R300 Econometrics

Metrics Enjoyers

Michaelmas Term, 2023-2024

Contents

1 Basic Probability. Conditional expectation function.

1.1 Random Variables

Definition 1.1.1: Cumulative distribution function

The cumulative distribution function of X is defined as $F_X(x) \equiv P(X \leq x)$. A function F is a cdf iff:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$;
2. $F(\cdot)$ nondecreasing;
3. $F(\cdot)$ right-continuous; i.e., $\forall x_0$, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

Definition 1.1.2: Probability density function

For a continuous r.v., $f_X(x)$ defined as the function which satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$ for all x. A function f_X is a pdf iff:

1. $\forall x$, $f_X(x) \geq 0$;
2. $\int_{\mathbb{R}} f_X(x)\,dx = 1$.

f_X gives the probability of any event: $P(X \in B) = \int_{\mathbb{R}} 1_{(x \in B)} f_X(x)\,dx$.
A continuous (in all dimensions) random vector X has joint pdf $f_X(x_1, \ldots, x_n)$ iff $\forall A \subseteq \mathbb{R}^n$, $P(X \in A) = \int \cdots \int_A f_X(x_1, \ldots, x_n)\,dx_1 \cdots dx_n$.

Exercise 1.1.1. Show that the standard normal density integrates to unity by showing (when $u > 0$):

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}uy^2} dy = \frac{1}{\sqrt{u}}.$$

Solution:-

$$\left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy\right] \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx\right] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} dxdy.$$

By changing to polar coordinates, $x^2 + y^2 = r^2$ and $dxdy = rdrd\theta$. Thus, the desired integral becomes:

$$\frac{1}{2\pi} \int_{0}^{2\pi} \int_{0}^{\infty} e^{-\frac{1}{2}ur^2} rdrd\theta = \frac{1}{u}$$

Setting $u = 1$ yields the desired result.

Definition 1.1.3: τ-th quantile

Let X be a random variable with distribution function F_X. The τ-th quantile of X is defined as the value x_τ such that

$$F_X^{-1}(\tau) = \inf\{x : F_X(x) \geq \tau\}$$

where $0 \leq \tau \leq 1$.

Why inf and not min?
Because F is right-continuous and nondecreasing, the superlevel sets of F are of the form $[a, \infty]$ where $a > -\infty$ or else the entire real line. When the superlevel set is the whole line, there is no min (among the reals), while the inf is -∞. For a = $+\infty$ the superlevel set is empty and so the inf = $+\infty$. These cases can potentially arise when $\tau = 0$ or $\tau = 1$ respectively. *If $\tau \in (0,1)$ then we can replace inf with min.*
If X is discrete, then using minimum and infimum are equivalent, since the support is finite and attains a minimum at some point. However, a contimuous X with infinite support will not achieve a minimum, hence the infimum is needed.

Example. The CDF of an Exponential distribution with parameter λ is given by

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

The quantile function for Exponential(λ) is derived by finding the value of Q for which $1 - e^{-\lambda Q} = p$:

$$Q(p;\lambda) = \frac{-\ln(1-p)}{\lambda},$$

for $0 \leq p < 1$. The quartiles are therefore:

- First quartile ($p = 1/4$): $-\ln(3/4)/\lambda$
- Median ($p = 1/2$): $-\ln(1/2)/\lambda$
- Third quartile ($p = 3/4$): $-\ln(1/4)/\lambda$.

Definition 1.1.4: Expectation

For a function g, the expectation of $g(X)$ is defined as $\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$.

Exercise 1.1.2. Suppose that Y is a continuous random variable with density $f(y)$ that is positive only if $y \geq 0$. If $F(y)$ is the distribution function, show that

$$\mathbb{E}(Y) = \int_0^\infty [1 - F(y)]dy$$

Solution:-

$$E(Y) = \int_0^\infty yf(y)dy = \int_0^\infty \left(\int_0^y dt\right) f(y)dy = \int_0^\infty \left(\int_t^\infty f(y)dy\right) dt$$
$$= \int_0^\infty P(Y > y)dy = \int_0^\infty [1 - F(y)]dy$$

Definition 1.1.5: Moment

For $n \in \mathbb{Z}$, the nth moment of X is $\mu'_n \equiv \mathbb{E}X^n$. Also denote $\mu'_1 = \mathbb{E}X$ as μ. The nth central moment is $\mu_n \equiv \mathbb{E}(X-\mu)^n$.

Two different distributions *can* have all the same moments, but only if the variables have unbounded support sets. Note that $\mathbb{E}X^n$ may not exist (the integral might be infinite), then we say the nth moment does not exist.

Notable moments and properties:

- The first raw moment is the mean, $\mu = \mathbb{E}[X]$
  - $\mathbb{E}[ag_1(X)+bg_2(X)+c] = a\mathbb{E}(g_1(X))+b\mathbb{E}(g_2(X))+c$ (i.e., expectation is a linear operator)
  - The mean is the MSE minimizing predictor for X; i.e., $\min_b \mathbb{E}(X-b)^2 = \mathbb{E}(X-\mathbb{E}X)^2$
  - If $X_1, \ldots, X_n$ mutually independent, then $\mathbb{E}[g_1(X_1)\cdot \cdots \cdot g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdot \cdots \cdot \mathbb{E}[g_n(X_n)]$.
- The second central moment is the variance, $\mathbb{E}[(x-\mu)^2]$
  - $Var(aX+bY) = a^2VarX + b^2VarY + 2abCov(X,Y)$
  - $\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$ (i.e.: residual variance + regression variance)
  - $Var\mathbf{X} \equiv \mathbb{E}[\mathbf{X}\mathbf{X}'] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]'$
  - $Var(\mathbf{X}+\mathbf{Y}) = Var(\mathbf{X}) + Cov(\mathbf{X},\mathbf{Y}) + Cov(\mathbf{X},\mathbf{Y})' + Var(\mathbf{Y})$;
  - $Var(\mathbf{AX}) = \mathbf{A}Var(\mathbf{X})\mathbf{A}'$.
  - $Cov(\mathbf{AX},\mathbf{BY}) = \mathbf{A}Cov(\mathbf{X},\mathbf{Y})\mathbf{B}'$;
  - $Cov(\mathbf{X},\mathbf{Y}) = Cov(\mathbf{Y},\mathbf{X})'$.
- The third central moment is the measure of lopsideness of the distribution. When standardised by the standard deviation it is known as the skewness. Any symmetric distribution will have skewness of 0.
- The fourth central moment is a measure of the heaviness of the tail. When standardised by the standard deviation, it is known as the kurtosis:

$$\text{Kurt}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\mu_2^2}.$$

Example. Find μ'_n for the uniform random variable with $\theta_1 = 0$ and $\theta_2 = \theta$.
By definition,

$$\mu'_n = E(Y^n) = \int_{-\infty}^{\infty} y^n f(y)\,dy = \int_0^\theta y^n \left(\frac{1}{\theta}\right) dy = \frac{y^{n+1}}{\theta(n+1)}\bigg|_0^\theta = \frac{\theta^n}{n+1}.$$

