III: Simple regression, complicated patterns and messy data

Alexandra Avdeenko

Data Analysis

2021

Literature

- Slides 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 lecture: Chapters 7 and 8

Topics of last week: Generalization and Testing

- Inference
- Repeated samples
- The Confidence Interval
- The bootstrap SE
- External validity
- Hypothesis
- The t-test
- Making a decision
- Multiple test

Motivation

- Spend a night in Vienna and you want to find a good deal for your stay.
- Travel time to the city center is rather important.
- Looking for a good deal: as low a price as possible and as close to the city center as possible.
- Collect data on suitable hotels

Topics for today: Simple Regression and Messy Data

Regression basics
Linear regression
Residuals
OLS Modeling
Causation
Messy data
Functional form
Log transformation
Take log?
Splines, polynomials
Selection
Messy data
Measurement error

Introduction

- Regression is the most widely used method of comparison in data analysis.
- Simple regression analysis amounts to comparing average values of a dependent variable (y) for observations that are different in the explanatory variable (x).
- Simple regression: *comparing conditional means*.
- Doing so uncovers the pattern of association between y and x. What you use for y and for x is important and not inter-changeable!

Section 1

Regression basics

Regression

- Simple regression analysis uncovers mean-dependence between two variables.
  - It amounts to comparing average values of one variable, called the dependent variable (y) for observations that are different in the other variable, the explanatory variable (x).
- Multiple regression analysis involves more variables -> later.

Regression - uses

- Discovering patterns of association between variables is often a good starting point even if our question is more ambitious.
- Causal analysis: uncovering the *effect* of one variable on another variable. Concerned with a parameter.
- Predictive analysis: what to expect of a y variable (long-run polls, hotel prices) for various values of another x variable (immediate polls, distance to the city center). Concerned with predicted value of y using x.

Regression - names and notation

- Regression analysis is a method that uncovers the average value of a variable y for different values of another variable x.

$$E[y|x] = f(x) \tag{1}$$

We use a simpler shorthand notation

$$y^E = f(x) \tag{2}$$

- dependent variable or left-hand-side variable, or simply the y variable,
- explanatory variable, right-hand-side variable, or simply the x variable
- "regress y on x," or "run a regression of y on x"= do simple regression analysis with y as the dependent variable and x as the explanatory variable.

Regression - type of patterns

Regression may find

- Linear patterns: positive (negative) association - average y tends to be higher (lower) at higher values of x.
- Non-linear patterns: association may be non-monotonic - y tends to be higher for higher values of x in a certain range of the x variable and lower for higher values of x in another range of the x variable
- No association or relationship

Non-parametric and parametric regression

- Non-parametric regressions describe the $y^E = f(x)$ pattern without imposing a specific functional form on f.
  - Let the data dictate what that function looks like, at least approximately.
  - Can spot (any) patterns well
- Parametric regressions impose a functional form on f. Parametric examples include:
  - linear functions: $f(x) = a + bx$;
  - exponential functions: $f(x) = ax^b$;
  - quadratic functions: $f(x) = a + bx + cx^2$,
  - or any functions which have parameters of a, b, c, etc.
  - Restrictive, but they produce readily interpretable numbers.

Non-parametric regression

- Non-parametric regressions come (also) in various forms.
- When x has few values and there are many observations in the data, the best and most intuitive non-parametric regression for $y^E = f(x)$ shows average y for each and every value of x.
- There is no functional form imposed on f here.
  - The most straightforward example if you have ordered variables.
  - For example, Hotels: average price of hotels with the same numbers of stars and compare these averages = non-parametric regression analysis.

Non-parametric regression: bins

- With many x values - two ways to do non-parametric regression analysis: bins and smoothing.
- Bins - based on grouped values of x
  - Bins are disjoint categories (no overlap) that span the entire range of x (no gaps).
  - Many ways to create bins - equal size, equal number of observations per bin, or bins defined by analyst.

Non-parametric regression: lowess (loess)

- Produce “smooth" graph - both continuous and has no kink at any point.
- also called smoothed conditional means plots = non-parametric regression shows conditional means, smoothed to get a better image.
- Lowess = most widely used non-parametric regression methods that produce a smooth graph.
  - *locally weighted scatterplot smoothing* (sometimes abbreviated as “loess").
- A smooth curve fit around a bin scatter.

