Linear Regression-1

Prof. Asim Tewari
IIT Bombay

Supervised learning

- Other factors:

1. Data heterogeneity: some algorithms require input to be numerical and scaled to similar range (eg. SVM, NN methods. On the other hand decision tree etc. Can handle heterogeneous data.
2. Redundancy in data (highly correlated input variables)
3. Presence of interactions and non-linearities

<u>Eg. Of supervised learning :</u>linear regression, logistic regression, SVM, naive Bayes, linear discriminant analysis, decision tree, k-nearest neighbor algo, neural network etc.

Unsupervised learning

- Test of inferring a function to describe hidden structure.
- Eg. - Clustering (K-mean, mixture models, hierarchies clustering)

  - Anomaly detection etc.

Regression

$\boxed{y = f(x)}$ $(\bar{x}_i, y_i)$

$$\bar{x}_i = \begin{Bmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^p \end{Bmatrix}$$

- Extract a relationship from data

Learning arbitrary function is intractable

H is a class of functions

→ parametric function $w \in R^d$

Parameters or the weights

$$h_w(x) = \boxed{g(x, w)}$$

?

Learning is to find w from data (x, y)

Regression

$$h_w(x) = g(x, w)$$

Case 1 : g_1 : $w_0 + w_1 x$ | $w_0, w_1,$

Case 2 : g_2 : $w_0 + w_1 x + w_2 x^2$ | w_0, w_1, w_2

Case 3 : g_3 : $w_0 + w_1 x + w_2 x^2 + w_3 x^3$ | w_0, w_1, w_2, w_3

Linear Regression

$y = f(x) + \epsilon$ ← Error / Noise