Thus,

$$\mu'_1 = \mu = \frac{\theta}{2}, \quad \mu'_2 = \frac{\theta^2}{3}, \quad \mu'_3 = \frac{\theta^3}{4},$$

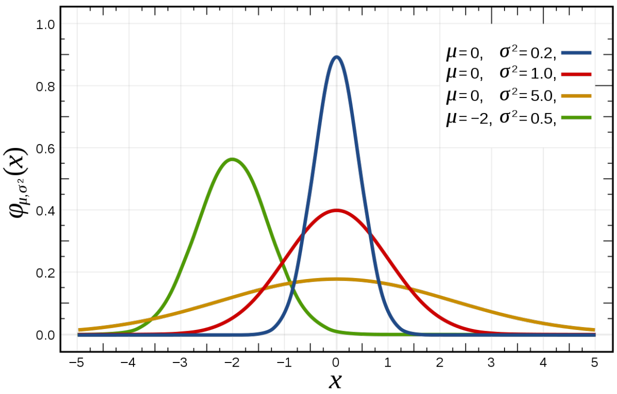
and so on.

1.2 Common Distributions

Normal (Gaussian)

PDF:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- $\mathbb{E}[X] = \mu$
  $\mathbb{E}[(X-\mu)] = 0$
- $\mathbb{E}[X^2] = \mu^2 + \sigma^2$
  $\mathbb{E}[(X-\mu)^2] = \sigma^2$
- $\mathbb{E}[X^3] = \mu^3 + 3\mu\sigma^2$
  $\mathbb{E}[(X-\mu)^3] = 0$
- $\mathbb{E}[X^4] = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$
  $\mathbb{E}[(X-\mu)^4] = 3\sigma^4$

Properties

- The distribution is entirely characterised by the first two moments
- Square of standard normal is χ^2_1.
- If $X \sim N(\mu,\sigma^2)$, $Y \sim N(\gamma,\tau^2)$, and $X \perp\!\!\!\perp Y$, then $X+Y \sim N(\mu+\gamma,\sigma^2+\tau^2)$ (i.e., independent normals are additive in mean and variance).
- For a standard normal: $\mathbb{E}[Z^k] = 0$ if k odd, $\mathbb{E}[Z^k] = 1\cdot 3\cdot 5\cdots(n-1)$ if k even.
- Ratio of independent standard normals is Cauchy ($\sigma = 1$, $\theta = 0$)

Lemma 1.2.1 (Stein's Lemma). If $g(\cdot)$ is differentiable with $\mathbb{E}|g'(X)| < \infty$, then $\mathbb{E}[g(X)(X-\mu)] = \sigma^2\mathbb{E}g'(X)$.

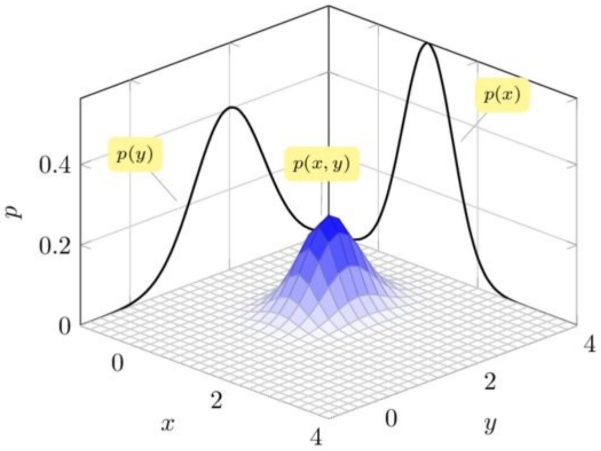
Proof. We shall prove in the case of a standard normal: $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
Since $\int x\exp(-x^2/2)\,dx = -\exp(-x^2/2)$ we get from integration by parts:
$E[g(X)X] = \frac{1}{\sqrt{2\pi}}\int g(x)x\exp(-x^2/2)\,dx = \frac{1}{\sqrt{2\pi}}\int g'(x)\exp(-x^2/2)\,dx = E[g'(X)]$. □

Multivariate Normal

PDF:

$$\frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where $\mu = \mathbb{E}[\mathbf{X}]$ and $\mathbf{\Sigma}_{ij} = Cov(X_i, X_j)$



Bivariate Case

- $\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$
- $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$

Properties

- A linear transformation of a normal is normal: if $\mathbf{X} \sim N_p(\mu, \boldsymbol{\Sigma})$, then for any $\mathbf{A} \in \mathbb{R}^{q\times p}$ with full row rank ($\Rightarrow q \leq p$), and any $\mathbf{b} \in \mathbb{R}^q$, we have $\mathbf{AX} + \mathbf{b} \sim N_q(\mathbf{A}\mu + \mathbf{b}, \mathbf{A\Sigma A'})$. In particular, $\boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \mu) \sim N(\mathbf{0}, \mathbf{I})$.
- The following transformations of $\mathbf{X} \sim N_p(\mu, \boldsymbol{\Sigma})$ are independent iff $\mathbf{A\Sigma B'} = Cov(\mathbf{AX}, \mathbf{BX}) = \mathbf{0}$:
  - $\mathbf{AX} \sim N(\mathbf{A}\mu, \mathbf{A\Sigma A'})$ and $\mathbf{BX} \sim N(\mathbf{B}\mu, \mathbf{B\Sigma B'})$,
  - $\mathbf{AX} \sim N(\mathbf{A}\mu, \mathbf{A\Sigma A'})$ and $\mathbf{X'BX} \sim \chi^2_{rk(\mathbf{B}\Sigma)}$ (where $\mathbf{B\Sigma}$ is an idempotent matrix),
  - $\mathbf{X'AX} \sim \chi^2_{rk(\mathbf{A}\Sigma)}$ and $\mathbf{X'BX} \sim \chi^2_{rk(\mathbf{B}\Sigma)}$ (where $\mathbf{A\Sigma}$ and $\mathbf{B\Sigma}$ are idempotent matrices).
- If X and Y are both normal and independent, this implies they are jointly normally distributed (i.e. (X, Y) is multivariate normal). However, a pair of jointly normal distributed variables need not be independent (would only be of if uncorrelated, $\rho = 0$).
- Independence and zero-covariance are equivalent for linear functions of normally distributed r.v.s.

