Week 1

• Resource website: buddhanandа_stat (syllabus, Lecture Notes, timetable, booklist); shalabh (IIT kanpur) (simulation, data demonstration)

Prediction → with less "error" → notion of distance/norm (eg. L_1 norm, L_2 norm etc.) → applied/using given data

---

(true) $Y = \beta_0 + \beta_1 X + \epsilon$ → random variable ('noise')

Y: dependent variable; X: independent of Y, β_0, β_1 and non-stochastic (ie non-random)

(A simple linear regression model)

* Linear in parameters (β's variables)

(eg: $Y = \beta_0 + \beta_1 X$ → linear model
$Y = \beta_0 + \beta_1 X^2 + \beta_2 X^3 + \beta_3 \log X$ → "linear" model)
or, $Y = (1 \; X^2 \; X^3 \; \log X) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$
or, $Y = X^T \beta$ (matrix representation)

(prediction) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$ *

Estimate β_0, β_1 (ie get $\hat{\beta}_0, \hat{\beta}_1$)

Error: (squared error) or L_2 norm (aka Euclidean norm)

$\underset{\sim}{y} = (y_1, y_2, \cdots y_n)^T$, then $\|y\|_2^2 = \sum y_i^2 = \underset{\sim}{y}^T \underset{\sim}{y}$ (vector notation)

suppose, $\underset{\sim}{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$ and $\underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ then, $\|z - y\|_2^2 = \sum (y_i - z_i)^2 = (\underset{\sim}{y} - \underset{\sim}{z})^T (\underset{\sim}{y} - \underset{\sim}{z})$

we want to minimize $S = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$ wrt β_0, β_1 & consequently the optimal value for β_0, β_1 are called the least square estimates $(\hat{\beta}_0, \hat{\beta}_1)$.

Sol^n:
- $\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
- $\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \Rightarrow \frac{\sum x_i y_i}{n} = \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \frac{\sum x_i^2}{n}$ (substituting)

(aka Normal Eqn)

or, $\frac{\sum x_i y_i}{n} = (\bar{y} - \hat{\beta}_1 \bar{x}) \bar{x} + \hat{\beta}_1 \frac{\sum x_i^2}{n}$

(solving) or, ⊛ $\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$, where $x_i = x_i - \bar{x}$, $y_i = y_i - \bar{y}$

(& $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$) ✓

But, no reliable measure of error, no testing of hypothesis, no check of model adequacy, no comparisons etc. can be done unless we approach the problem statistically (v/s ad-hoc math, as done above)

∴ we need to develop the foundations rigorously to answer those questions. (which can't be done by merely "2 normal equations")

⊛ we can predict: $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$
⊛ estimate error: $\sum (\hat{y}_i - y_i)^2 = e^2$

(In the above ad-hoc setup, ⊛, ⊛, ⊛ is all that we can do. Nothing more.) → point predictions/estimates

* we don't know the distrn of $\hat{\beta}_0, \hat{\beta}_1, e^2$ & hence we can't know any interval estimates (& hence no hyp. testing)

* Further, what if we have n-variables $(x_1, x_2, \cdots x_n)$; hence a need to impose Linear Algebra structure. Its going to get cumbersome.

xx (End of simplicity!)

Results from Linear Algebra:

Data: $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$; $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \forall i = 1, 2, \cdots n$ → A linear eqn → Linear s/M

we define → $\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}$, $\underset{\sim}{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n \times 2}$, $\underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2 \times 1}$, $\underset{\sim}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n \times 1}$

$\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ (Linear s/M in vector-matrix form)

where, $\underset{\sim}{Y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times 2}$, $\underset{\sim}{\beta} \in \mathbb{R}^{2 \times 1}$, $\underset{\sim}{\epsilon} \in \mathbb{R}^n$

* Assumptions: (a foresight)
① $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
② $\sigma^2, \beta_0, \beta_1$ are unknown (but indep of X, Y)
③ X_i's are non-stochastic

Vector space: A vector space V over $\mathbb{R}$, denoted by $(V, +, \cdot, \mathbb{R})$ has the following properties

$\forall \alpha, \beta \in \mathbb{R}$ and $\underset{\sim}{x}, \underset{\sim}{y}, \underset{\sim}{z} \in V$

1.) a) $+ : V \times V \to V$

b) $(\underset{\sim}{x} + \underset{\sim}{y}) + \underset{\sim}{z} = \underset{\sim}{x} + (\underset{\sim}{y} + \underset{\sim}{z})$

c) $\exists\ \underset{\sim}{0} \in V$ such that: $\underset{\sim}{0} + \underset{\sim}{x} = \underset{\sim}{x} + \underset{\sim}{0} = \underset{\sim}{x}\ \ \forall \underset{\sim}{x} \in V$

d) $\exists\ -\underset{\sim}{x}\ \forall \underset{\sim}{x} \in V$ such that: $-\underset{\sim}{x} + \underset{\sim}{x} = \underset{\sim}{0}$

e) $\underset{\sim}{x} + \underset{\sim}{y} = \underset{\sim}{y} + \underset{\sim}{x}\ \ \forall \underset{\sim}{x}, \underset{\sim}{y} \in V$

2.) a) $\cdot : \mathbb{R} \times V \to V$

b) $\alpha \cdot (\beta \cdot \underset{\sim}{x}) = (\alpha * \beta) \cdot \underset{\sim}{x}$

c) $1 \cdot \underset{\sim}{x} = \underset{\sim}{x}$

d) $(\alpha + \beta) \cdot \underset{\sim}{x} = \alpha \cdot \underset{\sim}{x} + \beta \cdot \underset{\sim}{x}$

e) $\alpha \cdot (\underset{\sim}{x} + \underset{\sim}{y}) = \alpha \cdot \underset{\sim}{x} + \alpha \cdot \underset{\sim}{y}$

Eg.
① $\mathbb{R}^n$ } (dim = is n)
② $\mathbb{C}^n$ }
③ $\mathbb{P}_n$ (polynomials of degree $\leq n$) (but note, the dim is $n+1$) (∵ $n+1$ coefficients)

Subspace: Let $(V, \cdot, +, \mathbb{R})$ is a vector space, and • $\boxed{S \subseteq V}$ such that • $\boxed{(S, +, \cdot, \mathbb{R})}$ is also a vector space

then S is known as subspace of V

Eg: ① $\mathbb{R}^2$: (a) origin $(0,0)$
(b) x axis, y axis, all lines passing through origin. } are subspaces of $\mathbb{R}^2$

② $\mathbb{R}^3$: (a) origin $(0,0,0)$
(b) all lines passing thr. origin
(c) all planes passing thr. origin } are subspaces of $\mathbb{R}^3$

③ $\mathbb{P}_n$: (a) $S = \mathbb{P}_r$, $r \leq 7$ are subspaces of $\mathbb{P}_7$.
(i.e. $\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_2, \ldots \mathbb{P}_7$)

Span: (defined for a set of vectors $\{\underset{\sim}{v}_1, \underset{\sim}{v}_2, \ldots \underset{\sim}{v}_n\} = A$)

Span(A) = set of all linear combinations of $\underset{\sim}{v}_1, \underset{\sim}{v}_2, \ldots \underset{\sim}{v}_n$ (space) $= \{\underset{\sim}{x} \mid \exists\ c_1, c_2, \ldots c_n \in \mathbb{R}$ s.t. $c_1 \underset{\sim}{v}_1 + \ldots c_n \underset{\sim}{v}_n = \underset{\sim}{x}\}$

Basis: (defined for a vector space) A basis of a vector space (V) or subspace ($S \subseteq V$) is defined as a set of linearly independent (non-zero) vectors s.t. they span V or S respectively.

* Need not be unique; but the no. of (non-zero) vectors in the basis is unique for a given V for a given V
↳ a.k.a Dimension of V (∵ dim(V) = cardinality of (|B|) basis of V any)

Eg: $V = \{(x, y, 0) \mid x \in \mathbb{R}, y \in \mathbb{R}\}$

Now, consider $B = \left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}\right\}$ ($\underset{\sim}{x}_1$, $\underset{\sim}{x}_2$) and clearly $\underset{\sim}{x}_1$ & $\underset{\sim}{x}_2$ are lin. indep. & span V

∴ B is a basis for V ∴ dim(V) = 2 (≠ "components" in vector of V)

(Note: This situation shall repeatedly occur in regression analysis) ← i.e. $\boxed{V \subset \mathbb{R}^3,\ \dim(V) = 2}$

orthogonal vectors: vector $\underset{\sim}{x}, \underset{\sim}{y} \in V$ are said to be orthogonal if $\underset{\sim}{x}^T \underset{\sim}{y} = \underset{\sim}{y}^T \underset{\sim}{x} = \underset{\sim}{0}$

* (Notation: $\underset{\sim}{x} \perp \underset{\sim}{y}$)

Orthogonal basis: (special choice of basis) If the vectors of basis set B are orthogonal to each other then the basis is known as a orthogonal basis.

* we can obtain an orthogonalized basis from a non-orthogonal one (via, say, Gram-schmidt orthogonalization)

(In most situations of RTSM, data set given will not be orthogonal, however we can transform them so as to become orthogonal)

orthogonal complement : If S is a subspace of V, then the orthogonal complement of S is defined as

$$S^{\perp} = \{ \underset{\sim}{v} \mid \underset{\sim}{v} \in V \text{ and } \underset{\sim}{v}^T \underset{\sim}{u} = 0 \ \forall \ \underset{\sim}{u} \in S \}$$

* $S^{\perp} \cap S = \underset{\sim}{0}$ ($\neq$ empty set)
* dimension($S^{\perp}$) + dimension(S) = dimension(V)

(**Note**: we shall use this heavily in RTSM as subspace of error (S_1) and subspace of data set's prediction (S_2) will be orthogonal complements & both shall together form entire V; however, S_2 alone cannot span entire V).

