Summary

1. Regression model / function

$\Rightarrow$ Linear model of conditional mean

2. Estimation of linear model

2.1 Simple linear regression (SLR)

[SLR 1] Model.

[SLR 2] Definition of least-squares loss function.

[SLR 3] Ordinary least-squares estimates and alternate forms.
(Derivation through differentiation $= 0$)

[SLR 4] Show that when OLS estimate is the unique minimizer of least-squares loss.

① Two key properties of convex functions
(do not require proof, only need to use these conclusions)

(1.1) A convex function $f: \mathbb{R}^d \to \mathbb{R}$ has a unique global minimizer that is the stationary point of f (i.e. $\bar{x}$ such that $\nabla f(\bar{x}) = 0$)

(1.2) Twice-differentiable f
is strictly convex if its Hessian is positive definite;
is convex if its Hessian is semipositive definite.

② Quadratic function $v^T A v$ (A is symmetric w.l.g.)
is strictly convex if A is positive definite;
is conve if A is semi-positive definite.

③ Show when the least-squares loss function is convex or strictly convex by deriving the Hessian matrix.

[SLR5] Properties of OLS estimates

① Identities on sample residuals $\sum_{i=1}^{n} \hat{R}_i = \sum_{i=1}^{n} \hat{R}_i X_i = 0$

② Data center

③ Relationship with Pearson correlation

④ Non-symmetricity in X & Y

2.2 Multiple linear regression (MLR)

[MLR 1] Model.

[MLR 2] Ordinary least-squares estimates

(1) Least-squares loss function ✓

(2) Derivation through multivariate differentiation ✓

(3) Numerical OLS solution via QR decomposition ✓

Fall 2020 Qualifying Exam - Option B

3. This problem investigates fixed design linear regressions in high-dimensional settings. Suppose that we observe a sample of n observations, $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \ldots, n$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ denote the design matrix and $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ be the response vector. Consider a linear model with i.i.d. mean-zero Gaussian noise

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n}), \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the unknown coefficient vector, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ is the noise term, and $\mathbf{I}_{n\times n}$ is an n-by-n identity matrix. Many modern applications of this model are high-dimensional, in that the number of features d is comparable to, or even larger than, the sample size n. Assume that $d = n$, and $\mathbf{X}$ has orthonormal columns such that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_{d\times d}$. Consider the following regularized estimator for $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \right\}, \tag{2}$$

where λ is an unknown tuning parameter, $\|\cdot\|$ denotes the vector 2-norm; i.e., $\|\mathbf{a}\| = \left(\sum_{j=1}^d |a_j|^2\right)^{1/2}$ for a vector $\mathbf{a} = (a_1, \ldots, a_d)^T \in \mathbb{R}^d$.

(a) Let $\lambda = 0$. $(2) \Rightarrow \hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2$ ✓ when $\lambda = 0$

i. Derive the expression for $\widehat{\boldsymbol{\beta}}_0$ by solving the optimization problem in (2) and find its distribution.

ii. Consider the prediction error for a new observation of the form $y_{\text{new}} = \mathbf{x}_{\text{new}}^T\boldsymbol{\beta} + \varepsilon_{\text{new}}$, where $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$ is a fixed covariate vector and $\varepsilon_{\text{new}} \sim \mathcal{N}(0, 1)$ is a noise term independent of $\{\varepsilon_i\}_{i=1}^n$. Find the expected squared prediction error, $\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T\widehat{\boldsymbol{\beta}})^2$, for this new observation.

