

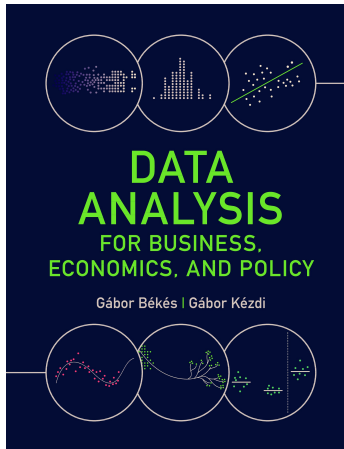
1 Simple regression

Alice Kügler

Data Analysis 2 – **MS Business Analytics**: Regression Analysis

2023

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 set of slides is for **Chapter 7**

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.



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!

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 -> covered later in this course.

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 .

$$E[y|x] = f(x) \quad (1)$$

We use a simpler shorthand notation

$$y^E = f(x) \quad (2)$$

- **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 the 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.

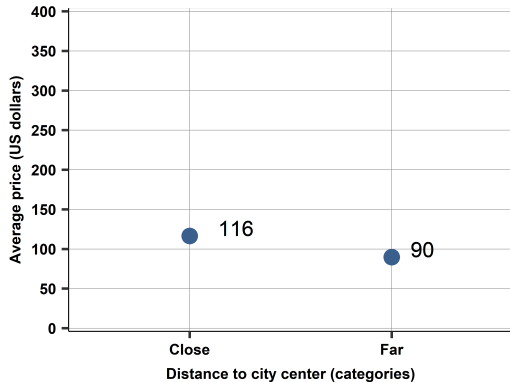
Non-parametric regression: lowess (loess)

- ▶ 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.

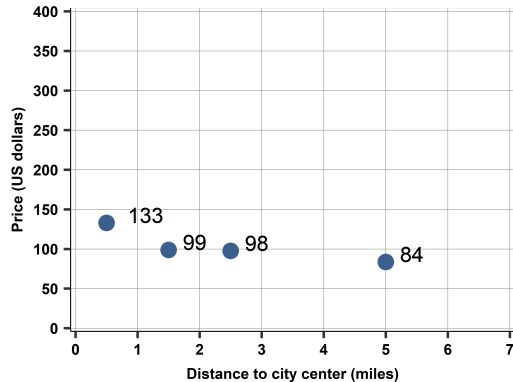
Case Study: Finding a good deal among hotels

- ▶ We look at Vienna hotels for a 2017 November weekday.
- ▶ We focus on hotels that are (i) in Vienna, (ii) not too far from the center, (iii) classified as hotels, (iv) 3-4 stars, and (v) do not have 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

