Lecture 3

Structural Modeling Workflow

Tyler Ransom
ECON 6343, University of Oklahoma

Today

- What steps are required to estimate a structural model?
- Go through each step on an example model

Steps we won't discuss today

- The material we discuss today will already assume you have data
- And that you have sufficient understanding of your data
- It also assumes you have an understanding of your preferred coding language
- These are all non-trivial steps, but they are typically covered in other classes
- I will (indirectly) try to help you develop these skills throughout the course

Steps to performing structural estimation

Mike Keane gave a `talk` at the University of Chicago in 2015 and listed these steps:

1. Theoretical Model Development
2. Practical Specification Issues
3. Solving the Model
4. Understanding How the Model Works
5. Estimation
6. Validation
7. Policy Experiments

An example model

To help fix ideas, let's revisit a commonly used model in introductory econometrics:

Mincer equation:

$$\log(w_i) = \beta_0 + \beta_1 s_i + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i \tag{1}$$

where we have cross-sectional data and where

- i indexes individuals
- w_i is employment income
- s_i is years of schooling
- x_i is years of work experience (or, more commonly, *potential* work experience)
- ε_i is anything else that determines income

(how many years have you been out of school)

We want to estimate $(\beta_1, \beta_2, \beta_3)$, which are **returns to human capital investment**

Quick review

- It is nearly certain that (1) suffers from omitted variable bias
    - i.e. there are lots of factors in ε_i that are correlated with both $\boldsymbol{s}_i$ and $\boldsymbol{w}_i$
- Thus, our estimates of $(\beta_1, \beta_2, \beta_3)$ will not be causal
- We could try to get causal estimates using a variety of identification strategies:
- find a valid instrument for $\boldsymbol{s}_i$ (Angrist and Krueger, 1991; Card, 1995)
- exploit a discontinuity in $\boldsymbol{s}_i$ (Ost, Pan, and Webber, 2018)
- randomize $\boldsymbol{s}_i$ (Attanasio, Meghir, and Santiago, 2011)
- etc.

A structural view of Equation (1)

- We know that (1) will produced biased estimates, but *why*? Some possibilities:
- **ability bias**
  - s_i and w_i are both positively correlated with unobservable cognitive ability
- **comparative advantage**
  - multidimensional unobservable ability $\implies$ self-selection into schooling
- **credit constraints**
  - s_i is a costly investment; some people may not be able to borrow enough
- **preference heterogeneity** (differing tastes for s_i, differing discount rates)

(differences in cultural norms)

1. Theoretical Model Development

- Since schooling has an up-front cost and long-term benefit, need a dynamic model
  - period 1: decide how much schooling to get
  - period 2: choose whether or not to work; if working, receive income by (1)
  - individuals choose schooling level to maximize lifetime utility
- Preferences (denote utility in period t by u_t, with s, x and w defined previously)

$$\begin{aligned} u_1\left(z, c, \eta_1\right) &= f\left(z, c, \eta_1\right) \\ u_2\left(w\left(s, x\right), k, \eta_2\right) &= g\left(w\left(s, x\right), k, \eta_2\right) \end{aligned} \tag{2}$$

where z is family background, c is schooling costs, k is number of kids in adult household and η_t are unobservable preferences [similar to ε in (1)]

1. Theoretical Model Development

With discount factor $\delta \in [0, 1]$, the discounted lifetime utility function is then

$$V = u_1 (z, c, \eta_1) + \delta u_2 (w (s, x), k, \eta_2) \tag{3}$$

delta = 1: weigh future and present equally
delta = 0: weighs present more over future

- Equations (1)–(3) define our model
- This model is still **laughably unrealistic**, but at least we have something
- A number of important questions arise (But we'll ignore these for today)
  - Where is cognitive ability? What exactly does c represent? Where are loans?
  - Maybe people should care about *consumption* in period 2, not income
  - Does family background really only enter u_1 and not $\log(w)$?
  - Should x in (1) be a function of s? (Lower $s \implies$ longer working life)
  - What are people's beliefs about future k when deciding s?

