Linear Regression: Ordinary Least Square

Xiang Zhou

School of Data Science
Department of Mathematics
City University of Hong Kong

Ordinary Linear Regression

Review of linear regression (univariate and multivariate)

- Least-square: is usually credited to Carl Friedrich Gauss (1795), but it was first published by Adrien-Marie Legendre (1805). history note. The approach was first successfully applied to problems in **astronomy**.
- Loss function: squared error loss $\ell(y, \hat{y}) = |y - \hat{y}|^2$
- Hypothesis space (model class): linear function (affine function with intercept)

History note : "method of least squares" by Gauss and Legendre

Based on d'Alembert's principle, Gauss derived *Principle of least constraint*:

$$Z = \sum_{i=1}^{N} \frac{1}{2m_i} (\boldsymbol{F}_i - m_i \boldsymbol{A}_i)^2$$

$\boldsymbol{F}_i$ and $\boldsymbol{A}_i$ are the forces and accelerations, respectively. For free particles, it recovers the classic Newton's motion $\boldsymbol{F}_i = m_i \boldsymbol{A}_i$. If constraints prevent the free choice of the $\boldsymbol{A}_i$, we can still minimize Z under the given auxiliary conditions. The solution obtained yields the actual motion of the system realized in nature.

Example

A particle is forced to stay on the surface $z = c(x, y)$ by the action of the force $\boldsymbol{F}$. Find the motion of the equation. Hint: $\dot{z} = c_x \dot{x} + c_y \dot{y}$ and $\ddot{z} = c_x \ddot{x} + c_{xx} \dot{x}^2 + c_{yy} \ddot{y} + c_{yy} \dot{x}^2 \approx c_x \ddot{x} + c_y \ddot{y}$. The constraint for $\boldsymbol{A} = (\ddot{x}, \ddot{y}, \ddot{z})$ is the linear equation $\ddot{z} = c_x \ddot{x} + c_y \ddot{y}$.

Data $(x_1, y_1), \ldots, (x_n, y_n)$, where

- x_i is the predictor (independent variable, input, feature)
- y_i is the response (dependent variable, output, outcome)

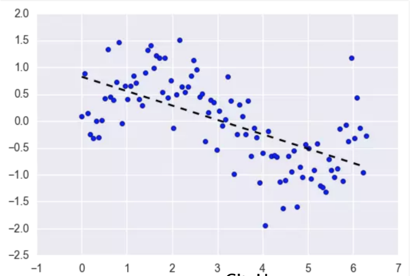
We denote the *regression function* as

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

The linear regression model assumes a specific linear form for f,

$$f(x) = \beta_0 + \beta x,$$

which is usually thought of as an approximation to the truth.



Least squared fitting

Minimize:

$$(\hat{\beta}_0, \hat{\beta}) = \underset{\beta_0, \beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i)^2.$$

Solution is:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \bar{x}\hat{\beta}. \end{aligned}$$

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta} x_i$ are the fitted values
- $r_i = y_i - \hat{y}_i$ are the residuals

Standard errors and confidence intervals

Assume further that

$$y_i = \beta_0 + \beta x_i + \epsilon_i,$$

where $E(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Then

$$se(\hat{\beta}) = \left(\frac{\sigma^2}{\sum (x_i - \bar{x})^2} \right)^{1/2},$$

where σ^2 can be estimated by $\hat{\sigma}^2 = \sum (y_i - \hat{y})^2/(n-2)$.

Under additional normality assumption of ϵ_i's, a $(1-\alpha)100\%$ confidence interval of β is

$$\hat{\beta} \pm z_{\alpha/2} \widehat{se}(\hat{\beta}).$$

Ordinary Least Square (OLS)

- The predictor variable $x = (x_0 \equiv 1, x_1, \ldots, x_p)$ and **Design Matrix**

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ & & \ldots & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \in \mathbb{R}^{n\times(p+1)}.$$

  n is the number of samples. The first column $x_{i0} \equiv 1$.

