Probability distributions in regression modeling

Contents

*Disclaimer: sections and lines in brown correspond to content which is **very much** 'under construction'.*

1 Motivation – Probability distributions in regression modeling

In regression modeling, we need random variables – and their probability distributions– because our models do not fit our data exactly.
A regression model is composed of:

1. a deterministic model which captures as much of the data variation as possible;
2. an error term e characterized by a probability distribution which captures the unexplained variation (the variation that remains *after* predicting the average).

Regression models are therefore not deterministic. Whether they are used for parameter estimation or prediction, **we assess how uncertain our estimates and/or predictions are using probability distributions.**

Note: is X a random or fixed variable?

In a regression model $y \sim f(X, e|\theta)$, e is a random variable and therefore y, a transformation of e, is itself a random variable. However, the explanatory variables X, may be considered random or fixed.

The implications include which CLTs and LLNs we'll draw sampling-based inference from, and the formulation of the model as y's individual or conditional distribution. The details go beyond the point of this document; we will simply make the following summary:

- In experimental studies, the researcher "sets" the values of X, therefore X are typically considered fixed, i.e., real vectors $X \in \mathbb{R}^p$.
  - model: a function f s.t. $y \sim f(X, e|\theta)$
  - estimation: in OLS, we assume $\mathbb{E}[e] = 0$ and look for $\mathbb{E}[y]$.
- In observational studies, we draw a sample $\{y, X\}$ from the population, therefore X are typically considered random.
  - model: f now gives the conditional probability: $y|X \sim f(X, e|\theta)$
  - estimation: in OLS, we assume in addition $\mathbb{E}[e|X] = 0$ and look for $\mathbb{E}[y|X]$.

In general, we will assume the context of an observational study, and therefore consider X an r.v.

2 Definitions

2.1 Random variables & probability distributions

A **random variable (r.v.) X** is a function that maps the outcome of a random process to a numeric value. It is either:

- discrete – ex: the outcome of a coin toss: $X = \{1 \text{ if tails}, 0 \text{ if heads}\}$; or
- continuous – ex: a continuous uniform r.v. $X \sim \mathrm{U}[a, b]$

It has a **probability distribution** which gives the probability of occurrence of each value. We can characterize the distribution using:

- the **probability function** $\boldsymbol{f_X(x) \equiv P[X = x]}$
- the **cumulative distribution function** $\boldsymbol{F_X(x) \equiv P[X < x]}$, i.e., the area under the density function $f_X()$ over $]-\infty, x[$.

A set of random variables, e.g., X, Y, have each their *individual* or *marginal* probability distribution, but also a *joint* probability distribution $f_{XY}()$, and *conditional* probability distributions:

$$f_{XY}(x, y) = f_X(x) f_{Y|X}(y|x) = f_{X|Y}(x|y) f_Y(y)$$

2.2 Every probability distribution has its moments...

We often focus on a few moments of a random variable's probability distribution: its 1st, 2nd central, and 3rd standardized moments, and when looking at multiple random variables, their 2nd central mixed moment[1]:

For X:

Expectation $\mathbb{E}[X] = \int x f_X(x)\, dx = \int x\, dF_X(x)$

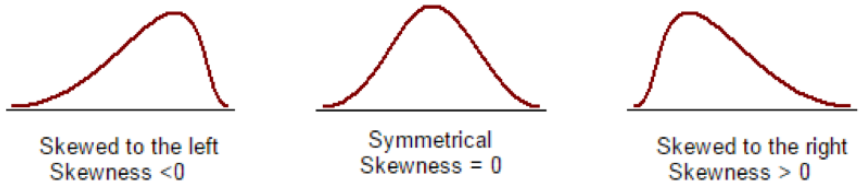
Variance $\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$

Standard deviation $\sqrt{\mathbb{V}[X]}$ is a measure of variability. It is commonly used as about 95% of observations of any distribution usually fall within 2 SDs. We may choose a different summary statistic, however, when data have a skewed distribution.

Skewness $Skew[X] = \dfrac{\mathbb{E}[(X - \mu_X)^3]}{\sigma_X^3}$

Is a measure of the distribution's asymmetry[2]. For a unimodal continuous r.v.:

- Skew < 0 = "skewed to the left" = "left-tailed" (fat left tail)
- Skew > 0 = "skewed to the right" = "right-tailed" (fat right tail)



[1] The r^{th} moment of X's probability distribution is $\mathbb{E}[X^r]$. Its r^{th} *central* moment is $\mathbb{E}[(X - \mu_X)^r]$. Its r^{th} *standardized* moment is $\mathbb{E}[(X - \mu_X)^r]/\sigma_X^r$.

