Lecture VI: More regression

STA9750

Fall 2018

Baruch College CUNY

Info on project + midterm

- Midterm
  - Assigned on Oct 25th, due Oct 31st @ 11:59pm
  - A few problems at the level of the HWs
  - Topics? One-sample tests, two-sample tests, ANOVA, and regression
- Project
  - I will give you a dataset and a prompt. You will analyze the data and write an 8 page max report as if you were reporting to a client.
  - It'll be assigned when we're done with SAS, and it'll be due at the end of the semester

Today

- Intervals for regression mean vs prediction intervals
- Multiple linear regression

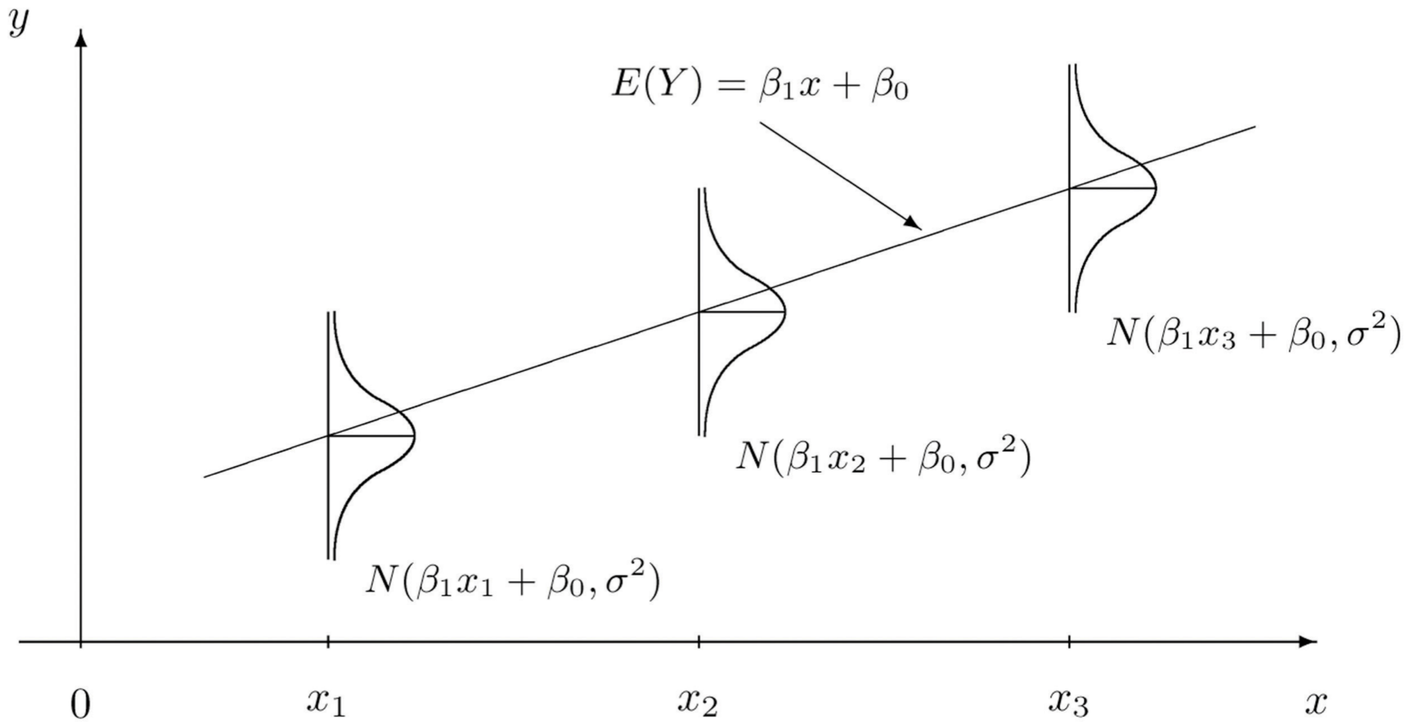
Simple linear regression

$$y_i \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Assumptions:

- Independence of outcomes y_i for i in 1:n (given the x_i).
- Normality
- Homoscedasticity (equal variance across observations, which doesn't depend on x_i)
- Linearity [i.e. $E(Y \mid X)$ is a line]



Source: https://stats.libretexts.org/Textbook_Maps/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)

CIs for regression mean and predictions

- Given a specific value of *x*, a $100(1-\alpha)\%$ CI for the *regression mean* $\beta_0 + \beta_1 x$ is

$$(b_0 + b_1 x) \pm t_{\alpha/2, n-2}\, s \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- Given a specific value of *x*, a $100(1-\alpha)\%$ CI for a *new observation (a prediction interval)* $\beta_0 + \beta_1 x + \varepsilon$ is

$$(b_0 + b_1 x) \pm t_{\alpha/2, n-2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

Strictly wider because we're taking into account the idiosyncratic error term ε

Multiple linear regression

- The same, but with more variables
- Least squares: find the b_j that minimize

$$\sum_{i=1}^{n}[y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1})]^2$$

- We can find CIs and hypothesis tests if we make assumptions
- We assume

$$y_i \overset{\text{ind}}{\sim} N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Independence of outcomes y_i for i in 1:n (given the x_{ij}).
- Normality
- Homoscedasticity (equal variance across observations, which doesn't depend on x_{ij})
- Linearity (i.e. E[Y | X] is a linear comb. of the Xs)

