02. Generalizing from Regressions and Messy Data

Agoston Reguly

Data Analysis 2: Regression analysis

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 $\sqrt{n-2}$.

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 estimated errors' square 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. are remain the same.

Case Study: Gender gap in earnings?

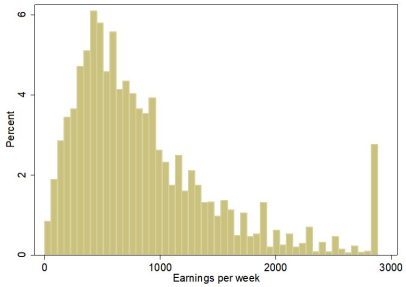
- Earning determined by many factor
- The idea of gender gap:
  - Is there a systematic wage differences between male and female workers?

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

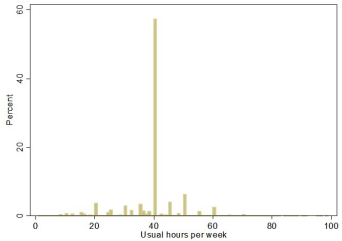
- Download data for 2014 (316,408 observations) with implemented restrictions $N = 149,316$
- Weekly earnings in CPS
  - Before tax
  - Top-coded very high earnings
    - at \$2,884.6 (top code adjusted for inflation, 2.5% of earnings in 2014)
  - Would be great to measure other benefits, too (yearly bonuses, non-wage benefits). But we don't measure those.



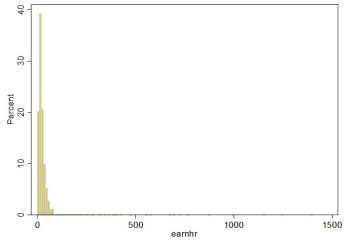
Case Study: Gender gap - control for hours

- Need to control for hours
  - Women may work systematically different in hours than men.
- Measure usual weekly working hours.
- Divide weekly earnings by 'usual' weekly hours (part of questionnaire)

Usual Weekly Hours



Earning per hour



Case Study: Gender gap - conditional descriptives

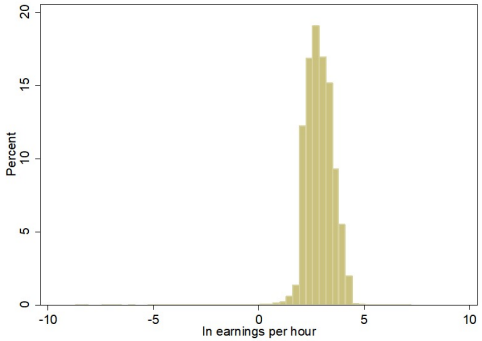
Gender	mean	p25	p50	p75	p90	p95
Male	$ 24	13	19	30	45	55
Female	$ 20	11	16	24	36	45
% gap	-17%	-16%	-18%	-20%	-20%	-18%

- 17% difference on average in per hour earnings between men and women

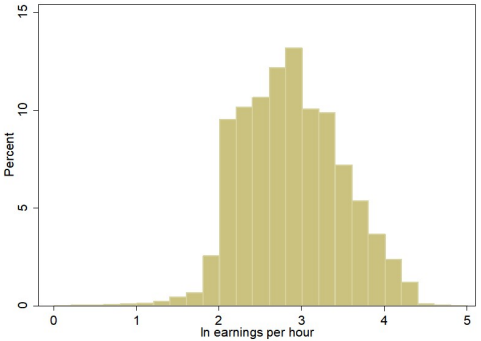
Case Study: Gender gap transform w ln

- To use simple linear regression we take logs.

Usual Weekly Hours in logs



Earning per hour in logs



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, $N = 4,740$

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

- We regressed log earnings per hour on D 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.

