1.4 Multilinear Regression

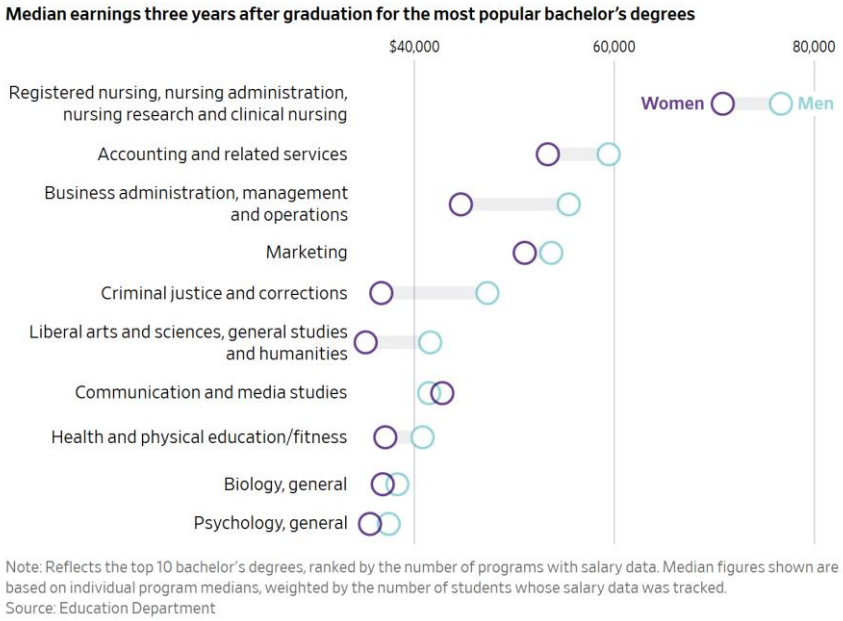
How do we model the conditional mean and interpret the results when there are multiple variables of interest?

How large is the gender wage gap?

THE WALL STREET JOURNAL.

Data Show Gender Pay Gap Opens Early

Disparities among male and female college graduates appeared within three years, a WSJ analysis of federal data for 2015 and 2016 graduates shows



Link to article (2022)

"Nationally, women across the workforce earn an average of 82.3 cents for every dollar a man earns, according to the Labor Department."

The New York Times

Women Earn $2 Million Less Than Men in Their Careers as Doctors

A survey of more than 80,000 physicians estimated that women make 25 percent less than men over a 40-year career.

Link 2021

"The salary gaps began at the beginning of a doctor's career and continued to widen until around Year 10 without recovering, the study found."

Background Case Study

Case II:
Understanding the Gender Pay Gap

- How large is the gender pay gap?
- What important factors explain this gap?

	lnw			
	(1)	(2)	(3)	(4)
female	-0.224***	-0.212***	-0.150***	-0.140***
	(0.012)	(0.012)	(0.012)	(0.012)
Constant	3.590***	3.542***	3.852***	-56.715
	(0.008)	(0.049)	(0.079)	(37.108)
Observations	9816	9816	9816	9816
S.E. type	hetero	hetero	hetero	hetero
R^2	0.036	0.043	0.182	0.195
Adj. R^2	0.035	0.043	0.168	0.181

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

How do we go from initial question to running these regressions?

Stata:
reg lnw female , robust
reg lnw female age edProf edPhd, robust
reg lnw female $DEMOG $FAMILY $WORK, robust
reg lnw female $DEMOG $FAMILY $WORK agesq-agequ hourssq-hoursqu, robust

Lecture Outline

1) What is a multiple linear regression? How do we interpret the β_i coefficients?

2) How do multiple linear regression and simple linear regression differ?

→ **Omitted Variable Bias**

3) Notes on standard errors and hypothesis testing

4) Modifications to your multilinear regression

→
- **Including polynomials and non-linear patterns**
- **Qualitative independent variable**
- **Interaction terms**

Multilinear Regression (I)

Start with the two-variable version of multilinear regression:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- How do we interpret β_0? The conditional mean of y when both x_1 and x_2 are 0.

