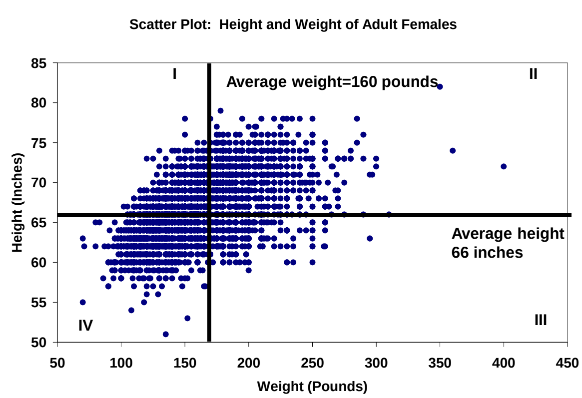
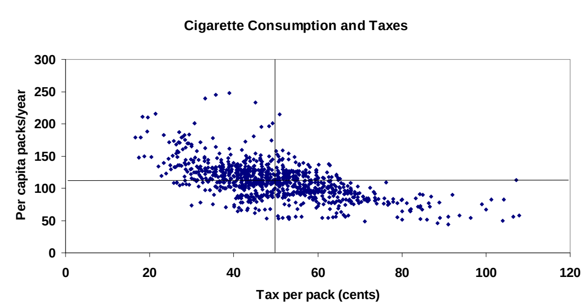
Moving from correlation to causation

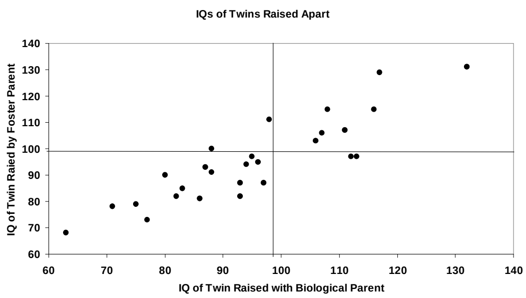
ECON 30331
Bill Evans

Scatter plot

- Sample of N observations
  - Students, doctors, state, countries etc.
- For each observation, 2 pieces of data (X,Y)
- Plot each point for all observations in sample







5

Covariance

- Measure of co-movement between variables
- Does the realization that X is above average convey any information about the likely value of Y?
- Identifies whether variables are 'statistically' related

6

Covariance

- x and y are random variables
- E[x]= μ_x Var(x) = σ^2_x
- E[y]= μ_y Var(y) = σ^2_y
- Cov(x,y) = E[(x - μ_x)(y - μ_y)] = σ_{xy}
  = E[xy] - $\mu_x\mu_y$ = σ_{xy}

7

$$\textit{If } \operatorname{cov}(x,y) > 0 \textit{ and } y > \overline{y},$$
$$\textit{then, on average, } x > \overline{x}$$

$$\textit{If } \operatorname{cov}(x,y) < 0 \textit{ and } y > \overline{y},$$
$$\textit{then, on average, } x < \overline{x}$$

8

Problem

- Covariance is scale dependent
  - Covariance between height and weight will differ if measured in centimeters & kilograms or inches & pounds
- Not an attractive property for a measure of co-movement

9

Demonstrate: Can show yourself

$$\text{cov}(x,y) = E[(x-\mu_x)(y-\mu_y)] = \sigma_{xy}$$
$$define: z = a + bx$$
$$\text{cov}(z,y) = E[(z-\mu_z)(y-\mu_y)]$$
$$z = a + bx$$
$$\mu_z = a + b\mu_x$$
$$(z-\mu_z) = b(x-\mu_x)$$
$$\text{cov}(z,y) = E[b(x-\mu_x)(y-\mu_y)]$$
$$= bE[(x-\mu_x)(y-\mu_y)] = b\sigma_{xy}$$

10

Correlation coefficient

$$\rho(x,y) = \sigma_{xy} / (\sigma_x \sigma_y)$$

$$-1 \le \rho(x,y) \le 1$$

- Unlike the covariance, the correlation coefficient is NOT scale dependent
- The value is the same regardless of how x and y are measured

11

Sample estimates

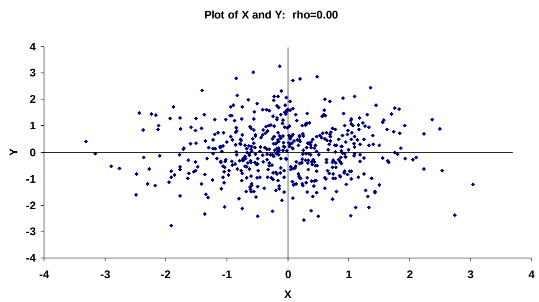
$$\hat{\sigma}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\hat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

