1 Linear Regression

1.1 Linear regression and Least Square Solution

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

Where Y is a n × 1 matrix, X is a n × k matrix, beta is k × 1 vector and ϵ is nx1 vector with ϵ_i begin iid with normal distribution.

Assumptions

1. Linear
2. X matrix has full rank. In other words, no multicollinearity.
2. error term has zero mean $E[\epsilon|X] = 0$
3. Homescedasticity or equal variance of ϵ. In other words, no autocorrelation between disturbances.$cov(\epsilon_i, \epsilon_j) = 0$.
6. Number of obsearvations n must be greater than the number of parameters.

Least Square Solution

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

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

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

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

The derivative of 3rd term

let $A = X^TX$,

$$\beta^TX^TX\beta = \beta^T\begin{pmatrix} \Sigma_iA_{1k}\beta_k \\ \Sigma_iA_{2k}\beta_k \\ ... \\ \Sigma_kA_{pk}\beta_k \end{pmatrix} = \Sigma_j\beta_j(\Sigma_kA_{jk}\beta_k)$$

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

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

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 **normal equation** and the estimation of β
Normal equation

$$(X^TX)\hat{\beta} = X^TY$$

Estimator of β

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

Using the rule of matrix mutiplication, we can rewrite X^TX as the f

$$\begin{aligned}X^TX &= \begin{pmatrix} \sum_{i=1}^N x^T_{1i}x_{i1} & \sum_{i=1}^N x^T_{1i}x_{i2} & ... & \sum_{i=1}^N x^T_{1i}x_{ik} \\ ... & ... & ... & \\ \sum_{i=1}^N x^T_{ki}x_{i1} & \sum_{i=1}^N x^T_{ki}x_{k2} & ... & \sum_{i=1}^N x^T_{ki}x_{ik} \end{pmatrix} \\ &= \sum_{i=1}^N \begin{pmatrix} x_{i1}x_{i1} & x_{i1}x_{i2} & ... & x_{i1}x_{ik} \\ ... & ... & ... & \\ x_{ik}x_{i1} & x_{ik}x_{k2} & ... & x_{ik}x_{ik} \end{pmatrix} \\ &= \sum_{i=1}^N \begin{pmatrix} x_{i1} \\ ... \\ x_{ik} \end{pmatrix} \begin{pmatrix} x_{i1} & ... & ...x_{ik} \end{pmatrix} \\ &= \sum_{i=1}^N X_iX_i^T\end{aligned}$$

where X_i is $K \times 1$ vector, and $X_i X_i^T$ is $K \times K$ matrix. Similarly,

$$X^T Y = \sum_{i=1}^{N} X_i y_i$$

where X_i is $K \times 1$ vector, y_i is a scalor, and $X_i y_i$ is $K \times 1$ vector. So the estimator of β is

$$\hat{\beta} = (\sum_{i=1}^{N} X_i X_i^T)^{-1}(\sum_{i=1}^{N} X_i y_i)$$

Least Square Estimator for Simple Linear Regression

$$y = \beta_0 + \beta_1 X + \epsilon$$

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

So

$$\beta_0 = \frac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i (\Sigma x_i y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2} \tag{1}$$

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

β_1 can also be written using the covariance

$$\beta_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})(x_i - \bar{x})} \tag{3}$$

And it is easy to show

$$\begin{aligned}
\beta_2 &= \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})(x_i - \bar{x})} \\
&= \frac{\sum_i^n (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\bar{y})}{\sum_i^n (x_i^2 - 2\bar{x} x_i + (\bar{x})^2)} \\
&= \frac{\sum_i^n x_i y_i - \sum_i^n \bar{x} y_i - \sum_i^n x_i \bar{y} + \sum_i^n \bar{x}\bar{y}}{\sum_i^n x_i^2 - \sum_i^n 2\bar{x} x_i + \sum_i^n (\bar{x})^2} \\
&= \frac{\sum_i^n x_i y_i - (\frac{1}{n}\sum_j^n x_j)(\sum_i^n y_i) - (\sum_i^n x_i)(\frac{1}{n}\sum_j^n y_j) + \sum_i^n (\frac{1}{n}\sum_j^n x_i)(\frac{1}{n}\sum_k^n y_i)}{\sum_i^n x_i^2 - \sum_i^n 2(\frac{1}{n}\sum_j^n x_j) x_i + \sum_i^n (\frac{1}{n}\sum_j^n x_j)^2} \\
&= \frac{\sum_i^n x_i y_i - (\frac{1}{n}\sum_j^n x_j)(\sum_i^n y_i) - (\sum_i^n x_i)(\frac{1}{n}\sum_j^n y_j) + n(\frac{1}{n}\sum_j^n x_i)(\frac{1}{n}\sum_k^n y_k)}{\sum_i^n x_i^2 - \sum_i^n 2(\frac{1}{n}\sum_j^n x_j) x_i + n(\frac{1}{n}\sum_j^n x_j)^2} \\
&= \frac{\sum_i^n x_i y_i - \frac{1}{n}(\sum_i^n x_i)(\sum_j^n y_j) - \frac{1}{n}(\sum_i^n x_i)(\sum_j^n y_j) + \frac{1}{n}(\sum_j^n x_i)(\sum_k^n y_k)}{\sum_i^n x_i^2 - \frac{2}{n}(\sum_j^n x_j)(\sum_i^n x_i) + \frac{1}{n}(\sum_j^n x_j)^2} \\
&= \frac{\sum_i^n x_i y_i - \frac{1}{n}(\sum_i^n x_i)(\sum_j^n y_j)}{\sum_i^n x_i^2 - \frac{1}{n}(\sum_j^n x_j)(\sum_i^n x_i)} \\
&= \frac{n\sum_i^n x_i y_i - (\sum_i^n x_i)(\sum_j^n y_j)}{n\sum_i^n x_i^2 - (\sum_j^n x_j)(\sum_i^n x_i)} \\
&= \frac{n\sum x_i y_i - (\sum x_i)(\sum_j^n y_j)}{n\sum x_i^2 - (\sum x_i)^2}
\end{aligned}$$

which is the same as Eq.2. We can interpret β as ratio of the covariance of x and y to the variance of x.

1.2 Projection matrix

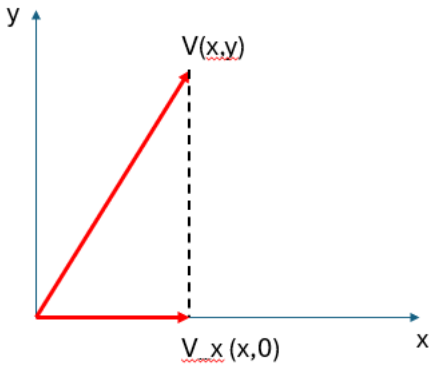
Given $\hat{\beta} = (X^T X)^{-1} X^T Y$, we have the predictor value of $y = X\beta$

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

The matrix $P = X(X^T X)^{-1} X^T$ 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.