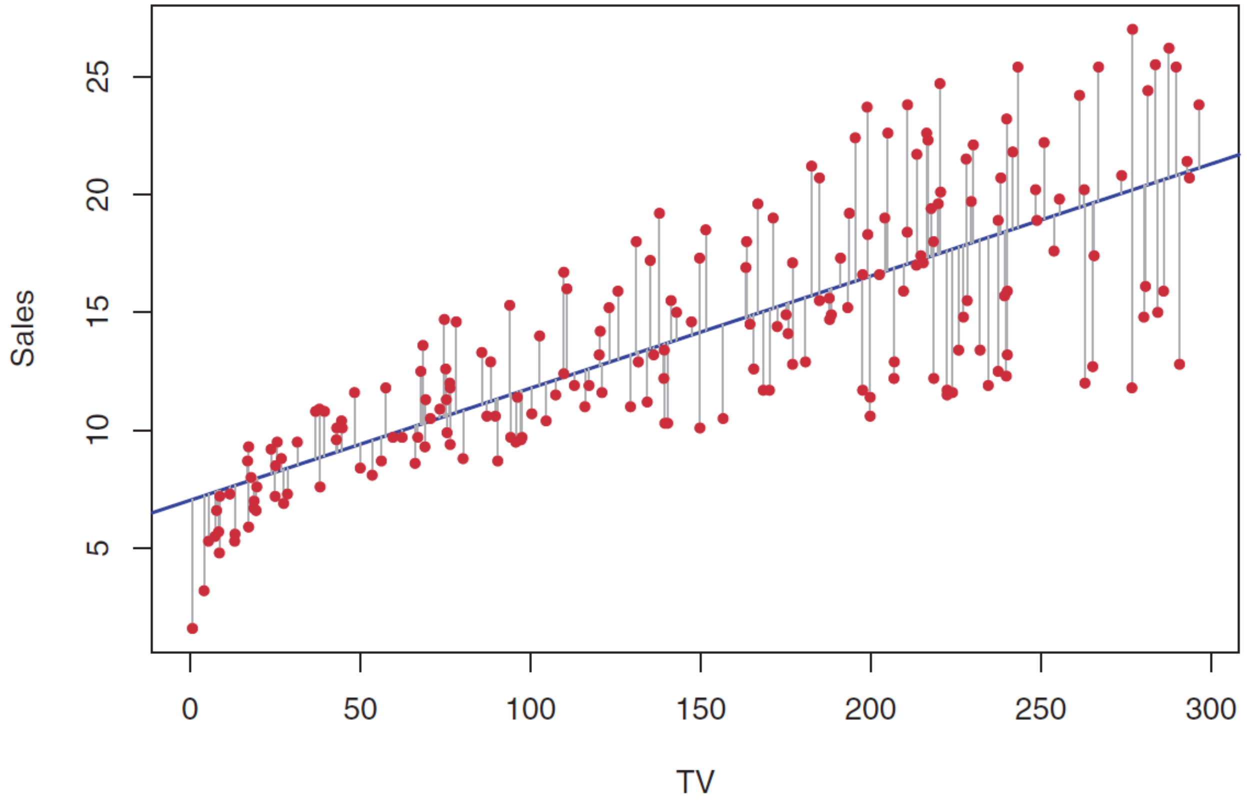
$y = \beta_0 + \beta_1 x + \epsilon$

- It assumes a linear relation between input x and output y

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X$$

Approximately modeled

$\hat{\beta}0$ and $\hat{\beta}1$ are unknown coefficients or parameters which are estimated from training data.



Linear Regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$y = \beta_0 + \beta_1 x + \epsilon$

error (zero mean)

Prediction (estimate) of Y

$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}, x_i \in R^p$

Estimating the coefficients: $p = 1$

- Training data (***x1***, ***y1***), (***x2***, ***y2***)... and (***xn***, ***yn***)
- n data pair
- we need $\hat{\beta}_0$ and $\hat{\beta}$1 ^ such that the linear model fits the data well
- measure of data fits the data well or closeness?

- One possible closeness measure is least square criterion

$e_i = y_i - \hat{y}_i$ - represents ith residual

Linear Regression

Residual sum of squares (RSS)

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

$e_i = y_i - \hat{y}_i$

$\hat{y}_i \rightarrow \hat{\beta}_0 + \hat{\beta}_1 x_i$

Minimize RSS is least square criterion

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

By minimizing RSS we can find

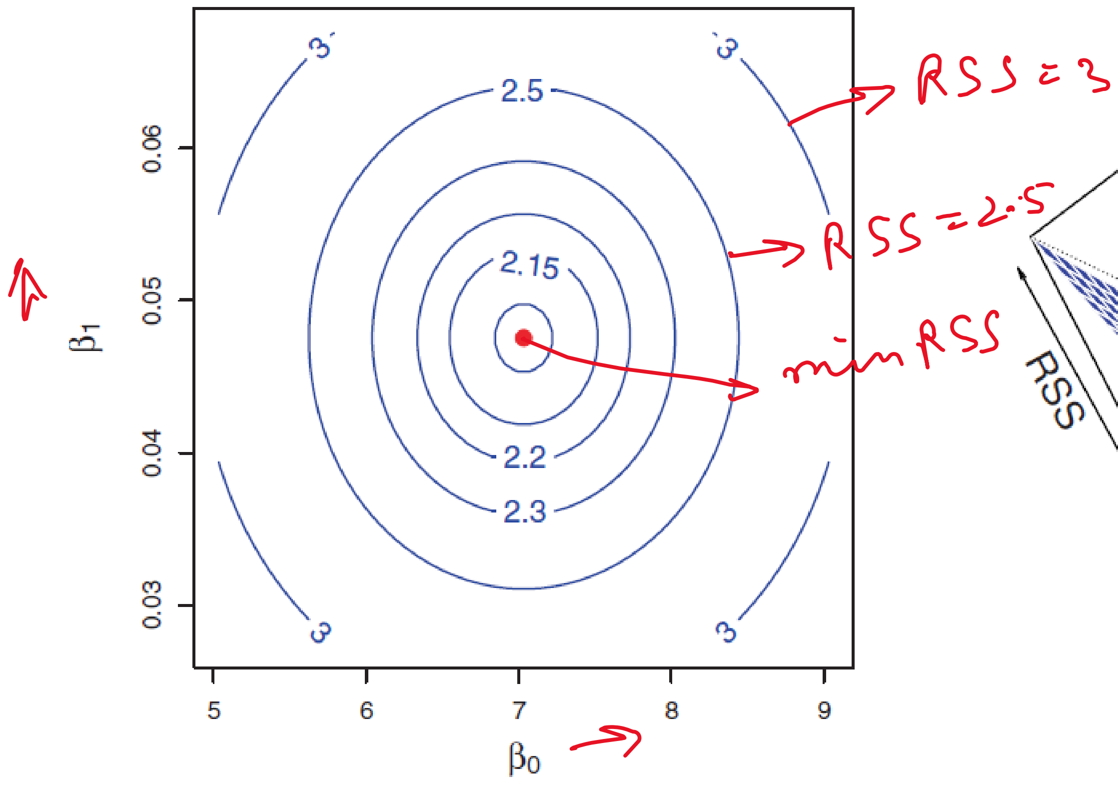
$\frac{\partial \mathrm{RSS}}{\partial \hat{\beta}_1} = 0, \quad \frac{\partial \mathrm{RSS}}{\partial \hat{\beta}_0} = 0$