Example (Individual normality $\not\Rightarrow$ joint normality). Consider $X \sim N(0, 1)$, and:

$$Y = \begin{cases} X, & \text{if } |X| \leq c \\ -X, & \text{if } |X| > c \end{cases} \quad \text{where } c > 0$$

When c is very small, $\text{corr}(X, Y) \approx -1$ and when c is very large, $\text{corr}(X, Y) \approx 1$. If the correlation is a continuous function of c, then there exists some c such that the correlation is 0. X and Y are uncorrelated, but clearly not independent since X completely determines Y. To show Y is normal:

$$\begin{aligned} P(Y \leq x) &= P(|X| < c \text{ and } X \leq x) + P(|X| > c \text{ and } -X \leq x) \\ &= P(|X| < c \text{ and } X \leq x) + P(|X| > c \text{ and } X \geq -x) \\ &= P(X \leq x) \end{aligned}$$

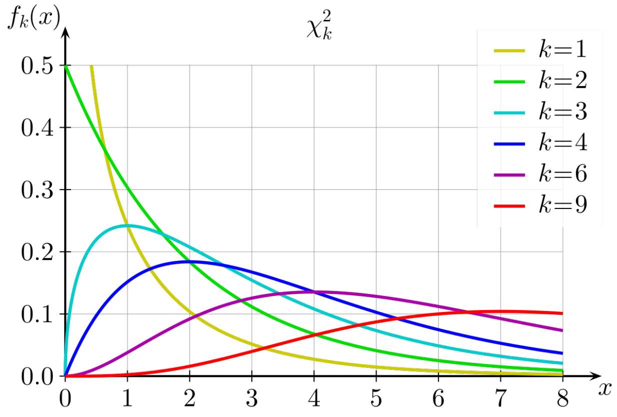
using the symmetry of $|X|$ and $|X| \leq c$. Note that $X - Y$ is not normally distributed due to the non-zero probability of $X - Y = 0$. However, a normal has no discrete part, i.e.: the probability of any point is 0. Thus, X and Y are not jointly normally distributed, even though they are individually normally distributed.

Chi-Squared (χ^2)

PDF:

$$\chi_k^2 = \sum_{i=1}^{k} Z_i^2$$

where $Z_i \overset{\text{iid}}{\sim} N(0,1)$



- $\mathbb{E}[X] = k$
$\mathbb{E}[(X-k)] = 0$
- $\mathbb{E}[X^2] = k(k+2)$
$\mathbb{E}[(X-k)^2] = 2k$
- $\mathbb{E}[X^3] = k(k+2)(k+4)$
$\mathbb{E}[(X-k)^3] = 8k$
- $\mathbb{E}[X^4] = k(k+2)(k+4)(k+6)$
$\mathbb{E}[(X-k)^4] = 12k^2 + 48k$

Properties

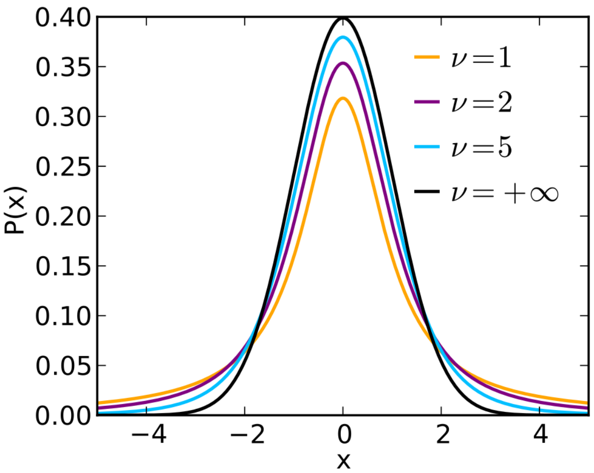
- If $X_1, \ldots, X_n$ are independent with $X_i \sim \chi^2_{p_i}$, then $\sum X_i \sim \chi^2_{\sum p_i}$ (i.e., independent chi squared variables add to a chi squared, and the degrees of freedom add).
- If $\mathbf{X} \sim N_n(\mu, \mathbf{\Sigma})$, then $(\mathbf{X}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu) \sim \chi^2_n$.
- If $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{I})$ and $\mathbf{P}_{n\times n}$ is an idempotent matrix, then $\mathbf{X}'\mathbf{P}\mathbf{X} \sim \chi^2_{rk(\mathbf{P})} = \chi^2_{\text{tr}(\mathbf{P})}$.
- If $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{I})$ then the sum of the squared deviations from the sample mean $\mathbf{X}'\mathbf{M}_\iota\mathbf{X} \sim \chi^2_{n-1}$.

Student's t

PDF:

$$t_\nu = \frac{Z}{\sqrt{X/\nu}} = c\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $Z \sim N(0,1)$, $X \sim \chi^2_\nu$



- Mean: 0 for $\nu > 1$
- Variance: $\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \le 2$
- Skewness: 0 for $\nu > 3$
- Ex. kurtosis: $\frac{6}{\nu-4}$ for $\nu > 4$, ∞ for $2 < \nu \le 4$

Why does the ν-th moment of t_ν not exist?

Consider the ν-th raw moment: $\int x^\nu c\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx \approx \int c\nu^{\frac{\nu+1}{2}} x^{-1} dx$ when x is large. This integral diverges, meaning the ν-th raw moment does not exist. A more rigourous proof requires the use of the Beta and Gamma functions.

