Resampling Methods for Uncertainty

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

Prediction and linear regression

- We have data $\{y_i, X_i\}$
- Interest on predicting y

$$y = f(X) + u \tag{1}$$

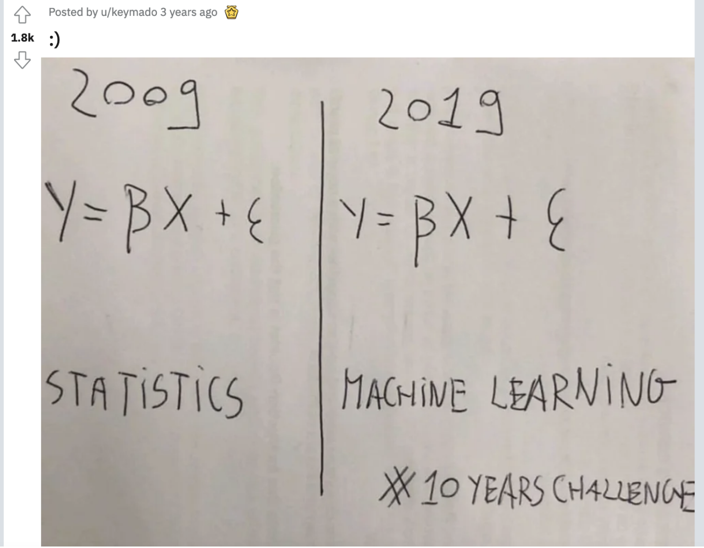
- When making a prediction we want to minimize the prediction errors
- A common loss function is the squared loss $L(e) = e^2$

Minimizing our losses

- The $E[L(e)]$ of using an estimate: $\hat{f}(x)$ can be decomposed

$$E(y-\hat{y})^2 = \underbrace{Bias^2(\hat{f}(X)) + V(\hat{f}(X))}_{Reducible} + \underbrace{Var(u)}_{Irreducible} \tag{2}$$

Prediction and linear regression



Prediction and linear regression

- We proposed

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{3}$$

- were estimating $f(X)$ boils down to finding β

Linear Regression

- Choose the estimators $\hat{\beta}$ such that we minimize the $E[L(e)]$ (SSR)

$$\hat{\beta} = \underset{\tilde{\beta}}{argmin}\, SSR(\tilde{\beta}) \tag{4}$$

- Compute β
  - QR: Householder transformation, Gram-Schmidt process (similar to FWL)
  - Gradient Descent
- Numerical Properties

Numerical Properties

- Numerical properties have nothing to do with how the data was generated
- These properties hold for every data set
- Helps in computing with big data

Agenda

Frisch-Waugh-Lovell (FWL) Theorem

- Lineal Model: $y = X\beta + u$
- Split it: $y = X_1\beta_1 + X_2\beta_2 + u$
  - $X = [X_1\ X_2]$, X is $n \times k$, X_1 $n \times k_1$, X_2 $n \times k_2$, $k = k_1 + k_2$
  - $\beta = [\beta_1\ \beta_2]$

Theorem

1. The OLS estimates of β_2 from these equations

$$y = X_1\beta_1 + X_2\beta_2 + u \tag{5}$$

$$M_{X_1}y = M_{X_1}X_2\beta_2 + residuals \tag{6}$$

   are numerically identical

2. the OLS residuals from these regressions are also numerically identical

Projection

OLS Residuals:

$$e = y - \hat{y} \tag{7}$$
$$= y - X\hat{\beta} \tag{8}$$

replacing $\hat{\beta}$

$$e = y - X(X'X)^{-1}X'y \tag{9}$$
$$= (I - X(X'X)^{-1}X')y \tag{10}$$

Define two matrices

- Projection matrix $P_X = X(X'X)^{-1}X'$
- Annihilator (residual maker) matrix $M_X = (I - P_X)$

Projection

- $P_X = X(X'X)^{-1}X'$
- $M_X = (I - P_X)$
- Both are symmetric
- Both are idempotent $(A'A) = A$
- $P_X X = X$ hence projection matrix
- $M_X X = 0$ hence annihilator matrix

We can write

$$SSR = e'e = u'M_X u \tag{11}$$

So we can relate SSR to the true error term u

Applications

- Why FWL is useful in the context of big volume of data?
- An computationally inexpensive way of
  - Removing nuisance parameters
    - E.g. the case of multiple fixed effects. The traditional way is either apply the within transformation with respect to the FE with more categories then add one dummy for each category for all the subsequent FE
  - Computing certain diagnostic statistics: Influential Observations, R^2, LOOCV.

