Data Science Methods for Clean Energy Research

Week 7 L1
(lost week because of snow and forming groups)

Feb 13, 2017

UNIVERSITY of WASHINGTON

My dog and his friend. I don't think he is still friends with this dog…

Breaking news

My dog got a haircut and a bath, now has boot-shaped feet

SEDS/DSCMER project work (day 3)

UNIVERSITY of WASHINGTON

Today

- Presentations from groups on concept
  - Order below, thanks random numbers!
- Instructor approval and comments via email from Dave and Jim today or tomorrow
- Questions?

Team	Member 1	Member 2	Member 3	Member 4	Order (lowest 1st)
1	Sarah Floris	Hongbin Liu	Joe Kasper	**Coco Mao**	0.1817972293
4	Jerry Chen	Ryan Kastilani	Yanbo Qi	**YuanYuan Shi**	0.3855352511
8	Jeff Harrison				0.5175930031
3	Yatong Ge	Nick Montoni	Arushi Prakash	**Wes Tatum**	0.564191776
5	Rahul Avadhoot	**Kejia Wu**	Hanyang Xu	Tong Zhang	0.6804034033
7	**Wesley Beckne**	Jessica Kong	Garrett Davidson		0.6890603786
2	Varun Aduru	Heidi Nelson	Khushmeen Sak	**Ethan Wang**	0.6917310163
6	**Ivan Cui**	Jiayuan Guo	**Daniel Pan**	Yongquan Xie	0.8141867342

Project milestones

- **Wed 2/8: Project pitches and forming teams**
- **Mon 2/13: Team pitches of basic idea**
- **SEDS HW4**
- **SEDS HW5**
- **Poster session: Presentation w/DIRECT faculty, (maybe) project sponsors, your labmates and your classmates. Location and time:** Thur March 16 3-5 pm
- **Final commit on GitHub: Turning in the project. Due Wed March 15 at 5pm**

