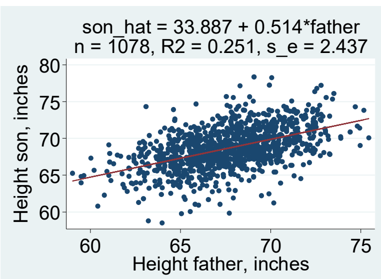
Simple Regression Model (Assumptions)

Lecture 18

Reading: Sections 18.1, 18.2, "Logarithms in Regression Analysis with Asiaphoria," 19.6 – 19.8
(Optional: "Normal probability plot" pp. 607-8)

1

Remember Regression?



s_e (s.d. of residuals) 2.437 inches: measures scatter about OLS line

R^2 0.251: 25.1% of variation in sons' heights explained by variation in their fathers' heights

OLS intercept 33.887: No interpretation b/c father cannot be 0 inches tall

OLS slope 0.514: For every extra 1 inch of father's height, son is on average about ½ inch taller

$\hat{y}$ (y-hat): Predicted y, given x; E.g. son of a 72 inch tall father predicted to be 70.895 inches (= 33.887 + 0.514*72)

e (residual): $e_i = y_i - \hat{y}_i$; E.g. if $\hat{y}_i$ is 70.895 but y_i is 68.531, then residual is -2.364 inches

2

Descriptive & Inferential Statistics

- Chap. 6: Scatterplots, Association, and Correlation
  - $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$
  - $r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n-1}$
- Chap. 7: Introduction to Linear Regression
  - $b = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$
  - $a = \bar{y} - b\bar{x}$
  - $e_i = y_i - \hat{y}_i$
  - $s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$
  - $R^2 = SSR/SST$

- Simple Reg.: Chaps. 18 & 19 (*Inference for Regression* & *Understanding Regression Residuals*)
- Multiple Reg.: Chaps. 20 & 21 (*Multiple Regression* & *Building Multiple Regression Models*)

BUT, multiple regression is also a new way to *describe* data: descriptive statistics

3

Questions and Data: Still Important

- Which kind of question?
  - *Research question*: What is causal effect of a change in X (e.g. match) on Y (e.g. amount given)
  - *Descriptive question*: what are patterns in data (e.g. how does household spending on food vary with income?)
- Which kind of data?
  - Observational or experimental data
    - Correlation ≠ causation is a cliché
    - Instead, *apply* understanding of data and specific context to interpret quantitative results
  - Cross-sectional, time series, or panel data

"The economic impact of universities: Evidence from across the globe"

Excerpt, p. 55: For further description of the data at the national level, we examine the cross sectional correlations of universities with key economic variables. Unsurprisingly, we find that higher university density is associated with higher GDP per capita levels. It is interesting that countries with more universities in 1960 generally had higher growth rates over the next four decades. Furthermore, there are strong correlations between universities and average years of schooling, patent applications and democracy. These correlations provide a basis for us to explore further whether universities matter for GDP growth within countries, and to what extent any effect operates via human capital, innovation or institutions.

Observational or experimental data?

Valero and Van Reenen (2019), https://doi.org/10.1016/j.econedurev.2018.09.001

Figure A3: Scatter Plots at Country Level, Cross Section in 2000
Panel A: Universities and income in 2000

