# ECONOMICS AND MACHINE LEARNING

# Identification

- Model does not allow causal effect (or a particular parameter) to be determined

- Lack of identification of causal parameters from data e.g. because <u>E (X'U) =/= 0</u>
  - OLS $\tilde{\beta}$ is biased i.e. E($\tilde{\beta}$)=/= β ; $\tilde{\beta}$ is also inconsistent
    - X is no longer exogenous
  - Simultaneous equation i.e. some X is also endogenous (Simultaneity bias)
    - P determines Q and Q determines P. Observed data is equilibrium $P^*$ and $Q^*$
  - Unobserved variable correlated to observed (Omitted variable bias)
    - $\ln wage_i = \tilde{\alpha} + \tilde{\beta} * schooling_i + \tilde{\gamma} * experience_i + u_i$ ;

      Missing variable ability / motivation
    - $\ln wage_i = \tilde{\alpha} + \tilde{\beta}_0 * schooling_i + \tilde{\gamma}_0 * experience_i + \tilde{\theta} * Ability + \varepsilon_i$ ;

      Cannot identify $\tilde{\beta}_0$ from joint distribution of wages, schooling and experience alone
  - Selection bias
    - Sample selection based on Y
    - Self selection e.g. into a training program

      $\ln wage_i = \tilde{\alpha} + \tilde{\beta} * schooling_i + \tilde{\gamma} * experience_i + \tilde{\theta} * D_i + u_i$ ; where $D_i$ =1 if participated in training program

  - Measurement error

# Treatment Effect

- Counterfactual: a situation different from the current
  - An objective of policy analysis: examining outcomes under different hypothetical states of the world
    - New policy being considered: interested in outcomes under the new policy
    - Evaluation of current policy: interested in what would have happened in the absence of a policy
  - Non-policy intervention/treatment
    - Hospitalization
    - Age?
- Treatment effect (Rubin's Causal Model)
  - Considers causal effect through definition of potential outcomes or counterfactuals rather than parameters of a regression model
  - Causal effect of treatment (assignment status) $D_i$ =0 or 1 on $Y_i$
    - $Y_{i1}$ is outcome for i if $D_i$ =1; $Y_{i0}$ if $D_i$=0
    - Individual causal (or treatment) effect is $Y_{i1}(D_i =1) – Y_{i0}(D_{i\ 0}=0)$
    - But both $Y_{i1}$ and $Y_{i0}$ not observed for the same individual

      $Y_i^{obs}= Y_{i0}(D_i)* (1-D_i) + Y_{i1}(D_i) *D_i$ potential outcomes under two states 0,1 based on observed $D_i$=0 or 1
    - Instead compare outcomes across different individuals to get average treatment effect
    - Non-parametric approach to evaluating impact of treatment i.e. treatment effect
  - Assignment of treatment
    - Experimental data: Random or assignment mechanism is a known function of observable characteristics of the units

      Outcome ($Y_i$) and assignment status ($D_i$) are independent

      Identification achieved through randomization
    - Observational data: Parts of "assignment" are unknown, based on unobservable characteristics. Choice decision.

      Find instrumental variable ($Z_i$) that is independent of outcomes but related to participation indicator ("assignment") $D_i$

      Natural experiment: program or treatment is implemented in a way that results in an exogenous instrument that enables identification of causal effect

# Measurement of Treatment Effect

- Measures of treatment effect
  - Average treatment effect (ATE) $E[Y_{i1} – Y_{i0}]$
  - Average treatment effect on the treated (ATET) $E[Y_{i1} – Y_{i0} \mid D_i = 1] = E[Y_{i1} \mid D_i = 1] - E[Y_{i0} \mid D_i = 1]$
    - Actual ($E[Y_{i1} \mid D_i = 1$ ) and counterfactual outcome ($E[Y_{i0} \mid D_i = 1)]$
  - Counterfactual not observed but use a control group or econometric technique to get a consistent estimate
  - Law of large numbers: Use sample averages for population averages e.g. ATE $\hat{\tau} = \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}$

- Corresponding regression approach where α is average causal effect
  - $Y_i = \alpha D_i$ ; $\boldsymbol{\alpha = Y_{i1} – Y_{i0}}$ when random assignment
  - If $Y_i$ depends on observables $X_i$ that may be correlated with assignment status $D_i$; assume independence of Y and D given X
    - i.e. $E[Y_{i0} \mid X_i, D_i] = X'_i\beta$ then regression $Y_i = \alpha D_i + X'_i\beta$ gives

$$E\{Y_i(D_i - E[D_i \mid X_i])\} / E\{D_i(D_i - E[D_i \mid X_i])\} = \alpha.$$

This is a weighted average of treatment – control contrasts for cells defined by X

  - If $Y_i$ depends on unobservables that may be correlated with treatment $D_i$?
    - Is there a way to estimate the causal effect α?

