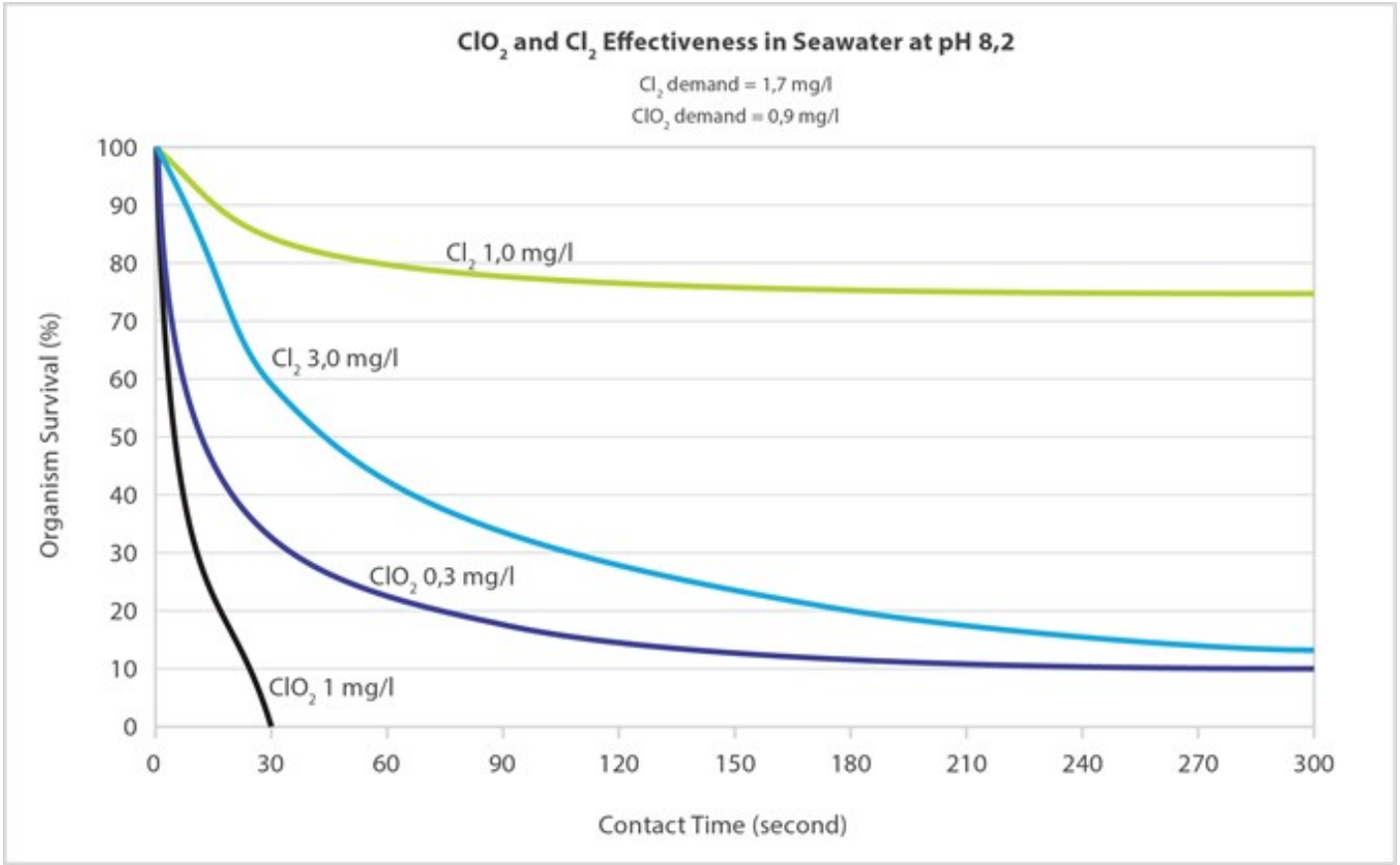
Applied Microeconometrics II , Lectures 9 and 10

Ciprian Domnisoru
Aalto University

Outline

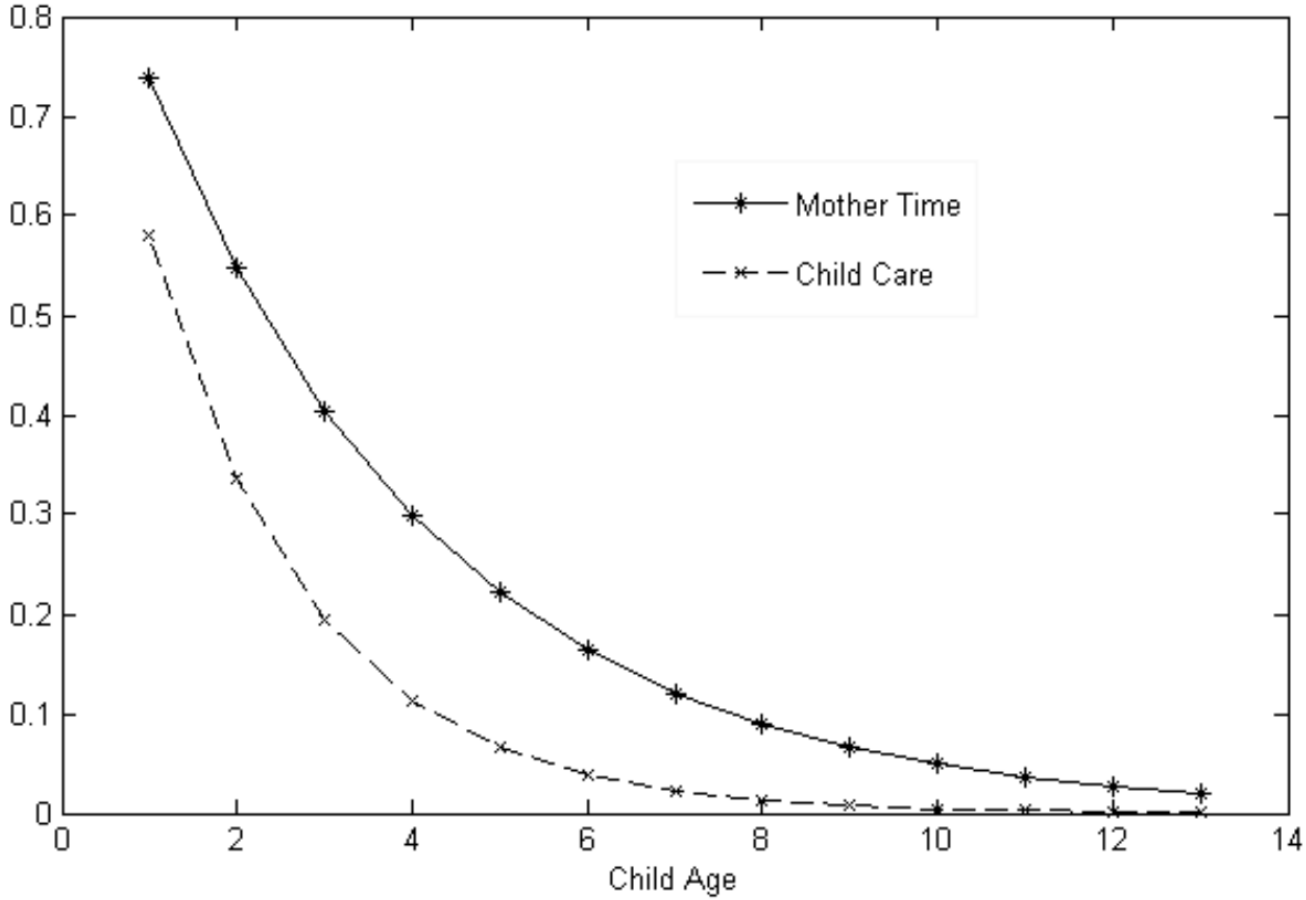
- Two related threads:
  - Structural modeling of data generating processes
  - Econometric techniques that go beyond OLS
- Logits and probits as *basic* models of utility maximization and examples of maximum likelihood estimation
- Broadening the perspective on maximum likelihood estimation and other strategies and application in the Duflo, Hanna and Ryan paper "Getting teachers to come to school"

DGP for chlorine dioxine effects on bacteria



Ylenia Brilli, "Mother's time allocation, childcare and child cognitive development", Journal of Human Capital, volume 16, number 2, summer 2022.

