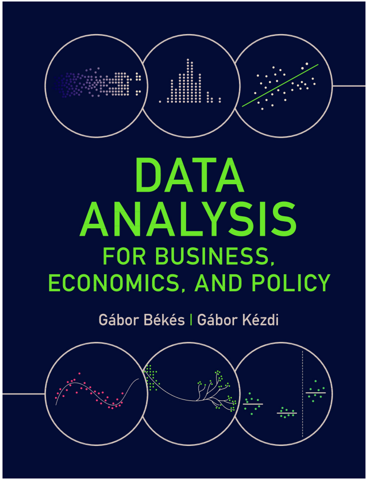
09 Generalizing regression results

Gabor Bekes

Data Analysis 2: Regression analysis

2019

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021 January
- Available in paperback, hardcover and e-book
- Slideshow be used and modified for educational purposes only
- **gabors-data-analysis.com**
  - Download all data and code
  - Additional material, links to references

Generalizing: reminder

- We have uncovered some pattern in our data. We want to generalize it
- Then the question to answer: Is the pattern we see in our data
  - true *in general*?
  - or is it just a chance event?
- 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 contained in /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 2x)
- 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 $\sqrt{n-1}$ does not matter. We'll ignore this.

Standard Error of the Slope

The simple SE formula of the slope is

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

- 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 $\sqrt{n-1}$ does not matter. We'll ignore this.

- 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,
  - larger the standard deviation of the explanatory variable,
  - 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 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
- Heteroskedasticity-robust SE formula (*White or Huber*) that 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

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**.
  - Statistical software compute both
- Coefficient estimates, R^2 etc. are the same
- In many cases, similar. In some cases, robust SE is larger – and rightly so.

Tech detour

- Always use robust standard errors ...

- For Stata, just use reg y x**, r**
- R (and Python) - bit more cumbersome
- R we use estimatR package lm_robust method
- Heteroskedasticity-robust SE formula - in practice 3 version with minor difference. Either ok.
  - Stata default is not the same as R/Python, but is what most people use.
  - in my R code, use the Stata (HC1) version
  - You can ignore and use R default (HC2, i think).

Case Study: Gender gap (in earnings)

- Earning determined by many aspects
- The idea of gender gap

Case Study: Gender gap (in earnings)

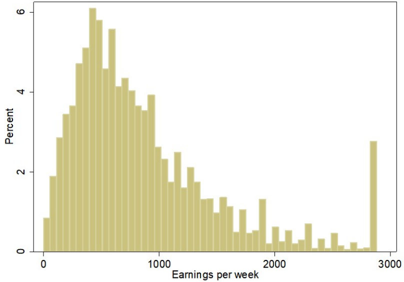
- Current Population Survey (CPS) of the U.S.
- Large sample of households
- Monthly interviews
  - Rotating panel structure: interviewed in 4 consecutive months, then not interviewed for 8 months, then interviewed again in 4 consecutive months
  - Weekly earnings asked in the "outgoing rotation group"
    - In the last month of each 4-month period
  - «morg: "Merged outgoing rotation group"
    - `http://www.nber.org/data/morg.html`
- Sample restrictions used:
  - Sample includes individuals of age 16-65
  - Employed (has earnings); self-employed excluded

Case Study: Gender gap (in earnings - data)

- Download data for 2012 (316,408 observations)
- Implement sample restrictions
  - Usual working hours non-missing and more than zero
    - (employed all that worked more than zero hour)
  - Weekly earnings non-missing and more than zero
    - (all that worked for pay)
  - Age at least 16 at most 64
  - Not self-employed
- 149,316 observations in total

Case Study: Gender gap (in earnings - data)

- Weekly earnings in CPS
  - Before tax
  - However reported (hourly, monthly, yearly etc.) converted to weekly earnings
    - Using information on hours per week, weeks per month, year, etc.
  - Top-coded very high earnings
    - at $2,884.6 (top code adjusted for inflation)
    - 2.5% of earnings in 2012
  - Would be great to measure other benefits, too (yearly bonuses, non-wage benefits). But we don't measure those.



