08. Complicated patterns and messy data

Agoston Reguly

Data Analysis 2: Regression analysis

2020

Functional form

- Relationships between y and x are often complicated!
- When and why care about the shape of a regression?
- How can we capture function form better?
  - This class is about transforming variables in a simple linear regression.

Motivation

- Interested in the pattern of association between life expectancy in a country and how rich that country is.
  - Uncovering that pattern is interesting for many reasons: discovery and learning from data.
- 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.

Functional form - linear approximation

- Linear regression – linear approximation to a regression of unknown shape:

$$y^E = f(x) \approx \alpha + \beta x$$

.

- Modify the regression to better characterize the nonlinear pattern if,
  - we want to make a prediction or analyze residuals - better fit
  - we want to go beyond the average pattern of association - good reason for complicated patterns
  - all we care about is the average pattern of association, but the linear regression gives a bad approximation to that - linear approximation is bad
- Not care
  - if all we care about is the average pattern of association,
  - if linear regression is good approximation to the average pattern

Functional form - types

There are many types of non-linearities!

- Linearity is one special cases of functional forms.
- We are covering the most commonly used transformations:
  - Ln of natural log transformation
  - Piecewise linear splines
  - Polynomials - quadratic form
  - Ratios

Functional form - decision

- Non-parametric methods great to get functional form, but no parameters.
    - Hard to interpret
- Need model functional form for interpretation! Implications:
    - Simplify the original pattern
    - Make assumption/restriction on the functional form
    - Accept that it will be far from perfect
- Many options how to choose! Decisions is needed:
    - Use theory to pick a model
    - Use statistical reasons (e.g. fit)
    - 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 (GDP increased by 3%)
- In cross-sectional data usually 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 we *approximate relative differences*.
  - Log differences works because differences in natural logs approximate percentage differences!

Logarithmic transformation - interpretation

- $ln(x)$ = 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 – percentages!

  Claim:

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

- The difference between the natural log of two numbers is approximately the relative difference between the two for small differences.

Logarithmic transformation - derivation

- From calculus we know:

$$\lim_{x \to x_0} \frac{ln(x) - ln(x_0)}{x - x_0} = \frac{1}{x_0}$$

- By definition it means a small change in x or $\Delta x = x - x_0$. Manipulating the equation, we get:

$$\lim_{\Delta x \to 0} ln(x_0 + \Delta x) - ln(x_0) = \lim_{\Delta x \to 0} \frac{\Delta x}{x_0}$$

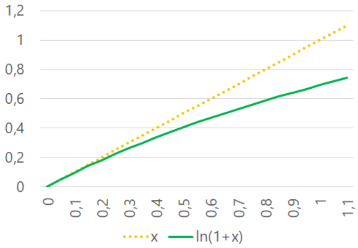
- If Δx is not converging to 0, this is an approximation of percentage changes.

$$ln(x_0 + \Delta x) - ln(x_0) \approx \frac{\Delta x}{x_0}$$

- Numerical examples ($x_0 = 1$):
  - $\Delta x = 0.01$ or 1% larger: $\ln(1{+}0.01) = \ln(1.01) = 0.0099 \approx 0.01$
  - $\Delta x = 0.1$ or 10% larger: $\ln(1{+}0.1) = \ln(1.1) = 0.095 \approx 0.1$

Log approximation: what is considered small?

- Log differences are good approximations for small relative differences!
- When Δx is considered small?
  - Rule of thumb: 0.3 (30% difference) or smaller
- But for larger x, there is a considerable difference,
  - A log difference of +1.0 corresponds to a +170 percentage point difference
  - A log difference of -1.0 corresponds to a -63% percentage point difference
- In case of large differences you may have to calculate percentage change by hand



When to take logs?

- Comparison makes mores sense in relative terms
  - Percentage differences
- Variable is positive value
  - There are some tricks to deal with 0s and negative numbers, but these are not so robust techniques.
- Most important examples:
  - Prices
  - Sales, turnover, GDP
  - Population, employment
  - Capital stock, inventories
- You may take the log for y or x or both!
  - These yield different models!

Interpreting parameters of regressions with log variables

$ln(y)^E = \alpha + \beta x_i$ - ‘log-level’ regression

- 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.