Dealing with 5-10%

- Sometimes regression result will not be significant at 5% but will be at 10%.
- What not to do? Well avoid:
  - a barely detectable statistically significant difference (p=0.073)
  - a margin at the edge of significance (p=0.0608)
  - not significant in the normally accepted statistical sense (p=0.064)
  - slight tendency toward significance (p=0.086)
  - slightly missed the conventional level of significance (p=0.061)
- More here

Dealing with 5-10%

- Sometimes regression result will not be significant at 1% (5%) but will be at 10%.
- What to take? It depends. (our view...)
- Sometimes you work on a proposal. Proof of concept.
  - To be lenient is okay.
  - Say the point estimate and note the 95% confidence interval.
- Sometimes looking for a proof. Beyond reasonable doubt.
  - Gender equality to be defended for a judge.
  - Here you wanna be below 1%
  - If not, say the p-value and note that at 1% you cannot reject the null of no difference.
- 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: [-0.1823 -0.1128]
- Market research analysts and marketing specialists
  - 281 employees, Female: 61%
- The regression estimate of slope is -0.113; 95% CI: [-0.23 +0.01]
- Why the difference?

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: [-0.1823 -0.1128]
- Market research analysts and marketing specialists
  - 281 employees, Female: 61%
- The regression estimate of slope is -0.113; 95% CI: [-0.23 +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?

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: [-0.1823 -0.1128]
- Market research analysts and marketing specialists
  - 281 employees, Female: 61%
- The regression estimate of slope is -0.113; 95% CI: [-0.23 +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

Reminder I: Multiple testing

- You are interested to find patterns
- There are hundred options
  - Many examples in medicine
- By chance you may find a significant relationship at 1%
- Hence: be very conservative
  - Some theory suggests using a very small p-value
  - Bonferroni correction - too conservative

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:

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

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

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

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

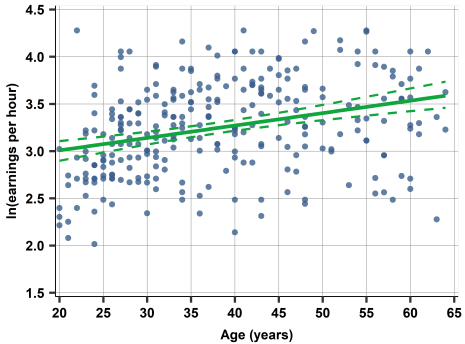
Model:

- 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

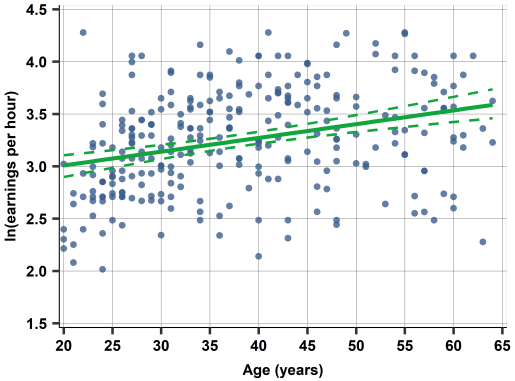


Prediction interval

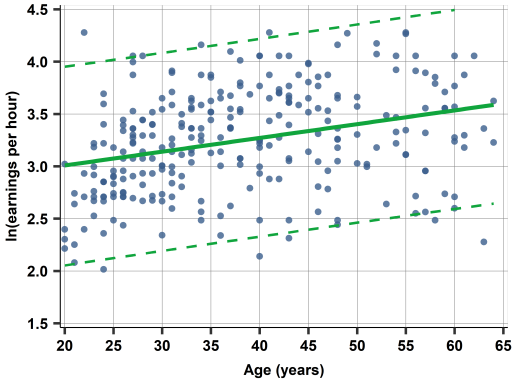
- *Prediction interval* answers:
  - Where to expect the particular y_j value if we know the corresponding x_j value and the estimates of the regression coefficients from the data.
- Difference between CI and PI.
  - The CI of the predicted value is about $\hat{y}_j$: where to expect the average value 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