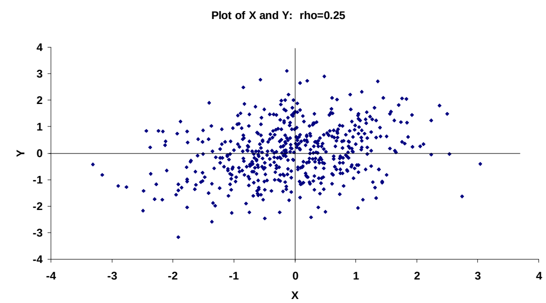
$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\hat{\rho} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$$

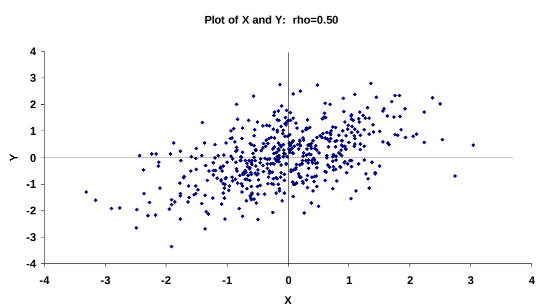
12


Plot of X and Y: rho=0.00
Y
X

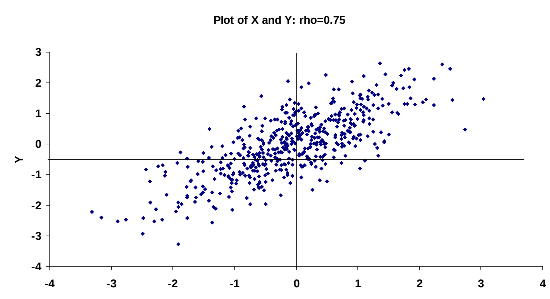
13


Plot of X and Y: rho=0.25
Y
X

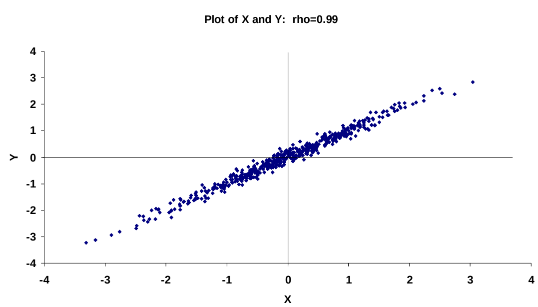
14


Plot of X and Y: rho=0.50
Y
X

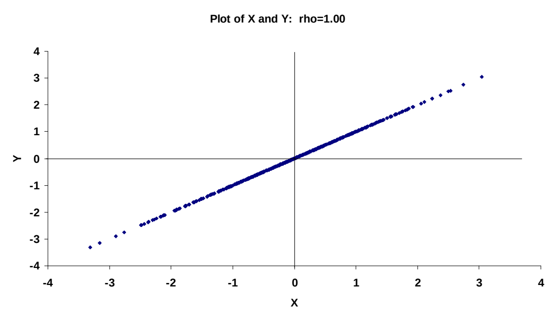
15


Plot of X and Y: rho=0.75
Y
X

16



17



18

Cross-Sectional data

- Height and weight, men
- Height/weight, women
- Log(wages)/educ (m)
- Log(wage)/age (m)

19

Cross-Sectional Data

- Husband/wife age
- Husband/wife educ
- Husband/wife height
- Father/son income
- Father/son educ.

20

Cross-Sectional Data

- IQ's of Identical twins
- IQ's of fraternal twins
- IQ's of identical twins raised apart
- IQ's of siblings
- IQ's of unrelated children reared together

21

Among undergrads in Intro Micro

- Math SAT/verbal SAT
- HS rank/total SAT
- GPA in micro/SAT
- GPA in micro/HS percentile

