MTH 686

Non-linear Regression.

What are we expecting from this course

(i) Beyond Linear Regression (Linear Models)

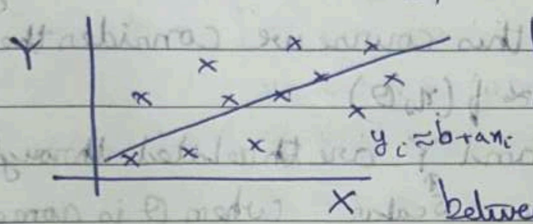
What is simple linear Regression?

There are 2 variables (X, Y) $\{(x_1, y_1), \dots, (x_n, y_n)\}$

Suppose the two variables are linearly related,

What is the relation b/w them?

I want to find best a & b.



We try to find the "best" straight line which explains the relation between the two.

Multiple Linear Regression $(y_1, x_{11}, \dots, x_{1p})$

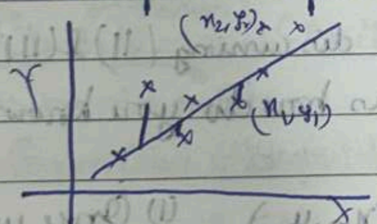
$(Y, \underbrace{X_1, X_2, \dots, X_p})$ $\quad (y_n, x_{n1}, \dots, x_{np})$

↑ independent variables

dependent variable.

$$y_i \approx b + a_1 x_{i1} + \dots + a_p x_{ip}$$

I want to find best possible $b, a_1, \dots, a_p$



$$\sum_{i=1}^{n} (y_i - a - b x_i)^2$$

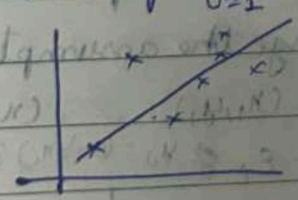
I want to minimize wrt a, b

"Least Squares Method" (i) Why Squares??

(ii) Why not some other distance??

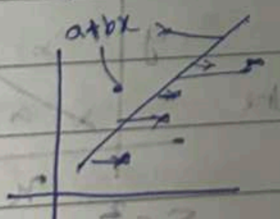
Instead of $\sum_{i=1}^{n} (y_i - a - b x_i)^2$, why not $\sum_{i=1}^{n} |y_i - a - b x_i|$.

Least absolute deviation estimators



$$\sum \rho(y_i - a - b x_i)$$

$$\rho(z) \geq 0 \quad \forall z \quad \rho(0) = 0$$



Total least squares

Beyond linear Regression

$y \approx a + bx \rightarrow$ Linear Regression

$y \approx f(x;\theta) \rightarrow$ Non linear Regression

Now $f()$ can be known or unknown.

(I) f is unknown, it is called nonparametric regression

(II) f is known it is called non linear regression

- Remember in this course we consider the following issues (1) $y \approx f(x,\theta)$

Suppose X and Y are related through $y \approx f(x;\theta)$
$\underset{\text{vector}}{X}$ $\underset{\text{Scalar}}{Y}$ when θ is some parameter (it can be a vector also)

(I) f is unknown [Non-parametric regression]

(II) f is known, θ is unknown & f is linear $\rightarrow$ Linear Regression

(III) f is known, θ is unknown & f is non linear $\rightarrow$ Non-linear Regression

We will be mainly discussing (II) & (III)

Natural Question is how do you know the functional form "f" ??

$(x_1, y_1), \ldots \ldots, (x_n, y_n)$

$y \approx f(x;\theta)$

(I) Once we make the assumption there should be a way to verify it!

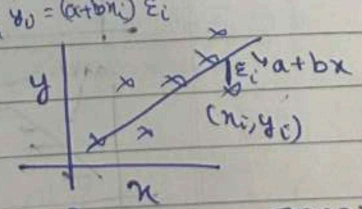
After verification

If we are satisfied with the assumption it is fine, otherwise we have to change the assumption.

Simple linear regression $(x_1,y_1), \ldots, (x_n,y_n)$

$y \approx a + bx \quad y_i = a + bx_i + \varepsilon_i$ or $y_i = (a+bx_i)\varepsilon_i$

You estimate a & b based on certain assumptions on ε_i & also the relation between x_i & y_i & then verify the assumptions.



