

Lecture 2:  
The classic and the predictive paradigms  
Decision Theory  
Big Data and Machine Learning for Applied Economics  
Econ 4676

Ignacio Sarmiento-Barbieri

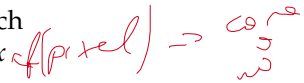
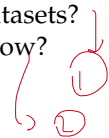
Universidad de los Andes

August 12, 2021

# Agenda

- 1 Review
- 2 Shifting Paradigms
- 3 How to Evaluate Estimators?
- 4 Statistical Decision Theory
- 5 Linear Regression
- 6 Recap

## Motivation

- ▶ We discussed the examples of Google Flu and Facebook face detection
  - ▶ Take away, the success was driven by an empiric approach
  - ▶ Given data estimate a function  $f(x)$  that predicts  $y$  from  $x$  
- ▶ This is basically what we do as economists everyday so:
  - ▶ Are these algorithms merely applying standard techniques to novel and large datasets?
  - ▶ If there are fundamentally new empirical tools, how do they fit with what we know?
  - ▶ As empirical economists, how can we use them? 

# Big vs Small, Classic vs Predictive

## ► Classical Stats (small data?)

- Get the most of few data (Gosset)
- Lots of structure, e.g.  $X_1, X_2, \dots, X_n \sim t_v$
- Carefully curated → approximates random sampling (expensive, slow) but very good and reliable

$$x_1, \dots, x_n \sim f(\theta)$$

42 V's of big data

## ► Big Data (the 4 V's)

- Data Volume
- Data Variety
- Data Velocity
- Data Value

# The Classic Paradigm

*modelo*

$$Y = \underbrace{f(X)}_{\text{obs}} + \underbrace{u}_{\mu^0} \quad \text{ob}$$

$$y = \frac{\alpha + \beta X}{f(X)} + u \quad (1)$$

*$\beta \neq 0$*

- ▶ Interest lies on inference
- ▶ "Correct"  $f()$  to understand how  $Y$  is affected by  $X$
- ▶ Model: Theory, experiment
- ▶ Hypothesis testing (std. err., tests)

*$f(X) = X\beta$*

# The Predictive Paradigm

$$f(x) = \underline{x} \underline{\beta}$$

$$\boxed{Y} = \underline{f(X)} + u$$

(2)

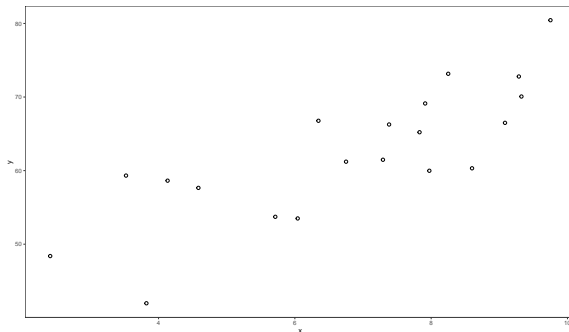
- ▶ Interest on predicting  $Y$
- ▶ "Correct"  $f()$  to be able to predict (no inference!)
- ▶ Model?

