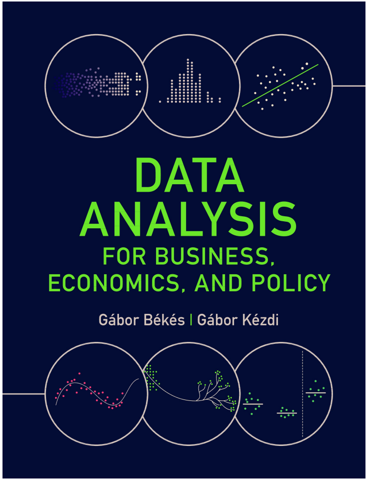
10. Multiple regression

Gabor Bekes

Data Analysis 2: Regression analysis

2019

Slideshow for the Békés-Kézdi Data Analysis textbook



DATA ANALYSIS

FOR BUSINESS, ECONOMICS, AND POLICY

Gábor Békés | Gábor Kézdi

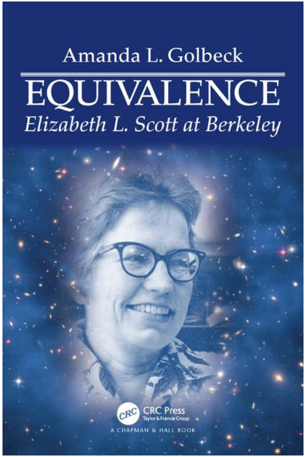
- Cambridge University Press, 2021 January
- Available in paperback, hardcover and e-book
- Slideshow be used and modified for educational purposes only
- **gabors-data-analysis.com**
  - Download all data and code
  - Additional material, links to references

Motivation

- *Find a good deal on a hotel to spend a night in a European city- analyzed the pattern of hotel price and distance and many other features to find hotels that are underpriced not only for their location but also those other features.*
- *Interested in finding evidence for or against labor market discrimination of women. Compare wages for men and women who share similarities in wage relevant factors such as experience and education.*

Motivation II

- Elizabeth Scott, a Berkeley statistics professor spent two decades analysing inequalities in academic salaries and advocating for change.
- "How one woman used regression to influence the salaries of many" by Amanda Golbeck (in Significance, Dec 2017)
  - `http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2017.01092.x/full`



Multiple regression analysis

- Multiple regression analysis uncovers average y as a function of more than one x variable: $y^E = f(x_1, x_2, ...)$.
- It can lead to better predictions $\hat{y}$ by considering more explanatory variables.
- It may improve the interpretation of slope coefficients by comparing observations that are different in terms of one of the x variables but similar in terms of other x variables.
- Multiple linear regression specifies a linear function of the explanatory variables for the average y.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$

Multiple regression

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{2}$$

- β_1 -the slope coefficient on x_1 shows difference in average y across observations with different values of x_1, *but the same value of* x_2.
    - β_2 shows difference in average y across observations with different values of x_2, *but the same value of* x_1.
- Can compare observations that are similar in one explanatory variable to see the differences related to the other explanatory variable.

Multiple regression - mechanics

Compare slope coefficient in simple (β) and multiple regression (β_1):

$$y^E = \alpha + \beta x_1 \tag{3}$$

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{4}$$

Intermediary step: regression of x_2 on x_1 (" $x - x$ regression") - δ is slope parameter:

$$x_2^E = \gamma + \delta x_1 \tag{5}$$

Multiple regression - mechanics

Compare slope coefficient in simple (β) and multiple regression (β_1):

$$y^E = \alpha + \beta x_1 \tag{3}$$

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{4}$$

Intermediary step: regression of x_2 on x_1 (" $x - x$ regression") - δ is slope parameter:

$$x_2^E = \gamma + \delta x_1 \tag{5}$$

Plug this back

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2(\gamma + \delta x_1) = \beta_0 + \beta_2 \gamma + (\beta_1 + \beta_2 \delta) x_1 \,. \tag{6}$$

It turns out that

$$\beta - \beta_1 = \delta \beta_2 \tag{7}$$

Multiple regression - mechanics

- The slope of x_1 in a simple regression is different from its slope in the multiple regression, the difference being the product of its slope in the regression of x_2 on x_1 and the slope of x_2 in the multiple regression.
- The slope coefficient on x_1 in the two regressions is different

