Resampling Methods for Uncertainty. Out of Sample Performance.

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Review
2. Uncertainty: Motivation
   - What are resampling methods?
3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline
4. Train and Test Sets. In-Sample and Out-of-Sample Prediction.
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
5. Review

Agenda

Predicting Well

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

- Interest on predicting y
- Under quadratic loss $\Rightarrow E[y|X = x]$

Linear Regression

$$
\begin{aligned}
y &= f(X) + u && (2)\\
&= \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u && (3)\\
&= X\beta + u && (4)
\end{aligned}
$$

- If $f(X) = X\beta$, obtaining $f(.)$ boils down to obtaining β

Linear Regression

- OLS says we should choose the estimators $\hat{\beta}$ such that we minimize the Sum of Square Residual (SSR)

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

$$= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ji} \right)^2 \tag{6}$$

$$= (y - X\hat{\beta})'(y - X\hat{\beta}) \tag{7}$$

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

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
    - Not feasible in certain instances.
  - Computing certain diagnostic statistics: Leverage, R^2, LOOCV.
  - Way to add more data without having to compute everything again

Applications: Fixed Effects

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

$$\ln w_{ijft} = x_{it}\beta + \lambda_i + \theta_j + \gamma_f + u_{ijft} \tag{8}$$

$$Y = X\beta + D_1\lambda + D_2\theta + D_3\gamma + u \tag{9}$$

  - 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: Outliers and High Leverage Data

- Note the following

$$\hat{\beta} = (X'X)^{-1}X'y \tag{10}$$

- 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

$$\hat{\beta}_j = c_j y \tag{11}$$

- App

Applications: Outliers and High Leverage Data

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

$$y = X\beta + \alpha e_j + u \tag{12}$$

using FWL we can do

$$M_{e_j} y = M_{e_j} X\beta + r \tag{13}$$

- 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: Outliers and High Leverage Data

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

$$y = X\beta + \alpha e_j + u \tag{14}$$

$$y = Z\theta + u \tag{15}$$

then the fitted values and residuals

$$y = P_Z y + M_z y \tag{16}$$

$$= X\hat{\beta}^{(j)} + \hat{\alpha} e_j + M_Z y \tag{17}$$

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

$$P_X y = X\hat{\beta}^{(j)} + \hat{\alpha} P_X e_j \tag{18}$$

$$X\hat{\beta} = X\hat{\beta}^{(j)} + \hat{\alpha} P_X e_j \tag{19}$$

$$X(\hat{\beta} - \beta^{(j)}) = \hat{\alpha} P_X e_j \tag{20}$$

Applications: Outliers and High Leverage Data

How to calculate α? FWL once again

$$M_X y = \hat{\alpha} M_X e_j + res \tag{21}$$

$$\hat{\alpha} = (e_j' M_X e_j)^{-1} e_j' M_X y \tag{22}$$

- $e_j' M_X y$ 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

$$\hat{\alpha} = \frac{\hat{u}_j}{1 - h_j} \tag{23}$$

where h_j is the j diagonal element of P_X

Applications: Outliers and High Leverage Data

Finally we get

$$(\hat{\beta}^{(j)} - \hat{\beta}) = -\frac{1}{1-h_j}(X'X)^{-1}X_j'\hat{u}_j \tag{24}$$

Agenda

1. Review
2. Uncertainty: Motivation
   - What are resampling methods?
3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline
4. Train and Test Sets. In-Sample and Out-of-Sample Prediction.
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
5. Review

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 and hence “overfit.”
- The ability to deal with this mess and noise is the most important skill you need.

Agenda

1. Review
2. Uncertainty: Motivation
   - What are resampling methods?
3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline
4. Train and Test Sets. In-Sample and Out-of-Sample Prediction.
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
5. Review

What are resampling methods?