450 millones  
de modelos

→ predecir bien!  
→ fuera de muestra !!  
≠  $R^2$  dentro de muestra

→  $R^2$

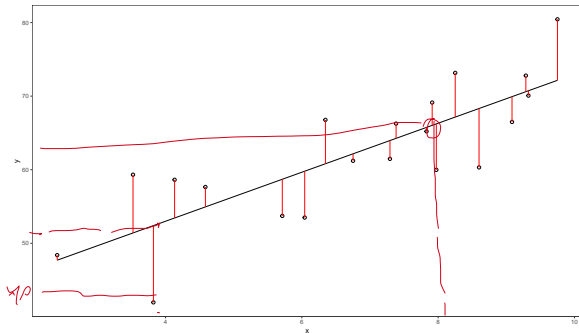
# How to choose $f(\cdot)$



Source: simulated data, see figures folder for scripts

# How to choose $f(\cdot)$

- Linear  $f(X) = X\beta$

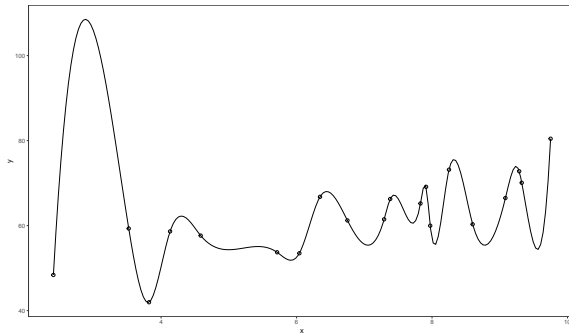


Source: simulated data, see figures folder for scripts



# How to choose $f(\cdot)$

- Spline  $f(X) = g(X)$ , where  $g$  is a spline



Source: simulated data, see figures folder for scripts

# Statistical Decision Theory: A bit of theory

- ▶ We need a bit of theory to give us a framework for choosing  $f$
- ▶ A decision theory approach involves an **action space**  $\mathcal{A}$
- ▶ The **action space**  $\mathcal{A}$  specify the possible "actions we might take"
- ▶ Some examples

Table 1: Action Spaces

Inference	Action Space
Estimation $\theta, g(\theta)$	$\mathcal{A} = \Theta$
Prediction	$\mathcal{A} = \text{space of } X_{n+1}$
Model Selection	$\mathcal{A} = \{\text{Model I, Model II, ...}\}$
Hyp. Testing	$\mathcal{A} = \{\text{Reject}   \text{Accept } H_0\}$

$$f(\theta)$$
$$p$$
$$N(\mu, \sigma)$$

# Statistical Decision Theory: A bit of theory

- ▶ After the data  $X = x$  is observed, where  $X \sim f(X|\theta)$ ,  $\theta \in \Theta$
- ▶ A decision is made
- ▶ The set of allowable decisions is the action space ( $\mathcal{A}$ )
- ▶ The loss function in an estimation problem reflects the fact that if an action  $a$  is close to  $\theta$ ,
  - ▶ then the decision  $a$  is reasonable and little loss is incurred.
  - ▶ if it is far then a large loss is incurred

$$a = \theta$$

$$L: \mathcal{A} \rightarrow [0, \infty]$$

(3)

# Statistical Decision Theory: A bit of theory

## Loss Function

- ▶ If  $\theta$  is real valued, two of the most common loss functions are

- ▶ Squared Error Loss:

$$L(a, \theta) = (a - \theta)^2 \quad (4)$$

- ▶ Absolute Error Loss:

$$L(a, \theta) = |a - \theta| \quad (5)$$

- ▶ These two are symmetric functions. However, there's no restriction. For example in hypothesis testing a "0-1" Loss is common.
- ▶ Loss is minimum if the action is correct

# Statistical Decision Theory: A bit of theory

## Risk Function

In a decision theoretic analysis, the quality of an estimator is quantified by its risk function, that is, for an estimator  $\delta(x)$  of  $\theta$ , the risk function is

$$R = \int L(X)$$

$$R(\theta, \delta) = E_{\theta}(L(\theta, \underbrace{\delta(X)}_{\theta})) \quad (6)$$

at a given  $\theta$ , the risk function is the average loss that will be incurred if the estimator  $\delta(X)$  is used

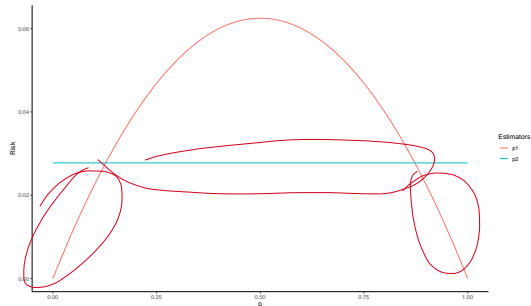
- ▶ Since  $\theta$  is unknown we would like to use an estimator that has a small value of  $R(\theta, \delta)$  for all values  $\theta$
- ▶ Loss is minimum if the action is correct
- ▶ If we need to compare two estimators ( $\delta_1$  and  $\delta_2$ ) then we will compare their risk functions
- ▶ If  $R(\delta_1, \theta) < R(\delta_2, \theta)$  for all  $\theta \in \Theta$ , then  $\delta_1$  is preferred because it performs better for all  $\theta$