Properties

- If $X_1, \dots, X_n$ are iid $N(\mu, \sigma^2)$, then $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$. However, we will generally not know σ. Using the sample variance rather than the true variance gives $\sqrt{n}(\bar{X} - \mu)/s \sim t_{n-1}$.
- If a t distribution has ν degrees of freedom, there are only $\nu - 1$ defined moments. ν has thicker tails than normal.
- t_1 is Cauchy distribution (the ratio of two independent standard normals). t_∞ is standard normal.

Example (Derive variance of Student's t)**.** Consider $X \sim t_\nu$. When $\nu > 1$:

$$E(X) = 0$$

$$(t_\nu)^2 \sim F_{1,\nu} \Rightarrow E(X^2) = E(Y)$$

with $Y \sim F_{1,\nu}$, where $F_{1,\nu}$ is the F-distribution with $(1, \nu)$ degrees of freedom. $E(Y)$ exists if and only if $\nu > 2$:

$$E(Y) = E(X^2) = \frac{\nu}{\nu - 2}$$

We therefore have:

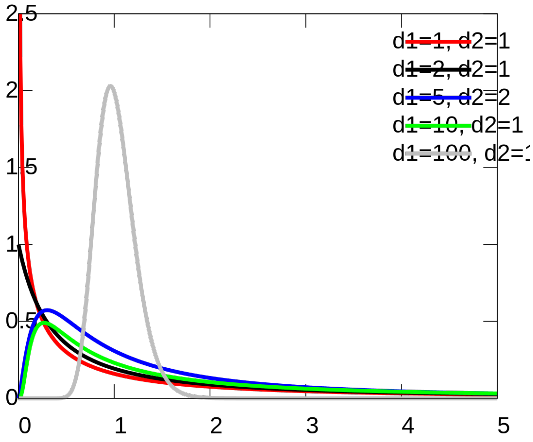
$$\text{var}(X) = E(X^2) - (E(X))^2 = \frac{\nu}{\nu - 2}$$

Snedecor's F

PDF:

$$F_{d_1,d_2} = \frac{X_1/d_1}{X_2/d_2}$$

where $X_1 \sim \chi^2_{d_1}$, $X_2 \sim \chi^2_{d_2}$



- Mean: $\frac{d_2}{d_2 - 1}$ for $d_2 > 2$
- Variance: $\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$, for $d_2 > 4$

Properties

- $1/F_{p,q} \sim F_{q,p}$ (i.e., the reciprocal of an F r.v. is another F with the degrees of freedom switched);
- $(t_q)^2 \sim F_{1,q}$;
- If $X \sim F_{p,q}$ then $Y = \lim_{q \to \infty} pX \sim \chi^2_p$

1.3 Conditional expectation function

Definition 1.3.1: Conditional distribution

Conditional distribution of Y given X is defined as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \text{if } f_X(x) \neq 0$$

Conditional expectation $E(Y|X=x)$ is defined as

$$E(Y|X=x) = \int_y y f_{Y|X}(y|x)dy$$

Often, we will skip $X = x$ having in mind that $E(Y|X)$ is a function of random variable X. Hence, it is itself a random variable.
We can also condition for/on multiple coordinates: e.g., for (X_1, X_2, X_3, X_4) a continuous random vector, $f(x_3, x_4|x_1, x_2) \equiv f(x_1, x_2, x_3, x_4)/f_{X_1X_2}(x_1, x_2)$, where f is a joint pdf, and $f_{X_1X_2}$ is the marginal pdf in X_1 and X_2.

Note:-

Borel Paradox: Be careful when we condition on events of probability zero: two events of probability zero may be equivalent, but the probabilities conditional on the two events is different!

Theorem 1.3.1 (Law of Iterated Expectations). $\mathbb{E}X = \mathbb{E}[\mathbb{E}(X|Y)]$, provided the expectations exist. More generally, when $\mathcal{L} \subseteq \mathcal{M}$ (i.e., $\mathcal{L}$ contains less information, $\mathcal{M}$ contains more),

$$\mathbb{E}[X|\mathcal{L}] = \mathbb{E}[\mathbb{E}(X|\mathcal{M})|\mathcal{L}] = \mathbb{E}[\mathbb{E}(X|\mathcal{L})|\mathcal{M}].$$

Proof.

$$\begin{aligned} E(Y) &= \int_y y f_Y(y)dy = \int_x \int_y y f_{XY}(x,y)dxdy = \int_x \int_y y f_{YX}(x,y)dydx \\ &= \int_x \int_y y f_{Y|X}(y|x) f_X(x) dydx = \int_x E(Y|X=x) f_X(x)dx = E(E(Y|X)). \end{aligned}$$

□

Theorem 1.3.2. $\mathbb{E}(Y|X)$ is the MSE $= E(Y - g(X))^2$ minimising predictor of Y based on knowledge of X.

Proof.

$$\begin{aligned} E(Y - g(X))^2 &= E[Y - E(Y|X) + E(Y|X) - g(X)]^2 \\ &= E[Y - E(Y|X)]^2 + 2E[(Y - E(Y|X))(E(Y|X) - g(X))] + E[E(Y|X) - g(X)]^2 \end{aligned}$$

Using the law of iterated expectaions: $E(Z) = E(E(Z|X))$

$$E[(Y - E(Y|X))(E(Y|X) - g(X))] = E(E[(Y - E(Y|X))(E(Y|X) - g(X))]|X)$$

Bring terms explained fully by X outside expectation

$$= E([E(Y|X) - g(X)]E\{[Y - E(Y|X)]|X\})$$

Expand conditional expectation

$$\begin{aligned}&= E([E(Y|X) - g(X)]\{E(Y|X) - E(Y|X)\})\\ &= 0\end{aligned}$$

Therefore,

$$\begin{aligned}2E[(Y - E(Y|X))(E(Y|X) - g(X))] &= 0 \Rightarrow\\ E(Y - g(X))^2 &= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2\\ &\geq E[Y - E(Y|X)]^2.\end{aligned}$$

and CEF is the best conditional predictor of Y □

Lemma 1.3.1 (Leibniz Rule). Let $f(x,t)$ be a continuously differentiable function then, for the function

$$F(t) = \int_{a(t)}^{b(t)} f(x,t)\, dx$$

the derivative of $F(t)$ with respect to t is given by

$$\frac{dF}{dt} = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}\, dx + f(b(t),t) \cdot \frac{db}{dt} - f(a(t),t) \cdot \frac{da}{dt}$$

Theorem 1.3.3. The conditional median $med(Y|X)$ is the expected absolute error $= E(|Y - g(X)||X = x)$ minimizing predictor of Y based on knowledge of X.

