Data, Environment and Society: Lecture 8: Introduction to Models

Instructor: Duncan Callaway
GSI: Seigi Karasaki

September 18, 2018

What is a mathematical model?

A system of equations that relates one set of variables to another set of variables.

Examples

1. The distance a cheetah travels in h hours at 65 miles per hour.

$$d(h) = 65h$$

2. The height of a rock thrown in straight up, after t seconds:

$$h(t) = \frac{1}{2}at^2 + v_0 t + h_0$$

... with gravity acceleration a, at initial velocity v_0, from initial height h_0.

3. The mean surface temperature of the Earth in 2100:

$$T_{\text{surf}} = f(\text{a lot of different variables!})$$

Models don't have to be "first principles"

1. Number of ER visits for cardiac problems per day

$$N_{ER} = \beta_0 + \beta_1 \cdot PM25$$

   where $PM25$ is PM 2.5 concentrations.
2. HDI as a function of energy access:

$$HDI = \beta_0 + \beta_1 r + \beta_2 r^2$$

   where r is the percentage of households in a country with access to electricity.

What is model *estimation*?

The process of choosing a model's parameters using a data set of measurements.

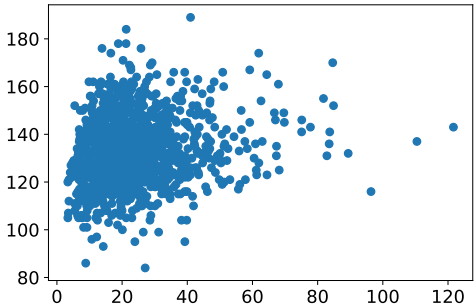
For example:

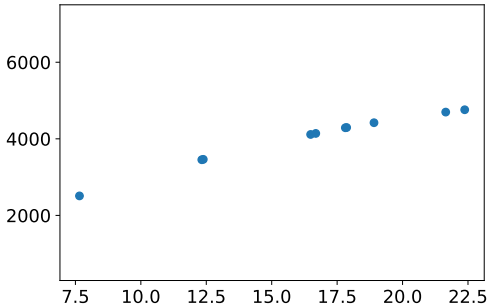
1. Record the height of a rock, and the time you made the measurement, several times as it flies through the air. Then use those data to choose the parameters of your "first principles" model so that its output matches your observations.

2. Obtain an administrative database of daily ER visits and the corresponding PM2.5 concentrations for each day. Use those data to choose β_0 and β_1 in $N_{ER} = \beta_0 + \beta_1 \cdot PM25$ so you can predict N_{ER} well from PM2.5.

Which is which?

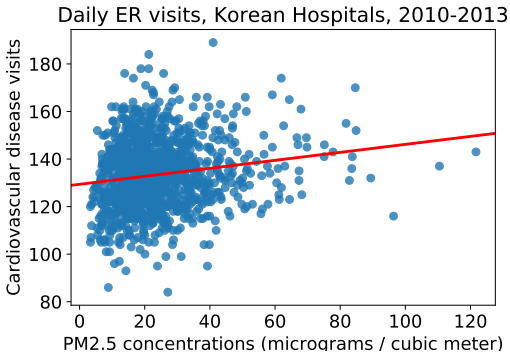
One is ER visits as a function of PM2.5 concentrations. One is height of a rock as a function of time.

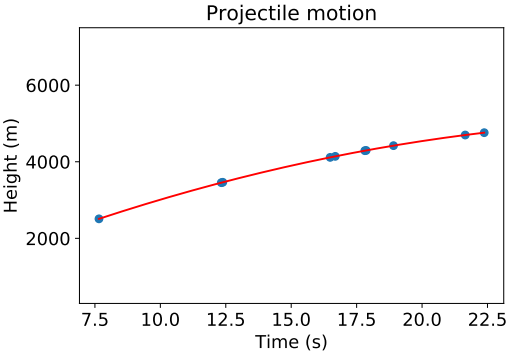




Which is which?

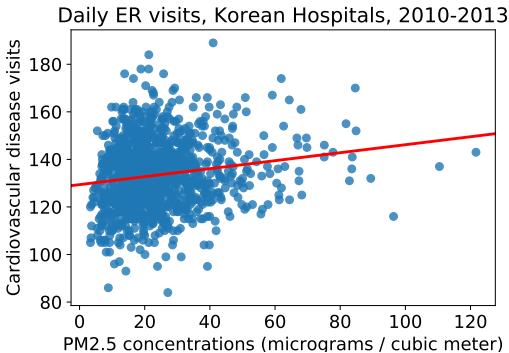
One is ER visits as a function of PM2.5 concentrations. One is height of a rock as a function of time.

