FOR A GOOD DESCRIPTION OF JOINT PDFs and LAW of ITERATED EXPECTATIONS ⇒ YOUTUBE : "INTROMEDIATE ECON"

Videos E1.4, E1.5 and E1.7

MSc Economics - Econometrics II

Review of projection and partitioned linear regression

[redacted]

Introduction

Recall from the Autumn term:

$$y = X\beta + e \tag{1}$$

where e is a vector of serially uncorrelated mean-zero disturbances with variance $\mathbb{E}(e^2) = \sigma^2$. OLS creates fitted vectors, $\hat{\mu}$, and $\hat{e}$, by performing orthogonal projection of y onto $Col(X)$. I.e. the OLS estimator $\hat{\beta}$ is the estimator that makes the fitted residuals, $\hat{e}$, orthogonal to the data, X; that is, solve $X'\hat{e} = 0$:

$$\begin{aligned} X'(y - X\hat{\beta}) &= 0 \\ \Rightarrow \hat{\beta} &= (X'X)^{-1}X'y \end{aligned}$$

so that we experss y as a vector of fitted values and fitted residuals

$$\begin{aligned} y &= \hat{\mu} + \hat{e} \\ &= X\hat{\beta} + \hat{e} \\ &= X(X'X)^{-1}X'y + \hat{e} \\ &= P_X y + (I - P_X)y \end{aligned} \tag{2}$$

The orthogonal projector, which achieves this decomposition $y = \hat{\mu} + \hat{e}$ is denoted P_X and is called the projector onto the column space of X. From (2) we see $P_X = X(X'X)^{-1}X'$

Properties of P_X:

- Symmetric → $P_X = P_X'$
- Idempotent → $P_X P_X = P_X$ and $P_X'(1 - P_X) = 0$
- if $z \in Col^{\perp}(X)$, $P_X z = 0$
- $P_X X = X$

$\hat{\mu}$ tells us about how y responds to all the information in X. This can be very useful, for example fior the model in (1) our best forecast of y given X is $\mathbb{E}(y|X) = \hat{\mu}$. Often we

are interested in the more subtle question, 'how does y respond to the unique variation in X_i, holding all other X_j, $j = 1 \ldots K \neq i$ constant'. We are familiar from experience with the idea that each element of $\hat{\beta} = (\hat{\beta}_1 \ldots \hat{\beta_K})'$ is an estimate of the variation in y uniquely attributable to X_i. The purpose of this lecture is to show geometrically, and algebraically, how OLS does this.

OLS as a two step procedure

Imagine $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$. First regress y on X_2 only. The result will be $\hat{\mu}_2$, that part of y explained only by X_2, and a residual orthogonal to $Col(X_2)$, which we call $y_{\perp 2}$ (where the notation reads 'the part of y orthogonal to X_2'). Now let us also regress X_1 on X_2, to remove the part of X_1 that covaries linearly with X_2, call this residual $X_{1\perp 2}$. If we now perform a second stage regression of the y-residual on the X_1-residual we will find the part of y in $Col(X_1)$ which cannot be explained by X_2; that is, we will have found $\hat{\mu}_1$, the part of y uniquely explained by X_1. If we then define $\hat{\mu}$ to be the part of y in $Col(X)$, it must be that $\hat{\mu} = \hat{\mu}_1 + \hat{\mu}_2$.[1]

The interesting thing is that this is *exactly what OLS does anyway.* We don't need to bother with the 2-stage procedure becasue we find $\hat{\mu}_1$ and $\hat{\mu}_2$, and the associated estimators $\hat{\beta}_1$, $\hat{\beta}_2$, directly from the regression equation:

$$\begin{aligned} y &= X\hat{\beta} + \hat{e} \\ &= \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} + \hat{e} \\ &= X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e} \\ &= \hat{\mu}_1 + \hat{\mu}_2 + \hat{e} \\ &= \hat{\mu} + \hat{e} \end{aligned} \qquad (3)$$

Exercise Derive the estimators $\hat{\beta}_1$, $\hat{\beta}_2$ from the description of the 2-stage procedure above. We will do this together in class.