- Let's now look at an example: $\ln wage^E = \beta_0 + \beta_1 years\_edu + \beta_2 age$

What is the predicted conditional mean of log wage for a person age 30 and 14 years of schooling?

What about age 31 when you have 14 years of schooling?

Multilinear Regression (I)

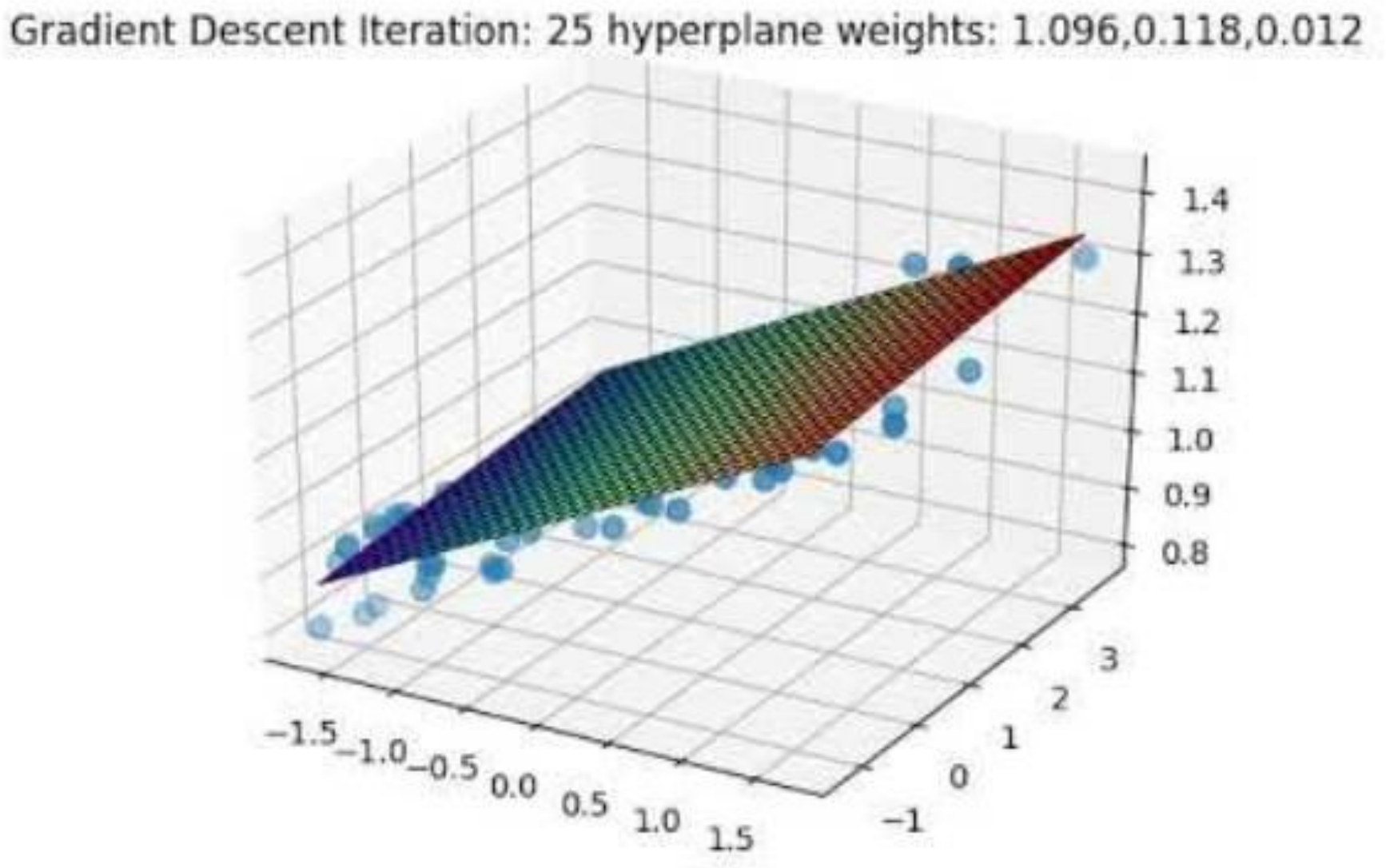
Start with two variable version of multilinear regression:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Interpretation of β_1

- On average, y is β_1 units larger for observations with one unit larger x_1 but with the same x_2.
- Differences in expected mean of y conditional on or controlling for other explanatory variables (covariates or confounder variables)

See video of hyperplane example.