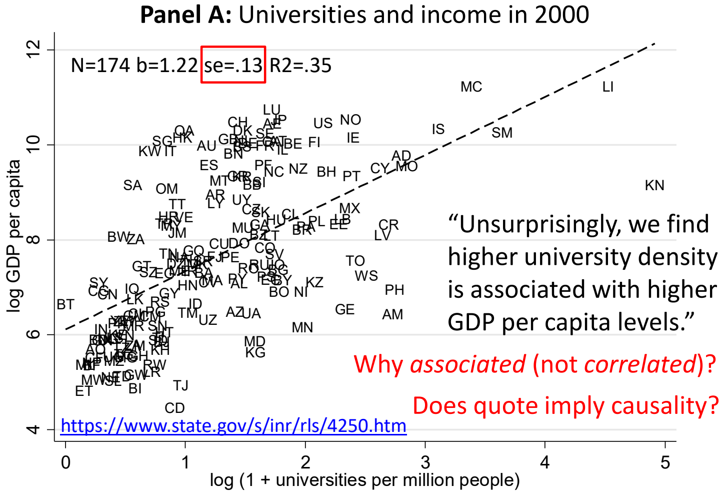


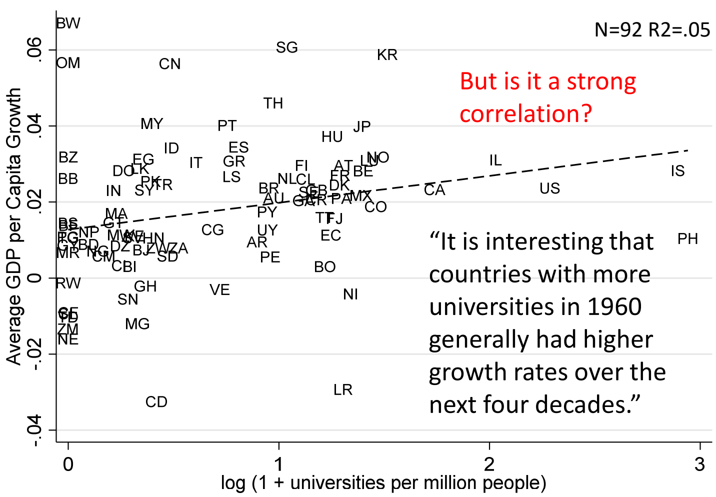
Figure is from appendix of Valero and Van Reenen (2019) and includes: "*Notes*: Each observation is a country in 2000. *Source*: WHED and World Bank GDP per capita"

X-variable is defined as Log(1 + universities per million people)

- Logs can straighten curved scatter plot
  - Plus one addresses countries with 0 universities
  - Example 1: x-value of <u>1</u> is a country with ≈ <u>1.72</u> universities per million: ln(1 + 1.72) ≈ 1
    - E.g. 10 universities w/ pop. 5.82 million: 1.72≈10/5.82
  - Example 2: x-value of <u>3</u> is a country with ≈ <u>19.09</u> universities per million: ln(1 + 19.09) ≈ 3
    - E.g. 25 universities w/ pop. 1.31 million: 19.09≈25/1.31
  - University density is over 11 times bigger in Example 2, but x-value only 3 times as big (diminishing returns)

7

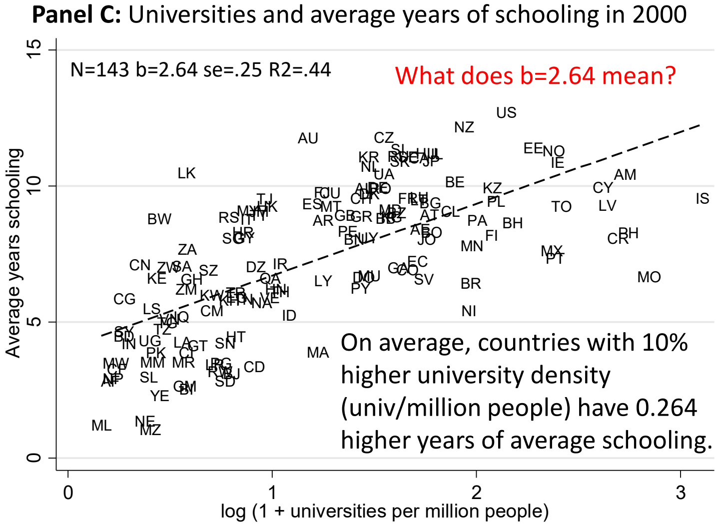
Panel B: Universities in 1960 and GDP/capita growth (1960-2000)



Notes: Each observation is a country. Average annual growth rates over the period 1960-2000 on the y axis. *Source*: WHED and World Bank GDP per capita

8

Panel C: Universities and average years of schooling in 2000



Notes: Each observation is a country. *Source*: WHED and years of schooling obtained from Barro-Lee dataset

9

Frozen Pizza (p. 627)

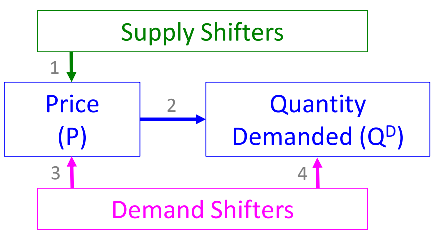
- How does the volume of sales depend on the price of frozen pizza?
  – What is the economic name of this relationship?
- Weekly data on price and quantity for each of four cities (1994 – 1996); 156 weeks
  – Raw data: ch18_MCSP_Frozen_Pizza.csv
  – Cross-sectional, time series, or panel?
  – Are these data observational or experimental?

