PS5841

Data Science in Finance & Insurance

Method of Least Squares

Yubo Wang

Spring 2022

Columbia University

Squared Error Loss for Prediction

- Find $f(\boldsymbol{X})$ for predicting Y given values of $\boldsymbol{X}$, where
  - Random input vector $\boldsymbol{X} \in \mathcal{R}^p$
  - Random output variable $Y \in \mathcal{R}$
- Squared error loss

$$Loss = [Y - f(\boldsymbol{X})]^2$$

- Expected (squared) prediction error

$$EPE(f) = E([Y - f(\boldsymbol{X})]^2)$$
$$= E_{\boldsymbol{X}} E_{Y|\boldsymbol{X}}([Y - f(\boldsymbol{X})]^2 | \boldsymbol{X})$$

COLUMBIA UNIVERSITY

Regression Function

- Expected (squared) prediction error

$$EPE(f) = E([Y - f(\boldsymbol{X})]^2)$$
$$= E_{\boldsymbol{X}} E_{Y|\boldsymbol{X}}([Y - f(\boldsymbol{X})]^2|\boldsymbol{X})$$

- Minimize EPE pointwise

$$f(\boldsymbol{x}) = \operatorname*{argmin}_c E_{Y|\boldsymbol{X}}([Y - c]^2|\boldsymbol{X} = \boldsymbol{x})$$

- Regression function

$$f(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$$

COLUMBIA UNIVERSITY

Scenario 1

- When an analytical solution for loss minimization is available
  - e.g. Linear Regression Function

COLUMBIA UNIVERSITY

Basic/Simple Linear Regression Model

- Training set of size n: $\{(x_i, y_i)\}$
  - y_i is the observed value of Y_i
  - The Y_is are independent
- Regression function

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

COLUMBIA UNIVERSITY

Fitted by the Method of Least Squares

- Assume $Var(Y_i) = \sigma^2 \; \forall \, i$ for now
- Minimize loss $\sum_i (y_i - \beta_0 - \beta_1 x_i)^2$
- Estimated parameters

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

$$\hat{\beta}_1 = \sum_i \omega_i y_i$$

$$\omega_i = \frac{x_i - \bar{x}}{(n-1) s_X^2} \rightarrow \sum_i \omega_i = 0$$

$$s_X^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

- Fitted model $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

COLUMBIA UNIVERSITY

Useful Results

- $\bar{\hat{y}} = \bar{y}$
- $\sum_i \hat{e}_i = 0, \ \hat{e}_i = \hat{y}_i - y_i$
- $\sum_i x_i \hat{e}_i = 0$
- TSS = ESS + RSS

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - y_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

- Estimating σ^2: $s^2 = \frac{1}{n-2} \sum_i \hat{e}_i^2$
- Coefficient of Determination: $R^2 = \frac{RSS}{TSS}$
- ANOVA Table: keep track of variability

COLUMBIA UNIVERSITY

Multiple Linear Regression Model

- Training set of size n: $\{(\boldsymbol{x}_i, y_i)\}$
  - y_i is the observed value of Y_i
  - The Y_is are independent
  - $Var(\boldsymbol{Y}) = V$
- Regression function

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$$

Columbia University

Fitted by the Method of Least Squares (1)

- Minimize loss

$$\sum_i \frac{1}{\sigma_i^2}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2$$

$$= (\boldsymbol{y}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$$

  - Matrix of 2nd derivatives positive definite
  - Solutions of the normal equations vs local minima at parameter space boundaries

COLUMBIA UNIVERSITY

Fitted by the Method of Least Squares (2)

- Estimated parameters (X_0 is the intercept if used)

$$\widehat{\boldsymbol{\beta}} = (X^T V^{-1} X)^{-1} X^T V^{-1} \boldsymbol{y}$$

$$X = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = (X_0 \quad X_1 \quad \cdots \quad X_k)$$