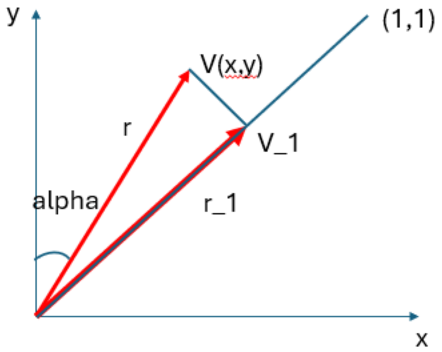
If we have a vector that is along the x axis

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

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

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

Define residual vector $\hat{\epsilon}$

$$\hat{\epsilon} = (I - P)y = (I - X(X^T X)^{-1} X^T)y$$

And it is easy to show $\hat{e}$ and X are orthogonal.

$$X^T \hat{\epsilon} = X^T(I - P)y = X^T(I - X(X^T X)^{-1} X^T)y = (X^T - X^T X(X^T X)^{-1} X^T)y = 0y = 0$$

For the above example, we define $M = I - \frac{1}{N} ii^T$, 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 $PP = P$.

1.3 Partitioned Regression and Regression

$$y = X\boldsymbol{\beta} + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

The normal equation is

$$\begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1^T y \\ X_2^T y \end{pmatrix}$$

If X1 and X2 are orthogonal, namely, $X_1^T X_2 = 0$, then

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathbf{1}} &= (X_1^T X_1)^{-1} X_1^T y \\ \hat{\boldsymbol{\beta}}_{\mathbf{2}} &= (X_2^T X_2)^{-1} X_2^T y \end{aligned}$$

If X1 and X2 are not orthogonal, we can solve for β_2 in the above normal equation set and get β_2

$$\begin{aligned} \hat{\beta}_2 &= [X_2^T(I - X_1(X_1^T X_1)^{-1} X_1^T)X_2]^{-1}[X_2(I - X_1(X_1^T X_1)^{-1} X_1^T)y] \\ &= (X_2^T M_1 X_2)^{-1}(X_2^T M_1 y) \end{aligned}$$

Given the fact that M_1 is symmetrical and idempotent, we can rewrite the above expression

$$\begin{aligned} \hat{\beta}_2 &= (X_2^T M_1 M_1 X_2)^{-1}(X_2^T M_1 M_1 y) \\ &= (X_2^T M_1^T M_1 X_2)^{-1}(X_2^T M_1^T M_1 y) \\ &= ((M_1 X_2)^T M_1 X_2)^{-1}((M_1 X_2)^T M_1 y) \qquad (4) \end{aligned}$$

The above uses the property that $M_1^T = M_1$ and $M_1 M_1 = M_1$
The $\hat{\beta}_2$ is also the solution of

$$M_1 Y = M_1 X_2 \beta_2 + \epsilon$$

where $M_1 y$ is the residual of y regressed on X_1 and $M_1 X_2$ is the residual of X_2 regressed on X_1. For example, in simple linear regression

$$Y = \beta_0 + x\beta_1$$

Where $X_1 = 1_N$, so its projection matrix is $i\frac{1}{N}i^T$, and the corresponding M matrix is $I - \frac{1}{N}ii^T$. We tries to calculate β using partition regression.

$$MY = (I - \frac{1}{N}ii^T)Y = Y - \bar{Y}MX = (I - \frac{1}{N}ii^T)Y = X - \bar{X}$$

Then

$$\beta_1 = ((MX)^T(MX))^{-1}((MX)^T(MY)) = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^N (x_i - \bar{x})^2} \tag{5}$$

which is the same as Eq.3

1.4 Variance componet identity

If we define our mean projection matrix P

$$P = i\frac{1}{N}i^T$$

and silimarly we define mean deviation project matrix

$$M = I - P = i\frac{1}{N}i^T$$

We have

$$y = \hat{y} + \hat{\epsilon} = X\hat{\beta} + \hat{\epsilon}$$

Multiplying M matrix on the left, we have

$$My = MX\hat{\beta} + M\hat{\epsilon} = MX\hat{\beta} + \hat{\epsilon}$$

$$\begin{aligned}(My)^2 &= (MX\hat{\beta} + \hat{\epsilon})^T(MX\hat{\beta} + \hat{\epsilon}) \\ &= (\beta^T X^T M^T + \hat{\epsilon}^T)(MX\hat{\beta} + \hat{\epsilon}) \\ &= (MX\hat{\beta})^2 + \beta^T X^T M^T \hat{\epsilon} + (\beta^T X^T M^T \hat{\epsilon})^T + (\hat{\epsilon})^2\end{aligned}$$

The 2nd and 3rd terms are zero because that 1)$\hat{\epsilon}$ has zero mean, so $M^T\hat{\epsilon} = M\hat{\epsilon} = \hat{\epsilon}$ and 2) $X^T\hat{\epsilon} = 0$, so

$$(My)^2 = (MX\hat{\beta})^2 + (\hat{\epsilon})^2$$

Rewriting the above equation using summation, we have

$$\sum_i (y_i - \bar{y})^2 = \sum_i^N (\bar{y}_i - \bar{\hat{y}})^2 + \sum_i (y_i - \hat{y})^2$$

Define

$$\begin{aligned} SST &= \sum_i (y_i - \bar{y})^2 \\ SSR &= \sum_i (\bar{y}_i - \bar{\bar{y}})^2 \\ SSE &= \sum_i (y_i - \hat{y})^2 \end{aligned}$$

SSE can also be written as

$$\begin{aligned} SSE &= SST - SSR \\ &= SST - (MX\hat{\beta})^2 \\ &= SST - ((MX)((MX)^T(MX))^{-1}(MX)^T(MY))^2 \end{aligned}$$

Let $U = MX$, and$V = MY$ then

$$\begin{aligned} SSE &= SST - (Z(Z^TZ)^{-1}Z^T)^2 \\ &= SST - (U(U^TU)^{-1}U^TV)^T(U(U^TU)^{-1}U^TV) \\ &= SST - V^TU(U^TU)^{-1}U^TU(U^TU)^{-1}U^TV \\ &= SST - V^TU(U^TU)^{-1}U^TV \\ &= SST - (MY)^T(MX)((MX)^T(MX))^{-1}(MX)^T(MY) \end{aligned}$$

Define

$$\begin{aligned} S_{xx} &= (MX)^T(MX) = \sum_i^N (x_i - \bar{x})^2 \\ S_{xy} &= (MX)^T(MY) = \sum_i^N (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

So

$$SSE = SST - S_{xy}^T S_{xx}^{-1} S_{xy}$$

Then we have

$$SST = SSR + SSE$$

1.5 Variance of $\hat{\beta}$ and σ^2 estimation

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

The above derivation use the fact that ϵ has a normal distribution with mean 0 and variance σ^2. For simple linear regression

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

$$Var(\hat{\beta}_0) = \frac{\Sigma x_i^2 \sigma^2}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

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

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

$$Var(\hat{\beta}_0) = \frac{\Sigma x_i^2 \sigma^2}{n\Sigma(x_i - \bar{x})^2}$$

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

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

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

We obtain the unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{SSE}{n-k}$$

Therefore the estimator of variance of β

$$\hat{V}ar(\hat{\beta}_i) = \hat{\sigma}^2(X^TX)^{-1}_{ii}$$

and the stanard error of β_i is

$$SE(\hat{\beta}_i) = \sqrt{\hat{\sigma}^2(X^TX)^{-1}_{ii}}$$

2 Properties of Least Square Estimators

When we have a estimator, we need to evaluate how good our estimator is? A few questions we can ask is 1): how far is the value of our estimator away from the true value, even in the ideal case when the sample size is inifinite? 2) when 1) is true, with finite sample size, does the value of our estimator approach to the true value as the sample size increase? In other words, does the estimator converge to the true value as sample size goes to inifinity? 3) when 1) and 2) is true, as the sample size increase, how fast does our estimator converges to true value? 4) with 1) 2) and 3), what is the asymptotic distribution of the estimator? If the distribution is normal, it can be used to do interval estimation such as confidence interval. The 1st question defines unbiasness, the 2nd one defines consistency, and the 3rd one defines efficiency.

