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

Instructor: Duncan Callaway
GSI: Salma Elmallah

September 24, 2019

Announcements and agenda

Reading

- Today: DS100 Ch 10, ISLR Ch 2
- Thursday: ISLR Ch 3.1

Announcements and agenda

Reading

- Today: DS100 Ch 10, ISLR Ch 2
- Thursday: ISLR Ch 3.1

Today:

- Visualization finish-up – types of plots
- What is a model, how do we estimate its parameters?
- Terminology going forward
- What's the bias-variance tradeoff
- Lighting review: estimating least squares regression coefficients

What is a mathematical model?

What is a mathematical model?

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

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*?

What is model *estimation*?

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

What is model *estimation*?

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

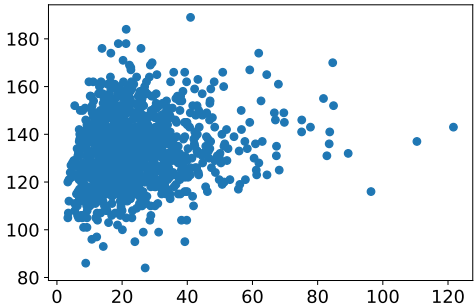
For example:

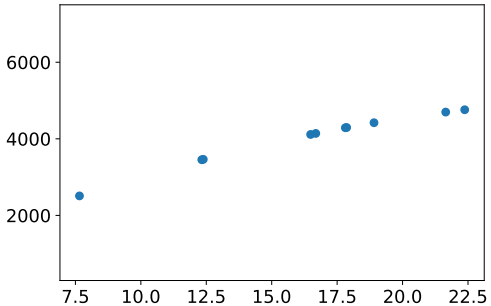
1. Record the height of a rock, and the time you made the measurement, several times as it travels through the air. Then use those data to choose the parameters of your "first principles" model so that its output matches your observations.

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?

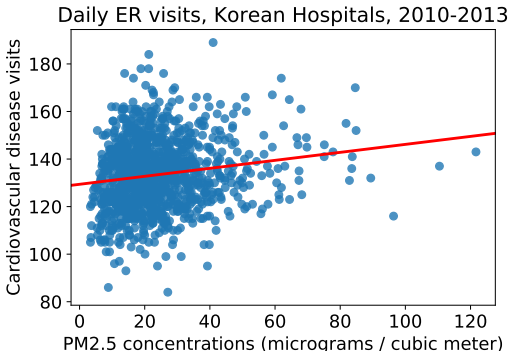
One is ER visits as a function of PM2.5 concentrations. One is height of a rock as a function of time.

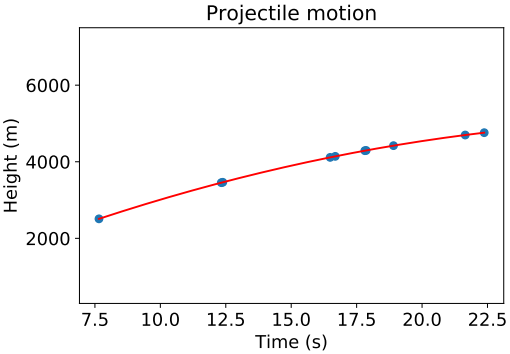




Which is which?

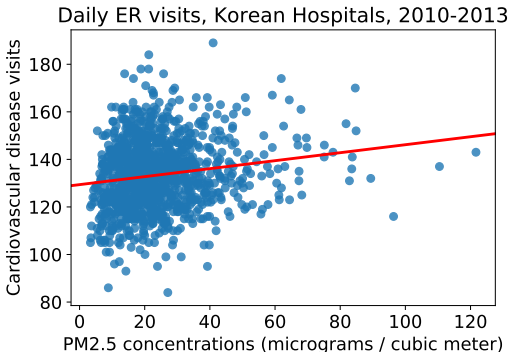
One is ER visits as a function of PM2.5 concentrations. One is height of a rock as a function of time.