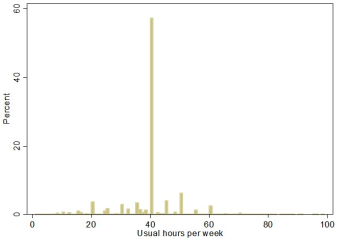
Case Study: Gender gap (in earnings - all)

Gender	mean	p25	p50	p75	p90	p95
Male	$ 988	481	800	1303	1962	2558
Female	$ 735	360	600	961	1442	1854
% gap	-26%	-25%	-25%	-26%	-26%	-28%

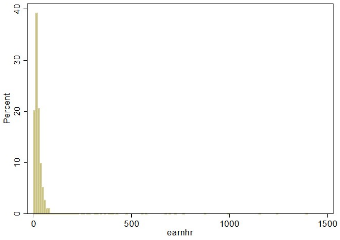
Case Study: Gender gap (in earnings - all)

- Need to control for hours
  - Women may work different hours than men
- Measure usual weekly working hours
- A lot of measurement error is likely (earnings and hours)
- Divide weekly earnings by usual weekly hours

Usual Weekly Hours



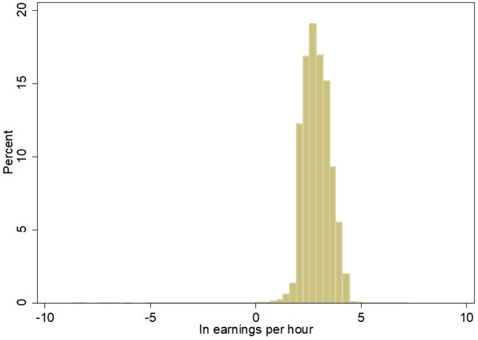
Earning per hour



Case Study: Gender gap (in earnings - all)

- Taking log
  - and keeping all observations

Usual Weekly Hours in logs



Earning per hour in logs

Case Study: Gender gap (in earnings - all)

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 (in earnings - comp science occup.)

- One key reason for gap could be women being sectors / occupations that pay less. Focus on a single one.
- Computer science occupations, $N = 4740$
- $\ln(w)^E = \alpha + \beta \times \mathbf{G}_{female}$
- We regressed log earnings per hour on G binary variable that is one if the individual is female and zero if male.
- This is a log-level regression.
  - The slope shows average differences in relative wages by gender in the dataset
- The 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.

Case Study: Gender gap (in earnings - comp science occup.)

- Our data is a random sample of all market analysts working in the U.S. in 2014.
  - The CPS is a high-quality sample with careful random sampling and high response rates.
- Use the standard tools of statistical inference to estimate the standard error, and then, the confidence interval
  - The estimated slope coefficient is -0.1475.
  - SE: .0177; 95% CI: [-.182 -.112]
    - Simple vs robust SE - Here no practical difference.

Case Study: Gender gap (in earnings - comp science occup.)

- 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 -11.3%.
- 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 earnings - mkt analyst occup.)

- Market research analysts and marketing specialists, $N = 281$
- Female: 61%
- Average hourly wage (earnings per hour) \$29 (sd:14.7)
  - Average log wage: 3.2

Case Study: Gender gap (in earnings - mkt analyst occup.)

- The regression estimate is **-0.113:**
  - female market research analyst employee earns 11 percent less, on average, than men with the same occupation in this dataset.
  - SE: .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 Beta (true) is Zero

- Testing hypotheses = decide if a statement about a general pattern is true.
- Often: Dependent variable and the explanatory variable are related at all? The null $H_0 : \beta_{true} = 0$ and the alternative $H_A : \beta_{true} \neq 0$, the t-statistic is:

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

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

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

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, ($p < 0.01$)
- Background: test theory

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]
  - Some fun: here (+R code)
- 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: [-.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?

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?

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, for which only the value of x is known.
- Linear regression – need coefficient estimates in the *general pattern* that is relevant for the observations we want to predict y for. In other words, true in the population.
- The estimated statistic here is a predicted value for a particular observation. 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
  - Prediction interval

Confidence interval of the regression line

- **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 $\hat{\beta}$.
  - 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.

Case Study: Gender gap (in earnings)

- Now look at earnings and age
- Only one industry: market research, N=281
- First look at patterns
- Then confidence interval

Case Study: Gender gap (in earnings) Regression table

- Log earnings and age
- Computer science occupation only.
- Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1.
- Source: `cps-earnings` dataset. 2014 CPS Morg.

