SOSC 5340: Overview of Statistical Inference and Prediction

Han Zhang

Feb 2, 2021

Outline

Instructors

Instructor: ZHANG, Han

- Office: 2379
- Email: zhangh@ust.hk
- Office Hour: Tue 2-3PM
  - In office (if you are on campus)
  - Zoom: `https://hkust.zoom.us/j/6522716568`

Teaching Assistant: YIN, Yabin

- Office: 3001
- Email: yyinak@connect.ust.hk
- Office Hour: TBD

Components

- Second course in SOSC's statistics sequences for research graduate students (after SOSC 5090);
- Three core goals of social sciences:
  1. Description: describing one variable
  2. Prediction: correlation between two social phenomena.
  3. Explanation: are the correlation causal?
- Three set of knowledge/skills
  1. Statistical estimation and inference
  2. Applied regression modeling
  3. Causal inference (second half of the semester)

Grading

Attendance	5%
Assignments	35%
Presentation of an academic article (15 min)	10%
Presentation of your final paper (20 min)	15%
Write-up of your final paper	35%

Attendance

- Please turn on your video
- Online teaching can be challenging; please do ask questions whenever you are not clear!

Assignments

- Homework assignment: short coding homework to make sure that you know how to run models we covered in the lectures.
- Our TA will hold tutorial sections to teach you
  - how to run these models before assignments.
  - discuss solutions of previous assignment
  - 3-4 times

Presentation of an academic article

- As a researcher, you will need to present your own research at academic conferences.
- This exercise prepares you with relevant skills
- A list of academic publications will be distributed later
- These publications apply what we have learned in the class to real social science problems, or discuss methodological pitfalls in current applied research
- You are required to select on article, and present it to the entire class (15 minutes)

Final Paper

- Most importantly, as a researcher, you will need to apply what you have learnt to a real social science problem, and write an academic article.
- It is very important to write and present your own work.
- You need to
  - Present your own final paper to the class (15%)
  - Write it down (35%)
- Treat this as a real paper that has the potential to be published at academic journals/presented at academic conferences.

Materials

- No formal textbooks; I will remind further readings in each class
- I recommend three books
- Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics* . Cambridge University Press, 2019. (more mathematical; mostly used for the first half of the class).
- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009. (more applied; mostly used for the second half of the class).
- Hansen, Bruce. *Econometrics*, 2020. Free at the author's website `https://www.ssc.wisc.edu/~bhansen/econometrics/`

Coding

- We will use R for lectures and tutorials
- If you prefer Stata, that is okay

Social science's goals

1. Description: describing one variable
2. Prediction: correlation between two social phenomena.
3. Explanation: are the correlation causal?

- Today's lecture focuses on the first two
- How do we use statistics to do description and prediction

Random variables

- Random variable: abstraction of some concept we care about.
- Examples:
  - define random variable X as gender; it can take several values from *male*, *female*, *transgender*, ...
  - define random variable X as height; it can take numeric values.

Probability distribution

- Probability density function (PDF): $f(x)$
  - How likely does random variable X take a particular value x
  - $f(x) = P(X = x)$
- Cumulative distribution function (CDF): $F(X) = P(X \leq x)$
  - What is the probability that a random variable X takes a value equal to or less than x?

Normal Distribution

- Probability density function of the normal distribution

$$f(x) = P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- Two key parameters
  - μ, mean
  - σ, standard deviation
  - Hence normal density is often written as $X \sim N(\mu, \sigma)$; $\sim$ means follows.
- Standard normal: $\mu = 0, \sigma = 1$

Joint and Conditional Probability

- Joint probability density function: $f(X = x, Y = y)$
  - Probability that X takes value x and Y takes value y
- Conditional probability density function
  - Probability that Y takes value y, give that X takes value x.

