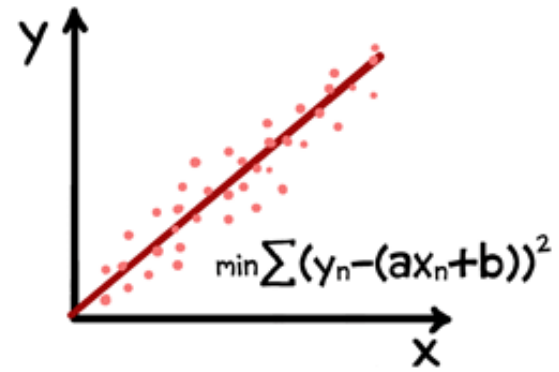
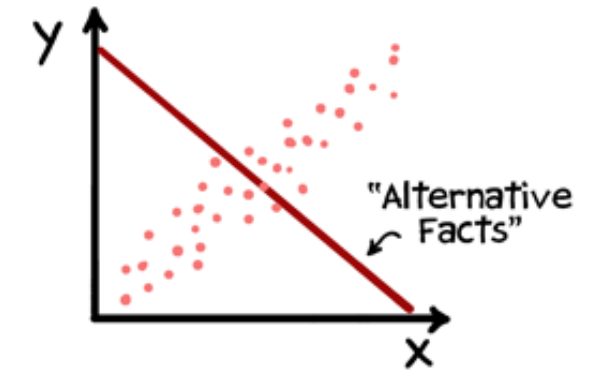


Linear Regression



JORGE CHAM © 2016

Societal Regression



WWW.PHDCOMICS.COM

INFO 251: Applied Machine Learning

Regression and Impact Evaluation










Announcements

- PS2 posted
- Enrollment updates
 - 95 enrolled, 8 waitlist, 10 CE
 - Roster will likely be updated tomorrow
 - Undergrads should email course staff asap if still interested
- Note: it might help to have a pen and paper handy for today's lecture, in order to do some basic arithmetic








Which of the following topics are you most interested in? Choose exactly 2:








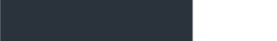




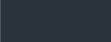


About the class

What field/discipline is your degree in?

MIMS 2025	5 respondents	5 %	 ✓
MIMS 2026	10 respondents	11 %	
Public Policy (incl. MDP, MDE)	8 respondents	9 %	
Social Science	3 respondents	3 %	
Natural Science / Engineering	27 respondents	29 %	
Haas	3 respondents	3 %	
Statistics	22 respondents	24 %	
Humanities	1 respondent	1 %	
Other	19 respondents	20 %	

What was your undergraduate major?

Math / Statistics	29 respondents	31 %	 ✓
Computer Science	24 respondents	26 %	
Engineering	30 respondents	32 %	
Social Science	11 respondents	12 %	
Humanities	5 respondents	5 %	
Natural Sciences	4 respondents	4 %	
Other	14 respondents	15 %	

Experimental methods for causal inference (impact evaluation, A-B testing)	22 respondents	24 %	 ✓
Non-experimental methods for causal inference (instrumental variables, regression discontinuity)	7 respondents	8 %	 ✓
Gradient descent	9 respondents	10 %	
Regularization and linear models	6 respondents	6 %	
Naive Bayes	7 respondents	8 %	
Decision trees and random forests	14 respondents	15 %	
Neural networks	25 respondents	27 %	
"Deep" learning	37 respondents	40 %	
LLMs	23 respondents	25 %	
Practical issues with machine learning (features, missing/imbalanced data, multi-label classification)	32 respondents	34 %	
Cluster analysis	12 respondents	13 %	
Dimensionality reduction	10 respondents	11 %	
Recommender systems	14 respondents	15 %	
Machine learning for causal inference	31 respondents	33 %	
Fairness and bias in ML	12 respondents	13 %	

Course Outline

- Causal Inference and Research Design
 - **Experimental methods**
 - Non-experiment methods
- Machine Learning
 - Design of Machine Learning Experiments
 - Linear Models and Gradient Descent
 - Non-linear models
 - Fairness and Bias in ML
 - Neural models
 - Deep Learning
 - Practicalities
 - Unsupervised Learning
- Special topics

Key Concepts (last lecture)

- Random selection and assignment
- Internal and external validity
- Counterfactuals
- Identifying assumptions
- Control groups
- Power Calculations
- Single difference design
- Pre vs. Post research design
- Difference-in-Difference (Double Difference) design
- Differential Trends
- Encouragement designs

Outline

- **Lecture 1 wrap-up: Multiple hypothesis testing**
- A complete impact evaluation example: Progresa
- Regression recap
- Regression and Impact Evaluation

Randomization: Pitfalls

- In a perfectly randomized experiment:
 - *If treatment is randomized, a simple difference between outcomes in treated and control units gives an unbiased estimate of impact*
- What could go wrong? Common threats to internal validity
 - Spillovers, externalities, and interference
 - An indirect or unintended effect which often does not comply with treatment assignment
 - Can be positive and negative
 - Examples?
 - Non-compliance
 - Attrition (esp. differential attrition)

What about multiple testing?