Link to video.

Case Study: Measuring the Gender Wage Gap

Variables	(1) ln wage	(2) ln wage	(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

How do you interpret the coefficients of model 2?

- Women earn, on average, about 19.5% less than men, holding other variables constant.

Note: Since the logs are larger, the approximation is not very precise. Difference of 19.5 log points is about 21%

- People who are one year older tend to have a 0.7% higher wage on average.

Lecture Outline

1) What is a multiple linear regression? How do we interpret the β_i coefficients?

2) How do multiple linear regression and simple linear regression differ?

→ Omitted Variable Bias

3) Notes on standard errors and hypothesis testing

4) Modifications to your multilinear regression

→
- Including polynomials and non-linear patterns
- Qualitative independent variable
- Interaction terms

Multilinear vs Simple (I)

Let's compare our multilinear regression

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad \text{Eq. A}$$

with a simple regression $y^E = \alpha + \beta x_1$ Eq. B

To give more context, suppose that:

y is the log of sales price.

x_1 is the log price of your product.

x_2 is the log price of your competitor's product.

Relationship between prices $x_2^E = \gamma + \delta x_1$ Eq. C

Multilinear vs Simple (II)

Combining the three equations and simplifying:

$$y^E = \beta_0 + \beta_2\gamma + (\beta_1 + \beta_2\delta)x_1$$

Recall that the slope of the simple one was β. Thus

$$\beta = \beta_1 + \beta_2\delta$$

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

The difference between the two slopes is $\beta_2\delta$, also called the omitted variable bias.

Example 1: Pricing and Volume of Sales

$$\ln sales_{own}{}^{E} = \alpha - 0.5 \ln price_{own}$$

$$\ln sales_{own}{}^{E} = \beta_0 - 3 \ln price_{own} + 3 \ln price_{comp}$$

$$\ln price_{own}{}^{E} = \gamma + 0.83 \ln price_{comp}$$

Discussion Questions

1) Interpret the key slope estimates from the three equations.
2) The bias $\beta - \beta_1 = \beta_2 \delta = 3 * 0.83 \approx 2.5$ is a positive bias. Intuitively does this make sense?

- Recall: $\beta = \beta_1 + \beta_2 \delta$

Example 2: How Large is the Gender Wage Gap?

Variables	(1) ln wage	(2) ln wage	(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

Békés & Kézdi (2021) Table 10.1

Discussion Questions

1) Interpret the key slope estimates from the three equations.
2) What is the bias? Intuitively does the sign of the bias make sense given the third column?

Lecture Outline

1) What is a multiple linear regression? How do we interpret the β_i coefficients?

2) How do multiple linear regression and simple linear regression differ? → Omitted Variable Bias

3) Notes on standard errors and hypothesis testing

4) Modifications to your multilinear regression →
- Including polynomials and non-linear patterns
- Qualitative independent variable
- Interaction terms

Simple Standard Error Formula

The SE in the multilinear regression case:

$$SE(\widehat{\beta_1}) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$$

The blue highlight is different from the SE in the simple linear regression case.

- Why R_1^2? This is the R^2 of the regression of x_1 on x_2 (when we have two variables).
  - Measures correlation between x_1 and x_2 (and other variables if more than two variables).
  - The stronger the correlation, the larger the SE.
  - Why does this intuitively make sense?