Applications: Fixed Effects

- For example: Carneiro, Guimarães, & Portugal (2012) *AEJ: Macroeconomics*

$$\ln w_{ijft} = x_{it}\beta + \lambda_i + \theta_j + \gamma_f + \delta_t + u_{ijft} \tag{12}$$

  - Data set 31.6 million observations, with 6.4 million individuals (i), 624 thousand firms (f), and 115 thousand occupations (j), 11 years (t).
  - Storing the required indicator matrices would require 23.4 terabytes of memory
  - From their paper
    "In our application, we first make use of the Frisch-Waugh-Lovell theorem to remove the influence of the three high- dimensional fixed effects from each individual variable, and, in a second step, implement the final regression using the transformed variables. With a correction to the degrees of freedom, this approach yields the exact least squares solution for the coefficients and standard errors"

Applications: Influential Observations

Note the following

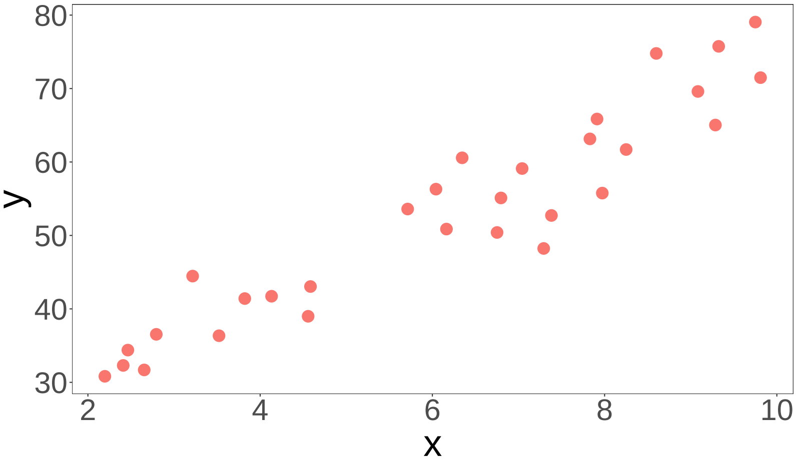
$$\hat{\beta} = (X'X)^{-1}X'y \tag{13}$$

each element of the vector of parameter estimates $\hat{\beta}$ is simply a weighted average of the elementes of the vector y

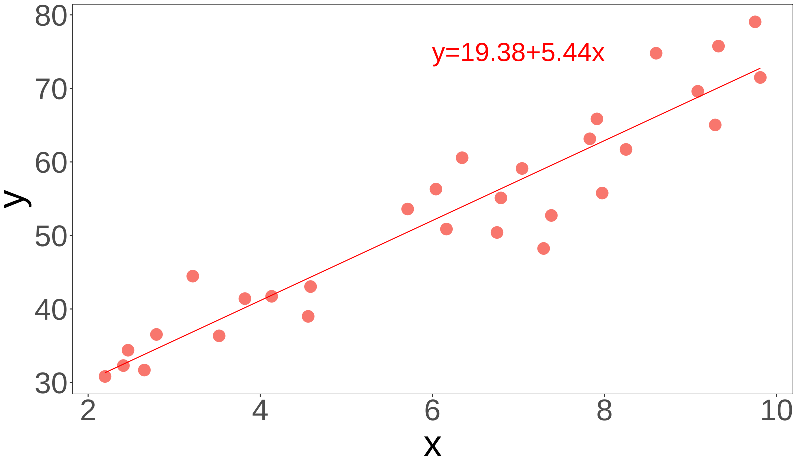
Let's call c_j the j-th row of the matrix $(X'X)^{-1}X'$ then

