Econometrics
Part I: Basic regression.
Session 2.

References on Bruce Hansen's textbook will be abbreviated with H. Say, H 1.1 means chapter 1.1 from Bruce Hansen's textbook.

Gauss-Markov theorem. Violations of GM theorem. (H 4.8, 4.20)
Estimating σ^2. (H 4.11)
Partitioned regression. (H 3.16, 3.18).
Maximum Likelihood Estimation. Cramer-Rao Lower Bound. (H 5.7-5.8, 5.18)

Linear combination of parameters

Often, our interest lies in accurate estimation of a linear combinations of regression parameters. Consider, for example, a regression of the logarithm of the total cost (TC) of a firm's production on the logarithm of the quantity (Q) produced, and the logarithms of the prices of the factors of production, say, capital and labour, (PC and PL) (see Ch. 1.7 of Hayashi (2000))

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PC_i + \beta_4 \log PL_i + \varepsilon_i.$$

Prediction example. The prediction of $\log TC$ for the firm with, say, $\log Q = 25, \log PC = 2$ and $\log PL = 1$ is the linear combination

$$\beta_1 + \beta_2 25 + \beta_3 2 + \beta_4.$$

Policy example. The effect of a subsidy that reduces PC by 10% and PL by 5% on the total cost is the linear combination

$$\beta_3 \log 0.9 + \beta_4 \log 0.95.$$

Testing example. To test the homogeneity of total cost with respect to the prices of production, we need to know whether

$$\beta_3 + \beta_4 = 1$$

In general, we may be interested in $\sum_{j=1}^{k} \gamma_j \beta_j = \gamma' \beta$, where γ is a $k \times 1$ vector of weights. If $\hat{\beta}$ is conditionally unbiased for β, then $\gamma' \hat{\beta}$ is conditionally unbiased

for $\gamma'\beta$. And the quality of $\gamma'\hat{\beta}$ as a conditionally unbiased estimator of $\gamma'\beta$ will be measured by

$$Var(\gamma'\hat{\beta}|X) = \gamma' Var\left(\hat{\beta}|X\right)\gamma,$$

where $Var\left(\hat{\beta}|X\right)$ is the so-called conditional variance-covariance matrix of $\hat{\beta}$.

Variance-covariance matrix

For a scalar variable Z variance is defined as

$$VarZ = EZ^2 - E^2Z$$

For vector variable we have variance-covariance matrix

$$\begin{aligned}
& VarZ \\
= \; & EZZ' - (EZ)(EZ)' \\
= \; & E\begin{pmatrix} Z_1^2 & Z_1Z_2 & ... & Z_1Z_n \\ Z_2Z_1 & Z_2^2 & ... & Z_2Z_n \\ \vdots & \vdots & \ddots & \vdots \\ Z_nZ_1 & Z_nZ_2 & ... & Z_n^2 \end{pmatrix} - E\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} E\begin{pmatrix} Z_1 & Z_2 & ... & Z_n \end{pmatrix} \\
= \; & \begin{pmatrix} varZ_1 & cov(Z_1Z_2) & ... & cov(Z_1Z_n) \\ cov(Z_1Z_2) & varZ_2 & ... & cov(Z_2Z_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(Z_1Z_n) & cov(Z_2Z_n) & ... & varZ_n \end{pmatrix}
\end{aligned}$$

For any deterministic $m \times n$ matrix M,

$$Var(MZ) = M(VarZ)M'.$$

Indeed, by definition,

$$\begin{aligned}
Var(MZ) &= E\left[(MZ)(MZ)'\right] - E(MZ)\left[E(MZ)\right]' \\
&= E\left[MZZ'M'\right] - E(MZ)E(Z'M') \\
&= ME\left[ZZ'\right]M' - ME(Z)\left[E(Z')M'\right] \\
&= M\left(E\left[ZZ'\right] - E(Z)E(Z')\right)M' = MVar(Z)M'.
\end{aligned}$$

Variance conditional on realizations of another random variable, X is defined as

$$\begin{aligned} Var(Z|X) &= E\left[(Z - E(Z|X))(Z - E(Z|X))'|X\right] \\ &= E(ZZ'|X) - E(Z|X)E(Z'|X) \end{aligned}$$

A useful formula (law of total variance) connecting conditional and unconditional variances is

$$Var(Z) = E\{Var(Z|X)\} + Var\{E(Z|X)\}.$$

The Gauss-Markov Theorem

Theorem 1 *Consider an $n \times 1$ random vector Y and an $n \times k$ random matrix X. Assume that:*
(GM1) rank(X)=k
(GM2) $E(Y|X) = X\beta$ (equivalently, $E(\varepsilon|X) = 0$)
(GM3) $Var(Y|X) = \sigma^2 I$ (equivalently, $Var(\varepsilon|X) = \sigma^2 I$)
Then $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$ has minimum variance in the class of estimators that, conditional on every X, are linear in Y and unbiased. In other words, $\hat{\beta}_{OLS}$ is the Best Linear conditionally Unbiased Estimator (BLUE).

Remark 2 *We may interpret Y and X as n i.i.d. observations of dependent variable y and explanatory variables $x_1, x_2, ..., x_k$ but this is not necessary. In particular, the observations may be not i.i.d. In fact, in classical regression model X is usually considered to be non-stochastic (that is random variable with degenerate distribution) so that the values of X are repeated in different samples. Further, you do not need assumptions that n (which may be interpreted as a number of observation in a regression) to go to infinity or that errors ε are normally distributed. The Gauss-Markov theorem is finite sample and distribution-free result. All we need for the Gauss-Markov theorem to hold is GM1-3.*

Proof: *Linearity:*

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

which is a linear function of Y.