Probability (exercise)

- Two treatments for kidney stones

	Treatment A		Treatment B	
Kidney Stone	cured	patient	cured	patient
Small	81	87	234	270
Large	192	263	55	80
Total	273	350	289	350

- X is whether the patient is cured (1) or not (0)
- What is the conditional probability of being cured, given treatment A and B?
  - $P(X = 1|treatment = A)$
  - $P(X = 1|treatment = B)$
- $P(X = 1|treatment = A) = 273/350 = 78\%$
- $P(X = 1|treatment = B) = 289/350 = 83\%$
- treatment B is more effective in the entire population

Probability (exercise)

Kidney Stone	Treatment A cured	Treatment A patient	Treatment B cured	Treatment B patient
Small	81	87	234	270
Large	192	263	55	80
Total	273	350	289	350

- What is the conditional probability of being cured, conditional on treatment status and stone size?
- Small Kidney Stone:
  - $P(X = 1|treatment = A, size = small) = 81/87 = 93\%$
  - $P(X = 1|treatment = B, size = small) = 234/270 = 87\%$
- Large Kidney Stone:
  - $P(X = 1|treatment = A, size = large) = 192/263 = 73\%$
  - $P(X = 1|treatment = B, size = large) = 55/80 = 69\%$
- B is more effective in the entire population, but A is more effect for both patients with small and large kidney stones.
- This is known as the Simpson's Paradox. Why?

Expected Value

- Expectation (expected value) $E(X)$:
  - The average value of a random variable X
- Categorical variable's expectation:
  - Let X be a random variable with a finite number of finite outcomes $x_1, x_2, \ldots, x_k$ occurring with probabilities $p_1, p_2, \ldots, p_k$
  - $E(X)$ is the weighted average of X, with probability as weights
  - $E[X] = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k = \sum_{i=1}^{k} x_i \, p_i$
- Continuous variable's expectation

$$E(X) = \int x \cdot f(x) dx$$

Expected Value (exercise)

- What is the $E(X)$ of the random variable X?

X	P(X)
0	0.8
1	0.1
2	0.06
3	0.03
4	0.04

Expected Value

- Useful formula of expected values
  1. Linearity of expectation:
     $E(aX + bY + c) = aE(X) + bE(Y) + c$
  2. Constant's expectation is constant: $E(c) = c$

Conditional Expectation

- Conditional expectation $E(Y|X = x)$:
  - On average, what is the value of a random variable Y, give that we already know that random variable X takes value x
  - Note that X is fixed to a determined value x (e.g., X is gender and x is male).
- Useful formula 3: Law of Iterated Expectation (Law of Total Expectation)

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \begin{cases} \sum_x \mathbb{E}[Y|X = x]P(X = x) & \text{discrete } X \\ \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x]f(x)dx & \text{continuous } X \end{cases} \tag{1}$$

- Basically, this theorem says that if we have knowledge about $P(X)$, and the conditional probability of $P(Y|X)$, we can calculate the average of Y.

Variance

- The variance measures the dispersion or the "spread" of a probability distribution.
- The variance of a random variable X, denoted $V(Y)$, is the expected value of the square of the deviation of Y from its mean:
- $V(X) = E\left[\left(X - E(X)\right)^2\right]$
- Standard deviation: $\sigma = \sqrt{V(X)}$

Variance

Definition (Alternative Formula for Variance)

$V(X) = E[X^2] - E[X]^2$

Proof.

$$\begin{aligned}
V(X) &= E\left[\left(X - E(X)\right)^2\right] && (2)\\
&= E[X^2 - 2XE(X) + E(X)^2] && (3)\\
&= E(X^2) - 2E[XE(X)] + E[E(X)^2] && (4)\\
&= E(X^2) - 2E(X)E(X) + E(X)^2 && (5)\\
&= E(X^2) - E(X)^2 && (6)
\end{aligned}$$

$\square$

Probability and Statistics

- Probability is defined on population
- Probability of population is often very hard to obtain;
  - We need to have information of every unit in the population
  - But it's often unrealistic
- Goal of Statistics: inferring properties of population from samples