The following proof is a complete version of the outline Alexey presents in the notes. A brief (similar) proof is given at the end.

Proof.

$$\begin{aligned}E(|Y - g(X)||X = x) &= \int_{-\infty}^{\infty} |y - g(x)| f_{Y|X}(y|x) dy\\ &= \int_{g(x)}^{\infty} (y - g(x)) f_{Y|X}(y|x) dy + \int_{-\infty}^{g(x)} (g(x) - y) f_{Y|X}(y|x) dy.\end{aligned}$$

Assume that $f_{Y|X}$ is zero to the left of some constant A, and is unity to the right of some constant B. The problem is:

$$\min_{g(x)} \left\{ \phi = \int_{g(x)}^{A} (y - g(x)) f_{Y|X}(y|x) dy + \int_{-B}^{g(x)} (g(x) - y) f_{Y|X}(y|x) dy \right\}$$

Applying Leibniz rule, we have:

$$\begin{aligned}\frac{d\phi}{dg(x)} &= \int_{A}^{g(x)} (1) f_{Y|X}(y|x) dy + (g(x) - g(x))(1) - (g(x) - A)(0)\\ &+ \int_{g(x)}^{B} (-1) f_{Y|X}(y|x) dy + (B - g(x))(0) - (g(x) - g(x))(1)\end{aligned}$$

> FOC:
>
> $$0 = \int_A^{g(x)} f_{Y|X}(y|x)dy - \int_{g(x)}^B f_{Y|X}(y|x)dy \Rightarrow \int_A^{g(x)} f_{Y|X}(y|x)dy = \int_{g(x)}^B f_{Y|X}(y|x)dy$$
>
> Hence, $g(x)$ must be the value of Y such that $P(Y \leq g(x)|X = x) = P(Y > g(x)|X = x)$.
> That is, $g(x)$ must be the median of the conditional distribution $F_{Y|X}$.
> To verify that we have minimized $E(|Y - g(x)||X = x)$:
>
> $$\frac{d^2\phi}{dg(x)^2} = \frac{\partial}{\partial g(x)} \left(\int_A^{g(x)} f_{Y|X}(y|x)dy - \int_{g(x)}^B f_{Y|X}(y|x)dy \right)$$
>
> $$= \int_A^{g(x)} 0 f_{Y|X}(y|x)dy + 1\left(\frac{dg(x)}{dg(x)}\right) - 1\left(\frac{dA}{dg(x)}\right) - \int_{g(x)}^B 0 f_{Y|X}(y|x)dy + 1\left(\frac{dB}{dg(x)}\right) - 1\left(\frac{dg(x)}{dg(x)}\right)$$
>
> $= [0 + 1 - 0] - [0 + 0 - 1] = 2(> 0)$ so we are characterising a minimum. □

Also, note that if we let $A \to -\infty$ and $B \to \infty$, the support of F can be taken to be the whole real line, so there is no loss of generality in establishing the above result with a support of $[A, B]$.

Alternative Proof

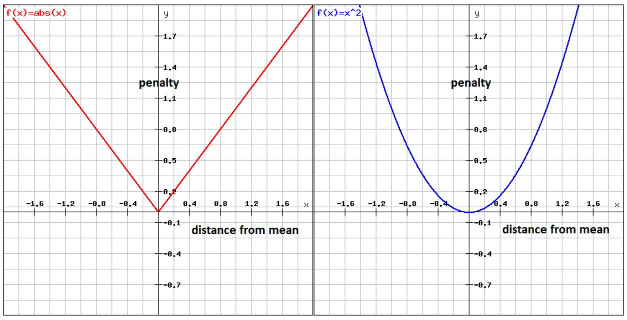
$$\frac{d}{dc}E(|X - c|) = E\left(\frac{d}{dc}|X - c|\right) = E\left(\frac{-(X - c)}{|X - c|}\right)$$

$$= E\left[1_{\{X<c\}} - 1_{\{X>c\}}\right] = P(X < c) - P(X \geq c)$$

$$\frac{d}{dc}E(|X - c|) = 0 \Rightarrow P(X < c) = P(X > c) = \frac{1}{2}$$

By definition of the median, $c = med(X)$ □

MAE vs MSE



- MAE $= E|Y - g(X)|$
- MSE $= E(Y - g(X))^2$

- MAE imposes a linear penalty on errors, i.e.: each deviation from the mean is given a proportional corresponding error.
- MSE is a squared proportional relationship between deviation and penalty. This will make sure that the further you are away from the mean, the proportionally more you will be penalized. Using this penalty function, outliers are deemed proportionally more informative than observations near the mean.

Because the MAE is a more robust estimator of scale than the sample variance or standard deviation, it works better with distributions without a mean or variance, such as the Cauchy distribution.

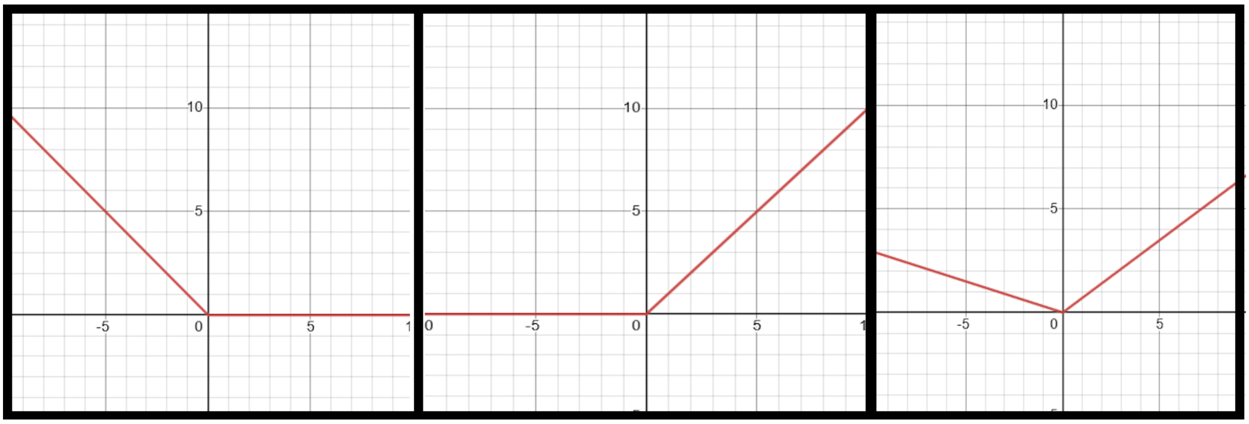
Weighted MAE

If underprediction is marginally less or more costly as overprediction, it makes sense to minimize the expectation of