# Experimental Data

- Examining the impact of  a policy decision using observational data
  - Policy variable becomes an explanatory variable
  - Estimate coefficient using appropriate techniques
  - Concern about identification of causal effects

- Randomized Control Trials: perform experiment and collect data from sampled units before and after the intervention
  - Define intervention as treatment
    - Policy variable
    - Price?
  - Randomly selected sample
    - Randomly selected units are treated
    - Distribution of  characteristics in treatment and control groups similar on unobservable and observable dimensions other than for the intervention.
    - $E(X'u)= 0$
    - Broadly speaking, counterfactual outcomes captured by outcomes for control group
    - In experiments econometrics less important compared to using field data: the econometrics is built into the design
    - Simple measure of impact of the program is the difference in means for the two groups
      - Add covariates to improve precision of estimates or correct for impact of missing data
- Practical challenges in conducting RCTs
  - Compliance
  - Contamination
  - Ethical concerns for some policy related interventions
  - Effective randomization is key
- Preferred approach now for many development economics research questions

# Examples of Experimental Data

- Man made experiments
  - Job training programs
  - RCT in the areas of health , education etc.

https://www.povertyactionlab.org/evaluations

  - The Impact of Training Informal Healthcare Providers in India
  - An Entrepreneurial Model of Community Health Delivery in Uganda
  - Graduating Microenterprises to Larger Loans in Egypt

- Natural experiments
  - Outcomes independent of tastes, abilities
    - Twins
    - Mandatory schooling laws
    - Introduction of a policy in selective states

# Instrumental Variables Estimation

- When RCT is not possible use econometric techniques to get unbiased estimates

- Instrumental variables (IV), 2SLS

  $y = \beta_0 + \beta_1 x + u,$ $\qquad \mathrm{Cov}(x,u) \neq 0.$

  – Find a variable z that is correlated with x but not u; Z is an instrumental variable (IV)

  $\mathrm{Cov}(z,x) \neq 0.$ $\qquad \mathrm{Cov}(z,u) = 0;$

    - Can check by regressing X on Z; sign of coefficient should be used to justify use of z as an instrument
    - Choice of Z based on economic reasoning.

  – Example

    ln $wage_i = \tilde{\alpha} + \tilde{\beta} *$ $schooling_i$ $+ \tilde{\gamma} *$ $experience_i + u_i$ ;

    ln $wage_i = \tilde{\alpha} + \tilde{\beta} *$ $schooling_i$ $+ \tilde{\gamma} *$ $experience_i + \tilde{\theta} *$ Ability $+\varepsilon_{i\,;}$

    - IQ is proxy for ability but not an instrument for schooling
    - Mothers schooling? Number of siblings?
    - **IV estimator** is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})}.$$

- IV estimator is biased in small samples but consistent.
- Randomization can be shown to equivalent to an instrument
- IV estimate is a function of LATE under certain conditions

# 2SLS Estimator

- Two stage least squares (2SLS) is a form of IV estimator

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1.$$

- Cov($y_{2,}$ $u_i$) =/= 0
  - OLS for $\beta_1$ does not give causal effect

- Consider instruments $Z_2$ and $Z_3$ for $y_2$ ($Z_1$ cannot be used as IV for $y_1$ as it appears in the structural equation)
  - First stage reduced form

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2,$$

$$\mathrm{E}(v_2) = 0, \mathrm{Cov}(z_1,v_2) = 0, \mathrm{Cov}(z_2,v_2) = 0, \text{and } \mathrm{Cov}(z_3,v_2) = 0.$$

  - Predicted value from first stage

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

  - Regress $y_1$ on $\hat{y}_2$ and $z_1$. OLS will give the 2SLS estimator for $\beta_1$

## Measurement of Treatment Effect – IV Approach

$$\text{potential earnings} = \begin{cases} Y_i(1,1) & \text{if } D_i = 1, Z_i = 1 \\ Y_i(1,0) & \text{if } D_i = 1, Z_i = 0 \\ Y_i(0,1) & \text{if } D_i = 0, Z_i = 1 \\ Y_i(0,0) & \text{if } D_i = 0, Z_i = 0 \end{cases}$$

- If $Y_i$ depends on unobservables that may be correlated with treatment $D_i$
  - Assume constant effects model $\alpha = Y_{i1} - Y_{i0}$ and let $Y_{i0} = \beta + \varepsilon_i$

    $Y_i = \beta + \alpha D_i + \varepsilon_i$
  - Consider instrument $Z_i$ for $D_i$ such that $E[\varepsilon_i \mid Z_i] = 0$
  - If $Z_i = 1$ or 0 then taking expectation of above equation with $Z_i = 1$ or 0

    $$\{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]\} / \{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]\} = \alpha.$$
  - IV estimator: Sample analog is Wald Estimator of (average) treatment effect

- If heterogeneous treatment effects e.g. positive for some and negative for others
  - IV methods fail to capture ATE or ATET
  - Instead under fairly general assumptions IV methods capture local average treatment effect (LATE) = ATE for individuals whose treatment status is influenced by an exogenous regressor that satisfies an exclusion restriction

$$E(Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}) = \frac{E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)}{E(D_i \mid Z_i = 1) - E(D_i \mid Z_i = 0)}.$$

  - LATE is average treatment effect for compliers (instrument specific sub-population):

    Population has compliers ($D_{i1} = 1$ and $D_{i0} = 0$); always-takers ($D_{i1} = 1$ and $D_{i0} = 1$) and never-takers ($D_{i1} = 0$ and $D_{i0} = 0$)
  - ATET is a weighted average of effects on always-takers and compliers (LATE).
  - ATE s a weighted average of effects on never–takers, always-takers and compliers
  - Conditional local average treatment effect (CLATE)
  - Heterogeneity in treatment effect may result in large number of covariates and their interactions

