6.s077 — INTRODUCTION TO DATA SCIENCE
EECS, MIT, Spring 2018

Lecture 4

Confidence Intervals
Predictions, and multidimensional data

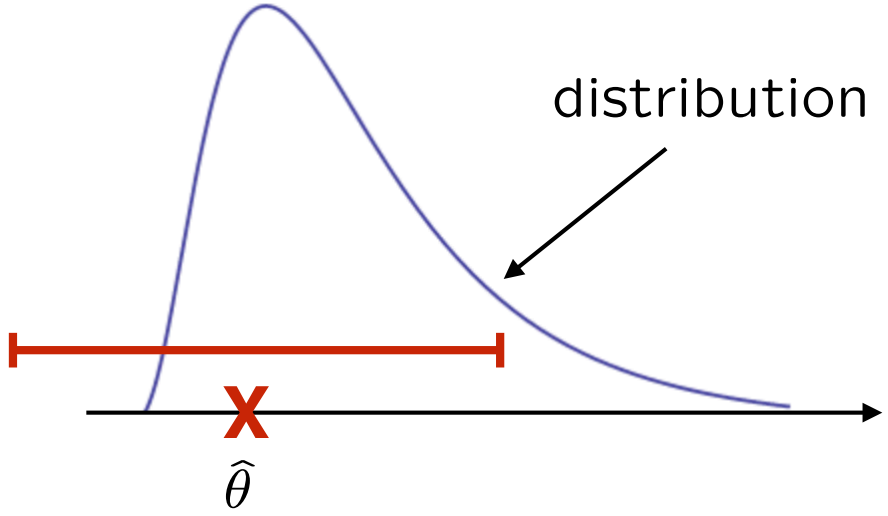
Today's agenda

- Confidence intervals
  - definition and interpretation
  - based on normal (and other) approximations
  - via bootstrap

Today's agenda

- Predicting an unknown random variable
  - expectation as a predictor
  - loss functions
  - empirical risk minimization (ERM)
- Multidimensional data
  - visualization
  - i.i.d. models
  - estimation (plugin, feature matching, ML)
- Predicting one variable from another
  - model-based, $\mathbb{E}[Y \mid X]$
  - based on data (X_i, Y_i)
- Structured (e.g., linear predictors)
  - Model-based; plugin; ERM
- Caution

Confidence intervals (CIs)



- The value of an estimator $\widehat{\Theta}$ may not be informative enough

constructed using the data

- An $1-\alpha$ **confidence interval** is an interval $[\widehat{\Theta}^-, \widehat{\Theta}^+]$, s.t. $\mathbb{P}^\theta(\widehat{\Theta}^- \le \theta \le \widehat{\Theta}^+) \ge 1-\alpha$, for all θ
  - often $\alpha = 0.05$, or 0.025, or 0.01

- Interpretation (for $\alpha = 0.05$). Suppose CI$= [-3.2, 4.7]$
  - "There is probability 95% that $\theta \in [-3.2, 4.7]$" **WRONG!!!**
  - if 100 statisticians produce 100 such CIs (for different data sets) we expect that 95 of them will capture θ
  - unless a certain extreme (5%) event has occurred, the CI captures θ

Formulas for CIs – "nice" but common cases

- Assume sampling distribution: $\widehat{\Theta} \approx N(\theta, \mathrm{se}^2) \approx N(\theta, \widehat{\mathrm{se}}^2)$

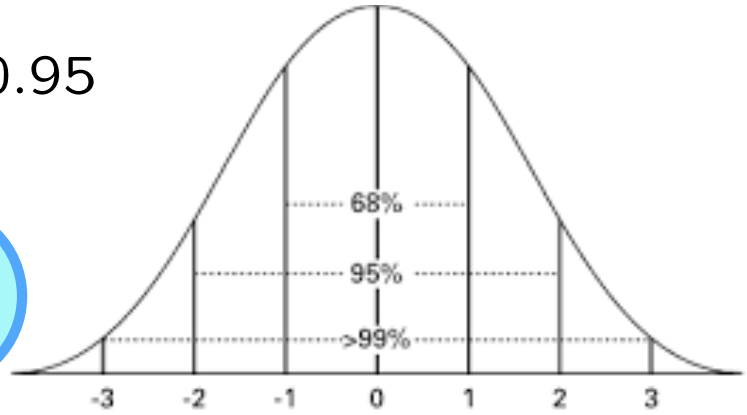
standard normal

exact, or CLT

bootstrap or other approximation (ML)