Multiple regression - mechanics

- The slope of x_1 in a simple regression is different from its slope in the multiple regression, the difference being the product of its slope in the regression of x_2 on x_1 and the slope of x_2 in the multiple regression.
- The slope coefficient on x_1 in the two regressions is different
  - unless x_1 and x_2 are uncorrelated $(\delta = 0)$ OR
  - the coefficient on x_2 is zero in the multiple regression $(\beta_2 = 0)$.
- The slope in the simple regression is larger if x_2 and x_1 are positively correlated and β_2 is positive
  - or x_2 and x_1 are negatively correlated and β_2 is negative

Multiple regression - mechanics

- If x_1 and x_2 are correlated, comparing observations with or without the same x_2 value makes a difference.
- If they are positively correlated, observations with higher x_2 tend to have higher x_1.
- In the simple regression we ignore differences in x_2 and compare observations with different values of x_1.
- But higher x_1 values mean higher x_2 values, too.
- Corresponding differences in y may be due to differences in x_1 but also differences in x_2.

Multiple regression - some language

- Multiple regression with two explanatory variables (x_1 and x_2),
- we measure differences in expected y across observations that differ in x_1 but are similar in terms of x_2.
- Difference in y by x_1, **conditional on** x_2. OR **controlling for** x_2.
- We condition on x_2, or control for x_2, when we include it in a multiple regression that focuses on average differences in y by x_1.

Multiple regression - some language

- Multiple regression with two explanatory variables (x_1 and x_2),
- When we are interested in a regression with x_2, but we have one without: x_2 is an **omitted variable** in the simple regression.
- The slope on x_1 in the sample is confounded by omitting the x_2 variable, and thus x_2 is a **confounder**.

Multiple regression - mechanics – SE

- Inference, confidence intervals in multiple regressions is analogous to those in simple regressions.

$$SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1 - R_1^2}} \tag{8}$$

- Same: the SE is small - small Std of the residuals (the better the fit of the regression); large sample, large the Std of x_1.
- New: $\sqrt{1 - R_1^2}$ term in the denominator - the R-squared of the regression of x_1 on x_2 - correlation between x_1 and x_2.
- The stronger the correlation between x_1 and x_2 the larger the SE of $\hat{\beta}_1$.
- Note the symmetry: the same would apply to the SE of $\hat{\beta}_2$.
- Also: in practice, use robust SE

Multiple regression - mechanics – collinearity

- **Perfectly collinearity** is when x_1 is a linear function of x_2.

Multiple regression - mechanics – collinearity

- **Perfectly collinearity** is when x_1 is a linear function of x_2.
- Consequence: cannot calculate coefficients.
  - One will be dropped by software
-
- Strong but imperfect correlation between explanatory is sometimes called **multicollinearity**.
- Consequence: We can get the slope coefficients and their standard errors,
  - The standard errors may be large.

Multiple regression - mechanics – collinearity

- Multicollinearity – recognized also by standard errors may be large.
- Reason: Few variables that are different in x_1 but not in x_2. Not enough observations for comparing average y across them.

- This is a small sample problem.

- May look at pair-wise correlations when start working with data
- Drop one or the other, or combine them (score).

Multiple regression - joint significance

- *Testing joint hypotheses*: null hypotheses that contain statements about more than one regression coefficient.
- We aim at testing whether a subset of the coefficients (such as all geographical variables) are all zero.
- F-test answers this.
  - Individually they are not all statistically different from zero, but together they may be.

- We may ask if *all slope coefficients are zero* in the regression.
- "Global F-test", and its results are often shown by statistical software by default. Don't use it, R-squared is fine.