FIGURE D.3
Elasticity of child's ability with respect to mother's time with the child and non-parental child care if maternal time includes also time when the father is around.



NOTE. This graph represents the productivity parameters for maternal time (τ_t) and non-parental child care (i_t) as a function of child's age $t = 1, 2, \ldots 13$. τ includes all time spells when the mother is with the child and also those when the mother is present and the father is around but not involved in child's activities.

Structural estimation

- Estimate parameters of a data generating process (DGP) which are assumed to be invariant to policy changes or other counterfactuals.
- a structure: "a set of functional or probabilistic relationships between observable and latent variables which implies a joint distribution of the observables"
- The goal of structural estimation is to estimate the parameters of the DGP
- Often nonlinear problems (nonlinearity in coefficients) arise as solutions of differential equations or growth or decay problems. Example: $Y_i = C \cdot 2^{-\frac{X_i}{\theta}} + \epsilon_i$ or $y = \theta_1 + \theta_2 e^{x\theta_3} + \epsilon$
- In setting where experiments (or past policies) are infeasible
- Example: child skill formation model. Cunha, Heckman, and Schennach, 2010. Use model to simulate counterfactual policies (e.g. where reading is reduced)

Cunha, Heckman, and Schennach, 2010

Skills evolve in the following way. Each agent is born with initial conditions $\theta_1 = (\theta_{C,1}, \theta_{N,1})$. Family environments and genetic factors may influence these initial conditions (see Olds (2002) and Levitt (2003)). We denote by $\theta_P = (\theta_{C,P}, \theta_{N,P})$ parental cognitive and noncognitive skills, respectively. $\theta_t = (\theta_{C,t}, \theta_{N,t})$ denotes the vector of skill stocks in period t. Let $\eta_t = (\eta_{C,t}, \eta_{N,t})$ denote shocks and/or unobserved inputs that affect the accumulation of cognitive and noncognitive skills, respectively. The technology of production of skill k in period t and developmental stage s depends on the stock of skills in period t, investment at t, $I_{k,t}$, parental skills, θ_P, shocks in period t, $\eta_{k,t}$, and the production function at stage s,

$$\theta_{k,t+1} = f_{k,s}(\theta_t, I_{k,t}, \theta_P, \eta_{k,t}) \tag{2.1}$$

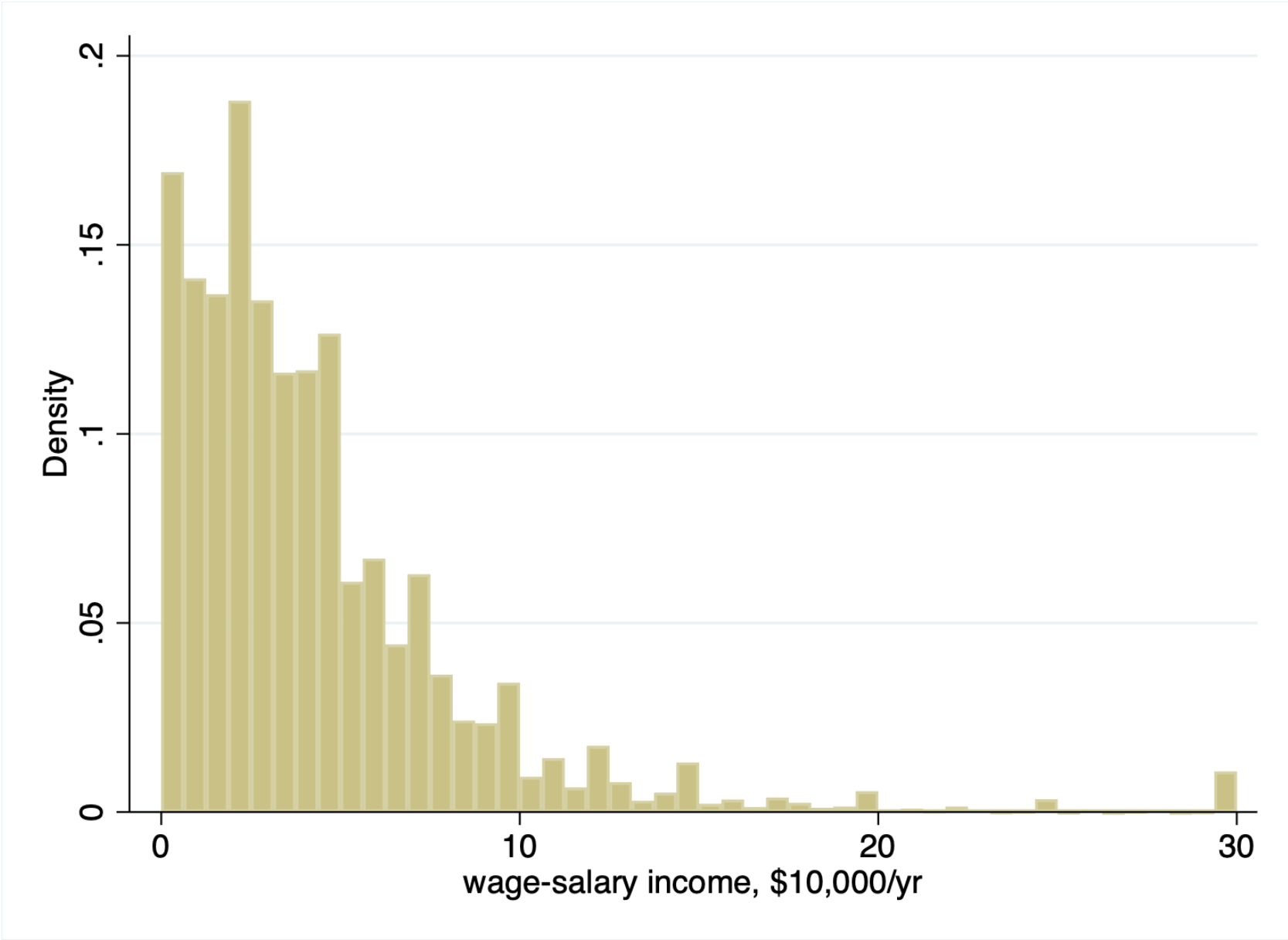
for $k \in \{C, N\}$, $t \in \{1, 2, \ldots, T\}$, and $s \in \{1, \ldots, S\}$. We assume that $f_{k,s}$ is monotone increasing in its arguments, twice continuously differentiable, and concave in $I_{k,t}$. In this model, stocks of current period skills produce next period skills and affect the current period productivity of investments. Stocks of cognitive skills can promote the formation of noncognitive skills and vice versa because θ_t is an argument of (2.1).

- How do we know we have the right DGP?
- Identification: would other parameters produce the same data?

Maximum Likelihood Estimation: Introduction

Maximum likelihood example

Suppose you want to model the wage distribution - which is right skewed...