- Joseph Rhine was a famous parapsychologist in the 1950's
 - Founder of Journal of Parapsychology, affiliate of AAAS
- Experiment: subjects had to guess whether 10 hidden cards were red or blue
 - He found that about 1 person in 1000 had ESP, i.e. they could guess the color of all 10 cards
 - He called back the “psychic” subjects and had them do the same test again. They all failed.
 - He concluded that the act of telling psychics that they have psychic abilities causes them to lose it...
- Publication bias and “results bias” exacerbate the problem
 - If we run 20 experiments, we expect to observe significant effects ($p < 0.10$) in 2 experiments
- If anything, “big data” exacerbate the problem
 - Data mining technologies and large datasets make it incredibly easy to ask questions and test hypotheses – you can effectively run hundreds of experiments in a few hours

What can you do?

- In general, we care whether a result is “unusual” relative to chance
 - Confidence intervals and p-values can help quantify this
 - (Tomorrow’s lab will cover hypothesis testing)
- Report everything you tried, not just successes
- Present & interpret effect size
 - “Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude – not just, does a treatment affect people, but how much does it affect them.”
 - Gene V. Glass

Functional form	Effect size interpretation (where β is the coefficient)
Linear f	
$y = f(x)$	A unit change in x is associated with an average change of β units in y .
$\ln(y) = f(x)$	For a unit increase in x , y increases on average by the percentage $100(e^\beta - 1)$ ($\cong 100 \beta$ when $ \beta < 0.1$).
$y = f(\ln(x))$	For a 1% increase in x , y increases on average by $\ln(1.01) \times \beta (\cong \beta/100)$.
$\ln(y) = f(\ln(x))$	For a 1% increase in x , y increases on average by the percentage $100(e^{\beta \cdot \ln(1.01)} - 1)$ ($\cong \beta$ when $ \beta < 0.1$).
Logistic f	
Numerical x	A unit change in x is associated with an average change in the odds of $Y = 1$ by a factor of β .
Binary x	The odds of $Y = 1$ at $x = 1$ are higher than at $x = 0$ by a factor of β .

Interpreting regression coefficients

What else can you do?

- Bonferroni Adjustments
 - With k tests, reduce the significance threshold for each test to $0.05 / k$
- Dunn-Sidak
 - With k tests, reduce the significance threshold to $1 - (.95)^{1/k}$
- Other options:
 - Family error rates
 - Pre-analysis plans
 - Randomization inference
 - See Duflo et al (2006) reading

Key Concepts (today's lecture)

- Progres
- Interpreting regression coefficients
- Dummy variables, “one-hot” vectors
- Heterogeneous treatment effects
- Regression and impact evaluation
 - Estimating treat vs. control
 - Interaction variables
 - Estimating difference-in-difference
- Cross-sectional vs. panel data
- Between vs. within variation
- Difference regressions
- Normalization
- Fixed effects

Outline

- Lecture 1 wrap-up
- **A complete impact evaluation example: Progresa**
- Regression recap
- Regression and Impact Evaluation

Progresa

- Goals of Progresa?
 - Increase education of the poor
 - Improve living conditions of the poor
- How to do it?
 - “Conditional cash transfers”
 - Provide subsidies to poor households if they send their children to school
- Sidebar: Why might this work?
 - Credit and liquidity constraints are a major obstacle to education
 - Reducing the cost of schooling is fundamental to many policies designed to improve educational outcomes