Multiple regression - many explanatory variables

- Having more explanatory variables is no problem.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... \tag{9}$$

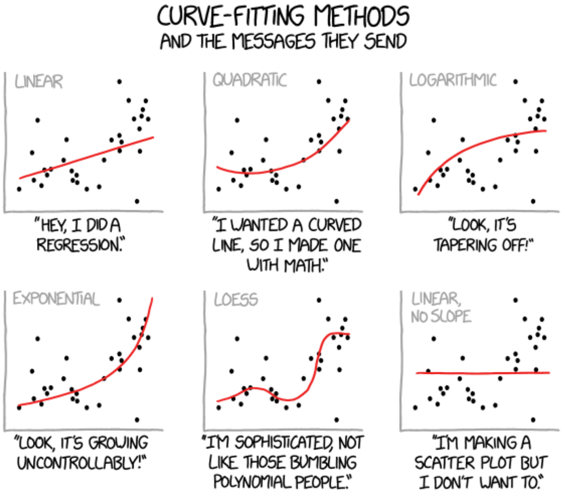
- Interpreting the slope of x_1: on average, y is β_1 units larger in the data for observations with one unit larger x_1 but the same value for all other x variables.
- SE formula - small when R_k^2 is small - R^2 of regression of x_k on all *other* x variables.

$$SE(\hat{\beta}_k) = \frac{Std[e]}{\sqrt{n} Std[x_k] \sqrt{1 - R_k^2}} \tag{10}$$

Multiple regression - non-linear patterns

- Uses splines, polynomials - actually like multiple regression - we have multiple coefficient estimates
- No fear of multicollinearity - not *linear* combinations
- Non-linear function of various x_i variables may be combined.

Multiple regression - non-linear patterns


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."

Multiple regression - qualitative variables

- Can have binary variables as well as other qualitative variables (factors)
- Consider a qualitative variable like continents. How to add it to the regression model?

Multiple regression - qualitative variables

- Can have binary variables as well as other qualitative variables (factors)
- Consider a qualitative variable like continents. How to add it to the regression model?
- Create binary variables (dummy variables) for all options. Add them - all but one.
- This one will be the base

Multiple regression - qualitative variables

- x is a categorical variable with three values *low*, *medium* and *high*
- binary variable m denote if $x = medium$, h variable denote if $x = high$.
- for $x = low$ is not included. It is called the *reference category* or left-out category.

$$y^E = \beta_0 + \beta_1 x_{medium} + \beta_2 x_{high} \tag{11}$$

Multiple regression - qualitative variables

$$y^E = \beta_0 + \beta_1 x_{med} + \beta_2 x_{high} \tag{12}$$

- Pick $x = low$ as the reference category. Other values compared to this.
  - This is the omitted variable
- β_0 shows average y in the reference category. Here, β_0 is average y when both $x_{med} = 0$ and $x_{high} = 0$: this is when $x = low$.
- β_1 shows the difference of average y between observations with $x = medium$ and $x = low$
- β_2 shows the difference of average y between observations with $x = high$ and $x = low$.

Multiple regression - qualitative variables

How to pick a reference category?

- Substantive guide: choose the category to which we want to compare the rest.
  - Examples include the home country, the capital city, the lowest or highest value group.
- The statistical guide: chose a category with a large number of observations.
  - Important when inference is important.
  - If reference category has few observations - coefficients will have large SE / wide CI.

Multiple regression - Interactions

- Many cases, data is made up of important groups: male and female workers or countries in different continents.
- Some of the patterns we are after may vary across these groups.
- The strength of a relation may also be altered by a special variable.
  - In medicine, a *moderator variable* can reduce / amplify the effect of a drug on people.
  - In business, financial strength can affect how firms may weather a recession.
- All of these mean different patterns for subsets of observations.

Multiple regression - Interactions

- Regression with two explanatory variables: x_1 is continuous, D is binary denoting two groups in the data (e.g., male or female employees).
- We wonder if the relationship between average y and x_1 is different for observations with $D = 1$ than for $D = 0$. How?