$$\tau 1(Y > g(X))(Y - g(X)) + (1 - \tau)1(Y \leq g(X))(g(X) - Y)$$

with $\tau \in (0,1)$. For example, parameter $\tau < 1/2$ would correspond to situations where the underprediction is less costly than overprediction. Following the same logic as above, we can show that *the corresponding best predictor would be τ-th quantile $\tau(X)$ of the conditional distribution of Y given X.*

Below we have (from left to right): $\tau = 1$ (no cost to overprediction), $\tau = 0$ (no cost to underprediction) and $\tau = 0.3$ (cost to both, but relatively more to overprediction.)



2 Causal interpretation of regression. Least Squares.

2.1 Regression and Causality

A variable x_1 can be said to have a causal effect on the response variable y if the latter changes when all other inputs are held constant. We can write a full model for the response variable y as:

$$y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$$

where x_1 and $\mathbf{x_2}$ are the observed variables, $\boldsymbol{\varepsilon}$ is an $\ell \times 1$ unobserved random factor and h is a functional relationship.

Definition 2.1.1: Causal effect

In the model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ the **causal effect** of x_1 on y is

$$C(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon}) = \nabla_1 h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon}),$$

the change in y due to a change in x_1, holding $\mathbf{x_2}$ and $\boldsymbol{\varepsilon}$ constant.

Note:-

This is just a definition, and does not necessarily describe causality in a fundamental or experimental sense. It might be more appropriate to label this a structural effect (the effect within the structural model).

Example. Suppose firms have Cobb-Douglas production functions:

$$y = AK^{\alpha}L^{\beta}$$

where K, L are observed capital and labour, A is an unobserved production technology and y is output. Here $x_1 = K, x_2 = L, \varepsilon = A$. Then the causal effect of capital on output is

$$C(K, L, A) = y'(K, L, A) = \alpha A K^{\alpha-1} L^{\beta}.$$

Even for firms with identical inputs, this effect differs due to unobserved A.

Sometimes it is useful to write this relationship as a potential outcomes function

$$y(x_1) = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$$

where the notation implies that $y(x_1)$ is holding $\mathbf{x_2}$ and $\boldsymbol{\varepsilon}$ constant. A popular example arises in the analysis of treatment effects with a binary regressor x_1. Let $x_1 = 1$ indicate treatment (e.g., a medical procedure) and $x_1 = 0$ indicate non-treatment. In this case $y(x_1)$ can be written

$$y(0) = h(0, x_2, \boldsymbol{\varepsilon}), \;\; y(1) = h(1, x_2, \boldsymbol{\varepsilon})$$

where $y(0)$ and $y(1)$ are known as the latent outcomes associated with non-treatment and treatment, respectively. The causal effect of treatment for the individual is the change in their health outcome due to treatment; the change in y as we hold both x_2 and $\boldsymbol{\varepsilon}$ constant:

$$C(x_2, \boldsymbol{\varepsilon}) = y(1) - y(0).$$

This is random as both potential outcomes $y(0)$ and $y(1)$ are different across individuals.

Example. Suppose there are two individals Yinfeng and Charles, and both have the possiblity of being a PhD graduate or dropping out. Suppose Yinfeng would earn £8/hour without a PhD and £12/hour as a PhD grad, while Charles would earn £20/hour without and £30/hour with a PhD. The causal effect of a PhD on wages is £4/hour for Yinfeng and £10/hour for Charles.

In a sample, we cannot observe both outcomes from the same individual, we only observe the realised value. As the causal effect varies across individuals and is not observable, it cannot be measured on the individual level. We therefore focus on aggregate causal effects, in particular what is known as the average causal effect.

Definition 2.1.2: Average causal effect

In the model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ the **average causal effect** of x_1 on y conditional on $\mathbf{x_2}$ is

$$\begin{aligned} ACE(x_1, \mathbf{x_2}) &= \mathbb{E}(C(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \,|\, x_1, \boldsymbol{x_2}) \\ &= \int_{\mathbb{R}^\ell} \nabla_1 h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \end{aligned}$$

where $f(\boldsymbol{\varepsilon} \,|\, x_!, \boldsymbol{x_2})$ is the conditional density of $\boldsymbol{\varepsilon}$ given $x_1, \boldsymbol{x_2}$.

Example. In the Cobb-Douglas example, the ACE of capital on output will be:

$$ACE(K, L) = \mathbb{E}(\alpha A K^{\alpha-1} L^{\beta} | K, L) = \alpha \mathbb{E}(A|K, L) K^{\alpha-1} L^{\beta}$$

Example. Considering again Yinfeng and Charles, suppose half our population are Yinfeng's and the other half Charles's, then the average causal effect of a PhD is $(10 + 4)/2 =$£7/hour. This is not the individual causal effect, it is the average of the causal effect across all individuals in the population.

We can think of $ACE(x_1, \mathbf{x_2})$ as the average effect in the general population. When we conduct regression analysis we might hope that regression reveals the ACE, i.e.: what is the relationship between $ACE(x_1, \mathbf{x_2})$ and the regression derivative $\nabla_1 m(x_1, \mathbf{x_2})$? The model $h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ implies that the CEF is

$$\begin{aligned} m(x_1, \boldsymbol{x_2}) &= \mathbb{E}(h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \,|\, x_1, \boldsymbol{x_2}) \\ &= \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon}, \end{aligned}$$

the average causal equation, averaged over the conditional distribution of the unobserved component $\boldsymbol{\varepsilon}$.