- If $V = Var(\boldsymbol{Y}) = \sigma^2 I$

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} = \sum_i \boldsymbol{\omega}_i y_i$$

$$\boldsymbol{\omega}_i = (X^T X)^{-1} (1, x_{i1}, \ldots, x_{ik})^T$$

- Fitted model $\hat{y}(\boldsymbol{x}) = \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{y}}(X) = X\widehat{\boldsymbol{\beta}}$

COLUMBIA UNIVERSITY

Useful Results (1)

When $E(\boldsymbol{Y}) = X\boldsymbol{\beta}, Var(\boldsymbol{Y}) = \sigma^2 I, Y_i$ s are independent RV

- $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
- $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(X^T X)^{-1}$
- LSE is BLUE (Gauss-Markov)
- $s^2 = \hat{\sigma}^2 = \frac{1}{n-p}\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)$

COLUMBIA UNIVERSITY

Useful Results (2)

- Hat Matrix $H = X(X^TX)^{-1}X^T$

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X^TX)^{-1}X^T\boldsymbol{y} = H(\boldsymbol{y})$$

$$H^T = H, \qquad HH = H$$

$$HX = X, \qquad HX_j = X_j$$

- $\hat{\boldsymbol{e}}^T X_j = 0, \hat{\boldsymbol{e}} = (I - H)\boldsymbol{y}$
- $\boldsymbol{y}^T\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}}$
- TSS = ESS + RSS

$$TSS = \boldsymbol{y}^T\boldsymbol{y} - n(\bar{y})^2$$

$$ESS = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\hat{\boldsymbol{y}}$$

$$RSS = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} - n(\bar{y})^2$$

- Coefficient of Determination: $R^2 = \frac{RSS}{TSS} = \left(r_{\boldsymbol{y},\hat{\boldsymbol{y}}}\right)^2$
- ANOVA Table: keep track of variability

COLUMBIA UNIVERSITY

Normal Linear Model: Inference (1)

- $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}, (X^T X)^{-1} \sigma^2)$
- $\hat{\sigma}^2 = \frac{ESS}{n-p}$
- $H_0: \hat{\beta}_j = 0$ vs $H_1: \hat{\beta}_j \neq 0 \rightarrow t = \frac{\widehat{\beta}_j}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \sim t_{n-p}$
- $100(1-\alpha)\%$ confidence interval

$$\hat{\beta}_j \pm t_{n-p,1-\alpha/2}\, \hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}}$$

- $100(1-\alpha)\%$ prediction interval for y at $\boldsymbol{x}$

$$\boldsymbol{x}^T \widehat{\boldsymbol{\beta}} \pm t_{n-p,1-\alpha/2}\, \hat{\sigma}(1 + \boldsymbol{x}^T (X^T X)^{-1} \boldsymbol{x})^{1/2}$$

COLUMBIA UNIVERSITY

Normal Linear Model: Inference (2)

$$H_0\colon \widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \dots, \hat{\beta}_{p_0}\right)^T$$
$$H_1\colon \widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \dots, \hat{\beta}_{p_1}\right)^T$$
$$p_0 < p_1 < n$$

Equivalently, the general linear hypothesis

$$H_0\colon C\widehat{\boldsymbol{\beta}} = \mathbf{0}$$
$$C = \left(0_{p_0}, I_{p_1-p_0}\right), \qquad \widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \dots, \hat{\beta}_{p_1}\right)^T$$

$$F = \frac{ESS_0 - ESS_1}{p_1 - p_0} \div \frac{ESS_1}{n - p_1} \sim \mathcal{F}_{p_1-p_0, n-p_1}$$

COLUMBIA UNIVERSITY

Power Transforms

- To make the assumption of normality (if desired) more plausible

COLUMBIA UNIVERSITY

Strictly Positive Data

- Box-Cox family of transforms $(y > 0)$

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

- Estimate λ via maximum likelihood
- In practice,
  - Typically, $\lambda = 1, 0.5, 0, -1$
  - No need to -1 and $\div \lambda$ for operations unaffected by location and scale shifts (e.g. some regressions)

