Lecture 3: Statistics Review and Linear Regression

Pierre Biscaye

Fall 2022

Distribution functions

Properties of *PDFs* $f(x)$

- $f(x) = Pr(X = x) \geq 0$ for all values of x
- $\sum_x f(x) = \int_x f(x)dx = 1$

CDFs $F(x)$ characterize $Pr(X \leq x)$

- $F(x) = \int_{-\infty}^{x} f(x)dx$ for a continuous random variable
- $F(x) = \sum_{min(X)}^{x} f(x)$ for a discrete random variable

Expected value

- The *Expected Value* of X is the weighted average of all possible realizations of X.

$$E[X] = \mu = \sum_x xf(x) \tag{1}$$

or

$$E[X] = \int_x xf(x)dx \tag{2}$$

Expected value of functions

- We can also take the expected value of functions of x, say $g(x)$.

$$E[g(X)] = \sum_x g(x)f(x) \tag{3}$$

or

$$E[g(X)] = \int_x g(x)f(x)dx \tag{4}$$

Example

- Suppose $P(X = 1) = 1/8, P(X = 2) = 1/2, p(X = 3) = 3/8$. Then
- $E[X] = \frac{1}{8}(1) + \frac{1}{2}(2) + \frac{3}{8}(3) = 18/8$; and
- $E[X^2] = \frac{1}{8}(1) + \frac{1}{2}(4) + \frac{3}{8}(9) = 5\frac{1}{2}$

Properties of expectations

- Expectations are linear: $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$
- Suppose $g(X) = aX + b$ where a and b are constants

$$\begin{aligned} E[g(X)] &= E[aX + b] = E[aX] + E[b] \\ &= \sum_x axf(x) + \sum_x bf(x) \\ &= a\sum_x xf(x) + b\sum_x f(x) \\ &= aE[X] + b \end{aligned}$$

Variance

- The *Variance* of X is defined as

$$\sigma^2 = E[(X-\mu)^2] = \sum_x f(x)(x-\mu)^2 \tag{5}$$

Note that

$$\sigma^2 = E[X^2 - 2X\mu + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - \mu^2 \tag{6}$$

- The *standard deviation* of X is the square root of the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{E[X^2] - \mu^2} \tag{7}$$

2 helpful rules

$$E[aX + b] = aE[X] + b \tag{8}$$
$$var(aX + b) = a^2 var(X) \tag{9}$$

Multiple Random Variables

- For most interesting problems, we will need more than one random variable.
  - e.g. wages and education; CO_2 and GDP.
  - often, a *dependent variable* Y and one (or more) *independent variable* X.
- With many random variables, need to consider the *joint distribution*. For Discrete Variables:

$$f_{x,y} = P(X = x, Y = y) \tag{10}$$

Joint PDF Example

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

What is the probability that an individual is the head of household with an incomplete primary education?

Conditional Density

- With a well-defined joint density, can define the conditional density

$$f_{x|y}(x|Y=y) = P(X=x|Y=y) \tag{11}$$

$$f_{x|y}(x|Y=y) = \frac{f_{x,y}(x,y)}{f_y(y)} = \frac{P(X=x, Y=y)}{P(Y=y)} \tag{12}$$

Conditional Density Example

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

What is the probability that an individual has an incomplete primary education *given* that they are the head of household?

$$f_{Ed|Head}(Ed = IncompletePrimary|Head = Yes)$$

Conditional Density Example

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

What is the probability that an individual has an incomplete primary education *given* that they are the head of household?

$$
\begin{aligned}
& f_{Ed|Head}(Ed = IncompletePrimary | Head = Yes) \\
&= \frac{P(Ed = IncompletePrimary, Head = Yes)}{P(Head = Yes)} \\
&= \frac{0.05}{0.05 + 0.07 + 0.19} \\
&= \frac{0.05}{0.31} \approx 0.15
\end{aligned}
$$

Independent Random Variables

- If 2 variables are independent (for discrete, continuous is analogous):

$$\begin{aligned} P(X = x, Y = y) &= f_{x,y}(X = x, Y = y) \\ &= f_x(x) * f_y(y) \\ &= P(X = x) * P(Y = y) \\ P(X = x|Y = y) &= f_{x|y}(X = x|Y = y) \\ &= \frac{f_{x,y}(x, y)}{f_y(y)} = \frac{f_x(x) * f_y(y)}{f_y(y)} \\ &= f_x(x) = P(X = x) \end{aligned}$$

- This is not true in general for random variables

Relationships between variables

The *Covariance* of X and Y is defined

$$Cov_{x,y} = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - \mu_x \mu_y \tag{13}$$

The *Conditional Expectation* of X given Y is

$$E[X|Y] = \sum_x x f_{x|y}(x|y) \tag{14}$$

Example

Suppose Primary Ed = 1, Secondary = 2, Tertiary = 3. What education levels do household heads have on average?

