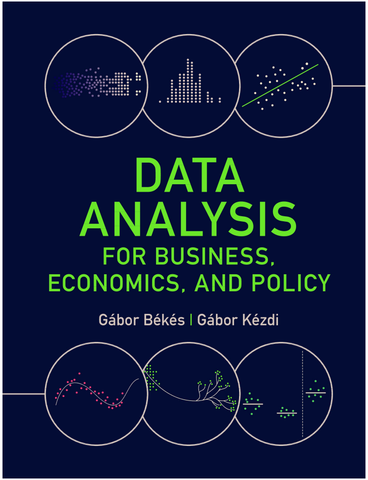
08 Complicated patterns and messy data

Gábor Békés (CEU)

Data Analysis 2: Regression analysis

2020-05-29

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021 January
- Available in paperback, hardcover and e-book
- Slideshow be used and modified for educational purposes only
- **gabors-data-analysis.com**
  - Download all data and code
  - Additional material, links to references

Function form thinking

- Relationships are complicated
- When and why care about the shape of a regression?
- How can we capture function form better?

Motivation

- Interested in the pattern of association between how long people live in a country and how rich that country is.
- Uncovering that pattern may be interesting for many reasons.
- Identify countries where people live longer than what we would expect based on their income, or countries where people live shorter lives.
- That would require analyzing regression residuals,
- For this purpose, getting a good approximation of the $y^E = f(x)$ function is important.

Functional form

- Linear regression – linear approximation to a regression of unknown shape: $y^E = f(x)$.
- Get the regression to better characterize the nonlinear pattern if
  - when we want to make a prediction or analyze residuals;
  - when we want to go beyond the average pattern of association;
  - when all we care about is the average pattern of association, but the linear regression gives a bad approximation to that.
- Not care
  - if all we care about is the average pattern of association,
  - if linear regression is goof approximation to that average pattern

Functional form

- Non-parametric methods great to get functional form, but no parameters – hard to replicate
- Need model functional form
  - Simplify
  - Make assumption
  - Accept that it will be far from perfect
- Many options, decisions will be needed
  - Sometimes, our use / theory will help pick a model
  - Sometimes executive decision which approach to use.

Functional form: ln transformation

- Frequent nonlinear patterns better approximated with y or x transformed by taking **relative differences**.
  - Example: time series of GDP per capita
- Not in cross-sectional data where there is no natural base for comparison.
- Taking the **natural logarithm** of a variable is often a good solution in such cases.
- When transformed by taking the natural logarithm, differences in variable values *approximate relative differences*.
- Examining log differences works because differences in natural logs approximate percentage differences.

Logarithmic Functions of y and/or x

- $ln(x)$ = the natural logarithm of x
  - Sometimes we just say log x and mean $ln(x)$. Could also mean log of base 10. Here we use $ln(x)$
- x needs to be a positive number
  - $ln(0)$ or $ln(negative\ number)$ do not exist
- Log transformation allows for comparison in relative terms (percentage), because:

$$ln(x+y) - ln(x) = ln(1 + \frac{\Delta x}{x}) \approx \frac{\Delta x}{x} \tag{1}$$

- Numerically:
  - $\ln(1.01) = 0.0099 \approx 0.01$
  - $\ln(1.1) = 0.095 \approx 0.1$

Logarithmic Functions of y and/or x

- Alternatively, from the relationship in calculus:

$$\frac{d ln(x)}{dx} = \frac{1}{x}$$

  - So that, for small Δx,

$$\frac{ln(x+\Delta x) - ln(x)}{\Delta x} \approx \frac{1}{x}$$

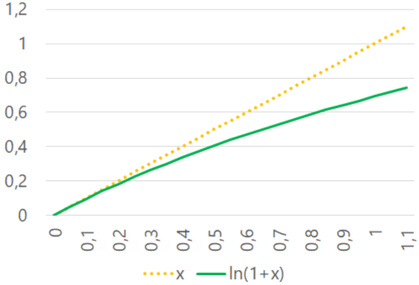
  - And thus

$$ln(x+\Delta x) - ln(x) \approx \frac{\Delta x}{x}$$

- i.e., the difference between the natural log of two numbers is approximately the relative difference between the two
  - For small differences

Log Approximation of Small Relative Diffs

- Log differences approximate small relative differences
- When x is small
  - 0.3 or smaller
  - the log approximation is close
- But for larger x, there is a difference,
  - And may have to calculate percentage change by hand



Ln(x) vs percentage

- Log differences approximate relative differences (percent)
- log difference - we mean $ln(x)$
    - x needs to be a positive number
- Log differences approximate small relative differences - say below 0.3
    - A difference of 0.1 log units corresponds to a 10% difference
    - For larger positive differences, the log difference is smaller
        - A log difference of +1.0 corresponds to a +170% difference
    - For larger negative differences, the log difference is larger in absolute value
        - A log difference of -1.0 corresponds to a -63% difference

Taking logs

- When to take logs?
- When comparison makes mores sense in relative terms
  - Percentage differences
- Most important examples
  - Prices
  - Sales, turnover, GDP
  - Population, employment
  - Capital stock, inventories

Interpreting parameters of regressions with log variables

$ln(y)^E = \alpha + \beta x_i$

- log y, level x
- α is average ln(y) when x is zero. (Often meaningless.)
- β :y is $\beta * 100$ percent higher, on average for observations with one unit higher x.