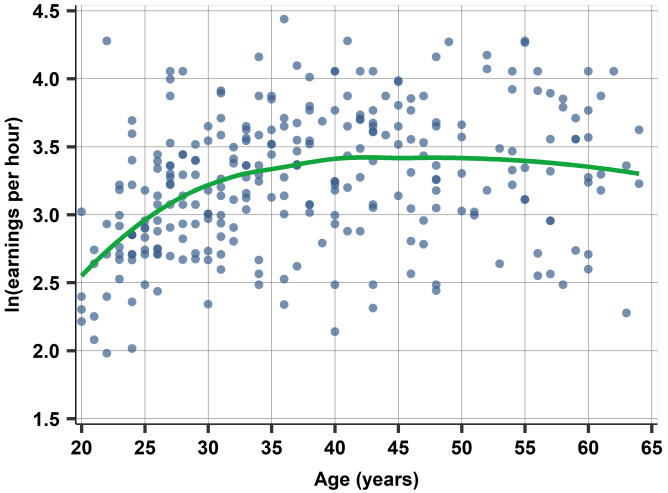
$$95\%PI(\hat{y}_j) = \hat{y} \pm 2SPE(\hat{y}_j)$$

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

- 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.

Case Study: Earnings and age - different fn forms

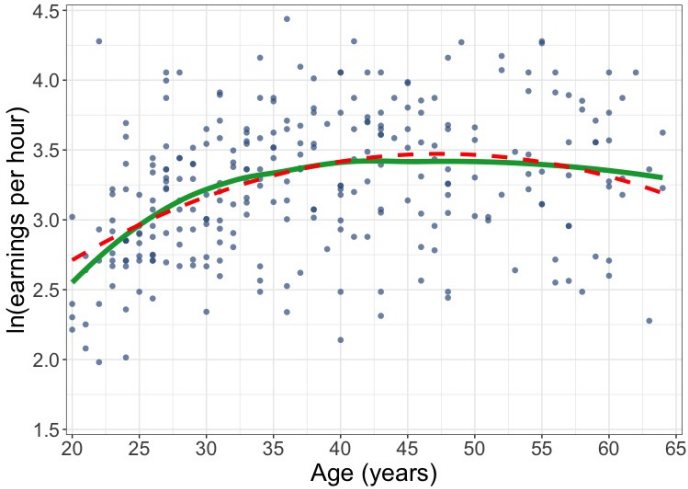
- Lets have a look at log earnings and age with lo(w)ess
- Pattern does not look like linear, but rather something else...



Case Study: Earnings and age - polynomial

- Another approach is to use polinomials!
- Start with simple quadratic pattern!

$$y^E = \alpha + \beta_1 x + \beta_2 x^2$$



Polynomials

- Quadratic function of the explanatory variable
  - Allow for a smooth change in the slope
  - Without any further decision from the analyst
- Technically: quadratic function is not a linear function (a parabola, not a line)
  - Handles only nonlinearity, which can be captured by a parabola.
  - Less flexible than a piecewise linear spline, but easier interpretation!

$$y^E = \alpha + \beta_1 x + \beta_2 x^2$$

- Can have higher order polynomials, in practice you may use cubic specification: $y^E = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
- General case

$$y^E = \alpha + \beta_1 x + \beta_2 x^2 + \ldots \beta_n x^n$$

Quadratic form - interpretation I.

$$y^E = \alpha + \beta_1 x + \beta_2 x^2$$

- α is average y when $x = 0$,
- β_1 has no interpretation in itself,
- β_2 shows whether the parabola is
  - U-shaped or convex (if $\beta_2 > 0$)
  - inverted U-shaped or concave (if $\beta_2 < 0$).

Quadratic form - interpretation II.

$$y^E = \alpha + \beta_1 x + \beta_2 x^2$$

- Difference in y, when x is different. This leads to (partial) derivative of y^E w.r.t. x,
$$\frac{\partial y^E}{\partial x} = \beta_1 + 2\beta_2 x$$
- the slope is *different for different values of* x
  - Compare two observations, j and k, that are different in x, by one unit: $x_k = x_j + 1$.
- Units which are one unit larger than x_j are higher by $\beta_1 + 2\beta_2 x_j$ in y on average.
  - Usually we compare to the average of x: $x_j = \bar{x}$.
    - Units which are one unit larger than the average of x are higher by $\gamma = \beta_1 + 2\beta_2 \bar{x}$ in y on average.

