Boring but important disclaimers:

- If you are not getting this from the GitHub repository or the associated Canvas page (e.g. CourseHero, Chegg etc.), you are probably getting the substandard version of these slides Don't pay money for those, because you can get the most updated version for free at

  `https://github.com/julianmak/academic-notes`

  The repository principally contains the compiled products rather than the source for size reasons.
- Associated Python code (as Jupyter notebooks mostly) will be held on the same repository. The source data however might be big, so I am going to be naughty and possibly just refer you to where you might get the data if that is the case (e.g. JRA-55 data). I know I should make properly reproducible binders etc., but I didn't...
- I do not claim the compiled products and/or code are completely mistake free (e.g. I know I don't write Pythonic code). Use the material however you like, but use it at your own risk.
- As said on the repository, I have tried to honestly use content that is self made, open source or explicitly open for fair use, and citations should be there. If however you are the copyright holder and you want the material taken down, please flag up the issue accordingly and I will happily try and swap out the relevant material.

OCES 3301 :

basic Data Analysis in ocean sciences

Session 3: regression

Outline

(Just overview here; for actual content see Jupyter notebooks)

- linear regression

  $\rightarrow$ mismatches

  $\rightarrow$ minimiser
- higher degree fitting

  $\rightarrow$ issues?
- (Pearson) correlation coefficient
- more variables (next session)

Linear regression

Figure: The eternal bendy boi.

Mismatches and linear regression

Given samples $\{x_i\}$, relative to mean, we could have

$$\text{err}_1 = |x_i - \overline{x}|, \qquad \text{err}_2 = (x_i - \overline{x})^2, \qquad \text{etc.}$$

Mismatches and linear regression

Given samples $\{x_i\}$, relative to mean, we could have

$$\mathrm{err}_1 = |x_i - \overline{x}|, \qquad \mathrm{err}_2 = (x_i - \overline{x})^2, \qquad \text{etc.}$$

More generally, given samples (x_i, y_i), we may postulate a linear model where

$$y = f(x) = ax + b.$$

- we have actual sample y_i
- we also have prediction $f(x_i)$

Mismatches and linear regression

Given samples $\{x_i\}$, relative to mean, we could have

$$\mathrm{err}_1 = |x_i - \overline{x}|, \qquad \mathrm{err}_2 = (x_i - \overline{x})^2, \qquad \text{etc.}$$

More generally, given samples (x_i, y_i), we may postulate a linear model where

$$y = f(x) = ax + b.$$

- we have actual sample y_i
- we also have prediction $f(x_i)$
- measure "skill" by $(y_i - f(x_i))^2$ (or whatever?)

Mismatches and linear regression

The linear regression model or the line of best fit is some

$$y = f(x) = ax + b$$

where the choice of a and b minimises a mismatch to be specified.

- only relative to mismatch, usually use the ℓ^2 mismatch

$$\ell_2 \sim \sum_{i=1}^{N} (y_i - f(x_i))^2$$

  $\rightarrow$ actually a closed form analytical solution here for a and b given x_i and y_i
- other choices possible (see Jupyter notebook exercise, but probably quite hard)

Higher degree fitting

I don't have to just stop at linear, could have for example (assuming ℓ_2 minimising)

$$y = g(x) = a_0 x^n + a_1 x^{n-1} + \cdots a_{n-1} x + a_n = \sum_{i=0}^{n} a_i x^{n-i}$$

Higher degree fitting

I don't have to just stop at linear, could have for example (assuming ℓ_2 minimising)

$$y = g(x) = a_0 x^n + a_1 x^{n-1} + \cdots a_{n-1} x + a_n = \sum_{i=0}^{n} a_i x^{n-i}$$

Figure: The eternal bendy boi again.

Issues

$$y = g(x) = a_0 x^n + a_1 x^{n-1} + \cdots a_{n-1} x + a_n = \sum_{i=0}^{n} a_i x^{n-i}$$

- need enough data points
- numerically ill-conditioned as n increases
- just because you could doesn't mean you should

  $\rightarrow$ Occam's razor: all things being equal, simplicity wins?

  $\rightarrow$ overfitting (see notebook)

(Pearson) correlation coefficient

Normally denoted r and given for a sample by

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

- the top part is the covariance

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

- $r = \pm 1$ means x_i and y_i are perfectly correlated or anti-correlated
- $r = 0$ means no correlation whatsoever
- other values could mean high or suggestions of correlation
- this is only **linear** correlation

Understanding: linear regression done badly

nature publishing group

ARTICLES
EPIDEMIOLOGY

Will All Americans Become Overweight or Obese? Estimating the Progression and Cost of the US Obesity Epidemic

Youfa Wang[1], May A. Beydoun[1], Lan Liang[2], Benjamin Caballero[1] and Shiriki K. Kumanyika[3]

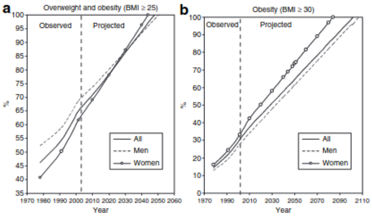

Figure 1 Prevalence of obesity and overweight among US adults: Observed during 1976–2004 and projected. The projected prevalence presented here are those based on our linear regression models.

Figure: Wang *et al.* (2008), Obesity. Name at least four things wrong with this graph.

Jupyter notebook

Go to 03 Jupyter notebook to play around with some artificial and “real” data