Lecture 9 : Generalising Regression Results

Sakib Anwar

AN7914 Data Analytics and Modelling

University of Winchester

2024

Generalizing: reminder

- We have uncovered some pattern in our data. We are interested in generalize the results.

- Question: Is the pattern we see in our data
  - True *in general*?
  - or is it just a special case what we see?
- Need to specify the situation
  - to what we want to generalize
- Inference - the act of generalizing results
  - From a particular dataset to other situations or datasets.

- From a sample to population/ general pattern = statistical inference
- Beyond (other dates, countries, people, firms) = external validity

Generalizing Linear Regression Coefficients from a Dataset

- We estimated the linear model
- $\hat{\beta}$ is the average difference in y *in the dataset* between observations that are different in terms of x by one unit.
- $\hat{y}_i$ best guess for the expected value (average) of the dependent variable for observation i with value x_i for the explanatory variable *in the dataset*.
- Sometimes all we care about are patterns, predicted values, or residuals, *in the data we have*.
- Often interested in patterns and predicted values in situations that are not limited to the dataset we analyze.
  - To what extent predictions / patterns uncovered in the data generalize to a situation we care about.

Statistical Inference: Confidence Interval

- The 95% CI of the slope coefficient of a linear regression
  - similar to estimating a 95% CI of any other statistic.

$$CI(\hat{\beta})_{95\%} = \left[\hat{\beta} - 2SE(\hat{\beta}), \hat{\beta} + 2SE(\hat{\beta})\right]$$

  - Formally: 1.96 instead of 2. (computer uses 1.96 – mentally use 2)
- The standard error (SE) of the slope coefficient
  - is conceptually the same as the SE of any statistic.
  - measures the spread of the values of the statistic across hypothetical repeated samples drawn from the same population (or general pattern) that our data represents

Standard Error of the Slope

The simple SE formula of the slope is

$$SE(\hat{\beta}) = \frac{Std\,[e]}{\sqrt{n} Std\,[x]}$$

- Where:
  - Residual: $e = y - \hat{\alpha} - \hat{\beta}x$
  - Std[e], the standard deviation of the regression residual,
  - Std[x], the standard deviation of the explanatory variable,
  - $\sqrt{n}$ the square root of the number of observations in the data.
    - Smaller sample – may use $\sqrt{n-2}$.

- A smaller standard error translates into
  - narrower confidence interval,
  - estimate of slope coefficient with more precision.
- More precision if
  - smaller the standard deviation of the residual – better fit, smaller errors.
  - larger the standard deviation of the explanatory variable – more variation in x is good.
  - more observations are in the data.
- This formula is correct assuming *homoskedasticity*

Heteroskedasticity Robust SE

- Simple SE formula is not correct in general.
  - Homoskedasticity assumption: the fit of the regression line is the same across the entire range of the x variable
  - In general this is not true
- Heteroskedasticity: the fit may differ at different values of x so that the spread of actual y around the regression is different for different values of x
- Heteroskedastic-robust SE formula (*White or Huber*) is correct in both cases
  - Same properties as the simple formula: smaller when $Std[e]$ is small, $Std[x]$ is large and n is large
  - E.g. White formula uses the squared estimated error from the model and weight the observations when calculating the $SE[\hat{\beta}]$
  - Note: there are many heteroskedastic-robust formula, which uses different weighting techniques. Usually referred as 'HC0', 'HC1', ... , 'HC4'.

The CI Formula in Action

- Run linear regression
- Compute endpoints of CI using SE
- 95% CI of slope and intercept
  - $\hat{\beta} \pm 2SE\left(\hat{\beta}\right)$; $\hat{\alpha} \pm 2SE\left(\hat{\alpha}\right)$
- In regression, as default, use robust SE.
  - In many cases homoskedastic and heteroskedastic SEs are similar.
  - However, in some cases, robust SE is larger – and rightly so.
- Coefficient estimates, R^2 etc. remain the same.

Case Study: Gender gap in earnings?

- Earning determined by many factor
- The idea of gender gap:
  - Is there a systematic wage differences between male and female workers?

Case Study: Gender gap - How data is born?