COLUMBIA UNIVERSITY

General Data

- Box-Cox family of transforms $(y > -\alpha)$

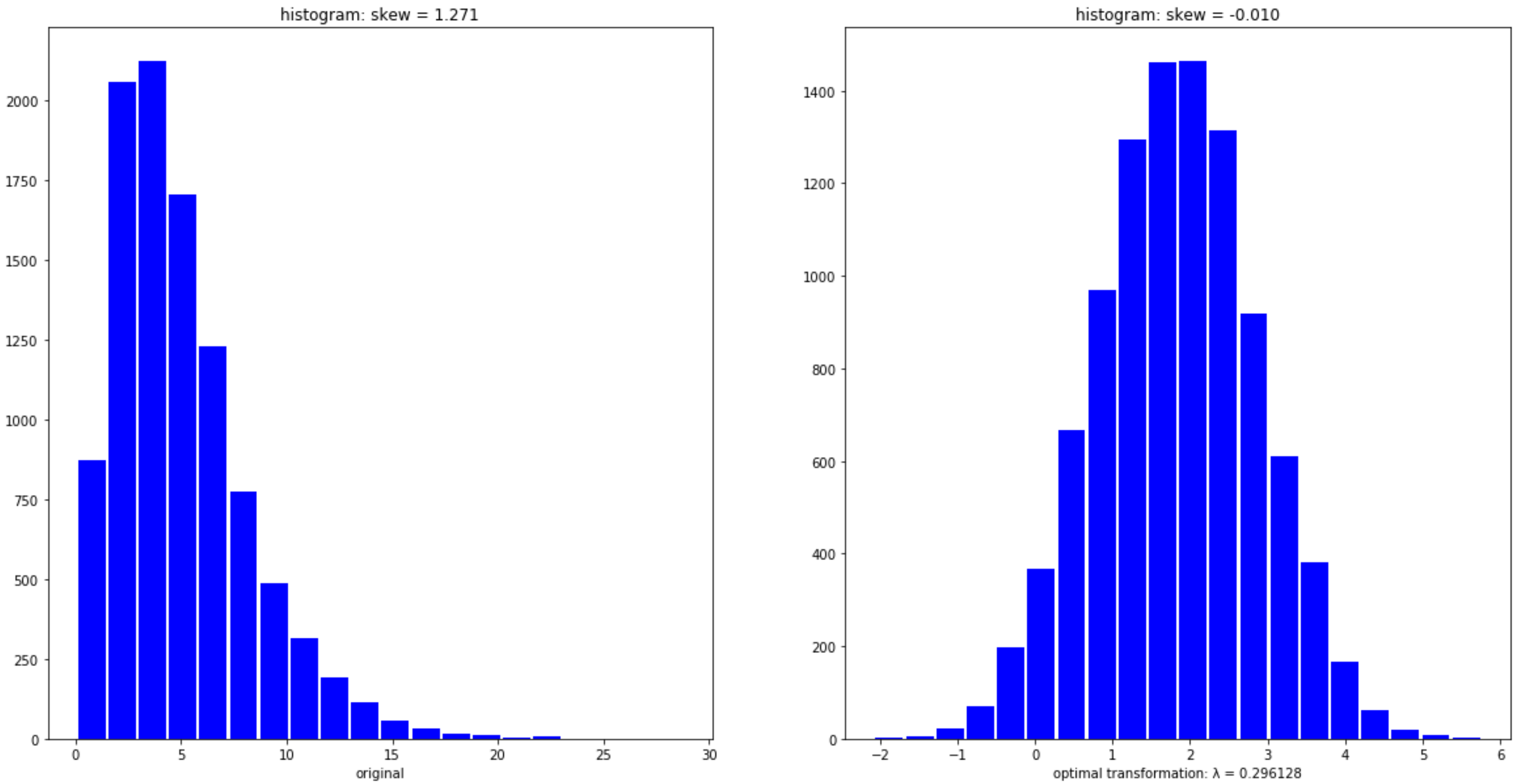
$$y^{(\lambda)} = \begin{cases} \dfrac{(y+\alpha)^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(y+\alpha), & \lambda = 0 \end{cases}$$

