

*Modelling Tobacco
Consumption with a Double
Hurdle Model for Ordered Data*

Mark Harris

&

Xueyan Zhao

**Dept. of Econometrics & Business
Statistics**

Monash University

Introduction

- Often in economics we are interested in modelling a **discrete ordinal** dependent variable:
 - Levels of household saving
 - Survey responses on opinions
 - Health status levels
 - Obesity levels
 - Employment status levels
 - Bond ratings
 - Job classifications by skill level
 - And so on!
- Typically how would we model this?

Introduction



1. Ordered Probit/Logit
 - i. **Only one latent variable** →
 - ii. Model restrictive
 - iii. Inconsistent with RUM

Introduction

2. OGEV

- i. **Several latent variables** with correlated errors
- ii. Consistent with RUM
- iii. **But**, weak identification (?) → problems in estimation
- iv. **And** same variables in each equation →
- v. Not common in the literature
- vi. Empirical example: Harris, Ramful and Zhao
forthcoming JHE

Introduction

- However, such data is often characterised by a build-up of “zero” observations
 - Zero consumption
 - “No change”
 - i.e., build-up is typically at one end of the scale
- Moreover, often such “zero” observations, **may come from two distinct sources**

Introduction

- Take a demand equation; “zero” may relate to:
 - An infrequent purchaser not purchasing in the period in question (income/price effects)
 - And from non-participants
- In labour supply; zero hours worked may relate to:
 - Unemployed individuals
 - And those not in labour force

Introduction

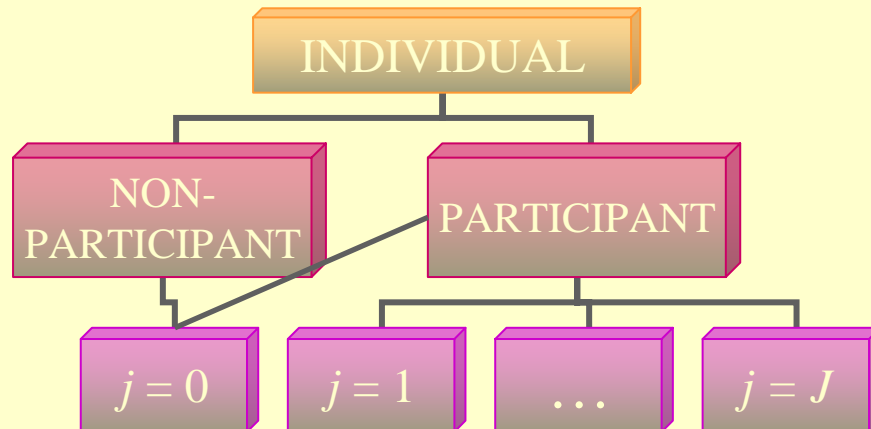
- In a health status model, a zero may relate to:
 - Someone who is inherently/genetically unhealthy
 - And someone who is unhealthy due to lifestyle factors
- **In each example, these relate to two distinctly different sets of individuals**

Introduction

- Why is this important?
- For policy purposes it's potentially very important to split the observed zeros
 - Labour supply and unemployment benefits: *positive* effect on unemployment, but negative on NILF
 - *E.g.* smoking and income: is likely to exert a *negative* effect on participation and *positive* on the amount consumed
- And, in general, you'll be estimating a misspecified model

Generalising the Ordered Probit

- So, graphically we have:



A “Selection” or “Participation” Equation

- In other words we first have a selection equation
- Consider two Regimes, $r = 0$ (“non-participation”) and 1 (“participation”)
 - non-smoker/smoker
 - non-labour market participant/participant
 - inherently unhealthy/inherently not unhealthy

An “Amount” or “Intensity” Equation

- Conditional on $r = 1$ (Regime 1) consumption levels under Regime 1 are represented by the usual ordered probit formulation
- So, implicitly we have two latent variables

$$\text{Probit : } r^* = x'\beta + \varepsilon, \quad \varepsilon \sim N(0,1)$$

and

$$\text{OP : } y^* = z'\gamma + u, \quad u \sim N(0,1)$$

Generalising the Ordered Probit

- So, analogously to double-hurdle models
 - to observe a positive amount of consumption ($y > 0$), the individual must overcome **two** hurdles
 1. $r = 1$; the individual is a “participant”
 2. **And** $y^* > 0$; the underlying latent propensity to consume, for participants, is positive

Generalising the Ordered Probit

- Moreover, are good reasons to expect that different variables affect the different equations
 - Price and income are likely to dominate the “demand” equation, y^*
 - Demographics likely to dominate “participation” equation, r^*
- Or, the same variables with different effects in each latent equation