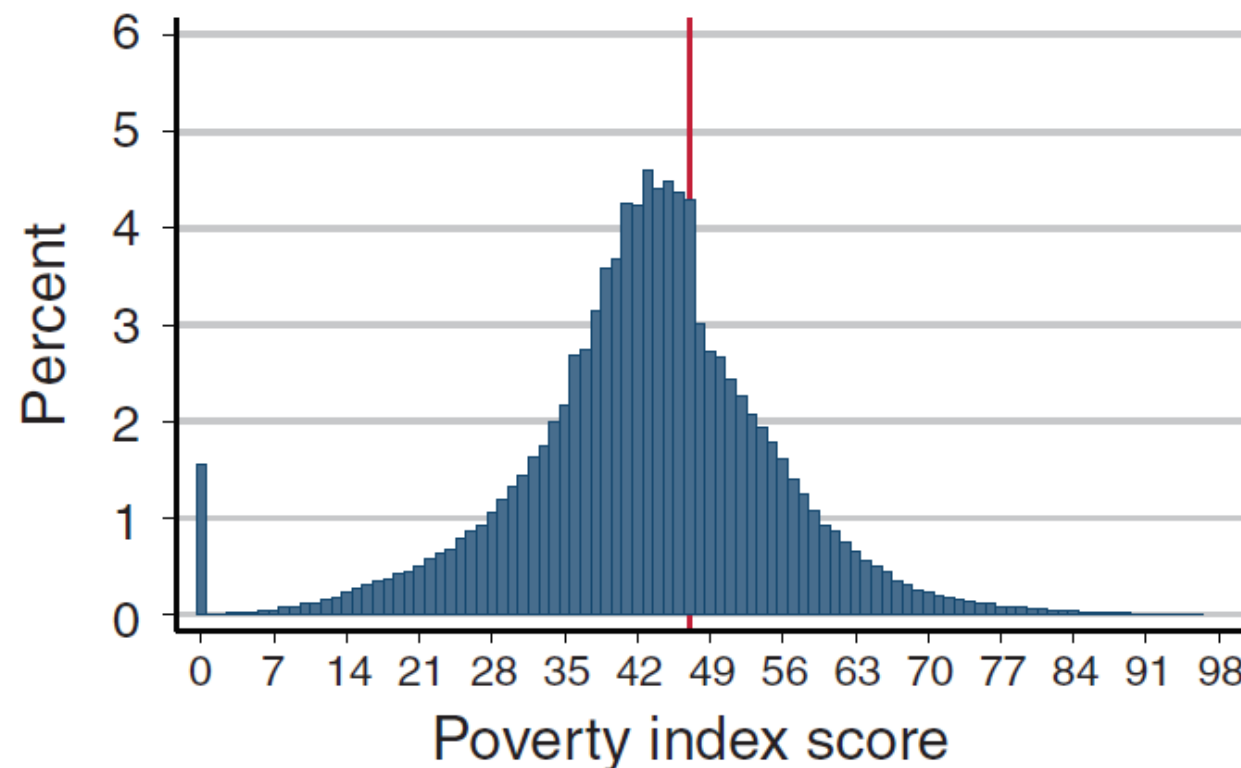
Progresas: Details

- Distinguishing features of Progresas
 - Targeted to poor households with primary school children
 - Money given to women in household
 - Randomized evaluation built in
 - 506 villages randomly assigned to treatment and control
- Scale and scope
 - Started in 1997, continues today (first as Oportunidades, now as Prospera)
 - Now covers 3 million people, 0.3% of Mexico GDP
 - More broadly, CCT's have now been used in dozens of countries
 - Fiszbein, A.S., Norbert R., 2009. Conditional Cash Transfers: Reducing Present and Future Poverty.
 - Baird et al. 2013. Relative Effectiveness of Conditional and Unconditional Cash Transfers for Schooling Outcomes in Developing Countries: A Systematic Review.

Progresa: Eligibility

- Eligibility for Progresa was established at the household level, following a two-step targeting procedure
 1. Must live in a “poor rural village”
 - These villages were selected on the basis of a marginality index, established in 1995 using information from the Census
 - For the impact evaluation, a subset of these communities were selected to be eligible for Progresa *before* other communities (these are the “treatment villages”)
 2. To receive benefits, household must be classified as “poor” (within a poor village)
 - In treatment villages, a household census was conducted prior to program launch
 - The census collected a “wealth index”, which was used to identify poor households
 - The details of this welfare index were kept secret. Why?