$y^E = \alpha + \beta ln(x_i)$

- level y, log x
- α is : average y when ln(x) is zero (and thus x is one).
- β y is $\beta/100$ units higher, on average, for observations with one percent higher x.

$ln(y)^E = \alpha + \beta ln(x_i)$

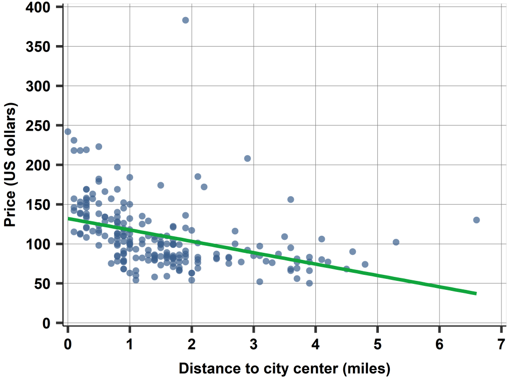
- log y, log x
- α: is average $ln(y)$ when $ln(x)$ is zero. (Often meaningless.)
- β: **y is** β **percent** higher on average for observations with one percent higher x.

Interpreting parameters of regressions with log variables

- Precise interpretation is key
- The interpretation of the slope coefficient differs in each case!
- Often verbal comparison is made about a 10% difference in x and a $\beta/10$ percent difference in y

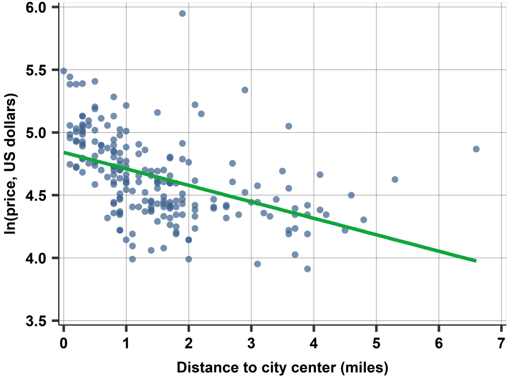
Hotel price-distance regression and functional form

- price =132.02-14.41*distance
- Issue ?



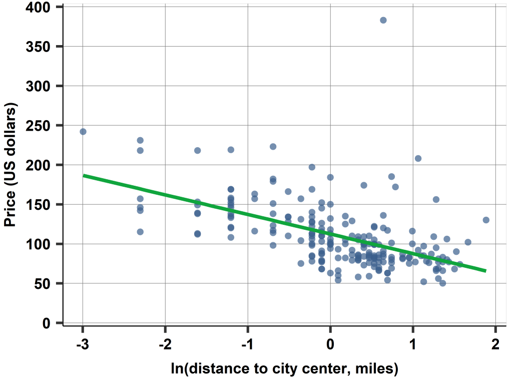
Hotel price-distance regression and functional form

- ln_price =4.84-0.13*distance
- Better approximation to the average slope of the pattern.
    - Distribution of log price is closer to normal than the distribution of price itself.
    - Scatterplot is more symmetrically distributed around the regression line



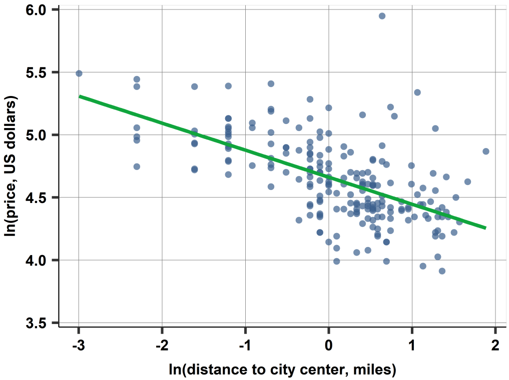
Hotel price-distance regression and functional form

- price =116.29-28.30*ln_distance
- We now make comparisons in terms percentage difference in distance



Hotel price-distance regression and functional form

- ln_price =4.70-0.25*ln_distance
- Comparisons relative terms for both price and distance



Hotel ratings and measurement error

táblázat. Hotel price and distance regressions

	(1)	(2)	(3)	(4)
VARIABLES	price	ln(price)	price	ln(price)
Distance to city center, miles	-14.41	-0.13		
ln(distance to city center)			-24.77	-0.22
Constant	132.02	4.84	112.42	4.66
Observations	207	207	207	207
R-squared	0.157	0.205	0.280	0.334

Robust standard errors in parentheses

Source: `hotels-vienna` dataset. Prices in US dollars, distance in miles.

