Causal inference examples

Jussi Keppo
keppo@nus.edu.sg

Modeling steps

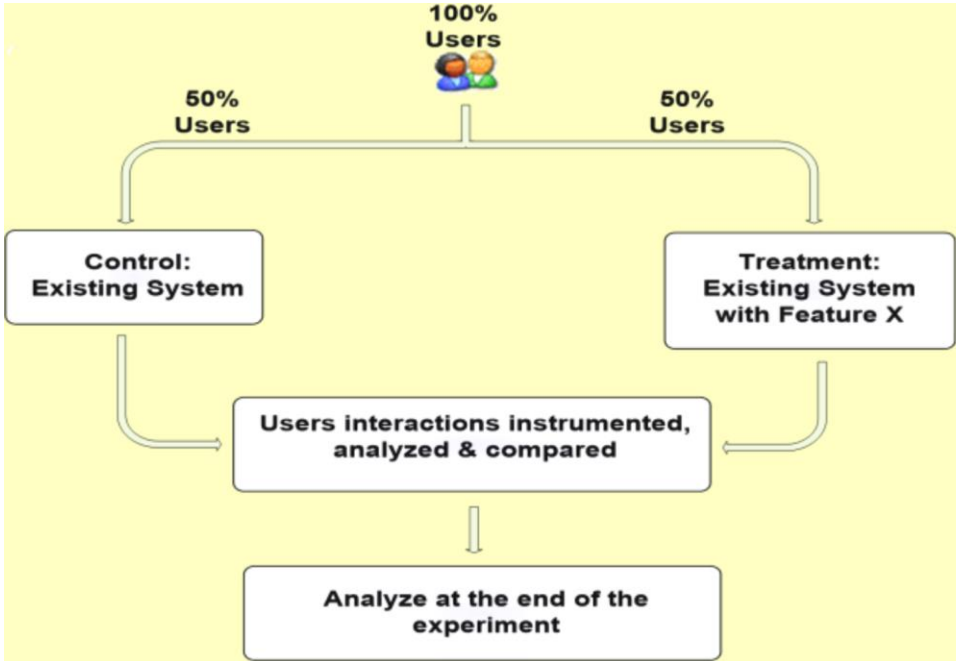
- Stage 1: Define the research problem
  - What is research or business problem and objectives? For instance, forecasting or finding a causal relationship
  - What techniques/models should be used?
- Stage 2: Develop the analysis plan
  - Implementation issues (sample sizes, allowable variable types, estimator)
  - This includes data collection
- Stage 3: Check the model assumptions
  - Are the underlying assumptions of the chosen model(s) satisfied?
  - E.g., normality, linearity, independence of error terms, equality of variances
- Stage 4: Evaluate overall model fit
  - Does the model(s) achieve acceptable levels on statistical criteria (e.g., significance)?
  - Are the proposed relationships identified? Is the result practically significant?
- Stage 5: Interpret the variables
  - Analyzing effects for individual variables by examining the estimated coefficients
  - Is there empirical evidence of relationships that can be generalized?
- Stage 6: Validate the model
  - How does the model(s) work with out-of-sample data?
  - Demonstrate the generalizability of the results to total population

Identification of causal relationships

- Four identification strategies:
  - Randomized experiment
  - Regression discontinuity
  - Difference in differences
  - Instrumental variables
- Estimating causal relationship is important e.g. when we change our policy
- If the objective is to just forecast (e.g. future returns or volatility) then we do need the above identifications
  - Just have a “horse race” between model candidates and check which works the best in out-of-sample

Randomized experiment example: A/B Testing

- A/B Testing, also known as bucket testing, split testing, is controlled/randomized experiment in the Internet Age
  - Online controlled experiment
- It is a crucial step in the Lean Startup Method and also heavily used in large online service companies including Amazon, EBay, Facebook, Google, Microsoft, Netflix, etc.
- Like clinical trial for drugs, decisions on website features are heavily based on A/B testing



