Applied Econometrics - 8205/8206 Matrix Algebra

Sergio Ocampo Díaz

Contents

1 Petrin's Notes

1.1 Matrix Multiplication, Transpose and Inverse

Let A be dimension $n \times k$ and B be dimension $k \times m$. Then the product matrix, $C = AB$ is well defined and has elements,

$$c_{ij} = \sum_{s=1}^{k} a_{is} b_{sj},$$

which is the inner product of the ith row of A with the jth column of B.

Proposition: Multiplication is distributive: $A(B + C) = AB + AC$

Proof: The element of $D = A(B + C)$ corresponding to the i^{th} row and j^{th} column is:

$$d_{ij} = \sum_{s=1}^{k} a_{is} \left(b_{sj} + c_{sj}\right) = \sum_{s=1}^{k} a_{is} b_{sj} + \sum_{s=1}^{k} a_{is} c_{sj} = (AB)_{ij} + (AC)_{ij}$$

Proposition: Transpose over matrix product: $(AB)^{'} = B^{'} A^{'}$

Proof:

$$(AB)^{'}_{ij} = (AB)_{ji} = \sum_{s=1}^{k} a_{js} b_{si} = \sum_{s=1}^{k} b_{si} a_{js} = \sum_{s=1}^{k} b^{'}_{is} a^{'}_{sj} = \left(B^{'} A^{'}\right)_{ij}$$

This can be extended to cases where there are more than two matrices using the associative property of matrix multiplication:

$$(ABC)^{'} = ((AB)\, C)^{'} = C^{'} (AB)^{'} = C^{'} B^{'} A^{'}$$

Proposition: Transpose of an inverse is the inverse of the transpose: $\left(A^{-1}\right)^{'} = \left(A^{'}\right)^{-1}$

Proof: By definition of inverse:

$$\begin{aligned} A^{-1} A &= I \\ \left(A^{-1} A\right)^{'} &= I \\ A^{'} \left(A^{-1}\right)^{'} &= I \end{aligned}$$

Then, again by definition of inverse:

$$\left(A^{-1}\right)^{'} = \left(A^{'}\right)^{-1}$$

The last two propositions are particularly useful for computing the covariance matrix of the OLS estimator. The estimator is given by:

$$\hat{\beta} = \left(X^{'} X\right)^{-1} X^{'} y$$

Its (conditional) covariance is defined as:

$$\begin{aligned}
E_X\left[\hat{\beta}\hat{\beta}'\right] &= E_X\left[\left(X'X\right)^{-1}X'y\left(\left(X'X\right)^{-1}X'y\right)'\right] \\
&= E_X\left[\left(X'X\right)^{-1}X'y\left(X'y\right)'\left(\left(X'X\right)^{-1}\right)'\right] \\
&= E_X\left[\left(X'X\right)^{-1}X'yy'X\left(\left(X'X\right)'\right)^{-1}\right] \\
&= E_X\left[\left(X'X\right)^{-1}X'yy'X\left(X'X\right)^{-1}\right] \\
&= \left(X'X\right)^{-1}X'E_X\left[yy'\right]X\left(X'X\right)^{-1} \\
&= \left(X'X\right)^{-1}X'\left(\sigma^2 I\right)X\left(X'X\right)^{-1} \\
&= \sigma^2\left(X'X\right)^{-1}\left(X'X\right)\left(X'X\right)^{-1} \\
&= \sigma^2\left(X'X\right)^{-1}
\end{aligned}$$

1.2 Trace

Proposition: If A is $m \times n$ and B is $n \times m$, then $Tr(AB) = Tr(BA)$

Proof:

$$Tr(AB) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{ji}a_{ij} = Tr(BA)$$

It follows by repeated application of the above theorem that,

$$Tr(ABC) = Tr(CAB) = Tr(BCA)$$

A useful application of this theorem is,

$$Tr(X(X'X)^{-1}X') = Tr(X'X(X'X)^{-1}) = Tr(I_k) = k$$

where X is $n \times k$ having full column rank k.

1.3 Quadratic Forms and Eigenvalues

Let $\mathbf{A}$ be a real symmetric $k \times k$ matrix. A quadratic form is defined as,

$$q = \mathbf{b}'\mathbf{A}\mathbf{b}$$

where $q \in \Re$ and $\mathbf{b}$ is any non-null $k \times 1$ vector.

- $\mathbf{A}$ is positive definite if $q > 0$ $\forall$ non-null $\mathbf{b}$.
- $\mathbf{A}$ is positive semi-definite if $q \geq 0$ $\forall$ non-null $\mathbf{b}$.

Given a $k \times k$ matrix $\mathbf{A}$, the solutions to,

$$\mathbf{A}\mathbf{c} = \lambda\mathbf{c},$$

are the (scalar) eigenvalues (λ) and eigenvectors ($\mathbf{c}$). There are k (potentially non-unique) pairs $(\lambda_i, \mathbf{c_i})$. In matrix form the k solution pairs $(\lambda_i, \mathbf{c_i})$ are written as,

$$\mathbf{AC} = \mathbf{C\Lambda},$$

where the eigenvectors are the columns of $\mathbf{C}$ and the eigenvalues form the diagonal of $\mathbf{\Lambda}$ such that,

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

Since $\mathbf{A}(r\mathbf{c}) = \lambda(r\mathbf{c})\ \forall\ r \in \Re$, we normalize the eigenvectors $\mathbf{c}$such that $\mathbf{c'c} = 1$.

Proposition The eigenvalues of a real symmetric positive definite matrix $\mathbf{A}$ are all positive.

Proof

$$\begin{aligned} \mathbf{Ac} &= \lambda\mathbf{c}, \quad \forall(\lambda, \mathbf{c}) \\ \mathbf{c'Ac} &= \mathbf{c'}\lambda\mathbf{c} \\ &= \lambda\mathbf{c'c} \\ &= \lambda \end{aligned}$$

Since $\mathbf{b'Ab} > 0\ \forall$ non-null $\mathbf{b}$, $\lambda > 0$.

Proposition If $\mathbf{A}$ is symmetric and positive definite, a non-singular matrix $\mathbf{P}$ can be found s.t. $\mathbf{A} = \mathbf{PP'}$.

Proof From Spectral decomposition of $\mathbf{A}$, $\mathbf{A} = \mathbf{C\Lambda C'}$. Let $\mathbf{P} = \mathbf{C\Lambda^{\frac{1}{2}}}$(possible because the eigenvalues are all positive). Then,

$$\begin{aligned} \mathbf{A} &= \mathbf{C\Lambda C'} \\ &= \mathbf{C}(\mathbf{\Lambda^{1/2}\Lambda^{1/2}})\mathbf{C'} \\ &= (\mathbf{C\Lambda^{1/2}})(\mathbf{\Lambda^{1/2}C'}) \\ &= (\mathbf{C\Lambda^{1/2}})(\mathbf{C\Lambda^{1/2}})' \\ &= \mathbf{PP'} \end{aligned}$$

This is specially relevant when A is a covariance matrix, by definition it is symmetric and positive definite.

2 Magnus & Neudecker - Matrix Differentials

2.1 Differentials, Partial Derivatives, Jacobians, Gradients (pg93)

Let $f : S \to \mathbb{R}^m$ be a function defined on a set $S \in \mathbb{R}^n$. Let c be an interior point of S, and let $B(c, r)$ be an n-ball lying in S.

2.1.1 Differentials

Let $u \in \mathbb{R}^n$ with $\|u\| < r$, so that $c + u \in B(c, r)$. If there exists a real $m \times n$ matrix A, depending on c but not on u, such that:

$$f(c+u) = f(c) + A(c)\,u + r_c(u) \qquad \lim_{u \to 0} \frac{r_c(u)}{\|u\|} = 0$$

then the function f is differentiable at c. Matrix A is the first derivative of f at c, and the $m \times 1$ vector

$$df(c; u) = A(c)\,u$$

its (first) differential. Note that the differential is a linear function of u and maps to the same space as f.

If f is differentiable it can be approximated by an affine function[1] of u. Note that this can be done only at an interior point of an open set.

2.1.2 Partial Derivatives

Let $f_i : S \to \mathbb{R}$ $(i = 1, \ldots, m)$ be the i^{th} component function of f, and e_j the j^{th} unit vector in $\mathbb{R}^n$. Consider $t \in \mathbb{R}$ such that $c + te_j \in B(c, r)$ for all j.

Consider the limit:

$$D_j f_i(c) = \frac{\partial f_i}{\partial x_j}(c) = \lim_{t \to 0} \frac{f_i(c + te_j) - f_i(c)}{t}$$

When the limit exists it is called the j^{th} partial derivative of f_i.

The Jacobian matrix of f at c is defined as the matrix of partial derivatives, where each row corresponds to a component of the function $(i = 1, \ldots, m)$ and each row to a component of the argument $(j = 1, \ldots, n)$:

$$Df(c) = \begin{bmatrix} D_1 f_1(c) & D_2 f_1(c) & \cdots & D_n f_1(c) \\ D_1 f_2(c) & D_2 f_2(c) & \cdots & D_n f_2(c) \\ \vdots & & \ddots & \\ D_1 f_m(c) & & & D_n f_m(c) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & & \ddots & \\ \frac{\partial f_{2m}}{\partial x_1} & & & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

When a function f is differentiable the matrix A of the definition above is given by $A(c) = Df(c)$, so that $df(c; u) = (Df(c))\,u$.

The gradient ∇f is an $n \times m$ matrix defined as the transpose of the Jacobian:

$$\nabla f(c) = [Df(c)]'$$

When f is a real valued function the Jacobian is a $1 \times n$ vector and the gradient a $n \times 1$ vector.

2.1.3 The Chain Rule

Consider now a function $g : \mathbb{R}^m \to \mathbb{R}^k$ such that it is differentiable at a point $b = f(c)$, let its Jacobian be $Dg(b)$ then:

$$D(g \circ f)(c) = (Dg(b))\,(Df(c))$$

So that the Jacobian of the new function is the product of the Jacobian of the original two.

[1] An affine function has the form $g(x) = a + bx$.

2.2 Scalar Functions of a Vector (pg200)

The two most important cases of a scalar function of an $n \times 1$ vector x are the linear form $\phi(x) = a^{'}x$, where a is an $n \times 1$ vector of constants, and the quadratic form $\psi(x) = x^{'}Ax$, where A is an $n \times n$ matrix of constants.

Fuction	Differential	Jacobian	Gradient
$a^{'}x$	$a^{'}dx$	$a^{'}$	a
$x^{'}Ax$	$x^{'}\left(A + A^{'}\right)dx$	$x^{'}\left(A^{'} + A\right)$	$\left(A + A^{'}\right)x$

For the differential of the quadratic function:

$$\begin{aligned} d\psi(x) &= dx^{'}Ax \\ &= (dx)^{'}Ax + x^{'}Adx \\ &= \left((dx)^{'}Ax\right)^{'} + x^{'}Adx \\ &= (Ax)^{'}dx + x^{'}Adx \\ &= x^{'}A^{'}dx + x^{'}Adx \\ &= x^{'}\left(A^{'} + A\right)dx \end{aligned}$$

Note that if A is symmetric then $d\psi(x) = 2x^{'}Adx$.

2.3 Examples

Chain rule:

$$\phi(x) = e^{x^{'}x}$$

$$d\phi(x) = de^{x^{'}x} = e^{x^{'}x}d\left(x^{'}x\right) = e^{x^{'}x}\left((dx)^{'}x + x^{'}dx\right) = e^{x^{'}x}\left(x^{'}dx + x^{'}dx\right) = 2e^{x^{'}x}x^{'}dx$$

$$D\phi(x) = 2e^{x^{'}x}x^{'}$$

Note that $x^{'}dx = (dx)^{'}x$ since they are or order 1×1, hence symmetric.

OLS:

$$\phi(\beta) = (y - X\beta)^{'}(y - X\beta) = e^{'}e$$

$$d\phi(\beta) = de^{'}e = 2e^{'}de = 2(y - X\beta)^{'}d(y - X\beta) = -2(y - X\beta)^{'}Xd\beta$$

$$D\phi(\beta) = -2y^{'}X + 2\beta^{'}X^{'}X$$

$$\nabla\phi(\beta) = -2X^{'}y + 2X^{'}X\beta$$

3 Ruud (Chapter 2)

Consider a system of variables where one variable y is explained by k variables $x = (x_1, \ldots, x_k)^{'}$. There are n (ordered) observations of the variables:

$$\begin{matrix} y_1 & x_{1,1} & \ldots & x_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{1,n} & \ldots & x_{k,n} \end{matrix}$$

Observations of the variables are contained in a $n \times 1$ vector y and a $n \times k$ matrix X, so that the observations are $[y|X]$.

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{1,1} & \ldots & x_{k,1} \\ \vdots & \ddots & \vdots \\ x_{1,n} & \ldots & x_{k,n} \end{bmatrix} = \begin{bmatrix} x_1^{'} \\ \vdots \\ x_n^{'} \end{bmatrix}$$

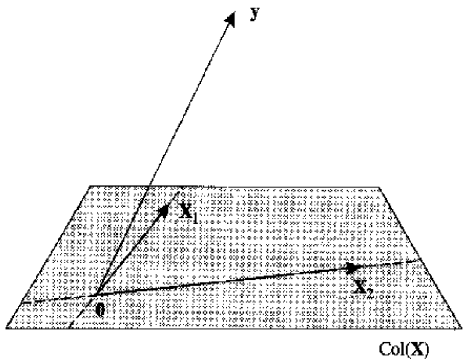


Figure 2.2 Vector representation of data.

The way in which y is explained by x is assumed to be linear. The relation is then for a given observation i:

$$y_i = x_{1,i}\beta_{1,i} + \ldots x_{k,i}\beta_{k,i} + \epsilon_i = x_i^{'}\beta_i + \epsilon_i$$

It is furthermore assumed that the relation between x and y does not depend in the observation: $\forall_{i,j}\beta_i = \beta_j = \beta$. Then it is possible to write:

$$y = X\beta + \epsilon$$

where β is a $k \times 1$ vector and $\epsilon = (\epsilon_1, \ldots \epsilon_n)^{'}$ is a $n \times 1$ vector or residuals.

The fit of the relation between y and x depends on β and is measured by how "small" ϵ is. If ϵ where to be equal to zero ($\forall_i \epsilon_i = 0$) then $y = X\beta$, this would be the best possible fit. In that case each observation of y would be a linear combination of the variables in the corresponding observation of x. This fact will be used below.

Even though the system has $k + 1$ variables, given the sample, they all belong to the same space. All the variables can be understood as vectors of the observation space $\mathbb{R}^n$, each variable is just an ordered tuple of n observations, hence all variables belong to a space with the same dimension (n). Since $\mathbb{R}^n$ is a vector space scalar multiplication and addition are well defined between its elements. The linear relation between y and x can then be understood as:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \beta_1 \begin{bmatrix} x_{1,1} \\ \vdots \\ x_{1,n} \end{bmatrix} + \ldots + \beta_k \begin{bmatrix} x_{k,1} \\ \vdots \\ x_{k,n} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Or more compactly:

$$y = \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

where $y, x_1, \ldots, x_k, \epsilon \in \mathbb{R}^n$.

It becomes clear that x can perfectly explain any vector (of observations) that has the form $X\gamma$ for arbitrary $\gamma \in \mathbb{R}^k$. That is, any vector that belongs to the subspace (of $\mathbb{R}^n$) formed by linear combinations of the variables $x_1, \ldots, x_k \in \mathbb{R}^n$, since these variables are the columns of matrix X that subspace is called the column space of X, or $\text{col}(X)$. Formally:

$$\text{col}(X) = \left\{ z \in \mathbb{R}^n | \exists_{\gamma \in \mathbb{R}^k} z = X\gamma \right\}$$

The procedure of varying β only allows movement across the elements of the subspace $\text{col}(X)$ and not in whole space $\mathbb{R}^n$. When $y \in \text{col}(X)$ there exists $\beta \in \mathbb{R}^k$ such that $y = X\beta$, the relation is (by definition) perfectly linear and the best fit is achieved.

When $y \notin \text{col}(X)$ it is not possible to get a perfect fit ($\epsilon \neq 0$), but it is possible to choose $\beta \in \mathbb{R}^k$ as to find an element in $\text{col}(X)$ that is "close" to y. Closeness in $\mathbb{R}^n$ can be captured by the Euclidean norm, so the problem is to find a β that minimizes the distance between y and $X\beta$ as measured by the norm $\|z\| = \sqrt{z'z} = \sqrt{\sum_{i=1}^{n} z_i^2}$. Since the argmin of the problem is invariant to monotone increasing transformation of the objective function this problem is equivalent to:

$$\min_{\beta \in \mathbb{R}^k} \|y - X\beta\|^2 = \min_{\beta \in \mathbb{R}^k} (y - X\beta)' (y - X\beta) = \min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \left(y_i - x_i'\beta \right)^2$$

Or in terms of the "error" ϵ:

$$\min_{\beta \in \mathbb{R}^k} \|\epsilon\|^2 = \min_{\beta \in \mathbb{R}^k} \epsilon'\epsilon = \min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \epsilon^2$$

That is, finding β such that it minimizes the sum of square residuals.

Since varying β in $\mathbb{R}^k$ is equivalent to moving in the column space of X then the problem can be also stated as:

$$\min_{\mu \in \text{col}(X)} \|y - \mu\|^2$$

This is a known problem, the solution to it is given by $\hat{\mu}$ which is called the orthogonal projection of y onto X. The minimum distance between y and its projection $\hat{\mu}$ is obtained when their difference $(y - \hat{\mu})$ is orthogonal to the column space of X.

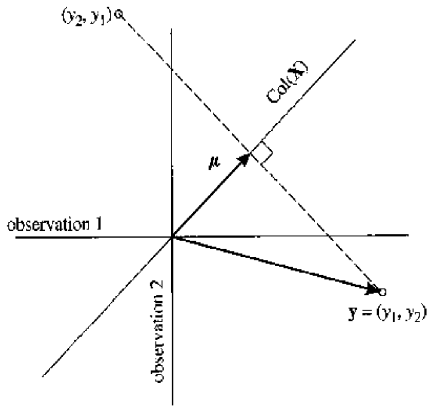


Figure 2.5 Ordinary least-squares projection in two dimensions.

When X has full column rank the projection can be obtained by means of the projector matrix:

$$P_X = X\left(X^{'}X\right)^{-1}X^{'}$$

The solution to the above problem is given by:

$$\hat{\mu} = P_X y \qquad X\hat{\beta} = P_X y \qquad X\hat{\beta} = X\left(X^{'}X\right)^{-1}X^{'}y \qquad X^{'}X\hat{\beta} = \left(X^{'}X\right)\left(X^{'}X\right)^{-1}X^{'}y \qquad X^{'}X\hat{\beta} = X^{'}y$$

$$\hat{\beta} = \left(X^{'}X\right)^{-1}X^{'}y$$

Note that P_X projects vector in the column space of X to themselves and vectors that are in $\text{col}\,(X)^{\perp}$, the orthogonal complement of $\text{col}\,(X)$, to zero:

- Let $z \in \text{col}\,(X)$ then there exists $\gamma \in \mathbb{R}^k$ such that $z = X\gamma$, moreover:

$$P_X z = X\left(X^{'}X\right)^{-1}X^{'}z = X\left(X^{'}X\right)^{-1}X^{'}X\gamma = X\gamma = z$$

- Let $z \in \text{col}\,(X)^{\perp} = \left\{z \in \mathbb{R}^n | X^{'}z = 0\right\}$, then:

$$P_X z = X\left(X^{'}X\right)^{-1}X^{'}z = X\left(X^{'}X\right)^{-1}0_{k\times 1} = 0_{n\times 1}$$

The projector matrix can also be used to project vectors to the orthogonal complement of a space by means of the annihilator matrix:

$$M_X = I - P_X$$

In this way $M_X z \in \text{col}\,(X)^{\perp}$, moreover if $z \in \text{col}\,(X)$ then $M_X z = 0$ and if $z \in \text{col}\,(X)^{\perp}$ then $M_X z = z$. The residual (the part of y that is not explained by a linear combination of X) is then determined by the part of y that is orthogonal to the elements in X, hence:

$$e = M_X y = y - P_x y = y - X\hat{\beta}$$

If y is orthogonal to x the solution to the problem is to set $\hat{\beta} = 0$, and the more close is y to be a linear combination of the variables in X the lower the error, as shown before if $y \in \text{col}\,(X)$ then $e = M_X y = 0$.

When X does not have full column rank the dimension of the subspace $\text{col}\,(X)$ is less than k, there is at least one redundant column (that is contained in the subspace formed by the other columns). Then the orthogonal projection is still defined, but there is more than one way to express the projector matrix.

4 Wooldridge (Chapter 2 - Simple Linear Model)

Consider a random sample of variables $\{y_i, x_i\}_{i=1}^n$ that was originated following the linear relation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

under the restrictions that $E[\epsilon] = 0$ and $\text{Cov}[\epsilon, x] = E[\epsilon x] = 0$ (note that there are no is in the restrictions since they hold for the random variable and not for a particular realization). Given the relation between y and x the relation between the variables the restrictions can be written as:

$$\begin{aligned} E[y - \beta_0 - \beta_1 x] &= 0 \\ E[(y - \beta_0 - \beta_1 x) x] &= 0 \end{aligned}$$

The method of moments can be then used to estimate $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ such that the sample counterparts of the restrictions hold:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right] &= 0 \\ \frac{1}{n}\sum_{i=1}^n \left[\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i\right] &= 0 \end{aligned}$$

The same conditions are obtained from minimizing the sum of squared residuals.

From the first equation:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

where $\overline{y} = \sum y_i / n$ and $\overline{x} = \sum x_i / n$.

Replacing on the second equation:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \left[\left((y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x})\right) x_i\right] &= 0 \\ \sum_{i=1}^n [(y_i - \overline{y}) x_i] &= \hat{\beta}_1 \left[\sum_{i=1}^n (x_i - \overline{x}) x_i\right] \end{aligned}$$

Note that:

$$\begin{aligned} \sum_{i=1}^n (x_i - \overline{x})^2 &= \sum_{i=1}^n \left(x_i^2 - 2\overline{x} x_i + \overline{x}^2\right) \\ &= \sum_{i=1}^n \left(\left(x_i^2 - \overline{x} x_i\right) - \left(\overline{x} x_i - \overline{x}^2\right)\right) \\ &= \sum_{i=1}^n (x_i - \overline{x}) x_i - \left(\sum_{i=1}^n \overline{x} x_i - n\overline{x}^2\right) \\ &= \sum_{i=1}^n (x_i - \overline{x}) x_i - \left(\overline{x}\sum_{i=1}^n x_i - n\overline{x}^2\right) \\ &= \sum_{i=1}^n (x_i - \overline{x}) x_i \end{aligned}$$

and that in a similar way:

$$\begin{aligned}\sum_{i=1}^{n}\left[\left(y_i-\overline{y}\right)\left(x_i-\overline{x}\right)\right] &= \sum_{i=1}^{n}\left(y_i-\overline{y}\right)x_i-\sum_{i=1}^{n}\left(y_i-\overline{y}\right)\overline{x}\\ &= \sum_{i=1}^{n}\left(y_i-\overline{y}\right)x_i-\left[\overline{x}\sum_{i=1}^{n}y_i-n\overline{yx}\right]\\ &= \sum_{i=1}^{n}\left(y_i-\overline{y}\right)x_i\end{aligned}$$

Replacing in the original equation

$$\hat{\beta}_1=\frac{\sum\limits_{i=1}^{n}\left(y_i-\overline{y}\right)\left(x_i-\overline{x}\right)}{\sum\limits_{i=1}^{n}\left(x_i-\overline{x}\right)^2}=\frac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(y_i-\overline{y}\right)\left(x_i-\overline{x}\right)}{\frac{1}{n}\sum\limits_{i=1}^{n}\left(x_i-\overline{x}\right)^2}=\frac{\widehat{\mathrm{Cov}}\left(y,x\right)}{\widehat{\mathrm{Var}}\left(x\right)}$$

In way the coefficient on x is a measure of the correlation between x and y, it has the same sign and its level is adjusted by the variance of x. The more variance x has the lower the value (since a unit increase in x means less when x varies in the thousands).