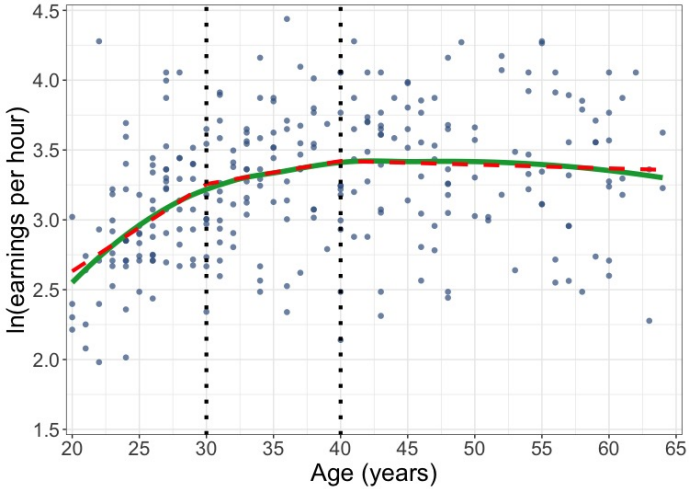
Case Study: Earnings and age - polynomials

- $\hat{\alpha} = 1.193$ is the average log-wage when age is 1. (not meaningful)
- $\hat{\beta}_1$ no interpretation in intself.
- $\hat{\beta}_2$ shows that it is inverted U-shape (concave)
- $\frac{\partial y^E}{\partial x} = \hat{\beta}_1 + 2\hat{\beta}_2\bar{x} = 0.096 - 2 * 0.001 * 38 = 0.02$
  - People who are one year older than the *average age* earn 2% more on average.
- ... but it depends on the age!
  - E.g., for people who are one year older than 30, earns 3.6% more on average.
  - but people who are one year older than 60, earns 2.4% less on average.

VARIABLES	ln wage
Constant	1.193** (0.341)
age	0.096** (0.018)
age squared	-0.001** (0.000)
Observations	281
R-squared	0.168

Case Study: Earnings and age - PLS

- It is not straightforward, why older people earns less!
- Maybe quadratic functional form mechanically introduce this decline!
- Lets use another nonlinear functional form that has easy interpretation!



Piecewise Linear Splines

- A regression with a piecewise linear spline of the explanatory variable.
  - Results in connected line segments for the mean dependent variable.
  - Each line segment corresponding to a specific interval of the explanatory variable.
- The points of connection are called knots,
  - the line may be broken at each knot so that the different line segments may have different slopes.
  - A piecewise linear spline with m line segments is broken by $m - 1$ knots.
- The places of the knots (the boundaries of the intervals of the explanatory variable) need to be specified by the analyst.
  - Written function for Python to calculate the rest.

Piecewise Linear Splines - formula

- A piecewise linear spline regression results in connected line segments, each line segment corresponding to a specific interval of x.
- The formula for a piecewise linear spline regression with m line segments (and $m-1$ knots in-between) is:

$$\begin{aligned} y^E =& \alpha_1 + \beta_1 x \mathbb{1}(x < k_1) + \ldots + (\alpha_j + \beta_j x)\mathbb{1}(k_{j-1} \leq x < k_j) + \ldots \\ &+ (\alpha_m + \beta_m x)\mathbb{1}(x \geq k_{m-1}), \qquad j = 2, \ldots, m-1 \end{aligned}$$

Piecewise Linear Splines - interpretation

$$y^E = \alpha_1 + \beta_1 x \mathbb{1}(x < k_1) + \ldots + (\alpha_j + \beta_j x)\mathbb{1}(k_{j-1} \leq x < k_j) + \ldots$$
$$+ (\alpha_m + \beta_m x)\mathbb{1}(x \geq k_{m-1}), \qquad j = 2, \ldots, m-1$$

Interpretation of the most important parameters:

- α_1 : average y when x is zero, if $k_1 > 0$ (Otherwise: $\alpha_1 + \alpha_j$, where $k_{j-1} \leq 0 < k_j$)
- β_1 : When comparing observations with x values less than k_1, y is β_1 units higher, on average, for observations with one unit higher x value.
- β_j : When comparing observations with x values between k_{j-1} and k_j, y is β_j units higher, on average, for observations with one unit higher x value.
- β_m : When comparing observations with x values greater than k_{m-1}, y is β_m units higher, on average, for observations with one unit higher x value.