(b) Let $\lambda > 0$.

i. Derive the expression for ridge regression estimator $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ by solving the optimization problem in (2).

ii. Consider the prediction error for a new observation of the form $y_{\text{new}} = \mathbf{x}_{\text{new}}^T\boldsymbol{\beta} + \varepsilon_{\text{new}}$, where $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$ is a fixed covariate vector and $\varepsilon_{\text{new}} \sim \mathcal{N}(0, 1)$ is a noise term independent of $\{\varepsilon_i\}_{i=1}^n$. Find the expected squared prediction error, $\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T\widehat{\boldsymbol{\beta}}^{\text{ridge}})^2$, for this new observation.

iii. For this part of the question only, assume that $\|\boldsymbol{\beta}\| = 1$. Derive the optimal λ that minimizes the mean squared error for the ridge estimator, $\mathbb{E}\|\widehat{\boldsymbol{\beta}}^{\text{ridge}} - \boldsymbol{\beta}\|^2$. Discuss how you would find λ in practice when $\|\boldsymbol{\beta}\|$ is unknown.

(c) Suppose that a prior distribution $\boldsymbol{\beta}^{\text{prior}} \sim \mathcal{N}(\mathbf{0}, \sigma^2\Phi)$ is imposed to the model (1), where σ^2 is an unknown variance parameter, and Φ is a known positive definite matrix. Furthermore, assume that $\boldsymbol{\beta}^{\text{prior}}$ and $\boldsymbol{\varepsilon}$ are independent.

i. Find the marginal distribution of $\mathbf{y}$.

ii. Derive the method-of-moments estimator for σ^2 based on the marginal distribution of $\mathbf{y}$.

When X has full column rank, $\Rightarrow$ [Assumed all below.]

① OLS algebraic solution $\hat{\beta} = (X^TX)^{-1}X^TY$

② OLS numerical solution through $X = QR$

$\Rightarrow \hat{\beta} = R^{-1}Q^TY$

Exercise 1:

(a) Express hat matrix $X(X^TX)^{-1}X^T$ though $X = QR$

(b) $\underset{3\times 2}{Q} = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \\ 1 & 0 \end{pmatrix}$ Derive QQ^T and Q^TQ

(c) Solve $\begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$

Solutions

(a) $X = QR$ ① $Q^TQ = I_p$ ② R upper triangular matrix

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

$= QR\left(\underbrace{(QR)^T}_{\downarrow}QR\right)^{-1}(QR)^T \qquad (AB)^T = B^TA^T$

$= QR\left(R^T\underbrace{Q^TQ}_{I_p}R\right)^{-1}R^TQ^T$

$= QR(R^TR)^{-1}R^TQ^T$

$= \underbrace{QRR^{-1}}_{I} \times \underbrace{(R^T)^{-1}R^T}_{I}Q^T$

X has full column rank

R invertible

R^T invertible

$(R^TR)^{-1} = R^{-1}\cdot(R^T)^{-1}$

$= \underbrace{\underset{n\times p}{Q}\,\underset{p\times n}{Q^T}}_{n\times n}$ (NOT NECESSARILY IDENTITY)

$(Q^TQ = I_p)$

(b) $\underset{3\times 2}{Q} = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \\ -1 & 0 \end{pmatrix}$

$$\underbrace{\underset{2\times 3}{Q^T}\underset{3\times 2}{Q}}_{2\times 2} = \begin{pmatrix} 0 & 0 & -1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_{2\times 2}$$

$$\underbrace{\underset{3\times 2}{Q}\,\underset{2\times 3}{Q^T}}_{3\times 3} = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \neq \text{Identity matrix}$$

(c) $$\underbrace{\begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}}_{\text{Upper triangular } 3\times 3} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 1 \end{pmatrix}$$

$$\Leftrightarrow \begin{cases} 2\beta_1 + \beta_2 + \beta_3 = 2 \\ \beta_2 + 2\beta_3 = 4 \\ \beta_3 = 1 \end{cases} \Rightarrow \begin{matrix} 2\beta_1 + 2 + 1 = 2 \\ \beta_2 + 2\times 1 = 4 \\ \end{matrix} \begin{matrix} \Rightarrow \\ \Rightarrow \\ \end{matrix} \begin{cases} \beta_1 = -\frac{1}{2} \\ \beta_2 = 2 \\ \beta_3 = 1 \end{cases}$$