From the matrix form of the problem one also has:

$$\begin{aligned}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} &= \left(\begin{bmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_n\end{bmatrix}\begin{bmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_n\end{bmatrix}\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}\\ &= \begin{pmatrix}n & n\overline{x}\\ n\overline{x} & \sum x_i^2\end{pmatrix}^{-1}\begin{bmatrix}n\overline{y}\\ \sum y_i x_i\end{bmatrix}\\ &= \frac{1}{n\sum x_i^2-n^2\overline{x}}\begin{bmatrix}\sum x_i^2 & -n\overline{x}\\ -n\overline{x} & n\end{bmatrix}\begin{bmatrix}n\overline{y}\\ \sum y_i x_i\end{bmatrix}\\ &= \frac{1}{(\sum x_i^2-n\overline{x})}\begin{bmatrix}\sum x_i^2\overline{y}-\overline{x}\sum y_i x_i\\ -n\overline{x}\overline{y}+\sum y_i x_i\end{bmatrix}\\ &= \frac{1}{\sum\left(x_i-\overline{x}\right)^2}\begin{bmatrix}\sum x_i^2\overline{y}-\overline{x}\sum y_i x_i\\ \sum y_i x_i-\sum x_i\overline{y}\end{bmatrix}\\ &= \frac{1}{\sum\left(x_i-\overline{x}\right)^2}\begin{bmatrix}\sum x_i^2\overline{y}-\overline{x}\sum y_i x_i\\ \sum\left(y_i-\overline{y}\right)x_i\end{bmatrix}\\ &= \frac{1}{\sum\left(x_i-\overline{x}\right)^2}\begin{bmatrix}\left[\sum\left(x_i-\overline{x}\right)^2\right]\overline{y}-\overline{x}\left(\sum\left(y_i-\overline{y}\right)x_i\right)\\ \sum\left(y_i-\overline{y}\right)x_i\end{bmatrix}\\ &= \begin{bmatrix}\overline{y}-\overline{x}\left(\frac{\sum(y_i-\overline{y})x_i}{\sum(x_i-\overline{x})^2}\right)\\ \frac{\sum(y_i-\overline{y})x_i}{\sum(x_i-\overline{x})^2}\end{bmatrix}\end{aligned}$$

Where the first element of the second to last step is obtained as follows:

$$\begin{aligned}
\sum x_i^2 \overline{y} - \overline{x} \sum y_i x_i &= \left[\sum \left(x_i - \overline{x}\right)^2\right] \overline{y} - \left[\sum \left(x_i - \overline{x}\right)^2\right] \overline{y} + \sum x_i^2 \overline{y} - \overline{x} \sum y_i x_i \\
&= \left[\sum \left(x_i - \overline{x}\right)^2\right] \overline{y} + \left(\sum x_i^2 - \left[\sum \left(x_i - \overline{x}\right) x_i\right]\right) \overline{y} - \overline{x} \sum y_i x_i \\
&= \left[\sum \left(x_i - \overline{x}\right)^2\right] \overline{y} + \overline{x}\overline{y} \sum x_i - \overline{x} \sum y_i x_i \\
&= \left[\sum \left(x_i - \overline{x}\right)^2\right] \overline{y} - \overline{x} \left(\sum y_i x_i - \overline{y} \sum x_i\right) \\
&= \left[\sum \left(x_i - \overline{x}\right)^2\right] \overline{y} - \overline{x} \left(\sum \left(y_i - \overline{y}\right) x_i\right)
\end{aligned}$$

5 FWL and Correlations

5.1 FWL with two explanatory variables

Consider a random sample of variables $\{y_i, x_i, z_i\}_{i=1}^{n}$ for $n = 200$ that was originated following the linear relation:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$

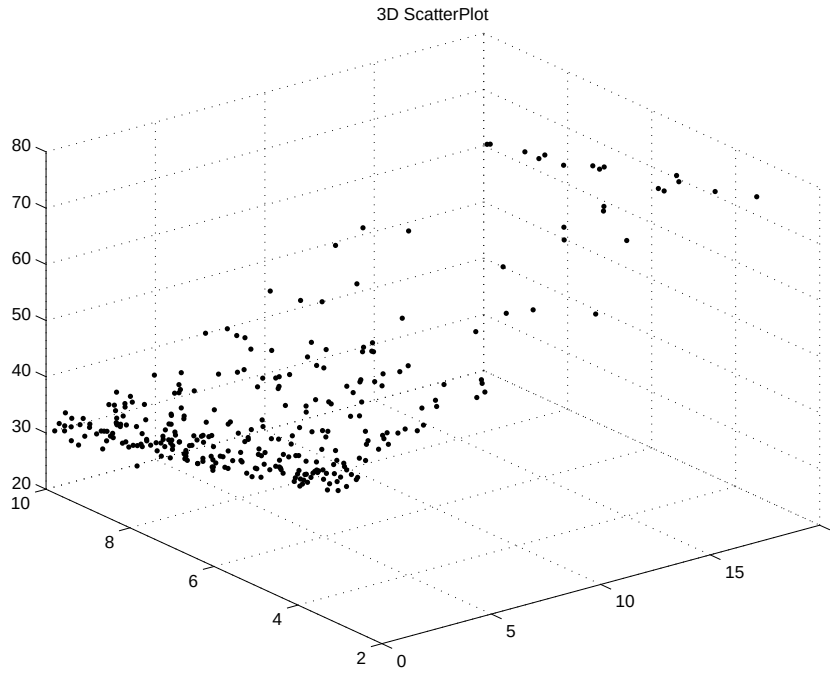
The idea is to present different ways of visualizing the data and obtaining the parameter values. There is an inherent difficulty in presenting data when there are multiple dimensions of it, and doing two dimensional cuts of the data does not reveal all the information needed to interpret the results. The use of FWL facilitates the analysis of the data.

Let $x \sim LN(0.5, 1.5)$, $z \sim U(3, 10)$, $\epsilon \sim N(0, 1)$ and the true parameters be given by: $\beta_0 = 50$, $\beta_1 = 1.5$ and $\beta_2 = -2$.

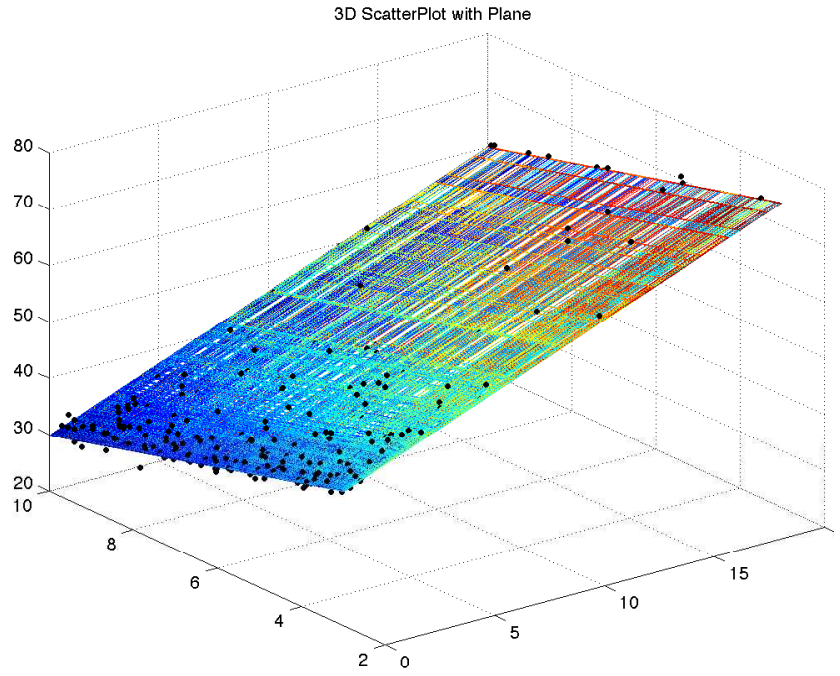
Given the sample, the correlation between x, z and y is given by:

	x	z	y
x	1	0.0696	0.8690
z	0.0696	1	-0.4181
y	0.8690	-0.4181	1

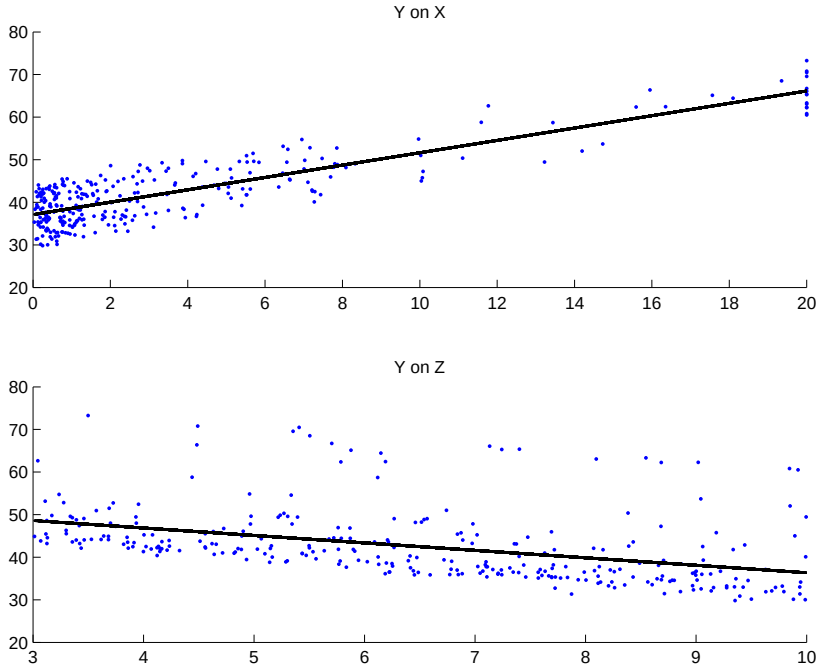
The relation between the variables can be examined through a scatter plot, yet the dimensionality of the problem makes it difficult to analyze:



It becomes even more confusing when trying to present the fitted plane in the same graph (recall that OLS will try to fit a plane to the observations the same way it fits a line when there is only one explanatory variable):



One alternative to the above presentation is to present two dimensional plots showing pairs (x, y) or (z, y). It is possible to fit a line to those pairs by projecting y onto x (or z), that is the points in the line are obtained as $\hat{y} = P_{\iota x} y = [\iota\, x] \left([\iota\, x]' [\iota\, x] \right)^{-1} [\iota\, x]' y$, where $\iota = [1, \ldots, 1]'$. Note that this line is not the same as the one obtained when regressing y on x and z, instead it is the line obtained when regressing y on x (or z) alone.



The difference between the estimates of the two dimensional approach and the three dimensional graph is presented in the following table:

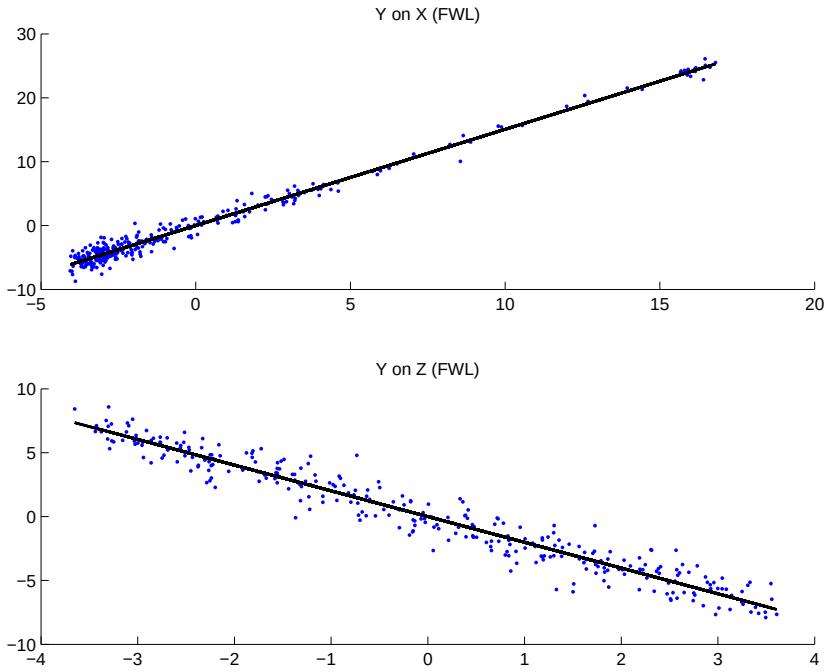
	β	$\hat{\beta}_{xz}$	$\hat{\beta}_x$	$\hat{\beta}_z$
β_0	50	49.9527	37.1273	53.8641
β_1	1.5	1.5068	1.4509	0
β_2	-2	-2.0138	0	-1.7508

The difference between the estimated coefficients of the two dimensional cases and the true coefficient is explained by the "omitted variable bias".

A better way to show the relation between x and y and z and y in a two dimensional setting is to use FWL to eliminate the variation in the interesting variables explained by the omitted variables. The following figure plots pairs of $(M_{\imath z}x, M_{\imath z}y)$ and $(M_{\imath x}z, M_{\imath x}y)$ where $M_{\imath x} = I - P_{\imath x}$. The fitted lines are the projection of $M_{\imath z}y$ on $M_{\imath z}x$ (or $M_{\imath x}y$ on $M_{\imath x}z$) given by:

$$\hat{y} = P_{M_{\imath z}x}M_{\imath z}y = M_{\imath z}x\left(\left(M_{\imath z}x\right)'M_{\imath z}x\right)^{-1}\left(M_{\imath z}x\right)'M_{\imath z}y = M_{\imath z}x\left(x'M_{\imath z}x\right)^{-1}x'M_{\imath z}y$$

The result is:



The OLS estimator is then not just a measure of correlation between variables, but rather a measure of marginal correlation, the correlation between y and an explanatory variable x after taking out the correlation induced by the possible relation between x and other explanatory variables (z). This marginality allows to decompose y into the effect of each of the variables that explains its movements, while a mere correlation can only indicate the direction of the relation and how strong it is.

As will be shown below the strength of the relation between two variables can be induced by a third variable that relates them, when the third variable is included OLS can identify its effect over the relations between variables concentrating in first hand effects only (the direct effects of one variable over another).

5.2 FWL with irrelevant variables

Consider a random sample of variables $\{y_i, x_i, z_i\}_{i=1}^{n}$ for $n = 200$ that was originated following the linear relation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad z_i = x_i + u_i$$

In this case there are two explanatory variables for y, but only one of them is in fact "causing" y. The irrelevant variable is nevertheless related to the relevant one, the correlation between them makes the irrelevant variable correlated with the independent one as well. If the variables were simply plotted in pairs the correlation between x and z will make z appear to have a relation to y. The use of FWL to present the variables eliminates the part of z that is correlated with x, thus eliminating its correlation with y, in this way the irrelevance of z for the regression becomes apparent.

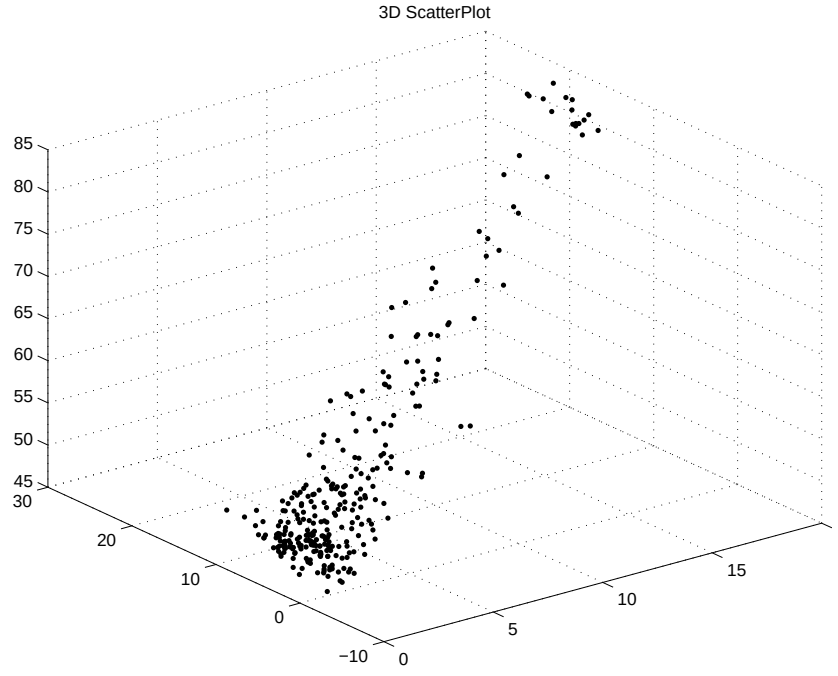
Let $x \sim LN(0.5, 1.5)$, $u \sim N(0, 3)$, $\epsilon \sim N(0, 1)$ and the true parameters be given by: $\beta_0 = 50$, $\beta_1 = 1.5$ and $\beta_2 = 0$.

Given the sample, the correlation between x, z and y is given by:

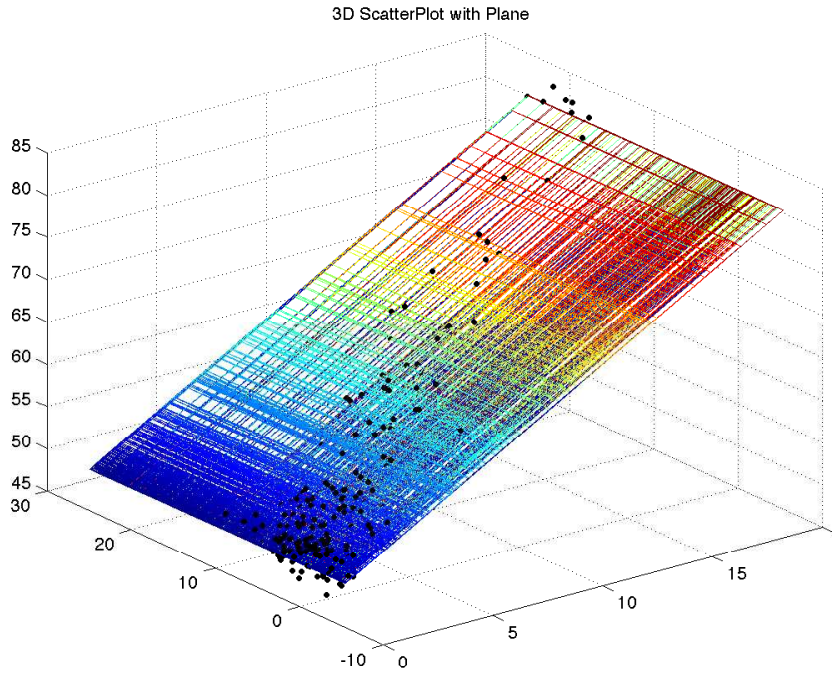
	x	z	y
x	1	0.8785	0.9918
z	0.8785	1	0.8692
y	0.9918	0.8692	1

The correlation between y and x is close to one (as expected) but the correlation between y and z is also high (0.87), a correlation this high is sometimes taken as an early indicator of a connection between variables. As before the

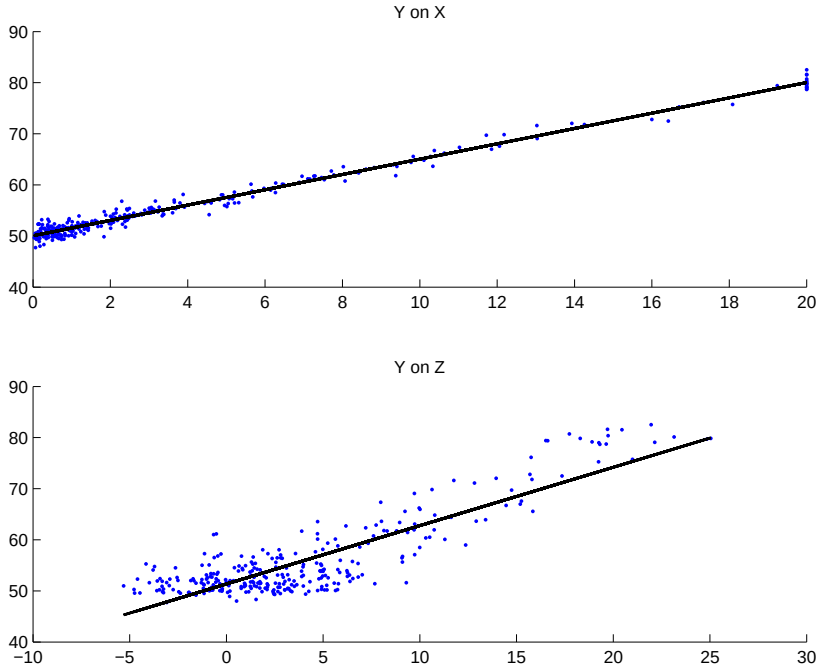
The relation between the variables can be examined through a scatter plot, yet the dimensionality of the problem makes it difficult to analyze, it is however possible to observe that all the variation in the vertical axis (y) is explained by variation in the right axis (x):



When one presents the fitted plane in the same graph it can be observed that the plane does not change along the z axis (left) when holding a value of x fixed (right), this effect becomes more difficult to appreciate when there are more than two potential explanatory variables:



One alternative to the above presentation is to present two dimensional plots showing pairs (x, y) or (z, y) along with its fitted lines. In this case it would seem like variable y is related (in a causal sense) to both x and z, yet this is only a product of the positive correlation induced between y and z by the underlying relations $x - y$ and $x - z$. Yet the fit of the line to the relation between y and z is remarkably good in the sample (recall that the line is obtained by regressing y on x (or z) alone).

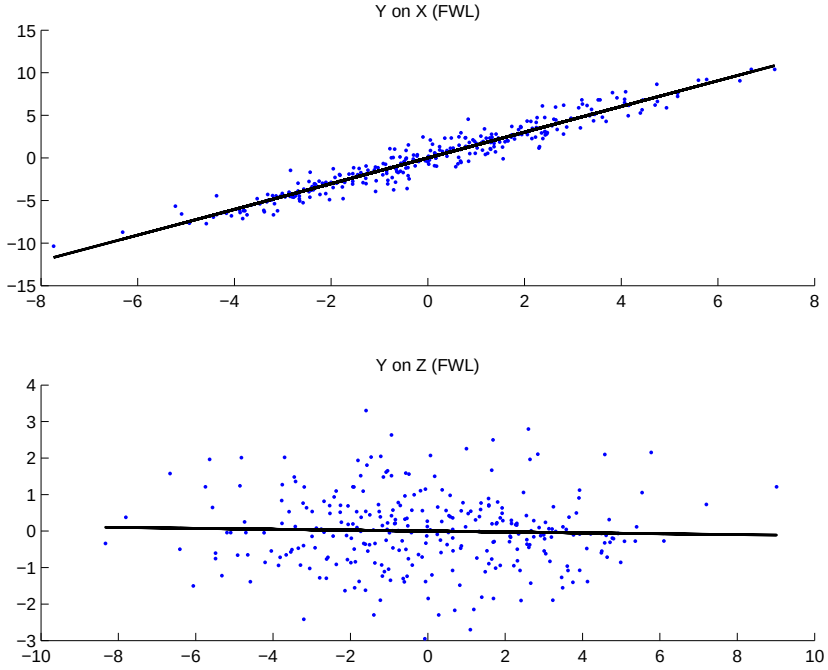


The difference between the estimates of the two dimensional approach and the three dimensional

graph is presented in the following table:

	β	$\hat{\beta}_{xz}$	$\hat{\beta}_x$	$\hat{\beta}_z$
β_0	50	50.0554	50.0547	51.357
β_1	1.5	1.511	1.4985	0
β_2	0	-0.0123	0	1.1415

When FWL is used to present the relation between x and y and z and y result changes dramatically, while the relation between x and y remains present the relation between z and y disappears, the fitted lines also indicate that there is no relation between z and y once the variation of x is taken out. As before the figure plots pairs of $(M_{\iota z}x, M_{\iota z}y)$ and $(M_{\iota x}z, M_{\iota x}y)$, and fitted lines given by $\hat{y} = P_{M_{\iota z}x}M_{\iota z}y$ and $\hat{y} = P_{M_{\iota x}z}M_{\iota x}y$.



