# Linear Regression and Resampling Methods for Uncertainty

## Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

# Agenda

# Agenda

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

# Predicting Well

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

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

# Linear Regression

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

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

# 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 $\beta$
  - QR (Gram-Schmidt process, similar to FWL)
  - Gradient Descent

# Agenda

# Uncertainty in Linear Regression

- To get a measure of the uncertainty, precision or variability of our estimates we need a measure
- We can estimate the Variance of our estimators
- Linear regression

$$Var(\hat{\beta}) = Var((X'X)^{-1}X'y) \tag{8}$$

# Uncertainty and Resampling

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

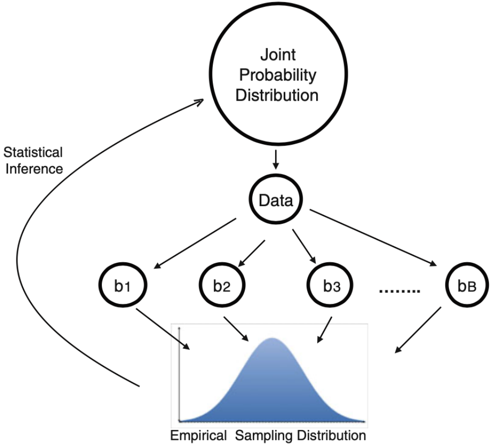
# Agenda

# The Bootstrap

## Introduction

- The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- 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



# The Bootstrap
## Introduction

- There are 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:
  - $Y_i \; i = 1, \ldots, n$
  - $\theta$ is the magnitude of interest
- To calculate it's variance
  1. Sample of size $n$ with replacement (*bootstrap sample*)
  2. Compute $\hat{\theta}_j \; j = 1, \ldots, B$
  3. Repeat $B$ times
  4. Calculate

$$\hat{V}(\hat{\theta})_B = \frac{1}{B}\sum_{j=1}^{B}(\hat{\theta}_j - \bar{\hat{\theta}})^2 \tag{9}$$

# Example: Elasticity of Demand for Gasoline


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

# Agenda

# 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:

$$y = f(X) + u \tag{10}$$

- 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

$$Err_{\mathcal{T}rain} = MSE[(y, \hat{y})|\mathcal{T}rain] \tag{11}$$

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

$$Err_{\mathcal{T}est} = MSE[(y, \hat{y})|\mathcal{T}est] \tag{12}$$

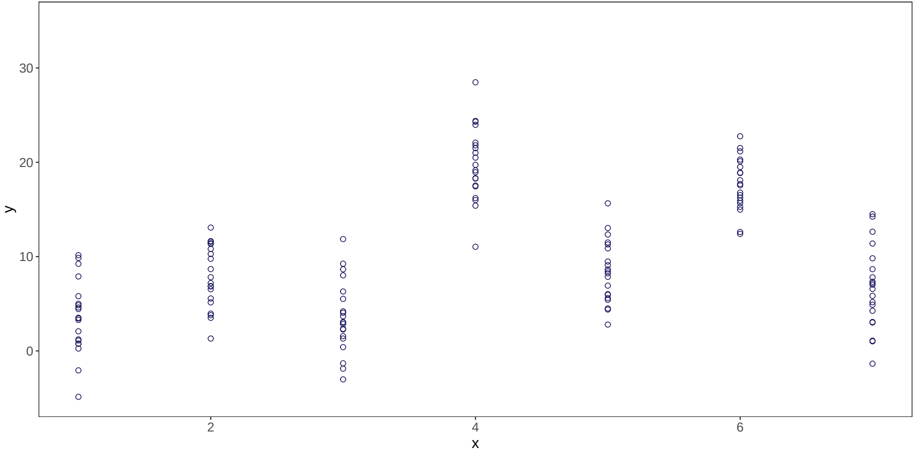
# Train and Test Sets. In-Sample and Out-of-Sample Prediction.