Applying the marginal effect operator [1], the regression derivative is:

$$\begin{aligned} \nabla_1 m(x_1, \boldsymbol{x_2}) &= \int_{\mathbb{R}^\ell} \nabla_1 h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} + \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \\ &= ACE(x_1, \boldsymbol{x_2}) + \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \end{aligned}$$

In general we see that the regression dervative does not equal the average causal effect. They are only equal in the special case when the second term equals zero, which occurs when the conditional

[1] Alexei uses $\frac{\partial}{\partial x_1}$ throughout, this is equivalent to the marginal effect operator used here with continuous x_1.

density of $\boldsymbol{\varepsilon}$ given $(x_1, \boldsymbol{x_2})$ does not depend on x_1 ($\nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) = 0$). When this condition holds then the regression derivative equals the ACE, which means that regression analysis can be interpreted causally, in the sense that it uncovers average causal effects.

Definition 2.1.3: Condiional Independence Assumption (CIA)

Conditional on $\mathbf{x_2}$, the random variables x_1 and $\boldsymbol{\varepsilon}$ are statistically independent.

The CIA implies $f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) = f(\boldsymbol{\varepsilon} \,|\, \boldsymbol{x_2})$ does not depend on x_1, and thus $\nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) = 0$. Thus the CIA implies that the regression derivative equals the ACE.

Theorem 2.1.1. In the structural model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$, the CIA implies

$$\nabla_1 m(x_1, \boldsymbol{x_2}) = ACE(x_1, \boldsymbol{x_2})$$

the regression derivative equals the average causal effect for x_1 on y conditional on $\mathbf{x_2}$.

Example (Nerlove: Returns to scale in electriciy supply)**.** Nerlove investigated returns to scale in a regulated industry (U.S. electricity) using Cobb-Douglas production. The market had the following features:

1. Privately owned local monopolies supply electricity on demand
2. These local monopolies face competitive factor prices
3. Electricity prices are set by the government

Notably Y is exogenously given (by consumer demand). Nerlove assumes firms pick K, L to minmise the cost of producing $Y = AK^\alpha L^\beta$, i.e. K, L both depend on A, Y, in particular $f(A|K, L)$ depends on K. Thus a regression of Y on K, L will not identify the ACE.

$$\min_{K,L} p_K K + p_L L \;\; s.t. \;\; Y = AK^\alpha L^\beta$$

The Lagrangian and FOCs for this problem are:

$$\mathcal{L} = p_K K + p_L L + \lambda(Y - AK^\alpha L^\beta)$$

$$\frac{\partial \mathcal{L}}{\partial K} = p_K - \lambda \alpha A K^{\alpha-1} L^\beta = 0, \quad \frac{\partial \mathcal{L}}{\partial L} = p_L - \lambda \beta A K^\alpha L^{\beta-1} = 0$$

$$\Rightarrow K = \frac{\alpha p_L}{\beta p_K} L$$

We can subsititute this into the production function to solve for L and K, giving:

$$TC = p_K \left(\frac{\alpha p_L}{\beta p_K} \left(\frac{Y}{A \left(\frac{\alpha p_L}{\beta p_K} \right)^\alpha} \right)^{\frac{1}{\alpha+\beta}} \right) + p_L \left(\left(\frac{Y}{A \left(\frac{\alpha p_L}{\beta p_K} \right)^\alpha} \right)^{\frac{1}{\alpha+\beta}} \right)$$

$$TC = p_L \left(\frac{Y \left(\frac{p_L \alpha}{p_K \beta} \right)^{-\alpha}}{A} \right)^{\frac{1}{r}} \left(\frac{r}{\beta} \right) = r \alpha^{-\alpha/r} \beta^{-\beta/r} A^{-1/r} Y^{1/r} p_K^{\alpha/r} p_L^{\beta/r}$$

Taking logs we obtain the following log-linear relationship for each firm:

$$\log(TC_i) = \mu_i + \frac{1}{r}\log(Y_i) + \frac{\alpha}{r}\log(p_K) + \frac{\beta}{r}\log(p_L)$$

where $\mu_i = \log[r(A_i\alpha^\alpha\beta^\beta)^{-\frac{1}{r}}]$. Coefficients in this equation are elasticities, for example $\frac{\beta}{r}$ is the elasticity of total cost with respect to the wage rate, i.e.: the percentage change in in total cost when the wage rate changes by 1%. The degree of returns to scale (the reciprocal of the output elasticity of total costs), is independent of the level of output.
To estimate this define $\mu \equiv \mathbb{E}[\mu_i]$, $\varepsilon_i \equiv \mu - \mu_i$ so $\mathbb{E}[\varepsilon_i] = 0$, firms with positive ε_i are high-cost firms.

$$\log(TC_i) = \beta_0 + \beta_1 log(Y_i) + \beta_2\log(p_K) + \beta_3\log(p_L),$$

where

$$\beta_0 = \mu, \beta_1 = \frac{1}{r}, \beta_2 = \frac{\alpha}{r}, \beta_3 = \frac{\beta}{r}$$

This equation is overidentified, the 4 coefficients are not free parameters, they are a function of three technology parameters (α, β, μ). Clearly $\beta_2 + \beta_3 = 1$ (as expected, cost function is linearly homogenous in factor prices). To fix this we can subtract p_L from each side and consider relative prices:

$$\log\left(\frac{TC_i}{p_L}\right) = \beta_0 + \beta_1 log(Y_i) + \beta_2\log\left(\frac{p_K}{p_L}\right)$$

To test constant returns to scale $(r = 1)$, just t-test $\beta_1 = 1$ in this restricted model.

2.2 Estimating population regression by least squares

If CIA holds, regression captures the causal effect of x's on y. However even if it doesn't, it still provides the best predictor of y given x's. We assume the regression function $\mathbb{E}[y|x] = m(x)$ is parametrised by a finite dimensional vector $\beta = [\beta_1, ..., \beta_k]^T$, so that estimating the population regression $m(x;\beta)$ is equivalent to estimating β. One approach to estimation is using the analogy principle.

Definition 2.2.1: Analogy principle

Consider finding an estimator that satisfies the same properties in the sample that the parameter satisfies in the population; i.e., seek to estimate $\beta(P)$ with $\beta(P_n)$ where P_n is the empirical distribution which puts mass $\frac{1}{n}$ at each sample point. Note this distribution converges uniformly to P.