Regression discontinuity

- Example: The impact of minimum drinking age on mortality

Age Profiles for Death Rates in the United States

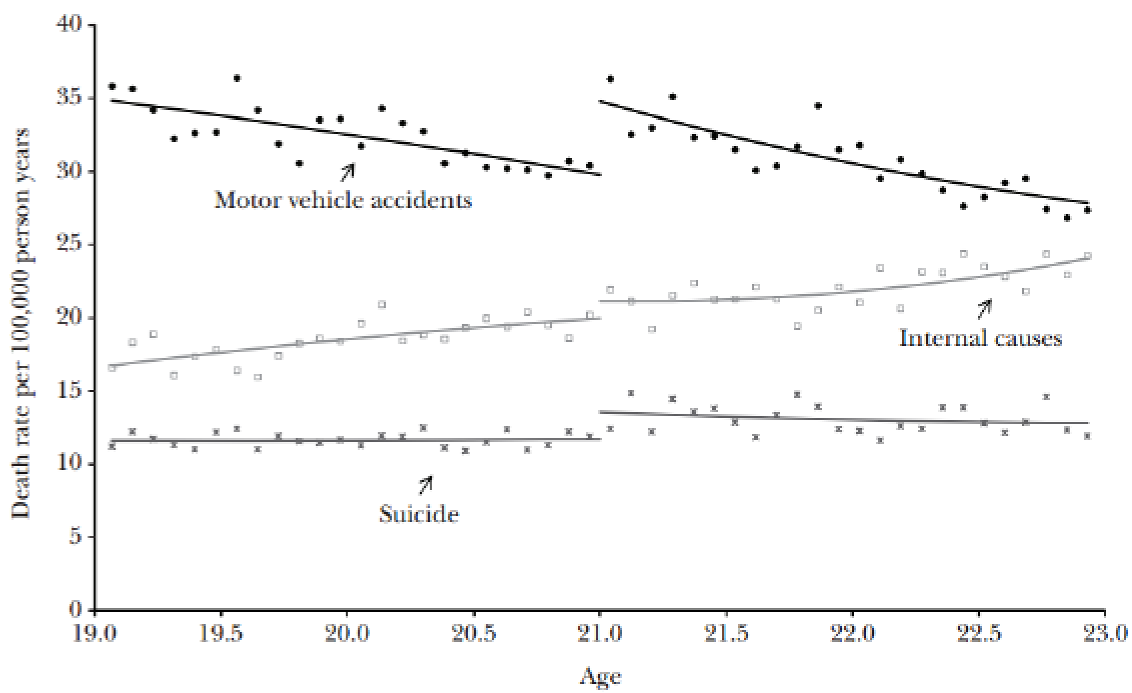
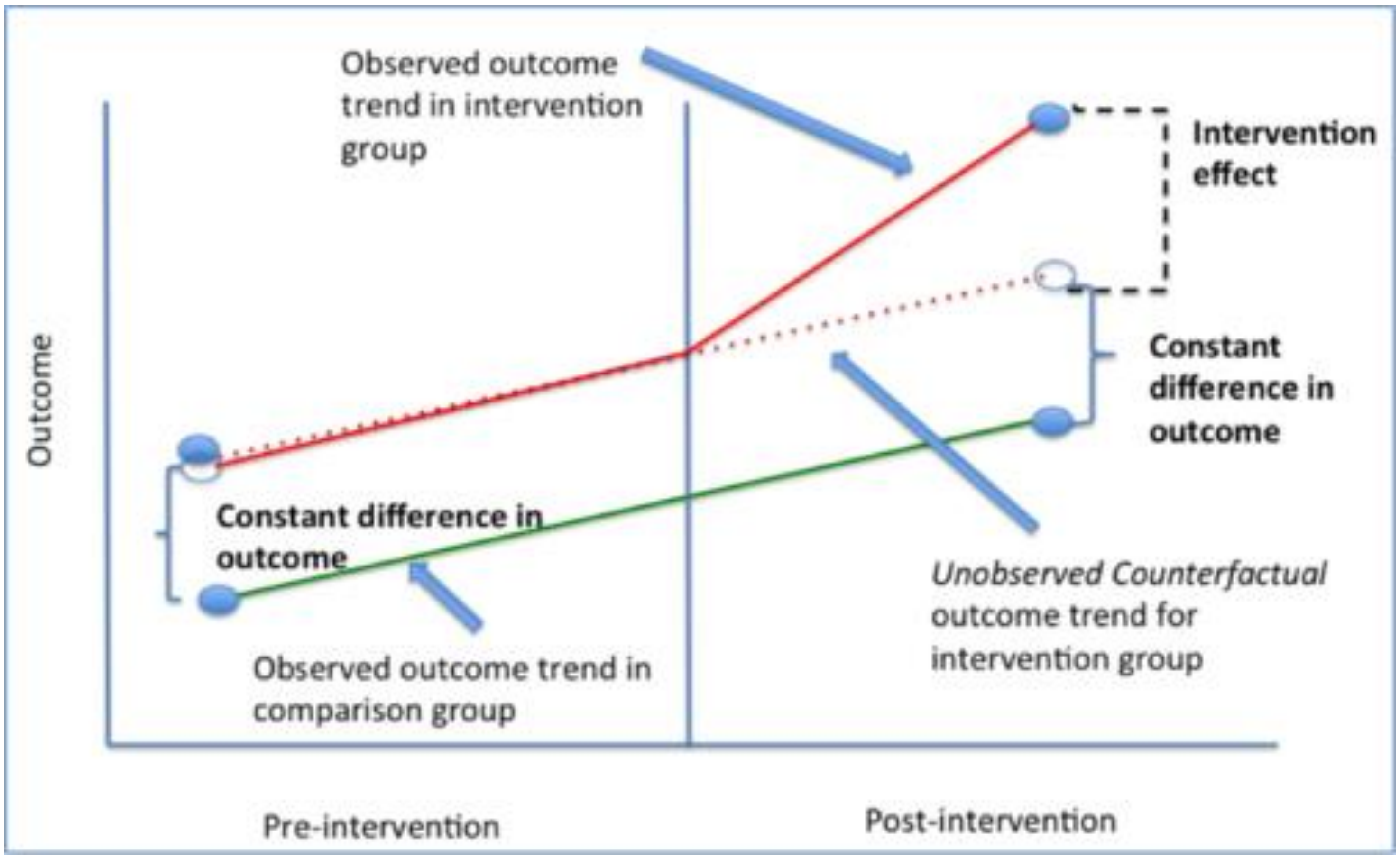


Figure 1: Death rates by age by type.

Difference in differences (DiD)



Regression implementation of DiD

- Can use unit fixed effects regression
  - *post* = dummy for "after" (post-treatment) period
  - f_i = unit dummies
  - t=0 (before) or 1 (after)

$$y_{it} = \alpha + f_i + \beta \cdot t + \delta \cdot w_i \cdot t + \varepsilon_{it}$$