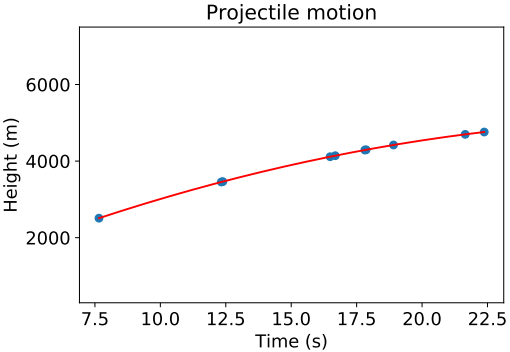


Source for hospital data: Hwang, S. H. *et al* (2017), PloS one, 12(8).

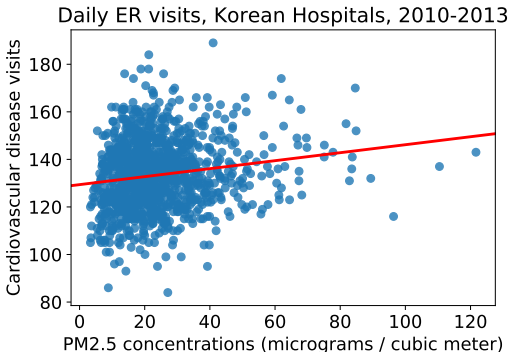
How did you choose?

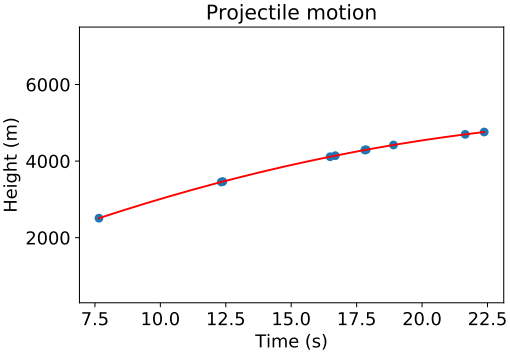


Source for hospital data: Hwang, S. H. *et al* (2017), PloS one, 12(8).



How did you choose?

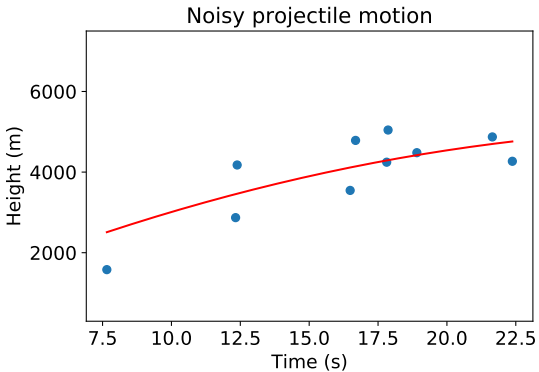




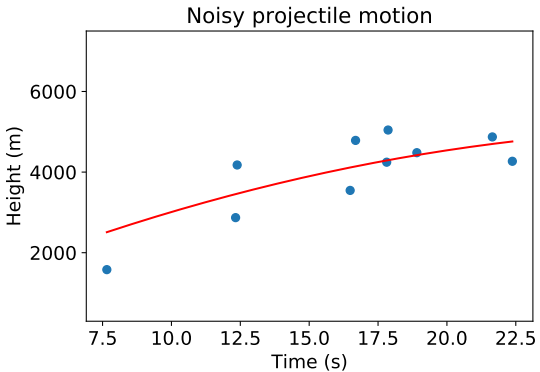
Source for hospital data: Hwang, S. H. *et al* (2017), PloS one, 12(8).

- 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?



Could the projectile plot ever look like this? How?



- Where could this noise come from?
  - Measurement error: the height numbers are erroneous
  - Model error: we didn't capture the whole process, for example a windy environment

Why might we build a model?

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		
y		

You say regressor, I say feature

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

You say regressor, I say feature

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

In this class, I'll stick to

1. "independent variable" or "feature" and
2. "dependent variable", "output variable" or "target"

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 ("epsilon")	i^{th} random error, uncorrelated with x_i, and mean zero

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

ϵ (epsilon) or e?

$y_i = f(x_i) + \epsilon_i$	the "true" model, if one exists.
$y_i = \hat{f}(x_i) + e_i$	the relationship between the data and the estimate.