# Scope of ML Applications in Econometric Analysis

ML techniques are in the prediction space

- Use in economic problems focussed on prediction: Making a prediction
  - Would traditional econometric estimation techniques necessarily be good predictors?
    - In sample and out of sample prediction
  - Techniques particularly useful for high dimensional models
    - Allows inclusion of a large number of covariates/predictors in non-linear and highly interactive ways
  - Applications
    - Identifying a target population
    - Finance: predicting stock prices
- Apply the ML approach to estimation problems: Making a decision
  - Address problems such as the fact that unlike true Y in validation sample being observed, true parameter is never observed
  - The techniques should give an estimate and as associated distribution for inference
    - Standard errors?
  - Applications
    - Use in instrumental variable estimation: prediction at the first stage

      Selecting covariates to control for heterogeneity

# ML Applications in Prediction Problems in Economics

- Hiring decisions: Use ML to predict worker productivity in order to make better decisions about hiring or promoting workers
  - Teacher quality predictions based on characteristics: used to improve hiring practices
  - Hiring decisions for police personnel
- Health care decisions: Use ML to predict health outcomes
  - Mortality rate predictions based on health indicators are used to better target the beneficiaries of health care such as surgeries, palliative care
  - Personalized medicine
- Policy targeting decisions:  predicting economic well-being at a granular level for purpose of policy targeting
  - Identifying the poor based on luminosity, mobile usage data, crop images etc.
  - Targeted transfer programs based on predicted poverty when there are measurement issues

# Applications- Predicting House Values

- Sendhil Mullainathan and Jann Spiess. *Machine Learning : An Applied Econometric Approach* Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 87–106

*Table 1*

**Performance of Different Algorithms in Predicting House Values**

| Method | Prediction performance ($R^2$) Training sample | Hold-out sample | Relative improvement over ordinary least squares by quintile of house value 1st | 2nd | 3rd | 4th | 5th |
|---|---|---|---|---|---|---|---|
| Ordinary least squares | 47.3% | 41.7% [39.7%, 43.7%] | – | – | – | – | – |
| Regression tree tuned by depth | 39.6% | 34.5% [32.6%, 36.5%] | –11.5% | 10.8% | 6.4% | –14.6% | –31.8% |
| LASSO | 46.0% | 43.3% [41.5%, 45.2%] | 1.3% | 11.9% | 13.1% | 10.1% | –1.9% |
| Random forest | 85.1% | 45.5% [43.6%, 47.5%] | 3.5% | 23.6% | 27.0% | 17.8% | –0.5% |
| Ensemble | 80.4% | 45.9% [44.0%, 47.9%] | 4.5% | 16.0% | 17.9% | 14.2% | 7.6% |

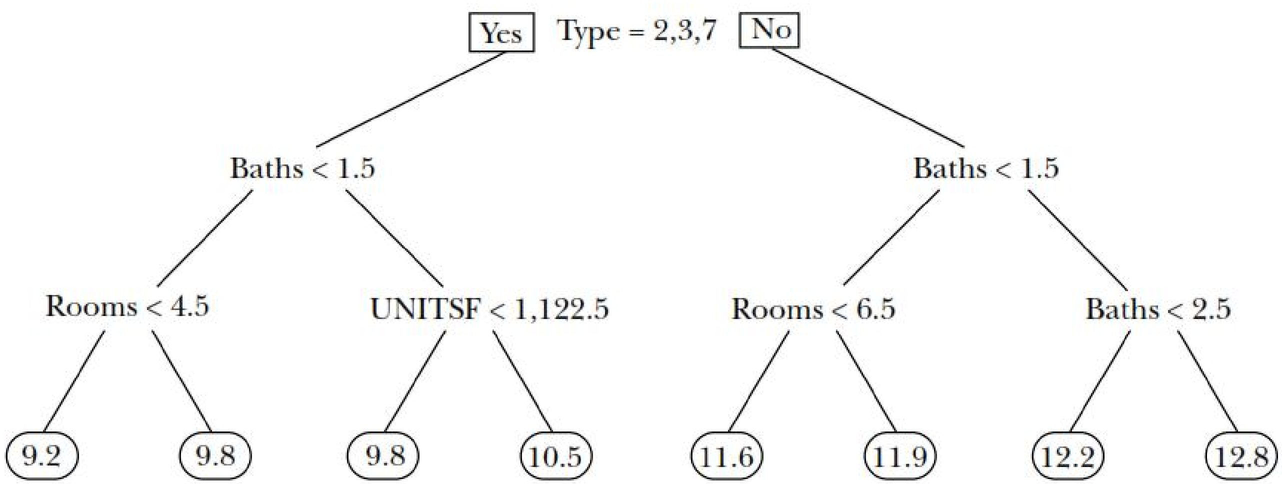
*Note:* The dependent variable is the log-dollar house value of owner-occupied units in the 2011 American Housing Survey from 150 covariates including unit characteristics and quality measures. All algorithms are fitted on the same, randomly drawn training sample of 10,000 units and evaluated on the 41,808 remaining held-out units. The numbers in brackets in the hold-out sample column are 95 percent bootstrap confidence intervals for hold-out prediction performance, and represent measurement variation for a fixed prediction function. For this illustration, we do not use sampling weights. Details are provided in the online Appendix at http://e-jep.org.

# Applications- Predicting House Values

- Sendhil Mullainathan and Jann Spiess. *Machine Learning : An Applied Econometric Approach* Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 87–106

*Figure 1*

**A Shallow Regression Tree Predicting House Values**




*Note:* Based on a sample from the 2011 American Housing Survey metropolitan survey. House-value predictions are in log dollars.