I.I.D. random variables

- Example: X is height and we want its probability distribution of all HK residents
- We can collect every HKer's height and tabulate as we did earlier; this costs a lot of money
- Or we can sample a HKer and record his/her height X_1, and then sample another HKer get height X_2.
- This process continues 100 times, we get $(X_1, X_2, \cdots, X_{100})$
- $(X_1, X_2, \cdots, X_{100})$ are independent and identically distributed (I.I.D.)
  - independent: our i th draw does not depend on the j th draw;
    - in math: $P(X_i, X_j) = P(X_i)P(X_j)$
  - identically distributed: they all come from the same probability distribution: HKer's height.
    - They are not coming from a different distribution, say, heights of desks

I.I.D. random variables (exercise)

- When the independent assumption may be violated?
  - e.g., samples are not random, but HKUST students.
- When the identically distributed assumption may be violated?
  - e.g., population changes during the sampling process.

Sample Mean of I.I.D. random variables

- Let $X_1, \cdots, X_n$ be i.i.d. random samples of random variable X

Definition (Sample Mean)

The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Theorem (The Expected Value of the Sample Mean is the Population Mean)

$E(\bar{X}) = E(X)$

- Implication:
  - Expectation of the sample mean equals population mean, which we cannot directly obtain
  - Sample mean is something we can obtain

Sample Mean of I.I.D. random variables

The Expected Value of the Sample Mean is the Population Mean.

$$
\begin{aligned}
E(\bar{X}) &= E(\frac{1}{n}(X_1 + \cdots + X_n)) && (7)\\
&= \frac{1}{n}E(X_1 + \cdots + X_n) && (8)\\
&= \frac{1}{n}[E(X_1) + \cdots + E(X_n)] && (9)\\
&= \frac{1}{n}[E(X) + \cdots E(X)] && (10)\\
&= E(X) && (11)
\end{aligned}
$$

□

Sampling Variance of the Sample Mean

- Let $X_1, \cdots, X_n$ be i.i.d. random samples of random variable X, with finite variance $V(X)$

Theorem (Sampling Variance of the Sample Mean)

The sampling variance of the sample mean is $V(\bar{X}) = \frac{V(X)}{n}$

- That is, sampling variance of the same mean decreases, as n increases.
- Note that population variance $V(X)$ is an unknown but fixed quantity.

Law of Large Numbers

- Let $X_1, \cdots, X_n$ be i.i.d. random samples of random variable X

Theorem (Weak Law of Large Numbers, Jacob Bernoulli, 1713)

The sample mean $\bar{X}$ converges in probability to the population mean $E(X)$, as $n \to \infty$.

- Convergence in probability:
  - If a and b convergence in probability, it is very likely that their difference will be very small.
  - $\lim_{n\to\infty} P(|a-b| \leq \epsilon) = 1$, for all $\epsilon > 0$.
- Implication of the Weak Law of Large Numbers:
  - As n gets large, the sample mean X becomes increasingly closer to the population mean $E(X)$.
  - It suggests that we can use sample mean to estimate population mean
  - And perhaps other sample quantities to population quantities?

Estimator

- Let $X_1, \cdots, X_n$ be i.i.d. random samples of random variable X
- We care about some population quantity of interest θ (e.g., mean, variance, median, etc)

Definition (Estimate and Estimator)

Estimator of a population quantity θ is a function of the samples, $\hat{\theta} = h(X_1, \cdots, X_n)$; $\hat{\theta}$ is the estimate of θ.

- In a nutshell, statistics uses estimator to provide estimate of population quantity
- Example: an estimator of population mean $E(X)$ is sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Note: there are usually many different estimators of the same quantity.
  - E.g., X_1 is also an estimator of $E(X)$. But intuitively it is not as good as the sample mean.
  - How can we say one estimator is better than the other? What properties should good estimators have?

Desirable Property: Unbiasedness

- For an estimator $\hat{\theta}$, its bias is $E(\hat{\theta}) - \theta$

Definition (Unbiased Estimator)

An estimator $\hat{\theta}$ of θ is an unbiased estimator if $E(\hat{\theta}) = \theta$ or bias is 0

- Question: sample mean $\bar{x}$ is an unbiased estimator of population mean $E(X)$. Why?
- Answer: because the expectation of sample mean equals to population mean ($E(\bar{X}) = E(X)$)

see proof here

Desirable property: Consistency

Definition (Consistent Estimator)

An estimator $\hat{\theta}$ is an consistent estimator if $\hat{\theta}$ converges in probability to θ, as $n \to \infty$.

