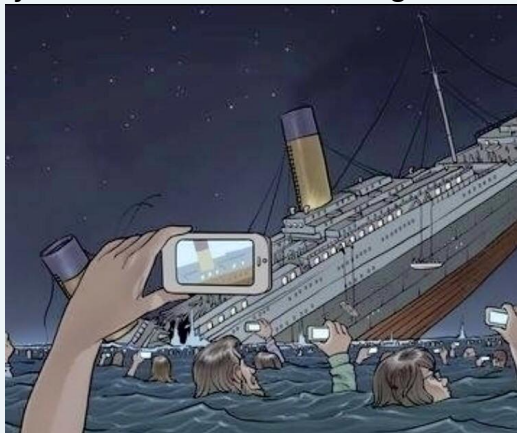


Would you have survived the sinking of the Titanic?



Econometrics: Computer Modelling

Felix Pretis

Institute for New Economic Thinking at the
Oxford Martin School, University of Oxford

Lecture 2: Micro-Econometrics:
Limited Dep. Variable Models

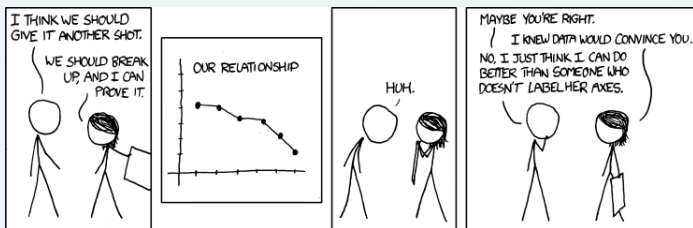
- 1: Intro to Econometric Software & Cross-Section Regression
- 2: **Micro-Econometrics: Limited Indep. Variable**
- 3: Macro-Econometrics: Time Series

Last time:

- Introduce econometric modelling in practice
- Introduce OxMetrics/PcGive Software

Today:

- Binary dependent variables & Count data
 - Probability of being accepted into a Masters/PhD Programme (between $[0,1]$)
 - Number of arrests (count)
 - Prob. of surviving the Titanic sinking, participating in labour force
- Additional functions in OxMetrics/PcGive



- Economies high dimensional, interdependent, heterogeneous, and evolving: comprehensive specification of all events is impossible.
- Economic Theory
 - likely wrong and incomplete
 - **meaningless** without empirical support
 - Econometrics to discover new relationships from data
 - Econometrics can provide empirical support. . . or refutation.

Structure of data

	admit	gre	gpa	rank	rank1	rank2	rank3	rank4
1	0	380	3.61	3	0	0	1	0
2	1	660	3.67	3	0	0	1	0
3	1	800	4	1	1	0	0	0
4	1	640	3.19	4	0	0	0	1
5	0	520	2.93	4	0	0	0	1
6	1	760	3	2	0	1	0	0
7	1	560	2.98	1	1	0	0	0
8	0	400	3.08	2	0	1	0	0
9	1	540	3.39	3	0	0	1	0
10	0	700	3.92	2	0	1	0	0
11	0	800	4	4	0	0	0	1
12	0	440	3.22	1	1	0	0	0
13	1	760	4	1	1	0	0	0

Data on admission to graduate school (US) as a function of:

- GPA
- GRE score
- Rank of undergraduate institution

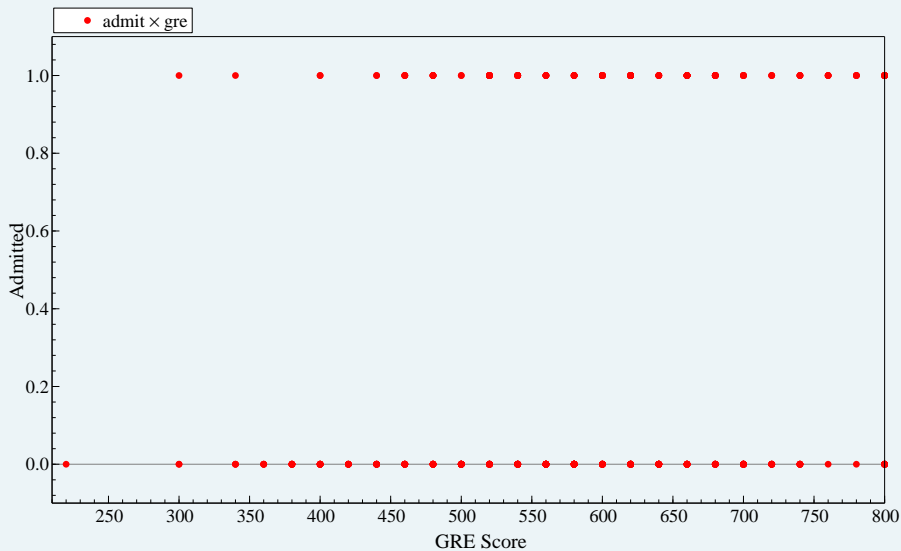
Dataset: "gradschool.xlsx"