22

Limitation

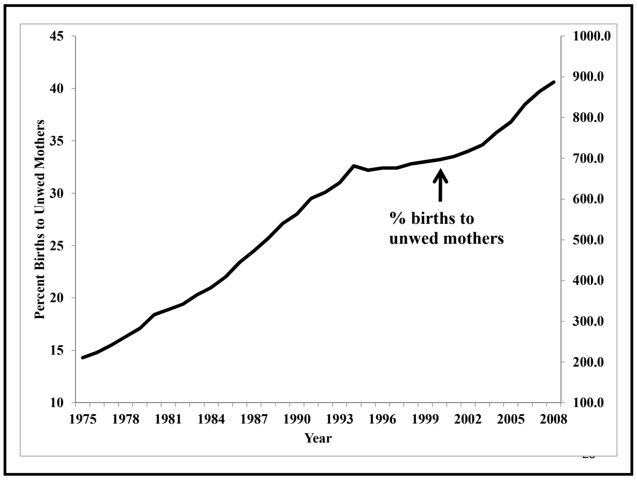
- Correlation coefficient is a convenient way to measure a statistical relationship between two variables
- It does not however signify anything more than statistical observation
- It also does no get us any closer to saying whether something is causally related
- Correlation does not equal causation

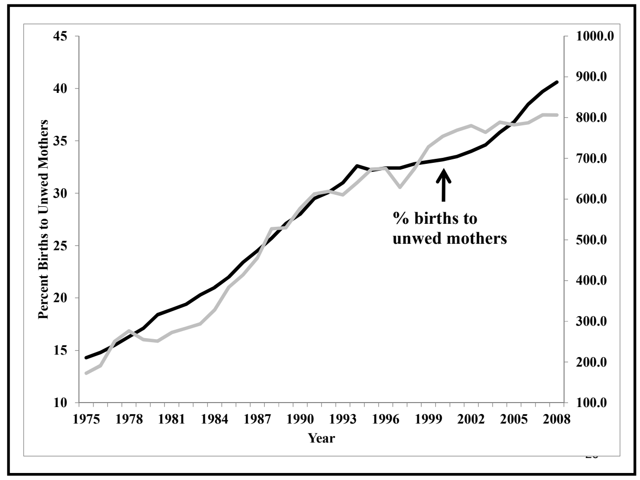
23

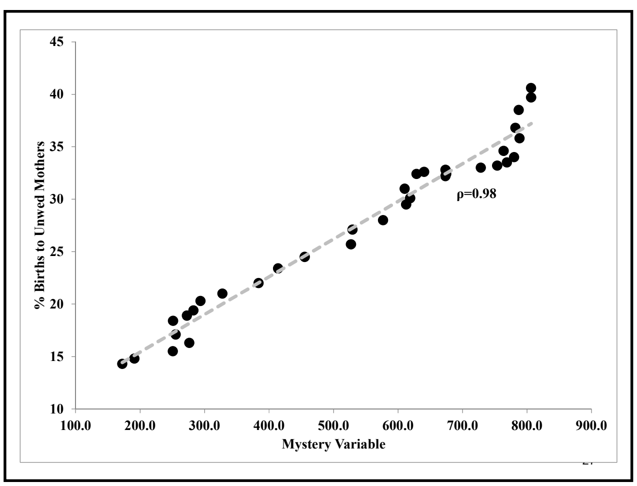
Births to unwed mothers

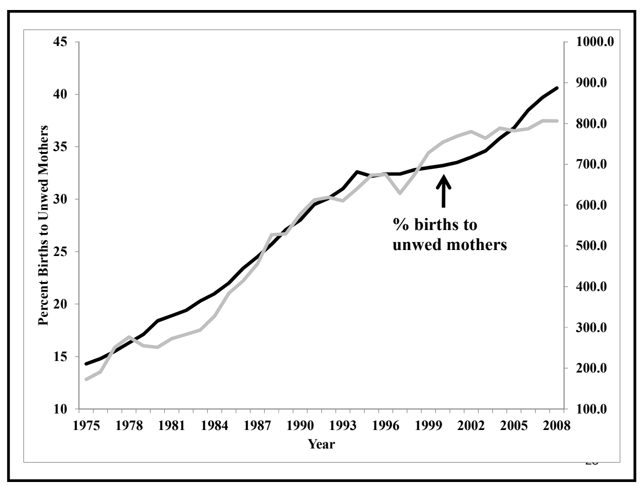
- Risen from 5% in 1960 to 37% in 2006
- Predictive of many child outcomes
  - Low birth weight, increased mortality, poor performance in schools, etc.
- Many potential explanations
  - Poor performance of male wages, rising divorce, availability of abortion
- Is there a magic bullet explanation?