- Or first difference form:

$$\Delta y_{it} = \beta + \delta \cdot w_i + \varepsilon_i$$

DiD Example: Banking regulation

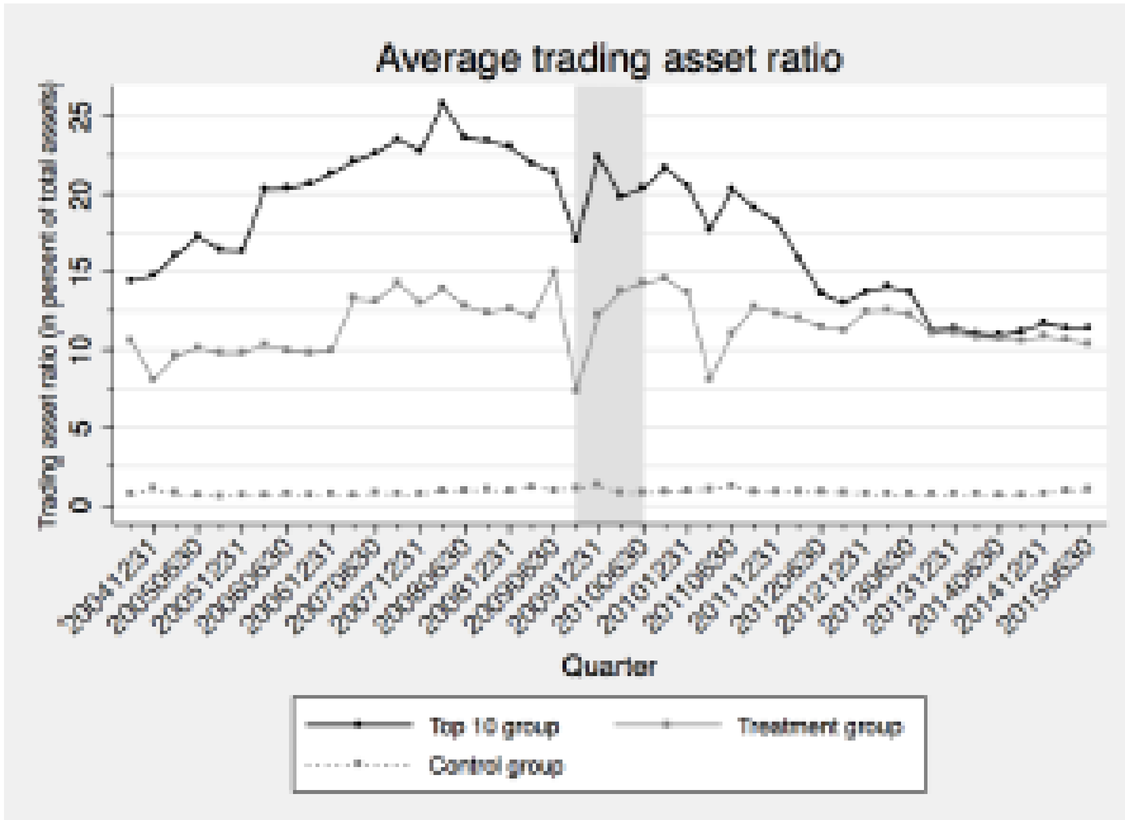
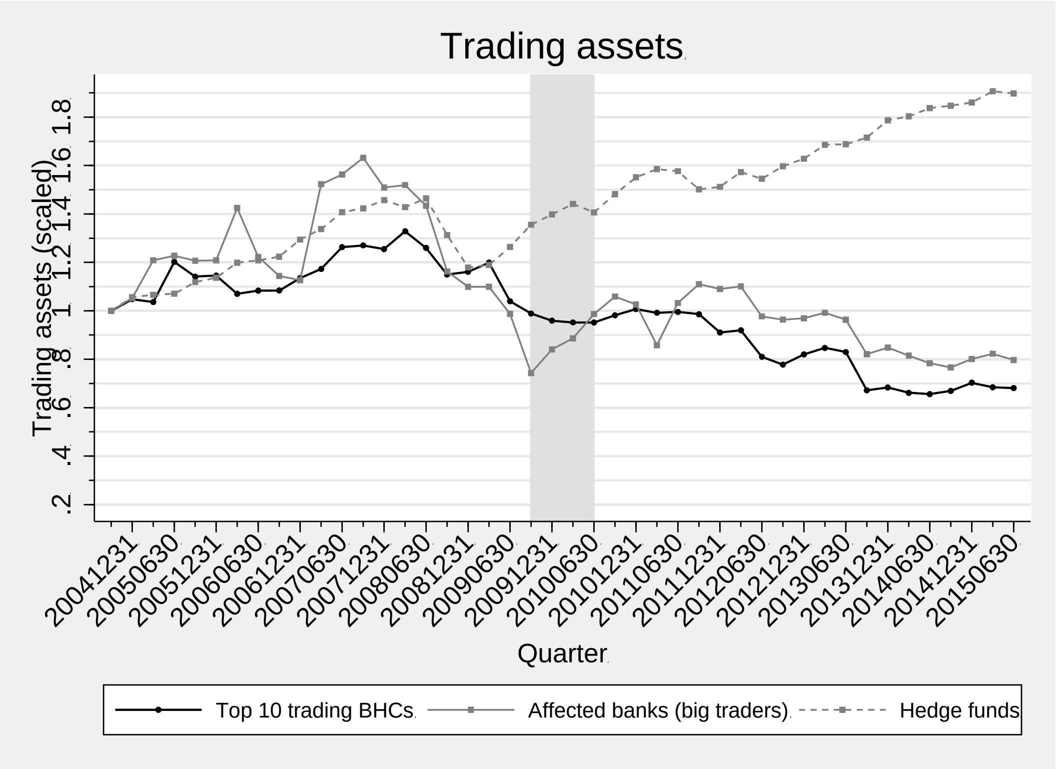
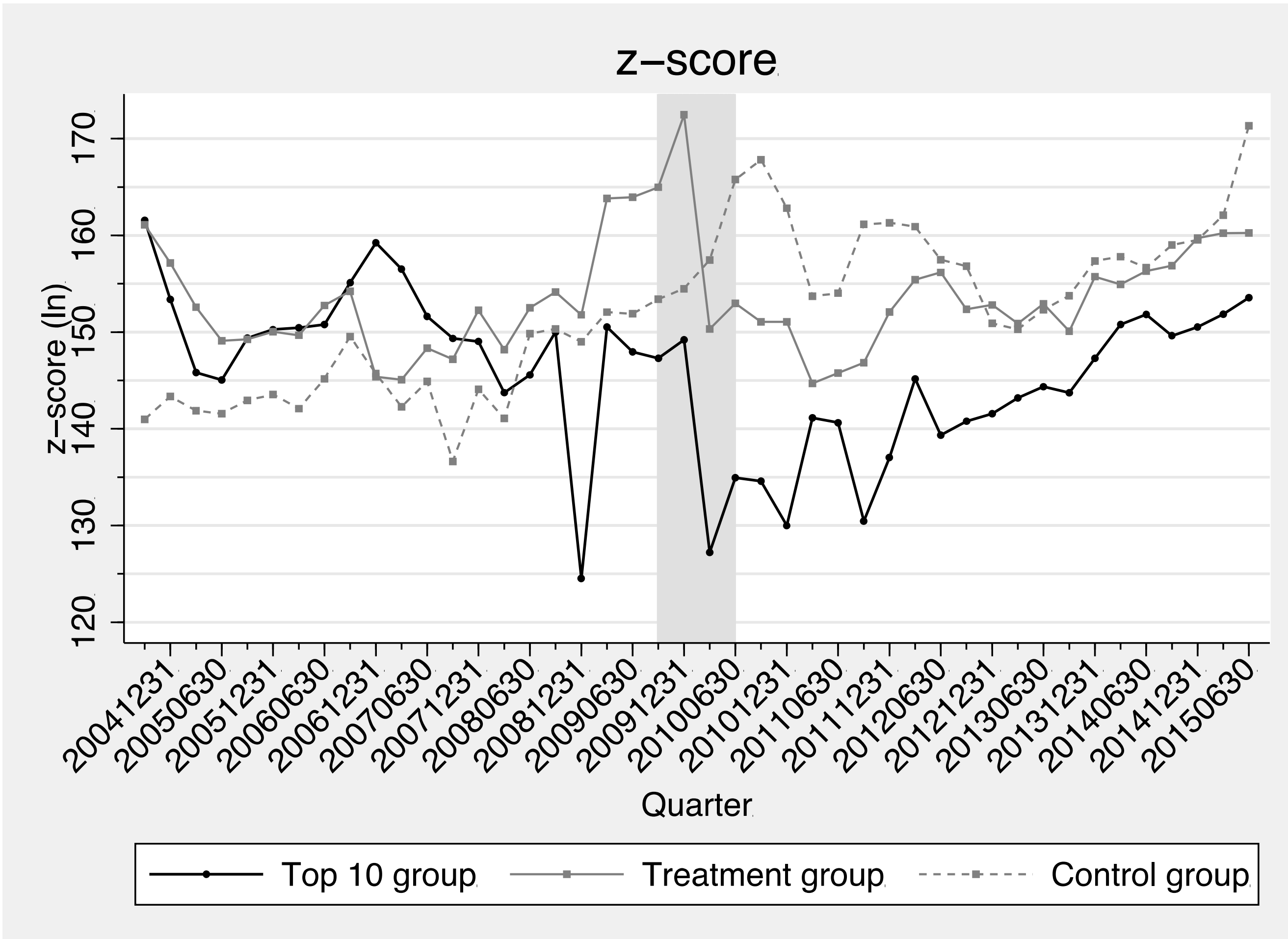