A Zero Inflated Ordered Probit Model

- Following the set-up of the Zero-Inflated count models (ZIPs and ZAPs), we can formulate the probabilities as:

$$\Pr_j = \begin{cases} \Pr(y = 0|z, x) = [1 - \Phi(x'\beta)] + [\Phi(x'\beta)\Phi(-z'\gamma)] \\ \Pr(y = 1|z, x) = \Phi(x'\beta)[\Phi(\mu_1 - z'\gamma) - \Phi(-z'\gamma)] \\ \Pr(y = 2|z, x) = \Phi(x'\beta)[\Phi(\mu_2 - z'\gamma) - \Phi(\mu_1 - z'\gamma)] \\ \vdots \\ \Pr(y = J|z, x) = \Phi(x'\beta)[1 - \Phi(\mu_{J-1} - z'\gamma)] \end{cases}$$

- So, equivalently could be termed a Zero Inflated Ordered Probit (ZIOP) model, or a Double Hurdle Model for Ordered Data

A Correlated Zero Inflated Ordered Probit Model

- But, as these two equations are for the **same individuals** → the two unobserved elements are likely to be **correlated** →
- Probabilities are now **bivariate** normal cdf's:

$$\Pr_j = \begin{cases} \Pr(y = 0|z, x) = [1 - \Phi(x'\beta)] + \Phi_2(x'\beta, -z'\gamma, -\rho) \\ \Pr(y = 1|z, x) = \Phi_2(x'\beta, \mu_1 - z'\gamma, -\rho) - \Phi_2(x'\beta, -z'\gamma, -\rho) \\ \Pr(y = 2|z, x) = \Phi_2(x'\beta, \mu_2 - z'\gamma, -\rho) - \Phi_2(x'\beta, \mu_1 - z'\gamma, -\rho) \\ \vdots \\ \Pr(y = J|z, x) = \Phi_2(x'\beta, z'\gamma - \mu_{J-1}, -\rho) \end{cases}$$

Estimation & Hypothesis Testing

1. Estimate by standard Maximum Likelihood techniques
2. Testing correlated ZIOP *versus* ZIOP
 - Easy! t -test of $\rho = 0$
3. Standard Wald and Likelihood Ratio tests for individual and joint significance of variables
4. ZIOP *versus* OP
 - Much harder!!
 - Under null hypothesis of OP, ZIOP is not identified
 - Try LR, Vuong non-nested test, Hausman, Information Criteria

Monte Carlo Evidence: OP True Model

- As this is a new model, we check to see if we're getting it right by generating some data...
- Firstly we generate according to a OP specification, but estimate a ZIOP one:
- If we use an Information Criteria approach, we'll correctly choose OP when true, virtually every time

Monte Carlo Evidence: ZIOP

True Model



- Next generate under zero-inflated ordered probit, and again we do fine:
 1. Closely estimate the “unknown” true parameters
 2. OP parameters are way off
 3. Using information criteria will always select ZIOP model (although does have difficulties differentiating between ZIOP and ZIOPC)
- That is, again will correctly accept ZIOP if true

Exclusion Restrictions?

- Also tried the same variables in x and z
- Sensible model??
- In any case, ZIOP still works well...

An Application: Tobacco Consumption

- Huge amount of literature on smoking is vast – a brief summary:
- Consumption is a natural two step process
 - Participation
 - Conditional, then how much?
- Deal with these in turn...

An Application: Tobacco Consumption

1. Participation:

- Studies have focused on effects of:
 - Parental smoking behaviour
 - Family background
 - Socio-economic status
 - Standard demographics
- Also much on decision to start smoking amongst teenagers (most “at risk” time)
 - In particular, recent rise in participation rates of young females

Literature Review

2. Consumption levels

- *i.e.*, how much people smoke?
- Literature largely focused on addictive nature of tobacco

Literature Review

- Psychological part of a “script”
- Script is a set of patterns which are used to guide behavior *i.e.* a “utility”
- Might be
 - tobacco and alcohol
 - the post modern
 - Or might be



consumption as

consumption
g quality

Literature Review

- Economists traditionally measure addiction through price inelastic demand schedules
- More recent approach – Becker and Murphy's (1988) theory of rational addiction
 - Current levels of addiction primarily affected by stock of past consumption (sort of “ratchet-effect”)
 - Essentially necessitates a lagged dependent variable in estimations *e.g.* Becker and Stiger (1977) and Chaloupka (1991)
- Can approximate this effect by judicious use of “age” variables

The Data

- We use the *Australian National Drug Strategy Household Survey*
 - Years 1995, 1998 and 2001
- Dependent variable takes the form:
 - Not currently smoking ($y = 0$):
 - ◆ Non-participants;
 - ◆ Recent quitters
 - ◆ Infrequent smokers
 - ◆ Under-reporters
 - ◆ Potential smokers
 - Smoking less than daily ($y = 1$)
 - Less than 20 a day ($y = 2$)
 - 20+ a day ($y = 3$)

What Variables to Include in “Participation” (x) & “Amount” (z)?