Datasets: .in7 & .bn7 files

Other file formats?

Build a **Linear Probability Model** (LPM) for gradschool admission:

- Construct appropriate variables to take the rank of the university into account (in Excel)
- Create a new database in OxMetrics
 - Open OxMetrics
 - Go to File, New, OxMetrics Data
 - Set start period = 1 (undated for cross-sectional data)
 - Copy & Paste Data from Excel file: "gradschool.xlsx"
 - Save as .in7 data file on your computer
- Plot the observed and predicted values against GRE. Based on the model output highlight shortcomings of the LPM



Fit a simple Linear probability model (OLS)

EQ(1) Modelling admit by OLS-CS

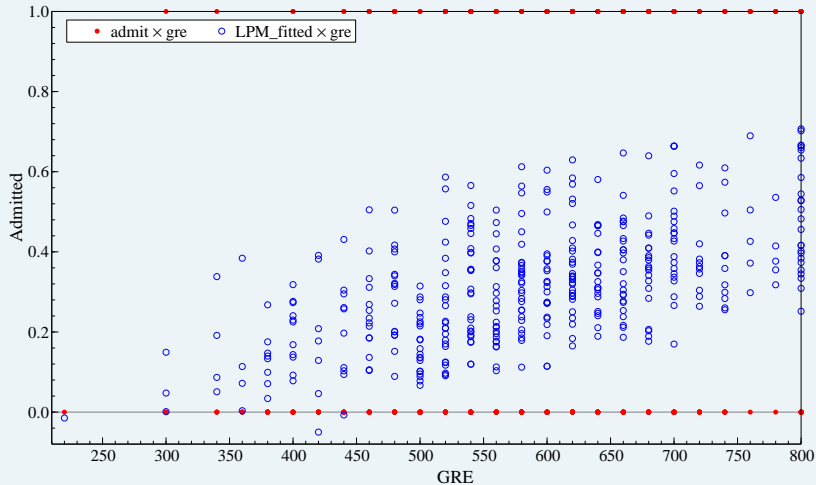
The dataset is: gradschool.in7

The estimation sample is: 1 - 400

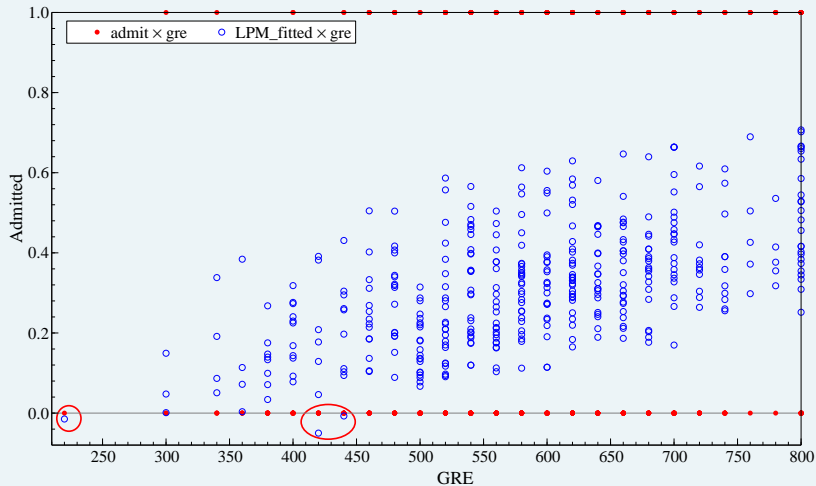
	Coefficient	Std.Error	t-value	t-prob	Part.R ²
Constant	-0.258910	0.2160	-1.20	0.2314	0.0036
gre	0.000429572	0.0002107	2.04	0.0422	0.0104
gpa	0.155535	0.06396	2.43	0.0155	0.0148
rank2	-0.162365	0.06771	-2.40	0.0170	0.0144
rank3	-0.290570	0.07025	-4.14	0.0000	0.0416
rank4	-0.323026	0.07932	-4.07	0.0001	0.0404
sigma	0.444866	RSS		77.9750245	
R ²	0.100401	F(5,394) =	8.795	[0.000]**	
Adj.R ²	0.0889844	log-likelihood		-240.56	
no. of observations	400	no. of parameters		6	
mean(admit)	0.3175	se(admit)		0.466087	
Normality test:	Chi ² (2) =	212.46	[0.0000]**		
Hetero test:	F(7,392) =	3.8513	[0.0005]**		
Hetero-X test:	F(8,391) =	3.6183	[0.0004]**		
RESET23 test:	F(2,392) =	0.19773	[0.8207]		