Figure 1 Average trading asset ratio of banks in three different groups.
This figure plots the average trading asset ratio of the 10 bank holding companies with the highest trading asset ratio in the 15 quarters before 2007. Banks with average trading asset ratio greater than 3% during the same period are in the treatment group. Banks with non-zero but less than 3% average trading asset ratio with the closest propensity score with the banks in treatment group are in the control group. The vertical gray area is the Volcker Rule's announcement time period, 2009 Q3 – 2010 Q2.

DiD Example: Banking regulation, Cont’d



DiD Example: Banking regulation, Cont'd



DiD Example: Minimum wages

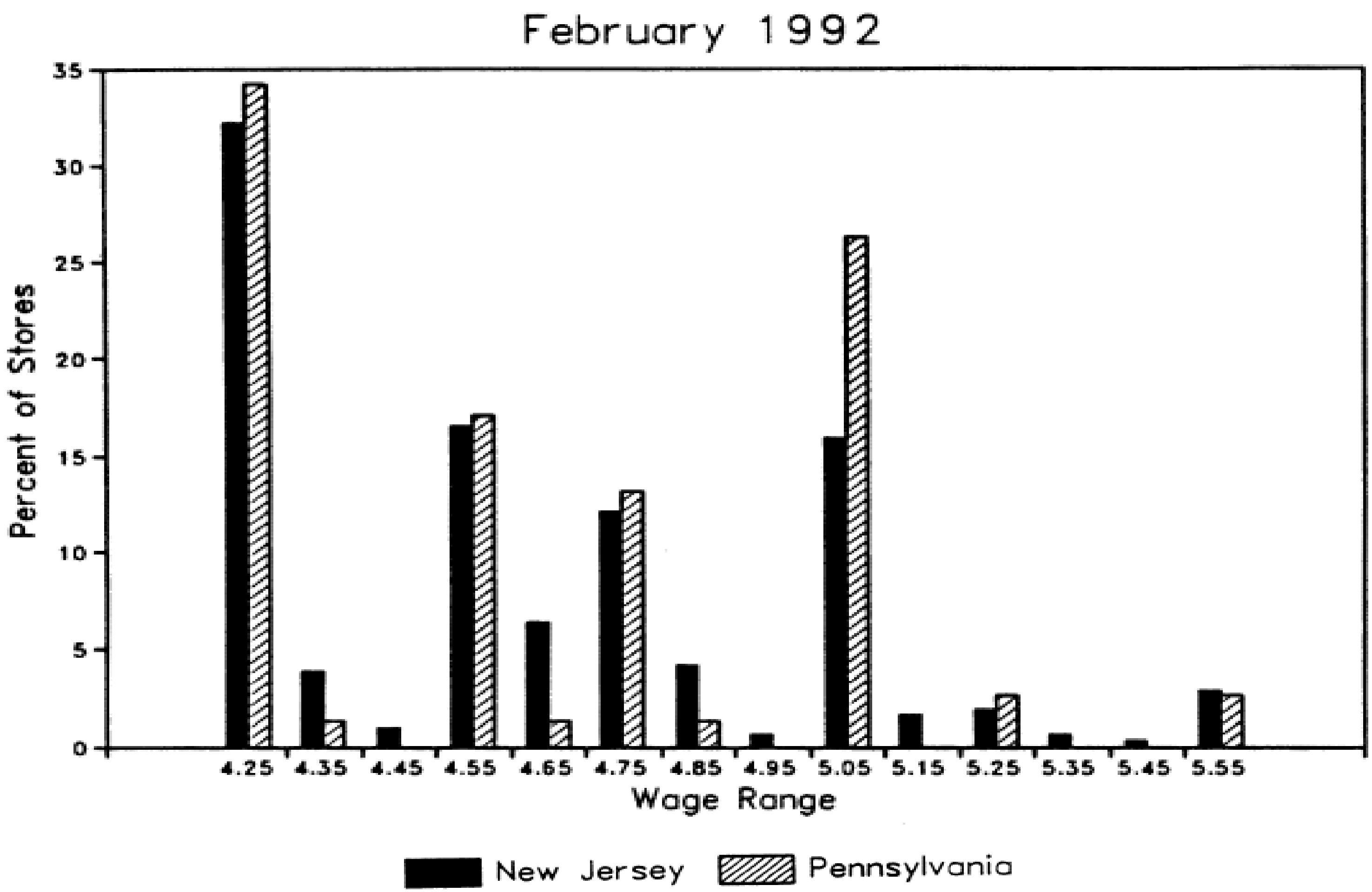
- Research question: Do (modestly) higher minimum wages reduce low-wage employment?
  - Card & Krueger (1994)