6 Monte Carlo

The objective of a Monte Carlo simulation is to verify finite and large sample results. Given a model for how variables are determined one can test the performance and properties of sample estimators by generating several samples drawn from the same population (same model with variables following the same distribution).

Monte Carlo simulation takes advantage of the law(s) of large numbers. In general all results on the law of large numbers specify conditions over a random variable (z) that guarantee that (given a sample $\{z_i\}_{i=1}^m$):

$$\frac{1}{m}\sum_{i=1}^{m} z_i \longrightarrow E[z_i]$$

Depending on the conditions imposed over z different types of convergence can be obtained. In general one needs that $E[z_i] = \mu$ for all i and that the random variable is bounded (in some sense), it is sufficient to impose that $E[|z_i|] < \infty$, but stronger convergence results can be obtained if the random variable has bounded second moments, i.e. $V[z_i] = \sigma^2 < \infty$.

In a Monte Carlo simulation the random variable of interest is, for example, the estimator of a parameter $\left(\hat{\beta}\right)$, or a statistic for hypothesis testing. Given a population each random sample, drawn from that population, gives rise to an estimator, if there are N such samples one can compute N estimators $\left\{\hat{\beta}_i\right\}_{i=1}^{N}$, this set constitutes a sample of a random variable, in this case of $\hat{\beta}$, this random variable satisfies $E\left[\hat{\beta}_i\right] = \beta$ for each i and has bounded second moments (provided that the underlying process for x and ϵ has bounded second moments), in particular $V\left[\hat{\beta}_i\right] = \sigma_\epsilon^2 E\left[\left(X^{'}X\right)^{-1}\right] < \infty$ for all i. One can use this sample (of estimators) to compute sample moments, by increasing N (the number of samples), this sample moments will converge to the true population values for $\hat{\beta}$.

There is another dimension in which a Monte Carlo simulation treats a given random variable. If all the random variables are drawn from the same population, then they are realizations of the same distribution. Given a theoretical distribution for the random variable the Monte Carlo simulation allows to verify it by examining the behavior of the realizations. This two uses of the Monte Carlo are presented below.

Let $x \sim N\left(\mu, \sigma_x^2\right)$ and $\epsilon \sim N\left(0, \sigma_\epsilon^2\right)$ be two real valued random variables such that $E[\epsilon|x] = 0$, and let y be such that:

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

for given parameters β_0 and β_1.

Consider now a random sample $\{(x_i, y_i)\}$ of sample n. Define $y = [y_1, \ldots, y_n]^{'}$ and $X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}^{'}$ as the sample observations. An econometrician can compute estimates for $\beta = [\beta_0, \beta_1]^{'}$ through OLS, she can also compute an estimator for σ_ϵ^2. The estimators are given by:

$$\hat{\beta} = \left(X^{'}X\right)^{-1} X^{'}y \qquad s^2 = \frac{e^{'}e}{n-2}$$

where $e = y - X\hat{\beta}$.

Given the distribution of ϵ and the model for y it is known that:

$$E\left[\hat{\beta}\right] = \beta \qquad V\left[\hat{\beta}\right] = \sigma_\epsilon^2 E\left[\left(X^{'}X\right)^{-1}\right] \qquad \hat{\beta} \sim N\left(\beta, \sigma^2 E\left[\left(X^{'}X\right)^{-1}\right]\right) \qquad E\left[s^2\right] = \sigma_\epsilon^2$$

These properties can be tested by computing averages across different samples, all drawn from the same distribution, and all following the same model. For this, first define the size of each sample (n) and the number of samples to be simulated (N), along with the parameters of the model: $\mu, \sigma_x^2, \sigma_\epsilon^2, \beta_0, \beta_1$.

```
%% Parameter Values

n = 100 ; % Sample size
N = 10000 ; % Number of samples

beta_0 = -5 ; % True intercept
beta_1 = 2 ; % True intercept

% Error term is distributed normal with mean 0 and std sigma_e
sigma_e = 0.8 ; % Error Variance

% Independent variable is distributed normal with mean mu_x and std dev  sigma_x
mu_x= 7 ; % Mean of independent variable
sigma_x= 0.6 ; % Std of independent variable
```

Given parameter values the variables can be simulated. Note that it is more efficient to simulate all the samples at once.

```
%% Variable simulation

% Error
eps_tot = normrnd(0,sigma_e,n,N);
% Independent variable
x_tot = normrnd(mu_x,sigma_x,n,N) ;
% x_tot = rand(n,N);
% Dependent variable
y_tot = beta_0+beta_1*x_tot+eps_tot ;
```

Then one can compute the estimators for the parameters, residuals and variance of the residuals:

```
%% Estimation

beta_ols = NaN(2,N) ; % Estimated Parameters
res_ols = NaN(n,N) ; % Residuals of the regression
s2 = NaN(N,1) ; % Estimates of error variance
XX_inv = NaN(2,2,N);
for i=1:N
    X = [ones(n,1),x_tot(:,i)] ;
    beta_ols(:,i) = ((X'*X)\X')*y_tot(:,i) ;
    res_ols(:,i) = (eye(n)-X*((X'*X)\X'))*y_tot(:,i) ;
    s2(i) = res_ols(:,i)'*res_ols(:,i)/(n-2) ;
    XX_inv(:,:,i) =  (X'*X)\eye(2) ;
end

beta_mean = mean(beta_ols,2) ; % This should be a vector of the treu parameters

res_mean   = mean(res_ols) ; % This should be a vector of zeros
```

```
s2_mean     = mean(s2) ; % This should be sigma_e^2

XX_inv_mean = mean(XX_inv,3) ;

beta_cov     = cov(beta_ols') ; % This should be equal to sigma_e*E[(X'X)^-1]

beta_cov_0 = sigma_e^2*XX_inv_mean ;
```

The code also computes the mean of the estimators across samples and the theoretical variance of $\hat{\beta}$ (approximated with the $E\left[\widehat{(X'X)^{-1}}\right] = \frac{1}{n}\sum\left(X_j'X_j\right)^{-1}$).

The results are reported in the following tables:

	β_0	β_1	σ_ϵ^2
True Value	-5	2	0.64
Mean of Estimates	-5.0021	2.0002	0.6412

The results in the table indicate that the estimators are in fact unbiased.

And for the covariance matrix:

TrueValue	β_0	β_1	Estimates	β_0	β_1
β_0	[0.9042]	[-0.1283]	β_0	[0.9018]	[-0.1278]
β_1	[-0.1283]	[0.0183]	β_1	[-0.1278]	[0.0182]

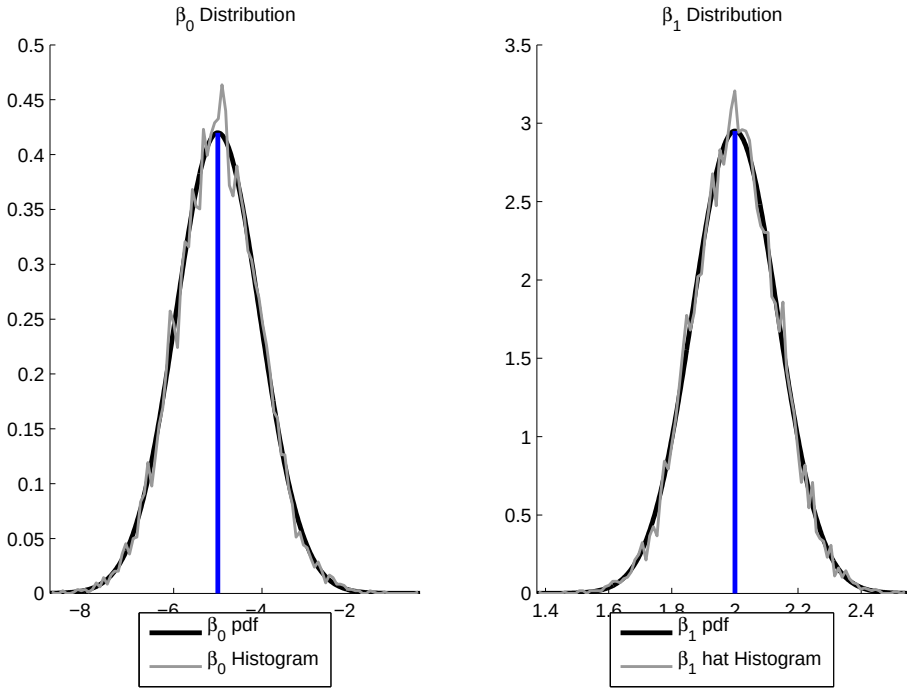
The results hold for the variance covariance matrix of the parameters. This will also show up in the graph below that presents the (actual) distribution of the estimators across samples.

The tables can be generates as follows:

```
% Reporting
    b_mat = [beta_0 beta_1 sigma_e^2;
                    beta_mean'       s2_mean ];
    col={'','b0','b1','sigma_e'};
    row={'true';'mean'};
    D=[col; [row , num2cell(b_mat)]];
    disp('Parameters')
    disp(D)

    c_mat = [beta_cov_0 NaN(2,1) beta_cov];
    col={'True','b0','b1','sample','b0','b1'};
    row={'b0';'b1'};
    D=[col; [row , num2cell(c_mat)]];
    disp('Covariance Matrix')
    disp(D)
```

Finally one can plot the distribution of the estimated parameters and the theoretical normal distribution that they are supposed to follow:



The distribution of the parameters in the (simulated) samples resembles the normal distribution that the parameter is supposed to take.

The code below generates the graph.

```
%% Ploting

% Beta_Hat

    b0_points = linspace(min(beta_ols(1,:)),max(beta_ols(1,:)),500) ;
    b0_pdf      = normpdf(b0_points,beta_0,(beta_cov_0(1,1))^0.5) ;

    b1_points = linspace(min(beta_ols(2,:)),max(beta_ols(2,:)),500) ;
    b1_pdf      = normpdf(b1_points,beta_1,(beta_cov_0(2,2))^.5) ;

    figure;
    subplot(1,2,1)
        hold on;
        [a,b] = hist(beta_ols(1,:),100) ;
        plot(b0_points,b0_pdf,'k','linewidth',2.5);
%         bar(b,a/N/diff(b(1:2)),0.3) ;
        plot(b,a/N/diff(b(1:2)),'color',[0.6,0.6,0.6],'linewidth',1.5) ;
        line([beta_0 beta_0],[0,max(b0_pdf)],'linewidth',2.5)
        xlim([min(beta_ols(1,:)),max(beta_ols(1,:))])
        legend('\beta_0 pdf','\beta_0 Histogram','location','southoutside');
        title('\beta_0 Distribution')
        hold off;
```

```
    subplot(1,2,2)
        hold on;
        [a,b] = hist(beta_ols(2,:),100) ;
        plot(b1_points,b1_pdf,'k','linewidth',2.5);
%         bar(b,a/N/diff(b(1:2)),0.3) ;
        plot(b,a/N/diff(b(1:2)),'color',[0.6,0.6,0.6],'linewidth',1.5) ;
        line([beta_1 beta_1],[0,max(b1_pdf)],'linewidth',2.5)
        xlim([min(beta_ols(2,:)),max(beta_ols(2,:))])
        legend('\beta_1 pdf','\beta_1 hat Histogram','location','southoutside');
        title('\beta_1 Distribution')
        hold off;
figurename='Monte_Carlo_Beta';
print('-dpdf',figurename)
```

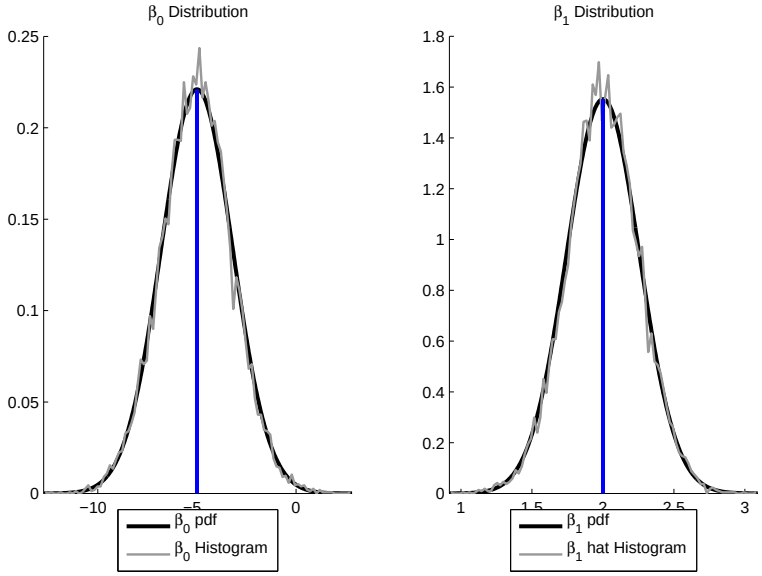
The properties of the estimator that have been discussed so far are all finite sample properties, hence they are valid for any value of n. There is a difference between the role played by n and the role played by N. The results above hold for any n provided that N is big, that is, the higher N is (the larger the sample of the random variable of interest) the more exact are the results to what the theory predicted. If N is not too big there might not be enough observations of $\hat{\beta}$ to get $\frac{1}{n}\sum\hat{\beta}_i \sim E\left[\hat{\beta}_i\right] = \beta$, or not enough to get $\hat{\beta}$ to look normal.

Consider two examples, in the first maintain the number of replications N but lower n, in the second one maintain the n but lower N. It is expected that the first one maintains all the results above, while adjusting for a new covariance matrix of the estimator, while in the second one the approximation to the true moments of the parameter should be less precise.

The results for the same simulations with $n = 30$ are presented below:

	β_0	β_1	σ_ϵ^2
True Value	-5	2	0.64
Mean of Estimates	-4.9993	1.9999	0.6448

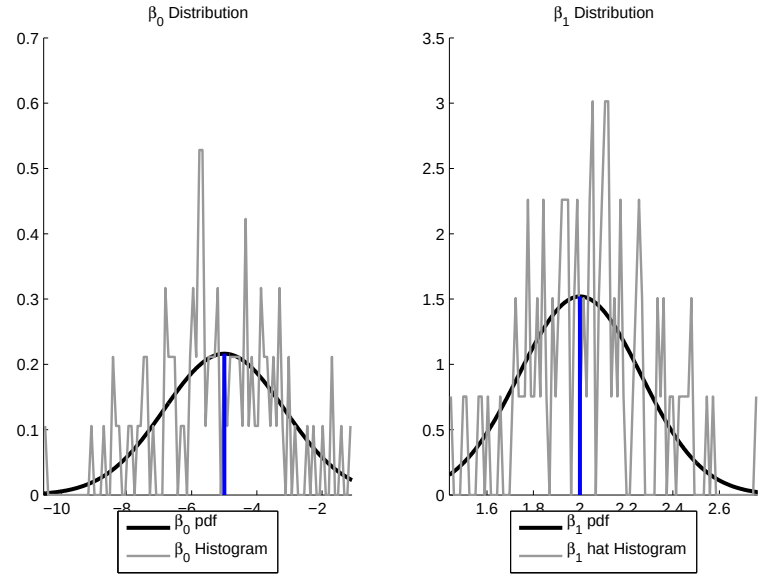
TrueValue	β_0	β_1	Estimates	β_0	β_1
β_0	[3.2296]	[-0.4582]	β_0	[3.1896]	[-0.4528]
β_1	[-0.4582]	[0.0655]	β_1	[-0.4528]	[0.0647]



The normal distribution still holds and the estimates are all unbiased. This changes when one takes only 100 simulations ($N = 100$).

	β_0	β_1	σ_ϵ^2
True Value	-5	2	0.64
Mean of Estimates	-5.1887	2.0297	0.6323

TrueValue	β_0	β_1	Estimates	β_0	β_1
β_0	[3.2804]	[-0.4664]	β_0	[3.8532]	[-0.5450]
β_1	[-0.4664]	[0.0667]	β_1	[-0.5450]	[0.0775]



7 Selected Exercises - Homework 2

7.1 Hayashi (p46) Exercise 3

For $F = (Rb-r)^{'}\left[R\widehat{V[b]}R^{'}\right]^{-1}(Rb-r)\,|\#r$ to be well defined matrix $R\left(X^{'}X\right)^{-1}R^{'}$ must be non-singular, prove that it is also positive definite.

- For $R\left(X^{'}X\right)^{-1}R^{'}$ to be positive definite we need $\forall_{z\neq 0} z^{'}R\left(X^{'}X\right)^{-1}R^{'}z > 0$ with $z \in \mathbb{R}^{\#r}$.
- We know $X^{'}X$ is positive definite then so is $\left(X^{'}X\right)^{-1}$.
- Since R is of full (row) rank $\forall_{z\neq 0} R^{'}z \neq 0$.
- Define $w = R^{'}z$ then $z^{'}R\left(X^{'}X\right)^{-1}R^{'}z = w^{'}\left(X^{'}X\right)^{-1}w > 0$ where the inequality follows from $\left(X^{'}X\right)^{-1}$ being p.d.
- Since the above works for any $z \in \mathbb{R}^{\#r}\setminus\{0\}$ then the result is proven.

7.2 Hayashi (p46) Exercise 7

Under A1.1 to A1-5 $V\left[s^2|X\right] = \frac{2\sigma^4}{n-k}$

- Note $s^2 = \frac{e^{'}e}{n-k} = \frac{\varepsilon M_X \varepsilon}{n-k}$.
- Under A1.5 $\epsilon \sim N\left(0,\sigma^2\right)$. Then $\frac{\epsilon}{\sigma} \sim N(0,1)$. Then $\left(\frac{\varepsilon}{\sigma}\right)M_X\left(\frac{\varepsilon}{\sigma}\right)|X \sim \chi^2_{(n-k)}$.
- Then: $V\left[s^2|X\right] = V\left[\frac{\varepsilon M_X \varepsilon}{n-k}|X\right] = V\left[\left(\frac{\sigma^2}{n-k}\right)\left(\frac{\varepsilon}{\sigma}\right)M_X\left(\frac{\varepsilon}{\sigma}\right)|X\right] = \frac{\sigma^4}{(n-k)^2}V\left[\left(\frac{\varepsilon}{\sigma}\right)M_X\left(\frac{\varepsilon}{\sigma}\right)|X\right] = \frac{2\sigma^4}{n-k}$

7.3 Johnston & Dinardo (Chapter 1 - Relationship Between Two Variables) I

Consider a model $y_i = \beta x_i + u_i$ without an intercept and a random sample for $\{y_i, x_i\}_{i=1}^n$. There are two estimators for β:

$$\beta_1 = \frac{\sum x_i y_i}{\sum x_i^2} \qquad \beta_2 = \frac{\overline{y}}{\overline{x}} = \frac{\sum y_i}{\sum x_i}$$

All usual assumptions apply, in particular $E[u_i|X] = 0$ and $V[u_i|X] = \sigma^2$.

Note that:

$$\beta_1 = \beta + \frac{\sum x_i u_i}{\sum x_i^2} \qquad \beta_2 = \beta + \frac{\sum u_i}{\sum x_i}$$

1. Expected value of the estimators:

$$\begin{aligned} E[\beta_1|X] &= \frac{1}{\sum x_i^2}\sum x_i E[y_i|X] = \frac{1}{\sum x_i^2}\sum x_i\left(\beta x_i + E[u_i|X]\right) = \beta \\ E[\beta_2|X] &= \frac{1}{\sum x_i}\sum E[y_i|X] = \frac{1}{\sum x_i}\sum\left(\beta x_i + E[u_i|X]\right) = \beta \end{aligned}$$

2. Variance of the estimators:

$$\begin{aligned} V[\beta_1|X] &= V\left[\beta + \frac{\sum x_i u_i}{\sum x_i^2}|X\right] = \frac{1}{\left(\sum x_i^2\right)^2} V\left[\sum x_i u_i|X\right] = \frac{1}{\left(\sum x_i^2\right)^2}\left(\sum x_i^2 V[u_i|X]\right) = \frac{\sigma^2}{\sum x_i^2} \\ V[\beta_2|X] &= V\left[\beta + \frac{\sum u_i}{\sum x_i}|X\right] = \frac{1}{\left(\sum x_i\right)^2} V\left[\sum u_i|X\right] = \frac{n\sigma^2}{\left(\sum x_i\right)^2} \end{aligned}$$

3. Efficiency:

$$\begin{aligned} V[\beta_1|X] &\leq V[\beta_2|X] \\ \frac{\sigma^2}{\sum x_i^2} &\leq \frac{n\sigma^2}{\left(\sum x_i\right)^2} \\ \left(\sum x_i\right)^2 &\leq n\left(\sum x_i^2\right) \end{aligned}$$

This can be verified with the Cauchy-Schwarz inequality: $\left(\sum x_i y_i\right)^2 \leq \left(\sum x_i^2\right)\left(\sum y_i^2\right)$, choosing $y_i = 1$ one gets the result.

7.4 Johnston & Dinardo (Chapter 1 - Relationship Between Two Variables) II

Consider a random sample of variables $\{y_i, x_i\}_{i=1}^n$ that was originated following the linear relation:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

This can be written as:

$$y_i = \beta_0 + \beta_1 \overline{x} + \beta_1 (x_i - \overline{x}) + u_i$$

for some value $\overline{x}$. Rename variables such that $\gamma_0 = \beta_0 + \beta_1 \overline{x}$, $\gamma_1 = \beta_1$ and $z_i = x_i - \overline{x}$. The model is then:

$$y_i = \gamma_0 + \gamma_1 z_i + u_i$$

Note that $\overline{z} = 0$ and $E[u_i|z] = 0$.

Given parameter estimates define the OLS residual:

$$e_i = y_i - \hat{\gamma}_0 - \hat{\gamma}_1 z_i$$

1. Verify that if one where to regress y on a constant and z ($y_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i + e_i$) the estimated coefficients would be: $\hat{\gamma}_0 = \overline{y}$ and $\hat{\gamma}_1 = \hat{\beta}_1$.

$$\begin{aligned} \hat{\gamma}_0 &= \overline{y} - \hat{\gamma}_1 \overline{z} = \overline{y} \\ \hat{\gamma}_1 &= \frac{\sum (y_i - \overline{y})(z_i - \overline{z})}{\sum (z_i - \overline{z})^2} = \frac{\sum (y_i - \overline{y}) z_i}{\sum z_i^2} = \frac{\sum (y_i - \overline{y})(x_i - \overline{x})}{\sum (x_i - \overline{x})^2} = \hat{\beta}_1 \end{aligned}$$

2. Verify that $\text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) = 0$. (Hint: Use the variance-covariance matrix of the OLS estimator).

$$\begin{aligned} V(\gamma) &= \sigma^2 \left(\begin{bmatrix} 1 & z_1 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix}' \begin{bmatrix} 1 & z_1 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix} \right)^{-1} \\ &= \sigma^2 \begin{bmatrix} n & \sum z_i \\ \sum z_i & \sum z_i^2 \end{bmatrix}^{-1} \\ &= \frac{\sigma^2}{n\sum z_i^2 - \left(\sum z_i\right)^2} \begin{bmatrix} \sum z_i^2 & -\sum z_i \\ -\sum z_i & n \end{bmatrix} \end{aligned}$$

Yet since $\sum z_i = n\overline{z} = 0$ we have:

$$V(\gamma) = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{\sum z_i^2} \end{bmatrix}$$

This shows that $\text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) = 0$.