- Tools that involves repeatedly drawing samples from a training set 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: estimate test error rates
  - Model Selection: select the appropriate level of model flexibility
  - They are computationally expensive! But these days we have powerful computers

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

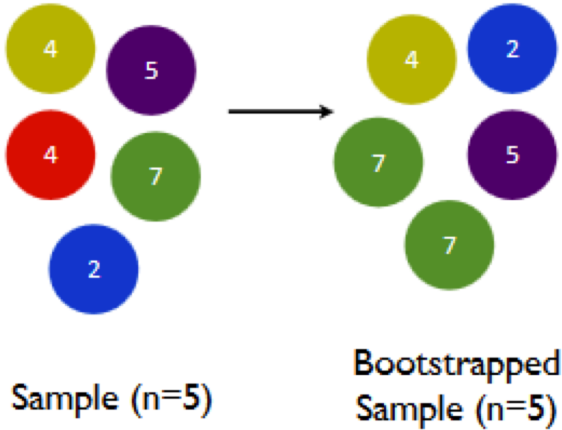
$$Var(\bar{Y}) \tag{25}$$

The Bootstrap

Introduction

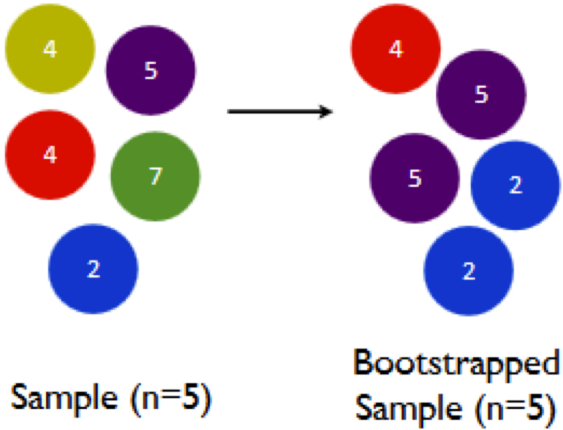
- Alternative way (no formula!)
  1. From the n original data points $y_1, y_2, \ldots, y_n$ take a sample *with replacement* of size n
  2. Calculate the sample average of this *"pseudo-sample"*
  3. Repeat this B times.
  4. Compute the variance of the B means

Sampling with replacement



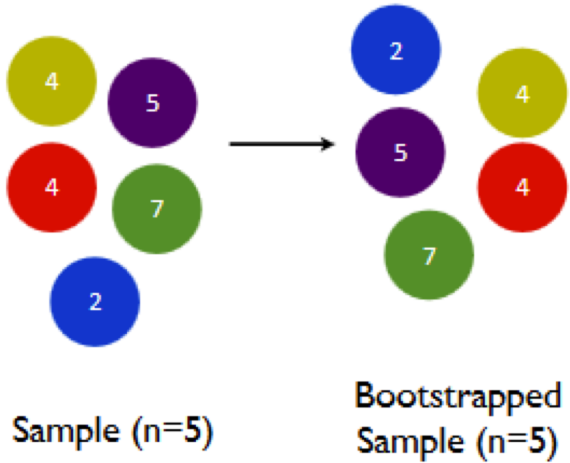
Sampling with replacement

Resampling creates synthetic variability



Sampling with replacement

Resampling creates synthetic variability



Uncertainty in Linear Regression

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

Uncertainty and Resampling

- Sometimes the analytical expression of the variance can be quite complicated.
- In these cases bootstrap can be useful
- The bootstrap provides a way to perform statistical inference by resampling from the sample.
- 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

Introduction

- 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, for example the variance:

$$\hat{V}(\hat{\theta})_B = \frac{1}{(B-1)} \sum_{j=1}^{B} (\hat{\theta}_j - \bar{\hat{\theta}})^2 \tag{26}$$

Example: Elasticity of Demand for Gasoline

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

Agenda

1. Review
2. Uncertainty: Motivation
   - What are resampling methods?
3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline
4. Train and Test Sets. In-Sample and Out-of-Sample Prediction.
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
5. Review