- Normal tables: $\mathbb{P}\left(\left| \dfrac{\widehat{\Theta} - \theta}{\widehat{\mathrm{se}}} \right| \leq 1.96 \right) \approx 0.95$

$$\mathbb{P}\left(\widehat{\Theta} - 1.96\,\widehat{\mathrm{se}} \leq \theta \leq \widehat{\Theta} + 1.96\,\widehat{\mathrm{se}} \right) = 0.95$$



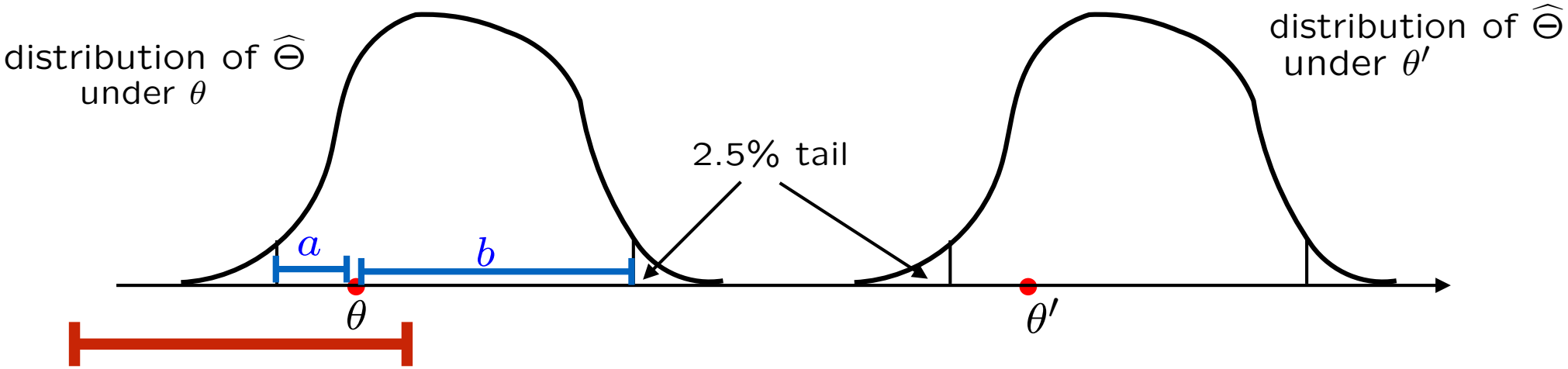
- Even if $\widehat{\Theta}$ is normal (no approximation)
  the approximation $\mathrm{se} \approx \widehat{\mathrm{se}}$
  may cause significant departure from normality ($\rightarrow$ t-distribution)

- **Caution:** We are dealing with the tails of the distributions;
  these can be very sensitive to approximations

Constructing CIs more generally (assume $\alpha = 0.05$)

- Want $\mathbb{P}^\theta(\widehat{\Theta}^- \le \theta \le \widehat{\Theta}^+) \ge 0.95$, for all θ (Hard!)
- Data confine θ and $\mathbb{P}$ to a "small" range