Progresa: Eligibility

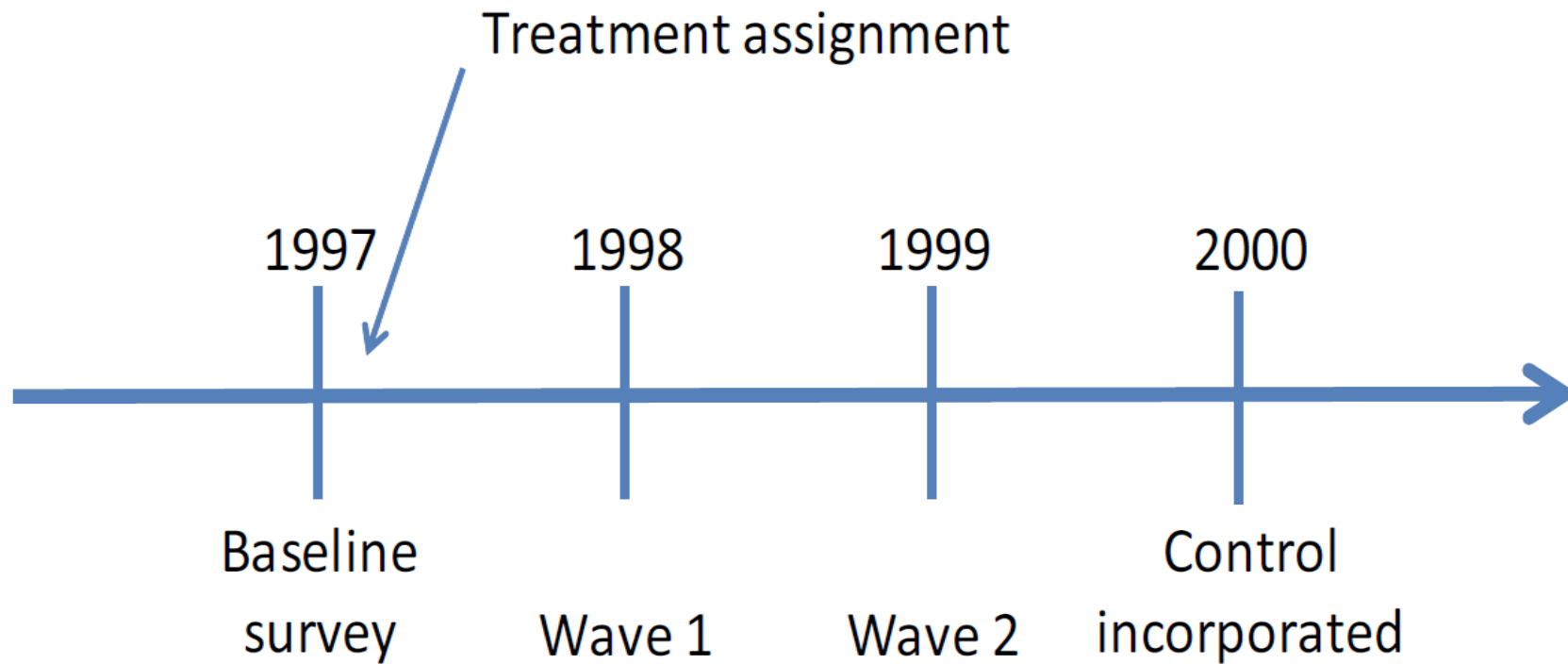


- Camacho, A., Conover, E., 2011. Manipulation of Social Program Eligibility. American Economic Journal: Economic Policy 3, 41–65.

Progresa: Implementation

- Nationwide, 50,000 communities were selected to receive Progresa
 - 78% of the households in those communities were deemed eligible
- For the impact evaluation, which occurred prior to the national roll-out, a subset of 506 communities in 7 states were selected to participate (i.e., ~1% of all communities)
 - Program was rolled out to 320 (“treatment”) communities in May 1998
 - Program was rolled out to remaining 186 (“control”) communities in late 1999
 - This is a classic example of a “staggered roll-out design”

