A brief introduction to regression models

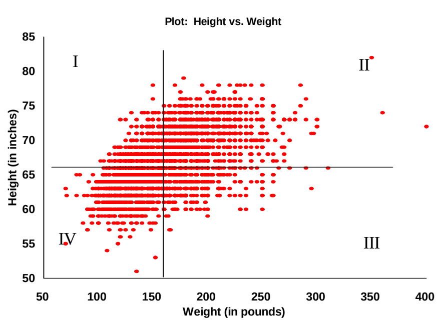
Freshman Honors Seminar
Bill Evans
Spring 2008

1

Scatter plot

- Sample of N observations
  - Students, workers, doctors, etc.
- For each observation, 2 pieces of data (X,Y)
- Plot each point for all observations in sample
- Graphical presentation of the statistical relationship between the two variables

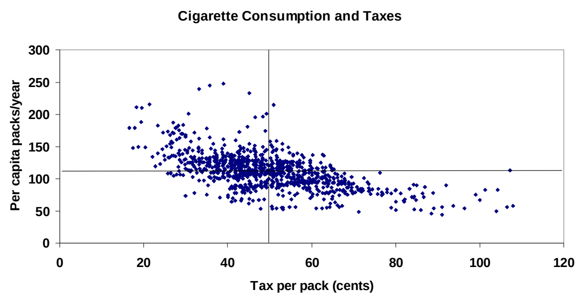
2

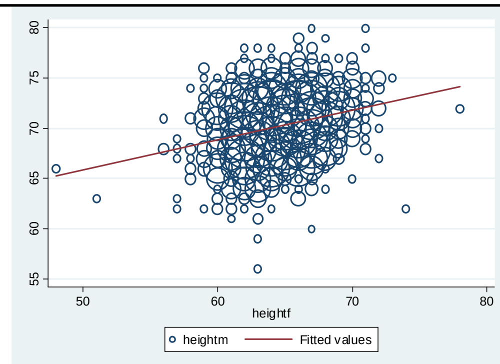


3

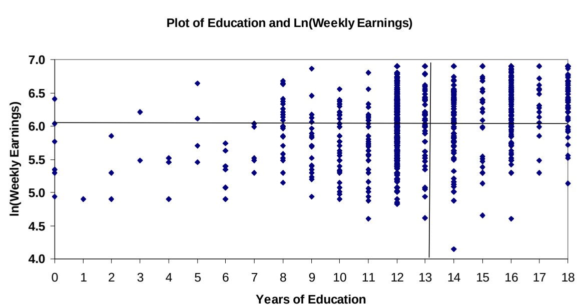
- The shape of the cloud will tell whether the variables are negatively or positively related
- The horizontal and vertical lines are the means for Y and X
- When the variables are + related
  - If X > average, we expect Y > average
  - If X < average, we expect Y < average
- When the variables are – related
  - If X < average, we expect Y > average
  - If X > average, we expect Y < average

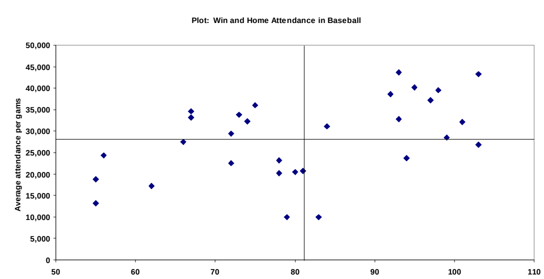
4

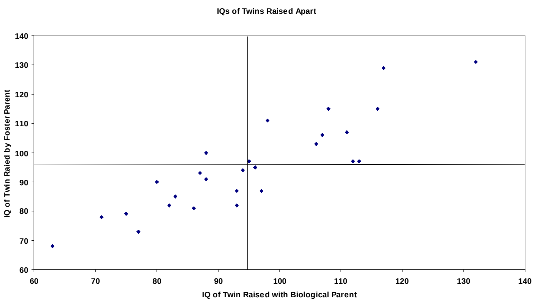




Height in inches of husband (y) and wife (x)







