Resampling Methods for Uncertainty

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Review
   - FWL
2. Uncertainty
   - Resampling methods
   - Parameter Assessment
     - Example: Elasticity of Demand for Gasoline
   - Model Assessment
     - Generalization. Out-of-sample Performance
     - Out-of-Sample Error Estimation
3. Recap

Agenda

1. Review
   - FWL
2. Uncertainty
   - Resampling methods
   - Parameter Assessment
     - Example: Elasticity of Demand for Gasoline
   - Model Assessment
     - Generalization. Out-of-sample Performance
     - Out-of-Sample Error Estimation
3. Recap

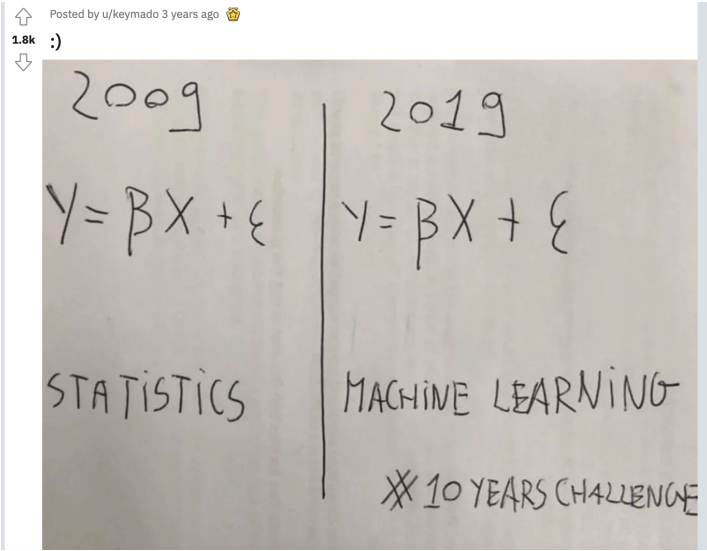
Prediction and linear regression

- 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 $L(y, \hat{y}) = (y - \hat{y})^2$

Prediction and linear regression



Prediction and linear regression

- We proposed

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2}$$

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

Linear Regression

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

$$\hat{\beta} = \underset{\tilde{\beta}}{argmin}\, SSR(\tilde{\beta}) \tag{3}$$

- 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

1. Review
   - FWL
2. Uncertainty
   - Resampling methods
   - Parameter Assessment
     - Example: Elasticity of Demand for Gasoline
   - Model Assessment
     - Generalization. Out-of-sample Performance
     - Out-of-Sample Error Estimation
3. Recap

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

$$y = X_1\beta_1 + X_2\beta_2 + u \tag{4}$$
$$res_{y \to X_1} = res_{X_2 \to X_1}\beta_2 + \epsilon \tag{5}$$

   are numerically identical

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

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*

$$\ln w_{ijft} = x_{it}\beta + \lambda_i + \gamma_f + \theta_j + \alpha_0 t + u_{ijft} \tag{6}$$

- Data set 31.6 million observations (n), with 6.4 million individuals (i), 624 thousand firms (f), and 115 thousand occupations (j).

Applications: Fixed Effects

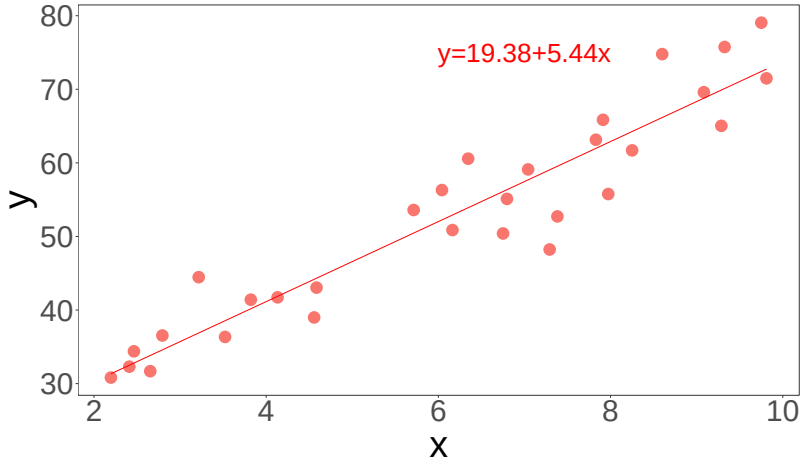
- For example: Carneiro, Guimarães, & Portugal (2012) *AEJ: Macroeconomics*

$$\ln w_{ijft} = x_{it}\beta + \lambda_i + \gamma_f + \theta_j + \alpha_0 t + u_{ijft} \tag{6}$$