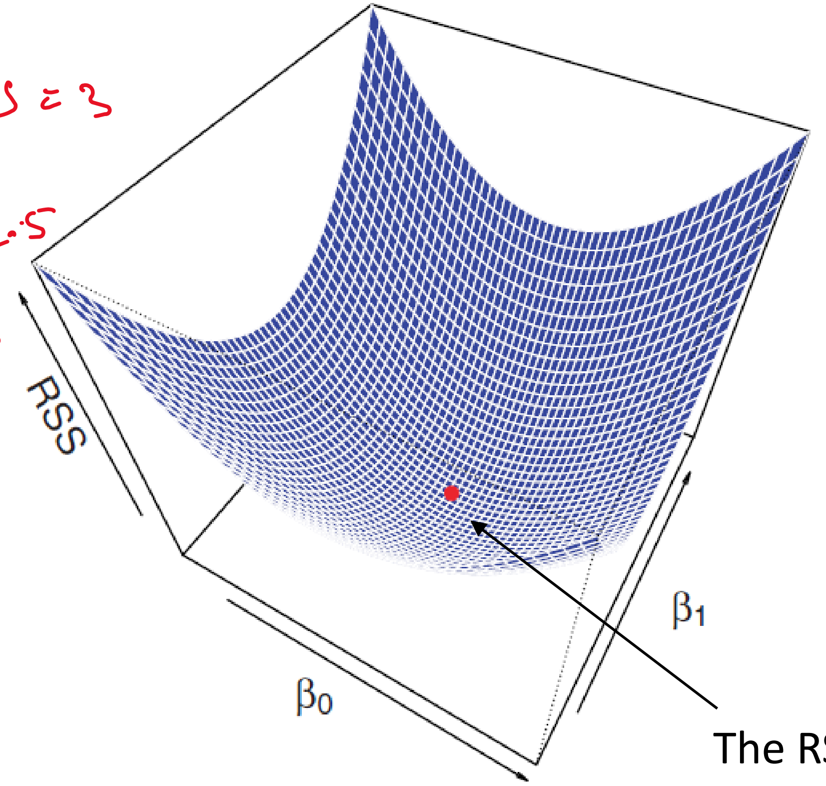
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \text{where } \bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i \text{ and } \bar{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$$

Linear Regression

$RSS = f(\hat{\beta}_0, \hat{\beta}_1)$;