Assume: sampling distribution does not change much within that small range



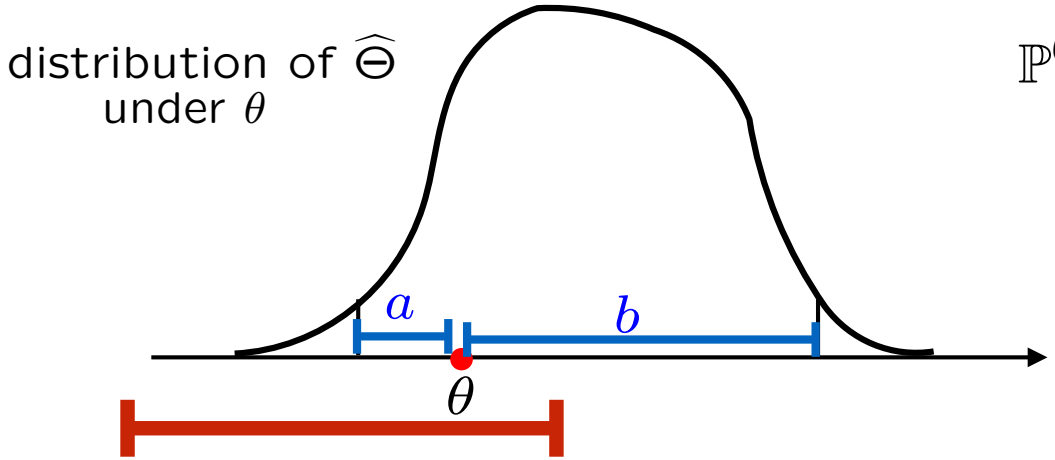
– unless a certain extreme (5%) event has occurred, the CI captures θ

$$\mathbb{P}^\theta(\theta - a \le \widehat{\Theta} \le \theta + b) = 0.95, \quad \text{for all } \theta$$

$$\mathbb{P}^\theta(\widehat{\Theta} - b \le \theta \le \widehat{\Theta} + a) = 0.95$$

- earlier normal example: $a = b = 1.96 \cdot \text{se}$

Constructing CIs more generally (assume $\alpha = 0.05$)



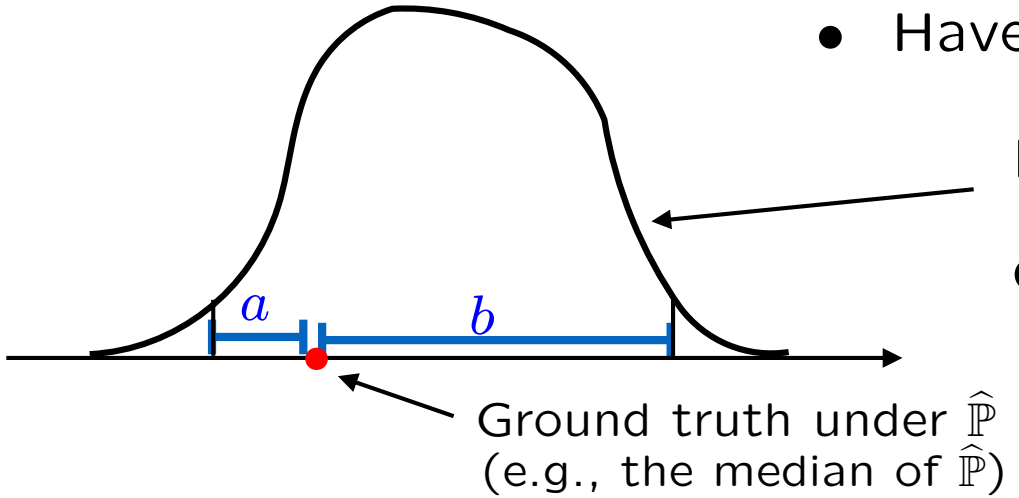
$$\mathbb{P}^{\theta}(\widehat{\Theta} - b \leq \theta \leq \widehat{\Theta} + a) = 0.95$$

Assume: sampling distribution does not change much within that small range

---

- Simulate under some "reasonable" θ: use $\widehat{\theta}$, and simulate $\mathbb{P}^{\widehat{\theta}}$ (parametric bootstrap)

- General case; e.g., estimate the median of $\mathbb{P}$, know nothing about $\mathbb{P}$
  - Have $\widehat{\mathbb{P}}$



Recap

- Underlying assumptions (must always be questioned):
  - independent, identically distributed X_i
  - finite mean and variance

- Key properties of an estimator $\widehat{\Theta}$:
  bias, variance, standard eror, sampling distribution
  (affected by the unknown true θ)

- need to know something about these properties in order to "interpret" $\widehat{\Theta}$