2.1 Unbiasness

Unbiased

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

Then the expectation of $\hat{\beta}$ condition on X is

$$E[\hat{\beta}|X] = \beta + (X^TX)^{-1}X^TE(\epsilon|X)$$
$$t$$

The last term is zero by assuption of linear regression. So

$$E[\hat{\beta}] = \beta$$

The expectation of the estimator is the same as true value, this is called **unbiased**.

Bias due to omission of relevant variables

Suppose we have a model

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

If we regression y on X_1 only, our estimator is

$$\hat{\beta}_1 = (X_1^TX_1)^{-1}X_1^Ty = \beta_1 + (X_1^TX_1)^{-1}X_1^TX_2\beta_2 + (X_1^TX_1)^{-1}X_1^T\epsilon$$

On the second term, we see unless 1)X_1 and X_2 are orthogonal, or 2)β_2 =0, β_1 is biased.

2.2 Consistency

The unbiasness gives us a metric of measuring how good our esitimator is, from population perspetive. In reality, as our sample size is finite, we need ask ourselves does our estimator converges to true value when sample size is sufficiently large. We know

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

$$\begin{aligned}\hat{\beta} &= \beta + (X^T X)^{-1} X^T \epsilon \\ &= \beta + (\Sigma_{i=1}^N X_i X_i^T)^{-1} X^T \epsilon \\ &= \beta + (\Sigma_{i=1}^N \frac{1}{N} X_i X_i^T)^{-1} (\frac{X^T \epsilon}{N})\end{aligned}$$

To show $\hat{\beta}$ converges to β, we need to show two things:
(1)$\frac{1}{n}\sum_i X_i X_i^T$ converges to Q in probability when N is large. Also the inverse of Q exists.
(2)$\frac{X^T \epsilon}{N}$ converges to zero in probability when N is large.
For (1), we write

$$Q^{(n)} = \frac{1}{n}\sum_i X_i X_i^T$$

which is a $K \times K$ matrix. Its element Q_{kl} is

$$Q_{kl}^{(n)} = \frac{1}{n}\sum_i x_{ik} x_{il}$$

Based on Law of large numbers, Q_{kl}^n converges to its expectation value $E[x_{ik}x_{il}]$. Let

$$E[x_{ik}x_{il}] = Q_{kl}$$

So each element of $Q^{(n)}$ converges Q_{kl}. Therefore, $Q^{(n)}$ converges to Q. To show the inverse of Q exists, we can show Q is a symmetric positive definite matrix. For any k dimenional vector v, we have

$$v^T Q v = E[v^T x_i x_i^T v]$$

Since $v^T x_i = \sum_k v_k x_{ik} = x_i^T v$ which is a scalar, so

$$v^T Q v = E[v^T x_i x_i^T v] = E[(v^T x_i)^2]$$

if for any i, x_{ik} are linear independent, in other words, no multicollinearity. Then $v^T x_i = \sum_k v_k x_{ik} \neq 0$ and $(v^T x_i)^2 > 0$. Therefore Q is SPD matrix and its inverse exists.

$$\begin{aligned}&\frac{X^T\epsilon}{N}\\ =&\begin{pmatrix}\frac{1}{N}\Sigma_1^N x_{i1}\epsilon_i\\ \frac{1}{N}\Sigma_1^N x_{i2}\epsilon_i\\ \frac{1}{N}\Sigma_1^N x_{i3}\epsilon_i\\ \dots\\ \frac{1}{N}\Sigma_1^N x_{ik}\epsilon_i\end{pmatrix}\\ =&\frac{1}{N}\Sigma_{i=1}^N X_i\epsilon_i = \bar{w}\end{aligned}$$

Where $\bar{w}$ is a k × 1 vector. To see the asymptotical behavior of w, we consider its mean and asymptotical varance. The mean is

$$E[w_i] = E_X[E[w_i|x_i]] = E_X[X_i E[\epsilon|X_i]] = 0$$

$$Var[\bar{w}] = E[Var[\bar{w}|X]] + Var[E[\bar{w}|X]] = E[Var[\bar{w}|X]] + 0 = E[Var[\bar{w}|X]]$$

$$Var[\bar{w}|X] = E[\bar{w}\bar{w}^T|X] = \frac{1}{n}X^T E[\epsilon\epsilon^T]X\frac{1}{n} = \frac{\sigma^2}{n}\frac{X^TX}{n}$$

$$E[Var[\bar{w}|X]] = \frac{\sigma^2}{n}E(\frac{X^TX}{n})$$

We have shown that $X^TX = \sum_i x_i x_i^T$, so

$$\frac{X^TX}{n} = \frac{1}{n}\sum_i x_i x_i^T$$

When $\frac{X^TX}{n}$ converges to Q,

$$E[Var[\bar{w}|X]] = 0$$

So $\bar{w}$ converges to $\mathbf{0}(k\times 1)$ vector. Then when N is sufficiently large, $\hat{\beta}$ converges to β. This is the proof of consistency.
There are certain conditions in which the estimators become inconsitent.
1) X is not full rank, or X has multicollinearity 2) $cov[X,\epsilon] \neq 0$

2.3 Efficiency

The least equare estimator has the smallest varaince, and this can be proved by Gauss-Markov theorem.

2.4 Multicollinearity

Suppose we have a regression model that contains two parameters

$$y = \beta_0 + X_1\beta_1 + X_2\beta_2$$

From above, we know variance of $\hat{\beta}$ is

$$Var(\hat{\beta}) = \frac{\sigma^2}{(X^T X)^{-1}}$$

When X only contains 2 variables, $X = (X_1, X_2)$

$$\begin{aligned}
Var(\hat{\beta}_1) &= \sigma^2 \frac{S_{22}}{S_{11}S_{22} - S_{12}^2} = \frac{1}{S_{11}(1 - \frac{S_{12}^2}{S_{11}S_{22}})} = \frac{1}{S_{11}(1 - r_{12}^2)} \\
Var(\hat{\beta}_2) &= \sigma^2 \frac{S_{11}}{S_{11}S_{22} - S_{12}^2} = \frac{1}{S_{22}(1 - \frac{S_{12}^2}{S_{11}S_{22}})} = \frac{1}{S_{22}(1 - r_{12}^2)}
\end{aligned}$$

Where

$$\begin{aligned}
S_{11} &= \Sigma(x_{1i} - \hat{x}_1)^2 \\
S_{22} &= \Sigma(x_{2i} - \hat{x}_2)^2 \\
S_{12} &= \Sigma(x_{1i} - \hat{x}_1)(x_{2i} - \hat{x}_2) \\
r_{12} &= \frac{S_{12}}{\sqrt{S_{11}S_{22}}}
\end{aligned}$$

r_{12} is the correlation coefficient. In extreme case, when X_1 and X_2 are perfectly correlated, the variance becomes infinite.