Issues with Linear Probability Model

- Assumes continuous dep. variable & constant effect of covariates on probability of success
- Predicted values outside $[0,1]$ range: *Test - Store...*
- Heteroskedasticity by construction



Fitted Values of LPM



$$P(y = 1|x) = G(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\beta) \quad (1)$$

Logistic Regression:

$$P(y = 1|x) = \frac{\exp(\beta_0 + \mathbf{x}\beta)}{1 + \exp(\beta_0 + \mathbf{x}\beta)} \quad (2)$$

- Maximum Likelihood Estimation
- No analytical solution

1) Log-Odds Ratio

Note that the odds ratio (probability of success over probability of failure) in the logit model is given as:

$$\frac{P(y = 1|x)}{1 - P(y = 1|x)} = \exp(\beta_0 + \mathbf{x}\beta) \quad (3)$$

Therefore, taking logs:

$$\log\left(\frac{P(y = 1|x)}{1 - P(y = 1|x)}\right) = \beta_0 + \mathbf{x}\beta \quad (4)$$

Thus, $100 \times \beta_k$ has the interpretation as % increase in odds ratio for a one-unit increase in x_k

2) Marginal Effects (ME)

$$\frac{\partial P(y = 1|x)}{\partial x_k} = \frac{\partial}{\partial x_k} \left(\frac{\exp(\beta_0 + \mathbf{x}\beta)}{1 + \exp(\beta_0 + \mathbf{x}\beta)} \right) \quad (5)$$

$$= \beta_k P(y = 1|x) (1 - P(y = 1|x)) \quad (6)$$

- ME_k same sign as coefficient β_k
- Marginal effects are largest when $P = 0.5$, i.e. largest for individuals whose outcomes have the highest variance.

- *Models for Discrete Data*
- *Binary Discrete Choice using PcGive*
- *Logit*
- *Newton's Method* (solution algorithm)

What are the effects of rank, gpa, gre, on the probability of being admitted to Grad School?

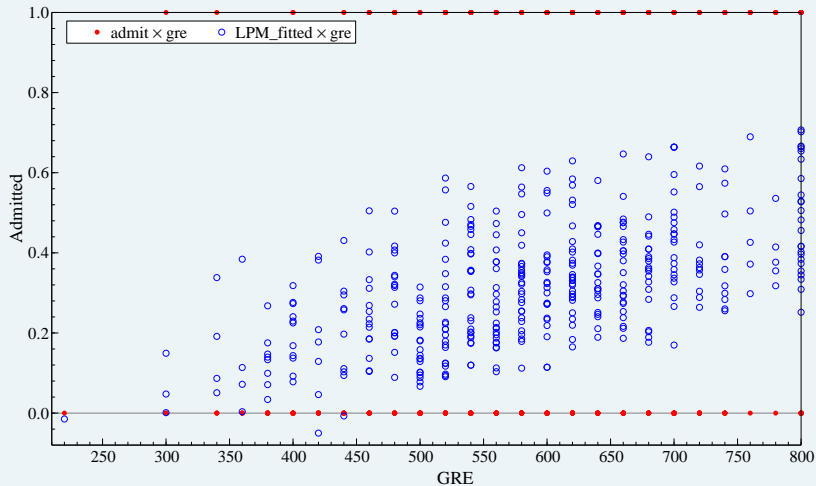
CS(1) Modelling admit by LOGIT
The dataset is: gradschool.in7
The estimation sample is 1 - 400