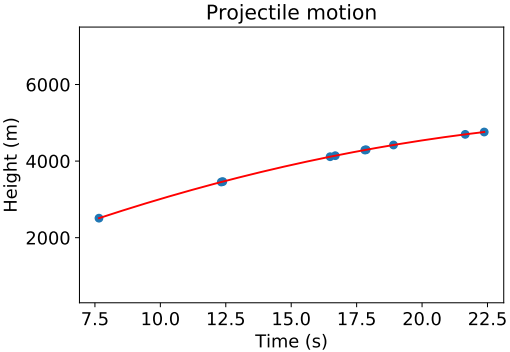




Source for hospital data: Hwang, S. H. *et al* (2017), PloS one, 12(8).

How did you choose?

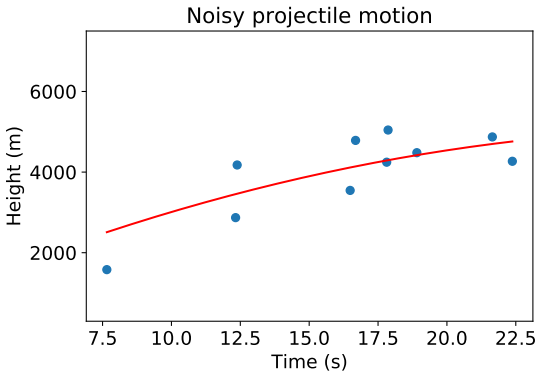




Source for hospital data: Hwang, S. H. *et al* (2017), PloS one, 12(8).

- The right plot has much less scatter
- The right plot seems to be describing a more systematic process

Could the projectile plot ever look like this? How?



- Where could this noise come from?
  - Measurement error
  - “process noise,” for example a windy environment.

Why might we build a model?

Prediction.

- Where will the projectile be in 5 seconds?
- I'm building a hospital in a city where I know the air quality trends as well as a bunch of other variables. How big should the ER be?

Inference: Estimating a parameter

- What is the acceleration of gravity?
- What is the correlation between air quality and ER visits?

Causal Inference:

- Does PM2.5 cause heart attacks?

Expectations for the model and data...

Prediction:

- Low expectations! As long as the independent variables are correlated with the dependent variables, we can make predictions.

Inference:

- Moderate expectations on the model: It needs to be sufficiently interpretable that we can understand what parameters mean

Causal Inference:

- Very high expectations! We need to be confident that *only* the independent variable is changing systematically across measurements.
- Otherwise we can't rule out the possibility that some other unobserved variable is impacting our observations.

You say regressor, I say feature

Math	Machine learning	Statistics
x	predictor, input variable, **independent variable**, feature	regressor, covariate, **independent variable**, explanatory variable, right hand variable
y	output variable response variable, **dependent variable**	**dependent variable,** outcome variable, left-hand side variable.

A little notation

Moving forward, we'll use this notation and terminology:

x_i	i^{th} observation of an independent variable
y_i	i^{th} observation of a dependent variable
ϵ_i	i^{th} random error, uncorrelated with x_i, and mean zero

$y_i = f(x_i) + \epsilon_i$	the "true" model, if one exists.
$\hat{y}_i = \hat{f}(x_i)$	our estimate of y_i using an estimate of f

Error or residual?

$$y_i = f(x_i) + \epsilon_i$$ the "true" model, if one exists.

$$y_i = \hat{f}(x_i) + e_i$$ the relationship between the data and the estimate.

So:

$$\epsilon_i$$ variation in y that is uncorrelated with x.

$$e_i = y_i - \hat{f}(x_i)$$ the "residual" between the data and the estimate.

Error v. Residual, continued

Important!

- ϵ and e could be very different.
- Because we'll rarely know the "true" model, we'll rarely know ϵ.
- On average, e will never be smaller than ϵ

How to evaluate how well a model performs?

Generic term: the *Cost function.*

- Cost functions can be used to describe how much of the variation in the data can be captured by the model.
- Example: The mean squared error:

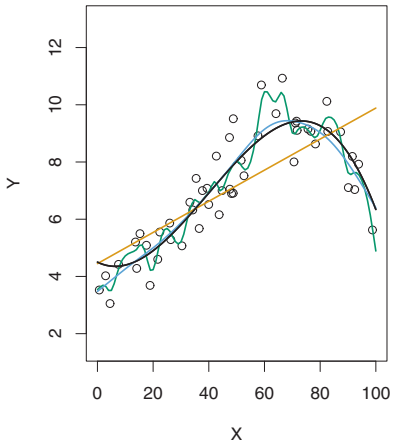
$$
\begin{aligned}
MSE &= \frac{1}{n}((y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2) \\
&= \frac{1}{n}(e_1^2 + e_2^2 + \cdots + e_n^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} e_i^2
\end{aligned}
$$

A thought experiment from ISLR Ch 2

Suppose you have four different model forms to choose from. When you fit them to the data, you get this figure.