Maximum likelihood example

- Maximum likelihood estimation works by choosing parameters to maximize the "likelihood" of the observed data, given a fully specified model for the data generating process.
- The likelihood is essentially the probability that you would observe the sample, based on particular parameter values.
- L(parameters | data) "the likelihood that the parameters take certain values given that we've observed some data."
- You are planning to model the wage distribution using a two-parameter gamma distribution with shape parameter α and rate parameter β (sometimes parametrized $1/\beta$ and called a scale parameter).
- The density function is:

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} exp(-\beta y)$$

The log likelihood for an observation is:
$ln(l_i) = \alpha ln(\beta) - ln\Gamma(\alpha) + (\alpha - 1) ln(y_i) - \beta y_i$
, where $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}$

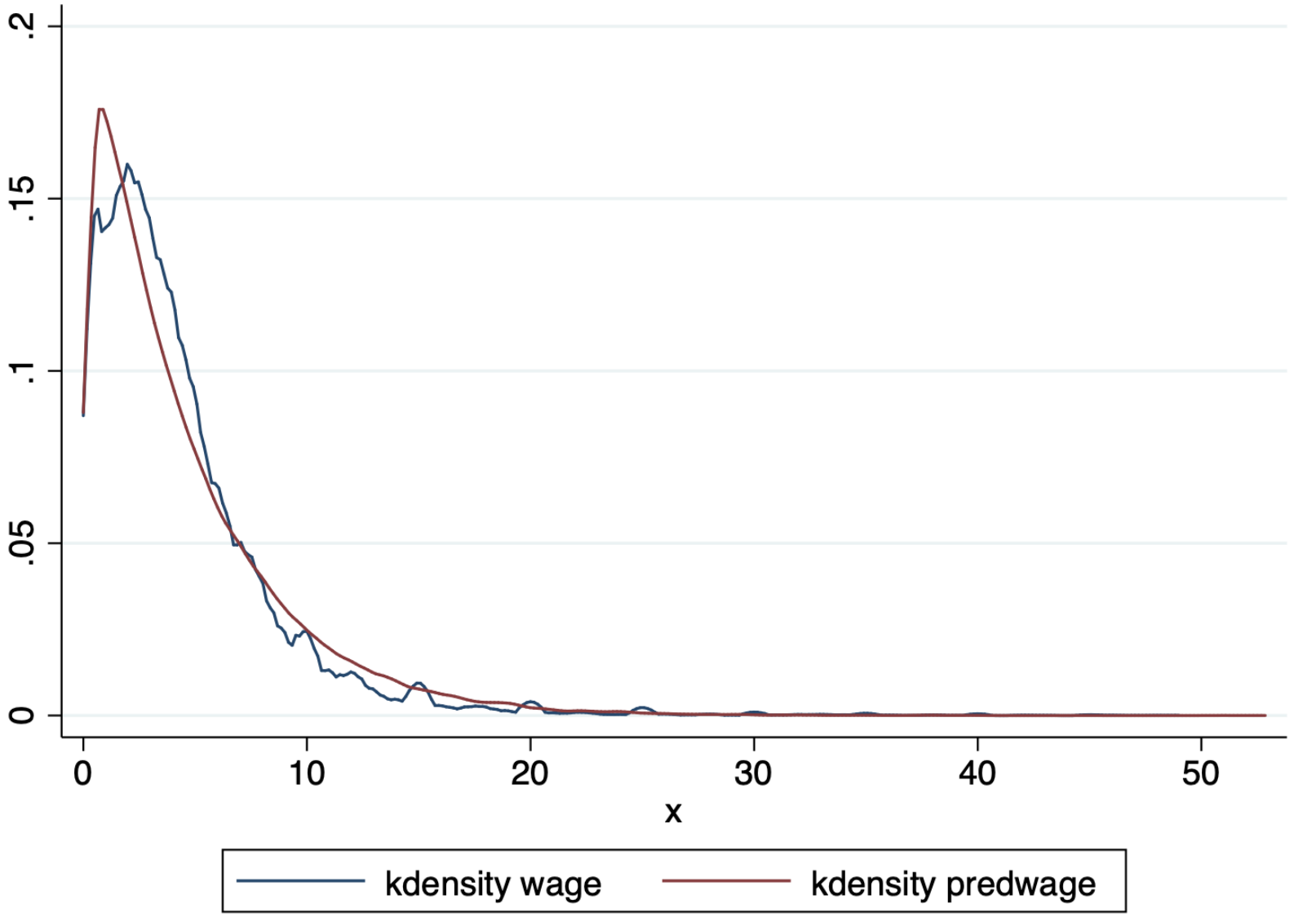
mlexp function in Stata

```
. mlexp ({a=1}*ln({b=.1}) - lngamma({a}) + ({a}-1)*ln(wage)  - {b}*wage)

initial:       log likelihood = -178608.12
rescale:       log likelihood = -178608.12
rescale eq:    log likelihood = -173389.94
Iteration 0:   log likelihood = -173389.94
Iteration 1:   log likelihood = -163081.69
Iteration 2:   log likelihood = -162813.54
Iteration 3:   log likelihood = -162808.55
Iteration 4:   log likelihood = -162808.55

Maximum likelihood estimation

Log likelihood = -162808.55                     Number of obs     =     64,748

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          /a |   1.097287   .0054134   202.70   0.000     1.086677    1.107897
          /b |   .2406719   .0014917   161.34   0.000     .2377483    .2435955
------------------------------------------------------------------------------
```

Distribution fit

gen predwage=rgamma(1.0972,1/.2406719)

twoway (kdensity wage if wage <50 || kdensity predwage if wage<50)



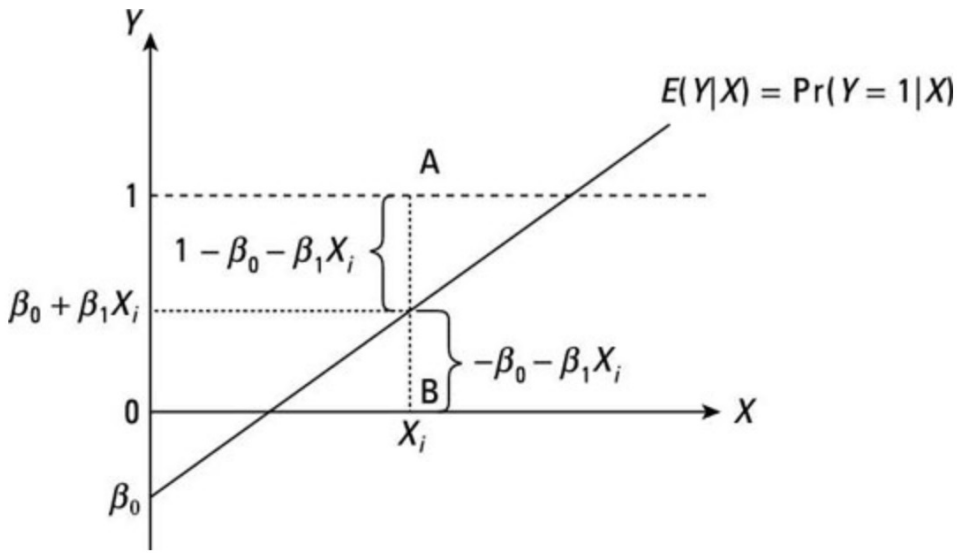
Random Utility Models: Logit and Probit

Fully specified data generating processes

- Utility when choice=bus. $U_0 = \beta_0 X + u_0$
- Utility when choice=car. $U_1 = \beta_1 X + u_1$
- Systematic taste variation, not random taste variation
- Choose car when $U_1 > U_0$
- Let y_i^* (the "latent variable") represent the difference in utility y_i^* $= U_1 - U_0 = X(\beta_1 - \beta_0) + (u_{1i} - u_{0i}) = X_i\beta + u_i$
- We observe a binary decision y_i, which takes the value 1 if $y_i^* > 0$ and 0 if $y_i^* <= 0$
- In the dataset, we only observe X_i and y_i, but not y_i^*