Unbiasedeness: We have shown in Session 1 that $\hat{\beta}_{OLS}$ is conditionally (and unconditionally) unbiased for β when CEF is linear, which is what GM2 assumes.

Minimum variance:

$$Var(\hat{\beta}_{OLS}|X) = Var((X'X)^{-1}X'Y|X) = (X'X)^{-1}X'Var(Y|X)X(X'X)^{-1}$$

Using GM3 we get

$$Var(\hat{\beta}_{OLS}|X) = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

The unconditional variance of $\hat{\beta}_{OLS}$ is then equal to

$$\begin{aligned} Var(\hat{\beta}_{OLS}) &= \sigma^2 E(X'X)^{-1} + Var(E(\hat{\beta}_{OLS}|X)) \\ &= \sigma^2 E(X'X)^{-1} + Var(\beta) = \sigma^2 E(X'X)^{-1} \end{aligned}$$

Now let us consider any linear conditionally unbiased estimator $\tilde{\beta} = AY$. We have

$$E(\tilde{\beta}|X) = E(AY|X) = AX\beta$$

This must be equal to β because $\tilde{\beta}$ is assumed to be conditionally unbiased. That is, for any β

$$AX\beta = \beta$$

and hence, $AX = I$. What is conditional variance of $\tilde{\beta}$?

$$Var(\tilde{\beta}|X) = Var(AY|X) = \sigma^2 AA'$$

Now, let us decompose A into two parts

$$A = A - (X'X)^{-1}X' + (X'X)^{-1}X' = W + (X'X)^{-1}X'$$

so that the conditional variance of $\tilde{\beta}$ is equal to

$$\begin{aligned} & \sigma^2(W + (X'X)^{-1}X')(W + (X'X)^{-1}X')' \\ = \; & \sigma^2\left[WW' + WX(X'X)^{-1} + (X'X)^{-1}X'W' + (X'X)^{-1}\right] \end{aligned}$$

But

$$WX = AX - (X'X)^{-1}X'X = I - I = 0$$

Therefore,

$$\begin{aligned} Var(\tilde{\beta}|X) &= \sigma^2 WW' + \sigma^2 (X'X)^{-1} \\ &= \sigma^2 WW' + Var(\hat{\beta}_{OLS}|X) \end{aligned}$$

so $Var(\tilde{\beta}|X)$ is $Var(\hat{\beta}_{OLS}|X)$ plus some positive semi-definite matrix, where matrix Q is positive semi-definite if and only if

$$x'Qx \geq 0$$

for any vector x. In this sense, conditional variance of any conditionally unbiased linear estimator is larger than that of the OLS estimator. Note that conditional variances of all components of $\hat{\beta}_{OLS}$: $\hat{\beta}_{1,OLS}$, $\hat{\beta}_{2,OLS}$, ..., $\hat{\beta}_{k,OLS}$ is less than the variance of the corresponding components of $\tilde{\beta}$. Moreover, conditional variance of any linear combination of the components of $\hat{\beta}_{OLS}$ is less than that of the corresponding linear combination of components of $\tilde{\beta}$.

What about unconditional variance?

$$\begin{aligned} Var(\tilde{\beta}) &= E(Var(\tilde{\beta}|X)) + Var(E(\tilde{\beta}|X)) \\ &= \sigma^2 EWW' + Var(\hat{\beta}_{OLS}) + Var\beta \\ &= \sigma^2 EWW' + Var(\hat{\beta}_{OLS}) \end{aligned}$$

That is, the unconditional variance of $\hat{\beta}_{OLS}$ is also less that that of $\tilde{\beta}$.

The theorem was proved by Gauss as early as in 1821. Later the theorem was mistakenly credited to Markov by Jerzy Neyman. Since then the theorem is called the Gauss-Markov theorem.

Violations of the GM theorem: GM1

Let us see what happens when GM1 is violated.

GM1: rank(X)$\neq k$. What does it mean that rank(X)<k? This means that there exists an exact linear dependence between the columns of X. If we write $X = [X^1, X^2, ..., X^k]$ where X^i denote the i-th column of X then there exist α_i, i=1,...,k, not all equal to zero and such that

$$\alpha_1 X^1 + ... + \alpha_k X^k = 0$$

or in matrix notations, there exists a $k \times 1$ vector α such that

$$X\alpha = 0$$

When rank(X)<k matrix $X'X$ is not invertible and hence $\hat{\beta}_{OLS}$ is not defined. What essentially is going on? We know that $Y = X\beta + \varepsilon$ but since $X\alpha = 0$ we may write that

$$\begin{aligned} Y &= X(\beta + \lambda\alpha) + \varepsilon \\ Y &= X\tilde{\beta} + \varepsilon \end{aligned}$$

for any λ. Hence any $\tilde{\beta}$ of the above form satisfies GM theorem and it is not clear what particular $\tilde{\beta}$ we are trying to estimate. We say that β is not possible to identify from the given data. A situation when rank(X)<k is called exact (or perfect) multicollinearity.

A very well known example of multicollinearity is so called dummy variable trap. Suppose that you are estimating a linear consumption function $C = \alpha + \beta * DI$ where DI stands for disposable income. And you want to get rid of possible seasonal factors affecting consumption, so you include dummy variables for Winter, Spring, Summer and Autumn and estimate regression

$$C = \alpha_0 + \alpha_1 W + \alpha_2 Sp + \alpha_3 Su + \alpha_4 A + \beta DI + \varepsilon$$

W, Sp, Su, and A are $n \times 1$ vectors of 0 and 1. For example $W_i = 1$ if i-th observation was made during winter and $W_i = 0$ otherwise. Parameters of the above model can not be identified. Indeed, we can increase all α_i, i=1,..,4 by the same amount and decrease α_0 by this amount and the model will remain valid.

