

1 Linear Regression

1.1 Linear regression and Least Square Solution

$$Y = X\beta + \epsilon$$

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

Assumptions

1. Linear
2. X matrix has full rank. In other words, no multicollinearity.
3. Homoscedasticity or equal variance of ϵ . In other words, no autocorrelation between disturbances. $cov(\epsilon_i, \epsilon_j) = 0$.
6. Number of observations n must be greater than the number of parameters.

Least Square Solution

The cost function is given by

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

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 derivative of 2nd term

Consider the derivative of $\alpha^T\beta$ with respect to β .

$$\begin{aligned}\alpha^T\beta &= \sum \alpha_i\beta_i \\ \frac{\partial \alpha^T\beta}{\partial \beta_i} &= \alpha_i\end{aligned}$$

Write the derivative in matrix form

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

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 $A = X^TX$,

$$\beta^TX^TX\beta = \beta^T \begin{pmatrix} \sum_i A_{1k}\beta_k \\ \sum_i A_{2k}\beta_k \\ \dots \\ \sum_k A_{pk}\beta_k \end{pmatrix} = \sum_j \beta_j (\sum_k A_{jk}\beta_k)$$

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

1) $l = j = k$

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

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

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

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

$$\frac{f(\beta)}{\partial \beta_l} = \sum_{j, j \neq l} A_{jl}\beta_j = \sum_{j, j \neq l} A_{lj}^T \beta_j$$

Therefore

$$\begin{aligned} \frac{f(\beta)}{\partial \beta_l} &= A_{ll}\beta_l + \sum_{k, k \neq l} A_{lk}\beta_k + A_{ll}\beta_l + \sum_{j, j \neq l} A_{lj}^T \beta_j \\ &= \sum_k A_{lk}\beta_k + \sum_j A_{lj}^T \beta_j \end{aligned}$$

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

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

which is a $p \times 1$ 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

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

Estimator of β

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

Least Square Estimator for Simple Linear Regression

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

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

So

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

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

1.2 Projection matrix

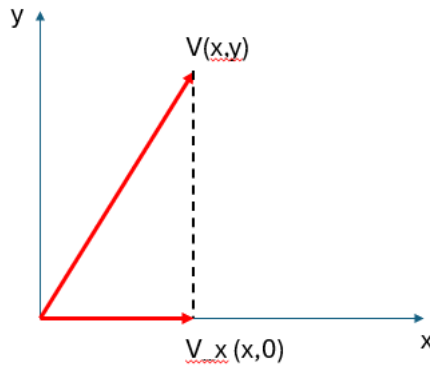
Given $\hat{\beta} = (X^T X)^{-1} X^T Y$, we have the predictor value of $y = X\beta$

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

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

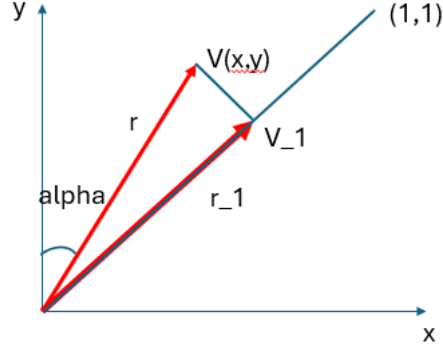
$$P_x = x(x^T x)^{-1} x^T$$

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

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

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

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

$$V_{1x} = r_1 \cos(\pi/4) = \frac{x+y}{2}$$

$$V_{1y} = r_1 \sin(\pi/4) = \frac{x+y}{2}$$

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

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

The projection matrix P is

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

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 $M y$ 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 Partial Regression

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

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

$$\begin{aligned} \hat{\beta}_1 &= (X_1^T X_1)^{-1} X_1^T y \\ \hat{\beta}_2 &= (X_2^T X_2)^{-1} X_2^T y \end{aligned}$$

If X_1 and X_2 are not orthogonal, we can solve for β_2 in the above normal equation set and get β_2

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

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

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

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

$$M_1 y = M_1 X_2 \beta + \epsilon$$

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 .