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

$$E[?|?] = ...$$

Example

Suppose Primary Ed = 1, Secondary = 2, Tertiary = 3. What education levels do household heads have on average?

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

$$E[Ed|Head = Yes] = \sum_{Ed} Ed \cdot f_{Ed|Head}(Ed|Head = Yes)$$
$$= 1\frac{0.05}{0.31} + 2\frac{0.07}{0.31} + 3\frac{0.19}{0.31} = 2.3$$

Characteristics of Random Variables in the population vs. sample

- So far, all of the descriptives we've built (PDF, CDF, μ, etc.) are the true parameters
  - These apply to the population: characteristics of a random variable that we should expect on average
- In the sample, we never actually oberve these true population parameters.
- Suppose the sample is a random draw from the popualtion, $i = 1, ..., n$. For each i we observe x_i, y_i.

Sample mean and the Law of Large Numbers

- In the sample we observe the *Sample Mean*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{15}$$

- $\bar{x}$ is a sample characteristic $\Rightarrow$ $\bar{x}$ is a random variable

$$E[\bar{x}] = E[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}E[\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}nE[x_i] = \mu$$

- the *Law of Large Numbers* tells us that if n is large enough, $\bar{x} \approx \mu$.
- Example with GDP data

Variance in the sample

- We also don't observe σ_x^2, we observe the *Sample Variance*

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - N\bar{x}^2) \tag{16}$$

- We also observe the *Sample Covariance*

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - N\bar{x}\bar{y}) \tag{17}$$

- For large n, $s_x^2 \approx \sigma_x^2$ and $s_{xy} \approx \sigma_{xy}$.

Relating sample characteristics to regression models

- We want to understand how (say) Carbon Emissions relate to GDP
- This is something like a conditional expectation
- We have data from a sample with random variables (at least one x and one y)
  - We can construct sample characteristics (Sample Mean, Sample Variance, Sample Covariance)
- How do we recover $E[Y|X]$, or in this case $E[CO_2/cap|GDP/cap]$?

Statistical models for conditional expectations

In this class, we'll mainly start by assuming

$$E[Y|X] = \beta_0 + \beta_1 X \tag{18}$$

In the sample, this means

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{19}$$

- the ϵ_i's are how each observation i differs from the Conditional Expectation.
  - We should want to make these small (and on average 0)

Ordinary Least Squares

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Least squares regression assumes: let's find small ϵ_i's by solving the *objective function*

$$min_{\beta_0,\beta_1} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

An optimization problem: take first derivatives and set equal to 0

Deriving Least Squares Estimators

$$\frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} = -2 \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{20}$$

$$\frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} = -2 \sum_i x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{21}$$

- Moment interpretation: first FOC: $E[\epsilon] = 0$
- second FOC: $Cov(X, \epsilon) = 0$ (Why? Work this out on your own)

Solve for βs by substitution

Deriving $\hat{\beta}_0$

$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{1}{N} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Deriving $\hat{\beta}_1$

$$\sum_i x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_i x_i(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})) = 0$$

$$\sum_i (x_i y_i - x_i \bar{y}) = \hat{\beta}_1 \sum_i (x_i^2 - x_i \bar{x})$$

And therefore

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2}$$

Intuition behind $\hat{\beta}_1$

Recalling that

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}^2) = \frac{1}{n-1}(\sum_{i=1}^{n} x_i - N\bar{x}^2)$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - N\bar{x}\bar{y})$$

We have

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2} \\
&= \frac{s_{xy}}{s_x^2} = \frac{\widehat{Cov(x,y)}}{\widehat{Var(x)}}
\end{aligned}$$

Excel example

Using these expressions

$$\hat{\beta_1} = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2} \tag{22}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{23}$$

to recover $\hat{\beta}$ estimates for

$$\frac{CO_{2i}}{Pop_i} = \hat{\beta_0} + \hat{\beta_1}\frac{GDP_i}{Pop_i}$$

using real data

Tying back to conditional expectations

We proposed a simple linear model for conditional expectations

$$E[Y|X = x] = \beta_0 + \beta_1 x \tag{24}$$

We derived equations to estimate the β parameters for a particular sample

$$\hat{\beta_1} = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2} \tag{25}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{26}$$

We can use these estimates to generate *predicted values*

$$\widehat{E[y_i|X = x_i]} = \hat{y}_i = \hat{\beta_0} + \hat{\beta_1} x_i \tag{27}$$