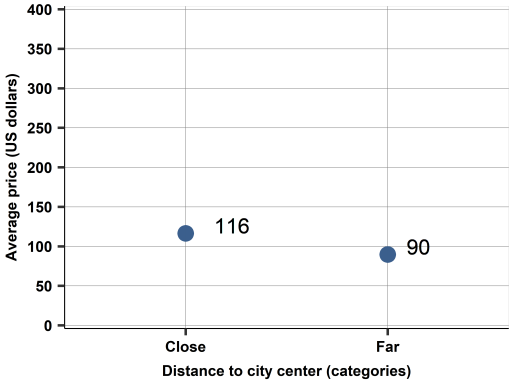
Non-parametric regression: lowess (loess)

- Smooth non-parametric regression methods, including lowess, do not produce numbers that would summarize the $y^E = f(x)$ pattern.
- Provide a value y^E for each of the particular x values that occur in the data, as well as for all x values in-between.
- Graph – we interpret these graphs in qualitative, not quantitative ways.
- They can show interesting shapes in the pattern, such as non-monotonic parts, steeper and flatter parts, etc.
- Great way to find relationship patterns

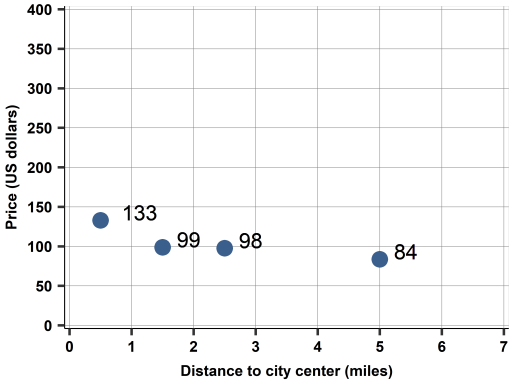
Case Study: Finding a good deal among hotels

- We look at Vienna hotels for a 2017 November weekday.
- we focus on hotels that are (i) in Vienna actual,(ii) not too far from the center, (iii) classified as hotels, (iv) 3-4 stars, and (v) have no extremely high price classified as error.
- There are 428 hotel prices for that weekday in Vienna, our focused sample has $N = 207$ observations.

Case Study: Finding a good deal among hotels

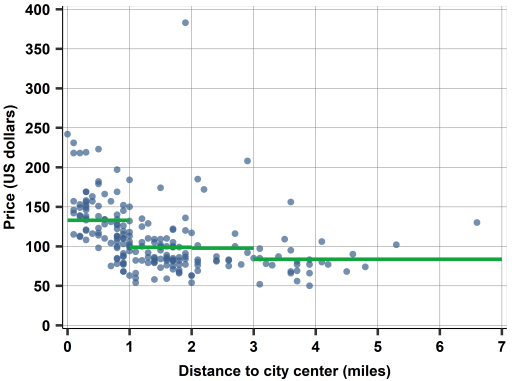


Bin scatter non-parametric regression, 2 bins

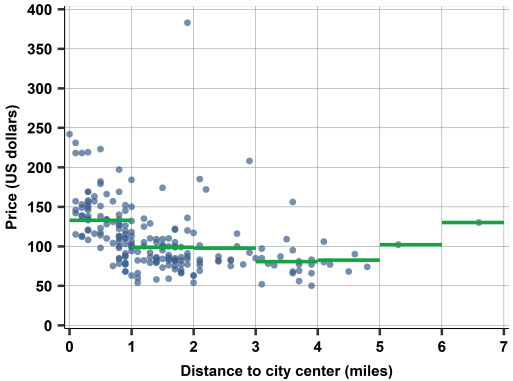


Bin scatter non-parametric regression, 4 bins

Case Study: Finding a good deal among hotels



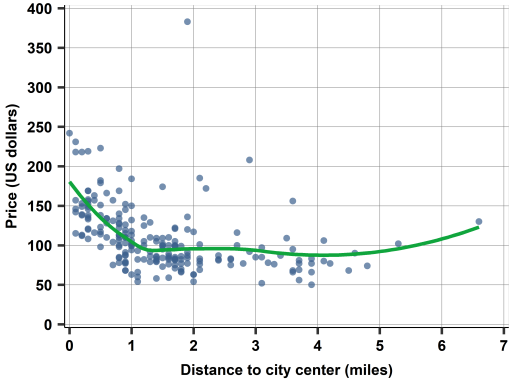
Scatter and bin scatter non-parametric regression, 4 bins



Scatter and bin scatter non-parametric regression, 7 bins

Case Study: Finding a good deal among hotels

- lowess non-parametric regression, together with the scatterplot.
- bandwidth selected by software is 0.8 miles.
- The smooth non-parametric regression retains some aspects of previous bin scatter – a smoother version of the corresponding non-parametric regression with disjoint bins of similar width.