OLS and Partitioned Regression

[1] We are saying that $Col(X)$ is the space spanned by the two vectors X_1, X_2, so that the projection of y onto $Col(X)$ is the vector addition of the projection of y onto $Col(X_2)$ and the projection of the residual from this operation onto $Col(X_1)$. In general $Col(X_1)$ and $Col(X_2)$ are not orthogonal subspaces (there is some covariance between the r.h.s. variables) and μ_1 and μ_2 are not then orthogonal.

Assumption: The $N \times K$ matrix X is of full column rank $rk(X) = K$, $N > K$.

Result: Partitioned Regression

Assume **A1** holds and then let

a $P_{X_2} = X_2(X_2'X_2)^{-1}X_2'$ be the projector onto $Col(X_2)$

b $X_{1\perp 2} = (I - P_{X_2})X_1$ be the part of X_1 which is orthogonal to $Col(X_2)$

c $y_{\perp 2} = (I - P_{X_2})y$ be the part of y which is orthogonal to $Col(X_2)$

then, for the OLS regression (3)

1 The fitted vector $\hat{\mu}_1 = X_1\hat{\beta}_1$ is the unique projection of y, or $\hat{\mu}$, onto $Col(X_1)$ such that elements of $Col(X_2)$ and $Col^{\perp}(X)$ are annihilated. Further $\hat{\mu}_1 = P_{12}y = P_{12}\hat{\mu}$ where the unique projector P_{12} is

$$P_{12} = X_1(X_{1\perp 2}'X_1)^{-1}X_{1\perp 2}'$$

2 The OLS fitted coefficients $\hat{\beta}_1$ are

$$\hat{\beta}_1 = (X_{1\perp 2}'X_{1\perp 2})^{-1}X_{1\perp 2}'y_{\perp 2}$$

Discussion

This is exactly what we found from the two-stage regression exercise. This surprising feature of OLS is known as partitioned regression, or the Frisch-Waugh-Lovell Theorem. To start understanding this result, first consider the projectors P_{X_2} and P_{12}. We are familiar with P_X, the projector onto $Col(X)$. Clearly P_{X_2} is then the projector onto $Col(X_2)$, a subspace of $Col(X)$. It is clear then that $(I - P_{X_2})$ is the projector onto $Col^{\perp}(X_2)$. Note, $X_1 \in Col(X)$, therefore $(I - P_{X_2})X_1$ must lie in the portion of $Col(X)$ which remains after $Col(X_2)$ has been removed. P_{12} is a direct way of projecting onto this subspace: it preserves X_1 while annihilating anything in $Col(X_2)$

$$\begin{aligned} P_{12}X_1 &= X_1(X_{1\perp 2}'X_1)^{-1}X_{1\perp 2}'X_1 = X_1 \\ \text{for } \alpha \in Col(X_2), \quad P_{12}\alpha &= X_1(X_{1\perp 2}'X_1)^{-1}X_{1\perp 2}'\alpha \\ &= X_1(X_{1\perp 2}'X_1)^{-1}[(I - P_{X_2})X_1]'\alpha \\ &= X_1(X_{1\perp 2}'X_1)^{-1}X_1'(I - P_{X_2})\alpha = 0 \end{aligned}$$

If you can't see the second part, replace α with X_2 itself.

Notice that $\hat{\beta}_1$ is precisely the estimator we derived earlier in the 2-step procedure. We interpret $\hat{\beta}_1$ as capturing the component of y which is colinear with X_1 but that cannot be fitted with X_2. Also $\hat{\mu_1} = P_{1\perp 2}y = P_{1\perp 2}\hat{\mu}$ is the part of $\hat{\mu}$ that cannot be explained by $\hat{\mu_2} \in Col(X_2)$. It is a fit of the variation is y uniquely due to variation in X_1. We see this geometrically in the figure 1.

