UCLA Anderson
SCHOOL of MANAGEMENT

Prescriptive Models & Data Analytics

Francisco Castro

Plan for today

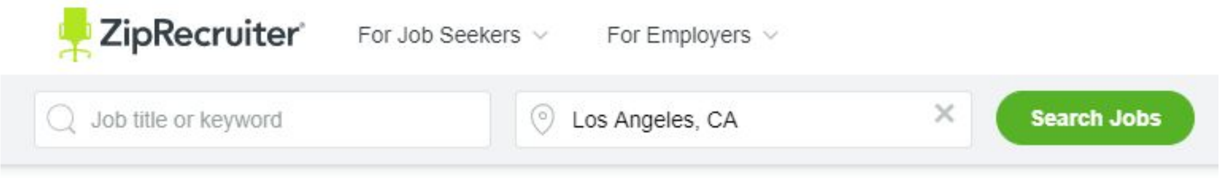
- Motivating examples of analytics influencing business decisions
- Course structure & logistics
- Recap of regression and some basic stats
- Practice of concepts using Jupyter and Python

Motivating examples

- Purpose of examples:
    - Highlight business decisions where analytics can make a difference
    - Highlight some methods that we will cover in this course

Example (1): ZipRecruiter optimal pricing

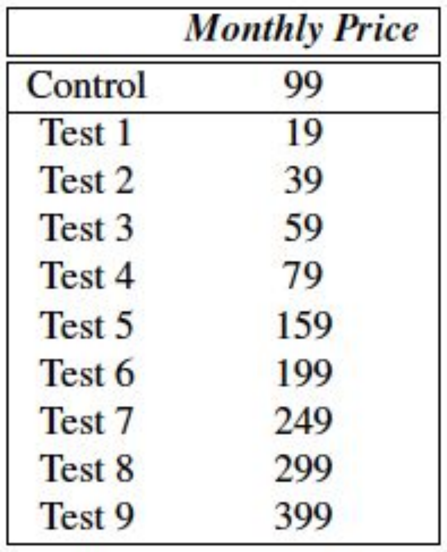
- ZipRecruiter is a platform / two-sided market
  - Firms pay monthly subscription to post jobs
  - Job seekers can find jobs for free



- Managerial question: <u>How much should we charge firms?</u>
  - Before this study ZipRecruiter had set a price of $99.
  - **Q:** Is this too high / too low / about right?
  - **A:** Depends on what is the demand function (high price = high revenue but may lose customers)

Experiment

- The firm ran a pricing experiment to estimate demand:
  - Randomly assign 10 different prices to new potential customers



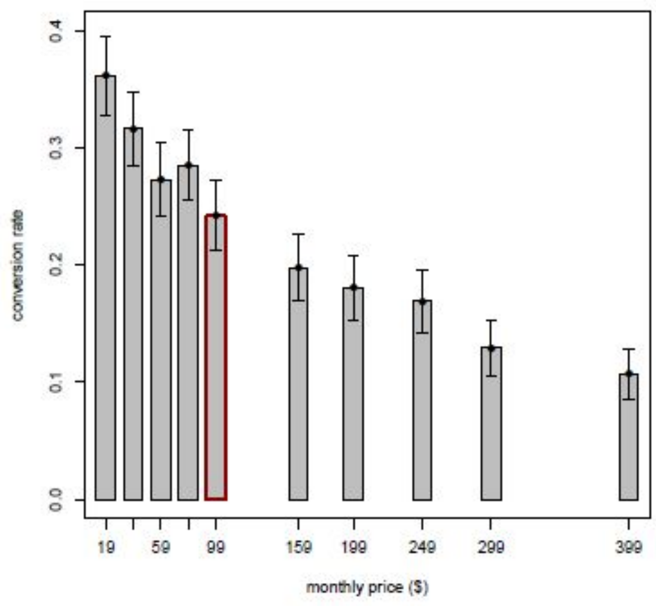
	Monthly Price
Control	99
Test 1	19
Test 2	39
Test 3	59
Test 4	79
Test 5	159
Test 6	199
Test 7	249
Test 8	299
Test 9	399