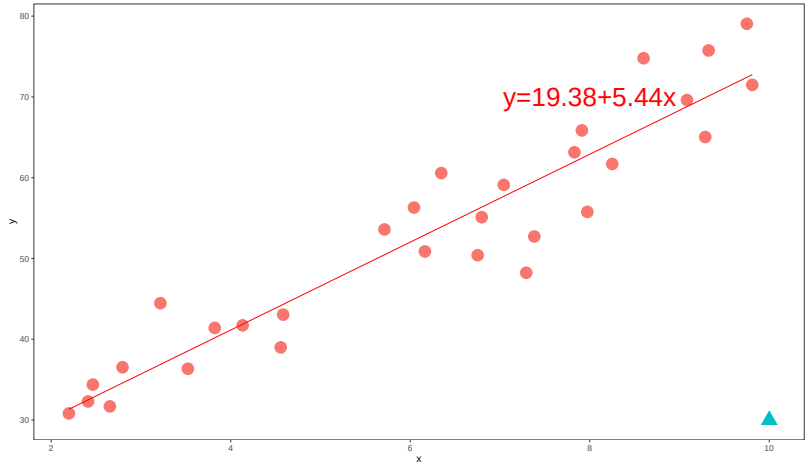
- Data set 31.6 million observations (n), with 6.4 million individuals (i), 624 thousand firms (f), and 115 thousand occupations (j).
- Storing the required indicator matrices would require 23.4 terabytes of memory

"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"

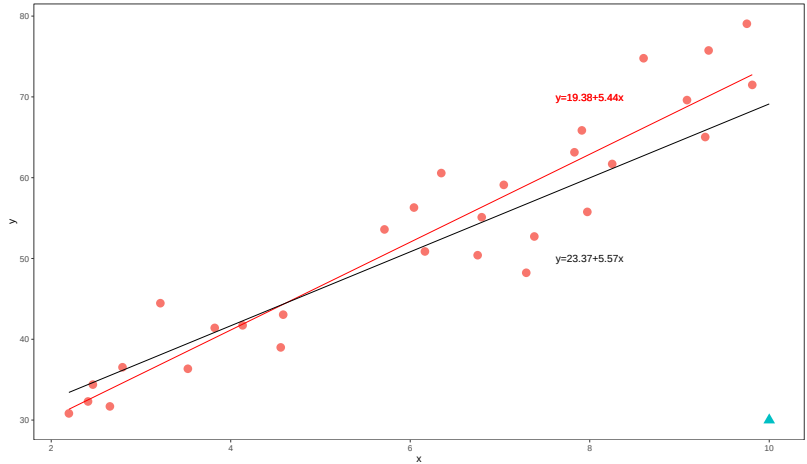
Applications: Influential Observations



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

$$y = X\beta + \alpha e_j + u \tag{7}$$

using FWL we can do

Applications: Influential Observations

Detour: Projection Matrices

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

Applications: Influential Observations

FWL shortcut with matrices

$$M_{e_j} y = M_{e_j} X\beta + \epsilon \tag{8}$$

- 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 $\beta^{(-j)}$

Check for HW:

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

Applications: Influential Observations

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

$$y = X\beta + \alpha e_j + u \tag{9}$$

$$y = Z\theta + u \tag{10}$$

we can write it as

$$y = P_Z y + M_z y \tag{11}$$

$$= X\hat{\beta}^{(-j)} + \hat{\alpha} e_j + M_Z y \tag{12}$$

Pre-multiply by P_X (HW show $M_Z P_X = 0$)