- Standard demographics in both except:
- Age in participation
 - Age (-)
 - Most smokers start at a young age but older individuals more likely to quit
 - Ln(age)
- Age in amount
 - “n”-shaped
 - Addiction built up in informative years, consumption reduces at older ages
 - Include age and age²
 - Proxies *rational addiction*

What Variables to Include in “Participation” (x) & “Amount” (z)?

- Income in both equations
 - Demand schedule reasons for amount
 - Social class proxy for participation
- Own price and cross-drug prices (alcohol and marijuana) only in amount equation

Results: General

- Prices were only significant in “amount” equation and not in participation decision
 - All negative (marijuana, alcohol and tobacco)
 - Suggesting tobacco is a normal good and that these drugs are compliments (in line with previous evidence)
- ρ was small, negative and insignificant:
- Negative ρ ? Might be picking-up the effects of parental smoking status
 - +tve on participation, but –tve on amount?
- Other ZIOPC parameters v. close to ZIOP
- Full results in paper, but let’s focus on a couple of interesting marginal effects...

Marginal Effects: Probit vs ZIOPC

	Probit	ZIOPC
	P(Y>0)	P(r=1)
Ln(Income)	0.0139	-0.0303
Age	-0.0061	-0.0169
Young Female	-0.0583	0.1733
Male ×1	0.0183	0.0943
Married ×1	-0.0816	-0.1593
Pre-School ×1	0.0119	-0.0537
Capital ×1	-0.0065	0.0132
Work ×1	0.0025	0.0136
Unemployed ×1	0.1285	0.0605
Study ×1	-0.0983	0.1885
English Speaking ×1	0.0458	0.0611
Degree ×1	-0.1497	-0.0791
Diploma ×1	-0.0382	-0.0283
Year12 ×1	-0.0557	-0.0195
School ×1	-0.1734	0.0020