- Response vector : $Y = \left[y_1, y_2, \ldots, y_n\right]^\mathsf{T}$.
- Linear model $\mathcal{H} = \left\{f : f(x) = \beta^\mathsf{T} x, \beta = (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}\right\}$.
- Risk minimization view:

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \|Y - X\beta\|_2^2 = \left(X^\mathsf{T} X\right)^{-1} X^\mathsf{T} Y.$$

- Model-based interpretation:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Standarlization of Data

The standarlization processing is helpful in many cases:

1. Centering
   - $x_{ij} \to x_{ij} - \bar{x}_{\cdot j}$, where $\bar{x}_{\cdot j} = \frac{1}{n}\sum_i x_{ij}$
   - $y_i \to y_i - \bar{y}$

   Then $\sum_i x_{ij} = \sum_i y_i = 0$. Then the intercept in OLS β_0 vanishes. For centered data: $\frac{1}{n}X^\top X = \frac{1}{\sum_i}(x_{ij}x_{ik})$ is the covariate matrix of the predictor.

2. Standardization (after centering):

$$x_{ij} \to \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_i x_{ij}^2}}.$$

   Then $\frac{1}{n}\sum_i x_{ij}^2 \equiv 1, \ \forall j$.

1. Understanding OLS from the perspective of MLE and Bayes
2. Understanding OLS from the perspective of linear algebra: orthogonal project, pseudo-inverse, Gram-Schmidt procedure; QR, SVD
3. Understanding uncertainty in $\hat{\beta}$: variance analysis
4. Understanding OSL as the minimum variance unbiased estimator of the response : Gauss-Markov theorem

Maximum log-*likelihood function*

$\varepsilon \sim \mathcal{N}(0, \sigma^2)$ leads to the log-likelihood function

$$
\begin{aligned}
\log \mathcal{L}(\beta; x_i, y_i) &= \log \prod_{i=1}^{n} p(y_i|x_i)p(x_i) = \sum_{i=1}^{n} \log p(y_i|x_i) + \sum_{i=1}^{n} \log p(x_i) \\
&= \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta^\intercal x_i)^2}{2\sigma^2}} \right] + \sum_{i=1}^{n} \log p(x_i) \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^\intercal x_i)^2 + \text{terms not depend on } \beta.
\end{aligned}
$$

Therefore $\hat{\beta}^{\mathsf{MLE}} = \hat{\beta}^{\mathsf{OLS}}$.

- Understanding OLS from the perspective of linear algebra: orthogonal project, pseudo-inverse, Gram-Schmidt procedure; QR, SVD

OLS prediction as the orthogonal projection

- The optimal prediction

$$\hat{Y} = X\hat{\beta} = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y =: \boxed{\mathrm{Proj}_{\mathsf{X}}\, Y} \tag{1}$$

  is the orthogonal projection of the vector $Y \in \mathbb{R}^n$ onto the subspace spanned by the $p+1$ column vectors of the matrix X

$$\mathsf{X} = \mathrm{span}\{X_0, X_1, \ldots, X_p\}$$

- $\hat{Y}$ is the point in $\mathbb{R}^n$ with the shortest Euclidian distance to this subspace X.
- It would be nice if we have a set of $p+1$ *orthonormal basis vector* of X. This can be done by Gram-Schmidt procedure (Sec. 3.2.3. in [ESL] under the name "sequential linear regression") .
- In addition, one can use QR, SVD decomposition of $X^{\mathsf{T}}X$. To efficiently find the orthogonal projection of the vector Y onto a subspace spanned by X_i in $\mathbb{R}^n$ is a classic topic in numerical linear algebra.

Properties of Projection matrix

$$P = \mathrm{Proj}_{\mathsf{X}} = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$$

satisfies

- symmetric: $P = P^{\mathsf{T}}$;
- idempotent: $P^2 = \mathbf{I}_n$ identity matrix;
- rank = $\dim(\mathsf{X}) = p + 1$
- eigenvalues: $p + 1$ ones and $n - (p + 1)$ zeros;
- trace = $\dim(\mathsf{X})$.