- Current Population Survey (CPS) of the U.S.
  - Administrative data
- Large sample of households
- Monthly interviews
  - Rotating panel structure: interviewed in 4 consecutive months, then not interviewed for 8 months, then interviewed again in 4 consecutive months
  - Weekly earnings asked in the "outgoing rotation group"
    - In the last month of each 4-month period
  - See more on MORG: "Merged outgoing rotation group"
- Sample restrictions used:
  - Sample includes individuals of age 16-65
  - Employed (has earnings)
  - Self-employed excluded

Case Study: Gender gap - the data

- Download data for 2014 (316,408 observations) with implemented restrictions $N = 149,316$
- Weekly earnings in CPS
  - Before tax
  - Top-coded very high earnings
    - at $2,884.6 (top code adjusted for inflation, 2.5% of earnings in 2014)
  - Would be great to measure other benefits, too (yearly bonuses, non-wage benefits). But we don't measure those.
- Need to control for hours
  - Women may work systematically different in hours than men.
- Divide weekly earnings by 'usual' weekly hours (part of questionnaire)

Case Study: Gender gap - conditional descriptives

Gender	mean	p25	p50	p75	p90	p95
Male	$ 24	13	19	30	45	55
Female	$ 20	11	16	24	36	45
% gap	-17%	-16%	-18%	-20%	-20%	-18%

- 17% difference on average in per hour earnings between men and women
- For linear regression analysis, we will use ln wage to compare relative difference.

Case Study: Gender gap in comp science occupation - Analysis

- One key reason for gap could be women being sectors / occupations that pay less. Focus on a single one: Computer science occupations, $N = 4,740$

$$\ln(w)^E = \alpha + \beta \times G_{female}$$

- We regressed log earnings per hour on G binary variable that is one if the individual is female and zero if male.
- The log-level regression estimate is $\hat{\beta} = -0.1475$
  - female computer science field employee earns 14.7 percent less, on average, than male with the same occupation in this dataset.
- Statistical inference based on 2014 data.
  - SE: .0177; 95% CI: [-.182 -.112]
    - Simple vs robust SE - Here no practical difference.

Case Study: Gender gap in comp science occupation - Generalizing

- In 2014 in the U.S.
  - the population represented by the data
- we can be 95% confident that the average difference between hourly earnings of female CS employee versus a male one was -18.2% to -11.2%.
- This confidence interval does not include zero.
- Thus we can rule out with a 95% confidence that their average earnings are the same.
  - We can rule this out at 99% confidence as well

Case Study: Gender gap in market analyst occupation

- Market research analysts and marketing specialists, $N = 281$, where females are 61%.
  - Average hourly wage is $29 (sd:14.7)
- The regression estimate is $\hat{\beta} = -0.113$:
  - Female market research analyst employee earns 11.3 percent less, on average, than men with the same occupation in this dataset.
- Generalization:
  - $SE[\hat{\beta}]$: .061; 95% CI: [-.23 +0.01]
    - We can be 95% confident that the average difference between hourly earnings of female CS employee versus a male one was -23% to +1% in the total US population
  - This confidence interval does include zero. Thus, we can not rule out with a 95% confidence that their average earnings are the same. $(p = 0.068)$
  - More likely, though, female market analysts earn less.
    - we can rule out with a 90% confidence that their average earnings are the same

Testing if (true) beta is zero

- Testing hypotheses: decide if a statement about a general pattern is true.
- Most often: Dependent variable and the explanatory variable are related at all?
- The null and the alternative:

$$H_0 : \beta_{true} = 0, \;\; H_A : \beta_{true} \neq 0$$

- The t-statistic is:

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}$$

- Often $t = 2$ is the critical value, which corresponds to 95% CI. $(t = 2.6 \rightarrow 99\%)$

Language: *significance* of regression coefficients

- A coefficient is said to be "significant"
  - If its confidence interval does not contain zero
  - So true value unlikely to be zero
- Level of significance refers to what % confidence interval
  - Language uses the complement of the CI
- Most common: 5%, 1%
  - Significant at 5%
    - Zero is not in 95% CI, Often denoted $p < 0.05$
  - Significant at 1%
    - Zero is not in 99% CI, $(p < 0.01)$

Ohh, that p=5% cutoff

- When testing, you start with a critical value first
- Often the standard to publish a result is to have a p value below 5%.
  - Arbitrary, but... [major discussion]
- If you find a result that cannot be told apart from 0 at 1% (max 5%), you should say that explicitly.
- Key point is: publish the p-value. Be honest...

Our two samples. What is the source of difference?