  - Estimate demand (conversion rate for each price) and revenue per customer

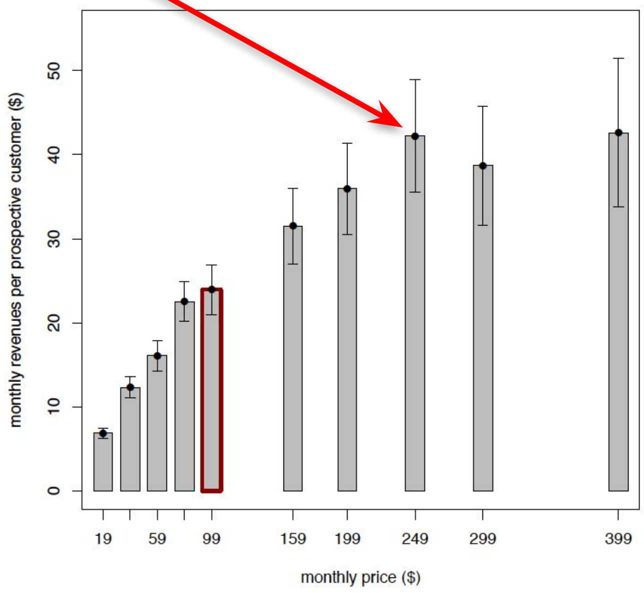
Findings

- Demand doesn’t drop very fast
- Optimal price: $249
- Price personalization (via LASSO regression) increases profits by an additional 19% relative to uniform pricing

Demand



Revenue per customer



Example (2): eBay ad effectiveness

- eBay tried to measure the effectiveness of their advertising, in particular the ROI on “paid links”
- In the past, eBay had tracked people that clicked on such a link and computed total revenue from those customers

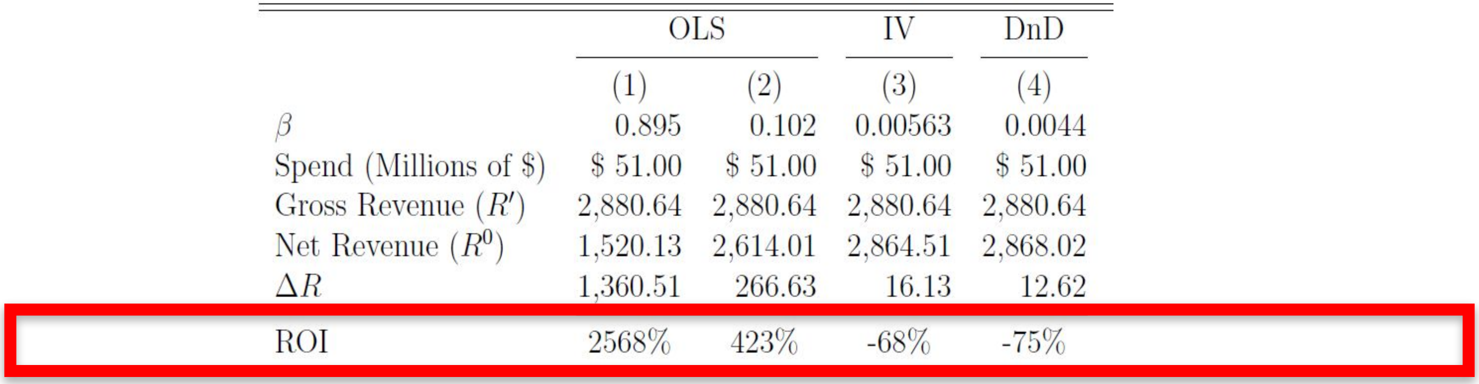
Q: Are there any issues with counting the number of users clicking on the paid link?

A: We need to ask: ”What would these users have done in the absence of the paid link?”

If users would have clicked on the natural link, you are picking up that effect as well

This is an example of the problem of “endogeneity” in empirical studies that you will learn about in the course

Correcting for endogeneity



	OLS		IV	DnD
	(1)	(2)	(3)	(4)
β	0.895	0.102	0.00563	0.0044
Spend (Millions of \$)	\$ 51.00	\$ 51.00	\$ 51.00	\$ 51.00
Gross Revenue (R')	2,880.64	2,880.64	2,880.64	2,880.64
Net Revenue (R^0)	1,520.13	2,614.01	2,864.51	2,868.02
ΔR	1,360.51	266.63	16.13	12.62
ROI	2568%	423%	-68%	-75%

- Column (1) is the naïve estimate. Columns (3) and (4) use methods we learn in this course to correct for endogeneity
  - Good data analysis can have a HUGE impact on profit

Q: The approach of column (1) is still very commonly used. Why?

A: Marketing professionals may have an incentive to convince you that their ads have large effects

→Understanding this is important for managers that rely on statistical analysis made by others!