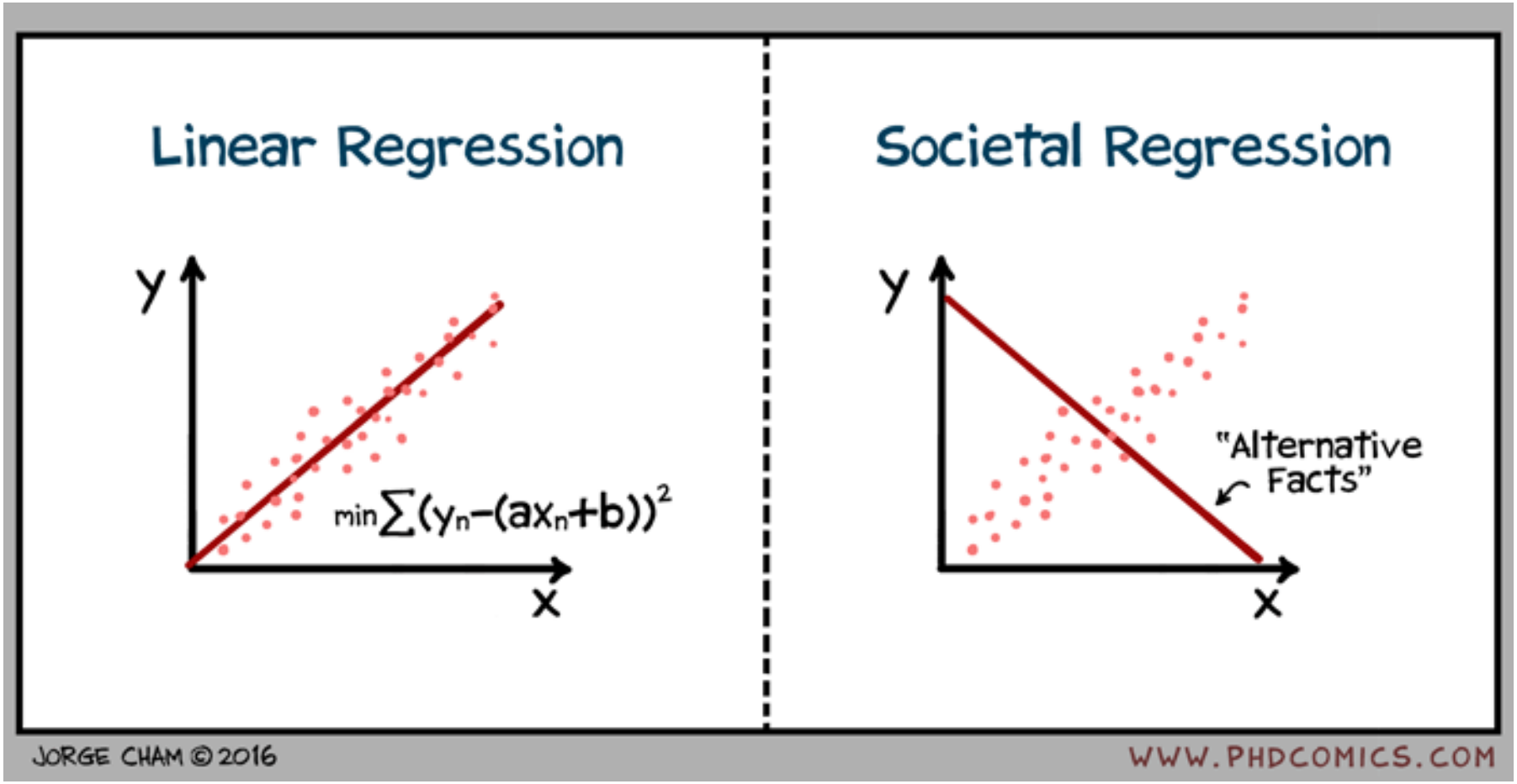
Plan for next 2 DSMCER sessions

- **Wed 2/8:** 40 min of class time (2:10-2:50). Opportunity for enterprising students to “pitch” ideas and recruit team members. We especially ESL students to participate and note we recognize their skills / abilities even if they are still new at English – great opportunity to practice and get feedback! Students will contribute to a Google Sheet if they like a project and want to sign up. Dave and I will form project teams @ end of class

- **Monday 2/13:** 40 min of class time (2:10-2:50). Teams have to pitch ideas to class/Dave/Jim for feedback. Two slide overview of project using Google slides template (so we can just have one laptop). We will provide a template

Regression!

PhD comics on point



Motivation/warmup: In groups of 2–3, 5 min max

- **Describe a time when you used linear (or otherwise) regression for research or class**
- **How did you determine the accuracy of the coefficients (parameters) you fit?**
- **How did you determine the accuracy of the model?**

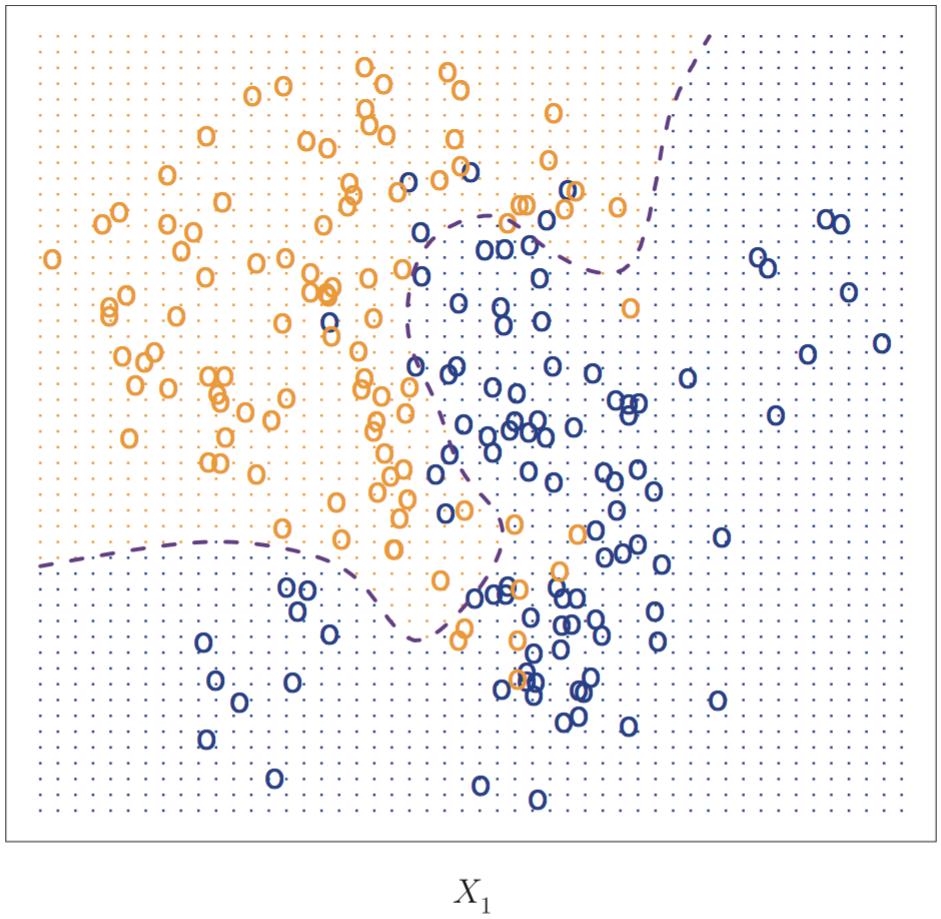
- **n.b., if you don't know the difference between the last two questions – you should also discuss this!**
  - n.b., is a great latin acronym to use regularly in your writing (e.g., above)