Progresa: Implementation



Progresa: Implementation

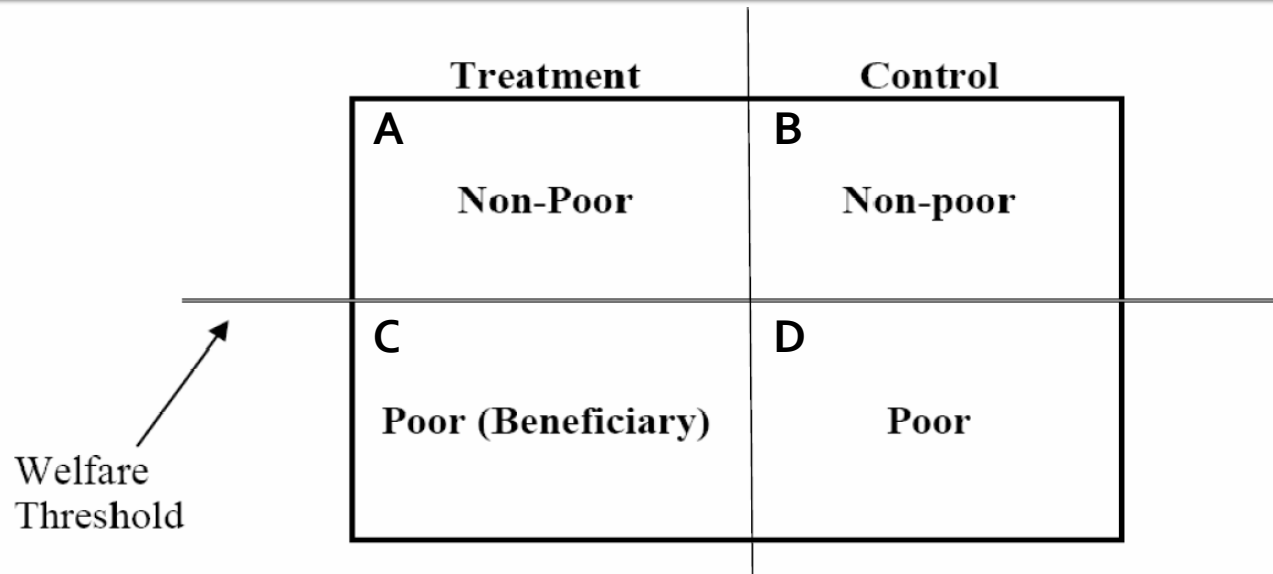


Figure 1: Program Evaluation Design

- How to measure effect of Progresa on enrollment, assuming we only have access to (post-intervention) data from 1998?
 - Compare poor in Treatment to poor in Control (i.e., C - D)

Progresa: Implementation

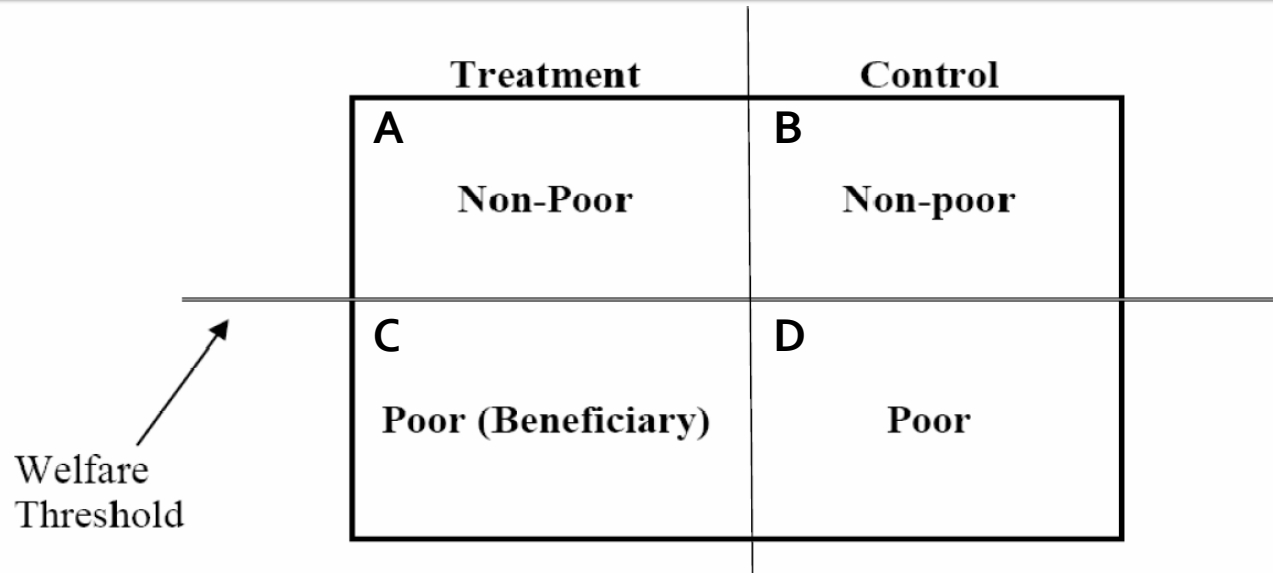


Figure 1: Program Evaluation Design

- The counterfactual (for enrollment of poor in treated villages) is...
 - Enrollment of poor in control villages (in absence of Progresa)
- The key identifying assumption is...
 - In the absence of treatment, enrollment in C would have been same as enrollment in D
- Does this assumption seem reasonable in this context?
- What evidence might we provide to support using this identifying assumption?