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

- El objetivo es predecir y dadas otras variables X. Ej: salario dadas las características del individuo
- Asumimos que el link entre y and X esta dado por el modelo:

$$y = f(X) + u \tag{27}$$

- donde $f(X)$ por ejemplo es $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
- u una variable aleatoria no observable $E(u) = 0$ and $V(u) = \sigma^2$

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

- Dos conceptos importantes
  - *Training error*: es el error de predicción en la muestra que fue utilizada para ajustar el modelo

$$Err_{\mathcal{T}rain} = MSE[(y, \hat{y})|\mathcal{T}rain] \tag{28}$$

  - *Test Error*: es el error de predicción fuera de muestra

$$Err_{\mathcal{T}est} = MSE[(y, \hat{y})|\mathcal{T}est] \tag{29}$$

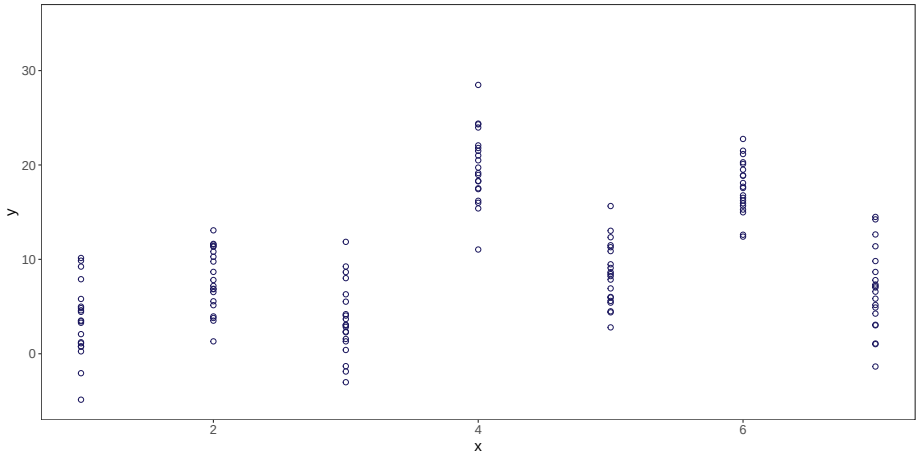
Train and Test Sets. In-Sample and Out-of-Sample Prediction.

- Como seleccionamos la especificación que minimize el error de predicción?

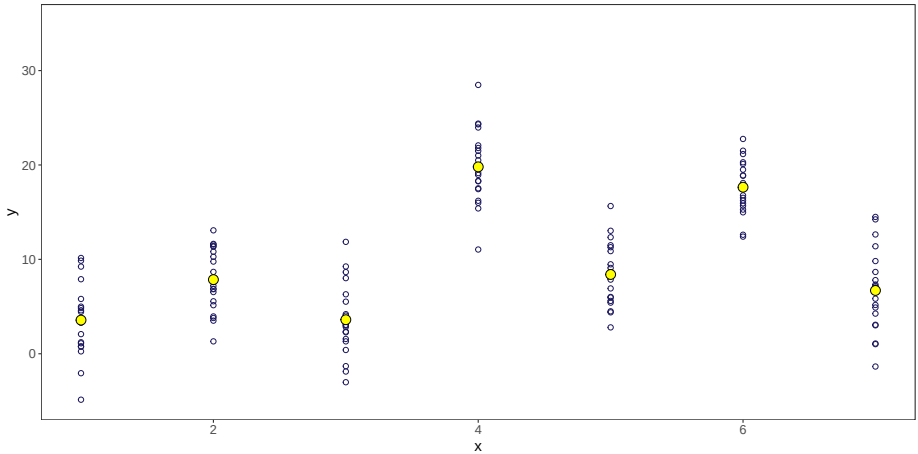
Train and Test Sets. In-Sample and Out-of-Sample Prediction.