Interpreting parameters of regressions with log variables

$ln(y)^E = \alpha + \beta x_i$ - 'log-level' regression

- 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-log' regression

- 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.

Interpreting parameters of regressions with log variables

$ln(y)^E = \alpha + \beta x_i$ - 'log-level' regression

- 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-log' regression

- 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-log' regression

- 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.

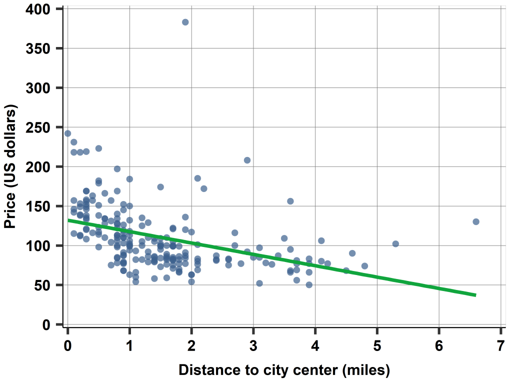
Interpreting parameters of regressions with log variables

- Precise interpretation is key
- The interpretation of the slope (and the intercept) coefficient(s) differs in each case!

- Often verbal comparison is made about a 10% difference in x if using level-log or log-log regression.

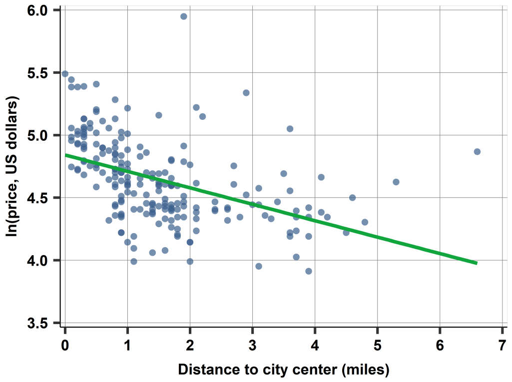
Hotel price-distance regression and functional form

- $price_i = 132.02 - 14.41 * distance_i$
- Issue ?



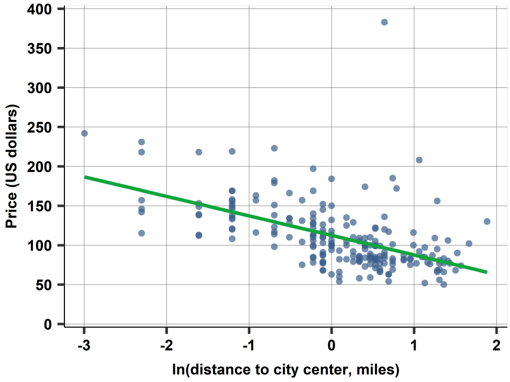
Hotel price-distance regression and functional form - log-level

- $\ln(price_i) = 4.84 - 0.13 * distance_i$
- 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



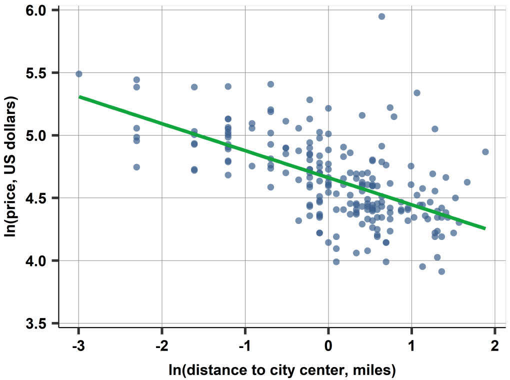
Hotel price-distance regression and functional form - level-log

- $price_i = 116.29 - 28.30 * \ln(distance_i)$
- We now make comparisons in terms percentage difference in distance
  - This transformation focuses on the lower and upper part of the domain in x: smaller values have even smaller log-values, while large values become closer to the average value.



Hotel price-distance regression and functional form - log-log

- $\ln(price_i) = 4.70 - 0.25 * \ln(distance_i)$
- Comparisons relative terms for both price and distance



Comparing different models

Table: Hotel price and distance regressions

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

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 2.477 US dollars less expensive, on average.
- ln(price) - ln(distance): hotels that are 10 percent farther away from the city center are 2.2 percent less expensive, on average.

