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

Review

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)
  - Lost of structure, e.g. $X_1, X_2, ..., X_n \sim t_v$
  - Carefully curated $\rightarrow$ approximates random sampling (expensive, slow) but very good and reliable

- Big Data (the 4 V's)
  - Data **V**olume
  - Data **V**ariety
  - Data **V**elocity
  - Data **V**alue

The Classic Paradigm

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

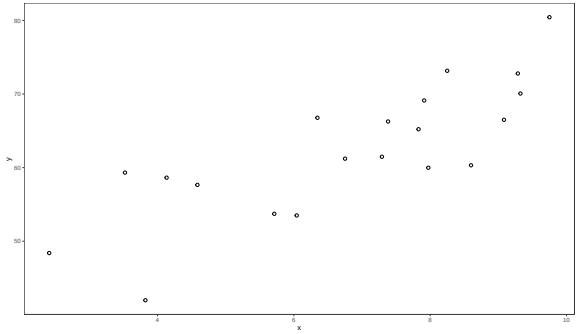
- Interest lies on inference
- "Correct" $f()$ to understand how Y is affected by X
- Model: Theory, experiment
- Hypothesis testing (std. err., tests)

The Predictive Paradigm

$$Y = f(X) + u \tag{2}$$

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

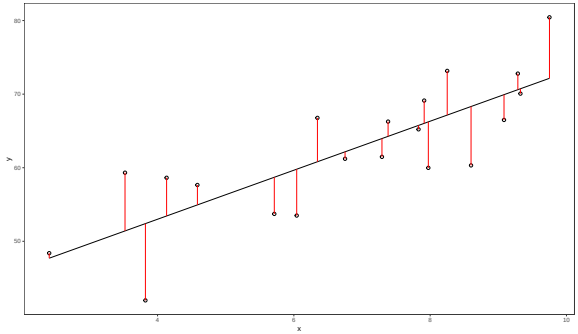
How to choose $f(.)$



Source: simulated data, see `figures` folder for scripts

How to choose $f(.)$

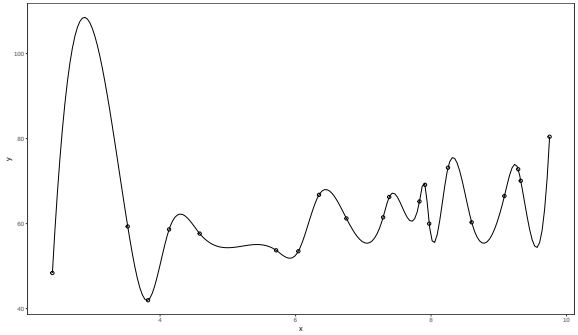
- Linear $f(X) = X\beta$



Source: simulated data, see `figures` folder for scripts

How to choose $f(.)$

- 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 θ, $g(\theta)$	$\mathcal{A} = \Theta$
Prediction	$\mathcal{A} = space\ of\ X_{n+1}$
Model Selection	$\mathcal{A} = \{Model\ I, Model\ II, ...\}$
Hyp. Testing	$\mathcal{A} = \{Reject \vert Accept\ H_0\}$

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 θ,
  - then the decision a is reasonable and little loss is incurred.
  - if it is far then a large loss is incurred

$$L : \mathcal{A} \to [0, \infty] \tag{3}$$

Statistical Decision Theory: A bit of theory

Loss Function

- If θ is real valued, two of the most common loss functions are
  - Squared Error Loss:

$$L(a, \theta) = (a - \theta)^2 \tag{4}$$

  - Absolute Error Loss:

$$L(a, \theta) = |a - \theta| \tag{5}$$

- 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 θ, the risk function is

$$R(\theta, \delta) = E_\theta(L(\theta, \delta(X)) \tag{6}$$

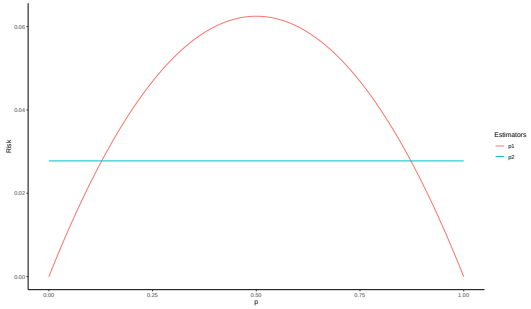
at a given θ, the risk function is the average loss that will be incurred if the estimator $\delta(X)$ is used