- estimating standard error and bias:
  exact analysis, CLT, parametric and general bootstrap
- main focus is usually on the standard error (as bias is often smaller)

Today's agenda

- Predicting an unknown random variable
  - expectation as a predictor
  - loss functions
  - empirical risk minimization (ERM)
- Multidimensional data
  - visualization
  - i.i.d. models
  - estimation (plugin, feature matching, ML)
- Predicting one variable from another
  - model-based, $\mathbb{E}[Y \mid X]$
  - based on data (X_i, Y_i)
- Structured (e.g., linear predictors)
  - Model-based; plugin; ERM
- Caution

Why do we care about data or models?

- To make predictions [today, we focus on this]
- To understand how data were generated

θ? $\mathbb{P}_X^\theta$

X_1

i.i.d. $\vdots$

$\vdots$

X_n

X?

- Make a "good" prediction $\widehat{x}$ of X
  - based on what information?
  - with what objective?

Using data and/or models for predictions

$\theta?$ $\mathbb{P}_X^\theta$

X_1

i.i.d. $\vdots$

$\vdots$

X_n

$X?$

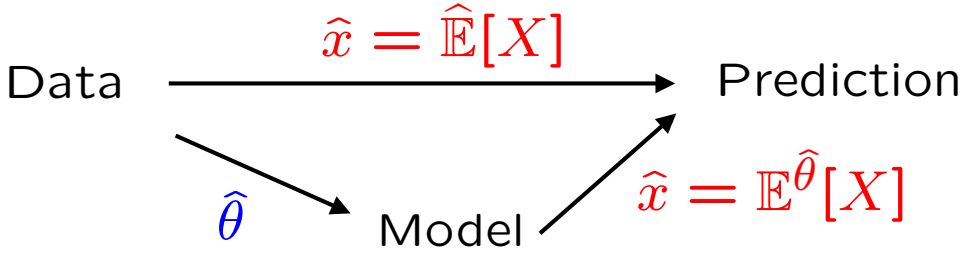
- Make a "good" prediction $\hat{x}$ of X
  - based on what information?
  - with what objective? $\min_{\hat{x}} \mathbb{E}^\theta\left[(X-\hat{x})^2\right]$

A. Have $\mathbb{P}_X^\theta$, know θ

$\hat{x} = \mathbb{E}[X]$ (data are irrelevant)

B. "Have" $\mathbb{P}_X^\theta$, don't know θ

1. Model-based: estimate $\hat{\theta}$ of θ $\longrightarrow \hat{x} = \mathbb{E}^{\hat{\theta}}[X]$
2. Plugin: $\hat{x} = \hat{\mathbb{E}}[X] = (X_1 + \cdots + X_n)/n$



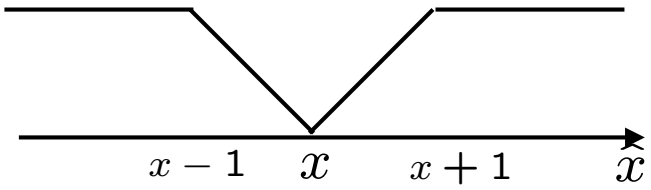
Empirical risk minimization (ERM)

- Variation of the plugin idea
  - Model-based: minimize risk $\mathbb{E}\Big[(X-\widehat{x})^2\Big]$ $\quad\widehat{x}=\mathbb{E}[X]$
  - Data-based: minimize empirical risk $\widehat{\mathbb{E}}\Big[(X-\widehat{x})^2\Big]$ $\quad\widehat{x}=\widehat{\mathbb{E}}[X]$

- Approach extends to other risk functions (estimation objectives)
  - Model-based: minimize risk $\mathbb{E}\Big[|X-\widehat{x}|\Big]$ $\quad\widehat{x}=\text{median}(\mathbb{P})$
  - Data-based: minimize empirical risk $\widehat{\mathbb{E}}\Big[|X-\widehat{x}|\Big]$ $\quad\widehat{x}=\text{median}(\widehat{\mathbb{P}})$

