101.11231
age_sq	1	-0.00043674	0.00022363	-1.95	0.0511	-0.54021	101.26851
edu	1	0.06668	0.00720	9.26	<.0001	0.27633	1.17899
hr	1	-0.00014321	0.00002515	-5.70	<.0001	-0.16813	1.15358
married	1	0.16284	0.07279	2.24	0.0255	0.06259	1.03596
salaried	1	0.29272	0.04410	6.64	<.0001	0.20423	1.25321
selfempl	1	-0.35059	0.05196	-6.75	<.0001	-0.19647	1.12215

On performing regression between dependent variable wage (log term) and; age and age squared term as independent variable, they came out to be statistically significant at 95% confidence interval. 2

But on running regression between significant variables (obtained in question 1-1) including age squared term, age squared term is not very significant at 95% confidence interval.

b. Wage (log term) as dependent variable and education, education squared term as independent variable:
ln(wage) = b0 + b1*edu + b2*(edu*edu)   ln(wage) = b0 + b1*age + b2* edu+b3*(edu*edu) +b4*hr+b5*mar+b6*sal+b7*self

Nonlinear OLS Summary of Residual Errors							
Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq
wage_lm	3	998	435.2	0.4361	0.6604	0.1363	0.1345

Nonlinear OLS Parameter Estimates				
Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
b0	2.603955	0.1784	14.60	<.0001
b1	-0.1214	0.0294	-4.12	<.0001
b2	0.008504	0.00124	6.86	<.0001

Number of Observations		Statistics for System	
Used	1001	Objective	0.4348
Missing	0	Objective*N	435.2120

Root MSE	0.60796	R-Square	0.2716
Dependent Mean	2.59390	Adj R-Sq	0.2665
Coeff Var	23.43800		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2.13154	0.20642	10.33	<.0001	0
age	1	0.01045	0.00196	5.34	<.0001	1.04552
edu	1	-0.08531	0.02733	-3.12	0.0019	17.48640
edu_sq	1	0.00677	0.00116	5.81	<.0001	17.90609
hr	1	-0.00013646	0.00002476	-5.51	<.0001	1.15190
married	1	0.14623	0.07178	2.04	0.0419	1.03782
salaried	1	0.24950	0.04415	5.65	<.0001	1.29329
selfempl	1	-0.36134	0.05123	-7.05	<.0001	1.12367

On performing regression between dependent variable wage (log term) and; education and education squared term as independent variable, they came out to be statistically significant at 95% confidence interval.

Also, on running regression between significant variable (obtained in question 1-1) including education squared term, education squared term is still significant at 95% confidence interval.

c. Wage (log term) as dependent variable and work hours, work hours squared term as independent variable:
ln(wage) = b0 + b1*hr + b2*(hr*hr)   ln(wage) = b0 + b1*age + b2* edu+b3*hr +b4*(hr*hr)+b5*mar+b6*sal+b7*self

Nonlinear OLS Summary of Residual Errors							
Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq
wage_lm	3	998	480.8	0.4817	0.6941	0.0459	0.0440

Nonlinear OLS Parameter Estimates				
Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
b0	2.555756	0.0847	30.18	<.0001
b1	0.000214	0.000071	3.03	0.0025
b2	-7.91E-8	1.542E-8	-5.13	<.0001

Root MSE	0.61311	R-Square	0.2592
Dependent Mean	2.59390	Adj R-Sq	0.2540
Coeff Var	23.63647		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.06282	0.15818	6.72	<.0001	0
age	1	0.01236	0.00199	6.22	<.0001	1.05892
edu	1	0.06700	0.00713	9.40	<.0001	1.16920
hr	1	0.00010539	0.00006542	1.61	0.1075	7.90794
hr_sq	1	-5.68673E-8	1.396758E-8	-4.07	<.0001	7.56017
married	1	0.13534	0.07268	1.86	0.0629	1.04617
salaried	1	0.27591	0.04406	6.26	<.0001	1.26686
selfempl	1	-0.33233	0.05176	-6.42	<.0001	1.12764

On performing regression between dependent variable wage (log term) and; work hours and work hours squared term as independent variable, they came out to be statistically significant at 95% confidence interval.