Other names used in statistics literature for the projection matrix $\mathrm{Proj}_{\mathsf{X}}$

- influence matrix;
- hat matrix

Singular Value Decomposition

- Assume $X = UDV^{\mathsf{T}}$ is a SVD of the design matrix X, then $D = \text{diag}\{d_0, \ldots, d_p\}$, d_i is the singular value of X.
- The column vectors of U, $\{U_i, 0 \leq i \leq p\}$, is a set of orthonormal basis of X.
- Then $X^{\mathsf{T}}X = VD^2V^{\mathsf{T}}$, and $\mathrm{Proj}_{\mathsf{X}} = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}} = (UDV^{\mathsf{T}})VD^{-2}V^{\mathsf{T}}VDU^{\mathsf{T}} = UU^{\mathsf{T}}$.
- $$\hat{Y} = \mathrm{Proj}_{\mathsf{X}}\, Y = UU^{\mathsf{T}}Y = \sum_{i=0}^{p} \alpha_i U_i, \quad \text{where} \quad \alpha_i = U_i \cdot Y.$$

Exercise

The projection matrix $\mathrm{Proj}_{\mathsf{X}}$ *has the trace* $p+1$.

(Hint $\mathrm{Trace}(AB) = \mathrm{Trace}(BA)$. The eigenvalues of the projection matrix are either 0 or 1.)

Exercise

Exercise 3.4 in [ESL].

The decomposition of sum-of-squares

For the OLS predicted response $\hat{Y} = X\hat{\beta}$, we have

$$\boxed{SST = SSR + SSE}$$

- SST= total sum of squares for the response variable

$$SST = \sum_i (y_i - \bar{y})^2 = \left\|Y - \bar{Y}\right\|_2^2$$

- SSE=sum of squares of errors [1]

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \|Y - \mathsf{proj}_{\mathsf{X}} Y\|_2^2$$

- SSR = sum of squares explained by regression

$$SSR = \sum_i (\hat{y}_i - \bar{\hat{y}})^2 = \left\|\hat{Y} - \bar{Y}\right\|_2^2$$

Note that the average of the training response $\bar{y}$ is equal to the average of predicted response $\bar{\hat{y}}$

[1] [ISL] [ESL] name this as RSS= residual sum of squares

Proof of $SST = SSE + SSR$: Exercise! (consider $Z = Y - \bar{y}1_n$ and $1_n = X_0 \in \mathsf{X}$. consider f centered data where $\bar{y} = 0$.)

Exercise

Show that

$$SSE = \|(\mathbf{I}_n - \mathrm{Proj}_{\mathsf{X}})\varepsilon\|_2^2 = \|\mathrm{Proj}_{\mathsf{X}^\perp}(\varepsilon)\|_2^2$$

$\mathbf{I}_n - \mathrm{Proj}_{\mathsf{X}}$ *is called residual marker matrix sometimes.*

by using $Y = X\beta + \varepsilon$ and $\hat{Y} = \mathrm{Proj}_{\mathsf{X}} Y$.
Draw a picture to illustrate this result.

- Understanding uncertainty in $\hat{\beta}$: unbiasedness, consistence, variance analysis

The distribution of the OLS coefficient $\hat{\beta}$

Since $Y = X\beta + \varepsilon$, then

$$\begin{aligned}\hat{\beta} &= (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y = (X^\mathsf{T}X)^{-1}X^\mathsf{T}(X\beta + \varepsilon) \\ &= \beta + (X^\mathsf{T}X)^{-1}X^\mathsf{T}\varepsilon\end{aligned}$$