For those of you interested in advertising…

- (Optional) Check the podcast: “Does advertising actually work?”, where you’ll also hear from the author of the study:
  - https://freakonomics.com/podcast/does-advertising-actually-work-part-1-tv-ep-440/
  - https://freakonomics.com/podcast/does-advertising-actually-work-part-2-digital-ep-441/

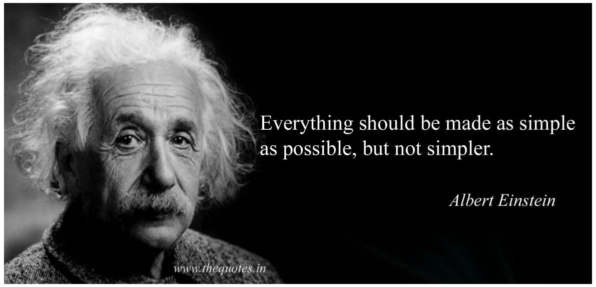
- You’ll learn:
  - Difficult to measure because of endogeneity
  - Incentives of marketing industry (and maybe of published empirical studies) to overestimate effect of ads
  - When done properly, research shows that many firms over-spend
  - What to do if marketing professionals try to “out-jargon” you…

Summary

- Big impact of analytics on managerial decisions and profits
- The examples use methods we discuss in this course:
    - A/B testing (experiments)
    - Ways to control for endogeneity when you can’t run experiments: Control variables, Diff-In-Diff
    - Big-data methods (LASSO)

Goal of this course

- Learn advanced analytics techniques to tackle business decisions
  - Focus on applications to real business problems
  - Statistics / coding only means-to-an-end
  - But: hard to understand complex issues without understanding what is “under the hood”



Everything should be made as simple as possible, but not simpler.

Albert Einstein

- At the end of the course you will be able to
  - Understand and critique analyses done by a data scientist
  - Use an appropriate empirical strategy when faced with a business challenge

Plan for today

- Motivating examples of analytics influencing business decisions
- Course structure & logistics
- Recap of regression and some basic stats
- Practice of concepts using Jupyter and Python

Course structure

1. Causal inference
   - First best approach: A/B tests (weeks 2/3)
   - If not possible: Controls for other influences (week 4), Panel Data (week 5), Diff-in-Diff (week 6), Instrumental Variables (week 7), Regression Discontinuity Design (week 8), Causal Lasso and Causal Tree (weeks 9 and 10)

2. Big data methods
   - Conceptual issues: Over-fitting, cross-validation (week 9)
   - LASSO (week 9)

3. Wrapping up
   - When to use which method (week 10)
   - Exam review (week 10)

Course administration and logistics

- Workflow:
  - No required readings before class. Plan on 30-60 minutes prep time to review key concepts from previous lectures
  - 1st part of class: lecture-based material (usually with application)
  - 2nd part of class (for most classes): you will be divided in small groups and work together on data analyses
  - After class: 3 problem sets (roughly every two weeks)
  - Classes build on each other, so be careful not to fall behind !!!

- Logistics:
  - All material will be available through Canvas
  - I will provide replication files with code for lectures and problem sets

Course administration and logistics (2)

- Grading:
  - Homeworks: 50%
  - Class attendance: 10% (taken at a random time during class)
  - Final Exam: 40%

- Homeworks: work in groups (randomly assigned for each homework), but hand-in individually
  - Problem set 1: A/B tests Jan 27th
  - Problem set 2: Control Variables Feb 11th
  - Problem set 3: Panel Data & DnD Feb 25th
  - Problem set 4: Lasso Regression Mar 16th

- Final exam: individual 3-hour open book / open notes exam administered in person
  - March 22, 8:30-11:30 am

Attendance

https://forms.gle/YH3cbaMU5X799ZLW7

Link will expire soon!

Readings

- Lectures are self-contained and understanding the material from lectures, workshops and homeworks is enough for the exam

- All other suggested readings are optional
  - Some are academic papers (if you want to discuss the papers please send me an email and we will find some time chat)
  - Others are on methodology (from textbooks)

- Optional textbooks:
  - Business Data Science, By Taddy
  - Mastering ’Metrics: The Path from Cause to Effect, by Angrist and Pischke
  - Mostly Harmless Econometrics: An Empiricist’s Companion, by Angrist and Pischke
  - Introduction to Econometrics (1st Edition), by Stock and Watson

Office hours

- By appointment. Please feel free to send me an email to set up a meeting any time!
- You should ask me all **conceptual questions**
- The TAs, Jian (Section 1) and Martin (Section 2), will hold office hours:
    - Section 1: See Canvas (starting this week)
    - Section 2: See Canvas (starting this week)
- For TAs office hours, do you prefer Zoom or in-person?
- Jian and Martin will answer all questions about **homeworks + coding**

Plan for today

- Motivating examples of analytics influencing business decisions
- Course structure & logistics
- Recap of regression and some basic stats
    - Interpreting regression coefficients
    - Precision: standard errors, t-stats, p-values

- Goals:
    - Be comfortable with interpreting regression output (coefficients, standard errors, p-values, etc.)
    - Secondary: (re-)familiarize with python's statistical tools

Data Example: Red Meat Consumption

From the NY Times (citing Harvard medical school study):

"Eating red meat is associated with a sharply increased risk of death from cancer and heart disease. The analysis … involved 121,342 men and women who filled out questionnaires about health and diet from 1980 through 2006."

