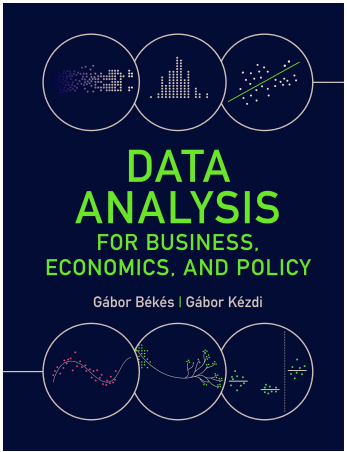
3 Generalizing regression results

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 9

Regression and statistical inference

Generalizing: reminder

- We have uncovered some pattern in our data. We are interested in generalizing 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
  - What do we want to generalize to
- Inference - the act of generalizing results
  - From a particular dataset to other situations or datasets.

- From a sample to the 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.
- $\hat{y}_i$ is the 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 we are 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 (CI)

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

Formulas: recap

- As introduced in Chpt.5:

$$Var(\bar{x}) = \frac{Var\,[x]}{n}$$

$$SE(\bar{x}) = \sqrt{\frac{Var\,[x]}{n}} = \frac{Std\,[x]}{\sqrt{n}}$$

- Variance of $\hat{\beta}$:

$$Var(\hat{\beta}) = \frac{Var\,[e]}{nVar\,[x]}$$

- SE formula of $\hat{\beta}$:

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

Heteroskedasticity robust standard error

- The 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 weights the observations when calculating the $SE[\hat{\beta}]$
  - Note: there are many heteroskedastic-robust formula, which use different weighting techniques. Usually referred as 'HC0', 'HC1', ... , 'HC4'.
    - E.g. 'HC0' in Python.

The confidence interval 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 a regression, as default, use robust SE.
  - Sometimes: homoskedastic and heteroskedasticity robust SEs are similar.
  - Sometimes: heteroskedasticity robust SE is larger – and rightly so.
- Coefficient estimates, R^2 etc. remain the same.

Testing if (true) beta is zero

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

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

- The t-statistic is:

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

- Often $t = 1.96$ (think 2) is the critical value, which corresponds to 95% CI. ($t = 2.6 \rightarrow 99\%$)

Testing if (true) beta is zero

Practical guidance:

- Choose a critical value
  - p-value, the probability of a false positive in our dataset
  - Balancing act: false positive (FP) and false negative (FN)
- Higher critical value
  - FP: less likely (less likely rejection of the null)
  - FN: more likely (high risk of not rejecting a null even though it is false)

Language: *significance* of regression coefficients

- A coefficient is said to be "statistically significant"
  - If its confidence interval does not contain zero
  - So the true value is 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 as $p < 0.05$
  - 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 the regression result will not be significant at 5% but will be at 10%.
- What not to do? Avoid language like...
  - "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

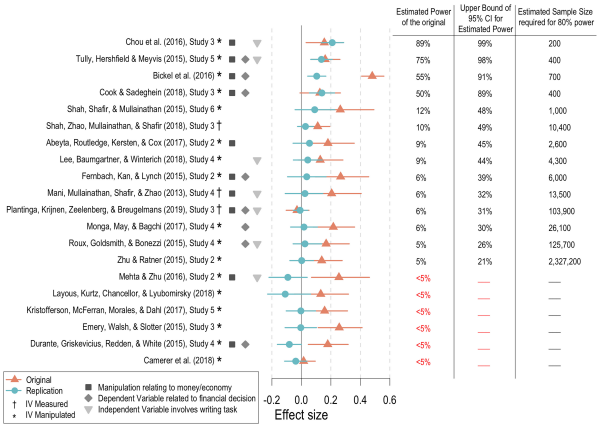
Dealing with 5-10%

- 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.
  - State 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 want to be below 1%.
  - If not, state the p-value and note that at 1% you cannot reject the null of no difference.
- Publish the p-value. Be honest ...

p-hacking

- Very often many steps lead to a regression analysis
  - Many steps: arbitrary decisions
- Often we work with a bias: looking to reinforce expectations
  - Show a "significant" result.
- p-hacking = do many versions, only showing results significant at, say, 5%
  - Danger: methods are fine, but not everything is presented
- Present your most conservative result first
  - Example: if uncertain, keep extreme values in.
- Show robustness checks: many additional regressions with different decisions
  - May add that keeping extreme values weakens findings
- p-hacking is linked to publication bias
- More: Chapter 6.9

Publication bias: example of replicating 20 psychology papers



- O'Donnell et al., "Empirical audit and review and an assessment of evidentiary value in research on the psychological consequences of scarcity", PNAS 2021
- Replication often produces a smaller effect size / wider CI

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 an 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 population/*general pattern*. 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.