(* These dimensions will give us the "degrees of freedom" in statistical analysis)

(* $\text{corr}(X,Y) = \dfrac{E(X \cdot Y) - E(X)E(Y)}{\sqrt{V(X) \cdot V(Y)}}$

Projection Matrix (because geometrically as we shall see, regression is nothing but a projection)

If S is a subspace of a vector space V, then the "projection matrix" P_S for $S \subseteq V$ satisfies the following: (definition)

① If $\underset{\sim}{v} \in S$, $P_S \underset{\sim}{v} = \underset{\sim}{v}$

② $\forall \ \underset{\sim}{v} \in V$, $P_S \underset{\sim}{v} \in S$

eg. $V = \mathbb{R}^3$ and $S = \{ (x, y, 0)^T \mid \underset{y \in \mathbb{R}}{x \in \mathbb{R}} \}$ (clearly S is a valid subspace of V — plane through origin.)

we already tell you the projection matrix

$$P_S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Now, i) $\underset{\sim}{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \therefore P_S \underset{\sim}{v} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} \in S$.

ii) $\underset{\sim}{v} = \begin{pmatrix} 4 \\ 5 \\ 0 \end{pmatrix} \therefore P_S \underset{\sim}{v} = \begin{pmatrix} 4 \\ 5 \\ 0 \end{pmatrix} = \underset{\sim}{v}$ ($\because \underset{\sim}{v} \in S$).

* iii) $\underset{\sim}{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \therefore P_S \underset{\sim}{v} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$, Now, $P_S (P_S \underset{\sim}{v}) = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} = P_S \underset{\sim}{v}$

ie $P_S^2 \underset{\sim}{v} = P_S \underset{\sim}{v}$ ($\because P_S$ is idempotent)

(* If P_S is a projection matrix for $S \subseteq V$, and $(I_n - P_S)$ is the projection matrix for $S^{\perp} \subseteq V$, then P_S & $(I - P_S)$ are orthogonal projection matrices of S and $S^{\perp}$

Note: we shall split a vector into two orthogonal components or spaces via multiplying it once with P_S & other time via $(I - P_S)$

Proof: $P_S \underset{\sim}{v} \perp (I - P_S) \underset{\sim}{v}$

$[(P_S \cdot \underset{\sim}{v}) \cdot (I - P_S) \cdot \underset{\sim}{v} = 0$ ($\because P_S^2 = P_S$; idempotent) (hence, $P_S \cdot (I - P_S) = P - P^2 = 0$)

(Proof: $(P_S \cdot v)^T (I - P_S) v = \underset{\sim}{v}^T (P_S^T I - P_S^T P_S) \underset{\sim}{v} = 0$

$P_S^T = P_S$

↑ dot product

* In regression analysis, these 2 components will be prediction and error (orthogonal) we shall/our job will be to simply (!) find the projection matrix.

Th^m: Projection matrix is an idempotent matrix. (Proof: $P_S \underset{\sim}{v} \in S \ \forall \ \underset{\sim}{v} \in V$ (by ②) Now, $P_S (P_S \underset{\sim}{v}) = P_S v$ (by ①) ($\because P_S^2 = P_S$))

Th^m: The eigen values of projection matrix are 1 & 0 (Proof: $P_S \underset{\sim}{v} = \lambda \underset{\sim}{v}$

or, $P_S (P_S \underset{\sim}{v}) = \lambda P_S \underset{\sim}{v}$

or, $P_S^2 \underset{\sim}{v} = \lambda P_S v$

or, $P_S \underset{\sim}{v} = \lambda P_S v$ or, $\lambda \underset{\sim}{v} = \lambda^2 \underset{\sim}{v} \Rightarrow \lambda^2 = \lambda \Rightarrow \lambda = 0, 1$

($\because \underset{\sim}{v}$ is a non-zero vector)

(* Th^m: If $\{ \underset{\sim}{v}_1, \underset{\sim}{v}_2, \ldots, \underset{\sim}{v}_k \}$ is an orthonormal basis of a subspace $S \subseteq V$ then, $P_S = \sum_{i=1}^{k} \underset{\sim}{v}_i \underset{\sim}{v}_i^T$ is an orthogonal projection matrix for $S \subseteq V$

(outer-product) (⇒ we'll get a matrix)

we use this theorem to find projection matrix via orthonormal basis.

(& To find orthonormal basis, we use Gram-Schmidt orthonormalization)

Column space: $\mathcal{C}(A)$ = linear combinations of the columns of a matrix $A = [\underset{\sim}{a}_1\ \underset{\sim}{a}_2 \dots \underset{\sim}{a}_n]$ all possible

$$\text{i.e. } \mathcal{C}(A) = \text{span}(\underset{\sim}{a}_1, \underset{\sim}{a}_2, \dots \underset{\sim}{a}_n) = \left\{ \sum_{i=1}^{n} x_i \underset{\sim}{a}_i \mid \forall x_i \in \mathbb{R} \right\} = \left\{ A\underset{\sim}{x} \mid \underset{\sim}{x} \in \mathbb{R}^n \right\}$$

Similarly, Row space $= \mathcal{R}(A) = \mathcal{C}(A^T)$

Properties:

① $\mathcal{C}(A:B) = \mathcal{C}(A) + \mathcal{C}(B)$; $\mathcal{C}(AB) \subseteq \mathcal{C}(A)$ (A:B partitioned)

② $\dim(\mathcal{C}(A)) = \text{Rank}(A)$

③ $\mathcal{C}(AA^T) = \mathcal{C}(A) \Rightarrow \text{Rank}(AA^T) = \text{Rank}(A)$

(Proof: ⟨i⟩ $\mathcal{C}(AA^T) \subseteq \mathcal{C}(A)$ by ① ; ⟨ii⟩ $\mathcal{C}(A) \subseteq \mathcal{C}(AA^T)$ → tedious proof.) $\Leftrightarrow \mathcal{C}(AA^T) = \mathcal{C}(A)$

1.) **Positive definite (PD) matrix**: (analogue of positive nos & negative nos) A is said to be a PD matrix iff $\underset{\sim}{x}^T A \underset{\sim}{x} > 0 \ \forall \underset{\sim}{x} \neq \underset{\sim}{0}$ $(\Leftrightarrow |A| > 0)$

Positive semidefinite matrix: A is said to be a PSD matrix iff $\underset{\sim}{x}^T A \underset{\sim}{x} \geq 0 \ \forall \underset{\sim}{x} \neq \underset{\sim}{0}$ $(\Leftrightarrow |A| \geq 0)$

2.) **Generalized inverse**: A^- has property: $AA^-A = A$

- (i.e. even for non-square matrices)
- but NOT unique

3.) For any matrix A, the projection matrix of $\mathcal{C}(A)$ is (AA^-) (but doesn't split into orthogonal components) and the orthogonal projection matrix is $\boxed{A(A^TA)^-A^T}$

Multivariate Analysis (// Prob. stats will mostly for 1 or 2 variables)

$$\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = (X_1\ X_2 \dots X_n)^T$$

1) $E(\underset{\sim}{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \underset{\sim}{\mu}$ (if $E(|X_i|) < \infty \ \forall i = 1, 2, \dots n$) (// check of absolute convergence for each X_i)

2) "Dispersion matrix" of $\underset{\sim}{X}$ (vector analogue of variance) denoted by $D(\underset{\sim}{X}) \equiv (\text{Cov}(X_i, X_j))_{ij}$

(a.k.a variance-covariance matrix) $= (E(X_iX_j) - E(X_i)E(X_j))_{ij} = (E(X_iX_j) - \mu_i\mu_j)_{ij}$

$\equiv \underset{n\times n}{\Sigma}$ or $\underset{n\times n}{\Sigma_X}$ $= \boxed{E(\underset{\sim}{X}\underset{\sim}{X}^T) - \underset{\sim}{\mu}\underset{\sim}{\mu}^T}$

(i) Note: Diagonal elements are the variances of the corresponding random variables.

(ii) Note: Covariance b/w vectors of 2 different dimensions (p, q)

$\text{Cov}(\underset{\sim}{U}_p, \underset{\sim}{V}_q) = (\text{Cov}(u_i, v_j))_{ij}$ (dimension: $p \times q$) $= \underbrace{E(\underset{\sim}{U}_p \underset{\sim}{V}_q^T)}_{p\times q} - \underbrace{\underset{\sim}{\mu}_U \underset{\sim}{\mu}_V^T}_{p\times q}$ (Generalized) for non-square matrices too.

(iii) $E(\underset{\sim}{X} + \underset{\sim}{b}) = E(\underset{\sim}{X}) + \underset{\sim}{b}$ where, $\underset{\sim}{b}$ is a constant/non-random vector.

(iv) $D(\underset{\sim}{X} + \underset{\sim}{b}) = D(\underset{\sim}{X})$ (v) $\text{Cov}(\underset{\sim}{X} + \underset{\sim}{b}, \underset{\sim}{Y} + \underset{\sim}{c}) = \text{Cov}(\underset{\sim}{X}, \underset{\sim}{Y})$

~~Some further results:~~ (vi) $E(\underset{\sim}{l}^T \underset{\sim}{X}) = \underset{\sim}{l}^T \underset{\sim}{\mu}$ ← finally a scalar

(vii) $D(\underset{\sim}{l}^T \underset{\sim}{X}) = \underset{\sim}{l}^T \Sigma \underset{\sim}{l}$ ← finally a scalar.

(aka $var(\underset{\sim}{l}^T \underset{\sim}{X})$)

(general result)

(viii) $E(A\underset{\sim}{X}) = A\underset{\sim}{\mu}$

(ix) $D(A\underset{\sim}{X}) = A \Sigma A^T$

(x) $cov(\underset{\sim}{U}_p, \underset{\sim}{V}_q) = \Gamma$ ('gamma' symbol), then: $cov(A\underset{\sim}{U}, B\underset{\sim}{V}) = A\Gamma B^T$ (general result) #

$\Rightarrow cov(A\underset{\sim}{X}, B\underset{\sim}{X}) = A \Sigma B^T$ (special case)

$\Rightarrow cov(\underset{\sim}{l}^T\underset{\sim}{X}, \underset{\sim}{l}^T\underset{\sim}{X}) = \underset{\sim}{l}^T \Sigma \underset{\sim}{l}$ (special case)

($= var(\underset{\sim}{l}^T\underset{\sim}{X})$)

(# Nothing but linearity of expectation.)

ie $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$

~~(which also implies...)~~

** Thm: show that $D(\underset{\sim}{X})$ is a p.s.d (positive semi definite matrix) (ie $\underset{\sim}{l}^T \Sigma \underset{\sim}{l} \geq 0 \;\; \forall \underset{\sim}{l} \neq \underset{\sim}{0}$) (p.s.d).

