# 01. Simple regression and nonlinear patterns

Agoston Reguly

Data Analysis 2: Regression analysis

2024

# Motivation

- What's data analysis?
- We build some model to get answers to our questions.
- Define a problem
  - Collect data (manage, wrangle, clean, etc) <— DA1
- Learn about patterns
- Use information to help decision in business, politics, economic policy
- Regression analysis is basic tool to do that
- In the end: *"All models are wrong, but some are useful."* - George Box

# Case study 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, compare average prices for various distances from center.
- Look for hotels where price is cheap relative to what being close to the center would normally cost.

# 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!

# Chapter 7

# 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.
  - Related to density plots, set the bandwidth for smoothing
    - ‘Bias-variance trade-off’: wider bandwidth results in a smoother graph but may miss important details of the pattern (higher bias, smaller variance); narrower bandwidth produces a more rugged-looking graph (small bias, higher variance)

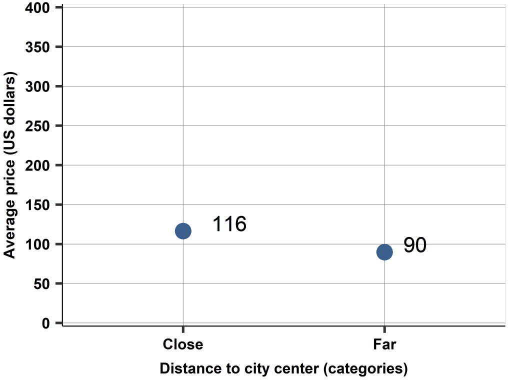
# Non-parametric regression: lowess (loess)

- 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

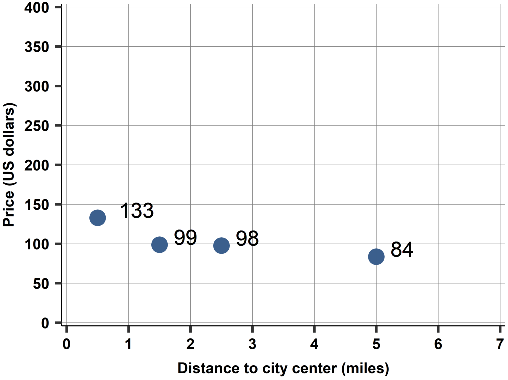
# Case Study: Finding a good deal among hotels

- 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

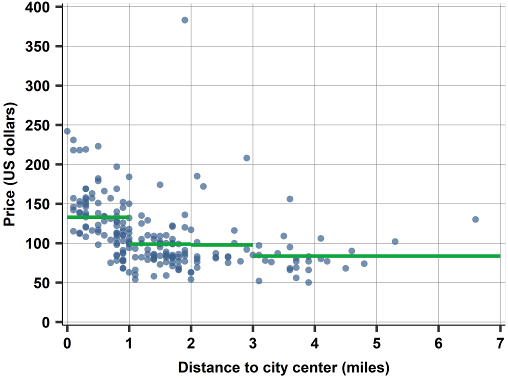



Bin scatter non-parametric regression, 2 bins

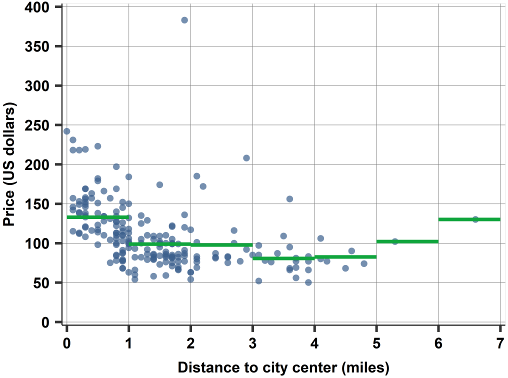



Bin scatter non-parametric regression, 4 bins

# Case Study: Finding a good deal among hotels



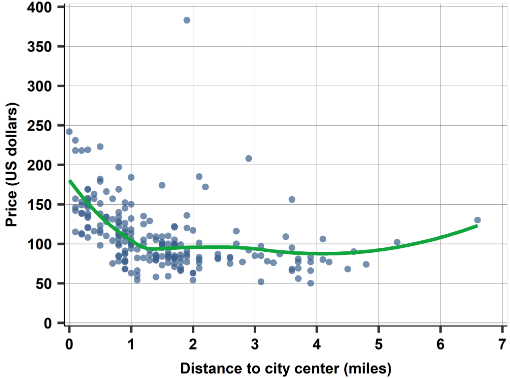

Scatter and bin scatter non-parametric regression, 4 bins




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.