$$P_X y = X\hat{\beta}^{(-j)} + \hat{\alpha} P_X e_j \tag{13}$$

$$X\hat{\beta} = X\hat{\beta}^{(-j)} + \hat{\alpha} P_X e_j \tag{14}$$

$$X(\hat{\beta} - \beta^{(-j)}) = \hat{\alpha} P_X e_j \tag{15}$$

Applications: Influential Observations

How to calculate α? FWL once again

$$M_X y = \hat{\alpha} M_X e_j + res \tag{16}$$

$$\hat{\alpha} = (e_j' M_X e_j)^{-1} e_j' M_X y \tag{17}$$

- $e_j' M_X y$ 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

$$\hat{\alpha} = \frac{\hat{u}_j}{1 - h_j} \tag{18}$$

where h_j is the j diagonal element of P_X

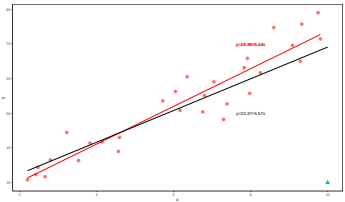
Applications: Influential Observations

Finally we get

$$(\hat{\beta}^{(-j)} - \hat{\beta}) = -\frac{1}{1-h_j}(X'X)^{-1}X_j'\hat{u}_j \tag{19}$$

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
     - Generalization. Out-of-sample Performance
     - Out-of-Sample Error Estimation
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.
- Two approaches
  - Analytical
  - Simulation: Resampling Methods

Motivation

Analytical: Example 1

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

$$Var(\bar{Y}) \tag{20}$$

Motivation

Analytical: Example 1

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

$$Var(\bar{Y}) \tag{21}$$

Motivation

Analytical: Example 2

- Suppose we have $y_i = \beta X_i + \epsilon_i \; i = 1, \ldots, n \; E(\epsilon|X) = 0 \; V(\epsilon|X) = \sigma^2 I$

$$\hat{\beta} = (X'X)^{-1}X'y \tag{22}$$

- What is the variance and sampling distribution?

Motivation

Analytical: Example 3

- Suppose we have wages w_i $i = 1, \ldots, n$ iid and we want to estimate an inequality measure:

$$Gini = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left| w_i - w_j \right|}{2n^2 \bar{w}} \tag{23}$$

- What is the variance and sampling distribution?

Agenda

1. Review
   - FWL
2. Uncertainty
   - Resampling methods
   - Parameter Assessment
     - Example: Elasticity of Demand for Gasoline
   - Model Assessment
     - Generalization. Out-of-sample Performance
     - Out-of-Sample Error Estimation
3. Recap

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, e.g. estimate standard errors
  - Model Assessment, e.g. finding the best model
  - They are computationally expensive! But these days we have powerful computers

Agenda

1. Review
   - FWL
2. Uncertainty
   - Resampling methods
   - Parameter Assessment
     - Example: Elasticity of Demand for Gasoline
   - Model Assessment
     - Generalization. Out-of-sample Performance
     - Out-of-Sample Error Estimation
3. Recap

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

$$Var(\bar{Y}) \tag{24}$$

- 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

- 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 $\{y_i, X_i\}$ $i = 1, \ldots, n$ iid
  - θ is the magnitude of interest

1. Sample of size n with replacement (*bootstrap sample*)
2. Compute $\hat{\theta}_j$ $j = 1, \ldots, B$
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?

- This method allows us to understand the variability of a statistic without relying on theoretical distribution assumptions.
- 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.

The Bootstrap

How Many Bootstrap Replications?

- The number of bootstrap replications B is a trade-off between accuracy and computation cost.
- Computation cost is linear in B, while accuracy (e.g., standard errors or p-values) improves as $B^{-1/2}$.
- In empirical research, preliminary calculations can use a modest B, but final results should use a larger B.
- Recommended values: $B = 10,000$ for final calculations, $B = 1,000$ as a minimal choice, and $B = 100$ for rough, quick estimates.

Agenda

1. Review
   - FWL
2. Uncertainty
   - Resampling methods
   - Parameter Assessment
     - Example: Elasticity of Demand for Gasoline
   - Model Assessment
     - Generalization. Out-of-sample Performance
     - Out-of-Sample Error Estimation
3. Recap

Generalization Overview

- In ML we care in out-of-sample prediction
- Generalization refers to a model's performance on unseen data.
- The ultimate goal is **not** minimizing the in-sample loss, but achieving low error out-of-sample on unseen data.