- Question: sample mean is an consistent estimator of population mean. Why?
- Answer: because of the Law of Large Numbers.

Estimation basics

- Let $X_1, \cdots, X_n$ be i.i.d. random samples of random variable X with variance $V(X)$
- We see that sample mean $\bar{X}$ is an unbiased and consistent estimator of population mean $E(X)$
- It is tempting to extend this method to other population quantities, by:
  1. express the (unestimated) poulation quantity as some population quantity that are estimatable.
  2. plug-in the sample estimator.
- This is called plug-in principle.

Application: plug-in estimator for population variance

- We want to apply the plug-in principle to estimate population variance $V(X)$.
- Step 1: express $V(X) = E[X^2] - E[X]^2$, because we already know how to estimate $E(X)$: $\bar{X}$
- Step 2: plug-in $\bar{X}$ in place of $E(X)$

Definition (Plug-in Variance Estimator)

$\hat{V}(X) = \overline{X^2} - \overline{X}^2$

- Is this plug-in variance estimator a good estimator?
    - As we have learned, an good estimator should have two good properties
    - unbiased?
    - consistent?

Plug-in estimator for population variance

- Unbiased estimator means $E(\hat{\theta}) - \theta = 0$
- Our variance estimator of $V(X)$ is $\widehat{V(X)} = \overline{X^2} - \overline{X}^2$
- Unbiasedness:

$$
\begin{aligned}
\mathrm{E}\left(\widehat{V(X)}\right) = \mathrm{E}\left[\overline{X^2} - \bar{X}^2\right] &= \mathrm{E}[\overline{X^2}] - \mathrm{E}\left[\bar{X}^2\right] && (12)\\
&= \mathrm{E}\left[X^2\right] - \left(\mathrm{E}[X]^2 + \mathrm{V}[\bar{X}]\right) && (13)\\
&= \left(\mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2\right) - \frac{\mathrm{V}[X]}{n} && (14)\\
&= \mathrm{V}[X] - \frac{\mathrm{V}[X]}{n} && (15)\\
&= \frac{n-1}{n}\mathrm{V}[X] && (16)
\end{aligned}
$$

Unbiased Estimator for Population Variance

- Plug-in population variance estimator $\hat{V}(X) = \overline{X^2} - \overline{X}^2$ is biased
- Plug-in population variance estimator $\hat{V}(X) = \overline{X^2} - \overline{X}^2$ is consistent (as $n \to \infty$, $\frac{n-1}{n}$ goes to 1)
- In general, plug-in estimator is consistent, but may be biased (advanced topic).

Theorem (Unbiased Estimator of Population Variance)

$\hat{V}(X) = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right)$ *is an unbiased and consistent estimator of population variance* $V(X)$

Estimator for variance of the sample mean

- The previous slides derives $\hat{V}(X)$, the estimator for population variance.
- But it is not estisampling mator of the variance of the sample mean $V(\bar{X})$
- Try plug-in estimator thisampling s time
  - Step 1: express the quantity of interest $V(\bar{X}) = \frac{V(X)}{n}$
  - Step 2: plug-in (since we just shown how to estimate $V(X)$ in the previous slide)

Theorem (Estimator of the Sampling Variance of the Sample Mean)

$\hat{V}(\bar{X}) = \frac{\hat{V}(X)}{n}$

- Plug-in estimator is an unbiased and consistent estimator this time (proof omitted)
- $\sqrt{\hat{V}(\bar{X})}$ is called standard error.