But on running regression between significant variable (obtained in question 1-1) including work hours squared term, work hours squared term is still significant at 95% confidence interval.

d. Final regression model with all squared terms:
ln(wage) = b0 + b1*age + b2*(age*age) +b3*edu+b4*(edu*edu)+b5*hr+ b6*(hr*hr)+b7*mar+b8*sal+b9*self

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.08407	0.42883	2.53	0.0116	0
age	1	0.05401	0.01905	2.84	0.0047	101.28698
age_sq	1	-0.00048804	0.00021795	-2.24	0.0254	101.37849
edu	1	-0.09551	0.02710	-3.52	0.0004	17.58912
edu_sq	1	0.00710	0.00115	6.15	<.0001	17.96569
hr	1	0.00013032	0.00006425	2.03	0.0428	7.93800
hr_sq	1	-6.21855E-8	1.371702E-8	-4.53	<.0001	7.58744
married	1	0.11073	0.07135	1.55	0.1210	1.04912
salaried	1	0.22330	0.04396	5.08	<.0001	1.31244
selfempl	1	-0.34777	0.05080	-6.85	<.0001	1.13020

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	145.89886	16.21098	44.88	<.0001
Error	991	357.97867	0.36123		
Corrected Total	1000	503.87752			

Root MSE	0.60102	R-Square	0.2896
Dependent Mean	2.59390	Adj R-Sq	0.2831
Coeff Var	23.17070		

From the above model we can see that coefficient of Dummy for married variable again is not significant at 95% confidence level.

After removing married variable from the final model:
ln(wage) = b0 + b1*age + b2*(age*age) +b3*edu+b4*(edu*edu)+b5*hr+ b6*(hr*hr)+ b7*sal+b8*self

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.15719	0.42654	2.71	0.0068	0
age	1	0.05487	0.01906	2.88	0.0041	101.20213
age_sq	1	-0.00049410	0.00021807	-2.27	0.0237	101.34598
edu	1	-0.09794	0.02707	-3.62	0.0003	17.53029
edu_sq	1	0.00719	0.00115	6.23	<.0001	17.92027
hr	1	0.00014542	0.00006356	2.29	0.0223	7.75598
hr_sq	1	-6.43713E-8	1.36542E-8	-4.71	<.0001	7.50745
salaried	1	0.21886	0.04390	4.99	<.0001	1.30689
selfempl	1	-0.34470	0.05079	-6.79	<.0001	1.12849

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	145.02888	18.12861	50.11	<.0001
Error	992	358.84865	0.36174		
Corrected Total	1000	503.87752			

Root MSE	0.60145	R-Square	0.2878
Dependent Mean	2.59390	Adj R-Sq	0.2821
Coeff Var	23.18714		

Now coefficient of all the variables are significant at 95% confidence level. Hence this is our final model.

2 Using the same model, run fixed effects models and random effects models i.e., FIXEDONE, FIXEDTWO, RANONE, RANTWO.
Create a table of coefficients side-by side with significant coefficients shown in bold (you may do this in Excel).