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

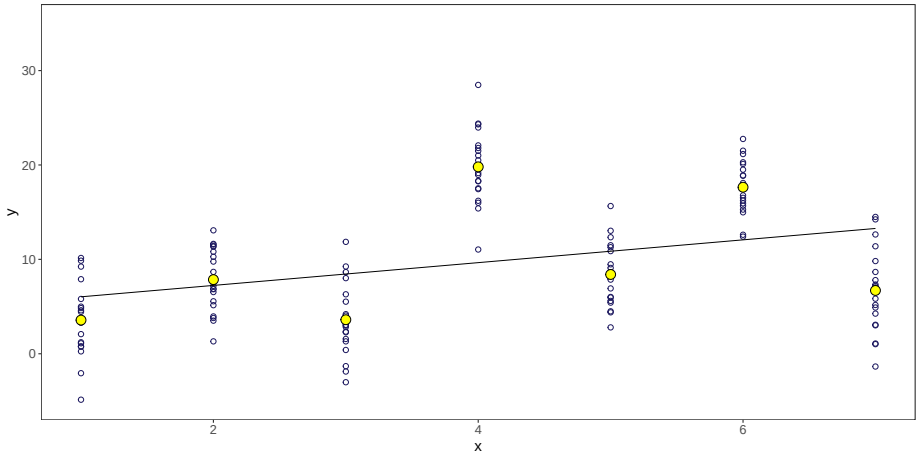
In-Sample Prediction and Overfit



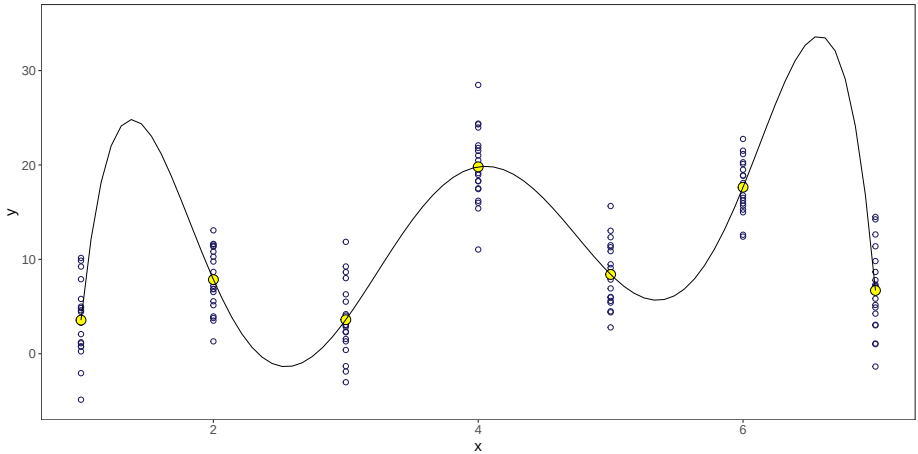
In-Sample Prediction and Overfit



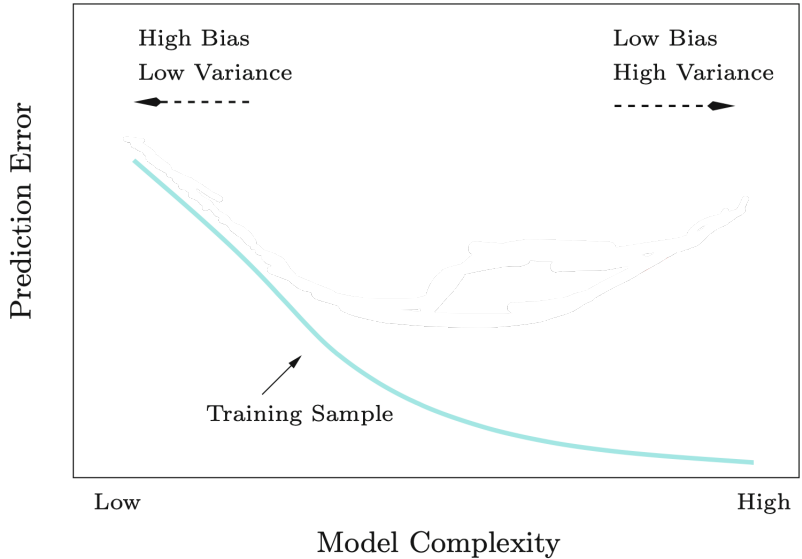
In-Sample Prediction and Overfit



In-Sample Prediction and Overfit



In-Sample Prediction and Overfit



In-Sample Prediction and Overfit

- Notemos que el MSE no es otra cosa que la suma de los residuales al cuadrado

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(X))^2 \tag{30}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{31}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(e)^2 \tag{32}$$