$$\hat{\beta}_j = c_j y \tag{14}$$

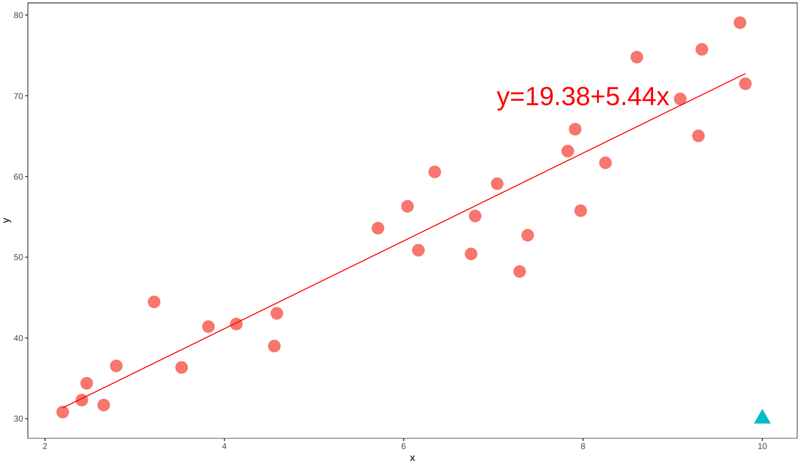
Applications: Influential Observations



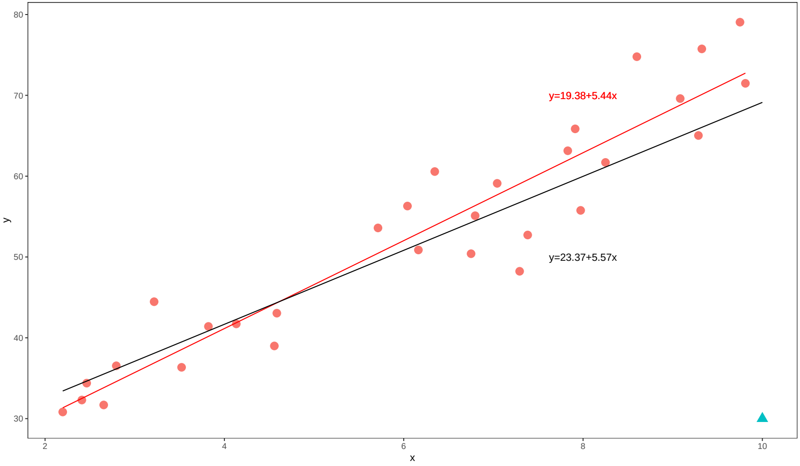
Applications: Influential Observations



Applications: Influential Observations



Applications: Influential Observations



Applications: Influential Observations

Consider a dummy variable e_j which is an $n - vector$ with element j equal to 1 and the rest is 0. Include it as a regressor

$$y = X\beta + \alpha e_j + u \tag{15}$$

using FWL we can do

$$M_{e_j} y = M_{e_j} X\beta + r \tag{16}$$

- β and *residuals* from both regressions are identical
- Same estimates as those that would be obtained if we deleted observation j from the sample. We are going to denote this as $\beta^{(j)}$

Note:

- $M_{e_j} = I - e_j(e_j' e_j)^{-1} e_j'$
- $M_{e_j} y = y - e_j(e_j' e_j)^{-1} e_j' y = y - y_j e_j$
- $M_{e_j} X$ is X with the j *row* replaced by zeros

Applications: Influential Observations

Let's define a new matrix $Z = [X, e_j]$

$$y = X\beta + \alpha e_j + u \tag{17}$$

$$y = Z\theta + u \tag{18}$$

we can write it as

$$y = P_Z y + M_z y \tag{19}$$

$$= X\hat{\beta}^{(j)} + \hat{\alpha} e_j + M_Z y \tag{20}$$

Pre-multiply by P_X (remember $M_Z P_X = 0$)

$$P_X y = X\hat{\beta}^{(j)} + \hat{\alpha} P_X e_j \tag{21}$$

$$X\hat{\beta} = X\hat{\beta}^{(j)} + \hat{\alpha} P_X e_j \tag{22}$$

$$X(\hat{\beta} - \beta^{(j)}) = \hat{\alpha} P_X e_j \tag{23}$$

Applications: Influential Observations

How to calculate α? FWL once again

$$M_X y = \hat{\alpha} M_X e_j + res \tag{24}$$

$$\hat{\alpha} = (e_j' M_X e_j)^{-1} e_j' M_X y \tag{25}$$

- $e_j' M_X y$ is the j element of $M_X y$, the vector of residuals from the regression including all observations
- $e_j' M_x e_j$ is just a scalar, the diagonal element of M_X
  Then

$$\hat{\alpha} = \frac{\hat{u}_j}{1 - h_j} \tag{26}$$

where h_j is the j diagonal element of P_X

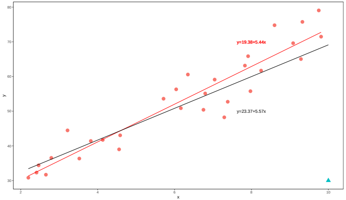
Applications: Influential Observations

Finally we get

$$(\hat{\beta}^{(j)} - \hat{\beta}) = -\frac{1}{1-h_j}(X'X)^{-1}X_j'\hat{u}_j \tag{27}$$

Influence depends on two factors

- $\hat{u}_j$ large residual $\rightarrow$ related to y coordinate
- $\hat{h}_j$ $\rightarrow$ related to x coordinate



HW. case of $y = \alpha + \beta x + u$ (ISLR)

Agenda

1 Review
- FWL

2 Uncertainty
- Resampling methods
- Parameter Assessment
  - Example: Elasticity of Demand for Gasoline
- Model Assessment
  - AIC: Akaike Information Criterion
  - SIC/BIC: Schwarz/Bayesian Information Criterion
  - Cross-Validation

3 Recap

Motivation

- The real world is messy.
- Recognizing this mess will differentiate a sophisticated and useful analysis from one that is hopelessly naive.
- This is especially true for highly complicated models, where it becomes tempting to confuse signal with noise.
- The ability to deal with this mess and noise is the most important skill you need.

