Error

PLSC 309

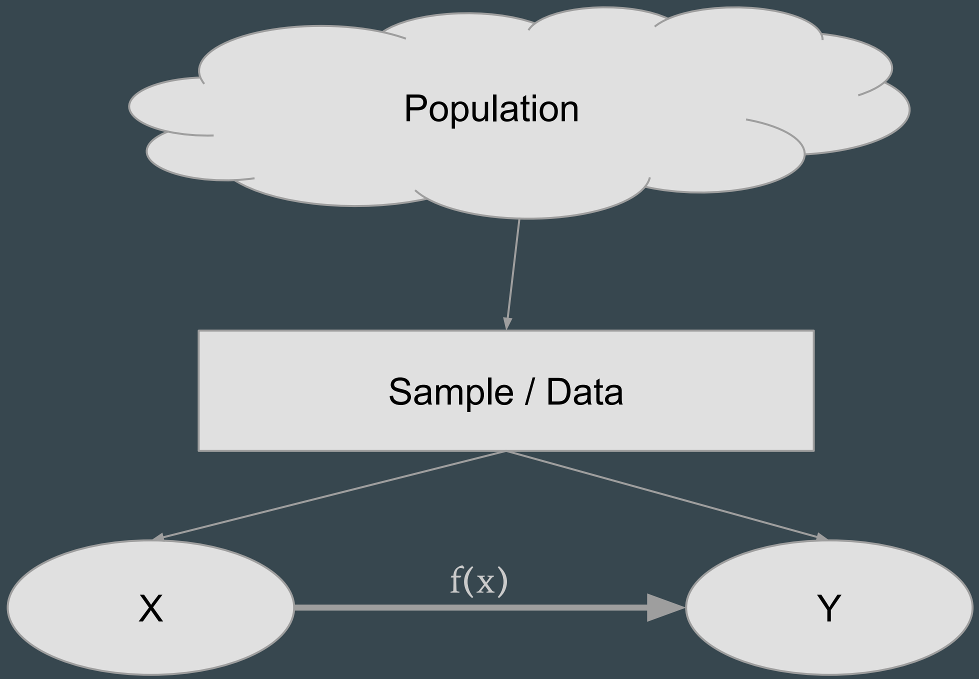
27 March 2019

Review

Last week we learned how to evaluate linear regression models:

- Independence
- Constant variance / normal / independent errors
- Temporal and spatial dependence
- Interactions
- Omitted variable bias
- Endogeneity

Review: statistical modelling



Review: systematic vs randomization error

- Both prevent your statistical model from saying something meaningful about the real world
- Systematic error is a problem with your system
  - E.g sampling bias; flaws in the data-generating process
  - You flip a weighted coin ten times and you get ten heads
- Randomization error is when you randomly have an extreme sample
  - Likelihood increases with small sample size
  - You flip a fair coin ten times and you get ten heads
- If you change nothing and just repeat your process, randomization errors will go away; systematic errors will not

Our big problem

- How do we know that the function our statistical model learns is actual what's going on in the real world?
- We can only observe the data that's in our sample
- Unobservable data is infinitely large!

A small bit of terminology

- Say we estimate a statistical model like linear regression and get a function
- We can determine how well this model meets its assumptions with the data we have
    - *In-sample error*
    - *Internal validity*
- But what about how well this model works with the rest of the world
    - *Out-of-sample error*
    - *External validity*
- Let's call the function we estimate g(X) and the function that exists in the real world f(X) to avoid confusion

Estimators

- g(X) is an *estimator* of f(X)
- *estimator*: a function that estimates
- For linear regression, g(X) is an estimator that produces estimates for
  - β
  - α
- If we knew the real-value of f(X) we could compare g(X) directly, and wouldn't need any in-sample estimates

Justification for linear regression

- If all of our assumptions are met...
- ...in other words if we have a *random sample* of sufficient size with *independent* observations
- In-sample error = out-of-sample error
- This is because the sample is *perfectly representative* of the population

In the real world we face two problems

1. Bias
    a. Our assumptions are not met
    b. Dependency between variables
    c. Non-random samples
2. Too much noise (variance)
    a. X might explain Y, but what if that explanation is very weak?
    b. Lots of moving parts, lots of random variation
    c. Leads to spurious correlations

What kind of g(X) do we want?