3. Verify that $\text{Cov}(e_i, \hat{\gamma}_0) = 0$.

$$\begin{aligned}
\text{Cov}(e_i, \hat{\gamma}_0) &= E\left[(y_i - \hat{\gamma}_0 - \hat{\gamma}_1 z_i)(\hat{\gamma}_0 - \gamma_0)\right] \\
&= E\left[(u_i - (\hat{\gamma}_0 - \gamma_0) - (\hat{\gamma}_1 - \gamma_1) z_i)(\hat{\gamma}_0 - \gamma_0)\right] \\
&= E\left[u_i(\hat{\gamma}_0 - \gamma_0)\right] - E\left[(\hat{\gamma}_0 - \gamma_0)^2\right] - E\left[(\hat{\gamma}_1 - \gamma_1)(\hat{\gamma}_0 - \gamma_0)\right] z_i \\
&= E\left[u_i \hat{\gamma}_0\right] - \frac{\sigma^2}{n} - 0 \\
&= E\left[u_i(\gamma_0 + \gamma_1 \overline{z} + \overline{u})\right] - \frac{\sigma^2}{n} \\
&= E\left[u_i \overline{u}\right] - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0
\end{aligned}$$

Note: $E\left[u_i \hat{\gamma}_0\right] = \frac{\sigma^2}{n}$.

4. Verify that $\text{Cov}(e_i, \hat{\gamma}_1) = 0$.

$$\begin{aligned}
\text{Cov}(e_i, \hat{\gamma}_1) &= E\left[(y_i - \hat{\gamma}_0 - \hat{\gamma}_1 z_i)(\hat{\gamma}_1 - \gamma_1)\right] \\
&= E\left[(u_i - (\hat{\gamma}_0 - \gamma_0) - (\hat{\gamma}_1 - \gamma_1) z_i)(\hat{\gamma}_1 - \gamma_1)\right] \\
&= E\left[u_i(\hat{\gamma}_1 - \gamma_1)\right] - E\left[(\hat{\gamma}_1 - \gamma_1)(\hat{\gamma}_0 - \gamma_0)\right] - E\left[(\hat{\gamma}_1 - \gamma_1)^2\right] z_i \\
&= E\left[u_i \hat{\gamma}_1\right] - 0 - \frac{\sigma^2}{\sum z_i^2} z_i \\
&= E\left[u_i \left(\frac{\sum y_j z_j}{\sum z_j^2}\right)\right] - \frac{\sigma^2}{\sum z_i^2} z_i \\
&= \frac{1}{\sum z_i^2}\left[\sum \left(E\left[u_i y_j\right] z_j\right)\right] - \frac{\sigma^2}{\sum z_i^2} z_i \\
&= \frac{\sigma^2}{\sum z_i^2} z_i - \frac{\sigma^2}{\sum z_i^2} z_i = 0
\end{aligned}$$

Note: $E\left[u_i \hat{\gamma}_1\right] = \frac{\sigma^2 z_i}{\sum z_i^2}$.

5. Verify that $\text{Cov}(e_i, \overline{u}) = 0$

$$\begin{aligned}
\operatorname{Cov}(e_i, \overline{u}) &= E\left[\left(u_i - (\hat{\gamma}_0 - \gamma_0) - (\hat{\gamma}_1 - \gamma_1) z_i\right) \overline{u}\right] \\
&= E\left[u_i \overline{u}\right] - E\left[(\hat{\gamma}_0 - \gamma_0) \overline{u}\right] - E\left[(\hat{\gamma}_1 - \gamma_1) \overline{u}\right] z_i \\
&= \frac{\sigma^2}{n} - E\left[\hat{\gamma}_0 \overline{u}\right] - E\left[\hat{\gamma}_1 \overline{u}\right] z_i \\
&= \frac{\sigma^2}{n} - \frac{1}{n^2}\left(\sum_i \sum_j E\left[y_j u_i\right]\right) - \frac{1}{n}\left(\sum_i \sum_j E\left[y_j u_i\right] z_j\right) \frac{z_i}{\sum z_j^2} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} - \frac{1}{n}\left(\sigma^2 \sum_j z_j\right) \frac{z_i}{\sum z_j^2} \\
&= 0
\end{aligned}$$

Recall that $\sum z_j = 0$.

6. Verify that $\operatorname{Cov}(e_i, \sum u_i z_i) = 0$

$$\begin{aligned}
\operatorname{Cov}\left(e_i, \sum u_j z_j\right) &= E\left[\left(u_i - (\hat{\gamma}_0 - \gamma_0) - (\hat{\gamma}_1 - \gamma_1) z_i\right)\left(\sum u_j z_j\right)\right] \\
&= \left(\sum_j E\left[u_i u_j\right] z_j\right) - E\left[(\hat{\gamma}_0 - \gamma_0)\left(\sum u_j z_j\right)\right] - E\left[(\hat{\gamma}_1 - \gamma_1)\left(\sum_j u_j z_j\right)\right] z_i \\
&= \sigma^2 z_i - E\left[\hat{\gamma}_0\left(\sum u_j z_j\right)\right] - E\left[\hat{\gamma}_1\left(\sum_j u_j z_j\right)\right] z_i \\
&= \sigma^2 z_i - \left(\sum_j E\left[\hat{\gamma}_0 u_j\right] z_j\right) - \left(\sum_j E\left[\hat{\gamma}_1 u_j\right] z_j\right) z_i \\
&= \sigma^2 z_i - \left(\frac{\sigma^2}{n} \sum_j z_j\right) - \left(\sum_j \left(\frac{\sigma^2}{\sum z_k^2} z_j\right) z_j\right) z_i \\
&= \sigma^2 z_i - \left(\frac{\sigma^2}{\sum z_k^2} \sum_j z_j^2\right) z_i \\
&= 0
\end{aligned}$$

7. Verify that $\operatorname{Cov}(\overline{u}, \sum u_i z_i) = 0$.

$$\operatorname{Cov}\left(\overline{u}, \sum u_i z_i\right) = \frac{1}{n} E\left[\left(\sum_j u_j\right)\left(\sum_i u_i z_i\right)\right] = \frac{1}{n}\left(\sum_j \sum_i E\left[u_j u_i\right] z_i\right) = \frac{\sigma^2}{n} \sum_i z_i = 0$$

7.5 Binary explanatory variable

Consider the following model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where y_i measures the income of individuals and x_i is a dummy variable indicating whether the individual is above average in height ($x_i = 1$) or below average in height ($x_i = 0$).

The OLS estimation solves:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = \min_{\beta_0,\beta_1} \left[\sum_{x_i=0} (y_i - \beta_0) + \sum_{x_i=1} (y_i - \beta_0 - \beta_1) \right]$$

The first order conditions are:

$$\begin{aligned} \sum_{x_i=0} \left(y_i - \hat{\beta}_0 \right) + \sum_{x_i=1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \right) &= 0 \\ \sum_{x_i=1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \right) &= 0 \end{aligned}$$

Using the second condition in the first:

$$\sum_{x_i=0} \left(y_i - \hat{\beta}_0 \right) = 0 \quad \rightarrow \quad \hat{\beta}_0 = \frac{\sum_{x_i=0} y_i}{n_0}$$

Then the estimator for β_0 is the average income given $x_i = 0$, that is the average income for individuals below average height.

Replacing on the second equation:

$$\hat{\beta}_1 = \frac{\sum_{x_i=1} y_i}{n_1} - \hat{\beta}_0$$

So that the estimator for β_1 is the difference between the average income of agents above average height and agents below average height.

As in the regression with only intercept the best prediction of the model is the mean of the variable, in this case average income. OLS separates the sample and gives as fitted value the average income of the type of agent.

8 OLS and GLS

Consider a model given by:

$$Y = X\beta + \varepsilon$$

where $E[\epsilon|X] = 0$ and $V[\varepsilon|X] = \Sigma$.

- The OLS estimator is given by:

$$\beta_{ols} = \left(X^{'}X\right)^{-1} X^{'}Y = \beta + \left(X^{'}X\right)^{-1} X^{'}\varepsilon$$

  Its expected value is given by:

$$E[\beta_{ols}|X] = \beta + \left(X^{'}X\right)^{-1} X^{'}E[\varepsilon|X] = \beta$$

  Then its variance is:

$$V[\beta_{ols}|X] = V\left[\beta + \left(X^{'}X\right)^{-1} X^{'}\varepsilon|X\right] = \left(X^{'}X\right)^{-1} X^{'}\Sigma X \left(X^{'}X\right)^{-1}$$

- The GLS estimator is given by:

$$\beta_{gls} = \left(X^{'}\Sigma^{-1}X\right)^{-1} X^{'}\Sigma^{-1}Y = \beta + \left(X^{'}\Sigma^{-1}X\right)^{-1} X^{'}\Sigma^{-1}\varepsilon$$

  Its expected value is given by:

$$E[\beta_{gls}|X] = \beta + \left(X^{'}\Sigma^{-1}X\right)^{-1} X^{'}\Sigma^{-1}E[\varepsilon|X] = \beta$$

  Then its variance is:

$$\begin{aligned} V[\beta_{gls}|X] &= V\left[\beta + \left(X^{'}\Sigma^{-1}X\right)^{-1} X^{'}\Sigma^{-1}\varepsilon|X\right] \\ &= \left(X^{'}\Sigma^{-1}X\right)^{-1} X^{'}\Sigma^{-1}\Sigma\Sigma^{-1}X \left(X^{'}\Sigma^{-1}X\right)^{-1} \\ &= \left(X^{'}\Sigma^{-1}X\right)^{-1} \end{aligned}$$

Both estimators are unbiased. It can be shown that β_{gls} is more efficient than β_{ols} in the sense that: $V[\beta_{ols}|X] - V[\beta_{gls}|X]$ is a positive definite matrix.

Note that if A and B are positive semi definite and invertible matrices then the following implication holds:

- If $A - B$ is positive semi definite then $B^{-1} - A^{-1}$ is also positive semi definite. That is, for all $c \neq 0$

$$c^{'}(A - B)c \geq 0 \longrightarrow c^{'}\left(B^{-1} - A^{-1}\right)c \geq 0$$

Set $A = V[\beta_{gls}|X]^{-1}$ and $B = V[\beta_{ols}|X]^{-1}$, then we want to show:

$$c^{'}\left(V[\beta_{gls}|X]^{-1} - V[\beta_{ols}|X]^{-1}\right)c \geq 0$$

Which implies the desired result:

$$c^{'}\left(V[\beta_{ols}|X] - V[\beta_{gls}|X]\right)c \geq 0$$

- Since Σ is positive definite there exists P such that $\Sigma = PP^{'}$ and $\Sigma^{-1} = P^{-1^{'}}P^{-1}$ with P and P full ranked.
- Then:

$$\begin{aligned} V\left[\beta_{gls}|X\right]^{-1} - V\left[\beta_{ols}|X\right]^{-1} &= X^{'}\Sigma^{-1}X - \left(X^{'}X\right)\left(X^{'}\Sigma X\right)^{-1}\left(X^{'}X\right) \\ &= X^{'}\left[\Sigma^{-1} - X\left(X^{'}\Sigma X\right)^{-1}X^{'}\right]X \\ &= X^{'}\left[P^{-1^{'}}P^{-1} - X\left(X^{'}PP^{'}X\right)^{-1}X^{'}\right]X \\ &= X^{'}P^{-1^{'}}\left[I - P^{'}X\left(X^{'}PP^{'}X\right)^{-1}X^{'}P\right]P^{-1}X \\ &= X^{'}P^{-1^{'}}\left[I - \left(P^{'}X\right)\left(\left(P^{'}X\right)^{'}\left(P^{'}X\right)\right)^{-1}\left(P^{'}X\right)^{'}\right]P^{-1}X \\ &= X^{'}P^{-1^{'}}M_{P^{'}X}P^{-1}X \\ &= \left(M_{P^{'}X}P^{-1}X\right)^{'}\left(M_{P^{'}X}P^{-1}X\right) \end{aligned}$$

- Let $c \in \mathbb{R}^{k}\backslash\{0\}$ then:

$$\begin{aligned} c^{'}\left(V\left[\beta_{gls}|X\right]^{-1} - V\left[\beta_{ols}|X\right]^{-1}\right)c &= c^{'}\left(M_{P^{'}X}P^{-1}X\right)^{'}\left(M_{P^{'}X}P^{-1}X\right)c \\ &= \left(\left(M_{P^{'}X}P^{-1}X\right)c\right)^{'}\left(M_{P^{'}X}P^{-1}X\right)c > 0 \end{aligned}$$

- The above is the desired result.
- Note that this was obtained also from the expression $X^{'}P^{-1^{'}}M_{P^{'}X}P^{-1}X$
  - Since $M_{P^{'}X}$ is idempotent then it is positive definite.
  - Since X and P have full rank (k) then $P^{-1}X$ has also rank k.
  - Then $P^{-1}Xc \neq 0$ for $c \neq 0$ which give $\left(P^{-1}X\right)^{'}M_{P^{'}X}\left(P^{-1}X\right)$.
  - (In general if B is full ranked and A is positive definite then $B^{'}AB$ is positive definite).

9 Feasible GLS

9.1 General setting

Consider a model given by:

$$Y = X\beta + \varepsilon$$

where $E[\epsilon|X] = 0$ and $V[\varepsilon|X] = \Sigma$. Matrix Σ can have an arbitrary form and can in general be a function of X.

If Σ is known then the GLS estimator can be implemented as:

$$\beta_{gls} = \left(X^{'}\Sigma^{-1}X\right)^{-1} X^{'}\Sigma^{-1}Y$$

If Σ is not known then GLS is not feasible, for it to be implemented an estimator of Σ has to be constructed. Provided a random sample $\{y_i, x_i\}_{i=1}^{n} = \{Y, X\}$ and since the OLS estimator is still unbiased (and consistent) the residuals it generates are good approximations for ε. Then the residuals e (of size $n \times 1$) can be used to construct an estimator $\hat{\Sigma}$.

A first approximation for $\hat{\Sigma}$ is a method of moments estimator, if one had access to N independent samples $\{Y_j, X_j\}_{j=1}^{N}$ one could obtain N residuals e_j and create the estimator:

$$\hat{\Sigma} = \frac{1}{N}\sum_{j=1}^{N} e_j e_j^{'} \longrightarrow E\left[\varepsilon\varepsilon^{'}\right] = \Sigma$$

the convergence to the variance of the disturbances ε is obtained by the law of large numbers and the fact that β_{ols} is still a good estimator for β.

When there is only one sample available the method of moments estimators cannot be constructed. There is an additional problem in this case, matrix Σ is an $n \times n$ symmetric matrix, hence it has $n(n-1)/2$ different elements, yet the OLS regression only generates n observations for the residuals. Then, in order to obtain $\hat{\Sigma}$ one needs to impose restrictions over Σ so that it can be estimated. Once one imposes enough restrictions it is possible to use e_i as a "sample estimator" of ϵ_i and compute $\hat{\Sigma}$.

Note that the problem of estimating $\hat{\Sigma}$ with only one sample does not go away when the sample size increases since the number of parameters to estimate increases with the sample. An alternative is to impose further restrictions, a leading example, explored below, is to assume that residuals are uncorrelated and that the variance follows a parametric form, so $\sigma_i^2 = f(\gamma, z_i)$ for a common parameter vector γ and a (potentially) observation specific set of variables z_i, doing this reduces the dimensionality of the problem from an estimation of the n σ_i^2 to the estimation of the parameter γ with the n observations of e in the sample.

Finally note that imposing too many restrictions is potentially hazardous for the estimation, if the restrictions don't hold in the DGP then the F-GLS estimator computed from $\hat{\Sigma}$ does not generate reliable results, in particular the estimator fails to be consistent for β.

9.2 Parametric assumption example

In order to identify all the elements in Σ one has to impose at least $\frac{n(n-1)}{2} - n = \frac{n(n-3)}{2}$ restrictions. Assume that:

$$E\left[\epsilon_i \epsilon_j\right] = \begin{cases} \sigma_i^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

so that the model is heteroskedastic but has zero correlation between errors. Then there are n parameters to estimate $\left\{\sigma_i^2\right\}_{i=1}^n$. In order to determine what σ_i^2 is (for each i) a parametric assumption is placed over the form of the variance, in particular assume that $\sigma_i^2 = f\left(x_i\right)$, so that all variances follow the same common function evaluated at the given characteristics of the individual.

One common specification for f_i is for it to be linear it (some of) the explanatory variables, it can also be that it depends on the square of the variables. For the example in this section consider:

$$f\left(z_i\right) = z_i^{'}\gamma$$

Since γ is an unknown parameter it has to be inferred from the sample. The following regression provides a (consistent) estimator of γ - provided that the assumptions over $E\left[\epsilon_i \epsilon_j\right]$ and f are correct:

$$e^2 = Z\gamma + \xi \longrightarrow \hat{\gamma} = \left(Z^{'} Z\right)^{-1} Z^{'} e^2$$

Where $e^2 = \left(e_1^2, \ldots, e_n^2\right)^{'}$ and e is the vector of OLS residuals from the regression of Y onto X. Since the OLS estimates are still unbiased (and consistent) e_i is a point estimator of ϵ_i and e_i^2 a point estimator of σ_i^2.

Then $\hat{\sigma}_i^2 = z_i^{'}\hat{\gamma}$ and:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_1^2 & & 0 \\ & \ddots & \\ 0 & & \hat{\sigma}_n^2 \end{bmatrix}$$

With the above results the F-GLS estimator can be computed as:

$$\beta_{fgls} = \left(X^{'}\hat{\Sigma}^{-1}X\right)^{-1} X^{'}\hat{\Sigma}^{-1}Y$$

9.2.1 Monte Carlo simulation

Consider the following DGP:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad x_i \sim U\left(a, b\right) \qquad \epsilon_i \sim N\left(0, \sigma_i^2\right) \qquad \sigma_i^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2$$

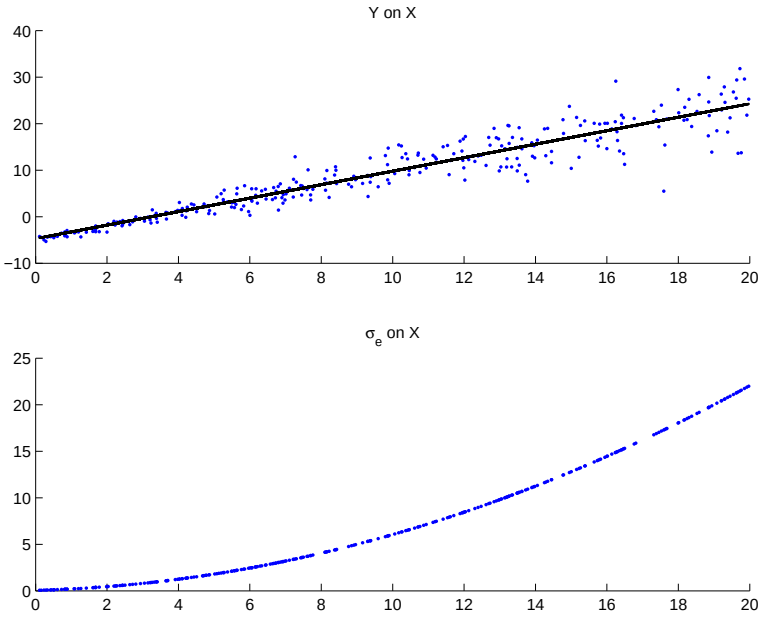
This model exhibits heteroskedasticity if $\gamma_1, \gamma_2 \neq 0$. Set the following parameter values:

$$a = 0 \quad b = 15 \quad \beta_0 = -5 \quad \beta_1 = 1.5 \quad \gamma_0 = 0.05 \quad \gamma_1 = 0.1 \quad \gamma_2 = 0.05$$

N independent samples of n independent observations are generated where $N = 10000$ and $n = 300$. Since observations in a sample are independent of one another it follows that for $\varepsilon = \left(\epsilon_1, \ldots, \epsilon_n\right)^{'}$ one has:

$$\Sigma = E\left[\varepsilon\varepsilon^{'}\right] = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$$

The following graph presents the relation between y and x in a given sample and the standard deviation of the disturbances of that sample. It is evident that for higher values of x the standard deviation of the error term is larger and the dispersion of y around the fitted line (black line in the top panel) increases for larger values of x.



Four different estimator of $\beta = (\beta_0, \beta_1)'$ are considered, the OLS estimator (β_{ols}), the GLS estimator (β_{gls}) that uses the true specification of σ_i^2 and the true parameter values, a F-GLS estimator (β_{fgls}) that uses the true specification of the error variance but has to estimate the parameters, and a F-GLS estimator $\left(\tilde{\beta}\right)$ that has a wrong specification for the error variance. The formulas for the estimators are:

$$\begin{aligned}
\beta_{ols} &= \left(X'X\right)^{-1} X'Y \\
\beta_{gls} &= \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y \\
\beta_{fgls} &= \left(X'\hat{\Sigma}^{-1}X\right)^{-1} X'\hat{\Sigma}^{-1}Y \\
\tilde{\beta} &= \left(X'\tilde{\Sigma}^{-1}X\right)^{-1} X'\tilde{\Sigma}_0^{-1}Y
\end{aligned}$$

Where Σ was already defined, $\hat{\Sigma} = \text{diag}\left(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2\right)$ with $\hat{\sigma}_i^2 = \hat{\gamma}_0 + \hat{\gamma}_1 x_i + \hat{\gamma}_2 x_i^2$ and $\tilde{\Sigma} = \text{diag}\left(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2\right)$ with $\tilde{\sigma}_i^2 = \tilde{\gamma}_0 + \tilde{\gamma}_1 x_i$. Provided e, the vector of OLS residuals, $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)$ is computed from the regression of e^2 on a constant, x and x^2, and $(\tilde{\gamma}_0, \tilde{\gamma}_1)$ is computed from the regression of e^2 on a constant and x.