$\sum \varepsilon_i^2 \quad \sum |\varepsilon_i| \quad \sum \rho(\varepsilon_i)$

In case of Linear Regression the methods are more or less same for a large class of problems. The same is ~~true~~ not true in case of non linear regression

$$y \approx f(x:\theta) \quad f \text{ is non-linear}$$

Remember f is non linear in the parameter θ and NOT in x.

For example $y_i \approx a + bx_i$ — linear

$y_i \approx a + b\ln x_i$ — linear

$y_i \approx a + e^{bx_i}$ — Non-linear $\quad i=1,\dots,n$

In this course we assume that $y_i = f(x_i;\theta) + \varepsilon_i$

Here f is known, θ is unknown, the additive error ε_i follows certain assumptions. f can be non linear in parameters. Our aim is to estimate the unknown parameter θ.

$\mathbb{R}^d$ $(\underset{\uparrow \text{ Vector valued}}{x_1}, \overset{\text{Scalar}}{y_1}), \dots, (x_n, y_n)$ $\quad \mathbb{R}^1$

(i) Different method of estimation

(ii) Numerical issues

(iii) Properties of these estimate

(iv) Geometric interpretations

Linear Regression Model.

We assume we have the data of the form

$$\{(\underset{\text{output}}{y_i}, \underbrace{x_{i1}, x_{i2}, \dots, x_{ip}}_{\text{input}}), i=1,\dots,n\}$$

Assume $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \quad i=1,\dots,n$

Non linear Regression

$y_i = f(x_i, \theta) + \varepsilon_i$. Here f is known maybe linear or non linear.

θ is unknown, ε_i's are errors

If $f(x_i, \theta)$ is linear then it is a linear regression.
The problem here we have data (observation)
$\{(x_1, y_1), \ldots, (x_n, y_n)\}$ We want to estimate θ.
($x_n \rightarrow$ vector)

Applications:

$y_i \rightarrow$ price of house $\quad x_{i1} \rightarrow$ locations $\; x_{i2} \rightarrow$ age
$x_{i3} \rightarrow$ area

Based on the data once you estimate the parameter θ, you can use it for the prediction.

Medical data

$y_i \rightarrow$ Fasting Blood Sugar level $\quad x_{i1} \rightarrow$ amount of exercise
$x_{i2} \rightarrow$ Weight $\; x_{i3} \rightarrow$ BMI

Error term

$y_i = f(x_i, \theta) + \varepsilon_i \rightarrow$ Additive error

log

$y_i = (f(x_i, \theta))\, \varepsilon_i \rightarrow$ Multiplicative error

To do "analysis" we need to make certain assumptions
(i) On f. (ii) On ε_i $\quad$ Minimum assumptions we need: $f: f(x_i, \theta) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
$\hookrightarrow$ Linear
ε_i: Random variable

Minimum assumption on ε_i
(1) We want $E(\varepsilon_i) = 0$ wlog
(Think, What if $E(\varepsilon_i) = \mu \leftarrow$ unknown) (We can't)

Just based on (1) can I estimate (reasonable way) $\beta_0, \beta_1, \ldots, \beta_p$. Yes we can. It may not have "nice" properties. We make the following assumptions
$Var(\varepsilon_i) = C$, & ε_i's are independent.

Model: $y_i = f(x_i, \theta) + \varepsilon_i$. f is known, θ is unknown, ε_i's are random variables. $E(\varepsilon_i) = 0$, ε_i's are independent, $Var(\varepsilon_i) = C$.

(i) First problem we want to solve, Estimate $\theta \cdot (\beta_1, \dots, \beta_p)$
(ii) Properties of the estimators.
Properties
Confidence intervals of estimators, Standard error
How close is the estimator w.r.t the true parameter?
Linear Model
$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$
We can always write in matrix form $Y = X\beta + \varepsilon$.
$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ $X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$ $\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$ $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

Linear Regression
$\underset{\text{observed output}}{y_i} = \beta_0 + \beta_1 \underset{\text{independent variable input}}{x_{i1}} + \dots + \beta_p x_{ip} + \underset{\text{random}}{\varepsilon_i}$