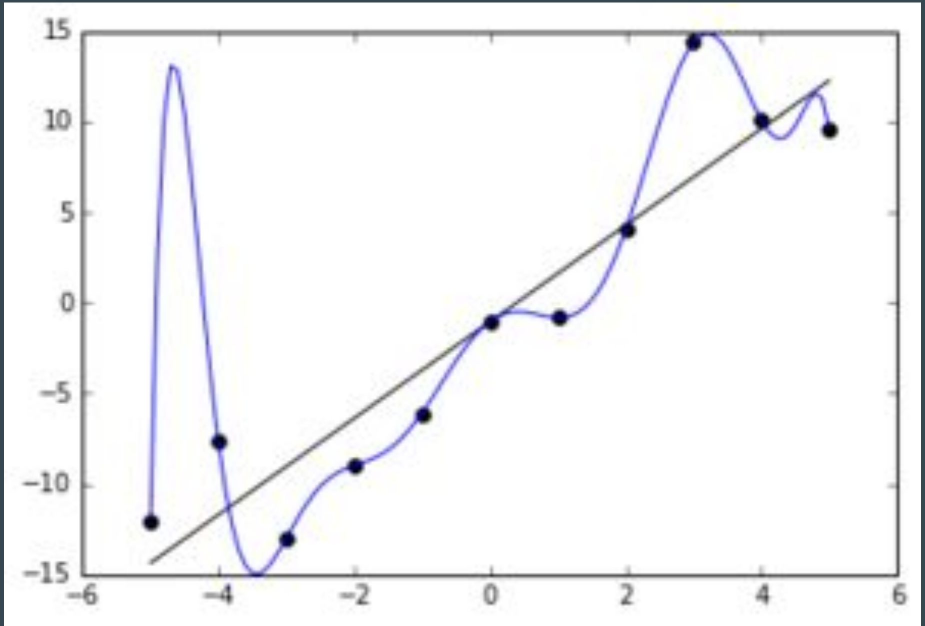
We want g(X), the function we learn from our model, to approximate some real world function f(X)

Why don’t we just connect the dots?

In other words, why don’t we just have $\hat{Y} = Y$?

- We are interested in understanding the real world
- If we understand the real world, we will be able to predict future values of Y
- The data that we work with has two problems
    - Bias
    - Variance

Overfitting



Overfitting: an intuitive definition

- Overfitting happens when our model is sticking too close to our sample
- The entire logic of statistics relies on *infinite, repeated sampling*
- If we base our entire model just on our single sample, this logic falls apart
- In that case, we “overfit” our data *---* we miss the forest for the trees

Let’s go back to our regression equation

- $Y = \beta X + \alpha$ now becomes…
- $Y = \beta X + \alpha + \varepsilon$
- ε = error term
- In other words, we’re saying the data we get is some combination of a true, linear relationship, plus some random noise

Linear regression: theory vs. reality

- $Y = \beta X + \alpha + \varepsilon$
- ε = bias + variance
- In other words, we're saying the data we get is some combination of a true, linear relationship, plus some random noise and systematic error

Error = bias + variance

First, some terminology

- $\hat{Y}$ = our estimated Y from our model
- $E(\hat{Y})$ = the expected estimate from our model if we repeated the sample process an infinite number of times
- Remember, when we put E in front of something, we are taking it's expected value
    - I.e. what that quantity would be if we repeated our whole process infinitely
- When we use Y in this context, we are talking about the entire population, not just our sample

Error = bias + variance

$$\text{Error} = E(Y - \hat{Y})^{2.}$$

- In other words, the true error for our model will be...
- The difference between our predicted values and actual values
- Squared to remove direction
- Take the expected value
- The squared difference between our predicted and actual values if we were to repeat our model infinitely

Error = bias + variance

$$E(Y - \hat{Y})^{2.} = E(Y - E(\hat{Y}))^2 + E(E(\hat{Y}) - \hat{Y})$$

- The highlighted term is the *model bias*
- It is the difference between our predicted and actual values when our model is repeated an infinite number of times
- Due to the law of large numbers, if all our assumptions are met, this should be zero!