LPM estimation

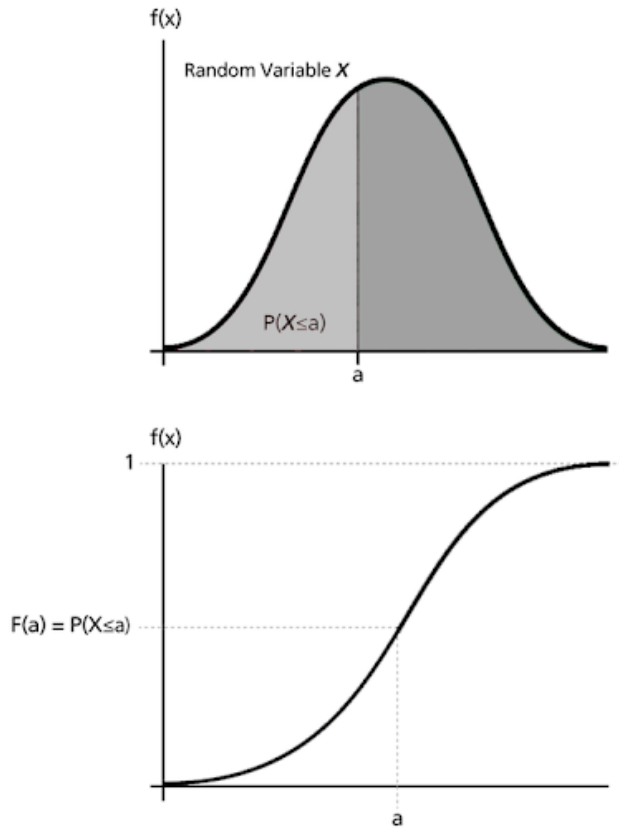
With OLS, you could estimate: $y = X\beta + u$



- What are the OLS predicted values?
- What is the marginal effect of a change in X?

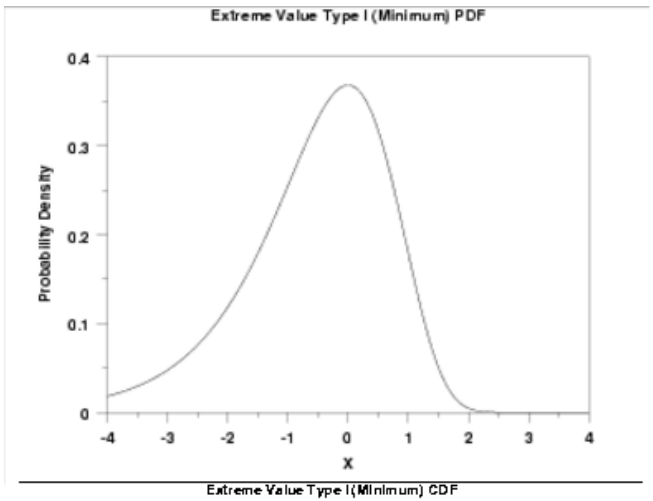
Choices of G

- $P(y_i = 1|X_i) = P(y_i^* > 0) = P(X_i\beta + u_i) = P(u_i > -X_i\beta) = 1 - G(-X_i\beta)$
- G is the cumulative density function of u (second graph below)



Choices of G

- If we assume u_i follows the standard normal distribution, because of its symmetry, $1 - \Phi(-X_i\beta) = \Phi(X_i\beta)$
- Then, $P(y_i|X_i) = \Phi(X_i\beta)$, which we call a probit model
- We can also assume u_i follows a Type I extreme value distribution.



Choices of G

- Given the type 1 extreme value distribution, we have
- The cdf of the error term is the logistic function: $\Lambda(z) = \frac{exp(z)}{1+exp(z)}$
- 1-Λ(-$X_i\beta$)=1-$\frac{exp(-X_i\beta)}{1+exp(-X_i\beta)} = \frac{exp(X_i\beta)}{1+exp(X_i\beta)}$
- We observe X_i in the dataset, and the probability that we observe y_i=1 is $\Phi(X_i\beta)$ for the probit model, and $\Lambda(X_i\beta)$ for the logit model.
- The probability that we observe $y_i = 0$ is $1 - \Phi(X_i\beta)$ for the probit model and $1 - \Lambda(X_i\beta)$ for the logit model.

Maximum likelihood function

- We want to find the parameters that determine $P(y_i = 1|X_i) = 1 - G(-X_i\beta)$ and we observe Y_i and X_i, but we can't run OLS. Our parameters concern the latent variable y*.
- Maximum likelihood estimation works by choosing parameters to maximize the "likelihood" of the observed data, given a fully specified model for the data generating process.
- The likelihood is essentially the probability that you would observe the sample, based on particular parameter values.
- L(parameters | data) "the likelihood that the parameters take certain values given that we've observed some data."
- P(data | parameters) "the probability density of observing the data with certain parameters"

Maximum likelihood function

- Note that since the outcome is binary, the calculation of the likelihood function is based on the Bernoulli distribution, which has pdf $p^y(1-p)^{1-y}$
- We can then construct the likelihood function- for the probit: $\prod (\Phi(X_i\beta))^{y_i}(1-\Phi(X_i\beta))^{1-y_i}$
- Log likelihood function for the probit $\sum y_i log(\Phi(X_i\beta)) + (1-y_i)(log(1-\Phi(X_i\beta)))$
- Log likelihood function for the logit $\sum y_i log(\Lambda(X_i\beta)) + (1-y_i)(log(1-\Lambda(X_i\beta)))$
- The estimator is typically found using numerical optimization methods to find the parameter vector that max/minimizes the objective function. These are iterative methods that try various parameters until they reach a point that they can't improve on. Some simple closed form solutions: the MLE for a linear model with a normally distributed error is OLS.

Logit example: McFadden(1974): transit choice

Table 2

Binary logit response curves; *dependent variable*: Auto–bus mode choice (zero if bus is usual or frequent mode, one otherwise); *estimation method*: Maximum likelihood on individual observations; *sample size*: 160; *T*-statistics in parentheses.

Independent variable	Model 1	Model 2	Model 3	Model 4
Family income with ceiling of $10,000, in $ per year	0.000065 (0.518)	0.000064 (0.517)	0.000095 (0.774)	0.000074 (0.601)
Car–bus cost, in cents per round trip	−0.00920 (3.085)	−0.00915 (3.184)	−0.01022 (3.726)	−0.01165 (4.506)
Car–bus on-vehicle time times post-tax wage, in min. per 1-way × $ per hr.	−0.00858 (1.263)	−0.00852 (1.273)	−0.01479 (2.460)	—
Bus walk time times wage, in min. per 1-way × $ per hr.	−0.000092 (0.021)	−0.000080 (0.018)	—	—
Bus first wait time times wage, same units	−0.01713 (0.771)	—	—	—
Bus transfer wait time times wage, same units	−0.01902 (1.365)	—	—	—
Bus total wait time times wage, same units	—	−0.01838 (1.947)	—	—
Bus total access time times wage, same units	—	—	−0.00314 (0.818)	—
Bus total travel time times wage, same units	—	—	—	−0.00728 (2.480)
Pure auto mode preference effect (constant)	0.1499 (0.165)	0.1483 (0.163)	0.3832 (0.428)	0.5516 (0.561)
Likelihood ratio index	0.30626	0.30623	0.2794	0.2633
R^2 index	0.92	0.93	0.66	0.61

"A positive coefficient indicates that when the remaining variables are zero, more than half the population will choose auto."

Reporting results from discrete outcome models

The coefficients themselves don't have a direct quantitative interpretation. Instead, we are interested in the *partial effect* or *marginal effect* of each X_k variable on P(y=1): $\frac{\partial P(y=1|x)}{\partial X_k} = g(X'\beta)\beta_k$
As you can see, the partial effects depend on the value of $x'\beta$. There are two common ways to calculate them:

$$g(\bar{x}'\beta)\beta_k \text{ or } \frac{1}{N}\sum g(x_i'\beta)\beta_k$$

For a dummy variable ($x_k = 0$ or 1), the partial effect is

$$G(\beta_0 + \beta_1 x_1 + ... + \beta_k + ...) - G(\beta_0 + \beta_1 x_1 + ... + 0 + ...)$$

This can also be calculated once at the means of the other variables or this difference can be averaged over all observations.