# 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 (think of logarithms, squares, etc.).
- 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*:
  - by doing linear regression analysis we assume that the regression function is linear in its coefficients.
  - this may be true or not.
- *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: $\alpha$ = average value of $y$ when $x$ is zero:
- $E[y|x = 0] = \alpha + \beta \times 0 = \alpha$.

- slope: $\beta$. = 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: $\beta$ = expected difference in $y$ corresponding to a one unit difference in $x$.
- *$y$ is higher, on average, by $\beta$ 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 $\beta$ higher for the observation with one unit higher $x$.

- Be careful...
  - "decrease/increase" – not right, unless time series or causal relationship only
  - "effect" – not right, unless causal relationship
  - comparing conditional means – always true whether or not the more ambiguous interpretations are true

# Regression: binary explanatory

Simplest case:

- $x$ is a binary variable, zero or one.
- $\alpha$ is the average value of $y$ when $x$ is zero ($E[y|x = 0] = \alpha$).
- $\beta$ 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 ($\alpha$) and average $y$ when $x$ is one ($\alpha + \beta$).

# Regression coefficient formula

Notation:

- General coefficients are $\alpha$ and $\beta$.
- 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

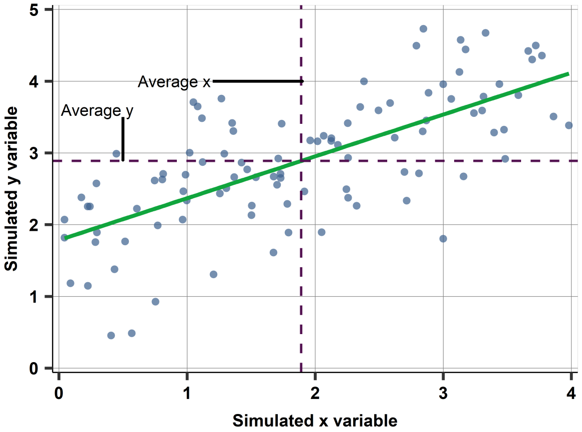
- The intercept – average $y$ minus average $x$ multiplied by the estimated slope $\hat{\beta}$.

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

- 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, $\hat{\alpha} + \hat{\beta}x$.

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

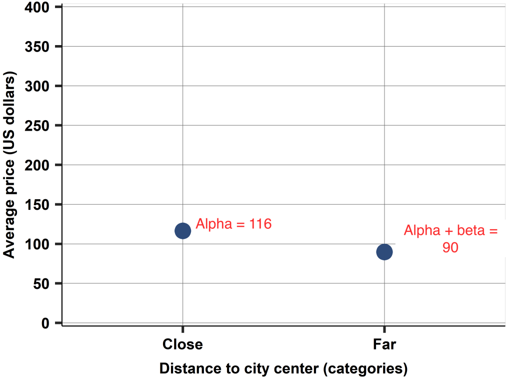
- For this minimization problem, we can use calculus to give $\hat{\alpha}$ and $\hat{\beta}$, the values for $\alpha$ and $\beta$ that give the minimum.
- HW: show the formula which minimize $\alpha, \beta$ and prove that this is indeed a minimum!

