Linear Regression

1. Overview and Basic Theory

2. R stuff and problems.

OLS

SLR

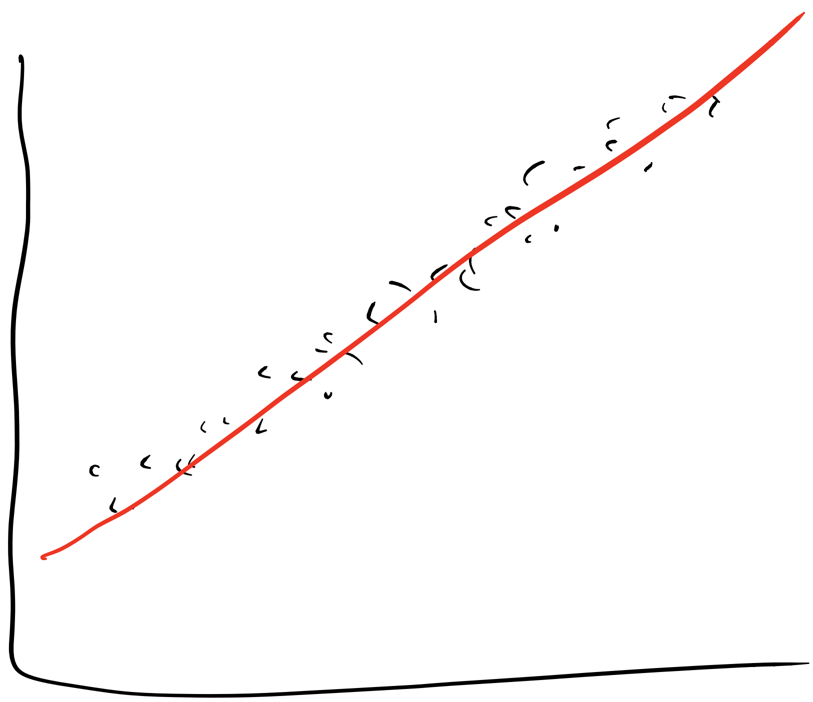
Least Squares

Ld regression

Linear regression w/ Least Squares

Linear Regression

We want to fit a straight line through data.



given (x_i, y_i) data pairs.

Setup:

Assume we have a relationship

$$y_i \approx \beta_0 + \beta_1 x_i$$

response

covariate

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$ ← noise

$$\varepsilon_i \overset{iid}{\sim} [0, \sigma_\varepsilon^2]$$

mean

variance

How do we find

$\hat{\beta}_0$, $\hat{\beta}_1$ (estimates of the coefficients)

our line estimate becomes

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$ "fitted value"

we can minimize how bad we do.

$y_i - \hat{y}_i$ is how bad we do.

A simple choice

"Least squares"

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

$$\Downarrow$$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

again the setup

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} [0, \sigma_\varepsilon^2]$$

often we assume

$$\varepsilon_i \overset{ind}{\sim} N(0, \sigma_\varepsilon^2)$$

$$y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma_\varepsilon^2)$$
$$i = 1, \dots, n$$

In our traditional

$$X_i \sim N(\theta, 1) \quad ; i = 1, \dots, n$$

The connection between Least squares and normal MLE.

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

MLE of β_0, β_1

$$y_i \overset{ind}{\sim} N\left(\beta_0 + \beta_1 x_i,\ \sigma_\varepsilon^2\right)$$

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(\frac{-1}{2} \frac{\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2}{\sigma_\varepsilon^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\right)^n \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right)$$

$$\ell_n(\beta_0, \beta_1) = \log\left(\left(\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\right)^n\right)$$

$$- \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

equiv to maximizing

$$- \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

equiv to minimizing

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2$$

Correlation

For random variables

$$\rho = cor(X,Y)$$

$$= \frac{cov(X,Y)}{sd(X)\cdot sd(Y)}$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$$\hat{\beta}_1 = r_{xy} \frac{sd_y}{sd_x}$$

$$\underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left| y_i - (b_0 + b_1 x_i) \right|$$

(L1)

vs.

$$\underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i) \right)^2$$

(L2)

Residuals

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$ "fitted values"

Residuals

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

$$\left(\text{Least squares} \quad \min \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \right)$$