Hotel price-distance regression interpretations

- price-distance: hotels that are 1 mile farther away from the city center are 14 US dollars less expensive, on average.
- ln_price - distance: hotels that are 1 mile farther away from the city center are 13 percent less expensive, on average.
- price - ln_distance: hotels that are 10 percent farther away from the city center are 2.830 US dollars less expensive, on average.
- ln_price - ln_distance: hotels that are 10 percent farther away from the city center are 2.50 percent less expensive, on average.

To Take log or Not to Take log

- Decide for substantive reason
- Take logs if variable is likely affected in multiplicative ways
  - i.e. increased or decreased by certain percentages
  - Often when variable is price, GDP, population
  - Sometimes even if variable is already a ratio, such as GDP/population
- Don't take logs if variable is likely affected in additive ways
  - i.e., increased or decreased by absolute values
  - Often when variable is a count, or percentage cannot be interpreted (grades)

To Take log or Not to Take log

- Decide for statistical reason
- Linear regression is better at approximating average differences if distribution of dependent variable is closer to normal
- Take logs if skewed distribution with long right tail
  - Don't take logs if already symmetric
    - Or skewed distribution with long left tail

- Most often the substantive and statistical arguments are aligned
  - the distribution of variables that are the results of multiplicative factors is usually skewed with a long right tail.
  - In case of conflict of these reasons focus on the interpretation of the slope coefficient and go with what seems best

To Take log or Not to Take log

- Log needs variable to be positive
  - Never negative, never zero
- What if want log but variable can be 0 or negative?
  - For example, wealth
- Create binary indicator(s) and log
  - For example, is wealth negative? (0 or 1) Is wealth zero? (0 or 1) If positive, log wealth (and replace this log value with zero if log cannot be taken)
- Sometimes adding a constant seems to do the trick
  - ln(x+1) if x is positive or zero
  - Not a good solution in principle
  - May be fine if x is either zero or takes on large values
    - If x is large ln(x) is almost the same as ln(x+1)

Hotel price-distance regression and model choice

- Log price models capture patterns better, this could be preferred.
- Ultimately, it depends on the goal of the analysis.
  - We are after a good deal on a single night – absolute price differences are meaningful,
  - Percentage differences in price may remain valid if inflation and seasonal fluctuations affect prices proportionately.

Hotel price-distance regression and model choice

- Compare Level-level and log-level regression: R-squared of the log-level regression is higher, suggesting a better fit.
- But: should not compare R-squared of two regressions with different dependent variables – compares fit in different units.
- log-log vs log-level - can compare and log-log wins on fit.

Hotel price-distance regression and model choice

- Compare Level-level and log-level regression: R-squared of the log-level regression is higher, suggesting a better fit.
- But: should not compare R-squared of two regressions with different dependent variables – compares fit in different units.
- log-log vs log-level - can compare and log-log wins on fit.
- Distance makes more sense in miles than in relative terms – given our purpose
- Could be that log-log is the right model - if e.g. prediction is the goal.
  - Note that prediction with log dependent variable is tricky

Other transformations: splines

- A regression with a piecewise linear spline of the explanatory variable
  - results in connected line segments for the mean dependent variable,
  - each line segment corresponding to a specific interval of the explanatory variable.
- The points of connection are called knots,
  - the line may be broken at each knot so that the different line segments may have different slopes.
  - A piecewise linear spline with m line segments is broken by m-1 knots.
- The places of the knots (the boundaries of the intervals of the explanatory variable) need to be specified by the analyst.
  - R, Stata have built-in routines/library to do that.

Other transformations: splines

- A piecewise linear spline regression results in connected line segments, each line segment corresponding to a specific interval of x.
- We can interpret parameters!
- The formula for a piecewise linear spline regression with m line segments (and $m-1$ knots in-between) is:

$$y^E = \alpha_1 + \beta_1 x[\text{if } x < k_1] + (\alpha_2 + \beta_2 x)[\text{if } k_1 \leq x \leq k_2] + \ldots + (\alpha_m + \beta_m x)[\text{if } x \geq k_{m-1}] \tag{2}$$

Other transformations: splines

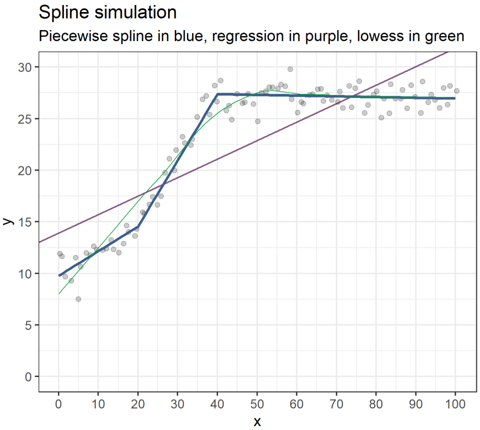
$$y^E = \alpha_1 + \beta_1 x[\text{if } x < k_1] + (\alpha_2 + \beta_2 x)[\text{if } k_1 \leq x \leq k_2] + \ldots + (\alpha_m + \beta_m x)[\text{if } x \geq k_{m-1}]$$

Interpretation of the most important parameters

- α : average y when x is zero.
- β_1 : When comparing observations with x values less than k_1, y is β_1 units higher, on average, for observations with one unit higher x value.
- β_2 : When comparing observations with x values between k_1 and k_2, y is β_2 units higher, on average, for observations with one unit higher x value.
- β_m : When comparing observations with x values greater than k_{m-1}, y is β_m units higher, on average, for observations with one unit higher x value.