Error = bias + variance

$$E(Y - \hat{Y})^{2.} = E(Y - E(\hat{Y}))^2 + E(E(\hat{Y}) - \hat{Y})$$

- The highlighted term is the *model variance*
- This is how much your predicted values differ from the possible range of predicted values you would get if you repeated your process an infinite number of times

Error = bias + variance

$$E(Y - \hat{Y})^{2.} = E(Y - E(\hat{Y}))^2 + E(E(\hat{Y}) - \hat{Y})$$

- Error is broken down into two components:
  - Bias
  - Variance
- Bias means you get the wrong answer
- Variance means that while your average guess might be right, any single guess could be way off from the true value

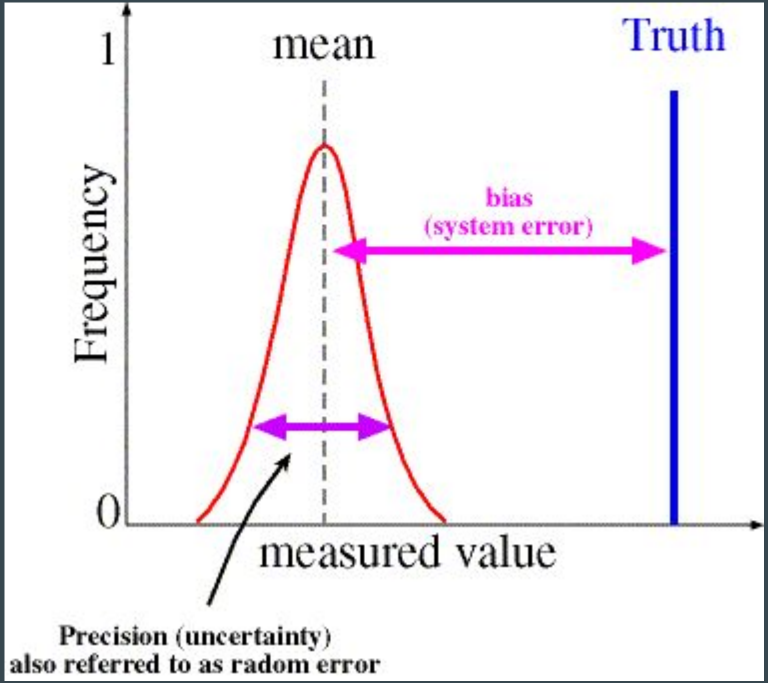
What do we want in an estimator?

- We want the estimator that approximates f(X) with the least amount of error
- There is a special term for this: Minimum Variance Unbiased Estimator (MVUE)
  - Unbiased
  - Smallest variance possible
- The MVUE is our best possible guess for f(X)

Bias

- Bias occurs when the parameter our model produces is different from the true parameter
- This is due to problems with the sample
  - Not large enough
  - Not truly random
- Repeated sampling won't fix systematic error!
- Can also happen when our assumptions are too strict

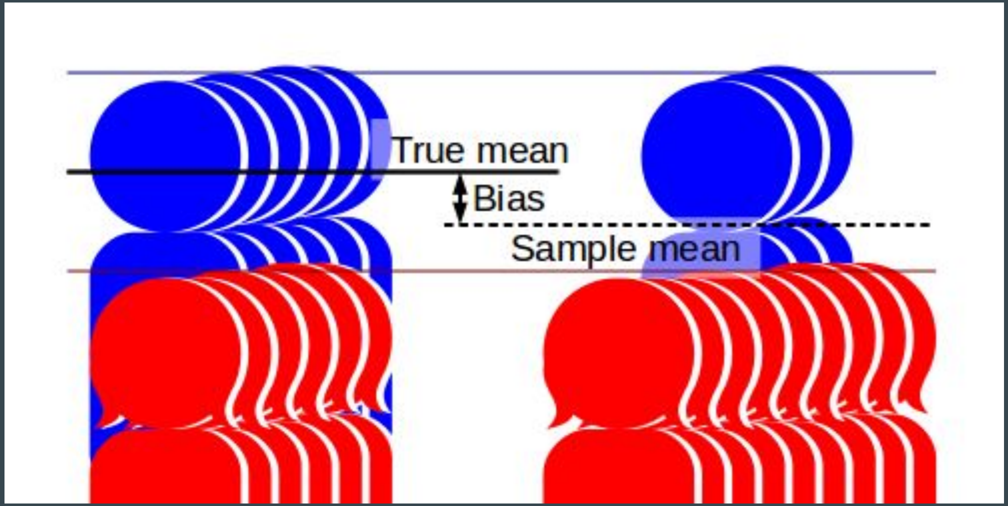
Bias


1
mean
Truth
Frequency
bias
(system error)
0
measured value
Precision (uncertainty)
also referred to as radom error

Bias, example

Bias, example

- Say you want to estimate the average height, but your sample includes more women than men