Here matrix $X = [1, W, Sp, Su, A, DI]$ and if $c = [-1, 1, 1, 1, 1, 0]$, then $Xc = 0$. In other words $W + Sp + Su + A = 1$ identically. User friendly programs, such as STATA will automatically drop one of the dummy variables, but if you use, say, MATLAB, then computer will say to you that matrix $X'X$ is ill-conditioned or poorly scaled (in short that X'*X cannot be inverted with a reasonable numerical accuracy).

Another example of perfect multicollinearity is situation when $n < k$ that is you have too few data points. If say you have only two observations and two regressors than you will perfectly fit the data. If you have three regressors than

you can perfectly fit data in many different ways.

A remedy for perfect multicollinearity would be to get rid of one dummy (or the constant) in the dummy trap example and get more data in the second example.

Often multicollinearity will not be perfect in that there will exist a linear combination of columns of X that is almost but not exactly equal to zero

$$Xc \approx 0$$

We will study effects of such non-perfect multicollinearity in more detail soon. Now I only mention that if there is multicollinearity then variance of the parameter estimates is likely to be high. Often, you may notice that estimates change sign when just a single observation is added to the data set. Also, the signs and the sizes of the estimates may be well off those theoretically expected. Why estimates are so variable in the multicollinearity case? Let me postpone this question until we study partitioned regression.

How to detect multicollinearity in the case when you have many regressors? One way is to look at the matrix of correlations of X. If some correlations are close to 1 or -1 then this indicates multicollinearity. However, this method will only detect linear dependence of two variables among the explanatory variables. Often, as in the dummy example, we will have linear dependence that involves many variables. In this case one can analyze eigenvalues of $X'X$.

Sometimes, instead of defining multicollinearity as a situation when $Xc \approx 0$ we define it as a situation when the condition number (ratio of the largest to the smallest eigenvalue) of $X'X$ is larger than a certain big number. Imagine situation when i-th eigenvalue of $X'X$ is extremely small. Is there a vector c such that $Xc \approx 0$? Denote i-th eigenvector of $X'X$ as e_i. Then $c = e_i$ because $e_i'X'Xe_i = \lambda_i e_i'e_i$ and since λ_i is very small $Xe_i \approx 0$. To detect multicollinearity we therefore analyze eigenvalues and eigenvectors of matrix $X'X$. the eigenvector of $X'X$ corresponding to small eigenvalue gives us components of the vector c.

How to combat multicollinearity? Well this is a little bit tricky question. Remember, GM theorem says that as long as multicollinearity is not perfect, β_{OLS} is still the best among the unbiased estimators of β even though variance of β_{OLS} is very large. If multicollinearity is present we might want to abandon the unbiasedness requirement and consider biased estimators. As we know, biased estimators may be better than OLS in terms of minimizing mean squared error of the estimation (because they can trade off bias and variance in a smart way). Hoerl and

Kennard (1970) proposed to use the so-called ridge regression instead of OLS in case of multicollinearity

$$\hat{\beta}_r = (X'X + \lambda I)^{-1} X'Y$$

Note that

$$E\left(\hat{\beta}_r | X\right) = (X'X + \lambda I)^{-1} X'X\beta$$

so $\hat{\beta}_r$ is biased if $\lambda \neq 0$. When λ rises bias increases but variance of the estimator falls.

Ridge and LASSO regressions as penalized LS.

It is easy to see that $\hat{\beta}_r$ is solving the Least Squares problem when size of $\|\beta\|^2 = \beta'\beta$ is penalized:

$$\min \|Y - X\beta\|^2 \text{ s.t. } \|\beta\|^2 \leq t$$

for some $t \geq 0$. Indeed, this constrained minimization problem cast in the Lagrange form is

$$\min \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|^2 \right\}$$

for some $\lambda \geq 0$. The first order conditions lead to solution

$$\hat{\beta}_r = (X'X + \lambda I)^{-1} X'Y.$$

Recently, there has been a lot of attention to a different LS penalization method, called LASSO (Least Absolute Shrinkage and Selection Operator), introduced by Tibshirani (1996). The problem solved by LASSO is

$$\min \|Y - X\beta\|^2 \text{ s.t. } \sum\nolimits_{j=1}^{k} \left|\beta_j\right| \leq t$$

In Lagrange form it looks like

$$\min \left\{ \|Y - X\beta\|^2 + \lambda \sum\nolimits_{j=1}^{k} \left|\beta_j\right| \right\}$$

for some $\lambda \geq 0$. Similarly to $\hat{\beta}_r$, LASSO estimates $\hat{\beta}_{LASSO}$ are more stable than OLS. In addition, many components of $\hat{\beta}_{LASSO}$ are often zero. Hence, LASSO not only estimates the parameters, but also excludes some variables from the model, making it more parsimonious and, thus, better interpretable.