24

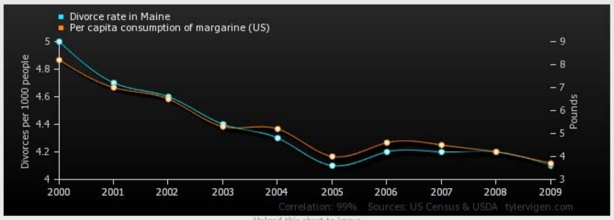

45
40
35
30
25
20
15
10
1000.0
900.0
800.0
700.0
600.0
500.0
400.0
300.0
200.0
100.0
Percent Births to Unwed Mothers
1975 1978 1981 1984 1987 1990 1993 1996 1999 2002 2005 2008
Year
% births to unwed mothers


45
40
35
30
25
20
15
10
1000.0
900.0
800.0
700.0
600.0
500.0
400.0
300.0
200.0
100.0
Percent Births to Unwed Mothers
1975 1978 1981 1984 1987 1990 1993 1996 1999 2002 2005 2008
Year
% births to unwed mothers


45
40
35
30
25
20
15
10
% Births to Unwed Mothers
100.0 200.0 300.0 400.0 500.0 600.0 700.0 800.0 900.0
Mystery Variable
ρ=0.98


45
40
35
30
25
20
15
10
1000.0
900.0
800.0
700.0
600.0
500.0
400.0
300.0
200.0
100.0
Percent Births to Unwed Mothers
1975 1978 1981 1984 1987 1990 1993 1996 1999 2002 2005 2008
Year
% births to unwed mothers

Divorce rate in Maine
correlates with
Per capita consumption of margarine (US)

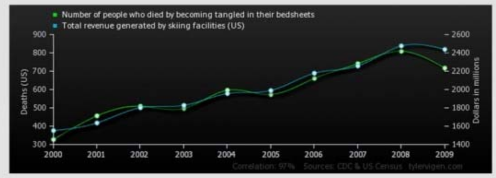


Upload this chart to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine Divorces per 1000 people (US Census)	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) Pounds (USDA)	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Number of people who died by becoming tangled in their bedsheets
correlates with
Total revenue generated by skiing facilities (US)

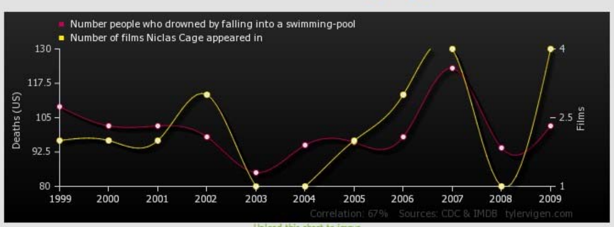


	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number of people who died by becoming tangled in their bedsheets Deaths (US) (CDC)	327	456	509	497	596	573	661	741	809	717
Total revenue generated by skiing facilities (US) Dollars in millions (US Census)	1,551	1,635	1,801	1,827	1,956	1,989	2,178	2,257	2,476	2,438

Correlation: 0.969724

Permalink - Not interesting

Number people who drowned by falling into a swimming-pool
correlates with
Number of films Nicolas Cage appeared in



Upload this chart to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool Deaths (US) (CDC)	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in Films (IMDB)	2	2	2	3	1	1	2	3	4	1	4

Economics as a science

- Utilize (more so than most social sciences) the scientific method
- Build models – test them with data – refine the models based on results
- Unless theory (models) can be tested, not much of a theory
- Economics has produced extensive statistical tools to test models

Basic economic model

- People/firms/organizations are purposeful
- Examples
  - Firms maximize profits
  - People maximize happiness/utility
- There are however limits or constraints on behavior
  - Consumers must pay prevailing prices
  - Firms have competitors

33

Break variables into 2 groups

- Exogenous (external conditions)
  - Constraints on behavior
  - "Treatments" Factors that can be altered
  - "Independent" variables
- Endogenous outcomes
  - Choice variables
  - Outcomes of systems
  - "Dependent" variables

34

Link between models/data

- Basic economic model has a prediction:
  - How quickly will demand fall when prices rise
  - What happens to outcomes (endogenous) when an external condition is changed (exogenous)
- Statistical goal: estimate the slope of the demand curve $\partial X / \partial P_x$

35

Theory of Demand

- Core model of intermediate micro
- Model set up
  - Consumers derive utility from consumption of 2 goods (x,y)
    - $U = U(x,y)$
    - Utility function has specific properties
  - Pick utility maximizing bundle of (x,y) subject to constraints
    - Fixed prices for goods: P_x and P_y
    - Fixed income, I

36

- Two implicit functions:
  $X = f(P_x,P_y,I)$
  $Y = g(P_x,P_y,I)$
- 3 "exogenous" variables: P_x, P_y and I
- 2 "endogenous" variables: x and y
- Comparative statics: $\partial X/\partial P_x$ or $\partial X/\partial I$

