Data analysis course of professor Murthi @ UTD: first session

Meisam Hejazinia

1/14/2013

Professor B.P.S. Murthi empirical research, 19 years. Empirical tools for using to used in data analysis for the summer paper. Lots of practice, homework and mistakes. Homework are important since you learn from them. There are many errors that you do while practice, and you will learn from.

Book is handbook of statistical analysis.

After first two classes it would be more about applied multivariate techniques, and econometric models and econometric forecasts. Refer to them as necessary.

Each homework will have the real databsed. You will try to reach your own conclusion from.

%40 is homework, and Quizzes have %20, and final %30 and class participation is %10.

What are interpretation of each output, and why other methods can not be used are things you will learn.

If you through the way the seed then there would be no apples, but you can narture it to get as many apples you want for the seed. This class also gives you the seed, and you need to do the excercise, and you can watch thousands of times excercises and that would not help. Exercising in really necessary.

Next session will be held in the lab, next to elevator.

F3 or run option will help you run.

The number of observations will be shown in the log window, and if it did not match it, you need to find the problem with it.

f5,f6, f7, editor, log, output helps you to navigate between windows.

www.ats.ucla.edustatsasdefault.htm is a fari amount of useful informations. It will tell you waht each these tests means.

If things were not covered, it will tell you what these things mean.

You can purchase utaustin to purchase, and you need $105 license for one year.

Why learn SAS?

Data available, and you can get job easily as database analyst.

Combination with MBA makes it powerful job candidate as backup plan.

Grocery scanner data, internet transaction and clickstream, and stock transactions data. For a long time more than %50 of market share was of SAS.

Stata are good for time series. Finance people usually use it.

R needs more programming than this. It is fully programming like fortran.

If new model cames that you have not used it, you can use Gauss, but you can use MATLAB and R.

Every SAS statement must end with semicolon.

Old version need eight characters.

For missing numerical data, use a single period of dot (.)

Two type of datasets:
temporary: when you close the program it is gone
permonent dataset: stay in the system for future

```
Data a1;
Input age income score;
Cards;
10 20 30
40 50 60
70 80 90
;
Run;
```

when you want to make sure that when there is fixed length you read it, and you do not use space.

```
Data bb;
infile "c:.data";
input name $ 1-9 var2 10-15 var 22-26;
run;
```

$ is used for the characters.

These are called format statements.

Input name $ 9. var2 6. @22 var3 5.2;
you must put dot after 6 in 6., since it is the convention

for 126.64 you need to write 6.2, so you need to count dot as well, and the first part is the total number of characters, and the second part is the number of floats.

for 0.1634 would be 6.4
for 110.1634 would be 8.4

@ will tell to jump to which location.

This is a short way of saying:
Input name $ 1-9 (var2 var3) (6. 5.2);

you can summerize by saying (var1 - var40)

You can use compound statements such that (39*2. 6.2) saying that the first 39 is in the form of 2. and the last one is in the form of 6.2.

The default is numeric.

when you have a separator, for comma seperated variables, you will have the following format:

```
Data a1;
infline "c:.csv" DLM=',';
input name $ var2 var 3;
run;
```

Over real data you need to be careful, since data may have comma, and lot of preprocessing would be needed.

To skip data you can use firstobs number,

Infile "c:.csv" DLM=',' Firstobs=4 missover;

missover says don't worry about the space, and if there is no dots it will go over the next line.

Missover is very useful. It is only useful for the blanks at the end not at the beggining.

To create permanent SAS dataset you will need LIBNAME :

```
libname cc 'c:';
data cc.billion;
Infile ....;
input ...;
run;
```

read data of sas into temporary
libname q2 'c:'; data a1;
set q2.billion;
run;

creating new data in the text file

Data a2;
set a1;
file 'c:.dat';
put name 1-9 var3 12-16;
run;

Using if then statement for creating some parts:

data a1; set q4;
if cost="." then cost='NA';
if sex="F" then delete;
if income="Low";
if type = 'tragedy' or type='comedy' or type='history' then keep;
if service='high' then ds1=1; else ds1=0;
if service='medium' then ds2=1; else ds2=0;
run;
qoute " would be case sensitive, but when you use "" then both will be taken into account.

to keep you do not need the if command, and only for deleting you need to use the if statement.

You can highlight portion of the code and run it.

you can pring part of dataset by:

proc print data=a1 (obs=20); run;

Sample excercise:

data a1;
infile 'path';
input city $ var1-var7; run;

problem is that there is space between the name of the variables:

To resolve this you can put the number of characters:
input city $ 15. var1-var7; run;

These are practical problems that can create problem for you.

If the format would not be the same, then you would have problem.

Always test to make sure that data is read properly.

Proc mean gives you mean std, minimum and maximum:

proc means; var var1-var7; run;

plotting graphs and charts:

proc gplot data=bb;
plot day* rain /haxis = 'mon' 'tues' ... 'fri' vaxis= 10 to 100 by 10;

The first one is vertical and second one is horizontal.

Overlay is to put multiple on the same sheet:

plot therapy*month trtests*month/ overlay;

use options statesment:

options obs=10;

You can use it before the data, and all the statements will work on the first 10 observations.

Just to test before run over all database you can use this.

Don't forget to modify it for real check after the test.

You should comment yourself for the code:

* this program reads the dataset "billion" and does a regression using sales as dependent variable and price as the independent variable N=576*

Make it clear by them.

You can put title and footnote for results:

```
title 'regression analysis';
footnote ' for data analysis course';
footnote2 'summer 1999';
```

For labling the output:

```
label mpg = 'miles per gallon';
```

helps to make the output look nicer;

```
proc sort data= a1 out=b1; by age;
run;
proc sort data a1; by price descending mpg; run;
```

descending should proceed the variable that you want to sort

print only fiew variables:

```
proc print data=bb; var mpg midprice; by manufact; run;
```

This will print the data that is sorted.

```
proc means data=a1 options
```

default is n, mean, std, dev, min, max, but you can ask for nmiss, range, sum, var, etc.

```
proc sort; by domestic; sort by doestic
proc means; var mpg; by domestic;
or
proc means; var mpg; class domestic;
```

using by you need to sort it, but when you use class it will sort by itself.

```
proc means; var mpg midprice; class domestic;
output out=stats mean=mmpg mpr std=smpg;
```

stats in above command is temporary sas dataset.

Calculate frequency is:

```
proc freq; table domestic; run;
proc freq; table domestic*stkshftchisq out=newdata;
run;
```

computer correlations:

```
proc corr;
var ...; run;
```

```
proc reg;
model y= x1 x2 x3;
run;
```

merging data:

```
data new;
set a1 a2 a3; run;
```

this will put one on the top of the other. This will be stacking form.

interleaving form would be in the form of:

```
pproc sort data a1; by name year;
proc sort data=a2; by name year;
data new;
set a1 a2;
by id;
run;
```

It is like stacking and then sorting.

One-to-one match merge:

```
Data new;
merge a1 a2;
by id; run;
```

It needs sorting since it does horizonal concatination.

This code assumes there is same number of observation on both sides.

proc sort data=a1; by id;
proc sort data=a2; by id;
data new;
merge a1 (in=aa) a2; if aa; by id; run;

This merges one to many. if two matches mean they are equal to 1.

This will work for two side merging:

merge a1 (in=aa) a2 (in=bb); if aa and bb; by id; run;

Go home read this comments, and next class you will have the data and you need to practice.

Suppose that the store managers asks you to check whether sales is greater than 100.

We always reject the null hypothesis. What you need to show is alternative hypothesis. Opposite of that we reject.

Null: $\mu_s \leq 100$
Alternative: $\mu_s > 100$

We use t-test or z-test.

z-test is for large samples $n \geq 30$
t-test is for samll sample $n < 30$

you check the mean of two population $\mu_f < \mu_m$

we want to reject %95 to reject this means that each tail should be 2.5%.

When you are testing on the two side mean you are checking the equality you use the two tail test. such mas $\mu_1 = k$ and $\mu_1 = \mu_2$

When inequality is checked one tail test is checked $\mu \leq 100$

%95 is the confidence interval

It is the commonly agreed benchmark. It is the default for papers.

two tail is more conservative test, since benchmark has higher threshold.

why we use standard error?

we get the sample with mean and standard error. If we use this a lot of time then we will get different values, and the mean of this would be population.

This is called population mean which is μ. Standard deviation of all the samples would be called the standard error which is $\sqrt{\frac{\sigma^2}{N}}$.

This is the central limit theorem, saying that no matter what is the distribution of populationt he sample mean will have normal distribution.

This is why we are using the normal distribution for our test.

standard error as a result is standard deviation of the means.

χ^2 is used two by two, saying that man and women are more likely to buy or not buy.

The null hypothesis is that there is no relationship between gender and purchase.

The alternative hypothesis is there is relationship between gender and purchas.

Correlation between two discrete variable could be checked with the discrete variable.

Correlaiton is used for contineus variables.

Data analysis course of professor Murthi @ UTD: first session

Meisam Hejazinia

1/23/2013

Class held in the lab.

SMVA01SAS for server.

F-test chesk the variance of two populations.

The variance of these two are unequql.

We should use the second one when the null is rejected. ttest test of variance here is correct; therefore we need to use Satter thwaits one.

In case t-test was not rejected then you need to use the pooled. Here we had 0.12 that is not lower than 0.5 we are not able to reject that they are the same. The conclusion is that they means are equal, and we can not say that they are not different.

Confidence is 1-"p-value"; here with 88% we can say that mean are different.

expand tab option should be used on infile in order to be able to not mistake tab with space:
infile "C:109420.SASFirstSession.txt" expandtabs firstobs=22;

proc univariate is similar to proc mean, but gives you more data such as quantiles.

proc contents could be used to know what is the content of the file.

The other way is to use proc mean.

In the following data: classT Method Mean 95% CL Mean Std Dev 95% CL Std Dev

Da 3.9701 3.8476 4.0926 0.5432 0.4693 0.6450
EV 3.7135 3.4633 3.9637 0.6940 0.5564 0.9227
Diff (1-2) Pooled 0.2566 0.0109 0.5023 0.5905 0.5211 0.6812
Diff (1-2) Satterthwaite 0.2566 -0.0195 0.5326

Method Variances DF t Value Pr ¿ —t—

Pooled Equal 108 2.07 0.0408
Satterthwaite Unequal 47.339 1.87 0.0677

Equality of Variances

Method Num DF Den DF F Value Pr ¿ F

Folded F 31 77 1.63 0.0865

We could not say that they are not different, first you look at the f-test and here it tells you to use pooled method result, and pooled method tells you that the difference is significant for the mean, yet not significant for variance due to f-test result

Data analysis course of professor Murthi @ UTD: first session

Meisam Hejazinia

1/30/2013

SAS practice is sample of the exam that you need to know this simple things. Broad stuff you need to be able to do without referring to the database. Do SAS practice for next session, then you will be asked next session to answer by the board.

Linesize=80 means 80 character per line.

option obs=10 means that the program work only in limited to the couple of observation before submit to the whole database.

Make sure you do not forget to remove it for the full dataset.

Center is for centering the output. Linesize is for controlling max length of output.

Long records up to 120 characters. Suppose you have 80 rows, mean two 40 rows.

0...80 and each will go the next line.

the screen would be 240 columns, then you need to use the option of $LRECL = 200$ needs to be put in the INFILE command. It is actually when the distance between them are $1 - 6$ ... in the input statement.

Label *mpg* ='Mile per Gallon';

proc sort data=a1 out=b1; puts output in p1

PROC SORT DATA=a1; by midprice DESCENDING mpg; In this case the car with higher mile per gallon would be higher.

PROC PRINT LABEL; VAR mpg midprice;BY manufact; run; You will use the LABEL option to format that.

PROC PRINT; TITLE 'Sum of Flower Data by Month'; FORMAT mp ms mm 3.; Means in three columns.

in PROC MEAN you can ask for NMISS, RANGE, SUM, VAR. the options should be put after the data statement. PROC MEANS DATA'='a1 options;

options referred to the specific workds, and you need to replace these words with word options above.

other options could be $alpha = value$ for confidence level. KURTOSIS, MODE, CV, STD ERROR and LCLM could also be used.

Go to help, index, and check what they mean.

output out=flowers mean(petunia snapdrag marigold)=mp ms mm; mean creating new dataset call flowers.

Every procedure has an option to output the file.

For chisquare

PROC FREQ; TABLE $domestic * stkshft/$ CHISQ OUT=newdata;

The out option is when you want to save it and

merge it in future.

it will test whether average income of region 1 and region 2 would be different:

Proc TTEST; class region; var income; run;

ttest only compares two regions.

If you want to do multiple, you need to separate it, and do two by two. You need to create dummy variables for this.

If you have four regions, you need to do ftest.

proc reg;

'outest' option means put the result of test in file.

'Noprint' means to suppress printed output.

'noint' will not put the intercept.

'STB' compares the standard regression coefficient.

'VIF' and 'TOL' could be used together, and they are redundant.

'PCORR1' will give you the predicted values.

Test $A1 + A2 = 1$ or $A1 - A2 = 1$ is testing the hypothesis. Checking whether effects are the same for both coke and pepsi for example.

'R' will give you residuals.

Output $Out = SASdataset$ $P = names$ $R = names$; can save residuals and predicted values in the separate file.

You can run regression for each, or do the regression only once, but use 'By' option will only be run against the categories.

for example 'By Category', can run the regression for each of them, and calculate the elasticities on one file, and do the regression and save outputs, and then use it.

Suppose your variable is Sales, and you are regressing on price, deal and so on. Deal is 0 when there is no promotion, and 1 means there is a promotion for that product. It is the test tha $H0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_6 = 0$ means all the coefficients are zero is the main theme that we are testing.