Odds ratios

Logit

Another feature of the logit model is that the results can be easily reported as effects on the *odds*,

$$\text{Odds} = \frac{P(y=1|x)}{P(y=0|x)}$$

This is because the odds has a simple expression:

$$\frac{\exp(x'\beta)/[1+\exp(x'\beta)]}{1/[1+\exp(x'\beta)]} = \exp(x'\beta)$$

Consider the partial effect of a dummy variable x_k:

$$\frac{\text{Odds}(x_k=1)}{\text{Odds}(x_k=0)} = \frac{\exp(\sum_{j\neq k}\beta_j x_j)\exp(\beta_k)}{\exp(\sum_{j\neq k}\beta_j x_j)\exp(0)} = \exp(\beta_k)$$

The proportional change in the odds (which is a ratio of the odds) for a one-unit change in x_k is simply $\exp(\beta_k)$. This is reported as "Odds ratio" in Stata.

Multiple outcome models

(a) Labor Force Outcomes - Let the utility of the different labor force outcomes for individual i be formulated as:

$$\begin{aligned} U_{i,Work} &= x_i'\beta_W + \epsilon_{iW} \\ U_{i,School} &= x_i'\beta_S + \epsilon_{iS} \\ U_{i,Unemp} &= x_i'\beta_U + \epsilon_{iU} \end{aligned}$$

Individual i chooses work iff $U_{i,W} = \max\{U_{i,W}, U_{i,S}, U_{i,U}\}$.

(b) Consumer Product Choice - Let the utility of the different products for individual i be formulated as:

$$\begin{aligned} U_{i,Honda} &= x_H'\beta + \epsilon_{iH} \\ U_{i,Chevy} &= x_C'\beta + \epsilon_{iC} \\ U_{i,Ford} &= x_F'\beta + \epsilon_{iF} \end{aligned}$$

Individual i chooses a Ford iff $U_{i,F} = \max\{U_{i,H}, U_{i,C}, U_{i,F}\}$.

Multiple outcome models

Multinomial logits for employed, disabled, or retired, ages 57-61.

	Males		Females	
Variable	Disabled	Retired	Disabled	Retired
	- - - RELATIVE RISK RATIOS - - -			
100-point health scale (/ 10):				
- linear term	0.51***	0.90*	0.56***	0.93
- squared term	0.98	0.98	1.03	1.05**
- cubic term	1.00	0.99	1.01	1.00
ADLs / IADLs	6.24***	3.43**	1.64	1.95*
Vision impairment	1.68	1.33	1.63*	0.99
Hearing impairment	0.81	1.21	1.50	1.50*
Physical lim.	2.53***	1.37	4.63***	1.42**
Cognitive lim.	1.69	0.85	2.05*	1.48
Social lim.	2.69**	2.79***	2.17**	1.64
Diabetes	1.02	0.61*	1.87**	1.63**
Asthma	1.63	1.29	0.93	1.01
High BP	0.92	0.98	1.44	1.09
Heart condition	2.17***	1.45*	1.23	0.79
Stroke	3.04**	1.07	2.02	1.31
Age 60 or above	N/A	N/A	N/A	N/A
Some college	0.42***	1.10	0.47***	0.76**
Black	1.53	1.26	0.86	0.84
Hispanic	0.92	0.87	0.96	1.43*
Divorced, sep. or widowed	2.31***	1.02	1.03	0.30***
Never married	5.47***	1.59	4.17***	0.59
Metropolitan area	2.24***	1.06	0.82	1.00

Models include dummies for region.
P-values: * p<.1; ** p<.05; *** p<.01.

IIA assumption

- These logit models have an assumption called "independence from irrelevant alternatives," which results from the assumption that the unobservables ϵ_{ij} are independent across options
- The assumption is that the ratio of probabilities between two options is independent of any other option. To see this, note that $\frac{P(y=j|x)}{P(y=h|x)} = exp[(x_j - x_h)\beta]$
- restaurant options: Mexican, Japanese, Sushi. If you're in the mood for sushi (a mood which is unobserved by the econometrician), there would be no change in the ratio of P(Japanese) / P(Mexican).
- You can simplify the problem by aggregating, or use a **nested logit**, in which the choices are grouped into nests (clusters) such that IIA holds within a nest but not necessarily between nests. e.g. A nest containing buses.

Stata implementation- more in review session

- **logit votedyes age education gender**
- To report effects on the odds, **logit votedyes age education gender, or**
- **margins sex** Margins for a categorical variable
- Margins for a continous variable: **margins, at(age=(10(10)80))**
- Marginal effects of all independent variables at the mean of other covariates **margins, dydx(_all) atmeans**
- Mean marginal effects of all covariates **margins, dydx(_all)**
- **mlogit brand age sex class, baseoutcome (2)**
- To report effects on the odds ("relative risk ratios") **mlogit brand age sex class, baseoutcome (2) rrr**
- **margins sex, predict(outcome(3))**

Dynamic problems. Application: Duflo, Hanna, Ryan: Getting Teachers to Come to School

Duflo, Hanna, Ryan paper: Getting Teachers to Come to School

Duflo, Hanna, Ryan: Incentives Work: Getting Teachers to Come to School

- Seva Mandir, an NGO in rural Rajasthan, who runs 150 "non-formal education center" (NFE): single teacher school for students who do not attend regular school.
- Students are 7-14 year old, illiterate when they join.
- Teacher absence rate 35%
- Schools teach basic hindi and math skills and prepare students to "graduate" to primary school.
- In 1997, 20 million children were served by such NFEs

Duflo, Hanna, Ryan: Incentives Work: Getting Teachers to Come to School

- Teachers in intervention schools received a camera with non-temperable time and date stamp.
- Instructed to take two pictures of themselves and the children every day (pictures separated by at least 5 hours, at least 8 children per picture).
- Payment is calculated each month and is a non-linear function of attendance: • Up to 10 days: Rs 500. • Each day above 10 days: Rs 50. • In non-intervention schools, teachers receive Rs 1000, and are reminded that attending at least 20 days is compulsory.

Incentives Work: Getting Teachers to Come to School

- "While the reduced form results inform us that this program was effective in reducing absenteeism, they do not tell us what **the effect of another scheme with a different payment structure would be**."
- "Thus, these findings suggest that teachers respond to the incentives. **Without more structure, however, it is not possible to conclude what part of the effect of the program was due to financial incentives per se.** To analyze this problem, we set up a dynamic labor supply model and we use the additional restrictions that the model provides to estimate its parameters."

Incentives Work: Getting Teachers to Come to School

Let m signify the month and t the day within the month, where $t = \{1, \ldots, T_m\}$.[14] The teacher's utility function over consumption, C_{tm}, and leisure, L_{tm}, each day in the month is as follows:

$$U_{tm} = U(C_{tm}, L_{tm}) = \beta C_{tm}(\pi_m) + (\mu_{tm} - P)L_{tm}, \tag{2}$$

where P is the nonpecuniary cost of missing work.[15] We have assumed that utility is linear in consumption and that consumption and leisure are additively separable. This formulation implies that there will not be a dependency in behavior between months. For example, a teacher would not decide to work more in one month because she worked little in previous months.[16]

Consumption is a function of earned income, π_m. Since we assume that there is no discounting within months and utility is linear in consumption, we can assume that the teacher consumes all her income on the last day of the month, when she is paid.[17] The parameter β converts consumption, measured in rupees, into utility terms. We let L_{tm} equal one if the teacher does not attend work on that day and zero otherwise.