Training and Test Loss

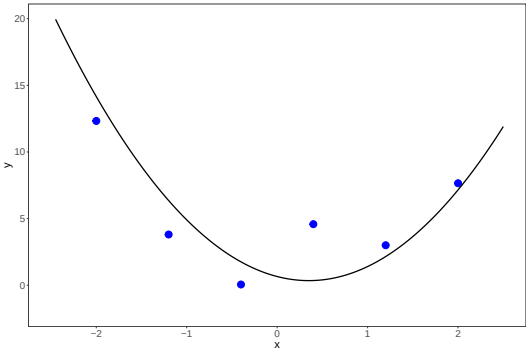
- Unseen data is typically referred as **test data**,
- While the sample data is called the **training data**.
- The expected loss over the test distribution is called the test loss.
- When the loss function is quadratic we have the EMSE:

$$\mathbb{E}[L(\theta)] = \mathbb{E}[(y - f_\theta(X))^2] \tag{25}$$

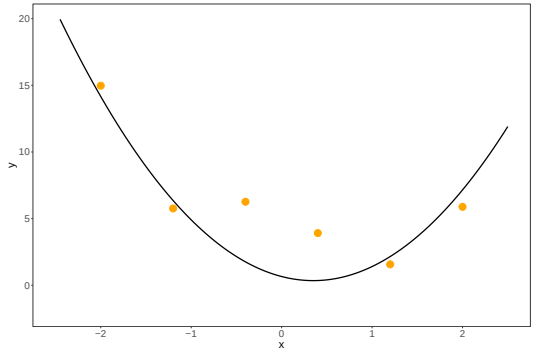
Overfitting and Underfitting

- Successfully minimizing training error does not always result in a small test error.
- A model is said to overfit if it predicts accurately on training data but poorly on test (unseen) data.
- A model underfits if its training error is relatively large, which usually means test error is also large.
- Understanding overfitting and underfitting helps in choosing appropriate model parameterizations.

Overfitting and Underfitting. Bias-Variance Tradeoff



(a) Training Data



(b) Testing Data

Overfitting and Underfitting. Bias-Variance Tradeoff

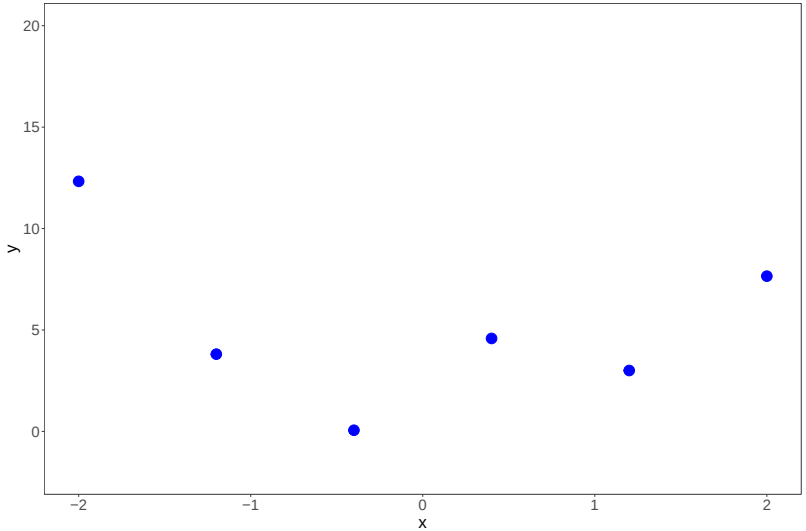


Figure 2: Training Data

Overfitting and Underfitting. Bias-Variance Tradeoff

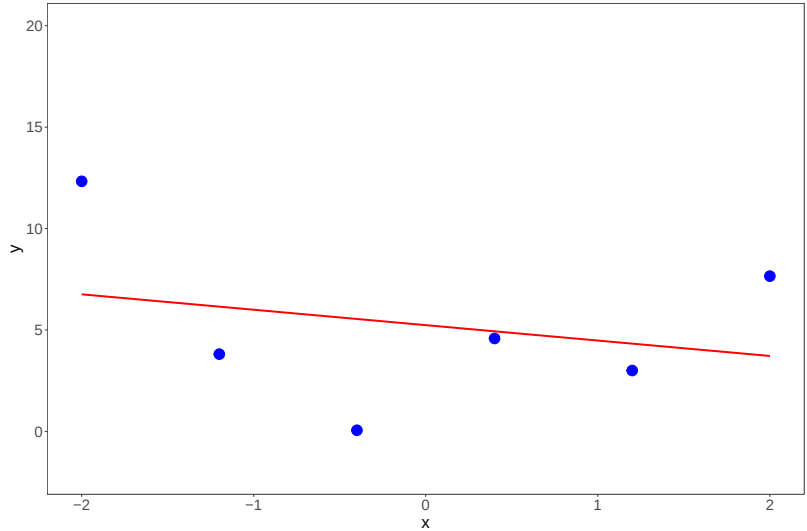
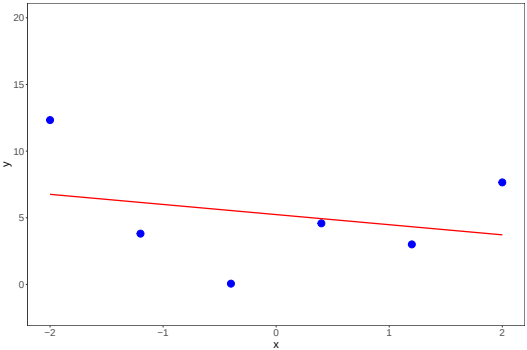