Section 2

Linear regression

Linear regression

Linear regression is the most widely used method in data analysis.

- imposes linearity of the function f in $y^E = f(x)$.
- Linear functions have two parameters, also called coefficients: the intercept and the slope.

$$y^E = \alpha + \beta x \tag{3}$$

- Linearity in terms of its coefficients.
  - can have any function, including any nonlinear function, of the original variables themselves
- linear regression is a line through the $x - y$ scatterplot.
  - This line is the best-fitting line one can draw through the scatterplot.
  - It is the best fit in the sense that it is the line that is closest to all points of the scatterplot.

Linear regression - assumption vs approximation

- *Linearity as an assumption*:
  - assume that the regression function is linear in its coefficients.
- *Linearity as an approximation*.
  - Whatever the form of the $y^E = f(x)$ relationship, the $y^E = \alpha + \beta x$ regression fits a line through it.
  - This may or may not be a good approximation.
  - By fitting a line we approximate the average slope of the $y^E = f(x)$ curve.

Linear regression coefficients

Coefficients have a clear interpretation – based on comparing conditional means.

$$E[y|x] = \alpha + \beta x$$

Two coefficients:

- intercept: α = average value of y when x is zero:
- $E[y|x = 0] = \alpha + \beta \times 0 = \alpha$.

- slope: β. = expected difference in y corresponding to a one unit difference in x.
- $E[y|x = x_0 + 1] - E[y|x_0] = (\alpha + \beta \times (x_0 + 1)) - (\alpha + \beta \times x_0) = \beta$.

Regression - slope coefficient

- slope: β = expected difference in y corresponding to a one unit difference in x.
- *y is higher, on average, by β for observations with a one-unit higher value of x.*
- Comparing two observations that differ in x by one unit, we expect y to be β higher for the observation with one unit higher x.

- Avoid “decrease/increase" – not right, unless time series or causal relationship only

Regression: binary explanatory

Simplest case:

- x is a binary variable, zero or one.
- α is the average value of y when x is zero ($E[y|x = 0] = \alpha$).
- β is the difference in average y between observations with $x = 1$ and observations with $x = 0$
  - $E[y|x = 1] - E[y|x = 0] = \alpha + \beta \times 1 - \alpha + \beta \times 0 = \beta$.
  - The average value of y when x is one is $E[y|x = 1] = \alpha + \beta$.
- Graphically, the regression line of linear regression goes through two points: average y when x is zero (α) and average y when x is one ($\alpha + \beta$).

Regression coefficient formula

Notation:

- General coefficients are α and β.
- Calculated *estimates* - $\hat{\alpha}$ and $\hat{\beta}$ (use data and calculate the statistic)
- The slope coefficient formula is

$$\hat{\beta} = \frac{Cov[x, y]}{Var[x]} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Slope coefficient formula is normalized version of the covariance between x and y.
  - The slope measures the covariance relative to the variation in x.
  - That is why the slope can be interpreted as differences in average y corresponding to differences in x.

Regression coefficient formula

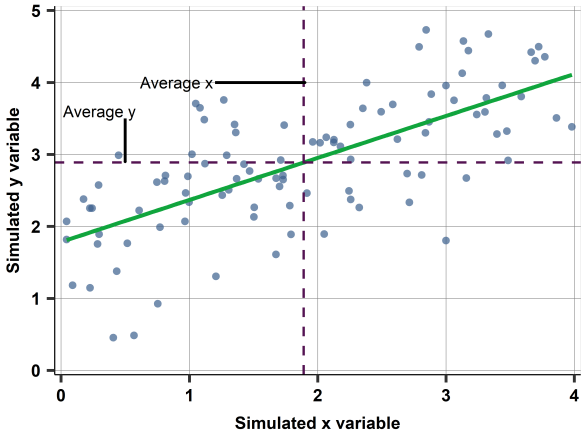
- The intercept – average y minus average x multiplied by the estimated slope $\hat{\beta}$.

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

- The formula of the intercept reveals that the regression line always goes through the point of average x and average y.
- Note, you can manipulate and get: $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$.

Ordinary Least Squares (OLS)

- OLS gives the best-fitting linear regression line.
- A vertical line at the average value of x and a horizontal line at the average value of y. The regression line goes through the point of average x and average y.