lower employment

- Card and Krueger consider impact of New Jersey's 1992 minimum wage increase from $4.25 to $5.05 per hour
  - In 2013$: Equivalent to an Increase from $7.00 to $8.30
    - Compare US minimum wage [$7.25]; IL minimum wage: S8.25
- Compare 410 fast-food restaurants in New Jersey (treated) to eastern Pennsylvania (control) before and after the increase
- Data on wages and employment:
  - March & Dec 1992, one month before; 8 months after increase
- Note the "local" nature of the question:
  - Microeconomic theory: raise minimum **enough** → lower employment
    - Higher prices → lower equilibrium demand
    - Over time, higher labor cost → substitute capital for labor

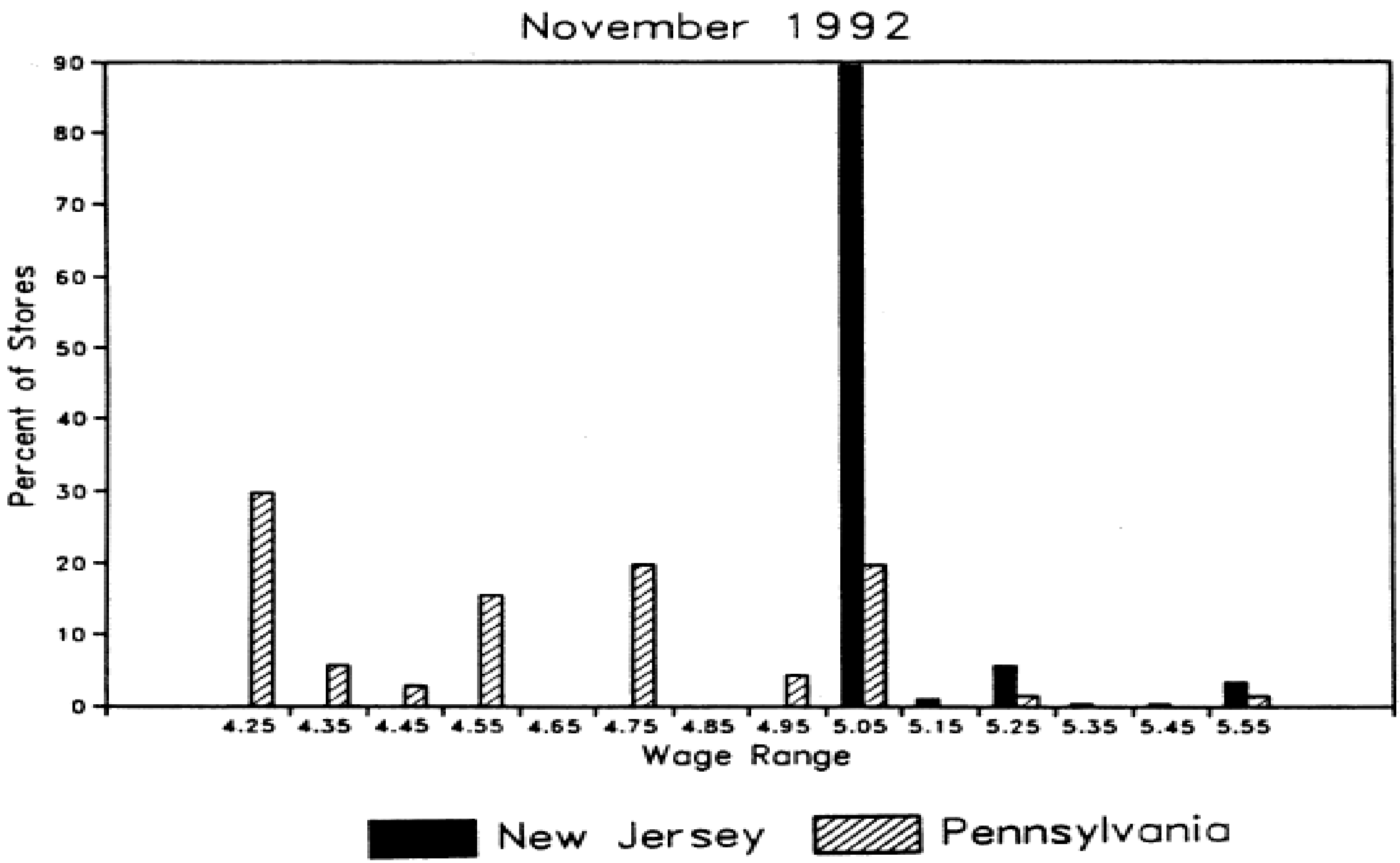
Minimum wages, Cont'd

- Figure below gives wages before the minimum wage increase
- Note:
  - National minimum = $4.25
  - new minimum is within range that some already pay.