9

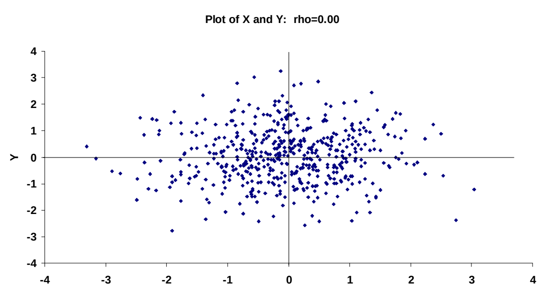
Correlation coefficient

- **The degree to which two variables is measured by the correlation coefficient**
- **Measures how much 'co-movement' there is between the variables**
- **ρ = correlation coefficient**
- **-1 < ρ < 1**

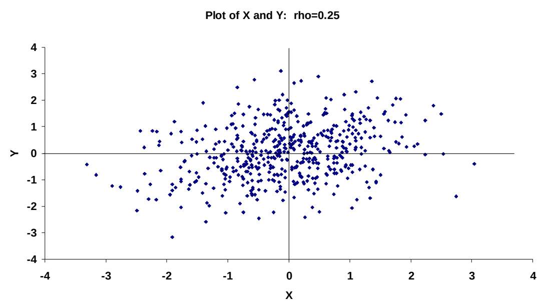
10

- **If ρ=1, perfectly + correlated -- if you know X you know exactly what Y will be and vice verse**
- **If ρ=0, no correlation between variables at all, Y does not tell you anything about the likely vaue of X (and vice versa)**
- **If ρ=-1, perfectly + correlated -- if you know X you know exactly what Y will be and vice verse**