- Yeo-Johnson family of transforms

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0, y \geq 0 \\ \ln(y+1), & \lambda = 0, y \geq 0 \\ -\dfrac{(-y+1)^{2-\lambda}-1}{2-\lambda}, & \lambda \neq 2, y < 0 \\ -\ln(-y+1), & \lambda = 0, y < 0 \end{cases}$$

COLUMBIA UNIVERSITY

Box-Cox Transform Example



COLUMBIA UNIVERSITY

Categorical Features

- Often modeled by binary (dummy) variables
- For a feature (factor) with J levels
  - Need J binary variables if there is no intercept
  - Need (J-1) binary variables if there is intercept
  - Baseline is the level without a dummy variable
- Example: 1 factor with 3 levels (A,B,C)
  - with baseline A (A as the reference level)

$$y = \beta_0 + \beta_1 x_B + \beta_2 x_C + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

  - How are $\hat{\beta}_j$'s interpreted?

Example

- 1 factor with 3 levels (A,B,C)
  - with baseline A (A as the reference level)

$$y = \beta_0 + \beta_1 x_B + \beta_2 x_C + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

- If level A is present, all else being equal

$$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

- If level B is present, all else being equal

$$\hat{y}_B = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

- Interpretation of $\hat{\beta}_1$

$$\hat{y}_B - \hat{y}_A = \hat{\beta}_1$$

Columbia University

Application

- Examples
  - Equity beta
    - Market model or CAPM
  - Demand for life insurance
    - Features are family characteristics that influence the amount of insurance purchased

COLUMBIA UNIVERSITY

Scenario 2

- When the loss needs to be numerically minimized
  - e.g. Non-Linear Regression Function
- (Stochastic) Gradient Descent
- Newton Raphson

COLUMBIA UNIVERSITY

Directional Derivative

- Directional derivative of $f: \mathcal{R}^p \to \mathcal{R}$ at $\boldsymbol{x}$ in the direction of a unit vector $\boldsymbol{u}$, the rate of change of f at $\boldsymbol{x}$ in the direction of $\boldsymbol{u}$

$$\boldsymbol{D}_{\boldsymbol{u}} f(\boldsymbol{x}) = \lim_{h \to 0} \frac{f(\boldsymbol{x} + h\boldsymbol{u}) - f(\boldsymbol{x})}{h} = \boldsymbol{\nabla} f(\boldsymbol{x}) \cdot \boldsymbol{u}$$

- Gradient operator (del operator)

$$\boldsymbol{\nabla} = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_p}\right)^T$$

- Gradient vector

$$\boldsymbol{\nabla}(f) = \boldsymbol{\nabla} f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_p}\right)^T = \frac{\partial f}{\partial \boldsymbol{x}}$$

COLUMBIA UNIVERSITY

Maximum Rate of Change

- When $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$
  - the maximum rate of increase of f is $|\nabla f(\boldsymbol{x})|$ and is in the direction of $\nabla f(\boldsymbol{x})$
  - the maximum rate of decrease of f is $|\nabla f(\boldsymbol{x})|$ and is in the direction of $-\nabla f(\boldsymbol{x})$

COLUMBIA UNIVERSITY

Example (1)

- $f(\boldsymbol{x}) = x_1 + x_2^2, \boldsymbol{\nabla} f(\boldsymbol{x}) = \begin{pmatrix} 1 \\ 2x_2 \end{pmatrix}$
- The unit vector in the direction of the gradient at $\boldsymbol{x}_0 = (1,1)^T$ is $\boldsymbol{u} = \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right)^T$ since

$$\boldsymbol{\nabla} f(\boldsymbol{x}_0) = (1,2)^T, \qquad |\boldsymbol{\nabla} f(\boldsymbol{x}_0)| = \sqrt{5}$$

