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 $\ell(y, \hat{y}) = |y - \hat{y}|^2$
- Many "fancy" machine learning algorithms *today* in literature are still based on this simple least square method.
- 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 $\ddot{z} = c_x \ddot{x} + c_y \ddot{y}$.

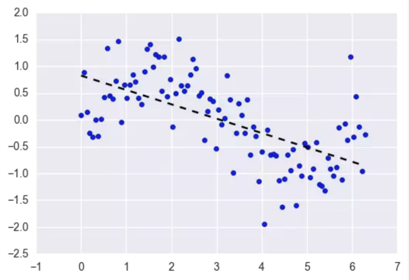
Data $(x_1, y_1), \ldots, (x_n, y_n)$.
The linear regression model assumes a specific linear form for f,

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

which is usually thought of as an approximation to the truth.
The loss is also called residual sum of square (RSS)

$$\mathcal{E}(f) = L(\beta_0, \beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

with the prediction $\hat{y}_i = \beta_0 + \beta x_i$



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 the standard deviation of $\hat{\beta}$ is

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

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**

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

  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:

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

- Model-based interpretation:

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

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 $\sum_i x_{ij} = \sum_i y_i = 0$. the intercept in OLS β_0 vanishes.
   - For centered data: the sample means of the predictor variable x and the response variable y are both zero;
   - The (j,k)-th entry of $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$ is $\frac{1}{n}\sum_{i=1}^n (x_{ij}x_{ik}) \approx \mathrm{cov}(X_j, X_k)$. So $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$ the (sample) variance-covariance matrix of the predictor variable x.

2. Standardization (after centering):

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

   Then $\frac{1}{n}\sum_i x_{ij}^2 \equiv 1, \ \forall j$.
   - For standardized data, the variance of each factor X_j is unit.
   - It follows that $\mathrm{Trace}(\mathbf{X}^\mathsf{T}\mathbf{X}) = \sum_{ij}(x_{ij}^2) = n^2$

Check the linear regression assumption !!!

- The true relationship is linear
- Errors are normally distributed
- Homoscedasticity of errors (or, equal variance around the line).
- Independence of the observations

Read https://towardsdatascience.com/how-do-you-check-the-quality-of-your-regression-model-in-pytho

Theories on OLS

[1]

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

[1] may be skipped for the first reading

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

Assume the measurement error ε follows other distribution, the other type of loss function [1] instead of sum of square errors will arise.

[1] In statistics, it is called "deviance". e.g., the Tweedie deviance

- 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

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

  is the orthogonal projection of the vector Y onto the column space of the matrix $\mathbf{X}$ in $\mathbb{R}^n$

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

- $\hat{Y}$ 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 $\mathbf{X}^\mathsf{T}\mathbf{X}$. 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

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

- influence matrix;
- hat matrix

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

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 $= \dim(\mathsf{X})$.

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

Decomposition of Total Sum of Squares

notations

(x_i, y_i) are the data and $\hat{y}$ are the predicted response. For any regression method, define

- SST= total sum of squares for the response variable (proportional to the variance of the response)

$$SST = \sum_i (y_i - \bar{y})^2$$

- SSReg = sum of squares explained by regression

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

- SSE = sum of squares of errors [1]

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

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

Definition (coefficient of determination)

$$R^2 = 1 - \frac{SSE}{SST}$$

For OLS with the optimal prediction $\hat{Y} = \mathbf{X}\hat{\beta}$, we have [1]

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

For OLS, the coefficient of determination [2] is

$$R^2 = \frac{SSReg}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\text{explained sum of squares by regression}}{\text{total sum of square}}$$

[1] proof: https://en.wikipedia.org/wiki/Explained_sum_of_squares#Partitioning_in_the_general_ordinary_least_squares_model

[2] https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

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

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

Since $Y = \mathbf{X}\beta + \varepsilon$, then

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

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

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

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

Therefore,

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

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

Consistency of $\hat{\beta}$

Assume that

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

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

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

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

- The distribution of $\hat{Y} = X\hat{\beta}$ is then $\mathcal{N}(\mathbf{X}\beta, \sigma^2 X(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})$
- 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

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

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

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

which gives the prediction interval.

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

Gauss-Markov theorem (Rao, 1973)

- Recall that given a training dataset $\mathtt{D}$ for supervised learning, the regression function $\hat{f}_{\mathtt{D}} \in \mathcal{H}$. In OLS, we assumed that $\hat{f}_{\mathtt{D}}(x)$ is a linear function of x.
- Now, if we fix a test input $x = a$, $\hat{f}_{\mathtt{D}}(a)$ 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 $\beta^{\mathsf{T}} a$ at $x = a$?
- Fix the design matrix $\mathbf{X}$, then this estimator takes the linear form in the response of training examples Y:

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

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 $x = a$, and u is in the space of linear transformations from the response training data $Y = \mathbf{X}\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I_n)$. This is to say that $u = c^T Y$ for some vector $c \in \mathbb{R}^n$ satisfying $\mathbb{E}\, u = a^T\beta$ for any β in $\mathbb{R}^{p+1}$. Prove

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

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

Proof.

$\mathbb{E}\, u = c^\mathsf{T}\, \mathbb{E}\, Y = c^\mathsf{T}\mathbf{X}\beta$ must equal $a^\mathsf{T}\beta$ for any β, then

$$\mathbf{X}^\mathsf{T} c = a.$$

To minimize $\mathrm{Var}(u) = c^\mathsf{T}\, \mathbb{V}(Y) c = \sigma^2 \|c\|_2^2$, the optimal c is the L_2-minimal solution of the linear system $\mathbf{X}^\mathsf{T} c = a$ (which is exactly the "pseudo-inverse" of $\mathbf{X}^\mathsf{T}$). The remaining is left as an exercise. □

Cramer-Rao low bound

This exercise is optional. If you know Cramer-Rao bound, it is worth trying.

Exercise

Find the Fisher information matrix I, 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* $\hat{\beta}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$ *is the lower bound* $I^{-1}(\beta)$