Note that all the estimator are unbiased for the true β since it is always true that $E[\epsilon|X] = 0$ (by construction of the data). This can be verified with the mean of the estimators across the N samples:

Mean	β	β_{ols}	β_{gls}	β_{fgls}	$\tilde{\beta}$
β_0	-5	-4.9975	-4.9995	-4.9239	-4.8764
β_1	1.5	1.4997	1.5	1.4934	1.4838

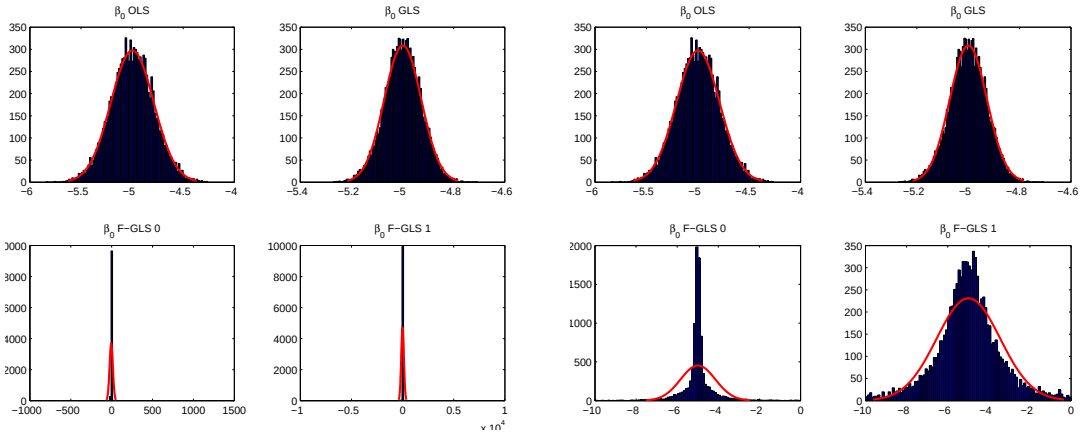
Yet the accuracy of the estimation, measured by the variance of the estimator across samples varies dramatically between estimators:

Std. Dev.	β_{ols}	β_{gls}	β_{fgls}	$\tilde{\beta}$
β_0	0.2099	0.0724	19.9983	129.8905
β_1	0.0302	0.0181	2.2529	13.2272

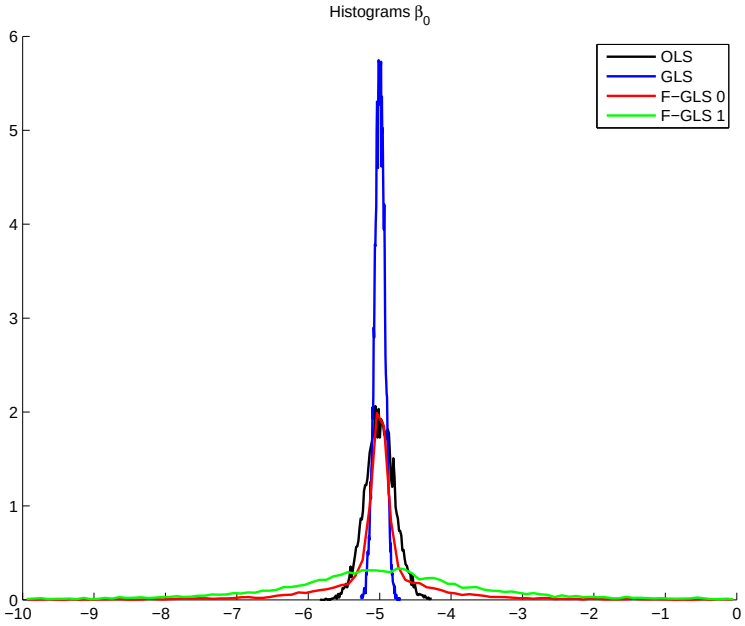
Note that the GLS estimator is by far the most efficient, but that the OLS estimator is (up to two orders of magnitude) more efficient than the F-GLS estimator under the true specification of the error. On the other hand the F-GLS estimator that omits x^2 from the error variance has the highest variance.

The results of this Monte Carlo simulation are in line with the theory referenced above by which F-GLS is only asymptotically efficient, its small sample properties are not determined and hence it might fail to be more efficient than OLS in a given sample. Also note that under a misspecification of the error variance the F-GLS fails to even be consistent, this is represented in the above statics by the large variance of the estimator.

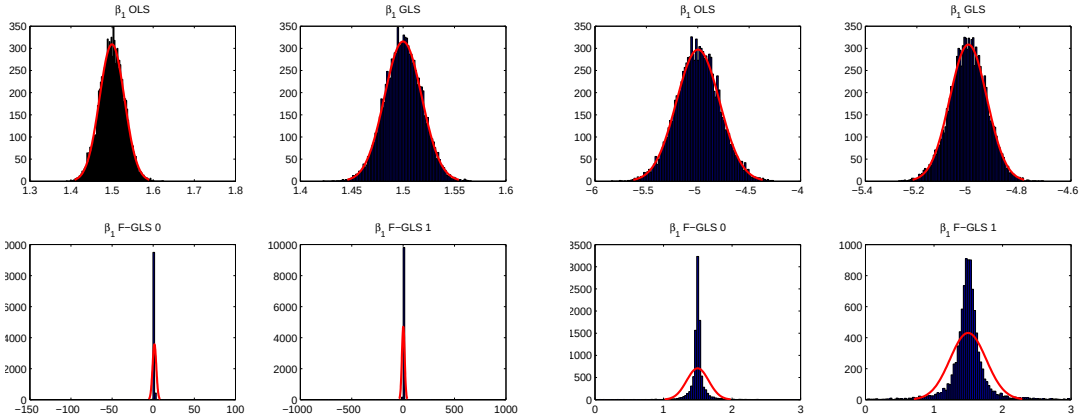
The figure below presents the histograms for β_0 (in its different estimators). The left panel presents the histograms of the whole sample and the right panel censors the sample to consider only realizations of the estimators in the interval $[-10, 0]$. Note that in the left panel the scale of the F-GLS estimators is in the thousands and tens of thousands for the two lower panels.

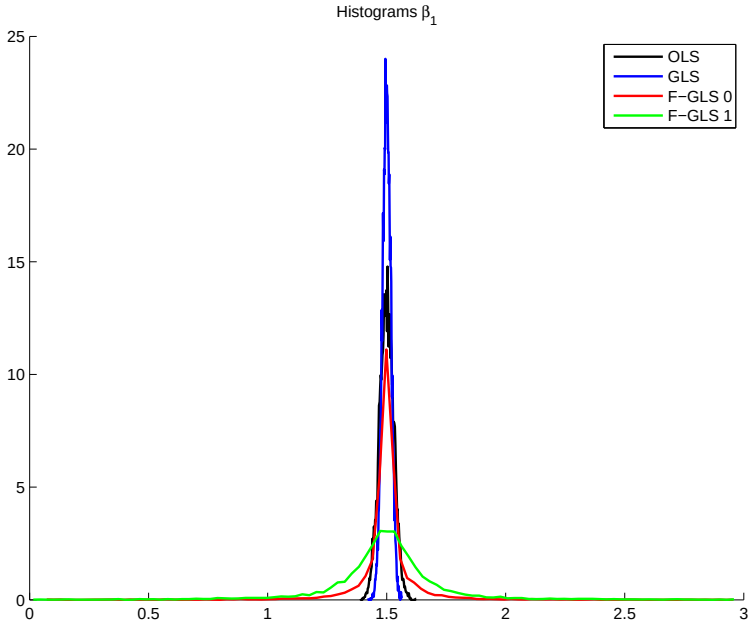


The following figure presents a joint plot of the histograms of the right hand panel. Its clear how the more precise estimator is the GLS estimator (blue line), and even though the mass of estimators of the OLS and F-GLS estimators in the area between -6 and -4 is similar the distribution of the F-GLS estimator presents much heavier tails, that is, it is much more likely to face a sample for which the estimator of β_0 lies far away from the true value. Finally note that the specification error in $\tilde{\beta}$ (green line) makes its distribution flatter and with heavy and long tails, this is how inconsistency looks like, adding more observations to the sample doesn't collapse the distribution on the true value.



Similar results are obtained for β_1. Note that the tails of the F-GLS estimators are more extended than the OLS ones.

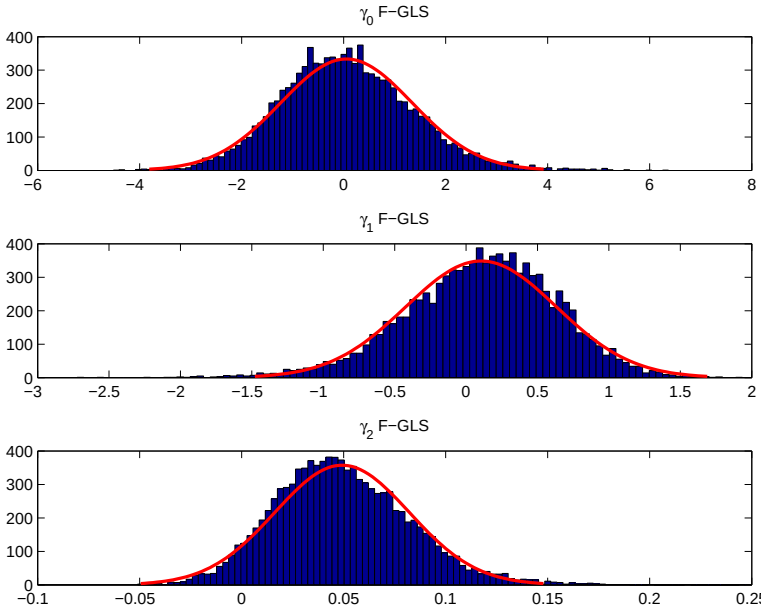




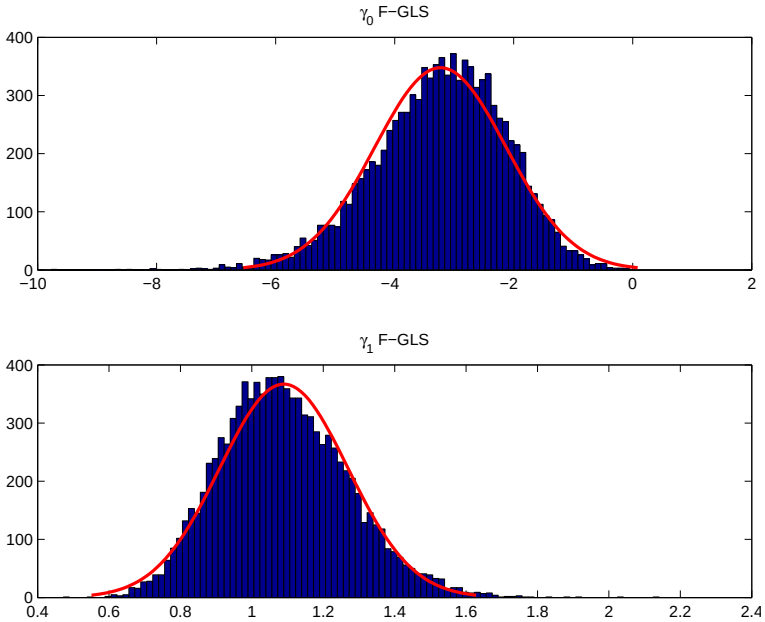
The reason OLS outperforms F-GLS in a given (small) sample is that the task of correctly estimating γ using the residuals from the OLS regression is a difficult one. Yet the estimates for γ are not (on average) far from the truth, the problem consists rather in the lack of precision with which γ can be estimated.

Mean	γ	$\hat{\gamma}$	$\tilde{\gamma}$
γ_0	0.05	0.0551	-3.23
γ_1	0.1	0.1052	1.0896
γ_2	0.05	0.0492	

The following figure presents the histograms for $\hat{\gamma}$. Note that the estimates are skewed and are sometimes negative.



The estimates for $\tilde{\gamma}$ move in a similar way, but due to the misspecification they are biased and also present a higher variance.



9.3 The linear probability model

Consider now a variable $y_i \in \{0,1\}$ such that $P(y=1|x) = x'\beta$. In this model $\beta_j = \frac{\partial P(y=1|x)}{\partial x_j}$ measures the change in the probability of $y=1$ given a change in a given variable x_j.

The following regression model can be used to estimate (consistently) β, provided a random sample $\{y_i, x_i\}_{i=1}^n$

$$y_i = x_i'\beta + \epsilon_i$$

Since y is only 0 or 1 ϵ is then given by:

$$\epsilon = \begin{cases} -x'\beta & \text{if } y=0 \\ 1-x'\beta & \text{if } y=1 \end{cases}$$

Then $\epsilon = -x'\beta$ with probability $1-x'\beta$ and $\epsilon = 1-x'\beta$ with probability $x'\beta$. This implies that the strict exogeneity assumption holds:

$$E[\epsilon|x] = -x'\beta\left(1-x'\beta\right) + \left(1-x'\beta\right)x'\beta = 0$$

It also implies that

$$V[\epsilon|x] = E\left[\epsilon^2\right] = \left(x'\beta\right)^2\left(1-x'\beta\right) + \left(1-x'\beta\right)^2\left(x'\beta\right) = \left(x'\beta\right)\left(1-x'\beta\right)$$

which indicates that the model presents heteroskedasticity. Note that in an *iid* sample $E[\epsilon_i\epsilon_j] = 0$ since x_i and x_j are independent. This gives:

$$\Sigma = \begin{bmatrix} \left(x_1'\beta\right)\left(1-x_1'\beta\right) & & 0 \\ & \ddots & \\ 0 & & \left(x_n'\beta\right)\left(1-x_n'\beta\right) \end{bmatrix}$$

Note: This same result can be obtained by noting that $y|x \sim \text{Bernoulli}\left(x'\beta\right)$ thus $V[y|x] = \left(x'\beta\right)\left(1-x'\beta\right)$ also, since $y = x'\beta + \epsilon$ it follows that $V[y|x] = V\left[x'\beta + \epsilon|x\right] = V[\epsilon|x]$.

Since the form of the heteroskedasticity is known F-GLS can be implemented provided a estimator for $P(y_i=1|x)$, the estimator is given by the OLS fitted value $\hat{y}_i = x_i'\hat{\beta}$. Then:

$$\hat{\Sigma} = \begin{bmatrix} \left(x_1'\hat{\beta}\right)\left(1-x_1'\hat{\beta}\right) & & 0 \\ & \ddots & \\ 0 & & \left(x_n'\hat{\beta}\right)\left(1-x_n'\hat{\beta}\right) \end{bmatrix}$$

9.4 Robust standard errors

An alternative to GLS when the homoskedasticity assumption fails and $E\left[\varepsilon\varepsilon^{'}\right] = \Sigma$ is to use the OLS estimator for β, while providing a correct estimate of the variance of the estimator $V\left[\hat{\beta}\right]$. When homoskedasticity fails the usual formula for the variance and its estimator fail to be true:

$$V\left[\beta_{ols}|X\right] = \sigma^2\left(X^{'}X\right)^{-1} \qquad \widehat{V\left[\beta_{ols}|X\right]} = s^2\left(X^{'}X\right)^{-1}$$

Instead the variance of the estimator is given by:

$$V\left[\beta_{ols}|X\right] = \left(X^{'}X\right)^{-1}\left(X^{'}\Sigma X\right)\left(X^{'}X\right)^{-1}$$

There is however a problem in obtaining an estimate for $V\left[\beta_{ols}|X\right]$ since from a given sample X is available, but, as before, one cannot get an estimate of Σ without extra assumptions (restrictions on the form of Σ), and in almost all cases the true form of Σ is unknown.

With genera homoskedasticity (different variances but zero correlation), matrix Σ has the form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$$

Even in this case there are only n observations available to estimate n unknowns. As before increasing n doesn't solve the problem since the number of parameters to be estimated increases with n.

White (1980) proposes an alternative, it consists on estimating the matrix $V = \frac{1}{n}X^{'}\Sigma X = E\left[\frac{1}{n}X^{'}\varepsilon\varepsilon^{'}X|X\right]$, note the following:

$$\begin{aligned} X^{'}\varepsilon\varepsilon^{'}X &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}\begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}\begin{bmatrix} x_1^{'} \\ \vdots \\ x_n^{'} \end{bmatrix} \\ &= \left(\sum_{i=1}^{n} x_i\epsilon_i\right)\left(\sum_{j=1}^{n} x_j^{'}\epsilon_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j x_i x_j^{'} \end{aligned}$$

Since it is assumed that the ϵ are uncorrelated:

$$E\left[\frac{1}{n}X^{'}\varepsilon\varepsilon^{'}X\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\epsilon_i^2 x_i x_i^{'}\right]$$

So matrix V is an average of expectations, in order to estimate it it is not needed to estimate each expectation separately, instead the average can be estimated as a whole by: $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 x_i x_i^{'}$, yet, since ϵ_i is not observed it can be estimated by:

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i^{'} = \frac{1}{n}X^{'}\mathrm{diag}\left(e^2\right)X$$

This can be computed from any sample and the precision increases as more data becomes available (e_i is a better estimator ϵ_i the more data is available since $\hat{\beta}$ becomes a better estimator of β). Then the

variance of the OLS estimator can be estimated in-sample as:

$$\begin{aligned} \widehat{V[\beta_{ols}|X]} &= \frac{1}{n}\left(\frac{1}{n}X^{'}X\right)^{-1}\hat{V}\left(\frac{1}{n}X^{'}X\right)^{-1} \\ &= \left(X^{'}X\right)^{-1}X^{'}\mathrm{diag}\left(e^2\right)X\left(X^{'}X\right)^{-1} \end{aligned}$$

White shows that under standard conditions:

$$\widehat{V[\beta_{ols}|X]} \to V[\beta_{ols}|X]$$

This variance can be used to construct the usual Wald statistic (and hence the F and t statics) for hypothesis testing.

9.4.1 Algebra with two regressors

Let $y_i = x_i^{'}\beta + \epsilon_i$ with $x_i = (w_i, z_i)^{'}$, then:

$$X^{'}X = \begin{bmatrix} w_1 & \cdots & w_n \\ z_1 & \cdots & z_n \end{bmatrix}\begin{bmatrix} w_1 & z_1 \\ \vdots & \vdots \\ w_n & z_n \end{bmatrix} = \begin{bmatrix} \sum z_i^2 & \sum w_i z_i \\ \sum z_i w_i & \sum z_i \end{bmatrix} = \sum_{i=1}^{n}\begin{bmatrix} z_i^2 & w_i z_i \\ z_i w_i & z_i \end{bmatrix} = \sum_{i=1}^{n} x_i x_i^{'}$$

$$\begin{aligned} X^{'}\varepsilon\varepsilon^{'}X &= \begin{bmatrix} w_1 & \cdots & w_n \\ z_1 & \cdots & z_n \end{bmatrix}\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}\begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}\begin{bmatrix} w_1 & z_1 \\ \vdots & \vdots \\ w_n & z_n \end{bmatrix} \\ &= \begin{bmatrix} \sum w_i\epsilon_i \\ \sum z_i\epsilon_i \end{bmatrix}\begin{bmatrix} \sum w_j\epsilon_j & \sum z_j\epsilon_j \end{bmatrix} \\ &= \begin{bmatrix} \sum\sum w_i w_j \epsilon_i\epsilon_j & \sum\sum w_i z_j \epsilon_i\epsilon_j \\ \sum\sum z_i w_j \epsilon_i\epsilon_j & \sum\sum z_i z_j \epsilon_i\epsilon_j \end{bmatrix} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j\begin{bmatrix} w_i w_j & w_i z_j \\ z_i w_j & z_i z_j \end{bmatrix} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j x_i x_j^{'} \end{aligned}$$

10 Selected Exercises - 8205 Final 2012

10.1 Mean square forecast error

Given n observations generated under the OLS ideal conditions we want to forecast y fora given value of x. I f we refer to this as a forecast of y_{n+1} on x_{n+1}, then given OLS regression coefficient $\hat{\beta}$, a natural point estimate for $E[y_{n+1}|x_{n+1}] = x_{n+1}^{'}\hat{\beta} = \hat{y}$. The forecast error is $\hat{e}_{n+1} = y_{n+1} - \hat{y}_{n+1}$. Calculate the mean square forecast error defines as:

$$\begin{aligned}
E\left[\hat{e}_{n+1}^2|x_{n+1}\right] &= E\left[\left(y_{n+1} - \hat{y}_{n+1}\right)^2\right] \\
&= E\left[\left(x_{n+1}^{'}\left(\beta - \hat{\beta}\right) + \epsilon_{n+1}\right)^2\right] \\
&= E\left[x_{n+1}^{'}\left(\beta - \hat{\beta}\right)\left(\beta - \hat{\beta}\right)^{'} x_{n+1} + x_{n+1}^{'}\left(\beta - \hat{\beta}\right)\epsilon_{n+1} + \epsilon_{n+1}^2\right] \\
&= x_{n+1}^{'}V\left[\hat{\beta}\right]x_{n+1} + x_{n+1}^{'}E\left[\left(\beta - \hat{\beta}\right)\epsilon_{n+1}\right] + \sigma_\epsilon^2 \\
&= x_{n+1}^{'}V\left[\hat{\beta}\right]x_{n+1} + x_{n+1}^{'}E\left[\left(X^{'}X\right)^{-1}X^{'}\varepsilon\epsilon_{n+1}\right] + \sigma_\epsilon^2 \\
&= x_{n+1}^{'}V\left[\hat{\beta}\right]x_{n+1} + x_{n+1}^{'}E_X\left[\left(X^{'}X\right)^{-1}X^{'}E\left[\varepsilon\epsilon_{n+1}|X\right]\right] + \sigma_\epsilon^2 \\
&= x_{n+1}^{'}V\left[\hat{\beta}\right]x_{n+1} + \sigma_\epsilon^2
\end{aligned}$$

10.2 Consistency and asymptotic bias

Consider an estimator $\hat{\beta}_n$ that for any n is equal to β with probability $1 - 1/n$ and equal to n with probability $1/n$.

1. Is this estimator asymptotically unbiased $\left(E\left[\hat{\beta}_n\right] \to \beta\right)$:

$$\begin{aligned}
E\left[\hat{\beta}_n\right] &= \beta\left(1 - \frac{1}{n}\right) + n\left(\frac{1}{n}\right) = \beta\left(1 - \frac{1}{n}\right) + 1 \\
E\left[\hat{\beta}_n\right] &\to \beta + 1
\end{aligned}$$

   The estimator is asymptotically biased. The bias is one.

2. What is the limiting value of the variance of $\hat{\beta}_n$?

$$\begin{aligned}
V\left[\hat{\beta}_n\right] &= E\left[\left(\hat{\beta}_n - E\left[\hat{\beta}_n\right]\right)^2\right] \\
&= \left(1 - \frac{1}{n}\right)\left(\beta - \left(\beta\left(1 - \frac{1}{n}\right) + 1\right)\right)^2 + \frac{1}{n}\left(n - \left(\beta\left(1 - \frac{1}{n}\right) + 1\right)\right)^2 \\
&= \left(1 - \frac{1}{n}\right)\left(\beta\frac{1}{n} - 1\right)^2 + \frac{1}{n}\left(n + \beta\frac{1}{n} - (\beta + 1)\right)^2 \\
&= \left(\beta\frac{1}{n} - 1\right)^2 - \left(\beta\frac{1}{n^{3/2}} - \frac{1}{\sqrt{n}}\right)^2 + \left(\sqrt{n} + \beta\frac{1}{n^{3/2}} - \frac{\beta + 1}{\sqrt{n}}\right)^2 \\
V\left[\hat{\beta}_n\right] &\to \infty
\end{aligned}$$

3. Is $\hat{\beta}_n$ consistent?

$$\forall_{\epsilon<1} \Pr\left\{\left|\hat{\beta}_n - \beta\right| > \epsilon\right\} = \frac{1}{n}$$
$$\lim_{n\to\infty} \Pr\left\{\left|\hat{\beta}_n - \beta\right| > \epsilon\right\} = 0$$

Then the estimator is consistent.

10.3 Gauss-Markov for residuals

Under the assumptions of the Gauss-Markov theorem show that the residual vector $e = M_X Y$ is a BLUE estimator of the stochastic disturbance vector ε.

Since the objective is the disturbance vector ε associated with data Y and X, the data is taken as given, a known non-stochastic quantity.

- First note that e is linear in Y.
- For unbiasedness $e = M_X Y = Y - X\hat{\beta}$. Then $E[e] = Y - XE\left[\hat{\beta}\right] = Y - X\beta = \varepsilon$. Thus e is unbiased.
- $V[e] = E\left[(e-\varepsilon)^2\right] = E\left[\left(Y - X\hat{\beta} - \varepsilon\right)^2\right] = E\left[\left(X\left(\beta - \hat{\beta}\right)\left(\beta - \hat{\beta}\right)' X'\right)\right] = XV\left[\hat{\beta}\right]X'$
- Let $\tilde{\beta}$ be a linear unbiased estimator of β. Then $\tilde{\beta} = C'Y$ where $C'X = I$. Let $\tilde{e} = Y - X\tilde{\beta}$. Since $\tilde{\beta}$ is unbiased for β, so is $\tilde{e}$ for ε.
- $V[\tilde{e}] = E\left[(\tilde{e}-\varepsilon)^2\right] = E\left[\left(Y - X\tilde{\beta} - \varepsilon\right)^2\right] = E\left[\left(X\left(\beta - \tilde{\beta}\right)\left(\beta - \tilde{\beta}\right)' X'\right)\right] = XV\left[\tilde{\beta}\right]X'$
- Consider:
$$V[\tilde{e}] - V[e] = X\left(V\left[\tilde{\beta}\right] - V\left[\hat{\beta}\right]\right)X'$$
Since $\hat{\beta}$ is BLUE the above expression is positive definite, then e is BLUE.