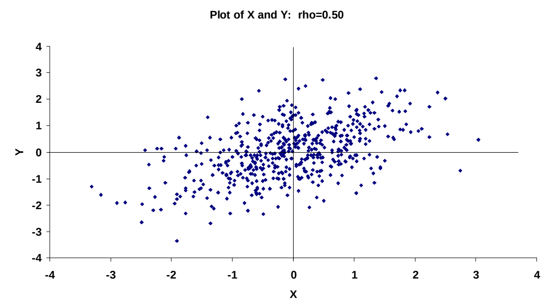
11



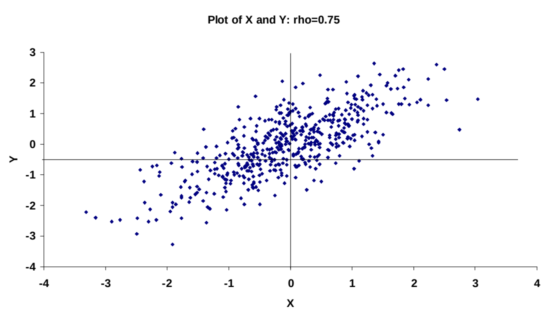
12


Plot of X and Y: rho=0.25
Y
X

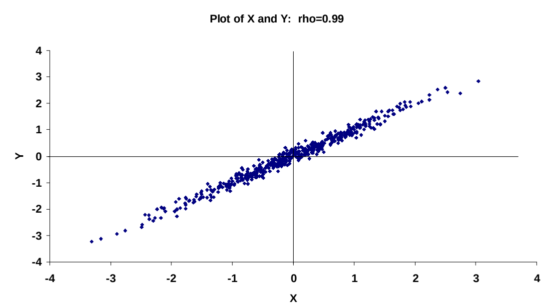
13


Plot of X and Y: rho=0.50
Y
X

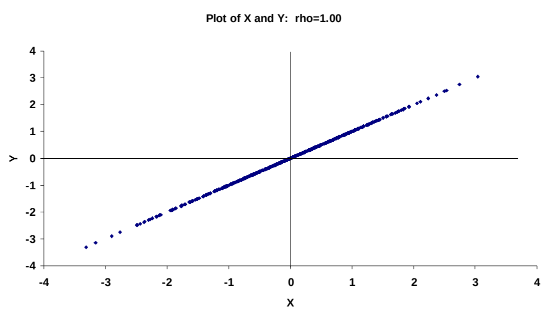
14


Plot of X and Y: rho=0.75
Y
X

15


Plot of X and Y: rho=0.99
Y
X

16



17

How to calculate

- N observations
- $\bar{X} = (x_1+x_2+\ldots x_n)/n = (1/n)\Sigma_i x_i$
- $\bar{Y} = (y_1+y_2+\ldots y_n)/n = (1/n)\Sigma_i y_i$

- $s^2_x = [1/(n-1)]\ \Sigma_i(x_i - \bar{X})^2$
- $s^2_y = [1/(n-1)]\ \Sigma_i(y_i - \bar{Y})^2$

- $\rho = [\Sigma_i(x_i - \bar{X})(y_i - \bar{Y})/(n-1)]/[s_x s_y]$

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
- 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

- Math/verbal SAT
- HS/college GPA
- Math SAT/Coll GPA
- Verbal SAT/Coll GPA

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
- Finally, does not provide for us measure of what we want (dy/dx)

23

Recall

- Define two types of variables
  - Exogenous factors: external conditions
  - Endogenous variables: outcomes" of a system
- Specifics:
  - Y, endogenous, dependent variable
  - X, exogenous, independent variables
- Y=f(x), as we change x, we change y
- dy/dx is the variable we are 'looking' for

24

Example: Theory of Demand

- Consumers derive utility by consuming two types of good: Y_1 and Y_2
- Their 'happiness' follows a number of rules and we can model this with a particular functional form
- The key assumption is declining marginal utility
- $U = U(Y_1,Y_2)$
- $dU/dY_1 > 0$
- $d^2U/dY_1^2 < 0$

25

- Holding Y_2 constant, person always values more of Y_1, but, the 1st unit generates more satisfaction than the 2nd
- $U(Y_1+1,Y_2) > U(Y_1,Y_2)$
- $dU(Y_1+1,Y_2)/dY_1 < dU(Y_1,Y_2)/dY_1$
- What are the constraints?
- Prices and income
- P_1 and P_2 are the prices
- I is income

26

- $I = P_1Y_1 + P_2Y_2$
- Maximize utility $U(Y_1,Y_2)$ subject to the fact that you must pay prevailing prices and cannot spend more than income
- Result: demand curves
- $Y_1=f(P_1,P_2,I)$
- $Y_2=g(P_1P_2,I)$

27

- Key empirical question:
- What are the 'comparative statics'
- dY_1/dP_1, dY_1/dP_2, dY_1/dI

28

- 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

29

- Hypothesize that "x and y are related"
  - Changes in external values of x will alter value of y
  - "comparative statics"
  - Place some structure on the relationship between x and y
- Linear model
    - $y_i = \alpha + \beta x_i + \varepsilon_i$

30

Linear model

- Sample of n observations, labeled as I

- $y_i = \alpha + \beta x_i + \varepsilon_i$

- α and β are "population" values – represent the true relationship between x and y
- Unfortunately – these values are unknown
- The job of the researcher is to estimate these values

31

- Notice that if we differentiate y with respect to x, we obtain
- $dy/dx = \beta$
- β represents how much y will change for a fixed change in x
  - Increase in income for more education
  - Change in crime or bankruptcy when casinos are opened
  - Increase in test score if you study more

32

Put some concreteness on problem

- Suppose a state is experiencing a significant budget shortfall
- Short-term solution – raise tax on cigarettes by 35 cents/pack
- Problem – a tax hike will reduce consumption (theory of demand)
- Question for state – as taxes are raised, how much will cigarette consumption fall