- Moving d units from $\boldsymbol{x}_0 = (1,1)^T$ in the direction of the gradient vector will land at $\boldsymbol{x}_1 = \boldsymbol{x}_0 + \boldsymbol{u}d = \left(1 + \frac{d}{\sqrt{5}}, 1 + \frac{2d}{\sqrt{5}}\right)^T$, with

rate of increase

$$f(\boldsymbol{x}_1) = f(\boldsymbol{x}_0) + \sqrt{5}d + \frac{4d^2}{5}$$

predicted increase

"error" on large steps

Example (2)

Same as predicted rate

gradient ascent					
d	f[1+(1/sqrt(5))d, 1+(2/sqrt(5))d]	f(1,1)	chg (f)	sqrt(5)*d	error
1	5.036067977	2	3.036068	2.23606798	0.8
0.1	2.231606798	2	0.231607	0.2236068	0.008
0.01	2.02244068	2	0.022441	0.02236068	8E-05
non-optimal direction					
d	f[1+(1/sqrt(2))d, 1+(1/sqrt(2))d]	f(1,1)	chg (f)		See! It's true!
1	4.621320344	2	2.62132	<	3.036068
0.1	2.217132034	2	0.217132	<	0.231607
0.01	2.021263203	2	0.021263	<	0.022441

Columbia University

Gradient Descent

- To minimize the loss function

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{n} R_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i(\boldsymbol{\beta}))^2$$

with learning rate $\eta > 0$, updating involves all observations

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \eta \boldsymbol{\nabla} R\left(\boldsymbol{\beta}^{(r)}\right)$$

$$= \boldsymbol{\beta}^{(r)} - \eta \sum_{i=1}^{n} \boldsymbol{\nabla} R_i(\boldsymbol{\beta}^{(r)})$$