Case Study: Earnings and age - polynomials

VARIABLES	ln wage
Constant	1.383** (0.369)
age < 30	0.062** (0.014)
$30 \leq$ age < 40	0.017 (0.010)
age ≥ 40	-0.003 (0.006)
Observations	281
R-squared	0.173

- $\hat{\alpha} = 1.383$ is the average log-wage when age is 1. (not meaningful)
- $\hat{\beta}_1$ within people less than 30 years old, we see that people who are one year older, on average earns 6.2% more.
- $\hat{\beta}_2$ within people more (or equal) than 30 years old, but less than 40, we see that people who are one year older, on average earns 1.7% more. However it is statistically not significant.
- $\hat{\beta}_2$ within people more (or equal) than 40 years old, we do not see any statistically significant differences in earnings.

Overview of piecewise linear spline

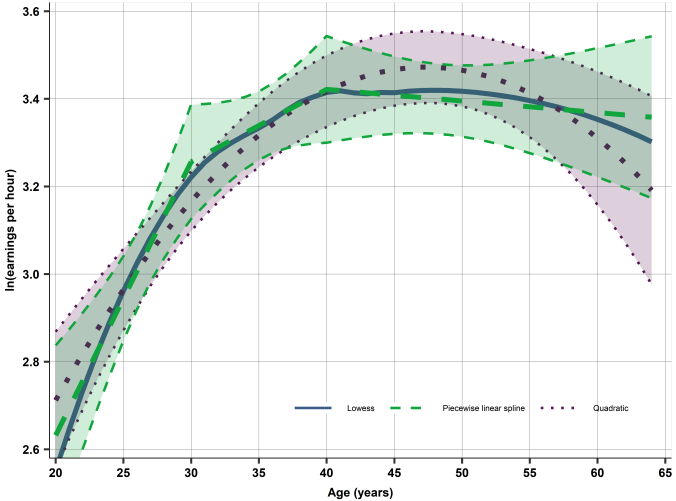
- A regression with a piecewise linear spline of the explanatory variable
- Handles any kind of nonlinearity
  - Including non-monotonic associations of any kind
- Offers complete flexibility
- But requires decisions from the analyst
  - How many knots?
  - Where to locate them
  - Decision based on scatterplot, theory / business knowledge
  - Often several trials.
- You can make it more complicated:
  - Quadratic, cubic or B-splines $\rightarrow$ rather a non-parametric approximation: interpretation-fit trade-off
  - Example: term-structure modelling (y: zero-coupon interest rate, x: maturity time) cubic spline is used. Link

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?



Which functional form to choose? - guidelines

Start with deciding whether you care about nonlinear patterns.

- Linear approximation OK if focus is on an average association.
- Transform variables for a better interpretation of the results (e.g. log), and it often makes linear regression better approximate the average association.
- Accommodate a nonlinear pattern if our focus is
  - on prediction,
  - analysis of residuals,
  - about how an association varies beyond its average.
  - Keep in mind - simpler the better!

Which functional form to choose? - practice

To uncover and include a potentially nonlinear pattern in the regression analysis:

1. Check the distribution of your main variables (y and x)
2. Uncover the most important features of the pattern of association by examining a scatterplot or a graph produced by a *nonparametric* regression such as lowess or bin scatter.
3. Think and check what would be the best transformation!
   - 3.1 Choose one or more ways to incorporate those features into a linear regression (transformed variables, piecewise linear spline, quadratic, etc.).
   - 3.2 Remember for some variables log transformation or using ratios is not meaningful!
4. Compare the results across various regression approaches that appear to be good choices. –> *robustness check*.

Reminder II: External validity

- Statistical inference helps us generalize to the population or general pattern
- Is this true beyond (other dates, countries, people, firms)?

Reminder II: 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 slope is similar!