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

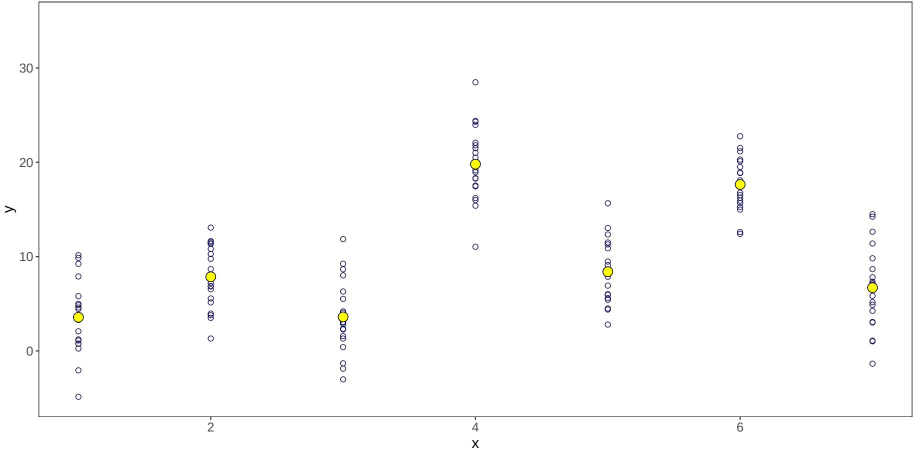
# Train and Test Sets. In-Sample and Out-of-Sample Prediction.