# Case Study: Finding a good deal among hotels - binary $x$

- When $x$ is binary: close = 0, far = 1, we have simple discrete conditioning for the expected value:

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

- where $\hat{\alpha} = 116$ is the average price for hotels that are close,
- and $\hat{\beta}$ is the *difference* between the average price for close vs far hotels.
  - $\hat{\beta} = 90 - 116 = -26$



# Case Study: Finding a good deal among hotels - continuous $x$ with 4 observation

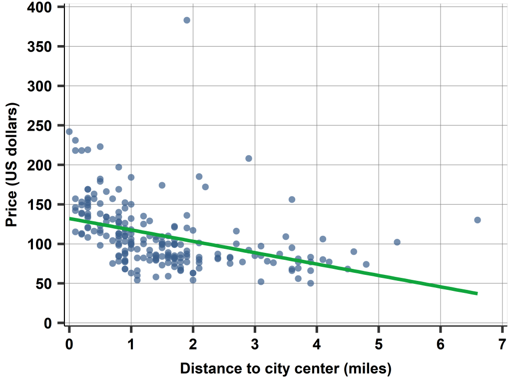
- Averages
  - price: $\bar{y} = 116.75 EUR$, distance: $\bar{x} = 2.65$ miles
- Variance and covariance:
  - Variance of distance: $\hat{V}[x] = 4.88$
  - Covariance of price and distance: $\widehat{Cov}[y, x] = -55.64$
- OLS estimators
  - $\hat{\beta} := \frac{\widehat{Cov}[y,x]}{\hat{V}[x]} = \frac{-55.64}{4.88} \approx -11.4$
  - $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 116.75 - (-11.4) * 2.65 \approx 147.23$
- To connect the dots, at $x = 2$ $y^E = 147.23 - 11.4 * 2 = 124.43$.

| hotel_id | price | distance |
|---|---|---|
| 22348 | 77 | 3.7 |
| 22125 | 104 | 0.7 |
| 22250 | 184 | 1.0 |
| 22349 | 102 | 5.3 |

Table: Randomly selected 4 hotels in Vienna

# Case Study: Finding a good deal among hotels - continuous $x$ with all data

- 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

- Compare linear model and non-parametric ones
- Linear is an average that fails to capture steep decline close to center
- Not bad approximation overall, but can be improved...

# 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 $y$.
  - Interpolation: predict *within* observed $x$ values - feasible if good model.
  - Extrapolation: predict *outside* observed $x$ values - adventurous, only if meaningful and have high external validity

# 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.
  - If we certain about our model: identifies observations that are special in that they have a dependent variable that is much higher or much lower than "it should be" as predicted by the regression.
  - If we are not certain in our model: how our predicted errors look like - can use it as a measure of model fit.
- 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 – 4 observations only

| hotel_id | price | distance | $\hat{y}_i$ | $e_i$ |
|---|---|---|---|---|
| 22348 | 77 | 3.7 | 105.07 | -28.07 |
| 22125 | 104 | 0.7 | 139.26 | -35.26 |
| 22250 | 184 | 1.0 | 135.84 | 48.16 |
| 22349 | 102 | 5.3 | 86.84 | 15.16 |

Table: Randomly selected 4 hotels in Vienna

- Remember, we have only 4 observation as a toy example.
- The estimated parameters are:
  - $\hat{\alpha} = 147.23$
  - $\hat{\beta} = -11.4$
- The predicted values are given by:

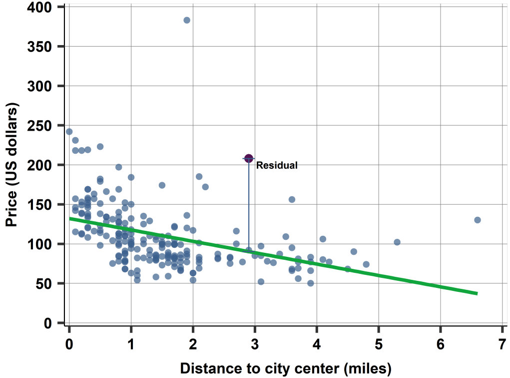
$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