Inference vs Estimation

- Estimation is about getting the (point) estimate $\hat{\theta}$ of quantity of interest θ
  - With unbiased estimator, $\hat{\theta}$ on average equals to θ
  - With consistent estimator, $\hat{\theta}$ converges to θ with more and more sample data
  - But in reality, we have only one $\hat{\theta}$
- Inference is about how certain we are about the estimate $\hat{\theta}$

Central Limit Theorem

- Let $X_1, \cdots, X_n$ be i.i.d. random samples of random variable X, with finite $E(X) = \mu$ and $V(X) = \sigma^2 > 0$

Definition (Standardized Sample Mean)

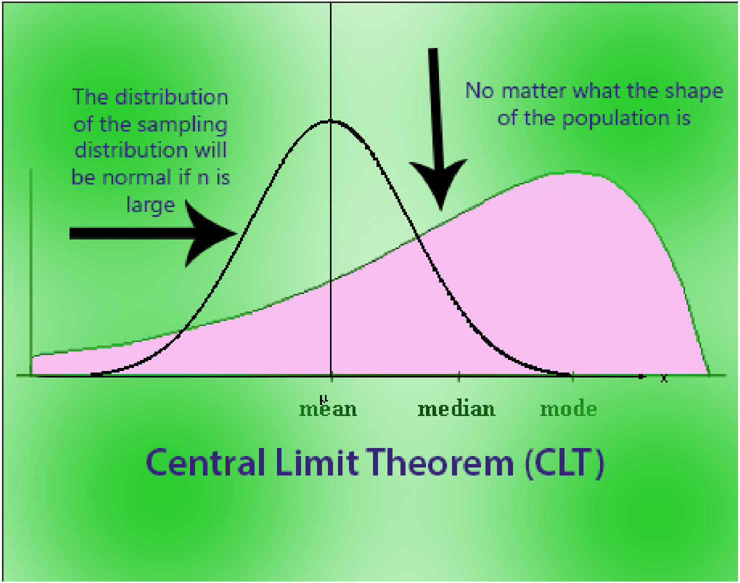
$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$

- $E(Z) = 0$; $V(Z) = \sigma(Z) = 1$; hence the name standardized sample mean.

Theorem (Central Limit Theorem)

- *The distribution of Z converges to a standard normal distribution ($Z \sim N(0, 1)$), as $n \to \infty$.*
- *Or equivalently,* $\sqrt{n}(\bar{X} - \mu) \sim N(0, \sigma^2)$

Central Limit Theorem



Implications of Central Limit Theorem

- Question: what's the difference between Law of Large Numbers and the Central Limit Theorem?
- Answer:
  - Law of large numbers suggests that sample mean converges to population mean. It's a property of the sample mean.
  - Central limit theorem suggest that (sample mean - population mean) roughly follows a normal distribution centered around 0. It's a property of the distribution of sample mean
- CLT is a stronger statement that LLN.
- Sampling distribution of the sample mean will tend to be approximately normal, even when the population distribution is not distributed normally;
- Thus, central limit theorem provides a general way for us to infer the uncertainty around our estimate of sample mean (and other quantities)

Desirable Property: asymptotically normal

- Central Limit Theorem means that sampling distribution of the sample mean will tend to be approximately normal
- This is another desirable property of estimator, called asymptotic normal

Definition (Asymptotic Normal Estimator)

An estimator $\hat{\theta}$ is an asymptotically normal estimator, if $\sqrt{n}(\hat{\theta} - \theta) \sim N(0, \phi^2)$ for finite $\phi > 0$, as $n \to \infty$.

- Many estimators you will learn in this course is asymptotically normal
  - But not all estimators have this good property
- The good thing about asymptotically normal estimator is that we can obtain confidence interval easily

Confidence Interval

Definition (Confidence interval)

A α confidence interval for quantity of interest θ is an estimated interval that covers the true value of θ with at least α probability

- Example: in social sciences, we often uses $\alpha =$95% confidence interval that looks like $[\theta_{min}, \theta_{max}]$. The probability that the true θ falls between $[\theta_{min}, \theta_{max}]$ is at least 95%.
- Note 1: wide confidence intervals are valid, but not useful
  - e.g., θ is the average height of HKers; $[0, 2.5]$ is a valid confidence interval but it is not very useful.
- Note 2: confidence interval does not need to be symmetric