Minimum assumption we make on ε_i: $E(\varepsilon_i) = 0$
$\underset{n\times 1}{Y} = \underset{n\times p}{X}\,\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$ — (*) $V(\varepsilon_i) = c$ (some constant)
$Y = (Y_1, \dots, Y_n)^T$ $X = ((x_{ij}))$
$\beta = (\beta_1, \dots, \beta_p)^T$ $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$
$Y, X \to$ known, we want to estimate β
It is a well studied problem when $n >> p$ & when p is known
Recently the following two problems become quite important
(1) If p is unknown?? $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$
(2) In $n \not> p$ (3) If $\text{Rank}(X) < p$
$y_i = f(\underbrace{x_{i1}, \dots, x_{ip}}_{\text{independent variables}}, \beta) + \varepsilon_i$

In (*) another standard assumption is

$$\text{Rank}(X) = p < n$$

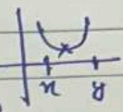
Based on the standard assumptions the most popular estimator of β is the least squares estimators (LSE)

$$Q(\beta) = \sum (y_i - \beta_0 - \beta_1 x_{i1} \cdots - \beta_p x_{ip})^2 \quad (**)$$

The LSE of β can be obtained as the argument minimum of $Q(\beta)$

$$Q(\beta) = (Y - X\beta)^T (Y - X\beta) \quad (**)$$

Note that (**) is a convex function



If $x, y \in \mathbb{R}$
$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$

~~If f is thrice differentiable & $f''(x) > 0 \ \forall x$~~

Note that if f is convex then f has unique minimum. A function f is convex if $f''(x) > 0 \ \forall x \in \mathbb{R}$ (if f'' exists)

The same result holds for $p > 1$ i.e. if $\frac{\partial^2 f(x)}{\partial x \partial x^T}$ exists and it is positive a ~~pos~~ definite matrix, then f is convex.

$x = (x_1, \ldots, x_p)^T \quad f: \mathbb{R}^p \to \mathbb{R} \quad \frac{\partial^2 f}{\partial x \partial x^T} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_p} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_p \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_p^2} \end{bmatrix}$

$$Q(\beta) = (Y - X\beta)^T (Y - X\beta)$$

$$= \underset{1\times n \ n\times 1}{Y^T Y} - 2\underset{1\times p \ p\times n \ n\times 1}{\beta^T X^T Y} + \underset{1\times p \ p\times p \ p\times 1}{\beta^T (X^T X) \beta}$$

$\underset{(p \times 1)}{\frac{\partial Q(\beta)}{\partial \beta}} \quad \underset{p\times 1}{\frac{\partial Q(\beta)}{\partial \beta}} = -2X^T Y + 2(X^T X)\beta$

$$\underset{p\times p}{\frac{\partial^2 Q(\beta)}{\partial \beta \partial \beta^T}} = \underset{p\times p}{2(X^T X)}$$

$\text{Rank}(X) = \text{Rank}(X^T X) = p$

Since $\underset{p\times p}{\text{Rank}(X^T X)} = p$ therefore $2(X^T X) > 0$ (positive definite matrix)

Therefore, the function $Q(\beta)$ has unique minimum.
Question is how to find the minimum ??
We will get it by solving $\frac{\partial Q(\beta)}{\partial \beta} = 0$

$$\frac{\partial Q(\beta)}{\partial \beta} = -2X^TY + 2(X^TX)\beta = 0$$

$$\Rightarrow (X^TX)\beta = X^TY \Rightarrow \beta = (X^TX)^{-1}X^TY$$

Under the standard assumptions LSE of β exists & it is unique

Linear Regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$ We have standard assumptions

$E(\varepsilon_i) = 0 \quad V(\varepsilon_i) = c > 0$

$R(X) = p < n$

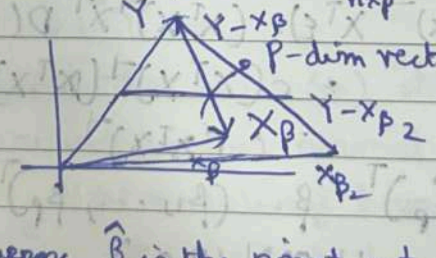
We have obtained the LSE of β:

$$\hat{\beta} = \underset{p\times p}{(X^TX)^{-1}} \; \underset{p\times n}{X^T} \; \underset{n\times 1}{Y} \qquad (Y - X\beta)^T(Y - X\beta)$$