Parameter Estimates									
		FIXONE		RANONE		FIXTWO		RANTWO	
Variable	DF	Estimate	Pr > |t|	Estimate	Pr > |t|	Estimate	Pr > |t|	Estimate	Pr > |t|
Intercept	1	-0.78993	0.5139	1.322901	0.0299	4.970931	<.0001	1.322944	0.0299
age	1	**0.134429**	0.0111	**0.068983**	0.0102	0	.	**0.068978**	0.0102
age_sq	1	**-0.00141**	0.0211	**-0.00068**	0.0265	**-0.0014**	0.022	**-0.00068**	0.0265
edu	0	0	.	**-0.09432**	0.0291	0	.	**-0.09432**	0.0291
edu_sq	0	0	.	**0.007299**	<.0001	0	.	**0.007299**	<.0001
hr	1	**-0.00041**	<.0001	**-0.00022**	0.0001	**-0.00041**	<.0001	**-0.00022**	0.0001
hr_sq	1	2.17E-08	.	-7.33E-09	.	2.16E-08	.	-7.33E-09	.
salaried	1	**0.125724**	0.0125	**0.18169**	<.0001	**0.127231**	0.0117	**0.1817**	<.0001
selfempl	1	**-0.22899**	0.0005	**-0.27204**	<.0001	**-0.22991**	0.0005	**-0.27205**	<.0001

3

FIXONE - One way fixed effects – cross section (CS) heterogeneity

From the above results we can observe that as in one way fixed effect model we analyze only the cross section to which the observation belong as in our case individuals. The above results shows that education of an individual does not change frequently and during the given years there is no change in education, hence its coefficient is zero.

In random effects model, variance components are calculated, they are used to standardize the data and OLS is performed.

4 3. Write a report on your findings. Interpret model fit, t-values, meaning of coefficients, collinearity diagnostics, White test, Breusch-Pagan test etc.

HAUSMAN Test: To decide which model to use for the given dataset we can check the statistics value from the Hausman Test.

Null Hypothesis: No correlation between error term u_i and X variables

Alternate Hypothesis: Correlation is present.
As the p-value (0.001) is less than alpha (0.05), we reject the Null Hypothesis and conclude that there is no correlation between error terms u_i and X variables, so use Fixed Effects model for the given dataset.

Interpretation of Model Fit:

Fixed One-Way Estimates

Fit Statistics			
SSE	64.3861	DFE	662
MSE	0.0973	Root MSE	0.3119
R-Square	0.8723		

One way random effects

Fit Statistics			
SSE	99.2813	DFE	993
MSE	0.1000	Root MSE	0.3162
R-Square	0.2196		

Two way fixed effects

Fit Statistics			
SSE	64.3632	DFE	661
MSE	0.0974	Root MSE	0.3120
R-Square	0.8723		

Two way random effects

Fit Statistics			
SSE	99.2928	DFE	993
MSE	0.1000	Root MSE	0.3162
R-Square	0.2196		

From the above statistics, we can observe that R-square value for Fixed one way estimates is same as two way fixed estimates. Also, R-square value for Random one way estimates is same as two way random estimates. The main difference between one way and two way is whether the given cross sectional dataset is time invariant or not. Also, from the dataset we can see that it is a balanced panel dataset. So, probably because of less number of observations or less number of observations across time periods, we do not see much difference in the model in-terms of model fit for one way and two way model.
Among the given variables we can see that age and education are quite time invariant, whereas out of other variables only hours worked in a year is time variant, rest two are dummy. So we do not have much variables which vary with time to capture the effect of time invariance in the dataset.

t-values: From the t-values we can see that age squared term, Work hours per year, Dummy for salaried and Dummy for self-employed are significant in all four models, where as age is not significant in fix two model and education, education squared term in fix one and fix two model. Work hours squared term is not significant in any model.

Use correct language (%s) when interpreting effect sizes of log dependent variable.

Meaning of coefficients:
Age: For a unit increase in age, wage will increase by 0.13 units ? as per Fixed effect model and by 0.06 units ? for Random effect model. So people as they grow old will have increase in wages.

Education: Education coefficient is zero in Fixed effect model and wage will decrease by 0.09 units as per Random effect model. The education level of a person does not vary much in 3 years time span.

Work hours per year: For a unit increase in work hours, wage will decrease by 0.00041 units as per Fixed effect model and by 0.09432 units for Random effect model.