For equality of more than two groups is F-test. When we have more than two variables we must have F-test. In this context this is the thefinition. The alternative explanation is H_a whether at least one β is not zero.

Intercept is the constant that we put. Slope is β and intercept is fixed effect. Intercept is nuisance parameter, and captures all the noise, not explained by $X_1..X_6$.

Effect of all other factors except this (average of them) is captured in the intercept.

R^2 is the percentage of variance explained by X_1 to X_6.

R^2 could be between zero and one, and zero means there is no explanatory.

There is no answer for good number, but with real life data, $0.2 - 0.4$ is acceptable. This means commercial data.

For experimental data, we have lot of data, and in every aspect are similar, then you want to get $0.8 - 0.9$. When ther is not so much variation in the data, then you want to get high R^2.

For most of search publication $0.2 - 0.4$ is good.

$sales_t = sales_{t-1}$, so $R^2 = 0.5$

In time series R^2 should be high, since over time the variation is not high.

Sumsquare is proportion of variance, and we have sum of squares of error, and sum of squares for variation.

$R^2 = SSR/SST$ where SST would be sum of SSR (Sum of square of Regression), and SSE (Sum of square of Error).

Customer satisfaction when we have service, product, completion of service ..., and these are noise.

R^2 can be increased by adding variables. Suppose there are 3 observations. $sales = a + b.price$.

If we add more variables, there would be no degree of freedom, and R^2 will reach to one. Means your degree of freedom increases to the point that $R^2 = 1$.

Parameter (p) relative to the number of observations will talk about the adjusted R^2. Means adding new variable will work as a penalty if it does not explain enough variance.

$t - value$ Null hypothesis is that $\beta_p = 0$ versus $\beta_p \neq 0$. p-value is $1 - confidence$ level.

Usually t-value should be 0.05.

t-value and p-value contain the same information. If it is greater it means H0 would not be rejected, if lower means it is rejected.

Each hypothesis is different between β and 0.

Yout take sample of two people (different people), and you calculate μ, and yout ake many two samples of two, and their standard deviation would be for example 2, 16 and so on for each of the samples s_1, s_2 , ..., S_1000, each with size two, then the population means would be average of means of all 1000 samples mean of size two, and standard error would be Standard Error = $\frac{\sigma_{sample}}{\sqrt{n}}$. The distribution of means of samples of any size therefore would converge to the population mean. n in above would be size of the samples.

$t = \frac{coeff}{stderror}$

$p - value = 1 - confid.level$

$sales = a + bp + c.Deal$

$b = -3$ means if price is increased by one unit, sales decreased by 3 units, all else sonstant.

If deal is one or zero, and the coefficient is $c = 7$, what does c means. This means if there is a deal, sales will increase by 7 units, then the sales will increase by one, all else would be constant. Dummy variable. Means $S_D - S_{ND} = 7$ units, and it means compare in the group, between two states.

t-distribution tells us that $t > 1.96$ means more than 95% confidence interval means 0.05 p-value.

Suppose, sales lower the price by 10%. If you lower your price sales will go up. Then you will put the deal and sales goes up by 20%, but no price change. Now when we do both, mean price would become lower, and would be deal, you will see 60%, and this means interaction or synnergy. They get angry, do bad things. They get drunk, they do bad things, and when they become angry and drunk they do much worse things. Interaction could be positive or negative. If it is positive it is called synergy, and if it would be negative it would be cancellation.

$sales = a + bp + c.Deal + d(p * Deal)$, when we put the multiplication here, the joint effect if gives you more than sum of the individual levels, then you will have the interaction.

As price increases sales goes on, if there would be no display. When there is a promotion (display), and the new line of price sales would not be parallel, then at different level of price you will have different level of sales. When you see converging, and diverging lines, mean there is no interaction (that means they are not parallel).

Price has negative effect, means lower price is good, deal is good for sales, and for synergy you should see minus for synergy.

This means people will become more price sensitive.

It depends on how you code the veriables.

In interpretation you need to be careful. You can fix things and say what would be good, means increasing, and multiplication of the signs will define it. In real life you don't know what is happening, you should do all this test, and find out whether the interaction works or not. This interaction could affect many things, and when you put sometimes things go off. This means the effect is not direct.

You can create new paper out of previous, because you can show that the interaction was the reason and not the main. Umbrella do not cause rain. People who eat breakfast lose weight. Strong correlation. Does it make sense? Suggestion could be active people get hungery, and active people lose weight.

Interaction helps you to give you the mechanism. Interaction will capture the moderating effect.

Causality could be showed only when you capture everything. In regression you capture correlation.

$procGLM$;

$ModelY = A|B|C$;

In one shot it will capture all the A, B, C, AB, AC, BC, ABC, yet this is overfitting. You must have the hypothesis, and why it makes sense. The right way is to think about it first, rather than writing small things, since anything you put in will affect the correlation of others.

Blind model is efficient, and you can make up the story after, but it is not recommended, and it is better to start from the hypothesis.

Which factor affects the sales more could only be found out based on scale of the variables. For example suppose the price is \$3, and the other did in cents \$300. When you scale coefficients change. Therefore you need to take care of the scales. Always look at the standardized beta correlation. It takes the values in the form of normal normal. This is called the scaling. When the request is to compare the coefficient, you need to use STB that standardizes the variables. You can multiply by mean and compare, yet it is less perfect and crude, since it does not take account of for the variance.

Suppose there is gender 0/1 and the cofficient is 0.7, and old versus young 0/1 and the coefficient is 0.6, we can not say aything since a mean is different, distribution of maile and female could be different. This could mean that 0.1 of population are femail, yet the number of old and young are the same. Mean for the categorical variable mean, which category is presented here more.

How to calculate the price elasticity?

$$P.E = \frac{\%chngS}{\%chngp} = \frac{\frac{\delta s}{s}}{\frac{\delta p}{p}} = \beta.\frac{p}{S} = (-3.\frac{\bar{P}}{\bar{S}})$$

This helps as point estimate to get the average of price elasticity. It is an estimate.

In case we have interaction we would have $\frac{ds}{dp} = (-3 - 2\bar{AD})$

Advertising and sales have saturation effect. It will not keep going up. You can put power to capture this saturation effect. You can put ${}_1.Ad + \beta_2.Ad^2$. Signs of this coefficient could be positive and negative. You can add higher powers as well, because it would be fourier. You need to look at the adjusted R^2, and the $t - values$ to understand when to stop.

Coup douglas $Y = aX_1^bX_2^c$ could be linearized in the form of $log(y) = log(a) + b.log(x) + c.log(x_2)$. In log-log b would be approximation of elasticity.

This is different math to create simple regression from different functional forms.

Prediction

$\hat{y} = a + bx_1 + cx_2$. Prediction within data is okay, but outside the data would be risky action.

If c is not significant, you should use it, because the coefficient will be affected when you use this model.

You can resubmit revised version. You can email it.

Last week there would be another homework.

Data analysis course of professor Murthi @ UTD: Fourth session

Meisam Hejazinia

02/06/2013

Regression hand out is finished.

We are doing SAS practice hand out.

Data A1;

Infile 'c:1.Dat';

input Fund $1-35 comp$ 36-70 NAV 71-77 Expart 5.2 ..;

Dollar sign is alphabet, and understanding the format is also important. If long you need to use:

Infile 'c:1.Dat' LRECL=200;

In this particular example you don't need it.

DATE has lot of informats, and it is dependent upon how it is written.

18/11/1958, means it is DD MM YYYY. The format then will say 'DDMMYYYY10.'

For time series you need to read data, and it is very useful, and you should pick it up by yourself.

It could be 'DDMMYY8.' time is also important.

It does not store the data like 18/11/58, but it stores in the form of # days form 1.1.60.

The net would be like $8 years*365+11*30+17 = ...$, means it is recoreded as contineous number of days.

These are called informat statements.

SAS knows how many space it will need, so we say 'DDMMYY', or 'MMDDYY10.'.

Given any day you can always calculate the day

Data B1;
SET A2;

IF Ticker=$'.'$ then delete;

It is not case sensitive.

PROC SORT DATA=B1 B2; BY TICKER;

DATA B3; Merge B1($IN = A$) B2($IN = B$);
IF A AND B; // If data exists in A and B; make sure that it exists in both database 'JOHN'
BY TICKER;
RUN;

$IN = A$ All missing datas will be kept out using this variable.

$REFDATA = DATE(7-99)MMYY4$. You put this in the data statement.

$TIME = intck('month', DATE, '01July1999'd)$;

Calculate the sum of assets for the whole group, and for each you calculate that and divide by month.

There are three objective categories for example

PROC MEANS;
VAR TASSETS;
OUTPUT $OUT = C1$ $SUM = S1$;
BY OBJ;
RUN;

The output would be objective $1, 2,$, and sume in separate variable called s1.

Dependeing upon how many variable you will ask, you will put it in the file.

You merge with original database, and then after matching by objective and then devide by themselves.

Professor is looking for the logic, and to make sure you get the logic, and the statement is not important.

PROC FREQ; TABLE OBJ; RUN;

Does the calculation on the number of observation.

For regression do the PROC REG; MODEL $Y = X1$ ...;

DATA CIO; SETCQ; IF ... For deleting the data with lower observations;

To run the regression multiple times rather than run multiple tyme write:

PROC REG; MODEL $y = X1$ ... ; BY OBJ;

In scanner data, we have household data, means you have household ID in the form of 1, 2, ..., and then you have for each weeks 1, 2, ..., multiple times.

The brand that they bought would be there for each time. A, B, A, A, A would be there. P_A, P_B, P_C, P_D would also be there. We would also have display for each brand. Display would be $D_A D_B D_C D_D$, mean the special flag that says item is on special price. Display would be like a flag saying what is special for today.

You would also have feature add. $F_A F_B F_C F_D$ would be like a newspaper advertising. It would be newspaper flyer.

Normally flaggs are in the form of 0, and 1. Either they are zero or one in a given week.

Even when there wasy feature add or display of other brand, the customer has purchased another brand.

What are uses of scanner data? How display, price, feature will affect the brand chosen?

You can find out how price sensitive each household, and each segment, or entire market.

Interaction effects of price display and features.

Seasonality effects.

When should be the time to promote.

Effect of competition.

Cross price elasticity.

Own price elasticity.

Every grocery store has these data, and you can do these analysis using this.

Fast moving items, and slow moving items for inventory policies.

Which brand are moving and what are not.

Which brand of Shampoo I should keep more, fast moving and slow moving items.

Effect of demographics.

Number of times they purchase and which brand they used.

Brand loyalty is also the thing that could be taken.

Typical random sample who they volunteer to give the data, and they did a survey, and they checked whether they have car or house. Use this data you can do analysis.

PROC MEANS and do SUM;

or you can do PROC FREQ and get COUNT;

to find how many times the brand is on display.

For part 2:

If $D_A = 1$ or F_A THEN $PROMO = 1$; else $PROMO1 = 0$;

You can do that for B and C. You need to show that you understood it.

You can create another database where $DISPLAY = 1$ and then do 'PROC MEAN'

The simplest is :

PROC MEAN; VAR p_A; By D_A; as long as the answer works it is fine.

For each person we want to know what is the frequency.

PROC FREQ; TABLE STORE; BY HHID;

Simplest is the way, so if it would be one line of code then it is better.

$log(sales) = a + b.log(pr) + c.log(ad)$

No need to do the interaction on this. It is about what does the intraction mean, and since it is in the form of multiplication. As a result this module has the interaction inside of it. Coup douglas has its own multiplicative.

When we do prediction consider non significant beta and we should not ignore non significant, and you should always consider it.

The other thing is 'INFLUENCE'. Suppose you have couple of observations. The fact that there is point there A or B influences the slope of the line. How much they would have influence is called influence. Influence talks about 'outliners'. You would need to compare $DFFITS > 2\sqrt{\frac{p}{n}}$. This will tell you what is the difference with respect to R-square, and you compare with r-square. Taking this how what would be different by fit. How much is R-square different. Is is large or smaller.

DF-BETAS is difference with betas. You compare this with $\frac{2}{\sqrt{n}}$. If you believe that those points do not exist. It happens so infrequently, then deleting can make sense. This will output the influentials. The answer is not that you have to delete it, but the question is whether it appears significantly in the data.

p is number of parameters. and n would be number of observations.

You must look at the absolute value, since it could be negative.

Outlier could be due to error, and it is keystore, and such value could not exist.

Multicollinearity:

You can call it collinearity or multicollinearity. It is $y = a + bx_1 + cx_2$ If the correlation between x_1, x_2 would be high the estimate would be biased. Question is that correlation is high. Suppose the effect of x_1 on y is 10. Because we have two x_1 we do not know where to put a, and all are true. You can not trust beta in this case.

Hight and weight are normally highly correlated. Income and education is also high. These things instigate the doubt that there is problem of mul-

ticollinearity. The variance of beta become large. $t - value$ in this case will go down, means that it has low effect. It may have an effect, yet you are assuming that it is not.

Another thing is linear dependency. Mean $x_1 + x_2 - x_3 = 0$. In combination there would be linear dependency. Most obvious reason is correlation. They are somehow connected in more than 2 at the time. Since, we do not see this, we would not be able to recognize.

To detect collinearity you need to 1. check correlation 2. check $VIF > 10$, and 3. collin that condition index > 10, and proportion of variance for two x's should be high more than 0.9. You should look at the proportion of variance in addition to checking that condition index > 10.

What is the solution? Delete one of the variables. If height and weight are correlated drop each one at the time and do two regression, check which one is more helpful.

Typically in surveys you measure satsifaction using Q_1, and Q_2, and I feel good about myself, so about everybody, halo effect $Q_1 + Q_2$, mean I feel good in one area, so I feel good in another. There is also something called principle component. This is also one way of removing collinearity. The last method is Ridge regression is also another way. It is part of advanced econometrics. The other methods was based on eign values. Eigen values are number of variance, and these joint variance are explained with this. You go along the rwo to the proportion of variance and you will find out that these three variables are correlated. If you want to check in your paper for collinearity you must check both check VIF and COLLIN.

