1 Linear Regression

1.1 simple linear regression

$$y = \beta_1 + \beta_2 x + \epsilon$$

Assumptions
1. Linear
2. X matrix is fixed.
3. error term has zero mean
4. Homescedasticity or equal variance of ϵ
5. No autocorrelation between disturbances.$cov(\epsilon_i, \epsilon_j) = 0$.
6. Zero covariance between ϵ_i, X_i.
7. Number of obsearvations n must be greater than the number of parameters.
8. $Var(X) > 0$.
9. no multicollinearity.

$$Y = X\boldsymbol{\beta} + \epsilon$$

Where Y is a nx1 matrix, X is anxp matrix, beta is px1 vector and ϵ is nx1 vector with ϵ_i begin iid with normal distribution

1.2 Least Square Regression Model

The cost function is given by

$$f(\boldsymbol{\beta}) = ||Y - X\beta||^2 = (Y - X\beta)^T(Y - X\beta) \quad = Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta$$

Since third term are scalar,

$$\beta^TX^TY = (\beta^TX^TY)^T = Y^TX\beta$$

$$f(\beta) = Y^TY - 2Y^TX\beta - \beta^TX^TX\beta = Y^TY - 2(X^TY)^T\beta + \beta^TX^TX\beta$$

The first term is a constant and its derivative is zero.
The deriviative of 2nd term
Consider the derivative of $\alpha^T\beta$ with respect to β.

$$\boldsymbol{\alpha}^T\boldsymbol{\beta} = \Sigma\alpha_i\beta_i$$
$$\frac{\partial\boldsymbol{\alpha}^T\boldsymbol{\beta}}{\partial\beta_i} = \alpha_i$$

Write the derivative in matrix form

$$\begin{pmatrix} \frac{\partial\boldsymbol{\alpha}^T\boldsymbol{\beta}}{\partial\beta_1} \\ \frac{\partial\boldsymbol{\alpha}^T\boldsymbol{\beta}}{\partial\beta_2} \\ \cdots \\ \frac{\partial\boldsymbol{\alpha}^T\boldsymbol{\beta}}{\partial\beta_3} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_p \end{pmatrix}$$

So if we let $\alpha = X^T Y$, we have

$$\frac{\partial 2(X^T Y)^T \beta}{\partial \beta} = 2X^T Y$$

The derivative of 3rd term
let $A = X^T X$,

$$\beta^T X^T X \beta = \beta^T \begin{pmatrix} \Sigma_i A_{1k}\beta_k \\ \Sigma_i A_{2k}\beta_k \\ ... \\ \Sigma_k A_{pk}\beta_k \end{pmatrix} = \Sigma_j \beta_j (\Sigma_k A_{jk}\beta_k)$$

To calculate the derivative of $f(\beta)$, we note there are only 3 cases that the derivative does not vanish
1) l = j = k

$$\frac{f(\boldsymbol{\beta})}{\partial \beta_l} = 2A_{ll}\beta_l$$

2) l=j, j $\neq$ k

$$\frac{f(\boldsymbol{\beta})}{\partial \beta_l} = \Sigma_{k,k\neq l} A_{lk}\beta_k$$

3) l=k, j $\neq$ k

$$\frac{f(\boldsymbol{\beta})}{\partial \beta_l} = \Sigma_{j,j\neq l} A_{jl}\beta_j = \Sigma_{j,j\neq l} A^T_{lj}\beta j$$

Therefore

$$\begin{aligned}\frac{f(\boldsymbol{\beta})}{\partial \beta_l} &= A_{ll}\beta_l + \Sigma_{k,k\neq l} A_{lk}\beta_k + A_{ll}\beta_l + \Sigma_{j,j\neq l} A^T_{lj}\beta j \\ &= \Sigma_k A_{lk}\beta_k + \Sigma_j A^T_{lj}\beta j\end{aligned}$$