Lets look at geometrically? $x_i \in \mathbb{R}^n$

We want to minimize $(Y - X\beta)^T(Y - X\beta)$

$Y \in \mathbb{R}^n \quad \beta \in \mathbb{R}^p \quad \underset{n\times p}{X} = [\underset{n\times 1}{x_1} \ldots \underset{n\times 1}{x_p}]$



Suppose $\hat{\beta}$ is the point s.t $(Y - X\hat{\beta})^T(Y - X\hat{\beta})$ is minimum $\Rightarrow (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \leq (Y - X\beta)^T(Y - X\beta)$ $\forall \beta \in \mathbb{R}^p$

$\beta^TX^T(Y - X\hat{\beta}) = 0 \quad \forall \beta$

$\beta^TX^T(Y - X\hat{\beta}) = 0 \quad \forall \beta$

$\Rightarrow X^T(Y - X\hat{\beta}) = 0 \Rightarrow X^TY - (X^TX)\hat{\beta} = 0$

$\hat{\beta} = (X^TX)^{-1}X^TY$!

I would like to obtain standard error of estimation of the above model.

How to obtain confidence interval of β based on a sample Y & based on the model assumptions. We make some further assumption on ε_i & it is as follows $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = c > 0$, ε_i's are independent identically distributed normal random variable

$$\hat{\beta} = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\beta + \varepsilon)$$

$$= (X^TX)^{-1}(X^TX)\beta + (X^TX)^{-1}X^T\varepsilon.$$

$$= \beta + (X^TX)^{-1}X^T\varepsilon$$

$$\underset{p\times 1}{\hat{\beta}} = \underset{p\times 1}{\beta} + \underset{p\times p}{(X^TX)^{-1}}\,\underset{p\times n}{X^T}\,\underset{n\times 1}{\varepsilon} \qquad \varepsilon, \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \text{normal } (0, c)$$

$$E((X^TX)^{-1}X^T\varepsilon) = (X^TX)^{-1}X^T E(\varepsilon) = 0$$

$$(X^TX)^{-1}X^T\varepsilon = \begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix} \text{normal distributions} \quad \begin{bmatrix} \times & \times & \cdots \\ & & \end{bmatrix}_{p\times n}$$

$$D((X^TX)^{-1}X^T\varepsilon) = (X^TX)^{-1}X^T D(\varepsilon) X (X^TX)^{-1}$$

$$= c(X^TX)^{-1}(X^TX)(X^TX^{-1})$$

$$= c(X^TX)^{-1}$$

$$\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T \qquad \beta = (\beta_1, \dots, \beta_p)^T$$

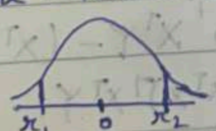
$$\hat{\beta}_1 = \beta_1 + U_1 \qquad U_1 \sim N(0, a_1)$$

where $a_1 = (1,1)$ - th element of $c(X^TX)^{-1}$

$$\hat{\beta}_1 - \beta_1 \sim N(0, a_1)$$

Suppose I know the variance c, then I know a_1

$$P[r_1 < \hat{\beta}_1 - \beta_1 < r_2] = 0.75$$



95%.
Confidence interval becomes
$(\hat{\beta}_1 - r_2, \hat{\beta}_1 - r_1)$ Here r_1 & r_2 are known
The second approach is called bootstrapping.
It is simple as a concept but computer intensive. You need less assumptions

Some preliminaries:
Suppose $Y = (Y_1, \ldots, Y_n)^T$ is a random vector then the mean vector of Y is $E(Y) = (E(Y_1), \ldots, E(Y_n))^T$
Dispersion matrix of Y / Variance-Covariance Matrix.

$$D(Y) = \begin{bmatrix} V(Y_1) & Cov(Y_1, Y_2) & \cdots & Cov(Y_1, Y_n) \\ & & & \\ & & & V(Y_n) \end{bmatrix}$$

$D_{ij}(Y) = Cov[Y_i, Y_j]$
$D(Y)$ is symmetric & positive definite ($l^T D(Y) l > 0$ $\forall l \neq 0$)
The random vector $Y = (Y_1, \ldots, Y_n)^T$ is said to be multivariate normal if any linear combination of $Y_1, \ldots, Y_n$ is univariate normal distribution ($*$)