37

- To build a statistical model that will allow us to predict the changes in outcomes, we need to assume a direction of causation
  - Prices alter how much you will purchase
  - Hours of study impact grades
  - Years of education alter earnings ability
- Our model will only accurately measure the impact of "x on y" if this assumption is correct

38

Basic model: OLS

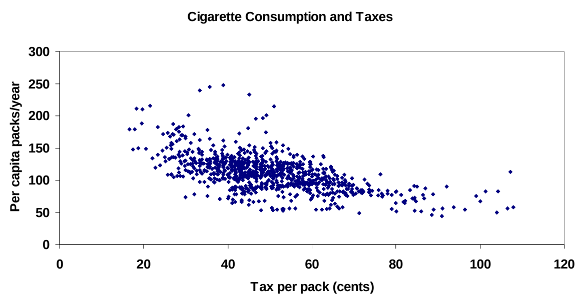
- Ordinary least squares regression
- Maybe 95% of statistics in social sciences
- Highly stylized models with tremendous capacity
  - Capacity comes from assumptions
  - If assumptions are correct – huge rewards
  - If assumptions are wrong, model is piece of junk

39

Example

- State running a budget deficit
- Can raise taxes on cigarettes to cover shortfall
- Problem: when tax rate (t) increase, demand falls (Q) and will impact revenues
- Rev = tQ
- $\partial Rev/\partial t = t[\partial Q/\partial t] + Q$
- Key question: what is $\partial Q/\partial t$

40

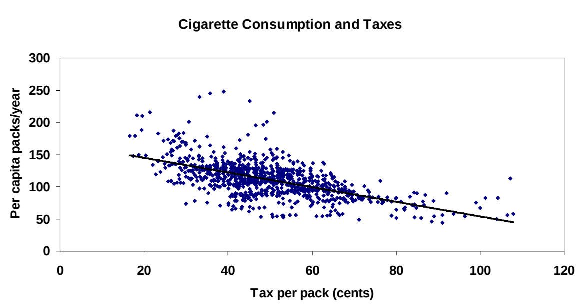


41

Model

- $Y_i = \beta_0 + X_i \beta_1 + \varepsilon_i$
  - Linear
  - One input/one output
  - Y=quantity of cigarettes
  - X=taxes on cigarettes
- Parameter of interest
  - $\partial Y / \partial X = \beta_1$

42



43

Problem

- Can always estimate basic model

  $Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i$
- This does not mean the estimate for β_1 is any good
- Two typical problems that invalidate the estimate of β_1
  - Reverse causation (x may cause y but y may also cause x)
  - Omitted variables bias (some third factor may explain both y and x and hence, explain at least part of the reason why they are statistically related).

44

Reverse Causation: An Economic Example

- Public finance economists are interested in the productivity of government spending
- Two largest components of local spending are schools and public safety
- Will hiring more police reduce crime?

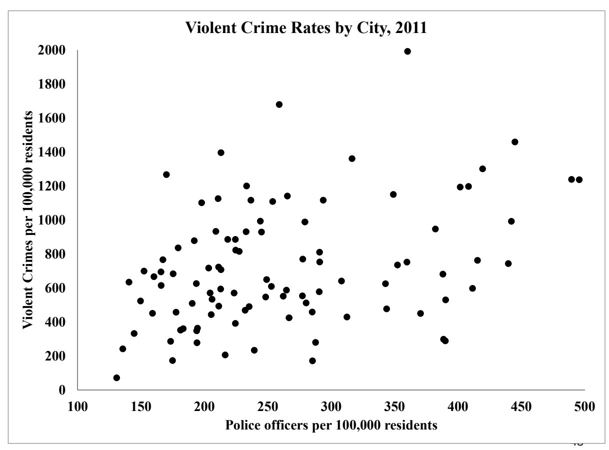
45

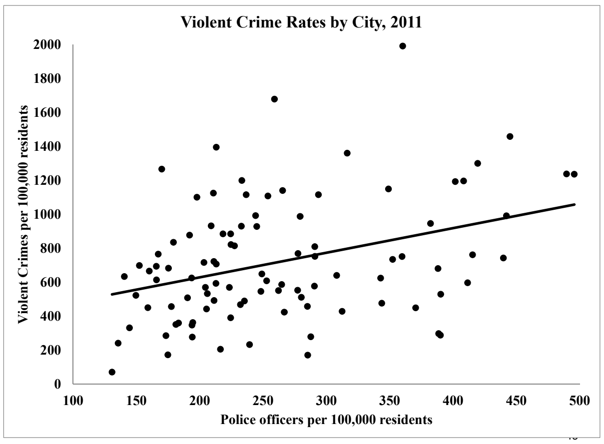
- Let y=crime rate (crime per person)
- Let x=police employed per person
- Interested in estimating the gradient
- $\partial y/\partial x$ how will crime change when a city hires more police