- and the errors are:

$$e_i = y_i - \hat{y}_i$$

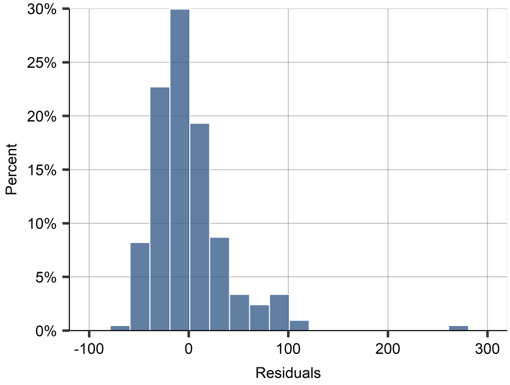
# Case Study: Finding a good deal among hotels

- The same applies for all the observations!
- Residual is vertical distance
- Positive residual shown here - price is above what predicted by regression line



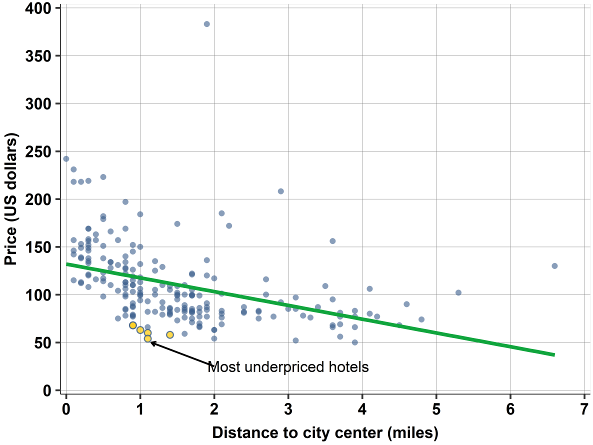
# Case Study: Finding a good deal among hotels

- 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

# 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 $\hat{y}_i = \bar{y}_i, \forall i$.
    - In linear regression, this corresponds to a slope of zero: the regression line is completely flat. $\beta = 0$ thus $y$ and $x$ are mean-independent.

# 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 $x$
  - We would like to understand how $x$ and $y$ are related and we want to describe this pattern with interpretable coefficients.
  - The regression that best approximates that pattern may have a high R-squared or a low R-squared.

# Case Study: Finding a good deal among hotels – 4 observations only

| hotel_id | price | distance | $\hat{y}_i$ | $e_i$ |
|---|---|---|---|---|
| 22348 | 77 | 3.7 | 105.07 | -28.07 |
| 22125 | 104 | 0.7 | 139.26 | -35.26 |
| 22250 | 184 | 1.0 | 135.84 | 48.16 |
| 22349 | 102 | 5.3 | 86.84 | 15.16 |

Table: Randomly selected 4 hotels in Vienna

- $R^2$ is a statistics using variance of $y$ and the predicted values $\hat{y}$.
  - $Var[\hat{y}] = 634.10$
  - $Var[y] = 2160.92$

$$R^2 = \frac{Var[\hat{y}]}{Var[y]} = \frac{634.10}{2169.92} = 0.29$$

# Case Study: Finding a good deal among hotels – all data

- Now, lets use all the data!

- Regression line and the pattern
- The R-squared of the regression is $0.16 = 16\%$.
  - This means that of the overall variation in hotel prices, 16% is explained by the linear regression with distance to the city center; the remaining 84% is left unexplained.
- 16% - good for cross-sectional regression with a single explanatory variable.
  - In any case it is the fit of the best-fitting line.

Chapter 8: 8.1, 8.2, 8.3, 8.4, 8.A1, 8.5, 8.13

# Functional form

- Relationships between y and x are often complicated!
- When and why care about the shape of a regression?
- How can we capture functional form better?
  - Can we do better than off-the-shelf 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 or log stands for natural log transformation- today
  - Ratios - today
  - Weighted OLS - today
  - Piecewise linear splines - next class
  - Polynomials - quadratic form - next class

# 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 are 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:
- 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!

