Regression Analysis

Last Updated:

August 9, 2023

Contents

Chapter 1

Preliminaries

1.1 Random Variables

We start with a review of basic concepts and definitions from probabaility.

Definition 1.1.1 (Random Experiments). *A* random experiment *is process with a sample space $\mathcal{S}$ for which it is impoosible to predict the outcome with certainty.*

Note that the sample space $\mathcal{S}$ is the set of all possible outcomes from a random experiment.

Definition 1.1.2 (Random Variable). *A random variable X is associated to a random experiment (or process) that is a function which maps the sample space to the real numbers, i.e $X : \mathcal{S} \mapsto \mathbb{R}$. If the set*

$$X(\mathcal{S}) = \{X(s) : s \in \mathcal{S}\}$$

is countable, then X is a discrete random variable, and if it is uncountable then X is a continuous random variable.

Definition 1.1.3 (Probability Functions). *Let $(\Omega, \mathcal{F}, P)$ be a probabaility space. A function $f : \mathbb{R} \to \mathbb{R}$ is called a probability density function if*

1. *$f(x) \geq 0$ for all x in the sample space, and*
2. *$\int_{-\infty}^{\infty} f(x)dx = 1$*

For discrete random variables, we use a probability mass function which follows the same properties with the integral replaced by a sum.

Theorem 1.1.1. *Two random variables X and Y are indepedent if their joint probabaility function $f(x, y)$ is the product of their marginal distributions*

$$f(x, y) = f_X(x)f_Y(y)$$

1.1.1 Expectation, Variance, and Covariance

Definition 1.1.4 (Expected Value)**.** *The expectation (or expected value) of a random variable X is defined as*

$$E(X) = \mu = \int_A xf(x)dx$$

The expectation is the average value, or the value expected to be observed by performing an experiment a very large number of times. We denote the expected value by $E(X) = \mu$ where μ is known as the mean. Note that $\bar{X}$ is the sample mean, which is the observed expected value from a population with mean μ given by

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

where n is the sample size.

Definition 1.1.5 (Variance)**.** *The variance of a random variable X is given by*

$$\operatorname{Var}(X) = \sigma^2 = E\left[(Y - E(Y))^2\right] = E(Y^2) - E(Y)^2$$

Note that the standard deviation σ is the square root of the variance σ^2.

Definition 1.1.6 (Covariance)**.** *The covariance of two random variables X, Y is defined by*

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

Definition 1.1.7 (Correlation)**.** *The correlation of two random variables ρ is defined by*

$$\rho(X, Y) = \frac{Cov(X, Y)}{\operatorname{Var}(X)\operatorname{Var}(Y)}$$

with $\operatorname{Var}(X)\operatorname{Var}(Y) \neq 0$.

When the correlation of two random variables is 0, we say that they are uncorrelated.

1.1.2 Properties

(i) The expectation is a linear operator, i.e

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

(ii) The variance is not linear in the same way,

$$\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$$

$$\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(Y_i, Y_j)$$

(iii) $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ and $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

(iv) If X_i are uncorrelated, then

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} a_i c_i \,\mathrm{Var}(X_i)$$

(v) $\mathrm{Cov}(X, Y) < 0$ if and only if observations of X above it's sample mean $\bar{X}$ tend to accompany corresponding observations of Y below it's sample mean $\bar{Y}$ and vice-versa.

(vi) $\mathrm{Cov}(X, Y) > 0$ if and only if observations of X above $\bar{X}$ tend to accompany corresponding observations of Y above $\bar{Y}$ and vice-versa.

(vii) $\mathrm{Cov}(X, Y) = 0 \implies X$ and Y are uncorrelated

(viii) If X and Y are indepedent, then they are uncorrelated. The opposite is not true however.

(ix) $|\rho(X, Y)| \leq 1$.

(x) $\rho(X, Y)| = 1 \iff X = aY + b$ for some $a, b \in \mathbb{R}$, in otherwords X is a linear combination of Y.

1.1.3 Random Vectors

Definition 1.1.8 (Random Vectors)**.** *If $X_1, \ldots, X_n$ are random variables, then*

$$\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

is a random vector, and the expected value for $\vec{X}$ is

$$E(\vec{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix}$$

Note that X_i's do not necessarily have the same distributions.

Definition 1.1.9 (Variance-Covariance Matrix)**.** *The* variance-covariance *matrix of $\vec{X}$ is the symmetric matrix*

$$\mathrm{Var}(\vec{X}) = [g_{ij}]$$

where the entries g_{ij} are

$$g_{ij} = \begin{cases} \mathrm{Var}(X_i) & i = j \\ \mathrm{Cov}(X_i, X_j) & i \neq j \end{cases}$$

In otherwords, the matrix looks like

$$\text{Var}(\vec{X}) = \begin{pmatrix} \text{Var}(X_1) & \cdots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_1, X_n) & \cdots & \text{Var}(X_n) \end{pmatrix}$$

If $X_1, \ldots, X_n$ are indepedent with the same variance σ^2, then

$$\text{Var}(\vec{X}) = \sigma^2 I_n$$

where I_n is the $n \times n$ identity matrix.