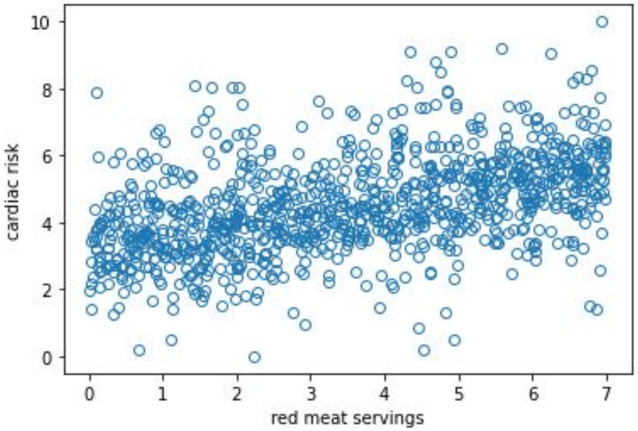
Q: Could there be any issues with the study?

A: People that eat red meat could be different in many other ways:

"People who ate more red meat were less physically active and more likely to smoke and had a higher body mass index, researchers found."

Univariate regression (one X-variable)

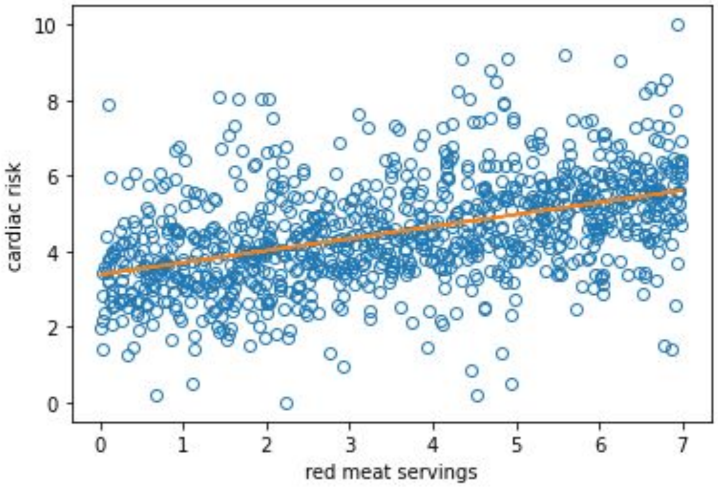
- For each person in the dataset, plot their value of red meat on the X (horizontal) axis and their cardiac risk score on the Y (vertical) axis



- The graph (scatterplot) suggests:
  - Consumption of red-meat is positively correlated with cardiac risk
  - Careful: we cannot say red-meat causes increase in cardiac risk: we will discuss this in depth next week

The regression line

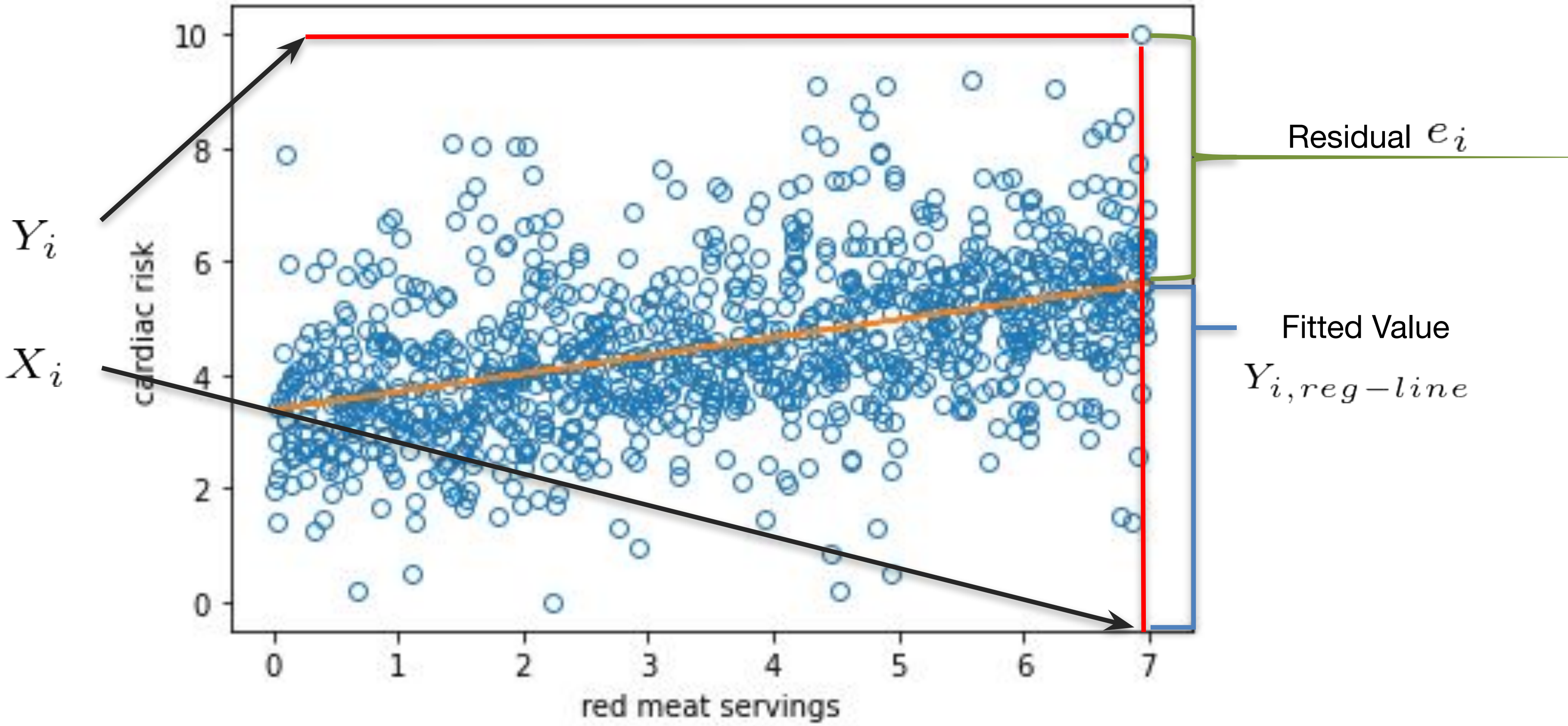
- We want to fit a line to the scatterplot