- Each leaf corresponds to a set of interactive dummies
  - For left most leaf

$$x_1 = 1_{TYPE=2,3,7} \times 1_{BATHS<1.5} \times 1_{ROOMS<4.5}$$

$$\alpha_1 = 9.2$$

# Applications

- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. *Prediction Policy Problems.* American Economic Review, May 2015
- Consider two (policy) decisions
  - Given drought conditions: Should one invest in rain dance? ($X_0$)
  - Seeing clouds: should one take an umbrella? ($X_0$)
  - Both require study of rain (Y)
    - Do rain dances cause rain? Vs Likelihood of rain high enough?

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial \pi}{\partial X_0} \underbrace{(Y)}_{\text{prediction}} + \frac{\partial \pi}{\partial Y} \underbrace{\frac{\partial Y}{\partial X_0}}_{\text{causation}}$$

    - Benefits of a policy depend on the (predicted) level of a variable
- MSE of predicton = $E_D[(\hat{f}(x) - y)^2] = \underbrace{E_D[(\hat{f}(x) - E_D[\hat{y}_0])^2]}_{\text{Variance}} + \underbrace{(E_D[\hat{y}_0] - y)^2}_{\text{Bias}^2}.$
- ML techniques
  - Use data to decide how to make bias variance trade-offs
  - Allow search over a large set of variable and functional forms

- Application to predicting which joint replacement surgeries would be futile
  - Costs of surgery (monetary and non-monetary) vs Benefits
  - Payoff depends upon how long patient lives after surgery
  - Predict mortality rates post surgery
    - Average mortality rates :1.2% die within a month, 4.2% within a year
    - Question is if the surgeries on predictably riskiest were futile
    - Beyond that, how do we allocate scarce resources to the patients for whom the causal effect on patient welfare is largest ?

Table 1—Riskiest Joint Replacements

| Predicted mortality percentile | Observed mortality rate | Futile procedures averted | Futile spending ($ mill.) |
|---|---|---|---|
| 1 | 0.435 (0.028) | 1,984 | 30 |
| 2 | 0.422 (0.028) | 3,844 | 58 |
| 5 | 0.358 (0.027) | 8,061 | 121 |
| 10 | 0.242 (0.024) | 10,512 | 158 |
| 20 | 0.152 (0.020) | 12,317 | 185 |
| 30 | 0.136 (0.019) | 16,151 | 242 |

*Notes:* We predict 1–12 month mortality using an $L_1$ regularized logistic regression trained on 65,395 Medicare beneficiaries undergoing joint replacement in 2010, using 3,305 claims-based variables and 51 state indicators. $\lambda$ was tuned using ten-fold cross-validation in the training set. In columns 1 and 2 we sort a hold-out set of 32,695 by predicted risk into percentiles (column 1) and calculate actual 1–12 month mortality (column 2). Columns 3 and 4 show results of a simulation exercise: we identify a population of eligibles (using published Medicare guidelines: those who had multiple visits to physicians for osteoarthritis and multiple claims for physical therapy or therapeutic joint injections) who did not receive replacement and assign them a predicted risk. We then substitute the high risk surgeries in each row with patients from this eligible distribution for replacement, starting at *median* predicted risk. Column 3 counts the futile procedures averted (i.e., replaced with non-futile procedures) and column 4 quantifies the dollars saved in millions by this substitution.

# Predicting Outcomes vs. Determination of Causal Effects

Susan Athey*. Beyond prediction: Using big data for policy problems.* SCIENCE, Feb 2017

- Increasing use of off-the-shelf ML prediction techniques
  - Few assumptions: stability of environment; no interactive effects across units being studied
  - Applications do not need knowledge of context

- What would happen under a variety of alternate policies that have never been implemented?

- Heterogeneity of effects
  - Can customer churn be reduced by outreach to those most likely to churn based on a predictive model
    - Focus on those most likely to respond to outreach – causal effect of outreach
    - Overlap in group being targeted and those who respond may small
  - Impact of increasing prices on quantity demanded
    - ML can predict what hotel occupancy rates if prices on a given day were high
    - Can not determine impact of e.g. a 5% price increase everywhere

# Application for use of ML in IV Estimation

- Children and their Parents Labour Supply (Angrist and Evans 1998)
- Is there a causal link between female labour supply and fertility (number of children)?
  - Many studies showed negative correlation between female labour supply and fertility
  - But economic models have considered hours worked being function of number of children and fertility a function of wages
  - Both are endogenous and jointly determined i.e. may be the case that men/women who desire a bigger family size also have a weaker preference for labour
  - Paper uses a unique IV to estimate causal effect: gender mix of children with two or more children
    - Parental preference for a sibling gender mix, parents of same-gender siblings more likely to have an additional child
    - Exogenous source of variation in family size
    - Compliers : those who had a third child because they had same sex children; always takers: those who would have had an additional child irrespective of gender mix of first two children; never-takers: those who would not have had a third child irrespective of gender mix of first two children

- US data
  - 1980 Census and 1990 census: Public Use Micro Samples (PUMS)
  - Sample restricted to households with mothers between 21-35, with two or more children and eldest child being less than 18
  - Gender of the eldest sibling defined same-sex sibling pairs
  - No fertility history, linking of information on household members