3 Model Testing

Lagrange Multiplier(LM) test
Suppose we have two models, one is restriced, the other is unrestricted: Restricted(R): $y = X_1\beta_1 + \epsilon$
Unrestricted (U): $y = X_1\beta_1 + X_2\beta_2 + \epsilon$
Given the unrestricted model, the likehood function is

$$L(\beta_1, \beta_2, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}(y - X_1\beta_1 - X_2\beta_2)^T(y - X_1\beta_1 - X_2\beta_2))$$

$$S_2 = \frac{\partial L}{\partial \beta_2} = \frac{1}{\sigma^2} X_2^T (y - X_1\beta_1 - X_2\beta_2)$$

When $\beta_2 = 0$, define

$$M_1 = I - X_1(X_1^T X_1)^{-1} X_1^T$$

and $M_1 X_1 = 0$.

$$S_2 = \frac{1}{\sigma^2} X_2^T (y - X\hat{\beta}_1) = \frac{1}{\sigma^2} X_2^T M_1 y = \frac{1}{\sigma^2} X_2^T M_1 (X_1\beta_1 + \epsilon) = \frac{1}{\sigma^2} X_2^T M_1 \epsilon$$

The last equal sign uses the fact $M_1 X_1 = 0$.

$$\begin{aligned} Var(X_2^T M_1 \epsilon) &= Var(X_2^T M_1 \epsilon) \\ &= X_2^T M_1 Var(\epsilon)(X_2^T M_1)^T \\ &= X_2^T M_1 M_1^T X_2 Var(\epsilon) \\ &= \sigma^2 X_2^T M_1 X_2 \end{aligned}$$

Define

$$V = X_2^T M_1 X_2$$

Then

$$Var(X_2^T M_1 \epsilon) = \sigma^2 V$$

So $X_2^T M_1 \epsilon$ follows normal distribution with mean 0 and variance $\sigma^2 X_2^T M_1 X_2$. Define

$$Z = \frac{X_2^T M_1 \epsilon}{\sqrt{\sigma^2 X_2^T M_1 X_2}} = \frac{S_2}{\sqrt{\sigma^2 V}}$$

then Z follows standard normal distribution. The **Lagrange Multiplier (LM) test** is defined

$$LM = Z^2 = \frac{(X_2^T M_1 \epsilon)^2}{\sigma^2 X_2^T M_1 X_2} = \frac{S_2^2}{\sigma^2 V}$$

which follows χ^2 distribution with degree of freedom 1.

F test

We define

$$\begin{aligned} SSE_U &= ||y - X_1 \hat{\beta}_1 - X_2 \hat{\beta}_2||^2 \\ SSE_R &= ||y - X_1 \hat{\beta}_1||^2 \end{aligned}$$

F test is defined as

$$F = \frac{\frac{Extra\ expalined\ variation}{Degree\ of\ Freedom}}{\frac{Remaining\ unexplained\ variation}{Degree\ of\ Freedom}} = \frac{SSE_R - SSE_U}{\frac{SSE_U}{n-1}}$$

Let $X = (X_1, X_2)$, and we define two projection matrices

$$\begin{aligned} P_U &= X(X^T X)^{-1} X^T \\ P_R &= X_1 (X_1^T X_1)^{-1} X_1^T \end{aligned}$$

$$SSR_R - SSR_U = y^T P_R y - y^T P_U y = y^T (P_R - P_U) y$$

recall

$$\hat{\beta}_2 = (X_2^T M_1 X_2)^{-1} X_2^T M_1 y$$

The corresponding projection matrix is

$$M_1X_2(X_2^TM_1X_2)^{-1}X_2^TM_1$$

The SSR_R - SSR_U is the additional variance explained by X_2 after removing the linear space of X_1 on X_2. This means the projection matrix corresponding to β_2 is P_U - P_R. So we can get P_U - P_R using the interpretation of projection matrix instead solving for the projection matrix itself.

$$P_U - P_R = M_1X_2(X_2^TM_1X_2)^{-1}X_2^TM_1$$

The extra expained sum of squares by the unrestricted model is

$$SSR_R - SSR_U = y^T(P_R - P_U)y = (X_2^TM_1y)^T(X_2^TM_1X_2)^{-1}X_2^TM_1y$$

with 1 degree of freedom as X_2 only contains 1 parameter.

$$F = \frac{SSR_R - SSR_U}{\frac{SSR_U}{n-k}} = \frac{(X_2M_1y)^T(X_2^TM_1y)}{\hat{\sigma}^2(X_2^TM_1X_2)} = LM$$

We see that F test and LM test are equivalent.

Wald Test

Recall the estimator for β_2 in Eq.4,

$$\hat{\beta}_2 = (X_2^TM_1X_2)^{-1}(X_2^TM_1y)$$

Substitue

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

we get

$$\hat{\beta}_2 = (X_2^TM_1X_2)^{-1}(X_2^TM_1\epsilon)$$

Since ϵ $N(0, \sigma^2I)$, we obtain

$$\beta_2 \sim N(0, \sigma^2(X_2^TM_1X_2)^{-1})$$

Thus, scaling by $1/\sigma^2$, we arrive at

$$\frac{1}{\sigma^2}X_2^TM_1\epsilon \sim N(0, X_2^TM_1X_2)$$

Construct Wald test W

$$W = \frac{\hat{\beta}_2}{\sqrt{\hat{V}ar(\hat{\beta}_2)}}$$

W follows t distribution. We now show W test is equivalent to LM test. Consider W^2

$$\begin{aligned}W^2 &= \hat{\beta}^T(Var(\hat{\beta}))^{-1}\hat{\beta} = \frac{1}{\sigma^2}\hat{\beta}^TV\hat{\beta} = \frac{1}{\sigma^2}(V^{-1}S_2)^TVV^{-1}S_2 = \frac{1}{\sigma^2}S_2^TV^{-1}VV^{-1}S_2\\ &= \frac{1}{\sigma^2}S_2^TV^{-1}S_2\\ &= LM\end{aligned}$$

4 Panel Data Model

We can view panel data as a "two dimensional" data set in which the sample does not only come from different individuals, but also same individual across different time point. We can write the regression model as

$$y_{it} = \alpha_{it} + \sum_k x_{itk}\beta_{itk} + u_{it}$$

where $1 < i < N$, $1 < t < T$, and $1 < k < K$. The equation has total sample size of NT with total number of parameter $NT(K+1)$, therefore it is not estimable. So we will make the following few assumptions

	$\alpha_{it} = \alpha_{is}$	$\alpha_{it} = \alpha_{jt}$	$\beta_{it} = \beta_{is}$	$\beta_{itk} = \beta_{jtk}$
Pooled	yes	yes	yes	yes
Fixed Effect	yes	no	yes	yes
Unrestricted	yes	no	yes	no

4.1 The unrestriced model

$$y_{it} = \alpha_i + \sum_k x_{itk}\beta_{ik} + u_{it}$$

The above equation can be written in matrix form:
for $i = 1$,

$$\begin{pmatrix} y_{11} \\ y_{12} \\ ... \\ y_{1T} \end{pmatrix} = \begin{pmatrix} 1 & x_{111} & x_{112} & ... & x_{11K} \\ 1 & x_{121} & x_{122} & ... & x_{12K} \\ 1 & ... & ... & ... & ... \\ 1 & x_{1T1} & x_{1T2} & ... & x_{1TK} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_{11} \\ \beta_{12} \\ ... \\ \beta_{1K} \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ ... \\ u_{1T} \end{pmatrix}$$

which we can also write as for $i = 2$,

$$\begin{pmatrix} y_{21} \\ y_{22} \\ ... \\ y_{2T} \end{pmatrix} = \begin{pmatrix} 1 & x_{211} & x_{212} & ... & x_{21K} \\ 1 & x_{221} & x_{222} & ... & x_{22K} \\ 1 & ... & ... & ... & ... \\ 1 & x_{2T1} & x_{2T2} & ... & x_{2TK} \end{pmatrix} \begin{pmatrix} \alpha_2 \\ \beta_{21} \\ \beta_{22} \\ ... \\ \beta_{2K} \end{pmatrix} + \begin{pmatrix} u_{21} \\ u_{22} \\ ... \\ u_{2T} \end{pmatrix}$$

So for each i, we can write

$$Y_i = 1_T\alpha_i + X_i\beta_i + U_i$$

where $Y_i = (y_{i1}, y_{i2}, ..., y_{iT})^T$, 1_T is a one vector of length T, X_i is $K \times T$ matrix, $\beta_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{iK})^T$, and $U_i = (u_{i1}, u_{i2}, ..., u_{iT})^T$.