Multicollinearity

$$SE(\widehat{\beta_1}) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1 - R_1^2}}$$

- *What happens when variables are perfectly colinear?*
- *What if they are very strongly correlated?*

Example: Running categorical variables

```
.  reg lnw female male age double_age agesq
note: female omitted because of collinearity.
note: age omitted because of collinearity.
```

Source	SS	df	MS
Model	149.353881	3	49.7846269
Residual	626.797473	2,351	.266608878
Total	776.151354	2,354	.329715953

Number of obs	= 2,355
F(3, 2351)	= 186.73
Prob > F	= 0.0000
R-squared	= 0.1924
Adj R-squared	= 0.1914
Root MSE	= .51634

lnw	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	0	(omitted)				
male	.3947785	.0221001	17.86	0.000	.3514408	.4381163
age	0	(omitted)				
double_age	.0429699	.0040565	10.59	0.000	.0350153	.0509245
agesq	-.0008502	.0000938	-9.06	0.000	-.0010342	-.0006663
_cons	1.001249	.1670579	5.99	0.000	.6736527	1.328845

For illustration purpose, “double_age” is age multiplied by 2.

Software randomly drops one of the variables.

Hypothesis Testing in Multiple Linear Regression

To test $H_0: \beta_1 = 0, H_A: \beta_1 \neq 0$, we use same procedure as the simple linear regression!

- Calculate the t-statistic
- Calculate the p-value
- Reject or accept

$\Rightarrow$ significant levels at 0.05 (5% or *), 0.01 (1% or **), and 0.001 (0.1% or ***).

Lecture Outline

1) What is a multiple linear regression? How do we interpret the β_i coefficients?

2) How do multiple linear regression and simple linear regression differ? → Omitted Variable Bias

3) Notes on standard errors and hypothesis testing

4) Modifications to your multilinear regression →
- Including polynomials and non-linear patterns
- Qualitative independent variable
- Interaction terms

Adding polynomials

Variables	(1) ln wage	(2) ln wage	(3) ln wage	(4) ln wage
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)
age^2			−0.001**	−0.017**
			(0.000)	(0.004)
age^3				0.000**
				(0.000)
age^4				−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

Békés & Kézdi (2021) Table 10.2

Interpretation does not change from the simple linear regression.

- Except now we add "keeping other variables constant".

Note: in this example, our coefficient on female does not change too much with changes in the functional form of age.

Adding qualitative variables in the regression

Variables	(1) ln wage	(2) ln wage	(3) ln wage
female	−0.195**	−0.182**	−0.182**
	(0.008)	(0.009)	(0.009)
ed_Profess		0.134**	−0.002
		(0.015)	(0.018)
ed_PhD		0.136**	
		(0.013)	
ed_MA			−0.136**
			(0.013)
Constant	3.514**	3.473**	3.609**
	(0.006)	(0.007)	(0.013)
Observations	18 241	18 241	18 241
R-squared	0.028	0.038	0.038

Békés & Kézdi (2021) Table 10.3

- First, choose a reference category (otherwise you have issues with multicollinearity).
- Interpretation will be based on the referenced category.
- From Model 2:

"Having a professional degree is associated with having an average wage that is 13.4% higher than those with just a master's degree."

Is this consistent with Model 3?

When should we add interaction terms?

Variables	(1) Women	(2) Men	(3) All
female			−0.036
			(0.035)
age	0.006**	0.009**	0.009**
	(0.001)	(0.001)	(0.001)
female × 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

Békés & Kézdi (2021) Table 10.4

When to do:

- if we think different groups have different slopes, captured in one regression.
- From article in the beginning of slides: “The salary gaps began at the beginning of a doctor’s career and continued to widen until around Year 10 without recovering, the study found.”

Math behind the interaction terms

$$y^E = \beta_0 + \beta_1 female + \beta_2 age + \beta_3 age * female$$

What happens when female = 0?

$$y^E(female = 0) = \beta_0 + \beta_2 age$$

What happens when female = 1?

$$y^E(female = 1) = \beta_0 + \beta_1 female + \beta_2 age + \beta_3 age * female$$

Note that there are two additions: one is due to the sole “female” variable. The second is due to differences of the slope on “age” for females.

Case Study Practice: The Best Deal Hotel

You'll continue the gender wage gap case study for your homework.

