Assumption: $E(Y|X)$ is linear in the inputs $X_1, \dots, X_p$.

* best for : small no. of training cases
low signal-to-noise ratio
Sparse.

§ 2. Linear Regression models and least squares.

$X^T = (X_1, X_2, \dots, X_p)$: input $\longrightarrow$ Y : output.

linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

β_j : unknown parameters or coefficients

X_j : come from different sources.

(i) quantitative inputs.
(ii) transforms of quantitative inputs (log, $\sqrt{}$, square)
(iii) basic expansions such as: $X_2 = X_1^2$, $X_3 = X_1^3$
leading to polynomial representation.
(iv) numeric or "dummy" coding of the levels of qualitative inputs.
(v) interactions between variables ($X_3 = X_1 \cdot X_2$)

- Model is linear in parameters, irrespective of source.

data : $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$

parameters: $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$

features : $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$

Residual Sum of Squares : $$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j\right)^2$$

- From statistical p.o.v., this criterion is reasonable if the training observations (x_i, y_i) represent independent random draws from their population.

- Even if x_i's are not drawn randomly drawn, criterion is still valid if $y_i | x_i$ are independent (i.e. y_i's are conditionally independent)

- RSS: measures the average error of fit.

$$RSS(\beta) = (y - X\beta)^T (y - X\beta) \longrightarrow \text{quadratic function in } p+1 \text{ parameters}$$

Diff wrt β:

$$\frac{\partial RSS}{\partial \beta} = -2X^T(y - X\beta)$$

$$\Rightarrow \frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2X^TX$$

Assuming X has full column rank,

$\Rightarrow X^TX$ is +ve definite.

& We set first ~~fir~~ derivative to zero

$$X^T(y - X\beta) = 0$$

To obtain unique solution:

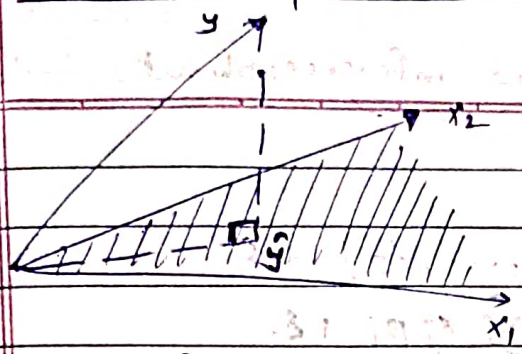
$$\Rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty$$

fitted values of training input:

$$\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty$$

$\hat{y} = Hy$ where $H = X(X^TX)^{-1}X^T \longrightarrow$ as it puts hat on y.

* Geometrical representation:



We minimize $RSS(\beta) = \|y - X\beta\|^2$ by choose $\hat{\beta}$ so that,

$y - \hat{y}$ is orthogonal to $\mathbb{R}^N$.

$\Rightarrow$ $\hat{y}$ is orthogonal projection of y onto ~~$\mathbb{R}^N$~~ subspace (column space of X)

As H computes $\hat{y}$, thus H is orthogonal projection matrix!

- If ~~not~~ all columns of X are linearly independent, then, then X^TX is singular $\Rightarrow$ $\hat{\beta}$ are not uniquely defined.
but, $\hat{y} = X\hat{\beta}$ ~~but~~ are still projection of y onto the column space of X, but there is more than one way to express that projection in terms of column vectors of X.

- To make X^TX non singular: recoding/ dropping redundant columns in X.

- If no of features $>$ no of observations:
  - ↳ filtering
  - ↳ regularization.

* To get sampling properties of $\hat{\beta}$:

Assumptn: (1) y_i uncorrelated
(2) constant variance: $Var(y_i) = \sigma^2$
(3) x_i are fixed (non random).

$Var(\hat{\beta}) = (X^TX)^{-1}\sigma^2$ ⟶ var cov matrix of $\hat{\beta}$.

$$\sigma^2 = E\left[\frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]$$

$\sigma^2 = E(\hat{\sigma}^2)$

* Additional assumptions to draw inferences about parameters of model

(1) $\varepsilon \sim N(0, \sigma^2)$

(2) $Y = E(Y | X_1, \ldots, X_p) + \varepsilon$

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon.$$

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

$$\Rightarrow (N-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}$$

$\Rightarrow \hat{\beta}$ & $\hat{\sigma}^2$ are statistically independent.