Principle analysis is variation of factor analysis.

Always check for multicollinearity, and write a statement telling that there is no multicollinearity in the data.

Once you do that, you will know that there is no bias. The other problem is: one of the assumptions is that the variance is constant across the observations.

We assume that the variance is constant, what if it is not?

For exampe regression on sales of large companies and small companies. Then the mean μ would be high to low. The variance of large companies in terms of sales would be higher. Suppose you did the sample of all airline firms. It would be in the same dataset. The variance is not constant. This means we do not have homdiasticity, and we would have problem of heteroscedadsticity.

Suppose you are doing survey on satisfaction, and all your questions are one to five. Across the students, since the scale is fixed it would not be a problem, yet for other data scales are not fixed. What if the variance is not constant: heteroscedasticity. One solution could be transform it, and by taking the $\sqrt{(}y)$ up to the point it could be used. Transformation of $ln(y)$, and $\frac{1}{y}$ the problem will be reduced. If the variance is decreasing with variable then you can use y, y^2, e^y.

This is easy solution. By looking at the plot you can not see whether it is shrinking or not.

We need a proper test to see whether it is true or not. There are a lot of test. 1. white test 2. Gold fild Quandt (GQ test). We go to these tests to see whether there is heteroscedasticity.

Suppose you have a variable that are either buy or do not buy. This is sure sign that there would be heteroscedasticity, since these variance are not constant.

Logit, probit, discriminant analysis. These are the reason that we are using heteroscedasticity. These are why we use other methods than regression.

Three desirable properties of estimators:

Every possible methods, we have thousands of methods, are:

1. Unbiasedness: $E(\hat{\beta}) = \beta$. As long as the method guids us through that you would be happy. Mean is always around the main. Multicollinearity does not give yout hat.

2. Efficiency: Variance of the stimate is the lowest among competing methods, obtaining β. If I use all the information I would get much better estimate. Even if I remove the outlier I would improve the efficiency.

3. Consistencey: in the limit as $n \to \infty$ then $\hat{\beta} \to \beta$, or $plim_{n\to\infty}\bar{\beta} \to \beta$.

Mean square error is combination of consistency and unbiasedness. Whether unbiasedness is better or consistency has not clear answer.

Once you understood that there is heteroscedasticity, you need to use weighted least square. Suppose that variance is function of size. if we believe that $var \propto (size)$, how do you know which to plug in $x_1, x_2,$, and it would be sequential gain. then if $y = a + bx_1 + cx_2$. Then we get regression of $\frac{y,}{=}\frac{x_1}{x_1} + b + c\frac{x_2}{x_1}$. It means any heteroscedasticity would be done by weighted least square. Then your new intercept would be b. The coefficient of x_1, you need to be careful how you interprete it. If you think about where you use it, it is the way.

You can use weight x_1 in the regression to correct it. Weighting does not lead to the same thing as original model.

We are saying that x_1 is offender. The error term is devided by x_1, if the variance is correlated with the variable, then $\frac{\epsilon}{x_1}$ would be a solution.

For your research you need to keep in mind to check:

1. is it linear? Check for non-linearity.

Put the polinomial terms, and check whether correltion works. You can also take log.

2. Interaction effect.

3. Collinearity.

4. Heteroscedasticity.

These things should be checked without anyone telling you.

For white test you can test it in GLM procedure.

If we check both x, and x^2 you need not to look for correlation, since it would approximately would be $ln(x)$. The more important thing is correlation between variables.

This comes in VIF, and it is nothing other than variation of $ln(x)$. This can always be explained by that.

Collinearity could be expected and happen. You can argue to people why you are not worrying about.

When you are doing multiple things, do loop will be useful.

Every do loop should have the 'end' statement.

if $y > 5$ then DO;
$month = years * 12$;
END;

Do until $n \geq 5$;

$n + 1$;

End;

The value of n when it is out would be 5.

Do $i = 1$ to 10;

Do $i = 0.1$ to 0.9 by 0.1, 1 to 10 by 1 will do the following:

0.1 0.2 0.9 and then 1 2 3 4 5 10

Changes increment after passing a certain point.

Array statement also would be good to work with matrixes.

$Array(.)$

If you have 36 month of data.

You want to be able to write Do $i = 1$ to 36;. then you can use array.

Array TIME(36) MTH1-MTH36;

Then you would be able to say $Time(i) = ...$ within the do loop.

You can also do the multidimensional array.

Quiz 1 would be SAS programming, merging, proc mean, regression, ttest and chisq test, and it would be one hour.

Data analysis course of professor Murthi @ UTD: Fifth session

Meisam Hejazinia

02/013/2013

Proc GLM can help you to do the regression, but it is mostly used for analysis of variance or ANOVA.

How to test for heteroscedasticity?

Proc model helps, also for non linear estimation also is used.

You try to do the white test.

Other than checking the plot you can use 'proc model;quit;'

White test result would be significant means there is heteroscedasticity.

To correct it you use weighted square. for example use $inc2_inv = 1/inc2$; and 'weight $inc2_inv;'$.

You will see that tests are not significant as a result, so by weighted the heteroscedasticity.

You need to google yourself and fine many other things, since class time is limited.

$Q-Q$ plot is basically if the distribution is normal.

It checks whether the data follows normal distributions. You can do that for exponential as well, by changing the theoritical parameters. Q-Q plot is very useful specially when you check for the normality. Most of the time normality is not important in regression.

The plots come automatically so to not allow it you need to write $plots = none$.

Proportion of variation for collinearity should be high to have multicollinearity.

VIF when high shows multicollinearity.

Multi collinearity removal by dropping the variables is kind of trial and error process.

If heteroscedasticity exists use weighted sum of square.

$\sigma^2 = k.x_1^2$ and you need to shrink it so use weight $\frac{1}{x_1^2}$

Outliers have strong influence, and you want to rule them out. Remove top 5% of observations and bottom 5% for the removal of outliers.

Whenever doing regression have an eye on the following issues:
1. outliers
2. collinearity
3. heterodescedasticity

Chi-sq is like correlation, but T-test tests both magnitude and significance.

When order is clear you can run t-test.

Table $(a5 - a21) * (Q5 - Q7)$;

Try to find these tricks by yourself, and be better than professor muthi as he asked.

t-test can tell you more about the relation whether international or U.S is more, you can use t-test, and you can talk about their difference.

Survey design tells you about how you can do with data. Scale are nominal (e.g red, blue), ordinal like (A B C D), and then you have interval (1 2 3 4), and finally you have Ratio variables. On ratio scale we have absolute zero.

Ordinal, interval and ratio scale. This data is interval. $1, 2, 3, 4$ and we can take the differences but not ratio. As long as comparing the differences it is okay, but you can not take ratios here.

When number of observation are low, for example 3, for 25% and 18 for 50% you need to merge cells, since in chi square the categories should be greater than 3.

F-value lower means the equality is rejected then for t-test you need to look at setterthwaite, and if F-value is not significant you need to use pooled.

You have order in your mind how the order is maintained for the text of in your coding, and you need to cut and paste.

He used ttest for this calculation for the second answer of how the imprtance affects the second one.

ttest is to test means between two independent groups.

Suppose they are couples then they are not independent, then we have to use paired t-test.

Pair t-test: right leg and left leg. Managers and employees in the firm can be dependent. On that case you need to use pair t-test.

proc ttest;

paired SBbefore*SBafter;

run;

Try independent test, and then dependent, and then check correlation.

Bass Model and Conjoint model.

Bass Model

Initial model was for durable goods: things that you do not purchase frequently.

The goal of Bass model is to predict maximum sales. How much capacity should be.

Download could be treated as Bass model.

Adoption / diffusion comes from disease.

p = fraction of people affected directly (innovators)

$2 - 5\%$ of people are usually innovators.

q: fraction of immitators.

m: refers to market potential.

In period one $p.m = N_1$ of the market is affected, and purchased the product. At time $t = 2$ there is always that still will be affected, and now the remaining would be $p(M - N_1) + q(N_1/M)(M - N_1)$.

$Sales_t = a + b.N_{(t-1)} + cN^2_{(t-1)}$

t = 0 1 2 ... 4

S_t = 0 20 60 160 460

N_t = 0 20 60 160 460

N_{t-1} =0 0 20 60 160

$(a, b, c) \rightarrow (p, q, m)$

$p = 0.02$ $q = 0.30$ $M = 10^6$

At least you need three data points (p, q, m) to calculate this, although the degree of freedom is zero.

If I know parameters of similar products, then I can forecast.

Fore example mp3 players is like walkman, tampe recorder, and raido.

Non linear least square could be used using 'proc model'.

Block busters start with high for adoption and it dies during time. Inverse Bass Model.

Sleeper movies like home alone and it was s-shaped.

Three stage Bass Model when DVD comes and when rental comes and so on.

Adoption of DB technologies and so on also use the same form of parameters.

Homework 3, create a code for Bass Model on the data on the last page.

This Bass model was on 1969.

Then on 1993 Bass, Jian, Kirishnan created generalized bass model and included advertising and price.

To estimate Horskey-Simon model of generalized Bass model you need to use non-linear least square for estimation.

For specific company you need to multiply to market share.

Data analysis course of professor Murthi @ UTD: Sixth session

Meisam Hejazinia

02/20/2013

Be careful about the size of group that you are doing chi-sq on, since they should be more than 5.

$'+'$ will be ignored.

In the for loop, jumping out of the do loop will happen when it becomes 7.

DLM and LRECL should be included when comma separated, and when record length exceeds an amount.

When you read, we put infile and output. For output you will say 'file' and 'put'.

If you need to specify format you need to specify.

Specify libname a'..';

In the data command you say 'q.name'.

These are for ASCII file, and create permanent file dataset.

Price elasticity: percentage change in market share, over percentage change in price. $\frac{\Delta MS}{\Delta P}\frac{p}{MS}$

You have to think about interactions as well.

You have think about that, you are smart, but professor is also smart based on common knowledge, so the questions would not be easy.

$(-1.28 - 0.47D - 1.07L)\frac{\bar{P}}{MS}$ For both display and price you need to account for the interaction as well.

T value is nothing, but coefficient divided by standard error. $t = \frac{coeff}{stdError}$, it does not need to be expressed explicitly you just need to check quickly by your eye, whether it is greater than 2 or not.

Most important has to do with scaling. You can not compare the betas. First solution was to use STB. Sometimes are not reported on the papers. At least then we scaled it by $\beta * mean$. It is not perfect, but it is useful, on the other hand you need to care about significance as well.

We need managerial explanation, and not statistical terms. Like you are explaining to your father and mother. 'There is a joint effect, if there is an effect loyal customers will respond by increasing market share.'

You need to use words, and not say that if loyalty increase by one, since it is mathematics.

When both of them mean display and loyal customer, on top of two effect market share would be increased.

If there is a display and price is increased the sales will go out.

Lower price is positive, and if display goes up is also good, so interaction since the coefficient is negative, is synergy.

If price reduced, and display increased then the sales/market share will go up.

Data A;

INFILE ONE DSD; * for delim. in the same file reading *

INPUT 10$ NAME $ DEPT $ PROF $;

CARDS;

....
.....

RUN;

Take this course seriously, since professor also puts considerable amount of time on the course.

You have to look at other features, while doing the homeworks. Do not do them mechanically.

Chi-sq, ttest, regression and so on, we have done. You need to pick up by yourself.

Proc content is useful to know what variables are in the data.

Proc print separated observation numbers Proc Print DATA=a1 (obs=10); RUN;

MMDDYY10. means the data would be in the form of 12/31/2011

If we want to used comma to read we use $COMMAw.d$ for example after each 3 we use ',', in the form of '12,300,10$.

DOLLARw.d is also useful when you want to use dollar format.

if type eq " airforce" then output airforce.

Select (Type);
When ("Army") output army;
When ...;
otherwise;
run;

To separate data to multiple datasets.

DATA army (KEEP=airport city state country);

set a1(firstobs= 3 obs= 100);

To start from specific observation.

Retain statement says keep specific data.

The initial value of month to date is zero, and you would be able to calculate cumulative sum;

data a2; set 1;
RETAIN mth2date 0;
mth2date=mth2date+saleamt;run;

if first.product then ptot=0; ptot+saleamt;

You can also use first or last then the output will be kept only.

Retain says hold this one.

input date1: date10. date2:date9. amt:dollar6.;
12312009, 29feb2000, $r,242

These are for non standard formats.

For warning you need to stop and examine what it really means.

When you put different lines of input then it will jump to different lines.

input Lname $ Fname $;
input City $ State $;

'' also says go to the next row:

input Lname Fname $
City $ State $;

You can also use row number in the form of:

input # Lname Fname $
#2 City $ State $;

You can use this to jump to the x'th row.

total= sum(of q1−q4); if total ge 50; run;

This does not print out the one lower than 50.

newvar=SUBSTR(string, start, length)

You can also use scan command for the substr until the delimeter.

Trim could also be used to delete the space before and end of statement.

Output Delivery system (ODS)

It can creat HTML reports that you can publish on the web.

You should read about them yourself as well, and on your next homework show it in your result, showing that you have done something on it.

Conjoint analysis is another application of regression.

1. New product development.

You want to know wat is the value of particular feature to consumer.

It gives you:
1. the value of each attribute to the customer.

2. Which attribute are important

We want to design orange juice. Price colories, nutrition.

Attributes of the products and levels.

I know that it price would be $2, or $3.

Colories could be 100 or 150.

Nutrition could be 20% or 30%.

If you want to do these things and you don't know anything from conjoint.