ϵ (epsilon) or e?

$$y_i = f(x_i) + \epsilon_i$$ 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.

ϵ (epsilon) v. e, 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? 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: ← residual

$$MSE = \frac{1}{n}\left((y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \dots (y_n - \hat{y}_n)^2\right)$$

$$= \frac{1}{n}\left(e_1^2 + e_2 \dots e_n^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

How to evaluate how well a model performs? 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:

$$MSE = \frac{1}{n}((y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2)$$

$$= \frac{1}{n}(e_1^2 + e_2^2 + \cdots + e_n^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

How to evaluate how well a model performs? 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:

$$MSE = \frac{1}{n}((y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2)$$

$$= \frac{1}{n}(e_1^2 + e_2^2 + \cdots + e_n^2)$$

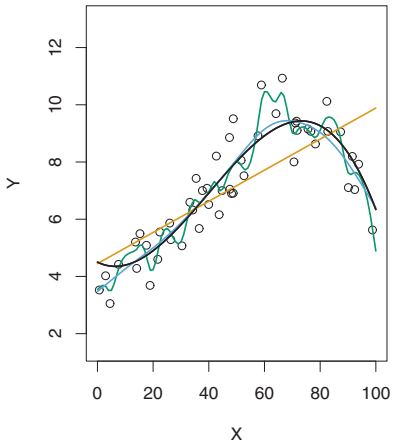
$$= \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

A major part of statistical learning lies in how the cost function is defined.

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?

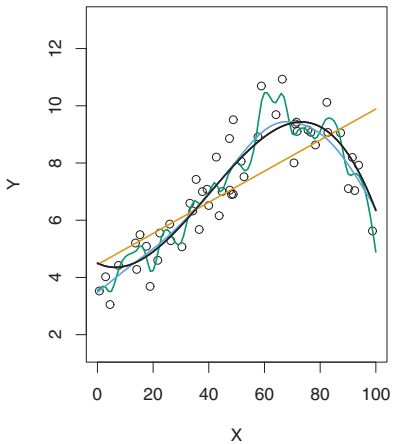


A thought experiment from ISLR Ch 2

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?

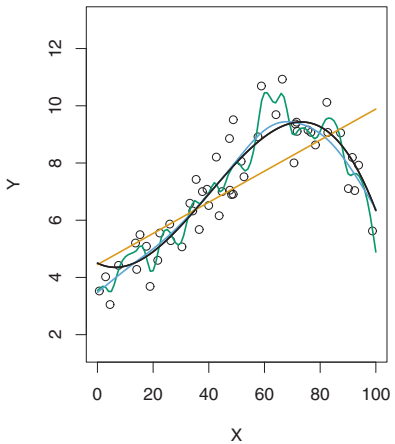


A thought experiment from ISLR Ch 2

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?

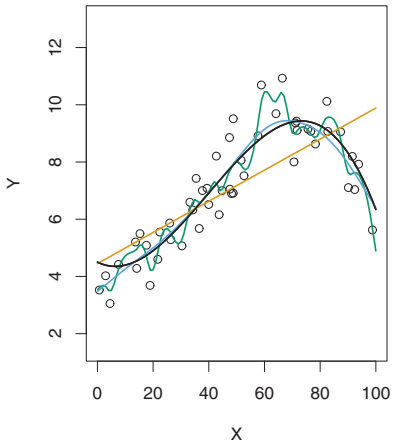


A thought experiment from ISLR Ch 2

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

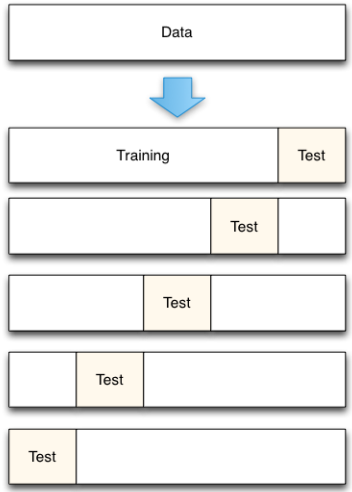
Choosing between different models can be done by partitioning your data in to “training” and “test” data.