$D(\underset{\sim}{X}) = \Sigma_X$

~~Solution/proof:~~

(hint: easily obtained via the fact: $V(X) \geq 0$)

Proof: We need to show that for every $\underset{\sim}{l} \neq \underset{\sim}{0}$; $\underset{\sim}{l}^T \Sigma \underset{\sim}{l} \geq 0$ ← scalar.

$\Leftrightarrow var(\underset{\sim}{l}^T \underset{\sim}{X}) \geq 0$ which is true! (since, variance of a random variable is non-negative).

**(Very important)

Thm: Let $E(\underset{\sim}{X}) = \underset{\sim}{\mu}$, $D(\underset{\sim}{X}) = \Sigma_{n\times n}$

then, $P((\underset{\sim}{X} - \mu) \in \mathcal{C}(\Sigma)) = 1$

(In simple words, the random component of $\underset{\sim}{X}$ is $(\underset{\sim}{X} - \underset{\sim}{\mu})$ and this theorem says that we can always express that random component as a linear combination of column vectors of $\Sigma_{n\times n}$.)

$P(\cdot\cdot) = 1$ (random vector) (in multivariate)

3) Theorem: Let X be a random vector with n components such that $E(\underset{\sim}{X}) = \underset{\sim}{\mu}$ and, $D(\underset{\sim}{X}) = \Sigma$ and $rank(\Sigma) = r \leq n$ (ie possibility of linear dependence in columns of Σ matrix)

If we assume that $\Sigma = BB^T$, where B is a $(n \times r)$ matrix and $C_{(r\times n)}$ is a left inverse of $B_{(n\times r)}$ (ie $CB = I_r$) then defining $\underset{\sim}{Y} = C(\underset{\sim}{X} - \underset{\sim}{\mu})$ we get

(i) $E(\underset{\sim}{Y}) = \underset{\sim}{0}$

(ii) $D(\underset{\sim}{Y}) = I_r$

(iii) $X = \mu + B\underset{\sim}{Y}$ with probability 1

(for full rank (ie $r = n$) $B = I_{n\times n}$)

Analogy:

let $E(Y) = 0$, $V(Y) = 1$ in $\mathbb{R}^1$

$X = \mu + \sigma Y \Rightarrow E(X) = \mu$, $V(X) = \sigma^2$

$E(\underset{\sim}{Y}) = \underset{\sim}{0}$, $D(\underset{\sim}{Y}) = I_n$ in $\mathbb{R}^n$

$X = \underset{\sim}{\mu} + \Sigma^{1/2} \underset{\sim}{Y} \Rightarrow E(\underset{\sim}{X}) = \underset{\sim}{\mu}$, $D(\underset{\sim}{X}) = (\Sigma^{1/2})^T \Sigma^{1/2} = \Sigma$

? (what do you mean by "$\Sigma^{1/2}$" bcoz Σ is a matrix)

* (Recall SVD: singular value decompⁿ)

For $\Sigma_{n\times n} = \Sigma^T$ (symmetric)

$\Sigma = PDP^T$, where P is an orthogonal matrix and, $D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$; $\Sigma \underset{\sim}{P_i} = \lambda_i P_i$

(we shall return here, when we do Principal component regression)

* $\Sigma^{1/2} = PD^{1/2}P^T$, $D^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_n} \end{pmatrix}$

4) Multivariate Normal: A random vector $\underset{\sim}{X}_n$ is said to be following multivariate normal distrⁿ if it has a p.d.f:-

$$f(\underset{\sim}{x}) = \frac{\exp\left\{-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu})\right\}}{(\sqrt{2\pi})^n \sqrt{|\Sigma|}}$$

$|\Sigma|$ ← determinant of Σ

* $E(\underset{\sim}{X}) = \underset{\sim}{\mu}$

* $D(\underset{\sim}{X}) = \Sigma_{n\times n}$

Scanned by CamScanner

Q.) Derive Univariate Normal distn ($n=1$) $(X \in \mathbb{R}^1)$

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

Q.) Derive Bivariate Normal distn ($n=2$) $(X \in \mathbb{R}^2)$

→ $f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \dfrac{\exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu})\right\}}{(\sqrt{2\pi})^2 \sqrt{|\Sigma|}}$, $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

$\therefore$ we get $|\Sigma| = w$ and, $\Sigma^{-1} = w$ (substitute)

$$= \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$
$$(2\pi)\,\sigma_1\sigma_2\sqrt{1-\rho^2}$$

Thm: (For any distribution) Let $\underset{\sim}{X}$ be a random vector in $\mathbb{R}^n$ with $E(\underset{\sim}{X}) = \underset{\sim}{\mu}$ and $D(\underset{\sim}{X}) = \Sigma$

then → $E(\underset{\sim}{X}^T A \underset{\sim}{X}) = \text{trace}(A\Sigma) + \underset{\sim}{\mu}^T A \underset{\sim}{\mu}$ ($\mathbb{R}^n$)

(scalar)

a.k.a "Quadratic Form"

$(\because \underset{\sim}{x}^T A \underset{\sim}{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j)$

Analogy: $\mathbb{R}^2$

$E(ax_1^2 + 2bx_1x_2 + cx_2^2)$

$= \text{trace}\left[\begin{pmatrix} a & b \\ b & c \end{pmatrix}\begin{pmatrix} V(x_1) & \text{cov}(x_1,x_2) \\ \text{cov}(x_1,x_2) & V(x_2) \end{pmatrix}\right] + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}^T \begin{pmatrix} a & b \\ b & c \end{pmatrix}\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$

$E(aX^2)$ ($\mathbb{R}^1$) $= aV(X) + a\mu$

$\because$ let $X_1, X_2, \ldots X_n$ iid $N(0,1)$ ↔ $\underset{\sim}{X} \sim N(\underset{\sim}{0}, I_n)$

Therefore, this is equivalent to

* then, $E(\sum X_i^2) = E(\underset{\sim}{X}^T I_n \underset{\sim}{X}) = tr(I_n I_n) + 0 \; (\because \underset{\sim}{\mu} = \underset{\sim}{0})$
$= n$

If, $X_i \sim N(\mu_i, 1)$ and indep.

then, $E(\sum X_i^2) = n + \underset{\sim}{\mu} I_n \underset{\sim}{\mu}$

or, $E(\sum X_i^2) = (n + \sum \mu_i^2)$

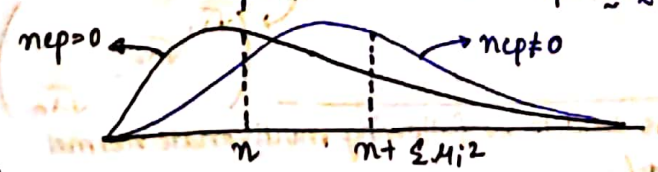
If $X_i \overset{iid}{\sim} N(0,1)$ then $\sum X_i^2 \sim \chi^2_n$ (already known) (a.k.a "central χ^2")

New: $X_i \overset{iid}{\sim} N(\mu_i, 1)$, then $\sum X_i^2 \sim \chi^2_{df=n}$ and $ncp = \sum \mu_i^2$ (or, $\underset{\sim}{\mu}^T\underset{\sim}{\mu}$) ($\because \sum \mu_i^2 = 0$)

(a.k.a "non-central χ^2") (more general)

(Note: $n \le n + \underset{\sim}{\mu}^T\underset{\sim}{\mu}$.)

i.e. $E(\chi^2_{df=n,\, ncp=0}) \le E(\chi^2_{df=n,\, ncp=\underset{\sim}{\mu}^T\underset{\sim}{\mu}})$



(Remark: whether $\mu_i > 0$ or < 0, $\sum \mu_i^2 > 0$ & hence push the mean rightward & as we shall see later that under null hypo. $ncp = 0$, whereas under alternate hyp. $\sum \mu_i^2 > 0$, & help us reject the null).

Even more general: If $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, I_n)$ then

$\underset{\sim}{X}^T A \underset{\sim}{X} \sim \chi^2_{df = \text{rank}(A)}$

$ncp = \underset{\sim}{\mu}^T A \underset{\sim}{\mu}$

iff A is an idempotent matrix.

(As we shall see later, A strongly governs the ncp & df; no matter what the $\underset{\sim}{\mu}$ vector's rank & μ_i's are).

Results:

① If $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, \Sigma)$
then $A\underset{\sim}{X} \sim N(A\underset{\sim}{\mu}, A\Sigma A^T)$

② If $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, \Sigma)$, and if $\exists$ B and its left inverse C, such that $\underset{\sim}{Y} = C(\underset{\sim}{X} - \underset{\sim}{\mu})$ follows $N(\underset{\sim}{0}, I_r)$
then, $X = \underset{\sim}{\mu} + B\underset{\sim}{Y}$ (Can always be written). (i.e. $\underset{\sim}{X}$ can be represented as a linear transformation of $\underset{\sim}{Y}$)

③ If A_1, A_2 symmetric and idempotent matrix such that $Q = (A_1 - A_2)$ is p.s.d, then $\underset{\sim}{X}^T Q \underset{\sim}{X}$ and $\underset{\sim}{X}^T A \underset{\sim}{X}$ are independent.
$A_1 = Q + A_2$, $|Q| \geq 0$, for normally distributed $\underset{\sim}{X}$

④ If $A = A^T$ and $CA = O$ matrix, then $X^T A X$ and CX are independent

(Special case: $X \sim N(0,1)$, $Y \sim \chi^2_n$ ⟩ indep.; then $\frac{X}{\sqrt{Y/n}} \sim t_n$; ← definition of t_n

↳ using ④, we can show that $\bar{X}$ and s^2 are independent.
(Hence $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_n$.)