If we consolidate equation set for all the value of i

$$\begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_N \end{pmatrix} = \begin{pmatrix} 1_T & 0 & 0 & ... & 0 \\ 0 & 1_T & 0 & ... & 0 \\ 0 & 0 & 1_T & ... & 0 \\ 0 & 0 & 0 & ... & 1_{NT} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_N \end{pmatrix} + \begin{pmatrix} X_1 & 0 & 0 & ... & 0 \\ 0 & X_2 & 0 & ... & 0 \\ 0 & 0 & X_3 & ... & 0 \\ 0 & 0 & 0 & ... & X_N \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_N \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ ... \\ U_N \end{pmatrix}$$

To solve for β_i, we can use the strategy of partition regression. β_i is the solution of the the regression

$$MY_i = MX_i\beta + U$$

where

$$\begin{aligned} M &= I - \frac{1}{T}1_T1_T^{'} \\ MY_i &= y_{it} - \bar{y}_{i.} \\ MX_i &= x_{it} - \bar{x}_{i.} \end{aligned}$$

Here to avoid duplicate notation, we denote the transposed matrix using $'$. The estimate of β is

$$\hat{\beta}_i = ((MX_i)^T(MX_i))^{-1}((MX_i)^T(MY_i)) = W_{xx,i}^{-1}W_{xy,i}$$

where

$$\begin{aligned} W_{xy,i} &= \sum_t^T (x_{it} - \bar{x}_{i.})(y_{it} - \bar{y}_{i.}) \\ W_{xx,i} &= \sum_t^T (x_{it} - \bar{x}_{i.})(x_{it} - \bar{x}_{i.})^T \end{aligned}$$

4.2 The pooled model

$$y_{it} = \alpha + \sum_k x_{itk}\beta_k + u_{it}$$

The above equation can be written in matrix form:
for $i = 1$,

$$\begin{pmatrix} y_{11} \\ y_{12} \\ ... \\ y_{1T} \\ y_{21} \\ y_{22} \\ ... \\ y_{2T} \\ ... \\ y_{NT} \end{pmatrix} = \begin{pmatrix} 1 & x_{111} & x_{112} & ... & x_{11K} \\ 1 & x_{121} & x_{122} & ... & x_{12K} \\ 1 & ... & ... & ... & ... \\ 1 & x_{1T1} & x_{1T2} & ... & x_{1TK} \\ 1 & x_{211} & x_{212} & ... & x_{21K} \\ 1 & x_{221} & x_{222} & ... & x_{22K} \\ 1 & ... & ... & ... & ... \\ 1 & x_{2T1} & x_{2T2} & ... & x_{2TK} \\ 1 & ... & ... & ... & ... \\ 1 & x_{NT1} & x_{NT2} & ... & x_{NTK} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_K \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ ... \\ u_{1T} \\ u_{21} \\ u_{22} \\ ... \\ u_{2T} \\ ... \\ u_{NT} \end{pmatrix}$$

Similarly, using the solution of β from Eq.5, the estimated β can be written as

$$M = I - \frac{1}{NT}1_{NT}1_{NT}^{'}$$

$$
\begin{aligned}
\hat{\beta} &= ((MX)^T(MX))^{-1}((MX)^T(MY_i)) \\
&= (\sum_i^N \sum_t^T (x_{it} - \bar{x}_{..})(x_{it} - \bar{x}_{..})^T)^{-1} \sum_i^N \sum_t^T (x_{it} - \bar{x}_{..})(y_{it} - \bar{y}_{..}) \\
&= T_{xx}^{-1} T_{xy}
\end{aligned}
$$

where

$$
\begin{aligned}
T_{xy} &= \sum_i^N \sum_t^T (x_{it} - \bar{x}_{..})(y_{it} - \bar{y}_{..}) \\
T_{xx} &= \sum_i^N \sum_t^T (x_{it} - \bar{x}_{..})(x_{it} - \bar{x}_{..})^T \\
T_{yy} &= \sum_i^N \sum_t^T (y_{it} - \bar{y})^2
\end{aligned}
$$

We call T_{xx}, T_{yy} and T_{xy} the total sum square of x, total sum square of y, and total sum of cross product. The sum of square error for the pooled model is

$$SSE_{pooled} = T_{yy} - T_{xy}^{'} T_{xx}^{-1} T_{xy}$$

with N-1-K degrees of freedom.

4.3 The fixed effect model

4.3.1 Model formulation and estimator

$$y_{it} = \alpha_i + \sum_k x_{itk}\beta_k + u_{it}$$

The above equation can be written in matrix form:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ ... \\ y_{1T} \\ y_{21} \\ y_{22} \\ ... \\ y_{2T} \\ ... \\ y_{NT} \end{pmatrix} = \begin{pmatrix} 1_T & 0 & 0 & ... & 0 \\ 0 & 1_T & 0 & ... & 0 \\ 0 & 0 & 1_T & ... & 0 \\ 0 & 0 & 0 & ... & 1_T \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_N \end{pmatrix}$$

$$+ \begin{pmatrix} x_{111} & x_{112} & ... & x_{11K} \\ x_{121} & x_{122} & ... & x_{12K} \\ ... & ... & ... & ... \\ x_{1T1} & x_{1T2} & ... & x_{1TK} \\ x_{211} & x_{212} & ... & x_{21K} \\ x_{221} & x_{222} & ... & x_{22K} \\ ... & ... & ... & ... \\ x_{2T1} & x_{2T2} & ... & x_{2TK} \\ ... & ... & ... & ... \\ x_{NT1} & x_{NT2} & ... & x_{NTK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_K \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ ... \\ u_{1T} \\ u_{21} \\ u_{22} \\ ... \\ u_{2T} \\ ... \\ u_{NT} \end{pmatrix}$$

Where 1_T is a $1 \times T$ vector. Similarly, using the solution of β from Eq.5, the M matrix is

$$M = I - \frac{1}{T} \begin{pmatrix} 1_{T\times T} & 0 & 0 & ... & 0 \\ 0 & 1_{T\times T} & 0 & ... & 0 \\ 0 & 0 & 1_{T\times T} & ... & 0 \\ 0 & 0 & 0 & ... & 1_{T\times T} \end{pmatrix}$$