You can ask people, and people can tell you.

People can not answer in most of the time, but they can talk about their product.

When people can not tell you exact answer, this technique would be helpful.

Whenever people can not actually or verbally tell you, you need to use this.

Many times people have problem telling you what they want.

If a combination of sample is given, then people can rank them.

Key is use ranking information to answer these informations.

If I want to create all possible combinations you need to create $2^3 = 8$ combinations.

You can create 8 samples or choices, and ask the customers to rank them.

You can also give them on the piece of card.

Which card is most prefered and least preferred.

Dividing pilings to two, and then you can combine most to the least.

It is useful for web design in IS.

For example in designing a shoe that has 1. sole (rubber, poly, plastic) 2. Upper (leader, cotton,

naylon), 3. Price (15, 30, 45)$.

We can create dummies in the form of:

If $S = R$ then $D_{S1} = 1$ else $DS_1 = 0$;

Rank could be preferences, and 27 combination would be there, and for each attribute we created two dummies D_1, D_2 for the first and so on till $D_1 - D_6$.

They take the combinations, and they rank them.

Rank have to first be converted to the preference.

We want to make $1 \rightarrow 27$, and $27 \rightarrow 1$. You are just flipping the scales.

Then you take preference which is inverted rank, and then we need to do the regression.

The regression output will be calculated on regressing the preference on the the dummies.

You need to run regression for each of the persons. Using do loop you will do this.

$b_1 = 1.0 = u_{plastic} - u_{rubber}$ (1)

$b_2 = -.33 = u_{pol} - u_R$ (2)

We get the difference from the base level.

$u_{pl} + u_{pol} + u_r = 0$ mean the mean equal to zero. (3)

Utility has no zero, and can be shift right or left, means sum would be zero. We do this so that we would be able to compare them, by standardizing.

By setting to zero we can compare utility of sole with upper with price, since ach would be zero.

This is scaling constraint.

This numbers are aribtrary.

From this three equations you would be able to find utility of three different levels.

To find which attribute is less important, now that we have $(U_{pl}, U_{po}, U_R = (.222, -.556, .778)$, stipper means my utility will jump. We use $maxU - MinU$ as an importance rate so $I = |MaxU - MinU|$. Here you can calculate importance for sole by $.778 - (-.556) = 1.334$

You can calculate for upper, and price.

The logic is that if the difference is higher it is more important.

All adds up to value 4.658, and then you can find relative importance by dividing each importance by the sum of importance.

Now for each customer we can say which value is important. We can for this customer for example say that $p > s > u$

Then we can calculate utility for each of the options for this specific customer. $U_{PL,L,15} = -.222 + .445 + 1.111$. These are called part worth, mean the value of each of the levels.

We made the assumption that the utilities are additive. This is the reasonable assumption, and we are making it here.

For every combination of attributes, you can find the utility value.

Now the question is how we can calculate the market share.

Nobody is interested in one person.

We want to know what does the market want. This helps me what each person wants. We now want to calculate market share.

choices	A	B	C
1	2.5	1.4	2.4
2	1.8	2.3	1.8

Lets think about three combination of choices (A,B,C), and for the person one the values are:

Maximum utility rule says that the customer uses the choice that has the maximum utility.

Maximum utility rule can tell you that the first person selects A, and second one selects B.

The loose choice is that says (A,B,C) will be selected with probability $(.39, .22, 38)$

$pr(A) = \frac{U_A}{U_A+U_B+U_C}$
Loose has the problem since, the utility of some are negative, and positive and negative does not make sense.

As a result behaving well could be expected from the logit rule $Pr(A) = \frac{e^{U_A}}{e^{U_A}+e^{U_B}+e^{U_C}}$

Then you can get the mean to say what would be market share.

You can do the what if analysis with different kind of things to say what if I got problem in one aspect.

You can also use this for the segmentation.

People with sole as the most important, and segment with highest upper, and the price as the highest important. This will help you to segment market.

Conjoint is the most usefule, and there are over 4,000 applications.

This is all dummy variable regression, and intelligent modification. Due to the repitition you need to do this, and save on file and do it again.

In many cases people can not think about each attribute, and if they could you do not need this.

On the website you can have click as an indication of preference.

The same analysis can be done with logit, when your dependant variable is click or not.

This is choice based conjoint. We given the set of choices what you have selected then we use logit.

Basically dependent variable is discrete variable.

We are using regression.

When preference is more in the form of discrete we will use this.

Size of the ad, and changing location can be done easily on the website.

Dimond are based on cut, color, clarity.

If you regress price on cut, clarity, color and canof, you can think about how the price should increase. In the website you can use this.

You can do this dummy variable regression, and understand how much you want to pay.

Blueline have these data, and you can use this.

You will understand which one is most important.

Doing this regression you will understand which one is more important. Price is market price. This price is equilibrium, and is measure of preference. Do loop, and dummy variable regression and arrays can be helpful to create code for conjoint and calculate market share.

Read ODS by yourself.

Data analysis course of professor Murthi @ UTD: Seventh session

Meisam Hejazinia

02/27/2013

Download new sylabus from the website, and there would be a homework due next week.

Bass model discussion. The code is for different dataset.

Prediction is important.

Data set is read in the form of cards.

Cumulative is calculated, and lag of cumulative is calculated, and square of the lag is taken.

$lagd = lag(cumd)$;

$Cumd + download$;

You get some regression parameters. You get $outest = coeff$;

Then you can use 'proc print'. Typically very small p around .02, and around .25 would be for q.

Now question is how do you forecst?

An array is created, since we wanted to forecast 'array nt[21] t1-t21 (0 ... 0);'

Then a do loop and a preduction is taken and the formula is put. Each time SP is calculated, and then they calculated $s(t)$. You just need to M, p, q, and using zero you calculate $S(t)$, and from that you calculate cumulative, and again you use it to calculate nt, and then plotting happens from there on.

You have to put zero in the data, since you want to estimate from that point. There would be two zeros. You must include zero for estimation.

Another way to estimate Bass model is using non-linear least square.

For non-linear least square you use 'PROC MODEL'. Gradian search starts from one point, and keeps looking, should the function maximizes here or not, and it continues to reach the improvement. If it doesn't move, it stops.

When improvement is not more than the value of $Converge = 0.0000001$ is the value of convergence, and then we have $Maxiter = 10000$ and you can define *exogeneous* variables, and then you define boundary.

We as FIT $Missing = pairwise$ OLS OLS outpredict out=Modelout COVB OUTCOV $OUTTEST = TEXT$. Weak and Download are exogeneous. You define Params $M = 3600000$ $p = 0.27$, and so on, are starting values. It starts, and iterates to the point that it would no more be able to improve the estimates.

$\%f(f, p, q)$ is a macro.

For prediction you should not use data. As far as you know M, p, q you can draw the graph.

For new product that you don't have data, you will take the M, p, q of similar products.

Explore 'proc model' yourself, since we will use GAUSS for likelihood estimation.

Genetic algorithm is used to reach global maximum.

For every combination you find utility; find out maximum utility, and predict for each what would be market share.

If utilities are close, then predictions are wrong. In this case we need to use logit to calculate probability, and then we take average probability as predicted market share. Otherwise you still have to do the same, to use do loop. Think about this excersize in the form of do loop.

Up to now the estimation we have done was cross sectional, and not panel structure. What you do with panel data?

CS: Cross section, could be households, firms

TS: Time series: weakly, yearly monthly

When we have panel data we can solve problem of unobserved heterogeneity.

Heterogeneity can be unobserved or observed.

Let's start with scanner data. Analysis on households. Households are different. If we don't account for it, we will have biases.

Two kind of people: 1. Low price sensitive 2. High price sensitive . Let say each have 50%.

For high price sensitivity would be -3, and for low price sensitivity we will have -1.

The fact that people are different. What coefficient will you get? $\beta = -2$ the average of both sensitivity, which is neither of those.

This is bias since you took the middle and that made nobody happy.

This $\beta = -2$ is for no heterogeneity. If you realize that the world is divided into two, then you will have bias of estimate. We could have more complex world with multiple segments.

Price sensitivty could be related to age and income, where age is in the form of two degree polynomial model. This will be called observed heterogeneity.

$sales = a + b.price + Inc + Age + Age^2$

a can be different across people. These three variables account for variation in intercept.

We controled for differences in a, but how will we control for differences in b?

$b = \gamma + \delta.Inc + age + age^2$

If we plug in we will ahve interaction effect. If $b = -3$ then we will should have $\delta = 1$ in the form of $a + [\gamma + \delta.Inc].price$. This is hierarchical model. When we talk about observed heterogeneity, then we will have hierarchical model.

Intercept heterogeneity, and slope hetrogeneity. Intercept hetrogeneity is α is different. Slope heterogeneity is β is different.

I am allowing for the fact that different people have different price sensitivity.

In marketing we also think about unobserved heterogeneity. We think about unobserved heterogenity. This means we don't observe it, but we want to control for it.

On panel data you can do something.

Observed heterogeneity could be taken into account by checking interactions.

$S_{it} = \alpha + \beta.price_{it}$

This is OLS regression.

We can develop models $S_{it} = \alpha_i + \beta.price_{it}$. Mean for everybody we have put intercept. Everything unique to specific person unexplained by price is put into α_i. This is called fixed intercept. Generally it is called fixed effect, but here we have slope fixed.

We can not have α_{it}, since we have ϵ_{it} as well. If we put α_i for $N = 100$, and for five years, we will have $\alpha+$ 99 dummies. You can put α_t in as well, mean for each year there is an intercept. You loose degree of freedom, but you can not do both in the same model.

α_t could be interpreted as the effect of government rule change.

We want to know the effect of Craig list on the classified advertising. It should be negative. You can put for free, or cheaper than newspaper. We found out how craiglist come, and as they put α_t they found out that the result is zero, since classified avertising were going down, and craiglist just came incidently. This α_t means trend, and something that is going down during time. There was no effect of craiglist despite this trend. Could have been effect of internet, email and so on. It could have been effect of many of these things.

The bias is very high, so you need to worried about them. You need to have two observation or more from the person. You have only one observation you can not do anything other than OLS.

These models are called fixed effect model.

You can put separate β for each as well. $\alpha_i + \beta_i.pric$. We are running regression for each person separately. You are loosing degree of freedom here. $\alpha_t + \beta_t$ could also be used.

This is big deal for panel data. Try to get panel data to control for unobserved heterogeneity. You can challange big paper by these.

Fixed effect we quickly loose a lot of degree of freedom, since you are estimating for example 100 parameters.

Cross section variation, and time series variation. Always focus on cross function variation. It has generally much larger variation than time series.

Cross sectional variation is much more important than time series variation. If you have degree of freedom do both, but if you do not have degree of freedom use cross section.

Fixed effect versus Random effect

$S_{it} = \alpha_i + \beta.P_{it} + \epsilon_{it}$

Assume $\alpha_i = \alpha + \xi_i$

Normal is generally used. We only need to estimate α and σ_{ξ_i}, and you save a lot of degree of freedom. Assumption: there is no correlation and it would be random distribution of α_i.

We have to assume that ξ_i and ϵ_{it} are not correlated. and ξ_i and p_it are also uncorrelated.

Now assume $\alpha_i = \bar{\alpha} + \xi_i + \nu_t$. In this case we need to estimate the variation $\sigma_{\xi_i}, \sigma_{\nu_t}, \bar{\alpha}$

This is intercept heterogeneity. We can do this for slope heterogeneity as well.

$\beta_i = \phi + \varphi_i$ and so on.

Unobserved heterogeneity creates bias. Cross section variation is more important than time series.

Software you need to use

PROC TSCSREG does it.

In term of coding it is not difficult. There are couple of things you need to understand.

You need to write cross section first and then time in the id statement of 'proc tscsreg'.

proc tscsreg data=a;

id CS time;

It is very important to sort data first based on cross section.

MODEL Y= X1 X2 FIXONE (α_i)
FIXTWO (α_i, α_t)
RANONE (α_i)
RANTWO (α_i, α_t)
FIXONETIME (α_t)

FIXONE is to add 99 dummies.

TSCSREG is for intercept effect.

PROC PANEL should be used for slope heterogeneity. Also PROC PANEL can do both intercept and slope unobserved heterogeneity.

Overview of the model of TSCSREG in the hand out.

Unbalanced data: Different cross section have different time series.

$TS = number$ tells us how many time series are there in the data. Standard output estimates.

We can do regression by variable.

ID statement is important. you need to put cross section first, and then time series.

Fuller-Battese is not different from Fixtow.

PARKS allows corrlation between time series.

You can control for fixed effect and random effect.

Moving average on time series you can use $DASILVA$

You can whether intercepts are the same across models or not.

Go over panel also in the SAS help.

One way and two way models. Auto regressive models, and moving average models could be used.

Model statement has the same form Fuller, PARKS, DASILVA, FIXONE, and FIXONETIME.

There would be homework that asks you to do this. Go over them and read by yourself.

Variance component for cross section is σ^2

In the table you can see that cross sections variance is much higher than time series.

Housman is the test of whether the random effect is correct model. No corrlation between variance and standard error.

Housman test:

Null hypothesis: Random effect is correct model. No corrlation between ξ_i, ϵ_i)

So, use fixed effect.

if $p < 0.05$ mean you must use fixed effect.

If you reject the null hypothesis then you must use fixed effect.

If slope heterogeneity is not there in 'proc panel', you should use 'proc mixed'. check it by yourself as part of your homework.

'proc mixed' is generally for slope heterogeneity.

Dynamic panel estimator paper. You need to understand what is the structure.

Time series heterogeneity. Cross panel heterogeneity are used in proc panel. 'Arrelan' method could be used. It is relatively new, and you should go to the paper and review it by youself to find out how this should happen.

Book on econometric models Pndick and Rubinfeld.