(Elaborate: X_i iid $N(\mu, \sigma^2)$. Now, $\bar{X} = \frac{1}{n}\Sigma_i X_i = \left(\frac{1}{n}\ \frac{1}{n}\ \frac{1}{n} \cdots \frac{1}{n}\right)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

and, $s^2 = \Sigma_i (x_i - \bar{x})^2$

$= \frac{1}{n}\underbrace{\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}^T}_{\underset{\sim}{1}^T}\underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}}_{\underset{\sim}{X}} = \underset{\sim}{1}^T\underset{\sim}{X}$, where $\underset{\sim}{1} = \frac{1}{n}\underset{\sim}{1}$ #(1 ≡ C) (in ④)

$= \underset{\sim}{X}^T(I_n - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T)\underset{\sim}{X}$ (verify) (≡ $\underset{\sim}{X}^T A \underset{\sim}{X}$) (in ⑧)

hint: expand $\Sigma(x_i - \bar{x})^2$ and use $a^T b = \Sigma_i a_i x_i$ again n again

- Also, $CA = \underset{\sim}{1}^T(I_n - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T)$ (*)
$= \frac{1}{n}\left(\underset{\sim}{1}^T I_n - \frac{1}{n}\underset{\sim}{1}^T\underset{\sim}{1}\underset{\sim}{1}^T\right)$
$= \frac{1}{n}\left(\underset{\sim}{1}^T - \underset{\sim}{1}^T\right)^{n} = \underset{\sim}{0}^T$

- Also, $A = A^T$ (verify yourself) —(**)
- Also, $A = A^2$ (idempotent) —(#)

Note: (*) & (**) are enough to show that $\bar{x}$ and s^2 are independent.
Note: (#) is needed because we want to show that s^2 will follow χ^2.

(Big picture: $\underset{\sim}{X}$ follows normal, then $A\underset{\sim}{X}$ follows normal
then, $(A\underset{\sim}{X})^T(A\underset{\sim}{X}) = \underset{\sim}{X}^T A^T A \underset{\sim}{X} = \underset{\sim}{X}^T A^2 \underset{\sim}{X} = \underset{\sim}{X}^T A \underset{\sim}{X}$
(if $A^T = A$) (if $A^2 = A$))

→ (As $\underset{\sim}{1}^T A = \underset{\sim}{0}^T$ and $A^T = A$, then ($\underset{\sim}{1}^T\underset{\sim}{X} = \bar{X}$ and, $\underset{\sim}{X}^T A \underset{\sim}{X} = s^2$) are independently distributed.)

→ ($\bar{X} = \underset{\sim}{1}^T\underset{\sim}{X} \sim N(\underbrace{\underset{\sim}{1}^T\underset{\sim}{\mu}}_{\text{scalar}}, \sigma^2\underbrace{\underset{\sim}{1}^T I_n \underset{\sim}{1}}_{\text{scalar}})$, where $\underset{\sim}{\mu} = \mu\underset{\sim}{1}$, $\underset{\sim}{1} = \frac{1}{n}\underset{\sim}{1}$

or, $\bar{X} \sim N(\mu, \sigma^2/n)$!

→ ($s^2 = \Sigma_i (x_i - \bar{x})^2 = \underset{\sim}{X}^T A \underset{\sim}{X}$; however, this is not χ^2: (because, that required $x_i \overset{iid}{\sim} N(\mu_i, 1)$
but we have, $x_i \overset{iid}{\sim} N(\mu, \sigma^2)$ ✗s.
#(∵ we need to scale it)

$\underset{\sim}{X} \sim N(\mu \cdot \underset{\sim}{1}, \sigma^2 I_n)$ or, $x_i \sim N(\mu, \sigma^2)$

define: $Y_i = (x_i/\sigma)$

$\underset{\sim}{Y} \sim N(\underbrace{(\mu/\sigma)}_{\mu^*}\underset{\sim}{1}, I_n)$ ⇐ ∴ $Y_i \sim N(\frac{\mu}{\sigma}, 1)$

$\therefore \frac{s^2}{\sigma^2} = \left(\frac{\underset{\sim}{X}}{\sigma}\right)^T A \left(\frac{\underset{\sim}{X}}{\sigma}\right) = \underset{\sim}{Y}^T A \underset{\sim}{Y}$ which does follow $\chi^2_{df = \text{rank}(A),\ ncp = (\mu^*)^2 \underset{\sim}{1}^T A \underset{\sim}{1}}$

Now, rank(A) $= n-1$ (trivial)
$ncp = (\mu^*)^2 \underset{\sim}{1}^T A \underset{\sim}{1} = 0$ (trivial)] just substitute A) or, open it up

$\therefore$ Finally, $\boxed{\left(\frac{s^2}{\sigma^2}\right) \sim \chi^2_{n-1,\ ncp=0} = df}$

→ And ultimately, we assimilate: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) \sim N(0,1)$

$\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$ and $\bar{x}, \frac{s^2}{\sigma^2}$ are independent

$\therefore t_{\text{stat}} \equiv \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{s^2/n-1}} \equiv \frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{s^2}{\sigma^2(n-1)}}}$

$\therefore \boxed{t_{\text{stat}} \sim t_{df=n-1,\ ncp=0}}$

→ (Key result used in ANOVA)
Let $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, I_n)$ and $\underset{\sim}{X}^T A \underset{\sim}{X} = \sum_i \underset{\sim}{X}^T A_i \underset{\sim}{X}$ (can be split like as)
when $A_i^T = A_i$ and $A_i^2 = A_i$
then A_i's are independent
and, $\underset{\sim}{X}^T A_i \underset{\sim}{X}$ $df = \text{rank}(A_i)$
$ncp = \underset{\sim}{\mu}^T A_i \underset{\sim}{\mu}$

$\left(\underset{\sim}{y} \cdot \begin{pmatrix} 1&0&0 \\ 0&1&0 \\ 0&0&1 \end{pmatrix}\right) = \underbrace{\begin{pmatrix} 1&0&0 \\ 0&0&0 \\ 0&0&0 \end{pmatrix}}_{A_1} + \underbrace{\begin{pmatrix} 0&0&0 \\ 0&1&0 \\ 0&0&0 \end{pmatrix}}_{A_2} + \underbrace{\begin{pmatrix} 0&0&0 \\ 0&0&0 \\ 0&0&1 \end{pmatrix}}_{A_3}$

$\left(\because A = A_1 + A_2 + A_3\right)$

Simple Linear Regression Model → $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
↓ Added assm of $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
"Gauss Markov" Model

$\therefore$ Unknowns: $(\beta_0, \beta_1, \sigma^2)$

In matrix notation (vector): $\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $\underset{\sim}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$, $\underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

$\therefore \underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\epsilon}$

Now, $\underset{\sim}{\epsilon} \sim N(\underset{\sim}{0}, \sigma^2 I_n)$

$\Rightarrow \boxed{\underset{\sim}{Y} \sim N(X\underset{\sim}{\beta}, \sigma^2 I_n)}$ (from multivariate analysis)

Note: ϵ_i are independent & identically distributed
but, Y_i are independent but Not identically distr (∵ they have different mean → $X\underset{\sim}{\beta}$)

Estimation of $\beta_0, \beta_1, \sigma^2$
1) Least squares method (LS)
2) Maximum Likelihood method (MLE)

* L.S. condition is to $\boxed{\text{minimize } S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}$ (a.k.a RSS)

$= \| \underset{\sim}{Y} - X\underset{\sim}{\beta} \|_2^2$ (square of Euclidean norm)

$= (\underset{\sim}{Y} - X\underset{\sim}{\beta})^T (\underset{\sim}{Y} - X\underset{\sim}{\beta})$

$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \arg\min_{(\beta_0, \beta_1)} S(\beta_0, \beta_1)$

✓ (Now, either we could do individual differentiation $\sum(\cdot)^2$ or, vector differentiation of last expression)

Let's try both!

Method 1: Vector differentiation.

Method 2: (Trivial)

(hint: $\frac{\partial \underset{\sim}{b}^T A}{\partial \underset{\sim}{b}} = A^T$; $\frac{\partial \underset{\sim}{b}^T A \underset{\sim}{b}}{\partial \underset{\sim}{b}} = \underset{\sim}{b}^T (A + A^T)$)

$$\frac{\partial S}{\partial \underset{\sim}{\beta}} = \frac{\partial (\underset{\sim}{Y}^T \underset{\sim}{Y} - \underset{1\times n}{\underset{\sim}{Y}^T} \underset{n\times 2}{X} \underset{2\times 1}{\underset{\sim}{\beta}} - \underset{\sim}{\beta}^T X^T \underset{\sim}{Y} + \underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta})}{\partial \underset{\sim}{\beta}}$$

(∵ $\underset{\sim}{Y}^T X \underset{\sim}{\beta}$ and $\underset{\sim}{\beta}^T X^T \underset{\sim}{Y}$ are scalars & (scalar)^T = scalar
∴ $(\underset{\sim}{Y}^T X \underset{\sim}{\beta})^T = \underset{\sim}{\beta}^T (\underset{\sim}{Y}^T X)^T = \underset{\sim}{\beta}^T X^T \underset{\sim}{Y}$.)

$$\text{or, } 0 = 0 - \frac{\partial}{\partial \underset{\sim}{\beta}} (2 \underset{\sim}{\beta}^T X^T \underset{\sim}{Y}) + \frac{\partial}{\partial \underset{\sim}{\beta}} (\underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta})$$

$$\text{or, } 0 = -2 (X^T \underset{\sim}{Y})^T + \hat{\underset{\sim}{\beta}}^T \underbrace{(X^T X + (X^T X)^T)}_{2 X^T X}$$

$$\text{or, } (X^T \underset{\sim}{Y})^T = \hat{\underset{\sim}{\beta}}^T (X^T X)$$

$$\text{or, } (X^T \underset{\sim}{Y}) = \underbrace{(X^T X)^T}_{(X^T X)} \hat{\underset{\sim}{\beta}} \Rightarrow \boxed{\hat{\underset{\sim}{\beta}} = (X^T X)^{-1} (X^T \underset{\sim}{Y})} \quad (\text{if } (X^T X)^{-1} \text{ exists})$$

$(n\times 1) \quad (n\times 1)$

(For 2 variable case (i.e. "simple" Regression): $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$,
$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \equiv \left(\frac{S_{xy}}{S_{xx}}\right)$)

Prediction

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \forall i = 1, 2, \dots n$

$$\text{or, } \hat{\underset{\sim}{Y}} = X \hat{\underset{\sim}{\beta}} = X (X^T X)^{-1} (X^T Y). = \boxed{X (X^T X)^{-1} X^T \underset{\sim}{Y}}.$$

(i.e. symmetric & idemp.)
(∵ $P_X^T = P_X^2 = P_X$)

Observations:
- we can identify this as the projection of $\underset{\sim}{Y}$ on the column space of X.
- Infact, this is not just any projection, but an orthogonal projection of $\underset{\sim}{Y}$ on the $\mathcal{C}(X)$

hence, we shall call it as: P_X

$\therefore \hat{\underset{\sim}{Y}} = P_X \underset{\sim}{Y}$

Further, the error vector in the LS prediction is: $\underset{\sim}{e} = \underset{\sim}{Y} - P_X \underset{\sim}{Y}$
$= (I_n - P_X) \underset{\sim}{Y}$ (Wow!)