11 Selected Exercises - 8205 Final 2013

11.1 True, partly true or false

$$E[\epsilon_i] = 0 \iff E[\epsilon_i|X_i] = 0$$

The expression is partly true since $E[\varepsilon_i|X_i] = 0 \to E[\epsilon_i] = 0$ but the implication does not hold in the other direction.

- Let $E[\varepsilon_i|X_i] = 0$, then: $E[\epsilon_i] = E_{X_i}[E[\epsilon_i|X_i]] = E_X[0] = 0$.
- Consider X_i such that $X_i = 1$ with probability $1/2$ and $X_i = -1$ with probability $1/2$ so that $E[X_i] = 0$. Let $\epsilon_i = X_i$, then $E[\epsilon_i] = 0$ but $E[\epsilon_i|X_i]$ is either 1 or -1, never equal to zero.

11.2 Time trend in errors

Suppose $y_t = x_t^{'}\beta + \epsilon_t$ and the ideal OLS assumptions hold except that $\epsilon_t = \delta t + u_t$ with $u_t \sim iid\left(0, \sigma^2\right)$. Suppose the data is first differentiated.

1. What is the mean and variance of the new error.

   $\Delta y_t = \Delta x_t \beta + v_t$ where $v_t = \epsilon_t - \epsilon_{t-1} = \delta + u_t - u_{t-1}$. Then $E[v_t] = \delta + E[u_t] - E[u_{t-1}] = \delta$ and $V[v_t] = E\left[\left(u_t - u_{t-1}\right)^2\right] = E\left[u_t^2\right] - 2E[u_t u_{t-1}] + E\left[u_{t-1}^2\right] = 2\sigma^2$

2. Is OLS in this transformed form the BLUE estimator?

   Assumption A1.4 fails to hold then OLS is no longer BLUE. Conditional homoskedasticity implies $E\left[(v-\delta)(v-\delta)^{'}\right] = kI$ for some positive constant k.

$$E\left[(v-\delta)(v-\delta)^{'}\right] = \begin{bmatrix} 2\sigma^2 & -\sigma^2 & 0 & \dots & 0 \\ -\sigma^2 & 2\sigma^2 & -\sigma^2 & & \vdots \\ 0 & -\sigma^2 & 2\sigma^2 & & 0 \\ \vdots & & & \ddots & -\sigma^2 \\ 0 & \dots & 0 & -\sigma^2 & 2\sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & & \vdots \\ 0 & -1 & 2 & & 0 \\ \vdots & & & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{bmatrix}$$

11.3 FWL

Suppose $y = x_1\beta_1 + x_2\beta_2 + u$ where y is $n \times 1$, x_1 is $n \times k_1$, x_2 is $n \times k_2$. Let $\hat{\beta}_1, \hat{\beta}_2$ be the OLS estimates from running this regression. Now consider the following regressions to be estimated by OLS, for which are the estimates of β_2 the same as for the original regression?

1. $y = x_2\beta_2 + u$

   (a) Only if x_2 is orthogonal to x_1 the two estimates are equal, in that case $M_1 x_2 = x_2$ and by FWL the OLS estimate is the same.

2. $P_x y = x_1\beta_1 + x_2\beta_2 + u$

   (a) The OLS estimate is the same for y or for its projection onto the column space of X. Note the following:

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

3. $M_1 y = x_2\beta_2 + u$

(a) Again, only if x_2 is orthogonal to x_1 will the two estimates be the same. Otherwise $M_1 x_2 \neq x_2$.

4. $M_1 y = x_1\beta_1 + M_1 x_2 \beta_2 + u$

   (a) Using FWL the estimate for β_2 doesn't change when estimating $M_1 y$ on $M_1 x_2$, in this case x_1 is an irrelevant variable (note that $M_1 x_1 = 0$) and including it doesn't change the results. Moreover $Cov(x_1, M_1 x_2) = 0$.

   (b) Alternatively by FWL this regression has the same estimates for β_2 as:

$$M_1 M_1 y = M_1 (x_1\beta_1 + M_1 x_2 \beta_2 + u) \longrightarrow M_1 y = M_1 x_2 \beta_2 + v$$

   noting that M_1 is idempotent. The second expression gives the same estimate for β_2 as the original regression by FWL.

5. $P_x y = M_1 x_2 \beta_2 + u$

   (a) As before the estimates from this regression and the regression $y = M_1 x_2 \beta_2 + u$ are the same. By FWL the estimate from the second expression is the same as the original estimate for β_2.

12 8205 Final 2014

12.1 (10 *points*). Collinearity and r^2

Collinearity between the regressors leads to a biased estimate of r^2 in the OLS setting.

Answer: Collinearity doesn't bias the r^2. The r^2 is defined in sample for any set of regressors, including collinear ones. Note that the r^2 only depends on the fitted values (or residuals) and those are always well defined, even in the presence of perfect collinearity, since they are determined by the projection of Y onto the column space of X.
The r^2 is then not biased in any sense since it is always defined for the sample (more collinearity doesn't affect the column space unless it goes to perfect collinearity).

12.2 *(20 points). Strict Exogeneity*

True, Partly True or False: OLS is BLUE if and only if the $kx1$ set of regressors X_i satisfies $Corr(X_i, \varepsilon_i) = 0 \ \forall i$. (That is Corr $\left(X_i^j, \epsilon_i\right) = 0$ for $j = 1, \dots k$).

Answer: Let $\hat{\beta}$ be BLUE under the Gauss Markov assumption. Under the assumption strict exogeneity of the residuals hold, $E[\varepsilon|X] = 0$, this is used for unbiasedness. Note that $Corr(X_i, \epsilon_i) = 0$ iff $Cov(X_i, \epsilon_i) = 0$. and that

$$Cov(X_i, \epsilon_i) = E[(X_i - E[X])\epsilon_i] = E_X[(X_i - E[X])E[\epsilon_i|X]] = 0$$

where the last equality follows from strict exogeneity. Since $Cov(X_i, \epsilon_i) = 0$ one gets $Corr(X_i, \epsilon_i) = 0$ as desired.
Let $Corr(X_i, \epsilon_i) = 0$ for all $i \in \{1, \dots, n\}$. Note that in general this doesn't imply any of the conditions stated in the Gauss Markov theorem. In particular it might be that all but the homoskedasticity assumption hold, but then OLS is not BLUE, it is less efficient than GLS (as shown in class GLS is BLUE).
Moreover one can show that $Corr(X_i, \epsilon_i) = 0$ doesn't imply strict exogeneity $E[\varepsilon|X] = 0$. Consider for example $X_i \sim iid\left(0, \sigma^2\right)$ and $\epsilon_i = X_{i+1}$ for $i \in \{1, \dots, n-1\}$ and $\epsilon_n = X_1$. Since X is iid $Corr(X_i, \epsilon_i) = Corr(X_i, X_{i+1}) = 0$ but $E[\varepsilon|X] = [X_2, X_3, \dots, X_n, X_1]'$. This proves that OLS is not BLUE since strict exogeneity is a necessary condition for OLS to be BLUE and is not implied by $Corr(X_i, \epsilon_i) = 0$.

12.3 *(20 points). Column Spaces*

Prove that $dim(col(X)) = k$ if and only if $rank(X'X) = k$.

Answer: In order to prove this I will establish two results: $\text{rank}(X) = \dim(\text{col}(X))$ and $\text{rank}(X) = k \iff \text{rank}\left(X^{'}X\right) = k$.
The first result follows from the definition of $\text{col}(X) = \{Z \in \mathbb{R}^n | \exists_{\alpha \in \mathbb{R}^k} Z = X\alpha\}$, it is clear that the columns of X form a basis for the space. The dimension of the space is defined as the number of linearly independent vectors in a basis of the space[2] and hence by definition the dimension is equal to the number of linearly independent columns of X, that is the rank of X.
The second result can be established as follows:

- Let $\text{rank}(X) = k$ then for all $\alpha \in \mathbb{R}^k \backslash \{0\}$ one has $X\alpha \neq 0_n$. To show that $X^{'}X$ is of full rank I show that it is a positive definite matrix. Let $\alpha \in \mathbb{R}^k \backslash \{0\}$, consider $\alpha^{'}X^{'}X\alpha = (X\alpha)^{'}(X\alpha) > 0$ where the inequality follows from $X\alpha \neq 0_n$. All positive definite matrix are of full rank.
- Let $\text{rank}\left(X^{'}X\right) = k$ then for all $\alpha \in \mathbb{R}^k \backslash \{0\}$ one has $X^{'}X\alpha \neq 0_k$, also since $X^{'}X$ is of full rank it is invertible, then this implies: $\left(X^{'}X\right)^{-1} X^{'}X\alpha \neq 0$ which is $\alpha \neq 0$. Alternatively one can argue for a contradiction that there exists $\hat{\alpha} \neq 0$ for which $X\hat{\alpha} = 0_n$, that implies that $0_k \neq X^{'}X\alpha = X^{'}0_n = 0_k$ which is a contradiction of $X^{'}X$ being full ranked.

[2] Usually the dimension is defined as the number of vectors in a basis of the space, since vectors in a basis are required to be linearly independent. The definition above is equivalent since one can always add a vector to a basis that is a linear combination of the other vectors without changing the space that the basis spans.

12.4 *(4 points each for 20 total). Interpreting Coefficients*

Consider a set of data where an observation is an individual's writing, math, and reading score on a standardized test, and a dummy variable equal to one when the individual is a female.
A researcher regresses the natural log of the writing score on the other variables and a constant. The writing and math score variables enter the regression in their natural log form and the reading score enters in its level (1-60). The results of the regression are in the table below.
Interpret the coefficient of every variable (including the intercept) and the 95% confidence interval for $\ln(\text{math})$ in the following table. Each interpretation should be one complete sentence that states the *exact* interpretation of the coefficient and nothing more.

$\ln(\text{write})$	Coefficient	Std. Err.	t	$P>\|t\|$	[95% Conf.	Interval]
female	0.1142399	0.0194712	5.87	0	0.07584	0.1526399
$\ln(\text{math})$	0.4085369	0.0720791	5.67	0	0.2663866	0.5506872
read	0.0066086	0.0012561	5.26	0	0.0041313	0.0090859
intercept	1.928101	0.2469391	7.81	0	1.441102	2.415099

Answer: First note that from the FOC of the OLS problem:

$$[\beta_0]0 = \sum_{female}\left(y-\hat{\beta}_0-\hat{\beta}_f-\hat{\beta}_m\ln(math)-\hat{\beta}_r read\right)+\sum_{male}\left(y-\hat{\beta}_0-\hat{\beta}_m\ln(math)-\hat{\beta}_r read\right)$$

$$[\beta_f]0 = \sum_{female}\left(y-\hat{\beta}_0-\hat{\beta}_f-\hat{\beta}_m\ln(math)-\hat{\beta}_r read\right)$$

Joining:

$$\hat{\beta}_0 = \overline{y}_m-\hat{\beta}_m\overline{\ln(math)}_m-\hat{\beta}_r\overline{read}_m$$

$$\hat{\beta}_f = \left(\overline{y}_f-\hat{\beta}_m\overline{\ln(math)}_f-\hat{\beta}_r\overline{read}_f\right)-\hat{\beta}_0$$

Thus, unlike when variables are demeaned, or there are no other explanatory variables besides the dummy, $\hat{\beta}_0$ and $\hat{\beta}_f$ are not by themselves the average score of a male and the difference between the score of a men and women.

- Female: In average a female has a 11.42% higher in score in writing than a male, all the rest equal, that is given a common value for $\ln(math)$ and $read$. (note that I use in average and not in expectation, since $\hat{\beta}$ gives sample instead of population values).
- $\ln(math)$: In average a 1% increase in the math score gives 0.4% more writing score, the effect doesn't distinguish gender.
- Read: In average one extra unit of reading score gives 0.66% more writing score, the effect doesn't distinguish gender.
- Intercept: There is no direct interpretation for the value of the intercept, unless the explanatory variables are demeaned or only the gender dummy is being used. Note that the intercept makes sure that the point formed by the average male is part of the regression. OLS always includes the average in the regression line.
- Confidence Interval: If one had a large number of samples of $\{(x,y)\}$(each of the same size), all drawn from the same population (following the same DGP), and if for each sample one computed $\hat{\beta}_m$ and its confidence interval, 95% of those confidence intervals would contain the true value of β_m. Note that the true value β_m is either inside this particular confidence interval or it isn't, generating more samples without changing the interval won't change this fact since β_m is not changing. Moreover there is no guarantee that the true parameter is in a particular confidence interval. Statements as "the true β is in the interval with 95% probability" are wrong, since the probability that the true β is in the interval is either 0 or 1.

12.5 *(30 points). GLS and WLS*

Suppose that the Data Generating Process (DGP) is that of standard OLS except that the variance of ϵ is equal to 1 for every odd observation and equal to 5 for every even observation. Otherwise all observations are independent of one another.

a) What is the objective function for the BLUE estimator for a fixed sample size of N equal to an even number? Write out explicitly $G_N(\beta)$ *for* $\hat{\beta} = argmin_{\beta \in R^k} G_N(\beta)$.

Answer: Let $E\left[\varepsilon\varepsilon^{'}\right] = \Sigma$. The BLUE estimator is the GLS estimator, $\hat{\beta} = \left(X^{'}\Sigma^{-1}X\right)^{-1} X^{'}\Sigma^{-1}Y$. This estimator solves:

$$\begin{aligned} G_N(\beta) &= \left(P^{-1}Y - P^{-1}X\beta\right)^{'}\left(P^{-1}Y - P^{-1}X\beta\right) \\ &= \left(Y - X\beta\right)^{'} P^{-1'}P^{-1}\left(Y - X\beta\right) \\ &= \left(Y - X\beta\right)^{'}\Sigma^{-1}\left(Y - X\beta\right) \end{aligned}$$

where $\Sigma = PP^{'}$ and $\Sigma^{-1} = P^{-1'}P^{-1}$. In this case we have:

$$\Sigma = \begin{bmatrix} 1 & 0 & \dots & & 0 \\ 0 & 5 & & & \vdots \\ \vdots & & \ddots & & \\ & & & 1 & 0 \\ 0 & \dots & & 0 & 5 \end{bmatrix} \qquad P = \begin{bmatrix} 1 & 0 & \dots & & 0 \\ 0 & \sqrt{5} & & & \vdots \\ \vdots & & \ddots & & \\ & & & 1 & 0 \\ 0 & \dots & & 0 & \sqrt{5} \end{bmatrix}$$

$$P^{-1} = \begin{bmatrix} 1 & 0 & \dots & & 0 \\ 0 & 1/\sqrt{5} & & & \vdots \\ \vdots & & \ddots & & \\ & & & 1 & 0 \\ 0 & \dots & & 0 & 1/\sqrt{5} \end{bmatrix} \qquad \Sigma^{-1} = \begin{bmatrix} 1 & 0 & \dots & & 0 \\ 0 & 1/5 & & & \vdots \\ \vdots & & \ddots & & \\ & & & 1 & 0 \\ 0 & \dots & & 0 & 1/5 \end{bmatrix}$$

b) What does the optimal Weighted Least Squares (WLS) estimator solve in this setting? Write down an expression for the estimator that solves the objective function.

Answer: The WLS estimator solves:

$$G_N(\beta) = \sum_{i=1}^{n} w_i e_i^2 = e^{'}We$$

where $W = \text{diag}\left([w_1, \dots, w_n]\right)$. From above we know that $W = \Sigma^{-1}$ and thus

$$w_i = \begin{cases} 1 & \text{if } i\text{is odd} \\ \frac{1}{5} & \text{if } i\text{is even} \end{cases}$$

The WLS estimator is equal to the GLS estimator. Note that they solve the same objective function.

$$\hat{\beta}_{WLS} = \left(X^{'}WX\right)^{-1} X^{'}WY = \left(X^{'}\Sigma^{-1}X\right)^{-1} X^{'}\Sigma^{-1}Y$$

Now suppose that every two adjacent observations have correlation equal to 0.8944 but otherwise all observations are independent of one another (that is, n=1,2 has the same variance as above, but now they are correlated, and similarly for n=3,4, n=5,6,..., In this setting observations 2 and 3 are not correlated, nor are observations 4 and 5 and so on).

c) Write down the variance covariance matrix $E[\epsilon\epsilon'|X] = \Omega$ for N=10.

Answer: First one needs to establish the covariance between two "adjacent" observations, $\text{Cov}(\epsilon_i, \epsilon_{i+1}) = \rho_{i,i+1}\sqrt{V(\epsilon_i)\,V(\epsilon_{i+1})} = 0.8944\sqrt{5} = 2$. Then:

$$\Omega = \begin{bmatrix} 1 & 2 & 0 & \dots & 0 \\ 2 & 5 & & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & & 1 & 2 \\ 0 & \dots & 0 & 2 & 5 \end{bmatrix}_{10\times 10}$$

d) Consider N=2. For $\Omega = PP'$ one can show in this case that P^{-1} - the square root of Ω^{-1} - is given by

$$P^{-1} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \tag{1}$$

What is the exact form of the objective function in this case? What is it for a sample of size N, with N even? Intuitively why does the weighting take this form?

Answer: From before the objective function (for two observations) is:

$$G_N(\beta) = e^{'}P^{-1'}P^{-1}e = e^{'}\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}e = e^{'}\begin{bmatrix} 5 & -2 \\ -2 & 1 \end{bmatrix}e = 5e_1^2 - 4e_1e_2 + e_2$$

When there are N observations, since there is no covariance between observations that are not "adjacent" the weighted sum of squares is:

$$G_N(\beta) = e^{'}\begin{bmatrix} 5 & -2 & 0 & \dots & 0 \\ -2 & 1 & & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & & 5 & -2 \\ 0 & \dots & 0 & -2 & 1 \end{bmatrix}e = 5e_1^2 - 4e_1e_2 + e_2 + \ldots + 5e_{n-1}^2 - 4e_{n-1}e_n + e_n$$

As in WLS the residuals of observations with less variance are weighted more than the residuals with more variance (odd observations have a direct weight of 5 and even observations of 1). Since "adjacent" observations are positively correlated it is expected that they co-move so that $e_ie_{i+1} > 0$, these co-movements are thus weighted less since they are, in a way, expected to happen. Notice that without the cross term a correlated movement of e_i and e_{i+1} will be counted "twice" while it arises from a single movement of the underlying errors.

e) Consider N=2. Now suppose every two adjacent observations have a negative correlation of -0.8994 but otherwise all observations are independent of one another. In this case

$$P^{-1} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{2}$$

What is the exact form of the objective function in this case? What is it for a sample of size N, with N even? Intuitively why does the weighting take this form?

Answer: From before the objective function (for two observations) is:

$$G_N(\beta) = 5e_1^2 + 4e_1e_2 + e_2$$

When there are N observations, since there is no covariance between observations that are not "adjacent" the weighted sum of squares is:

$$G_N(\beta) = 5e_1^2 + 4e_1e_2 + e_2 + \ldots + 5e_{n-1}^2 + 4e_{n-1}e_n + e_n$$

The intuition is the same as before.

13 GMM

13.1 Theory

13.1.1 Assumptions

1. The setup is the following: there is a linear relation between a dependent variable y and a set of L explanatory variables z the relation is given by:

$$y_i = z_i^{'}\delta + \epsilon_i$$

The interest of the econometrician is to find an estimate for the true value of δ a $L \times 1$ vector of parameters. As opposed to the usual linear regression it is not assumed that $E[\epsilon_i z_i] = 0$ (that is the predetermined regressors assumption). Some regressors can then be endogenous if it holds that $E[\epsilon_i z_{l,i}] \neq 0$.

2. There is also a set of K variables x_i which are referred to as instruments. It is assumed that the process $\{w_i\}$ where $w_i = (y_i, z_i, x_i)$ is jointly stationary. This allows to make inference from continuous functions of the sample that use y, z and x simultaneously.

3. The set of instruments x_i is assumed to be orthogonal to the errors, this is equivalent to the predetermined regressors assumptions, but applied to the instruments and the residuals. Then, defining $g_i = x_i \cdot \epsilon_i$

$$\begin{aligned} \forall_k \quad E[\epsilon_i x_{k,i}] &= 0 \\ E\left[x_i \cdot \left(y_i - z_i^{'}\delta\right)\right] &= 0 \\ E[g_i] &= 0 \end{aligned}$$

4. There is nothing preventing x_i and z_i from sharing variables. If an explanatory variable $z_{l,i}$ already satisfies $E[\epsilon_i z_{l,i}] = 0$ then it can be used as an instrument, then $x_{l,i} = z_{l,i}$. There is however a restriction on how many instruments are needed. The $K \times L$ matrix $E\left[x_i z_i^{'}\right] = \Sigma_{xz}$ must be of full column rank, a necessary condition for that is to have $K \geq L$.

   (a) Calling $\sigma_{xy} = E[x_i \cdot y_i]$ one can express the orthogonality condition as:

$$E\left[x_i \cdot \left(y_i - z_i^{'}\delta\right)\right] = 0 \longrightarrow E[x_i \cdot y_i] = E\left[x_i z_i^{'}\right]\delta \longrightarrow \underset{K\times 1}{\sigma_{xy}} = \underset{K\times L}{\Sigma_{xz}}\underset{L\times 1}{\delta}$$

   It turns out that there is a unique value of δ_0 that satisfies this equation if and only if Σ_{xy} is of full column rank.

5. For further results one can assume some structure over the random variable g_i. It is assumed that g_i is a martingale difference sequence with finite second moments, so that $E\left[g_i g_i^{'}\right] = E\left[\epsilon_i^2 x_i x_i^{'}\right] = S$ is finite.

13.1.2 Estimation

Let $\{y_i, z_i, x_i\}_{i=1}^n$ be a random sample of the variable described above. One can approximate σ_{xy} and Σ_{xz} by their sample counterparts and try to use the orthogonality condition to obtain $\hat{\delta}$. Defining $s_{xy} = \sum x_i \cdot y_i / n$ and $S_{xz} = \sum x_i z_i^{'} / n$ the equation is:

$$S_{xz}\delta = s_{xy}$$

If $K = L$ then by the full column rank condition S_{xz} is invertible (for sufficiently large n) and $\hat{\delta} = S_{xz}^{-1} s_{xy}$. If $K > L$ there might not be a solution for the equation above. GMM doesn't look for an exact match (a $\hat{\delta}$ that solves the equation) but instead for an estimate that makes the expression $g_n(\delta) = S_{xz}\delta - s_{xy}$ as close as possible to zero.

Let W be a positive definite $K \times K$ matrix, a positive definite matrix defines an square form. Then $g_n(\delta)' W g_n(\delta) \geq 0$ and its equal to zero only if $g_n(\delta) = 0$. So minimizing $g_n(\delta)' W g_n(\delta)$ over δ gives a value for $\hat{\delta}$ that brings $g_n(\delta)$ as close as possible to zero.