Dummy for self-employed: The interpretation of the coefficient of a self-employed is that when a person is self-employed, sales would increase by 0.12 units in Fixed effect model and 0.18 units as per Random effect model.

Dummy for salaried: The interpretation of the coefficient of salaried is that when a person is salaried, sales would decrease by 0.22 units in Fixed effect model and 0.27 units as per Random effect model.

Collinearity diagnostics:

Correlation of Estimates (RANONE)

Correlations of Parameter Estimates									
	Intercept	age	agesq	edu	edusq	hr	hrsq	salaried	selfempl
Intercept	1.00000	-0.88627	0.86347	-0.38176	0.35663	-0.08816	0.08565	0.03782	0.00692
age	-0.88627	1.00000	-0.99377	-0.03088	0.00857	-0.01323	-0.00856	-0.04318	-0.00774
agesq	0.86347	-0.99377	1.00000	0.04004	-0.01736	0.02619	-0.00099	0.04166	0.00558
edu	-0.38176	-0.03088	0.04004	1.00000	-0.96866	-0.03059	0.02872	0.05316	-0.00743
edusq	0.35663	0.00857	-0.01736	-0.96866	1.00000	0.02241	-0.02255	-0.11063	-0.01471
hr	-0.08816	-0.01323	0.02619	-0.03059	0.02241	1.00000	-0.92422	-0.12326	-0.00672
hrsq	0.08565	-0.00856	-0.00099	0.02872	-0.02255	-0.92422	1.00000	0.08207	-0.01666
salaried	0.03782	-0.04318	0.04166	0.05316	-0.11063	-0.12326	0.08207	1.00000	0.12658
selfempl	0.00692	-0.00774	0.00558	-0.00743	-0.01471	-0.00672	-0.01666	0.12658	1.00000

Correlation of Estimates (FIXONE)

Correlations of Parameter Estimates									
	Intercept	age	agesq	edu	edusq	hr	hrsq	salaried	selfempl
Intercept	1.00000	-0.96568	0.89245	.	.	-0.01386	0.03204	0.07426	-0.04029
age	-0.96568	1.00000	-0.97340	.	.	-0.04952	0.01854	-0.07148	0.03825
agesq	0.89245	-0.97340	1.00000	.	.	0.06589	-0.03119	0.06908	-0.03510
edu	.	.	.	.	.	.	.	.	.
edusq	.	.	.	.	.	.	.	.	.
hr	-0.01386	-0.04952	0.06589	.	.	1.00000	-0.92420	-0.08416	0.02122
hrsq	0.03204	0.01854	-0.03119	.	.	-0.92420	1.00000	0.05469	-0.00758
salaried	0.07426	-0.07148	0.06908	.	.	-0.08416	0.05469	1.00000	0.03720
selfempl	-0.04029	0.03825	-0.03510	.	.	0.02122	-0.00758	0.03720	1.00000

The correlation between estimates and their squared terms is high which is expected. Apart from that, no high correlation is observed between the estimates.

Breusch-Pagan test:

Hausman Test for Random Effects

Coefficients	DF	m Value	Pr > m
6	5	46.99	<.0001

Breusch Pagan Test for Random Effects (One Way)

DF	m Value	Pr > m
1	479.32	<.0001

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.322901	0.6084	2.17	0.0299
age	1	0.068983	0.0268	2.57	0.0102
agesq	1	-0.00068	0.000307	-2.22	0.0265
edu	1	-0.09432	0.0432	-2.19	0.0291
edusq	1	0.007299	0.00182	4.00	<.0001
hr	1	-0.00022	0.000056	-3.86	0.0001
hrsq	1	-7.33E-9 (?)	[illegible]	[illegible]	[illegible]

Variance Component Estimates

Variance Component for Cross Sections	0.268172
Variance Component for Time Series	0
Variance Component for Error	0.097372

Hausman Test for Random Effects

Coefficients	DF	m Value	Pr > m
5	4	45.84	<.0001

Breusch Pagan Test for Random Effects (Two Way)

DF	m Value	Pr > m
2	480.60	<.0001

2.5

From the above results we can see that variance of cross section is greater than variance of time series. The BP test checks whether the variance of the errors from a regression is dependent on the values of the independent variables. In that case, heteroskedasticity is present.
As the test statistic has a p-value below threshold (e.g. $p<0.05$) then the null hypothesis of homoskedasticity is rejected. Hence we can conclude with 95% confidence level that heteroskedasticity is present in our model.