Functional form: splines

- Simple spline
- Knots at 20, 40
- $\alpha = 10$
- $\beta_1 = 0.2$
- $\beta_2 = 0.7$
- $\beta_3 = 0.0$



Other transformations: splines

- The way it's formulated is that each segment has a slope we can interpret
- An alternative presentation of the same results shows the slope of a reference line segment and the difference of the slope from the reference for every other line segment
- Need to check output...

Other transformations: splines

- A regression with a piecewise linear spline of the explanatory variable
- Handles any kind of nonlinearity
  - Including non-monotonic associations of any kind
- Offers complete flexibility
- But requires decisions from the analyst
  - How many knots?
  - Where to locate them
  - Decision based on scatterplot, theory / business knowledge
  - Often several trials.
- Automatic models:
  - Look at patterns and optimize some penalty function for complexity
  - mfb in R, fp in Stata.

Other transformations: polynomials

$$y^E = \alpha + \beta_1 x + \beta_2 x^2 \tag{3}$$

- Quadratic function of the explanatory variable – allow for a smooth change in the slope
- Convex or concave relationship
- Technically: quadratic function, not linear function (a parabola, not a line)
- Handles only certain kinds of nonlinearity and offers substantially less flexibility than a piecewise linear spline
- But requires no further decision from the analyst

- Can have higher order polynomials: cubic with $\beta_3 * x^3$, etc

Other transformations: quadratic form

$$y^E = \alpha + \beta_1 x + \beta_2 x^2 \tag{4}$$

- α is average y when $x = 0$,
- β_1 has no interpretation,
- β_2 only shows whether the parabola is
  - U-shaped or convex (if $\beta_2 > 0$)
  - inverted U-shaped or concave (if $\beta_2 < 0$).
- the slope is different for different values of x,
- the slope is $\beta_1 + 2\beta_2 x$ (the first derivative of the quadratic function).
  - We can compare two observations, denoted by j and k, that are different in x, by one unit so that $x_k = x_j + 1$. y is higher by $\beta_1 + 2\beta_2 x_j$ units for observation k than for observation j.

Other transformations: ratios

- Ratios of variables - normalization of totals
- Most often: per capita: GDP/capita, revenues/employee, sales/shop
- Can take logs easily

Other transformations: ratios

- Ratios of variables - normalization of totals
- Most often: per capita: GDP/capita, revenues/employee, sales/shop
- Can take logs easily
- log of a ratio equals the difference of the two logs:
  $ln(GDP/Pop) = ln(GDP) - ln(Pop)$.

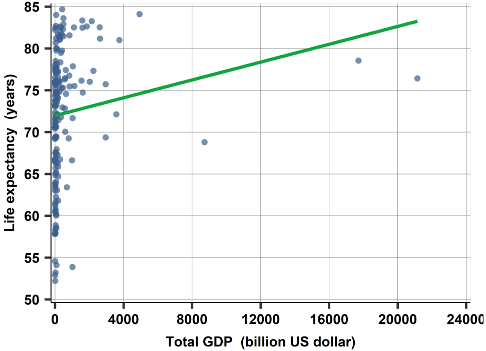
Life expectancy and income

- How long people live in a country and how rich that country is.
- To identify countries where people live longer than what we would expect based on their income, or countries where people live shorter lives.
- Analyzing regression residuals – getting a good approximation of the $y^E = f(x)$ function is important.

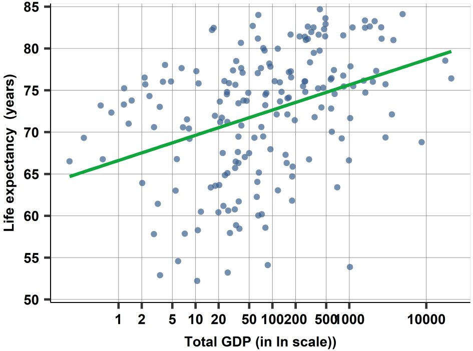
Life expectancy and income

- Life expectancy at birth in a given year is a measure of how long people live in a country on average. It is the average age at which people die in the given year.
- Data from World Development Indicators website, maintained by the World Bank.
- Massive panel data,we use 2017.
- There are 217 countries in this data table, but GDP and life expectancy is available for only 182
- Average life expectancy is 72 years, with a range of 52 to 85.
- Total GDP is 0.2 billion to 20 trillion,
  - Due to variation both in size (number of people) and income per person.

