5: Basic Regression

ModernDive r4ds book club

Fundamental premise of data modeling

To make explicit the relationship between:

- an outcome variable y (dependent variable, response variable)
- an explanatory/predictor variable x (independent variable, covariate)

Fundamental premise of data modeling

To make explicit the relationship between:

- an outcome variable y (dependent variable, response variable)
- an explanatory/predictor variable x (independent variable, covariate)

Explanatory Modeling

- quantify the relationship between y and x
- determine the significance of any relationships
- have measures summarizing these relationships
- identify any causal relationships between the variables

Predictive Modeling

- Can I make good predictions of y from x?

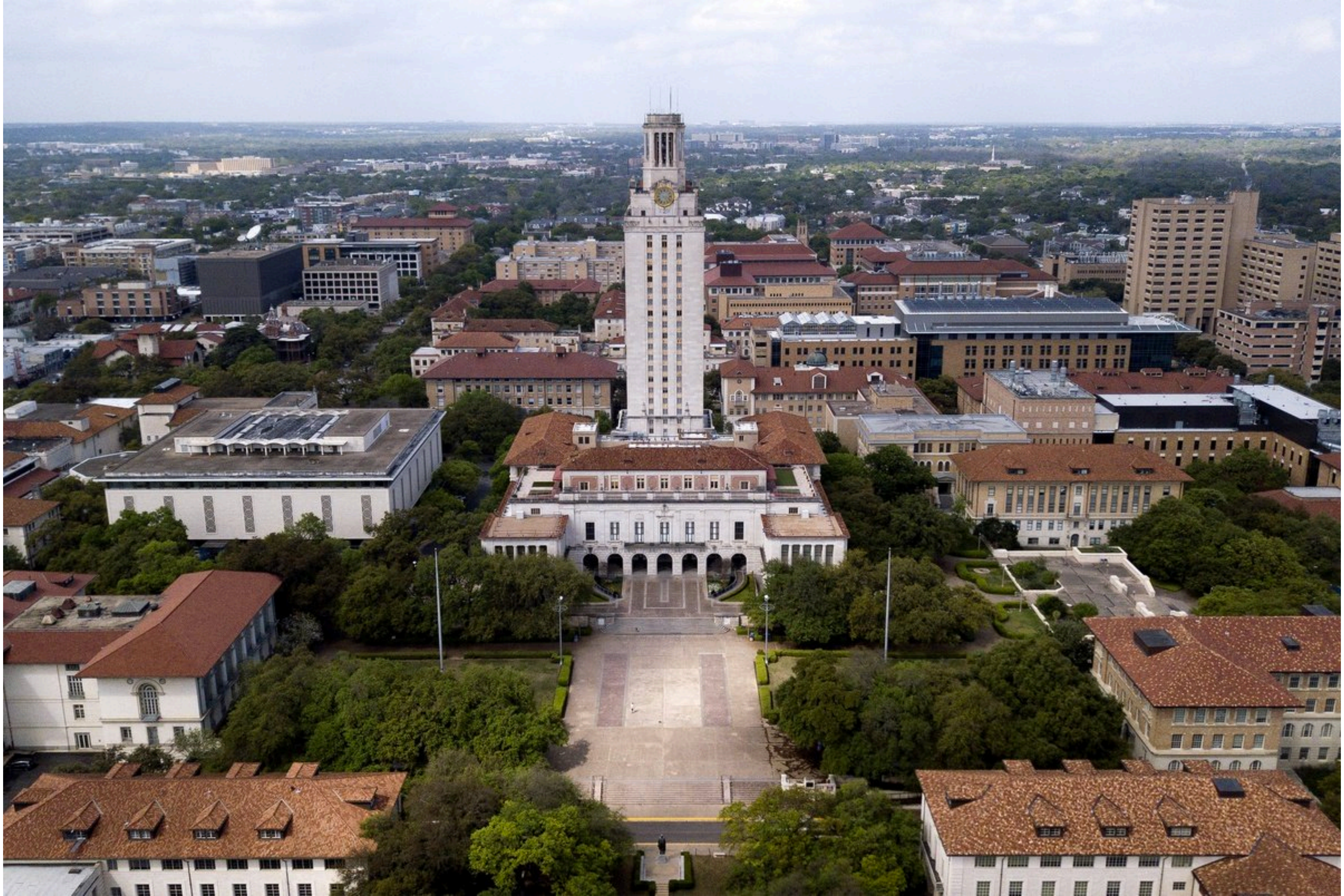

Lung
Cancerous
tumor
Affected
lymph node

explanatory: describing and quantifying smoking risk factors predictive: can we predict whether someone will develop lung cancer?

“Linear regression involves a *numerical* outcome variable y and explanatory variables x that are either *numerical* or *categorical*…the relationship between y and x is assumed to be linear.”