From bottom $\Leftarrow$

Numerical OLS   Normal equation $\Rightarrow$ Upper triangular system

Exercise 2 $X^\top X$

Two ways to think about this matrix product

(1) By row (observations)

$$\underset{n\times p}{X} = \begin{pmatrix} — & x_{(1)}^T & — \\ — & x_{(2)}^T & — \\ & \vdots & \\ — & x_{(n)}^T & — \end{pmatrix} \qquad X^T = \begin{pmatrix} | & | & & | \\ x_{(1)} & x_{(2)} & \cdots & x_{(n)} \\ | & | & & | \end{pmatrix}$$

($x_{(j)} \in \mathbb{R}^p$, $j = 1 \cdots n$ by default $x_{(j)}$ are column vectors)

(2) By columns (covariates)

$$\underset{n\times p}{X} = \begin{pmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(p)} \\ | & | & & | \end{pmatrix} \qquad \underset{p\times n}{X^T} = \begin{pmatrix} — & x^{(1)T} & — \\ — & x^{(2)T} & — \\ & \vdots & \\ — & x^{(p)T} & — \end{pmatrix}$$

$x^{(j)} \in \mathbb{R}^n$ $j = 1, \cdots, p$

Exercise: Express $X^\top X$ under two views

Solutions.

① $$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} | & & | \\ x_{(1)} & \cdots & x_{(n)} \\ | & & | \end{pmatrix} \begin{pmatrix} - & x_{(1)}^T & - \\ & \vdots & \\ - & x_{(n)}^T & - \end{pmatrix}$$

$$= \sum_{i=1}^{n} \underbrace{\underset{p\times 1}{x_{(i)}} \; \underset{1\times p}{x_{(i)}^T}}_{p\times p} \Rightarrow \text{sum of } n \text{ } p\times p \text{ matrices}$$

② $$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} - & x^{(1)T} & - \\ - & x^{(2)T} & - \\ & \vdots & \\ - & x^{(p)T} & - \end{pmatrix} \begin{pmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(p)} \\ | & | & & | \end{pmatrix}$$

$$= \begin{pmatrix} \underbrace{\underset{1\times p}{x^{(1)T}} \underset{p\times 1}{x^{(1)}}}_{1\times 1} & \underbrace{x^{(1)T} x^{(2)}}_{1\times 1} & \cdots & \underbrace{x^{(1)T} x^{(p)}}_{1\times 1} \\ x^{(2)T} x^{(1)} & x^{(2)T} x^{(2)} & \cdots & x^{(2)T} x^{(p)} \\ \vdots & \vdots & & \vdots \\ \underbrace{x^{(p)T} x^{(1)}}_{1\times 1} & \underbrace{x^{(p)T} x^{(2)}}_{1\times 1} & \cdots & \underbrace{x^{(p)T} x^{(p)}}_{1\times 1} \end{pmatrix}_{p\times p}$$

$$= \left(x^{(i)T} x^{(j)} \right)_{1\le i,j \le p} \quad (p\times p \text{ matrix})$$

(Use for HW1-Q3)

Exercise 3

Example: Suppose $Y_1 \cdots Y_n$ have common mean β

The least square estimate $\min_{\beta} \underbrace{\sum_{i=1}^{n} (Y_i - \beta)^2}_{L(\beta)}$

① Specific case

derivative

$$\frac{\partial L(\beta)}{\partial \beta} = -2 \sum_{i=1}^{n} (Y_i - \beta) = 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}$$

(Quadratic, unique minimizer)

② General Alternatively, we can linear regression model as

Plug-in

design matrix

$$\underset{n\times 1}{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \underbrace{\underset{n\times 1}{E(Y)} = \underset{n\times 1}{\begin{pmatrix} \beta \\ \beta \\ \vdots \\ \beta \end{pmatrix}} = \underset{n\times 1}{1_n} \times \beta} \qquad \underset{n\times 1}{1_n} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\underset{n\times 1}{X} = 1_n$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$= (\underset{1\times n}{1_n^T} \underset{n\times 1}{1_n})^{-1} 1_n^T Y$$

$$= n^{-1} \times 1_n^T Y = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y} \qquad \hat{\beta} = \bar{Y}$$

These two views should be equivalent.