10

Demand Estimation: Price Endogenous

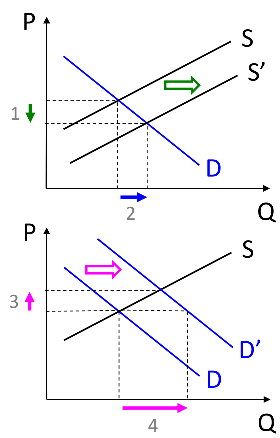
Supply shifters for frozen pizza?

Supply shifters NOT lurking/omitted/ unobserved/confounding variables



Demand shifters ARE lurking/omitted/ unobserved/confounding variables

Demand shifters for frozen pizza?

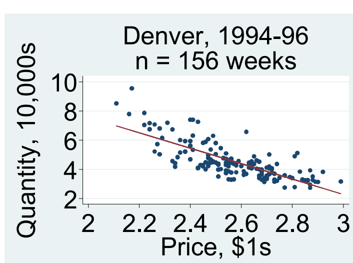


Even if price unchanged, shift directly affects Q^D

11

Frozen Pizza: OLS

- $r = -0.7697$
- $R^2 = 0.5924$
- $\hat{Q} = 18.12 - 5.28\,P$
  – *Interpret* the line?
  For frozen pizza sales in Denver from 1994-96, ___
  – Is the OLS line an estimate of the demand equation?



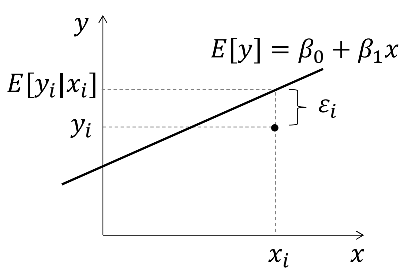
12

Simple Linear Regression: One x-variable

- Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
  - y_i: dependent var., regressand, y-var., LHS-var.
  - x_i: independent var., regressor, explanatory var., x-var., RHS-var. (i.e. right-hand side variable)
  - i: observation index (often i or j cross-sectional data; t time series data; it or jt panel data)
  - β_0: intercept (constant) parameter
  - β_1: slope parameter
  - ε_i: error term, residual, disturbance

13

Error term in $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



- Line is expected value: $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_i$
- Error explains deviations from expectations

- ε_i includes all other factors that affect y_i aside from x_i
  - Impossible to collect data on everything: some variables unobserved to the researcher
  - It reflects reality: model cannot control for everything

In the above graph is ε_i positive or negative?

14

Assumptions Tame Elusive Epsilon

- We *cannot* observe ε_i $\left(\varepsilon_i = y_i - (\alpha + \beta x_i)\right)$ but we *can* observe e_i $\left(e_i = y_i - (a + b x_i)\right)$
  - Notice how many of the six assumptions are about the unobservable ε
    - Some assumptions can be checked by analyzing e_i (the statistic tied to the parameter ε), but some cannot
    - In general, models make assumptions about unknowns
      - For example, a model could assume the outcome of the role of a die follows a discrete Uniform distribution: i.e. it's fair with a 1/6 probability of each outcome {1, 2, 3, 4, 5, 6}

15

<u>Six</u> Assumptions of Linear Regression Model

- Book gives only four:
  - One skipped b/c obvious
  - Another skipped b/c only required for a causal interpretation
  - To minimize confusion, list extra two as 5 & 6
- Econometrics addresses *substantial* violations of assumptions
  - ECO372H Applied Regression Analysis and Empirical Papers
  - ECO374H Forecasting and Time Series Econometrics
  - ECO375H Applied Econometrics I
  - ECO475H Applied Econometrics II