The workings of LASSO are particularly transparent in the so-called orthonor-

mal design case, when $X'X = I$. In such a case,

$$\|Y - X\beta\|^2 = \left\|\beta - \hat{\beta}_{OLS}\right\|^2 + \hat{\varepsilon}'_{OLS}\hat{\varepsilon}_{OLS}$$

so the problem is equivalent to

$$\min\left\{\sum_{j=1}^{k}\left(\beta_j - \hat{\beta}_{j,OLS}\right)^2 + \lambda\sum_{j=1}^{k}\left|\beta_j\right|\right\}.$$

Note that the function to minimize is a sum of functions

$$\left(\beta_j - \hat{\beta}_{j,OLS}\right)^2 + \lambda\left|\beta_j\right|,$$

which can be minimized separately. We have

$$\left(\beta_j - \hat{\beta}_{j,OLS}\right)^2 + \lambda\left|\beta_j\right| = \begin{cases} \left(\beta_j - \hat{\beta}_{j,OLS}\right)^2 + \lambda\beta_j & \text{for } \beta_j \geq 0 \\ \left(\beta_j - \hat{\beta}_{j,OLS}\right)^2 - \lambda\beta_j & \text{for } \beta_j < 0 \end{cases}$$

Using this, it is straightforward to obtain the following solution

$$\hat{\beta}_{j,LASSO} = \begin{cases} \hat{\beta}_{j,OLS} - \lambda/2 & \text{if } \hat{\beta}_{j,OLS} \geq \lambda/2 \\ \hat{\beta}_{j,OLS} + \lambda/2 & \text{if } \hat{\beta}_{j,OLS} \leq -\lambda/2 \\ 0 & \text{otherwise} \end{cases}$$

Such an estimator is also known as the soft thresholding estimator: when $\left|\hat{\beta}_{j,OLS}\right|$ is below the threshold $\lambda/2$ set $\hat{\beta}_{j,LASSO}$ to zero, when $\left|\hat{\beta}_{j,OLS}\right|$ is above the threshold, set the estimator to $sgn\left\{\hat{\beta}_{j,OLS}\right\}\left(\left|\hat{\beta}_{j,OLS}\right| - \lambda/2\right)$.

In contrast, the ridge estimator in the case of the orthonormal design has form

$$\hat{\beta}_r = (I + \lambda I)^{-1}X'Y = \frac{\hat{\beta}_{OLS}}{1+\lambda}.$$

So it does not set any $\hat{\beta}_{j,r}$ to zero. It just shrinks $\hat{\beta}_{OLS}$.

Violations of the GM theorem: GM2

Let us now turn to violation of GM2: $E(\varepsilon|X) \neq 0$

This happens if CEF of Y is not equal to $X\beta$. For example, when CEF is not linear. To quickly check linearity of $X\beta$, one can plot residuals versus values of Y predicted from regression, that is versus $X\hat{\beta}_{OLS}$. If graph exhibits patterns, this

is an indication of model misspecification. There exist a vast literature on testing different model specification. A good reference is Ed. Leamer's chapter in the Handbook of Econometrics. We will talk more about special forms of violation of $E(Y|X) = X\beta$ assumption later in the course.

What are the consequences of such a violation? It turns out that $\hat{\beta}_{OLS}$ becomes biased. Indeed

$$E\left(\hat{\beta}_{OLS}|X\right) = \beta + (X'X)^{-1}X'E(\varepsilon|X) \neq \beta$$

Note that even if E($\varepsilon|X$) = 0 is violated, $E(\varepsilon_i X)$ may be equal to zero, so that β are the parameters of the BLP. In such a case as we will see later in the course $\hat{\beta}_{OLS}$ will be consistent for β. A common remedy for violation of GM2 is instrumental variables estimator that we will discuss later in the course. The idea is to find instrumental variables Z, such that $E(Z'\varepsilon) = 0$ and E(Z'X) is invertible. Then since $Y = X\beta + \varepsilon$ we have

$$\begin{aligned} Z'Y &= Z'X\beta + Z'\varepsilon \\ E\left(Z'Y\right) &= E\left(Z'X\right)\beta \\ \beta &= E\left(Z'X\right)^{-1}E\left(Z'Y\right) \end{aligned}$$

and we can use analogy principle to define

$$\hat{\beta}_{IV} = \left(Z'X\right)^{-1}\left(Z'Y\right)$$

Violations of the GM theorem: GM3

Remember $\text{Var}(\varepsilon|X) = E(\varepsilon\varepsilon'|X) - E(\varepsilon|X)E(\varepsilon|X)' = E(\varepsilon\varepsilon'|X)$ if GM2 holds.

$$\begin{aligned} E(\varepsilon\varepsilon'|X) &= \begin{pmatrix} E(\varepsilon_1^2|X) & E(\varepsilon_1\varepsilon_2|X) & ... & E(\varepsilon_1\varepsilon_N|X) \\ E(\varepsilon_2\varepsilon_1|X) & E(\varepsilon_2^2|X) & ... & E(\varepsilon_2\varepsilon_N|X) \\ : & : & : & : \\ E(\varepsilon_N\varepsilon_1|X) & E(\varepsilon_N\varepsilon_2|X) & ... & E(\varepsilon_N^2|X) \end{pmatrix} \\ &= \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix} \end{aligned}$$

GM3 can be separated in two parts, either of which can be violated:

1) Homoskedasticity. $\text{Var}(\varepsilon_i|X) = \sigma^2$ for any i.

2) No Serial Correlation. $E(\varepsilon_i\varepsilon_j|X) = 0$ for any $i \neq j$.

Violating GM3 means either

1) Heteroskedasticity. Not all diagonal elements of the conditional variance matrix are equal.

or

2) Serial correlation. The conditional variance matrix is not diagonal.

In either of these two cases, $\hat{\beta}_{OLS}$ is unbiased, but its variance is not minimal.

$$Var(\hat{\beta}_{OLS}|X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

where Ω is equal to $\text{Var}(\varepsilon|X)$. A Generalized Least Square estimate would be more efficient in this case.

Estimating σ^2.