Define

$$\tilde{x}_{itk} = (MX)_{itk} = x_{itk} - \bar{x}_{i.k}$$

which is an NT x K matrix,

$$\tilde{y}_{it} = (MY)_{it} = y_{it} - \bar{y}_{i.}$$

which is an NT x 1 vector. The estimated β can be written as

$$\hat{\beta} = ((MX)^{'}(MX))^{-1}((MX)^T(MY_i))$$

and

$$((MX)^{'}(MX))_{kl} = \sum_i \sum_t \tilde{x}_{itk}\tilde{x}_{itl}$$

$$((MX)^{'}(MY))_k = \sum_i \sum_t x_{itk}\tilde{y}_{it}$$

Let $\tilde{X}_{it} = (x_{it1} - \bar{x}_{i.1}, x_{it2} - \bar{x}_{i.2}, ..., x_{itk} - \bar{x}_{i.k})$, $\tilde{y}_{it} = y_{it} - \bar{y}_{i.}$ Then

$$((MX)^{'}(MX)) = \sum_i \sum_t \tilde{X}_{it}\tilde{X}_{it}^{'}$$
$$((MX)^{'}(MY)) = \sum_i \sum_t \tilde{X}_{it}\tilde{y}_{it}$$

$$\begin{aligned}\hat{\beta} &= [\sum_i^N \sum_t^T \tilde{X}_{it}\tilde{X}_{it}^{'}]^{-1}[\sum_i^N \sum_t^T \tilde{X}_{it}\tilde{y}_{it}] \\ &= W_{xx}^{-1}W_{xy}\end{aligned}$$

where

$$W_{xy} = \sum_i^N \sum_t^T \tilde{X}_{it}\tilde{y}_{it}$$
$$W_{xx} = \sum_i^N \sum_t^T \tilde{X}_{it}\tilde{X}_{it}^T$$
$$W_{yy} = \sum_i^N \sum_t^T \tilde{y}_{it}^2$$

We call W_{xx}, W_{yy} and W_{xy} the within-groups sum square of x, the within-groups sum square of y, and within-groups of cross product. The name within-groups means they utilized the variation within group i. The sum of square error is

$$SSE_{fix} = W_{yy} - W_{xy}^{'}W_{xx}^{-1}W^{xy}$$

with NT - N - K degrees of freedom.
If we remind ourselves of the concept of partial regression, we let

$$Y^* = MYX^* = MXU^* = MU$$

then the fix effect model can also be written as

$$Y^* = X^*\beta + U^*$$

or

$$y_{it} - \bar{y}_{i.} = (X_{it} - \bar{X}_{i.})\beta + (u_{it} - \bar{u}_{i.})$$

4.3.2 Properties of Fixed effect estimator

1) Unbiasedness

$$\begin{aligned}
\hat{\beta} &= [\sum_i^N \sum_t^T \tilde{X}_{it} X_{it}^{'}]^{-1} [\sum_i^N \sum_t^T (\tilde{X}_{it})(\tilde{y}_{it})] \\
&= [\sum_i^N \sum_t^T \tilde{X}_{it} X_{it}^{'}]^{-1} [\sum_i^N \sum_t^T (\tilde{X}_{it})(\tilde{X}_{it}^{'} \beta + \tilde{u}_{it})] \\
&= [\sum_i^N \sum_t^T \tilde{X}_{it} X_{it}^{'}]^{-1} [\sum_i^N \sum_t^T (\tilde{X}_{it} \tilde{X}_{it}^{'} \beta + \tilde{X}_{it} \tilde{u}_{it})] \\
&= [\sum_i^N \sum_t^T \tilde{X}_{it} X_{it}^{'}]^{-1} [\sum_i^N \sum_t^T \tilde{X}_{it} \tilde{X}_{it}^{'}] \beta + [\sum_i^N \sum_t^T \tilde{X}_{it} X_{it}^{'}]^{-1} [\sum_i^N \sum_t^T \tilde{X}_{it} \tilde{u}_{it}]
\end{aligned}$$

So

$$E[\hat{\beta}] = \beta + E[\sum_i^N \sum_t^T \tilde{X}_{it} X_{it}^{'}]^{-1} [\sum_i^N \sum_t^T \tilde{X}_{it} \tilde{u}_{it}]$$

As u_{it} is uncorrelated with x_{it}, it is easy to prove the 2nd term above is zero. Therefore, the within group estimator is unbiased.
2) Consistency
To show consistency, we need to prove

$$[\sum_i^N \sum_t^T \tilde{X}_{it} X_{it}^{'}]^{-1} [\sum_i^N \sum_t^T \tilde{X}_{it} \tilde{u}_{it}] = [\frac{1}{NT} \sum_i^N \sum_t^T \tilde{X}_{it} X_{it}^{'}]^{-1} [\frac{1}{NT} \sum_i^N \sum_t^T \tilde{X}_{it} \tilde{u}_{it}] = 0$$

Similar to our previous consistency proof, the above converges to zero when 1) either T or N goes to infinity, 2) $\tilde{x}_{it}$ and $\tilde{u}_{it}$ are not correlated, 3) x_{it} and u_{it} has finite second moment.

4.3.3 Connections between fixed effect model estimator and pooled model estimator when T is large

To see the connection between fixed effect estimator and pooled model estimator, we first introduce between-groups estimator, which allows us to easily understand the relationship among differect estimators.
We define the between-groups sum of square is

$$\begin{aligned}
B_{xx} &= \sum_i^N T(\bar{x}_{i.} - \bar{x}_{...})(\bar{x}_{i.} - \bar{x}_{...})^{'} \\
B_{yy} &= \sum_i^N T(\bar{y}_{i.} - \bar{y}_{...})(\bar{y}_{i.} - \bar{y}_{...})^{'} \\
B_{xy} &= \sum_i^N T(\bar{x}_{i.} - \bar{x}_{...})(\bar{y}_{i.} - \bar{y}_{...})
\end{aligned}$$

It is easy to show that

$$\begin{aligned} T_{xx} &= W_{xx} + B_{xx} \\ T_{yy} &= W_{yy} + B_{yy} \\ T_{xy} &= W_{xy} + B_{xy} \end{aligned}$$

And similarly we can define between-groups estimator

$$\beta_{between} = \frac{B_{xy}}{B_{xx}}$$

We rewrite the estimator of β for the pooled model

$$\begin{aligned} \beta_{pooled} &= \frac{T_{xy}}{T_{xx}} \\ &= \frac{B_{xy}}{T_{xx}} + \frac{W_{xy}}{T_{xx}} \\ &= \frac{B_{xx}}{T_{xx}} \frac{B_{xy}}{B_{xx}} + \frac{W_{xx}}{T_{xx}} \frac{W_{xy}}{W_{xx}} \end{aligned}$$

because

$$T_{xx} = W_{xx} + B_{xx}$$

So if we define

$$\omega = \frac{W_{xx}}{T_{xx}}$$

then

$$\beta_{pooled} = \omega \frac{W_{xy}}{W_{xx}} + (1 - \omega) \frac{B_{xy}}{B_{xx}}$$

When $T \to \infty$, how do W_{xx} and B_{xx} behave? remember

$$\begin{aligned} B_{xx} &= \sum_{i}^{N} T(\bar{x}_{i.} - \bar{x}_{...})(\bar{x}_{i.} - \bar{x}_{...})' \\ W_{xx} &= \sum_{i}^{N} \sum_{t}^{T} (x_{it} - \bar{x}_{i.})(x_{it} - \bar{x}_{i.})^T \end{aligned}$$

When T increase, $\bar{x}_{i.}$ would move close to $E[x_{i.}]$. so its variance would decrease, so B_{xx} grows sub-linealy. In practice, most of the panel data noise comes from transitory noise. so W_{xx} grows faster than B_{xx} when T is large. Then $\omega \to 1$. So in large T, the pooled estimator converges to fix effect estimator.