① $\underset{\sim}{e}$ is orthogonal to $\hat{\underset{\sim}{Y}}$

② Although $\underset{\sim}{\varepsilon}$ has covariance matrix $\sigma^2 I_n$, $\underset{\sim}{e}$ does not have independent components.

(Note: orthogonality of $\underset{\sim}{e}$ and $\underset{\sim}{Y}$ implies their uncorrelation. (∵ an added assumption of Normality will imply their independence))

$\underset{\sim}{Y} \sim N(X \underset{\sim}{\beta}, \sigma^2 I_n)$

$\therefore P_X \underset{\sim}{Y} \sim N(\underbrace{P_X X \underset{\sim}{\beta}}_{X}, \sigma^2 \underbrace{P_X^T P_X}_{P_X})$ (or, $\hat{\underset{\sim}{Y}}$)

(since projection of X on its own columnspace will be X itself)

(since, P_X is a projection matrix & hence $P_X^T = P_X = P_X^2$)

and, $(I - P_X) \underset{\sim}{Y} \sim N(\underset{\sim}{0}, \sigma^2 (I_n - P_X))$ (or, $\underset{\sim}{e}$) (∵ $\underset{\sim}{e} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}}$)

(Alter: verify it as $(X (X^T X)^{-1} X^T)^T (X (X^T X)^{-1} X^T) = X (X^T X)^{-1} X^T$)

- $\hat{\underset{\sim}{\beta}} = [(X^T X)^{-1} X^T \underset{\sim}{Y}] \sim N(\underbrace{(X^T X)^{-1} X^T X}_{I_n} \underset{\sim}{\beta}, \underbrace{(X^T X)^{-1} X^T I_n X (X^T X)^{-1}}_{I_n} \sigma^2)$

$\therefore \hat{\underset{\sim}{\beta}} \sim N(\underset{\sim}{\beta}, (X^T X)^{-1} \sigma^2)$

↳ Linear ($\because \hat{\underset{\sim}{\beta}} = (X^T X)^{-1} X^T \underset{\sim}{Y}$)
unbiased estimator ($\because \hat{\underset{\sim}{\beta}} \sim N(\underset{\sim}{\beta}, (X^T X)^{-1} \sigma^2)$, $E(\hat{\beta}) = \beta$)

Essentially, we just use:-
given: $\underset{\sim}{Y} \sim N(X \underset{\sim}{\beta}, \sigma^2 I)$
- $A \underset{\sim}{Y} \sim N(A\mu, A \Sigma A^T)$ again & again

keep in mind
$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2}$

Also, • $\text{cov}(AY, BY) = (A \Sigma B^T) \sigma^2$

Note: $\underset{\sim}{\beta}$ need not be independent vector
(i.e. $\beta_1, \beta_2, \beta_0, \dots$ may not be indep.)
(i.e. $(X^T X)^{-1}$ need not be diagonal)
(**) (i.e. $\text{cov}(\beta_i, \beta_j)$ need not be zero for all $i \neq j$.)

Now, squared error / error sum of square / residual sum of square :-

$RSS \equiv \sum e_i^2 \equiv \underset{\sim}{e}^T \underset{\sim}{e} = [(I-P_X)\underset{\sim}{Y}]^T[(I-P_X)\underset{\sim}{Y}] = \underset{\sim}{Y}^T(I-P_X)\underset{\sim}{Y}$ ($\because P_X^2 = P_X$, $P_X^T = P_X$)

(I_n: rank $= n$ (obviously); P_X: constructed using X only, ↳ rank $= 2$ (assuming inverse exists))

$\therefore RSS \sim \chi^2_{df=}$ (NO)

* we need to scale it down

$\frac{RSS}{\sigma^2} \sim \chi^2_{df = \text{rank}(I_n - P_X)}$ (rank $= (n-2)$)

$ncp = \left[\left(\frac{X\underset{\sim}{\beta}}{\sigma}\right)^T (I_n - P_X)\left(\frac{X\underset{\sim}{\beta}}{\sigma}\right)\right]$

(Using Result: If A is an idempotent matrix and $\underset{\sim}{Z} \sim N(\underset{\sim}{\mu}, I_n)$ then, $\underset{\sim}{Z}^T A \underset{\sim}{Z} \sim \chi^2_{\text{rank}(A)}, \underset{\sim}{\mu}^T A \underset{\sim}{\mu})$

(rank(A) → df; $\underset{\sim}{\mu}^T A \underset{\sim}{\mu}$ → ncp.)

Now, $df = n-2$ ✓

$ncp = (\underset{\sim}{\beta}^T X^T (I_n - P_X) X \underset{\sim}{\beta}) \frac{1}{\sigma^2}$

$= \underset{\sim}{\beta}^T 0 \underset{\sim}{\beta} \frac{1}{\sigma^2} = 0$ ✓

For 2 variable case (i.e. $Y = \beta_0 + \beta_1 X + \epsilon$)

$\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}})\sigma^2)$

$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{s_{xx}})$

$\frac{RSS}{\sigma^2} \sim \chi^2_{df=n-2, ncp=0}$

$\hat{\sigma}^2 = \left(\frac{RSS}{n-2}\right) = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$

$E(\hat{\sigma}^2) = \frac{E(RSS)}{n-2} = \frac{(n-2)\sigma^2}{n-2} = \sigma^2$ ($\because$ unbiased estimator of σ^2)

① $\underset{\sim}{e}$ and $\underset{\sim}{\hat{Y}}$ are orthogonal (trivial) to each other, & ($\because (P_X \underset{\sim}{Y}) \cdot (I_n - P_X)\underset{\sim}{Y} = \underset{\sim}{Y}^T P_X^T (I_n - P_X)\underset{\sim}{Y} = 0$)

② they are distributionally independent (using, $cov(AY, BY) = (AIB^T)\sigma^2$)

(Proof: here, $cov(\underset{\sim}{\hat{Y}}, \underset{\sim}{e})$ ↵

$= cov(\underbrace{P_X \underset{\sim}{Y}}_{A}, \underbrace{(I_n - P_X)\underset{\sim}{Y}}_{B}) = [P_X I (P_X^T - I)]\sigma^2$

$= (P_X P_X^T - P_X)\sigma^2$

$= (P_X^2 - P_X)\sigma^2 = 0$ ($\because P_X^T = P_X = P_X^2$))

Prediction for new x_0 (or, x_0 for 2 var case)

Recall, we have data $\{(x_i, y_i) \mid i = 1, 2 \ldots n\}$ (2 variable case)

From there we get $\hat{\beta}_0, \hat{\beta}_1$ as LS estimate

so the prediction line or regression line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$

Now,

For a new value x_0 under the same model, we can predict the value of y_0 as follows:-

(that is, the same model from which the data (x_i, y_i) was obtained)

$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ (Not same as $y_0 = \beta_0 + \beta_1 x$)

* Apart from the $var(\epsilon) = \sigma^2$, we also have the new sources of uncertainty which impact the distribution of $\hat{y}_0$

(Although, $y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$; $\hat{y}_0$ will have different variance)

Reason: $\hat{\beta}_0$, $\hat{\beta}_1$

So, what is the distribution of $\hat{y}_0$?? (larger) In fact.

Ans: $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ (x_0 → constant)

($\hat{\beta}_0$ → Normal, $\hat{\beta}_1$ → Normal) linear comb^n.

$\therefore \hat{y}_0$ will follow Normal distr^n. So, just find $E(\hat{y}_0)$ and $V(\hat{y}_0)$ & you are done.

easy

$V(\hat{y}_0)$ involves $cov(\hat{\beta}_0, \hat{\beta}_1)$ (which need not be zero (as per previous (last page, last section))

- $E(\hat{y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = (\beta_0 + \beta_1 x_0)$
- $V(\hat{y}_0) = V(\hat{\beta}_0) + V(\hat{\beta}_1)\cdot x_0^2 + 2\,cov(\hat{\beta}_0, \hat{\beta}_1)\cdot x_0$

$$= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{x_0^2}{S_{xx}} - \frac{2\bar{x}x_0}{S_{xx}}\right]$$

$$= \sigma^2\left[\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}\right]$$

$\therefore \hat{Y}_0 \sim N\left(\beta_0+\beta_1 x_0,\ \sigma^2\left(\frac{1}{n} + \frac{(\bar{x}-\bar{x}_0)^2}{S_{xx}}\right)\right)$

↳ Not same as σ^2 (i.e. $V(Y_0)$)

As promised, we have different variance (infact larger).