Normal Approximation-based Confidence Interval

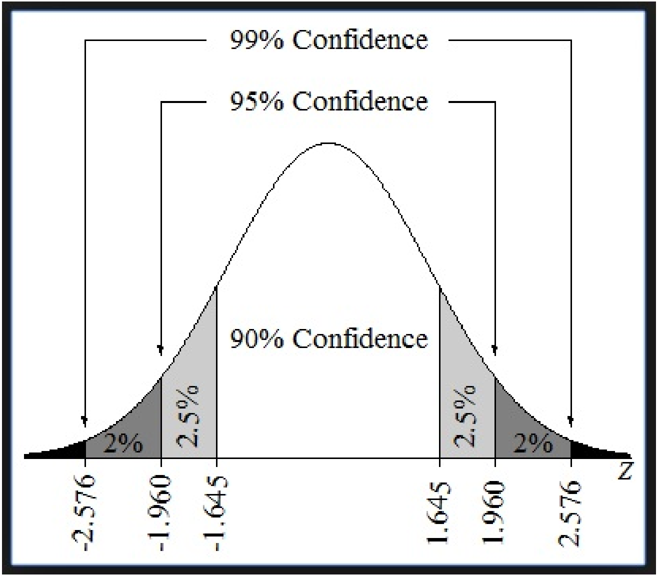
Definition (Estimating Normal Approximation-based Confidence Interval)

A normal approximation-based confidence interval for θ can be estimated by:

$$\left(\hat{\theta} - z_{\frac{1+\alpha}{2}}\sqrt{\hat{V}(\hat{\theta})}, \hat{\theta} + z_{\frac{1+\alpha}{2}}\sqrt{\hat{V}(\hat{\theta})}\right)$$

- Intuition: the standarized sample mean follows a standard normal distribution, given the Central Limit Theorem.
- z is the quantile function of a standard normal distribution
  - $\alpha = 0.95$; $z_{0.975} = 1.96$
  - $\alpha = 0.99$; $z_{0.995} = 2.58$
- Normal Approximation-based Confidence Interval is valid for asymptotically normal estimators

Illustration



Steps to empirically estimate confidence interval

- Steps to estimate the Normal Approximation-based Confidence Interval for sample mean in a given sample
- Step 1: calculate sample mean $\bar{X}$ and sampling variance of the sample mean $\hat{V}(\bar{X})$
- Step 2: construct confidence interval as

$$\left(\bar{X} - z_{\frac{1+\alpha}{2}}\sqrt{\hat{V}(\bar{X})}, \bar{X} + z_{\frac{1+\alpha}{2}}\sqrt{\hat{V}(\bar{X})}\right)$$

- E.g., for 95% confidence interval

$$\left(\bar{X} - 1.96\sqrt{\hat{V}(\bar{X})}, \bar{X} + 1.96\sqrt{\hat{V}(\bar{X})}\right)$$

Bootstrap

- Normal Approximation is not the only way to construct valid confidence intervals
  - Reason 1: it only works for asymptotic normal estimator
  - Reason 2: you have to know what $\hat{V}(\bar{X})$.
- The Bootstrap is more general method to construct confidence intervals; one of the most important modern statistical concept (Efron, 1979)
  - Drawback of Bootstrap: it's a data-driven method; slow; no analytical solutions.

Bootstrap procedures

Assume we already have $X_1, \cdots, X_n$ be i.i.d. random samples of random variable X). We are interested in estimating a α confidence interval for a population quantity θ

1. Take a with replacement sample of size n from $X_1, \cdots, X_n$
2. Calculate the sample analog of θ
3. Repeat 1 and 2 for m times. We end up having m estimates of θ, $(\hat{\theta}_1, \cdots, \hat{\theta}_m)$
4. Take the $\frac{1-\alpha}{2}$ and $\frac{1+\alpha}{2}$ quantile of the values $(\hat{\theta}_1, \cdots, \hat{\theta}_m)$. These two quantiles give us the bootstrap confidence intervals.

