Linear Regression

Wei Li

Syracuse University

Spring 2021

Linear Regression

Notations

Linear Regression Models

Review on matrix theory

Linear regression (details)

Linear regression with orthogonal design

Regression by Successive Orthogonalization

Some further remarks

Notations

Notations

A random variable or random vector:

- Y: response variable
- X: random variable or random vector
  - if a p-dim random vector, $X = (X_1, \ldots, X_p)^T$.

Suppose that we have a random sample, that is say n copies of (Y, X)'s from certain population.

- Subscript i sometimes used to emphasizes for the ith observation, say the pair (Y_i, X_i), where $X_i = (X_{i,1}, \cdots, X_{i,p})^T$.

Observed values:

- y_i: the value of response variable for ith observation
- $\boldsymbol{x}_i$: the ith observed value of X
  - $\boldsymbol{x}_i$ could be a scalar of a vector. If a scalar, just x_i.

- $\mathbf{y}$: the n-dim response vector consisting of y_i.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- $\mathbf{X}$: the $n \times p$ design matrix
  - ith row is $\boldsymbol{x}_i^T$
  - jth column is $\mathbf{x}_j$

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_p \end{bmatrix}$$

All vector are taken as column vectors by default. Generic capital letter or bold-face capital letter will often denote a matrix, e.g., A or $\mathbf{A}$.

Linear Regression Models

Linear Regression Models

Given a list of random variables $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$. Here $X = (X_1, \ldots, X_p)^p$ is the covariate vector.

The covariates may come from different sources

- quantitative inputs; dummy coding qualitative inputs.
- transformed inputs: $\log(X_1), X_1^2, \sqrt{X_1}, \ldots$
- basis expansion: $X_1, X_1^2, X_1^3, \ldots$ (polynomial representation)
- interaction between variables: $X_1 X_2, \ldots$

Suppose we have a random sample $\{(Y_i, X_i)\}_{i=1}^n$. A standard linear regression model assumes

$$Y_i = X_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \text{ i.i.d }, \quad E(\epsilon_i) = 0, \text{Var}(\epsilon_i) = \sigma^2$$

- Y_i is the response for the ith observation, $X_i \in R^p$ is the covariates

classical model assumptions for simplicity:

- independence of errors ϵ_i
- constant error variance (homoscedasticity)
- ϵ_i independent of X_i.

note:

- normality of ϵ is not needed provided sample size is large.
- violation of homoscedasticity (heteroscedasticity) can be dealt with in general
- ϵ_i (mean) independent of X_i is the key for interpreting coefficients.

Review on matrix theory

Notations

- $\mathbf{x}$ is a nonzero $m \times 1$ vector
- $\boldsymbol{O}$ is a zero vector of $m \times 1$
- $\mathbf{e}_i, i = 1, \ldots, m$ is $m \times 1$ unit vector, with 1 in the ith position and zeros elsewhere.
- A an $m \times m$ matrix.
- The ith column of A can be expresed as $A\mathbf{e}_i$, for $i = 1, \ldots, m$
- col(A) : the subspace of R^m spanned by the columns of A.
- I_m is the identity matrix of size m.

Basic concepts

- The determinant of A is $\det(A) = |A|$.
- The trace of A is $\mathrm{tr}(A) =$ the sum of the diagonal elements.
- The roots of the m th degree of polynomial equation in λ.

$$|\lambda I_m - A| = 0$$

denoted by $\lambda_1, \cdots, \lambda_m$ are called the **eigenvalues** of A. The collection $\{\lambda_1, \cdots, \lambda_m\}$ is called the **spectrum** of A.

- Any nonzero $m \times 1$ vector $\mathbf{x}_i \neq \mathbf{0}$ such that

$$A\mathbf{x}_i = \lambda_i \mathbf{x}_i$$

is an **eigenvector** of A corresponding to the eigenvalue λ_i.

Orthogonal matrix

An $m \times n$ matrix U has orthonormal columns if $U^T U = I$.

An $m \times m$ matrix P is called an orthogonal matrix if

$$PP^T = P^T P = I_m, \quad \text{or } P^{-1} = P^T.$$

Any square matrix with orthonormal columns is an orthogonal matrix, and such a matrix must have orthonormal rows too.