Outline

- Project work
- Warmup
- Simple linear regression
  - The model
  - The training
  - The error
    - In the coefficients
    - In the model
  - Making predictions with the model
- Multiple regression

One thing I forgot to mention…

A note about distances and identifying who is "nearest" ?



- In this example, and all others we looked at and discussed you all made a key assumption (*without realizing it*) when you were identifying neighbors
- The data in each dimension of X should be scaled similarly!
- We are concerned with relative distances in the parameter space, not absolute distances in the underlying units
- Otherwise units, dimensionality, etc. can drastically favor closeness in arbitrary dimensions of X

Simple linear regression (SLR)

- **Why are we spending time on this?**
- **The regression framework (**generic in supervised machine learning**)**
  - **Propose a model**
  - **Define the error metric**
  - **Estimate the coefficients and assess the estimate**
  - **Determine the error in the model**

$$Y \approx \beta_0 + \beta_1 X. \quad (3.1)$$

The humble SLR

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \quad (3.2)$$

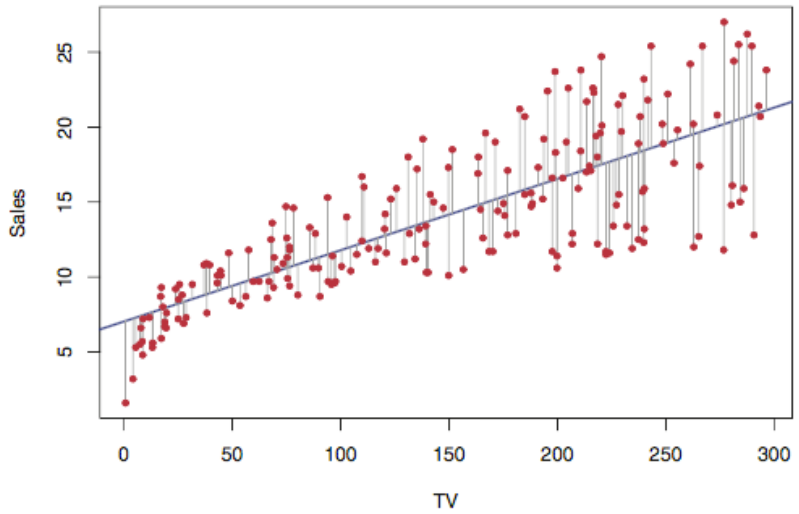
Distinguishing the estimates we make with the training data (hat) from the proposed actual coefficients/responses…

