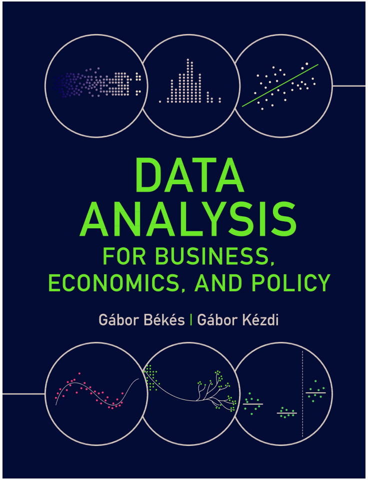
07. Simple regression

Gabor Bekes

Data Analysis 2: Regression analysis

2019

Slideshow for the Békés-Kézdi Data Analysis textbook



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

Motivation

- What's data analysis?
- We do not really know, but we know good data analysis (Roger Peng, Johns Hopkins)
- 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

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 that 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).
- Comparing conditional means
- Doing so uncovers the pattern of association between y and x.
- Regression is about comparing means.

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 -> week 3

Regression

- 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.
- **predictive analysis**: what to expect of a variable (long-run polls, hotel prices) for various values of another variable (immediate polls, distance to the city center).
- In both causal analysis and predictions we are often concerned with other variables that may exert influence.

Regression

- **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

Regression may find

- positive (negative) association - average y tends to be higher (lower) at higher values of x
- pattern of 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 / 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 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: $f(x) = a + bx + cx^2$, etc.
  - Functions have parameters a, b, c, etc.
  - Restrictive, but they produce readily interpretable numbers.

Non-parametric regression

- Non-parametric regressions come 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.
  - 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 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
    - wider bandwidth results in a smoother graph but may miss important details of the pattern.
    - narrower bandwidth produces a more rugged-looking graph

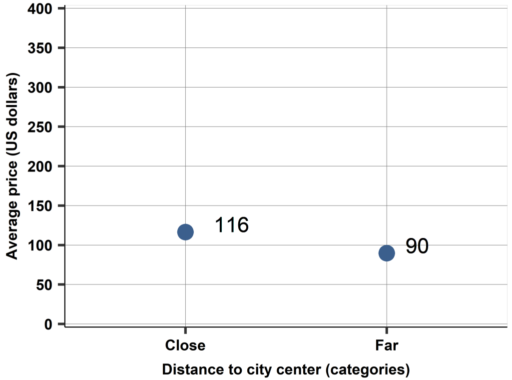
Non-parametric regression: lowess (loess)

- 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

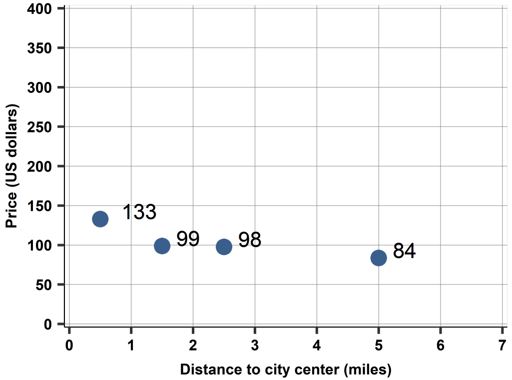
Case Study: Finding a good deal among hotels

- 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

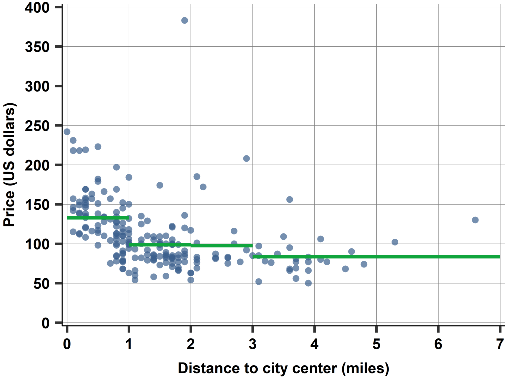


Bin scatter non-parametric regression, 2 bins

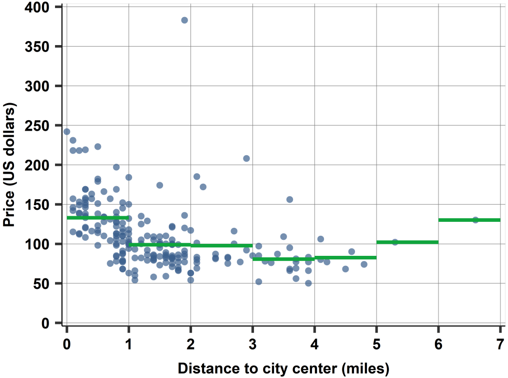


Bin scatter non-parametric regression, 4 bins

Case Study: Finding a good deal among hotels



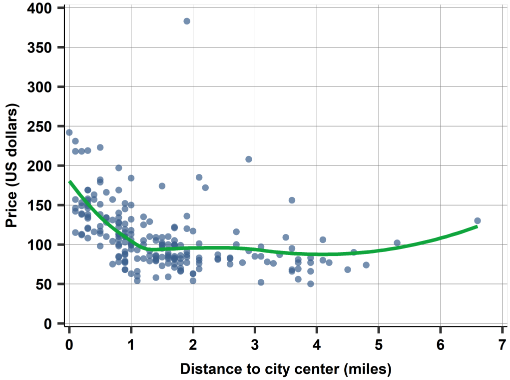
Scatter and bin scatter non-parametric regression, 4 bins



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.