Progresa: Externalities

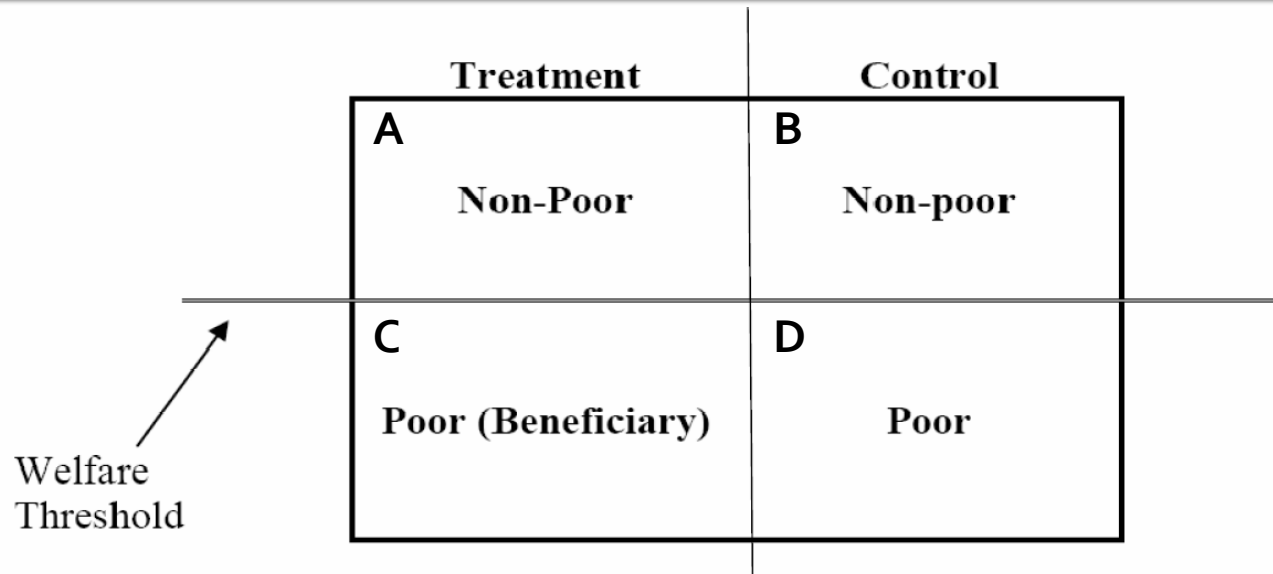


Figure 1: Program Evaluation Design

- The “basic” treat-vs-control design measures the treatment effect by looking at outcomes among the poor, and ignores the non-poor
- But what if we observe differences in the outcomes of the non-poor?
 1. Perhaps treatment and control villages were different to begin with (i.e., randomization failed)
 2. Perhaps there were spillover effects from poor to non-poor (within treated villages)

Progresa: Diff-in-Diff

- What if we have two rounds of survey data: baseline data (before intervention) and endline data (after intervention)?
 - Starting point: we can assess whether treatment and control villages differed before Progresa
- Difference-in-difference design for evaluating Progresa
 1. Focusing on poor, how do changes between 97 and 98 compare between T and C villages?
 2. Focusing on endline, how do differences between Poor and Non-Poor compare between T and C?
 - What are the identifying assumptions of these two designs?

1997 ("Baseline")		1998 ("Endline")	
Treatment	Control	Treatment	Control
E Non-Poor	F Non-poor	A Non-Poor	B Non-poor
G Poor (Beneficiary)	H Poor	C Poor (Beneficiary)	D Poor

Outline

- A complete impact evaluation example: Progresa
- **Regression recap**
- Regression and Impact Evaluation