Agenda

What are resampling methods?

- Tools that involves repeatedly drawing samples and refitting a model of interest on each sample in order to obtain more information about the fitted model
  - Parameter Assessment: estimate standard errors
  - Model Assessment: finding the best model

Agenda

The Bootstrap

Introduction

- Suppose we have $y_1, y_2, \ldots, y_n$ iid $Y \sim (\mu, \sigma^2)$ (both finite)
- We want to estimate

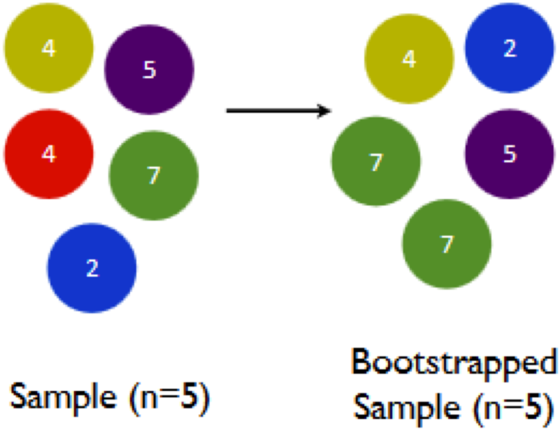
$$Var(\bar{Y}) \tag{28}$$

The Bootstrap

Introduction

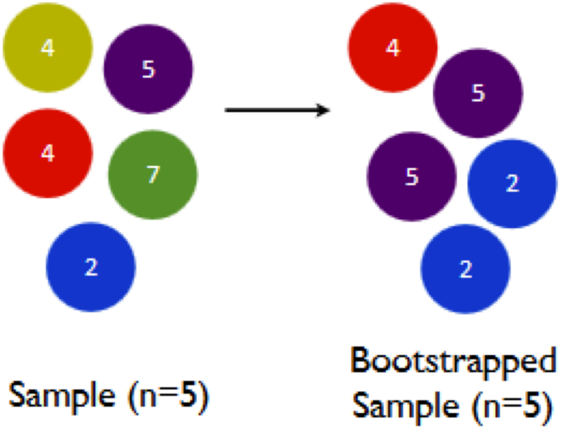
- Alternative way (no formula!)
  1. From the n original data points $y_1, y_2, \ldots, y_n$ take a sample *with replacement* of size n
  2. Calculate the sample average of this *"pseudo-sample"* (Bootstrap sample)
  3. Repeat this B times.
  4. Compute the variance of the B means

The Bootstrap
Sampling with replacement



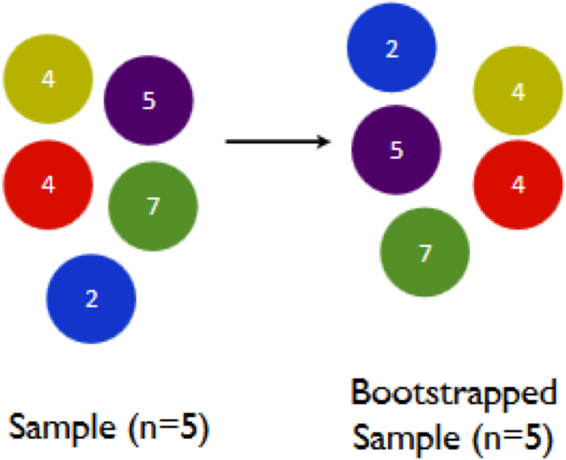
The Bootstrap

Sampling with replacement



The Bootstrap

Sampling with replacement



- Alternative way (no formula!)
  1. From the n original data points $y_1, y_2, \ldots, y_n$ take a sample *with replacement* of size n
  2. Calculate the sample average of this *"pseudo-sample"* (Bootstrap sample)
  3. Repeat this B times.
  4. Compute the variance of the B means

The Bootstrap

Variance in Linear Regression

- Suppose we have $\{y_i, X_i\}$ $i = \{1, \ldots, n\}$iid
- We want to estimate $Var(\hat{\beta})$

The Bootstrap

- Sometimes the analytical expression of the variance can be quite complicated.
- In these cases bootstrap can be useful
- In German the expression *an den eigenen Haaren aus dem Sumpf zu ziehen* nicely captures the idea of the bootstrap – *“to pull yourself out of the swamp by your own hair.”*

The Bootstrap

Two key properties

- Two key properties of bootstrapping that make this seemingly crazy idea actually work.
    1. Each bootstrap sample must be of the same size (n) as the original sample
    2. Each bootstrap sample must be taken with replacement from the original sample.