Overview of the theoretical model

- As you can see, it takes a lot of know-how to write down even a simple model
- Requires knowledge about the subject and about math/econ more generally

Exogenous variables

- family background (z)
- schooling costs (c)
- children in household (k)

Endogenous variables

- schooling (s)
- period-2 work decision

Outcome variable

- labor income (w)

Unobservables

- income (ε)
- preferences (η_t)

Model parameters

- returns to human capital (β)
- discount factor (δ)
- other parameters implied by $f(\cdot)$ and $g(\cdot)$

2. Practical Specification Issues

- Now that we have a model, we need to figure out how to take it to data
- This is where we apply knowledge about **our data** and **stats/econometrics**
- Key data questions:
  - Can we observe the variables of the model in our data set?
  - If so, are they reliably measured?
- Key specification questions:
  - How to model η_t and ε? (Need to make distributional assumptions) (non-linear specifications?)
  - Functional forms of $f(\cdot)$ and $g(\cdot)$
  - Should s be continuous (years of schooling) or discrete (college/not)?

2. Practical Specification Issues

- We won't get into too many details about this today, but specification is important!
- What determines the specification is often:
  - what is reliably measured in the data
  - what is computationally feasible to estimate
- Parameters of the model either need to be **estimated** or **calibrated** (assuming a certain value for a parameter)
- e.g. often we don't have reliable data to allow us to estimate δ; we must calibrate it
- Computational feasibility often governs how we specify the different functions
  - e.g. *linear-in-parameters* with *additively separable* unobservables [like (1)]

Example with real data

For an empirical paper, look at data to see what can feasibly be included in the model.

- Here is some real data from the most recent round of the NLSY97

```
using CSV, DataFrames, Statistics
df = CSV.read("Data/slides3data.csv"; missingstrings=["NA"])
size(df)
# outputs (6009, 12)

describe(df)
# outputs the below:
12×8 DataFrame
```

Row	variable Symbol	mean Float64	min Real	median Float64	max Real	nunique Nothing	nmissing Union…
1	id	4534.71	4	4544.0	9022		
2	female	0.52671	0	1.0	1		

- We have demographics/background, wages, employment status, education, fertility
- N=6009, age $\in \{\mathbf{33}, \ldots, \mathbf{37}\}$, and 35% of respondents graduated college
- 24% have at least one college-graduate parent

Example: setting up the specification

- It looks like we can estimate some form of our model
- We have family background, cost of college (this is the `efc` variable)
- We have employment status, wage and number of children
- It looks like we'll have to have s be binary (`collgrad` variable)
- Also need to assume $x = age - 18$ if non-grad, $x = age - 22$ if grad (Mincer, 1974)
- Then we just need to add some functional form assumptions, and we'll be ready
  - $\varepsilon \sim$ Normal, $\eta_t \sim$ Logistic
  - $u_{i1} = \alpha_0 + \alpha_1 \, \textbf{parent\_college} + \alpha_2 \, \textbf{efc} + \eta_1$
  - $u_{i2} = \gamma_0 + \gamma_1 \mathbb{E} \log w_i + \gamma_2 \, \textbf{numkids} + \eta_2$

Parameters of the empirical model

- We can now detail the parameters of the empirical model
- **wage parameters** $(\beta, \sigma_\varepsilon)$
  - The latter is the std. dev. of income shocks
- **schooling parameters** (α)
- **employment parameters** (γ, δ)
- Then write down a statistical objective function as a fn. of data and parameters
  - e.g. maximize the likelihood, or minimize the sum of the squared residuals
- We'll learn how to do this in later classes, but not today

3. Solving and 4. Understanding How the Model Works

- **Solving the model:**
  - solve the dynamic utility max problem for given parameter values
  - (we aren't estimating parameter values yet) Once the model is solved, we update it with parameters and solve the model again.
  - (we will talk about how to do this next week)
- **Understanding the model:**
  - simulate data from the model (Simulated data does not need to match the actual data if the parameters are set to certain values)
  - make sure the simulated data is consistent with the model's implications

Then, look at descriptive stats from the simulated data.