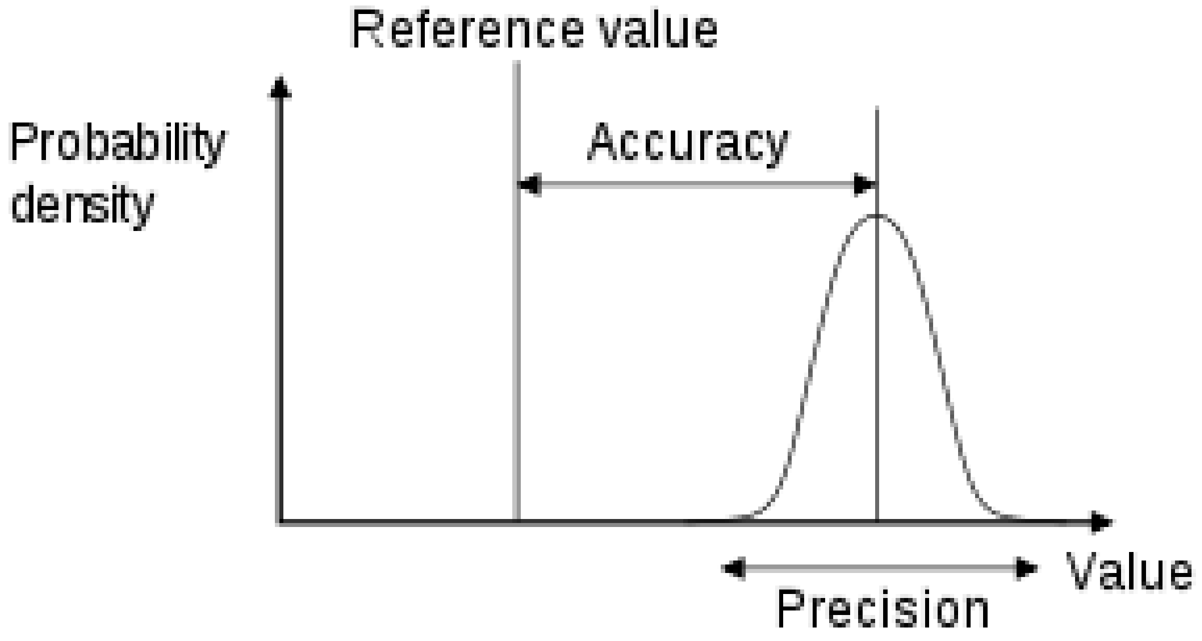




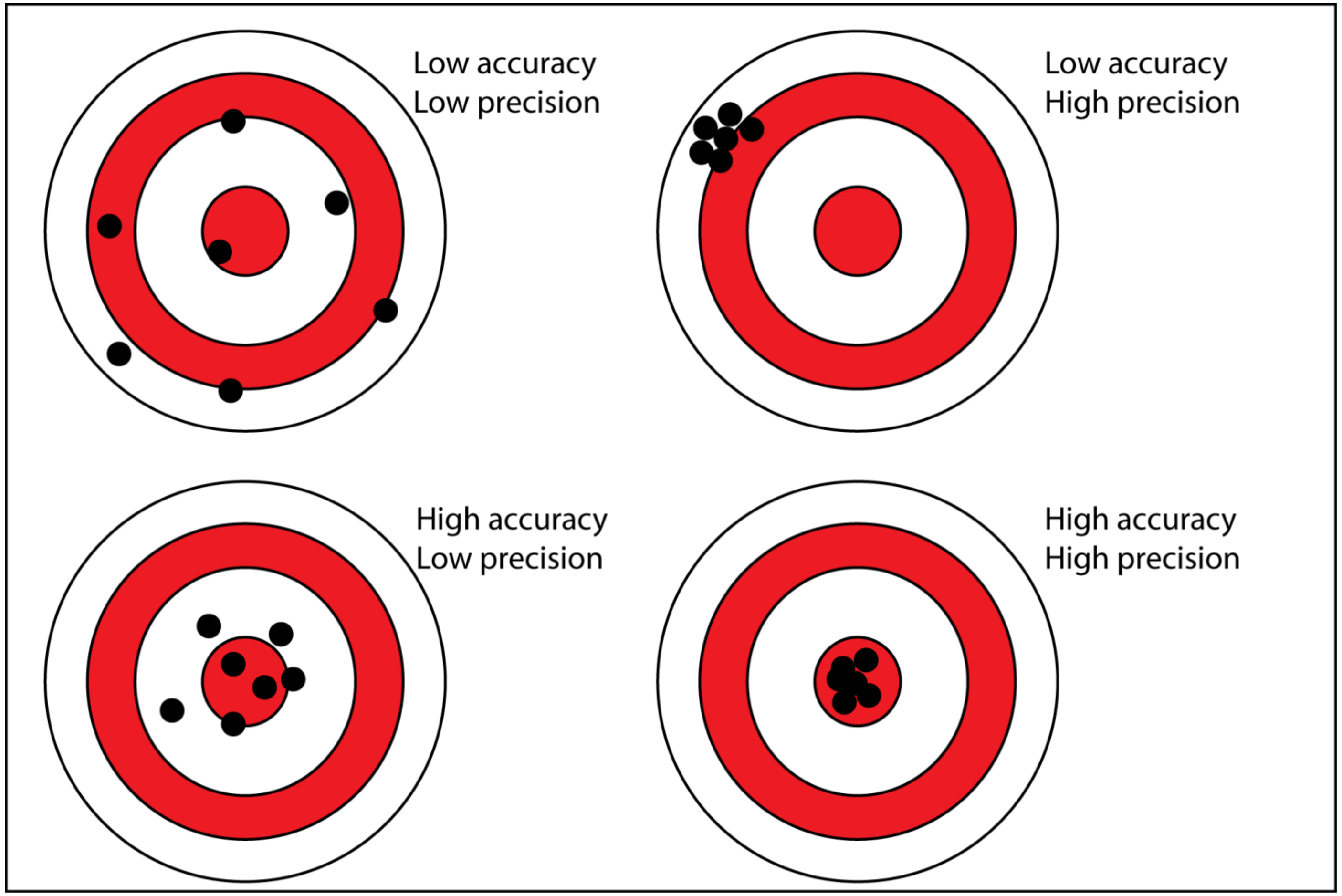
The RSS Value is minimum

Accuracy and Precision

- Accuracy refers to the closeness of a measured value to a standard or known value. Accuracy is a description of systematic errors, a measure of statistical bias.
- Precision refers to the closeness of two or more measurements to each other. Precision is a description of random errors, a measure of statistical variability.



Accuracy Vs. Precision



Linear Regression

- True relation between X and Y

$Y = f(x) + €$

↑ Mean zero

random error Independent of x

If f(x) is linear then $Y = \beta_0 + \beta_1 X + \epsilon.$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are analogous to estimation of population mean from sample mean