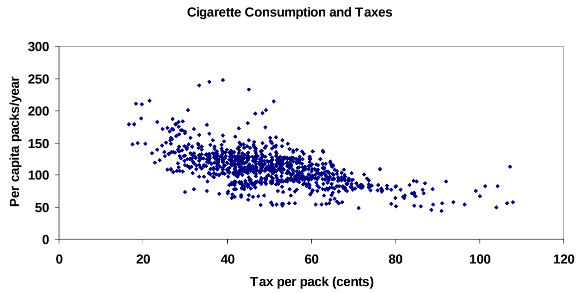
- Suppose y is a state's per capita consumption of cigarettes
- x represents taxes on cigarettes
- Question – how much will y fall if x is increased by 35 cents/pack?
- Note – there are many reasons why people smoke – cost is but one of them –

Benefits and Costs of Model

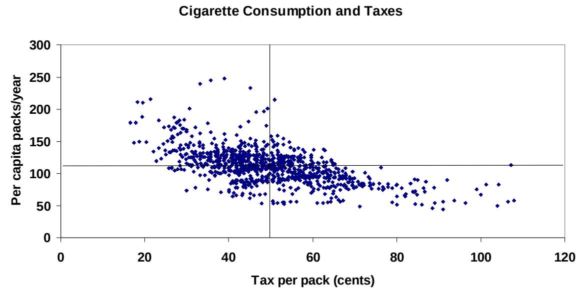
- Placed more structure on the model, therefore we can obtain precise statements about the relationship between x and y
- These statement will be true so long as the hypothesized relationship is true
- As you place more structure on any model, the chance that the assumptions of the model are correct declines.

Data

- (Y) State per capita cigarette consumption for the years 1980-1997
- (X) tax (State + Federal) in real cents per pack
- "Scatter plot" of the data
- Negative covariance between variables
  - When $x>\bar{X}$, more likely that $y<\bar{Y}$
  - When $x<\bar{X}$, more likely that $y>\bar{Y}$
- Goal: pick values of α and β that "best fit" the data
  - Define best fit in a moment



37



38

What is ε_i?

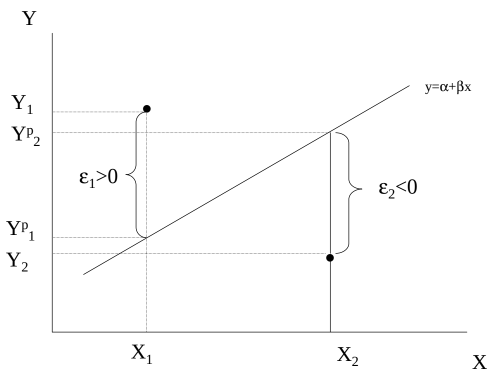
- There are many factors that determine a state's level of cigarette consumption
- Some of these factors we can measure, but for what ever reason, we do not have data
  - Education, age, income, etc.
- Some of these factors we cannot measure
  - Dislike of cigarettes, anti-smoking sentiment of your friends/neighbors/relatives
- ε_i identified what we cannot measure in our model

39

What is ε_i?

- Given linear model $y_i = \alpha + \beta x_i + \varepsilon_i$
- We can predict an level of consumption given parameter values
- $y^p_i = \alpha + \beta x_i$
- The predicted value will not always be accurate – sometimes we will over or under predict the true value
- Because of the linear relationship between x and y, predictions will lie along a line

40



What is ε_i?

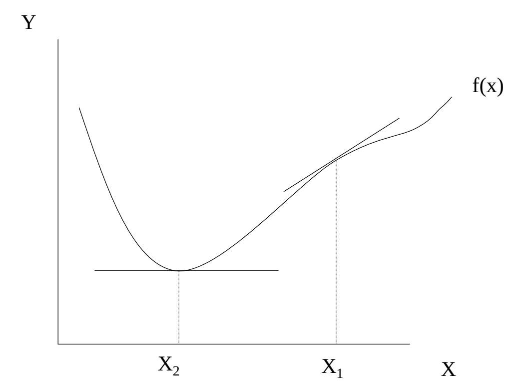
- The difference between the actual and predicted value is the error ε_i
- y_i - y^p_i= y_i - α + βx_i = ε_i
- We never actually observe $\varepsilon_{i.}$ This is the "true error" based on the population values of α and β. Because we do not know α and β, we never know ε_i.
- We can however estimate values of ε_I by estimating values of α and β.
- Our goal, is to choose values for α and β subject to some criteria