Which model should you choose?

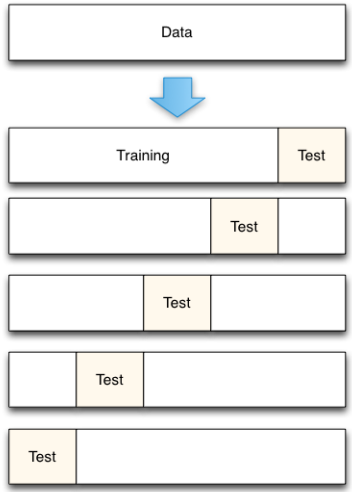
- The one that minimizes mean squared error?
- Careful! Doesn't the squiggly one minimize mean squared error?
- To do model selection we need to understand the concept of training and testing data.



Concept: Test and training data

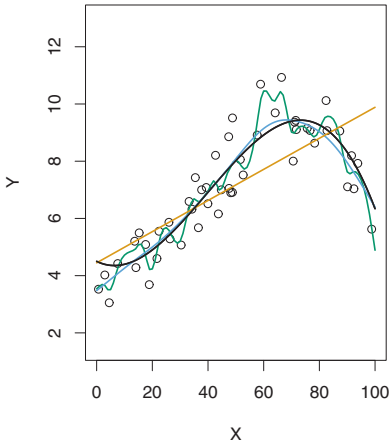
Choosing between different models can be done by partitioning your data in to “training” and “test” data.

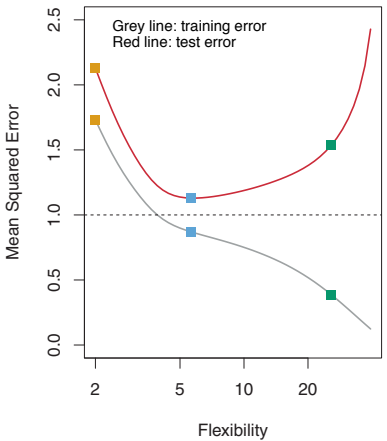
- “Training data”: The data we use to choose the parameters of an individual model.
- “Test data”: A set of data we withhold; it’s not for training. We use this data set to compare how different *models* perform relative to one another.



Source: kaggle.com

MSE for test and training data





What might a plot of MSE versus model "flexibility" look like?

Bias v. Variance

Bias:

- The propensity for a model to produce errors that are systematically high or low
- Bias can be positive in one range of the predictor and negative in another.

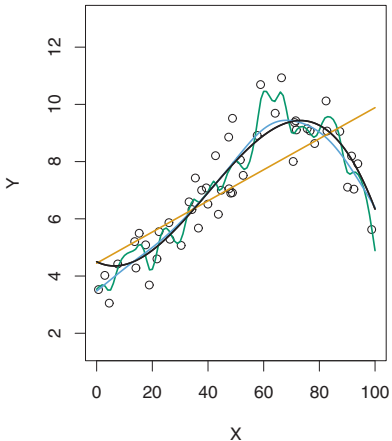
Variance

- The propensity for a model to make very different predictions if it is fit with two different training data sets from the same process.

Total error can be decomposed:

$$\text{Average } (y_0 - \hat{f}(x_0))^2 = \text{var}(\hat{f}(x_0)) + [\text{bias}(\hat{f}(x_0))]^2 + \text{var}(\epsilon_0)$$

Bias v. Variance, ctd.



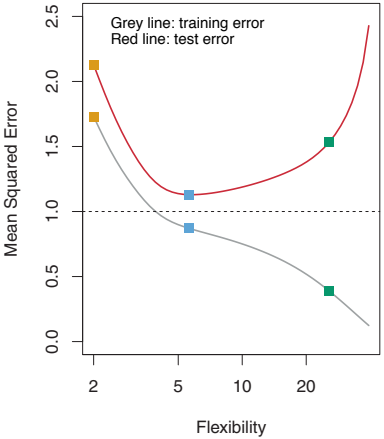
Which model has the greatest propensity for bias?

- The linear one. In ranges of x, it systematically under- or over-estimates.

Which model has the greatest propensity for variance?

- The squiggly one. If we drew another sample of data, we'd probably get very different squiggles.

Decomposing bias-variance



Take a moment to think about how bias and variance add up to make the red curve on the left. Try to draw bias and variance separately.