Lecture 2 Linear regression ; Least Squares

23 Feb 2018

- Recall our multiple regression model:

  - Script Ch 1
  - ISLR Ch 3.3.3

$$Y = X\beta + \varepsilon .$$

and that X is an $n \times p$ design matrix which we assume is fixed meaning that it is non-random

- How do we construct our model? (fitting ...)

(1) Regression through the origin, ie. y-intercept is zero i.e. there is no y-intercept as we assume the model goes through the origin.

(p=1) ie.

$$Y_i = \beta_i x_i + \varepsilon_i \quad (i = 1, \dots, n)$$

$$\hookrightarrow X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \beta = \beta_1$$

(2) Simple linear regression

ie. $p=2$

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

(3) Quadratic regression

$p=3$

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$$

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

- Estimating the coefficients β, σ^2 by using the LSQ criterion.

  - We do not know $\beta_1, \dots, \beta_p$, σ^2
  - We want to estimate them, ie. find the LSQ estimate $\hat{\beta}$ ... (params)

By the LSQ criterion, $\hat{\beta} = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n [Y_i - (b_1 x_{i1} + \dots + b_p x_{ip})]^2$

(find $\underset{b}{\operatorname{argmin}} \sum \text{RSS}_i$)

$$\equiv \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n [Y_i - x_i^T b]^2$$

$$\equiv \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \| Y - Xb \|^2$$

- In multiple linear regression we are trying to find the (hyper-) plane s.t. $\sum$ RSS is minimal. ie $\sum$ squared distances of data to prediction

*note for LSQ, we measure vertically ie wrt/along y axis

For PCA, we measure wrt/along the normal (of the line)

← quadratic regr. for example

pt of minimum RSS.

Comp Stats

Lecture 2: Linear and Multiple Regression; Least squares

23 Feb 2018

- **Convex Optimization Problem:**
  - idea: take (partial) derivative of RSS wrt b, set it to zero:
  - ie $RSS = \|Y - Xb\|^2 = (Y-Xb)^T(Y-Xb)$

$$\frac{\partial}{\partial b}\|Y-Xb\|^2 = -2X^T(Y-Xb)$$

  - We want $X^T(Y - X\hat{\beta}) = 0$ ← "pre-normal" eq'n

$$\equiv X^TY = (X^TX)\hat{\beta}$$

⇒ Estimate $\hat{\beta}$ by ⟹ $\boxed{\hat{\beta} = (X^TX)^{-1}X^TY}$ the normal eq'n

* Compute this using QR decomposition or Cholesky instead of the inverse

For the corresponding estimate of σ^2, we can use the $\hat{\beta}$ residuals, so estimate

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y - X\hat{\beta})^2$$

, where $n-p$ is equal to the degrees of freedom. If it was $\frac{1}{n}$ or $\frac{1}{n-1}$ we would get the sample variance.

s.t. $\boxed{E(\hat{\sigma}^2) = \sigma^2}$

  - Again we have that $\hat{\beta}$ and $\hat{\sigma}^2$ are random since they have distributions (generated from simulations).

Moments Q: what can we say about the distribution of $\hat{\beta}$?

(1) First moment of $\hat{\beta}$ (expectation / mean)

$$E(\hat{\beta}) = E\left((X^TX)^{-1}X^TY\right) \quad \text{from the normal eq'n}$$

$$= (X^TX)^{-1}X^T E(X\beta + \varepsilon)$$

- Since X is assumed to be fixed (ie non-random) we can factor it out and so ...
- By linearity of expectation

$$= (X^TX)^{-1}X^T(X^T\beta + E(\varepsilon))$$

* Using the assumption $E(\varepsilon) = 0$ which indicates unbiased model is ⟹

$$E(\hat{\beta}) = (X^TX)^{-1}(X^TX)\beta = \beta \;//$$

$(X^TX)^{-T} \equiv (X^TX)^{-1}$ since X is symmetric

(2) 2nd moment: $Var(\hat{\beta}) = Var\left((X^TX)^{-1}X^TY\right)$

$$= (X^TX)^{-1}X^T\, Var(X\beta + \varepsilon)\, X(X^TX)^{-T}$$

$Var(X\beta + \varepsilon) = Var(\varepsilon) = \sigma^2 \cdot I$

So $Var(\hat{\beta}) = \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1}$

⇐ so

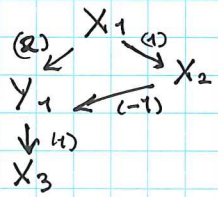
⇒ $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

⇐ if we assume errors are uncorrelated so we get a normal distribution.

- If we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
  ⇒ then $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X^TX)^{-1})$ by central limit theorem.

Lecture 2, Linear Regression, least Squares, Multiple Regression, cont'd 23 Feb 2018

- Interpreting parameters in multiple regression: (R code)
  - need/use a large sample to get precise estimates
  In the example, the generation of variable is represented as:



wherein edge weights are multipliers of the input variable to obtain the target variable.

- * A parameter has a meaning relative / in the context of the other variables in the model.
  (see code comparing β_2 vs. β_3 vs. $\beta_1 \vee \beta_2$ vs. $\beta_1 \vee \beta_2 \vee \beta_3$.

- How does one interpret β_i, given all the other variables in the model.

  $\hat{\beta}_{ij}$ = effect of x_j on Y_i after subtracting the effects of all other variables in the model.

  ($\hat{\beta}_{ij}$ depends only on response variable Y_i and the jth predictor x_{ij}.

* Remember these analyses do not support / imply causation!