Variance

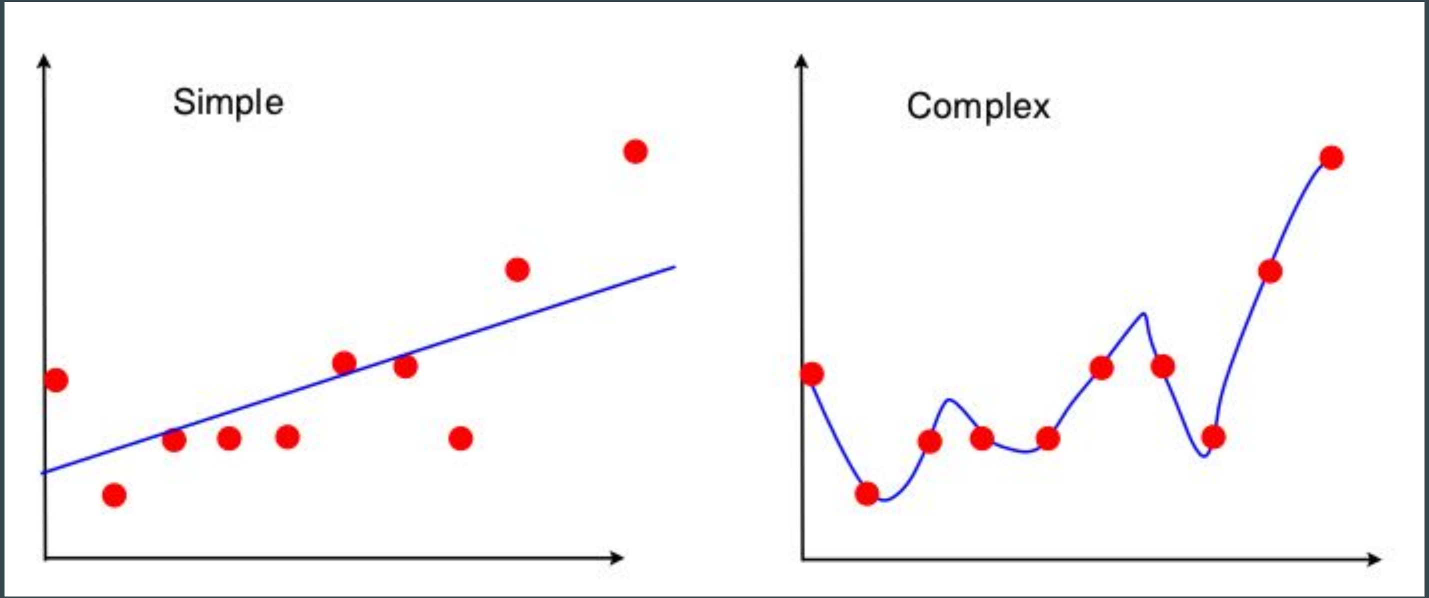
- Variance means that our function is imprecise
- g(X) jumps all over the place
- This happens when the model catches too much noise
  - Bad or imprecise measurements
  - Low sample size
  - Complex relationships

Variance, example

Say you want to determine the rate of medical bankruptcies in the U.S. You use a representative, random sample of 30 households.

- The true rate of medical bankruptcy is around 0.7%
- If one households in your sample went through a medical bankruptcy, you'd estimate 3.3%
- If zero households in your sample went through a medical bankruptcy, you'd estimate 0

Variance, example



Bias and variance in statistical modelling

- Bias vs. variance in modelling is about finding a model that balances between signal and noise
- We want our model to approximate the real world
  - That is, we want to reduce bias
- ...but we also want to get consistent predictions
  - That is, we want to reduce variance
- Statistical modelling must balance these two factors