The coefficient on the value of leisure, μ_{tm}, has a deterministic and stochastic component:

$$\mu_{tm} = \mu + \epsilon_{tm}. \tag{3}$$

The deterministic component, μ, is the difference between the value of leisure and the intrinsic value of being in school, including any innate motivation. To the extent that teachers value teaching, or do not want to disappoint students and parents, μ will be less positive. The stochastic shock, ϵ_{tm}, captures variation in the opportunity cost of attending work on a given day; we assume that it has a normal distribution.

Teachers' decision problem

- Each day, a teacher chooses whether or not to attend school, by comparing the value of attending school to that of staying home or doing something else.
- State space s = (t, d), where t is the current time and d is the days worked previously in the current month.
- Payoffs: if the teacher does not attend school: $\mu + \epsilon_t$
- Payoff of attending school is calculated at the end of the month according to:

$$\pi(d) = 500 + 50 * max\{0, d - 10\}$$

- T takes value between 1 and T = 25.
- Transitions: Each day, t increases by one, unless t = T, in which case it resets to t = 1. If a teacher has worked in that period d increases by one, otherwise it remains constant.

Value functions

- A sum of objective functions can be re-expressed recursively using a value function. S=state space (e.g. wealth). X=control variables (e.g. consumption).

$$\mathrm{V}=\sum_{t=1}^{T}\beta^{t}\ \mathrm{E}[\mathrm{U}(X_t|S_t)]$$

- Bellman Equation: breaking the sequence into a simpler problem:

$$V(S) = \max_X \left[U(X|S) + \beta E[V(S')\right]$$
$$S' = f(X, S)$$

- This allows us to solve complicated problems of forward looking agents maximizing discounted stream of utility.
- *"An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision"*. Bellman, 1957 Dynamic Programming
- Usually solved iteratively, given an initial guess of the value function.

Teachers' value function

- Given the salary payoff structure, for t $\leq$ T, we can write the value function for each teacher as follows:
- $V(t, d) = \max\{\mu + \epsilon_t + EV(t+1, d), EV(t+1, d+1)\}$
- At time T:
  $V(T, d) = \max\{\mu + \epsilon_T + \beta\pi(d) + EV(1, 0), \beta\pi(d+1) + EV(1, 0)\}$
- β is marginal utility of income
- EV(1, 0) enters both side and can thus be ignored: we can solve each month independently, backwards from time T

Log-likelihood

The log likelihood is:

$$LLH(\theta) = \sum_{i=1}^{N} \sum_{m=1}^{M_i} \sum_{t=1}^{T_m} [1(\text{work}) Pr(\text{work}|t, d, \theta)$$

$$+1(\text{not work})(1 - Pr(\text{work}|t, d, \theta)],$$

where:

$$\begin{aligned} Pr(\text{work}|t, d, \theta) &= Pr(\mu + \epsilon_t + EV(t+1, d) < EV(t+1, d+1)) \\ &= Pr(\epsilon_t < EV(t+1, d+1) - EV(t+1, d) - \mu) \\ &= \Phi(EV(t+1, d+1) - EV(t+1, d) - \mu), \quad (10) \end{aligned}$$

- At time T, the agent faces a static decision, namely work if: $\mu + \epsilon_T + \beta\pi(d) > \beta\pi(d+1)$
- The probability of this event is:

$$Pr(work|d,\theta) = Pr(\epsilon_T > \beta(\pi(d+1) - \pi(d)) - \mu) = 1 - \Phi(\beta(\pi(d+1) - \pi(d)) - \mu)$$

- When d<10, the difference between $\pi(d+1)$ and $\pi(d)$ is zero, and β does not enter the equation $Pr(work|d,\theta) = 1 - \Phi(\mu)$
- "If all teachers share same μ, μ is identified by teachers who are out of the money, and then β from teachers in the money."
- var(ϵ) normalized to be equal to 1.
- If teachers have different μ model still identified by comparing different teachers with themselves over time (teacher fixed effect).

Results from simple model

- Common β and μ for all teachers in Model I
- Teachers respond to financial incentives
- Predicted number of days worked in the treatment group, 17.23 very close to actual 17.16
- Estimated opportunity cost of working, $\mu = 1.564$ is positive- model underpredicts the number of days that teachers work in the control group. Model 1.31 days, data 12.9 days
- Model II accounts for heterogeneity in μ but still underpredicts control group.

Parameter	Model I (1)	Model II (2)	M
β	0.049 (0.001)	0.027 (0.000)	
μ_1	1.564 (0.013)		
ρ			
σ_1^2			
μ_2			
σ_2^2			
ρ			
Yesterday shifter			
Attendance			
Test score			
Heterogeneity	None	FE	
Three-day window	No	No	
LLH	10,269.13	9,932.71	
ϵ_{Bonus}	1.09 (0.147)	0.592 (0.062)	
ϵ_{bonus_cutoff}	−18.26 (2.023)	−1.90 (0.564)	−
Predicted days worked	17.23 (0.361)	17.30 (0.153)	
Days worked $BONUS = 0$	1.31 (0.041)	6.96 (0.101)	
Out-of-sample prediction	21.47 (0.046)	19.975 (0.164)	

Results from complex model

- Model correlation in error terms :
- $\epsilon_{mt} = \rho\epsilon_{m,t-1} + \nu_{mt}$
- Autocorrelation could be either positive (illness) or negative (teacher has a task to accomplish).
- Use method of simulated moments: simulate work history for different parameters, and try to match a distribution of days worked at the beginning of the month.
- Heterogeneity introduced by drawing p teachers from a distribution with high outside option, and 1 − p from distribution with low outside option.

Parameter	Model V (5)
β	0.013 (0.001)
μ_1	−0.428 (0.045)
ρ	0.449 (0.043)
σ_1^2	0.007 (0.019)
μ_2	1.781 (0.345)
σ_2^2	0.050 (0.545)
ρ	0.024 (0.007)
Yesterday shifter	
Attendance	
Test score	
Heterogeneity	RC
Three-day window	No
LLH	
ϵ_{Bonus}	0.196 (0.053)
ϵ_{bonus_cutoff}	−0.14 (0.144)
Predicted days worked	16.75 (0.391)
Days worked $BONUS = 0$	12.90 (0.281)
Out-of-sample prediction	17.77 (0.479)

Common (structural) estimation methods

- **Maximum likelihood:** maximize the likelihood of drawing that data x from a model, given parameters θ.

$$\max_\theta lnL = \sum_i^N ln(f(x_i|\theta))$$

- In cases where the data distribution function is unknown or difficult to derive analytically, **Generalized method of moments.** Minimize the distance between model moments $m(x|\theta)$ and data moments m(x)

$$min_\theta ||m(x|\theta) - m(x)||$$

- **Moment conditions**: functions of the model parameters and data, such that their expectation is zero at the true value of the parameters.
- Moments for OLS estimation. y=$\beta_0+\beta_1 X + \epsilon$
  Taking first order conditions of $min_{\hat{\beta}_0,\hat{\beta}_1} W = \sum_{i=1}^N (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ is equivalent to a sample moment condition $\frac{1}{N}\sum_{i=1}^N (y_i - x_i'\hat{\beta})x_i = 0$, which results from the orthogonality condition
  $E[\epsilon_i x_i] = E[(y_i - x_i'\beta)x_i] = 0$.

Common (structural) estimation methods

- If m(x $|\theta$) is not known analytically- **method of simulated moments**. Simulate the model data S times, and take the average of the moments of the simulated data as estimators of the model moments.

$$\hat{m} = \frac{1}{S}\sum_1^S m(\tilde{x}_s|\theta)$$

$$\text{Then run } min_\theta ||\hat{m}(\tilde{x}|\theta) - m(x)||$$

- Common norm: sum of squared errors. Minimize $\sqrt{(\frac{\hat{m}(\tilde{x}|\theta)-m(x)}{m(x)})^2}$

Out of sample prediction

- Seva Mandir changed rule after experiment was over (and model was estimated!)
- New rule: Rs 700 for 12 days of work. Increment of Rs 70 after the 13th day
- Model does well too.

