**Homework assignment 4**

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 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.
2. Develop a model to test if there are nonlinear effects for some variables.
3. 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).

1. Write a report on your findings. Interpret model fit, t-values, meaning of coefficients, collinearity diagnostics, White test, Breusch-Pagan test etc.
2. What is the effect of panel data models on the coefficients. What parameters have changed and by what percentage?
3. We are especially interested in the effect of education on wages. Notice how much (%) has this coefficient changed across the different models?

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.

|  |  |
| --- | --- |
| MS | Relative market share |
| QUAL | Relative quality |
| PRICE | Relative price |
| PLB | Product line width |
| DC | Relative direct costs |
| PION | Whether a firm is a pioneer (1) or not (0) |
| EF | Whether a firm is an early follower (1) or not (0) |
| PHPF | Pioneer high purchase frequency |
| PLPF | Pioneer low purchase frequency |
| PSC | Pioneer seasonal product change |
| PAPC | Pioneer annual/periodic product change |
| NCOMP | Number of competitors |
| MKTEXP | Relative marketing expenditures (similar to ‘relative advertising and promotion’) |
| TYRP | Twenty year pioneer |
| PNP | Percentage of new products |
| CUSTTYP | Relative customer type |
| NCUST | Relative Number of customers |
| CUSTSIZE | Relative customer size |
| PENEW | Plant and equipment newness |
| CAP | Capacity utilization |
| RBVI | Relative backward vertical integration |
| EMPRODY | Employee productivity |
| UNION | Percentage of employees unionized |

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

model MS=qual plb price pion tyrp 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).
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.

**Q3. Use the churn data to understand what factors are good predictors of churn (i.e. a customer leaving the firm).**

**Churn is the dependent variable that takes the value 1 if a customer has left (cancelled account) and 0 otherwise. There are 100,000 observations in the data. Take a random sample of 50% of the data using PROC SURVEYSELECT. Use one half of the sample for estimation of parameters (training sample) and the other half for prediction (test sample).**

**We need to use a logistic regression model, since Y is a binary discrete variable.**

**Use the following website as a reference for doing the exercise:**

[**http://www.ats.ucla.edu/stat/r/dae/logit.htm**](http://www.ats.ucla.edu/stat/r/dae/logit.htm)

**Make sure that there are no two explanatory variables that are highly correlated. Use correlation analysis to determine the correlation between the variables.**

**Write a report explaining what are the top 5 factors that predict churn. Comment on model fit, interpret the effect of a continuous variable and a discrete variable (at the least), interpret odds ratios, percentage concordant etc.**

**Predict whether a customer will churn in the test sample and report how accurate is your prediction in terms of percentage of accurate predictions (e.g. 70%).**