Exercise $E(Y|x) = \alpha + \beta x$ in SLR

$(\hat{\alpha}, \hat{\beta})$

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$ equivalent

Hint: (1) $$\underset{n\times 2}{\mathbf{X}} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

intercept

1-dim covariate

(2) $\underset{2\times n}{\mathbf{X}^T}\ \underset{n\times 2}{\mathbf{X}}$ 2×2 matrix

(3) Inverse: a general invertible 2×2 matrix

$$\underset{2\times 2}{\begin{pmatrix} a & b \\ c & d \end{pmatrix}}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$ (if $ad \neq bc$)

Solution (1) $\mathbf{X}^T\mathbf{X}$

$$= \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

(2) $(X^T X)^{-1}$

$$= \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

(3) $$\underbrace{\underset{2\times n}{X^T} \; \underset{n\times 1}{Y}}_{2\times 1}$$

$$= \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix}$$

(4) $$\underset{2\times 1}{\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}} = \underbrace{(X^T X)^{-1}}_{2\times 2} \; \underbrace{X^T Y}_{2\times 1}$$

$$= \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \underset{2\times 2}{\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}} \times \underset{2\times 1}{\begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix}}$$

$$= \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 \times \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i Y_i \\ -\sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i + n \sum_{i=1}^{n} x_i Y_i \end{pmatrix}$$

Use $\sum_{i=1}^{n} Y_i = n\bar{Y}$ $\quad \sum_{i=1}^{n} X_i = n\bar{X}$

$$= \frac{1}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} X_i^2 \times n\bar{Y} - n\bar{X} \cdot \sum_{i=1}^{n} X_i Y_i \\ - n^2 \bar{X}\bar{Y} + n \sum_{i=1}^{n} X_i Y_i \end{pmatrix}$$

Equivalent to

$$= \begin{pmatrix} \bar{Y} - \hat{\beta}\bar{X} \\ \dfrac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \end{pmatrix}$$

OLS estimates $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$ by differentiation

2.4 Unique minimizer ?

The LS solution of MLR is unique as long as the columns of X are linearly independent.

(has full column rank)

Proof: (1) Hessian of $L(\beta) = \|Y - X\beta\|^2$

$$\underset{p\times 1}{\frac{\partial L(\beta)}{\partial \beta}} = -2X^T(Y - X^T\beta)$$

$$= 2X^TX\beta - \underline{2X^TY}$$

Hessian
$$\underset{p\times p}{\frac{\partial^2 L(\beta)}{\partial \beta\, \partial \beta^T}} = 2X^TX - 0 \qquad (\text{Qua } 0\text{-}4)$$

(2) Hessian is positive definite $\Rightarrow$ $L(\beta)$ is strictly convex and has a unique global minimizer

(Remains to show)

$\Rightarrow$ Hessian matrix is p.d. is columns of X are linearly independent

For $v \neq 0 \in \mathbb{R}^p$,

$$\underbrace{\underset{1\times p}{v^T}\,(\underset{p\times n}{X^T}\,\underset{n\times p}{X})\,\underset{p\times 1}{v}}_{1\times 1} = (Xv)^T(Xv) \qquad (Xv)^T = v^TX^T$$

$$= \|Xv\|^2$$

Since X's columns are linearly independent,

$$\Rightarrow Xv \neq 0 \Rightarrow \|Xv\|^2 > 0$$

Therefore, the Hessian of $L(\beta)$: $2X^TX$ is positive definite.