- Fertility model
  - Utility from number of children and quality of children
  - Child quality function of inputs and parents times in home production; gender mix increases utility
  - Fertility measure – more than two children
    - IV is same sex siblings : Virtually random assignment
- Labour supply model: individual and household
  - Two sub-samples: females, married females
  - Other covariates – age, education, race of mother
  - Labour supply measures – work in year prior to Census
    - If Work for pay, used Weeks worked, hours worked, annual labour income

TABLE 2—DESCRIPTIVE STATISTICS, WOMEN AGED 21–35 WITH 2 OR MORE CHILDREN

| | Means and (standard deviations) | | | | | |
|---|---|---|---|---|---|---|
| | 1980 PUMS | | | 1990 PUMS | | |
| | | Married couples | | | Married couples | |
| Variable | All women | Wives | Husbands | All women | Wives | Husbands |
| *Children ever born* | 2.55<br>(0.81) | 2.51<br>(0.77) | — | 2.50<br>(0.76) | 2.48<br>(0.74) | — |
| *More than 2 children* (=1 if mother had more than 2 children, =0 otherwise) | 0.402<br>(0.490) | 0.381<br>(0.486) | — | 0.375<br>(0.484) | 0.367<br>(0.482) | — |
| *Boy 1st* ($s_1$) (=1 if first child was a boy) | 0.511<br>(0.500) | 0.514<br>(0.500) | — | 0.512<br>(0.500) | 0.514<br>(0.500) | — |
| *Boy 2nd* ($s_2$) (=1 if second child was a boy) | 0.511<br>(0.500) | 0.513<br>(0.500) | — | 0.511<br>(0.500) | 0.512<br>(0.500) | — |
| *Two boys* (=1 if first two children were boys) | 0.264<br>(0.441) | 0.266<br>(0.442) | — | 0.264<br>(0.441) | 0.265<br>(0.441) | — |
| *Two girls* (=1 if first two children were girls) | 0.242<br>(0.428) | 0.239<br>(0.427) | — | 0.241<br>(0.428) | 0.239<br>(0.426) | — |
| *Same sex* (=1 if first two children were the same sex) | 0.506<br>(0.500) | 0.506<br>(0.500) | — | 0.505<br>(0.500) | 0.503<br>(0.500) | — |
| *Twins-2* (=1 if second birth was a twin) | 0.0085<br>(0.0920) | 0.0083<br>(0.0908) | — | 0.012<br>(0.108) | 0.011<br>(0.105) | — |
| *Age* | 30.1<br>(3.5) | 30.4<br>(3.4) | 33.0<br>(4.6) | 30.4<br>(3.5) | 30.7<br>(3.3) | 33.4<br>(4.8) |
| *Age at first birth* (parent's age in years when first child was born) | 20.1<br>(2.9) | 20.8<br>(2.9) | 24.0<br>(4.0) | 21.8<br>(3.5) | 22.4<br>(3.5) | 25.1<br>(4.7) |
| *Worked for pay* (=1 if worked for pay in year prior to census) | 0.565<br>(0.496) | 0.528<br>(0.499) | 0.977<br>(0.150) | 0.662<br>(0.473) | 0.667<br>(0.471) | 0.968<br>(0.175) |
| *Weeks worked* (weeks worked in year prior to census) | 20.8<br>(22.3) | 19.0<br>(21.9) | 48.0<br>(10.5) | 26.2<br>(22.9) | 26.4<br>(22.9) | 47.1<br>(12.0) |
| *Hours/week* (average hours worked per week) | 18.8<br>(18.9) | 16.7<br>(18.3) | 43.5<br>(12.3) | 22.5<br>(19.1) | 22.2<br>(18.9) | 44.0<br>(13.3) |
| *Labor income* (labor earnings in year prior to census, in 1995 dollars) | 7,160<br>(10,804) | 6,250<br>(10,211) | 38,919<br>(25,014) | 9,550<br>(13,071) | 9,616<br>(13,238) | 36,623<br>(30,283) |
| *Family income* (family income in year prior to census, in 1995 dollars) | 42,342<br>(26,563) | 47,646<br>(25,821) | — | 42,558<br>(34,692) | 49,196<br>(34,740) | — |
| *Non-wife income* (family income minus wife's labor income, in 1995 dollars) | — | 41,635<br>(24,734) | — | — | 39,580<br>(31,892) | — |
| Number of observations | 394,835 | 254,654 | 254,654 | 380,007 | 301,588 | 301,588 |

*Notes:* The samples include women aged 21–35 with two or more children except for women whose second child is less than a year old. In the 1980 PUMS, the married women sample refers to women who were married at the time of their first birth, married at the time of the survey, and married once. In the 1990 PUMS, the married women are those married at the time of the Census.

TABLE 3—FRACTION OF FAMILIES THAT HAD ANOTHER CHILD BY PARITY AND SEX OF CHILDREN

| Sex of first child in families with one or more children | All women | | | | Married women | | | |
|---|---|---|---|---|---|---|---|---|
| | 1980 PUMS (649,887 observations) | | 1990 PUMS (627,362 observations) | | 1980 PUMS (410,333 observations) | | 1990 PUMS (477,798 observations) | |
| | Fraction of sample | Fraction that had another child | Fraction of sample | Fraction that had another child | Fraction of sample | Fraction that had another child | Fraction of sample | Fraction that had another child |
| (1) one girl | 0.488 | 0.694 (0.001) | 0.489 | 0.665 (0.001) | 0.485 | 0.720 (0.001) | 0.487 | 0.698 (0.001) |
| (2) one boy | 0.512 | 0.694 (0.001) | 0.511 | 0.667 (0.001) | 0.515 | 0.720 (0.001) | 0.513 | 0.699 (0.001) |
| difference (2) − (1) | — | 0.000 (0.001) | — | 0.002 (0.001) | — | 0.000 (0.001) | — | 0.001 (0.001) |