- 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 (depend 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

Confidence interval of the regression line - use

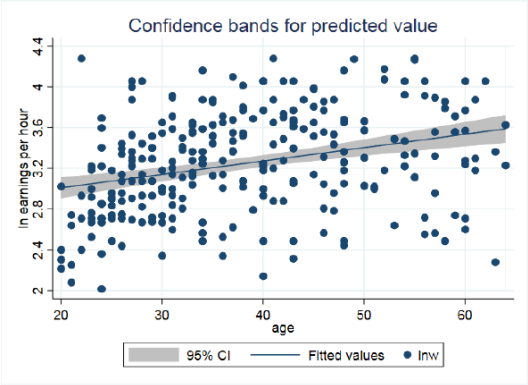
- Can be used for any model
  - Spline, polynomial
  - The way it is computed is different for different kinds of regressions
  - Always true that the CI is narrower if
    - the $Std[e]$ is smaller,
    - the n is larger and
    - the $Std[x]$ is larger
- In general, the CI for the predicted value is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

Prediction interval (PI)

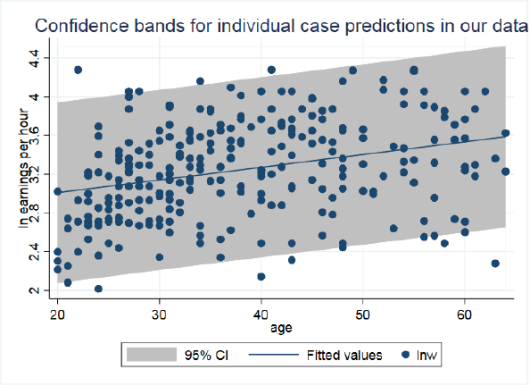
- *Prediction interval* answers:
  - Where to expect the particular y_j value if we know the corresponding x_j value and the estimates of the regression coefficients from the data.
- Difference between CI and PI:
  - The CI of the predicted value is about $\hat{y}_j$: where to expect the average value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

External validity

- Statistical inference helps us generalize to the population or to the general pattern

- Is this true beyond (other dates, countries, people, firms)?
- As external validity is about generalizing beyond what our data represents, we cannot assess it using our data.
  - We will never really know. 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, for:
  - Other dates (focus in class)
  - Other cities
  - Other type of accommodation: apartments
- Compare them to our benchmark model result.
- Learn about uncertainty when using model to check 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.

$$ln(y)^E = \alpha_1 + \beta_1 x \; 1_{x<2m} + (\alpha_2 + \beta_2 x) \; 1_{x \geq 2m}$$

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: hotels 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.

Stability of hotel prices - takeaway

- Fairly stable over time but the uncertainty is larger
- For more, read case study B in Chapter 9

- Evidence of some external validity in Vienna

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

Case study A

Case study: gender gap in earnings?

- Earnings are determined by many factors
- The idea of gender gap:
  - Is there a systematic earnings difference between male and female workers?

Case study: gender gap - how is data 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 do not measure those.
- Need to control for hours
  - Women may work systematically different 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 computer science occupation - analysis

- One key reason for gap could be women working in 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 the binary variable G 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 computer 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

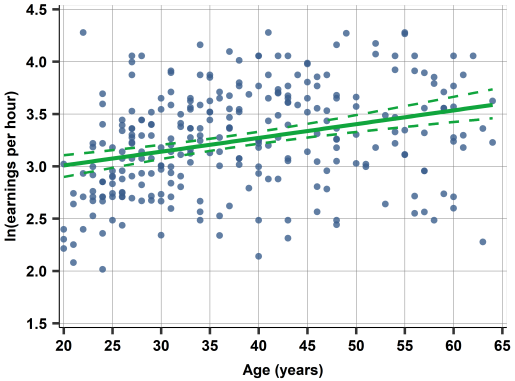
- Market research analysts and marketing specialists, $N = 281$, where females account for 61%.
  - Average hourly wage is \$29 (sd: 14.7)
- The regression estimate is $\hat{\beta} = -0.113$:
  - Female market research analyst employee earns 11.3 percent less, on average, than men with the same occupation 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 a female marketing 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.