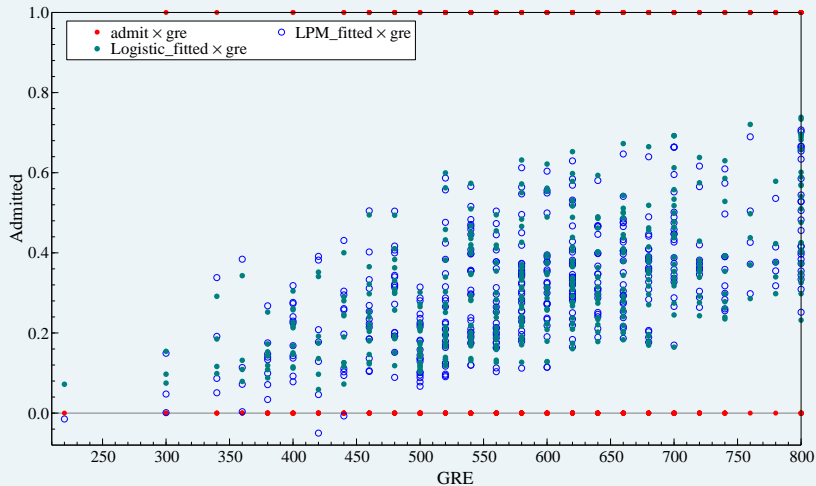
	Coefficient	Std.Error	t-value	t-prob
Constant	-3.98998	1.140	-3.50	0.001
gre	0.00226443	0.001094	2.07	0.039
gpa	0.804038	0.3318	2.42	0.016
rank2	-0.675443	0.3165	-2.13	0.033
rank3	-1.34020	0.3453	-3.88	0.000
rank4	-1.55146	0.4178	-3.71	0.000

log-likelihood	-229.258746	no. of states	2
no. of observations	400	no. of parameters	6
baseline log-lik	-249.9883	Test: Chi ² (5)	41.459 [0.0000]**
AIC	470.517492	AIC/n	1.17629373
mean(admit)	0.3175	var(admit)	0.216694

Newton estimation (eps1=0.0001; eps2=0.005): Strong convergence

	Count	Frequency	Probability	loglik
State 0	273	0.68250	0.68250	-97.40
State 1	127	0.31750	0.31750	-131.9
Total	400	1.00000	1.00000	-229.3





Replicability is important!

Easy to make mistakes/forget what you have done. Code to reproduce your modelling:

- Batch code (intuitive code, similar to STATA do-files)
- Ox code (matrix programming language, similar to Matlab, R)

Batch Code:

- .fl files
- ALT+B: Batch code for last model

```
//Lecture 2: Micro-Econometrics, Limited Dependent Variable

//Linear Probability Model

module("PcGive");
package("PcGive", "Cross-section");
usedata("gradschool.in7");
system
{
    Y = admit;
    Z = Constant, gre, gpa, rank2, rank3, rank4;
}
estimate("OLS-CS", 1, 1, 400, 1);

//Logistic Regression Model

module("PcGive");
package("LogitJD", "Binary");
usedata("gradschool.in7");
system
{
    Y = admit;
    X = gre, gpa, rank2, rank3, rank4;
    F = Constant;
}
estimate("LOGIT", 1, 1, 400, 1);
```

Using Batch Code, estimate and store the following models for Gradschool admissions:

- A linear probability model without an intercept
- A logistic regression without the GPA variable and using observations for the individuals $i = 50, \dots, 200$.
- In the form of comments in batch code, add the results of a test that all rank variables can be dropped from the model.

So far: Binary dependent variable $[0,1]$

Now: Count data – Poisson regression



- Dependent Variable: non-negative integers 0,1,2...
- $y \sim \text{Poisson}(\mu)$

Example:

- y = number of vehicle collisions in winter
- m = number of cars in Oxford (exposures)

$$E[y_i|x] = \mu_i = m_i\theta_i \text{ (where } \theta_i \text{ is the rate)} \quad (7)$$

$$E[y_i|x] = m_i \exp(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i}) \quad (8)$$

$$\log(E[y_i|x]) = \log(m_i) + \beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i} \quad (9)$$

Estimation using Maximum Likelihood:

$$y_i = E[y_i|x_i] + u_i \quad (10)$$

$$y_i = e^{x_i'\beta} + u_i \quad (11)$$

Interpretation:

- Approx: $100\beta_k \Delta x_k \approx \% \Delta E[y_i|x]$
- Exact proportional change: $\exp(\beta_k \Delta x_k) - 1$

Modelling Number of Arrests:

- Number of times a man is arrested in 1986: 'narr86'
- "arrests.in7"
- Plot the data!