This interesting thing is to show that regression actually manages to do the 2-step procedure in 1step. At the end of the note is a demonstration of why this result works, which is adapted from Ruud's completely formal proof. But first, look at an informative example.

Some useful examples

Constants. Consider the model in (1) with $X = \begin{bmatrix} \iota & x \end{bmatrix}$. ι is a $T.1$ column of ones - i.e. of regression intercepts. In line with the 2-step approach, imagine first regressing y on a column of ones: you would recover fitted values with the value $\bar{y}$. Do the same with x, and find its mean. Then regress the residuals against eachother and you will find your coefficient measures the covariation between y and x, scaled by the variance of x. This is because the covariance is based on de-meaning the data - removing the part of the data in the dimension $Col(\iota)$, i.e. the means. Now, lets look at the algebra of OLS, and we will see that $\hat{\beta}_2$ is precisely this estimator.

$$
\begin{aligned}
\beta &= (X'X)^{-1}X'y \\
&= \left[\begin{bmatrix}\iota & x\end{bmatrix}' \begin{bmatrix}\iota & x\end{bmatrix}\right]^{-1} \begin{bmatrix}\iota & x\end{bmatrix}' y \\
&= \begin{bmatrix} T & T\bar{x} \\ T\bar{x} & \sum x_t^2 \end{bmatrix}^{-1} \begin{bmatrix} T\bar{y} \\ \sum x_t y_t \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\beta_2 &= \frac{-T\bar{x}\bar{y} + \sum x_t y_t}{\sum x_t^2 - T\bar{x}^2} \\
&= \frac{\sum (x_t - \bar{x}) y_t}{\sum (x_t - \bar{x})^2} = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})^2}
\end{aligned}
$$

$$
\begin{aligned}
\beta_1 &= \frac{\bar{y}\sum x_t^2 - \bar{x}\sum x_t y_t}{\sum (x_t - \bar{x})^2} \\
&= \frac{\bar{y}\sum x_t^2 - \bar{x}\sum x_t y_t - T\bar{y}\bar{x}^2 + T\bar{y}\bar{x}^2}{\sum (x_t - \bar{x})^2} \\
&= \bar{y} - \bar{x}\beta_2
\end{aligned}
$$

You can prove the partitioned regression theorem directly in a similar way, but the algebra gets messy. The general argument at the end is more elegant.

Panel data. Imagine you have data on the earnings of 10000 people at five periods in time. you might think that individual earnings are explained by a set of K regressors, but also by individual specific effects such as luck or talent, which are not directly observed. For each individual n in time period t, this regression is given by

$$
y_{nt} = x_{nt}'\beta = \sum_{k=1}^{K} x_{ntk}\beta_{1k} + \sum_{n=1}^{10000} \delta_{ntk}\beta_{2n}
$$

$$
\text{where} \quad \delta_{ntk} = \begin{cases} 1 & \text{iff n=k,} \\ 0 & \text{otherwise} \end{cases}
$$

for individual n at time t. So each individual's earning at a given time are explained by that individuals realisation of K explanatory factors times the K common factorloadings β_{1k}, and the individual intercept β_{2n}.

However, it would be very inefficient to calculate (and read!) a regression with $10,000+K$ r.h.s. variables in it. Fortunately, with partitioned regression we don't have to do this. Group the dummy variables into $X_2 = [\delta_{nt,k}]$ which has nt rows and k columns. As with intercepts, regression against the dummy variables will create averages, but this time we will get 10,000 individual averages, relating to each person in the sample as this is how X_2 now groups the data. If we regress the X_1 data on common explanatory factors against this X_2, the residuals will give us a matrix $X_{1\perp 2}$ of differences of the form $(x_{ntk} - \bar{x}_{nk})$ where