Confidence Interval: Interpretations

- Interpretating confidence intervals carefully
- a 95% confidence interval $[\theta_{min}, \theta_{max}]$ contains the population quantity θ with at least 95% probability.
- In reality, we are estimating confidence intervals $[\theta_{min}, \theta_{max}]$.

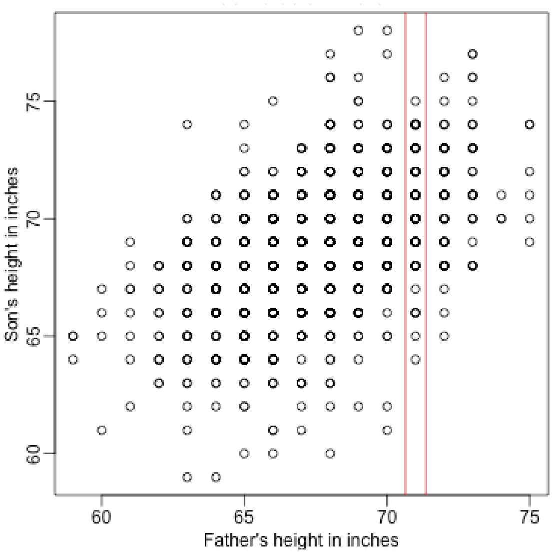
- How should we interpret 95% estimated confidence interval $[\hat{\theta}_{min}, \hat{\theta}_{max}]$ then?
  - Through repeated samples (each time we sample n units), 95% of estimated confidence intervals would contain the population quantity θ
  - If we know $[\theta_{min}, \theta_{max}]$, we do not need repeated sampling
  - note that in reality we only have one sample (of n units)

Prediction

- Now let us move on to two variable setting
- Given two variables Y and X, and we observed X takes the value x.
- A prediction of Y given X is a function $g(X)$ that approximate Y
- For instance, $E(Y|X)$ is a prediction of Y given X
- Again, there are tons of ways to predict Y given X (e.g., median of Y given X)

Prediction (example)

- Predicting son's height with father's height



Using Conditional Expectation as Prediction

- If $g(X) = E(Y|X)$, that is, we use the conditional expectation as the prediction
- The prediction error is $\epsilon = Y - E(Y|X)$
- This prediction error has some good properties

Property 1: $E(\epsilon) = 0$.

$$
\begin{aligned}
E(\epsilon) &= E\big[Y - E(Y|X)\big] && (17)\\
&= E(Y) - E\big[E(Y|X)\big] && (18)\\
&= E(Y) - E(Y), \text{(Law of Iterated Expectation)} && (19)\\
&= 0 && (20)
\end{aligned}
$$

$\square$

Conditional Expectation as Prediction (cont'd)

Property 2: $E(\epsilon|X) = 0$.

$$
\begin{aligned}
E(\epsilon|X) &= E\big[Y - E(Y|X)|X\big] && (21)\\
&= E(Y|X) - E\big[E(Y|X)|X\big] && (22)\\
&= E(Y|X) - E(Y|X), \text{(Law of Iterated Expectation)} && (23)\\
&= 0 && (24)
\end{aligned}
$$

$\square$

Using Conditional Expectation as Prediction

- Property 2 means that on conditional on X, the mean of prediction error is 0
- This property is also called mean independent because $E(\epsilon|X) = E(\epsilon) = 0$
  - That is, we only assume that on average X and error are independent
  - Recall independence means that $P(\epsilon|X) = P(\epsilon)$

Independent, mean independent, and uncorrelated

- Independent: $P(XY) = P(X)P(Y)$
- Mean independent: $P(Y|X) = E(Y)$
- Uncorrelated: $E(XY) = E(X)E(Y)$
- In general, we have the following relationship (the reverse is not true):

X, Y are independent $\implies$ X, Y are mean independent $\implies$ X, Y are uncorrelated.