| Sex of first two children in families with two or more children | All women | | | | Married women | | | |
|---|---|---|---|---|---|---|---|---|
| | 1980 PUMS (394,835 observations) | | 1990 PUMS (380,007 observations) | | 1980 PUMS (254,654 observations) | | 1990 PUMS (301,588 observations) | |
| | Fraction of sample | Fraction that had another child | Fraction of sample | Fraction that had another child | Fraction of sample | Fraction that had another child | Fraction of sample | Fraction that had another child |
| one boy, one girl | 0.494 | 0.372 (0.001) | 0.495 | 0.344 (0.001) | 0.494 | 0.346 (0.001) | 0.497 | 0.331 (0.001) |
| two girls | 0.242 | 0.441 (0.002) | 0.241 | 0.412 (0.002) | 0.239 | 0.425 (0.002) | 0.239 | 0.408 (0.002) |
| two boys | 0.264 | 0.423 (0.002) | 0.264 | 0.401 (0.002) | 0.266 | 0.404 (0.002) | 0.264 | 0.396 (0.002) |
| (1) one boy, one girl | 0.494 | 0.372 (0.001) | 0.495 | 0.344 (0.001) | 0.494 | 0.346 (0.001) | 0.497 | 0.331 (0.001) |
| (2) both same sex | 0.506 | 0.432 (0.001) | 0.505 | 0.407 (0.001) | 0.506 | 0.414 (0.001) | 0.503 | 0.401 (0.001) |
| difference (2) − (1) | — | 0.060 (0.002) | — | 0.063 (0.002) | — | 0.068 (0.002) | — | 0.070 (0.002) |

*Notes:* The samples are the same as in Table 2. Standard errors are reported in parentheses.

TABLE 4—DIFFERENCES IN MEANS FOR DEMOGRAPHIC VARIABLES
BY *SAME SEX* AND *TWINS*-2

| | Difference in means (standard error | |
|---|---|---|
| | By *Same sex* | |
| Variable | 1980 PUMS | 1990 PUMS |
| *Age* | −0.0147<br>(0.0112) | 0.0174<br>(0.0112) |
| *Age at first birth* | 0.0162<br>(0.0094) | −0.0074<br>(0.0114) |
| *Black* | 0.0003<br>(0.0010) | 0.0021<br>(0.0011) |
| *White* | 0.0003<br>(0.0012) | −0.0006<br>(0.0013) |
| *Other race* | −0.0006<br>(0.0005) | −0.0014<br>(0.0009) |
| *Hispanic* | −0.0014<br>(0.0009) | −0.0007<br>(0.0010) |
| *Years of education* | −0.0028<br>(0.0076) | 0.0100<br>(0.0074) |

*Notes:* The samples are the same as in Table 2. Standard errors are reported in

# IV Estimates

- For $y_i = \alpha + \beta x_i + \varepsilon_i$
  - Two endogenous regressors: “More than 2 children” or “ number of children”
- $Z_i$ =1 or 0

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})}.$$

  is $\beta_{IV} = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0)$

  –

  - Simple illustration of how instruments identify the effect of children on labour supply

TABLE 5—WALD ESTIMATES OF LABOR-SUPPLY MODEL

| Variable | 1980 PUMS: Mean difference by *Same sex* | 1980 PUMS: Wald estimate using as covariate: *More than 2 children* | 1980 PUMS: Wald estimate using as covariate: *Number of children* | 1990 PUMS: Mean difference by *Same sex* | 1990 PUMS: Wald estimate using as covariate: *More than 2 children* | 1990 PUMS: Wald estimate using as covariate: *Number of children* |
|---|---|---|---|---|---|---|
| *More than 2 children* | 0.0600<br>(0.0016) | — | — | 0.0628<br>(0.0016) | — | — |
| *Number of children* | 0.0765<br>(0.0026) | — | — | 0.0836<br>(0.0025) | — | — |
| *Worked for pay* | −0.0080<br>(0.0016) | −0.133<br>(0.026) | −0.104<br>(0.021) | −0.0053<br>(0.0015) | −0.084<br>(0.024) | −0.063<br>(0.018) |
| *Weeks worked* | −0.3826<br>(0.0709) | −6.38<br>(1.17) | −5.00<br>(0.92) | −0.3233<br>(0.0743) | −5.15<br>(1.17) | −3.87<br>(0.88) |
| *Hours/week* | −0.3110<br>(0.0602) | −5.18<br>(1.00) | −4.07<br>(0.78) | −0.2363<br>(0.0620) | −3.76<br>(0.98) | −2.83<br>(0.73) |
| *Labor income* | −132.5<br>(34.4) | −2208.8<br>(569.2) | −1732.4<br>(446.3) | −119.4<br>(42.4) | −1901.4<br>(670.3) | −1428.0<br>(502.6) |
| ln(*Family income*) | −0.0018<br>(0.0041) | −0.029<br>(0.068) | −0.023<br>(0.054) | −0.0085<br>(0.0047) | −0.136<br>(0.074) | −0.102<br>(0.056) |

*Notes:* The samples are the same as in Table 2. Standard errors are reported in parentheses.