$$= SSR \tag{33}$$

- Esta medida nos da una idea de *lack of fit* que tan mal ajusta el modelo a los datos

In-Sample Prediction and Overfit

- Un problema del SSR es que nos da una medida absoluta de ajuste de los datos, y por lo tanto no esta claro que constituye un buen SRR.
- Una alternativa muy usada en economía es el R^2
- Este es una proporción (la proporción de varianza explicada),
  - toma valores entre 0 y 1,
  - es independiente de la escala (o unidades) de y

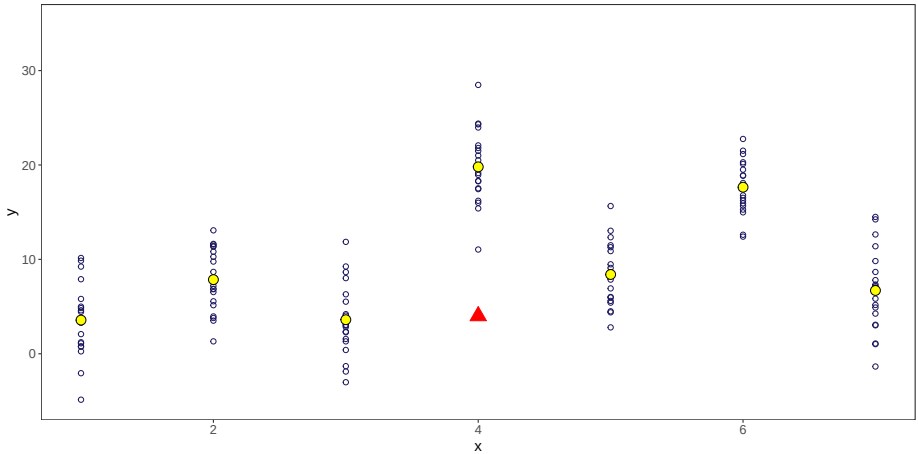
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{34}$$