Poisson Regression:

- *Models for Discrete Data*
- *Count Data using PcGive*

Model:

- Dependent variable: "narr86"
- Independent variables:
 - "pcnv" (prop. of prior arrests that led to conviction)
 - "avgsen" (avg sentence length)
 - "tottime" (time in prison since 18)
 - "ptime86" (months spent in prison)
 - "qemp86" (quarters employed)
 - "inc86" (income)
 - "black", "hispan"

CS(1) Modelling narr86 by POISSON
 The dataset is: arrests.in7
 The estimation sample is 1 - 2725

	Coefficient	Std.Error	t-value	t-prob
Constant	-0.617178	0.06365	-9.70	0.000
pcnv	-0.405258	0.08488	-4.77	0.000
avgsen	-0.0236365	0.01993	-1.19	0.236
totttime	0.0243425	0.01476	1.65	0.099
ptime86	-0.0985944	0.02071	-4.76	0.000
qemp86	-0.0361131	0.02892	-1.25	0.212
inc86	-0.00814627	0.001038	-7.85	0.000
black	0.660356	0.07383	8.94	0.000
hispan	0.499594	0.07392	6.76	0.000
log-likelihood	-2249.08013	not truncated		
no. of observations	2725	no. of parameters		9
baseline log-lik	-2441.921	Test: Chi ² (8)		385.68 [0.0000]**
AIC	4516.16026	AIC/n		1.65730652
mean(narr86)	0.404404	var(narr86)		0.737742

- Store the batch code as ".fl" file.
- Re-run the model having dropped the least significant variable using batch code.

What is the effect of being black/hispanic on the number of arrests?

Surviving the Titanic

What is your estimated probability of survival?



- Create variables that measure the cabin class (& clean data)
- Create a new database using "titanic_data.csv"
- Estimate the probability of survival ("survived") using
 - Cabin class
 - "sex": =1 if female
 - "age": in years
 - "num_sibs_sp": number of siblings or spouses on board
 - "num_par_ch": number of parents or children on board



Answering the following questions:

- What model is most appropriate for the dataset at hand?
 - Estimate models using **three** alternative methods and compare the results
 - Create a batch file to run your models.
 - Create plots illustrating your results.
- What is the unconditional probability of survival?
- What is the effect of cabin class/sex/age/having siblings or kids on-board on the probability of survival? How can the coefficients be interpreted? What difference do you find between the two methods used?
- What is your personal probability of survival for your assigned cabin class, given that you assume your parents were not on board, but your siblings/spouses (if you have any) would have been?

OxMetrics and PcGive Exercise: Female labour force participation.

- Create a new database using "labourforce.xlsx"
- "inlf" binary variable =1 if married woman in labour force in 1975.

hours	hours worked, 1975
kidslt6	# kids < 6 years
kidsge6	# kids 6-18
age	woman's age in yrs
educ	years of schooling
wage	estimated wage from earns., hours
repwage	reported wage at interview in 1976
hushrs	hours worked by husband, 1975
husage	husband's age
huseduc	husband's years of schooling
huswage	husband's hourly wage, 1975
faminc	family income, 1975
mtr	fed. marginal tax rate facing woman
motheduc	mother's years of schooling
fatheduc	father's years of schooling
unem	unem. rate in county of resid.
city	=1 if live in city
exper	actual labor mkt exper

Answering the following questions:

- Estimate models using two alternative methods and compare the results (create a batch file).
- What is the effect of age/educ/experience/having kids on the probability of being in the labour force? What difference do you find between the two methods used?
- Allow for diminishing marginal returns to experience. What are your findings?
- Build a more general model, including additional covariates. Which ones are significant? How could you reduce the number of variables?
- Using batch code, sequentially eliminate variables based on their significance (conduct backwards-elimination). What other model selection methods could you use? Advantages/disadvantages?
- Classification: Hold back 200 observations, predict the labour force participation for the hold-back sample. What proportion are correctly classified? Build a model that achieves the highest classification rate.