Random Explanatory variables[1]

STA 2101 Fall 2019

[1]See last slide for copyright information.

Overview

Preparation: Indicator functions

Conditional expectation and the Law of Total Probability

$I_A(x)$ is the *indicator function* for the set A. It is defined by

$$I_A(x) = \begin{cases} 1 & \text{for } x \in A \\ 0 & \text{for } x \notin A \end{cases}$$

Also sometimes written $I(x \in A)$

$$\begin{aligned} E(I_A(X)) &= \sum_x I_A(x) p(x), \text{ or} \\ & \quad \int_{-\infty}^{\infty} I_A(x) f(x)\, dx \\ &= P\{X \in A\} \end{aligned}$$

So the expected value of an indicator is a probability.

Applies to conditional probabilities too

$$
\begin{aligned}
E(I_A(X)|Y) &= \sum_x I_A(x)p(x|Y), \text{ or} \\
& \quad \int_{-\infty}^{\infty} I_A(x)f(x|Y)\,dx \\
&= Pr\{X \in A|Y\}
\end{aligned}
$$

So the conditional expected value of an indicator is a *conditional* probability.

Double expectation

$$E\left(X\right) = E\left(E[X|Y]\right) = E(g(Y))$$

Showing $E(X) = E(E[X|Y])$

Again note $E(E[X|Y])$ is an example of $E(g(Y))$

$$
\begin{aligned}
E(E[X|Y]) &= \int E[X|Y=y] f_y(y)\, dy \\
&= \int \left(\int x f_{x|y}(x|y)\, dx \right) f_y(y)\, dy \\
&= \int \left(\int x \frac{f_{x,y}(x,y)}{f_y(y)}\, dx \right) f_y(y)\, dy \\
&= \int \int x\, f_{x,y}(x,y)\, dx\, dy \\
&= E(X)
\end{aligned}
$$

Double expectation: $E\left(g(X)\right) = E\left(E[g(X)|Y]\right)$

$$E\left(E[I_A(X)|Y]\right) = E[I_A(X)] = Pr\{X \in A\}, \text{ so}$$

$$\begin{aligned} Pr\{X \in A\} &= E\left(E[I_A(X)|Y]\right) \\ &= E\left(Pr\{X \in A|Y\}\right) \\ &= \int_{-\infty}^{\infty} Pr\{X \in A|Y = y\} f_Y(y)\, dy, \text{ or} \\ & \quad \sum_y Pr\{X \in A|Y = y\} p_Y(y) \end{aligned}$$

This is known as the *Law of Total Probability*

Don't you think its strange?

- In the general linear regression model, the X matrix is supposed to be full of fixed constants.
- This is convenient mathematically. Think of $E(\widehat{\boldsymbol{\beta}})$.
- But in any non-experimental study, if you selected another sample you'd get different X values, because of random sampling.
- So X should be at least partly random variables, not fixed.
- View the usual model as *conditional* on $\mathcal{X} = X$.
- All the probabilities and expected values in the typical regression course are *conditional* probabilities and *conditional* expected values.
- Does this make sense?

$\widehat{\boldsymbol{\beta}}$ is (conditionally) unbiased

$$E(\widehat{\boldsymbol{\beta}}|\mathcal{X} = X) = \boldsymbol{\beta} \text{ for } \textit{any} \text{ fixed } X$$

It's *unconditionally* unbiased too.

$$E\{\widehat{\boldsymbol{\beta}}\} = E\{E\{\widehat{\boldsymbol{\beta}}|X\}\} = E\{\boldsymbol{\beta}\} = \boldsymbol{\beta}$$

Perhaps Clearer

$$
\begin{aligned}
E\{\widehat{\boldsymbol{\beta}}\} &= E\{E\{\widehat{\boldsymbol{\beta}}|X\}\} \\
&= \int \cdots \int E\{\widehat{\boldsymbol{\beta}}|\mathcal{X} = X\} \, f(X) \, dX \\
&= \int \cdots \int \boldsymbol{\beta} \, f(X) \, dX \\
&= \boldsymbol{\beta} \int \cdots \int f(X) \, dX \\
&= \boldsymbol{\beta} \cdot 1 = \boldsymbol{\beta}.
\end{aligned}
$$

Conditional size α test, Critical value f_α

$$Pr\{F > f_\alpha | \mathcal{X} = X\} = \alpha$$

$$\begin{aligned} Pr\{F > f_\alpha\} &= \int \cdots \int Pr\{F > f_\alpha | \mathcal{X} = X\} f(X)\, dX \\ &= \int \cdots \int \alpha f(X)\, dX \\ &= \alpha \int \cdots \int f(X)\, dX \\ &= \alpha \end{aligned}$$

The moral of the story

- Don't worry.
- Even though the explanatory variables are often random, we can apply the usual fixed X model without fear.
- Estimators are still unbiased.
- Tests have the right Type I error probability.
- Similar arguments apply to confidence intervals and prediction intervals.
- And it's all distribution-free with respect to X.

Copyright Information