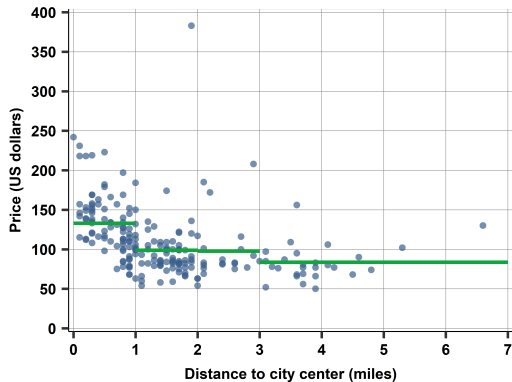


Bin scatter non-parametric regression, 2 bins

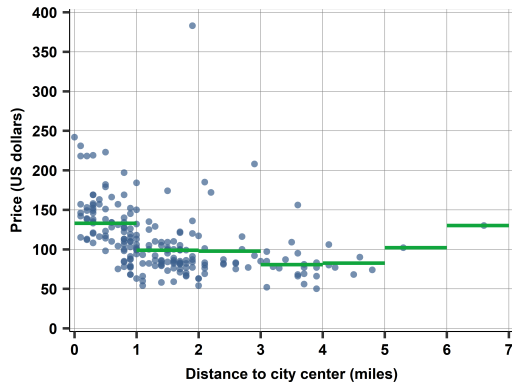


Bin scatter non-parametric regression, 4 bins

Case Study: Finding a good deal among hotels



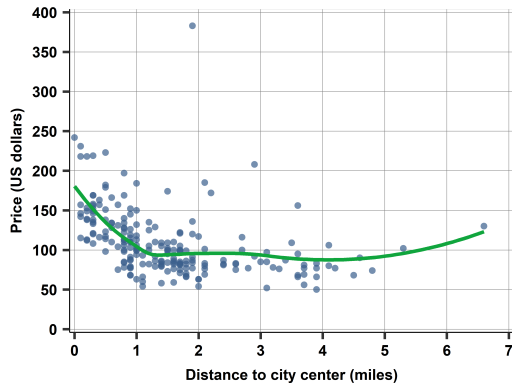
Scatter and bin scatter non-parametric regression, 4 bins



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.

$$y^E = \alpha + \beta x \quad (3)$$

- ▶ 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 .

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 .
- ▶ 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 variable

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

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

- ▶ 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

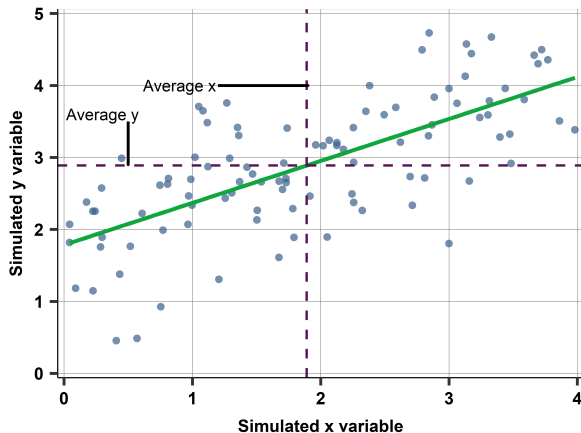
- The intercept – average y minus average x multiplied by the estimated slope $\hat{\beta}$.

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

- 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 .



Regression coefficient formula

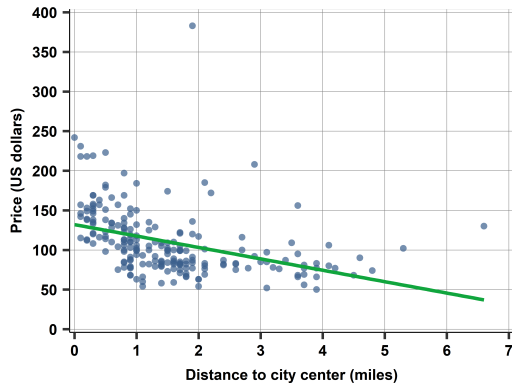
- **Ordinary Least Squares – OLS** is a method to find the best fit with a formula.
- The idea underlying OLS is to find the values of the intercept and slope parameters that make the regression line fit the scatterplot best.
- OLS method finds the values of the coefficients of the linear regression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

$$\min_{\alpha, \beta} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \quad (4)$$

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. **Please check out 7.U1.**

Case Study: Finding a good deal among hotels

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

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} .

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

- ▶ What about non-parametric regressions?

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} .

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

- ▶ What about 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.

Residuals

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

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

- ▶ The residual is meaningful only for actual observations.
 - ▶ While we can have predicted values for any x , actual y values are only available for the observations in our data.
- ▶ 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.

Use of residuals

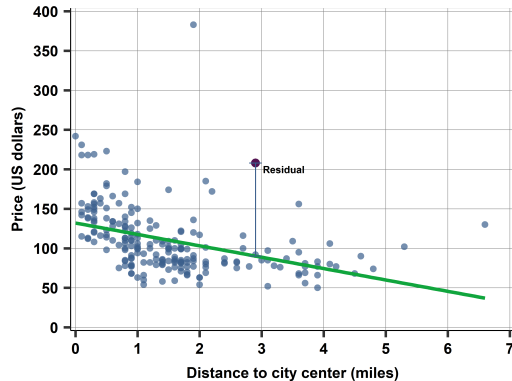
- ▶ The residual may be important on its own right.
 - ▶ Interested in identifying observations that are special in that they have a dependent variable that is much higher or much lower than “it should be” as predicted by the regression.