Columbia University

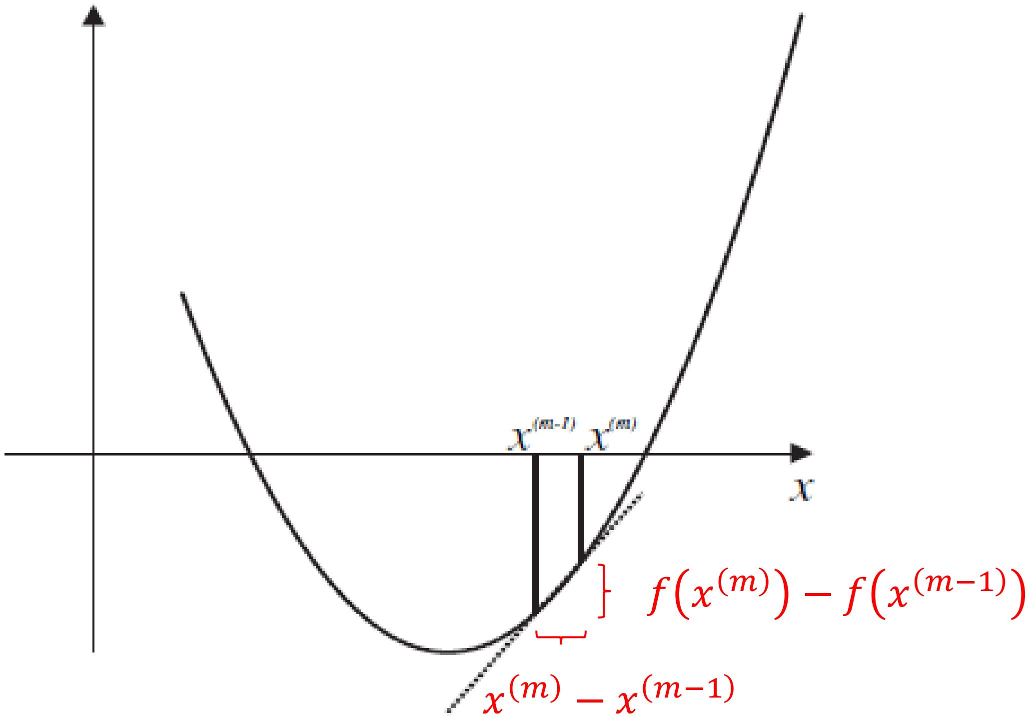
Stochastic Gradient Descent

- SGD is a stochastic approximation of GD.
- SGD uses randomly selected samples/subset from the training set for each iteration
- At extreme, updating would involve only a single (randomly selected) observation

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \eta \boldsymbol{\nabla} R_i\left(\boldsymbol{\beta}^{(r)}\right)$$

COLUMBIA UNIVERSITY

Find a root of a function



Root: $f\left(x^{(m)}\right) = 0$ for some m.

$$f'\left(x^{(m-1)}\right) = \frac{f\left(x^{(m)}\right) - f\left(x^{(m-1)}\right)}{x^{(m)} - x^{(m-1)}}$$

$$x^{(m)} = x^{(m-1)} - \frac{f\left(x^{(m-1)}\right)}{f'\left(x^{(m-1)}\right)}$$

COLUMBIA UNIVERSITY

Newton Raphson

- $f: \mathcal{R} \to \mathcal{R}$

$$x^{(r+1)} = x^{(r)} - \frac{f\left(x^{(r)}\right)}{f'\left(x^{(r)}\right)}$$

- $f: \mathcal{R}^p \to \mathcal{R}^q$

$$\boldsymbol{x}^{(r+1)} = \boldsymbol{x}^{(r)} - \left[J\left(\boldsymbol{x}^{(r)}\right)\right]^{-1} f\left(\boldsymbol{x}^{(r)}\right)$$

$$J = \frac{\partial f}{\partial \boldsymbol{x}}, \qquad J_{ij} = \frac{\partial f_i}{\partial x_j}$$

Columbia University

Minimizing Squared Loss

- To minimize the loss function

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{n} R_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i(\boldsymbol{\beta}))^2$$

- Updating

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - [J(\boldsymbol{\beta}^{(r)})]^{-1} \boldsymbol{\nabla} R(\boldsymbol{\beta}^{(r)})$$

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \left[H(\boldsymbol{\beta}^{(r)})\right]^{-1} \left(\frac{\partial R(\boldsymbol{\beta}^{(r)})}{\partial \boldsymbol{x}}\right)$$

$$J = \frac{\partial}{\partial \boldsymbol{\beta}}\left(\frac{\partial R}{\partial \boldsymbol{\beta}^T}\right), \qquad J_{ij} = \frac{\partial}{\partial x_i}\left(\frac{\partial \sum_k R_k}{\partial x_j}\right)$$

$$H = \frac{\partial^2 R}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}, \qquad H_{ij} = \frac{\partial^2 \sum_k R_k}{\partial x_i \partial x_j}$$

COLUMBIA UNIVERSITY

Example

- To minimize the loss function

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{n} R_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

- $\boldsymbol{\nabla} R(\boldsymbol{\beta}) = \begin{pmatrix} \sum_i e_i \\ \sum_i e_i x_{i1} \\ \sum_i e_i x_{i2} \end{pmatrix}$

- $J(\boldsymbol{\beta}) = \begin{pmatrix} X_0^T X_0 & X_0^T X_1 & X_0^T X_2 \\ X_1^T X_0 & X_1^T X_1 & X_1^T X_2 \\ X_2^T X_0 & X_2^T X_1 & X_2^T X_2 \end{pmatrix}$

COLUMBIA UNIVERSITY

Example

- To maximize a log-likelihood function (or minimize -1*)

$$l(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta};y_i)$$

COLUMBIA UNIVERSITY

That was

to be continued