4. What is the effect of panel data models on the coefficients? What parameters have changed and by what percentage?

	Fixed Effects		Regression		
Variable	Estimate	Pr > |t|	Estimate	Pr > |t|	%change
Intercept	-0.78993	0.5139	1.15861	0.0067	-168.179111
age	0.134429	0.0111	0.0548	0.0041	145.3083942
age_sq	-0.00141	0.0211	-0.00049	0.0238	185.835918
edu	0	.	-0.09797	0.0003	-100
edu_sq	0	.	0.00719	<.0001	-100
hr	-0.00041	<.0001	0.000145	0.0224	-382.135976
hr_sq	0.00	.	0.00	<.0001	-133.70128
salaried	0.125724	0.0125	0.21895	<.0001	-42.5786709
selfempl	-0.22899	0.0005	-0.34517	<.0001	-33.6587768

1

Comparing the panel data (Fixed Effect) and poolled data results for the given dataset, we can see there is change in parameter coefficients. Coefficient of Age was 0.0041 in regression equation (question 1-2) and has become 0.13 in fixed effect model. It coefficient has increased which will increase in the wage for unit change in the age. Similary the coefficient of age squared term is increased but the other coefficient values is decreased.

5. We are especially interested in the effect of education on wages. Notice how much (%) has this coefficient changed across the different models?

Education Estimate for ln(WAGE) using RANTWO

The estimate of education and lnwage is non-linear in nature. As the variable educ has a significant squared term, we can interpret the relation as:
Delta(ln(WAGE)) = 2*0.007299*EDUC – 0.09432

So,

at EDUC=10, Increase in one year in education increases wage by 5.166%
at EDUC=12, Increase in one year in education increases wage by 8.085%
at EDUC=14, Increase in one year in education increases wage by 11.005%

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.322944	0.6083	2.17	0.0299
age	1	0.068978	0.0268	2.57	0.0102
agesq	1	-0.00068	0.000307	-2.22	0.0265
edu	1	-0.09432	0.0432	-2.19	0.0291
edusq	1	0.007299	0.00182	4.00	<.0001
hr	1	-0.00022	0.000056	-3.85	0.0001
hrsq	1	-7.33E-9	0	.	.
salaried	1	0.1817	0.0424	4.28	<.0001
selfempl	1	-0.27205	0.0523	-5.20	<.0001

Q2. I have provided a dataset PIMS.dat which has data on industrial goods manufacturers. The variables in the data are in the following order. These variables and definitions are given in the paper by Robinson and Fornell (1985) on pioneering advantages (see Tables 1, 2 and 3). As in the paper by Robinson and Fornell (1985), we will estimate a simultaneous system of five equations. While the paper considered consumer goods industries, we are interested in replicating the analysis for industrial goods industries.

Please estimate a 2SLS model with the following five equations.

model MS=qual plb price pion ef phpf plpf psc papc ncomp mktexp
model Qual=price dc pion ef tyrp mktexp pnp
model PLB=dc pion tyrp ef pnp custtyp ncust custsize
model Price=ms qual dc pion ef tyrp mktexp pnp
model DC=ms qual pion ef tyrp penew cap rbvi emprody union