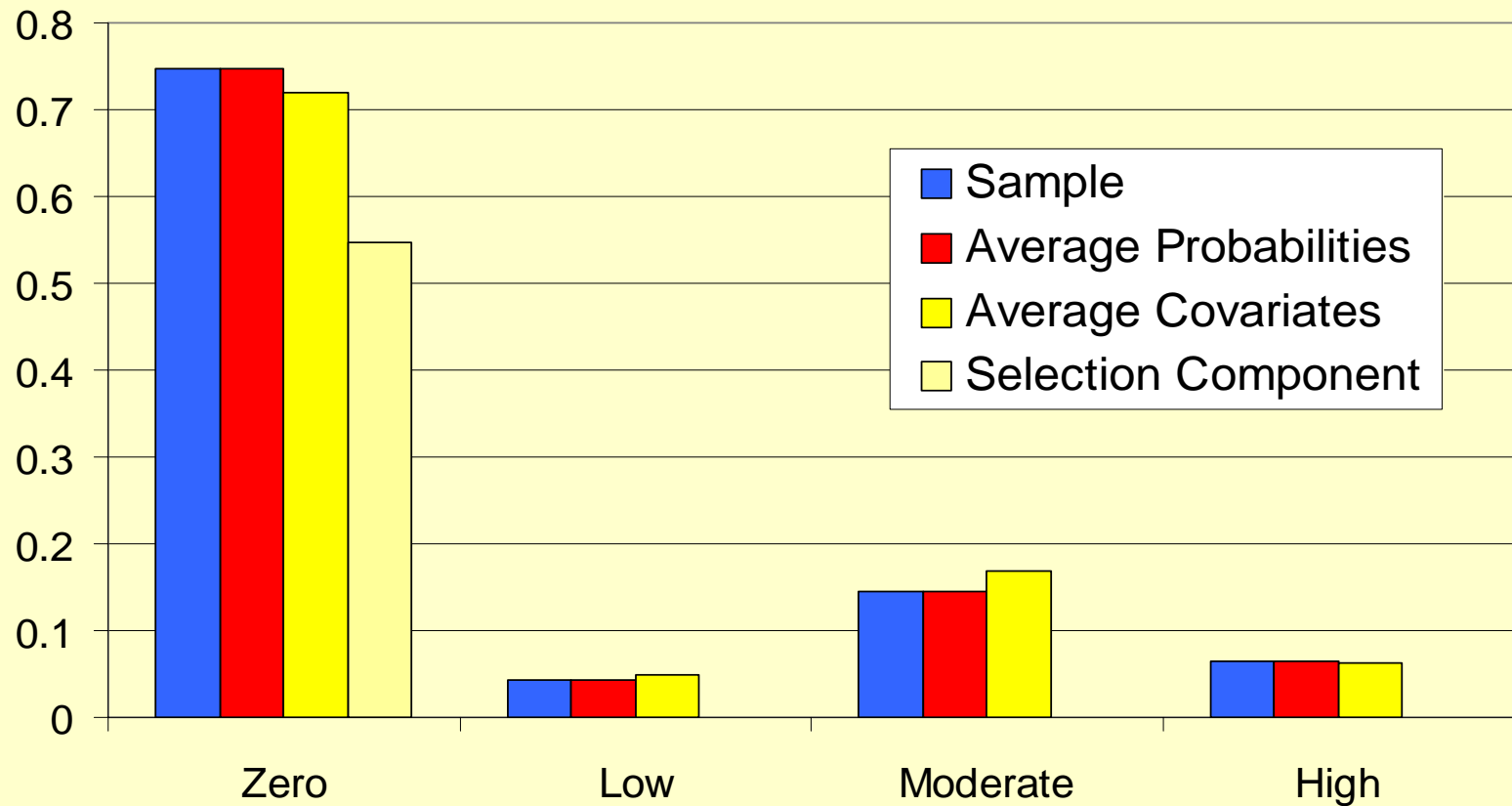
MEs: Ordered Probit vs ZIOPC

	OP	ZIOPC		
	Y=0	Y=0		
		Non- Participation	Zero Consumption	Full
Male ×1	-0.0478	-0.0183	-0.0561	-0.0744
Married ×1	0.1007	0.0816	0.0424	0.1241
Pre--School ×1	0.0091	-0.0119	0.0469	0.0350
Capital ×1	0.0126	0.0065	0.0031	0.0097
Work ×1	0.0542	-0.0025	0.0379	0.0354
Unemployed ×1	-0.0454	-0.1285	0.0685	-0.0600
Study ×1	0.1110	0.0983	-0.0737	0.0246
English Speaking ×1	-0.0532	-0.0458	-0.0183	-0.0641
Degree ×1	0.1617	0.1497	0.0483	0.1980
Diploma ×1	0.0465	0.0382	0.0260	0.0642
Year 12 ×1	0.0658	0.0557	0.0252	0.0810
School ×1	0.1353	0.1734	-0.0781	0.0953
Young Female	-	-0.1733	0.0629	-0.1104
Ln(P _A)	0.3217	-	0.2979	0.2979
Ln(P _M)	0.0020	-	-0.0036	-0.0036
Ln(P _I)	0.1555	-	0.1499	0.1499
Ln(Income)	-0.0030	0.0303	-0.0169	0.0134
Age	-	0.0169	0.0086	0.0254