$l^T Y$ $\forall l$

Using the def^n ($*$) it follows that if $Y_1, \ldots, Y_n$ are independent normal distributions with mean $0, \ldots, 0$, respectively & variance of $Y_1, \ldots, Y_n$ is $\sigma_1^2, \ldots, \sigma_n^2$ respectively, then $(Y_1, \ldots, Y_n)^T$ is multivariate normal.
Note that $E(Y) = (E(Y_1), \ldots, E(Y_n))^T$
$= (0, 0, \ldots, 0)$

$$\underset{n \times n}{D(Y)} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ & \sigma_2^2 & & \\ 0 & & \ddots & \sigma_n^2 \end{bmatrix}$$

If $Y = (Y_1, \dots, Y_n)^T$ is multivariate normal with mean vector 0, and dispersion matrix I

$\Rightarrow Y_1, \dots, Y_n$ are i.i.d. $N(0,1)$

If $Y_1, \dots, Y_n$ are i.i.d $N(0,1)$

$Y = (Y_1, \dots, Y_n)^T \sim N_n(0, I_{n\times n})$

$(\underset{n\times 1}{Y} + \underset{n\times 1}{\mu}) \sim N_n(\mu, I)$

Suppose $\underset{n\times 1}{Y} = (Y_1, \dots, Y_n)^T \sim N_n(0, I)$

$\begin{bmatrix} U_1 \\ \vdots \\ U_n \end{bmatrix} = \underset{n\times n}{A}\,\underset{n\times 1}{Y}$ $\quad l^T U = \underbrace{l^T A}\, Y = \nu^T Y$

$Y \sim N_n(0, I)$ $\quad \underset{n\times n}{A}\,\underset{n\times 1}{Y} \sim N_n(0, AIA^T)$

$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \\ a_{n1} & \cdots & & a_{nn} \end{pmatrix}$, $\quad D(AY) = \begin{bmatrix} V(U_1) & \cdots \\ & \ddots \end{bmatrix}$

If $Y \sim N_n(\mu, \Sigma)$ $\quad AY \sim N_n(A\mu, A\Sigma A^T)$

$Y \sim N_n(\mu, \Sigma)$ $\quad \Sigma$ is positive definite matrix

$\Rightarrow \exists$ a matrix A s.t. $A\Sigma A^T = I$

$\Rightarrow X = AY \sim N_n(A\mu, I)$

$= N_n(\theta, I)$ $\quad \theta = A\mu$

$\Rightarrow (X - \theta) \sim N_n(0, I)$ $\quad r(A)$

$X_1 - \theta_1, \dots, X_n - \theta_n$ i.i.d $N(0,1)$

Let $Z_1, \dots, Z_n$ iid $N(0,1)$, $\quad Z_1^2 + \dots + Z_n^2 \sim$ Chi-square with n-degrees of freedom. (χ^2_n)

$(X_1 - \theta_1)^2 + \dots + (X_n - \theta_n)^2 \sim \chi^2_n$

$(X - \theta)^T (X - \theta) \sim \chi^2_n$ $\quad A\Sigma A^T = I$

$(AY - A\mu)^T (AY - A\mu) \sim \chi^2_n$ $\quad \Sigma A^T = A^{-1}$

$\Sigma = A^{-1}(A^T)^{-1}$

$(Y - \mu)^T A^T A (Y - \mu) \sim \chi^2_n$ $\quad (A^T A)^{-1} = \Sigma$

$(Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi^2_n$

What is a confidence interval / set??

Suppose $X_1 \ldots X_n \sim f(x, \theta^0)$ $\theta^0 \to$ Scalar, $\theta^0 \to$ true value of θ

What is confidence interval of θ??

Suppose $\hat{\theta}(x_1, \ldots, x_n)$ is an estimate of θ,

What do we mean by confidence interval of θ^0??

$I(\hat{\theta})$

We say $I(\hat{\theta})$ is a 90% confidence interval

$P[I(\hat{\theta}(X_1, X_2, \ldots, X_n)) \ni \theta^0] = 0.90$

We are considering confidence interval / set in case of linear regression $i = 1, \ldots, n$

$y_i = \beta_0^0 + \beta_1^0 x_{i1} + \ldots + \beta_p^0 x_{ip} + \varepsilon_i$

ε_i's are iid $N(0,1)$. We have LSE of $\beta_0^0, \ldots, \beta_p^0$, say $\hat{\beta}_0, \ldots, \hat{\beta}_p$. Based on that we want to construct confidence intervals / confidence set of $\beta_0^0, \ldots, \beta_p^0$.