$$= 1 - \frac{SRR}{TSS} \tag{35}$$

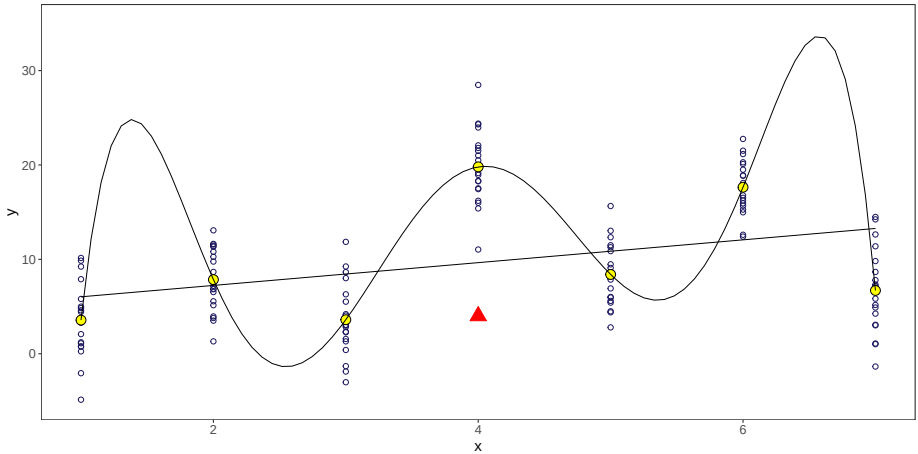
Out-of-Sample Prediction and Overfit

- ML nos interesa la predicción fuera de muestra

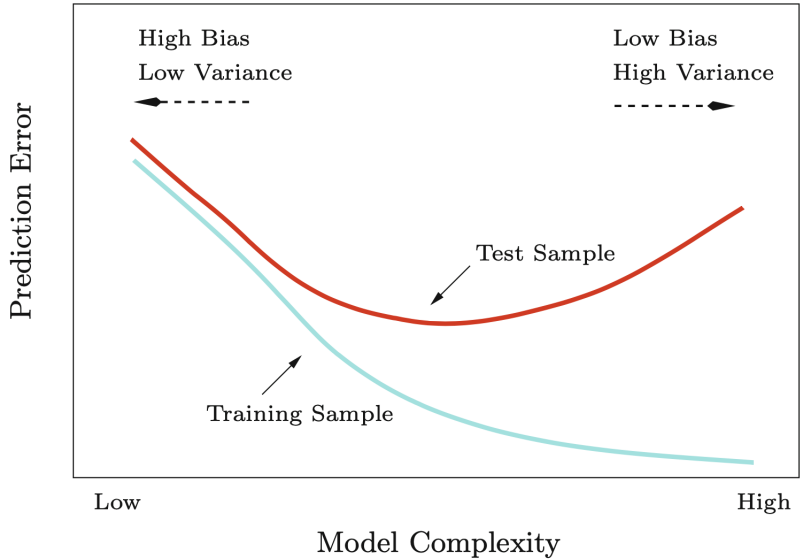
Out-of-Sample Prediction and Overfit



Out-of-Sample Prediction and Overfit



Out-of-Sample Prediction and Overfit



Out-of-Sample Prediction and Overfit

- ML nos interesa la predicción fuera de muestra
- Overfit: modelos complejos predicen muy bien dentro de muestra, pero tienden a hacer un trabajo fuera de muestra
- Hay que elegir el nivel adecuado de complejidad
- Como medimos el error de predicción fuera de muestra?
- R^2 no funciona: se concentra en la muestra y es no decreciente en complejidad

Test Error

- Para seleccionar el mejor modelo 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, R2 ajustado

Agenda

1. Review
2. Uncertainty: Motivation
   - What are resampling methods?
3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline
4. Train and Test Sets. In-Sample and Out-of-Sample Prediction.
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
5. Review

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:

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

Agenda

1. Review
2. Uncertainty: Motivation
   - What are resampling methods?
3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline
4. Train and Test Sets. In-Sample and Out-of-Sample Prediction.
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
5. Review

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:

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - \frac{1}{2}p_j log(n) \tag{37}$$

Test Error

AIC vs BIC

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{38}$$

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j\frac{1}{2}log(n) \tag{39}$$

- SIC tiende a elegir modelos más pequeños.
- En efecto, al dejar que la penalización tienda al infinito lentamente con n, eliminamos la tendencia de AIC a elegir un modelo demasiado grande.

Test Error

- Para seleccionar el mejor modelo 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, R2 ajustado
  - Levantarnos de nuestros bootstraps (resampling methods) y estimar directamente el Test Error (error de prueba)

Agenda

1. Review
2. Uncertainty: Motivation
   - What are resampling methods?
3. The Bootstrap
   - Example: Elasticity of Demand for Gasoline
4. Train and Test Sets. In-Sample and Out-of-Sample Prediction.
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
5. Review

Test Error

Cross-Validation

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

Agenda

Review

Hoy

- Dilema Sesgo/Varianza
- Sobreajuste y Selección de modelos
  - AIC y BIC
  - Enfoque de Validación
  - LOOCV
  - K-fold Cross-Validation (Validación Cruzada)