To Take log or Not to Take log - substantive reason

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, number of death due to covid
  - 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
  - E.g. number of guests in a hotel, grade for a course

To Take log or Not to Take log - statistical reason

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 (log makes it worse...)
- 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 reasons, focus on the interpretation and magnitude of the slope coefficient and go with the most reasonable setup.

Comparing different models - model choice

Table: Hotel price and distance regressions

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

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

Model choice - substantive reasoning

- It depends on the goal of the analysis!
- Prices
  - 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.
  - Or we are after relative differences - we do not mind about the magnitude that we are paying, we only need the best deal.
- Distance
  - Distance makes more sense in miles than in relative terms – given our purpose is to find a *relatively* cheap hotel.

Model choice - statistical reasoning

- Visual inspection
  - Log price models capture patterns better, this could be preferred.
- Compare fit measure (R^2)
  - Level-level and level-log regression: R-squared of the level-log regression is higher, suggesting a better fit.
  - Log-level and log-log regression: R-squared of the log-log regression is higher, suggesting a better fit.
- Should not compare R-squared of two regressions with *different dependent variables* – compares fit in different units!

Model choice - statistical reasoning

- Visual inspection
  - Log price models capture patterns better, this could be preferred.
- Compare fit measure (R^2)
  - Level-level and level-log regression: R-squared of the level-log regression is higher, suggesting a better fit.
  - Log-level and log-log regression: R-squared of the log-log regression is higher, suggesting a better fit.
- Should not compare R-squared of two regressions with *different dependent variables* – compares fit in different units!
- Final verdict:
  - log-log probably the best choice:
    - can interpret in a meaningful way and
    - gives good prediction as this is the goal!
    - Note: prediction with log dependent variable is tricky.

Piecewise Linear Splines

分段函数

- 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 has built-in routines calculate the rest.

Piecewise Linear Splines - formula

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

$$y^E = \alpha_1 + (\beta_1 x)\mathbb{1}_{x<k_1} + (\alpha_2 + \beta_2 x)\mathbb{1}_{k_1 \leq x < k_2} + \\ \ldots + (\alpha_{m-1} + \beta_{m-1} x)\mathbb{1}_{k_{m-2} \leq x < k_{m-1}} + (\alpha_m + \beta_m x)\mathbb{1}_{x \geq k_{m-1}}$$

Piecewise Linear Splines - interpretaton

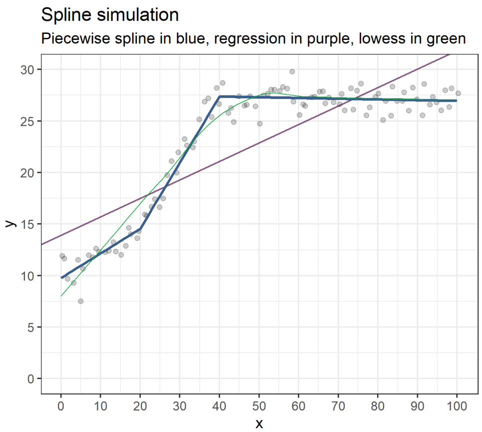
$$y^E = \alpha_1 + (\beta_1 x)\mathbb{1}_{x<k_1} + \ldots + (\alpha_j + \beta_j x)\mathbb{1}_{k_{j-1}\leq x<k_j} \ldots + (\alpha_m + \beta_m x)\mathbb{1}_{x\geq k_{m-1}}$$
$$j = 2, \ldots, m-1$$

Interpretation of the most important parameters:

- α_1 : average y when x is zero, if $k_1 > 0$ (Otherwise: $\alpha_1 + \alpha_j$, where $k_{j-1} \leq 0 < k_j$)
- 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_j : When comparing observations with x values between k_{j-1} and k_j, y is β_j 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.