If P is an orthogonal matrix, then $\left|PP^T\right| = |P|\left|P^T\right| = |P|^2 = |I| = 1$, so $|P| = \pm 1$

- For any A, we have $\operatorname{tr}\left(PAP^T\right) = \operatorname{tr}\left(AP^T P\right) = \operatorname{tr}(A)$.
- PAP^T and A have the same eigenvalues, since

$$\left|\lambda I_m - PAP^T\right| = \left|\lambda PP^T - PAP^T\right| = |P|^2 \left|\lambda I_m - A\right| = \left|\lambda I_m - A\right|$$

Spectral decomposition of a symmetric matrix

If A is symmetric, there exists an orthogonal matrix P such that

$$P^T AP = \Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_m\}$$

- λ_i 's are the eigenvalues of A.
- The eigenvectors of A are the column vectors of P.
- Denote the ith column of P by $\mathbf{p}_i$, then

$$PP^T = \sum_{i=1}^{m} \mathbf{p}_i \mathbf{p}_i^T = I_m$$

- The **spectral decomposition** of A is

$$A = P\Lambda P^T = \sum_{i=1}^{m} \lambda_i \mathbf{p}_i \mathbf{p}_i^T$$

- $\text{tr}(A) = \text{tr}(\Lambda) = \sum_{i=1}^{n} \lambda_i$ and $|A| = |\Lambda| = \prod_{i=1}^{m} \lambda_i$

Idempotent Matrices

An $m \times m$ matrix A is idempotent if

$$A^2 = AA = A$$

- The eigenvalues of an idempotent matrix are either zero or one

$$\begin{aligned} \lambda \mathbf{x} = A\mathbf{x} = A(A\mathbf{x}) = A(\lambda \mathbf{x}) = \lambda^2 \mathbf{x} \\ \Longrightarrow \lambda = \lambda^2 \end{aligned}$$

- If A is idempotent, so is $I_m - A$.

Projection Matrices

A symmetric, idempotent matrix A is called a **projection matrix**.

If A is a $m \times m$ symmetric idempotent , then

- If $\text{rank}(A) = r$, then A has r eigenvalues equal to 1 and $m - r$ zero eigenvalues.
- $\text{tr}(A) = \text{rank}(A)$
- $I_m - A$ is also symmetric idempotent, of rank $m - r$.

Given $\mathbf{x} \in R^m$, define $\mathbf{y} = A\mathbf{x}, \mathbf{z} = (I - A)\mathbf{x} = \mathbf{x} - \mathbf{y}$. Then

- $\mathbf{y} \perp \mathbf{z}$
- $\mathbf{y}$ is the orthogonal projection of x onto the subspace $\text{col}(A)$.
- $\mathbf{z} = (I - A)\mathbf{x}$ is the orthogonal projection of x onto the complementary subspace such that

$$\mathbf{x} = \mathbf{y} + \mathbf{z} = A\mathbf{x} + (I - A)\mathbf{x}$$

Linear regression (details)

Linear regression (details)

- The response vector $\mathbf{y} = (y_1, \cdots, y_n)^T$
  - The design matrix $\mathbf{X}$.
  - Assume the first column of $\mathbf{X}$ is $\mathbf{1}$
  - The dimension of $\mathbf{X}$ is $n \times (1 + p)$.
  - The regression coefficients $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}$.
  - The error vector $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$

The linear model is written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

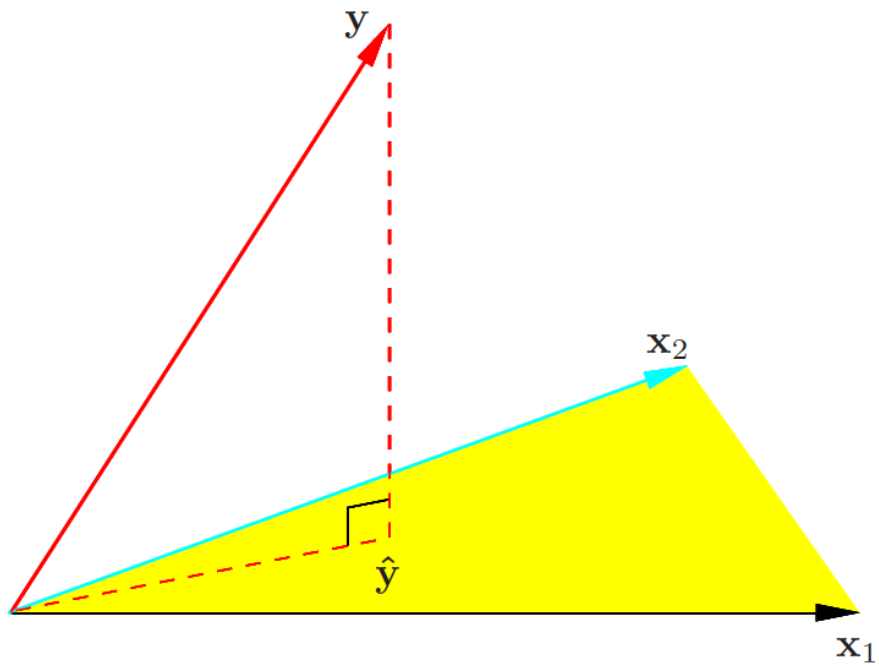
- the estimated coefficients $\hat{\boldsymbol{\beta}}$
- the predicted response $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

