4 Linear regression

4.1 Simple LR

See notes

4.2 Multiple LR

Setup: Features $X_1, \dots, X_p$, continuous response Y

MLR model is:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon \qquad (3)$$

random

We say Y is being regressed on $X_1, \dots, X_p$. Here, $\beta_0, \dots, \beta_p$ are regression parameters. This is the usual model

$$Y = f(\underline{x}) + \varepsilon$$

with a linear specification for f: $f(\underline{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

f is the population regression line & ε is the residual.

Interpret: β_0 = Avg value of Y when $X_1 = \cdots = X_p = 0$

β_k = Avg change in Y for a unit increase in X_k with all other features fixed

Given a set of data $y_1, \ldots, y_n$ with y_i having corresponding features $x_{i1}, x_{i2}, \ldots, x_{ip}$, we have n observation equations

$$y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + \varepsilon_1$$

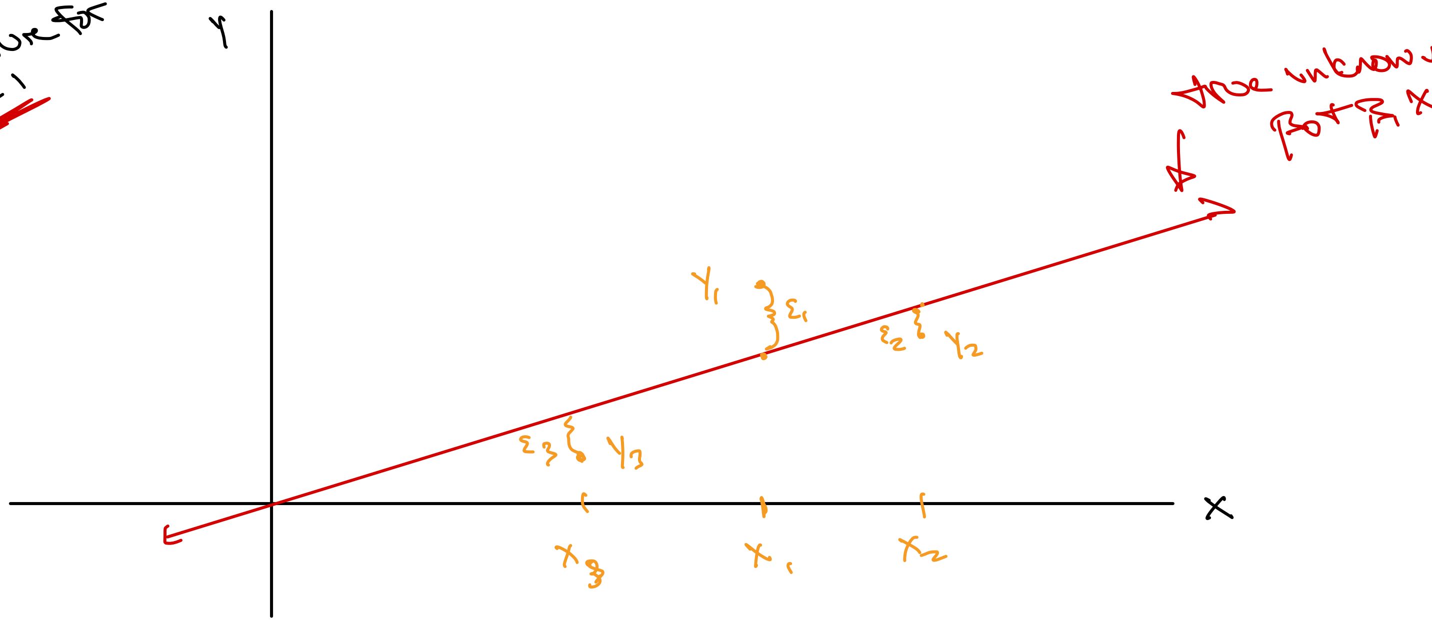
$$y_2 = \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} + \varepsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + \varepsilon_n$$

Picture for $p = 1$

x_1



Annoying to write

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad \text{for } i = 1, \ldots, n$$

$$\underset{n\times 1}{\underline{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \underset{(p+1)\times 1}{\underline{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \underset{n\times 1}{\underline{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\underset{n\times (p+1)}{X} = \begin{array}{c} \begin{matrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_p \end{matrix} \\ \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \end{array}$$