Simultaneous Systems of equations

$Sales = a + b.Price + c.Adv.$

When we do regression we are making an assumption that price, advertisng, and all the x variables are exogeneous. Mean it is set by someone outside of the system.

For example sample of rats I give them 8 miligram, and another sample 10 miligram. You are observing blood pressure. Dosage is set by the researcher. Variables are set by God, or someone out of the system.

Sales is function of price or advertising, people do not argue. Price itself depends on the sales. Y depends on X, and X depends on Y. We will have simultaneous effect. We have the problem called endogeneity. Endogeneity means, X variable is not set by outside the system, but process was that it is set by manager. Price is not exogeneous by defnition, but it is set by someone inside the system.

Profit that the customer gives to the bank is called profit. Profit depends on Direct mail, Tel sell, internet, and direct sales. These are called modes of acquisition.

Which of these effects profit more.

It turns out that $DM > INT > DS > TS$, mean depending on which mode of aquisition people come, this would be different.

There has to be theory. Why should it matter? Cost effect are passed out. Self initiated and other initiated. When I initiate it, I want it more. When company wants it it would be different. Self is more profitable than others. The other is targetting. Some methods are more targetable than others. If firm wants to target profitable customers, they send you fliers.

Direct male and tel sales are more targetted. Internet previously was not targetted.

Self initiated more profitable. Hypothesis is that it is the best situation.

If you apply for credit card you are in need, and you will go for it yourself.

Question is is this endogenous? If you do the regression, means you believe that DM is exogenous.

The customer when applying they select where they want to go through.

We have to build a model that $DM = f()$. We will have separate equations for $TS = f()$, and $INT = f()$.

We need instrument to solve this.

Pricing and advertising is also indogeneous.

We have to have another equation for $Adv = f(sales)$.

These are now called indogeneous variables, and each variable explains another.

The best way is to do the simultaneous estimation.

Supply is function of price, and demand is also function of price.

In equilibrium $Q^S = Q^D$.

I will observe price and quantity. There will be noises, but all I see is equlibrium prices.

We have identification problem. We have two unknowns, but four parameters. This is called identification problem. You can not get variation to identify for both variations.

Only variation comes from noise. Ideally we only should observe point.

We want to identify both supply and demand. You can think of this as $Q = a + bP$, and $P = c + dQ$, mean price is affected by quantity, and quantity is affected by price.

We need something more, mean instruments to solve these equations.

There are many possible demand functions.

We have income, because income affects demand.

For the demand equation, $Y = y1$, $y = y2$, $y = y3$. Y gives you additional variation. So you can identify it in this case. In this case you can see different equlibriums based on variation in income. Now if you have T, then you can identify both supply and demand. Identification means you need extra variables to solve this equations. As long as you have variation you would be able to solve these equations.

Heterogeneity is always the first problem, and indogeneity is always second problem. In real life, especially in management, lot of variables are endogeneous.

Price is indogenous. Today's sales can not depend on the last week price. Something that happened way back can not affect today's decision. This means predetermined.

Exogeneous variable that is correlated with price, but less with sales. These are instrument variables (IV). Should be correlated wtih indogeneous, but less correlated wtih sales. In other word its correlation with dependent is only through the indogeneous variable.

The reviewer usually makes counter argument. It is about our theory, and belief. You need to convince that this is good variable. It could be in different point in time. How many do we need is ordered condition.

three endogeneous: $n - 1 = 2$. We need two IV to be excluded from each equation for advertising-price equation. Two in the price equation that do not apear in the sales equation.

Therefore maximum is 6, and minimum could be managed to be four.

This could be done using 'proc syslin'.

Read the handout. If you don't do this your estimation will be biased.

Data analysis course of professor Murthi @ UTD: Eighth session

Meisam Hejazinia

03/06/2013

Simultaneous euqation we discussed.

The variable of x is correlated with ϵ error term is result of indogeneity, which jeopardizes the OLS assumption.

y variable is represented as the error term.

Correlation happens if something is varying. If x is fixed then we have no correlation. If x is not fixed, then x is called stochastic, and then it is possible that x and unexplained portion ϵ is supposed to not correlated.

x is indogeneous if $cov(x, \epsilon) \neq 0$.

Also in other word x is not independent of Y, and X is function of Y.

One way to fix this is write simultaneous equation. In this case OLS estimate would be biased.

Error term is any part of variation in y that is not explained by the x, means it is random variable that creates variation in y.

Error term has mean zero.

Find β that gives least sum of squares.

If the mean of error term is 0 then it means it is not biased, but it has variation.

Once we predict $\hat{y} = X\beta$.

We use normality, since CLT says for large sample, mean of mean is always normal.

If error would be normal, then it would not have bias.

If the error is not normal, you can use non-parametric regression.

We do test for normality of error term. You check with qq-plot to find how much deviation is from normality.

If you believe from theory that it is not normal, you can use log normal.

When you only have two equations, and there are more parameters than data, we will have unidentified equation. (four parameters including intercepts, but only two known p and q)

$P_t - \mu_{p_t}$ is called p_t.

11.3 called structural parameters, and 11.4 is reduced parameters. Yet you can not recover structural parameters from reduced prameters.

You can estimate π_{12} and π_{22}, but you can not recover $\alpha_2, \beta_2, \alpha_1, \beta_1$ from this, due to identification problem.

To identify the full system you will need more instruments.

We can recover α_2, as one parameter only.

The question is how much you can recover.

If you have no exogeneous variable, you will have cluster of points.

Suppose you put one instrument Y_t in demand, at each level of income, you will get the demand curve for each, and you can then identify the supply equation.

If you can also have instrument on supply, you can recover demand, so you will have exactly identified. The problem is who said Y is the unique exogeneous variable.

Overideintified means there are multiple solution possible. When you have one solution then you will have exactly identified.

Two state Least square is when you system of equation is overidentified. When you have multiple solution.

Order condition says, the number of predetermined variable excluded from the equation must be greater than or equal to the number of included indogenous variable minus one.

You can have $2 - 1$ instruments for each of the equation.

If it would be zero we will have over identification.

The curve 11.3 instrument on demand helps to identiy the supply curve, but we do not have any instrument in the supply curve to identify supply. Using instrument on demand we can identify supply, but demand could be any of these shapes.

This condition works if $\beta \neq 0$, else it would be meaningless.

At least to separate cluster of points should exist for identification.

When we have over instrument, we don't know which to use. How do I use both? In this case you should use 'two stage least square'.

On first stage one indogeneous variable is regressed against all predetermined variable in equation. p is regressed against y and t. then calculate $\hat{p}_t$, which is independent of error term, now use it in the first equation.

One endogeneous regressed over all the variables. Then calculate $\hat{p}$, and then calculate Q.

First you do regression from the first, and calculate it. Then you put it in the second equation, and you calculate the second one. Using this method you can remove indogeneity.

Predetermined variable means exogeneous variables.

Regress on both y, and t of p. Regress on all exogeneous variables.

Then you put $\hat{p}$ in the second and estimate with all other variables.

Where $\hat{p}$ would be the predicted value.

You can do either way, mean you can start with Q, and then substitute $\hat{Q}$ in the second.

The predicted value would not have noise around.

Oversupply of instrument is removed through 2SLS (two stage Least Square).

If another person is setting rule, for example government then we will have exogeneous. Inside firm it would be difficult to say something is exogeneous.

If you have great instrument, then this works, but the problem is always instrument.

2SLS removes the problem of overidentification, since it takes essense of all the variables.

Read the examples of note by yourself.

Hausman specification test checks for simultaneouty by checking the correlations.

If preicted value of error term is put in the second equation with the predicted mean in two different variables in the second equation, and the coefficients are different, it shows no simultaneity. $p_t = \hat{p}_t + \hat{v}_{2t}$, $q_t = \alpha\hat{p}_t + \beta v_{2t} + \epsilon_t$.

SUR: Seamingly Unrelated Regression

$Y_1 = \alpha + \beta.X_1 + \epsilon_1$

$Y_2 = \gamma + \delta X_2 + \epsilon_2$

$Cov(\epsilon_1, \epsilon_2) \neq 0$

To improve efficiently we use SUR.

Zellrer(62).

How in real life you have this?

$MS_coke = \alpha + \beta.p_{coke} + \ldots + \epsilon_1$.

$MS_{pepsi} = \gamma + \delta.p_{pepsi} + \ldots + \epsilon_2$.

They seem to be unrelated, but the correlation is in error term.

It could be about Sales of coke, and Sales of pepsi.

Efficiency has to do with the variance.

In this case we have no simultaneity, and only error terms are correlated.

Correlation is not because of simultaneity, but because of different process.

Simultaneity and $corr(\epsilon_1, \epsilon_2)$ makes us to use 3SLS.

You start with system of equation. If simultaneity only then 2SLS, and if you have $corr(\epsilon_1, \epsilon_2)$ then use SUR, and if both exist then you should use 3SLS.

Maximum likelihood also could be used to calculated from MLE (maximum likelihood). $\beta = \beta$. Then $LIML$(Limited information Maximum Likelihood) is used for simultaneity, and $FIML$(Full information maximum likelihood) is used for $3SLS$.

All is done through 'proc Syslin'.

In SUR you stack data to get better standard error. It is pooling, and allow error terms to be correlated.

Because of some external market condition they are not completely unrelated in the 'seemingly unrelated regression'.

In syslin help of SAS the data is simulated.

In system of equation you need to specify what is endogenous by 'endogenous p;', and instruments by 'instruments y u s;'.

R-square can not be interpreted as percentage of variance.

ANOVA also gives F-test, but still some caveat in interpretation.

In the parameter estimate you can compare OLS or 2SLS, and you will see for example 20% improvement showing how much the bias was, and how the sophisticated method removed it.

Indogeneity problem on the OLS caused the effect of p to be positive while it should be negative.

There is nothing magic about 95%, and you can report 94% as confidence interval.

Writing Macros in SAS

'Proc Transreg' which is transform regression for conjoint.

Macro is like subroutine, and it is module and you can use in different places.

Defining macro variables:

%Let dsn = clinics;

It will say wherever you saw this just replace it.

Using macro replaces all with the word.

You will define the variable and then it automatically replaces all with the value of this variable.

To link this variable and use it you need to use *&dsn* in multiple places in the code.

% Macro Look; * creates a macro called Look

$MEND <look>;

The ampersant looks for the macro variable.

If you want to call the macro you need to use:

%look(clinics,10);

By doing this you can do for all.

Industry, since it does this calculation on different products, they usually use macros. Mainly due to repition.

%Do i=1 %To 5;
PROC REG data=develop outtest=B&i;
...

Could be used.

Factor Analysis

How many twits have you done or reviews as measure of brand loyalty. Attitude toward brand could also be used for brand loyalty. Brand loyalty can not just have one measure, and could have multiple measures.

Quality of food, atmosphere, waiting and so on.

Service quality is measure of multiple parameters. Multiple measures are not independent and they are correlated.

How do you create one measure.

One can ask multiple questions. Can we created an index which is weighted scores of multiple measures? Yes, through factor analysis.

Factor analysis reduces the number of variables, and condenses the number of variables.

It helps to retain 90% of information. Information is usually about variation.

We have to reduce, since for example 200 piece of information is gained per person by census. Difficult to work with too much data. Even if you can the result would be distorted due to multicollinearity.

The first technique is principle factor analysis (PCA). It is linear composite of original variable. Consumer Price Index (CPI) is measure of basket of goods. How do you weight? How much weight is put on these is through factor analysis.

PCA ensures new variables are uncorrelated.

We need a composite of all the scores. GPA that has assumption that weights are equal. Why can not you weight differently?
Factor analysis assigns the weight so that principle components assigns most variance among the basket.

Once you create index, you can measure economy or certain type of stock. Like Dow Jones.

One of the beautiful thing is that suppose you come with two indexes. We want it to be uncorre-

lated. The first index capturing maximum variation in data, and second index as second measure of differences.

PCA summerizes information in data by reducing orginal set to smaller set of factors.

$P_1 = r_{11}X_1 + r_{12}X_2 + \ldots + r_{1N}X_N$.

Square of the weights should add up to one. This is nothing but scaling. This is weighting variables into one index.

Principle components are uncorrelated.

Conditions:

1. Maximum variance.

2. Scaling condition (weight square as measure of variance sum up to 1)

3. Uncorrelation of the factors.

The first component is counted for example for the bottle, and then the second one would be perpendecular to this. The third dimension will be perpendecular to all of these.

You will get diminishing return, so you discard the rest after some time. Three variables at most will have three principle component.

We can flip and write variance as function of factors.

In principle component there is no error term, where as factor analysis we allow the error term which is unique aspect to x_1, in the form of $x_1 = a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + \ldots + a_{1m}F_m + a_1U_1$

Think of it as variance unique to x_1.

Principle component is solution to multicollinearity.

How banks are doing: (bank survey)

1. Small banks charge less than large banks 2. large banks are more likely to make mistake than small banks
3. Tellers do not need to be extremely courteous and friendly- its enough for them to be civil
...

The questions would be correlated.

Factor loading are weights of each factor for each of the variables. Factor 1 when you look will show the the high weight of the last three variables. While second factor has higher weight on the first two variables. You can put whatever name you want on them.

You will look at the high loadings. Loadings (weights) are measure of of variance. We just look at the higher than .5, since it is squared will become .25.

Communality says these two factors combined how the capture the variation in main variable. for example for the first variable when it is .57 you are loosing 43% of the data.

Percentage of the varation explained also shows how much of the variance is explained by each of the factors. Factors are not also correlated.

Principle factor analysis will give you following data that you can use:

1. Factor loading

2. Communality

3. Percentage of variation

For each person we can calculate predicted value.

Factor score is predicted value for each value of the factors based on data.