$$\min_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- Normal equations: $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$
- $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = P_{\mathbf{X}}\mathbf{y}$
- Residual vector is $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (I - P_{\mathbf{X}})\,\mathbf{y}$.
- Residual sum squares $RSS = \mathbf{r}^T\mathbf{r}$.

Call the following square matrix the projection or hat matrix:

$$P_{\mathbf{X}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$



ESL: Fig 3.2

Properties:

- symmetric and non-negative definite
- idempotent: $P_{\mathbf{X}}^2 = P_{\mathbf{X}}$. The eigenvalues are 0 's and 1 's.
- $P_{\mathbf{X}}\mathbf{X} = \mathbf{X}, \quad (I - P_{\mathbf{X}})\,\mathbf{X} = 0$

We have

$$\mathbf{r} = (I - P_{\mathbf{X}})\,\mathbf{y}, \quad RSS = \mathbf{y}^T\,(I - P_{\mathbf{X}})\,\mathbf{y}$$

Note

$$\mathbf{X}^T\mathbf{r} = \mathbf{X}^T\,(I - P_{\mathbf{X}})\,\mathbf{y} = 0$$

The residual vector is orthogonal to the column space spanned by $\mathbf{X}$, col($\mathbf{X}$).

sampling properties of $\hat{\boldsymbol{\beta}}$

- $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$
- The variance σ^2 can be estimated as

$$\hat{\sigma}^2 = RSS/(n-p-1)$$

This is an unbiased estimator, i.e.,$\text{E}\left(\hat{\sigma}^2\right) = \sigma^2$

Inferences under Gaussian Errors:

Under the Normal assumption on the error ϵ, we have

- $\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$
- $(n-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p-1}$
- $\hat{\boldsymbol{\beta}}$ is independent of $\hat{\sigma}^2$

To test $H_0 : \beta_j = 0$, we use

- if σ^2 is known, $z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{v_j}}$ has a standard normal distribution under H_0 where v_j is the j th diagonal element of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$;
- if σ^2 is unknown, $t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$ has a t_{n-p-1} distribution under H_0.

Confidence intervals for coefficients:

- Under Normal assumption, the $100(1-\alpha)\%$ C.I. of β_j is

$$\hat{\beta}_j \pm t_{n-p-1,\frac{\alpha}{2}} \hat{\sigma} \sqrt{v_j}$$

  where $t_{k,\nu}$ is ν upper-percentile of t_k distribution.

- In practice, we use the approximate $100(1-\alpha)\%$ C.I. of β_j

$$\hat{\beta}_j \pm z_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{v_j}$$

  where $Z_{\frac{\alpha}{2}}$ is $\frac{\alpha}{2}$ upper percentile of the standard Normal distribution.

- Even if the Gaussian assumption does not hold, this interval is approximately right, with the coverage probability $1-\alpha$ as $n \to \infty$.

Confidence intervals and prediction intervals for means:

Let for some fixed values $\mathbf{x}_0$ for $\mathbf{x}$.

- The $100(1-\alpha)\%$ confidence interval for $E(Y|X=\mathbf{x}_0)$ is given by

$$\hat{y}_0 \pm z_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}$$

  where $\hat{y}_0 = \mathbf{x}_0^T\hat{\boldsymbol{\beta}}$.
- The $100(1-\alpha)\%$ prediction interval for the value of Y when $X=\mathbf{x}_0$ is given by

$$\hat{y}_0 \pm z_{\alpha/2}\hat{\sigma}\sqrt{1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}$$

F distribution:

Distributions of Quadratic Form:

- If $X \sim N_p(\boldsymbol{\mu}, I_p)$, then

$$W = X^T X = \sum_{i=1}^{p} X_i^2 \sim \chi_p^2(\lambda), \quad \lambda = \boldsymbol{\mu}^T \boldsymbol{\mu}$$

- Special case: If $X \sim N_p(\mathbf{0}, I_p)$, then $W = X^T X \sim \chi_p^2$
- If $U_1 \sim \chi_p^2, U_2 \sim \chi_q^2$ and $U_1 \perp U_2$, then

$$F = \frac{U_1/p}{U_2/q} \sim F_{p,q}$$

- If $U_1 \sim \chi_p^2(\lambda), U_2 \sim \chi_q^2$ and $U_1 \perp U_2$, then

$$F = \frac{U_1/p}{U_2/q} \sim F_{p,q}(\lambda), \quad (\text{ noncentral } F)$$

Example: Assume $\mathbf{y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n\right)$.

$$RSS = \mathbf{y}^T(I - P_{\mathbf{X}})\mathbf{y} = \|\mathbf{r}\|^2, \quad SSR = \mathbf{y}^T P_{\mathbf{X}}\mathbf{y} = \|\hat{\mathbf{y}}\|^2$$

Then

$$F = \frac{SSR/(p+1)}{RSS/(n-p-1)} \sim F_{p+1,n-p-1}(\lambda), \quad \lambda = \|X\beta\|^2/\sigma^2$$

Nested Model Selection:

To test for significance of groups of coefficients simultaneously, we use F-statistic

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1 - 1)}$$

where

- RSS_1 is the RSS for the bigger model with $p_1 + 1$ parameters
- RSS_0 is the RSS for the **nested** smaller model with $p_0 + 1$ parameter, have $p_1 - p_0$ parameters constrained to zero.

F-statistic measure the change in RSS per additional parameter in the bigger model, and it is normalized by $\hat{\sigma}^2$.

- Under the assumption that the smaller model is correct,

$$F \sim F_{p_1 - p_0, n - p_1 - 1}$$

Confidence set

- The approximate confidence set of $\boldsymbol{\beta}$ is

$$C_{\boldsymbol{\beta}} = \left\{ \boldsymbol{\beta} \mid (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \left(\mathbf{X}^T \mathbf{X} \right) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \hat{\sigma}^2 \chi^2_{p+1,\alpha} \right\}$$

  where $\chi^2_{k,\alpha}$ is α upper percentile of χ^2_k distribution.
- The confidence interval for the true function $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}$ is

$$\left\{ \boldsymbol{x}^T \boldsymbol{\beta} \mid \boldsymbol{\beta} \in C_{\boldsymbol{\beta}} \right\}$$

Linear regression with orthogonal design

Linear regression with orthogonal design

- If X is univariate, the least square estimate is

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{< \mathbf{x}, \mathbf{y} >}{< \mathbf{x}, \mathbf{x} >}$$

- if $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d]$ has orthogonal columns, i.e.,

$$< \mathbf{x}_j, \mathbf{x}_k >= 0, \quad \forall j \neq k$$

or equivalently, $\mathbf{X}^T\mathbf{X} = \text{diag}\left(\|\mathbf{x}_1\|^2, \ldots, \|\mathbf{x}_d\|^2\right)$. The OLS estimates are given as

$$\hat{\beta}_j = \frac{\langle \mathbf{x}_j, \mathbf{y} \rangle}{\langle \mathbf{x}_j, \mathbf{x}_j \rangle} \quad \text{for} \quad j = 1, \ldots, d$$

  - Each input has no effect on the estimation of other parameters.
  - Multiple linear regression reduces to univariate regression.

Regression by Successive Orthogonalization

To orthogonalize $\mathbf{X}$

Consider $\mathbf{y} = \beta_0\mathbf{x}_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \boldsymbol{\epsilon}$. $(\mathbf{x}_0 = \mathbf{1})$ Orthogonization process:

(1) We regress $\mathbf{x}_1$ onto $\mathbf{x}_0$, compute the residual

$$\mathbf{z}_1 = \mathbf{x}_1 - \gamma_{01}\mathbf{x}_0. \quad (\text{ note } \mathbf{z}_1 \perp \mathbf{x}_0)$$

(2) We regress $\mathbf{x}_2$ onto $(\mathbf{x}_0, \mathbf{z}_1)$, compute the residual

$$\mathbf{z}_2 = \mathbf{x}_2 - \gamma_{02}\mathbf{x}_0 - \gamma_{12}\mathbf{z}_1. \quad (\text{ note } \mathbf{z}_2 \perp \{\mathbf{x}_0, \mathbf{z}_1\})$$

Note: $\text{span}\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2\} = \text{span}\{\mathbf{x}_0, \mathbf{z}_1, \mathbf{z}_2\}$.