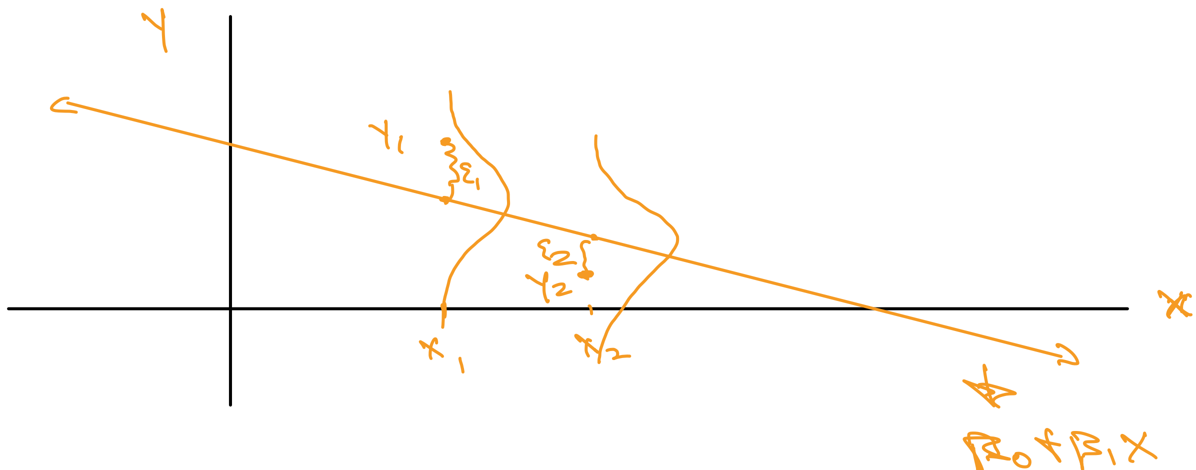
$$\underset{n\times 1}{\underline{Y}} = \underset{n\times(p+1)}{X}\ \underset{(p+1)\times 1}{\underline{\beta}} + \underset{n\times 1}{\underline{\varepsilon}} \qquad (4)$$

4.2 Assumptions

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad , i = 1, \ldots, n \qquad (4)$$

What do we assume about (4)?

$\boxed{A1}$ • (4) holds • $\{\varepsilon_i\}$ iid $N(0, \sigma^2)$



$\boxed{A2}$ • (4) holds • $E\varepsilon_i = 0 \ \forall i$ • $\mathrm{Var}\,\varepsilon_i = \sigma^2 \ \forall i$
(homoskedastic)

• $\{\varepsilon_i\}$ are iid

[relaxing normality]

$\boxed{A3}$ • (4) holds • $E\varepsilon_i = 0 \ \forall i$ • $\mathrm{Var}\,\varepsilon_i = \sigma^2 \ \forall i$

• $\{\varepsilon_i\}$ pairwise uncorrelated

[relaxing iid]