- “Training data”:
- “Test 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.

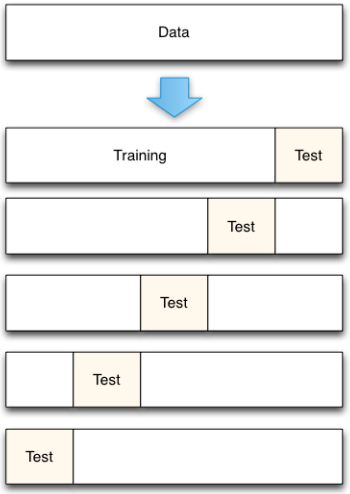


Source: kaggle.com

Concept: Test and training data

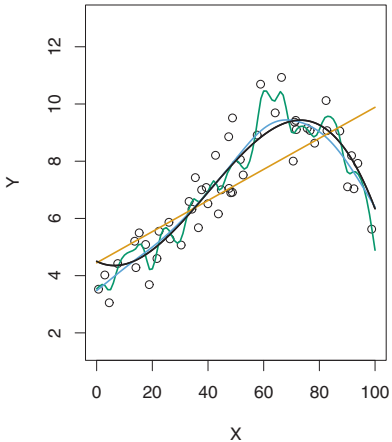
Choosing between different models can be done by partitioning your data in to “training” and “test” data.

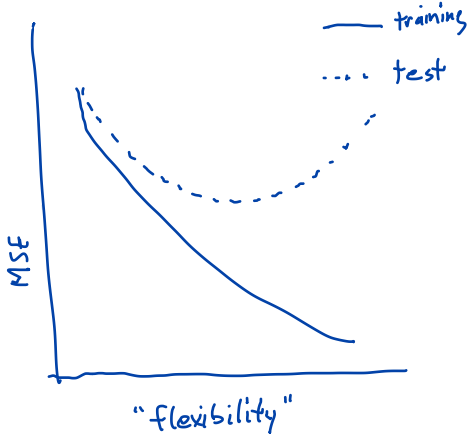
- “Training data”: The data we use to choose the parameters of an individual model.
- “Test data”: A set of data we withhold; it’s not for training. We use this data set to compare how different *models* perform relative to one another.
- Note: later in the course, we will talk about validation data for comparing completely different model forms.



Source: kaggle.com

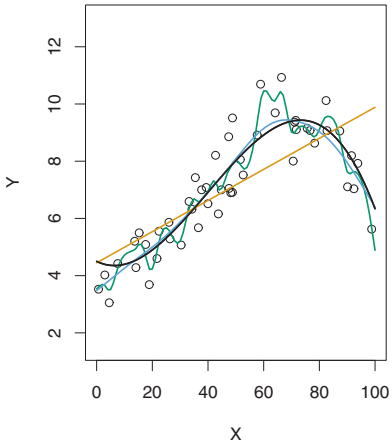
MSE for test and training data

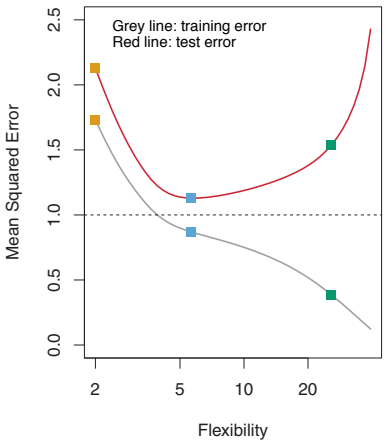




What might a plot of MSE versus model "flexibility" look like?

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 that are sampled from the same population.

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.