$$\bar{x}_{nk} = \frac{1}{5}\sum_{t=1}^{5} x_{ntk}$$

i.e. a matrix of differences between t realisations of x_{nk} and the time average of the kth factor that each individual experiences. Similarly, $y_{\perp 2} = y_{nt} - \bar{y}_n$, is the de-meaned data for each individual. Now to find the vetor of common factor loadings, β_1, we just perform the regression $\beta_1 = (X'_{1\perp 2}X_{1\perp 2})^{-1}X'_{1\perp 2}y_{\perp 2}$, which is straightforward for the computer to do.

Formal Discussion

I omit some details that would make this a genuine proof, notably the uniqueness of P_{12} and the demonstration that it annihilates $Col^{\perp}(X)$, the latter being geometrically intuitive. Details are found in Ruud, Chapter 3.3.

We first show that the propsed formula for $\hat{\beta}_1$ does indeed solve the OLS problem:

$$\min_{\hat{\beta}} \|y - X\hat{\beta}\|^2$$

For any β we can define the associated fitted values, $\mu = X\beta$. Our plan is to transform this minimum distance problem defined in terms of both elements of β into one in terms of only β_1.

To start, we can make an orthogonal decomposition of the error $e = y - \mu$:

$$\begin{aligned} y - \mu := e &= (I - P_{X_2})(y - \mu) + P_{X_2}(y - \mu) \\ \|y - \mu\|^2 := e'e &= \|(I - P_{X_2})(y - \mu)\|^2 + \|P_{X_2}(y - \mu)\|^2 \\ &= e'(I - P_{X_2})e + e'P_{X_2}e \end{aligned}$$

where the cross products in the squares disappear, because we are considering orthogonal components of e.[2] The term

$$\begin{aligned}(I - P_{X_2})(y - \mu) &= (I - P_{X_2})y - (I - P_{X_2})(\mu_1 + \mu_2) \\ &= y_{\perp 2} - (I - P_{X_2})X_1\beta_1 - (I - P_{X_2})X_2\beta_2 \\ &= y_{\perp 2} - (I - P_{X_2})X_1\beta_1 \\ &= y_{\perp 2} - X_{1\perp 2}\beta_1\end{aligned}$$

and

$$\begin{aligned}P_{X_2}(y - \mu) &= P_{X_2}y - P_{X_2}X_1\beta_1 - P_{X_2}X_2\beta_2 \\ &= P_{X_2}(y - X_1\beta_1) - X_2\beta_2\end{aligned}$$

thus

$$\|y - \mu\|^2 = \|y_{\perp 2} - X_{1\perp 2}\beta_1\|^2 + \|P_{X_2}(y - X_1\beta_1) - X_2\beta_2\|^2$$

but as $P_{X_2}(y - X_1\beta_1) \in Col(X_2)$ then $\forall \beta_1 \in \mathbb{R}^{K_1}$ we can choose β_2 such that[3]

$$P_{X_2}(y - X_1\beta_1) - X_2\beta_2 = 0$$

hence

$$\underset{\beta}{\text{Min}} \|y - \mu\|^2 = \underset{\beta_1}{\text{Min}} \|y_{\perp 2} - X_{1\perp 2}\beta_1\|^2 \tag{4}$$

and we have transformed the original OLS problem into a problem involving only β_1.

Of course the solution to (4), which has the standard OLS form, is:

$$\beta_1 = (X'_{1\perp 2}X_{1\perp 2})^{-1}X_{1\perp 2}y_{\perp 2}$$

This proves part **2** of the theorem.

[2]Equally, this is an application of Pythagoras to the dot product.

[3]Think of X_2 as an N.1 vector - a straight line from the origin. Then any two vectors in the space $Col(X_2)$ are just two lines of different lengths w the same orientation and we can always choose β_2 to make them the same length so that they cancel out.