Life expectancy and total GDP

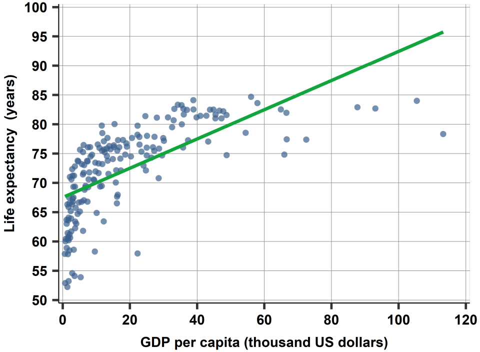


ábra. Life expectancy and total GDP

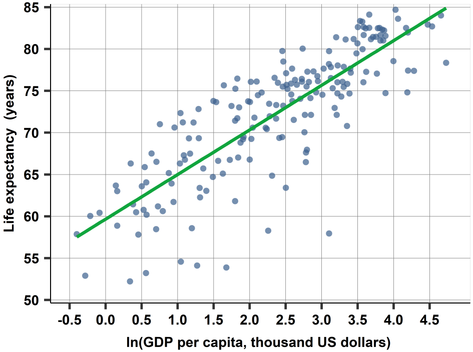


ábra. Life expectancy and ln total GDP

Life expectancy and GDP per capita

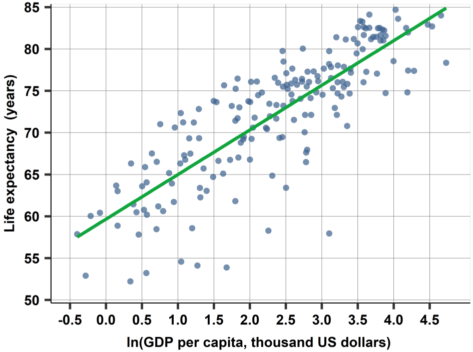


ábra. Life expectancy and GDP/capita

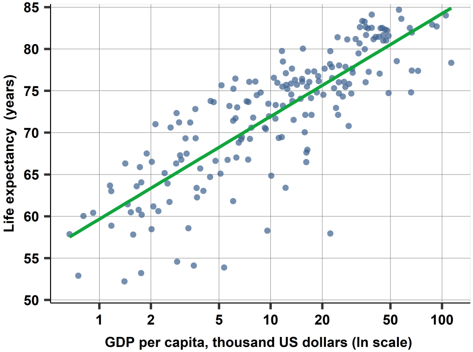


ábra. Life expectancy and ln GDP/capita

Life expectancy and GDP per capita - log representations



ábra. Life expectancy and GDP/capita



ábra. Life expectancy and ln GDP/capita

Model choice 1

- Taking ln GDP - because we typically care about percentage and not dollar differences
- Normalize with population as we care about per capita income to measure richness of a country

- level-log regression, slope is 5.3
  - 1 percent higher GDP per capita have life expectancy higher by 0.053 years, on average.
- Countries with a 10 percent higher GDP per capita have a half (0.53) year higher life expectancy on average.

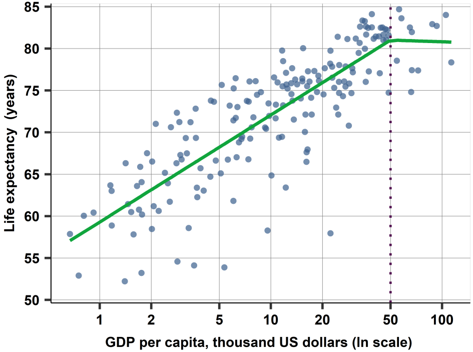
Life expectancy and income - findings

- Countries with the shortest lives given their income include Equatorial Guinea, Nigeria, and Cote d'Ivoire (about 11-17ys minus)
- Countries with the longest lives given their income include Vietnam, Nicaragua, and Lebanon (7ys more)
- Lives are more than two years shorter than expected in the U.S.A., and five years longer than expected in Japan.

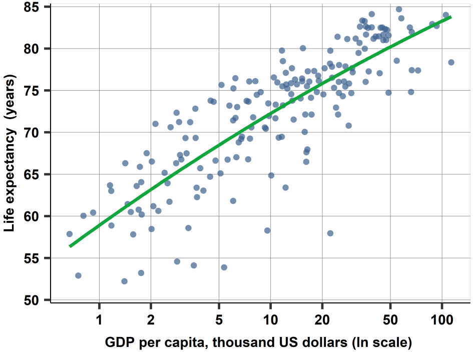
Life expectancy and income

- Improve model fit, as we care about residuals
- Already took rario and log
- Splines - capture flattening of at the end
  - Quadratic - similar purpose
- Can we compare R-squared of?
  - Life expectancy regressed on level GDP/capita
  - Life expectancy regressed on ln GDP/capita
  - Life expectancy regressed on splines of ln GDP/capita
  - Life expectancy regressed on level GDP/capita and its square

Life expectancy and GDP per capita



ábra. Ln GDP/capita - spline



ábra. Ln GDP/capita - quadratic

Life expectancy and income

- R-squared slightly higher for quadratic or spline
- The ranking of the top and bottom lists are similar across these various regressions, although the magnitudes of the residuals differ slightly.
- Expected: both make little change to what is basically a linear association.
  - Here: linear is indeed a good approximation. Not perfect, but good

Which functional form to choose?