Simulation for piecewise linear splines

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



Overview of piecewise linear spline

- 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.
- You can make it more complicated:
  - Quadratic, cubic or B-splines $\rightarrow$ rather a non-parametric approximation: interpretation-fit trade-off
  - Example: term-structure modelling (y: zero-coupon interest rate, x: maturity time) cubic spline is used. Link

Polynomials

- Quadratic function of the explanatory variable
  - Allow for a smooth change in the slope
  - Without any further decision from the analyst
- Technically: quadratic function is not a linear function (a parabola, not a line)
  - Handles only nonlinearity, which can be captured by a parabola.
  - Less flexible than a piecewise linear spline, but easier interpretation!

$$y^E = \alpha + \beta_1 x + \beta_2 x^2$$

- Can have higher order polynomials, in practice you may use cubic specification: $y^E = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
- General case

$$y^E = \alpha + \beta_1 x + \beta_2 x^2 + \ldots \beta_n x^n$$

Quadratic form - interpretation I.

$$y^E = \alpha + \beta_1 x + \beta_2 x^2$$

- α is average y when $x = 0$,
- β_1 has no interpretation in itself,
- β_2 shows whether the parabola is
  - U-shaped or convex (if $\beta_2 > 0$)
  - inverted U-shaped or concave (if $\beta_2 < 0$).

Quadratic form - interpretation II.

$$y^E = \alpha + \beta_1 x + \beta_2 x^2$$

- Difference in y, when x is different. This leads to (partial) derivative of y^E w.r.t. x,

$$\frac{\partial y^E}{\partial x} = \beta_1 + 2\beta_2 x$$

- the slope is *different for different values of* x
  - Compare two observations, j and k, that are different in x, by one unit: $x_k = x_j + 1$.
- Units which are one unit larger than x_j are higher by $\beta_1 + 2\beta_2 x_j$ in y on average.
  - Usually we compare to the average of x: $x_j = \bar{x}$.
    - Units which are one unit larger than the average of x are higher by $\gamma = \beta_1 + 2\beta_2\bar{x}$ in y on average.
- Why, higher order polynomial is rather non-parametric method?

Ratios

- Ratios of variables - normalization of totals
  - For many comparisons you need to use ratios to compare the same thing!
- Most often, *per capita measures*: GDP/capita, revenues/employee, sales/shop.
- For ratios, you can take logs as well.
  - Bear in mind the interpretation changes as well!
  - log of a ratio equals the difference of the two logs.

$$ln(GDP/Pop) = ln(GDP) - ln(Pop)$$

Weighted Regression

- Instead of transforming your variables, you may change your estimation method.
- Weighted regression:
  - By pre-specified weights (often an other variable) it weights the importance of each observation.
  - Weights can be manually given as well.
- Weighted OLS estimates:

$$\arg\min_{\alpha,\beta} \sum_{i=1}^{N} w_i \left(y_i - \alpha - \beta x_i\right)^2$$

- It weights the errors by w, thus both y and x are weighted
  - Interpretation changes, sometimes it is straightforward, other times it is not.
    - If w is a meaningful variable - change the interpretation
- Good method for robustness check.

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 $\rightarrow$ 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 year 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