More generally, we may use Gram-Schmidt procedure, to transform $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_p)$ to $\mathbf{Z} = (\mathbf{z}_0, \ldots, \mathbf{z}_p)$ where $\mathbf{z}_j$ is the residual of regress $\mathbf{x}_j$ on $\mathbf{x}_0, \ldots, \mathbf{x}_{j-1}$ In fact, such a $\mathbf{Z}$ has orthogonal columns. $\{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_p\}$ forms orthogonal basis for $\mathrm{Col}(\mathbf{X})$.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$
2. For $j = 1, \ldots, p$, successively perform the following: regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ to produce coefficients

$$\hat{\gamma}_{kj} = \frac{\langle \mathbf{z}_k, \mathbf{x}_j \rangle}{\langle \mathbf{z}_k, \mathbf{z}_k \rangle}$$

   for $k = 0, \ldots, j-1$, and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$.
3. Regress $\mathbf{y}$ on $\mathbf{z}_p$ to get

$$\hat{\beta}_p = \hat{\eta}_p = \frac{\langle \mathbf{y}, \mathbf{z}_p \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}.$$

4. To compute $\hat{\beta}_j$, for $j = p-1, \ldots, j = 0$: regress Regress $\mathbf{y}$ on $\mathbf{z}_j$ to get $\hat{\eta}_j$ for all $j = 0, \ldots, p-1$,

$$\hat{\eta}_j = \frac{\langle \mathbf{z}_j, \mathbf{y} \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle}.$$

Let Γ be the $(p+1) \times (p+1)$ upper triangular matrix with all diagonal elements equal to 1 and $\Gamma_{ij} = \hat{\gamma}_{i-1,j-1}$ for $j > i \geq 1$. Solve for $\hat{\beta}_j$, for $j = p-1, \ldots, j = 0$ recursively from $\Gamma \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\eta}}$.

In general, for arbitrary index j, we can put the j-th regression in the **last** column, then do the orthogonalizaiotn process to obtain $\hat{\beta}_j$.

Multicollinearity

For the term $j = p$ (the step 3 in above procedure), the p-th coefficient (the last coefficient)

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}$$

If $\mathbf{x}_p$ is highly correlated with some of the other $\mathbf{x}_j's$, then

- The residual vector $\mathbf{z}_p$ is close to zero
- The coefficient $\hat{\beta}_p$ will be very unstable
- The variance estimate

$$\text{Var}\left(\hat{\beta}_p\right) = \frac{\sigma^2}{\|\mathbf{z}_p\|^2}$$

The precision for estimating $\hat{\beta}_p$ depends on the length of $\mathbf{z}_p$, or, how much $\mathbf{x}_p$ is unexplained by the other (or previous) $\mathbf{x}_k$'s

Computational algorithms

Consider the Normal Equation

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}$$

We like to avoid computing $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ directly.

(1) QR decomposition of $\mathbf{X}$:

- $\mathbf{X} = QR$ where Q is orthonormal and R is upper triangular

(2) Cholesky decomposition of $\mathbf{X}^T\mathbf{X}$:

- $\mathbf{X}^T\mathbf{X} = \tilde{R}\tilde{R}^T$ where $\tilde{R}$ is lower triangular

QR algorithm

We can represent step 2 of the above Algorithm in matrix form:

$$\mathbf{X} = \mathbf{Z}\Gamma$$
$$\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_p] \text{ and } \mathbf{Z} = [\mathbf{z}_0, \ldots, \mathbf{z}_p]$$

Standardizing $\mathbf{Z}$ using $D = \text{diag}\{\|\mathbf{z}_0\|, \ldots, \|\mathbf{z}_p\|\}$,

$$\mathbf{X} = \mathbf{Z}\Gamma = \mathbf{Z}D^{-1}D\Gamma \equiv QR, \quad \text{with } Q = ZD^{-1}, \quad R = D\Gamma$$

- The columns of Q consists of an orthonormal basis for the column space of X.
- Q is orthonormal matrix of $n \times (p+1)$, satisfying $Q^TQ = I$.
- R is upper triangular matrix of $(p+1) \times (p+1)$, full-rank.