# Statistical Decision Theory: A bit of theory

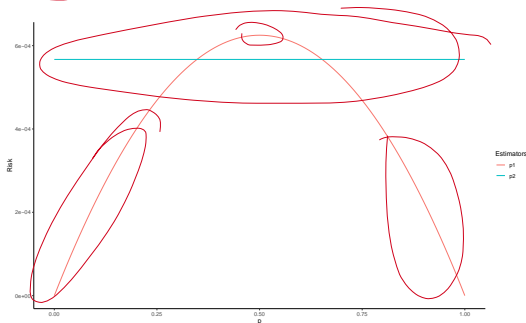
How to verify

## Example: Binomial Risk Function

- ▶ Let  $X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p)$
- ▶ Consider 2 estimators for  $p$ :  $\hat{p}^1 = \frac{1}{n} \sum X_i$  and  $\hat{p}^2 = \frac{\sum X_i + \sqrt{n/4}}{n + \sqrt{n}}$  *Bayes*
- ▶ Their risks are:  $R(\hat{p}^1, p) = \frac{p(1-p)}{n}$  and  $R(\hat{p}^2, p) = \frac{n}{4(n + \sqrt{n})^2}$



(a) n=4



(b) n=400

# Decision Theory for prediction

How to choose  $f$ ?

- ▶ In a prediction problem we want to predict  $\underline{Y}$  from  $\underline{f(X)}$  in such a way that the loss is minimum
- ▶ Assume also that  $X \in \mathbb{R}^p$  and  $Y \in \mathbb{R}$  with joint distribution  $Pr(X, Y)$

$$R(Y, f(X)) = E[(Y - f(X))^2] \quad (7)$$

$$= \int (y - f(x))^2 Pr(dx, dy) \quad (8)$$

conditioning on  $X$  we have that

$$R(Y, f(X)|X) = E_X E_{Y|X}[(Y - f(X))^2|X] \quad (9)$$

this risk is also known as the **mean squared (prediction) error**  $MSE(f)$

# Decision Theory for prediction

$$f(X=x) - m$$

It suffices to minimize the  $MSE(f)$  point wise so

$$f(x) = \operatorname{argmin}_m E_{Y|X}[(Y - \underline{m})^2 | X = x] \quad (10)$$

$Y$  a random variable and  $m$  a constant (predictor)

$$\min_m E(Y - m)^2 = \int (y - m)^2 f(y) dy \quad (11)$$

**Result:** The best prediction of  $Y$  at any point  $X = x$  is the conditional mean, when best is measured using a square error loss

$$E(Y)$$



# Decision Theory for prediction

**Proof**

FOC

$$\frac{\partial}{\partial m}$$

$$\int -2(y - m)f(y)dy = 0 \quad (12)$$

Dividing by -2 and reorganizing

$$\int y f(y) dy - \int m f(y) dy = 0$$

$$m \underbrace{\int f(y) dy}_{=1} = \underbrace{\int y f(y) dy}_{E(y)} \quad (13)$$

$$m = E(y)$$

# Decision Theory for prediction

$$m \int (y) dy = \int y f(y) dy \quad (14)$$

$$m = E(Y|X = x) \quad (15)$$

The best prediction of  $Y$  at any point  $X = x$  is the conditional expectation function (CEF), when best is measured using a square error loss

- ▶ What shape does the CEF take?
- ▶ Linear
  - ▶  $(y, X)$  are jointly normal
  - ▶ When models are saturated.

# Linear Regression

- Note the following from the *Regression-CEF Theorem*

The function  $X'\beta$  provides the minimum risk linear approximation to  $E(Y|X)$ , that is

$$\beta = \underset{b}{\operatorname{argmin}} E \{ (E(Y|X) - X'b)^2 \} \quad (16)$$

- Proof

$$(Y - X'b)^2 = (Y - E(Y|X)) + (E(Y|X) - X'b)^2 \quad (17)$$

$$= \underbrace{(Y - E(Y|X))^2}_{=0} + \underbrace{(E(Y|X) - X'b)^2}_{(Y - X'b)^2} + \underbrace{2(Y - E(Y|X))(E(Y|X) - X'b)}_{=0} \quad (18)$$

- The CEF approximation problem then has the same solution as the population least square problems

# Linear Regression

- ▶ Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable.
- ▶ The fact that Regression approximates the CEF is useful because it helps describe the essential features of statistical relationships, without necessarily trying to pin them down exactly.
- ▶ Linear regression is the “work horse” of econometrics and (supervised) machine learning.
- ▶ Very powerful in many contexts.
- ▶ Big ‘payday’ to study this model in detail.