I.e $\hat{\beta}_0$ an $\hat{\beta}_1$ are unbiased estimate of true β_0 and β_1

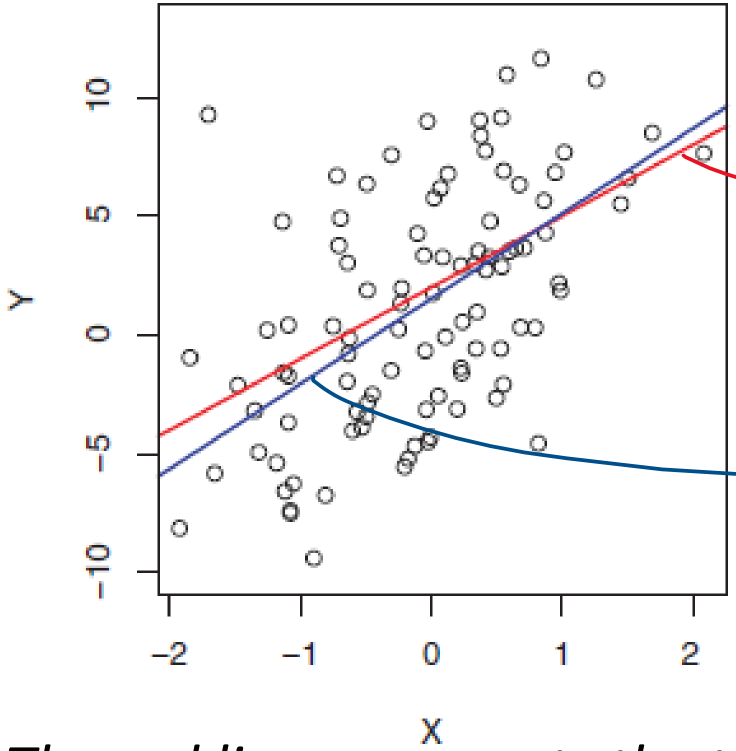
$Y = \beta_0 + \beta_1 x + \epsilon$

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

1.) How close is $\hat{\beta}_0$ to β_0 and $\hat{\beta}_1$ to β_1 ?

2.) How close is $\hat{y}$ to y ?

Linear Regression



β_0 , β_1

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

A simulated data set. Left: *The red line represents the true relationship, f*(*X*) = 2+3*X*, *which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for f*(*X*) *based on the observed data, shown in black*