# 2SLS Estimates

$$y_i = \alpha_0' \mathbf{w}_i + \alpha_1 s_{1i} + \alpha_2 s_{2i} + \beta x_i + \varepsilon_i$$

- $W_i$ demographic variables
- $S_{1i}$ , $S_{2i}$ indicators of the sex of the first two children of mother I
- *First stage for "More than 2 children"*

$$x_i = \pi_0' \mathbf{w}_i + \pi_1 s_{1i} + \pi_2 s_{2i} + \gamma(Same\ sex_i) + \eta_i$$

TABLE 6—OLS ESTIMATES OF *MORE THAN 2 CHILDREN* EQUATIONS

| | All women | | | Married women | | |
|---|---|---|---|---|---|---|
| Independent variable | (1) | (2) | (3) | (4) | (5) | (6) |
| 1980 PUMS | | | | | | |
| *Boy 1st* | — | −0.0080 (0.0015) | 0.0001 (0.0021) | — | −0.0111 (0.0018) | −0.0016 (0.0026) |
| *Boy 2nd* | — | −0.0081 (0.0015) | — | — | −0.0095 (0.0018) | — |
| *Same sex* | 0.0600 (0.0016) | 0.0617 (0.0015) | — | 0.0675 (0.0019) | 0.0694 (0.0018) | — |
| *Two boys* | — | — | 0.0536 (0.0021) | — | — | 0.0598 (0.0026) |
| *Two girls* | — | — | 0.0698 (0.0021) | — | — | 0.0789 (0.0026) |
| With other covariates | no | yes | yes | no | yes | yes |
| $R^2$ | 0.004 | 0.084 | 0.084 | 0.005 | 0.078 | 0.078 |

TABLE 7—OLS AND 2SLS ESTIMATES OF LABOR-SUPPLY MODELS USING 1980 CENSUS DATA

| | All women | | | Married women | | | Husbands of married women | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Estimation method | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS |
| Instrument for *More than 2 children* | — | *Same sex* | *Two boys, Two girls* | — | *Same sex* | *Two boys, Two girls* | — | *Same sex* | *Two boys, Two girls* |
| Dependent variable: | | | | | | | | | |
| *Worked for pay* | −0.176 (0.002) | −0.120 (0.025) | −0.113 (0.025) [0.013] | −0.167 (0.002) | −0.120 (0.028) | −0.113 (0.028) [0.013] | −0.008 (0.001) | 0.004 (0.009) | 0.001 (0.008) [0.013] |
| *Weeks worked* | −8.97 (0.07) | −5.66 (1.11) | −5.37 (1.10) [0.017] | −8.05 (0.09) | −5.40 (1.20) | −5.16 (1.20) [0.071] | −0.82 (0.04) | 0.59 (0.60) | 0.45 (0.59) [0.030] |
| *Hours/week* | −6.66 (0.06) | −4.59 (0.95) | −4.37 (0.94) [0.030] | −6.02 (0.08) | −4.83 (1.02) | −4.61 (1.01) [0.049] | 0.25 (0.05) | 0.56 (0.70) | 0.50 (0.69) [0.71] |
| *Labor income* | −3768.2 (35.4) | −1960.5 (541.5) | −1870.4 (538.5) [0.126] | −3165.7 (42.0) | −1344.8 (569.2) | −1321.2 (565.9) [0.703] | −1505.5 (103.5) | −1248.1 (1397.8) | −1382.3 (1388.9) (0.549) |
| ln(*Family income*) | −0.126 (0.004) | −0.038 (0.064) | −0.045 (0.064) [0.319] | −0.132 (0.004) | −0.051 (0.056) | −0.053 (0.056) [0.743] | — | — | — |
| ln(*Non-wife income*) | — | — | — | −0.053 (0.005) | 0.023 (0.066) | 0.016 (0.066) [0.297] | — | — | — |

*Notes:* The table reports estimates of the coefficient on the *More than 2 children* variable in equations (4) and (6) in the text. Other covariates in the models are *Age, Age at first birth,* plus indicators for *Boy 1st, Boy 2nd, Black, Hispanic,* and *Other race.* The variable *Boy 2nd* is excluded from equation (6). The *p*-value for the test of overidentifying restrictions associated with equation (6) is shown in brackets. Standard errors are reported in parentheses.