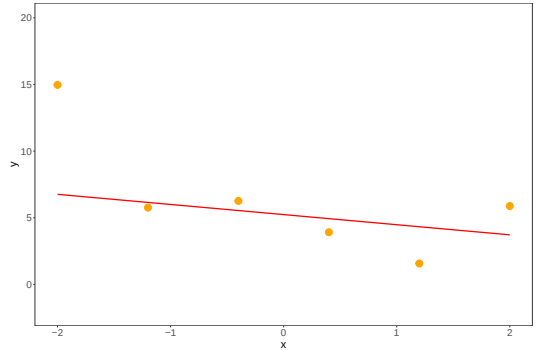
Figure 3: Training Data

Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Performance



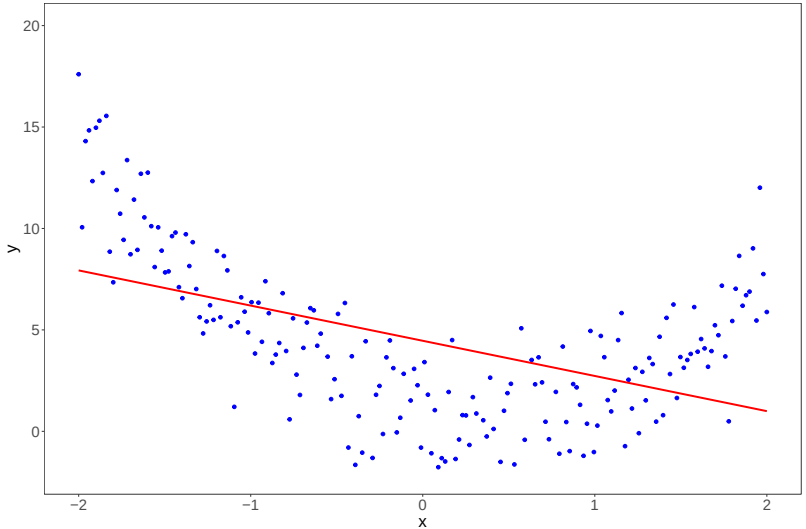
(a) Training Data



(b) Testing Data

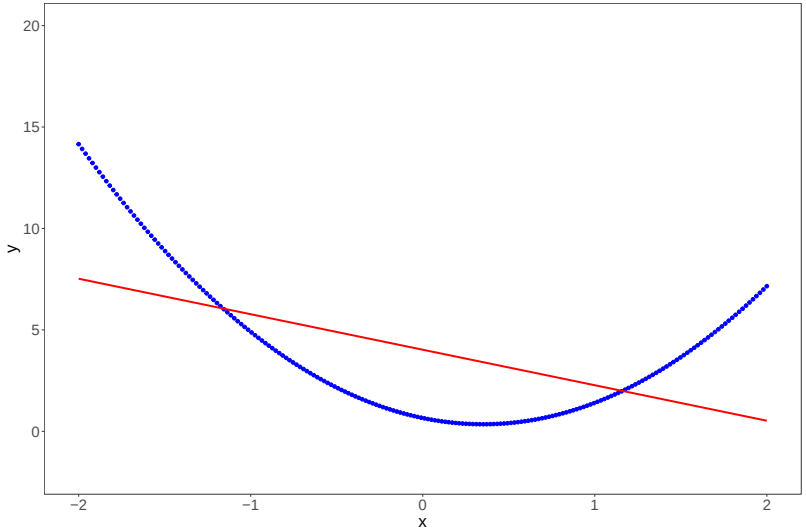
Overfitting and Underfitting. Bias-Variance Tradeoff

More data?