Stability of hotel prices - idea

- Here we ask different questions: whether we can infer something about the price–distance pattern for situations outside the data:
- Is the slope coefficient close to what we have in Vienna, November, weekday:
  - Other dates
  - Other cities
  - Other type of accommodation: apartments
- Compare them to our benchmark model result
- Learn about uncertainty when using model to some types of external validity.

Why carrying out such analysis?

- Such a speculation may be relevant:
  - Find a good deal in the future without estimating a new regression but taking the results of this regression and computing residuals accordingly.
  - Be able to generalize to other groups, date and places.

Benchmark model

The benchmark model is a spline with a knot at 2 miles.

$$ln(y)^E = \alpha_1 + \beta_1 x \mathbb{1}(x < 2\text{miles}) + (\alpha_2 + \beta_2 x)\mathbb{1}(x \geq 2\text{miles})$$

Data is restricted to 2017, November weekday in Vienna, 3-4 star hotels, within 8 miles.

Benchmark model

The benchmark model is a spline with a knot at 2 miles.

$$ln(y)^E = \alpha_1 + \beta_1 x \mathbb{1}(x < 2\text{miles}) + (\alpha_2 + \beta_2 x)\mathbb{1}(x \geq 2\text{miles})$$

Data is restricted to 2017, November weekday in Vienna, 3-4 star hotels, within 8 miles.

- Model has three output variables: $\alpha = 5.02$, $\beta_1 = -0.31$, $\beta_2 = 0.02$
- α: Hotel prices are on average 151.41 euro (exp(5.02)) at the city center
- β_1: hotels that are within 2 miles from the city center, prices are 0.31 log units or 36% ($exp(0.31) - 1$) cheaper, on average, for hotels that are 1 mile farther away from the city center.
- β_2: hotels in the data that are beyond 2 miles from the city center, prices are 2% higher, on average, for hotels that are 1 mile farther away from the city center.

Comparing dates

VARIABLES	(1) 2017-NOV-weekday	(2) 2017-NOV-weekend	(3) 2017-DEC-holiday	(4) 2018-JUNE-weekend
dist_0_2	-0.31**	-0.44**	-0.36**	-0.31**
	(0.038)	(0.052)	(0.041)	(0.037)
dist_2_7	0.02	-0.00	0.07	0.04
	(0.033)	(0.036)	(0.050)	(0.039)
Constant	5.02**	5.51**	5.13**	5.16**
	(0.042)	(0.067)	(0.048)	(0.050)
Observations	207	125	189	181
R-squared	0.314	0.430	0.382	0.306

Note: *Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1*
Source: `hotels-europe` data. Vienna, reservation price for November and December 2017, June in 2018

Comparing dates - interpretation

- November weekday and the June weekend: $\hat{\beta}_1 = 0.31$
  - Estimate is similar for December (-0.36 log units)
  - Different for the November weekend: they are 0.44 log units or 55% ($exp(0.44) - 1$) cheaper during the November weekend.
    - The corresponding 95% confidence intervals overlap somewhat: they are [-0.39,-0.23] and [-0.54,-0.34].
    - Thus we cannot say for sure that the price–distance patterns are different during the weekday and weekend in November.

Comparing across cities

VARIABLES	(1) Vienna	(2) Amsterdam	(3) Barcelona
dist_0_2	-0.31**	-0.27**	-0.06
	(0.038)	(0.040)	(0.034)
dist_2_7	0.02	0.03	-0.05
	(0.033)	(0.037)	(0.058)
Constant	5.02**	5.24**	4.67**
	(0.042)	(0.041)	(0.041)
Observations	207	195	249
R-squared	0.314	0.236	0.023

Note: *Robust standard errors in parentheses *** $p<0.01$, ** $p<0.05$, * $p<0.1$*
Source: `hotels` data. November 2017, weekday

Comparing across accommodation types

VARIABLES	(1) Hotels	(2) Apartments
dist_0_2	-0.31**	-0.26**
	(0.035)	(0.069)
dist_2_7	0.02	0.12
	(0.032)	(0.061)
Constant	5.02**	5.15**
	(0.044)	(0.091)
Observations	207	92
R-squared	0.314	0.134

Note: *Robust standard errors in parentheses*
*** $p<0.01$, ** $p<0.05$, * $p<0.1$
Source: `hotels` data. Vienna, November 2017 weekday

Stability of hotel prices - takeaway

- Fairly stable overtime but uncertainty is larger
- Variation across cities, may not transfer to other cities
- Apartments similar to hotels

- Evidence of some external validity in Vienna
- External validity in other cities may vary, we do not know

- External validity – if model applied beyond data, there is additional uncertainty!