So far our focus has been on estimation of β. Next we turn to estimation of the other parameter in the GM assumptions, σ^2. Estimating σ^2 is important for obtaining an estimate of

$$Var\left(\hat{\beta}_{OLS}|X\right) = \sigma^2 \left(X'X\right)^{-1}.$$

First note, that $\sigma^2 = E\left(\varepsilon_i^2|X\right)$. Since obviously $\sigma^2 = E\sigma^2$, we have

$$\sigma^2 = E\left(E\left(\varepsilon_i^2|X\right)\right) = E\left(\varepsilon_i^2\right).$$

The analogy principle suggests estimating $E\left(\varepsilon_i^2\right)$ by $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$. Unfortunately, because β is unknown, $\varepsilon_i = y_i - X_i\beta$ is unknown too. We can approximate ε_i with $\hat{\varepsilon}_i$ using our estimate $\hat{\beta} \equiv \hat{\beta}_{OLS}$

$$\hat{\varepsilon}_i = y_i - X_i\hat{\beta}.$$

Thus, σ^2 could be estimated as

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Is $\tilde{\sigma}^2$ unbiased? No, but it is easy to modify $\tilde{\sigma}^2$ to turn it into an unbiased estimator.

First, let us express the vector of residuals $\hat{\varepsilon}$ in terms of the vector of errors ε.

$$\hat{\varepsilon} = Y - X\hat{\beta} = Y - X\left(X'X\right)^{-1}X'Y = \left(I - X\left(X'X\right)^{-1}X'\right)Y = M_X Y$$

Matrix M_X is known as the residual maker matrix, and has the following properties.

Symmetric: $M_X' = M_X$

Idempotent: $M_X M_X = M_X$

Such matrices has a geometric interpretation of the projector operators. Multiplying a vector from the left by M_X results in the projection of the vector on the subspace spanned by all vectors orthogonal to the columns of matrix X. Note that $P_X = X\left(X'X\right)^{-1}X'$ is also a projector (this time on the subspace spanned by the columns of X)

We note that $M_X X = 0$ and $M_X Y = M_X \varepsilon$. Therefore,

$$\hat{\varepsilon} = M_X Y = M_X \varepsilon.$$

Now, let us compute $E\left(\tilde{\sigma}^2|X\right)$. We have

$$\begin{aligned} E\left(\hat{\varepsilon}'\hat{\varepsilon}|X\right) &= E\left(\varepsilon' M_X' M_X \varepsilon|X\right) = E\left(\varepsilon' M_X \varepsilon|X\right) = E\left(\operatorname{trace}\left(\varepsilon' M_X \varepsilon\right)|X\right) \\ &= E\left(\operatorname{trace}\left(M_X \varepsilon\varepsilon'\right)|X\right) = \operatorname{trace} E\left(\left(M_X \varepsilon\varepsilon'\right)|X\right) \\ &= \operatorname{trace}\left[M_X E\left(\varepsilon\varepsilon'|X\right)\right] = \operatorname{trace}\left[M_X \sigma^2 I\right] = \sigma^2 \operatorname{trace} M_X. \end{aligned}$$

What is $\operatorname{trace} M_X$?

$$\begin{aligned} \operatorname{tr} M_X &= \operatorname{tr}\left(I - X\left(X'X\right)^{-1}X'\right) \\ &= \operatorname{tr} I - \operatorname{tr}\left(X\left(X'X\right)^{-1}X'\right) \\ &= n - \operatorname{tr}\left(\left(X'X\right)^{-1}X'X\right) \\ &= n - \operatorname{tr}\left(I_k\right) = n - k \end{aligned}$$

Thus,

$$E\left(\tilde{\sigma}^2|X\right) = \frac{n-k}{n}\sigma^2$$

leaving the following as an unbiased estimator of σ^2 :

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}.$$

Partitioned regression

Suppose we wish to partition the independent variables, X, from our model into X_1 and X_2.

$$X = (X_1, X_2)$$

Note that X_1 is $n \times k_1$, X_2 is $n \times k_2$, and $k_1 + k_2 = k$. Partition β conformably as

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Thus we have

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Recall

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1} X'Y$$

and note

$$X'\hat{\varepsilon} = X'Y - X'X\hat{\beta} = 0.$$

Thus the residual is orthogonal to X, which corresponds to the population property $E(X_i\varepsilon_i) = 0$.

In the partitioned model, we have the following:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix}.$$

This yields two equations in two unknowns:

$$(X_1'X_1)\hat{\beta}_1 + (X_1'X_2)\hat{\beta}_2 = X_1'Y \quad (1)$$
$$(X_2'X_1)\hat{\beta}_1 + (X_2'X_2)\hat{\beta}_2 = X_2'Y \quad (2)$$

$$\hat{\beta}_1 = (X_1'X_1)^{-1}\left(X_1'Y - (X_1'X_2)\hat{\beta}_2\right)$$

Note that this is just as if $\hat{\beta}_1$ were used to estimate the following:

$$\underbrace{Y - X_2\hat{\beta}_2}_{\tilde{Y}} = X_1\beta_1 + \varepsilon$$

If we know $\hat{\beta}_2$, we can just subtract out its effect (via X_2) on Y, and then regress this new variable, $\tilde{Y}$, on X_1.

Substituting the latter expression for $\hat{\beta}_1$ into (2), we obtain

$$\begin{aligned}
&(X_2'X_1)(X_1'X_1)^{-1}\left(X_1'Y - (X_1'X_2)\hat{\beta}_2\right) + (X_2'X_2)\hat{\beta}_2 = X_2'Y \\
&\left((X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)\right)\hat{\beta}_2 = \left(X_2' - (X_2'X_1)(X_1'X_1)^{-1}X_1'\right)Y \\
&X_2'\left(I - X_1(X_1'X_1)^{-1}X_1'\right)X_2\hat{\beta}_2 = X_2'\left(I - X_1(X_1'X_1)^{-1}X_1'\right)Y
\end{aligned}$$

Recalling the definition of the residual maker matrix, M_X, we define M_1 as the residual maker matrix for X_1 :