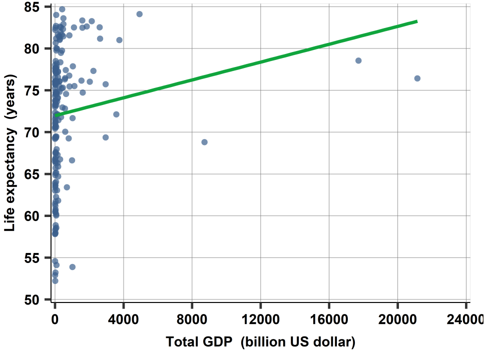


Figure: Life expectancy and total GDP

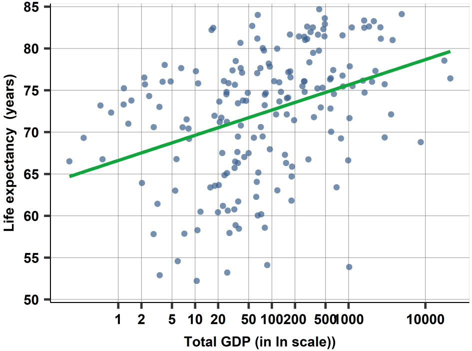


Figure: Life expectancy and total GDP w ln scale

Life expectancy and GDP per capita

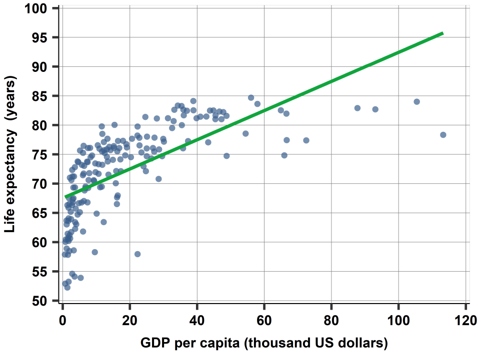


Figure: Life expectancy and GDP/capita

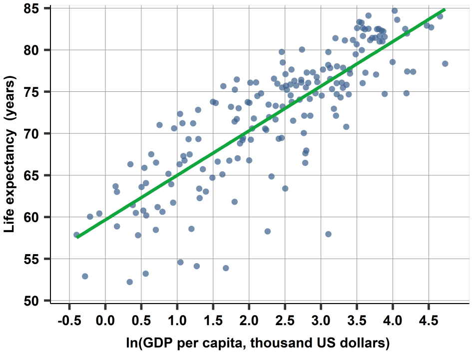


Figure: Life expectancy and ln GDP/capita

Life expectancy and GDP per capita - log representations

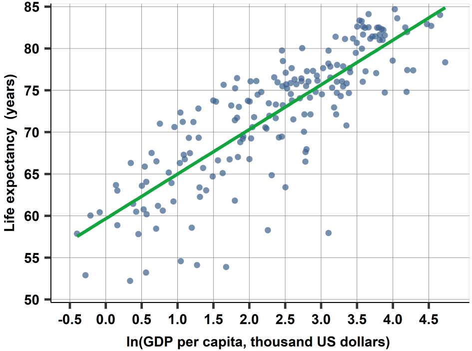


Figure: Life expectancy and ln GDP/capita

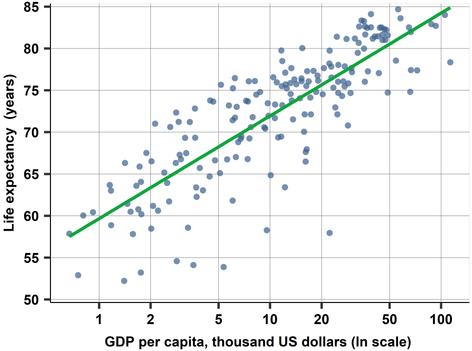


Figure: Life expectancy and GDP/capita w ln scale

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 11-17 years minus)
- Countries with the longest lives given their income include Vietnam, Nicaragua, and Lebanon (7 years 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 ratio and log
- 3 new models:
  - Splines - capture flattening of at the end: knot at 50 GDP/capita.
  - Quadratic - similar purpose
  - Weighted OLS - address different importance to observations.

Life expectancy and GDP per capita - PLS and quadratic

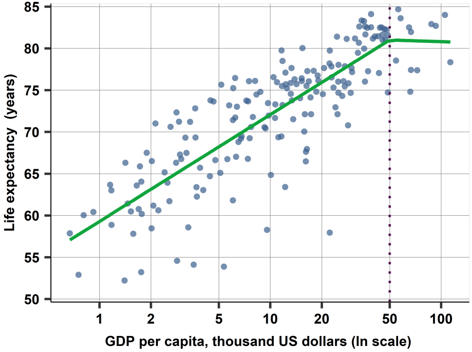


Figure: Ln GDP/capita - spline

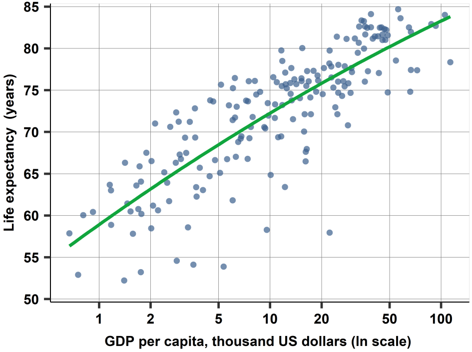
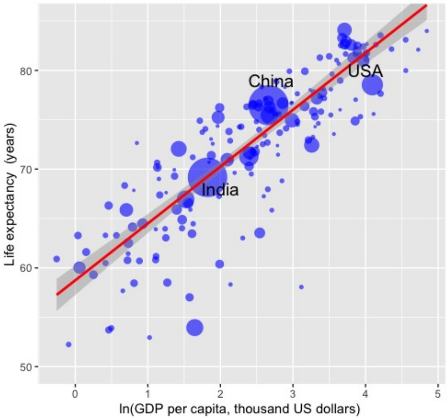


Figure: Ln GDP/capita - quadratic

Life expectancy and GDP per capita - WOLS

- Regress life expectancy on GDP per capita, weighted by population.
  - $w_i = population_i / \sum_{i=1}^{N} population_i$
- Slope parameter: 5.8 (simple level-log is 5.3)
- Interpretation: people who live in countries with 10 percent higher GDP per capita live, on average 0.6 years longer.