Multiple regression - qualitative variables

- Option 1: Two *parallel lines* for the y - x_1 pattern: one for those with $D = 0$ and one for those with $D = 1$.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 D \tag{13}$$

Two groups, $D = 0$ and $D = 1$. Difference is the level

$$y_0^E = \beta_0 + \beta_2 \times 0 + \beta_1 x_1 \tag{14}$$

$$y_1^E = \beta_0 + \beta_2 \times 1 + \beta_1 x_1 \tag{15}$$

Multiple regression - qualitative variables

- Option 2: If we want to *allow for different slopes* in the two D groups we have to do something different, add an interaction term.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 x_1 D \tag{16}$$

Intercepts different by β_2 AND slopes different by β_3.

$$y_0^E = \beta_0 + \beta_1 x_1 \tag{17}$$

$$y_1^E = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_1 \tag{18}$$

Multiple regression - Interactions

- Separate regressions in the two groups and the regression that pools observations but includes an interaction term yield *exactly the same* coefficient estimates.
- The coefficients of the separate regressions are easier to interpret.
- But the pooled regression with interaction allows for a direct test of whether the slopes are the same.

- Extension: D1, D2 are binaries, x continuous:

$$y^E = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 x + \beta_4 D_1 x + \beta_5 D_2 x \tag{19}$$

Interaction with two continuous variable

- Same model used for two continuous variables, x_1 and x_2:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{20}$$

- Example: Firm level data, 100 industries.
- y is change in revenue x_1 is change in global demand, x_2 is firm's financial health
- The interaction can capture that drop in demand can cause financial problems in firms, but less so for firms with better balance sheet.

A1 Understanding the gender difference in earnings

- In the USA (2014), women tend to earn about 20% less than men
- Aim 1: Find patterns to better understand the gender gap. Our focus is the interaction with age.
- Aim 2: Think about if there is a causal link from being female to getting paid less.

Case Study - Gender gap in earnings

- 2014 census data
  - Age between 15 to 65
  - Exclude self-employed (earnings is difficult to measure)
  - Include those who reported 20 hours more as their usual weekly time worked
- Employees with a graduate degree (higher than 4-year college)
- Use log hourly earnings (lnw) as dependent variable
- Use gender and add age as explanatory variables

Gender gap in earnings

We are quite familiar with the relation between earnings and gender:

$$\ln w^E = \alpha + \beta female, \qquad \beta < 0$$

Let's include age as well:

$$\ln w^E = \beta_0 + \beta_1 female + \beta_2 age$$

We can calculate the correlation between female and age, which is in fact negative.

What do you expect about β, β_1, δ?
Reminder:

$$age^E = \gamma + \delta female$$

Gender gap regression - baseline

VARIABLES	(1) lnw	(2) lnw	(3) age
female	-0.195**	-0.185**	-1.484**
	(0.008)	(0.008)	(0.159)
age		0.007**	
		(0.000)	
Constant	3.514**	3.198**	44.630**
	(0.006)	(0.018)	(0.116)
Observations	18,241	18,241	18,241
R-squared	0.028	0.046	0.005

Note: *All employees with a graduate degree. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1*

Source: `cps-earnings` dataset. 2014 CPS Morg.

Interpretations and connections

$$\beta - \beta_1 = \delta\beta_2$$

which can be calculated easily:

- $\beta - \beta_1 = -0.195 - (-0.185) = -0.01$
- $\delta\beta_2 = -1.48 \times 0.007 \approx -0.01$

Interpretation:

- Age is a confounder, it is different from zero and the beta coefficient changes.
- But a weak one.

Interpretations and connections

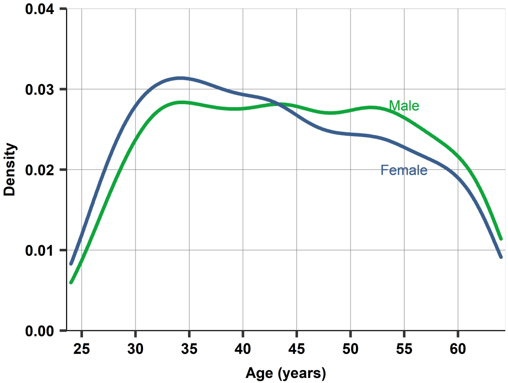
- women of the same age have a slightly smaller earnings disadvantage in this data because they are somewhat younger, on average
- employees that are younger tend to earn less
- part of the earnings disadvantage of women is thus due to the fact that they are younger.
  - This is a small part: around 1 percentage points of the 20% difference,
  - = a 5% share of the entire difference.