Minimum wages, Cont’d

- Figure below gives wages after the minimum wage increase



Minimum wages, Cont'd

Selected Card & Krueger results:

Full-time equivalent employment, per restaurant.

Drop 6 restaurants which closed from pre to post; 4 which temporarily closed.

Time	PA	NJ	NJ - PA
Before	23.33	20.44	-2.89
	(1.35)	(0.51)	(1.44)
After	21.17	21.03	-0.14
	(0.94)	(0.52)	(1.07)
After - Before	-2.16	0.59	**2.76****
	(1.25)	(0.54)	**(1.36)**

NJ looks better after minimum wage increase

But effect is **entirely** because PA employment declines

Puts great stress on the parallel changes assumption.

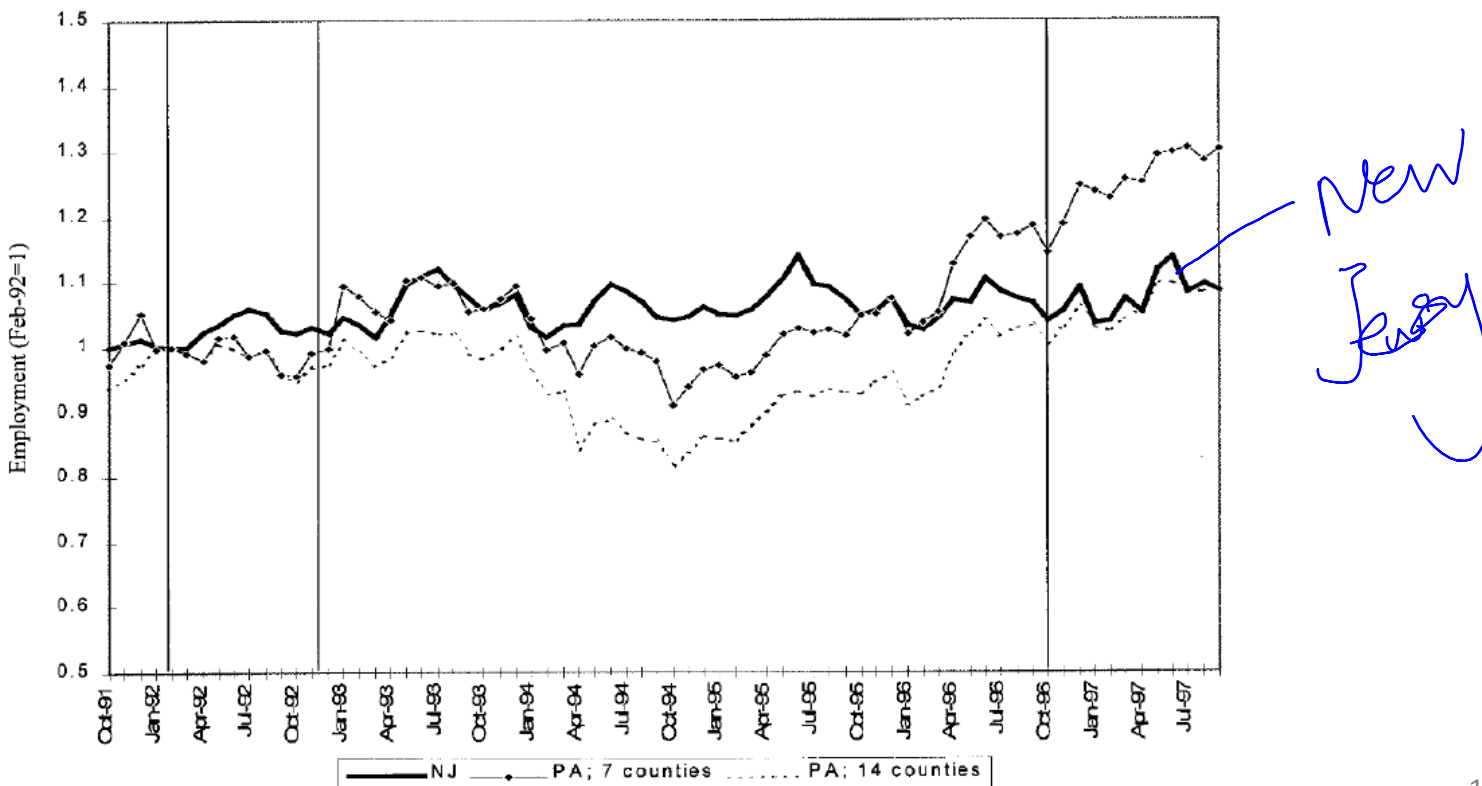
Come back to this assumption . . .