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.

Regression

- **linearity as an assumption**:
  - by doing linear regression analysis we 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.
  - By fitting a line, linear regression approximates the average slope of the $y^E = f(x)$ curve.
- The average slope has an important interpretation: it is the difference in average y that corresponds to different values of x, averaged across the entire range of x in the data.

Regression coefficients

- Coefficients have a clear interpretation – based on comparing conditional means.
- $y^E = \alpha + \beta x$ has two coefficients:
- **intercept**: $\alpha =$ 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**: $\beta. =$ 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.

- 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 ambitious interpretations are true

Regression: binary explanatory

- 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

- Calculated from data - $\hat{\alpha}$ and $\hat{\beta}$ = **estimates** of the general coefficients α and β.
- 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

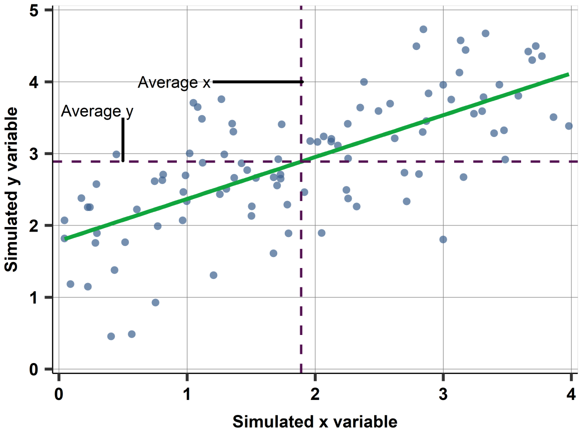
- The intercept – average y minus average x multiplied by the estimated slope $\hat{\beta}$.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{4}$$

- The formula of the intercept reveals that the regression line always goes through the point of average x and average y.
- $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$.
  - In linear regressions, the expected value of y for average x is indeed average y.

OLS

- Figure - scatterplot with the best-fitting linear regression found by OLS.
  - Artificial data
- 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.



Regression coefficient formula

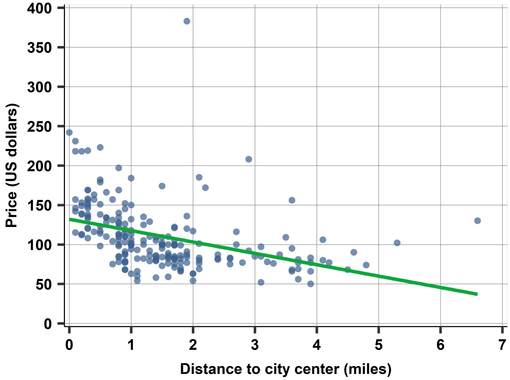
- The derivation of the formulae is called **Ordinary Least Squares** and is abbreviated as **OLS**
- The idea underlying OLS is to find the values of the intercept and slope parameters that make the regression line fit the scatterplot best.
- OLS method finds the values of the coefficients of the linear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

$$min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{5}$$

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.

Case Study: Finding a good deal among hotels

- The linear regression of hotel prices (in EUR) 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 EUR 133.
- The slope of the linear regression is -14. Hotels that are 1 mile further away from the city center are, on average, EUR 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

Predicted dependent variable and residuals

- The **predicted value** of the dependent variable = best guess for its average value if we know the value of the explanatory variable.
- The predicted value can be calculated from the regression for all x
- The predicted values of the dependent variable are the points of the regression line itself
- The predicted value of dependent variable y for observation i is denoted as $\hat{y}_i$.

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \tag{6}$$

- Non-parametric regressions

Predicted dependent variable and residuals

- The **predicted value** of the dependent variable = best guess for its average value if we know the value of the explanatory variable.
- The predicted value can be calculated from the regression for all x
- The predicted values of the dependent variable are the points of the regression line itself
- The predicted value of dependent variable y for observation i is denoted as $\hat{y}_i$.

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \tag{6}$$

- Non-parametric regressions
- Predicted dependent variables exist
  - Complete list of predicted values of the dependent variable for each value of the explanatory variable in the data.