1. Run the 2SLS model using SAS (PROC SYSLIN) and estimate the effect of pioneering on market share. Be sure to consider the direct effects as well as the indirect effects. (read the paper on pioneering advantages for this interpretation).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	42.00303	66.23322	0.63	0.5261	Intercept
qual	1	0.510213	0.123225	4.14	<.0001	qual
plb	1	-1.01115	0.413727	-2.44	0.0147	plb
price	1	0.852267	0.683350	1.25	0.2126	price
pion	1	7.543081	3.538605	2.13	0.0332	pion
tyrp	1	-0.37812	3.225028	-0.12	0.9067	tyrp
ef	1	5.787019	1.562470	3.70	0.0002	ef
phpf	1	0.584949	1.530377	0.38	0.7024	phpf
plpf	1	0.167317	3.955604	0.04	0.9663	plpf
psc	1	-30.8958	12.92588	-2.39	0.0170	psc
papc	1	-1.50612	2.252581	-0.67	0.5039	papc
ncomp	1	-7.54440	0.496102	-15.21	<.0001	ncomp
mktexp	1	-0.28543	0.172875	-1.65	0.0990	mktexp

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	-265.494	63.56925	-4.18	<.0001	Intercept
price	1	2.595316	0.638913	4.06	<.0001	price
dc	1	10.47285	1.958009	5.35	<.0001	dc
pion	1	-0.39839	4.642522	-0.09	0.9316	pion
ef	1	-2.23599	2.142150	-1.04	0.2968	ef
tyrp	1	0.187802	4.337466	0.04	0.9655	tyrp
mktexp	1	-0.48914	0.205769	-2.38	0.0176	mktexp
pnp	1	0.211277	0.061838	3.42	0.0007	pnp

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	109.0706	1.650573	66.08	<.0001	Intercept
dc	1	-8.73302	2.565446	-3.40	0.0007	dc
pion	1	1.715145	1.698292	1.01	0.3127	pion
tyrp	1	-0.29136	1.523126	-0.19	0.8483	tyrp
ef	1	-0.12958	0.772197	-0.17	0.8668	ef
pnp	1	0.054686	0.021919	2.49	0.0127	pnp
custtyp	1	3.940911	1.368804	2.88	0.0041	custtyp
ncust	1	0.225838	0.210697	1.07	0.2840	ncust
custsize	1	0.520397	0.638869	0.81	0.4155	custsize

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	100.3140	0.916136	109.50	<.0001	Intercept
ms	1	-0.01764	0.019380	-0.91	0.3628	ms
qual	1	0.141663	0.035232	4.02	<.0001	qual
dc	1	-0.45759	0.619927	-0.74	0.4606	dc
pion	1	1.661022	1.073669	1.55	0.1221	pion
ef	1	0.070665	0.517670	0.14	0.8914	ef
tyrp	1	-1.42106	0.979397	-1.45	0.1470	tyrp
mktexp	1	0.224952	0.030689	7.33	<.0001	mktexp
pnp	1	-0.02127	0.016541	-1.29	0.1988	pnp

The below table shows the 2SLS estimations for five models. Estimations for pion variable show the pioneering effect on market share. Pioneering has direct effect on market share. In models where dependent variable is not market share, the coefficient of "pion" are insignificant.

Dependent Variable	**ms**	**qual**	**plb**	**price**	**dc**
Intercept	42.00303	-265.494	109.0706	100.314	1.140754
ms				-0.01764	0.004963
qual	**0.510213**			**0.141663**	**0.035383**
price	0.852267	**2.595316**			
plb	**-1.01115**				

dc		**10.47285**	**-8.73302**	-0.45759	
pion	**7.543081**	-0.39839	1.715145	1.661022	-0.07697
ef	5.787019	-2.23599	-0.12958	0.070665	0.140887
phpf	0.584949				
plpf	0.167317				
psc	-30.8958				
papc	-1.50612				
ncomp	-7.5444				
mktexp	-0.28543	**-0.48914**		0.224952	
tyrp	-0.37812	0.187802	-0.29136	**-1.42106**	0.221522
pnp		**0.211277**	**0.054686**	-0.02127	
custtyp			**3.940911**		
ncust			0.225838		
custsize			0.520397		
penew					-0.0034
cap					0.000041
rbvi					-0.04885
emprody					**0.002523**
union					0.00146