Interpretations and connections

- A single linear interaction may not be enough.
- Next: drill down the impact of age

Age distributions

Age distribution of male and female employees with degrees higher than college



- Relatively few below age 30
- Above 30
  - close to uniform for men
  - for women, the proportion of female employees with graduate degrees drops above age 45, and again, above age 55
- Two possible things
  - fewer women with graduate degrees among the 45+ old than among the younger ones
  - fewer of them are employed

A3 Understanding the gender difference in earnings

- Maybe age as confounder is non-linear.
- Extend our analysis with including higher orders of age

A3 Understanding the gender difference in earnings

- Maybe age as confounder is non-linear.
- Extend our analysis with including higher orders of age

- Not much difference re female variable
- However R^2 increases as we include higher orders.

	(1)	(2)	(3)	(4)
VARIABLES	lnw	lnw	lnw	lnw
female	-0.195**	-0.185**	-0.183**	-0.183**
	(0.008)	(0.008)	(0.008)	(0.008)
age		0.007**	0.063**	0.572**
		(0.000)	(0.003)	(0.116)
agesq			-0.001**	-0.017**
			(0.000)	(0.004)
agecu				0.000**
				(0.000)
agequ				-0.000**
				(0.000)
Constant	3.514**	3.198**	2.027**	-3.606**
	(0.006)	(0.018)	(0.073)	(1.178)
Observations	18,241	18,241	18,241	18,241
R-squared	0.028	0.046	0.060	0.062

Note: **** $p<0.01$, ** $p<0.05$, * $p<0.1$*

Interaction between gender and age

- Why we assume that age has the same slope regardless of gender? We might want to check, whether they are different!

Interaction between gender and age

- Why we assume that age has the same slope regardless of gender? We might want to check, whether they are different!

- Are the slopes significantly different?
- Can one get the slope for age for female only from the regression with the interaction?
- How the gender dummy's coefficient changed?

Interaction between gender and age

- Look men and women separately. Earning for men rises faster with age
- Or, look at them pooled with interaction.
  - Observe that pooling with interaction is the SAME as two separate models.
- Constant is close to zero

	(1)	(2)	(3)
	WOMEN	MEN	ALL
VARIABLES	lnw	lnw	lnw
female			-0.036
			(0.035)
age	0.006**	0.009**	0.009**
	(0.001)	(0.001)	(0.001)
female X age			-0.003**
			(0.001)
Constant	3.081**	3.117**	3.117**
	(0.023)	(0.026)	(0.026)
Observations	9,685	8,556	18,241
R-squared	0.011	0.028	0.047

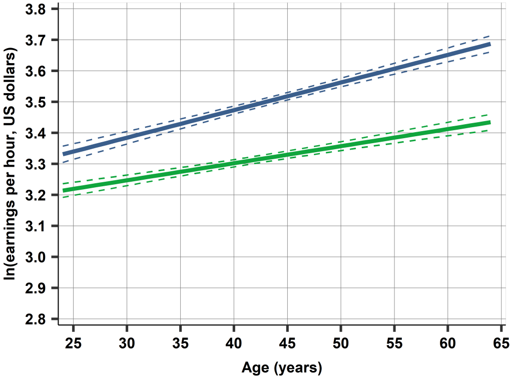
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Nonlinearities and interactions

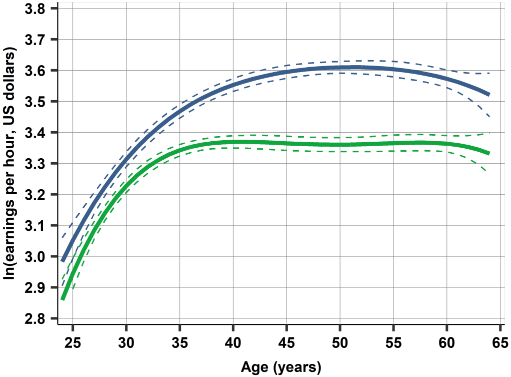
We estimate

$$
\begin{aligned}
lnw^E = \beta_0 &+ \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 age^4 \\
&+ \beta_5 female + \beta_6 female \times age + \beta_7 female \times age^2 \\
&+ \beta_8 female \times age^3 + \beta_9 female \times age^4
\end{aligned}
$$

Nonlinearities and interactions



Log earnings per hour and age by gender: predicted values and confidence intervals from a linear regression interacted with gender.