* Testing of hypothesis & confidence intervals of β_j

To test: $H_0: \beta_j = 0$ Vs $H_1: \beta_j \neq 0$.

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

where v_j: j^{th} diagonal element of $(X^T X)^{-1}$

Under H_0: $z_j \sim t_{N-p-1}$

→ rejected for large absolute value of t.

→ if $\hat{\sigma}$ is replaced by σ, z_j will have standard normal distribution.

→ as sample size T, tail quantiles of t & Z are almost same, so Z is used.

* Test for significance of groups of coefficients simultaneously.

$$F = \frac{(RSS_0 - RSS_1)/p_1 - p_0}{RSS_1 / N - p_1 - 1}$$

where :

RSS_1 : residual sum of squares for the least squares fit of the bigger model with p_1+1 parameters.

RSS_0 : —"— nested smaller model with p_0+1 parameters.

$p_1 - p_0$ parameters constrained to be 0.

- F statistics: measures the change in residual sum of squares per additional parameter in the bigger model, normalised by estimate of σ^2.

H_0 : smaller model is correct.

$$F \sim f_{p_1-p_0,\, N-p_1-1}.$$

For large N,

$$F \sim \chi^2_{p_1-p_0}$$

* confidence interval for β_j :

$$\left[\hat{\beta}_j - z^{(1-\alpha)} v_j^{1/2} \hat{\sigma},\ \beta_j + z^{(1-\alpha)} v_j^{1/2} \hat{\sigma}\right]$$

where $z^{(1-\alpha)}$ is the $(1-\alpha)$ percentile of normal distribution

* confidence set for β :

$$C_\beta = \left\{ \beta \,\middle|\, (\hat{\beta}-\beta)^T X^T X (\hat{\beta}-\beta) \le \hat{\sigma}^2 \chi^2_{p+1}{}^{(1-\alpha)} \right\}.$$

where : $\chi^2_\ell{}^{(1-\alpha)}$ is the $(1-\alpha)^{th}$ percentile of the chi-squared distribution on ℓ degrees of freedom.

3.2.2 : The Gauss - Markov Theorem.

The least squares estimates of the parameters β have the smallest variance among all linear unbiased estimates.

~~Explanation~~ :

~~Let $\theta = a^T\beta$ be~~

- focus on estimation of any linear combination of the parameters $\theta = a^T\beta$, e.g. predictions $f(x_0) = x_0^T\beta$ are of this form.

$$\hat{\theta} = a^T\hat{\beta} = a^T(X^TX)^{-1}X^Ty.$$

→ $a^T\hat{\beta}$ is unbiased for $a^T\beta$.

$$E(a^T\hat{\beta}) = E(a^T(X^TX)^{-1}X^Ty)$$
$$= a^T(X^TX)^{-1}X^TX\beta$$
$$= a^T\beta.$$

* Gauss Markov Thm :

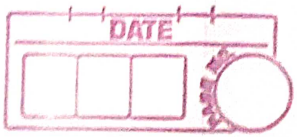
If we have any other linear estimator $\tilde{\theta} = c^Ty$ that is unbiased for $a^T\beta$ i.e. $E(c^Ty) = a^T\beta$ then:

$$Var(a^T\hat{\beta}) \leq Var(c^Ty)$$

⇒ least squares estimator has the smallest MSE of all linear estimators with no bias.

BUT: there might exist a biased estimator with smaller MSE.

→ Least MSE estimator is preferred.



- consider the prediction of the new response at input x_0:

$$Y_0 = f(x_0) + \varepsilon_0$$

Expected prediction error of estimate:

$$\hat{f}(x_0) = x_0^T \hat{\beta}$$

is:

$$E\left(Y_0 - \hat{f}(x_0)\right)^2 = \sigma^2 + E\left(x_0^T \hat{\beta} - f(x_0)\right)^2$$

$$= \sigma^2 + MSE\left(\hat{f}(x_0)\right)$$

(variance of Y_0).