Checking assumptions

- We can check the assumptions using the same plots we used for simple linear regression
- There are some additional plots/statistics that are useful for identifying ***influential*** observations
  - What does influential mean? It can potentially be defined in different ways…
  - A useful perspective is "how much does my *fit* change if I take out this observation?"
  - Different proposals in the literature: Cook's distance, DFFITS, DFBETAs, …

Cook's distance

- Cook's distance of observation *i* is

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2}$$

$\hat{y}_j$ predicted value for observation j with all the observations

$\hat{y}_{j(i)}$ predicted value for observation j after taking out the i-th observation

s^2 our usual estimator for the residual variance

- How big is big? Different recommendations… Some people say $D_i > 1$
- I recommend looking closely at any observation that seems to "stick out"

DFBETAs

- How much does the least squares estimate *b* change if I take out observation *i*?

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{bmatrix}$$ least squares estimate

$b_{(i)}$ least squares estimate after removing observation *i*

$$\text{DFBETA}_i = b - b_{(-i)}$$

It's a *p*-dimensional vector!

- Again, look for values that "stick out"

Leverage, outliers, and influence

- Leverage: measures how far away x_i is from the other x values [goes from 0 to 1, from "average x" to "very unusual x"]
- High leverage: unusual value of x_i, which may or may not be well predicted by our line
- Big residual e_i: point that is badly predicted by our line (outliers)
- Observations with high leverage and big residuals are highly influential… Cook's distance can be written as

$$D_i = \frac{e_i^2}{s^2 p}\left[\frac{h_i}{(1-h_i)^2}\right]$$

h_i: leverage of observation i

e_i: residual of observation i

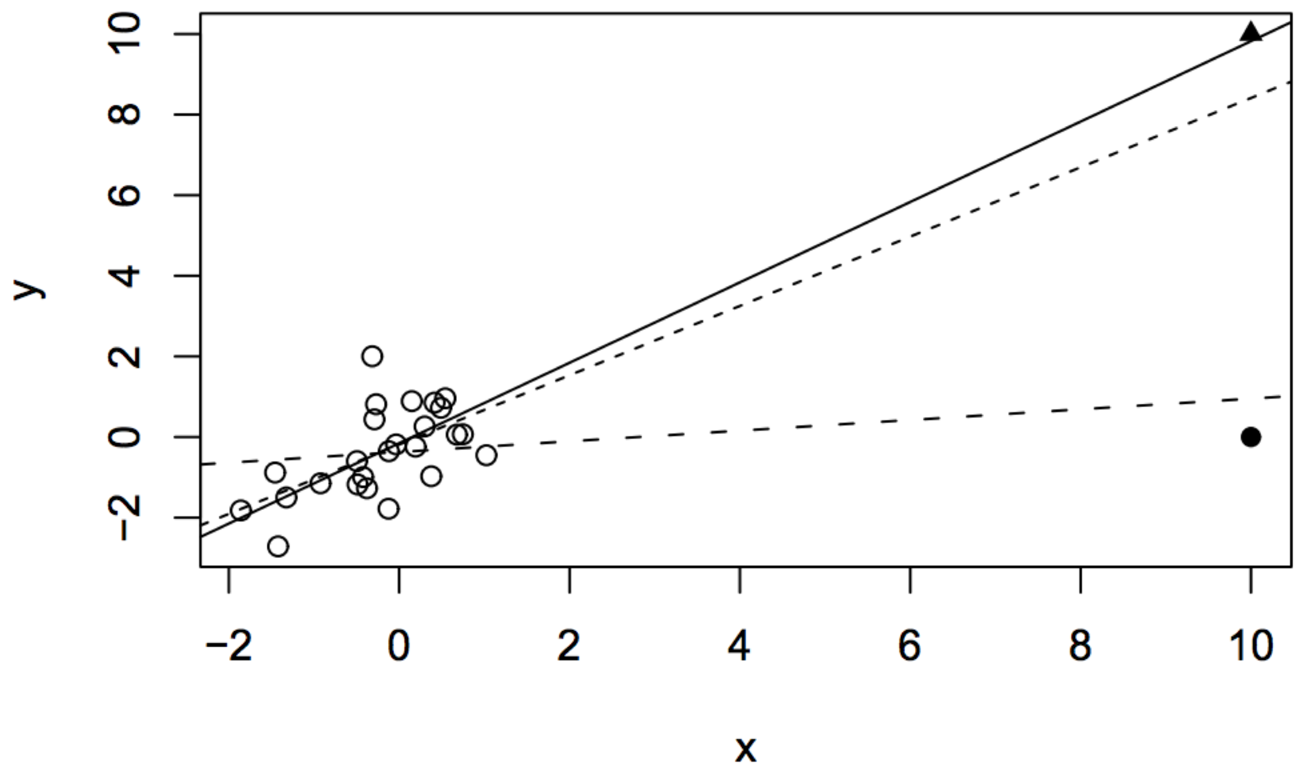


Figure 7.2: Outliers can conceal themselves. The solid line is the fit including the ▲ point but not the • point. The dotted line is the fit without either additional point and the dashed line is the fit with the • point but not the ▲ point.

Source: https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Next time

- Go through multiple linear regression handout
- More topics on regression
  - How to introduce categorical predictors
  - Pick the "best" model