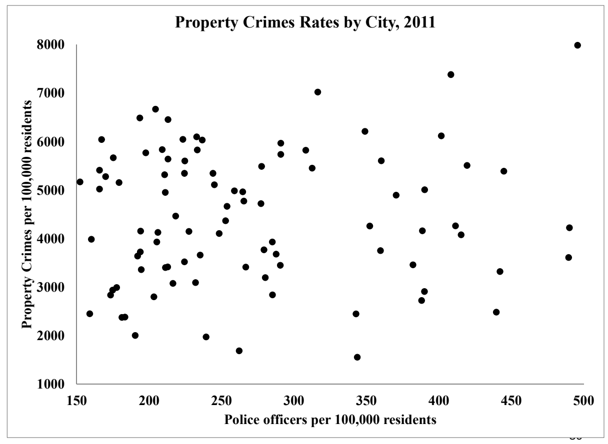
46

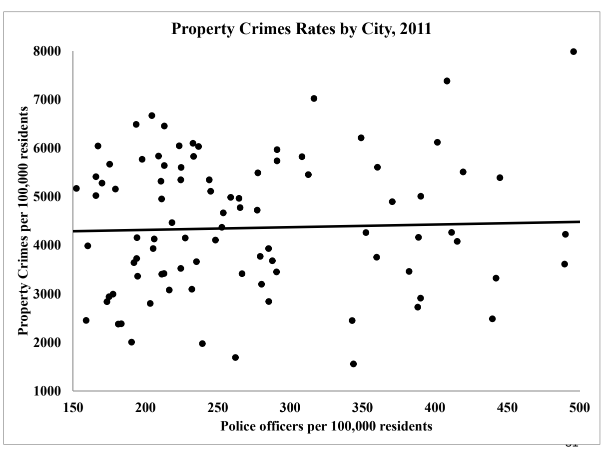
- Collect data on a cross section of cities
  - 61 cities with populations in excess of 250K
- Estimate basic model
  $Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i$
- What do you think is the most frequent sign (+ or -) on police?

47









Highest violent crime rates, largest 100 cities

- Crime Rank

- 1.
- 2.
- 3.
- 4.
- 5.
- 6.
- 7.

52

Omitted variables bias

- Teen childbearing is associated with a number of poor economics outcomes later in life
  - Lower education
  - Lower earnings
  - Higher rates of welfare participation

Outcomes of women aged 30-34 by Teen motherhood status

Outcome	Teen mother	Not a teen mother
< a HS degree	19.8%	6.6%
≥ college degree	9.0%	43.0%
In poverty	30.9%	13.0%
On welfare	6.9%	2.6%
Income from work	$23,884	$36,206

54

Omitted variables bias

- Teen childbearing is associated with a number of poor economics outcomes later in life
  - Lower education
  - Lower earnings
  - Higher rates of welfare participation
- Teen moms are not an random sample of the population – more likely from
  - Poor schools
  - Families with lower-educated moms
  - Families with teen mothers themselves

Washington Post, August 15, 1997, page A3

Lasting Effects Found From Spanking Children Antisocial Behavior Is Increased, Study Says

Spanking children is apt to cause more long-term behavioral problems than most parents who use that approach to discipline may realize, a new study reports.

56

Children who get spanked regularly are more likely over time to cheat or lie, to be disobedient at school and to bully others, and have less remorse for what they do wrong, according to the study by researchers at the University of New Hampshire. It is being published this month in the medical journal Archives of Pediatrics and Adolescent Medicine. "When parents use corporal punishment to reduce antisocial behavior, the long-term effect tends to be the opposite," the study concludes.

57

4 tasks

- Outline basic statistical models
  - How do we get the estimates?
- Demonstrate properties – we want to know
  - When do we get "good" estimates?
  - When do we not??
- Illustrate how they are used in research
  - Do the estimates provide good internal and external validity
- Demonstrate how to obtain results using STATA

58

Take away skills

- Some will use these techniques in the future – make your professor proud

- Some will not – your job is then to be a critical reader of the newspaper

59