- We can represent **any** line as: $Y_{i,reg-line} = \beta_0 + \beta_1 X_i$
- We can describe each person "i" in the scatterplot as

$$Y_i = (\beta_0 + \beta_1 X_i) + e_i = Y_{i,reg-line} + e_i$$

Fitted values and residuals



Y_i

X_i

Residual e_i

Fitted Value

$Y_{i,reg-line}$

How to determine regression coefficients and fit

- Estimate β_0 and β_1 by minimizing the distance of observations to the line, i.e. minimizing the sum of squared errors (Ordinary Least Squares estimator)

$$\min_{\beta_0,\beta_1} \sum e_i^2 = \min_{\beta_0,\beta_1} \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$$

- Measure the fit of the regression by the R-squared
- Equals fraction of the outcome variance explained by regression
- Varies from 0 (no explanatory power) to 1 (perfect fit)

$$R^2 = 1 - \frac{Var(e)}{Var(Y)}$$

Python and Jupyter Notebook

- We will use Jupyter Notebooks throughout the course

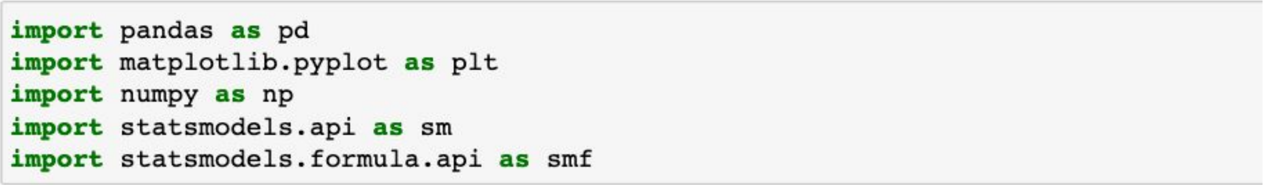
- Jupyter Notebook contains text and “code-chunks”

- I have uploaded a Notebook file that contains the code we used during today’s lecture

- How to operate Jupyter Notebook
  - You can run code chunks separately
  - More important: can compile a word or pdf document that contains text, commands, and python output (regression results etc.):
    - Install pandoc: https://pandoc.org/installing.html
    - Word: pandoc jupyter_file.ipynb -s -o new_word_file.docx
    - Pdf: File>Download> as Pdf via Latex
- You can use R in this environment too

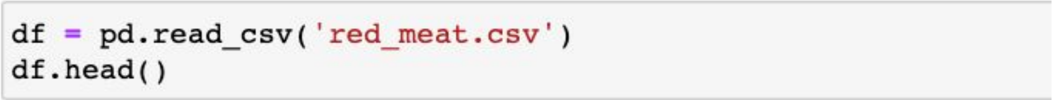
First steps in Jupyter Notebook

- Setup
  - Download red_meat.csv data to directory of your choice
  - Create a Notebook in the same directory
- Import:
  - modules



```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

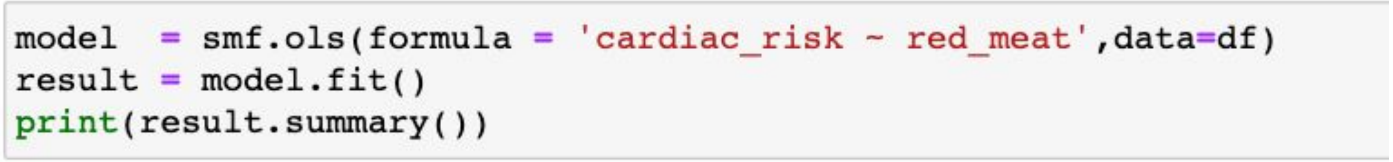
  - data set