The Bootstrap

- In general terms:
  - Sample $\{y_i, X_i\}$ $i = 1, \dots, n$ iid
  - θ is the magnitude of interest

1. Sample of size n with replacement (*bootstrap sample*)
2. Compute $\hat{\theta}_j$ $j = 1, \dots, B$
3. Repeat B times
4. Calculate the magnitude of interest

Example: Elasticity of Demand for Gasoline

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

The Bootstrap

Why it works?

- The key is that the distribution of any estimator or statistic is determined by the distribution of the data.
- While the latter is unknown it can be estimated by the empirical distribution of the data.

Agenda

Prediction

- Objective predict y given X.

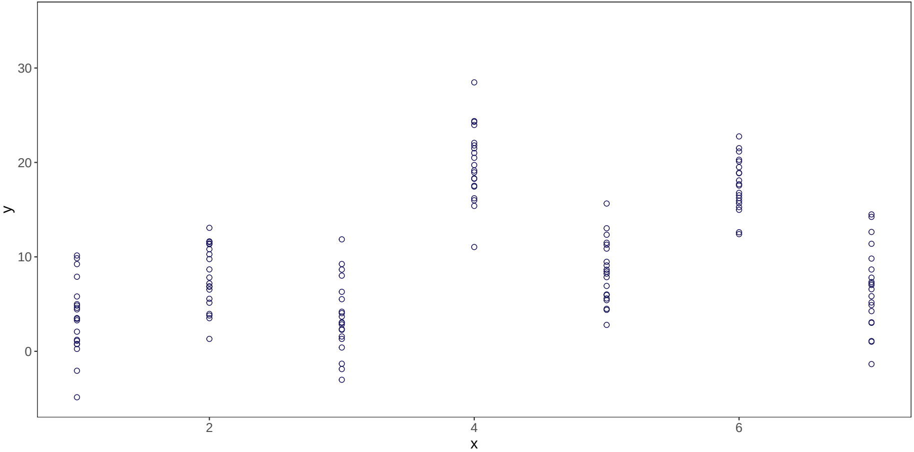
$$y = f(X) + u \tag{29}$$

- u rv $E(u) = 0$ and $V(u) = \sigma^2$

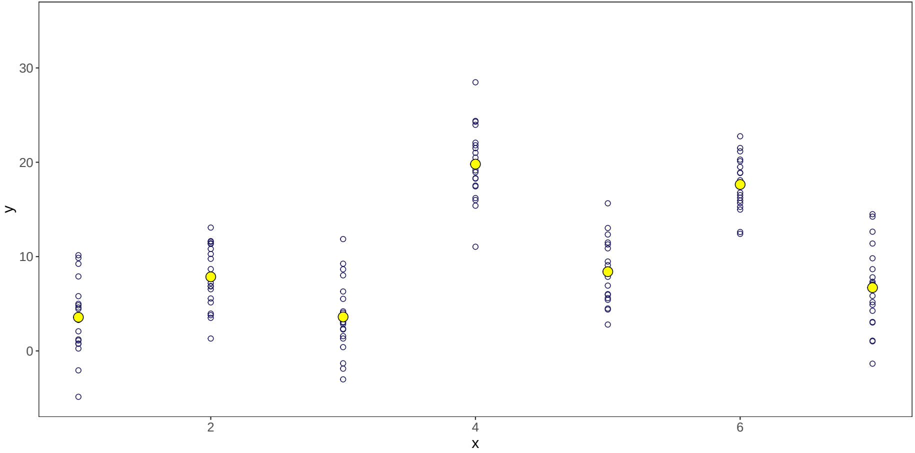
In-Sample Prediction and Overfit

- How do we select a model with the lowest prediction error?