4.3.4 Limitations of fixed effect model

1)Time-invariant variables are dropped
For a given individual i and regressor k, and any different time t and s($t \neq s$), if $x_{itk} = x_{isk}$, this x variable will be dropped during estimation. Because this condition leads to $x_{it} = \bar{x}_{i.}$
2)Loss degree of freedom
In fixed effect, each individual adds up one parameter, so total N individuals need N parameters. This can cause problem in short panel.

4.3.5 F test for fixed effect model

$$\begin{aligned} F &= \frac{(SSE_{pooled} - SSE_{fix})/((NT-1-K)-(NT-N-K))}{SSE_{fix}/(NT-N-K)} \\ &= \frac{(SSE_{pooled} - SSE_{fix})/(N-1)}{SSE_{fix}/(NT-N-K)} \end{aligned}$$

4.4 The random effect model

4.4.1 Model formulation and estimator

In fixed effect model, we treat individual mean α_i is a constant. In random effect model, we treat α_i as a random variable. We write our random effect model as

$$y_{it} = \sum_k x_{itk}\beta_k + \alpha_i + u_{it}$$

Where $\alpha_i \in Normal(0, \sigma_\alpha^2)$, $u_{it} \in Normal(0, \sigma_u^2)$, $E(\alpha_i u_{it}) = 0$. Define

$$v_{it} = \alpha_i + u_{it}$$

The variance of v_{it} for fixed i is

$$E(v_{it}v_{is}) = E(\alpha_i + u_{it})(\alpha_i + u_{is}) = \sigma_\alpha^2 + \delta_{ts}\sigma_u^2$$

The last term is non-zero only when t = s. Thus, the covariance matrix for individual i is

$$V_i = \begin{pmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & ... & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_u^2 & ... & \sigma_\alpha^2 \\ ... & ... & ... & ... \\ \sigma_\alpha^2 & \sigma_\alpha^2 & ... & \sigma_\alpha^2 + \sigma_u^2 \end{pmatrix}$$

where V_i is $T \times T$ matrix. Stack together for all individuals, the whole covariance matrix is

$$V = \begin{pmatrix} V_1 & 0 & ... & 0 \\ 0 & V_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & V_N \end{pmatrix}$$

Based on the solution of GLS, we define

$$V_i^{-1/2} = \frac{1}{\sigma_u}[I - \frac{\theta}{T}1_T1_T^{'}]$$

where

$$\theta = 1 - \frac{\sigma_u}{\sqrt{\sigma_u^2 + T\sigma_\alpha^2}}$$

and the transformation of X_i and y_i

$$\tilde{y_i} = V_i^{-1/2}y_i = y_i - \theta\bar{y_i}$$
$$\tilde{X_i} = V_i^{-1/2}X_i = X_i - \theta\bar{X_i}$$

where y_i is $T \times 1$ vector, and X_i is $T \times K$ matrix. The estimator of β becomes

$$\hat{\beta}_{RE} = (\tilde{X}^{'}V^{-1}\tilde{X})^{-1}(\tilde{X}^{'}V^{-1}\tilde{Y})$$

With a little derivation, we can prove β_{RE} is a combination of estimator of pooled model and fixed effect model:

$$\hat{\beta}_{RE} = (1 - \omega)\hat{\beta}_{pooled} + \omega\hat{\beta}_{FE}$$

where

$$\omega = \frac{T\sigma_\alpha^2}{T\sigma_\alpha^2 + \sigma_u^2}$$

4.4.2 Connections to estimator of pooled model and fixed effect model

1) when $\sigma_\alpha >> \sigma_u$, then $\omega \to 1$, $\theta \to 1$, this leads to fixed-effect model. In fixed effect model
a. Large σ_α means large viariation of α for different i, in other words, α_i is very different.
b. Since Y_it for fixed i is center around α_i, this means centers of Y_i given different i are far apart.
c. The smaller σ_u compared to σ_α means the variation of Y_it is small. Therefore, the probability density functions of Y_i for different i barely overlap, and their tails almost do not touch each other at all.
2) when $\sigma_\alpha << \sigma_u$, then $\omega \to 0$, $\theta \to 0$, this leads to pooled model. In pooled model,
a. Small σ_α means α_is are almost identitical .
b. Since Y_it for fixed i is center around α_i, this means centers of Y_i given different i are very close to each other.
c. The larger σ_u compared to σ_α means the variation of Y_it is large. Therefore, the probability density functions of Y_i for different i are completely overlapping. The difference between different individual i is hardly visible. So it is unnecessary to model the individual mean.

4.4.3 Estimation of σ_u^2 and σ_α^2

Estimation of σ_u^2 can be based on the regression formula

$$y_{it} - \bar{y}_{i.} = (X_{it} - \bar{X}_{i.})\beta + (u_{it} - \bar{u}_{i.})$$

and σ_u^2 can be estimated using sum of square error and its degrees of freedom.

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} [(y_{it} - \bar{y}_i) - \hat{\beta}_{fix}(x_{it} - \bar{x}_i)]}{N(T-1) - K}$$

σ_α^2 can be estimated by first taking the average of random effect model.

$$\bar{y}_i = \mu + \bar{X}_i\beta + \alpha_i + \bar{u}_i$$

Then because α and u_i are independent,

$$Var(\bar{y}_i - \mu - \bar{X}_i\beta) = Var(\alpha) + Var(\bar{u}_i) = Var(\alpha) + \frac{1}{T}Var(u)$$

Then

$$\sigma_\alpha^2 = \frac{\sum_{i=1}^{N} (\bar{y}_i - \hat{\mu} - X_i\hat{\beta})^2}{N - (K+1)} - \frac{1}{T}\hat{\sigma}_u^2$$

4.4.4 Coping with the Limitation of Random Model

The random model, like linear model in general, assumes α does not correlate with variable X. Mundlak intruduced auxiliary regression where we write

$$\alpha_i = \sum_k \bar{x}_{i.k} a_k + \omega_i$$

where $\omega_i \in Normal(0, \sigma_\omega^2)$. So

$$y_{it} = \sum_k x_{itk}\beta_k + \sum_k \bar{x}_{i.k} a_k + \omega_i + u_{it}$$

If we define the error term

$$v_{it} = \omega_i + u_{it}$$

The above equation can be written in matrix form:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ ... \\ y_{1T} \\ y_{21} \\ y_{22} \\ ... \\ y_{2T} \\ ... \\ y_{NT} \end{pmatrix} = \begin{pmatrix} 1_T & 0 & 0 & ... & 0 \\ 0 & 1_T & 0 & ... & 0 \\ 0 & 0 & 1_T & ... & 0 \\ 0 & 0 & 0 & ... & 1_T \end{pmatrix} \begin{pmatrix} \bar{x}_{1.1} & \bar{x}_{1.2} & ... & \bar{x}_{1.K} \\ \bar{x}_{2.1} & \bar{x}_{2.2} & ... & \bar{x}_{2.K} \\ \bar{x}_{3.1} & \bar{x}_{3.2} & ... & \bar{x}_{3.K} \\ ... & ... & ... & ... \\ \bar{x}_{N.1} & \bar{x}_{N.2} & ... & \bar{x}_{N.K} \end{pmatrix} \begin{pmatrix} \bar{a}_1 \\ \bar{a}_2 \\ \bar{a}_3 \\ ... \\ \bar{a}_K \end{pmatrix}$$

$$+ \begin{pmatrix} x_{111} & x_{112} & ... & x_{11K} \\ x_{121} & x_{122} & ... & x_{12K} \\ ... & ... & ... & ... \\ x_{1T1} & x_{1T2} & ... & x_{1TK} \\ x_{211} & x_{212} & ... & x_{21K} \\ x_{221} & x_{222} & ... & x_{22K} \\ ... & ... & ... & ... \\ x_{2T1} & x_{2T2} & ... & x_{2TK} \\ ... & ... & ... & ... \\ x_{NT1} & x_{NT2} & ... & x_{NTK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_K \end{pmatrix} + \begin{pmatrix} v_{11} \\ v_{12} \\ ... \\ v_{1T} \\ v_{21} \\ v_{22} \\ ... \\ v_{2T} \\ ... \\ v_{NT} \end{pmatrix}$$