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

Applying the residual maker matrix to another matrix yields the residual:

$$M_1Y = Y - X_1 \underbrace{(X_1'X_1)^{-1}X_1'Y}_{\text{LS coefficient of } Y \text{ on } X_1} = Y - X_1\tilde{\gamma}.$$

Naturally, the residual of X_1 regressed on X_1 should be 0 :

$$M_1X_1 = X_1 - X_1(X_1'X_1)^{-1}X_1'X_1 = X_1 - X_1 = 0.$$

Substitute in M_1 as follows

$$X_2'M_1X_2\hat{\beta}_2 = X_2'M_1Y$$

We have

$$\begin{aligned}
\hat{\beta}_2 &= (X_2'M_1X_2)^{-1}X_2'M_1Y \\
&= (X_2'M_1'M_1X_2)^{-1}X_2'M_1'M_1Y \\
&= \left[(M_1X_2)'(M_1X_2)\right]^{-1}(M_1X_2)'(M_1Y)
\end{aligned}$$

This result for $\hat{\beta}_2$ is as if we ran a regression of M_1Y on M_1X_2. Similarly,

$$\hat{\beta}_1 = \left[(M_2X_1)'(M_2X_1)\right]^{-1}(M_2X_1)'(M_2Y)$$

We could estimate $\hat{\beta}_1$ alone by the following procedure:

Step 1:

Regress Y on X_2 and form residual, which captures that portion of Y not correlated with X_2 :

$$\hat{e} = Y - X_2 \left(X_2'X_2\right)^{-1} X_2'Y = M_2 Y$$

Regress X_1 on X_2 and form residual, which captures that portion of X_1 not correlated with X_2

$$\hat{v} = X_1 - X_2 \left(X_2'X_2\right)^{-1} X_2'X_1 = M_2 X_1$$

Step 2:

Regress residuals $\hat{e}$ on $\hat{v}$, or equivalently, M_2Y on M_2X_1.

This coefficient measures the "effect" of X_1 on Y after controlling for X_2

$$\hat{\beta}_1 = \left[\left(M_2X_1\right)'\left(M_2X_1\right)\right]^{-1}\left(M_2X_1\right)'\left(M_2Y\right) = \left(\hat{v}'\hat{v}\right)^{-1}\hat{v}'\hat{e}.$$

The residual from this regression is $\tilde{\varepsilon} = \hat{e} - \hat{v}\hat{\beta}_1$

Standard errors in partitioned regressions

Let us calculate the variances for a partitioned regression coefficients, $Var\left(\hat{\beta}_j|X\right)$.

We have

$$\begin{aligned}
Var\left(\hat{\beta}_1|X\right) &= Var\left(\left(X_1'M_2X_1\right)^{-1}X_1'M_2Y|X\right) \\
&= \left(X_1'M_2X_1\right)^{-1}X_1'M_2Var\left(Y|X\right)M_2X_1\left(X_1'M_2X_1\right)^{-1} \\
&= \sigma^2\left(X_1'M_2X_1\right)^{-1}X_1'M_2M_2X_1\left(X_1'M_2X_1\right)^{-1} \\
&= \sigma^2\left(X_1'M_2X_1\right)^{-1}X_1'M_2X_1\left(X_1'M_2X_1\right)^{-1} \\
&= \sigma^2\left(X_1'M_2X_1\right)^{-1}.
\end{aligned}$$

Similarly, $Var\left(\hat{\beta}_2|X\right) = \sigma^2\left(X_2'M_1X_2\right)^{-1}$.

This sheds light on the problems with multicollinearity that we discussed above. If X_1 and X_2 are "almost" collinear, projection of X_1 on the spaces orthogonal to X_2 must be very small. Hence, we would have $X_1'M_2X_1$ close to zero and $\left(X_1'M_2X_1\right)^{-2}$ very large, making $Var\left(\hat{\beta}_1|X\right)$ also very large.

Now compare regression residuals. If the original regression were carried out,

we would have the following residual:

$$\hat{\varepsilon} = Y - X\hat{\beta} = Y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2$$

From Step 2 of the partitioned regression, we have the following residuals

$$\tilde{\varepsilon} = M_2\left(Y - X_1\hat{\beta}_1\right).$$

Note that $\tilde{\varepsilon} = \hat{\varepsilon}$.

$$\begin{aligned}\tilde{\varepsilon} &= M_2\left(Y - X_1\hat{\beta}_1\right) = M_2\left(Y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\right) \\ &= M_2\hat{\varepsilon} = \hat{\varepsilon}.\end{aligned}$$

Thus the variance estimate by, say, STATA from running Step 2 of the partitioned regression would only be off by a degree of freedom factor:

$$\begin{aligned}\widehat{Var}\hat{\beta}_1 &= \hat{\sigma}^2\left(X_1'M_2X_1\right)^{-1} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}\left(X_1'M_2X_1\right)^{-1} \\ &= \frac{n-k_1}{n-k}\frac{\tilde{\varepsilon}'\tilde{\varepsilon}}{n-k_1}\left(X_1'M_2X_1\right)^{-1} = \frac{n-k_1}{n-k}\widetilde{Var}\hat{\beta}_1\end{aligned}$$

The fact that $\tilde{\varepsilon} = \hat{\varepsilon}$ together with the formula $\hat{\beta}_1 = \left[\left(M_2X_1\right)'\left(M_2X_1\right)\right]^{-1}\left(M_2X_1\right)'\left(M_2Y\right)$ is known as the **Frisch-Waugh Theorem**.