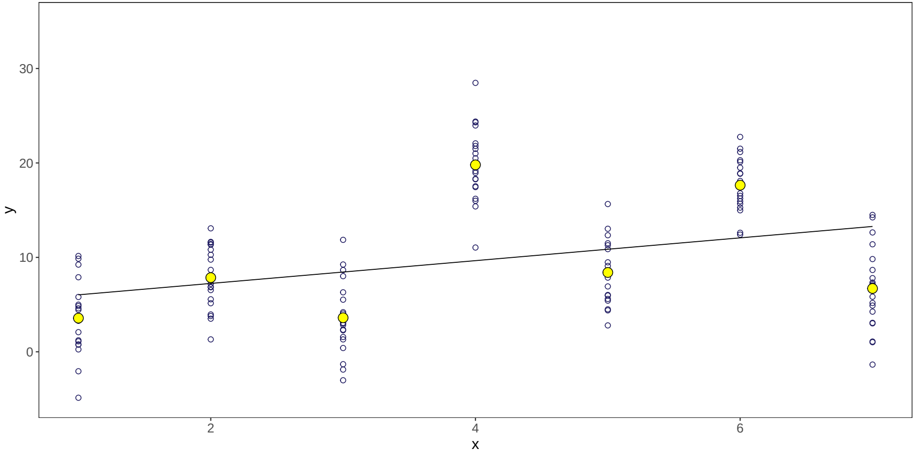
In-Sample Prediction and Overfit



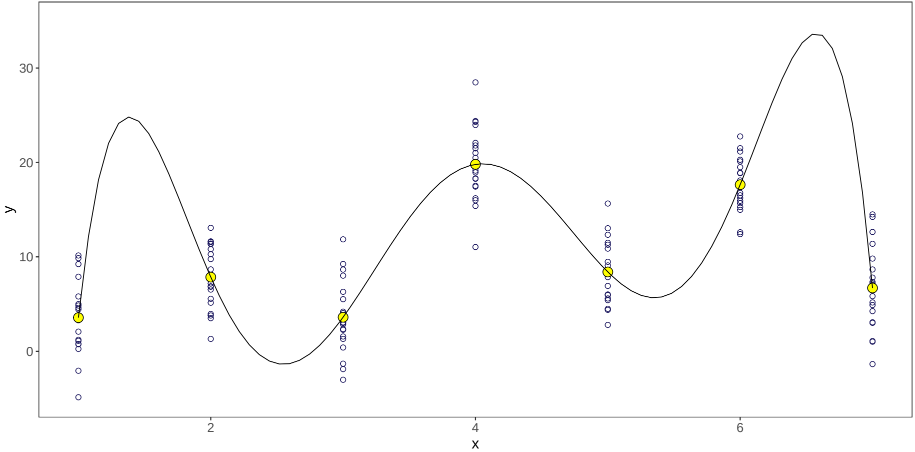
In-Sample Prediction and Overfit



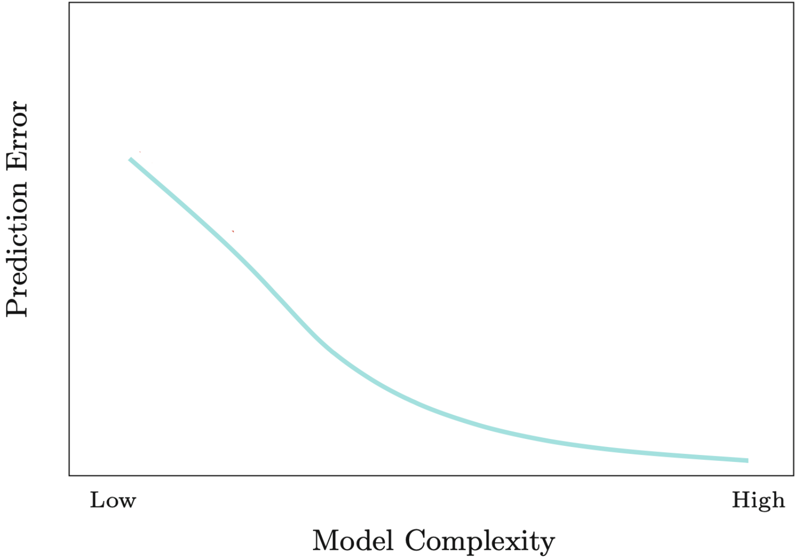
In-Sample Prediction and Overfit



In-Sample Prediction and Overfit



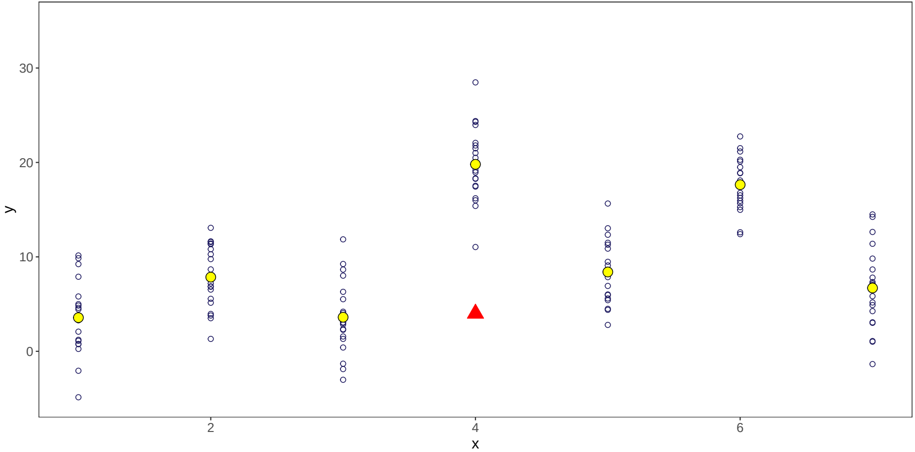
In-Sample Prediction and Overfit



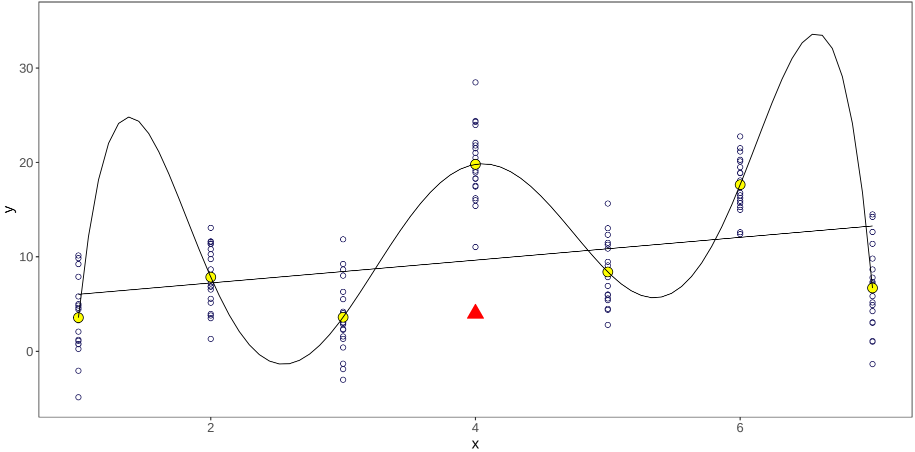
Out-of-Sample Prediction and Overfit

- ML we care about out of sample prediction

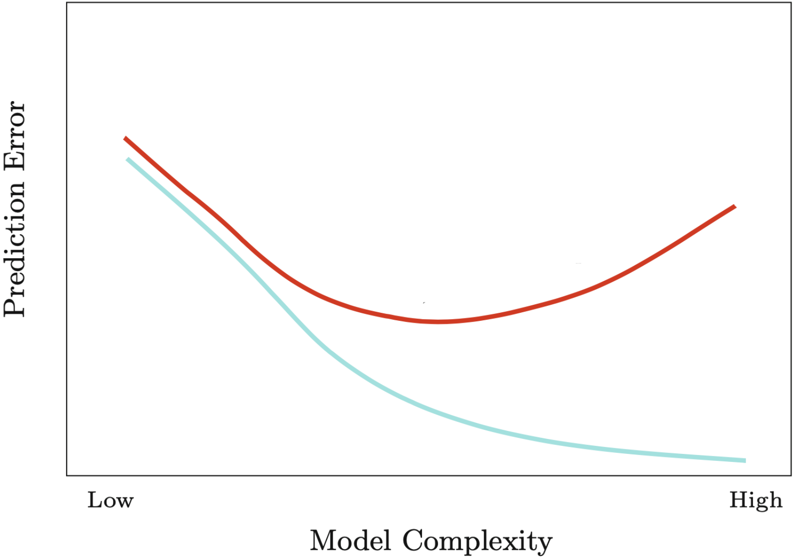
Out-of-Sample Prediction and Overfit



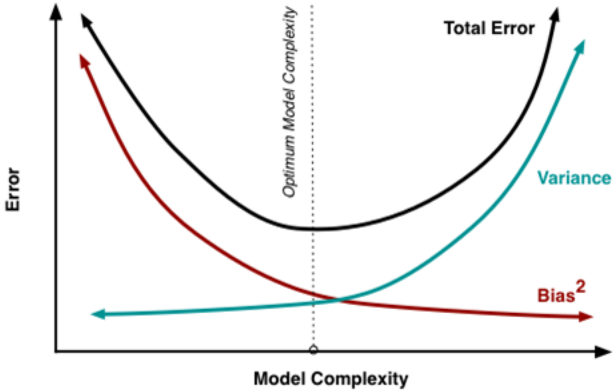
Out-of-Sample Prediction and Overfit



Out-of-Sample Prediction and Overfit

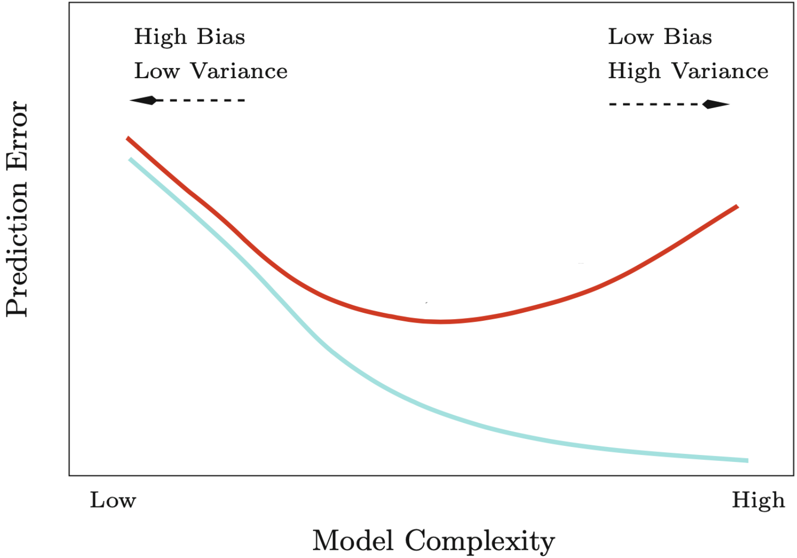


Out-of-Sample Error and the Bias Variance Trade-Off