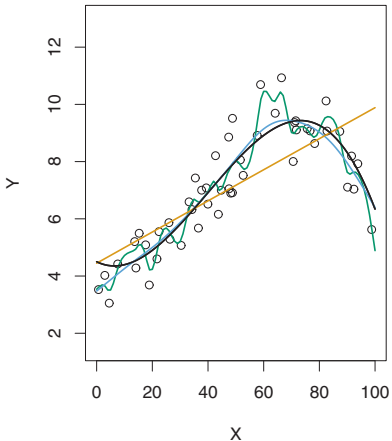
Variance

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

Side note: Total error can be decomposed:

$$\begin{aligned}\text{Avg } (y_0 - \hat{f}(x_0))^2 =& \text{(variance in a prediction, across different training data)} \\ &+ (\text{systematic bias})^2 + \text{(variance in } y \text{ that's uncorrelated with } x) \\ =& \text{var}(\hat{f}(x_0)) + [\text{bias}(\hat{f}(x_0))]^2 + \text{var}(\epsilon_0)\end{aligned}$$

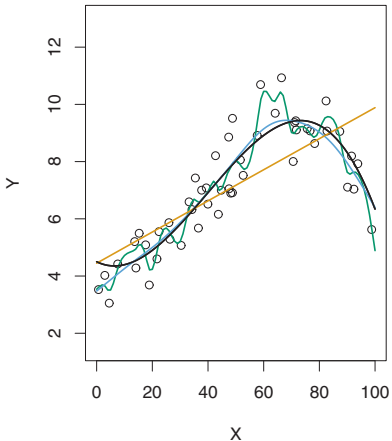
Bias v. Variance, ctd.



Which model has the greatest propensity for bias? (Systematically high or low estimates)

Which model has the greatest propensity for variance? (Different predictions if fit with different data sets)

Bias v. Variance, ctd.

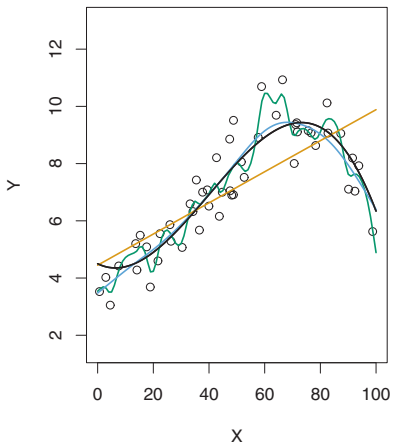


Which model has the greatest propensity for bias? (Systematically high or low estimates)

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

Which model has the greatest propensity for variance? (Different predictions if fit with different data sets)

Bias v. Variance, ctd.



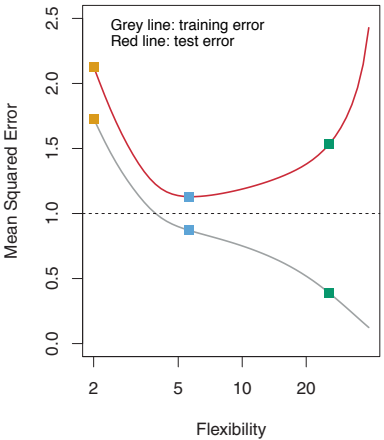
Which model has the greatest propensity for bias? (Systematically high or low estimates)

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

Which model has the greatest propensity for variance? (Different predictions if fit with different data sets)

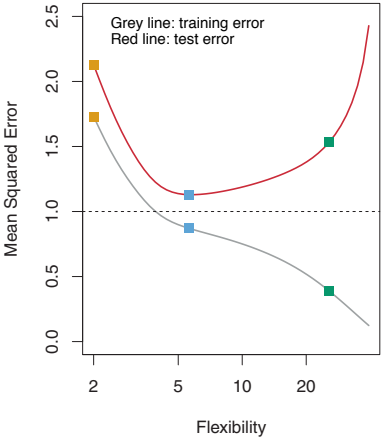
- The squiggly one. If we drew another sample of data, we'd probably get very different squiggles.

Decomposing bias-variance

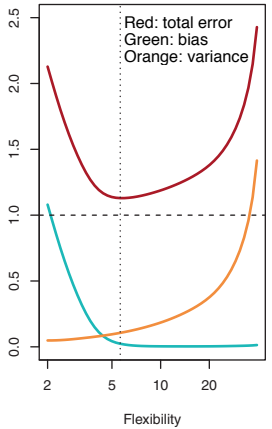


Take a moment to think about how bias and variance add up to make the red curve on the left. Try to draw bias and variance separately.

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.



Bias variance tradeoff is one of the most important concepts we'll learn.

As we learn to train different models, we'll always be seeking to balance these two sources of error.

