COMP 333 — Week 2 Example 1: Tipping

Tipping in a Restaurant: Dataset

The dataset was first used in the book:

> Bryant, P. G. and Smith, M (1995), *Practical Data Analysis: Case Studies in Business Statistics*. Richard D. Irwin Publishing, Homewood, IL.

It is also used in the wikipedia article on EDA.

Collection

One waiter recorded information about each tip he received over a period of a few months working in one restaurant.

Overall this made a data frame with 244 rows and 7 variables

He collected several variables:

- tip in dollars,
- bill in dollars,
- sex of the bill payer,
- whether there were smokers in the party,
- day of the week,
- time of day,
- size of the party.

Tipping in a Restaurant: Model

From the data, the model is created by fitting a straight line
to relate the tip to the size of the party.

First we **engineer a new feature** *tip_rate* which is the tip as a percentage of the bill.

The approach is to fit a regression model to predict the tip rate.

The fitted **model** is

- $tip\_rate = 0.18 - 0.01 \times party\_size$

Technically, "fitting a straight line" is called *linear regression*.

We could use the model to make predictions:

> if size of the dining party increases by one (leading to a higher bill),
> the tip rate will decrease by 1%.

Tipping in a Restaurant: EDA

The straight line regression model does not tell the whole story!

There are interesting relationships that can be discovered between other variables.

This is where Exploratory Data Analysis comes in.

The example looks at

- continuous or discrete nature of the actual tip amounts
  by binning tips either into bins of size \$1 or size 10c
- the variance in the amount of the tip as the amount of the bill increases
  using a scatter plot, and fin that more customers are "cheap" than "generous" in tipping
- the relationship between tipping and gender; and
- the relationship between tipping and smoking.

You can see the details of this example of EDA in:

- the slides
- the wikipedia article
- the book starting at page 4
  D. Cook and D.F. Swayne (with A. Buja, D. Temple Lang, H. Hofmann, H. Wickham, M. Lawrence) (2007), *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*, Springer, 978-0387717616
  which you can find a pdf online by searching for the book title.