1.1.4 Samples

In practice, we usually work with samples from the random variables, so we do not have the population mean μ and population variance σ^2. Instead, we used unbiased estimators $\bar{X}$ and s^2 to estimate $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$ respectively. Let $\{X, Y\}_{i=1}^n$ be a sequence of random variables, then the sample means are given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \ \bar{Y} = \sum_{i=1}^n Y_i$$

and the sample variance is

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2, \ s_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$$

Note that for the sample variance we divide by $n - 1$ because we have 1 less degree of freedom since to calculate the sample variance, we use the sample mean which also uses the n samples. The sample covariance is given by

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

1.1.5 Important Distributions

Definition 1.1.10 (Cumulative Distribution Functions)**.** *The cumulative distribution function (c.d.f) of a continuous random variable X is defined by*

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

where $f(x)$ is the probabaility density function of X.

Note that

$$f(x) = \frac{d}{dx}F(x)$$

Normal Distribution

We denote a random variable X with normal distribution with mean μ and variance σ^2 as $X \sim N(\mu, \sigma^2)$, with c.d.f

$$F(x) = P(X \leq x) = \Phi(x)$$

and p.d.f

$$f(x) = \Phi'(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

Note that we cannot calculate a formula for the c.d.f but compute values for it by evaluating the integral numerical for various values of $\Phi(x)$.

Chi-Squared Distribution

We say a random variable X has a χ^2 distribution with ν degrees of random $X \sim \chi^2(\nu)$, with a p.d.f

$$f(x;\nu) = \frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} I(y > 0)$$

where Γ is the gamma function. If we have $X_i \sim \chi^2(\nu_i)$, then

$$\sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} \nu_i\right)$$

There is an important relationship between the chi-squared distribution and the normal distrubtion. Let $Z \sim N(0,1)$, that is Z follows a standard normal distribution, then $Z^2 \sim \chi^2(1)$.

t-Distribution

If $Z \sim N(0,1)$, and $U \sim \chi^2(\nu)$ with Z and U indepedent, then

$$T = \frac{Z}{\sqrt{U/\nu}} \sim t(\nu)$$

This is the t-distribution with ν degrees of freedom.

F-Distribution

If $U_1 \sim \chi^2(\nu_1)$, and $U_2 \sim \chi^2(\nu_2)$ are indepedent, then

$$F = \frac{U_1/\nu_1}{U_2/\nu_2} \sim F(\nu_1, \nu_2)$$

which is known as the F-Distribution (or Fisher's distribution) with ν_1, ν_2 degrees of freedom.

Theorem 1.1.2. *Let $X_i, \dots, X_n$ be indepedent normal random variables with mean $\mu_1, \dots, \mu_n$ and variance $\sigma_1^2, \dots \sigma_n^2$, then*

$$X_1 + \cdots + X_n = N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2)$$

Theorem 1.1.3 (Central Limit Theorem)**.** *Let $\{X_i\}$ be a sequence of random variables with mean μ, and variance σ^2 and sample mean $\bar{X}$, then*

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \to N(0, 1)$$

Note that we can represent this in many equivalent ways, such as

$$\sqrt{n}(\bar{X} - \mu) \to \sigma N(0, 1) = N(0, \sigma^2)$$

Theorem 1.1.4. *Let $X_1, \dots, X_n$ be indepedent normal random variables with mean μ and variance σ^2. Let $\bar{X}$ be the sample mean and s^2 be the sample variance, then*

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(n-1)$$

1.2 Multivariate Calculus

Definition 1.2.1 (Gradient)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. if $\vec{X} = (X_1, \dots, X_n)$, then the gradient of f is*

$$\nabla f(\vec{X}) = \begin{pmatrix} \frac{\partial f(\vec{X})}{\partial X_1} \\ \frac{\partial f(\vec{X})}{\partial X_2} \\ \vdots \\ \frac{\partial f(\vec{X})}{\partial X_n} \end{pmatrix}$$

The gradient is a *linear operator*, so

$$\nabla(af + bg)(\vec{X}) = a\nabla f(X) + b\nabla g(X)$$

1.3 Matrix Algebra

Let $A \in M_{m,n}(\mathbb{R})$, and $\vec{X}$ be a random vector. Consider the linear transformation $W = A\vec{X}$. Then

$$E(W) = AE(\vec{X}), \ \operatorname{Var}(W) = A \operatorname{Var}(\vec{X}) A^T$$

Furthermore, if $\vec{X} \sim N(E(\vec{X}), \operatorname{Var}(\vec{X}))$, then

$$W \sim N(E(W), \operatorname{Var}(W)) = N(AE(\vec{X}), A \operatorname{Var}(\vec{X}) A^T)$$

Note that since $\vec{X}$ is a random vector, it is $n \times 1$, and A is $m \times n$, so $A\vec{X}$ is $m \times 1$ and the variance-covariance matrix $\operatorname{Var}(A\vec{X})$ is square $m \times m$.

Definition 1.3.1 (Trace)**.** *If $A \in M_{n,n}(\mathbb{R})$, that is, A is a square matrix, then the* trace *of A is defined as*

$$\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}$$

The trace is a linear operator

$$\operatorname{tr}(kA + B) = k \operatorname{tr}(A) + \operatorname{tr}(B)$$

and we also have that if AB and BA are square,

$$\operatorname{tr}(AB) = \operatorname{tr}(BA)$$

Definition 1.3.2 (Transpose)**.** *The transpose of a matrix A denoted by A^T is obtained by interchanging its rows and its columns. Or in otherwords, simply reflecting the matrix along its primary diagonal.*

Properties

If $A \in M_{n,n}(\mathbb{R})$ and $k \in \mathbb{R}$, then

(i) $(A^T)^T = A$

(ii) $k^T = k$

(iii) $(kA + B)^T = kA^T + B^T$

(iv) $(AB)^T = B^T A^T$

1.4 Quadratic Forms