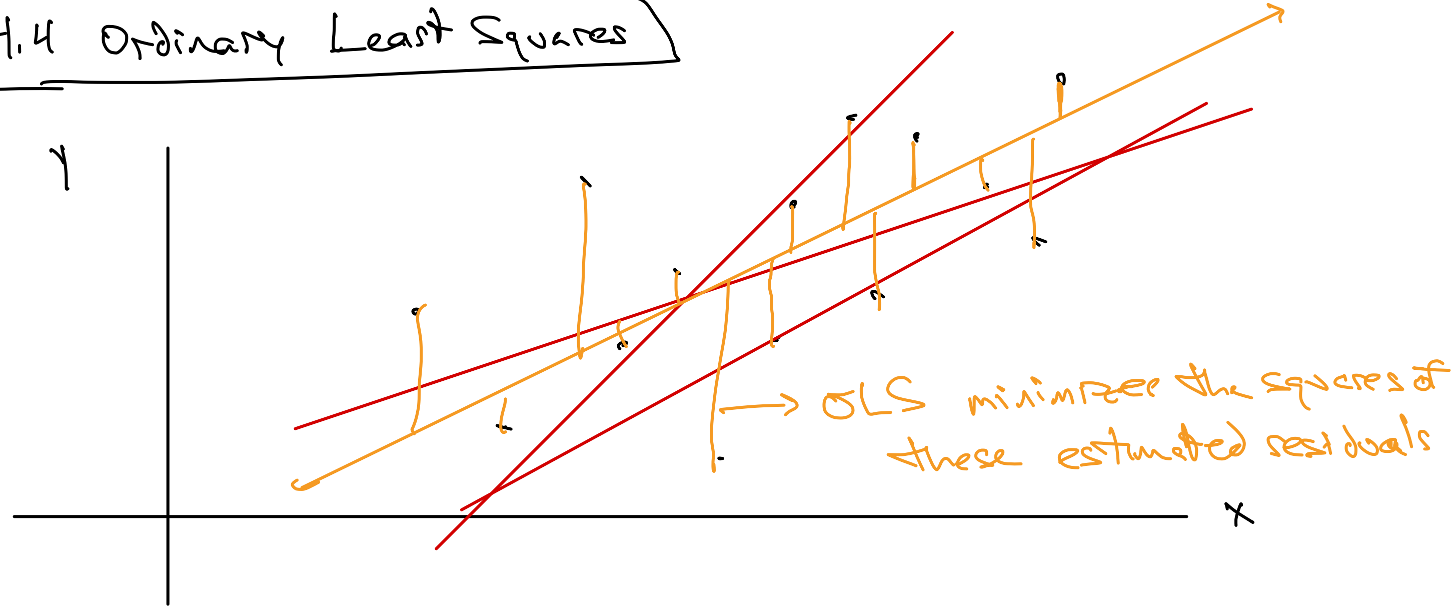
We will always assume $\boxed{A1}$ – $\boxed{A3}$, in which case

$$E\,\underline{\varepsilon} = \underline{0} \qquad \mathrm{Var}\,\underline{\varepsilon} = \sigma^2 I$$

so $\Rightarrow$ $E\underline{Y} = E(X\underline{\beta} + \underline{\varepsilon}) = E(X\underline{\beta}) + E\underline{\varepsilon} = X\underline{\beta}$

$\mathrm{Var}\,\underline{Y} = \mathrm{Cov}(\underline{Y}, \underline{Y}) = \mathrm{Cov}(X\underline{\beta} + \underline{\varepsilon}, X\underline{\beta} + \underline{\varepsilon}) = \mathrm{Cov}(\underline{\varepsilon}, \underline{\varepsilon}) = \mathrm{Var}\,\underline{\varepsilon} = \sigma^2 I$

4.4 Ordinary Least Squares



The OLS estimator for $\underline{\beta}$ minimizes residual sum of squares:

$$RSS(\underline{\beta}) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2$$

$$= (\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta})$$

The <u>OLS estimator</u> is:

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

Note!

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

$$= (X^TX)^{-1}X^TX\beta = I\beta = \beta \Rightarrow \hat{\beta} \text{ is unbiased}$$

$$\operatorname{Var}\hat{\beta} = \operatorname{Cov}(\hat{\beta}, \hat{\beta}) = \operatorname{Cov}\left((X^TX)^{-1}X^Ty, \left((X^TX)^{-1}X^Ty\right)^T\right)$$

$$= (X^TX)^{-1}X^T\operatorname{Cov}(y, y)X^{T^T}(X^TX)^{-T}$$

$$= (X^TX)^{-1}X^T\underbrace{(\operatorname{Var}y)}_{\sigma^2 I}X(X^TX)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} \underbrace{X^T X}\,\underbrace{(X^T X)^{-1}}$$

$$\boxed{\underbrace{\mathrm{Var}\,\hat{\beta}}_{(p+1)\times 1} = \sigma^2 \Big(\underbrace{\underbrace{X^T}_{(p+1)\times n}\;\underbrace{X}_{n\times(p+1)}}\Big)^{-1}} \longrightarrow (p+1) \times (p+1)$$

Standard errors for $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are the square roots of the diagonal of

$$\widehat{\mathrm{Var}}\,\hat{\beta} = \hat{\sigma}^2 (X^T X)^{-1}$$

where, for example,

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

Approx 95% CI for β_j is

$$\hat{\beta}_j \pm 2\,SE(\hat{\beta}_j)$$