Our two samples: what is the source of difference in CI?

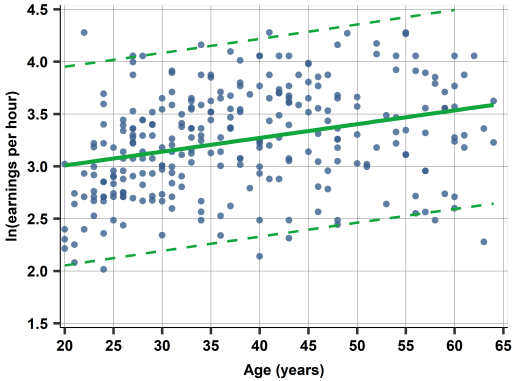
- Computer and mathematical occupations
  - 4,740 employees, female: 27.5%
  - The regression estimate of slope: -0.1475 ; 95% CI: [-.1823 -.1128]
- Market research analysts and marketing specialists
  - 281 employees, female: 61%
- The regression estimate of slope is -0.113; 95% CI: [-.23 +0.01]
- Why 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 the marketing industry.

Confidence vs prediction interval

Confidence interval

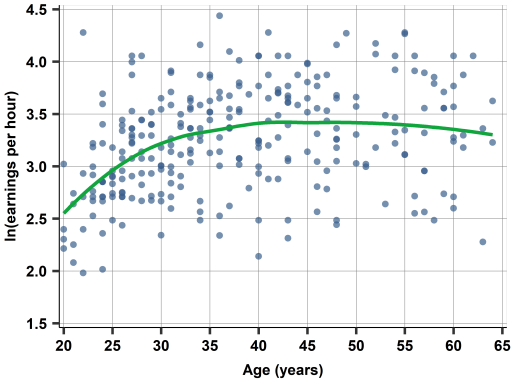


Prediction interval

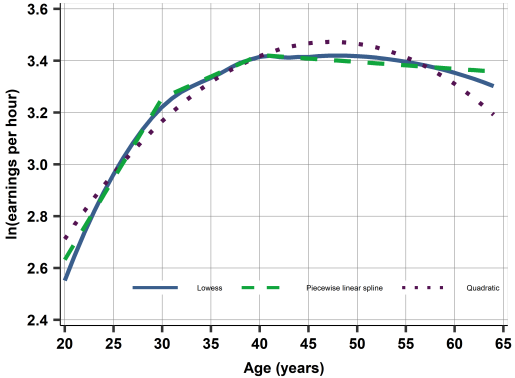


Capture non-linearity with functional form

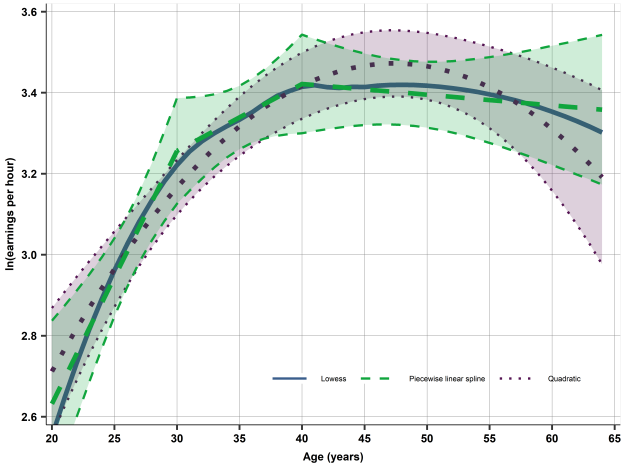
Lowess and scatterplot



Various functional forms overlaid



Average log earnings and age: regressions with CI



Testing hypotheses about regression coefficients

- Recall that the coefficient estimate was -0.11
- To formally test whether average earnings are the same by gender, we simply test if the coefficient on the binary variable is zero
  - against the alternative that it is not zero.
- Corresponding t-statistic is -1.8.
- The critical values for a 5% significance level are $\pm$ 2, and -1.8 falls within the critical values not outside of them.
- Thus, we cannot reject the null of equal average earnings at a 5% level of significance.
- Same result with the p-value: $p = 0.07 > 0.05$

Case study: earnings and age - presenting regression table

Model:

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

Variables	ln wage
Female	-0.11
	(0.062)
Constant	3.31**
	(0.049)
Observations	281
R-squared	0.012

Case study: earnings and age - presenting regression table

Model:

- ln *wage* $= \alpha + f(age)$
- 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