Predicted dependent variable and residuals

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

$$e_i = y_i - \hat{y}_i. \tag{7}$$

- The residual for i = difference of two y values: the value of y for the observation (the y value of the scatterplot point) minus its predicted value $\hat{y}$
  - $\hat{y}$ = the y value of the regression line for the corresponding x value
- The residual is the vertical distance between the scatterplot point and the regression line.
  - For points above (below) the regression line the residual is positive (negative).
- The residual may be important on its own right.
  - Interested in identifying 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.

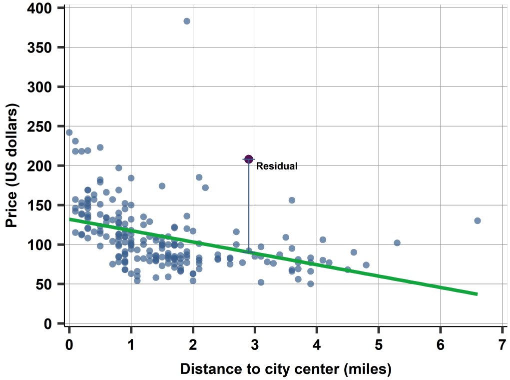
Predicted dependent variable and residuals

- Residuals can be computed for existing observations only
  - While we can have predicted values for any x, actual y values are only available for the observations in our data

- Residuals sum to zero if a linear regression is fitted by OLS.
- Sum is zero –> average of the residuals is zero, too.
- A related fact is that the predicted average is equal to the actual average of the left-hand-side variable: average $\hat{y}$ equals average y.
- Not exam, but may check textbook chapter AGTK section for details.

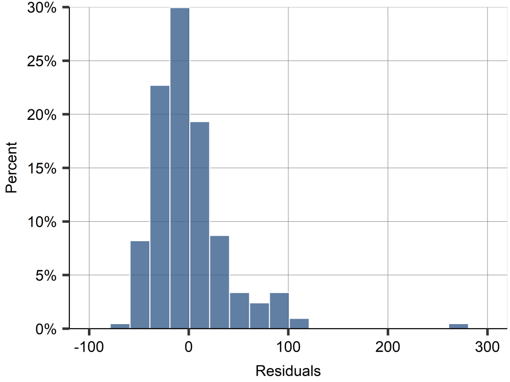
Case Study: Finding a good deal among hotels

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



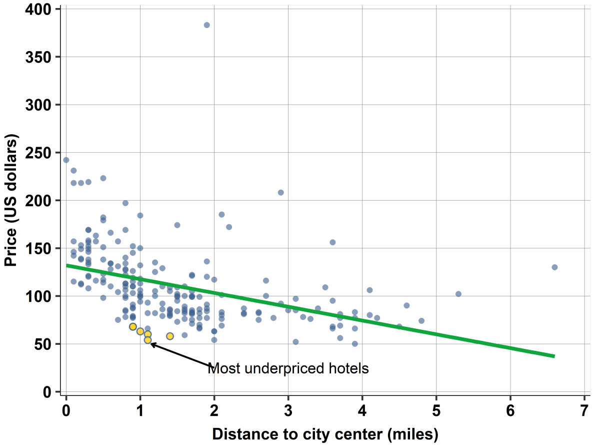
Case Study: Finding a good deal among hotels

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



Case Study: Finding a good deal among hotels

- Key graph of this exercise
- Scatterplot with regression line
- Capturing over and underpriced hotels



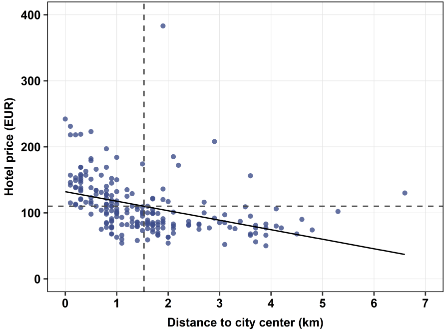
Case Study: Finding a good deal among hotels

- A list of the hotels with the five lowest value of the residual.
- Bear in mind, we can (and will) do better
  - Non-linear pattern
  - Functional form
  - Taking into account differences beyond distance

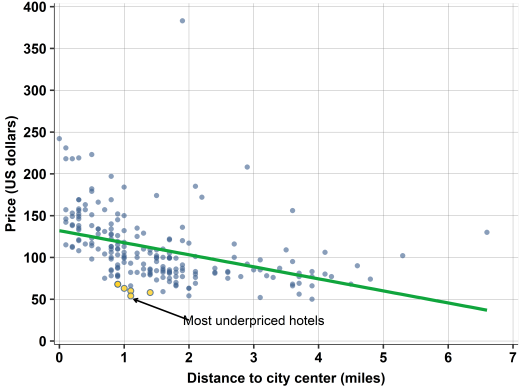
No.	hotel_id	distance	price	predicted price	residual
1	22080	1.1	54	116.17	-62.17
2	21912	1.1	60	116.17	-56.17
3	22152	1	63	117.61	-54.61
4	22408	1.4	58	111.85	-53.85
5	22090	0.9	68	119.05	-51.05

Source: `hotels` data. Vienna, November 2017, weekday.