VARIABLES	lnw
female	-0.11
	(0.062)
Constant	3.31**
	(0.049)
Observations	281
R-squared	0.012

Case Study: Gender gap (in earnings)

- Log earnings and age
  - linear
- Market research analysts
- Narrow as SE is small
- Hourglass shape
  - Smaller close mean x, mean y

Ch09_figures/Ch09_unused_figures/F9_ea

Standard error of predicted average

- Predicted average y has a standard error

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

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

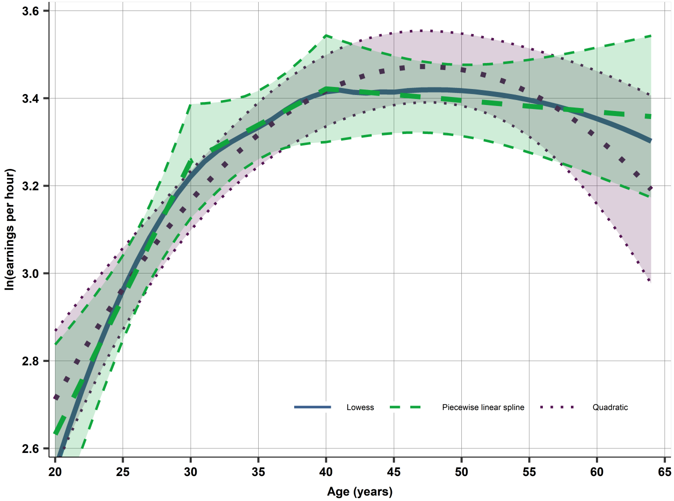
- Based on formula for regression coefficients
- It's small
  - if coefficient SE are small
  - Particular x_j coefficient is close to the mean of x
- $1/n$ emphasizing the role of sample size
- Use robust SE formula in practice, but a simple formula is instructive

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
    - 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: Gender gap (in earnings) - select fn form with CI

- Log earnings and age
  - Lowess
  - Piecewise linear spline
  - quadratic function
- Market research analysts
- 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 that actual y_j will be around its conditional mean.
- What shall we expect in graphs?

Confidence vs Prediction interval

Confidence interval

Ch09_figures/Ch09_unused_figures/F9_earnings_f5_1ci_sed.png

Prediction interval

Ch09_figures/Ch09_unused_figures/F9_ea

A bit 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)
  - PI: think about it *as if* we added Std[e] to the CI formula.

- It summarizes the additional uncertainty here: 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.
- In very large samples the standard error for average y is very small.
  - In contrast, 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 this new element.

Remember: 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

External validity: reminder

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

External validity: reminder

- 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

- But analyzing other data may 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.

B1 How stable is the hotel price - distance to center relationship?

- Here we ask a different question: 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
  - Apartments
- Compare them to our benchmark
- Learn about uncertainty when using model for beyond population

B1 How stable is the hotel price - distance to center relationship?

- Such a speculation may be relevant:
- Expand development services we offer for relatively low priced hotels.
- Find a good deal in the future without estimating a new regression but taking the results of this regression and computing residuals accordingly.

B1 How stable is the hotel price - distance to center relationship?

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

$$ln(y)^E = \alpha_1 + \beta_1 x[\text{if } x < 2m] + (\alpha_m + \beta_m x)[\text{if } x \geq 2m] \quad (1)$$

The benchmark November weekday Vienna model is

- Model has three output variables: $\alpha = 5.02$, $\beta_1 = -0.31$, $\beta_2 = 0.02$
- Hotel prices are on average 151.41 euro (exp 5.02) at no distance from center
- hotels in the data 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.
- 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.
- at 4 miles, we would have $lnprice = 5.02 - 0.31 * 2 + 0.02 * 2 = 5.60$

B1 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

B1 Comapring dates

- November weekday and the June weekend: $\beta = 0.31$
  - Among hotels in the data 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.
- Estimate is similar for December (-0.36 log units)
- It looks different for the November weekend: they are 0.44 log units or 55% ($exp(0.44) - 1$) cheaper during the November weekend.

B1 Comapring dates

- November weekday and the June weekend: $\beta = 0.31$
  - Among hotels in the data 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.
- Estimate is similar for December (-0.36 log units)
- It looks 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 dates 2 – same hotels

VARIABLES	(1) 2017-NOV-weekday	(2) 2017-NOV-weekend	(3) 2017-DEC-holiday	(4) 2018-JUNE-weekend
dist_0_2	-0.28**	-0.44**	-0.40**	-0.28**
	(0.058)	(0.055)	(0.045)	(0.053)
dist_2_7	-0.03	-0.02	-0.01	-0.03
	(0.049)	(0.041)	(0.031)	(0.039)
Constant	5.02**	5.52**	5.19**	5.12**
	(0.068)	(0.069)	(0.067)	(0.078)
Observations	98	98	98	98
R-squared	0.291	0.434	0.609	0.332

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. Same hotels only.

B1 comparing 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

B1 Comparing 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

B1 How stable is the hotel price - distance to center relationship?

- 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.