Data Is Messy

- Clean and neat data exist only in dreams and in some textbooks...
- Data may be messy in many ways!
- Structure, storage type differs from what we want
  - Needs cleaning $\Longrightarrow$ DA1

There are potential issues with the variable(s) itself:

- Some observations are influential
  - How to handle them? Drop them? Probably not but depends on the context.
- Variables measured with (systematic) error
  - When does it lead to biased estimates?

Extreme values vs influential observations

- Extreme values concept:
  - Observations with extreme values for some variable
- Extreme values examples:
  - Banking sector employment share in countries. Luxembourg: 10%
  - Number of foreign companies registered/population. Cyprus or US Virgin Island.
  - Hotel price of 1 US dollars or 10,000 US dollars
- Influential observations
  - Their inclusion or exclusion influences the regression line
  - Influential observations are extreme values
  - But not all extreme values are influential observations!
- Influential observations example
  - Wage regressed on size: small tech companies with large wages.

Extreme values and influential observations

- What to do with them?
- Depends on why they are extreme
  - If by mistake: may want to drop them ($ 1000+)
  - If by nature: don't want to drop them (other hotel)
  - Grey zone: patterns work differently for them for substantive reasons
    - General rule: avoid dropping observations based on value of y variable
- Dropping extreme observations by x variable may be OK
  - May want to drop observations with extreme x if such values are atypical for question analyzed.
  - But often extreme x values are the most valuable as they represent informative and large variation.

Classical Measurement Error

- You want to measure a variable which is not so easy to measure:
  - Quality of the hotels
  - Inflation
  - Other latent variables with proxy measures
- Usually these miss-measurement are present due to
  - Recording errors (mistakes in entering data)
  - Reporting errors in surveys (you do not know the exact value) or administrative data (miss-reporting)
- 'Classical measurement error':
  - One of the most common and 'best' behaving problem – but a problem.
  - It needs to satisfy the followings:
    - It is zero on average (so it does not affect the average of the measured variable)
    - (Mean) independent from all variables.
- There are many other 'non-classical' measurement error, which cause problems in modelling.

Is measurement error in variables a problem?

It depends...

- Prediction: your are predicting *with* the errors - not a particular problem, but need to be addressed when predicting or generalizing.
- Association:
  - Interested in the estimated coefficient value (not just the sign)
    - Spending and income; price and distance, etc.
  - Depends on whether it is in y or in x

Is measurement error in variables a problem?

Solution?

- Often cannot do anything about it!
  - The problem is with data collection/how data is generated.
  - Exceptions:
    - you have a variable which is correlated with the error term $\rightarrow$ use multiple regression $\rightarrow$ we will see it in Chapter 10.
- If cannot do anything, what is the consequence of such errors:
  - Does measurement error make a difference in the model parameter estimates?
  - Do we expect parameters (such as OLS coefficient) to be different from what they would be without the measurement error?

Two cases for classical Measurement Error

- Classical measurement error in the dependent (y or left-hand-side) variable
  - is not expected to affect the regression coefficients.
- Classical measurement error in the explanatory (x or right-hand-side) variable
  - will affect the regression coefficients.

- We are covering how to mathematically approach this problem.
  - Show general way of thinking about *any* type of measurement error.
  - There are lot of format for measurement errors, you may want to have an idea whether it affects your regression coefficient(s):
    - If yes we call it 'biased' parameter(s).