- Objective formulated in terms of a loss function $\ell(x,\widehat{x})$

  minimize $\mathbb{E}\Big[\ell(X-\widehat{x})\Big]$ $\qquad\widehat{\mathbb{E}}\Big[\ell(X-\widehat{x})\Big]=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\ell(x_i-\widehat{x})$



- Approach has solid theory behind

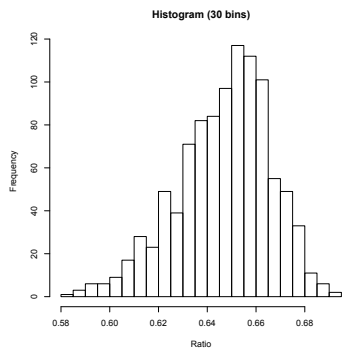
Multidimensional data

- ith sample (x_i, y_i, z_i)

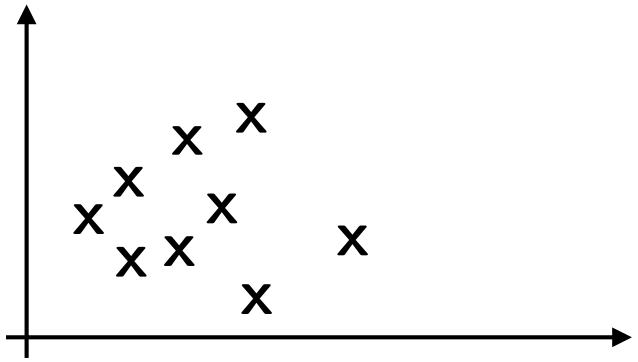
  e.g., (gender, height, weight) of ith person

- Visualize:

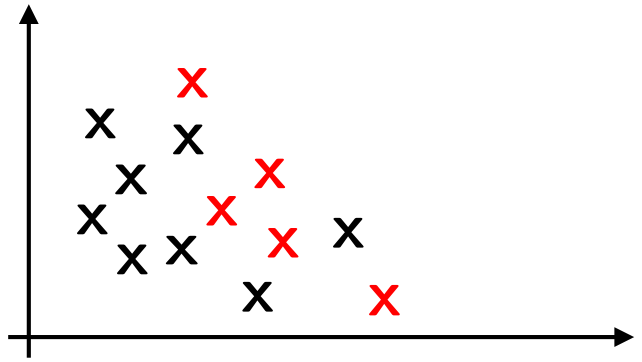
one at a time



two at a time



three, one categorical



"Standard model:" Independent random sampling

- Assume two-dimensional data (only for exposition purposes)
- $\mathbb{P}$: distribution on $\mathbb{R}^2$ (discrete; continuous; mixed)

$$p_{X,Y}(x,y) \quad f_{X,Y}(x,y)$$

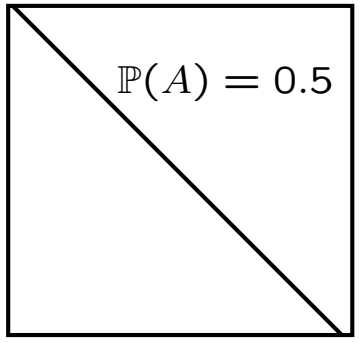
- $(X_i, Y_i) \sim \mathbb{P}$, independent for different i

  X_i and Y_i usually dependent

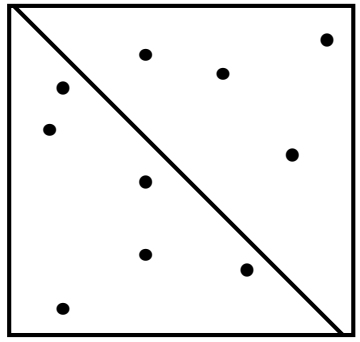
- $\widehat{\mathbb{P}}$: empirical distribution on $\mathbb{R}^2$
  - weight $1/n$ on each data point

$\widehat{\mathbb{P}}(A) =$ fraction of data that belong to A

uniform $\mathbb{P}$



$\widehat{\mathbb{P}}$



$\widehat{\mathbb{P}}(A) = 4/10$

Parameter estimation with multidimensional data

- Nothing new!