3. Solving and 4. Understanding How the Model Works

- Start with as simple of a model as possible; make sure things are working
- When introducing more complexities, do "numerical comparative statics"
- Make sure the parameters move in the correct directions
  - e.g. $\uparrow \beta_1 \implies \uparrow$ schooling (ceteris paribus)
- If they don't, you've likely got a bug somewhere

Example with real data

- How would we do this in Julia?

- We can simulate log wages and then see how close we got

- This is kind of silly in our simple model, but the workflow is there

First, set the parameters (beta and sigma)

```
N = size(df,1)
β = [1.65,.4,.06,-.0002]
σ = .4;
df.exper = df.age .- ( 18*(1 .- df.collgrad) .+ 22*df.collgrad )
df.lwsim = β[1] .+ β[2]*df.collgrad .+ β[3]*df.exper .+ β[4]*df.exper.^2 .+ σ*randn(N)
df.lw    = log.(df.wage)
```

- We can then compare how `df.lwsim` compares with `df.lw` in the data

  Comparing lwsim and lw, we might made wrong assumptions about sigma or other parameters. Variance of lwsim seems to be lower than lw.

```
describe(df;cols=[:lw,:lwsim])
# returns
│ Row │ variable │ mean     │ min      │ median   │ max      │ nunique │ nmissing │ eltype    │
```

5. Estimation

- Most structural models require **nonlinear estimation**
- e.g. MLE/GMM or their simulated counterparts
- In nonlinear optimization, starting values are crucial
- Initializing at random starting values is likely to give poor results
- Keane recommends calibrating the model by hand
    - e.g. match the intercept of each equation to the $\overline{Y}$'s in the data

For the logit models, play with simulations such as

- I recommend estimating an intercepts-only model (or with very few X's)
- But this advice is model-specific!

5. Estimation

- There are lots of algorithms for nonlinear optimization

- We'll talk more about these later in the course

- Your next problem set will show how to do this in Julia

Example using real data

- In our simple model, we can get good starting values by estimating OLS and logits
- The wage equation can be estimated by OLS (on the subsample who are employed)

```
using GLM
β̂ = lm(@formula(lw ~ collgrad + exper + exper^2), df[df.employed.==1,:])
# returns
Coefficients:
─────────────────────────────────────────────────────────────────────────────────
                 Estimate   Std. Error    t value  Pr(>|t|)    Lower 95%   Upper 95%
─────────────────────────────────────────────────────────────────────────────────
(Intercept)    2.94607     0.323145      9.11688     <1e-18    2.31255     3.57959
collgrad       0.534326    0.0271395    19.6881      <1e-82    0.481119    0.587532
exper         -0.0265561   0.0412115    -0.644386    0.5194   -0.107351    0.0542385
exper ^ 2      0.0014304   0.00132307    1.08112     0.2797   -0.00116346  0.00402426
─────────────────────────────────────────────────────────────────────────────────

df.elwage = predict(β̂, df) # generates expected log wage for all observations

r2(β̂)                                # reports R2
sqrt(deviance(β̂)/dof_residual(β̂))  # reports root mean squared error
```

Example using real data

- The u_t equations can be estimated as simple logits (on the full sample)

```
α̂ = glm(@formula(collgrad ~ parent_college + efc), df, Binomial(), LogitLink())
# returns
Coefficients:
────────────────────────────────────────────────────────────────────────────────
                 Estimate  Std. Error   z value  Pr(>|z|)   Lower 95%   Upper 95%
────────────────────────────────────────────────────────────────────────────────
(Intercept)     -1.20091   0.0364888   -32.9118    <1e-99  -1.27243    -1.1294
parent_college   1.47866   0.068433     21.6074    <1e-99   1.34453     1.61278
efc              0.0450253 0.00437704   10.2867    <1e-24   0.0364464   0.0536041
────────────────────────────────────────────────────────────────────────────────

γ̂ = glm(@formula(employed ~ elwage + numkids), df, Binomial(), LogitLink())
# returns
Coefficients:
────────────────────────────────────────────────────────────────────────────
```

Do these results make sense?