$$E(y-\hat{y})^2 = \underbrace{Bias^2(\hat{f}(X)) + V(\hat{f}(X))}_{Reducible} + \underbrace{Var(u)}_{Irreducible} \tag{30}$$

Out-of-Sample Error and the Bias Variance Trade-Off



Out-of-Sample Prediction and Overfit

- ML we care about out of sample prediction
- How we estimate the out of sample error?

In-Sample and Out-of-Sample Errors.

- Two important concepts
  - *Training error*:

$$Err_{\mathcal{T}rain} = MSE[(y, \hat{y})|\mathcal{T}rain] \tag{31}$$

  - *Test Error*:

$$Err_{\mathcal{T}est} = MSE[(y, \hat{y})|\mathcal{T}est] \tag{32}$$

Test Error

- How dowe estimate the $Err_{\mathcal{T}est}$
- Two ways
  - Ex post penalization: AIC, BIC, Adj R^2

Test Error

AIC

- Akaike (1969)
- Minimize

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{33}$$

Test Error

SIC/BIC

- Schwarz (1978)
- Minimize

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - \frac{1}{2}p_j log(n) \tag{34}$$

Test Error

AIC vs BIC

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{35}$$

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j\frac{1}{2}log(n) \tag{36}$$

Test Error

- How dowe estimate the $Err_{\mathcal{T}est}$
- Two ways
  - Ex post penalization: AIC, BIC, Adj R^2
  - Resampling methods

Test Error

Cross-Validation

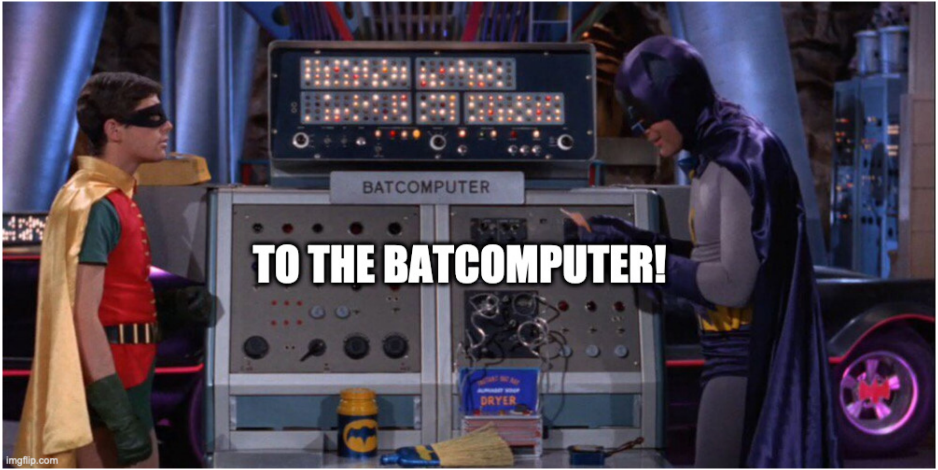


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

Recap

- Review + FWL
- Resampling Methods:
  - Parameter Assessment: estimate standard errors
  - Model Assessment: finding the best model