```
df = pd.read_csv('red_meat.csv')
df.head()
```

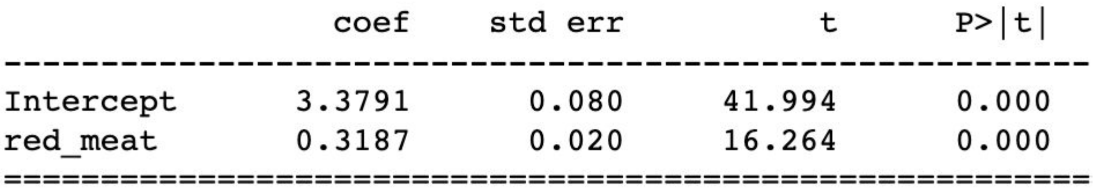
First regression in Python

Commands:



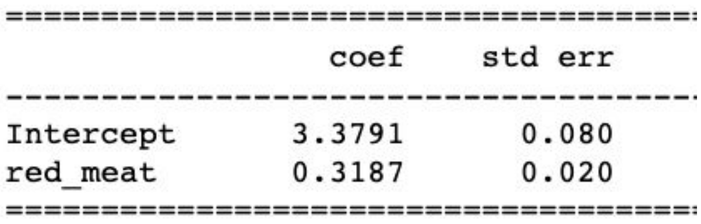
```
model  = smf.ols(formula = 'cardiac_risk ~ red_meat',data=df)
result = model.fit()
print(result.summary())
```

Output:

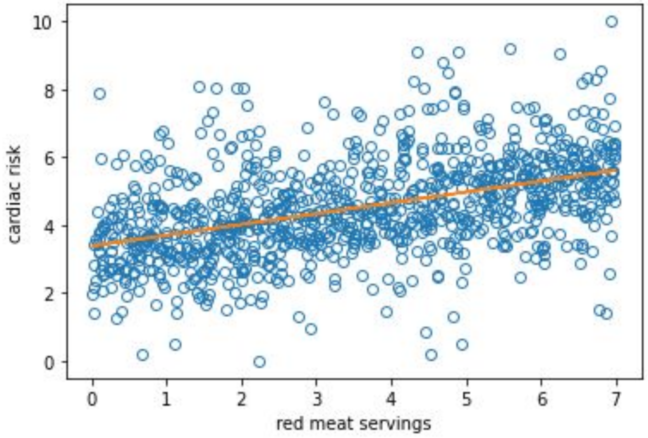


```
                 coef    std err          t      P>|t|
------------------------------------------------------
Intercept      3.3791      0.080     41.994      0.000
red_meat       0.3187      0.020     16.264      0.000
======================================================
```

Interpreting intercept and slope coefficient



```
=====================================
                 coef      std err
-------------------------------------
Intercept       3.3791       0.080
red_meat        0.3187       0.020
=====================================
```



Q: What is the interpretation of the intercept?

A: People who don’t eat red meat have an expected cardiac risk score of 3.379

Q: What is the interpretation of the red meat coefficient estimate?

A: We estimate that expected cardiac risk score increases by 0.31 points when consuming one additional serving of red meat per week

slido

What is the difference in risk between eating 3 and 4 servings of red meat?

ⓘ Start presenting to display the poll results on this slide.

Poll answer

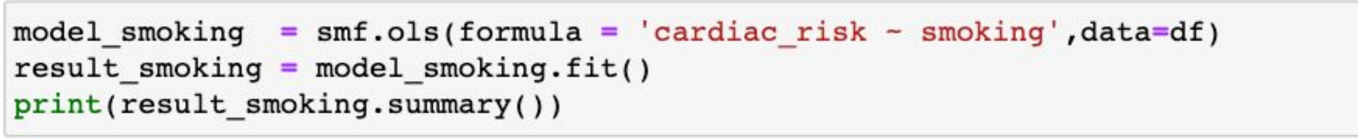
- Answer is 0.31869
- Because of linearity, the effect of increasing red meat consumption by one unit is the same regardless of the starting point

Q: How to capture the possibility that the effect of red meat is different at high levels than at low levels of consumption?

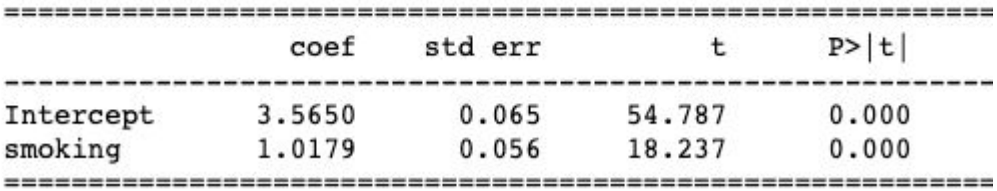
A: Fit a quadratic polynomial by including both red_meat and red_meat^2 as regressors

Other regressions: smoking & cardiac risk

Q: Run a regression of cardiac risk on smoking (packs per day). Interpret the intercept and the coefficient estimate on smoking