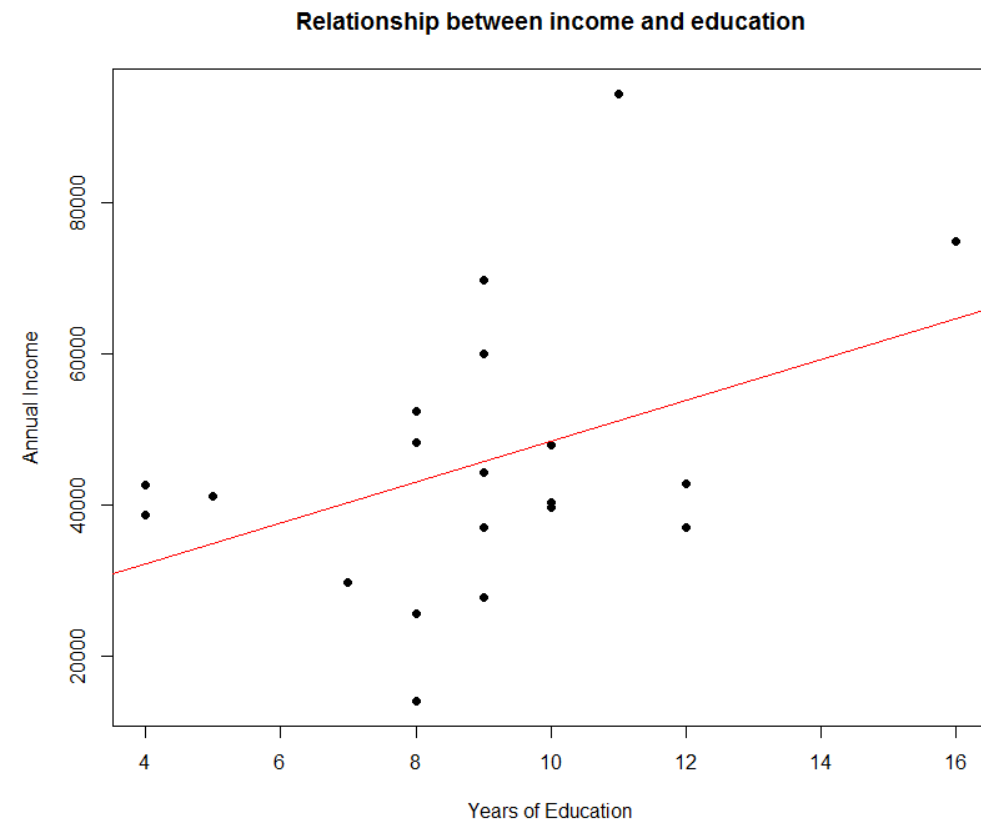
Regression and Impact Evaluation

- Thus far we've used cross-tabs to "eyeball" the impact of a treatment
- Advantages
 - Very simple to compute
 - Easily interpretable
- Disadvantages
 - How to measure statistical precision?
 - How to deal with known confounds (e.g., differential trends)?
 - How to estimate treatment effect heterogeneity?
 - May be unbiased, but may not be precise - controlling for additional factors may increase precision

Regression: Quick Recap

- Linear regression offers a concise summary of the mean of one variable as a function of the other variable through two parameters: the slope and the intercept of the regression line
 - Causality often *implied*, rarely *justified*

	Education	Age	Income
[1,]	8	35	30942.35
[2,]	8	23	37323.89
[3,]	8	58	49381.84
[4,]	5	41	31680.86
[5,]	13	35	81147.84
[6,]	9	43	38682.86
[7,]	8	35	34632.30
[8,]	7	56	14394.98
[9,]	11	62	22243.85
[10,]	14	24	51831.79
[11,]	12	25	23963.90
[12,]	12	32	66780.27
[13,]	4	41	26979.73
[14,]	8	49	38837.48
[15,]	10	21	40726.37
[16,]	8	33	40269.51
[17,]	4	36	34293.32
[18,]	10	38	61158.98
[19,]	11	36	64329.59
[20,]	9	48	51069.77



Regression: Quick Recap

- Simple bivariate (linear) regression

- The regression model

$$wages_i = \alpha + \beta * education_i + error_i$$

- The fitted model

$$wages_i = 12409 + 3310 * education_i + error_i$$

- Intuition check

- What does β tell us?
- What is 12409?
- What are the expected wages be for someone with 14 years of education?
 - $12409 + 14 * 3310 = 58,749$

Regression: Quick Recap

- Regression with binary predictor/independent variables

- The regression model

$$wages_i = \alpha + \beta * isForeign_i + error_i$$

- The fitted model

$$wages_i = 54212 - 2710 * isForeign_i + error_i$$

- Multiple (linear) regression

$$wages_i = \alpha + \beta * education_i + \gamma * isForeign_i + error_i$$

Regression: Categorical variables

- What if our control variables are categorical?
 - Example: We want to study the relationship between wages and education, controlling for country
 - $Wages_i = \alpha + \beta Education_i + \gamma Country_i + \epsilon_i$