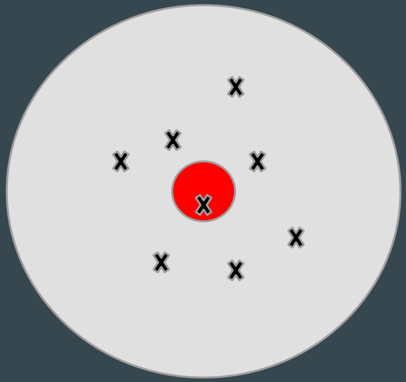
Bias and variance

Bias

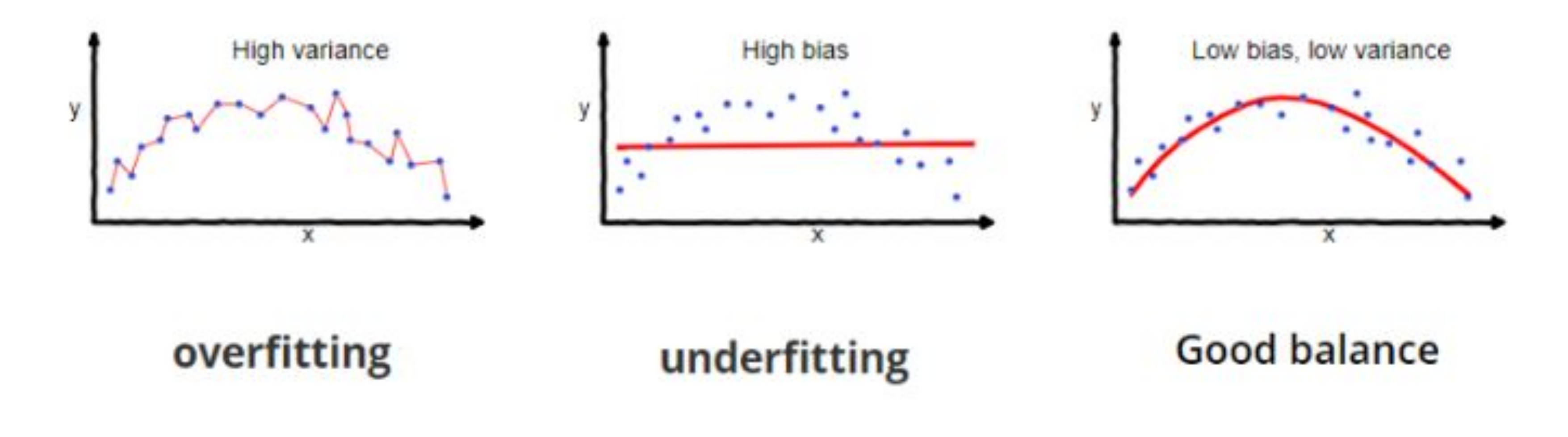
- g(x) makes wrong predictions
- g(x) should be *flexible*
- Being really good at making really bad predictions

Variance

- g(x) captures too much noise
- g(x) should be *simple*
- Being really bad about making really good predictions