More on OLS

- The idea underlying OLS is to find the values of the intercept and slope parameters that make the regression line fit the scatterplot 'best'.
- OLS method finds the values of the coefficients of the linear regression that minimize the sum of squares of the difference between actual y values and their values implied by the regression, $\hat{\alpha} + \hat{\beta}x$.

$$min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- For this minimization problem, we can use calculus to give $\hat{\alpha}$ and $\hat{\beta}$, the values for α and β that give the minimum.

Section 3

Residuals

Predicted values

- The predicted value of the dependent variable = best guess for its average value if we know the value of the explanatory variable, using our model.
- The predicted value can be calculated from the regression for any x.
- The predicted values of the dependent variable are the points of the regression line itself.
- The predicted value of dependent variable y is denoted as $\hat{y}$.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

- Predicted value can be calculated for any model of y.

Residuals

- The residual is the difference between the actual value of the dependent variable for an observation and its predicted value :

$$e_i = y_i - \hat{y}_i, \qquad \text{where} \qquad \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

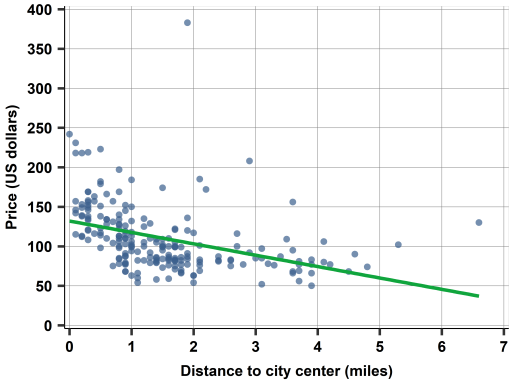
- The residual is meaningful only for actual observation. It compares observation i's difference for actual and predicted value.
- The residual is the vertical distance between the scatterplot point and the regression line.
  - For points above the regression line the residual is positive.
  - For points below the regression line the residual is negative.

Some further comments on residuals

- The residual may be important on its own right.
- Residuals sum up to zero if a linear regression is fitted by OLS.
  - It is a property of OLS: $E[e_i] = 0$
  - Remember: we minimized the *sum* of squared errors...

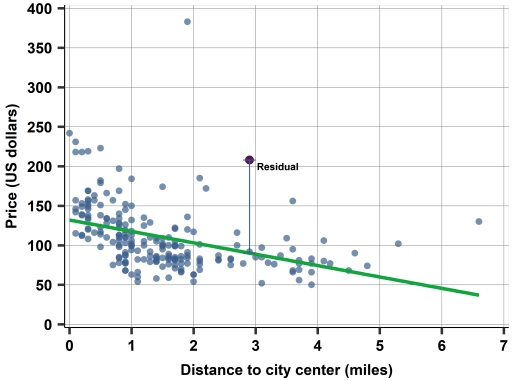
Case Study: Finding a good deal among hotels

- The linear regression of hotel prices (in $) on distance (in miles) produces an intercept of 133 and a slope -14.
- The intercept is 133, suggesting that the average price of hotels right in the city center is $ 133.
- The slope of the linear regression is -14. Hotels that are 1 mile further away from the city center are, on average, $ 14 cheaper in our data.



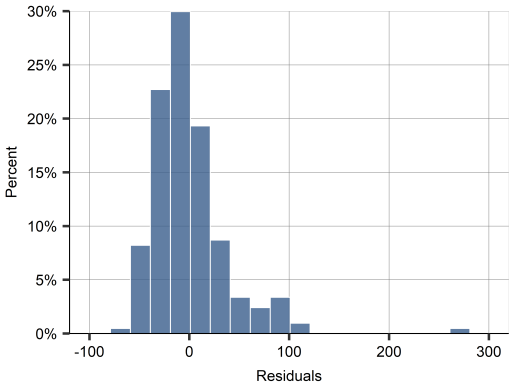
Case Study: Finding a good deal among hotels

- Residual is vertical distance
- Positive residual shown here - price is above what predicted by regression line



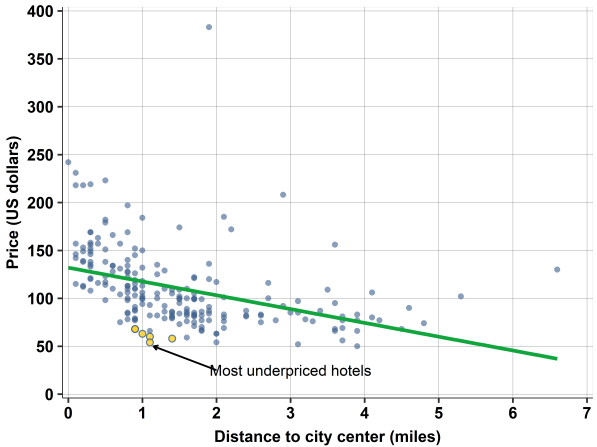
Case Study: Finding a good deal among hotels