Counterfactual analysis

"A primary benefit of estimating a structural model of behavior is the ability to calculate outcomes under economic environments not observed in the data. We are interested in finding the cost-minimizing combination of the two policy instruments, the size of the bonus and the threshold to get into the bonus, that lead to a minimum number of days worked in a month. "

TABLE 5—COUNTERFACTUAL COST-MINIMIZING POLICIES

Expected days worked (1)	Bonus cutoff (2)	Bonus (3)	Expected cost (4)	Test score gain over control group (13 days) (5)
14	0	0	500	0.04
15	21	25	521	0.07
16	22	75	664	0.11
17	21	75	672	0.15
18	20	75	755	0.18
19	20	100	921	0.22
20	20	125	1,112	0.26
21	16	225	2,642	0.29
22	11	275	4,604	0.33

The NGO could have induced higher work effort with approximately the same expenditure by doubling the bonus threshold and nearly tripling the per-day bonus... teachers in our sample appear to be more likely than not to attend school even without incentives and be forward-looking. A higher threshold avoids rewarding inframarginal days.

A word on calibration

Let the commuter value time at v_i dollars per hour. An extensive literature (Small and Verhoef 2007; Abrantes and Wardman 2011) concludes that commuters place a higher value on time spent waiting for transit, stuck in traffic, or walking than they do on the same amount of time in other circumstances. Defining a "delay multiplier" $c > 1$, we can write the commuter's problem as maximizing

$$U_i = X_i - v_i\left[R_i\left(\frac{m}{s_r} + c(a_{ri} + w_r)\right) + (1 - R_i)\left(\frac{m}{s_d} + c(a_d + w_{di})\right)\right] \tag{3}$$

$$s.t. \qquad Y_i = X_i + m \cdot (p_r R_i + p_d(1 - R_i)),$$

where s_r is rail speed, a_{ri} is rail access and egress time, w_r is average waiting time for the train, s_d is free-flow driving speed, a_d is car access and egress time, and w_{di} is driving delay time (i.e., the difference between driving time in free-flow traffic and actual driving time). For simplicity we do not include a mode-specific utility shock, although our conclusions are qualitatively robust to doing so (see online Appendix A2). Solving the commuter's problem leads to a decision rule under which the commuter takes rail if and only if

$$\left[c(a_{ri} + w_r) + \frac{m}{s_r}\right] - \left[c(a_d + w_{di}) + \frac{m}{s_d}\right] \leq \frac{m}{v_i}(p_d - p_r). \tag{4}$$

Rail is the more appealing choice if the difference between delay-penalized rail travel time $c(a_{ri} + w_r) + m/s_r$ and delay-penalized driving travel time $c(a_d + w_{di}) + m/s_d$ is less than the difference between the cost of driving and the cost of taking rail, converted from dollars to hours $\left((p_d - p_r)m/v_i\right)$. The share of commuters taking rail is thus determined by the probability that the inequality above holds. We calibrate the model under two scenarios. The first scenario assumes that, consistent with the existing literature, all peak-period drivers face the same average congestion delay, w_d. We set the value of time v_i at a fraction of the hourly wage and calibrate the model by varying the distribution of rail

- "We **set** the value of time v_i at a fraction of the hourly wage and calibrate the model by varying the distribution of rail access times until the probability of taking rail equals the observed rail market share in Los Angeles."
- "In cases with any ambiguity we tried to **choose** parameter values **consistent with the previous literature** (e.g., Parry and Small 2009). We assume a trip length of seven miles for commuters in rail catchment areas and five miles for commuters in bus catchment areas. Transit headways and fares come from historical MTA documents, and driving costs come from the American Automobile Association (LACMTA 2003; AAA 2004). "
- "We could not find authoritative data on parking costs or the share of commuters with free parking, so we **assumed** that 85 percent of commuters have free parking and that parking costs $5.00 per day for those with paid"

TABLE 1—PARAMETER VALUES FOR MODEL CALIBRATION

Parameter	Related variable	Rail	Bus	Source (where applicable)
General parameters				
Trip length	m	7 miles	5 miles	Parry and Small (2009)
Hourly wage (average)	v_i	\$21.60		BLS (2004)
Hourly wage (95 percent interval)	v_i	\$8.00–\$65.50		BLS (2004)
Wage multiplier for value of time	v_i	0.5		Parry and Small (2009)
Delay multiplier	c	1.8		Parry and Small (2009)
Transit travel time and costs				
Transit vehicle speed (average)	s_r, s_b	23 mph	11 mph	Parry and Small (2009)
Transit vehicle speed (95 percent interval)	s_r, s_b	23 mph	8.8–11.6 mph	
Avg. time between trains/buses	w_r, w_b	7 mins	8 mins	Los Angeles County MTA
Walking speed	a_r, a_b	2.5 mph		
Adult fare (per mile)	p_r, p_b	\$0.12	\$0.17	Los Angeles County MTA
Driving travel time and costs				
Free-flow driving speed	s_d	40 mph	35 mph	Parry and Small (2009)
Actual driving speed (average)	w_d	30 mph	27.1 mph	Parry and Small (2009)
Actual driving speed (95 percent interval)	w_d	14–40 mph	13.3–35 mph	PeMS data, Bing maps
Access, parking, and egress time	a_d	3 mins		
Operating costs (per mile)	p_d	\$0.15		AAA (2004)
Share commuters with free parking	p_d	85 percent		
Parking costs (per day)	p_d	\$5.00		

Notes: The delay multiplier applies to time spent waiting for transit, walking, or delayed in traffic. For time spent delayed in traffic, we calculate delay time as the difference between actual driving time and driving time under free-flow conditions.

Calibration results

- "We predict that ceasing transit service would increase average delays by 0.189 minutes per mile (38 percent)."
- "This effect is 5.9 times larger than the predicted effect in the homogeneous driving time model."
- "The implied fare elasticity in the heterogeneous model is -1.1, which is slightly larger than estimates from the literature."
- "Our preferred RD estimate ... finds that average delay increases 0.194 minutes per mile (47 percent)."

TABLE 2—MODEL CALIBRATION RESULTS

	Homogeneous driving speed	Heterogeneous driving speed
Outcomes		
Average delay for drivers	0.50 mins/mile	0.50 mins/mile
Average delay for rail passengers (if they chose to drive)	0.50 mins/mile	3.19 mins/mile
Average delay for bus passengers (if they chose to drive)	0.50 mins/mile	2.47 mins/mile
Effect of ceasing transit on average delay	0.032 mins/mile	0.189 mins/mile
Average consumer surplus for rail passengers	\$0.08/mile	\$0.24/mile
Average consumer surplus for bus passengers	\$0.04/mile	\$0.11/mile
Calibration parameters		
Share of population within two miles of rail line	51 percent	30 percent
Average bus line spacing in residential areas	0.4 miles	0.5 miles

Notes: Average delay is chosen to match Parry and Small (2009). Calibration parameter values are the values necessary to equate predicted ridership with observed ridership.

Summary of the "structural toolkit"

- Need to find the deep parameters (policy invariant parameters) in order to conduct counterfactual simulations. These parameters feature in a model of the data generating process.
- In the process of setting up the data generating process, fix some parameters (discount factor), likely make functional form assumptions on the error term that will simplify estimation
- Most choice problems rely on Type I extreme value error term distribution or some other convenient assumption on the error terms.
- Dynamic problems simplified through the use of value functions and the Bellman (functional) equation. $V(S) = \max_X [U(X|S) + \beta E[V(S')]$

Summary of the "structural toolkit"

- Try to form a likelihood function- analytical solutions hard to find, some integrals very difficult to evaluate.
- Method of moments : no clear guidance of what the "optimal" moments are, but they should be informative of the parameters you are trying to estimate- variations of parameters should induce variation in moments
- Moments are not known analytically/difficult to derive: simulated method of moments.
- Often as a first step, estimate/calibrate some parameters that don't require structural estimation (e.g. OLS). In the second step, estimate remaining parameters.