# Linear Regression Model

$f(X) = X\beta$ , estimating  $f(\cdot)$  boils down to estimating  $\beta$

$$y = X\beta + u \quad (19)$$

where

- ▶  $y$  is a vector  $n \times 1$  with typical element  $y_i$
- ▶  $X$  is a matrix  $n \times k$ 
  - ▶ Note that we can represent it as a column vector  $X = \begin{bmatrix} X_1 & X_2 & \dots & X_k \end{bmatrix}$   
 $\begin{matrix} n \times k & n \times 1 & n \times 1 & n \times 1 \end{matrix}$
- ▶  $\beta$  is a vector  $k \times 1$  with typical element  $\beta_j$

Thus

$$\begin{aligned} y_i &= X_i' \beta + u_i \\ &= \sum_{j=1}^k \beta_j X_{ji} + u_i \end{aligned}$$

*Terminated doc!*

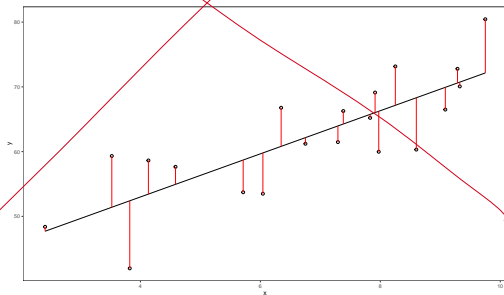
(20)

# Linear Regression Model

How do we estimate  $\beta$ ?

- ▶ Method of Moments (for HW)
- ▶ MLE (more on this later)
- ▶ OLS: minimize risk squared error loss  $\rightarrow$  minimizes SSR ( $e'e$ )
  - ▶ where  $e = Y - \hat{Y} = Y - X\hat{\beta}$
  - ▶ In the HW, you will show that  $\min$  SSR same as  $\max R^2$

OLS solution:  $\hat{\beta} = (X'X)^{-1}X'y$



# Gauss Markov Theorem

Gauss-Markov Theorem says that

$$\hat{\beta} = (X'X)^{-1}X'y \quad (21)$$

- ▶ The OLS estimator ( $\hat{\beta}$ ) is BLUE, the more efficient than any other linear unbiased estimator,
- ▶ Efficiency in the sense that  $Var(\tilde{\beta}) - Var(\hat{\beta})$  is positive semidefinite matrix.

Proof: HW. Tip: a matrix  $M_{p \times p}$  is positive semi-definite iff  $c'Mc \geq 0 \forall c \in \mathbb{R}^p$

# Gauss Markov Theorem

- ▶ Gauss Markov Theorem that says OLS is BLUE is perhaps one of the most famous results in statistics.
  - ▶  $E(\hat{\beta}) = \beta$
  - ▶  $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- ▶ However, it is essential to note the limitations of the theorem.
  - ▶ Correctly specified with exogenous Xs,
  - ▶ The term error is homoscedastic
  - ▶ No serial correlation.
  - ▶ Nothing about the OLS estimator being the more efficient than any other estimator one can imagine.



# Prediction vs Estimation

- ▶ **Predicting well in this context  $\rightarrow$  estimating well**

- ▶ Note that the prediction of  $y$  will be given by  $\hat{y} = X\hat{\beta}$

- ▶ Under Gauss-Markov framework

- ▶  $E(\hat{y}) = X\beta$

- ▶  $V(\hat{y}) = \sigma^2 X'(X'X)^{-1}X$

- ▶ Then if  $\hat{\beta}$  is unbiased and of minimum variance,

- ▶ then  $\hat{y}$  is an unbiased predictor and minimum variance, from the class of unbiased linear estimators/predictors

- ▶ Proof: for HW similar to  $\hat{\beta}$  proof

# Recap

- ▶ We start shifting paradigms
- ▶ Tools are not that different (so far)
- ▶ Decision Theory: Risk with square error loss  $\rightarrow$  MSE
- ▶ OLS is a "work horse" approximates the  $E[Y|X]$  quite well
- ▶ Next Class:
  - ▶ Next Class: OLS, Geometry, Properties

## Further Readings

- ▶ Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics. Princeton university press.
- ▶ Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury.
- ▶ Tom Shaffer The 42 V's of Big Data and Data Science.  
<https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html>