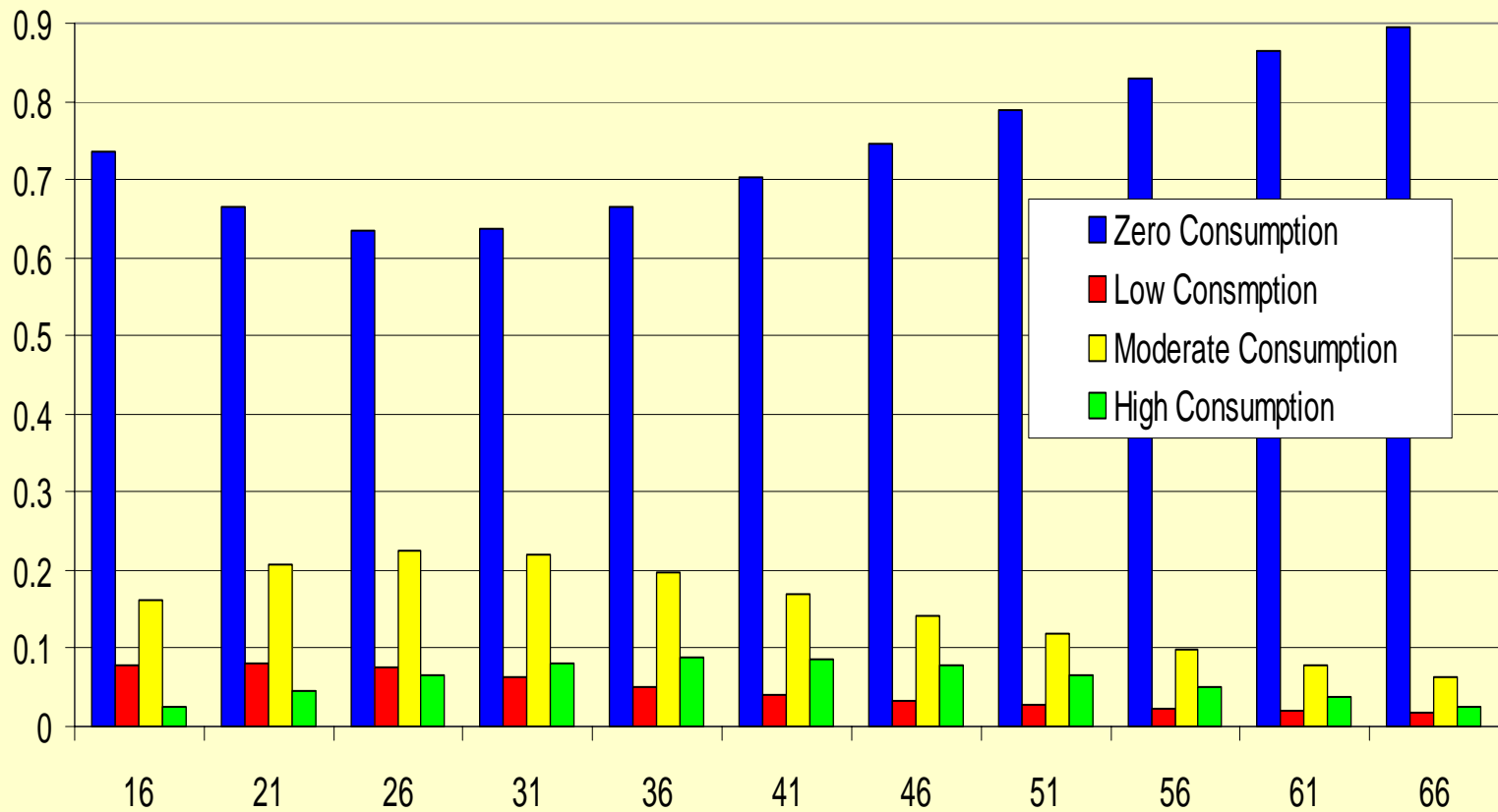
MEs : Ordered Probit vs ZIOPC

	OP	ZIOPC	OP	ZIOPC	OP	ZIOPC
	Y=1		Y=2		Y=3	
Male ×1	0.0056	0.0094	0.0261	0.0428	0.0161	0.0222
Married ×1	-0.0114	-0.0159	-0.0545	-0.0713	-0.0347	-0.0368
Pre--School ×1	-0.0011	-0.0057	-0.0050	-0.0209	-0.0030	-0.0085
Capital ×1	-0.0015	0.0021	-0.0069	-0.0035	-0.0042	-0.0083
Work ×1	-0.0062	0.0033	-0.0295	-0.0154	-0.0185	-0.0234
Unemployed ×1	0.0049	0.0053	0.0243	0.0327	0.0161	0.0220
Study ×1	-0.0155	0.0165	-0.0634	-0.0089	-0.0322	-0.0322
English Speaking ×1	0.0069	0.0061	0.0298	0.0362	0.0165	0.0218
Degree ×1	-0.0219	-0.0115	-0.0914	-0.1103	-0.0485	-0.0763
Diploma ×1	-0.0057	-0.0020	-0.0257	-0.0340	-0.0151	-0.0282
Year 12 ×1	-0.0083	-0.0016	-0.0367	-0.0433	-0.0208	-0.0360
School ×1	-0.0203	-0.0024	-0.0784	-0.0536	-0.0367	-0.0392
Young Female	-	0.0183	-	0.0660	-	0.0261
Ln(P _A)	-0.0379	0.0108	-0.1765	-0.1432	-0.1073	-0.1655
Ln(P _M)	-0.0002	-0.0001	-0.0011	0.0017	-0.0007	0.0020
Ln(P _I)	0.0183	0.0054	0.0853	0.0721	0.0518	0.0833
Ln(Income)	0.0004	-0.0034	0.0016	-0.0087	0.0010	-0.0013
Age		-0.0012		-0.0135		-0.0107