$cov(\hat{\beta}_0, \hat{\beta}_1) = ?$

$\hat{\beta}_1 = \sum_i \left(\frac{x_i-\bar{x}}{S_{xx}}\right) y_i = (k_1 y_1 + k_2 y_2 + \dots)$

$\hat{\beta}_0 = \sum_i \left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{S_{xx}}\right) y_i = (g_1 y_1 + g_2 y_2 + \dots)$ → g_i

only self terms will have non-zero covariance (i.e. $cov(y_i, y_j) = 0\ (i \neq j)$, $= \sigma^2\ (i=j)$)

$\therefore cov(\hat{\beta}_0, \hat{\beta}_1) = \sum k_i g_i \sigma^2$

$= \sigma^2\left(\frac{\sum k_i}{n} - \bar{x}\sum k_i^2\right)$ ($\sum k_i = 0$, $\sum k_i^2 = 1/S_{xx}$) $= -\sigma^2\frac{\bar{x}}{S_{xx}}$.

Testing of hypothesis.

- $H_0 \rightarrow \beta_0 = b_0$, $H_1 \rightarrow \beta_0 \neq b_0$

Now, $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$

↳ Under H_0: $\frac{\hat{\beta}_0 - b_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$

* Problem: We don't know the true value of σ^2

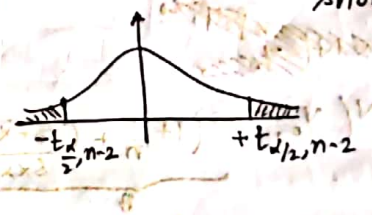
But, we just know an estimate → $\hat{\sigma}^2$

$\hat{\sigma}^2 = \frac{1}{n-2}\sum (y_i - \bar{y})^2$

(= $S_{yy} - \frac{S_{xy}^2}{S_{xx}}$ (final result, computationally shortcut))

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$

↳ Under H_0: $\frac{\hat{\beta}_0 - b_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$



$\therefore$ Now,

if $|t_{computed}$ (or t stat)$| > t_{\frac{\alpha}{2}, n-2}$,

we reject [null hypothesis (H_0)] at level of significance α.
[in favour of H_1]

- consequently, we can quickly obtain the $(1-\alpha)$ confidence Interval for β_0
- similarly, $H_0 \rightarrow \beta_1 = b_1$, $H_1 \rightarrow \beta_1 \neq b_1$

Now, $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2\left(\frac{1}{S_{xx}}\right)\right)$

but σ^2 is unknown,

so, we rather use: $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2\left(\frac{1}{S_{xx}}\right)}} \sim t_{n-2}$

↳ Under H_0: $\frac{\hat{\beta}_1 - b_1}{\sqrt{\hat{\sigma}^2\left(\frac{1}{S_{xx}}\right)}} \sim t_{n-2}$

$\therefore$ Now,

if $|t\text{-stat}| > t_{\frac{\alpha}{2}, n-2}$

we reject [null hypothesis (H_0)] at level of significance α.
[in favour of H_1]

- once again, we can obtain the $(1-\alpha)$ confidence interval of β_1 as follows:-

$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$ (T)

$\Rightarrow P\left(-t_{\frac{\alpha}{2}, n-2} < T < t_{\frac{\alpha}{2}, n-2}\right) = 1-\alpha$

$\Rightarrow P\left(\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right) = 1-\alpha$

* (Interval is a random variable) ⇒

- $$P\left(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right) = 1-\alpha$$

Note: For a given sample we have a given realization of this interval & in that context, the probability that β_1 is contained in it or not is either 1 or 0; but for the interval estimator the probability is $1-\alpha$ (ie if you keep taking out random samples & keep realizing 1 or 0, the frequency of 1 on avg./long run will be $1-\alpha$) ↩ Informally

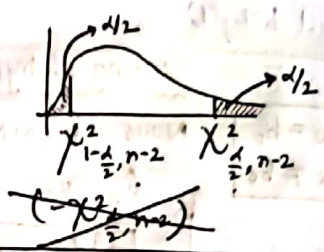
* similarly,

- $$P\left(\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)} < \beta_0 < \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}\right) = 1-\alpha$$

* what about σ^2?? (ie $(1-\alpha)$ CI for σ^2?)

Sol^n: Now, $\frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$ (caution: χ^2 is Not a symmetric distr^n (unlike t distr^n))

$\because P\left(\chi^2_{1-\frac{\alpha}{2},n-2} < \frac{RSS}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-2}\right) = 1-\alpha$



$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$ or, $P\left(\frac{1}{\chi^2_{1-\frac{\alpha}{2},n-2}} > \frac{\sigma^2}{RSS} > \frac{1}{\chi^2_{\frac{\alpha}{2},n-2}}\right) = 1-\alpha$

or, $P\left(\frac{RSS}{\chi^2_{\frac{\alpha}{2},n-2}} < \sigma^2 < \frac{RSS}{\chi^2_{1-\frac{\alpha}{2},n-2}}\right) = 1-\alpha$

(Note: $RSS = \sum_i (y_i - \hat{y}_i)^2$, $\hat{\sigma}^2 = RSS/n-2$)

$\left(\equiv P\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2},n-2}} < \sigma^2 < \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2},n-2}}\right) = 1-\alpha\right)$

* Finally, PREDICTION INTERVAL (we don't call it by the generic term "confidence interval" to distinguish that we are no longer constructing an interval for a parameter (ie Y_0 is Not a parameter) but a random variable itself.

Now, $\hat{Y}_0 \sim N\left(Y_0, \sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$

$\sigma^2_{\hat{Y}_0}$

or, $\hat{Y}_0 - Y_0 \sim N(0, \sigma^2_{\hat{Y}_0})$

or, $\frac{\hat{Y}_0 - Y_0}{\sqrt{\sigma^2_{\hat{Y}_0}}} \sim N(0,1)$ but σ^2 is unknown (& hence $\sigma^2_{\hat{Y}_0}$ is unknown)

∴ we rather use: $\frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2_{\hat{Y}_0}}} \sim t_{n-2}$ $\left(\hat{\sigma}^2_{\hat{Y}_0} = \hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$

$\Rightarrow P\left(-t_{\frac{\alpha}{2},n-2} < \frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2_{\hat{Y}_0}}} < t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha$

$\Rightarrow P\left(\hat{Y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2_{\hat{Y}_0}} < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2_{\hat{Y}_0}}\right) = 1-\alpha$

Carefully note:
$\sigma^2_{\hat{Y}_0}$ (v/s $\sigma^2_{Y_0}$)
&
$\hat{\sigma}^2_{\hat{Y}_0}$

Maximum Likelihood Estimation of regression parameters.

Here, 'even for point estimation' we need the distributional assumption;

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim$ iid $N(0, \sigma^2)$

$\Rightarrow$ **$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and independent** (already known)

Now,

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)_{mle} = \underset{\beta_0, \beta_1, \sigma^2}{\arg\max} \prod_{i=1}^{n} \frac{e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma}$$

$$= \underset{\beta_0, \beta_1, \sigma^2}{\arg\max} \frac{e^{-\frac{1}{2\sigma^2}\sum (y_i - \beta_0 - \beta_1 x_i)^2}}{(\sqrt{2\pi}\,\sigma)^n}$$

Observe: (if we take logarithm to simplify)

maximizing log likelihood $\equiv$ minimizing $\underbrace{\sum (y_i - \beta_0 - \beta_1 x_i)^2}_{RSS}$ $\equiv$ OLS estimate.

(In above case)

$\therefore \hat{\beta}_{0,LS} = \hat{\beta}_{0,mle}$, $\hat{\beta}_{1,LS} = \hat{\beta}_{1,mle}$; but $\hat{\sigma}^2_{LS} = RSS/n-2$ whereas $\hat{\sigma}^2_{MLE} = RSS/n$.

↳ unbiased estimator. ↳ ~~unbi~~ **biased**

When X too, is a random variable.

Suppose, (X, Y) is a paired random variable with Bivariate Normal distr^n $(N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}))$

Given, data set: $D = \{(x_i, y_i), i = 1, 2 \dots n\}$

1.) The regression model → **$E(Y|X=x)$**

2.) Similarly, we could've chosen to study the other potential regression model → $E(X|Y=y)$)

// Anyways, the results will be symmetric

Now,

$$f(x,y) = \frac{e^{\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y}\right\}}}{2\pi\,\sigma_x\,\sigma_y\sqrt{1-\rho^2}}$$

I we want $E(Y|X=x)$: [So, take out "$e^{\frac{-1}{2(1-\rho^2)}\left\{\frac{x-\mu_x}{\sigma_x}\right\}^2}$" and "complete remaining square" in terms of $e^{\frac{-1}{2(1-\rho^2)}\left\{y - \text{◎}\right\}^2}$ & obtain $E(Y|X=x)$]!

$$\text{or, } \underbrace{f(x,y)}_{\text{joint distr}^n} = \left[\frac{e^{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{y-\mu_y}{\sigma_y}\right) - \rho\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^2}}{\sqrt{2\pi}\left(\sqrt{1-\rho^2}\,\sigma_y\right)}\right]\left[\underbrace{\frac{e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2}}{\sqrt{2\pi}\,\sigma_x}}_{\text{marginal distr}^n \text{ of } X}\right]$$

Now, we need conditional distr^n of $(Y|X=x)$

$$f_{Y|X=x}(y|x) = \left(\frac{\text{joint of } X \& Y}{\text{marginal of } X}\right) = \left[\frac{e^{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{y-\mu_y}{\sigma_y}\right) - \rho\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^2}}{\sqrt{2\pi}\,\left(\sigma_y\sqrt{1-\rho^2}\right)}\right]$$

$$= \frac{e^{\frac{-1}{2(1-\rho^2)\sigma_y^2}\left[y - \left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)\right]^2}}{\sqrt{2\pi}\,\left(\sigma_y\sqrt{1-\rho^2}\right)} \longleftrightarrow N\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x),\ \sigma_y^2(1-\rho^2)\right)$$

(realize that)

$$\therefore \boxed{(Y|X=x) \sim N\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x),\ \sigma_y^2(1-\rho^2)\right)}$$

↳ $E(Y|X=x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$ and, $V(Y|X=x) = \sigma_y^2(1-\rho^2)$ (conditional variance)

$\star$ ($\leq$ uncond^l variance)

or, $\boxed{E(Y|X=x) = \underbrace{\left(\mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x\right)}_{\beta_0} + \underbrace{\left(\rho\frac{\sigma_y}{\sigma_x}\right)}_{\beta_1} x}$

Now, testing whether x is significant or not

≡ whether $\beta_1 = 0$ or not

≡ whether **$\rho = 0$ or not**

P.T.O.