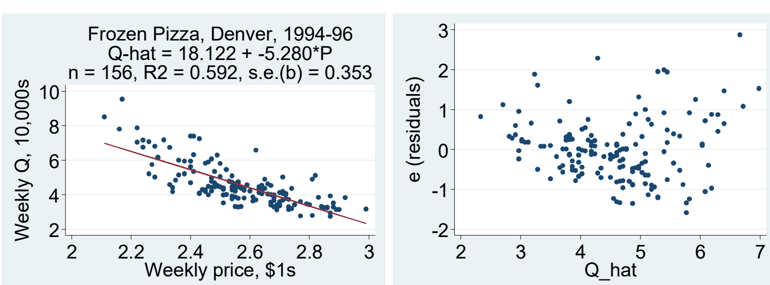
Assumption #1

- Regression equation is linear in the <u>error</u> and <u>parameters</u>; the variables (in boxes) are linearly related to each other

$$\boxed{} = \alpha + \beta \boxed{} + \varepsilon_i$$

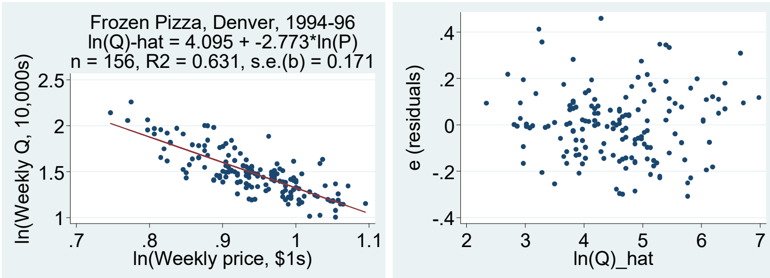
  - <u>Not</u> assuming that what is in boxes is linear (so long as no nonlinear functions of *parameters* or nonlinear functions of the *error*)
    - Example of a <u>linear</u> regression: $y_i = \alpha + \beta x_i^2 + \varepsilon_i$
    - Example of a <u>linear</u> regression: $\ln(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$

Diagnostic Plot: e versus $\hat{y}$



Which violations can we see?

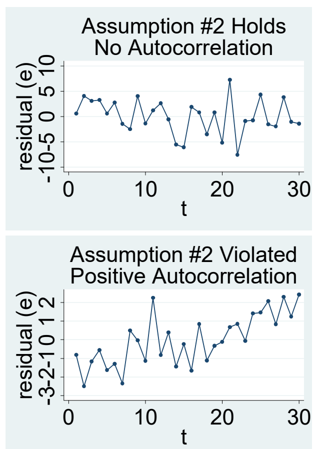
Natural Log Transformations



19

Assumption #2

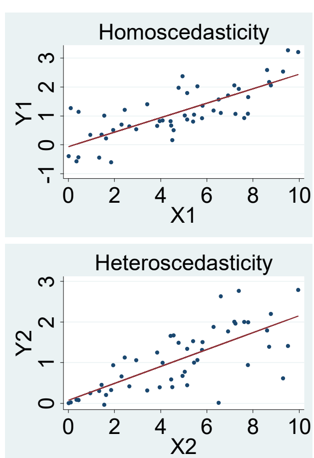
- No autocorrelation / no serial correlation: $COV[\varepsilon_i, \varepsilon_j] = 0$ if $i \neq j$
  - Common problem in time-series data
    - E.g. higher than expected inflation today, likely high tomorrow
  - Errors assumed not systematically related across observations



20

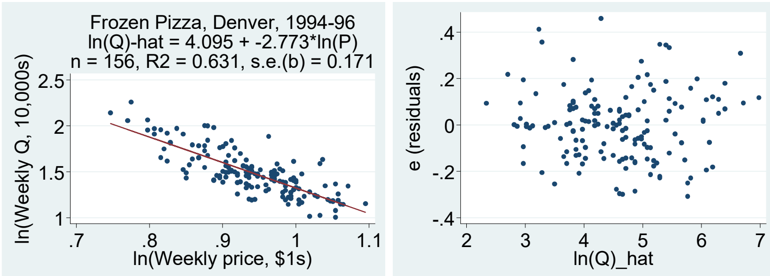
Assumption #3

- Homoscedasticity: $V[\varepsilon_i] = \sigma_\varepsilon^2, i = 1, \ldots, n$
  - "Equal variance assumption"
  - Error ε_i is just as "noisy" for all values of x
  - Violation is called heteroscedasticity
  - Common problem in cross-sectional data



21

Fix Assumption #1 issues before checking Assumption #3



Heteroscedasticity – unequal variance of the residuals – is often a byproduct of a violation of the linearity assumption

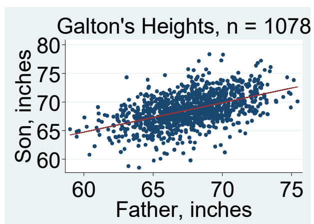
Is Denver pizza regression an example?