2.5

Endogenous Variables: MS, QUAL, PLB, PRICE, DC; hence 5 simultaneous equations.
Instrument Variables: PION, EF, PHPF, PLPF, PSC, PAPC, NCOMP, MKTEXP, TYRP, PNP, CUSTTYP, NCUST,

Product quality, Product Line Breadth, If a business is a market pioneer, Whether a firm is an early follower, Pioneer is selling seasonally changed goods/inventory, and Number of competitors are significant (i.e. |t-value|>1.96 and p-value<0.05) at 95% confidence level.

Direct Effect: Pioneering is a significant variable. (t value = 2.13 and p-value = 0.0332) so there is an effect of pioneering on market share. If the firm is a pioneer, the market share will increase by 7.54 units. Also, if a firm is a pioneer and sells goods that are changed seasonally then the market share will decrease by 30.89 units.

Indirect Effect: (where pion = 1)
PION-QUAL-MS: -0.39839* 0.510213= -0.20
PION-PLB-MS: 1.715145*(-1.01115) = -1.73
PION-PRICE-MS: 1.661022*0.852267 = 1.41
PION-DC-PLB-MS: -0.07697*(-8.73302)*(-1.01115)= -0.68
PION-DC-PRICE-MS: -0.07697*(-0.45759)*(0.852267)= 0.03
PION-DC-QUAL-MS: -0.07697*(10.47285)*(0.510213)= -0.41
Total indirect effect is -1.58
Total Effects = Direct Effect + Indirect effect = 5.9

DC?
DC is not in MS equation!
Interactions?

2. **Run a simple regression model of market share as given in the first equation. What is the effect of pioneering on market share using this simple model? How does this effect change across different models.**

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	47.42575	7.94019	5.97	<.0001
qual	qual	1	0.16785	0.01669	10.06	<.0001
plb	plb	1	-0.48829	0.06033	-8.09	<.0001
price	price	1	0.33321	0.07842	4.25	<.0001
pion	pion	1	12.60256	2.58479	4.88	<.0001
tyrp	tyrp	1	-2.92100	2.40991	-1.21	0.2257
ef	ef	1	4.96526	1.21261	4.09	<.0001
phpf	phpf	1	1.61160	1.20764	1.33	0.1823
plpf	plpf	1	1.14544	2.71817	0.42	0.6735
psc	psc	1	-20.77004	9.99043	-2.08	0.0378
papc	papc	1	-1.20403	1.55664	-0.77	0.4394
ncomp	ncomp	1	-7.48754	0.37961	-19.72	<.0001
mktexp	mktexp	1	-0.10667	0.07495	-1.42	0.1550

We observed following differences between regression and 2sls model-

- Pioneering is significant in both the models, but its coefficient is high in simple regression (12.6) as opposed to 2SLS (7.54 direct and 5.9 indirect). The reason is- there is endogeneity due to simultaneity occurring in MS, QUAL, PLB, PRICE, DC. That is why simple regression is violating the unbiased assumption. The solution is to use 2SLS model instead of simple regression which removes endogeneity issue.
- **Both models have common significant coefficients except price. Price** is significant in the simple regression model since it is correlated with Quality, however, Price is insignificant in 2SLS model. This endogeneity is removed by 2SLS and so it becomes insignificant.
- Considering 95% significance level, Product Quality, Product Line Breadth, Whether firm is an early follower, Pioneer Seasonal Product Change, Number of competitors are significant in both the models but have higher coefficients in simple regression that confirms endogeneity as variables are showing effects of relation with each other. The remaining variables remain insignificant in both 2SLS and simple regression model.

2 SLS model has resolved endogeneity and unobserved heterogeneity and is giving unbiased results.

4