Suppose we want 95% Confidence interval of β_k^0,

$\Rightarrow$ I want an interval (random) which depends on

* $\hat{\beta}_k$ s.t $P[I_k \ni \beta_k^0] = .95$; $k = 0, \ldots, p$

We can have $I_0, I_1, \ldots, I_p$

95% Confidence set of the parameter vector $\beta^0 = (\beta_0^0, \ldots, \beta_p^0)$

$\Rightarrow A(y_1, \ldots, y_n)$ $P[A(y_1, \ldots, y_n) \ni \beta^0] = 0.95$

meaning of probabilities of value in interval

* Suppose you draw 2 random numbers a & b from (0,1)

Let $c = \min\{a, b\}$ & $d = \max\{a, b\}$

Ex. $\underset{0}{[} \; \underset{c}{[} \;\; \underset{d}{]} \; \underset{1}{]}$

Basic Idea:

Suppose you want to construct confidence interval of β_k^0 based on $\hat{\beta}_k$ then you try to find a pivotal quantity based on $\hat{\beta}_k$ & β_k^0 whose distribution is independent of parameters

Let recall

$X_{p\times1} \sim N_p(\underset{p\times1}{\mu}, \underset{p\times p}{\Sigma}) \Rightarrow AX \sim N_p(A\mu, A\Sigma A^T)$

Suppose, $X_1, \dots, X_n$ are i.i.d $N(0,1) \Rightarrow X_1^2 + \dots + X_n^2 \sim \chi_n^2$ (Chi-square)

We want to form (construct) a 95% confidence set of β°

$$\underset{n\times1}{Y} = \underset{n\times p}{X}\,\underset{p\times1}{\beta^\circ} + \underset{n\times1}{\varepsilon} \qquad \varepsilon_i\text{'s are iid } N(0,1)$$

$Y = (Y_1, \dots, Y_n)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$

$\hat\beta = (X^TX)^{-1}X^TY$. We want to construct a 95% confidence set of β° based on $\hat\beta$

$$\hat\beta = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\beta^\circ + \varepsilon) = \beta^\circ + (X^TX)^{-1}X^T\varepsilon$$

$$(\hat\beta - \beta_0) = \underset{p\times p}{(X^TX)^{-1}}\,\underset{p\times n}{X^T}\,\underset{n\times1}{\varepsilon} \qquad \varepsilon = \begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{bmatrix} \sim N_n(0, I_n)$$

$$(X^TX)^{-1}X^T\varepsilon \sim N_p(0_{p\times1}, (X^TX)^{-1}X^TX(X^TX)^{-1})$$

$$(\hat\beta - \beta^\circ) \sim N_p(0_{p\times1}, \underset{p\times p}{(X^TX)^{-1}})$$

Note that X^TX is a positive definite matrix

Since (X^TX) is a positive definite matrix $\exists$ a matrix A s.t. $a^T(X^TX)a > 0 \; \forall a \neq 0$

$$A(X^TX)A^T = I$$

$$\Rightarrow (X^TX) = (A^TA)^{-1} \Rightarrow (X^TX)^{-1} = (A^TA)$$

$$\hat\beta - \beta^\circ \sim N_p(0, \underset{p\times p}{(X^TX)^{-1}})$$

$$(A^T)^{-1}(\hat\beta - \beta^\circ) \to (p\times1) \sim N_p(0, A^{T^{-1}}(X^TX)^{-1}A^{-1}) \sim N_p(0, I)$$

$$\begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} \qquad Z_1^2 + \dots + Z_p^2 \sim \chi_p^2 \text{ (Chi-square)}$$

$$(\hat\beta - \beta_0)^T \underbrace{A^{-1}(A^T)^{-1}}_{(A^TA)^{-1}} (\hat\beta - \beta_0) \sim \chi^2_p$$

How to use (✱) to form confidence set of β°.

$$\underbrace{(\hat\beta - \beta_0)^T (X^TX)(\hat\beta - \beta_0)}_{\text{Pivotal quantity}} \sim \chi^2_p \; (✱)$$

are iid $N(0,1)$
tion.
$X)^{-1})$
parameters
0.80 ?? No ✗
dent of
tion
& c_2 accordingly