- Computer and Mathematical Occupations
  - 4740 employees, Female: 27.5%
  - The regression estimate of slope: -0.1475 ; 95% CI: [-.1823 -.1128]
- Market research analysts and marketing specialists
  - 281 employees, Female: 61%
- The regression estimate of slope is -0.113; 95% CI: [-.23 +0.01]
- Why the difference?
  - True difference: gender gap is higher in CS.
  - Statistical error: sample size issue $\longrightarrow$ in small samples we may find more variety of estimates. (Why? Remember the SE formula.)
- Which explanation is true?
  - We do not know!
  - Need to collect more data in CS industry.

Chance Events And Size of Data

- Finding patterns by chance may go away with more observations
  - Individual observations may be less influential
  - Effects of idiosyncratic events may average out
    - E.g.: more dates
  - Specificities to a single dataset may be less important if more sources
    - E.g.: more hotels
- More observations help only if
  - Errors and idiosyncrasies affect some observations but not all
  - Additional observations are from appropriate source
    - If worried about specificities of Vienna more observations from Vienna would not help

Prediction uncertainty

- Goal: predicting the value of y for observations outside the dataset, when only the value of x is known.
- We predict y based on coefficient estimates, which are relevant in the *general pattern*/population. With linear regression you have a simple model:

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \epsilon_i$$

- The estimated statistic here is a predicted value for a particular observation $\hat{y}_j$. For an observation j with known value x_j this is

$$\hat{y}_j = \hat{\alpha} + \hat{\beta} x_j$$

- Two kinds of intervals:
  - Confidence interval for the predicted value/regression line - uncertainty about $\hat{\alpha}, \hat{\beta}$
  - Prediction interval - uncertainty about $\hat{\alpha}, \hat{\beta}$ *and* ϵ_i

Confidence interval of the regression line I.

- Confidence interval (CI) of the predicted value = the CI of the regression line.
- The predicted value $\hat{y}_j$ is based on $\hat{\alpha}$ and $\hat{\beta}$ only.
  - The CI of the predicted value combines the CI for $\hat{\alpha}$ and the CI for $\hat{\beta}$.
- What value to expect if we know the value of x_j and we have estimates of coefficients $\hat{\alpha}$ and $\hat{\beta}$ from the data.
- The 95% CI of the predicted value - $95\% CI(\hat{y}_j)$ is
  - the value estimated from the sample
  - plus and minus its standard error.

Confidence interval of the regression line II.

- Predicted average y has a standard error (homoskedastic case)

$$95\% CI(\hat{y}_j) = \hat{y} \pm 2SE(\hat{y}_j)$$

$$SE(\hat{y}_j) = Std[e]\sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{nVar[x]}}$$

- Based on formula for regression coefficients, it is small if:
  - coefficient SEs are small (depends on $Std[e]$ and $Std[x]$).
  - Particular x_j is close to the mean of x
  - We have many observations n
- The role of n (sample size), here is even larger.
- Use robust SE formula in practice, but a simple formula is instructive

Case Study: Earnings and age - regression table

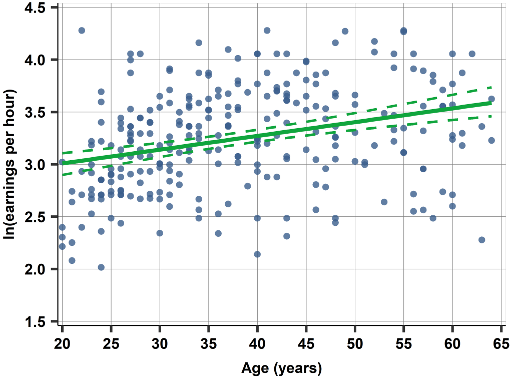
Model:

- $\ln wage = \alpha + \beta age$
- Only one industry: market analysts, $N = 281$
- Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1.

VARIABLES	ln wage
age	0.014**
	(0.003)
Constant	2.732**
	(0.101)
Observations	281
R-squared	0.098

Case Study: Earnings and age - CI of regression line

- Log earnings and age
  - linearity is only an approximation
- Narrow CI as SE is small
- Hourglass shape
  - Smaller as x_j is closer to the mean of x