```
model_smoking  = smf.ols(formula = 'cardiac_risk ~ smoking',data=df)
result_smoking = model_smoking.fit()
print(result_smoking.summary())
```



```
==============================================================
                 coef    std err          t      P>|t|
--------------------------------------------------------------
Intercept      3.5650      0.065     54.787      0.000
smoking        1.0179      0.056     18.237      0.000
==============================================================
```

A: We estimate that smoking one extra pack per day increases expected risk score by 1.01 points. Expected risk score for non-smokers is 3.56

Binary X (dummy variable)

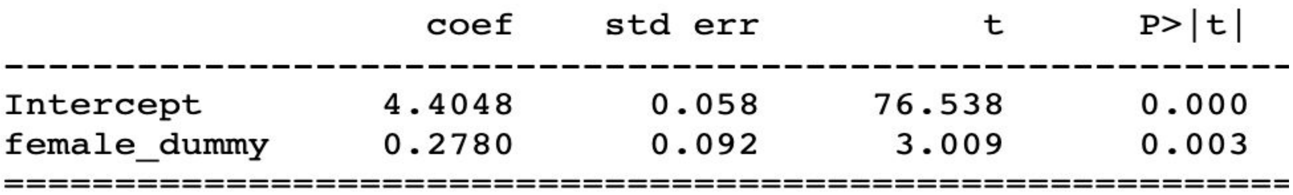
- Run a regression of cardiac risk on the female_dummy (variable=1 if respondent is female, 0 otherwise) OR on the male_dummy (not both, because of multicollinearity)

slido

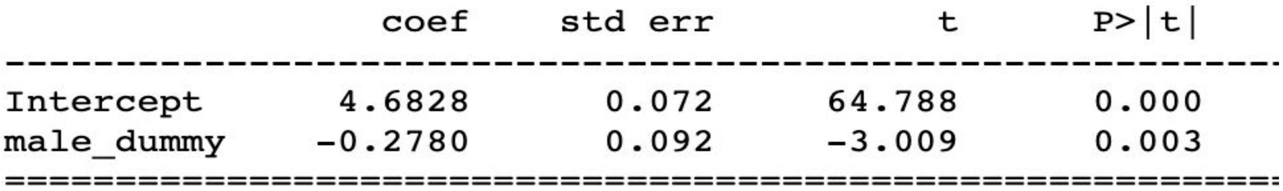
What are the expected risk scores?

ⓘ Start presenting to display the poll results on this slide.

Poll answer



```
                     coef    std err          t      P>|t|
----------------------------------------------------------
Intercept          4.4048      0.058     76.538      0.000
female_dummy       0.2780      0.092      3.009      0.003
==========================================================
```