- Can look at residuals from linear regressions
- Centered around zero
- Both positive and negative



Case Study: Finding a good deal among hotels

- If linear regression is accepted model for prices
- Draw a scatterplot with regression line
- With the model you can capture the over and underpriced hotels



Case Study: Finding a good deal among hotels

A list of the hotels with the five lowest value of the residual.

No.	Hotel_id	Distance	Price	Predicted price	Residual
1	22080	1.1	54	116.17	-62.17
2	21912	1.1	60	116.17	-56.17
3	22152	1	63	117.61	-54.61
4	22408	1.4	58	111.85	-53.85
5	22090	0.9	68	119.05	-51.05

- Bear in mind, we can (and will) do better - this is not the best model for price prediction.
  - Non-linear pattern
  - Functional form
  - Taking into account differences beyond distance

Section 4

OLS Modeling

Model fit - R^2

- *Fit of a regression* captures how predicted values compare to the actual values.
- *R-squared* (R^2) – how much of the variation in y is captured by the regression, and how much is left for residual variation

$$R^2 = \frac{Var[\hat{y}]}{Var[y]} = 1 - \frac{Var[e]}{Var[y]} \tag{4}$$

where, $Var[\hat{y}] = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, and $Var[e] = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2$.

- Decomposition of the overall variation in y into variation in predicted values "explained by the regression") and residual variation ("not explained by the regression"):

$$Var[y] = Var[\hat{y}] + Var[e] \tag{5}$$

Model fit - R^2

- R-squared (or R^2) can be defined for both parametric and non-parametric regressions.
- Any kind of regression produces predicted $\hat{y}$ values, and all we need to compute R^2 is its variance compared to the variance of y.
- The value of R-squared is always between zero and one.
- R-squared is zero, if the predicted values are just the average of the observed outcome $\hat{y}_i = \bar{y}_i, \forall i$.

Model fit - how to use R^2

- R-squared may help in choosing between different versions of regression for the *same data*.
  - Choose between regressions with different functional forms
  - Predictions are *likely* to be better with high R^2
- R-squared matters less when the goal is to characterize the association between y and x

Correlation and linear regression

- Linear regression is closely related to correlation.
- Remember, the OLS formula for the slope

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

- In contrast with the correlation coefficient, its values can be anything. Furthermore y and x are *not interchangeable*.
- Covariance and correlation coefficient can be substituted to get $\hat{\beta}$:

$$\hat{\beta} = Corr[x, y]\frac{Std[y]}{Std[x]}$$

- Covariance, the correlation coefficient, and the slope of a linear regression capture similar information: the degree of association between the two variables.

Correlation and R^2 in linear regression

- R-squared of the simple linear regression is the square of the correlation coefficient.

$$R^2 = (Corr[y, x])^2$$

- So the R-squared is yet another measure of the association between the two variables.

Section 5

Causation

Regression and causation

- Be very careful to use neutral language, not talk about causation, when doing simple linear regression!
- Think back to sources of variation in x
  - Do you control for variation in x? Or do you only observe them?
- Regression is a method of comparison: it compares observations that are different in variable x and shows corresponding average differences in variable y.
  - Regardless of the relation of the two variable.

Regression and causation - possible relations

- Slope of the $y^E = \alpha + \beta x$ regression is not zero in our data
- Several reasons, not mutually exclusive:
  - x causes y
  - y causes x
  - A third variable causes both x and y (or many such variables do):
- In reality if we have observational data, there is a mix of these relations.

Summary take-away

- Regression – method to compare average y across observations with different values of x.
- Non-parametric regressions (bin scatter, lowess) visualize complicated patterns of association between y and x, but no interpretable number.
- Linear regression – linear approximation of the average pattern of association y and x
- In $y^E = \alpha + \beta x$, β shows how much larger y is, on average, for observations with a one-unit larger x
- When β is not zero, one of three things (+ any combination) may be true:
  - x causes y
  - y causes x
  - a third variable causes both x and y.

- If you are to study more econometrics, advanced statistics - Go through textbook under the hood derivations sections!