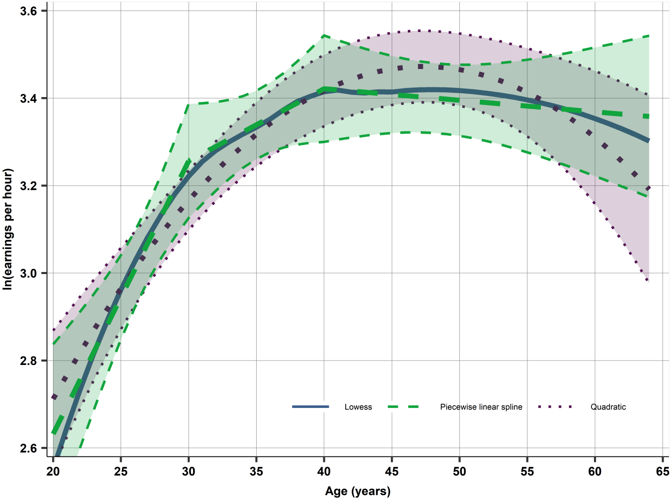


Confidence interval of the regression line - use

- Can be used for any model
  - Spline, polynomial
  - The way it is computed is different for different kinds of regressions (usually implemented in R packages)
  - always true that the CI is narrower
    - the smaller $Std[e]$,
    - the larger n and
    - the larger $Std[x]$
- In general, the CI for the predicted value is an interval that tells where to expect average y given the value of x in the population, or general pattern, represented by the data.

Case Study: Earnings and age - different fn form with CI

- Log earnings and age with:
  - Lowess
  - Piecewise linear spline
  - quadratic function
- 95% CI dashed lines
- What do you see?

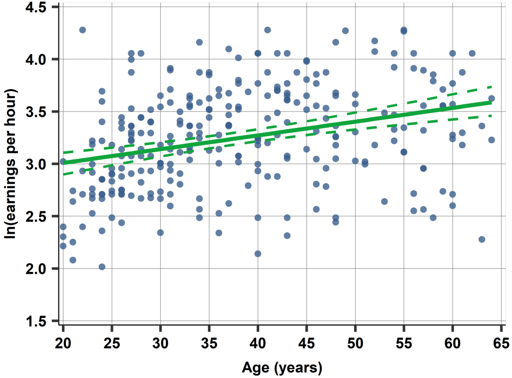


Prediction interval

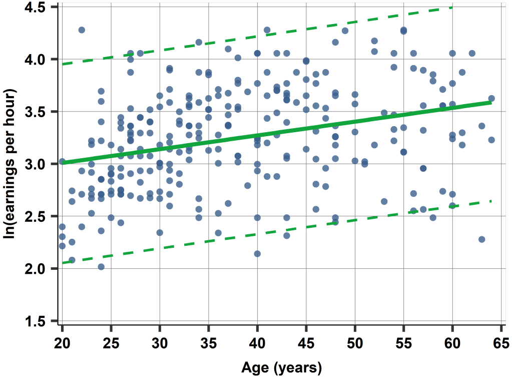
- *Prediction interval* answers:
  - Where to expect the particular y_j value if we know the corresponding x_j value and the estimates of the regression coefficients from the data.
- Difference between CI and PI.
  - The CI of the predicted value is about $\hat{y}_j$: where to expect the average value of the dependent variable if we know x_j.
  - The PI (prediction interval) is about y_j itself not its average value: where to expect the actual value of y_j if we know x_j.
- So PI starts with CI. But adds additional uncertainty $(Std[\epsilon_i])$ that actual y_j will be around its conditional.
- What shall we expect in graphs?

Confidence vs Prediction interval

Confidence interval



Prediction interval



More on prediction interval

- The formula for the 95% prediction interval is

$$95\%PI(\hat{y}_j) = \hat{y} \pm 2SPE(\hat{y}_j)$$

$$SPE(\hat{y}_j) = Std[e]\sqrt{1 + \frac{1}{n} + \frac{(x_j - \bar{x})^2}{nVar[x]}}$$

- SPE – Standard Prediction Error (SE of prediction)
  - It does matter here which kind of SE you use!
- Summarizes the additional uncertainty: the actual y_j value is expected to be spread around its average value.
  - The magnitude of this spread is best estimated by the standard deviation of the residual.
- With SPE, no matter how large the sample we can always expect actual y values to be spread around their average values.
  - In the formula, all elements get very small if n gets large, except for the new element.

External validity

- Statistical inference helps us generalize to the population or general pattern
- Is this true beyond (other dates, countries, people, firms)?
- As external validity is about generalizing beyond what our data represents, we can't assess it using our data.
  - We'll never really know. Only think, investigate, make assumption, and hope...

Data analysis to help assess external validity

- Analyzing other data can help!
- Focus on β, the slope coefficient on x.
- The three common dimensions of generalization are *time, space, and other groups*.

- To learn about external validity, we always need additional data, on say, other countries or time periods.
  - We can then repeat regression and see if the slope is similar!