For X, Y:

Covariance $cov[X,Y] = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Correlation $\rho_{X,Y} = \dfrac{cov[X,Y]}{\sigma_X \sigma_Y} \quad \in [-1, 1]$

Pearson's correlation coefficient is the normalized covariance. The covariance is not easy to interpret as its value depends on the values of the variables. Instead $\rho_{X,Y}$ normalizes the variables, s.t. its magnitude shows the *strength* of the *linear* relation.

⚠ Covariance and correlation are measures of *linear* dependence only. $\rho_{X,Y}$ is the slope of the regression of normalized Y on normalized X, i.e., the strength of the *linear* relation. A correlation of 0 would indicate only that there is no *linear* relationship between the variables. They may have a nonlinear relationship; always check using a scatterplot.

2.3 ... of which we can compute only sample equivalents

Given a sample $\{x_i, y_i\}_i$ of n realizations of the random variables X and Y, we can define sample equivalents of the moments and properties of X and Y's individual and joint probability distributions. It will be important to correct for bias (due to reduced degrees of freedom) when necessary[3]:

	Population parameter	*Bias-adjusted* sample equivalent
Expectation	$\boldsymbol{\mu_X} = \mathbb{E}[X]$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$\boldsymbol{\sigma^2_X} = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$	$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard deviation	σ_X	$s_x = \frac{1}{\sqrt{n-1}}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Covariance	$\boldsymbol{\sigma^2_{X,Y}} = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big]$	$s_{x,y}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

2.4 Robust measures: *median, interquantile range...*

The mean – as a measure of central tendency – and the standard deviation – as a measure of statistical dispersion – are not robust to outliers. They are influenced by the magnitude of values, as a result adding one outlier to our data can change the sample mean and sample standard deviation[4] drastically.

We can use instead measures that are rank-based and therefore *robust* to outliers, for example to measure:

- central tendency: the **median** $\tilde{\boldsymbol{X}}$. It has a breakdown point[5] of 50% (while the mean has a breakdown point of $\frac{1}{n}$: a single large observation can throw it off).
- statistical dispersion

[3]For example: by definition, an *unbiased* estimate of σ^2 is an s^2 s.t. $\mathbb{E}[s^2] = \sigma^2$. We can show that $\mathbb{E}[\sum_{i=1}^{n}(x_i - \bar{x})^2] = ... = (n-1)\sigma^2$. So we must divide the sum by $n-1$, for it to be an *unbiased* estimate of σ^2.
This bias has to do with *degrees of freedom*. Informally: what causes this bias is using the mean calculated from the sample $\bar{x}$ instead of the mean of the population $\mathbb{E}[X]$. (Of course, we don't know the mean of the population, so we are forced to use the sample mean. But we understand that and adjust for it.) Now, why does that bias manifest itself as $n-1$ and not n? Because we lost one degree of freedom when we calculated (then used) the sample mean. Remember that every time you calculate a statistic, you lose a degree of freedom.

[4]In the standard deviation, the distances from the mean are squared, so large deviations are weighted more heavily, and thus outliers can heavily influence it.

[5]The breakdown point of an estimator is the proportion of incorrect observations (e.g. arbitrarily large observations) it can handle before giving an incorrect (e.g., arbitrarily large) result.

- The **interquantile range (IQR)** is the difference between the upper and lower quartiles: $\text{IQR}[X] = Q_3[X] - Q_1[X] = F_X^{-1}(0.75) - F_X^{-1}(0.25)$. It has a breakdown point of 25%
- The **median absolute deviation (MAD)** is the median of the absolute deviations from the median: $\text{MAD}[X] = \text{median}\big[|X_i - \tilde{X}|\big]$

2.5 Standardizing: *z-score*

A standardized variable or "z-score" is a variable that has been rescaled to have mean 0 and standard deviation 1. The standardized z_i measures how many standard deviations from the mean the raw y_i is.

Population: $z = \dfrac{y - \mu}{\sigma_y}$

Sample: $z_i = \dfrac{y_i - \bar{y}}{s_y}$, where s_y = the sample standard deviation

2.6 Independence $\implies$ uncorrelation = orthogonality

Independence and correlation are statistical concepts, whereas orthogonality is a linear algebra concept.