$H_0: \rho = 0$
$H_1: \rho \neq 0$

Now, $\hat{\rho} = r = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$

and, under H_0 :

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2, ncp=0}$$

(Generalized result:

$H_0: \rho = \rho_0$
$H_1: \rho \neq \rho_0$

Now, define $z = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right) \left(\equiv \tanh^{-1}(r)\right)$

$\mu_z = \frac{1}{2} \log\left(\frac{1+\rho_0}{1-\rho_0}\right) \left(\equiv \tanh^{-1}(\rho_0)\right)$

$\sigma_z^2 = (n-3)^{-1}$

and, $\left(\frac{z - \mu_z}{\sigma_z}\right) \sim N(0,1)$ when $n \to \infty$

[(based on the famous variance stabilization formula (& its special cases Box-Cox method)

: $T_n \xrightarrow{p} \mu$

then, $\sqrt{n}(T_n - \mu) \xrightarrow{d} N(0, \sigma^2)$

then, $\sqrt{n}(g(T_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2 (g'(\mu))^2)$

* "We reject H_0 in favour of H_1" vs (~~we reject H_0~~) → meaningless!

There is no such thing as "Rejecting H_0"

It is always "in relative/ reference/ favour of H_1"

MULTIPLE LINEAR REGRESSION

$$\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & & & & & \\ \vdots & & & & & \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix}_{n\times(k+1)} \leftarrow \text{vars \& intercept}$$

$$\underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{(k+1)\times 1}, \quad \underset{\sim}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}, \quad \underset{\sim}{\epsilon} \sim N(\underset{\sim}{0}, \sigma^2 I_n)$$

LS condition: $S^2 = (\underset{\sim}{Y} - X\underset{\sim}{\beta})^T (\underset{\sim}{Y} - X\underset{\sim}{\beta})$

$* \quad \hat{\underset{\sim}{\beta}} = \underset{\beta \in \mathbb{R}^{k+1}}{\operatorname{argmin}} S^2(\underset{\sim}{\beta})$

we get $\hat{\underset{\sim}{\beta}}_{LS} = (X^TX)^{-1}X^T\underset{\sim}{Y}$ (assume $|X^TX| \neq 0$)

Now, $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ v/s H_1: atleast one is non-zero ↔ whether we should build a model or not??

→ Under H_0: $y_i \overset{iid}{\sim} N(\beta_0, \sigma^2)$ (NOT $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$)

Approach (to test this):

ANOVA (does the model explain any variability in response variable (y))

* The famous ANOVA identity equation

$TSS = ESS + RSS$

(Proof)↓

$TSS = \sum (y_i - \bar{y})^2$ → $\sum (y_i - \bar{y})^2$ (Total sum of squares)

$ESS = \sum (\hat{y}_i - \bar{y})^2$ → $\sum (\hat{y}_i - \bar{y})^2$ (Explained sum of squares)

$RSS = \sum (y_i - \hat{y}_i)^2$

In matrix notation:

$$TSS \equiv \underset{\sim}{Y}^T \left(I_n - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T\right)\underset{\sim}{Y}$$

$$ESS \equiv \underset{\sim}{Y}^T \left(P_X - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T\right)\underset{\sim}{Y}$$

and, $RSS \equiv \underset{\sim}{Y}^T (I_n - P_X)\underset{\sim}{Y}$

(show that $TSS = ESS + RSS$) : $ESS + RSS = \underset{\sim}{Y}^T (P_X - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T)\underset{\sim}{Y} + \underset{\sim}{Y}^T(I_n - P_X)\underset{\sim}{Y}$

$= \underset{\sim}{Y}^T(P_X - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T + I_n - P_X)\underset{\sim}{Y}$

Now, we are ready to apply Cochrane's th^m $= \underset{\sim}{Y}^T(I_n - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T)\underset{\sim}{Y} = TSS$ (hence proved)

* (Recall we had said, "key result in ANOVA").

- $\left(\frac{RSS}{\sigma^2}\right) \sim \chi^2_{n-(k+1)}$, ncp=0
- $\left(\frac{ESS}{\sigma^2}\right) \sim \chi^2_k$, ncp=$\lambda$
- $\left(\frac{TSS}{\sigma^2}\right) \sim \chi^2_{n-1}$, ncp=$\lambda$

(Note: $A = A_1 + A_2 + \dots A_k$. then, $rank(A) = \sum rank(A_i)$ only when A_i's are symmetric idempotent)

What is λ?

$ncp = \lambda = \frac{1}{\sigma^2}(X\underset{\sim}{\beta})^T(P_X - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T)(X\underset{\sim}{\beta})$

$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T(X^TP_XX - \frac{1}{n}X^T\underset{\sim}{1}\underset{\sim}{1}^TX)\underset{\sim}{\beta} = \frac{1}{\sigma^2}\underset{\sim}{\beta}^T(X^TX - \frac{1}{n}X^T\underset{\sim}{1}\underset{\sim}{1}^TX)\underset{\sim}{\beta}$ ($\because P_XX = X$)

(To conveniently solve this, we partition X as $X = (\underset{\sim}{1} : X_R)$ & $\underset{\sim}{\beta} = \begin{pmatrix}\beta_0 \\ \underset{\sim}{\beta}_R\end{pmatrix}$)

$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left(\begin{bmatrix} n & 1^TX_R \\ X_R^T1 & X_R^TX_R\end{bmatrix} - \frac{1}{n}\begin{pmatrix}1^T \\ X_R^T\end{pmatrix}(1\ 1^T)(1\ X_R)\right)\underset{\sim}{\beta}$

$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left(\begin{bmatrix} n & 1^TX_R \\ X_R^T1 & X_R^TX_R\end{bmatrix} - \frac{1}{n}\begin{pmatrix}1^T \\ X_R^T\end{pmatrix}(n\ 1^TX_R)\right)\underset{\sim}{\beta}$

$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left(\begin{bmatrix} n & 1^TX_R \\ X_R^T1 & X_R^TX_R\end{bmatrix} - \begin{bmatrix} n & 1^TX_R \\ X_R^T1 & \frac{1}{n}X_R^T1\,1^TX_R\end{bmatrix}\right)\underset{\sim}{\beta}$

$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\begin{pmatrix} 0 & \underset{\sim}{0}^T \\ \underset{\sim}{0} & X_R^TX_R - \frac{1}{n}X_R^T1\,1^TX_R\end{pmatrix}\underset{\sim}{\beta}$

$= \frac{1}{\sigma^2}\underset{\sim}{\beta}_R^T(X_R^TX_R - \frac{1}{n}X_R^T1\,1^TX_R)\underset{\sim}{\beta}_R$ (i.e. independent of β_0)

$= \frac{1}{\sigma^2}\underset{\sim}{\beta}_R^TX_R^T(I_n - \frac{1}{n}1\,1^T)X_R\underset{\sim}{\beta}_R$

Now, the ncp = 0 (λ) iff all $\beta_1, \beta_2 \dots \beta_k = 0$ (i.e. $\underset{\sim}{\beta}_R = \underset{\sim}{0}$) (Not β_0)

which is nothing but the null hypothesis of F test ($H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$)

Moving forward,

To test: $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
H_1: atleast one is non-zero

Now, $\frac{ESS}{\sigma^2} \sim \chi^2_{k, ncp=\lambda}$

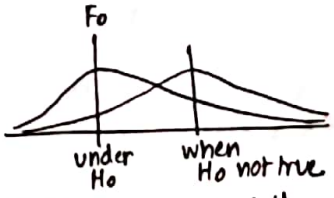
but, we don't know σ^2

so, instead we have: $\hat{\sigma}^2 = RSS/n-(k+1)$

Now, $\frac{ESS/k}{RSS/n-(k+1)} \sim F_{k, n-k-1, ncp=\lambda}$.

and, under H_0 ($\lambda = 0$): $F_0 = \left[\frac{ESS/k}{RSS/n-k-1}\right] \sim F_{k,n-k-1}$

(∴ central F distr^n)
(v/s. noncentral F).



∴ we reject null hypothesis in favour of H_1 iff $F_0 > F_{\alpha,k,n-k-1}$ at α level of ~~zero~~ significance

(Hence, we are studying 'mean' through variance, & hence the name is "Analysis of variance" for mean)

↳ heavily employed in Design & Analysis of Experiments

• Now, (testing individual variable's significance) → C

$X = \begin{bmatrix} 0 & 1 & 2 & 3 & \cdots & k \\ 1 & 1 & & & & \end{bmatrix}$

① $H_0: \beta_j = b_j$ v/s $H_1: \beta_j \neq b_j$

where, $\hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)$

we need to extract out the distribution of $\hat{\beta}_j$

$\Rightarrow \hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{j+1,j+1})$

and, under H_0 $(\beta_j = b_j)$: $T = \dfrac{\hat{\beta}_j - b_j}{\hat{\sigma}\sqrt{C_{j+1,j+1}}} \sim t_{n-k-1}$, where $\hat{\sigma}^2 = \dfrac{RSS}{n-k-1}$

$\therefore$ we reject H_0 in favour of H_1, iff $|T| > t_{\frac{\alpha}{2}, n-k-1}$

• Let's test more general hypotheses!

② $H_0: \beta_j - \theta\beta_i = b$ v/s $H_1: \beta_j - \theta\beta_i \neq b$

•, ③ $H_0: \beta_j - \theta\beta_i - k\beta_l = b$ v/s $H_1: \beta_j - \theta\beta_i - k\beta_l \neq b$,... or, in general any linear combination of β_i's.

$\therefore$ $H_0: l^T\beta = b$ v/s $H_1: l^T\beta \neq b$ (i.e. $l^T\beta = b$ v/s $l^T\beta \neq b$)

(In fact ①, ②, ③ are all special cases, where l is chosen accordingly;

→ ①: $l = (0, \ldots, 1 \leftarrow j^{th}\text{ element}, \ldots, 0)$

→ ②: $l = (0, \ldots, 1 \leftarrow j^{th}\text{ element}, \ldots, -\theta \leftarrow i^{th}\text{ element}, \ldots, 0)$ & so on.