Maximum likelihood estimation

We now consider an alternative method for estimating β : maximum likelihood estimation. Recall a typical maximum likelihood setup:

$$\begin{aligned}z_1, ..., z_n \overset{i.i.d.}{\sim} f\left(\cdot|\theta\right) \quad &\rightarrow \quad \mathcal{L} = \prod_{i=1}^{n} f\left(z_i|\theta\right) \\ L &= \sum_{i=1}^{n} \ln f\left(z_i|\theta\right) \\ \hat{\theta}_{ML} &= \arg\max_{\theta} L\end{aligned}$$

Least squares provides an estimator of the best linear predictor, *without restricting the form of the underlying distribution of the data.* In this sense, OLS is a semi-parametric estimator. (We call a model and an associated estimator parametric if the model completely specifies the distribution of the data up to a finite set

of unknown parameters). Now, we add another GM condition that specifies the (conditional) distribution of the data.

(GM1) $\text{rank}(X) = k$

(GM2) $E(Y|X) = X\beta$ (or alternatively) $E(\varepsilon|X) = 0$

(GM3) $Var(Y|X) = \sigma^2 I = Var(\varepsilon|X)$

(GM4) $Y|X \sim N(X\beta, \sigma^2 I)$

Note that (GM4) implies (GM2) and (GM3), and adds the (conditional) normality assumption to $Y|X$ as an additional condition.

$$\begin{aligned} f\left(Y_1, ..., Y_n|X, \beta, \sigma^2\right) \quad &: \quad \mathcal{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2}} \\ L \quad &= \quad -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{\sum_{i=1}^{n}(Y_i - X_i'\beta)^2}{2\sigma^2} \end{aligned}$$

Hence, the FOCs are:

$$\begin{aligned} \frac{\partial L}{\partial \beta} \quad &= \quad -\frac{\sum_{i=1}^{n}(Y_i - X_i'\beta)(-X_i)}{\sigma^2} = 0 \\ \frac{\partial L}{\partial \sigma^2} \quad &= \quad -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(Y_i - X_i'\beta)^2}{2\sigma^4} = 0 \end{aligned}$$

The solutions are

$$\hat{\beta}_{ML} = (X'X)^{-1}X'Y = \hat{\beta}_{OLS}$$

and

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\hat{\beta}_{ML}\right)^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$

Thus, $\hat{\beta}_{OLS}$ is actually the maximum likelihood estimator (MLE) for β, so it has the desirable consistency, asymptotic normality and asymptotic efficiency properties of MLE. The maximum likelihood variance estimator has a different form than our earlier unbiased variance estimator.

Cramér-Rao Lower Bound

Consider a formal statistical problem of estimating a parameter vector θ if you have a vector of data Z with the joint density of its elements given by $f(z|\theta)$. Suppose that $\hat{\theta}$ is an unbiased estimator of θ. Then we have the following Cramér-

Rao lower bound on its variance

$$Var\left(\hat{\theta}\right) \geq \mathcal{I}\left(\theta\right)^{-1}$$

in the sense that the difference between the left and the right hand sides is a positive semi-definite matrix. Here $\mathcal{I}\left(\theta\right) = Var\left(\frac{d}{d\theta}\log\left(f\left(Z|\theta\right)\right)\right)$ is the Fisher information.

Proof (non-examinable): Define

$$P = E\left(\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)'\right),$$

$$Q = E\left(\left(\hat{\theta}-\theta\right)\frac{d}{d\theta'}\log\left(f\left(Z|\theta\right)\right)\right),$$

and

$$R = E\left(\frac{d}{d\theta}\log\left(f\left(Z|\theta\right)\right)\frac{d}{d\theta'}\log\left(f\left(Z|\theta\right)\right)\right).$$

Note that $R = \mathcal{I}\left(\theta\right)$. Indeed, by definition of the variance

$$Var\left(\frac{d}{d\theta}\log\left(f\left(Z|\theta\right)\right)\right) = R - E\left(\frac{d}{d\theta}\log\left(f\left(Z|\theta\right)\right)\right)E\left(\frac{d}{d\theta}\log\left(f\left(Z|\theta\right)\right)\right)'$$

On the other hand,

$$\begin{aligned} E\left(\frac{d}{d\theta}\log\left(f\left(Z|\theta\right)\right)\right) &= \int\frac{d}{d\theta}\log\left(f\left(z|\theta\right)\right)f\left(z|\theta\right)dz = \int\frac{\frac{d}{d\theta}f\left(z|\theta\right)}{f\left(z|\theta\right)}f\left(z|\theta\right)dz \\ &= \int\frac{d}{d\theta}f\left(z|\theta\right)dz = \frac{d}{d\theta}\int f\left(z|\theta\right)dz = \frac{d}{d\theta}1 = 0. \end{aligned}$$

In particular,

$$\left[\begin{array}{cc} P & Q \\ Q' & R \end{array}\right] \geq 0$$

because the left hand side is a variance-covariance matrix. Assuming that $\mathcal{I}\left(\theta\right)$ is non-degenerate (positive definite), we can premultiply both sides of the above inequality by $[I, -QR^{-1}]$, and postmultiply by $[I, -QR^{-1}]'$ to obtain

$$P - QR^{-1}Q' \geq 0.$$

On the other hand,

$$\begin{aligned} Q &= E\left(\left(\hat{\theta}-\theta\right)\frac{d}{d\theta'}\log\left(f\left(Z|\theta\right)\right)\right) = E\left(\hat{\theta}\frac{d}{d\theta'}\log\left(f\left(Z|\theta\right)\right)\right) \\ &= \int \hat{\theta}\frac{d}{d\theta'}f\left(z|\theta\right)dz = \frac{d}{d\theta'}\int \hat{\theta}f\left(z|\theta\right)dz = \frac{d}{d\theta'}E\hat{\theta} = \frac{d}{d\theta'}\theta = I \end{aligned}$$