Section 6

Messy data

Motivation: Complicated patterns and messy data

- 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.

Section 7

Functional form

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

Section 8

Log transformation

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 $ln(\textit{negative number})$ 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:

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

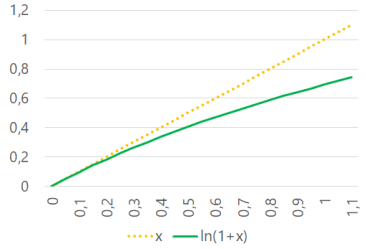
- If Δx is not converging to 0, this is an approximation of percentage changes.

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

- 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
- 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
- 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
- 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
- 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
- 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
- β: y is β percent higher on average for observations with one percent higher x.

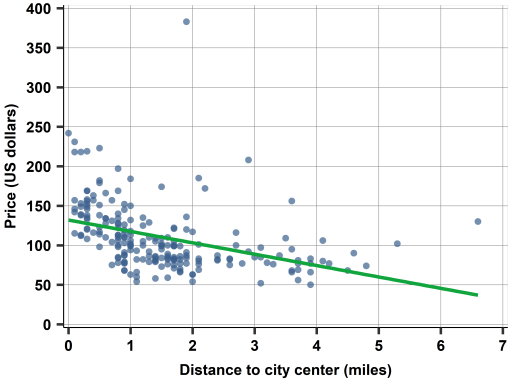
Interpreting parameters of regressions with log variables

- 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.

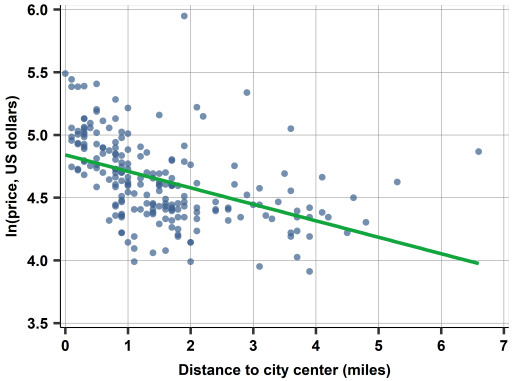
Hotel price-distance regression and functional form

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



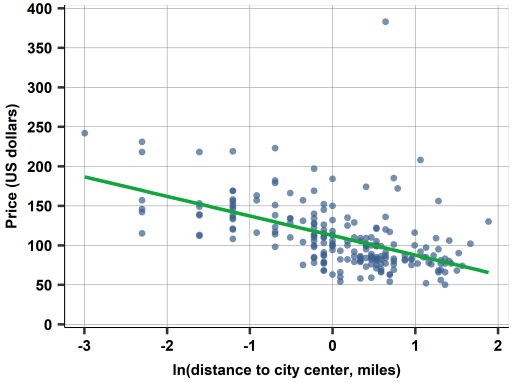
Hotel price-distance regression and functional form - log-level

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



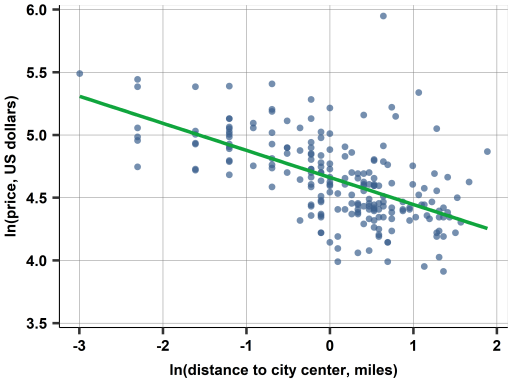
Hotel price-distance regression and functional form - level-log

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

- column (1) price-distance: hotels that are 1 mile farther away from the city center are 14 US dollars less expensive, on average.
- column (2) ln(price) - distance: hotels that are 1 mile farther away from the city center are 13 percent less expensive, on average.
- column (3) price - ln(distance): hotels that are 10 percent farther away from the city center are 2.477 US dollars less expensive, on average.
- column (4) ln(price) - ln(distance): hotels that are 10 percent farther away from the city center are 2.2 percent less expensive, on average.

Section 9

Take log?

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.

Section 10

Splines, polynomials

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.
  - 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:

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

Piecewise Linear Splines - interpretaton

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

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.

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?

Section 11

Selection

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*.

Section 12

Messy data

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.

Section 13

Measurement error

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 $Cov[e, x] = 0$

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

- 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:

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

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

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.