- 2 r.v. X and Y are **independent ($\boldsymbol{X \perp\!\!\!\perp Y}$)** iff $f_{XY}(x, y) = f_X(x) f_Y(y),\ \forall (x, y)$
- 2 r.v. X and Y are **uncorrelated** iff $cov[X, Y] = 0$, i.e., $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- 2 vectors u and v are **orthogonal** iff $< u, v >= 0$
  A space of random variables can be considered a vector space. We can therefore define an inner product in that space, in different ways. One common choice is to define it as the covariance: $< X, Y >:= cov[X, Y]$. 2 r.v. X and Y are therefore **orthogonal** iff $cov[X, Y] = 0$.

independent	$\implies$	uncorrelated, orthogonal
$\shortparallel$		$\shortparallel$
$X \perp\!\!\!\perp Y$		$\text{cov}[X,Y] = 0$

3 Common families of probability distributions

3.1 Discrete

- Bernoulli distribution $Ber(p)$
- binomial distribution $B(n, p)$
- Poisson distribution $Pois(\lambda)$
- negative binomial distribution $NB(r, p)$

3.2 Continuous

- Beta distribution on $[0, 1]$
- chi-squared distribution
- Weibull distribution
- logistic distribution on $[-\infty, \infty]$
- Normal distribution
- Student's t-distribution

4 Theorems (Probability Theory)

This section draws heavily from Colin Cameron's lecture notes "Asymptotic Theory for OLS".

Consider a sequence of real numbers, such as $a_n = 4 + \frac{5}{n}$. a_n has a limit, to which it converges[6] with certainty: $a_n \xrightarrow[n\to\infty]{} 4 \equiv a_\infty$.

Consider now a sequence of random variables b_n, e.g., a stochastic extension of a_n, that also converges. We are also interested in b_n's behavior at its limit; however, as it is stochastic, we can't be *certain* that it will converge to it. But we are *almost certain*. To capture this *almost certainty*, we introduce the concept of *convergence in probability*[7], and write: $\underset{n\to\infty}{\mathbf{p}\lim}\, b_n = b_\infty$ or $b_n \xrightarrow[n\to\infty]{p} b_\infty$.

Let us focus on a particular type of sequence of random variables: a sample average[8]. Consider an infinite number of random variables[9] $X_1, X_2, ...$ For any sample of size n, the sample average is $\bar{X}_n \equiv \frac{1}{n}(X_1 + ... + X_n)$.

We will consider three situations:

a. X_i are **not independent** over i
b. X_i are **independent** over i but not identically distributed: $X_i \sim (\mu_i, \sigma_i^2)$
c. X_i are **iid**: $X_i \overset{\text{iid}}{\sim} (\mu, \sigma^2)$

Two theorems tell us that, and how, $\bar{X}_n$ converges:

- **Laws of Large Numbers** say a sequence of sample averages $\bar{X}_n$ *converges in probability.*
- **Central Limit Theorems** say a sequence of sample averages $\bar{X}_n$ *converges in distribution.*

4.1 Law of Large Numbers (LLN)

Law of Large Numbers (LLN)

a. The average $\bar{X}_n$ of n random variables converges in probability:

$$\bar{X}_n - \mathbb{E}[\bar{X}_n] \xrightarrow[n\to\infty]{p} 0 \quad \Longleftrightarrow \quad \underset{n\to\infty}{\mathbf{p}\lim}\, \bar{X}_n = \lim_{n\to\infty} \mathbb{E}[\bar{X}_n]$$

b. If X_i **independent**, the probability limit simplifies to: $\underset{n\to\infty}{\mathbf{p}\lim}\, \bar{X}_n = \lim_{n\to\infty} \frac{1}{n} \sum_i \mathbb{E}[X_i]$

c. If X_i **iid**, it further simplifies to: $\underset{n\to\infty}{\mathbf{p}\lim}\, \bar{X}_n = \lim_{n\to\infty} \frac{1}{n} n\mu = \mu \quad \Longleftrightarrow \quad \bar{X}_n \xrightarrow[n\to\infty]{p} \mu$

4.2 Central Limit Theorem (CLT)

[6] A series converges to a_∞ if for every arbitrarily small distance $\varepsilon > 0$, there is an N s.t. $\forall n \geqslant N, |a_n - a_\infty| < \varepsilon$. In other words, the series will eventually be within any small distance ε of its limit.