Linear regression

Regression: A method to estimate the expected value of an output variable (y), *conditional* on one or more input values (x)

- KNN regression (end of these slides, and in textbook) does this by averaging nearby values.
- Linear regression does this by fitting a linear function to the data.
- Broadly speaking, *regression* can be used for prediction.
- *Linear* regression specifically can also be used for inference.
- Many of the methods we'll work with later in the semester will be rooted in linear regression.

The basic model

- x_i: one dimensional independent variable
- y_i: one dimensional dependent variable

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$
$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

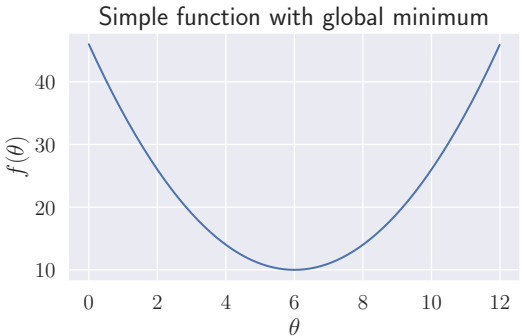
- We use the $\hat{\cdot}$ symbol to denote an estimate, or prediction

(extremely important) Side note: Optimality.

Define the "argument" that minimizes a function f with respect to θ as:

$$\theta^* = \arg\min_{\theta} f(\theta)$$

In the plot below, what's θ^*?

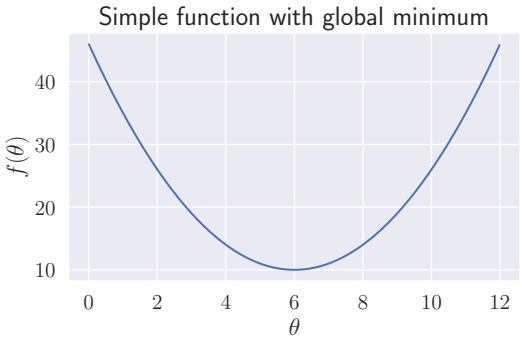


(extremely important) Side note: Optimality.

Define the "argument" that minimizes a function f with respect to θ as:

$$\theta^* = \arg\min_{\theta} f(\theta)$$

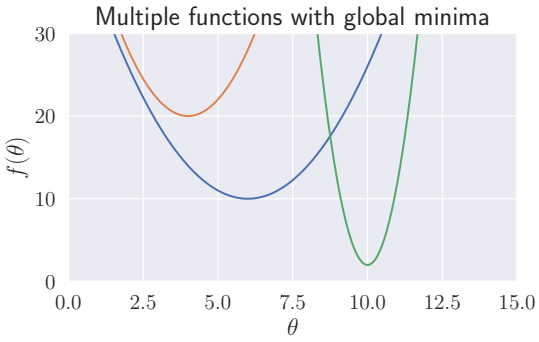
In the plot below, what's θ^*?



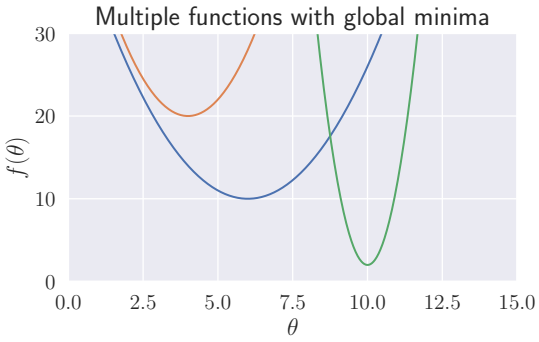
$$\theta^* = \arg\min_{\theta} f(\theta) = 6$$

$$f(\theta^*) = 10$$

What do the minima share in common?

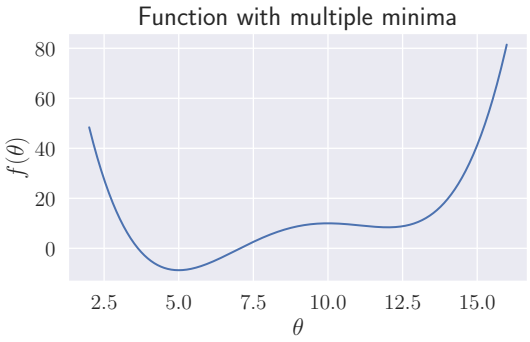


What do the minima share in common?