Log earnings per hour and age by gender: predicted values and confidence intervals from a regression with 4th-order polynomial interacted with gender.

Understanding the gender difference in earnings

- the average earnings difference is around 10% between ages 25 and 30
- increases to around 15% by age 40, and reaches 22% by age 50,
- from where it decreases slightly to age 60 and more by age 65.
- confidence intervals around the regression curves are rather narrow, except at the two ends.
- Conclusion?

Multiple regression - Causal analysis

- One main reason to estimate multiple regressions is to get closer to a causal interpretation.
- By conditioning on other observable variables, we can get closer to comparing similar objects – „apples to apples" – even in observational data.
- But getting closer is not the same as getting there.
- In principle, one may help that by conditioning on *every* potential confounder: variables that would affect y and the causal variable x_1 at the same time.
- Ceteris paribus = conditioning on **every** such relevant variable. .

Multiple regression - Causal analysis

- Ceteris paribus = conditioning on **every** such relevant variable. .
- *Ceteris paribus* prescribes what we want to condition on; a multiple regression can condition on **what's in the data** the way it is measured.
- Importantly, conditioning on everything is impossible in general.
- Multiple regression is never (hardly ever) ceteris paribus

Multiple regression - Causal analysis

- In randomized experiments, we use causal language, as treated and untreated units similar - by random grouping.
- In observational data, comparisons don't uncover causal relations.
  - Cautious with language. No use of "effect", "increase".
  - Regression, even with multiple x is just comparison. Conditional mean.

Multiple regression - Causal analysis

- Not all variables should be included as control variables even if correlated both with the causal variable and the dependent variable.
- *Bad conditioning variables* are variables that are correlated both with the causal variable and the dependent variable but are actually part of the causal mechanism.
- This is the reason to exclude them. .
- Example, when we want to see how TV advertising affects sales. Should control for how many people viewed the advertising?

Multiple regression - Causal analysis

- Not all variables should be included as control variables even if correlated both with the causal variable and the dependent variable.
- *Bad conditioning variables* are variables that are correlated both with the causal variable and the dependent variable but are actually part of the causal mechanism.
- This is the reason to exclude them. .
- Example, when we want to see how TV advertising affects sales. Should control for how many people viewed the advertising?
  - No. Part of why less advertising may hurt sales - fewer heads.
- Super hard

Multiple regression - Causal analysis

- A multiple regression on observational data is rarely capable of uncovering a causal relationship.
  - Cannot capture all potential confounder. (Not ceteris paribus)
  - Potential bad controls
  - We can never really know. BUT
- multiple regression can get us **closer** to uncovering a causal relationship
  - Compare units that are the same in many respects - controls

A6 Understanding the gender difference in earnings - Causal analysis

What may cause the difference in wages?

- Labor discrimination - one group earns less even if they have the same *marginal product*
- Try control for marginal product (or for variables which matters to marginal product)
  - Eg.: occupation (as an indicator for inequality in gender roles), or industry, union status, hours worked and other socio-economic characteristics
- Use variables as controls - does comparing apples to apple change coefficient of female variable?

Causal analysis - results

- More and more confounders added
- Female coefficient reduced from 22% to 14%
- Compare two people, with same age, hours, industry, occupation, geography, background (=confounders) - women earn 14% less, on average.