Case Study: Just discuss dataviz - maybe skip



Scatterplot and regression and means



Scatterplot and regression and best/worst deals

Model fit

- **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

$$R^2 = \frac{Var[\hat{y}]}{Var[y]} = 1 - \frac{Var[e]}{Var[y]} \tag{8}$$

  where $Var[y] = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, $Var[\hat{y}] = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, and $Var[e] = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2$. Note that $\bar{\hat{y}} = \bar{y}$, and $\bar{e} = 0$.
- Decomposition of the overall variation in y into variation in predicted values "explained by the regression") and residual variation ("not explained by the regression"):

$$Var[y] = Var[\hat{y}] + Var[e] \tag{9}$$

Model fit

- 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.
- If R-squared of zero - all predicted $\hat{y}$ values = overall average value $\bar{y}$ in the data regardless of the value of the explanatory variable x.
  - This corresponds to a slope of zero: the regression line is completely flat.

Model fit

- Fit depends (1): how well the particular version of the regression captures the actual function f in $y^E = f(x)$
  - Can be helped by modelling
- Fit depends (2): how far actual values of y are spread around what would be predicted using the actual function f.

Model fit

- R-squared may help in choosing between different versions of regression for the same data.
  - Choose between regressions with different functional forms
  - Predictions. (prediction quality on a different sample we estimated)
- R-squared matters less when the goal is to characterize the pattern $y^E = f(x)$.
  - R-squared can help finding the regression that best approximates the $f(x)$ pattern.
  - The regression that best approximates that pattern may have a high R-squared or a low R-squared.

Correlation and linear regression

- Linear regression is closely related to correlation.
- The OLS formula for the slope estimate of the linear regression $y^E = \alpha + \beta x$ is also a normalized version of the covariance, only here it is divided by the variance of the x variable: $\hat{\beta} = \frac{Cov[y,x]}{Var[x]}$.
- In contrast with the correlation coefficient, its values can be anything, and y are x are not interchangeable.
- Covariance, the correlation coefficient, and the slope of a linear regression capture similar information: the degree of association between the two variables.

$$\hat{\beta} = Corr[x, y]\frac{Std[y]}{Std[x]} \quad Corr[x, y] = \hat{\beta}\frac{Std[x]}{Std[y]} \tag{10}$$

Correlation and linear regression

- Another way to normalize the covariance: dividing it by the variance of y not x.
- = OLS estimator for the slope coefficient of the **reverse regression**: switching the role of y and x in the linear regression.

$$x^E = \gamma + \delta y \tag{11}$$

- 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 as $\hat{\beta} = \hat{\delta}\frac{Var[y]}{Var[x]}$.
  - Different unless $Var[x] = Var[y]$,
  - always have have the same sign
  - both are larger in magnitude the larger the covariance.
- What about R-squared?

Correlation and 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.
- The numerator of R-squared, $Var[\hat{y}]$, can be written out as $Var[\hat{\alpha} + \hat{\beta}x] = \hat{\beta}^2 Var[x]$, and thus

$$R^2 = \hat{\beta}^2 Var[x]/Var[y] = (\hat{\beta} Std[x]/Std[y])^2$$

- R^2 for our regression and the reverse regression is the same.

Regression and causation

- Were very careful to use neutral language, not talk about causation
- Think back to sources of variation in x
- When we have observational data, and we pick x and y and decide how to run the regression
- Regression is a method of comparison: it compares observations that are different in variable x and shows corresponding average differences in variable y.
- It is a way to find patterns of association by comparisons.
  - Can't, infer causation from regression analysis is not the fault of the method.

Regression and causation

- The key is the source of variation i x - the method will never do the causal claim.
- It is always the data that makes it. More precisely, how the data was collected, how variation in x was provided
- For example: advertising and sales
  - Observational data, regression, no causal claim.
- 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.

Regression and causation

- The proper interpretation of the slope is necessary whether the data is observational or comes from a controlled experiment.
- A positive slope in a regression of sales on advertising means that sales tend to be higher when advertising time 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

Regression and causation

- 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:
  1. x causes y. If this is the single one thing behind the slope, it means that we can expect y to increase by β units if we were to increase x by one unit.
  2. 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.
  3. 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).

Case Study: Finding a good deal among hotels

- Fit and causation
- 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.

Case Study: Finding a good deal among hotels

- Slope is -14
- Does that mean that a longer distance causes hotels to be cheaper by that amount?

Summary take-away

- Regression – method to compare avg 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 – 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 not zero, one of three things (+ any combination ű) 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 AGTK derivations sections carefully!