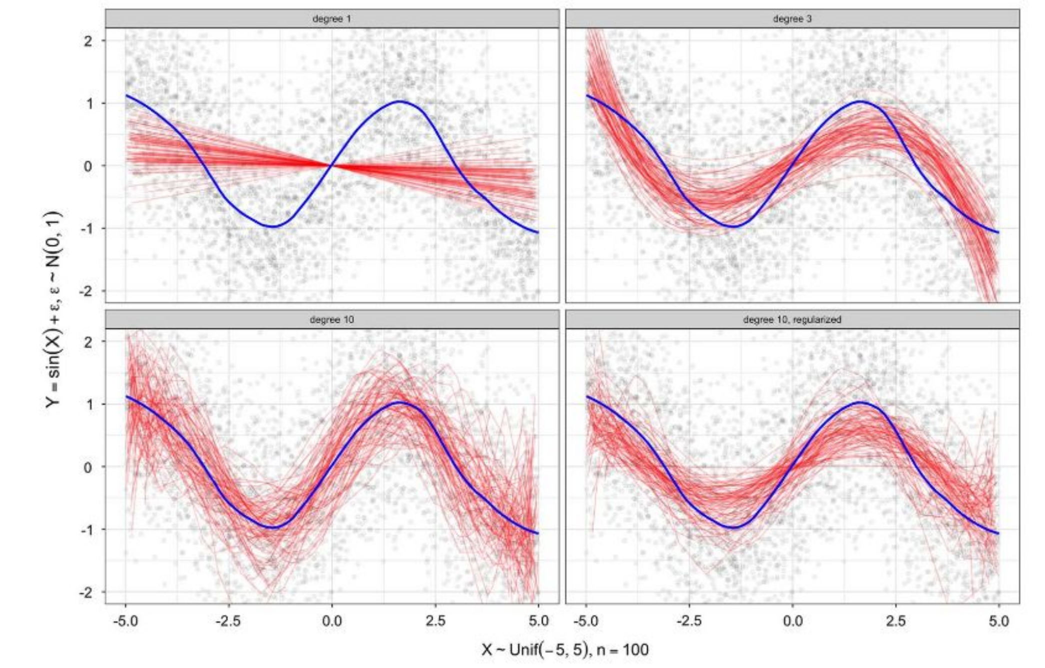
Bias and variance are a trade-off



Bias and variance

High Bias + Variance

High Variance

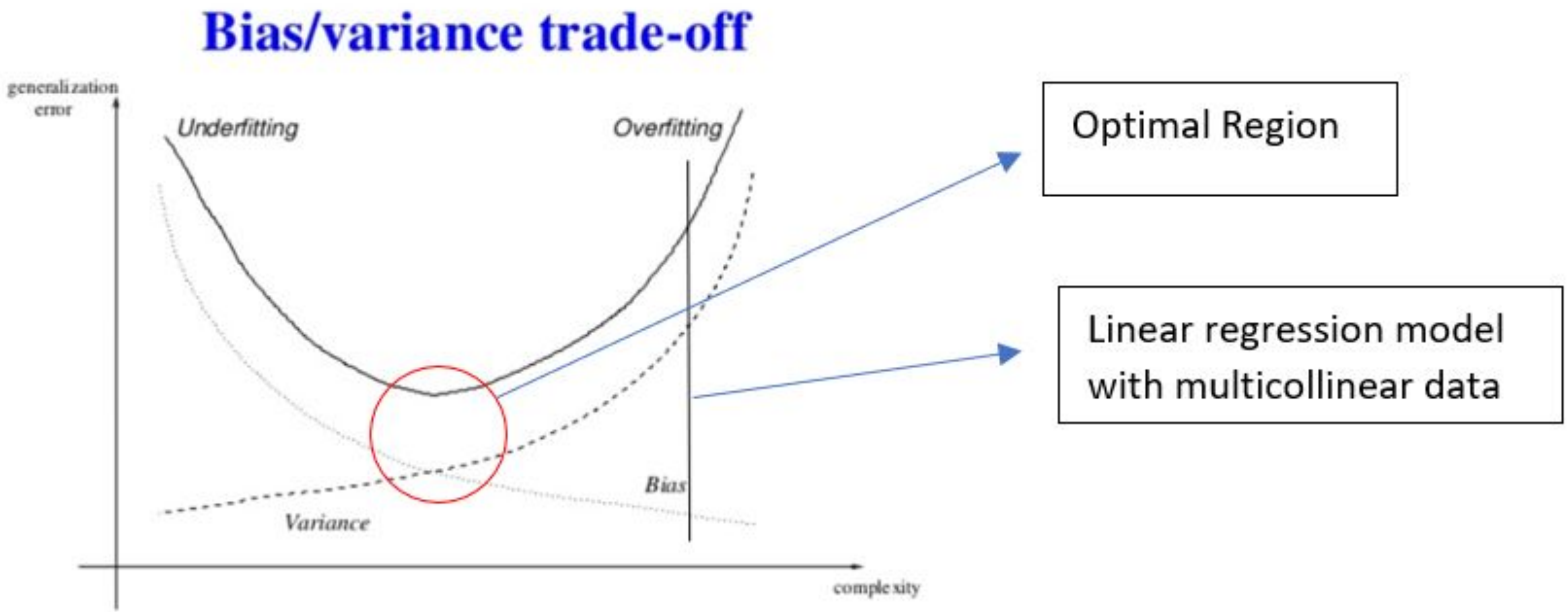


High Bias

...just right!

Source: Jones and Fariss (2015)

Bias and variance



So how do we balance bias and variance?

- Unfortunately, we can’t really calculate this directly, because we are missing a piece of crucial information
- We have no idea what the population values are
- We have are not really sure about our behavior under repeated samples
- Luckily, there is a solution to this!

Review

- We distinguished between out-of-sample and in-sample error
- While in-sample error approximates out-of-sample error if all of our assumptions are perfect, this is rarely the case!
- We found that error can be decomposed into two distinct categories
  - Bias
  - Variance

Review: bias and variance

- Bias measures how far away our models are from the true model
- Variance measures how scattered our guesses for the true model will be
- In a scenario where repeated sampling isn't possible (aka 99.9% of the time), both bias and variance will lead to greater error
- We must find *the bias variance trade-off*
    - A model that is flexible enough to get the right answers, but simple enough to give consistent predictions
- A model that solves the bias variance tradeoff is MVUE
    - Minimum variance, unbiased estimator