Square of the factor loading weights will give you the eigne value and that will help you calculate

variation explained.

Eignevalue is interpreted as percentage of variance explained.

Eigen vector is factor loadings.

The maximum eigne value by this could be n which is number of variabels, and $\frac{eigen}{n}$ is the the percentage of the variance explained by the factor.

Subash book of multivariate analysis has this topic.

Before factor analysis you should either do mean correction or standardized.

Standardization makes sure variance is one, but mean is zero.

The principle component is the angle that by projection creates the maximum variance.

Data analysis course of professor Murthi @ UTD: Eighth session

Meisam Hejazinia

03/20/2013

Principle component

For data reduction into uncorrelated principle components.

From book of Subhash and Schanma (Multivariate analysis)

Mean corrected value means you deduct mean from original values.

Sum of square cross product (SSCP): Degree of covariance

Given by this point we try to find axis that explains more of the variance.

$x_1^* = cos\theta \times x_1 + sin\theta \times x_2$

Means at one axis when you start to rotate data you will find maximum variance. Now that they are projected to the new factor you will get the factor scores. You try to find the axis that explains the most variances.

In the factor analysis we allow for error term which is not in the principle factor analysis.

Factor rotation is different method to get better interpretation by imposing additional constraints. These are categories of method, and there is no one way to do this. The key thing is whether this makes sense. In regression everybody get same answer, but in factor analysis it is more exploratory, and each person could have different answer to this. It is possible to have more than one answer.

Varimax rotation the goal is to get loadings close to zero or one. It tries to make it as much as possible.

The weight should be more than .5 so that this variable would be $\in$ special factor. You get fore example 2 and three factor, for 2 could be quantitative of qualitative GMAT score. Quartimax is another method. Verimax helps to put them into two category, mean the first three variables would be summerized in the first factor with weight of each greater than .5, and summerize the second three factors with high weight into the second category.

Rotation role is to redistribute the variance.

Example:
for 8 Variables
at most we will have 8 eigen value
Eigen value should be > 1
At least each should have $var > \frac{1}{8}$

Another alternative to Eiven value > 1, is select the elbow. These are percentage of the variance that is explained, and when you get diminishing return stop.

Positioning map will show the position in the space consisting of axis of factors. Brand map is useful to do the strategic thinking.

Create furthure regressions using $F_1 + F_2$ instead of X is another usage of factor analysis that could be used to solve the problem of multicolinearity.

PC	FA
Create a composite index (e.g. CPI)	$F_1 = w_1X_1 + \ldots + w_nX_n + \epsilon_p$
Data reduction	Allow for error term
Uncorrelated PC	May/May not be correlated (may not be orthogonal)
PC1 most var	Also data reduction
PC2 next most	understand underlying factors 'constructs'
$PC = w_1X_1 + w_2.X_2 + \ldots + w_nX_n$	

Factor loading	weights
Factor score	Predicted value of factors
Communality	Variance of X's explained by all factors = 40% or = 80%
Proportion of variance	Varianced explained by one factor
Eigen value	Measure of proportion of variance
Eigen vector	Meaure of factor loading $w_1^2 + w_2^2 + \ldots + w_n^2 = 1$

If the goal of the regression is prediction then multicolinearity is not an issue, but when we are talking about interpretation then the multicollinearity is problem (for causality, associations, and so on).

The paper of How to interpret output of SAS for factor analysis is the paper that you could read. Use this paper for your homework to come with the good analysis.

'Annotated factor analysis' means it explains what each of the terms mean.

Default option is eigen value = 1. You define the varimax here $method = printcomp$ will tell you principle components. If you want factor analysis output you need to used $method = printit\ prior = sms$. SMS: prior commonality constraint.

Correlation matrix is also given.

Some factors may not be loaded, and those factors may be loaded to all the factors.

The absolute value is important in terms of weight $> .5$, and the sign is not important.

The key here is does it make sense or not. Then it shows you inter factor correlated. In principle components you will not have correlation, but on the factor analysis there may be correlations.

In finance based on risk and reward, and the stock price the factor analysis is done.

Some review

Why endogeneity:

1. Error in measurement

2. Simultaneity

3. Ommited variables

Pepsi and Coke have correlated error terms, but since they do not have common customers with RC Cola the correlation in error terms would not exist.

3SLS is useful when you have both **2SLS** and **SUR**. Means we will have both simultaneity and correlation. $Cov(x, \epsilon) \neq 0$, and $cov(\epsilon_1, \epsilon_2) \neq 0$

Sometimes when β is random, then we would $corr(\epsilon_1, \epsilon_2)$.

2SLS is useful only when you have instruments available, or you would have weak instruments. If

2SLS	SUR(Seemingly Unrelated Reg)
If x is endogenous, $cov(X, \epsilon) \neq 0$	Correlation in error terms $cov(\epsilon_1, \epsilon_2) \neq 0$
Over identified system	

not what you will do? Then this method would not be useful anymore.

Instrument as census data.

We have another method called **Latent Instrumental Variable (LIV)** is state of art, and you can use maximum likelihood estimation for it. It is 2012 paper, read it by yourself.

Another thing that you can do is **Propensity Score Matching**. Suppose there is some indogeneity, means something that you did not observe. Experimental study you will randomly pick. In this case you will have affinity group versus non affinity group. We want to see whether for example profits of affinity customers are greater than non affinity, and whether both groups are the same? Customer choose whether they want to play role in the experiment. You match person from one group with the person from second group. You match them on everything. Means same age, same income and same education. We create something called score as function of all these scores and match them, then you will find what is the profit. Then the difference will not come from the difference of people. It is good to apply specially for group with endogeneity. What is the effect of wages on joining the union. The goal is to know whether joining the union affects your wages or not. If you do the matching then the effect of all the other control variables would become insignificant.

The SAS code is available online. The method is developed by Heckman.

$profit = \beta.D_{Affinity}$ It is logit or probit, the probability that you be affinity customers versus not joining affinity should become the same.

HW for next time bring examples on these three.

Cluster Analysis

Set of techniques to group similar objects.

For example for segmentation. Which customers are similar, and how many such groups are there.

In Bilogy: which plants are similar. 'Bean'.

What are other techniques available?

1. Decision trees:

2. AID: Automatic interaction Detection

3. CHAID: Chi-squre AID

The idea is to group objects. How to group objects based on similarity. The points are close to eachother but their distance is small.

This is attribute space, and the closer they are we would be able to much better group them.

Common notion is eucludian distance:

$\sqrt{(y_1 - y_2)^2 + (x_1 - x_2)^2}$

There is thing called city block distance. In which you can not go directly. In case you have special space like this you can thing of this.

Similarity of vactor:

1. $cost\theta$.

2. Mahalanobis distance. Statistical distance when factors are correlated.

It is set of techniques. Everybody is right. Which is right one to do nobody knows.

Calculate distance between all points, and groups the closest into one group.

The question is how do you calculate distance between two groups.

Distance of groups:

1. Minimum distance: Single linkage

2. Maximum distance

3. Average distance: Average linkage

4. Distance between centroids.

Output of cluster analysis will tell you:

1. How many groups to keep: 1, 2, 3, ...

2. Which group does each element blong to?

There are two kind of cluster analysis:

1. Hierarchical: Obtain a tree like diagram based on 'd'. (Dendogram). The question is how many cluster do you want, and you can cut this tree at each leve, and when the gain by cutting lower is not much you stop. It is mainly used for small samples ($< 50 obs$). If it would be thousands of observation there would be problem.

2. Non-hierarchical: We prespecify the number of clusters. Test when to stop. Iterative process will tell you how many clusters should be maintained.

Factor analysis: Groups variables that are similar

Cluster Analysis: groups observations that are similar.

Hierarchical initially used for the initial stage for exploratory, and then you go over all observation and use non hierarchical.

Criteria in selecting number of clusters:

RMSSTD should be small

SPR should be small

R^2 Should be large close to 1

Distance between two clusters should be small

Age and income are exogeneous so not good for clustering. You need to capture things that are interested for firm, such as profit, and generally behavioral variables. Probability of purchase on the website could be another interesting variable.

After doing clustering we do the discriminant analysis to find the variable to discriminates between groups.

So following are steps:

1. Do factor analysis on behavior

2. Cluster analysis (based on behavioral variables)

3. DiscriminantLogit Analysis (Use demographic and media habbit)

4. Targeting (Distinguishing characteristics)

For example AT&T churn is around 50%. We need to know how to detect churn, and what offer I should give them. We use churn and then run discriminant analysis to find out to whom I should send the promotion.

Discriminant Analysis

We want to find variables that best separate the groups. That is why the method is called discriminant analysis. We want to find the axis that best separates for example good stock from bad stock in term of their ROI.

We need to find the cutting score to classify them as good or bad.

Find discriminant function that best separates one group from the other.

Maximize Sum of squre between devided by sum of squres within the groups $\lambda = \frac{SS_b}{SS_w}$.

Each line is a new discriminant line.

Based on where they are located you many need at least, 2, or 3 for separating four groups.

Discriminant function is the line that separates the member of the two groups.

Once We get the classification function then we can do the prediction of the cluster for any new item.

Methods of Classification include:

1. Cut-Off Value Method
2. Decision Theory Approach
3. Classification Function Approach
4. Mahalanobis Distance Method

When every we have discrete dependent variables then you can not use OLS:

The reason is:

1. High hetroscedasticity, or BIAS

2. Prediction will lie between zero and one and beyond zero and one.

In this case you need to use Discriminant analysis or Logit.

Why we need two different methods?

Discriminant analysis X is distributed multivariate normal.

	1	2
1	80%	20%
2	28%	72%

In Logit we do not have that assumption.

If your dependent variable is discrete then you would not have multivariate normal, so logit would be better, since you would not have MVN distribution for them.

Before doing your homeworks read the annotated files of SAS output.

Discriminant analysis will give you the weight, and based on the weights we will classify. Actual is 1 or 0 and prediction would be 1 or 0 (say group 1 or group 0).

In SAS the commands are the following:

1. PROC CANDISC
2. PROC FACTOR
3. PROC FASTCLUS
4. PROC CLUSTER

Poisson Regression

The number of persons killed in Army (Count number of people)

You will get distribution but they would not be normal. OLS becomes problematic. Poisson regression is correct when we are dealing with 'Count Models'.

Number of hospital visits in last time.

If your goal is to predict count the poisson regression is correct

Mean and variance should be the same in the poisson distribution.

Options you have:

Poisson regression.

Negative binomial distribution. When variance is much greater than the mean.

Zero inflated regression model: when the number of zeros is high.

SAS would be:

proc genmod data=;

model . . .dist=negbin;

spend some time to look at the matrial that has been put online, at least once.

Data analysis course of professor Murthi @ UTD: Nineth session

Meisam Hejazinia

03/27/2013

Clustering variables are variables to identify behavior, and then connect them to the discriminant variables, which are demographic variables.

There could be two classification functions that try to find average score of each group. Discriminant function would be the axis by which the points are plotted.

Looking to the mean is useful to name the clausters.

Quiz would be on the material that have been covered until now. 'do loop', 'array', and macros are part of exam.

1 Logit and Probit

When variance is much larger than the mean, then you need to use negative bionomial.

If there is large proportion of zeros in data, then we use zero inflated poisson, or zero inflated negative binomial.

You are treating data as two segments, one with zero, and one with positive numbers.

Likelihood

Probability that you observe y given all the x variables.

$p(y|x) = \frac{1}{\sqrt{2.\pi.\sigma}}.exp(\frac{-(y-x\beta)^2)}{2\sigma^2})$

Likelihood: if observations are independent

$= \pi_{i=1}^{N} p_i(y|x)$

$p \in [0, 1]$

It will go to a very small value, and computer will assume it zero.

Log transformation is monotonic function. $Log - lik = \sum_{i=1}^{N} ln(p_i(y|x)) = -\frac{T}{2}ln(2.\pi) - \frac{T}{2}ln\sigma^2 - \sum \frac{(Y-X\beta)^2}{2\sigma^2}$

The assumption of observation are independent.

Maximum likelihood function (MLE) maximizes this function and finds β, σ^2 so that the likelihood function is maximized.

When we have correlated error terms then we will have $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2}, \sigma_2^2 \\ \sigma_{1,2} & \end{bmatrix}$

Coveraince could come from $cov(x, \epsilon) + cov(\epsilon_1, \epsilon_2)$, which means the first one shows 2SLS need to be used, and the second shows SUR, and if both exists would be 3SLS.

In all the methods so far we were minimizing least squares or weighted least squers. Another way of estimating is using GMM (General method of moments).

GMM is minimizing least square with weights

equal to covariances. GMM has different objective function.

If there is multiple maximum bayesian is the faster way to solve the problem since there could be only local maximum that MLE would be problematic to reach.

Autocorrelation or serial correlation is not also covered in this class. In correlation also GMM works better.

In the case of time series the elements of the large matrix would include correlation between each element of panel during time, and if between observations there is correlation with error term you will have non zero off diagonal.

Logit model:

1. Binary Logit

2. Multinomial Logit

If X follows multivariate normal distribution, means they could be correlated we can use discriminant analysis. If X is a discrete variable, yet Logit gives you better estimation.

Only when Y is discrete you will use discriminant analysis.

Examples:

1. Go prublic or stay private

2. Online versus offline purchase

3. Netflix rating $1-5$

4. Erning management or not

5. File for bankrupcy or not

6. Credit ratings AA AA

7. Brand choice

8. Churn

Why OLS is not good here?

1. Heterodescedasticity

2. Prediction

$U_1 = \alpha_1 + \beta.X_1 + \epsilon_1$

$U_2 = \alpha_2 + \beta.X_2 + \epsilon_2$

$P(1) = P(U_1 > U_2) = P(\alpha + \beta.X_1 + \epsilon_1 > \alpha_2 + \beta.X_2 + \epsilon_2) = P(\epsilon_1 - \epsilon_2 > \alpha_2 - \alpha_1 + \beta(X_2 - X_1))$

$\epsilon_1, \epsilon_2 \sim$ extreme value. Gaumbel.