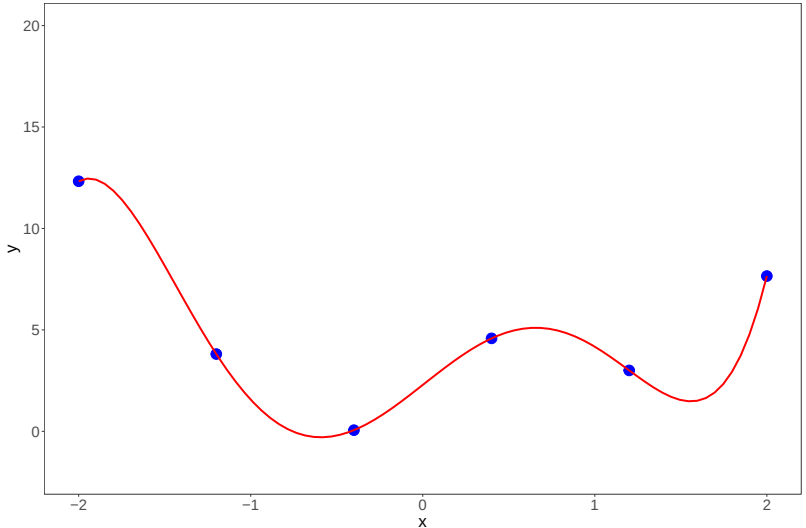
Overfitting and Underfitting. Bias-Variance Tradeoff

Noiseless data?



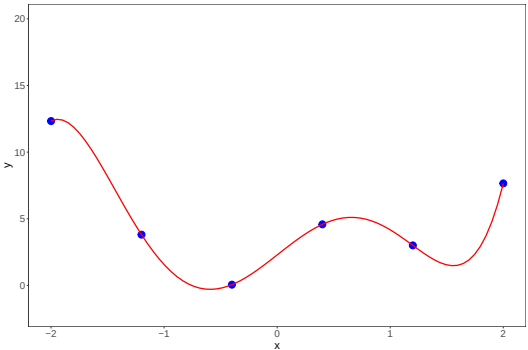
Overfitting and Underfitting. Bias-Variance Tradeoff