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

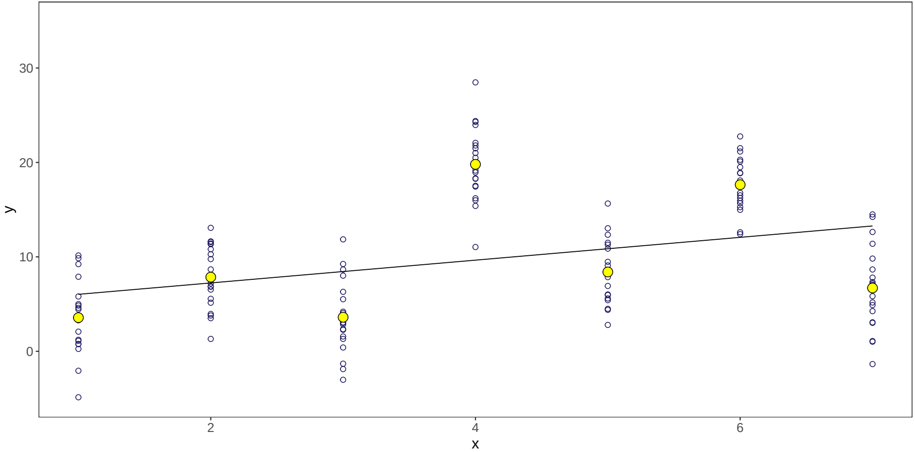
# In-Sample Prediction and Overfit



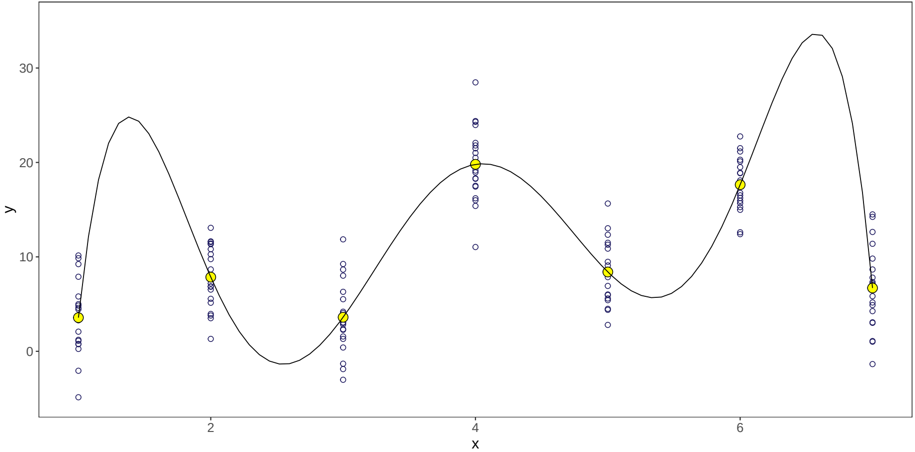
# In-Sample Prediction and Overfit



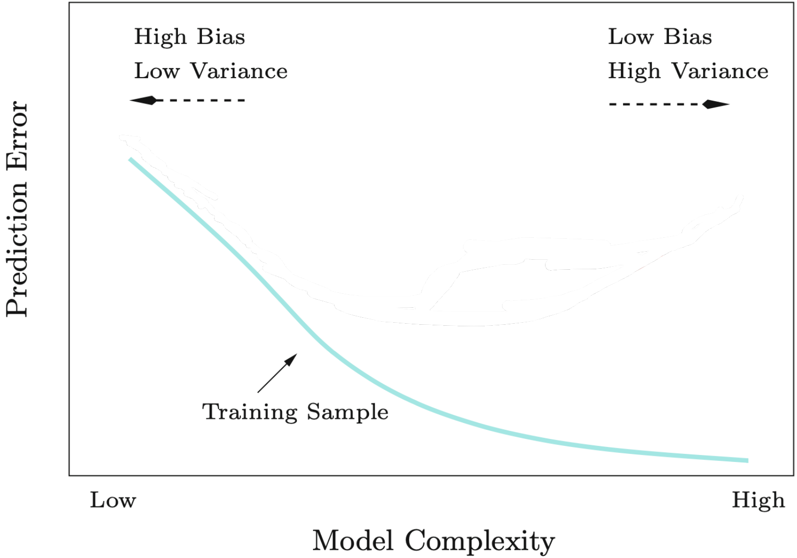
# In-Sample Prediction and Overfit



# 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

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(X))^2 \tag{13}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{14}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(e)^2 \tag{15}$$

$$= SSR \tag{16}$$

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

# In-Sample Prediction and Overfit

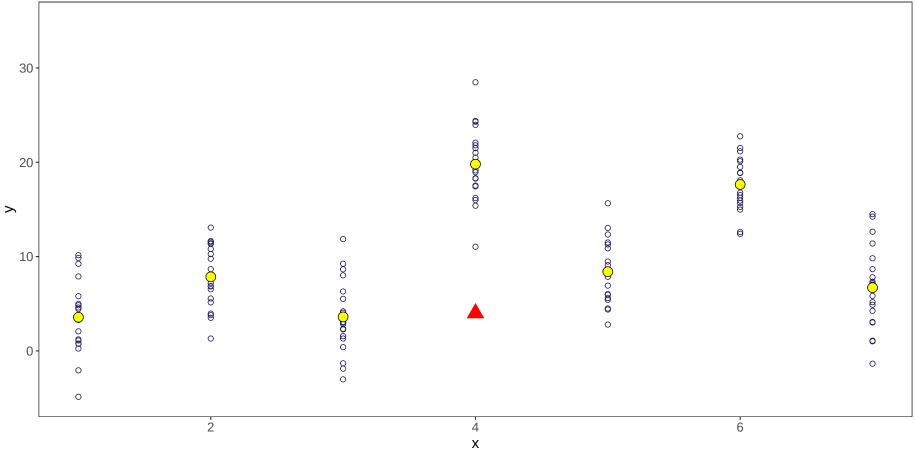
- Un problema del SSR es que nos da una medida absoluta de ajuste de los datos, y por lo tanto no esta claro que constituye un buen SRR.
- Una alternativa muy usada en economía es el $R^2$
- Este es una proporción (la proporción de varianza explicada),
  - toma valores entre 0 y 1,
  - es independiente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

$$
\begin{align}
R^2 &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{17}\\
&= 1 - \frac{SRR}{TSS} \tag{18}
\end{align}
$$