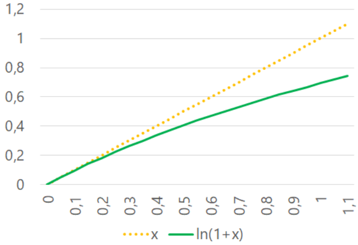
  Claim:

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

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

# Log approximation: what is considered small?

- Log differences are good approximations for small relative differences!
- When $\Delta x$ 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
- $\alpha$ is average ln(y) when x is zero. (Often meaningless.)
- $\beta$: 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
- $\alpha$ is average ln(y) when x is zero. (Often meaningless.)
- $\beta$: 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
- $\alpha$ is : average y when ln(x) is zero (and thus x is one).
- $\beta$: 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
- $\alpha$ is average ln(y) when x is zero. (Often meaningless.)
- $\beta$: 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
- $\alpha$ is : average y when ln(x) is zero (and thus x is one).
- $\beta$: 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

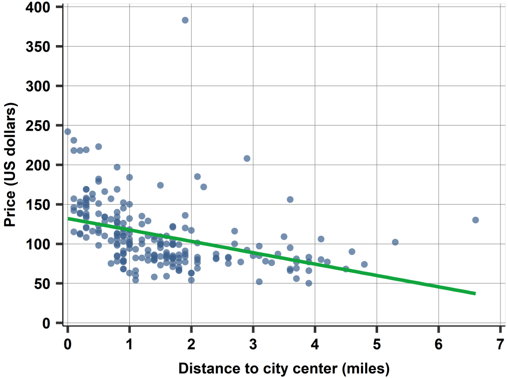
- log y, log x
- $\alpha$: is average $ln(y)$ when $ln(x)$ is zero. (Often meaningless.)
- $\beta$: y is $\beta$ percent higher on average for observations with one percent higher x.

# 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.

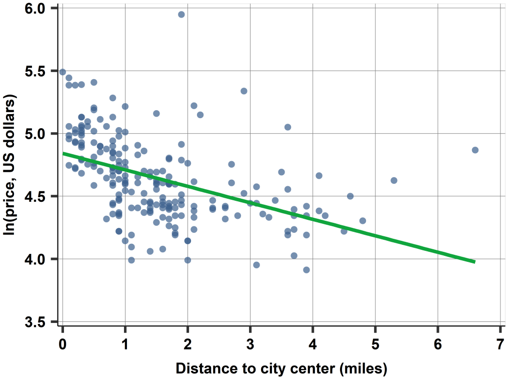
# Hotel price-distance regression and functional form

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



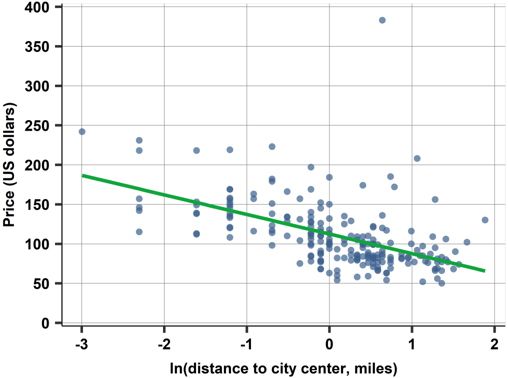
# Hotel price-distance regression and functional form - log-level

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



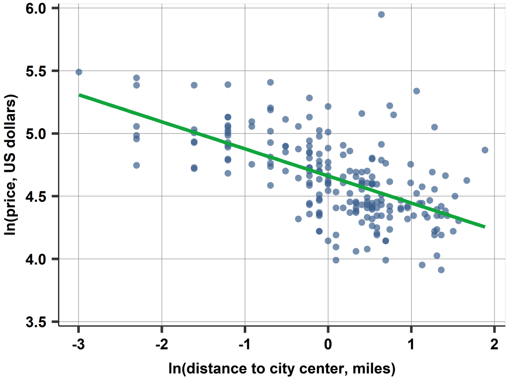
# Hotel price-distance regression and functional form - level-log

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

# Other transformations – 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
    - If $w$ is approx the same for all units - similar to simple linear regression, but add 'weighted by'.
- Good method for robustness check. – E.g., weight by ratings.

# 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 - variation in $x$

- The key is the source of variation in $x$ - the method will never do the causal claim.
- It is always the data that makes it possible to claim causal relationship. More precisely, how the data was collected, how variation in $x$ was provided.