“basic regression” refers to linear regression models with a single explanatory variable x

5.1 One numerical explanatory variable



y = teaching score
x = beauty score

“A crucial step before doing any kind of analysis or modeling is performing an *exploratory data analysis*”

3 steps of EDA

- Looking at raw values
- Computing summary statistics
- Creating data visualizations

LC5.1

Conduct a new exploratory data analysis with the same outcome variable y being `score` but with `age` as the new explanatory variable x

What can you say about the relationship between age and teaching scores based on this exploration?

$$y = m \cdot x + b$$

$$y = m \cdot x + b$$

$$\hat{y} = b_0 + b_1 \cdot x$$

$$y \mapsto \hat{y} \quad m \mapsto b_1$$

$$\hat{y} = b_0 + b_1 \cdot x$$

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	3.880	0.076	50.96	0	3.731	4.030
bty_avg	0.067	0.016	4.09	0	0.035	0.099

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	3.880	0.076	50.96	0	3.731	4.030
bty_avg	0.067	0.016	4.09	0	0.035	0.099

“…it has no *practical* interpretation since observing a bty_avg of 0 is impossible”

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	3.880	0.076	50.96	0	3.731	4.030
bty_avg	0.067	0.016	4.09	0	0.035	0.099

“For every increase of 1 unit in bty_avg, there is an *associated* increase of, *on average,* .067 units of score”

LC5.2

Fit a new simple linear regression using `lm(score ~ age, data = evals_ch5)` where `age` is the new explanatory variable x.

Get information about the “best-fitting” line from the regression table by applying the `get_regression_table()` function. How do the regression results match up with the results from your earlier exploratory data analysis?

Relationship of teaching and beauty scores

ID	score	bty_avg	score_hat	residual
21	4.9	7.33	4.37	0.531
22	4.6	7.33	4.37	0.231
23	4.5	7.33	4.37	0.131
24	4.4	5.50	4.25	0.153

"…a 'best-fitting' line refers to the line that minimizes the sum of squared residuals out of all possible lines we can draw through the points"

LC5.3

Generate a data frame of the residuals of the model where you used `age` as the explanatory x variable.

5.2 One categorical explanatory variable

Bubbles
FACTS
TEACH
ABOUT
HOW TO USE
Share
English
INCOME LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
years
90
80
70
60
50
40
30
20
Life expectancy
2018
per person (GDP/capita, PPP$ inflation-adjusted)
500
1000
2000
4000
8000
16k
32k
64k
Income
1800
1900
2000
Color
World Regions
Select
Search...
Afghanistan
Albania
Algeria
Andorra
Angola
Antigua and Barbuda
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Size
Population
Zoom
OPTIONS
EXPAND
PRESENT

LC5.4

Conduct a new exploratory data analysis with the same explanatory variable x being `continent` but with `gdpPercap` as the new outcome variable y. What can you say about the differences in GDP per capita between continents based on this exploration?

“Our model will not yield a ‘best-fitting’ regression line like [before], but rather *offsets* relative to a baseline for comparison”

```r
lifeExp_model <- lm(lifeExp ~ continent, data = gapminder2007)
get_regression_table(lifeExp_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	54.8	1.02	53.45	0	52.8	56.8
continentAmericas	18.8	1.80	10.45	0	15.2	22.4
continentAsia	15.9	1.65	9.68	0	12.7	19.2
continentEurope	22.8	1.70	13.47	0	19.5	26.2
continentOceania	25.9	5.33	4.86	0	15.4	36.5

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	54.8	1.02	53.45	0	52.8	56.8
continentAmericas	18.8	1.80	10.45	0	15.2	22.4
continentAsia	15.9	1.65	9.68	0	12.7	19.2
continentEurope	22.8	1.70	13.47	0	19.5	26.2
continentOceania	25.9	5.33	4.86	0	15.4	36.5

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	54.8	1.02	53.45	0	52.8	56.8
continentAmericas	18.8	1.80	10.45	0	15.2	22.4
continentAsia	15.9	1.65	9.68	0	12.7	19.2
continentEurope	22.8	1.70	13.47	0	19.5	26.2
continentOceania	25.9	5.33	4.86	0	15.4	36.5

LC5.5

Fit a new linear regression using `lm(gdpPercap ~ continent, data = gapminder2007)` where `gdpPercap` is the new outcome variable `y`.

Get information about the "best-fitting" line from the regression table by applying the `get_regression_table()` function. How do the regression results match up with the results from your previous exploratory data analysis?

LC5.6

Using either the sorting functionality of RStudio's spreadsheet viewer or using the data wrangling tools you learned in Chapter 3, identify the five countries with the five smallest (most negative) residuals? What do these negative residuals say about their life expectancy relative to their continents' life expectancy?

LC5.7

Repeat this process, but identify the five countries with the five largest (most positive) residuals. What do these positive residuals say about their life expectancy relative to their continents' life expectancy?

LC5.8

Note in Figure 5.13 there are 3 points marked with dots and:

- The “best” fitting solid regression line in blue
- An arbitrarily chosen dotted red line
- Another arbitrarily chosen dashed green line

Compute the sum of squared residuals by hand for each line and show that of these three lines, the regression line in blue has the smallest value

LC5.8