Consider the objective function:

$$J(\delta, W) = n g_n(\delta)' W g_n(\delta) = n (s_{xy} - S_{xz}\delta)' W (s_{xy} - S_{xz}\delta)$$

The objective function is convex since W is positive definite, then the first order conditions are necessary and sufficient for a solution:

$$\frac{\partial J(\delta, W)}{\partial \delta'} = \left(\frac{\partial g_n(\delta)}{\partial \delta'}\right)\left(\frac{\partial J(\delta, W)}{\partial g_n'}\right) = \left(-S_{xz}'\right)\left(2nW\left(s_{xy} - S_{xz}\hat{\delta}\right)\right) = 0$$

$$S_{xz}' W s_{xy} = S_{xz}' W S_{xz}\hat{\delta}$$

$$\hat{\delta} = \left(S_{xz}' W S_{xz}\right)^{-1}\left(S_{xz}' W s_{xy}\right)$$

Alternatively one can expand the objective function as:

$$\begin{aligned} J(\delta, W) &= n\left(s_{xy}' - \delta' S_{xz}'\right) W (s_{xy} - S_{xz}\delta) \\ &= n\left(s_{xy}' W (s_{xy} - S_{xz}\delta) - \delta' S_{xz}' W (s_{xy} - S_{xz}\delta)\right) \\ &= n\left(s_{xy}' W s_{xy} - s_{xy}' W S_{xz}\delta - \delta' S_{xz}' W s_{xy} + \delta' S_{xz}' W S_{xz}\delta\right) \\ &= n\left(s_{xy}' W s_{xy} - 2\delta' S_{xz}' W s_{xy} + \delta' S_{xz}' W S_{xz}\delta\right) \end{aligned}$$

Then one gets the FOC as:

$$\frac{\partial J(\delta, W)}{\partial \delta'} = 2n\left(-S_{xz}' W s_{xy} + S_{xz}' W S_{xz}\delta\right)$$

The estimate of the GMM procedure can be obtained from the data as:

$$\begin{aligned} \hat{\delta} &= \left(S_{xz}' W S_{xz}\right)^{-1}\left(S_{xz}' W s_{xy}\right) \\ &= \left(\frac{1}{n}\left(Z'X\right) W \frac{1}{n}\left(X'Z\right)\right)^{-1}\left(\frac{1}{n}\left(Z'X\right) W \frac{1}{n}\left(X'y\right)\right) \\ &= \left(\left(Z'X\right) W \left(X'Z\right)\right)^{-1}\left(\left(Z'X\right) W \left(X'y\right)\right) \end{aligned}$$

Where X is $n \times K$, Z is $n \times L$ and y is $n \times 1$.

Also note that:

$$s_{xy} = \frac{1}{n}X'y = \frac{1}{n}X'(Z\delta + \varepsilon) = \frac{1}{n}X'Z\delta + \frac{1}{n}X'\varepsilon = S_{xz}\delta + \overline{g}$$

Where $\overline{g} = \frac{1}{n}\sum x_i \cdot \epsilon_i$, so that:

$$\begin{aligned} \hat{\delta} &= \left(S_{xz}' W S_{xz}\right)^{-1}\left(S_{xz}' W s_{xy}\right) \\ &= \left(S_{xz}' W S_{xz}\right)^{-1} S_{xz}' W (S_{xz}\delta + \overline{g}) \\ &= \delta + \left(S_{xz}' W S_{xz}\right)^{-1} S_{xz}' W \overline{g} \end{aligned}$$

13.1.3 Two stage least squares

The two stage least square estimator follows the following algorithm:

1. Regress (or project) z onto x:

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

2. Regress (or project) y onto $\hat{Z}$ to obtain $\hat{\delta}$:

$$\begin{aligned}
\hat{\delta} &= \left(\hat{Z}'\hat{Z}\right)^{-1}\hat{Z}'y \\
&= \left(Z'P_X P_X Z\right)^{-1} Z'P_X y \\
&= \left(Z'P_X Z\right)^{-1} Z'P_X y \\
&= \left(Z'X\left(X'X\right)^{-1}X'Z\right)^{-1} Z'X\left(X'X\right)^{-1}X'y \\
&= \left(S_{xz}'\left(X'X\right)^{-1}S_{xz}\right)^{-1} S_{xz}'\left(X'X\right)^{-1}s_{xy} \\
&= \left(S_{xz}'S_{xx}^{-1}S_{xz}\right)^{-1} S_{xz}'S_{xx}^{-1}s_{xy}
\end{aligned}$$

There are two things to point out about this procedure. The first one is that all variables in z are being projected onto all variables of x, as opposed to saying that there are some instruments assigned to a particular regressor, that do not affect other regressors. The second one is that this estimator can be obtained in one step as the GMM estimator when $W = S_{xx}^{-1} = \frac{1}{n}\left(X'X\right)^{-1}$:

$$\begin{aligned}
\hat{\delta} &= \left(\hat{Z}'\hat{Z}\right)^{-1}\hat{Z}'y \\
&= \left(Z'P_X P_X Z\right)^{-1} Z'P_X y \\
&= \left(Z'P_X Z\right)^{-1} Z'P_X y \\
&= \left(Z'X\left(X'X\right)^{-1}X'Z\right)^{-1} Z'X\left(X'X\right)^{-1}X'y \\
&= \left(S_{xz}'\left(X'X\right)^{-1}S_{xz}\right)^{-1} S_{xz}'\left(X'X\right)^{-1}s_{xy} \\
&= \left(S_{xz}'S_{xx}^{-1}S_{xz}\right)^{-1} S_{xz}'S_{xx}^{-1}s_{xy}
\end{aligned}$$

13.2 Simulations (Measurement Error)

Consider the following situation: there is a variable y that follows the process:

$$y_i = \beta x_i + \epsilon_i$$

Yet the variable x is not observable, instead one observes $\hat{x}$ which is defined as:

$$\hat{x}_i = x_i + \nu_i$$

There is another variable z that is related to x. It follows the process

$$z_i = \gamma x_i + \eta_i$$

In the above x, ϵ, ν and η are all mean zero independent variables. Independence implies that z_i is correlated with x but not with ν.

An econometrician has access to a sample of size n of $\{y_i, \hat{x}_i, z_i\}$. The OLS estimator for β is:

$$\beta_{ols} = \left(\hat{x}'\hat{x}\right)^{-1}\hat{x}'y = \frac{\sum \hat{x}_i y_i}{\sum \hat{x}_i^2}$$

Note that if one replaces for the DGP:

$$\begin{aligned}\beta_{ols} &= \frac{\frac{1}{n}\sum \hat{x}_i\left(x_i\beta + \epsilon_i\right)}{\frac{1}{n}\sum \hat{x}_i^2} \\ &= \frac{\frac{1}{n}\sum \hat{x}_i x_i}{\frac{1}{n}\sum \hat{x}_i^2}\beta + \frac{\frac{1}{n}\sum \hat{x}_i \epsilon_i}{\frac{1}{n}\sum \hat{x}_i^2}\end{aligned}$$

The p-lim of β_{ols} is:

$$\begin{aligned}\text{plim}\beta_{ols} &= \frac{E\left[\hat{x}_i x_i\right]}{E\left[\hat{x}_i^2\right]}\beta + \frac{E\left[\hat{x}_i \epsilon_i\right]}{E\left[\hat{x}_i^2\right]} \\ &= \frac{E\left[x_i^2\right] + E\left[\nu_i x_i\right]}{E\left[x_i^2 + 2x_i\nu_i + \nu_i^2\right]}\beta + \frac{E\left[x_i\epsilon_i\right] + E\left[\nu_i\epsilon_i\right]}{E\left[x_i^2 + 2x_i\nu_i + \nu_i^2\right]} \\ &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}\beta\end{aligned}$$

So that β_{ols} is biased. The bias is given by:

$$\text{plim}\beta_{ols} - \beta = -\beta\frac{\sigma_\nu^2}{\sigma_x^2 + \sigma_\nu^2}$$

Note that the bias is also present in small samples:

$$\begin{aligned}E\left[\beta_{ols}\right] &= E\left[\frac{\frac{1}{n}\sum \hat{x}_i x_i}{\frac{1}{n}\sum \hat{x}_i^2}\beta + \frac{\frac{1}{n}\sum \hat{x}_i \epsilon_i}{\frac{1}{n}\sum \hat{x}_i^2}\right] \\ &= E\left[\frac{\frac{1}{n}\sum \hat{x}_i x_i}{\frac{1}{n}\sum \hat{x}_i^2}\right]\beta + E\left[\frac{\frac{1}{n}\sum \hat{x}_i \epsilon_i}{\frac{1}{n}\sum \hat{x}_i^2}\right] \\ &= E\left[\frac{\frac{1}{n}\sum \hat{x}_i x_i}{\frac{1}{n}\sum \hat{x}_i^2}\right]\beta + E_{\hat{x}}\left[\frac{\frac{1}{n}\sum \hat{x}_i E\left[\epsilon_i|\hat{x}\right]}{\frac{1}{n}\sum \hat{x}_i^2}\right] \\ &= E\left[\frac{\frac{1}{n}\sum \hat{x}_i x_i}{\frac{1}{n}\sum \hat{x}_i^2}\right]\beta\end{aligned}$$

An alternative to this is to use the variable z to instrument for $\hat{x}$. z is a good instrument because its independent of ϵ and ν but its related to x. The instrumental variable estimator is:

$$\beta_{iv} = S_{z\hat{x}}^{-1} s_{zy} = \left(z^{'}\hat{x}\right)^{-1} z^{'}y = \frac{\sum z_i y_i}{\sum z_i \hat{x}_i}$$

$$\beta_{iv} = \frac{\sum z_i x_i}{\sum z_i \hat{x}_i}\beta + \frac{\sum z_i \epsilon_i}{\sum z_i \hat{x}_i}$$

This estimator is biased in small samples:

$$E\left[\beta_{iv}\right] = E\left[\frac{\sum z_i x_i}{\sum z_i \hat{x}_i}\right]\beta + E_{z,\hat{x}}\left[\frac{\sum z_i E\left[\epsilon_i | z, \hat{x}\right]}{\sum z_i \hat{x}_i}\right] = E\left[\frac{\sum z_i x_i}{\sum z_i \hat{x}_i}\right]\beta$$

But its consistent:

$$\text{plim}\beta_{iv} = \frac{E\left[z_i x_i\right]}{E\left[z_i \hat{x}_i\right]}\beta + \frac{E\left[z_i \epsilon_i\right]}{E\left[z_i \hat{x}_i\right]} = \frac{E\left[z_i x_i\right]}{E\left[z_i x_i\right] + E\left[z_i \nu_i\right]}\beta = \frac{E\left[z_i x_i\right]}{E\left[z_i x_i\right]}\beta = \beta$$

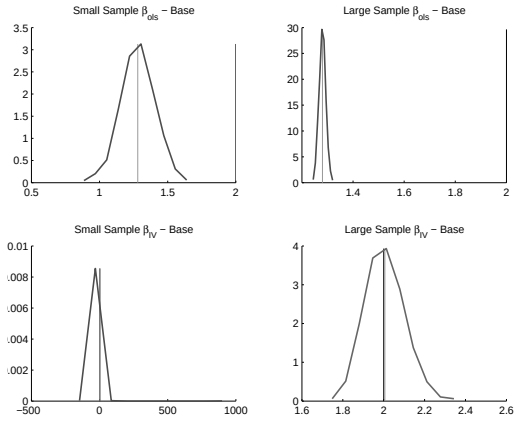
A simulation of the model described was carried out with the following parameters:

$$\beta = 2 \quad \gamma = 0.01 \quad \sigma_x = 0.2 \quad \sigma_\epsilon = 0.2 \quad \sigma_\nu = 0.15 \quad \sigma_\eta = 0.01$$

First 1000 samples of 100 observations where simulated and then 1000 samples of 10000 observations where simulated. The table below shows the mean of the estimators for the small and large samples using OLS and IV. The bias in the small sample is evident. Also the bias in the large sample is contrasted with the theoretical value of the value obtained above.

	β	Small Sample	Large Sample	Bias	T. Bias
OLS	2	1.2802	1.2803	-0.7197	-0.7200
IV	2	2.6278	2.0056	0.0056	-0.7200

The figure below shows the histograms for the simulation. The first row has the results for OLS estimation in the small sample and the large sample, its clear that β_{ols} is converging to the wrong value. The second row presents the results for the IV estimation. The small sample IV has a huge variance, so that the estimator takes on some very extreme values, although most of the values are concentrated in the proximities of the true parameter. The large sample properties are much better and one can see that the estimator is consistent for β.



The code below was used to generate the simulation.

```
% Econometrics 8206
% Monte Carlo to assess Measurement Error
% Sergio Ocampo Diaz

%% Parameter Values

bet = 2    ; % Parameter for regression
gam = 0.01 ; % Parameter for Instrument

s_x   = 0.2 ; % Sigma X
s_eps = 0.2 ; % Sigma epsilon
s_v   = 0.15 ; % Sigma V
s_eta = 0.01 ; % Sigma eta

N = 1000 ; % Number of samples
n_vec = [100,10000]; % Sample sizes

%% Small and large Sample Properties

beta_ols = NaN(N,2) ;
beta_iv  = NaN(N,2);

for j=1:2
    n=n_vec(j); % Sample Size
    x_tot   = s_x   * randn(n,N) ; % Matrix of real x
    eps_tot = s_eps * randn(n,N) ; % Matrix of epsilon
    v_tot   = s_v   * randn(n,N) ; % Matrix of v
    eta_tot = s_eta * randn(n,N) ; % Matrix of eta

    y_tot   = bet * x_tot + eps_tot ; % Dependent variable
    x_obs   = x_tot + v_tot         ; % Observerd x with measurement error
    z_tot   = gam * x_tot + eta_tot ; % Instrument

    for i=1:N
        beta_ols(i,j) = (x_obs(:,i)'*x_obs(:,i))\x_obs(:,i)'*y_tot(:,i);
        beta_iv(i,j)  = (z_tot(:,i)'*x_obs(:,i))\z_tot(:,i)'*y_tot(:,i);
    end
end

%% Moments and Distribution

% Mean of OLS estimator (small sample)
    mean_beta_ols_ss = mean(beta_ols(:,1)) ;
% Mean of OLS estimator (large sample)
    mean_beta_ols_ls = mean(beta_ols(:,2)) ;
% Mean of IV estimator (small sample)
    mean_beta_iv_ss = mean(beta_iv(:,1)) ;
% Mean of IV estimator (large sample)
    mean_beta_iv_ls = mean(beta_iv(:,2)) ;
% Asympthotic bias of OLS
    bias_ols = mean_beta_ols_ls - bet ;
    bias_iv  = mean_beta_iv_ls  - bet ;
```

```
% Attenuation bias
    attenaution_bias = -bet*s_v^2/(s_x^2+s_v^2) ;

% Reporting
    b_mat = [ bet mean_beta_ols_ss mean_beta_ols_ls bias_ols attenaution_bias ;
              bet mean_beta_iv_ss  mean_beta_iv_ls  bias_iv  attenaution_bias ];
    col={'','True','Small Sample','Large Sample','Bias','Bias Formula'};
    row={'OLS';'IV'};
    D=[col; [row , num2cell(b_mat)]];
    disp('Parameters Mean')
    disp(D)

%% Plotting

figure;
    subplot(2,2,1); hold on;
        [a,b] = hist(beta_ols(:,1)) ;
        line([bet bet],[0 max(a/N/diff(b(1:2)))],'color',[0 0 0])
        line([mean_beta_ols_ss mean_beta_ols_ss],[0 max(a/N/diff(b(1:2)))],'color',[0.6 0.6 0.6])
        plot(b,a/N/diff(b(1:2)),'color',[0.3 0.3 0.3],'linewidth',1.5) ;
        title('Small Sample \beta_{ols} - Base')
        hold off;
    subplot(2,2,2); hold on;
        [a,b] = hist(beta_ols(:,2)) ;
        line([bet bet],[0 max(a/N/diff(b(1:2)))],'color',[0 0 0])
        line([mean_beta_ols_ls mean_beta_ols_ls],[0 max(a/N/diff(b(1:2)))],'color',[0.6 0.6 0.6])
        plot(b,a/N/diff(b(1:2)),'color',[0.3 0.3 0.3],'linewidth',1.5) ;
        title('Large Sample \beta_{ols} - Base')
        hold off;
    subplot(2,2,3); hold on;
        [a,b] = hist(beta_iv(:,1)) ;
        line([bet bet],[0 max(a/N/diff(b(1:2)))],'color',[0 0 0])
        line([mean_beta_iv_ss mean_beta_iv_ss],[0 max(a/N/diff(b(1:2)))],'color',[0.6 0.6 0.6])
        plot(b,a/N/diff(b(1:2)),'color',[0.3 0.3 0.3],'linewidth',1.5) ;
        title('Small Sample \beta_{IV} - Base')
        hold off;
    subplot(2,2,4); hold on;
        [a,b] = hist(beta_iv(:,2)) ;
        line([bet bet],[0 max(a/N/diff(b(1:2)))],'color',[0 0 0])
        line([mean_beta_iv_ls mean_beta_iv_ls],[0 max(a/N/diff(b(1:2)))],'color',[0.6 0.6 0.6])
        plot(b,a/N/diff(b(1:2)),'color',[0.4 0.4 0.4],'linewidth',1.5) ;
        title('Large Sample \beta_{IV} - Base')
        hold off;
figurename='Beta_Base.pdf'; print('-dpdf',figurename)
```

14 ME-GMM and Homoskedasticity

14.1 ME-GMM Assumptions

Consider a set of variables $y = [y_1, \ldots, y_M]^{'}$, each of those variables, say y_m, relates to some explanatory variables $z_m = \left[z_m^{(1)}, \ldots, z_m^{(L_m)}\right]^{'}$ in a linear fashion, so that:

$$y_{m,i} = z_{m,i}^{'}\delta_m + \epsilon_{m,i}$$

where ϵ_i is a scalar.

Note that each variable y_m has its own vector of parameters, explanatory variables and error term, so that in principle there are no restrictions on each z_m being different, and even if there are common variables there are no restrictions on the parameters δ_m that relate them with y_m.

As in the GMM framework there are no restrictions over the relation between z_m and ϵ_m.

Each variable y_m also has associated a set of "instruments" $x_m = \left[x_m^{(1)}, \ldots, x_m^{(K_m)}\right]^{'}$, it is assumed that the whole set $\{y_m, z_m, x_m\}_{m=1}^{M}$ is jointly ergodic stationary. This is a much stronger assumption than that of asking each $\{y_m, z_m, x_m\}$ to be individually ergodic stationary, the strengthening of the condition is needed since in ME-GMM there will be interactions of variables across equations, inference has to be made over those interactions.

As in GMM the instruments are assumed to be orthogonal to the errors, so that for each $m \in \{1, \ldots, M\}$ $E[x_{m,i} \cdot \epsilon_{m,i}] = 0$ instruments only have to be valid for their own equation, an instrument of one equation is allowed to have correlation with the error of another equation. Then there are as many orthogonality conditions as there are instruments, a total of $\sum_m K_m$. Define

$$g_i = \begin{bmatrix} x_{1,i} \cdot \epsilon_{1,i} \\ \vdots \\ x_{M,i} \cdot \epsilon_{M,i} \end{bmatrix}$$

Then the moment conditions are $E[g_i] = 0_{\sum K_m \times 1}$. As before one can use the linearity of the model to get a better expression for $E\left[g_i\left(w_i, \overline{\delta}\right)\right]$ where $w_i = \{y_{m,i}, z_{m,i}, x_{m,i}\}_{m=1}^{M}$ is the collection of all data (for the i^{th} observation) and $\overline{\delta} = \left[\overline{\delta}_1^{'}, \ldots, \overline{\delta}_M^{'}\right]^{'}$ is a vector with all the individual parameter vectors stacked. The expression is:

$$\begin{aligned}
E\left[g_i\left(w_i, \overline{\delta}\right)\right] &= E\begin{bmatrix} x_{1,i} \cdot \left(y_{1,i} - z_{1,i}^{'}\overline{\delta}_1\right) \\ \vdots \\ x_{M,i} \cdot \left(y_{M,i} - z_{M,i}^{'}\overline{\delta}_M\right) \end{bmatrix} = \begin{bmatrix} E[x_{1,i} \cdot y_{1,i}] \\ \vdots \\ E[x_{M,i} \cdot y_{M,i}] \end{bmatrix} - \begin{bmatrix} E\left[x_{1,i}z_{1,i}^{'}\right]\overline{\delta}_1 \\ \vdots \\ E\left[x_{M,i}z_{M,i}^{'}\right]\overline{\delta}_M \end{bmatrix} \\
&= \begin{bmatrix} E[x_{1,i} \cdot y_{1,i}] \\ \vdots \\ E[x_{M,i} \cdot y_{M,i}] \end{bmatrix} - \begin{bmatrix} E\left[x_{1,i}z_{1,i}^{'}\right] & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & E\left[x_{M,i}z_{M,i}^{'}\right] \end{bmatrix} \begin{bmatrix} \overline{\delta}_1 \\ \vdots \\ \overline{\delta}_M \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{xy}^{(1)} \\ \vdots \\ \sigma_{xy}^{(M)} \end{bmatrix} - \begin{bmatrix} \Sigma_{xz}^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_{xz}^{(M)} \end{bmatrix} \begin{bmatrix} \overline{\delta}_1 \\ \vdots \\ \overline{\delta}_M \end{bmatrix} \\
&= \sigma_{xy} - \Sigma_{xz}\overline{\delta}
\end{aligned}$$

Just as before it is needed t hat Σ_{xz} has full column rank, this holds if and only $\Sigma_{xz}^{(m)}$ is of full column rank for all m. Note that this is the same as having each individual equation satisfy the full rank condition. In order to compute ME-GMM its necessary to be able to compute GMM in each equation separately.

Conditions over g_i can be imposed to be able to conduct inference and get large sample properties.

14.2 Estimation

As before the objective is to find a $\hat{\delta}$ such that $\sigma_{xy} - \Sigma_{xz}\hat{\delta} = 0$. In a given sample σ_{xy} and Σ_{xz} are replaced by their sample counterparts, given a sample of size n:

$$s_{xy} = \begin{bmatrix} s_{xy}^{(1)} \\ \vdots \\ s_{xy}^{(M)} \end{bmatrix} \qquad S_{xz} = \begin{bmatrix} S_{xz}^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S_{xz}^{(M)} \end{bmatrix}$$

where $s_{xy}^{(m)} = \frac{1}{n} X_m^{'} y_m$ and $S_{xz}^{(m)} = \frac{1}{n} X_m^{'} Z_m$. As before for a given sample there might not be a $\hat{\delta}$ such that $S_{xy} - S_{xz}\hat{\delta} = 0$ so instead one looks for

$$\hat{\delta} = \underset{\overline{\delta}}{\operatorname{argmin}} J\left(\overline{\delta}, W\right) = \underset{\hat{\delta}}{\operatorname{argmin}} \frac{1}{n} \left(S_{xy} - S_{xz}\overline{\delta}\right)^{'} W \left(S_{xy} - S_{xz}\overline{\delta}\right)$$

There are only two differences between this and the usual GMM, first the dimension of the parameter vector is not L (as before) but $\sum L_m$, second the dimension (and composition) of the weighting matrix W has changed, now:

$$W = \begin{bmatrix} W_{11} & \cdots & W_{1M} \\ \vdots & \ddots & \vdots \\ W_{M1} & \cdots & W_{MM} \end{bmatrix}$$

The estimator is then analogous to that of GMM:

$$\hat{\delta} = \left(S_{xz}^{'} W S_{xz}\right)^{-1} S_{xz}^{'} W s_{xy}$$

Equations are only related to each other through W, so if W is a block diagonal matrix the estimator for each $\hat{\delta}_m$ boils down to:

$$\hat{\delta}_m = \left(S_{xz}^{(m)'} W_{mm} S_{xz}^{(m)}\right)^{-1} S_{xz}^{(m)'} W_{mm} s_{xy}^{(m)}$$

the same expression as in the equation by equation GMM.