Predicted dependent variable and residuals

- ▶ Residuals sum to zero if a linear regression is fitted by OLS.
- ▶ Sum is zero \rightarrow average of the residuals is zero, too.
- ▶ Predicted average is equal to the actual average for y : average \hat{y} equals average y .
 - ▶ See 7.U2 for details.

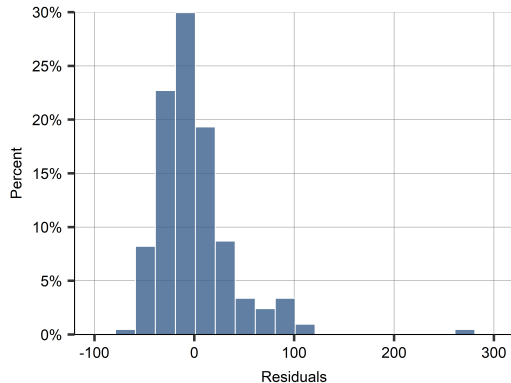
Case Study: Finding a good deal among hotels

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



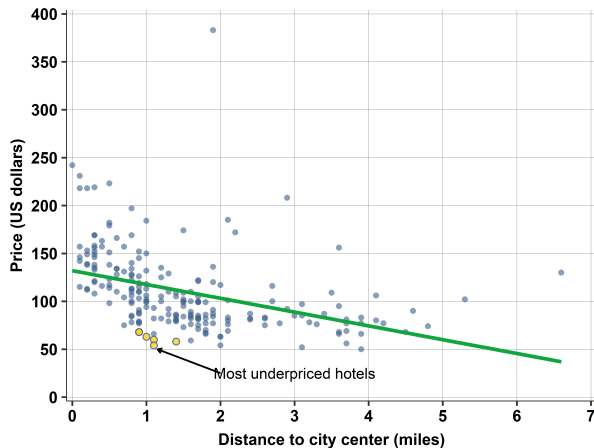
Case Study: Finding a good deal among hotels

- ▶ 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

$$R^2 = \frac{\text{Var}[\hat{y}]}{\text{Var}[y]} = 1 - \frac{\text{Var}[e]}{\text{Var}[y]} \quad (5)$$

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

$$\text{Var}[y] = \text{Var}[\hat{y}] + \text{Var}[e] \quad (6)$$

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 $\hat{y}_i = \bar{y}_i, \forall i$.
 - ▶ How does the regression line look like?

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 modeling
- Fit depends (2): how far actual values of y are spread around what would be predicted using the actual function f .
 - Given by data

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
 - ▶ More on this in DA3
- ▶ 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

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

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

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

- ▶ 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.

$$R^2 = (\text{Corr}[y, x])^2$$

- 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:

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

Reverse regression

- Consider two similar models

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

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

- What can we say about estimated coefficients?
- What can we say about the R^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 $\hat{\delta} = \frac{\text{Cov}[y,x]}{\text{Var}[y]}$.
- ▶ The OLS slopes of the original regression and the reverse regression are related:

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

- ▶ Different, unless $\text{Var}[x] = \text{Var}[y]$
- ▶ But always have the same sign.
- ▶ Both are larger in magnitude the larger the covariance.

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 $\hat{\delta} = \frac{\text{Cov}[y,x]}{\text{Var}[y]}$.
- ▶ The OLS slopes of the original regression and the reverse regression are related:

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

- ▶ Different, unless $\text{Var}[x] = \text{Var}[y]$
 - ▶ 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.

Regression and causation

- ▶ We 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.
 - ▶ If we cannot infer causation from regression analysis — not the fault of the method.

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.
 - ▶ Topic for DA4

Regression and causation

- ▶ Instead of: “correlation (regression) does not imply causation”
 - ▶ Use: we should not infer cause and effect from comparisons in observational data.
- ▶ Suggested approach consists of two steps:
 - ▶ First interpret the object (correlation of slope coefficient) precisely
 - ▶ Conclude and discuss causal claims, if any

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?

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