In the regression context:

$$\beta = \arg\min_b \mathbb{E}(y - m(x;b))^2$$

The sample analgoue of expectation is the average:

$$\hat{\beta} = \arg\min_b \frac{1}{n}\sum_{i=1}^{n}(y - m(x;b))^2$$

When $m(x;b)$ is linear in b, the method is called OLS. We assume that the pbservations of the data (y_i, x_i) are independent and come from the same joint distribution. Let

$$\underbrace{X_i}_{K\times 1} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{iK} \end{bmatrix}, \underbrace{\beta}_{K\times 1} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix},$$

$$\underbrace{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \underbrace{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \underbrace{X}_{n\times K} = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix}.$$

When our model contains a constant, one of the columns of X will contain only ones. Our linear model can thus be represented as:

$$Y = X\beta + \varepsilon$$

When estimating we select the $\hat{\beta}$ such that the sum of squared residuals ($e'e$) is minimised[1].

$$\begin{aligned} e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= (y' - \hat{\beta}'X')(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

Where $y'X\hat{\beta} = (y'X\hat{\beta})' = \hat{\beta}'X'y$ since the transpose of a scalar is itself.

Note:-

Matrix differentiation

$$\frac{\partial \mathbf{a}'\mathbf{b}}{\partial \mathbf{b}} = \frac{\partial \mathbf{b}'\mathbf{a}}{\partial \mathbf{b}} = \mathbf{a} \tag{2.1}$$

when $\mathbf{a}$ and $\mathbf{b}$ are $K \times 1$ vectors.

$$\frac{\partial \mathbf{b}'\mathbf{A}\mathbf{b}}{\partial \mathbf{b}} = 2\mathbf{A}\mathbf{b} = 2\mathbf{A}'\mathbf{b} \tag{2.2}$$

when $\mathbf{A}$ is any symmetric matrix.

$$\frac{\partial 2\mathbf{b}'\mathbf{X}'\mathbf{y}}{\partial \mathbf{b}} = \frac{\partial 2\mathbf{b}'(\mathbf{X}'\mathbf{y})}{\partial \mathbf{b}} = 2\mathbf{X}'\mathbf{y} \tag{2.3}$$

and

$$\frac{\partial \mathbf{b}'\mathbf{X}'\mathbf{X}\beta}{\partial \mathbf{b}} = \frac{\partial 2\mathbf{A}\beta}{\partial \mathbf{b}} = 2\mathbf{A}\beta = 2\mathbf{X}'\mathbf{X}\beta \tag{2.4}$$

when $\mathbf{X}'\mathbf{X}$ is a $K \times K$ matrix.

Solving for the minimum:

$$\begin{aligned} \frac{\partial e'e}{\partial \hat{\beta}} &= -2X'y + 2X'X\hat{\beta} = 0 \\ &\Rightarrow X'X\hat{\beta} = X'y \\ &\Rightarrow (X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y \\ &\Rightarrow I_K\hat{\beta} = (X'X)^{-1}X'y \\ &\Rightarrow \hat{\beta} = (X'X)^{-1}X'y \end{aligned}$$

Here we have assumed that the inverse of $X'X$ exists, i.e. X is full rank[2]. To check this is a minimum, take second derivative which gives us $2X'X$ which is clearly positive semi-definite (when X is full rank). Note that $X'X$ is always square ($k \times k$) and always symmetric.

[1] Note that $e \neq \varepsilon$, residuals e are observed, whilst disturbances ε are unobserved.

[2] The inverse of $X'X$ may not exist, it does not exist in the following two cases: 1) When $n < k$; we have more independent variables than observations 2) One or more of the independent variables are a linear combination of the other variables i.e. perfect multicollinearity.

We can further show that $X'e = 0$, consider the normal form equations $X'X\hat{\beta} = X'y$:

$$
\begin{aligned}
(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= \mathbf{X}'(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}) \\
(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{e} \\
\mathbf{X}'\mathbf{e} &= \mathbf{0}
\end{aligned}
$$

Proposition 2.2.1 (Properties of OLS)**.** From $X'e$=0 we can derive a number of properties.

1. The observed values of X are uncorrelated with the residuals.
2. The sum of the residuals is zero.
3. The sample mean of the residuals is zero.
4. The regression hyperplane passes through the sample means of observables.
5. The predicted values of y are uncorrelated with the residuals.

Where 2-5 hold when the regression includes a constant term.

Proof. Using $X'e = 0$

1. $\mathbf{X}'\mathbf{e} = 0$ implies that for every column $\mathbf{x}_k$ of $\mathbf{X}$, $\mathbf{x}_k'\mathbf{e} = 0$. In other words, each regressor has zero sample correlation with the residuals. Note that this does not mean that $\mathbf{X}$ is uncorrelated with the disturbances; we'll have to assume this.
2. If there is a constant, then the first column in $\mathbf{X}$ (i.e. $\mathbf{X}_1$) will be a column of ones. This means that for the first element in the $\mathbf{X}'\mathbf{e}$ vector (i.e. $\mathbf{X}_{11}e_1 + \mathbf{X}_{12}e_2 + \ldots + \mathbf{X}_{1n}e_n$), to be zero, it must be the case that $\sum_i e_i = 0$.
3. This follows straightforwardly from the previous property i.e. $\bar{e} = \frac{\sum e_i}{n} = 0$.
4. This follows from the fact that $\bar{e} = 0$. Recall that $e = y - \mathbf{X}\hat{\boldsymbol{\beta}}$. Dividing by the number of observations, we get $\bar{e} = \bar{y} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}} = 0$. This implies that $\bar{y} = \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}$.
5. $\hat{y}'e = (\mathbf{X}\hat{\boldsymbol{\beta}})'e = \hat{\boldsymbol{\beta}}'\mathbf{X}'e = 0$

□

Theorem 2.0.1. This is a theorem.

Proof. This is a proof. □

Example. This is an example.

Proof. This is an explanation. □

Claim 2.0.1. This is a claim.

Corollary 2.0.1. This is a corollary.

Proposition 2.0.1. This is a proposition.

Lemma 2.0.1. This is a lemma.

Question 1

This is a question.

Solution:-

This is a solution.

Question 2

This is another question.

Solution:-

This is another solution.

Exercise 2.0.1. This is an exercise.

Definition 2.0.1: Test

This is a definition.

Note:-

This is a note.