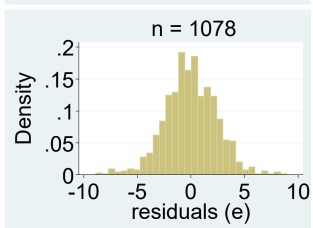
Remember that Chapter 18 advises you to check the assumptions in order: start with the linearity assumption

22

Assumptions #4 & #5

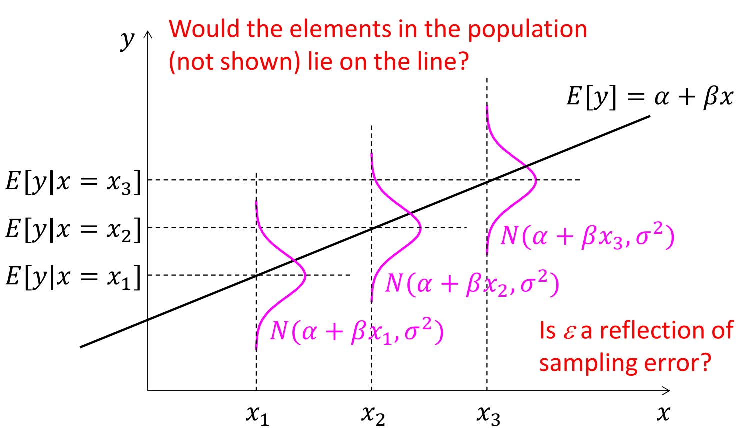
- Galton's data (Lec. 5)
  - Assumptions 1-3 hold?
- Normality: ε_i is Normal
  - ε_i is unobserved so check $e_i = y_i - \hat{y}_i$
- Error has mean zero: $E[\varepsilon_i] = 0, i = 1, \ldots, n$
  - Constant term (i.e. $\beta_0$or α) picks up any constant effects, not the error





23

Graphical Summary

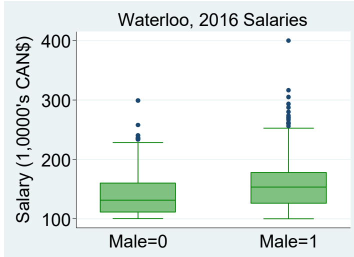


Assumptions #3, #4, and #5 combined: $\varepsilon_j \sim N(0, \sigma^2)$

24

2017 ON Public Sector Disclosure for University of Waterloo employees

Sex	n	Mean	S.d.
F	416	$139.74K	$33.74K
M	941	$155.36K	$36.96K



OLS Results:

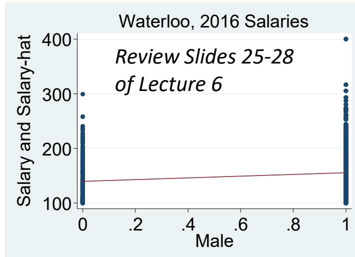
Salary-hat = 139.74 + 15.62*Male

R^2 = 0.0385, n = 1,357, s_e = 36.006

Assumption #1 violated?

Assumption #3 violated?

Assumption #4 violated?



Assumption #6

- x uncorrelated w/ error: $COV[x_i, \varepsilon_i] = 0$
  - Exogeneity: x variable(s) unrelated with error
    - Dosage is exogenous: $Sleep_i = \alpha + \beta dosage_i + \varepsilon_i$
    - Experimental data *can* est. *causal* effect: $E[b] = \beta$
  - Endogeneity: x variable(s) related with error
    - With observational data, lurking/unobserved/omitted/confounding variables mean x and error are related
    - Price of pizza is endogenous: $Q_t = \beta_0 + \beta_1 P_t + \varepsilon_t$
    - Endogeneity bias means: $E[b_1] \neq \beta_1$

In estimating $Salary_i = \beta_0 + \beta_1 Male_i + \varepsilon_i$ with $n = 1{,}357$ Waterloo employees, is $Male$ endogenous?

"Short-Hand" Assumptions

1) Linear relationship between variables (possibly non-linearly transformed)
2) No correlation amongst errors (no autocorrelation for time-series data)
3) Homoscedasticity (single variance) of errors
4) Normally distributed errors
5) Constant included (error has mean 0)
6) No relationship between x and error