Using Conditional Expectation as Prediction

- Property 3 says $g(X)$ and error is uncorrelated; it can be derived from Property 2 (mean independence) and Property 1 ($E(\epsilon) = 0$)

Property 3: $E\big(g(X)\epsilon\big) = 0$, for any $g(X)$.

$$
\begin{aligned}
E\big[g(X)\epsilon\big] &= E\big[g(X)\big(Y - E(Y|X)\big)\big] && (25)\\
&= E\big[g(X)Y - g(X)E(Y|X)\big] && (26)\\
&= E\big[g(X)Y\big] - E\big[g(X)E(Y|X)\big] && (27)\\
&= E\big[g(X)Y\big] - E\Big[E\big(g(X)Y|X\big)\Big], \text{(g(X) is a constant given } X) && (28)\\
&= E\big[g(X)Y\big] - E\big[g(X)Y\big], \text{(Law of Iterated Expectation)} && (29)\\
&= 0 && (30)
\end{aligned}
$$

Evaluating Predictions

- We have seen that $E(Y|X)$ is a good guess for Y:
  - Property 1: mean error is 0
  - Property 2: error and prediction $g(X)$ is mean independent
  - Property 3: error and prediction $g(X)$ is uncorrelated
- But Mean error $E(\epsilon) = E(Y - g(X))$ has one drawback: insensitive to the sign of error
- e.g., $Y = 0$; our guesses $g(X)$ are -100, 100, -100, 100
  - Intuitively these guesses are not good
  - But $E(\epsilon) = 0$

MAE and MSE

- Mean Absolute Error (MAE): $E[|Y - g(X)|]$
- Mean Square Error (MSE): $E[(Y - g(X))^2]$
- MSE make sure that you get penalized more, if the absolute error is large.
    - MSE is perhaps the most widely used error metric
- Both MAE and MSE ≥ 0; a good estimation thus should minimize MAE or MSE

Conditional Expectation As the Best Predictor

- There are some even better properties of $E(Y|X)$ that make it the best predictor, given Mean Squared Error

Theorem (Conditional Expectation as the Best Predictor)

Conditional Expectation Function $E(Y|X)$ is the best predictor of Y because it minimizes Mean Squared Error

- We have two predictions for Y, $E(Y|X)$ and any other $g(X)$
- We want to show that the MSE of any other $g(X)$ not smaller than the MSE of $E(Y|X)$
- In math term: $E[(Y - g(X))^2] \geq E[(Y - E(Y|X))^2]$
- Hint: use the conditional expectation error $\epsilon = Y - E(Y|X)$

Conditional Expectation As the Best Predictor

Conditional Expectation as the Best Predictor.

$$
\begin{aligned}
E\left[\left(Y - g(X)\right)^2\right] &= E\left[\left(\epsilon + E(Y|X) - g(X)\right)^2\right] && (31)\\
&= E\left[\epsilon^2 + 2\epsilon\left(E(Y|X) - g(X)\right) + \left(E(Y|X) - g(X)\right)^2\right] && (32)\\
&= E\left[\epsilon^2\right] + 2E\left[\epsilon\left(E(Y|X) - g(X)\right)\right] + E\left[\left(E(Y|X) - g(X)\right)^2\right] && (33)\\
&= E\left[\epsilon^2\right] + E\left[\left(E(Y|X) - g(X)\right)^2\right], \text{(Property 3)} && (34)\\
&\geq E\left[\epsilon^2\right] && (35)\\
&= E\left[\left(Y - E(Y|X)\right)^2\right] && (36)
\end{aligned}
$$

$\square$

Conditional Expectation As the Best Predictor

- Note that this says that the conditional expectation gives an upper bound on how well we can make a guess of Y based on X, if we want to minimize MSE
- If the conditional expectation itself is not a very good predictor, we can still make lots of errors
    - But in this case, other predictions can only be worse

Today's summary

- Population/sample
- Estimator; three good properties of estimator
- Inference; confidence interval; normal approximation vs. Bootstrap
- Conditional expectation is the best predictor in minimizing MSE

Today's readings

- Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics* . Cambridge University Press, 2019. (Chapter 2 - 4)
- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009.
  - Discuss Conditional expectation is the best predictor (Chapter 3.1)
  - Motivated differently from Aronow and Miller