TABLE 8—OLS AND 2SLS ESTIMATES OF LABOR-SUPPLY MODELS USING 1990 CENSUS DATA

| | All women | | | Married women | | | Husbands of married women | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Estimation method | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS |
| Instrument for *More than 2 children* | — | *Same sex* | *Two boys, Two girls* | — | *Same sex* | *Two boys, Two girls* | — | *Same sex* | *Two boys, Two girls* |
| Dependent variable: | | | | | | | | | |
| *Worked for pay* | −0.155 | −0.092 | −0.092 | −0.147 | −0.104 | −0.104 | −0.102 | 0.017 | 0.017 |
| | (0.002) | (0.024) | (0.024) | (0.002) | (0.024) | (0.024) | (0.001) | (0.009) | (0.009) |
| | | | [0.743] | | | [0.576] | | | [0.989] |
| *Weeks worked* | −8.71 | −5.66 | −5.64 | −8.25 | −5.76 | −5.76 | −1.03 | 1.01 | 1.01 |
| | (0.08) | (1.16) | (1.16) | (0.09) | (1.15) | (1.15) | (0.05) | (0.63) | (0.63) |
| | | | [0.391] | | | [0.670] | | | [0.708] |
| *Hours/week* | −6.80 | −4.08 | −4.10 | −6.39 | −3.94 | −3.95 | −0.06 | 0.85 | 0.83 |
| | (0.07) | (0.98) | (0.98) | (0.07) | (0.96) | (0.96) | (0.05) | (0.69) | (0.69) |
| | | | [0.489] | | | [0.665] | | | [0.180] |
| *Labor income* | −3984.4 | −2099.6 | −2096.2 | −3753.9 | −2457.5 | −2456.3 | 929.7 | 1348.7 | 1354.8 |
| | (44.2) | (664.0) | (663.8) | (50.7) | (669.7) | (669.7) | (114.9) | (1536.0) | (1535.9) |
| | | | [0.830] | | | [0.893] | | | [0.711] |
| ln(*Family income*) | −0.119 | −0.124 | −0.122 | −0.103 | −0.054 | −0.054 | — | — | — |
| | (0.005) | (0.071) | (0.071) | (0.004) | (0.051) | (0.051) | | | |
| | | | [0.270] | | | [0.878] | | | |
| ln(*Non-wife income*) | — | — | — | −0.004 | 0.020 | 0.020 | — | — | — |
| | | | | (0.005) | (0.068) | (0.068) | | | |
| | | | | | | [0.452] | | | |

*Notes:* The table reports the coefficient on the ***More than 2 children*** variable in equations (4) and (6) in the text estimated with 1990 Census data. Other covariates in the models are *Age, Age at first birth,* plus indicators for *Boy 1st, Boy 2nd, Black, Hispanic,* and *Other race*. The variable *Boy 2nd* is excluded from equation (6). The *p*-value for the test of overidentifying restrictions associated with equation (6) is shown in brackets. Standard errors are reported in parentheses.

# Treatment Effect Heterogeneity

- Effect may vary
  - with own and partners wages
  - with own and partners education
- OLS and 2SLS estimates conditional on

earning or earning potential

TABLE 9—2SLS ESTIMATES OF LABOR-SUPPLY MODELS WITH INTERACTION TERMS USING 1980 CENSUS DATA

| | *More than 2 children* | *Worked for pay* | | | *Weeks/year* | | |
|---|---|---|---|---|---|---|---|
| Sample/variables | First stage | Mean of dependent variable | OLS | 2SLS | Mean of dependent variable | OLS | 2SLS |
| A. Results for wives by *husband's earnings:* | | | | | | | |
| *Bottom third of husband's earnings distribution* | 0.057 (0.003) | 0.570 | −0.186 (0.003) | −0.122 (0.060) | 21.1 | −9.23 (0.15) | −7.55 (2.60) |
| *Middle third of husband's earnings distribution* | 0.072 (0.003) | 0.569 | −0.165 (0.004) | −0.185 (0.047) | 20.8 | −8.31 (0.15) | −7.11 (2.04) |
| *Top third of husband's earnings distribution* | 0.079 (0.003) | 0.448 | −0.152 (0.003) | −0.078 (0.042) | 15.2 | −6.76 (0.15) | −3.17 (1.82) |
| B. Results for wives by *wife's education:* | | | | | | | |
| *Wife < high-school graduate* | 0.071 (0.004) | 0.468 | −0.150 (0.005) | −0.121 (0.064) | 16.1 | −7.30 (0.20) | −7.12 (2.80) |
| *Wife high-school graduate* | 0.073 (0.003) | 0.524 | −0.156 (0.003) | −0.147 (0.038) | 19.2 | −7.74 (0.13) | −6.42 (1.65) |
| *Wife > high-school graduate* | 0.063 (0.003) | 0.567 | −0.179 (0.004) | −0.082 (0.054) | 20.4 | −8.33 (0.15) | −2.93 (2.33) |
| C. Results for wives by *wife's education* for women whose *husband's earnings* are in middle third: | | | | | | | |
| *Wife < high-school graduate* | 0.079 (0.008) | 0.481 | −0.138 (0.009) | −0.275 (0.109) | 16.7 | −7.10 (0.38) | −10.2 (4.83) |
| *Wife high-school graduate* | 0.076 (0.004) | 0.551 | −0.157 (0.003) | −0.189 (0.060) | 20.3 | −8.33 (0.21) | −7.78 (2.64) |
| *Wife > high-school graduate* | 0.062 (0.006) | 0.640 | −0.184 (0.006) | −0.125 (0.098) | 23.7 | −9.07 (0.28) | −3.98 (4.30) |
| D. Results for husbands by *wife's education:* | | | | | | | |
| *Wife < high-school graduate* | 0.071 (0.004) | 0.945 | −0.014 (0.001) | −0.013 (0.020) | 44.5 | −1.36 (0.10) | −0.21 (1.37) |
| *Wife high-school graduate* | 0.074 (0.003) | 0.981 | −0.005 (0.001) | 0.005 (0.012) | 48.4 | −0.53 (0.06) | 0.92 (0.81) |
| *Wife > high-school graduate* | 0.063 (0.003) | 0.987 | −0.002 (0.001) | 0.009 (0.016) | 49.2 | −0.23 (0.08) | 0.25 (1.14) |

*Notes:* The table reports estimates of the coefficient on *More than 2 children* in equation (4) in the text, modified to allow interactions with wives' schooling and husbands' education as indicated. Main effects for each interaction variable (*husband's earnings distribution* and *wife's education*) are included in the equation. Other covariates in the models are those listed in