Probability in logit would be in the density form.

If we assume these to be normal, mean normal error then the difference would also be normal so Probit would be the model. Would not be close form density but an integral.

Guble has fatter tails than normal distribution.

$P = \frac{1}{1+exp(\alpha+\beta.X)}$

$ln(\frac{p}{1-p}) = \alpha + \beta.X$

$\frac{p}{1-p}$ is called odds. This means log of odds would be linear. Odds means chance of winning against losing.

$p(1) = \frac{1}{1+exp(\alpha_1+\beta.X)}$

if $o = \frac{p}{1-p}$ then $p = \frac{o}{o+1}$

This saying that you can convert odds to probabilities.

$P(2) = 1 - P(1) = 1 - \frac{1}{1+exp()} = \frac{exp()}{1+exp()}$

This is logit, and to calculate it you will use proc logistic.

'class' statement in logistic procedure will tell sas that it is discrete variable.

The probability model says 'No' for pain. What ever value is coded with 1, you will use it as main probability, here was 'No'. To change this you need to use 'descending' option in the 'proc logistic'.

AIC: information criteria. SC: Shwartz criteria, and '-2 Log L' is log likelihood. If there is little improvement in likelihood compared to the intercept only, then there is not much useful covariance. The more covariance you put $log(p = 1) = 0$, generally log likelihood moves from $-\infty$ to '0'. $R^2 = \frac{}{LL_0}$. Here intercept only = 81.503, and without is 48.596, so the $R^2 = \frac{81-48}{81}$. Likelihood ratio test is to test improvement in model fit. Model fit is how good the model is. Likelihood ratio test sometimes is called (LRT).

$LRT = \frac{L_1}{L_0}$. Generally it is used for nested models. For non nested models this will not work.

(1) X_1 X_2 Y

(2) X_1 X_2 X_3 X_3

$LRT = \frac{L_2}{L_1}$

$-2(\frac{L_2}{L_1}\chi^2$ $df = N_2 - N_1$

$-2(logL_2 - logL_1)\chi^2_{df}$

AIC adds $2((k-1)+s)$ this $-2LL$. For AIC, and SC lower is better.

Schwarts criterian is also called Bayesian information criterion.

Here interpretation should be based on odds. You will for example say that $log(odd)$ for women is less than man.

Exponent of coefficient tells you odds ration after one unit increase of variable. As age increases the odd ratio will go up by 1.308.

In the lotistic regression first look at which direction you are going in the probability. Positive or negative.

Annotated output at UCLA website gives you how you can intrpret the results.

The percent concordant should be as high as possible, since it checks estimate value with real value, and if match it would be concordant, otherwise it would be discordant.

Multinomial Logit

V_1

$U_1 = \overbrace{\alpha + \beta.X_1}+\epsilon_1$

V means everything is now known to you.

$U_2 = V_2 + \epsilon_2$

$U_3 = V_3 + \epsilon_3$

$P(1) = \frac{e^{V_1}}{e^{V_1}+e^{V_2}+e^{V_3}} = \frac{1}{1+e^{V_2-V_1}+e^{V_3-V_1}}$

How many intercepts for 3 brands alternative
X = not choice specific. Choice invariant.

For example your characteristics are not choice specific. Price, promotion, and so on are on the other hand choice specific (Choice varying variables).

Suppose price is different $U_1 = \alpha_1 + \beta.Price_1$, and $U_2 = \alpha_2 + \beta.Price_2$. can allow. β is common to both brands.

If we have $\beta.Inc$ in both utility euqations, then in deduction it will cancel out. If X = choic invariant, you need to specify separate β's.

$V_1 = \alpha_1 + \beta.Inc$

$V_2 = \alpha + \beta.Inc$

To estimate this you have to put separate β for each choice to be able to estimate it.

Even for price you can put different coefficient, but it is not necessary since it will not cancel apart.

$V_1 = \alpha_1 + \beta_2.Inc$

$V_2 = \alpha + \beta_1.Inc$

Second you can only have one intercept, and the other should be zero, since the difference only could be identified and not each of 1 and α_2.

Suppose you are using 'proc logistic'. It actually calculates $log(\frac{p_2}{1}) = \alpha_1 + \beta.X$, and $log(\frac{p_3}{p_1}) = \alpha_2 + \beta.X$. Proc logistic gives you this two by two comparison. It is pairwise comparison. Mean comparison is two by two.

What we need to do is conditional logit. You need to use 'proc MDC'. This is what we call multinomial logit (MNL). The result would be $p(3) = \frac{e^{V_3}}{}$

$\epsilon - \epsilon$ would be logit

MDC requires to code this as dummy variables.

IIA: Indipendence from irrelevent alternatives. Logit has IIA problem. Suppose you hare choice of Bus, or Car. If you have two line bus. Although you may expect .25, .25, and .5, logit will give you .33 for each. Means the relation between brands is not taken into account.

This is independence between alternatives. In other word cross elasticities are same.

If the choices are not independent, then you can not use multinomial logit, due to IIA assumption. Solutions:

1. In this case you need to use MNP (multivariate normal probit). Number of integrals would be equal to the number of brands.

2. Use nested logit. You divide coffee into two, and within each category you run logit. Then you calculate probability that you are at each of the levels.

3. Use Random coefficients. Means β is not fixed but random. $\beta + \epsilon$. The random coefficient handles the correlation between brands. This is called RCL. Random coefficient logit. This is most commonly used. Barry Lebinson and Pakes (BLP) 95.

What factors really make sense, using logistic model for homework.

Data analysis course of professor Murthi @ UTD: eleventh session

Meisam Hejazinia

04/03/2013

If we have simultaneous equations with the form:

$y_1 = a_1 + b_1.y_2 + c_1.x_1$

$y_2 = a_2 + b_2y_2 + c_2x_1 + d_1x_2$

what is the effect of x_1 on y_1?

It would be in the form of $\frac{(c_1+c_2b_1)}{1-b_1b_2}$

This resulted from solving the equations.

These are indirect effects that came from other.

$AIC = -2ln(L) + 2k$ Lower would always be better. You log likelihood should improve as you add more variable, else you are getting penalized.

$SC = -2ln(L) + ln(n)k$

SBC shwartsi bayesian criterian, is the same as BIC bayesian information criterian.

R^2 would be $R^2 = 1 - \frac{L}{L_0}$

L_0 is null model with no coveriance.

L is full model.

L_0 is intercept only, or no covariance.

$Likelihood = \prod_{i=1}^{n} .P_i$

If we have $A = 36$, and $B = 64$ as market share. $P_A = (.36)^{36}$ $P_B = (.64)^{64}$ in the sample of one hundred observations.

$LL = \sum i = 1^n ln(p_i) = 36.ln(.36) + 64.ln(.64)$

AIC should be as small as possible. Akaike information criteria.

Set= $\{Bus, Car, Train\}$ Nominal variable, only identifies a category

$Set = \{A, B, C, D\}$ is called an ordinal variable. Order is known.

$service = (H, M, L)$

We will have cut off points that identifies how these could be classified.

IIA problem?

The utility of choice of one alternative does not depend on whatever the utility of other choices that are there.

$v_1 = a_1 + bp_1$

$v_2 = bp_2$

$pr(1) = \frac{e^{v_1}}{e^{v_1}+e^{v_2}}$

Own price elasticity is nothing but $\nu_{p_1} = \frac{\partial pr_1}{\partial p_1}\frac{p_1}{pr_1} = \beta.(1 - pr_1)p_1$

Crosss price elasticity would be in the form of $= \frac{\partial pr_1}{\partial p_2}\frac{p_2}{pr_1} = \beta.pr_1.p_2$

Solution for IIA:

1. Nested logit model: Start with one category and divide into multiple categories that are homogeneous. Br_1, and Br_2. Coffee (Regular(power Inst,(Br1,Br2),Instant),Decat) could be the form of nested logit. $pr(j) = p(j \in N_k).pr(j|j \in N_k)$. Error terms in this case would be GEV, mean generalized extreme value. This result in close form solution.

2. Random coefficient logit. $V_1 = a_1 + bp_1 + \epsilon_1 = \bar{a}_1 + \bar{b}p_1 + (\epsilon_1 + \gamma + \nu)$

from $\bar{a}_1 = (\bar{a}_1 + \gamma)$ $\bar{b} = \bar{b}_1 + \nu$

$\epsilon_1 \sim$ Gumbel EV

$\gamma \sim N(0, \sigma_\gamma^2)$

$\nu \sim N(0, \sigma_\nu^2)$

These should be independent. $Corr(re_b + re_c) \neq 0$ Means random error of b and c then you should use bivariate. $N(0, \Sigma)$. where Σ would be variance covariance matrix.

Random coefficients by themselves in this case create correlation between alternatives. This is shown by Hausman and Wise in 1982.

This is basically random effect in logit models.

3. Multivariate probit. Probit using multivariate normal. This will use multivariate normal, means allows correlation between alternatives.

The correlation means there is common factor that affects both error term of the first and second choice would be correlated.

For multinomial probit we have GHK estimator. This is simulation based method.

Another method for multinomial probit is Bayesian estimator.

2013 paper is FC-MNL which is flexible multinomial logit that you can use.

Independet probit has also the same problem as independent logit, means it has IIA problem.

There is a coefficient that converts probit to logit. $\beta_{probit} = k\beta_{logit}$

Tobit: Tobit's probit

Dependent variable is discrete and continues. Lots of zeros.

$wage : uneployed = 0$

$employed = cont.$

Credit card spending is the same. If you are active customer spending is positive, if you are inactive the spending would be zero.

This is called censored regression.

Below the certain treshold I am observing zeros.

All I can say about the zeros is that they have not crossed the threshold.

For purchases also when I do not buy the value would be zero. If I buy how much you would be able to see.

There is another thing called truncation.

Tobit is censored regression, but in truncation we only observe values $> c$, greater than c.

Truncation means you only observe below the threshold or above it.

Truncation tries to make the area under the curve of the distribution equal to one.

The difference is that in the censored data but all the destribution of below value will go to the zero density, and the proportionally making the area of the curve equal to one does not hold.

For truncation we will have:

$y_i = y_i^*$ if $y_i^* > 0$

c otherwise

$f(y^*|y^* > 0) =$
$\frac{c-\mu}{\sigma}$

For censored regression:

$\prod_{y_i^*>c} \frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma}) \prod_{y_i^* \leq c} \Phi(\frac{c-\mu}{\sigma})$

These are called limited dependent variable or censored dependent variable.

'Proc QLIM' is used for any kind of limited variable models.

'Proc MDC' is for conditional logit. As Mcfadden conditional logit.

The difference is that it has more than two choices.

Do this and run examples by yourself, and ask questions.

They used gradient search method.

Qlim handout shows likelihood functions. Qlim also does the ordered logit.

Bivariate: simultaenous logit, tobit, or probit. Means they are estimated jointly.

Hetroscedasticity could also happen here in limited equations.

Double censor models talks about top and bottom censoring.

There is homework due next week and submit it. It is for logit probit, and tobit.

Data analysis course of professor Murthi @ UTD: twelfth session

Meisam Hejazinia

04/10/2013

Discriminant analysis has the assumption of multivariate normal, so if we have discrete variables it is better to use logit.

ttest assumes that the variables are independent. ttest does not control for the other variables.

Heterogeneity, and endogeneity gives biased estimates, making ttest estimates bias. These make us to use these sophisticated techniques rather than simple ttests.

Factor analysis is not data reduction, but to find underlying constructs. Things that are not measurable on one dimensions.

The number of excluded variables from the equations should be greater than number of endogeneous variable minus one.

Fix one does the intercept effect, and not the coefficient effect.

In the homework that we had acquisition channels. We used dummy with respect to some base level. Here all was maseared was in comparison with internet.

Most of the time lots of data are missing, so always use 'proc freq' and 'proc mean' to find out how is the data.

You have to also focus on the big picture and know for example if the segment is profitable from t-statistics point of view, whether it is also important from the point of view of size.

In the tobit we had bias.

$\beta_{logit}/\beta_{probit}$ is 1.70.

'pico prog1.txt' helps you to edit the program.

'pwd' says where you are in the system.

Run the codes and check how it works. Proc mdc in sas should also give yout he same result.

Data analysis course of professor Murthi @ UTD: thirteen session

Meisam Hejazinia

04/17/2013

1 Survival analysis

In interpreting dummy variables you need to explain that with respect to the one that is left out.

Independent from irrelevant alternatives (IIA) means the probability of making choice only depends on their own characterstics, and not on third choices charactersitics.

$Likelihood = \prod_{i=1}^{n} \pi_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$

Log likelihood:

$\sum x_i log(\theta) + (1-x_i) log(1-\theta)$

$f(x) = \frac{1}{\sqrt{2\pi.\sigma^2}} exp(-\frac{1}{2\pi^2}(y - X\beta)^2)$

In writing the Gauss, you can ignore any constant, since it is calculating first order for the function to maximize.

$LL = -\frac{n}{2} ln(\sigma^2) + \sum[-\frac{1}{2\sigma^2}(Y - X\beta)^2] - \frac{n}{2} ln(2\pi)$

Mixture model: Person can be in state one or state 2. They may have either time to go to compare prices, or you are busy. We have evaluate state. You have time to evaluate, and second state when you do not evaluate. If you evaluate the utility would be function of prices, display, and charactersitics. If you are not evaulating. The utility function would be function of α.

You can be in first state with probability p or in other one with probability $1 - p$.