- **Plugin:**

$$\theta = \mathbb{E}\Big[g(X,Y)\Big] \qquad \widehat{\Theta} = \widehat{\mathbb{E}}\Big[g(X,Y)\Big] = \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i)$$

$$\theta = \mathsf{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

$$\widehat{\Theta} = \widehat{\mathbb{E}}[XY] - \widehat{\mathbb{E}}[X]\,\widehat{\mathbb{E}}[Y]$$

Parameter estimation with multidimensional data (ctd.)

- **Feature matching and the method of moments**
- Example:

  $\mathbb{P}^\theta \qquad \theta = (\theta_1, \theta_2, \theta_3)$

  Choose $\mathbb{E}^\theta[X],\ \mathbb{E}^\theta[Y],\ \mathbb{E}^\theta[XY]$ as features

$$\begin{aligned} \mathbb{E}^\theta[X] &= \widehat{\mathbb{E}}[X] \\ \mathbb{E}^\theta[Y] &= \widehat{\mathbb{E}}[Y] \\ \mathbb{E}^\theta[XY] &= \widehat{\mathbb{E}}[XY] \end{aligned} \qquad \text{solve for } \theta = (\theta_1, \theta_2, \theta_3)$$

  - easiest when we have closed-form expressions (as a function of θ) for left-hand side

Parameter estimation with multidimensional data (ctd.)

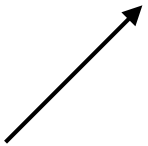
- **Maximum Likelihood**

$$\max_{\theta} \quad \prod_{i=1}^{n} \mathbb{P}^{\theta}(X_i = x_i) \qquad \prod_{i=1}^{n} f_X^{\theta}(x_i)$$

X_i, x_i are now vectors $\qquad$ f_X is a joint PDF

  - Nothing new
    - nice theoretical properties remain

- Sampling distribution, bootstrap, confidence intervals: again nothing new



one parameter at a time

What is new? New types of prediction problems

θ?	$\mathbb{P}_Y^\theta$
i.i.d.	Y_1
	$\vdots$
	$\vdots$
	Y_n
	Y?

$$\min_{\widehat{y}} \ \mathbb{E}^\theta\left[(Y-\widehat{y})^2\right]$$

θ?	$\mathbb{P}_{X,Y}^\theta$
X_1	Y_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
X_n	Y_n
X	Y?

$$\min_{g(\cdot)} \ \mathbb{E}^\theta\left[(Y-g(X))^2\right]$$

$\widehat{Y}=g(X)$ estimator

A. Have $\mathbb{P}_Y^\theta$, know θ

$\widehat{Y}=\mathbb{E}[Y]$

Have $\mathbb{P}_{X,Y}^\theta$, know θ

$\widehat{Y}=\mathbb{E}[Y\,|\,X]$

data $X_1, Y_1, \ldots, X_n, Y_n$ are irrelevant

Some details

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y]\big)^2\Big] \le \mathbb{E}\Big[\big(Y - c\big)^2\Big] \qquad \text{for all } c$$

apply to conditional universe, where $X = x$ has occurred

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x]\big)^2 \;\Big|\; X = x\Big] \le \mathbb{E}\Big[\big(Y - c\big)^2 \;\Big|\; X = x\Big], \qquad \text{for all } c$$

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x]\big)^2 \;\Big|\; X = x\Big] \le \mathbb{E}\Big[\big(Y - g(x)\big)^2 \;\Big|\; X = x\Big], \qquad \text{for all } g(\cdot)$$

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2 \;\Big|\; X\Big] \le \mathbb{E}\Big[\big(Y - g(X)\big)^2 \;\Big|\; X\Big], \qquad \text{for all } g(\cdot)$$

take expectation of both sides; use law of iterated expectations, and the property $X \le Y \;\Rightarrow \mathbb{E}[X] \le \mathbb{E}[Y]$

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2\Big] \le \mathbb{E}\Big[\big(Y - g(X)\big)^2\;\Big], \qquad \text{for all } g(\cdot)$$