Confidence interval / set

$$Y = X\beta^0 + \varepsilon$$

$(\varepsilon_1, \dots, \varepsilon_n)^T = \varepsilon$ here $\varepsilon_1, \dots, \varepsilon_n$ are iid $N(0,1)$

Note that based on that assumption

$$\hat{\beta} = (X^TX)^{-1}X^TY \sim N_p(\beta^0, (X^TX)^{-1})$$

LSE $\quad \hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$

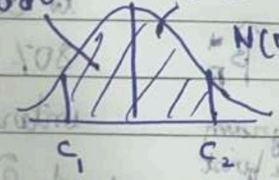
$\Rightarrow \hat{\beta}_k \sim N(\beta^0_k, a_{ii}) \quad k = 1, \dots, p$

when $(\beta^0_1, \dots, \beta^0_p)^T = \beta^0$

$a_{ii} = (i,i)$th element of $(X^TX)^{-1}$

$$\frac{\hat{\beta}_k - \beta^0_k}{\sqrt{a_{ii}}} \sim N(0,1)$$ ← independent of parameter

↑ Pivotal quantity



$P\left\{\left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{a_{kk}}}\right) \in (c_1, c_2)\right\} = 0.80$ (known: c_1, c_2)

$\Rightarrow P\left\{\beta^0_k \in \left(\hat{\beta}_k - \sqrt{a_{kk}}\,c_2,\ \hat{\beta}_k - \sqrt{a_{kk}}\,c_1\right)\right\} = 0.80$

For simplicity we use the following notation

$$P[\beta^0_k \in (b_{kL}, b_{kU})] = 0.80$$

$$P[\beta^0 \in (b_{1L}, b_{1U}) \times \dots (b_{pL}, b_{pU})] = 0.80\ ??$$

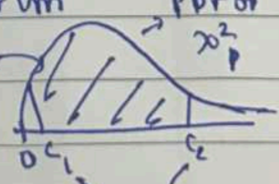
No ✗

Recall

$$\underbrace{(\hat{\beta} - \beta^0)^T (X^TX)^{-1} (\hat{\beta} - \beta^0)}_{\text{Pivotal}} \sim \chi^2_p$$ – independent of parameter

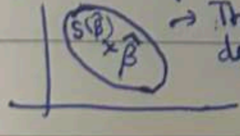
The PDF of χ^2_p is of the following form

$P[(\hat{\beta}-\beta^0)^T (X^TX)^{-1} (\hat{\beta}-\beta^0) \le c_2] = 0.85$ (known: c_2)



If we take $p = 2$

We choose $c_1 = 0$ & c_2 accordingly



→ The shape depends on $(X^TX)^{-1}$

$P[\beta^0 \in S(\hat{\beta})] = 0.85$

In case of confidence interval it looks like

$$(\hat{\beta}_k - b_{kL}, \hat{\beta}_k + b_{kU})$$

Natural question what will happen if ε_i's are not normal.

We can use bootstrap to compute confidence interval/set

$$Y = X\beta^0 + \varepsilon \qquad \varepsilon_1, \dots, \varepsilon_n \text{ are iid } (0,1)$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Compute $\hat{\varepsilon} = (Y - X\hat{\beta})$ $\quad \hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)$

$(\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \dots, \hat{\varepsilon}_n^*)$ You draw sample from $(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)$ at random with replacement

$$Y_1^* = X\hat{\beta} + \varepsilon_1^* \rightarrow \hat{\beta}_1^*$$

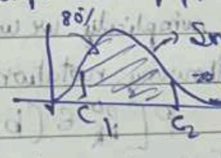
$$Y_2^* = X\hat{\beta} + \varepsilon_2^* \rightarrow \hat{\beta}_2^*$$

$$\vdots$$

$$Y_B^* = X\hat{\beta} + \varepsilon_B^* \rightarrow \hat{\beta}_B^*$$

Suppose I want to construct a 80% confidence interval of β_1^0

Histogram of the first component of $\hat{\beta}_1^*, \dots, \hat{\beta}_B^*$



Smooth histogram $\Rightarrow P[\beta_1^0 \in (c_1, c_2)] = 0.80$