Summary of the "structural toolkit"

- Common to test the robustness of models by increasing complexity, and comparing to predictions from static (or naive) models
- Model fit (usually unsurprising for moments you target) - use moments that are not targeted.
- Out-of-sample checks
- Compare parameters to literature; discuss magnitude of results, compare to reduce form estimates.
- Can calculate counterfactual analyses under different models and compare implications.

Structural estimation: what's the downside

- Estimation will be more challenging
- Sometimes need to interpret marginal effects rather than coefficients
- Relatively more conditions needed for identification :
  - structural form assumptions
  - normalizations
  - assuming separable preferences (over time and states) to simplify estimation
  - setting the discount factor to a constant.

Structural models: upside

- There is only so much we can learn from historical quasi-experimental variation
- Perform counterfactual experiments
- Test specific mechanisms
- Used to compare predictive power of competing theories
- Sometimes parameters have direct interpretation as economic concepts (elasticities, risk aversion parameters, discount factors, etc.)

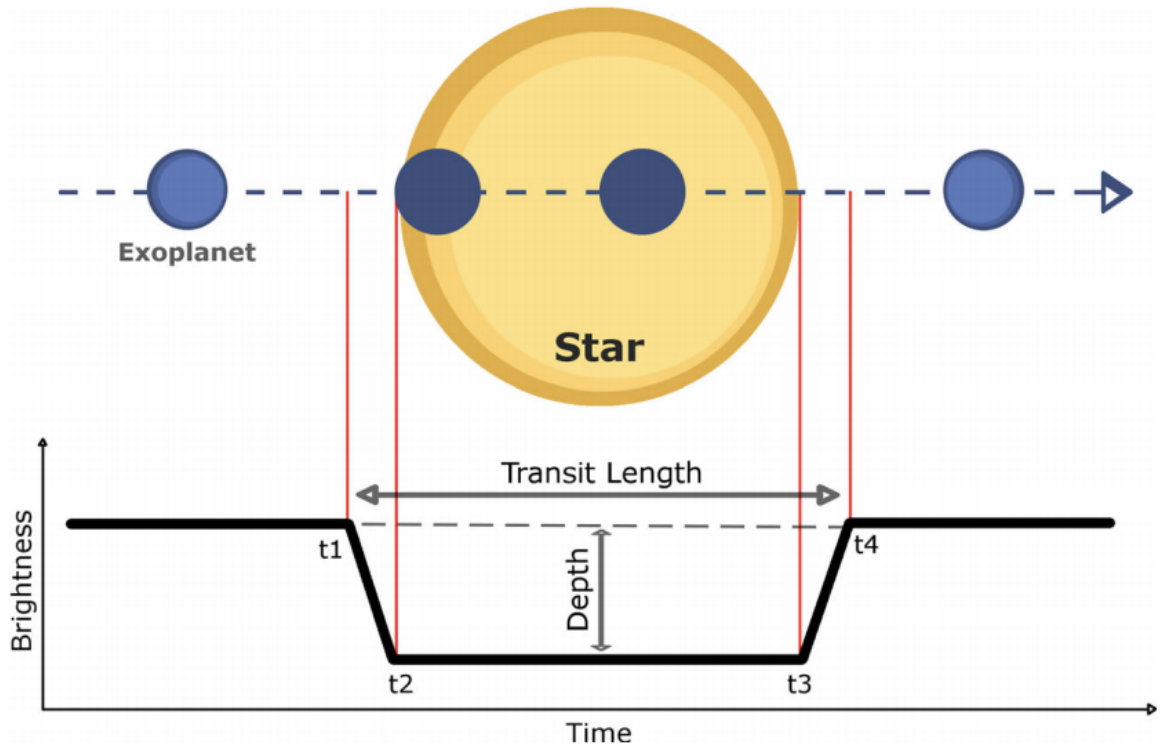


Fig. 2 Example of a light curve. As the exoplanet orbits the star, different brightness values are obtained. Some parameters that can be extracted from a light curve are: Beginning of ingress ($t1$); end of ingress ($t2$); beginning of egress ($t3$); end of egress ($t4$); transit length; and transit depth

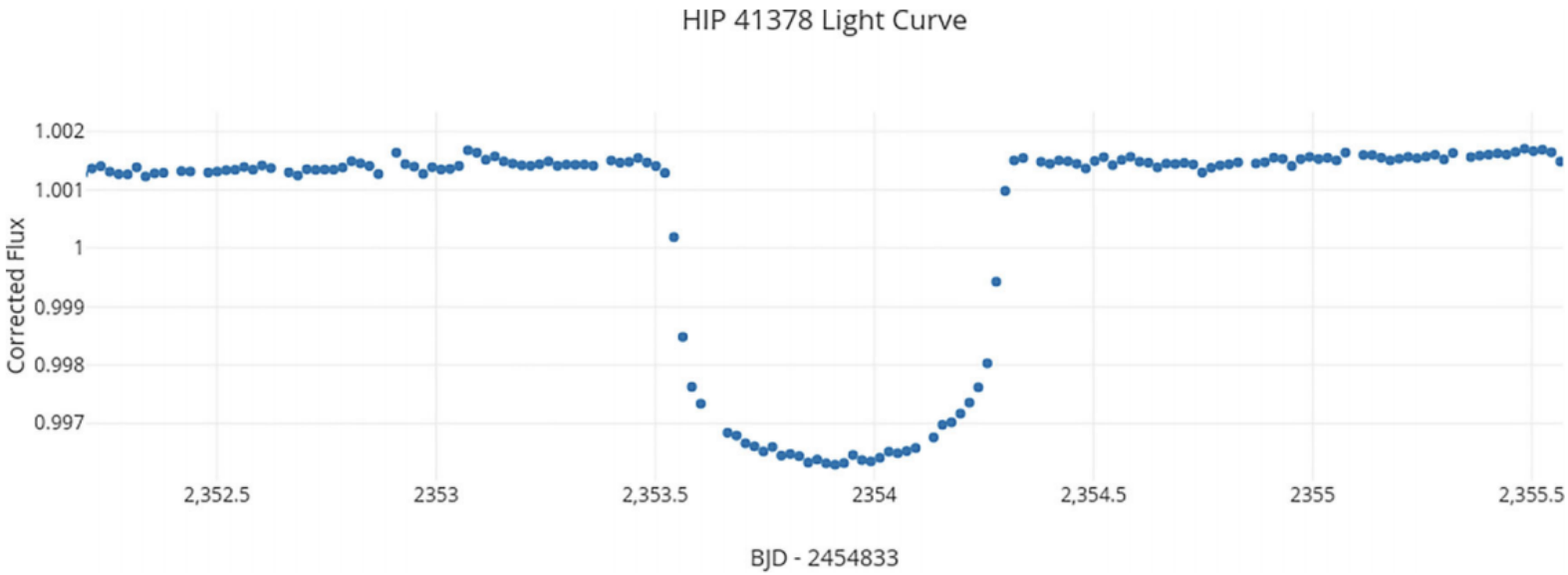


Fig. 4 Real light curve extracted from the planetary system around the star HIP 41378 in the MAST archive. The x-axis represents a measure of time called Barycentric Julian Day (*BJD*); the value 2454833 that accompanies the x-axis title, is to be summed to the x-axis value in order to calculate the BJD for each measurement. The y-axis represents the brightness of the star. This figure was created by following the Transit Light Curve Tutorial