Non-classical measurement error

- In real-life data measurement error in variables may or may not be classical
  - Very often, it is not
  - Variables measured with error may be less dispersed and have non-zero mean.
- Measurement error may be related to variables of interest
  - E.g., association between income and expenditure, where the expenditures measured by credit card spending. This measure contains systematic measurement errors for expenditure. E.g. poorer people's access to credit card is different and if they have access their expenditures based on credit card usage is quite different from e.g. rich people.
  - This often means that modelling needs to be redesigned
- Non-classical measurement error has consequences that are different.

Consequences of measurement error

- Most variables in economic and social data are measured with noise. So what is the practical consequence of knowing the potential bias?
- Estimate magnitude which affects regression estimates.
- Look for the source, think about it's nature and consider impact.
- Super relevant issue for data collection, data quality!

Hotel ratings and measurement error

- In this case study, we will try to understand how measurement error in hotel rating may be investigated with its impact somewhat understood.
- New linear regression specification: price (y) and customer rating (x). The price comparison website publishes the averages of ratings customers gave to each hotel.
  - Ratings vary between 1 and 5, 5 being excellent – measure of quality for hotels
- Show an association between price and a proxy for quality.
- The measure of customer rating is an average calculated from individual evaluations. That is a noisy measure of hotel quality
  - Theoretical reason - rating is a noisy proxy for quality
  - Technical reason - too few ratings gives large measurement errors for quality

Hotel ratings and measurement error

- The data includes the number of ratings that were used to calculate average customer ratings.
- If classical measurement error plays a role, it should play a larger role for hotels with few ratings than for hotels with many ratings.

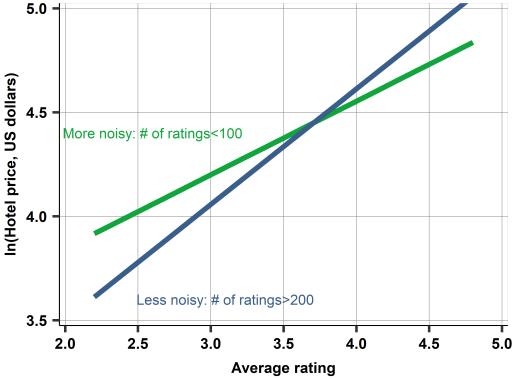
- Three groups: few, medium, many. Focus few vs many
- Few ratings – less than 100 ratings (77 hotels, with 57 ratings each on average).
- Many ratings – more than 200 ratings (72 hotels, with 417 ratings each on average).
- Average customer rating is rather similar (3.94 and 4.20). Standard deviation of the average customer ratings – lot larger among hotels with few ratings (0.42 versus 0.26).

Hotel ratings and measurement error

- We regressed (the log of) hotel price (y) on average ratings (x) separately for hotels with few ratings (less than 100) and hotels with many ratings (more than 200).
- If there is classical measurement error in average ratings, the error should be more prevalent among hotels with few ratings, and so the regression line should be flatter for few ratings

Hotel ratings and measurement error

- Log hotel price and average customer ratings.
- Hotels with noisier measure of ratings (# ratings < 100)
- Hotels with less noisy measure (# ratings > 200)



Hotel ratings and measurement error

- That is indeed what we find. The first slope coefficient is 0.35; the second one is 0.55
  - flatter, less positive slope and higher intercept among hotels with few ratings.
- There appears to be substantial measurement error in average customer ratings among hotels where that average rating is based on a few customers' reports.
  - We expect a regression with average customer ratings on the right-hand-side to produce an attenuated slope.
- Should we do anything about that? And if yes, what?

Hotel ratings and measurement error

- That is indeed what we find. The first slope coefficient is 0.35; the second one is 0.55
  - flatter, less positive slope and higher intercept among hotels with few ratings.
- There appears to be substantial measurement error in average customer ratings among hotels where that average rating is based on a few customers' reports.
  - We expect a regression with average customer ratings on the right-hand-side to produce an attenuated slope.
- Should we do anything about that? And if yes, what?
- If we are interested in the effect of ratings on prices, this is clearly an issue. Discard hotels with less than a minimum number of reviews (maybe 10 or 20 or 50 or 100 - depends on sample size)