	(1)	(2)	(3)	(4)
VARIABLES	ln wage	ln wage	ln wage	ln wage
Female	-0.224**	-0.212**	-0.151**	-0.141**
	(0.012)	(0.012)	(0.012)	(0.012)
Age and education		YES	YES	YES
Family background			YES	YES
Hours worked			YES	YES
Government or private			YES	YES
Union member			YES	YES
Not born in USA				YES
Age in polynomial				YES
Hours in polynomial				YES
Observations	9,816	9,816	9,816	9,816
R-squared	0.036	0.043	0.182	0.195

Restricted sample: employees of age 40 to 60 with a graduate degree that work 20 hours per week or more

Discussion

- Could not safely pin down the role of labor market discrimination and broader gender inequality
- Multiple regression - *closer* to causality
- but, broader gender inequality seem to matter: inequality in gender roles likely plays a role in the fact that women earn less per hour than men.
- we cannot prove that that the remaining 14% is due to discrimination - *plenty of remaining heterogeneity*
- Also: Selection and bad controls?

Multiple regression - prediction and benchmarking

- Reason to estimate a multiple regression is to make a *prediction*
- find the best guess for the dependent variable y_j for a particular *target observation* j

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \ldots \tag{21}$$

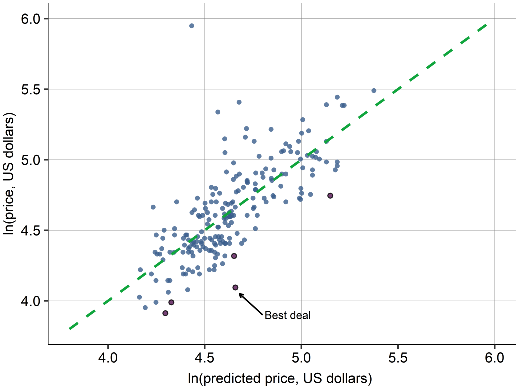
- When the goal is prediction we want the regression to produce as good a fit as possible.
- As good a fit as possible in the general pattern that is representative of the target observation j.
- A common danger is *overfitting* the data: finding patterns in the data that are not true in the general pattern.

Multiple regression - benchmarking

- The $\hat{y} - y$ *plot* has $\hat{y}$ on the horizontal axis and y on the vertical axis.
- The plot features the 45 degree line and the scatterplot around it = the regression line of y regressed on $\hat{y}$.
- The scatterplot around this line shows how actual values of y differ from their predicted value $\hat{y}$.

B1 Finding a good deal among hotels with multiple regression

- Hotel prices, many predictor variables
- Estimate model (R2=0.56), and get $\hat{y} - y$
- $\hat{y} - y$ plot for log hotel price
- What are the good deals?



Source: `hotels` dataset. Vienna, 2017 November, weekday.

Multiple regression - Variable selection

- How should one decide which variables to include and how?
- Depends on the purpose: prediction or causality.
- Lot of judgement calls to make
  - Very hard task. No super-duper solution.
- Non-linear fit - use a non-parametric first and if non-linear, pick a model that is close - quadratic, piecewise spline.
- If two or many variables strongly correlated, pick one of them. Sample size will help decide.

Multiple regression - Variable selection for causal questions

- Causal question in mind x impact on y. Having z variables to condition on, to get closer to causality.
- Our aim is to focus on the coefficient on one variable. What matter here are the estimated value of the coefficient and its confidence interval.
- Keep z – keep many variables that help comparing apples to apples
- Drop z if they not matter
- Functional form for z matters only for crucial confounders
- Present the model you judge is best, and then report a few other solutions – robustness.

Multiple regression - Variable selection – process

- Select control variables you want to include
- Select functional form one by one
- Focus on key variables by domain knowledge, add the rest linearly

- Key issue is sample size
  - For 20-40 obs, about 1-2 variables.
  - For 50-100 obs, about 2-4 variables
  - Few hundred obs, 5-10 variables could work
  - Few thousand obs, few dozen variables, including industry/country/profession etc dummmies, interactions.
  - 10-100K obs - many variables, polynomials, interactions

Multiple regression - Variable selection for predictiom

- IF Prediction - keep whatever works
- Balance is needed to ensure it works beyond the data at hand
- Overfitting: building a model that captures some patterns that may fit the data we have but would not generalize to the data we use the prediction for.
- Focus on functional form, interactions
- Value simplicity. Easier to explain, more robust.
- Formal way: BIC ("Bayesian Information Criterion"). Similar to R-squared but takes into account number of variables.
  - The smaller, the better
  - Do not use adjusted R-squared