Now, $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

$\therefore$ $l^T\hat{\beta} \sim N(l^T\beta, \sigma^2 l^T(X^TX)^{-1}l)$

#(using: the result in multivariate distribution analysis: if $Y \sim N(\mu, \Sigma)$ then $l^TY \sim N(l^T\mu, l^T\Sigma l)$)

Something crucial:-

$\hat{\beta} \sim N(\beta, (X^TX)^{-}\sigma^2)$ — $(X^TX)^{-}$: generalized inverse.

If $|X^TX| \neq 0$, then any linear parametric function of β in the form $p^T\beta$ is ESTIMABLE.

(i.e. only then, there exists a linear combination of Y (eg. l^TY) such that $E(l^TY) = p^T\beta$)

// Motivation: For testing, we first need to do estimation, (i.e. develop estimator(s)) & estimators exist only if the parametric function is ESTIMABLE) — unbiased

Best Linear unbiased estimator (BLUE)

An unbiased estimator of $p^T\beta$ is said to be the BLUE iff it has minimum variance among all linear unbiased estimators of $p^T\beta$

$$U = \{ l^T Y \mid E(l^T Y) = p^T\beta \quad \forall \beta \in \mathbb{R}^{k+1} \}$$

$$V(l_0^T Y) \le V(l^T Y) \quad \forall l, l_0 \in U$$

Theorem: A linear function of Y is BLUE of its expectation iff it is uncorrelated with all linear zero function (LZF)

$E(l^T Y) = p^T\beta$

iff $cov(l^T Y, l_*^T Y) = 0$ where, $l_*^T Y$ is a LZF

(*) **Very important result:** If $l^T Y$ is a LUE of $p^T\beta$ then the corresponding BLUE is $l^T P_X Y$

(gives us a "recipe" to go from LUE → BLUE) (Proof: skipped!)

* Every estimable linear parametric function has a unique BLUE

(Proof: assume $\exists\ l_1^T Y$ and $l_2^T Y$ which are BLUE of $p^T\beta$; & then prove by contradiction)

Polynomial Regression

↳ in variables, but NOT parameters. (i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ ✓ ; $Y = \beta_0 + \beta_1 X + \beta_0\beta_1 X^2 + \epsilon$ ✗)

① Eg. $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

② Eg. $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \beta_4 x_{1i}^2 + \beta_5 x_{2i}^2 + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Note, however that both are "linear" in parameters.
& hence can be analyzed in the same framework as Linear Regression.
just their X matrix shall be different. Eg. ①: $X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$

Eg. ②: $X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} & x_{11}^2 & x_{21}^2 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{1n}x_{2n} & x_{1n}^2 & x_{2n}^2 \end{bmatrix}$

Note: 1) If all x_is are equal then X^TX shall not be invertible.

2) If $\sum |x_i| < \epsilon$, small no. then with higher degrees of x_is the columns will become $\approx$ numerically zero!

3) At most $(n-1)$ degree polynomial, will allow estimation of $\beta_0, \beta_1, \ldots \beta_{n-1}$
At most $(n-2)$ degree polynomial, will allow estimation of " " ... " & σ^2.

• $k+2$ parameters (slopes & intercept & σ^2)

∴ At most: $k+2 \le n$

$k \le n-2$

Model adequacy checking

$E(\underset{\sim}{Y}) = X\underset{\sim}{\beta}$, $\underset{\sim}{\beta}$ is model parameters

$\underset{\sim}{Y} \sim N(X\underset{\sim}{\beta}, I_n\sigma^2)$, σ^2 unknown.

~~errors are uncorrelated~~

~~errors are normally distributed~~ errors are independent

coefficient of determination (R^2) = Explained variation by model (ESS) / Total variation in response (TSS) $\left(\equiv \boxed{\frac{ESS}{TSS}}\right)$

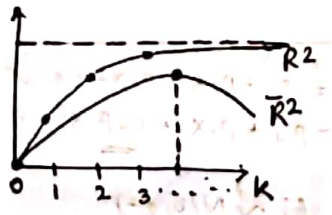
$$= \frac{\sum_i (\hat{y}_i - \bar{\hat{y}}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \quad (\because \bar{\hat{y}} = \bar{y}, \text{ Under OLS})$$

- Now, because
  $TSS = ESS + RSS$
  or, $1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$
  or, $1 = R^2 + \left(\frac{RSS}{TSS}\right)$ b/w 0 & 1 $\therefore \boxed{R^2 \in [0,1]}$ $\left(R^2 = 1 - \frac{RSS}{TSS}\right)$

- If we increase the number of regressors, (such that X^TX remains invertible)
  "R^2 cannot decrease" (ie either constant or increase)
  [(∵ the very objective of OLS estimation is Min. RSS $\equiv$ Min. $\frac{RSS}{TSS}$ (∵ TSS constant for given sample; irrespective of prediction or parameter estimates)
  $\equiv$ Max. $\left(1 - \frac{RSS}{TSS}\right)$
  $\equiv$ Max. R^2)
  ↳ becomes 1 when $k = n-1$

$\therefore$ we cannot use R^2 as a means to decide the adequate no. of regressor variables (k).

- Adjusted R^2 $(\bar{R}^2) = 1 - \frac{RSS/df(RSS)}{TSS/df(TSS)} = 1 - \frac{RSS/n-k-1}{TSS/n-1}$ $\leq R^2$ $(\because k \geq 0)$



Even though this (so to say) gives us an 'optimal' number of regressors, but how to "pick" which shall be those regressors from a set of several potential explanatory vars.? (Umm.. Try out all ^pC_k possibilities (say, p no. of potential regressors) & compare their $\bar{R}^2$?)

- Error Analysis / Residual Analysis

We try to get the best idea about the 'error' from our known 'residual':

* Assmns made about errors → can't be observed
  - uncorrelated
  - zero mean
  - constant variance

∴ Let's use 'residuals' as their proxy & verify these assmns on the sample:

$$\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} = (I - P_X)\underset{\sim}{Y} \sim N(0, \sigma^2(I - P_X))$$

$$\frac{\underset{\sim}{e}^T\underset{\sim}{e}}{\sigma^2} \sim \chi^2_{n-k-1}$$

$e_i = (y_i - \hat{y}_i) \sim N(0, \sigma^2(1 - h_{ii}))$; assuming $H = P_X$.

$Corr(e_i, e_j) = cov(y_i - \hat{y}_i, y_j - \hat{y}_j)$
$= \sigma^2(-h_{ij}) = \sigma^2(I - H)_{ij}$

∴ estimated errors are correlated in general.

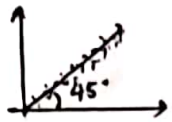
Define: $r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$ (Standardized)

$\Rightarrow r_i$ will follow $N(0,1)$ when $n \uparrow \uparrow \infty$
because $\hat{\sigma}^2 \to \sigma^2$ with prob. 1

If we plot the histogram then we are expected to get pdf of standard Normal.

we also can do "q-q plot for r_i's (i.e. quantile-quantile plot)

If the errors are normally distributed, then only q-q plot will give a straight line passing through origin with slope 1.



q = x such that $P(Z \le x) = p$

r_i such that $\frac{\#\{r_i < x\}}{n} = p$

|| or, we could do a "p-p" plot (probability-probability plot)

∴ If the plot is different at all, then this indicates Non-normality (i.e. violation of normality assumption)

- Now, suppose we define $e_{(i)} = y_i - \hat{y}_{(i)}$, prediction error of y_i but the same (y_i, x_i) has been removed to get $\hat{\beta}$, $\hat{y}_i$.

* If we remove (y_i, x_i) from the dataset, run the OLS, get $\hat{\beta}$ and hence "$\hat{y}_{(i)}$" (on the $(n-1)$ dataset)

$\hat{y}_{(i)}$ = predicted value of y_i based on $(n-1)$ observations.

$\sum_{i=1}^{n} e_{(i)}^2$ = Predicted Residual error sum of squares

or, $\sum_i (y_i - \hat{y}_{(i)})^2$ = PRESS. (a.k.a "Leave one out" or "Jack knife")

we can show that, $e_{(i)} = \frac{e_i}{1-h_{ii}} \sim N\left(0, \frac{\sigma^2}{1-h_{ii}}\right)$

standardize $\frac{e_{(i)}}{\sqrt{\sigma^2/1-h_{ii}}} = \frac{e_i/1-h_{ii}}{\sqrt{\sigma^2/1-h_{ii}}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$

using $\hat{\sigma}^2$ (based on n-1 obsns) ← but "not" really (∵ here, we have to get $\hat{\sigma}^2$ based on $(n-1)$ obsns v/s (n) obsns in here)

$\frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$ ← same as (previous page's)

$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$

* Q. Do we really have to estimate the model n times to obtain PRESS?

Ans: Obviously NOT! The results beside come to our aid.

(Ofcourse, you need to prove the statement "we can show that, $e_{(i)} = \frac{e_i}{1-h_{ii}} \sim N\left(0, \frac{\sigma^2}{1-h_{ii}}\right)$")

Long matrix based proof!! (Ofcourse, I've skipped it)

and, similarly we can show that:

$$\hat{\sigma}^{2*}_{\text{using } (n-1) \text{ obsns}} = \left[\frac{(n-k-1)MSRes - \frac{e_i^2}{1-h_{ii}}}{n-k-2}\right]$$

, where MSRes is obviously $\left(\frac{SSRes}{n-k-1}\right)$ the original SSRes using n obsns

---

(Q:) How would you test if i^{th} observation is an outlier?

→ $H_0: \mu = 0$

$H_1: \mu \neq 0$

$y_i - \hat{y}_{(i)} = e_{(i)} \sim N\left(0, \frac{\sigma^2}{1-h_{ii}}\right)$

$\frac{e_{(i)}}{\sqrt{\sigma^2/1-h_{ii}}} \sim N(0,1)$ under H_0.

$\Rightarrow \frac{e_{(i)}}{\sqrt{s_{(i)}^2/1-h_{ii}}} \sim t_{n-k-2}$