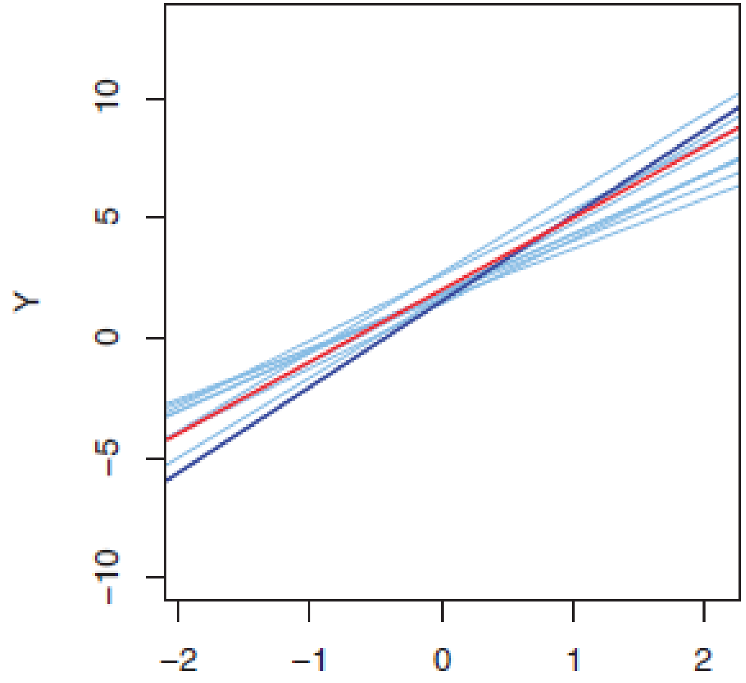
y_1	x_1
y_2	x_2
y_3	x_3
⋮	⋮
y_n	x_n

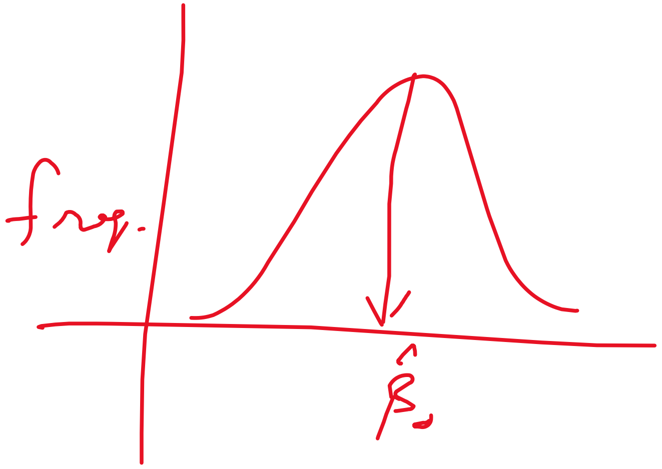
$\hat{\beta}_0, \hat{\beta}_1$

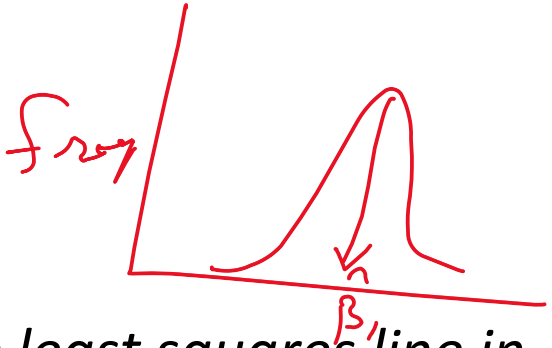
$\hat{\beta}_0, \hat{\beta}_1$

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

Linear Regression







The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a <u>separate random set of observations</u>. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.

Concepts of Population and Sample

- Mean
- Variance
- Covariance

- Population and Sample
- Population mean and variance
- Sample mean and variance

Mean and Variance

Expected value of $g(x)$

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) f(x)\, dx$$

X is a random variable with p.d.f. $f(x)$

Mean value of X is $\mu \equiv \int_{-\infty}^{\infty} x f(x)\, dx = E(X)$

Variance of X is $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$

$$= E\left[(X - E(X))^2\right] = E\left[X^2 - 2XE(X) + [E(X)]^2\right]$$

$$= E(X^2) - 2E(X)E(X) + [E(X)]^2 = E(X^2) - [E(X)]^2$$

$$\Rightarrow \mathrm{Var}(X) = E(X^2) - [E(X)]^2$$

Variance and Covariance

$$\mathrm{Cov}(X,Y) = E\left([X-E(X)][Y-E(Y)]\right)$$

$$= E(XY) - 2E(X)E(Y) + E(X)E(Y)$$

$$= E(XY) - E(X)E(Y)$$

$$\therefore \mathrm{Var}(X) = E(X^2) - [E(X)]^2$$

and $\mathrm{Cov}(X,Y) = E(X.Y) - E(X)E(Y)$

Variance and Covariance

1.) $Var(X) \geq 0$

2.) $P(X=a)=1 \iff Var(X)=0$

3.) $Var(X+a) = Var(X)$

4.) $Var(aX) = a^2 Var(X)$

5.) $Var(X) = Cov(X, X)$

6.) $Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X, Y)$

Variance of sum of random variables

$$\mathrm{Var}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i,j}^{N} \mathrm{Cov}(X_i, X_j)$$