[7] Precisely, a series b_n *converges in probability* to its limit b_∞ iff the probability that the series will be within any small distance ε of its limit, converges to 1 as $n \to \infty$.

[8] We are particularly interested in sample averages because many estimators and tests statistics are made of them. So by knowing how sample averages behave, we'll also know the behavior of estimators and tests statistics as the sample size $n \to \infty$.

[9] We can think of each X_i as the result of a trial; for example, the result of a coin flip. Arbitrarily setting $X_i = 1$ for heads and 0 for tails, the sample average, i.e., the sequence $X_n = \frac{1}{n}(X_1 + ... + X_n)$, represents the proportion of heads in the n coin flips. Intuitively, we know it will converge *most probably* to $\frac{1}{2}$ as $n \to \infty$. I.e., $\underset{n\to\infty}{\text{plim}}\, X_n = \frac{1}{2}$, or $X_n \xrightarrow[n\to\infty]{p} \frac{1}{2}$.

Central Limit Theorem (CLT)

a. The *normalized* average Z_n of n random variables is a random variable that converges to a normal distribution:

$$Z_n \equiv \frac{\bar{X}_n - \mathbb{E}[\bar{X}_n]}{\sqrt{\mathbb{V}[\bar{X}_n]}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

b. If X_i **independent**:

$$\frac{\frac{1}{n}\sum_i \left(X_i - \mathbb{E}[X_i]\right)}{\frac{1}{n}\sqrt{\sum_i \mathbb{V}[X_i]}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

c. If X_i **iid**[a]:

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

It can be convenient to express results in terms of $\bar{X}_n$. We say that *in large samples*[b] $\bar{X}_n$ is *asymptotically normally distributed*:

a. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\lim_{n\to\infty} \mathbb{E}[\bar{X}_n],\ \lim_{n\to\infty} \mathbb{V}[\bar{X}_n]\right)$

b. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\lim_{n\to\infty} \frac{1}{n}\sum_i \mathbb{E}[X_i],\ \lim_{n\to\infty} \frac{1}{n^2}\sum_i \mathbb{V}[X_i]\right)$

c. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

[a] Continuing the previous example where X_i is the probability that the i-th coin flip is heads. If one flips a coin n times, the probability of getting $\frac{n}{2}$ heads will get increasingly close to a normal distribution centered on 0 for higher n.

[b] The term "in large samples" means that n is large enough that it's reasonable to consider the approximation, but not so large that the asymptotic variance goes to zero, making the distribution degenerate.

Remarks:

- By an LLN, $\bar{X}_n$ has a degenerate distribution as it converges to a constant, $\mathbb{E}[\bar{X}_n]$. To apply the CLT, we first scale $\left(\bar{X}_n - \mathbb{E}[\bar{X}_n]\right)$ by its standard deviation $\sqrt{\mathbb{V}[\bar{X}_n]}$ to construct a random variable with variance 1, i.e., with a nondegenerate distribution.
- LLNs and CLTs are widely used in econometrics because extremum estimators involve averages.
  - LLNs give consistency. Ex: we can rewrite $\hat{\beta}_{\text{OLS}}$ to make two averages appear and apply LLNs:

$$\begin{aligned}\hat{\beta}_{\text{OLS}} &= \beta_0 + (X'X)^{-1}X'e \ = \ \beta_0 + \left(\tfrac{1}{n}X'X\right)^{-1}\tfrac{1}{n}X'e \\ &= \beta_0 + \underbrace{\left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1}}_{\xrightarrow[n\to\infty]{p} \text{ finite, } \neq 0} \underbrace{\frac{1}{n}\sum_i x_i e_i}_{\xrightarrow[n\to\infty]{p} 0} \qquad \text{as } \mathbb{E}[x_i e_i] = 0\end{aligned}$$

  - CLTs give limit distributions, after rescaling. Ex: we can center and rescale $\hat{\beta}_{\text{OLS}}$ to apply a CLT:

$$\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta_0) = \underbrace{\left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1}}_{\xrightarrow[n\to\infty]{p} \text{ finite, } \neq 0} \underbrace{\frac{1}{\sqrt{n}}\sum_i x_i e_i}_{\xrightarrow[n\to\infty]{d} \mathcal{N}(0,\ldots)} \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, M_{\text{X'X}}^{-1} M_{\text{X'}\Sigma\text{X}} M_{\text{X'X}}^{-1}\right)$$