Note that $\varepsilon \sim N(0, \sigma^2 I_n)$, thus

$$\mathbb{E}\,\hat{\beta} = \beta \qquad \text{(unbiased estimator)}$$

$$\begin{aligned}\mathbb{V}(\hat{\beta}) &= \mathbb{V}((X^\mathsf{T}X)^{-1}X^\mathsf{T}\varepsilon) \\ &= (X^\mathsf{T}X)^{-1}X^\mathsf{T}\,\mathbb{V}(\varepsilon)(X^\mathsf{T}X^{-1}X^\mathsf{T})^\mathsf{T} \\ &= \sigma^2(X^\mathsf{T}X)^{-1}X^\mathsf{T}I_nX(X^\mathsf{T}X)^{-1} \\ &= \sigma^2(X^\mathsf{T}X)^{-1}.\end{aligned}$$

Therefore,

$$\boxed{\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\mathsf{T}X)^{-1})},$$

from which the confidence interval of $\hat{\beta}$ can be calculated.

Consistency of $\hat{\beta}$

Assume that

$$\lim_{n\to\infty}\left(\frac{X^{\mathsf{T}}X}{n}\right)=\Delta$$

exists as a nonstochastic and nonsingular matrix (for example, $|x_{ji}|\leq c$ is bounded). Then

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\,|\hat{\beta}-\beta|^2 &= \lim_{n\to\infty}\mathbb{V}(\hat{\beta})\\ &= \sigma^2\lim_{n\to\infty}\frac{1}{n}\left(\frac{X^{\mathsf{T}}X}{n}\right)^{-1}\\ &= \sigma^2\lim_{n\to\infty}\frac{1}{n}\Delta^{-1}\\ &= 0\end{aligned}$$

This implies that OLSE $\hat{\beta}$ converges to in quadratic mean. Thus OLSE $\hat{\beta}$ is a consistent estimator of β.

- The distribution of $\hat{Y} = X\hat{\beta}$ is then $\mathcal{N}(X\beta, \sigma^2 X(X^\mathsf{T}X)^{-1}X^\mathsf{T})$
- When a new data of input x arrives, taking value $x_i = a_i, \quad i = 1, \ldots, p$, with $a = (1, a_1, a_2, \ldots, a_p)^\mathsf{T} \in \mathbb{R}^{p+1}$, then the prediction from the regression equation is

$$\hat{y} := a^\mathsf{T}\hat{\beta} \sim \mathcal{N}(a^\mathsf{T}\beta,\ \sigma^2 a^\mathsf{T}(X^\mathsf{T}X)^{-1})a)$$

  which can give the confidence interval of $\hat{y} = a^\mathsf{T}\hat{\beta}$.
- But remember that in our model $Y = X\beta + \varepsilon$, it is assumed that the data you *observe* inevitably is contaminated by the measurement error ε. By including this measurement error, the predicted value at this new input $x = a$ is

$$\hat{y} + \varepsilon_a = a^\mathsf{T}\hat{\beta} + \varepsilon_a$$

  where ε_a is $\mathcal{N}(0, \sigma_a^2)$ and independent of the training data you used to build the regression equation.
  It is clear that the distribution of $\hat{y} + \varepsilon_a$ is

$$\mathcal{N}(a^\mathsf{T}\beta,\ \sigma^2 a^\mathsf{T}(X^\mathsf{T}X)^{-1})a + \sigma_a^2),$$

  which gives the prediction interval.

The variance of the measurement error σ^2

- Recall SST is the sample variance of Y then $\mathbb{E}\, SST = (n-1)\sigma^2$ since $\mathbb{V}(Y) = \mathbb{V}(\varepsilon) = \sigma^2$.
- We show below that $\mathbb{E}\, SSE = (n-p-1)\sigma^2$
- Which one among SST and SSE should be used to define $\hat{\sigma}^2$, the estimate of the variance of ε?