$$= \sum_{i=1}^{N} \mathrm{Var}(X_i) + \sum_{i \neq j}^{N} \mathrm{Cov}(X_i, X_j)$$

$$\mathrm{Var}\left[\sum_{i=1}^{N} a_i X_i\right] = \sum_{i,j}^{N} a_i a_j \mathrm{Cov}(X_i, X_j)$$

$$= \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j}^{N} a_i a_j \mathrm{Cov}(X_i, X_j)$$

$$= \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) + \sum_{1 \leq i < j \leq N} a_i a_j \mathrm{Cov}(X_i, X_j)$$

Variance and Covariance

If $Cov(X_i, X_j) = 0 \quad \forall \; i \neq j$

$\Rightarrow$ X_i, X_j are uncorrelated

For N independent r.v. $X_1, X_2 \ldots\ldots X_N$

$$Var\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} Var(X_i)$$

If all the N r.v. have the same variance σ^2

then $$Var\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} Var(X_i) = N\sigma^2$$

Variance of mean

Mean of n r.V.s is $\frac{1}{n}\sum_{i=1}^{n} x_i$

Variance of mean of n r.V.s would be

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} var(x_i)$$ } Assuming x_i's are independent

$$= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$ } Assuming x_i's are i i d

$$\therefore Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{\sigma^2}{n}$$ if x_i's are i i d

Population and sample

Population : Population of size N with value X_i

Population mean : $\mu \equiv \frac{1}{N} \sum_{i=1}^{N} X_i$ $\Big\}$ $\mu = E(X)$

Population variance : $\sigma^2 \equiv \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2$ $\Big\}$ $\sigma^2 = E[X - E(X)]^2$

Sample : Take n random values (with replacement) from the population. $y_1, y_2, \ldots\ldots y_n$

Sample mean : $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Sample variance : ?

Sample mean

Expected value of Sample mean:

$$E(\bar{y}) = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(y_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu \qquad = \mu$$

$\therefore E(\bar{y}) = \mu$

Expected value of sample mean is equal to population mean.

$\Rightarrow$ Sample mean is an unbiased estimator of the population mean.

Sample Variance

How to define sample variance so that it is an unbiased estimator of the population variance.

Let squared deviation be defined as

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

then

$$E(\sigma_y^2) = E\left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2\right)$$

$$= E\left(\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2\right)$$

Sample Variance

$$\Rightarrow E(\sigma_y^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left(y_i^2 - \frac{2}{n} y_i \sum_{j=1}^{n} y_j + \frac{1}{n^2}\sum_{j=1}^{n} y_j \sum_{k=1}^{n} y_k\right)$$

$j=i$ $j\neq i$ $k\neq j$ $k=j$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{2}{n}\right)E(y_i^2) - \frac{2}{n}\sum_{j\neq i} E(y_i y_j) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j} E(y_j y_k) + \frac{1}{n^2}\sum_{j=1}^{n} E(y_j^2)$$

$$\Rightarrow E(\sigma_y^2) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(1-\frac{2}{n}\right)(\sigma^2+\mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2} n(n-1)\mu^2 + \frac{n}{n^2}(\sigma^2+\mu^2)\right]$$

Sample Variance

$$\Rightarrow E(\sigma_g^2) = \frac{1}{n}\sum\left(\frac{n-1}{n}\right)\sigma^2$$

$$= \frac{1}{n}\, n \left[\frac{n-1}{n}\right]\sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2$$

$$\therefore E(\sigma_g^2) = \frac{n-1}{n}\,\sigma^2$$

$\therefore$ σ_g^2 estimates the population variance with a bias of $\frac{n-1}{n}$ factor

Sample Variance

$\therefore$ if we define sample variance $s^2 = \frac{n}{n-1} \sigma_y^2$

then it will be an unbiased estimator of the population variance.

$\therefore$ Sample variance s^2 is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$