- since θ is unknown we would like to use an estimator that has a small value of $R(\theta, \delta)$ for all values θ
- Loss is minimum if the action is correct
- If we need to compare two estimators (δ_1 and δ_2) then we will compare their risk functions
- If $R(\delta_1, \theta) < R(\delta_2, \theta)$ for all $\theta \in \Theta$, then δ_1 is preferred because it performs better for all θ

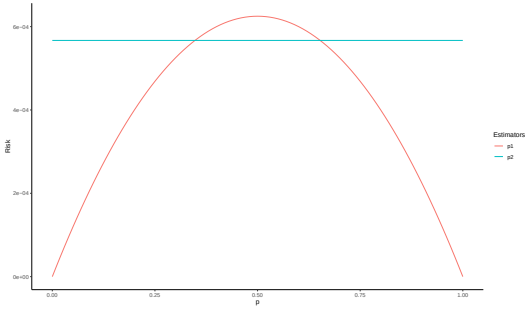
Statistical Decision Theory: A bit of theory

Example: Binomial Risk Function

- Let $X_1, X_2, ...X_n \sim Bernoulli(p)$
- Consider 2 estimators for p: $\hat{p}^1 = \frac{1}{n}\sum X_i$ and $\hat{p}^2 = \frac{\sum X_i + \sqrt{n/4}}{n+\sqrt{n}}$
- 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 Y from $f(X)$ 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 $Pr(X, Y)$

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

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

conditioning on X we have that

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

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

Decision Theory for prediction

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

$$f(x) = argmin_m E_{Y|X}[(Y-m)^2|X=x) \tag{10}$$

Y a random variable and m a constant (predictor)

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

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

Decision Theory for prediction

Proof

FOC

$$\int -2(y-m)f(y)dy = 0 \tag{12}$$

Dividing by -2 and reorganizing

$$m\int (y)dy = \int yf(y)dy \tag{13}$$

Decision Theory for prediction

$$m \int (y)dy = \int yf(y)dy \tag{14}$$

$$m = E(Y|X = x) \tag{15}$$

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

$$\beta = \underset{b}{argmin}\, E\left\{(E(Y|X) - X'b)^2\right\} \tag{16}$$

- Proof

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

$$= (Y - E(Y|X))^2 + (E(Y|X) - X'b)^2 + 2(Y - E(Y|X))(E(Y|X) - X'b) \tag{18}$$

- 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.
- On the other hand, if we prefer to think about approximating $E(Y|X)$, as opposed to predicting Y, the Regression-CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.
- The statement that regression approximates the CEF lines up with our view of empirical work as an effort to 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 $f(.)$ boils down to estimating β

$$y = X\beta + u \tag{19}$$

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 $\underset{n\times k}{X} = [\underset{n\times 1}{X_1}\ \underset{n\times 1}{X_2} \dots \underset{n\times 1}{X_k}]$
- β is a vector $k \times 1$ with typical element β_j

Thus

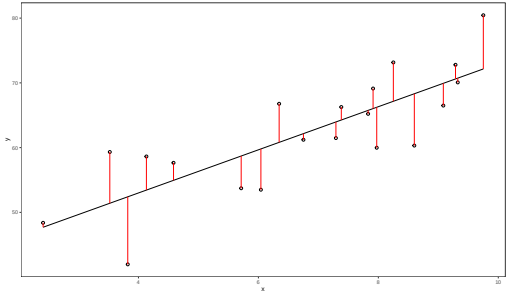
$$\begin{aligned} y_i &= X_i'\beta + u_i \\ &= \sum_{j=1}^{k} \beta_j X_{ji} + u_i \end{aligned} \tag{20}$$

Linear Regression Model

How do we estimate β?

- 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

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

- 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 $c'Mc \geq 0 \; \forall x \in \mathbb{R}^p$

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