The first term is the lth row of vector $A\beta = X^T X\beta$, and the 2nd term is the lth row of vector$A^T\beta = X^T X\beta$. So we put the whole derivative in matrix form

$$\frac{f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T Y + 2X^T X\beta$$

which is a px1 vector with each row corresponding to the derivative with respect to β_i letting the derivative equal to zero yields the estimation of β

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

1.3 Least Square Estimator for Simple Linear Regression

$$
\begin{aligned}
&\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\
=&(X^TX)^{-1}X^TY \\
=&\left(\begin{pmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_n \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_n \end{pmatrix}\right)^{-1}\begin{pmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_n \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix} \\
=&\frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\begin{pmatrix} \Sigma_i x_i^2 & -\Sigma_i x_i \\ -\Sigma_i x_i & n \end{pmatrix}\begin{pmatrix} \Sigma_i y_i \\ -\Sigma x_i y_i \end{pmatrix}
\end{aligned}
$$

So

$$
\beta_1 = \frac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i(\Sigma x_i y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2}
$$
$$
\beta_2 = \frac{-\Sigma x_i \Sigma y_i + n\Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}
$$

1.4 Projection matrix

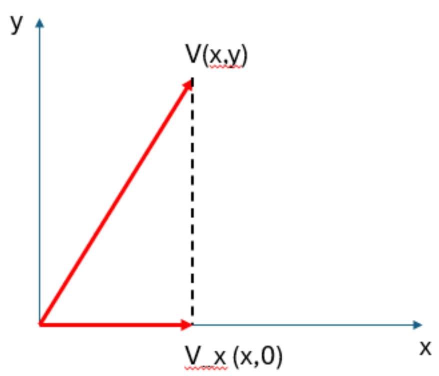
Given $\hat{\beta} = (X^TX)^{-1}X^TY$, we have the predictor value of $y = X\beta$

$$
\hat{y} = X(X^TX)^{-1}X^Ty
$$

The matrix $P = X(X^TX)^{-1}X^T$ is a projection matrix. It projects the vector of y into the column space of X.

Understand the word projection

Let us understand this first through geometry point of view. Consider a vector on 2 dimensional space, $V_1 = (x_1, y_1)^T$, where x_1 and y_1 are the x and y component, respectively. If we project the vector V into x-line, then apparently we get $V_x = (x_1, 0)^T$, see graph below.



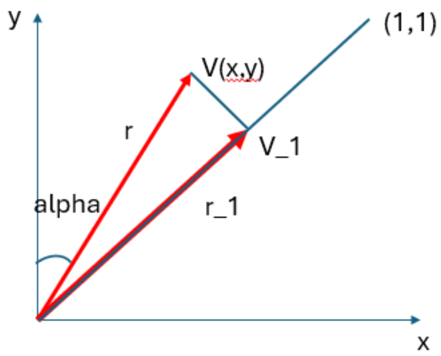
If we have a vector that is along the x axis

$$X = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

The projection matrix of a vector into x line is

$$\begin{aligned} P_x &= x(x^T x)^{-1} x^T \\ &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Applying this projection matrix to any 2 dimensional vector V gives $(V_x, 0)^T$. So it projects the vector into x line. Let us take another example. Imagine V_1 is vector if we project V onto the line that has 45 degree angle with x axis. See below.



In order to calculate V_1, we see

$$r_1 = rcos(\pi/4 - alpha) = r(\frac{\sqrt{2}}{2}\frac{y}{r} + \frac{\sqrt{2}}{2}\frac{x}{r}) = \frac{\sqrt{2}}{2}y + \frac{\sqrt{2}}{2}x$$

$$\begin{aligned} V_{1x} &= r_1 cos(\pi/4) = \frac{x+y}{2} \\ V_{1y} &= r_1 sin(\pi/4) = \frac{x+y}{2} \end{aligned}$$

After we understand this using geometry point of view, we can workout from algebra point of view. The vector we want to project onto is

$$i = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The projection matrix of a vector into x line is

$$\begin{aligned}
P_x &= x(x^T x)^{-1} x^T \\
&= \begin{pmatrix} 1 \\ 1 \end{pmatrix} \left(\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 \end{pmatrix} \\
&= \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}
\end{aligned}$$