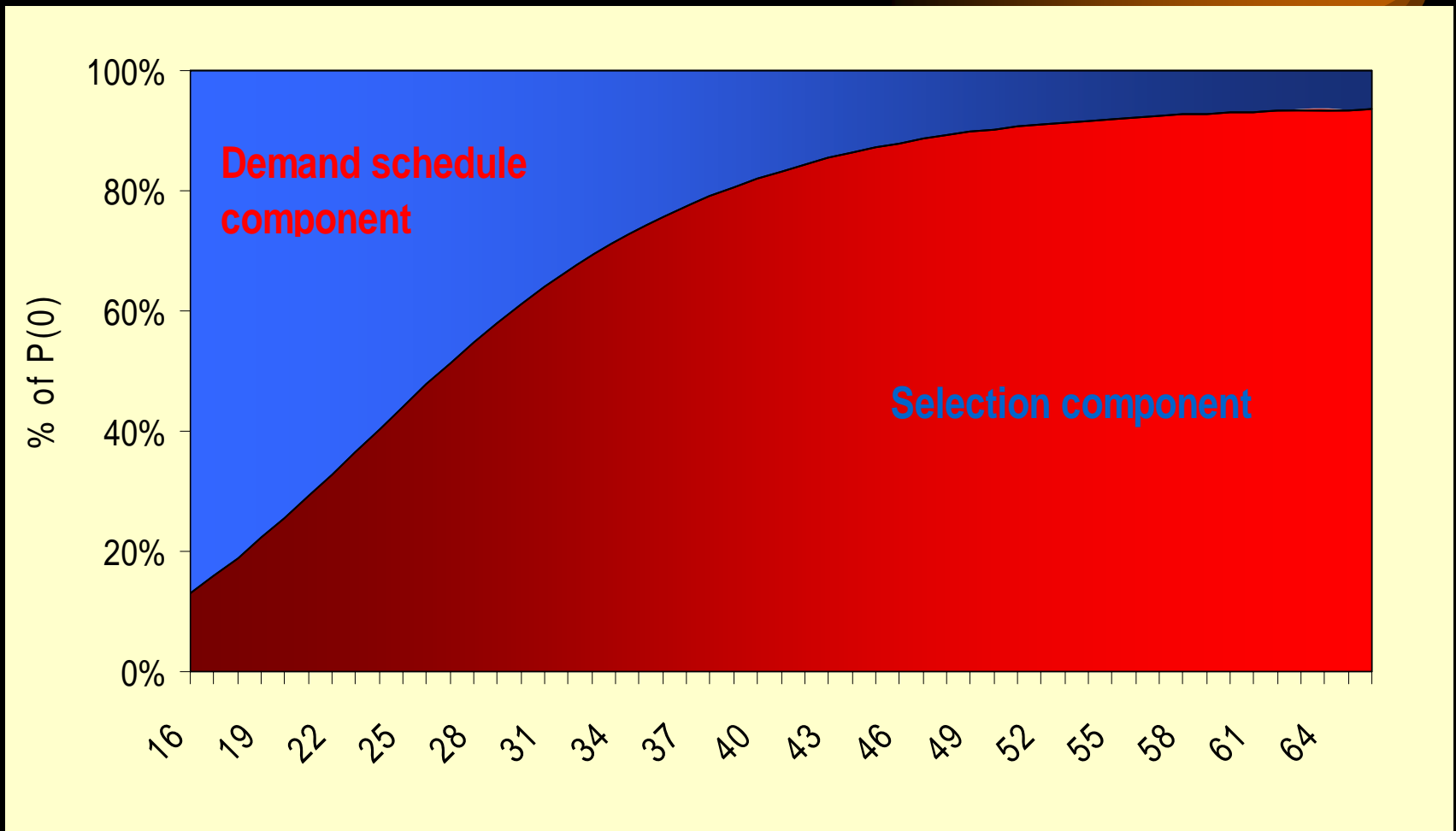
Results: Some Average Probs



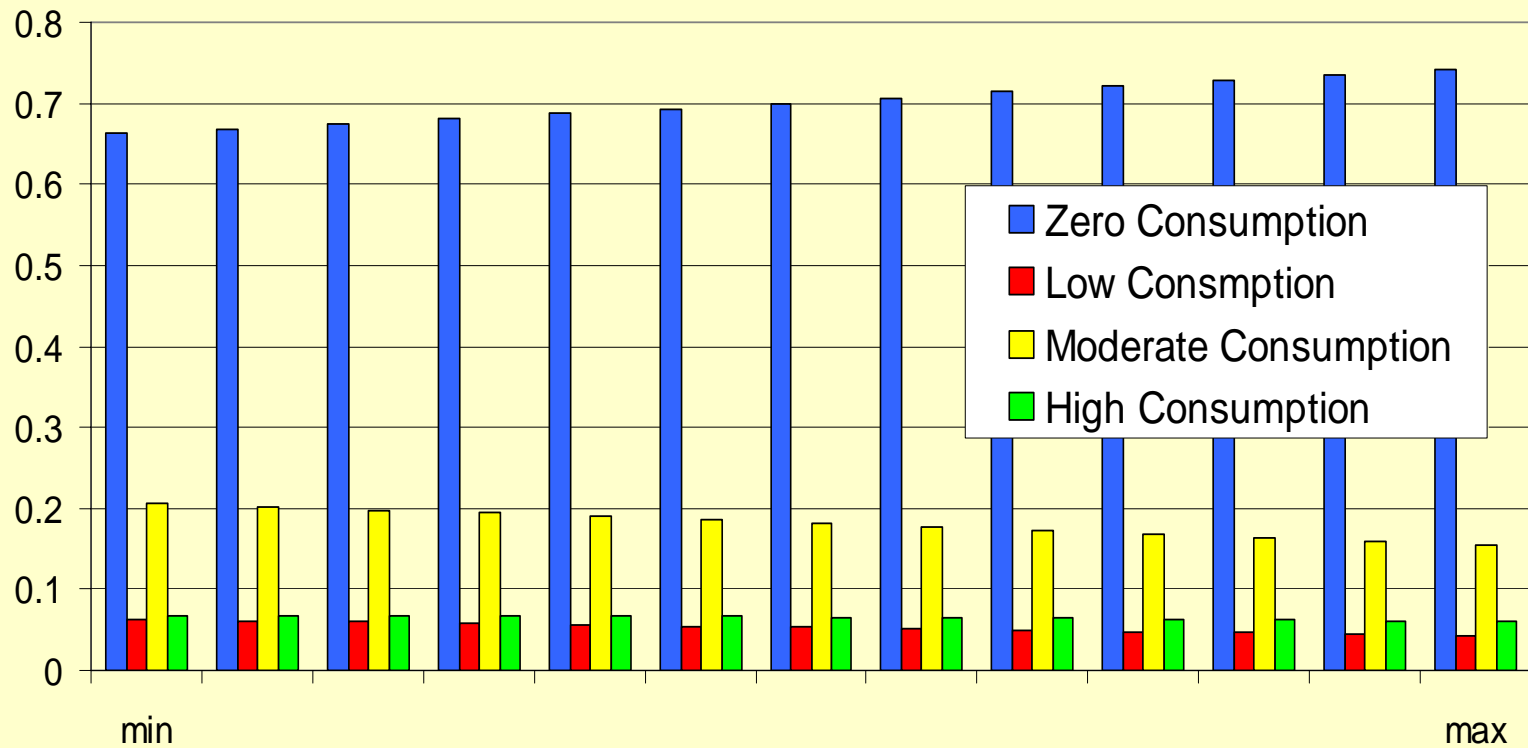
Results: Age



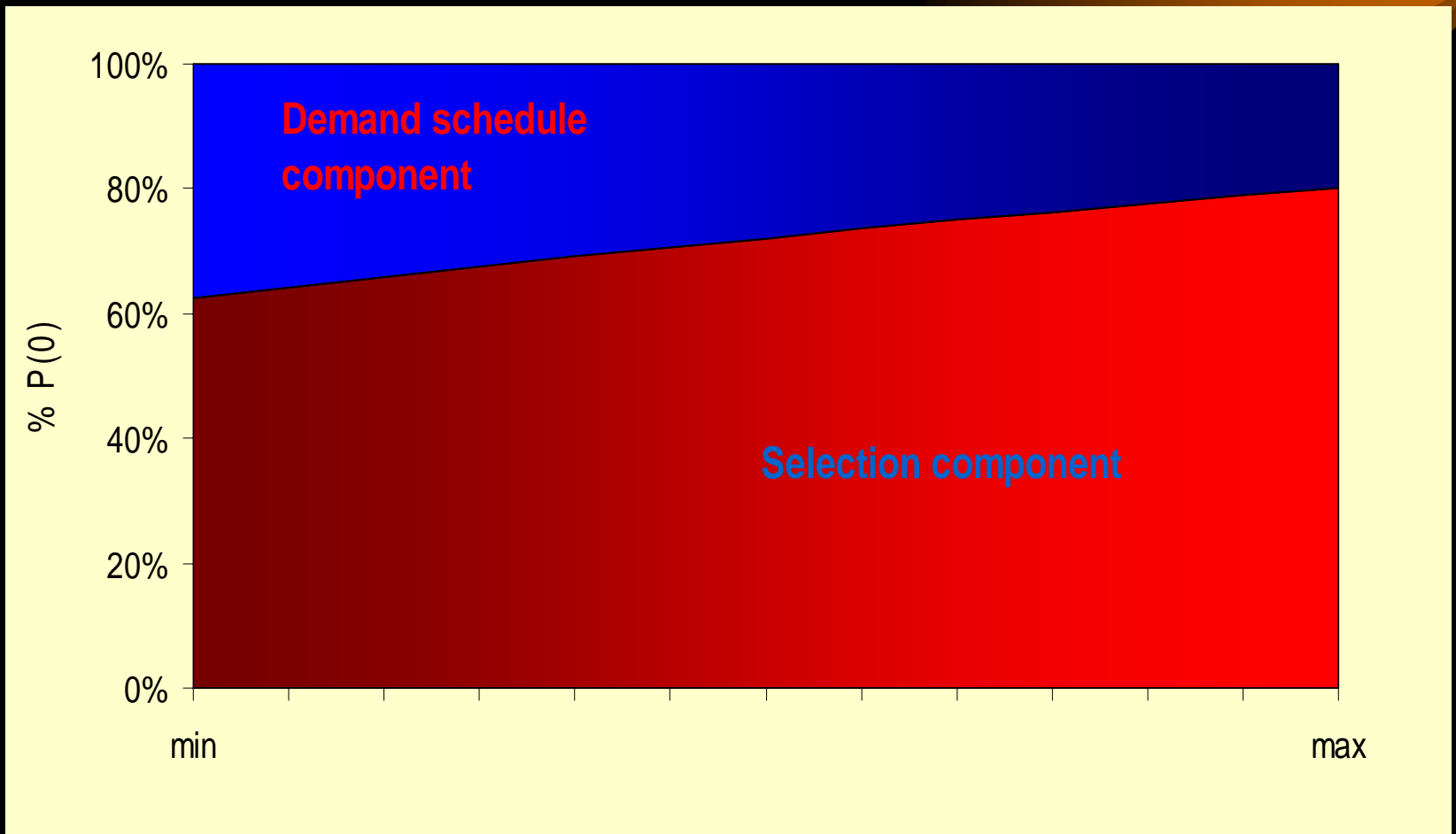
Results: Age & Zero Probs



Results: Income



Results: Income & Zero Probs



Further Potential Health Economics Uses...

1. Obesity

- Apparently, it's "the new smoking"...!!
- Often (?) coded from: obese (= 0) → underweight (= 4?) or *vice versa*
- So, might have a nature/nurture argument that says there's two "types" of obesity:
 - Genetically/inherently determined in "stage 1", so a function of parents weight, height *etc.*
 - In "stage 2", BMI levels more determined by lifestyle factors (social status, income, exercise *etc.*) ???

Further Potential Health Economics Uses...

2. (Self-Reported) Health Status

- Same kind of arguments here
- “Unhealthy” (or conversely “healthy”) individuals may be inherently so