```
                    coef    std err          t      P>|t|
---------------------------------------------------------
Intercept         4.6828      0.072     64.788      0.000
male_dummy       -0.2780      0.092     -3.009      0.003
=========================================================
```

- Y_i=β_0 +β_1*female_dummy_i+e_i

- Expected cardiac risk score for males is β_0 (because female_dummy_i= 0 if person i is male)

- Expected cardiac risk score for females is β_0+ β_1 (because female_dummy_i= 1 if person i is female)

Plan for today …

- On the increasing importance of marketing analytics
- Course logistics
- Regression (and some basic stats) re-cap
  - Interpreting regression coefficients: univariate / multivariate regression, continuous / discrete variables
  - Precision: standard errors, t-stats, p-values

Samples and Population

- There is a population relationship between two variables

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

- For a given sample, we can construct estimates (where the $\hat{e}_i$ are estimated residuals and different from the true residuals e_i)

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{e}_i$$

- What are the properties of the estimator $\hat{\beta}_1$ across different hypothetical samples drawn from the population?

Unbiasedness and precision of estimator

- If the errors (e_i) are truly random, the least squares estimator $\hat{\beta}_1$ is unbiased, i.e. correct in expectation / on average

$$E(\hat{\beta}_1) = \beta_1$$

- Under the same assumption we can derive the variance of the least square estimator (higher variance = lower precision)

$$Var(\hat{\beta}_1) = \frac{1}{N}\frac{Var(e)}{Var(X)} = \frac{1}{N}\frac{\sigma^2}{Var(X)}$$

- Intuitively:
  - More variation in the error makes the estimator less precise
  - More variation in X makes the estimator more precise

Standard error

- $Var(\hat{\beta}_1)$ depends on the unknown $\sigma^2 = Var(e)$ and therefore we have to estimate it by $s^2 = Var(\hat{e})$

$$s^2 = \frac{1}{N-2} \sum (\hat{e}_i - E(\hat{e}_i))^2$$

(the division by N-2 gives an unbiased estimator of σ^2)

- The standard error is an estimator of the standard deviation of $\hat{\beta}_1$ (that now only depends on the data):

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{N} \frac{s^2}{Var(X)}}$$

Hypothesis testing

- The central limit theorem gives an approximate distribution for $\hat{\beta}_1$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{1}{N}\frac{s^2}{Var(X)}\right)$$

- A standard test is for “statistical significance”. That is, a test of the null hypothesis that β_1= 0. Under this hypothesis we have

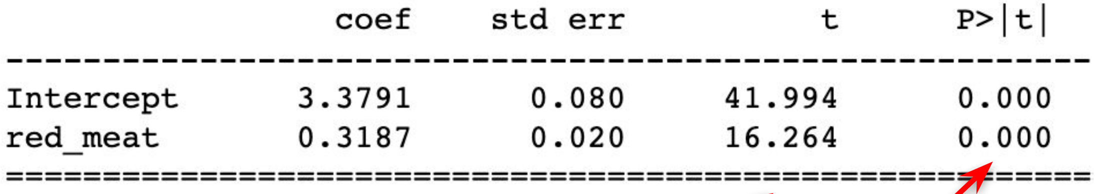
$$\hat{\beta}_1 \sim N\left(0, \frac{1}{N}\frac{s^2}{Var(X)}\right) \longrightarrow \boxed{(\hat{\beta}_1 - 0)/SE} \sim N(0,1)$$

- $SE = \sqrt{V\hat{a}r(\hat{\beta}_1)} = \sqrt{\frac{1}{N}\frac{s^2}{Var(X)}}$

Test statistic

- Technical reference:
    - Appendix 4.3 in Stock & Watson

Statistical significance



```
                    coef    std err          t      P>|t|
---------------------------------------------------------
Intercept         3.3791      0.080     41.994      0.000
red_meat          0.3187      0.020     16.264      0.000
=========================================================
```

Q: What is the hypothesis we are testing?

A: That the true regression coefficient is zero

Q: What is the test-statistic?

A: $(\hat{\beta}_1 - 0)/SE$ = 0.318688/0.019594 = 16.264

Q: How do we compute the p-value?

A: p-value = 2 * Pr(Z>16.264) = tiny! (where Z is a standard normal)

Statistical significance

Q: How do we interpret the test-statistic in words?

A: If the coefficient was truly zero then the estimate we obtained is over 16.2 standard deviations away from the truth. This is a highly unlikely event!

Q: How do we interpret the p-value?

A: If we took repeated samples, only "p-value" percent of the time would we get an estimate as far away from zero as 0.318688, if the true coefficient was zero

Statistical significance

- We reject the null hypothesis if p-value<0.05 = correlation between red meat and cardiac risk is statistically significant

- Careful:
    - Statistical significance and causality are different concepts
    - Neither necessary nor sufficient for one another

slido

The standard error is

ⓘ Start presenting to display the poll results on this slide.

Poll answer

- $SE = \sqrt{V\hat{a}r(\hat{\beta}_1)}$ the estimator of the standard deviation of the coefficient estimator $\hat{\beta}_1$
- The coefficient β_1 is a constant, so it has no standard deviation
- The standard deviation of $\hat{\beta}_1$, $\sqrt{V\hat{a}r(\hat{\beta}_1)}$ is unknown so it must be estimated

Re-cap

- Interpreting univariate regression output:
  - X continuous: expected change in Y for a one unit change in X
  - X discrete: difference in means between group with X=1 and group with X=0 (regression is logically equivalent to 2-sample mean test)

- Precision and hypothesis testing:
  - Standard error = uncertainty of coefficient estimate
  - T-stat: Under hypothesis of no effect, how far away from zero is the estimated coefficient
  - P-value: Under null of no effect, probability of observing an effect as far away from zero as the one estimated (reject if $p\text{-value} < 0.05$)

Workshop

- Some logistics for the workshop (that will last until the end of class):
    - Download this week’s workshop files from Canvas
    - Work in groups of 2 or 3
    - Work on and fill out one Jupyter notebook with the answers per team (these will not be graded)
    - I will be around to help with questions
    - Solution will be on Canvas after class