Minimum wages, Cont'd

Longer trends in NJ v. PA fast-food employment:
One surely wouldn't conclude that NJ employment ***rose***
Better research design: pre-period data for a longer time

    Data not available

Still, no strong evidence of NJ employment **drop**



Instrumental Variables: Earning-schooling

- Consider the following OLS regression:

$$y = \beta_1 x + \underbrace{(\beta_2 m + e)}_{u}$$

β_1 will be bia

where y = log of earnings
 x = years of schooling
 m = learning ability
 e, u = error terms

- We want to measure how an *exogenous* change in x affects y (i.e., b_1) :
  - there is association between x and u
  - as a consequence, there is both a direct effect via $b_1 x$ and an indirect effect via u affecting x which in turn effects y
  - thus, b_1 is biased

Earning-schooling, Cont’d

- Definition of an instrument:

$$\begin{tikzcd} z \arrow[r] & x \arrow[r] & y \\ & u \arrow[u] \arrow[ur] & \end{tikzcd}$$

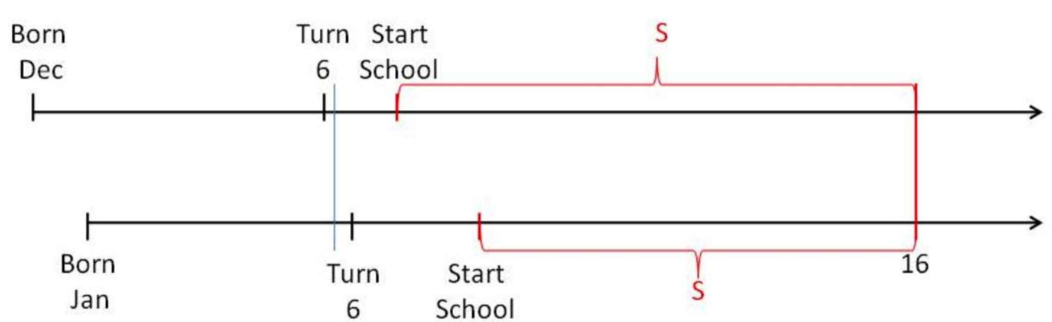
- A variable *z* is called an instrument or instrumental variable for the regressor *x* in the scalar regression model if

  (1) *z* is uncorrelated with error *u* (i.e. cov(*z*, *u*) = 0)

  (2) *z* is correlated with the regressor *x* (i.e. cov(*z*, *x*) ≠ 0)

Earning-schooling, Cont'd

- Selecting the instrument requires domain knowledge:
  - In the returns to education literature Angrist and Krueger (1991) used *quarter of birth* as an instrument al variable for schooling
  - In the US, you could drop out of school once you turned 16
  - Children have different ages when they start school and thus different lengths of schooling at the time they turn 16 when they can potentially drop out

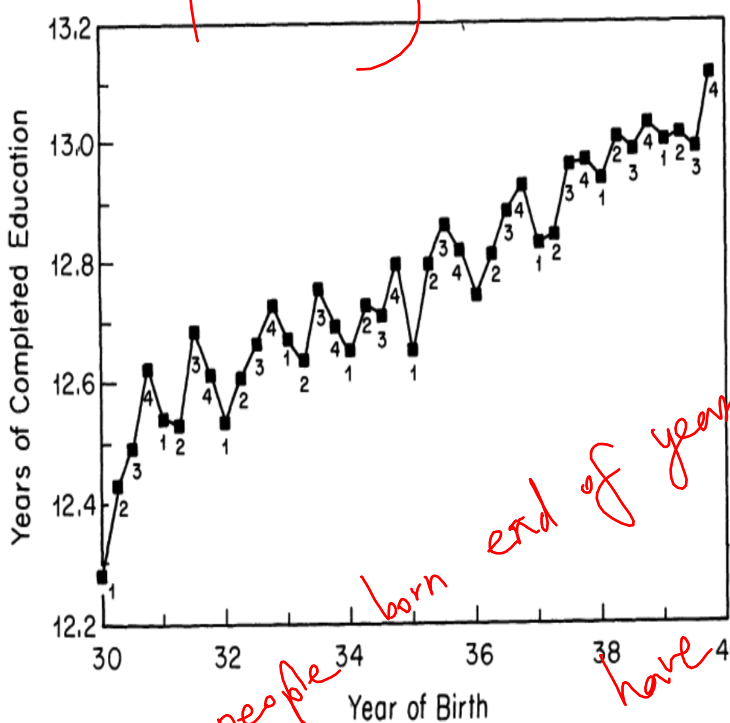


Earning-schooling, Cont'd

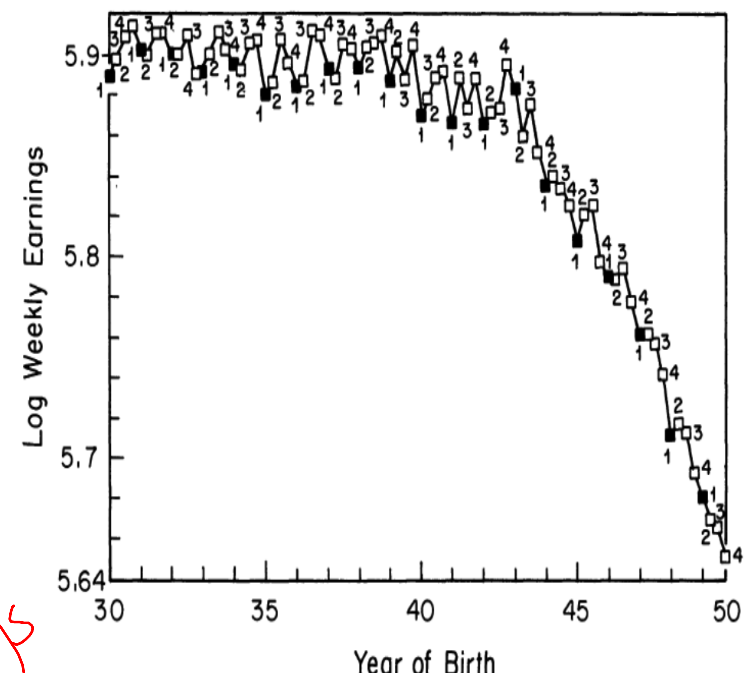
- First condition: cov(z,x)≠0
- Men born earlier in the year have lower schooling. This indicates that there is a first stage.