 - Stage 1 relates to “genetically determined” unhealthy, so determined by proxies for such
- In “stage 2”, inherently not unhealthy individuals may become less so due lifestyle factors...

Further Potential Health Economics Uses...

3. Mental Health Status

- Again potentially two types of mentally unhealthy individuals: inherently so, and “the rest”
- And so on!
- In all of these examples it may be useful for policy to disentangle effects of “common” variables as in the smoking example
- And in all there are potentially 2 distinct types of “zero” observations...

Conclusions

- We propose a new model for ordered discrete data
- Following double hurdle and zero inflated models, observed “zeros” can come from two distinct decisions/equations/regimes
- Can easily allow for the likely correlation between the two equations
- Monte Carlo results suggest that the model performs well

Conclusions

- Model was applied to tobacco consumption
 - Good fit (“sensible” signs and signif.), *e.g.*
 - price effects present only in the amount equation;
 - Are clear “n-shaped” (*rational addiction*) age effects for amount of consumption
 - And so on...

Conclusions

- If the d.g.p. is **not** “zero-split” → testing procedures will pick OP
- **But** if the d.g.p. **is** “zero-split”, ZIOP model has important advantages over conventional OP model:
 1. OP and/or simple Probit will “get it wrong”

Conclusions

2. Model can be used to estimate proportion of zeros coming from each regime
 - And how this changes with personal characteristics
3. Which variables are important in which regime
 - Potentially very important for policy analysis
- Model is forthcoming in next release of Limdep, so if anyone has any such data, tell me and get in quick!!!

THE END!