- Start with deciding whether you care about nonlinear patterns.
  - Linear approximation OK if focus is on an average association.
  - Transform variables for a better interpretation of the results (e.g. log), and it often makes linear regression better approximate the average association.
  - Accommodate a nonlinear pattern if our focus is
    - on prediction,
    - analysis of residuals,
    - about how an association varies beyond its average.

Which functional form to choose?

- Do the rest if you want to uncover and include a potentially nonlinear pattern in the regression analysis.
- Uncover the most important features of the pattern of association by examining a scatterplot or a graph produced by a nonparametric regression such as lowess or bin scatter.
- Choose one or more ways to incorporate those features into a linear regression (transformed variables, piecewise linear spline, quadratic, etc.).
- Compare the results across various regression approaches that appear to be good choices. –> **robustness check**.

Data Is Messy

- Clean and neat data exist only in dreams
  - and in some textbooks
- Data may be messy in many ways
- Structure, storage type differs from what we want
  - Needs cleaning — DA1
- Some observations are influential
  - Could drop them, but better not to
- Variables measured with error
- In aggregate data some observations may represent more individuals
  - May or may not use weights

Extreme values and influential observations

- Extreme values concept
  - Observations with extreme values for some variable
- Extreme values examples
  - Banking sector employment share in countries. Luxembourg: 10%
  - Number of foreign companies registered/population. Cyprus or US Virgin Island
  - Hotel price of 1 US dollars, 10,000 US dollarss
- Influential observations
  - Their inclusion or exclusion influences the regression line
  - Influential observations are extreme values
  - But not all extreme values are influential observations
- Influential observations example
  - Very large tech companies in a regression of size and average wage

Extreme values and influential observations

- What to do with them?
- Depends on why they are extreme
  - If by mistake: may want to drop them (EUR1000+)
  - If by nature: don't want to drop them (other hotel)
  - Grey zone: patterns work differently for them for substantive reasons
  - Here: general rule: avoid dropping observations based on value of y variable
- Dropping extreme observations by x variable may be OK
  - May want to drop observations with extreme x if such values are atypical for question analyzed
  - But often extreme x values are the most valuable as they represent informative and large variation

Measurement Errors In Variables

- Goal: measuring the association between variables
    - Spending and income; price and distance, etc.
- Interested in the estimated coefficient value (not just the sign)
- Observed variables have measurement error
    - Mistake, hard-to-measure data, created variables
- Often cannot do anything about it!
- So, question is what the consequence is of such errors
- Does measurement error make a difference in the model parameter estimates?
    - Whether we expect parameters (such as OLS coefficient) to be different from what they would be without the measurement error
- Does the answer depend on the type of measurement error?

Classical Measurement Error

Classical measurement error (also called noise)

1. is zero on average (so it does not affect the average of the measured variable) and
2. is independent of all other relevant variables, including the error-free variable.

Examples

Classical Measurement Error

Classical measurement error (also called noise)

1. is zero on average (so it does not affect the average of the measured variable) and
2. is independent of all other relevant variables, including the error-free variable.

Examples

- Recording errors
  - E.g., due to mistakes in entering data
- Reporting errors in surveys or administrative data
  - If they are random around the true quantities

Classical Measurement Error

- Classical measurement error in the dependent (y or left-hand-side) variable
  - is not expected to affect the regression coefficients.
- Classical measurement error in the explanatory (x or right-hand-side) variable
  - will affect the regression coefficients.

Classical measurement error in the dependent variable (y)

Compare the slope of model with an error-free dependent variable to the slope of the same regression where y is measured with error.

$$y = y^* + e$$
$$y^E = \alpha^* + \beta^* x^*$$
$$y^E = \alpha + \beta x$$

The slope coefficient in the first regression with error-free dependent variable is

$$\beta^* = \frac{Cov\left[y^*, x\right]}{Var\left[x\right]}$$

The coefficient in the second regression with a dependent variable measured with classical error is

$$\beta = \frac{Cov\left[y, x\right]}{Var\left[x\right]}$$

Classical measurement error in the dependent variable (y)

The two are equal because the measurement error is not correlated with all relevant variables, including x so that $Cov\left[e,x\right]=0$

$$\beta=\frac{Cov\left[y,x\right]}{Var\left[x\right]}=\frac{Cov\left[\left(y^{*}+e\right),x\right]}{Var\left[x\right]}=\frac{Cov\left[y^{*},x\right]+Cov\left[e,x\right]}{Var\left[x\right]}=\frac{Cov\left[y^{*},x\right]}{Var\left[x\right]}=\beta^{*}$$

- Classical measurement error in the dependent (LHS) variable makes the slope coefficient unchanged because the expected value of the error-ridden y is the same as the expected value of the error-free y.
- **Consequence**: classical measurement error in the dependent variable is not expected to affect the regression coefficients.