From exercise, we have

$$SSE = \|\mathrm{Proj}_{\mathsf{X}^\perp}(\varepsilon)\|_2^2 = \varepsilon^\mathsf{T}(\mathrm{Proj}_{\mathsf{X}^\perp})^\mathsf{T}(\mathrm{Proj}_{\mathsf{X}^\perp})\varepsilon.$$

where the Gaussian vector ε have variance matrix $\sigma^2 I_n$. Since the dimension $\dim X^\perp = n - \dim(\mathsf{X}) = n - (p+1)$, then $\mathrm{Trace}(\mathrm{Proj}_{\mathsf{X}^\perp}) = n - (p+1)$. Then we have the conclusion

$$\mathbb{E}\, SSE = \mathrm{Trace}((\mathrm{Proj}_{\mathsf{X}^\perp})\sigma^2 I_n) = (n-(p+1))\sigma^2.$$

Exercise

Let $\mu = \mathbb{E}(X)$ *and* $\Sigma = \mathbb{V}(X)$ *be the mean vector and the covariance matrix of the random vector* X *in* $\mathbb{R}^n$. M *is* $n \times n$ *symmetric matrix. Define the random variable* $z = (X - \mu)^{\mathsf{T}} M (X - \mu)$, *then*

$$\mathbb{E}(z) = \mathrm{Trace}(M\Sigma) = \mathrm{Trace}(\Sigma M)$$

and thus

$$\mathbb{E}(X^{\mathsf{T}} M X) = \mathrm{Trace}(M\Sigma) + \mu^{\mathsf{T}} M \mu.$$

- Understanding OSL as the best linear unbiased estimator (BLUE) with the smallest MSE.

Gauss-Markov theorem (Rao, 1973)

- Recall the ML basics: Given a training dataset $\mathtt{D}$, the function to approximate in the hypothesis space $\mathcal{H}$ $\hat{f}_{\mathtt{D}} \in \mathcal{H}$ is a function x. In OLS, we assumed that $\hat{f}_{\mathtt{D}} = \beta^{\mathsf{T}} x$ is a linear function of x parametrized by β.
- Now, if we fix a testing input $x = a$ now, $\hat{f}_{\mathtt{D}}(a)$ then is a mapping (statistics) from $\mathtt{D}$ to $\mathcal{Y}$. What if we assume this mapping is linear and consider the **MVU**(minimum variance unbiased) estimator of the ground truth $\beta^{\mathsf{T}} a$ when $x = a$?
- Fix the design matrix X, then this estimator takes the linear form

$$Y \to c^{\mathsf{T}} Y$$

with the coefficient $c \in \mathbb{R}^n$.

Theorem (Gauss-Markov Theorem)

Let u be an unbiased estimate of the ground truth response $a^T\beta$ at the new input $x = a$ in the space of linear transformations from the response training data $Y = X\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I_n)$. This is to say that $u = c^T Y$ for some vector $c \in \mathbb{R}^{(p+1)}$ satisfies $\mathbb{E}\, u = a'\beta$ for any *β in $\mathbb{R}^{p+1}$. Then*

$$\mathit{Var}(u) \geq \mathit{Var}(\hat{y}) = \sigma^2 a^T (X^T X)^{-1} a$$

where $\hat{y} = a^T \hat{\beta}^{OLS} = a^T (X^T X)^{-1} X^T Y$. (see Exercise 3.3 in [ESL].)

Proof.

$\mathbb{E}\, u = c^\mathsf{T}\, \mathbb{E}\, Y = c^\mathsf{T} X\beta$ must equal to $a'\beta$; then

$$X'c = a.$$

$\mathrm{Var}(u) = c^\mathsf{T}\, \mathrm{Var}(y) c = \sigma^2 c^\mathsf{T} c$. The optimal c is the minimizer of the L_2 norm $\|c\|_2$ subject to the n linear constraints: $X^\mathsf{T} c = a$. This is to search the L_2-minimal solution of the linear system $X^\mathsf{T} c = a$. The remaining is left as an exercise. $\square$

Cramer-Rao low bound

Exercise

Find the Fisher information matrix, which is the covariance matrix of the parameter-gradient of the log likelihood function $I(\beta) := \mathbb{V}(\partial_\beta \log p(Y;\beta))$ *and show that the variance matrix of* $\hat{\beta}^{OLS} = (X^\top X)^{-1} X^\top Y$ *is the lower bound* $I^{-1}(\beta)$