Lecture 7 : Simple Regression

Sakib Anwar

AN7914 Data Analytics and Modelling

University of Winchester

2024

Part 2 : Discovering

- We are now in Part 2 of the course.
- This part introduces uncovering patterns of associations with regression analysis.

- Modelling with cross-sectional data where dependent variable is continuous or binary.

Discovering

- We assume you have already an approximately well cleaned and organized data.
- Part 2 is about discovering patterns.

Discovering

- We assume you have already an approximately well cleaned and organized data.
- Part 2 is about discovering patterns.
  - This is the fun part!
  - Unfortunately, this is only a fraction of your working time.

Discovering

- We assume you have already an approximately well cleaned and organized data.
- Part 2 is about discovering patterns.
  - This is the fun part!
  - Unfortunately, this is only a fraction of your working time.
- Proper discovery means strong knowledge of statistical tools
  - Understanding the theory takes time.
  - Using theory in a computer takes a few seconds...

Data Analysis 2: Patterns - topics

1. Simple Regression (non-parametric and parametric, simple linear regression's anatomy, model summary)
2. Complicated patterns and messy data (transformations and more advanced functional forms, influential observations, measurement errors, weighted regression)
3. Generalizing results of a regression (SE of coeff, CI, prediction intervals, hypothesis testing, external validity)
4. Multiple linear regression (using more xs, omitted variable bias, inference, variable selection)
5. Probability models (binary regression models: LPM, probit, logit, non-linear regression, marginal differences, model evaluation)
6. Time series models (time series properties, (non)-stationarity and random walk, seasonality, type of trends, serial correlation, leads and lags, SARIMA models)[We will do this if we have enough time left]

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 the 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 the 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 is 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.

Linear regression

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

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

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

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

Linear regression - assumption vs approximation

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

Linear regression coefficients

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

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

Two coefficients:

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

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

Regression - slope coefficient

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

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

Regression: binary explanatory

Simplest case:

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

Regression coefficient formula

Notation:

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

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

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

Regression coefficient formula

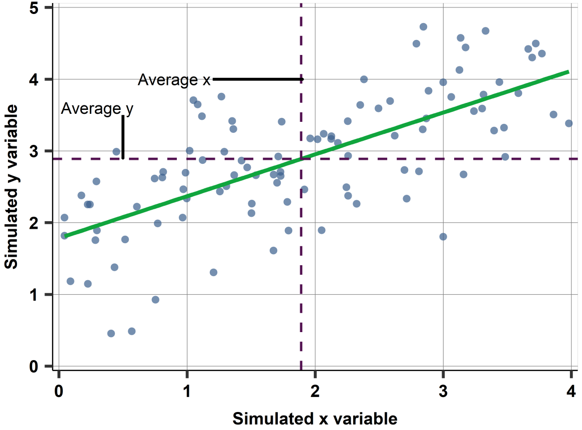
- The intercept – average y minus average x multiplied by the estimated slope $\hat{\beta}$.

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

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

Ordinary Least Squares (OLS)

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



More on OLS

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

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

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

Predicted values

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

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

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

Residuals

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

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

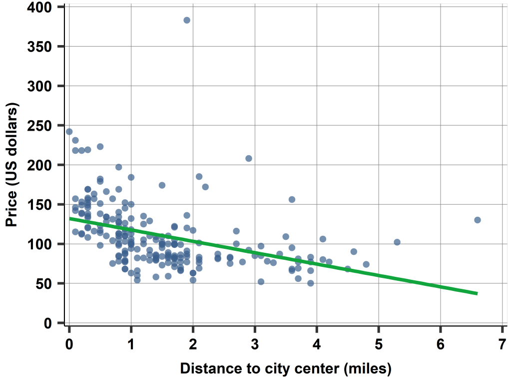
- The residual is meaningful only for actual observation. It compares observation i's difference for actual and predicted value.
- The residual is the vertical distance between the scatterplot point and the regression line.
  - For points above the regression line the residual is positive.
  - For points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

Some further comments on residuals

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

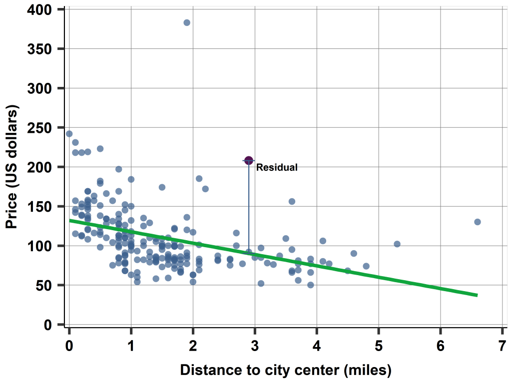
Case Study: Finding a good deal among hotels

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



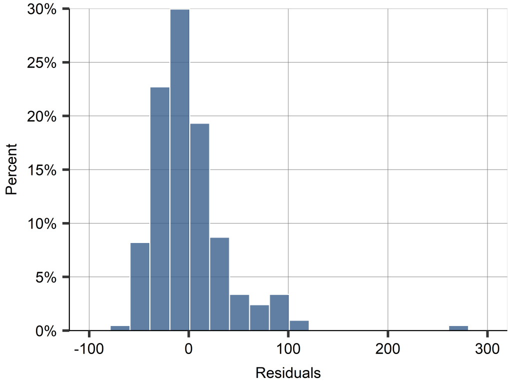
Case Study: Finding a good deal among hotels

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



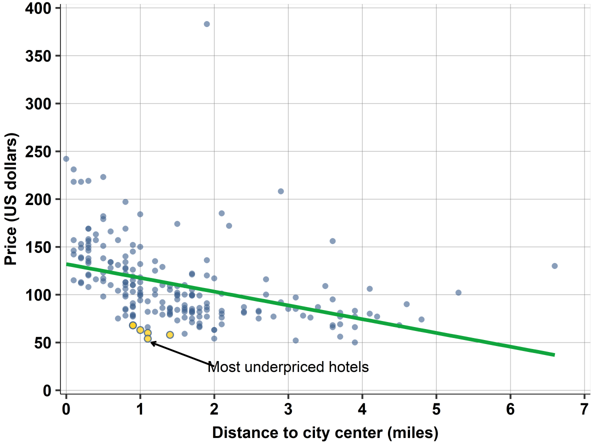
Case Study: Finding a good deal among hotels

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



Case Study: Finding a good deal among hotels

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



Case Study: Finding a good deal among hotels

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

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

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

Model fit - R^2

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

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

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

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

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

Model fit - R^2

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

Model fit - how to use R^2

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

Correlation and linear regression

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

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

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

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

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

Correlation and R^2 in linear regression

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

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

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

Regression and causation

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

Regression and causation - possible relations

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

Summary take-away

- Regression – method to compare average y across observations with different values of x.
- Linear regression – linear approximation of the average pattern of association y and x
- In $y^E = \alpha + \beta x$, β shows how much larger y is, on average, for observations with a one-unit larger x
- When β is not zero, one of three things (+ any combination) may be true:
  - x causes y
  - y causes x
  - a third variable causes both x and y.

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

Compulsory Reading: Please read Chapter 7 of Gabor Bekes Book