More Complex Model

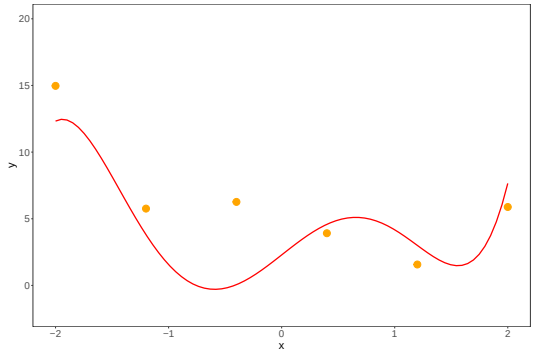


Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Performance



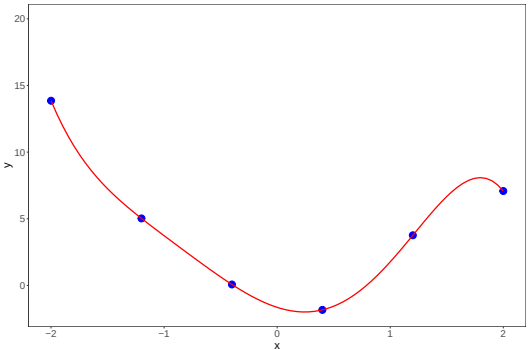
(a) Training Data



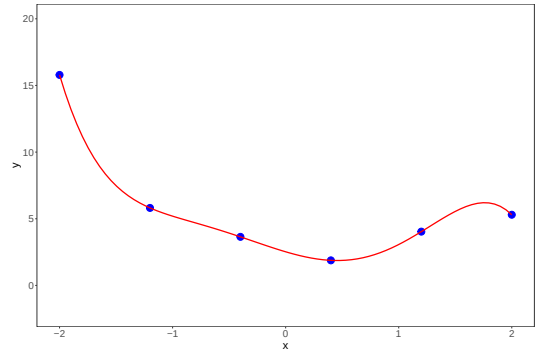
(b) Testing Data

Overfitting and Underfitting. Bias-Variance Tradeoff

Variance



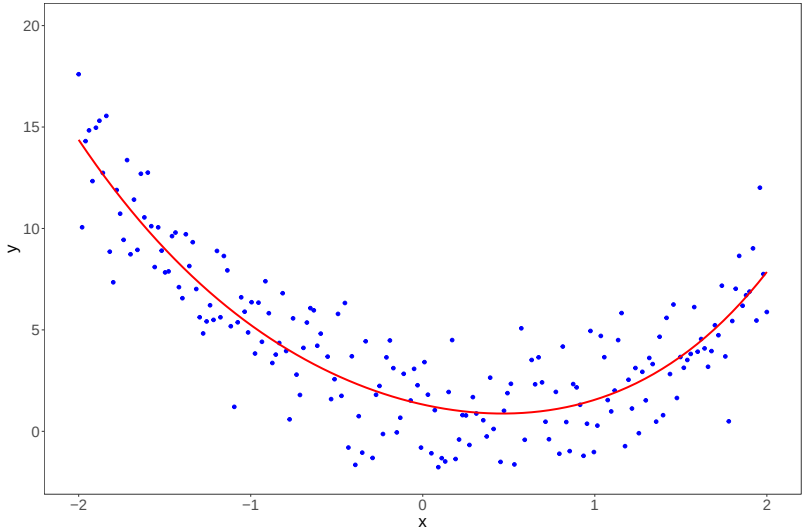
(a) Training Data 2



(b) Training Data 3

Overfitting and Underfitting. Bias-Variance Tradeoff

More Data

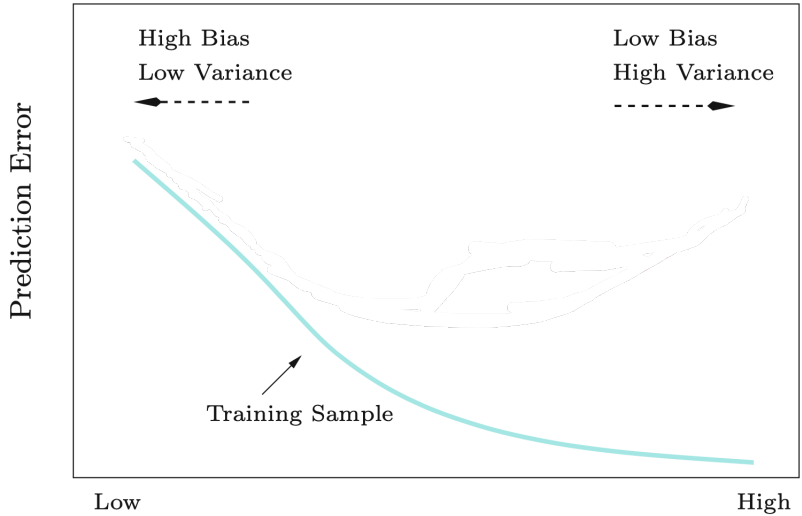


Key Insights on Bias-Variance Tradeoff

- The bias term reflects the error introduced by the model's inability to approximate the true function f^*.
- The variance term reflects the sensitivity of the model to the specific training set.
- As dataset size increases, variance generally decreases.

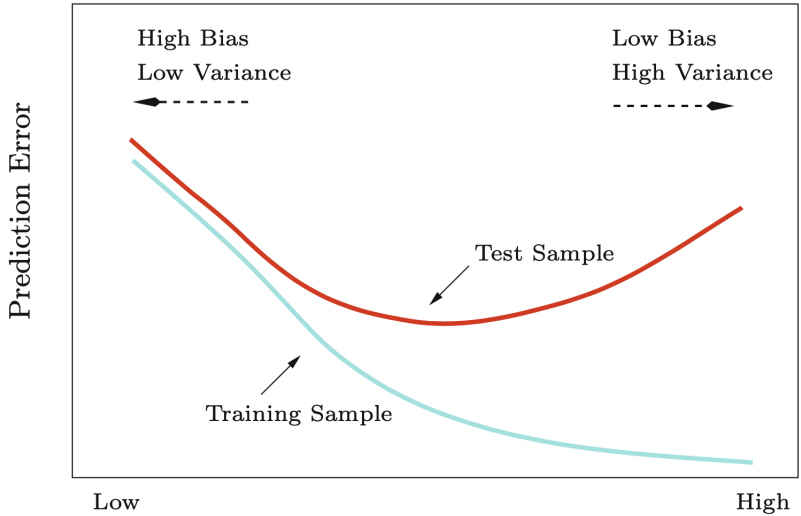
Overfitting and Underfitting. Bias-Variance Tradeoff