- How to deal with a categorical predictor?
 - Convert to a single binary variable:
 - $Wages_i = \alpha + \beta Education_i + \gamma USA_i + \epsilon_i$
 - $USA_i = 1$ iff worker i is from USA, $USA_i = 0$ otherwise
 - Makes sense if we care about the effect of one category relative to others
 - Convert to a set of binary variables:
 - $Wages_i = \beta Education_i + \gamma_1 USA_i + \gamma_2 CHINA_i + \dots + \gamma_M Country_m + \epsilon_i$
 - $Wages_i = \beta Education_i + \sum_{c=1}^M \gamma_c \mathbf{1}(Country_i = c) + \epsilon_i$
 - $Wages_i = \beta Education_i + Country_i + \epsilon_i \leftarrow$ this is an abuse of notation, but it is very common

Regression: “Dummy” variables

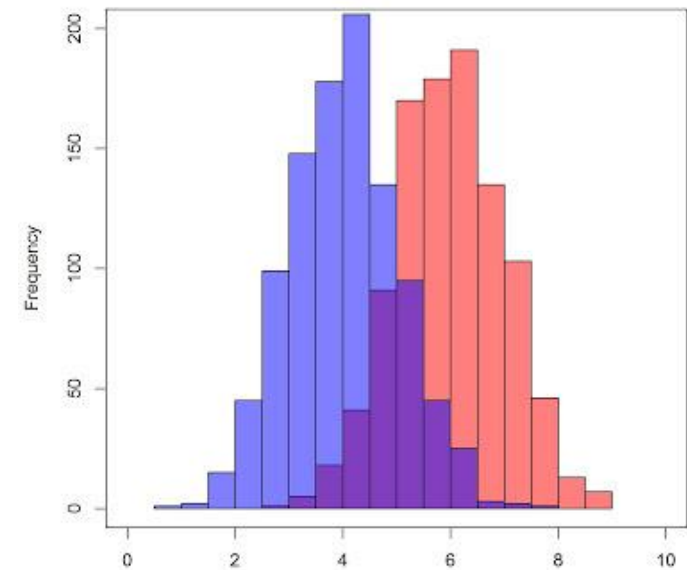
- Converting levels to a series of binary variables
 - $Y_i = \alpha + \beta Education_i + Country_i + \epsilon_i$
 - $Country_i$ is a “dummy variable” or “one-hot vector” or “fixed effect”
- Interpretation
 - Equivalent to creating a country-specific intercept
 - Each of the $Country$ coefficients indicates average wage for workers from that particular country (when $Education_i = 0$), *relative to the reference country*
 - With M countries, we have (M-1) coefficients and an intercept α
 - Alternatively, estimate with no intercept and M coefficients
 - $Y_i = \beta Education_i + Country_i + \epsilon_i$
 - Intuition check: Will the coefficients for “country” dummies be the same in both cases?

Outline

- A complete impact evaluation example: Progresa
- Regression recap
- **Regression and Impact Evaluation**

Regression and Impact: Basics

- How to measure the effect of treatment T on outcome Y in a regression?
 - The regression equation:
$$Y_i = \alpha + \beta T_i + \epsilon_i$$
 - Example: We estimate the effect of eating a cookie on happiness on a scale of 1-10. We estimate $\hat{\alpha} = 4.1, \hat{\beta} = 1.3$. What does this mean?
- If T is randomly assigned, $\hat{\beta}$ is an estimate of the **causal impact** of T on Y



Regression: “Control” variables

- How to simultaneously measure the effect of a treatment T and a non-experimental control variable X on an outcome Y in a regression setting?

$$Y_i = \alpha_1 + \beta_1 T_i + \gamma X_i + e_i$$

- How is this different from a version without control variables?

$$Y_i = \alpha_2 + \beta_2 T_i + \epsilon_i$$

- In a perfectly randomized experiment...
 - What, if anything, can we say about $Cor(T_i, X_i)$?
 - What, if anything, can we say about our estimates of β_1 and β_2 ?
 - What, if anything, can we say about γ ?

Control variables: Example

- Example: We are estimating the effect of eating a cookie (T_i) on happiness (Y_i) on a scale of 1-10, while controlling for years in grad school (X_i)

- Regression equation?

- $Y_i = \alpha + \beta T_i + \gamma X_i + \epsilon_i$

- Coefficient estimates

- $\hat{\alpha} = 3.4, \hat{\beta} = -0.5, \hat{\gamma} = 1.2$

- What do these results mean?