Life expectancy and income

- Fit of the models (R^2):
  - Life expectancy regressed on level GDP/capita - 0.44
  - Life expectancy regressed on ln(GDP/capita) - 0.68
  - Life expectancy regressed on ln(GDP/capita) weighted by population - 0.68
  - Life expectancy regressed on splines of ln(GDP/capita) - 0.69
  - Life expectancy regressed on level GDP/capita and its square - 0.69
- 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
  - Note: when modelling you want to have as simple model as possible with the best fit. Over-complicated models usually a bad idea.

Which functional form to choose? - guidelines

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.
  - Keep in mind - simpler the better!

Which functional form to choose? - practice

To uncover and include a potentially nonlinear pattern in the regression analysis:

1. Check the distribution of your main variables (y and x)
2. 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.
3. Think and check what would be the best transformation!
   3.1 Choose one or more ways to incorporate those features into a linear regression (transformed variables, piecewise linear spline, quadratic, etc.).
   3.2 Remember for some variables log transformation or using ratios is not meaningful!
4. 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 $\Longrightarrow$ DA1

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 $\Longrightarrow$ DA1

There are potential issues with the variable(s) itself:

- Some observations are influential
  - How to handle them? Drop them? Probably not but depends on the context.
- Variables measured with (systematic) error
  - When does it lead to biased estimates?

Extreme values vs 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 or 10,000 US dollars
- 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
  - Wage regressed on size: small tech companies with large wages.

Extreme values and influential observations

- What to do with them?
- Depends on why they are extreme
  - If by mistake: may want to drop them ($ 1000+)
  - If by nature: don't want to drop them (other hotel)
  - Grey zone: patterns work differently for them for substantive reasons
    extra value in y, maintain; in x, it may not ok to drop
    - 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.
- Do not drop extreme observations by y, only if you are certain it is a mistake.

Classical Measurement Error

- You want to measure a variable which is not so easy to measure:
  - Quality of the hotels
  - Inflation
  - Other latent variables with proxy measures
- Usually these miss-measurement are present due to
  - Recording errors (mistakes in entering data)
  - Reporting errors in surveys (you do not know the exact value) or administrative data (miss-reporting)
- 'Classical measurement error':
  - One of the most common and 'best' behaving problem – but a problem.
  - It needs to satisfy the followings:
    - It is zero on average (so it does not affect the average of the measured variable)
    - (Mean) independent of all variables.
- There are many other 'non-classical' measurement error, which cause problems in modelling.

Is measurement error in variables a problem?

It depends...

- Prediction: your are predicting *with* the errors - not a particular problem, but need to be addressed when generalizing.
- Association:
  - Interested in the estimated coefficient value (not just the sign)
    - Spending and income; price and distance, etc.
  - Depends on whether it is in y or in x

Is measurement error in variables a problem?

Solution?

- Often cannot do anything about it!
  - The problem is with data collection/how data is generated.
  - Exceptions:
    - you have a variable which is correlated with the error term → use multivariate regression.
    - You have a time-series dataset → some fancy method in state-space modelling.
- If cannot do anything, what is the consequence is of such errors
  - Does measurement error make a difference in the model parameter estimates? larger or smaller
  - Do we expect parameters (such as OLS coefficient) to be different from what they would be without the measurement error?

Two cases for classical Measurement Error

- Classical measurement error in the dependent (y or left-hand-side) variable
  - is not expected to affect the regression coefficients. I measurement error
- Classical measurement error in the explanatory (x or right-hand-side) variable
  - will affect the regression coefficients. I measurement error
- We are covering how to mathematically approach this problem.
  - Show general way of thinking about *any* type of measurement error.
  - There are lot of format for measurement errors, you may want to have an idea whether it affect your regression coefficient(s):
    - If yes we call it 'biased' parameter(s).