Notation

- True model
  - $y_i = \alpha + \beta x_i + \varepsilon_i$
  - We observe data points (y_i,x_i)
  - The parameters α and β are unknown
  - The actual error (ε_i) is unknown
- Estimated model
  - (a,b) are estimates for the parameters (α,β)
  - e_i is an estimate of ε_i where
  - $e_i=y_i-a-bx_i$
- How do you estimate a and b?

Objective: Minimize sum of squared errors

- Min $\Sigma_i e_i^2 = \Sigma_i (y_i - a - bx_i)^2$
- Minimize the sum of squared errors (SSE)
- Treat positive and negative errors equally
  - Over or under predict by "5" is the same magnitude of error
  - "Quadratic form"
  - The optimal value for a and b are those that make the 1st derivative equal zero
  - Functions reach min or max values when derivatives are zero



45

- SSE = $\Sigma_i(y_i - a - bx_i)^2$
- To minimize a function, choose values of a and b that force the 1st derivatives to zero
- d(SSE)/da = $-2\ \Sigma_i(y_i - a - bx_i) = 0$
  - $-2\ \Sigma_i(y_i - a - bx_i) = 0$
    - Multiply by –1/2
  - $\Sigma_i(y_i - a - bx_i) = 0$
    - Divide by n
  - $(1/n)\ \Sigma_i(y_i - a - bx_i) = 0$

46

- Rewrite all terms
- $(1/n)\Sigma_i(y_i) - (1/n)\Sigma_i(a) - (1/n)\Sigma_i(bx_i) = 0$
- Note that
  - » $(1/n)\Sigma_i(y_i) = \bar{Y}$
  - » $(1/n)\Sigma_i(a) = (1/n)(na) = a$
  - » $(1/n)\Sigma_i(bx_i) = (b/n)\ \Sigma_i(x_i) = b\ \bar{X}$
- Therefore
- $\bar{Y} - a - b\ \bar{X} = 0$
- And
- $a = \bar{Y} - b\ \bar{X}$

47

- What is derivative of SSE with respect to b?
- SSE = $\Sigma_i(y_i - a - bx_i)^2$
- d(SSE)/db = $-2\ \Sigma_i x_i(y_i - a - bx_i) = 0$
- From previous slide, we know that
  - $a = \bar{Y} - b\ \bar{X}$
- Substitute this into d(SSE)/db
- $-2\ \Sigma_i x_i[y_i - a - bx_i] = -2\ \Sigma_i x_i[y_i - (\bar{Y} - b\ \bar{X}) - bx_i] = 0$
- Collect like terms
- $-2\ \Sigma_i x_i[(y_i - \bar{Y}) - b(x_i - \bar{X})] = 0$

48

- Multiply both side by –(1/2)
- $\Sigma_i x_i[(y_i – \bar{Y}) – b(x_i - \bar{X})] = 0$
- Expand expression
- $\Sigma_i x_i[(y_i – \bar{Y})] – b\, \Sigma_i x_i[(x_i - \bar{X})] = 0$
- Solve for b
- $b\, \Sigma_i x_i[(x_i - \bar{X})] = \Sigma_i x_i[(y_i – \bar{Y})]$
- $b=\Sigma_i x_i[(y_i – \bar{Y})] / \Sigma_i x_i[(x_i - \bar{X})]$
- and
- $a= \bar{Y} - b\, \bar{X}$