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

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

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

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

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

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

Try

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

So

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

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

$$\begin{aligned}
SSE &= \sum_i (y - \hat{y}_i)^2 \\
&= (Y - X\beta)^T (Y - X\beta) \\
&= (Y - X(X^T X)^{-1} X^T Y)^T (Y - X(X^T X)^{-1} X^T Y) \\
&= (Y - PY)^T (Y - PY) \\
&= Y^T (1 - P)^T (1 - P) Y = Y^T (1 - P) Y \\
&= (X\beta + \epsilon)^T (1 - P) (X\beta + \epsilon) \\
&= \beta^T X^T (1 - P) X \beta + 2\beta^T X^T (1 - P) \epsilon + \epsilon^T (1 - P) \epsilon
\end{aligned}$$

$$E[SSE] = E[\epsilon^T (1 - P) \epsilon] = E[\epsilon^T \epsilon] \text{trace}(1 - H) = \sigma^2(n - k)$$

We obtain the unbiased estimator of σ^2

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

Therefore the estimator of variance of β

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

and the standard error of β_i is

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

2 Properties of Least Square Estimators

2.1 Unbiasness

When we have an estimator, we need to ask ourselves two questions, 1) how accurate is our estimator, can the estimator give us the true value of β , 2) Does it converge to the real value with reasonable speed as sample size increases? The metric we use to evaluate accuracy is unbiasedness, and the metric we measure the convergence speed is the efficiency.

Unbiased

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

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

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

The last term is zero by assumption 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

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \epsilon$$

On the second term, we see unless 1) X_1 and X_2 are orthogonal, or 2) $\beta_2 = 0$, β_1 is biased.

2.2 Consistency

The unbiasedness gives us a metric of measuring how good our estimator is, from population perspective. 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

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

$$\begin{aligned} X^T X &= \sum_{i=1}^N \begin{pmatrix} x_{1i}^T x_{i1} & x_{1i}^T x_{i2} & \dots & x_{1i}^T x_{ik} \\ \dots & \dots & \dots & \dots \\ x_{ki}^T x_{i1} & x_{ki}^T x_{i2} & \dots & x_{ki}^T x_{ik} \end{pmatrix} \\ &= \sum_{i=1}^N \begin{pmatrix} x_{i1} x_{i1} & x_{i1} x_{i2} & \dots & x_{i1} x_{ik} \\ \dots & \dots & \dots & \dots \\ x_{ik} x_{i1} & x_{ik} x_{i2} & \dots & x_{ik} x_{ik} \end{pmatrix} \\ &= \sum_{i=1}^N \begin{pmatrix} x_{i1} \\ \dots \\ x_{ik} \end{pmatrix} \begin{pmatrix} x_{i1} & \dots & \dots & x_{ik} \end{pmatrix} \\ &= \sum_{i=1}^N X_i X_i^T \end{aligned}$$

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

If X_i s are iid, then by law of large numbers

$$\sum_{i=1}^N \frac{1}{n} X_i^T X_i$$

converges to Q in probability.

$$\begin{aligned} & \frac{X^T \epsilon}{N} \\ &= \begin{pmatrix} \frac{1}{N} \sum_1^N x_{i1} \epsilon_i \\ \frac{1}{N} \sum_1^N x_{i2} \epsilon_i \\ \frac{1}{N} \sum_1^N x_{i3} \epsilon_i \\ \vdots \\ \frac{1}{N} \sum_1^N x_{ik} \epsilon_i \end{pmatrix} \\ &= \frac{1}{N} \sum_{i=1}^N X_i \epsilon_i = \bar{w} \end{aligned}$$

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

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

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

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

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

When $\frac{X^T X}{n}$ 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 inconsistent.

1) X is not full rank, or X has multicollinearity 2) $cov[X, \epsilon] \neq 0$

2.3 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.

2.4 Model Testing

Lagrange Multiplier(LM) test

Suppose we have two models, one is restricted, 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 likelihood function is

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

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

When $\beta = 0$, define

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

and $M_1 X_1 = 0$.

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

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

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

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

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

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

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

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

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

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

$$M_1 X_2 (X_2^T M_1 X_2)^{-1} X_2^T M_1$$

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.

$$P_U - P_R = M_1 X_2 (X_2^T M_1 X_2)^{-1} X_2^T M_1$$

The extra explained sum of squares by the unrestricted model is

$$SSR_R - SSR_U = y^T (P_R - P_U) y = (X_2^T M_1 y)^T (X_2^T M_1 X_2)^{-1} X_2^T M_1 y$$

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

$$F = \frac{SSR_R - SSR_U}{\frac{SSR_U}{n-k}} = \frac{(X_2 M_1 y)^T (X_2^T M_1 y)}{\hat{\sigma}^2 (X_2^T M_1 X_2)} = LM$$

We see that F test and LM test are equivalent.