The first state given that you are in state 1 would be multinomial logit. In state 2 we will have $v_i = f(\alpha_i)$. This is intrinsic preference. You will have logit for the second state, but we do not have any covariates.

Assuming that the customer could be in any of the two states, assuming q would be probability of being in first state, and $1 - q$ probability of being in the second state. $p(1) = q.p_{1|1} + (1-q).p_{1|2}$. You need to take product of two sides, or take log likelihood to calculate this. The question was how much is the bias?

q turned out to be 60%.

$q = f(income, weekday, weekend)$

How you will put this?
You can use $q = \frac{1}{1+e^{(a+bInc)}}$, for logistic transformation, since q should between zero and one.

To implement this in Gauss you need to do the following:

You already had logv1. For the second one logv2, and you will need $v_1 = exp(\gamma_1), \ldots, v_n = exp(\gamma_n)$

You will put $q = \frac{1}{1+exp()}$.

In writing the log likelihood you will put $q.logv1 + (1-q).logv2$.

Be careful that the number of intercept only

should be $n-1$, since it works as dummy, and everything is compared with the reference.

The main modification is adding more parameters, and caclulating q, and second portion conditioning on the first.

In this case you will have 9 parameters, four we had before. two for q as intercept, and the coefficient, and 3 for three brands in the case of not searching.

Now another model (model 2):

1. Calculate for all the brands: q

2. Evaulate favorite brands: r

3. No evaluation: $1-q-r$

Latent class models:

Let us assume thre are 2 segments.

Model is (s) for segment one and $(1-s)$ for second segment. For the first segment we have display and price, but on the second one we have separate prameter of price senstitivity. This one is segmentation model. The paper for this is Kamakura and Russel JMR.

This is variation of mixture model. The way they define likelihood is changed.

Prameters and sesntivity should be different.

When you have panel data for each customer you are able to have different intercepts, and idfferent slopes. This is one way to solve unobserved hetrogeneity in multinomial logit.

If you have 4 brands you would need 9 parameters. 4 for each of the branches and one for the segment selection, so 9 would be needed. You allow more degree of freedom, and you match things better.

You test with k number of segments (1,2,3,...), and see what number better explains this market.

In calculating likelihood you need to have number of households, and their purchases. In this case you have probability that person would be in the first segment, and the probability that the person would be in the second segment. $L_{i=1} = S(A_1) + (1-S)(A_2)$. In this case you need $\prod_{i=1}^{N} L_i$

The difference with cluster analysis is that we are segmenting based on unobserved parameters.

Finally we have random coefficient. Latent class is like fixed effect. In the utility function we have error component in coefficients: $v_i = (\alpha + \epsilon_i) + (\bar{\beta} + u_i)Pr$, we estimate intercepts and betas, and additional parameters of σ^2 to check the variations.

Here we assume the errors and normal. We will have combination of logit and normal distribution.

Gaussian quadratives are solution for this estimation. In this case you need to integrate out.

Random coefficient requires integration, while latent class does not require integration.

In the next homework we will have a group for likelihood function estimation.

Bayesian method does not use gradiant method, so local maximum would not be an issue there. You should use multiple draws to make sure that it works.

Survival analysis:

Mostly it is used to model time. It is for time to an event. Bilogy.

Time is contineous, so regression should be used.

Regression assumes that error term is normal.

Second problem is sensoring.

For sensoring we can use censor regression, but normality assumption still is not solved. Therefore we need model of survival analysis.

Model time in the form of $f(\alpha + \beta.X) + \epsilon$, This model is called accelerated failiure time model (AFT).

You can also call something a hazard. Hazard function is the probability of death, given that survived until T. Hazard function=$prob(death|survived.till.T)$.

For human being hazard function has bath tube form.

This is good for insurance rates.

Proportional hazards model (PHREG).

Hazard: $prob(die|personis45)$. Here the population is only people at 45.

P(a person dies at 45) is not hazard, since the population is all the people.

$h(t)Pr(t \leq T \leq t + dt|T \geq t) = \frac{P(t \leq T \leq t+dt, T \geq t)}{P(T \geq t)} = limit_{dt \to 0} \frac{F(t+dt)-F(t)}{1-F(t)dt}$

Survival function $= 1 - F(t)$

PROC lifereg is in SAS for accelerated failiure time model, and Proportional hazard model is called PHREG.

$Log(T) = \alpha + \beta_1.Inc + \ldots + \epsilon$ where ϵ error term can be treated as exponential, weibull, but Weibull is used more. Exponential would be constant hazard model, means hazard will not increase or decrease over time. Weibull depends on time. Gombertz is another distribution that you need to write the likelihood function by yourself.

$h(t) = \alpha + \beta.Inc + \ldots$

$h(t) = h_0(t).exp(x_\beta)$ base hazard rate. $h_0(t)$ is non parametric.

$\frac{h(t|X_F)}{h(t|Xz)} = \frac{exp(x_1\beta)}{exp(X_2\beta)}$ This is Cox model, since whole insurance is built over it.

What is the meaning of β here?

Hazard goes up means time to death has gone down. These two have opposite signs.

$(exp(\beta) - 1)100 =\%$ change in hazard

Last thing is 'proc lifetest', to check difference in hazard function for two groups. For example man has more hazard than woman.

Data analysis course of professor Murthi @ UTD: Fourtheen Session

Meisam Hejazinia

04/24/2013

1 Gaussian Quadrature, and Random Coefficient Logit Model

$v_i = a_i + b.x_i + \epsilon_i$

Slope fix, but random intercept. $a_i = \bar{a_i} + u_i$ where u_i is random error term, assumed to be normally distributed with $N(0, \sigma^2_{v_i})$.

$p(i|u_i) = \frac{e^{v_i}}{\sum e^{v_i}}$

We are interested in $p(i) = \int_{-\infty}^{\infty} p(i|v_i)\phi(v_i)$ tri-variate normal, since we have three intercepts.

$\int_{-1}^{1} f(x)dx = \sum w(x)g(x)$

Gauss Hermite Quadrature

$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx = \sum w(x)f(x)$

Function $f(x)$ is logit probability $p(i|u_i)$

Book of Numerical Recipes

Here the weight would be .886227, and the integral result would be $0.886227(f(\sigma) + f(-\sigma))$.

Two or three point is enough for approximation.

Abramovitz and Stegun had provided weightings in their book. For three points it is $(1.2, 0, -1.2)$

Trapezoid rule: area calculation for integration.

Gauss Hermite is saying that instead of breaking down into trapezoids, and rectangles and calculte area, just use two points and using weight you can calculate the integral.

Whole point here is that a_i is not fixed anymore and is random, so you need integration. This integration is estimated using Gauss Hermit method.

If they were not normal, but discrete, then you would have:

a_1 with probability p_1
a_2 with probability p_2
a_3 with proabbility p_3

Then you would have weighted average with weights of probability.

This is discrete heterogeneity.

If bimodal you would use two variables with normal distribution.

You only estimate a_1, a_2, p_1, and p_2.

When you have more than two variables and they are correlated then you need to use var-covar matrix, and use contineous distribution.

At household level you integrate here.

In other word for each brand at household level we have intercept, and coefficients. These intercepts

and coefficients could be different for households, and will have normal distribution. Per household you do weighted average given what the parameter of the household would be probabilisticly. Again here you aggregate at household level and then multiply to get the likelihood.

The only parameter that is added here is variation, and you calculate integration (mean getting weighted average) using weighting that is available in books.

2 DEA and SFA

Are used to estimate the production function.

SFA stands for Stochastic Frontier Analysis. DEA is known as Data Envelopment Analysis.

We allow the frontier to be noisy

SFA $y = \alpha + \beta_1 X_1 + \beta_2 . X_3 + \epsilon$

DEA:

Multiple outputs, multiple inputs, but not stochastic.

Both are used for measuring efficiency. Efficiency analysis of retailer, post offices, and banks.

OLS captures average, and tries to capture the best.

Frontier tries to capture the best possible outcome.

The idea is how to compare with the best possible outcome.

How to estimate frontier. Get the best points and join. There would be no stochastic element.

DEA fits efficient frontier, and any point is compared with them, and talks about how much improvement I can make.

We can have two categories: best in math and best in english for example.

Sales force:

Output: (1) # orders (2) $ value orders (3) customer satisfaction
Input: (1) hours (2) salary (3) travel costs

Retailers: (1) Sales (2) Profit (3) Number of sales (4) # Customers
Input: (1) Expenses (2) Inventory (3) # employees
Non controlable factors: (1) Ad (2) Tax

We want to take all these inputs and given the controlable factors get the relation with output.

DEA calculates measure of efficiency, which is nothing but weighted average of outputs as a ratio of weighted average of inputs.

$\nu_o = \frac{w_1 . O_1 + w_2 . O_2 + w_3 . O_3}{v_1 . I_1 + v_2 I_2 + ... + v_n . I_n}$ is observable efficiency.

So that $\nu_i \leq 1n$ there is $n - 1$ constraints.

DEA tries to solve this by LP. Linear programming solves this per person.

$w_i \geq 0$

$v_i \geq 0$

Weights are calculated as optimal weights.

Efficiency: Maximum output for given input. This is called output oriented DEA.

Think of efficiency as given output for minimum inputs. This is input oriented DEA.

Output and input are data, and wieghts are parameters we estimate. The data here is cross section.

In the graph form we can say that you are moving a line in the way that every point would be at its right. Then comparing to the efficient frontier fixing one variable, you will see the slack (horizontal or vertical distance) from the optimal level of output.

This hast the assumption of constant return to scales. (CCR model).

(BCC) model allows varying return to scale.

The example here is that big firms have higher efficiency. Slacks (distance horizontally, or vertically from frontier) would be compared with different frontiers in this case. Each person is compared with different benchmark and not the same benchmark.

These are useful when you have multiple outputs that can not be compared in terms of input. For example in school output. If you had dollar values then you could go over SFA method.

The pareto optimal is equivalent to the production function frontier.

Back to SFA (Stochastic Frontier Analysis)

$\epsilon = u + v$

u is random error term

v one sided error term of efficiency

Inefficiency term v could be exponential, truncated normal or gamma distribution

Greene 1993 reviewed all possible models of SFA in his paper.

Basically it is good to do both SFA and DEA and show that results are robust.

SFA means the error is there, but DEA means there is no error there.

We try to develop bivariate SFA or multivariate SFA.

Typically using PCA and Factor analysis to summerize outputs in the index.

Random coefficients and other methods could be matched to this model.

Banker for the long time worked on SDEA (Stochastic DEA), but the result is not published.

Gaussian quadrature is talking about weighted average calculation of integrals.

Addition is easy in computer, but integration is hard, so we try to use addition rather than integration to calculate results.

You can only compare likelihood within NESTED MODELS (e.g. intercept only, and with intercept). You can not compare across models.

Interpretation and contineous variables are different. Dummy variables the comparison is with the one that is left out.

Weibull shape tells us what kind of shape it takes.

Weibull is generalization of exponential distribution.

Model fit compares model with intercept to model with covariates.

When you get only '.' in SAS means the model does not scale.

In this case one term is so huge that that the numbers are not close. In this case you need to divide by thousands (This eixts in Tobit models).

In this case Hessian could not be constructed, so standard erros will not be there.

Life test is used to compare surival of two or more groups.

Always you can do things non parametric form. In this case you need to use MLE to make your own model.

Competing risks specification: exists when you have more than one hazard. Two hazards such as obesity and smoking that can kill you. You create two hazard one for each cause.

There are independent causes, and they are competing for the same reason.

Hazard ratio is $exp(\beta) - 1$, when β would be from $PHREG$.

One MLE using GAUSS will be in exam. Code will be given so that you change the relevent portion.

You need to know basic SAS.

Interpretation will be in the exam. Precise interpretation. When it is dummy always compare with out option. When continuous talk about how much increase or decrease.

We want to know when to use each of the models.

In logit and probit always think in term of elasticities. You can convert betas to elasticities.

When you have mixture model or latent model. How to calculate elasticity?

For price elasticity calculation. You put values inside the price for example and calculate how final probability is changed and integrate it (in the form of summation). In other word you do numerical integration. For example you say 5% increase. In this case you multiply to 1.05 the price, and then check what would be the change in the probability. Numerical integration will give you the average change.

Data analysis course of professor Murthi @ UTD: Fourtheen Session

Meisam Hejazinia

05/01/2013

1 Gaussian Quadrature, and Random Coefficient Logit Model

R^2 tells you which model is better, since log likelihood only is comparable with respect to intercept only model.

You should know under which condition you used these tests. When t-test, regression, factor, cluster, chi-square is important to know, and will be asked in the exam.

We will talk about SAS code, log likelihood, and Gauss for three last homeworks will mainly be asked in the exam.

Therefore last topic will only be asked in detail.

In logit homeworks, income number of members, and kids, these values will not moge much.

$P_1 = \frac{e^{V_1}}{e^{V_1}+e^{V_2}} = \frac{1}{1+e^{V_2-V_1}}$

In non choice specific covariates you have to have two separate coefficients to identify.

You need to assume different coefficients for different brands. For identification purposes you need to only have one of the coefficients. All esle will be compared with this one.

In this case you say effect of income depends on which brand you purchase.

This says elasticity of income is different with respect to the choice.

Mixture model is nested for the normal model. Here $-log(1)$ is zero, so going toward zero is better.

Print the values in Gauss:

$P1;;D1;;F;;L$ to printing the data.

Couple of things to check in Guass:

1. Variance should be positive. To make sure it is positive take the square root, and then square it, and you make sure that it is positive this way.

Variance should be positive definite or positive semidefinite.

$\Sigma = C'C$, where 'C' is known as choleskey decomposition.

To make sure that a prameter is always positive you put $\beta = \frac{e^a}{1+e^a}$, which is logistic form. In this case you will estimate a. This helps you to make sure that $\beta \in (0, 1)$