Similarly to the original random model, the variance matrix of v_{it} is

$$V_i = \begin{pmatrix} \sigma_\omega^2 + \sigma_u^2 & \sigma_\omega^2 & ... & \sigma_\omega^2 \\ \sigma_\omega^2 & \sigma_\omega^2 + \sigma_u^2 & ... & \sigma_\omega^2 \\ ... & ... & ... & ... \\ \sigma_\omega^2 & \sigma_\omega^2 & ... & \sigma_\omega^2 + \sigma_u^2 \end{pmatrix}$$

The way to estimator β should be GLM, but what happens if we just run OLS on this model by assuming v_{it} follows standard normail distribution? Remember the method of doing partial regression. Running OLS on this model is equivalently running the **residual** of Y after regressing on $\bar{X}_{i.}$ on the residual of X after regressing on $\bar{X}_{i.}$.

Obviously, regressing X_{it} on $X_{i.}$ leads the residual

$$X_{it}^* = X_{it} - \bar{X}_{i.}$$

Where each term is a K by 1 vector Let us regress Y_{it} on $\bar{X}_{i.}$ and compute the residual. The regression is

$$y_{it} = \mu + \lambda \bar{x}_{i.} + \eta_{it}$$

We realize that μ and $\bar{x}_{i.}$ are constant for a given i. So the fitted value on a constant is the mean of the y_{it} with respect t. We can write

$$\bar{y}_{i.} = \sum_k \beta_k \bar{x}_{i.k} + \sum_k \bar{x}_{i.k} a_k + \bar{v}_i = \sum_k (\beta_k + a_k) \bar{x}_{i.k} + \bar{v}_i$$

We know $E[v_i] = 0$. So when we regress y on $\bar{x}_{i.}$, the fitted value becomes

$$\hat{y}_{it} = \sum_k (\beta_k + a_k) \bar{x}_{i.k}$$

Then the residual is

$$\begin{aligned} y_{it} - \hat{y}_{it} &= \sum_k x_{ik}\beta_k + \sum_k \bar{x}_{i.k} a_k + \epsilon_{it} - (\sum_k (\beta_k + a_k)\bar{x}_{i.k}) \\ &= \sum_k \beta_k (x_{ik} - \bar{x}_{i.k}) + \epsilon_{it} - \bar{\epsilon}_{i.} \\ &= y_{it} - \bar{y}_{i.} \end{aligned}$$

So the OLS we try to run is regress $y_{it} - \bar{y}_{i.}$ on $x_{it} - \bar{x}_{i.}$, which is equivalent to fixed effect estimation.

4.5 Fixed Effect Model Generalization

We can generalize fixed effect model by including more individual specific variables which vary across different individuals and but do not vary over time. We can write the model as

$$\begin{aligned} y_{it} &= \mu + \alpha_i + z_{i1}\gamma_1 + z_{i2}\gamma_2 + ... + z_{ip}\gamma_p + \sum_k x_{itk}\beta_k + u_{it} \\ &= \mu + \alpha_i + \sum_p z_{ip}\gamma_p + \sum_k x_{itk}\beta_k + u_{it} \end{aligned}$$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ ... \\ y_{1T} \\ y_{21} \\ y_{22} \\ ... \\ y_{2T} \\ ... \\ y_{NT} \end{pmatrix} = \begin{pmatrix} 1_T & 0 & 0 & ... & 0 \\ 0 & 1_T & 0 & ... & 0 \\ 0 & 0 & 1_T & ... & 0 \\ 0 & 0 & 0 & ... & 1_T \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_N \end{pmatrix}$$

$$+ \begin{pmatrix} 1_T & 0 & 0 & ... & 0 \\ 0 & 1_T & 0 & ... & 0 \\ 0 & 0 & 1_T & ... & 0 \\ 0 & 0 & 0 & ... & 1_T \end{pmatrix} \begin{pmatrix} z_{11} & z_{12} & ... & z_{1p} \\ z_{21} & z_{22} & ... & z_{2p} \\ ... & ... & ... & ... \\ z_{N1} & z_{N2} & ... & z_{Np} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ ... \\ \gamma_N \end{pmatrix}$$

$$+ \begin{pmatrix} x_{111} & x_{112} & ... & x_{11K} \\ x_{121} & x_{122} & ... & x_{12K} \\ ... & ... & ... & ... \\ x_{1T1} & x_{1T2} & ... & x_{1TK} \\ x_{211} & x_{212} & ... & x_{21K} \\ x_{221} & x_{222} & ... & x_{22K} \\ ... & ... & ... & ... \\ x_{2T1} & x_{2T2} & ... & x_{2TK} \\ ... & ... & ... & ... \\ x_{NT1} & x_{NT2} & ... & x_{NTK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_K \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ ... \\ u_{1T} \\ u_{21} \\ u_{22} \\ ... \\ u_{2T} \\ ... \\ u_{NT} \end{pmatrix}$$

This model suffers multicollinearity so α and γ are not both estimable. We can set

$$\alpha_i^* = \alpha_i + \sum_p z_{ip}\gamma_p$$

Then the model is effectively the same as the original fixed effect model, so we can use the same trick of solving fixed effect model to estimate β. To estimate γ, we note

$$\bar{y}_{i.} - \sum \bar{x}_{i.k}\beta_k = \sum_p z_{ip}\gamma_p + \alpha_i + \bar{u}_i$$

Let $\epsilon_i = \alpha_i + \bar{u}_i$ and by minimizing $\sum_i \epsilon_i^2$, we obtain

$$\hat{\gamma} = [\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})^*]^{-1}\{\sum_{i=1}^{N}(z_i - \bar{z})[(\bar{y}_i - \bar{y}) - (\bar{x}_i - \bar{x})^{'}\beta]\}$$

Testing of σ_α We define the null and alternative hypothesis
$H_0 : \sigma_\alpha^2 = 0$
$H_1 : \sigma_\alpha^2 > 0$
We construct LM test, which uses the score function(gradient of log-likelyhood) with respect to σ_α^2 evaluated at $\sigma_\alpha^2 = 0$:

$$\begin{aligned} LM &= \frac{(\partial l/\partial \sigma_\alpha^2)^2}{Var(\partial l/\partial \sigma_\alpha^2)}|_{\sigma_\alpha^2} \\ &= \frac{nT}{2(T-1)}[\frac{[\sum_{i=1}^{n}(\bar{e}_i)^2]}{(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}e_{it}^2)} - 1]^2 \end{aligned}$$