- It can be informative to try and interpret even these simple results
- wage equation:
  - insignificant return to experience is surprising; otherwise makes sense
- schooling choice:
  - If `efc` captures college costs, it should have a negative sign
  - This suggests omitted variable bias in this equation
- employment choice:
  - These results check out; may want to introduce nonlinearities in `numkids`

6. Validation

- If you have a good model, it should be **valid** (i.e. predict well out of sample)
  Predict data on 80% of the sample, and use rest 20% of the sample for validation purpose.
- Validation is not always possible, but it's good to do if you can
- e.g. if experimental data, estimate on control group, validate on treatment group
- e.g. see if model can replicate major policy change in data
- More simply, you could throw out half your data, then try to predict other half
  - This is typically not done if the full sample isn't huge

7. Policy Experiments

- This is the main reason to do structural estimation!
- Structural estimation $\implies$ recovering the DGP of the model
- Once we know the DGP, we can simulate data from it and do policy experiments
    - requires having policy-invariant parameters!
- We can predict the effects of:
    - proposed policies
    - hypothetical policies
- Contrast with RCTs, which only reveal effects of implemented policies

Example using real data

- We have two policy variables we could play with

    1. `efc` (i.e. how much gov't subsidizes college tuition & fees)
    2. return to schooling (this could change due to e.g. technological change)

- Here's how we would look at a counterfactual with lower cost:

```
df_cfl      = deepcopy(df)
df_cfl.efc = df.efc .- 1          # change value of efc to be $1,000 less
df.basesch = predict(α̂, df)      # predicted collgrad probabilities under status quo
```

- Average likelihood of `collgrad` *declines* from 35% to 34.2%

- This doesn't make sense because the `efc` coefficient didn't make sense

Example using real data

- We can't assess the counterfactual of increasing the return to schooling
- Because `elwage` doesn't directly enter the `collgrad` logit model
- This is because we aren't really estimating the dynamic model yet
- We will learn how to do this in the near future

In summary: Why structural estimation?

- Want to examine effects of policies not yet implemented
- Learn more about economics by looking through the lens of a model
- Assess performance of theoretical models in explaining real-world data
- Can be used to build up long-run "canonical" models of behavior in many areas
- It can be really fun to do more complicated econometrics beyond simple regressions
- Observational data is much cheaper to collect than experimental data

In summary: Why *not* structural estimation?

- It's really difficult to write down and estimate a tractable, realistic model!
- It requires additional effort beyond data preparation and running regressions
- Understanding identification of the model takes a lot of effort, too
- It can be really miserable to try and debug the code of a structural estimation
- Many structural models can take weeks to estimate one specification
    - in addition to months spent coding/debugging beforehand
- As you can see, even with a simple model things have already gotten complicated!

References

Angrist, J. D. and A. B. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?" In: *Quarterly Journal of Economics* 106.4, pp. 979-1014. DOI: 10.2307/2937954.

Attanasio, O. P, C. Meghir, and A. Santiago (2011). "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA". In: *Review of Economic Studies* 79.1, pp. 37-66. DOI: 10.1093/restud/rdr015.

Card, D. (1995). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling". In: *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*. Ed. by L. N. Christofides, E. K. Grant and R. Swidinsky. Toronto: University of Toronto Press.

Mincer, J. (1974). *Schooling, Experience and Earnings*. New York: Columbia University Press for National Bureau of Economic Research.

Ost, B, W. Pan, and D. Webber (2018). "The Returns to College Persistence for Marginal Students: Regression Discontinuity Evidence from University Dismissal Policies". In: *Journal of Labor Economics* 36.3, pp. 779-805. DOI: 10.1086/696204.