Classical measurement error in the explanatory variable (x)

Compare the slope of model with an error-free explanatory variable to the slope of the same regression where **x** is measured with error.

$$x = x^* + e$$

$$y^E = \alpha^* + \beta^* x^*$$

$$y^E = \alpha + \beta x$$

The slope coefficient with error-free explanatory variable is

$$\beta^* = \frac{Cov\left[y, x^*\right]}{Var\left[x^*\right]}$$

The coefficient with a explanatory variable measured with classical error is

$$\beta = \frac{Cov\left[y, x\right]}{Var\left[x\right]}$$

Classical measurement error in the explanatory variable (x)

- Derivation

$$\beta = \frac{Cov\left[y,x\right]}{Var\left[x\right]} = \frac{Cov\left[y,(x^{*}+e)\right]}{Var\left[x^{*}+e\right]} = \frac{Cov\left[y,x^{*}\right]+Cov\left[y,e\right]}{Var\left[x^{*}\right]+Var\left[e\right]} = \frac{Cov\left[y,x^{*}\right]}{Var\left[x^{*}\right]+Var\left[e\right]}$$

$$= \frac{Cov\left[y,x^{*}\right]}{Var\left[x^{*}\right]}\frac{Var\left[x^{*}\right]}{Var\left[x^{*}\right]+Var\left[e\right]}$$

$$= \beta^{*}\frac{Var\left[x^{*}\right]}{Var\left[x^{*}\right]+Var\left[e\right]}$$

- So the attenuation bias occurs because the error inflates the variance in the explanatory (RHS) variable.

Classical measurement error in the explanatory variable (x)

- Slope coefficients are different in the presence of classical measurement error in the explanatory variable.
  - The slope coefficient in the regression with an error-ridden explanatory ((x)) variable is smaller in absolute value than the slope coefficient in the corresponding regression with an error-free explanatory variable.

$$\beta = \beta^* \frac{Var\left[x^*\right]}{Var\left[x^*\right] + Var\left[e\right]}$$

  - The sign of the two slopes is the same
  - But the magnitudes differ.
- **Attenuation bias**: the slope is attenuated because of the error in the x variable. The larger the **noise-to-signal ratio** the stronger the attenuation.
- Consequence: **on average β^* is closer to zero than it should be**.

Classical measurement error in the explanatory variable (x)

- Attenuation bias in the slope coefficient:

$$\beta = \beta^* \frac{Var\left[x^*\right]}{Var\left[x^*\right] + Var\left[e\right]}$$

  - So β is smaller in absolute value than β^*
- As a consequence α^* is also biased

$$\alpha^* = \bar{y} - \beta^* \bar{x}$$

- If one is biased the other one should be biased, too
  - The bias can go either way
  - If β^* is smaller α^* is larger
  - And vice versa, if β^* is larger α^* is smaller

Classical measurement error in the explanatory variable (x)

- Without measurement error,

$$\alpha^* = \bar{y} - \beta^* \overline{x^*}$$

- With measurement error,

$$\alpha = \bar{y} - \beta \bar{x}$$

- Classical measurement error leaves expected values (averages) unchanged so we can expect

$$\bar{x} = \overline{x^*}$$

Both regressions go through the same $(\bar{x}, \bar{y})$ point. Can derive that the difference in the two intercepts:

$$\alpha = \bar{y} - \beta \bar{x} = \bar{y} - \beta \bar{x}^* = \bar{y} - \beta \overline{x^*} + \beta^* \overline{x^*} - \beta^* \overline{x^*} = \alpha^* + (\beta^* - \beta)\, \overline{x^*}$$

$$= \alpha^* + \left(\beta^* - \beta^* \frac{Var\left[x^*\right]}{Var\left[x^*\right] + Var\left[e\right]} \right) \overline{x^*} = \alpha^* + \beta^* \frac{Var\left[e\right]}{Var\left[x^*\right] + Var\left[e\right]} \overline{x^*}$$

Classical measurement error in the explanatory variables (x)

- Noise to signal ratio is

$$\frac{Var\left[e\right]}{Var\left[x^{*}\right]}$$

- When the noise-to-signal ratio is low
  - we may safely ignore the problem.
  - this happens often when
    - when we are confident that recording errors are at not important
    - when our data has an aggregate variable estimated from very large samples.
- When the noise-to-signal ratio is substantial
  - we may be better of assessing its consequences.