- Second condition: cov(z,u)=0
- Difference in schooling due to different quarter of birth translates into different earning with a fixed age. This indicates that there is a second stage.

quarter by



people born end of year have higher earnings



Earning-schooling, Cont'd

baseline is quarter 4

all negative

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect[a] I	II	III	F-test[b] [P-value]
Total years of education	1930–1939	12.79	−0.124	−0.086	−0.015	24.9
			(0.017)	(0.017)	(0.016)	[0.0001]
	1940–1949	13.56	−0.085	−0.035	−0.017	18.6
			(0.012)	(0.012)	(0.011)	[0.0001]
High school graduate	1930–1939	0.77	−0.019	−0.020	−0.004	46.4
			(0.002)	(0.002)	(0.002)	[0.0001]
	1940–1949	0.86	−0.015	−0.012	−0.002	54.4
			(0.001)	(0.001)	(0.001)	[0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004	0.051	0.012	5.9
			(0.014)	(0.014)	(0.014)	[0.0006]
	1940–1949	14.28	0.005	0.043	−0.003	7.8
			(0.011)	(0.011)	(0.010)	[0.0017]
College graduate	1930–1939	0.24	−0.005	0.003	0.002	5.0
			(0.002)	(0.002)	(0.002)	[0.0021]
	1940–1949	0.30	−0.003	0.004	0.000	5.0
			(0.002)	(0.002)	(0.002)	[0.0018]

- The table reports estimates of each quarter of birth (main) effect relative to the fourth quarter
- The F-tests indicate that after removing trends, the small within-in-year-of-birth differences in average years of education are highly statistically significant
  - F-statistics is for a test of the hypothesis that the quarter-of-birth dummies jointly have no effects
- Quarter of birth is a strong predictor of total years of education
- Reassuring quarter of birth does not affect the probability of graduating from college

Earning-schooling, Cont’d

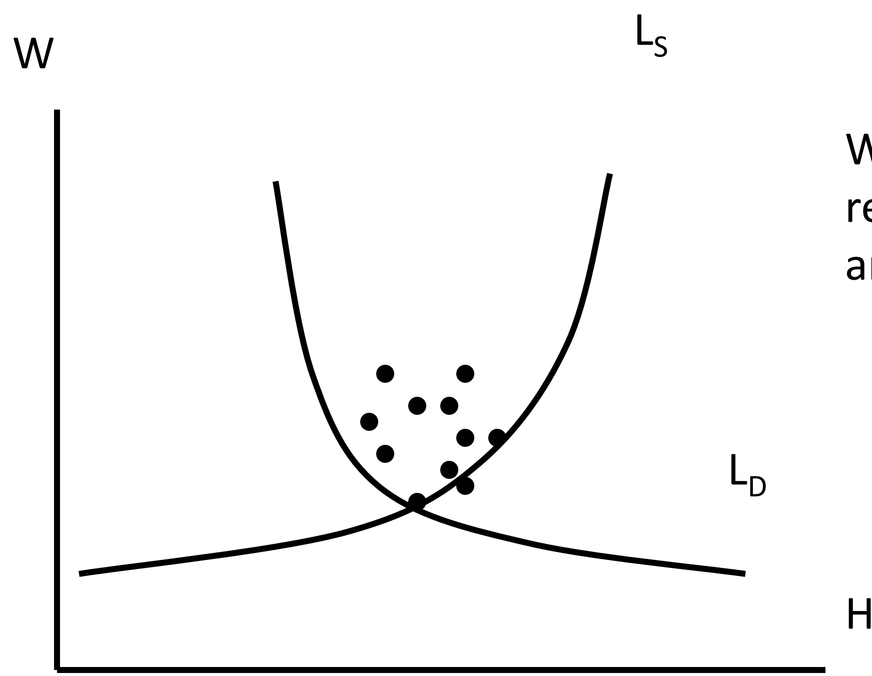
- Estimation results with quarter of birth IV dummies:

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS
Years of education	0.0711 (0.0003)	0.0891 (0.0161)	0.0711 (0.0003)	0.0760 (0.0290)
Race (1 = black)	—	—	—	—
SMSA (1 = center city)	—	—	—	—
Married (1 = married)	—	—	—	—
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No
Age	—	—	−0.0772 (0.0621)	−0.0801 (0.0645)
Age-squared	—	—	0.0008 (0.0007)	0.0008 (0.0007)
χ^2 [dof]	—	25.4 [29]	—	23.1 [27]

- The 2SLS estimates do not equal the OLS estimates in this over-identified model, and the standard errors from OLS regression is much smaller than those from 2SLS regression

Instrumental Variables: Labor demand and supply

- Two variables: hours and wages
- Can construct a graph of labor demand and supply



We want to identify the relationship between hours and wages for workers

Labor demand and supply, Cont'd

- In terms of labor demand and labor supply:

  (D): $H_i = a_1 + b_1 W_i + u_i$
  (S): $H_i = a_2 + b_2 W_i + c_2 X_i + d_2 NK_i + v_i$

  where X_i = age and NK_i = number of kids

- We assume labor market equilibrium, or $H_i^D = H_i^S$
- We also assume that hours and wages are endogenous
  - Both are determined at the same time: either could be placed as the dependent variable for the model
  - Since hours and wages are determined simultaneously, there is a relationship between wages and the error term u_i

Labor demand and supply, Cont'd

- Solution: Two stage least squares
- First stage: use OLS to estimate reduced form for the right-hand side endogenous variable W_i:

$$W_i = \pi_{10} + \pi_{11}X_i + \pi_{12}NK_i + \varepsilon_{1i}$$

- Second stage: use predictions from the first stage for $\hat{W}_i$:

$$H_i = a_1 + b_1\hat{W}_i + u_i$$

  – Then we estimate this model using the proxy for wages