We still need to show $\hat{\mu}_1 = P_{12}y = P_{1,2}\hat{\mu}$.

$$\begin{aligned}
\hat{\mu_1} &= X_1\hat{\beta}_1 \\
&= X_1(X_{1\perp 2}'X_{1\perp 2})^{-1}X_{1\perp 2}'y_{\perp 2} \\
&= X_1(X_1'(I-P_{X_2})'(I-P_{X_2})X_1)^{-1}X_1'(I-P_{X_2})'(I-P_{X_2})y \\
&= X_1(X_1'(I-P_{X_2})'X_1)^{-1}X_1'(I-P_{X_2})'y \\
&= X_1([(I-P_{X_2})X_1]'X_1)^{-1}[(I-P_{X_2})X_1]'y \\
&= X_1(X_{1\perp 2}'X_1)^{-1}X_{1\perp 2}'y \\
&= P_{12}y
\end{aligned}$$

Finally, we must show that $P_{12}y = P_{12}\hat{\mu}$

$$\begin{aligned}
P_{12}y &= P_{12}(y-\hat{\mu}+\hat{\mu}) \\
&= P_{12}(y-\hat{\mu}) + P_{12}\hat{\mu} \\
&= P_{12}\hat{\mu} \qquad (5)
\end{aligned}$$

as $(y-\hat{\mu}) \in Col^{\perp}(X) \Rightarrow P_{12}(y-\hat{\mu}) = 0$, P_{12} being a projector which annihilates $Col(X_2) \oplus Col^{\perp}(X)$.

This proves the assertion in part **1** of the theorem (except uniqueness which is more technical).

Reference: P. A. Ruud *Introduction to Classical Econometrics* 2000

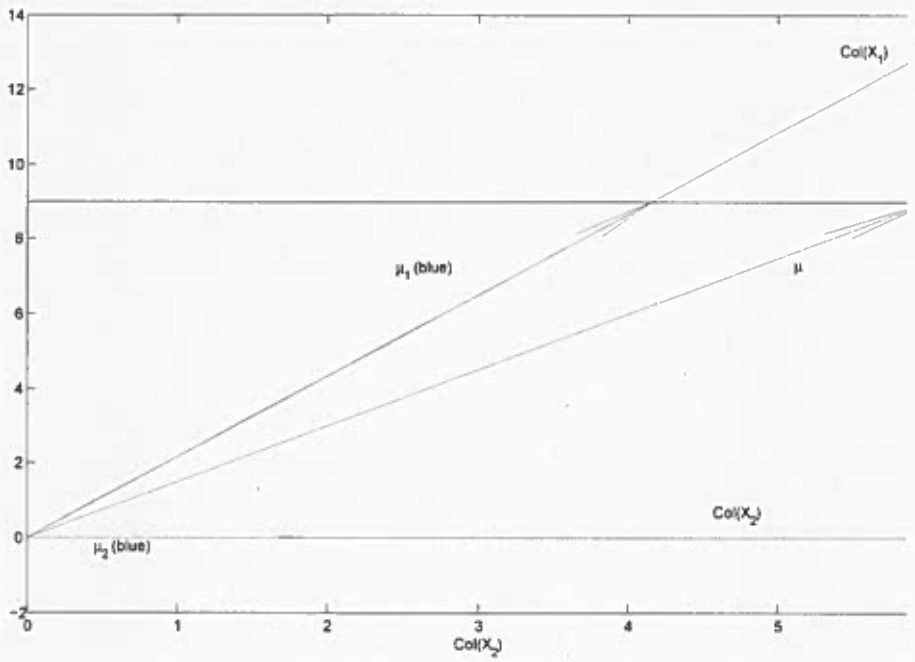

Figure 1: The projector P_{12} sends μ on a projection along the black line at $y = 9$ towards $Col^{\perp}(X_2)$, but it stops at $Col(X_1)$ as P_{12} is not an orthogonal projector.