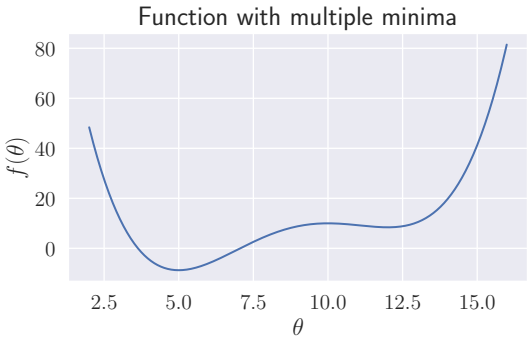


$$\frac{\partial f(\theta)}{\partial \theta}\Big|_{\theta^*} = 0$$

What's the challenge here?



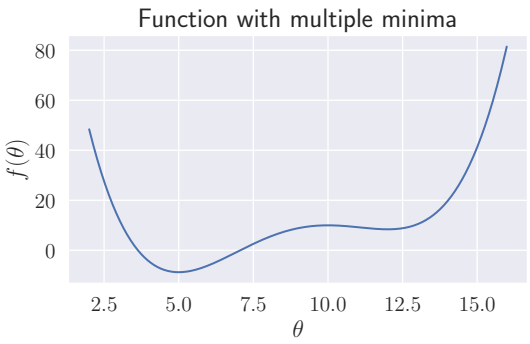
What's the challenge here?



$\frac{\partial f(\theta)}{\partial \theta} = 0$ at more than one point.

The function is said to be "non-convex"

What's the challenge here?



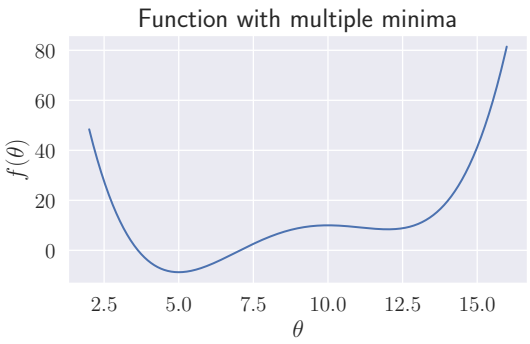
$\frac{\partial f(\theta)}{\partial \theta} = 0$ at more than one point.

The function is said to be "non-convex"

Which should we choose?

- We could enumerate all the solutions and choose the best.

What's the challenge here?



$\frac{\partial f(\theta)}{\partial \theta} = 0$ at more than one point.

The function is said to be "non-convex"

Which should we choose?

- We could enumerate all the solutions and choose the best.
- But that can get really tedious with complicated functions.

Estimation can be framed as an optimization problem

In many forms of estimation, we set up the problem as follows:

$$\{\hat{\beta}_0, \hat{\beta}_1\} = \arg\min_{\beta_0, \beta_1} J(\beta_0, \beta_1)$$

...where βs are the parameters we wish to identify.

In this course, we'll be looking at a broad variety of ways to define the *cost function*, J.

Linear regression as optimization

In “least squares” linear regression, the starting point for estimation is

$$\{\hat{\beta}_0, \hat{\beta}_1\} = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (e_i)^2$$

Linear regression as optimization

In “least squares” linear regression, the starting point for estimation is

$$\{\hat{\beta}_0, \hat{\beta}_1\} = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (e_i)^2$$
$$= \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Linear regression as optimization

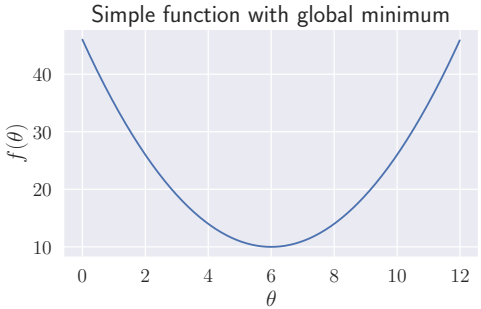
In “least squares” linear regression, the starting point for estimation is