Classical measurement error in the dependent variable (y) - I.

It means:

$$y = y^* + e$$

Where, $E[e] = 0$ and e is mean independent from x and y $(E[e \mid x, y] = 0)$.
Reminder if e is mean independent from x, y, then $Cov[e, x] = 0$, $Cov[e, y] = 0$)

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

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

Slope coefficients in the two regression are:

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

Classical measurement error in the dependent variable (y) - II.

Compering the two coefficients we show 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.
  - But it lowers R^2 by increasing the disturbance term $u = u^{*} + e$.

Classical measurement error in the explanatory variable (x) - I.

It means:

$$x = x^* + e$$

Where, $E[e] = 0$ and e is mean independent from y and x, thus $Cov[e, y] = 0$, $Cov[e, x] = 0$.

Again let us compare the slopes of the two models, where x^* is the error-free explanatory variable x is measured with error.

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

The slope coefficients for the two models are similar to the previous ones:

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

Classical measurement error in the explanatory variable (x) - II.

Let us relate β to β^*:

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

- $\beta \neq \beta^*$, thus it is a 'bias'.
- We call it the '*attenuation bias*', while the error inflates the variance in the explanatory (RHS) variable and makes β closer to zero.

Classical measurement error in the explanatory variable (x) - III.

- 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.
- Consequence: on average β^* is closer to zero than it should be.

Effect of a biased parameter

- 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 usually biased too
  - The value of intercept goes in the opposite direction!
  - β is closer to zero, α is further away from α^*

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:

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

Noise to signal ratio

- 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.

Review for classical measurement errors

- Classical measurement error in *dependent variable*
  - No bias, but nosier results.
- Classical measurement error in *explanatory variable*
  - Larger variation of x
  - Beta will be biased - attenuation bias
    - 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. (Larger variation in x)
    - Therefore we should expect smaller difference in y associated with differences in x, than with differences in the true variable x^*. (Biased parameter)
    - You can interpret your result as a lower (higher) bound of the true parameter if your sign is positive (negative).
- Most often you only speculate about classic measurement error.
  - Looking at how is data collected
  - Infer from what you learn about the process, sampling

Non-classical measurement error

- In real-life data measurement error in variables may or may not be classical
  - Very often, it is not
  - Variables measured with error may be less dispersed and have 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 which affects regression estimates.
- Look for the source, think about it's nature and consider impact.
- Super relevant issue for data collection, data quality!

Thinking frame for measurement errors

1. Define measurement free variable and variable with measurement error.
2. Define the nature of the measurement error.
   2.1 Expectation, variance (other moments) and relation to other variables.
3. Write your true models with and without measurement error.
4. Compare the estimated parameter of interest in the two models
   4.1 Are they the same or different?
      4.1.1 If same – good to go.
      4.1.2 Different – can you characterize the difference? Can you fix it? Does it affect your interpretation?

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.
- New linear regression specification: price (y) and customer rating (x). 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 for hotels
- Show an association between price and a proxy for quality.
- The measure of customer rating is an average calculated from individual evaluations. That is a noisy measure of hotel quality
  - Theoretical reason - rating is a noisy proxy for quality
  - Technical reason - too few ratings gives large measurement errors for quality

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.

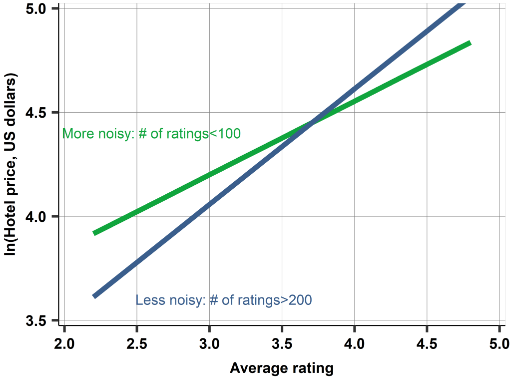
- Three groups: few, medium, many. Focus few vs many
- Few ratings – less than 100 ratings (77 hotels, with 57 ratings each on average).
- Many ratings – more than 200 ratings (72 hotels, with 417 ratings each on average).
- Average customer rating is rather similar (3.94 and 4.20). Standard deviation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