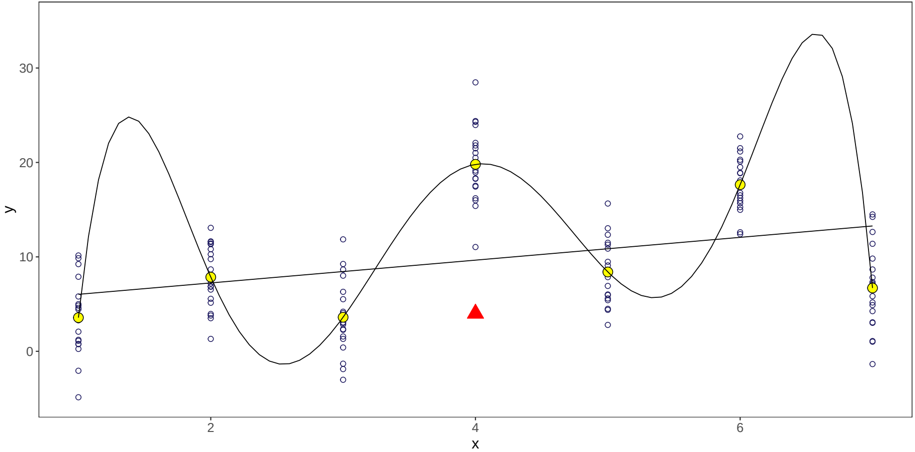
# Out-of-Sample Prediction and Overfit

- ML nos interesa la predicción fuera de muestra

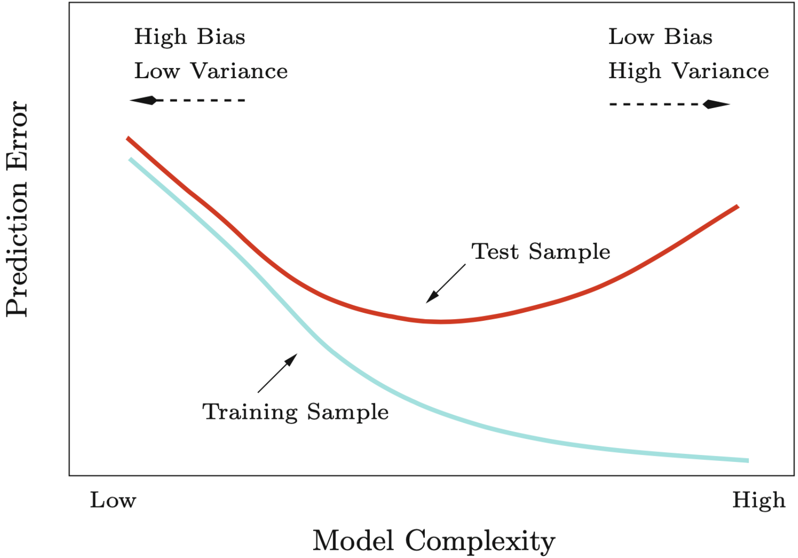
# Out-of-Sample Prediction and Overfit



# 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

# 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:

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{19}$$

# Agenda

# 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:

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - \frac{1}{2}p_j log(n) \tag{20}$$

# Test Error

## AIC vs BIC

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{21}$$

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j\frac{1}{2}log(n) \tag{22}$$

- 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

# Test Error

## Cross-Validation

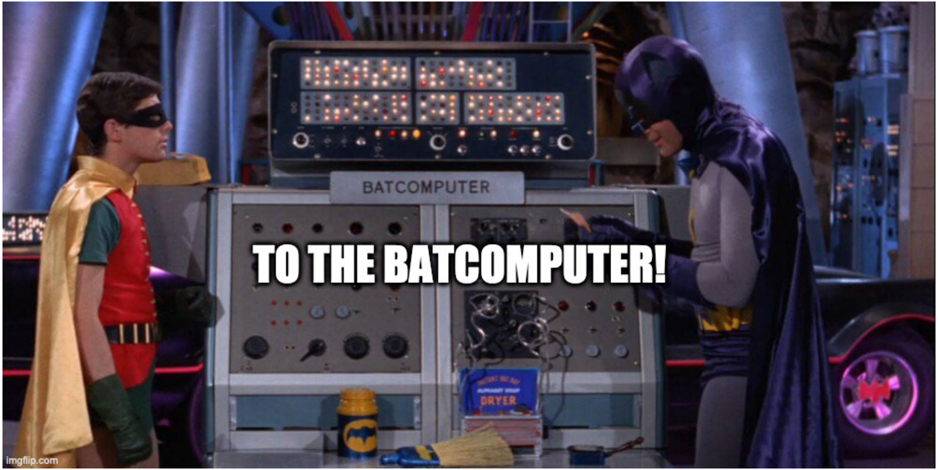



photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

# Agenda

# Review

Hoy

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