Generally one would want to use the efficient ME-GMM estimator, as before its given by $W = S^{-1}$ where

$$S = \begin{bmatrix} E\left[\epsilon_{1,i}^2 x_{1,i} x_{1,i}^{'}\right] & \cdots & E\left[\epsilon_{1,i}\epsilon_{M,i} x_{1,i} x_{M,i}^{'}\right] \\ \vdots & \ddots & \vdots \\ E\left[\epsilon_{M,i}\epsilon_{1,i} x_{M,i} x_{1,i}^{'}\right] & \cdots & E\left[\epsilon_{M,i}^2 x_{M,i} x_{M,i}^{'}\right] \end{bmatrix}$$

For a given sample one can use the estimator of S given by $\hat{S}$, where

$$\hat{S} = \frac{1}{n} \sum \begin{bmatrix} \hat{\epsilon}_{1,i}^2 x_{1,i} x_{1,i}^{'} & \cdots & \hat{\epsilon}_{1,i}\hat{\epsilon}_{M,i} x_{1,i} x_{M,i}^{'} \\ \vdots & \ddots & \vdots \\ \hat{\epsilon}_{M,i}\hat{\epsilon}_{1,i} x_{M,i} x_{1,i}^{'} & \cdots & \hat{\epsilon}_{M,i}^2 x_{M,i} x_{M,i}^{'} \end{bmatrix}$$

A consistent estimate of $\hat{\epsilon}_{m,i}$ can be obtained from $\hat{\epsilon}_{m,i} = y_{m,i} - z_{m,i}^{'}\tilde{\delta}_m$ where $\tilde{\delta}$ is a consistent estimator of δ_m, a candidate for $\tilde{\delta}$ is the (single equation) GMM estimator (2SLS or efficient GMM).

Note that single equation GMM was already consistent, what is to gain by ME-GMM is to make the estimator more efficient by using the information in cross moments across equations. If S is block diagonal there are no efficiency gains.

14.3 Conditional Homoskedasticity

Since the efficient ME-GMM estimator depends on the structure of S the form of the variance of the errors across equations is relevant for the method. In the (very) special case of conditional homoskedasticity the formulas given above simplify and allow for an easier computation of the results and, more importantly, they allow to estimate less parameters.

Conditional homoskedasticity in this environment is:

$$E\left[\epsilon_{m,i}\epsilon_{h,i}|x_{m,i}, x_{h,i}\right] = \sigma_{mh} \qquad m, h \in \{1, \ldots, M\}$$

This implies that :

$$S = \begin{bmatrix} \sigma_{11}^2 E\left[x_{1,i}x_{1,i}^{'}\right] & \cdots & \sigma_{1M}^2 E\left[x_{1,i}x_{M,i}^{'}\right] \\ \vdots & \ddots & \vdots \\ \sigma_{M1}^2 E\left[x_{M,i}x_{1,i}^{'}\right] & \cdots & \sigma_{MM}^2 E\left[x_{M,i}x_{M,i}^{'}\right] \end{bmatrix}$$

14.3.1 FIVE - Full information instrumental variable efficient (estimator)

The FIVE estimator takes advantage of the form of S under conditional homoskedasticity to get:

$$\hat{S}_{FIVE} = \frac{1}{n}\sum \begin{bmatrix} \hat{\sigma}_{11}^2 x_{1,i}x_{1,i}^{'} & \cdots & \hat{\sigma}_{1M}^2 x_{1,i}x_{M,i}^{'} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{M1}^2 x_{M,i}x_{1,i}^{'} & \cdots & \hat{\sigma}_{MM}^2 x_{M,i}x_{M,i}^{'} \end{bmatrix}$$

Where $\hat{\sigma}_{mh} = \frac{1}{n}\sum \hat{\epsilon}_{m,i}\hat{\epsilon}_{h,i}$ and, as before, $\hat{\epsilon}_{m,i}$ can be obtained from $\hat{\epsilon}_{m,i} = y_{m,i} - z_{m,i}^{'}\tilde{\delta}_m$ where $\tilde{\delta}$ is a consistent estimator of δ_m, a candidate for $\tilde{\delta}$ is the (single equation) GMM estimator (2SLS or efficient GMM).

The FIVE estimator is the ME-GMM estimator using $W = \hat{S}_{FIVE}^{-1}$. It inherits all the asymptotic properties of the ME-GMM estimator and is asymptotically efficient under conditional homoskedasticity.

14.3.2 3SLS - Three stage least squares (estimator)

Now suppose that besides conditional homoskedasticity the set of instruments across equations is the same, so that $x_m = x_h = x$ for all m and h. In this case $E\left[x_{m,i}x_{h,i}^{'}\right] = E\left[x_i x_i^{'}\right]$ and:

$$S = \begin{bmatrix} \sigma_{11}^2 E\left[x_i x_i^{'}\right] & \cdots & \sigma_{1M}^2 E\left[x_i x_i^{'}\right] \\ \vdots & \ddots & \vdots \\ \sigma_{M1}^2 E\left[x_i x_i^{'}\right] & \cdots & \sigma_{MM}^2 E\left[x_i x_i^{'}\right] \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1M}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{M1}^2 & \cdots & \sigma_{MM}^2 \end{bmatrix} \otimes E\left[x_i x_i^{'}\right] = \Sigma \otimes E\left[x_i x_i^{'}\right]$$

The 3SLS estimator takes advantage of this by computing:

$$\hat{S}_{3SLS} = \hat{\Sigma} \otimes \left(\frac{1}{n}\sum x_i x_i^{'}\right) = \hat{\Sigma} \otimes \left(\frac{1}{n}X^{'}X\right)$$

where

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{11}^2 & \cdots & \hat{\sigma}_{1M}^2 \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{M1}^2 & \cdots & \hat{\sigma}_{MM}^2 \end{bmatrix}$$

and $\hat{\sigma}_{mh} = \frac{1}{n}\sum \hat{\epsilon}_{m,i}\hat{\epsilon}_{h,i}$ is computed as before. The 3SLS estimator is the ME-GMM estimator using $W = \hat{S}_{3SLS}^{-1}$. It inherits all the asymptotic properties of the ME-GMM estimator and is asymptotically efficient under conditional homoskedasticity. The term 3SLS follows from using the 2SLS estimator to obtain $\hat{\Sigma}$.

14.3.3 SUR - Seemingly unrelated equations (estimator)

If all regressors are exogenous one can make use the z variables as instruments, then $x = \cup z_m$. The SUR estimator is an alternative to OLS that aims to gain in efficiency by using the relation between equations. Note that for this its necessary not only that $E[z_{m,i}\epsilon_{im}] = 0$ but also that $E[z_{m,i}\epsilon_{h,i}] = 0$ for all m and h, this follows since all z variables are going to be used as instruments for all equations.

This is just a special case of the 3SLS estimator, the formulas for the asymptotic variance and for the estimator are somewhat simplified.

15 Maximum Likelihood (Davidson & MacKinnon)

An alternative estimation method (to GMM) is maximum likelihood (ML), provided a complete characterization of the DGP of the data ML estimation provides consistent and asymptotically normal (CAN) estimators under few additional conditions, it also provides non-linear estimators and their properties in a unified manner.

The main drawback of ML estimation is that it requires a complete specification of the data through a parametric model. This means that given some set of parameters the probability distribution of the data is known. This distribution is represented in the likelihood function:

$$L(\theta|y) = f(y/\theta)$$

where $f(\cdot|\theta)$ is the PDF of y given θ. Note that one θ is known the PDF of y is known too, and that the PDF completely characterizes the behavior of y as a random variable, also note that y can be a vector valued random variable. When the PDF is evaluated at a certain y (say $y = [y_1, \ldots, y_n]'$ a random sample) it can no longer be interpreted as a PDF, instead it gives the probability of observing the given data for any value of θ (the parameters). The ML procedure seeks to find a value for θ that maximizes the probability of observing the observed sample.

The maximum likelihood estimator (MLE) of theta, $\hat{\theta}$, is the value of θ for which $L(\theta|y)$ is maximized. for $\hat{\theta}$ to be the MLE it must be the unique (global) maximizer of L. Note that this is equivalent to ask $\hat{\theta}$ to be the maximizer of $l(\theta|y) = \ln L(\theta|y)$ since the logarithm is a monotone increasing function. All the results for ML estimation are obtained with the log-likelihood function.

Note that in general the random variable of interest is a vector (since there are several observations of the realization of y), the likelihood is given by the joint probability of y which is in general a complicated object. In many cases if the realizations of y are independent (although perhaps not identically distributed) and then the joint distribution of y is given by:

$$f(y|\theta) = \prod_{i=1}^{n} f_i(y_i|\theta)$$

In this case the likelihood of the sample is obtained as the product of the likelihood of each observation. The log-likelihood is then:

$$l(y|\theta) = \sum_{i=1}^{n} l_i(y_i|\theta)$$

There is another special case that arises frequently in ML estimation, variables with serial correlation (for example if y follows an AR(p) process), in this case one can use the fact that the joint distribution of two variables can be written as:

$$f(y_1, y_2) = f(y_1) f(y_2|y_1)$$

Using this equality successively one gets:

$$f(y|\theta) = f(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} f\left(y_i|y^{i-1}, \theta\right)$$

Note that this requires conditioning in an initial history y^0. The log-likelihood is:

$$l(\theta|y) = \sum_{i=1}^{n} l_i\left(\theta|y^i\right)$$

where $l_i\left(y^i|\theta\right) = f\left(y_i|y^{i-1}, \theta\right)$. Note that this generalizes the case above allowing each term in the sum of likehoods to depend on the history up that observation and not only on the current realization of y.

15.1 Types of maximum likelihood estimators

Since there are two ways of characterizing the maximizer of a function there are two types of MLE. A maximizer can be defined either directly as:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} l\left(y|\theta\right)$$

or by means of the first order conditions of the problem, when the objective function differentiable and satisfies the second order conditions for FOC to be sufficient, in this case $\hat{\theta} \in \Theta$ is such that:

$$\frac{\partial l\left(\hat{\theta}|y\right)}{\partial \theta} = 0$$

For most cases an MLE is of both types. A type 2 MLE also requires that as $n \to \infty$ $l\left(\hat{\theta}|y\right) > l\left(\theta|y\right)$ for all $\theta \in \Theta$.

15.1.1 Score vector

The score vector plays an important role in ML theory, it is defined as the gradient of the likelihood function with respect to the parameters, that is, as the vector of first order conditions:

$$g\left(\theta, y\right) = \frac{\partial l\left(\theta|y\right)}{\partial \theta}$$

This vector has the same dimension as θ, the if there are K parameters $g : \mathbb{R}^K \to \mathbb{R}$.

Let $l\left(\theta|y\right) = \sum_{i=1}^{n} l_i\left(\theta|y^i\right)$ and define a $K \times n$ matrix G such that:

$$G_{k,i}\left(\theta, y^i\right) = \frac{\partial l_i\left(\theta|y^i\right)}{\partial \theta_k}$$

It follows that:

$$g_k\left(\theta, y\right) = \sum_{i=1}^{n} G_{k,i}\left(\theta, y^i\right)$$

Under the true GDP, that is, if y was actually generated by a GDP characterized by $f\left(y|\theta\right)$, the expectations of all the elements of G are zero at the true parameter θ_0. This is used to establish properties of the score vector.

To see this note that

$$\int e^{l_i\left(\theta_0|y^i\right)} dy_i = \int f\left(y_i|y^{i-1}, \theta_0\right) dy_i = 1$$

then differentiating both sides with respect to θ_k one gets:

$$\begin{aligned}
\frac{\partial \int e^{l_i\left(\theta_0|y^i\right)} dy_i}{\partial \theta_k} &= 0 \\
\int \frac{\partial e^{l_i\left(\theta_0|y^i\right)}}{\partial \theta_k} dy_i &= 0 \\
\int e^{l_i\left(\theta_0|y^i\right)} \frac{\partial l_i\left(\theta_0|y^i\right)}{\partial \theta_k} dy_i &= 0 \\
\int f\left(y_i|y^{i-1}, \theta_0\right) \frac{\partial l_i\left(\theta_0|y^i\right)}{\partial \theta_k} dy_i &= 0 \\
E_{y_i}\left[\frac{\partial l_i\left(\theta_0|y^i\right)}{\partial \theta_k}\right] &= 0 \\
E_{y_i}\left[G_{k,i}\left(\theta_0, y^i\right)|y^{i-1}\right] &= 0
\end{aligned}$$

Taking expectations over this with respect to y^{i-1} gives:

$$E\left[G_{k,i}\left(\theta_0, y^i\right)\right] = 0$$

Then:

$$E\left[g\left(\theta_0, y\right)\right] = 0$$

The covariance of the elements of the matrix G can also be computed. Consider $G_{ik}\left(\theta_0, y^i\right)$ and $G_{jl}\left(\theta_0, y^j\right)$, wlog let $i < j$ (note $i \neq j$), then the covariance between the two terms is:

$$\begin{aligned} E\left[G_{ik}\left(\theta_0, y^i\right) G_{jl}\left(\theta_0, y^j\right)\right] &= E\left[E\left[G_{ik}\left(\theta_0, y^i\right) G_{jl}\left(\theta_0, y^j\right)|y^i\right]\right] \\ &= E\left[G_{ik}\left(\theta_0, y^i\right) E\left[G_{jl}\left(\theta_0, y^j\right)|y^i\right]\right] \\ &= E\left[G_{ik}\left(\theta_0, y^i\right) E\left[E\left[G_{jl}\left(\theta_0, y^j\right)|y^{j-1}\right]\right]\right] \\ &= 0 \end{aligned}$$

Note that conditional on y^i the first term $G_{ik}\left(\theta_0, y^i\right)$ is known, and that $E\left[G_{jl}\left(\theta_0, y^j\right)|y^{j-1}\right] = 0$ from above, and so it is also 0 conditional on the subset of observations y^i. Knowing the covariance of the terms of G allows to compute the covariance matrix of the score vector. That is given by the information matrix.

15.1.2 Information matrix

Let I_i be the $K \times K$ covariance matrix of the elements of the i^{th} column of G (it has already been shown that there is no covariance between columns):

$$I_i\left(\theta_0\right) = E\left[G_i\left(\theta_0, y^i\right) G_i\left(\theta_0, y^i\right)'\right]$$

The k^{th} l^{th} element of the matrix is given by: $E\left[G_{ki}\left(\theta_0, y^i\right) G_{li}\left(\theta_0, y^i\right)\right]$ and is a measure, for history up to i, of how much information is about parameters k and l. The information matrix of a sample is:

$$I\left(\theta_0\right) = \sum_{i=1}^{n} I_i\left(\theta_0\right) = \sum_{i=1}^{n} E\left[G_i\left(\theta_0, y^i\right) G_i\left(\theta_0, y^i\right)'\right]$$

Because of the zero covariance between columns of G this is equivalent to:

$$I\left(\theta_0\right) = E\left[g\left(\theta, y\right) g\left(\theta, y\right)'\right]$$

The covariance matrix of the score vector. This matrix must be positive definite to have a strong identification of the parameters.

The asymptotic information matrix is $\mathcal{I}\left(\theta\right) = \text{plim}\frac{1}{n} I\left(\theta\right)$.

Closely related to the information matrix is the hessian of the likelihood function, a $K \times K$ matrix $H\left(\theta\right)$ obtained from the second derivatives of the log-likelihood with respect to the parameters. The relation is named information matrix equality:

$$\mathcal{I}\left(\theta\right) = -\mathcal{H}\left(\theta\right)$$

where $\mathcal{H}\left(\theta\right) = \text{plim}\frac{1}{n} H\left(\theta\right)$.

15.2 Consistency (for type 1 MLE)

Several conditions are needed for consistency of the MLE estimator.

1. Small sample identification: for any given sample y and parameter vectors θ_1, θ_2, $l\left(\theta_1, y\right) \neq l\left(\theta_2, y\right)$ so that the parameters can always be told apart.

2. Asymptotic identification: for all $\theta \in \Theta$ we have $\text{plim}\frac{1}{n}l(\theta) \neq \text{plim}\frac{1}{n}l(\theta_0)$ so that the true parameter can be told apart from any other parameter as the sample goes to infinity.

3. Regularity conditions: the function $l(\theta, y)$ is, for any given sample, continuous and (its expectation) converges uniformly to the true population function.

Let $L(\theta)$ be the likelihood function (without logs), and θ_0 be the true parameter vector. By Jensen's inequality (using the concaveness of the log function):

$$E\left[\log\left(\frac{L(\theta)}{L(\theta_0)}\right)\right] < \log E\left[\frac{L(\theta)}{L(\theta_0)}\right]$$

with strict inequality for all $\theta \neq \theta_0$. Note that:

$$E\left[\frac{L(\theta)}{L(\theta_0)}\right] = \int \frac{L(\theta)}{L(\theta_0)} L(\theta_0)\, dy = \int L(\theta)\, dy = 1$$

Then:

$$E[\log L(\theta)] < E[\log L(\theta_0)]$$

The expectation of the log-likelihood at the true parameter is strictly greater than the expectation at any other parameter. Under the regularity conditions it follows, applying the LLN to the elements of the log-likelihood, that:

$$\text{plim}\frac{1}{n}l(\theta) = \lim\frac{1}{n}E[l(\theta)]$$

Then the inequality above implies:

$$\text{plim}\frac{1}{n}l(\theta) \leq \text{plim}\frac{1}{n}l(\theta_0)$$

Since the MLE estimator $\hat{\theta}$ maximizes the log-likelihood it must be that:

$$\text{plim}\frac{1}{n}l\left(\hat{\theta}\right) \geq \text{plim}\frac{1}{n}l(\theta_0)$$

Then it must be that:

$$\text{plim}\frac{1}{n}l\left(\hat{\theta}\right) = \text{plim}\frac{1}{n}l(\theta_0)$$

The asymptotic identification assumption gives consistency since the only way the above condition holds is if:

$$\text{plim}\hat{\theta} = \theta_0$$

The proof is not as rigorous as it can be made but it shows the main elements of the process.

15.3 Asymptotic Normality -and efficiency- (for type 2 MLE)

For obtaining the asymptotic distribution of the MLE one uses the score vector, note that by definition:

$$g\left(\hat{\theta}\right) = 0$$

Using a first order Taylor expansion of this expression around the true parameter one gets:

$$g\left(\hat{\theta}\right) = g(\theta_0) + H\left(\overline{\theta}\right)\left(\hat{\theta} - \theta_0\right) = 0$$

Where the relation is exact by evaluating matrix H at $\overline{\theta} \in \left[\theta_0, \hat{\theta}\right]$ where the value of $\overline{\theta}$ is given by the intermediate value theorem. Note that $\left\|\bar{\theta} - \theta_0\right\| \leq \left\|\hat{\theta} - \theta_0\right\|$, then since $\hat{\theta}$ is consistent for θ_0 so is $\overline{\theta}$. The above condition gives:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\left(\frac{1}{n} H\left(\overline{\theta}\right)\right)^{-1} \sqrt{n} g\left(\theta_0\right)$$

Then one gets:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{a} -\left(\mathcal{H}\left(\theta_0\right)\right)^{-1} \sqrt{n} g\left(\theta_0\right)$$

If the model is correctly specified then the information identity holds and $-\left(\mathcal{H}\left(\theta_0\right)\right)^{-1} = \mathcal{I}^{-1}\left(\theta_0\right)$ holds, in general the condition above holds and the one below only holds if the model is correctly specified.

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{a} \mathcal{I}^{-1}\left(\theta_0\right) \sqrt{n} g\left(\theta_0\right)$$

Under regularity conditions for g, and recalling that it has expected value 0, the CLT applies and:

$$\sqrt{n} g\left(\theta_0\right) \overset{a}{\sim} N\left(0, E\left[g\left(\theta_0\right) g\left(\theta_0\right)'\right]\right)$$

Where, from before, $E\left[g\left(\theta_0\right) g\left(\theta_0\right)'\right] = \mathcal{I}\left(\theta_0\right)$. This gives the result:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{a}{\sim} N\left(0, \mathcal{I}^{-1}\left(\theta_0\right)\right)$$

if the model is misspecified the asymptotic variance is given by:

$$\operatorname{Avar}\left(\hat{\theta}\right) = \mathcal{H}^{-1}\left(\theta_0\right) \mathcal{I}\left(\theta_0\right) \mathcal{H}^{-1}\left(\theta_0\right)$$

One can further show that if the model is correctly specified then the MLE is efficient with respect to any other $\sqrt{n}$ consistent estimator that is asymptotically unbiased. This is a very general result since it encompasses a wide variety of alternative estimators.

Let $\tilde{\theta}$ be a competing estimator, it can be shown that:

$$\sqrt{n}\left(\tilde{\theta} - \theta_0\right) = \sqrt{n}\left(\hat{\theta} - \theta_0\right) + v$$

where v is a random variable with zero expectation and that is uncorrelated with the term $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$. Is the no-correlation result what gives the efficiency of $\hat{\theta}$ since then:

$$\operatorname{Avar}\left(\tilde{\theta}\right) = \operatorname{Avar}\left(\hat{\theta}\right) + \operatorname{var}\left(v\right)$$

and $\operatorname{var}\left(v\right)$ is positive definite. Moreover it can be shown that $\hat{\theta}$ attains (asymptotically) the Cramer-Rao lower bound. This is what makes MLE so attractive, although for a finite sample there is no guarantee that the MLE performs better than any other parameter.

15.4 Example: Exponential distribution

Suppose that each observation of y is generated according with PDF:

$$f\left(y|\theta\right) = \theta e^{-\theta y}$$

There are n independent observations of y. The log-likelihood is then:

$$l(\theta|y) = \sum_{i=1}^{n} (\log\theta - \theta y_i) = n\log\theta - \theta\sum_{i=1}^{n} y_i$$

Note that this function is globally concave in θ and then the FOC are necessary and sufficient for a global optimum:

$$\frac{n}{\hat{\theta}} - \sum_{i=1}^{n} y_i = 0 \longrightarrow \hat{\theta} = \frac{1}{\overline{y}}$$

This solution is unique and is the same as the solution to the method of moments estimator:

$$E[y_t] = \theta^{-1} \longrightarrow \hat{\theta} = \frac{1}{\overline{y}}$$

15.5 Example: Normal linear regression model

Consider the linear regression model:

$$y = X\beta + u \qquad u \sim N\left(0, \sigma^2 I\right)$$

where the explanatory variables X are considered exogenous and taken as given. Then $y \sim N\left(X\beta, \sigma^2 I\right)$ and:

$$f\left(y|\beta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{\frac{-n}{2}} e^{-\frac{1}{2}(y-X\beta)'\left(\sigma^2 I\right)^{-1}(y-X\beta)}$$

And the log-likelihood is:

$$l\left(\beta, \sigma^2|y\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

One can solve jointly for β and σ^2 or solve for one of them and then replace in the objective function to solve for the other one, this is called concentrating the likelihood.

Solving jointly the FOC are:

$$\begin{aligned} -\frac{1}{\hat{\sigma}^2} X'\left(y - X\hat{\beta}\right) &= 0 \\ \frac{-n}{2\hat{\sigma}^2} + \frac{1}{2\left(\hat{\sigma}^2\right)^2}\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right) &= 0 \end{aligned}$$

From the first equation:

$$\hat{\beta} = \left(X'X\right)^{-1} X'y$$

From the second equation:

$$\hat{\sigma}^2 = \frac{1}{n}\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right) = \frac{1}{n}\hat{u}'\hat{u}$$

Under normal residuals the OLS estimator for $\hat{\beta}$ coincides with the MLE and shares all its properties.

The only thing that is left to establish is the asymptotic distribution of the ML estimates.