Therefore we easily see

$$V_1 = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(x+y) \\ \frac{1}{2}(x+y) \end{pmatrix}$$

which is the same as what we get based on geometry. For n dimensional vector y, if our X matrix has rank of k, then the projection matrix P projects the vector y into k dimensional hyperplane. For example, if we define

$$i_N = \begin{pmatrix} 1 \\ 1 \\ ... \\ 1 \end{pmatrix}$$

The projection matrix P is

$$P = i \frac{1}{N} i^T = \frac{1}{N} \begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 1 \end{pmatrix}$$

Projection matrix into null space
If P is a projection matrix, the matrix $I - P$ is also a projection matrix. In linear regression model

$$\begin{aligned}
y &= X\beta + \epsilon \\
P &= X(X^T X)^{-1} X^T \\
\hat{\epsilon} &= (I - P)y = (I - X(X^T X)^{-1} X^T)y
\end{aligned}$$

For the above example, we define $M = I - \frac{1}{N} i i^T$, and My express the mean deviations of a vector.
Idempotent property of projection matrix
Consider the previous example that we project a vector V onto x axis, how about we do this projection twice, we would end up the same vector V_x. Using a little matrix algebra, it is easy to prove that for any project matrix P, we have $PP = P$.

1.5 Variance of $\hat{\beta}$

$$\begin{aligned}
Var(\hat{\beta}) = Var((X^T X)^{-1} X^T \epsilon) &= (X^T X)^{-1} X^T Var(\epsilon)((X^T X)^{-1} X^T)^T \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}
\end{aligned}$$

For simple linear regression

$$Var(\hat{\beta}) = \frac{\sigma^2}{n\Sigma x_i^2 - (\Sigma x_i)^2} \begin{pmatrix} \Sigma_i x_i^2 & -\Sigma_i x_i \\ -\Sigma_i x_i & n \end{pmatrix}$$

$$Var(\hat{\beta}_1) = \frac{\Sigma x_i^2 \sigma^2}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$Var(\hat{\beta}_2) = \frac{n\sigma^2}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

Try

$$\begin{aligned} \Sigma(x_i - \bar{x})^2 = \Sigma(x_i^2 - 2\bar{x}x_i + \bar{x}^2) &= \Sigma_i(x_i^2 - 2(\Sigma_j \frac{x_j}{n})x_i + \frac{(\Sigma_j x_j)^2}{n^2}) \\ &= \Sigma_i x_i^2 - \frac{2}{n}(\Sigma_i x_i)^2 + \frac{(\Sigma_i x_i)^2}{n} = \Sigma_i x_i^2 - \frac{1}{n}(\Sigma x_i)^2 \end{aligned}$$

So

$$Var(\hat{\beta}_1) = \frac{\Sigma x_i^2 \sigma^2}{n\Sigma(x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_2) = \frac{n\sigma^2}{n\Sigma(x_i - \bar{x})^2} = \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}$$

1.6 Sum of Square Error

$$\begin{aligned} SSE &= \Sigma_i(y - \hat{y}_i)^2 \\ &= (Y - X\beta)^T(Y - X\beta) \\ &= (Y - X(X^TX)^{-1}X^TY)^T(Y - X(XTX) - 1X^TY) \\ &= (Y - PY)^T(Y - PY) \\ &= Y^T(1-P)^T(1-P)Y = Y^T(1-P)Y \\ &= (X\beta + \epsilon)^T(1-P)(X\beta + \epsilon) \\ &= \beta^TX^T(1-P)X\beta + 2\beta^TX^TX^T(I-P)\epsilon + \epsilon^T(I-H)\epsilon \end{aligned}$$

$$E[SSE] = E[\epsilon^T(I-P)\epsilon] = E[\epsilon^T\epsilon]trace(I-H) = \sigma^2(n-p)$$