49

Descriptive Statistics

- x=taxes and y=consumption
- $\bar{X}$ =49.60816 (real cents/pack)
- $\bar{Y}$=111.21481 (packs per person per year)

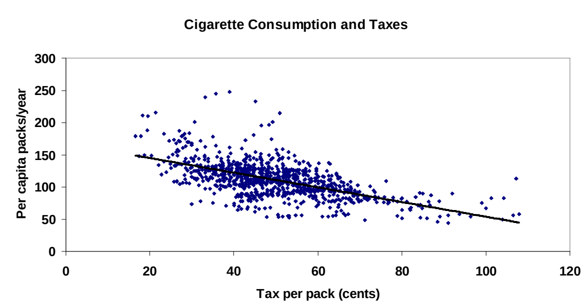
- b= -1.139
- $a= \bar{Y}-b\bar{X}$ = 111.21481 – (-1.139)(49.60816) = 167.72

50

Using the results

- b=dy/dx = -1.139
- For every penny increase in taxes, per capita consumption falls by 1.139 packs per year
- A 35 cent increase in taxes will reduce consumption by (35)(1.139) = 39.8 packs per person per year

51



52

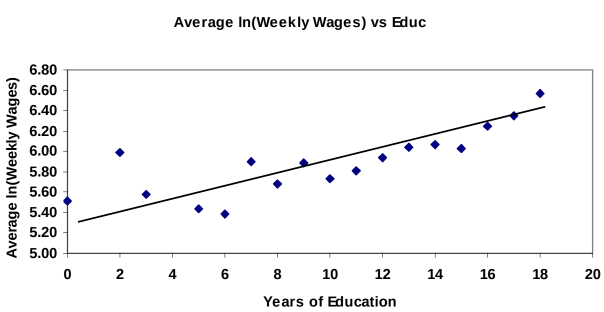
Example 2: Education and Earnings

- Stylized fact: log wages or earnings is linear in education (above a certain range)
- Interpreted as a "return to education"
- Theoretical models why this would be the case
- Linear model:
  - y=ln(weekly wages) – endogenous variable
  - x=years of education – exogenous factor
  - $y_i = \alpha + \beta x_i + \varepsilon_i$

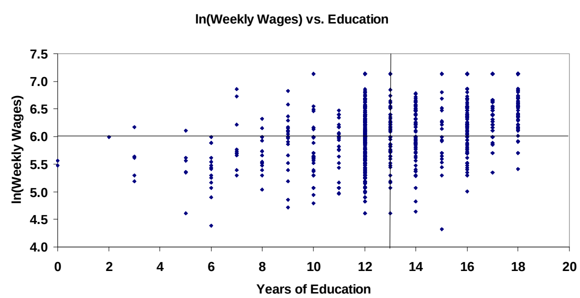
53

- Notice that β has a different interpretation
- β=dY/dX
- In this case, y=ln(Wages)
- dln(Wages)/dX = (1/wages)dWages/dX
- dWages/wages = % change in changes
  - (change in wages over base wages)
- when the endogenous variable is a natural log, β=dY/dX is interpreted as '% change in y for a unit change in x'

54



55

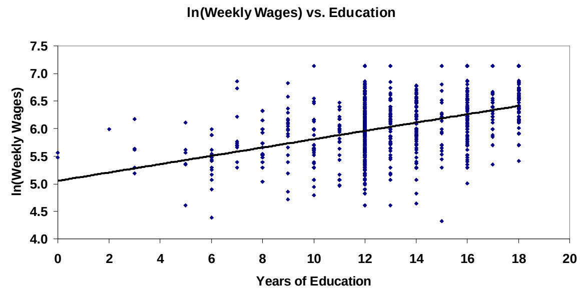


56

Descriptive Statistics

- x=education and y=ln(weekly wages)
- $\bar{X}$ =12.96
- $\bar{Y}$=6.03
- b = 0.075
- a= $\bar{Y}$-b$\bar{X}$ = 6.03 – (0.075)(12.96) = 5.05

57



58