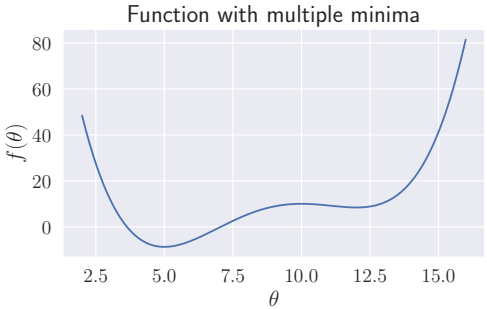
$$
\begin{aligned}
\{\hat{\beta}_0, \hat{\beta}_1\} &= \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (e_i)^2 \\
&= \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2
\end{aligned}
$$

Why choose a quadratic (squared) objective function?

Why choose a quadratic (squared) objective function?

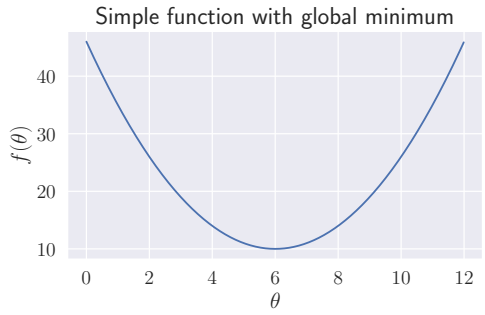
Hint:

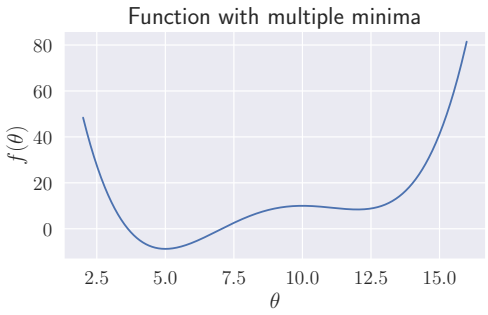




Why choose a quadratic (squared) objective function?

Hint:





With least squares, the cost function

- Has one global minimum
- Is differentiable – we can write an equation for $\frac{\partial f(\theta)}{\partial \theta} = 0$

Solving the estimation problem

$$\{\hat{\beta}_0, \hat{\beta}_1\} = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

So the optimal parameters must satisfy:

$$\frac{\partial \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2}{\partial \beta_0} = 0$$

$$\frac{\partial \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2}{\partial \beta_1} = 0$$

The solution:

$$\frac{\partial \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\partial \hat{\beta}_0} = 0 \qquad \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\partial \hat{\beta}_0} = 0 \qquad \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Parametric vs. non-parametric models

The model examples we discussed so far are **parametric**, meaning they relate inputs to outputs with a mathematical function defined by parameters.

But **non-parametric** models are also possible.

- These don't use functions with coefficients
- Instead the data *become* the model

It's easiest to see this by example using the K-nearest neighbors algorithm.

K-nearest neighbors (KNN)

We'll work with just a one-dimensional independent variable. For example,

- y_i could be NOx emissions from a power plant,
- x_i could be its coal use;
- different i would correspond to different power plants in different years.

Definitions:

- First, define proximity between two points as $|x_i - x_j|$
- Next, define $\mathcal{N}_i$ as the set of K points closest to x_i

K-nearest neighbors

The basic idea behind using KNN for regression (i.e. predicting a continuous variable or set of variables) is simple:

$$\hat{y}_j = \frac{1}{K} \sum_{i \in \mathcal{N}_j} y_i$$

In other words, the prediction equals the average of the K nearest points.

If you're working with KNN, what is your most important decision?

If you're working with KNN, what is your most important decision?

What is K?

Check of intuition: Would increasing K reduce or increase bias?

If you're working with KNN, what is your most important decision?

What is K?

Check of intuition: Would increasing K reduce or increase bias? **Increase!**

- Using a lower K would cause the estimates to more closely follow the underlying data.
- In the extreme, $K = 1$ would make the model equal the underlying data.
- At the other extreme, $K = n$ would make the model equal the sample mean.