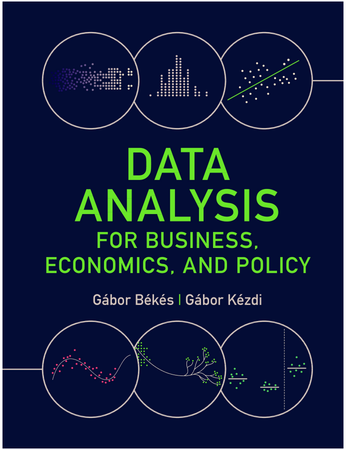
08. Complicated patterns and messy data

Davide Del Prete

Data Analysis 2: Regression analysis

2024

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- gabors-data-analysis.com
    - Download all data and code: gabors-data-analysis.com/data-and-code/
- This slideshow is for Chapter 08

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

- 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.

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: ln transformation

- Frequent nonlinear patterns better approximated with y or x transformed by taking relative differences:
- 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:

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

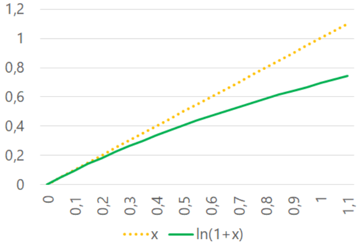
- If Δx is not converging to 0, this is an approximation of percentage changes.

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

- 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.

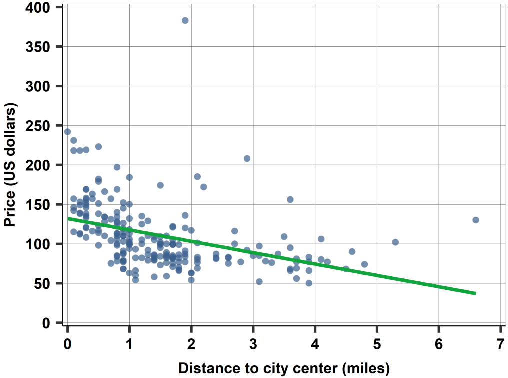
Interpreting parameters of regressions with log variables

- 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.

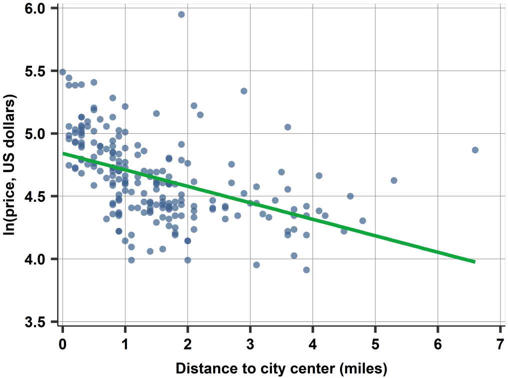
Hotel price-distance regression and functional form

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



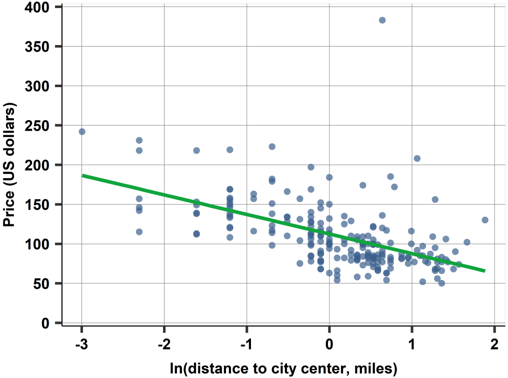
Hotel price-distance regression and functional form - log-level

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



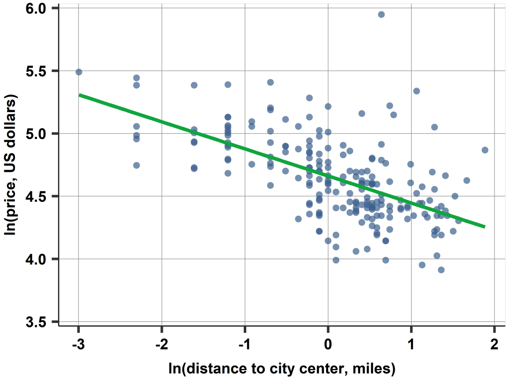
Hotel price-distance regression and functional form - level-log

- $price_i = 116.29 - 28.30 * \ln(distance_i)$
- 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, 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

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
- Don't take logs if variable is likely affected in additive ways

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

- Most often the substantive *and* statistical arguments are aligned

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, 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

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

- 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.

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:

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

Piecewise Linear Splines - interpretaton

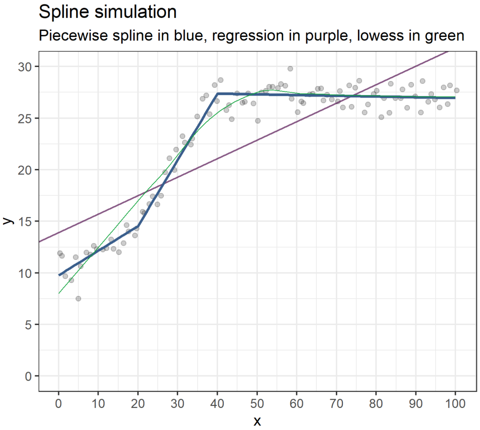
$$\begin{aligned} y^E =& \alpha_1 + (\beta_1 x)\mathbb{1}_{x<k_1} + \ldots + (\alpha_j + \beta_j x)\mathbb{1}_{k_{j-1} \leq x < k_j} \ldots + (\alpha_m + \beta_m x)\mathbb{1}_{x \geq k_{m-1}} \\ j =& 2, \ldots, m-1 \end{aligned}$$

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 → 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?

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

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:
- 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

Extreme values and influential observations

- What to do with them?
- Depends on why they are extreme
  - If by mistake: may want to drop them
  - If by nature: don't want to drop them
  - Grey zone: patterns work differently for them for substantive reasons
    - 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.

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 from 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 predicting or generalizing.
- Association:
    - Interested in the estimated coefficient value (not just the sign)

Solution?

- Often cannot do anything about it!
    - The problem is with data collection/how data is generated.
- If cannot do anything, what is the consequence of such errors:
    - Does measurement error make a difference in the model parameter estimates?

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.
- Classical measurement error in the explanatory (x or right-hand-side) variable
  - will affect the regression coefficients.

- 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 affects 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 any relevant variable(s), 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^*:

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

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

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

- $\beta \neq \beta^*$, 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 parameter is biased the other one usually biased too
  - The value of intercept changes 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:

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

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

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 sampling process.

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!

- Have a look at the case study on hotels in Chapter08!

Summary take-away

- Regression – functional form selection can help better capture relationships
- Several real life data problems may lead to estimation problems.