We then can show

$$R\boldsymbol{\beta} = Q^T\mathbf{y}$$

Based on this, we solve for $\hat{\boldsymbol{\beta}}$ as follows:

(1) Conduct QR decomposition of $\mathbf{X} = QR$. (Gram-Schmidt Orthogonalization)
(2) Compute $Q^T\mathbf{y}$
(3) Solve the triangular system $R\boldsymbol{\beta} = Q^T\mathbf{y}$.

Cholesky Decomposition algorithm

For any positive definite square matrix A, we have

$$A = RR^T$$

where R is a lower triangular matrix of full rank.

(1) Compute $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$
(2) Factoring $\mathbf{X}^T\mathbf{X} = RR^T$, then $\hat{\beta} = \left(R^T\right)^{-1} R^{-1}\mathbf{X}^T\mathbf{y}$
(3) Solve the triangular system $R\mathbf{w} = \mathbf{X}^T\mathbf{y}$ for $\mathbf{w}$.
(4) Solve the triangular system $R^T\boldsymbol{\beta} = \mathbf{w}$ for $\boldsymbol{\beta}$.

Choleksy decompostion algorithm can be faster than QR for small d, but can be less stable.

Some further remarks

The role of $E(Y|X)$ in our interpretation

It is common to interpret the coefficient, say β_1 as the "'effect" on the average value of Y from increasing X_1 by one unit while holding the other predictors or covariates unchanged. This interpretation however hings upon the assumption that ϵ_i is independent of all X's, or more precisely,

$$E(\epsilon|X) = 0.$$

It is important to emphasize that linear regression models seldom satisfy this assumption in practice.

It is always possible to write

$$Y = \mu(X) + \epsilon$$

where $\mu(X) = E(Y|X)$ and ϵ satisfies $E(\epsilon|X) = 0$.

Here $\mu(X)$ is called the **regression function**, with which we like to predict Y.

For a linear regression coefficients to have meaningful interpretation, one essentially believe that the linear model $X^T\beta$ is correctly specified for $E(Y|X)$.

The statistical meaning of $\mu(X)$

$$\begin{aligned}\text{MSE}(f) &= E\left[(Y - f(X))^2\right] \\ &= E\left[V[Y \mid X] + (E[Y - f(X) \mid X])^2\right]\end{aligned}$$

The optimal function $\hat{f}$ is given by

$$\mu(x) = E[Y \mid X = x]$$

In other words, given X, the best predictor for Y is the conditional expectation $E[Y \mid X]$ (in mean-squared sense).

Causal relationship?

In most classical courses in linear regression, X is viewed as "independent variable", while Y viewed as "dependent" variable, which may seems to suggest some **causal relationship** between them. However this is not necessarily so.

The conditional expectation $E(Y|X)$ or $E(X|Y)$ may be defined regardless of the actual causal relationship between X and Y.

In the so-called structural equations framework, $\mu(X)$ may have structural meaning (often suggested by theory), which means X is viewed as a **direct cause** of Y. In that case, it might only make sense to consider $E(Y|X)$. But a causal relationship can be studied in a different framework that subsumes the structural equations framework.

Linear Smoothers

To predict at some arbitrary point x using a simple linear regression,

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x = \sum_{i=1}^{n} \frac{1}{n} \left(1 + \frac{(x - \bar{x})(x_i - \bar{x})}{\hat{\sigma}_X^2} \right) y_i$$

where $\hat{\sigma}_X^2 = \sum_i (x_i - \bar{x})^2 / n$. Note that it takes the form

$$\hat{\mu}(x) = \sum_{i=1}^{n} w(x, x_i) \cdot y_i = \mathbf{w}(x)\mathbf{y}$$

Using a multiple linear regression,

$$\hat{\mu}(x) = x^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{w}(x)\mathbf{y}$$

We call a regression function of this form a **linear smoother** (note this is so-named because it is linear in y, but need not behave linearly as a function of x!).

Another example: k-nearest-neighbors regression

$$w\left(x, x_i\right) = \begin{cases} 1/k & \text{if } x_i \text{ is one of the } k \text{ nearest points to } x \\ 0 & \text{otherwise} \end{cases}$$

that is,

$$\hat{\mu}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$

where $N_k(x)$ gives the k nearest neighbors of x.

Linear smoothers include many important classes of estimators: kernel regression, locally weighted regression, Gaussian process regression, smoothing splines, series or sieve regression.