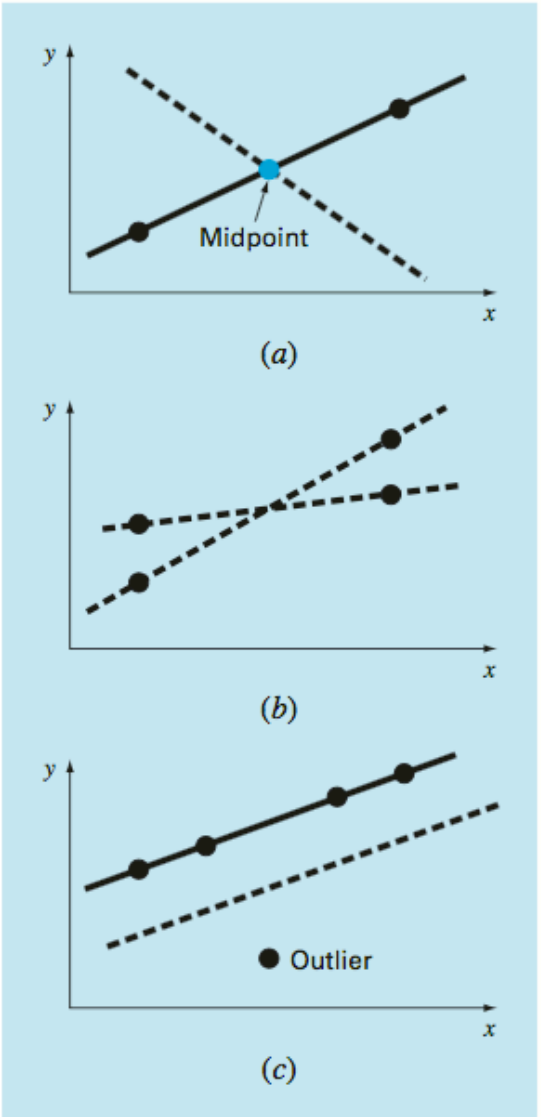
Once the SLR model is known, what is the error?

FIGURE 17.2
Examples of some criteria for "best fit" that are inadequate for regression: (*a*) minimizes the sum of the residuals, (*b*) minimizes the sum of the absolute values of the residuals, and (*c*) minimizes the maximum error of any individual point.

- **The SSR describes the sum of squared residual errors, very similar in concept to the MSE we discussed last week**

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \quad (3.3)$$





Top: Fig 3.1 ISL
Right: Figure 17.2 from Chapra and Canale numerical methods for Engineers

The method of "least squares" minimization

- Given our model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and some training data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we need to determine what values of our parameters $\widehat{\beta}_i$ minimize the error.
- This is accomplished by determining $\frac{\partial RSS}{\partial \widehat{\beta}_0}$ and $\frac{\partial RSS}{\partial \widehat{\beta}_1}$ setting equal to zero and solving:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Ways to estimate the uncertainty in the coefficients

- There is a simple estimate for the standard error

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (3.8)$$

  - Note the variance (σ^2) is estimated from the RSS as: $\sigma = RSE = \sqrt{RSS/(n-2)}$ (RSE = residual standard error)

- Given the standard errors (ISL, eq 3.8), we can estimate the confidence intervals as: $\widehat{\beta_0} \pm t^* SE(\widehat{\beta_0})$ or $\widehat{\beta_1} \pm t^* SE(\widehat{\beta_1})$

Hypothesis testing on the coefficients

- Recall from statistical hypothesis testing, I said that we would revisit P-values in this section of the class
- Once we have our estimated coefficient values we test the hypothesis $H_0 : \beta_1 = 0$
  - If this hypothesis true, then our model is simply: $y = \beta_0 + \epsilon$,i.e., there is no correlation between x and y!

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}, \qquad (3.14)$$

Accuracy of the model – how are we doing overall?

- **Residual standard error**

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad (3.15)$$

  - **A measure of the** lack of fit **of your model (**in units of Y!**)**

- **R^2 statistic**

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad (3.17) \qquad \mathrm{TSS} = \sum(y_i - \bar{y})^2$$

  - **A scale invariant measure (0-1 range) that explains "***the proportion of the variability of Y that is explained by X***"**
  - **Lets chat about TSS and what it means…**

Quick example of SLR and P-value calculation with scikit-learn

> Switch to python!

The correlation (we will need this in multiple regression)

- **Recall the basic descriptor – correlation coefficient or simply** correlation**, which we use to describe trends in our data and relationship between variables**

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}, \qquad (3.18)$$

Multiple regression

- Concept: independently assess the variation in Y with different values of X:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \qquad (3.19)$$

- As with SLR, the coefficients are determined by setting the analytical partial derivatives to zero and solving the resultant *p+1* linear equations
- As with SLR there is an exact solution