We have $P - R^{-1} \geq 0$ so that

$$P \geq R^{-1} = \mathcal{I}\left(\theta\right)^{-1}.\square$$

Remark 3 *Note that above we interchanged the order of differentiation and integration freely. For such an interchange to be valid, we need to impose some technical regulatory conditions on the density* $f\left(z|\theta\right)$*. For details, see Amemiya (1985, ch. 1.3.2).*

Computing the Fisher information can be facilitated by using the following **information equality**

$$Var\left(\frac{d}{d\theta}\log\left(f\left(Z;\theta\right)\right)\right) = -E\left(\frac{d^2}{d\theta d\theta'}\log\left(f\left(Z;\theta\right)\right)\right),$$

which can be proven as follows.

$$\begin{aligned} E\left(\frac{d^2}{d\theta d\theta'}\log\left(f\left(Z;\theta\right)\right)\right) &= \int \frac{d}{d\theta}\left(\frac{df\left(z;\theta\right)/d\theta'}{f\left(z;\theta\right)}\right)f\left(z;\theta\right)dz \\ &= \int\left(\frac{d^2f\left(z;\theta\right)/d\theta d\theta'}{f\left(z;\theta\right)} - \frac{\left(df\left(z;\theta\right)/d\theta\right)\left(df\left(z;\theta\right)/d\theta'\right)}{f^2\left(z;\theta\right)}\right)f\left(z;\theta\right)dz \\ &= \frac{d^2}{d\theta d\theta'}\int f\left(z;\theta\right)dz - \int \frac{d\log\left(f\left(z;\theta\right)\right)}{d\theta}\frac{d\log\left(f\left(z;\theta\right)\right)}{d\theta'}f\left(z;\theta\right)dz \\ &= -E\left(\frac{d\log\left(f\left(Z;\theta\right)\right)}{d\theta}\frac{d\log\left(f\left(Z;\theta\right)\right)}{d\theta'}\right) \\ &= -Var\left(\frac{d}{d\theta}\log\left(f\left(Z;\theta\right)\right)\right). \end{aligned}$$

Notation 4 $\frac{d^2}{d\theta d\theta'}\log\left(f\left(Z;\theta\right)\right)$ *is called the Hessian, and denoted as* H*.* $\frac{d}{d\theta}\log\left(f\left(Z|\theta\right)\right)$ *is called the score, and denoted as* S*.*

Cramér-Rao Lower Bound in the regression context

We will apply the CRLB conditionally on X. Derive expected Hessian $E(H)$:

$$E(H) = \begin{pmatrix} E\left(\frac{\partial^2 L}{\partial\beta\partial\beta'}|X\right) & E\left(\frac{\partial^2 L}{\partial\beta\partial\sigma^2}|X\right) \\ E\left(\frac{\partial^2 L}{\partial\sigma^2\partial\beta'}|X\right) & E\left(\frac{\partial^2 L}{\partial\sigma^2\partial\sigma^2}|X\right) \end{pmatrix}$$

Recall that

$$L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2}{2\sigma^2}$$

We have

$$\begin{aligned}
\frac{\partial^2 L}{\partial\beta\partial\beta'} &= \frac{\partial}{\partial\beta'}\frac{\sum_{i=1}^{n} X_i\left(Y_i - X_i'\beta\right)}{\sigma^2} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i X_i' = -\frac{1}{\sigma^2}X'X, \\
\frac{\partial^2 L}{\partial\beta\partial\sigma^2} &= \frac{\partial}{\partial\sigma^2}\frac{\sum_{i=1}^{n} X_i\left(Y_i - X_i'\beta\right)}{\sigma^2} = -\frac{\sum_{i=1}^{n} X_i\left(Y_i - X_i'\beta\right)}{\sigma^4} = -\frac{1}{\sigma^4}X'\left(Y - X\beta\right) \\
\frac{\partial^2 L}{\partial\sigma^2\partial\sigma^2} &= \frac{n}{2}\frac{1}{\sigma^4} - \frac{\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2}{\sigma^6} = \frac{n}{2}\frac{1}{\sigma^4} - \frac{1}{\sigma^6}\left(Y - X\beta\right)'\left(Y - X\beta\right).
\end{aligned}$$

Therefore,

$$E(H) = \begin{pmatrix} -\frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{n}{2}\frac{1}{\sigma^4} - \frac{n\sigma^2}{\sigma^6} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma^2}X'X & 0 \\ 0 & -\frac{n}{2}\frac{1}{\sigma^4} \end{pmatrix}$$

Hence,

$$\mathcal{I}(\theta)^{-1} = \begin{pmatrix} \sigma^2\left(X'X\right)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}$$

The variance of $\hat{\beta}_{OLS} = \hat{\beta}_{ML}$ meets the CRLB. Thus, we have the following theorem

Theorem 5 *Under GM 1-2-3-4, OLS is Best Unbiased Estimator (BUE).*

Remark 6 *GM4 allows us to compare the OLS estimator with all unbiased estimators, not only linear ones. For example, any unbiased estimator that would treat "outliers" differently from other observations cannot be better than OLS under GM1,2,3,4.*

References

[1] Amemiya, T. (1985) *Advanced Econometrics*, Basil Blackwell.

[2] Hayashi, F. (2000) *Econometrics*, Princeton University Press.

[3] Hoerl, A.E. and Kennard, R.W. (1970) "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics* 12, 55-67

[4] James, W. and Stein, C. (1971) "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1, 361–379

[5] Leamer, E. E. (1983) "Model Choice and Specification Analysis," ch.5 in Griliches and Intriligator (eds) Handbook of Econometrics, Vol. 1, 285-325.

[6] Tibshirani, R. (1996) "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1 pp. 267-288.