In-Sample Prediction and Overfit



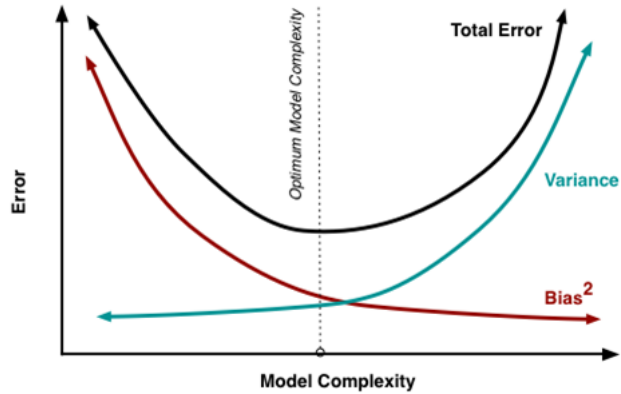
Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Prediction and Overfit



Mathematical Decomposition for Regression

Bias-Variance Tradeoff



Source: https://tinyurl.com/y4lvjxpc

- ML best kept secret: By tolerating some bias we can have significant gains in variance

Key Insights on Bias-Variance Tradeoff

- The bias term reflects the error introduced by the model's inability to approximate the true function f^*.
- The variance term reflects the sensitivity of the model to the specific training set.
- As dataset size increases, variance generally decreases.
- The noise term σ^2 is unavoidable and cannot be predicted.
- The decomposition for other problems is less clear than for regression, but still present.

Train and Test Sets. In-Sample and Out-of-Sample Prediction.

- Como seleccionamos la parametrización que minimize el error de predicción fuera de muestra?
- Problema: solo contamos con una muestra

Test Error

- Para seleccionar la mejor parametrización con respecto al Test Error (error de prueba), es necesario estimarlo.
- Hay dos enfoques comunes:
  - Podemos estimar indirectamente el error de la prueba haciendo un ajuste al error de entrenamiento para tener en cuenta el sesgo debido al sobreajuste $\Rightarrow$ Penzalización ex post: AIC, BIC, etc.

Test Error

AIC

- Akaike (1969) fue el primero en ofrecer un enfoque unificado al problema de la selección de modelos.
- Elegir el modelo j tal que se minimice:

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{26}$$

Test Error

SIC/BIC

- Schwarz (1978) mostró que el AIC es inconsistente, (cuando $n \to \infty$, tiende a elegir un modelo demasiado grande con probabilidad positiva)
- Schwarz (1978) propuso:

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - \frac{1}{2}p_j log(n) \tag{27}$$

Test Error

- Para seleccionar la mejor parametrización con respecto al Test Error (error de prueba), es necesario estimarlo.
- Hay dos enfoques comunes:
  - Podemos estimar indirectamente el error de la prueba haciendo un ajuste al error de entrenamiento para tener en cuenta el sesgo debido al sobreajuste $\Rightarrow$ Penzalización ex post: AIC, BIC, etc.
  - Levantarnos de nuestros bootstraps (resampling methods) y estimar directamente el Test Error (error de prueba)

Test Error

Cross-Validation

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

Agenda

1. Review
   - FWL
2. Uncertainty
   - Resampling methods
   - Parameter Assessment
     - Example: Elasticity of Demand for Gasoline
   - Model Assessment
     - Generalization. Out-of-sample Performance
     - Out-of-Sample Error Estimation
3. Recap

Recap

- Review + FWL
- Resampling Methods:
  - Parameter Assessment: estimate standard errors
  - Model Assessment: finding the best model
    - Bias-Variance Tradeoff (Dilema Sesgo/Varianza)
    - Sobreajuste y Selección de modelos: AIC y BIC, Enfoque de Validación, LOOCV, K-fold Cross-Validation (Validación Cruzada)