Example: advertising ($x$) and sales ($y$)

- Observational data, collected from a firm and using regression –> no causal claim.
  - In holidays more people go shopping and firms are increasing their advertisements also in these days –> Sales are not increased by advertisement but because of holiday.
- If firm consciously experiments by allocating varying resources to advertising, in a random fashion, and keep track of sales. A regression of sales on the amount of advertising can uncover the effect of advertising here. (More in DA4)
  - Same method, but can do causal claim because of variation in advertisement.

# Regression and causation - possible relations

- Slope of the $y^E = \alpha + \beta x$ regression is not zero in our data ($\beta \neq 0$) and the linear regression captures the $y$–$x$ association reasonably well, one of three things – which are not mutually exclusive – may be true:
  - $x$ causes $y$:
    - If this is the single one thing behind the slope, it means that we can expect $y$ to increase by $\beta$ units if we were to increase $x$ by one unit.
  - $y$ causes $x$.:
    - If this is the single one thing behind the slope, it means that we can expect $x$ to increase if we were to increase $y$:
  - A third variable causes both $x$ and $y$ (or many such variables do):
    - If this is the single one thing behind the slope it means that we cannot expect $y$ to increase if we were to increase $x$ (or the other way around).
- In reality if we have observational data, there is a mix of these relations. E.g. if $y$ has an effect on $x$ and $x$ has an effect on $y$ we call it 'endogeneity'.
  - E.g. hotel ratings ($x$) and price ($y$)

# Regression and causation

- The proper interpretation of the slope is necessary regardless the data is observational or comes from a controlled experiment.
    - Safe way: A positive slope in a regression of sales on advertising, means that sales tend to be higher on average when advertising is higher.
- Instead of "correlation (regression) does not imply causation"–> we should not infer cause and effect from comparisons in observational data.
- Suggested approach is two steps:
    - First interpret precisely the object (correlation ot slope coefficient)
    - Conclude and discuss causal claims if any

# Case Study: Finding a good deal among hotels

- Level-level regression: slope is -14
- Does that mean that a longer distance causes hotels to be cheaper by that amount?

# Summary take-away I

- 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 $x$
- Classical setup is level-level regression:
  - In $y^E = \alpha + \beta x$, $\beta$ shows how much larger $y$ is, on average, for observations with a one-unit larger $x$
- But you may use ln-transformation for better interpretation or fit
  - level-log
  - log-level
  - log-log

# Summary take-away II

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

- It is best advised to use these linear regressions as a descriptive tool!
  - It shows the average pattern of association.
- Why? Because, when $\beta$ 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!

# Further insights

# Model fit - truth vs model

The 'true model', that we do not know:

$$y_i = f(x) + \varepsilon_i$$

Fit depends:

1. How well the particular version of the regression captures the actual function of $f(x)$
   - Can be helped by choice of model (parametric vs non-parametric, use of variables, functional form, ect.)
2. How far the realizations of $y_i$ are spread around the true functional form of $f$ due to $\varepsilon_i$

# 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.
- To show this equality holds, the trick is to substitute the numerator of R-squared and manipulate:

$$R^2 = \frac{Var[\hat{y}]}{Var[y]} = \frac{Var[\hat{\alpha} + \hat{\beta}x]}{Var[y]} = \frac{\hat{\beta}^2 Var[x]}{Var[y]} = \left(\hat{\beta}\frac{Std[x]}{Std[y]}\right)^2 = (Corr[y,x])^2$$

# Reverse regression

- One can change the variables, but the interpretation is going to change as well!

$$x^E = \gamma + \delta y$$

- The OLS estimator for the slope coefficient here is $\hat{\delta} = \frac{Cov[y,x]}{Var[y]}$.
- The OLS slopes of the original regression and the reverse regression are related:

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

  - Different, unless $Var[x] = Var[y]$,
  - but always have the same sign.
  - both are larger in magnitude the larger the covariance.
- $R^2$ for the simple linear regression and the reverse regression is the same.

# 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 $\Delta x$ 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: $\ln(1{+}0.1) = \ln(1.1) = 0.095 \approx 0.1$