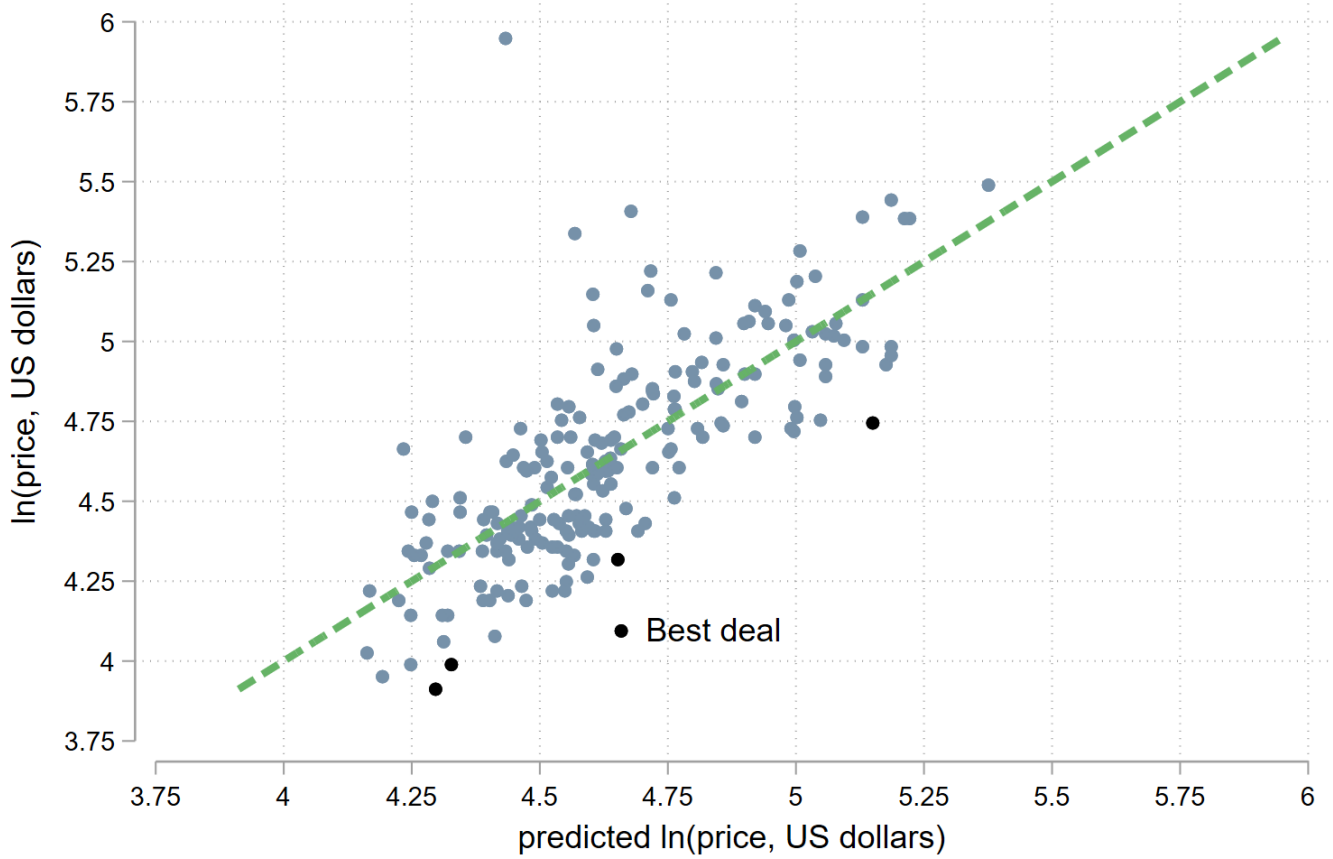
For today, let's go back to Case Study I!

Open Lecture Code 1.2 Hotel.

We will continue this code:

a) Replicate the following regressions
b) Replicate the picture
c) Get the list of best deal hotel

Follow instructions step-by-step & discussion.



How do we go from initial question to running this regression?

Stata: "reg lnprice distsp1 distsp2 distsp3 star35 star4 ratingsp1 ratingsp2, robust"

Reference

- Agresti, A. (2018). *Statistical methods for the social sciences* (5th ed.). Pearson.
- Békés, G., & Kézdi, G. (2021). In *Data Analysis for Business, Economics, and Policy* (pp. iii–iii). title-page, Cambridge: Cambridge University Press.