Intuition for the Attenuation Bias

- Classical measurement error induces extra variation in the explanatory variable
- Beta will be underestimated
  - closer to zero / smaller in absolute value
- Consequence:
  - When we compare two observations that are different in x by one unit, the true difference in x^* is likely less than one unit
  - Therefore we should expect smaller difference in y associated with differences in x, than with differences in x^*
- Most often you only speculate about classic measurement error.
  - Looking at how is data collected
  - Infer from what you learn about the process, sampling

Extra: non-classical measurement error

- In real-life data measurement error in variables may or may not be classical
  - Very often, it isn't
  - Variables measured with error may be less dispersed (non-zero mean)
- Measurement error may be related to variables of interest
  - E.g., the share of expenditures missing from records (e.g., credit card records) may be larger for poorer people that spend on different kinds of things
  - This often means that modelling needs to be redesigned
- Non-classical measurement error has consequences that are different

Consequences

- Most variables in economic and social data are measured with noise. So what is the practical consequence of knowing the potential bias?
- Estimate magnitude - affects regression estimates
- Look for the source, think about it's nature and consider impact
- Super relevant issue for data collection, data quality

Classical measurement error summary

1. Classical measurement error in the dependent (y) variable is not expected to affect the regression coefficients.
2. Classical measurement error in the explanatory (x) variable will affect the regression coefficients.
3. In particular, the estimated beta will be closer to zero than it would be without measurement error.
4. Almost all variables are measured with error. Need to think about consequences.

Hotel ratings and measurement error

- In this case study, we will try to understand how measurement error in hotel rating may be investigated with its impact somewhat understood.
- Let's investigate another association: price and customer rating. The price comparison website publishes the averages of ratings customers gave to each hotel.
- Ratings vary between 1 and 5, 5 being excellent – measure of quality: hotels
- Show an association between price and quality.
- However, the measure of customer rating is an average calculated from individual evaluations. That is a noisy measure of hotel quality
  - Fundamentally, quality is more than rating
  - Technical reason - too few ratings to express quality - personal experience too large

Hotel ratings and measurement error

- The data includes the number of ratings that were used to calculate average customer ratings.
- If classical measurement error plays a role, it should play a larger role for hotels with few ratings than for hotels with many ratings.

- Three groups: few, medium, many. Focus few vs many
- Few ratings – less than 100 ratings (77 hotels, with 57 ratings each on average).
- Many ratings – more than 200 ratings (72 hotels, with 417 ratings each on average).
- Average customer rating is rather similar (3.94 and 4.20). Standard deviation of the average customer ratings – lot larger among hotels with few ratings (0.42 versus 0.26).

Hotel ratings and measurement error

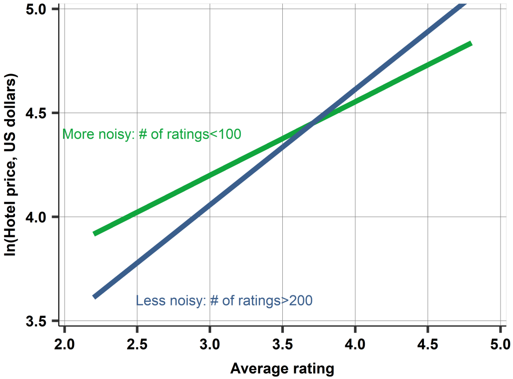
- We regressed (the log of) hotel price (y) on average ratings (x) separately for hotels with few ratings (less than 100) and hotels with many ratings (more than 200).
- If there is classical measurement error in average ratings, the error should be more prevalent among hotels with few ratings, and so the regression line should be...

Hotel ratings and measurement error

- We regressed (the log of) hotel price (y) on average ratings (x) separately for hotels with few ratings (less than 100) and hotels with many ratings (more than 200).
- If there is classical measurement error in average ratings, the error should be more prevalent among hotels with few ratings, and so the regression line should be... ... flatter for few ratings

Hotel ratings and measurement error

- Log hotel price and average customer ratings.
- Hotels with noisier measure of ratings (# ratings < 100)
- Hotels with less noisy measure (# ratings > 200)



Hotel ratings and measurement error

- That is indeed what we find. The first slope coefficient is 0.35; the second one is 0.55
- flatter, less positive slope and higher intercept among hotels with few ratings.
- There appears to be substantial measurement error in average customer ratings among hotels where that average rating is based on a few customers' reports.
- Thus, we can expect a regression with average customer ratings on the right-hand-side to produce an attenuated slope.
- Should we do anything about that? And if yes, what?

Hotel ratings and measurement error

- That is indeed what we find. The first slope coefficient is 0.35; the second one is 0.55
- flatter, less positive slope and higher intercept among hotels with few ratings.
- There appears to be substantial measurement error in average customer ratings among hotels where that average rating is based on a few customers' reports.
- Thus, we can expect a regression with average customer ratings on the right-hand-side to produce an attenuated slope.
- Should we do anything about that? And if yes, what?
- If we are interested in the effect of ratings on prices, this is clearly an issue. Discard hotels with less than a minimum number of reviews (maybe 10 or 20 or 50 or 100 - depends on sample size)