Ordinary Least Squares at advanced level

1. Review of the two-variate case with algebra

OLS is the fundamental technique for linear regressions. You should by now be aware of the two-variate case and the usual derivations. In this text we are going to review the OLS using matrix algebra, which is the right tool to have a more generalized (multivariate) view of the OLS methodology.

In the standard two-variate case we had the following model for the population:

$Y_i = \beta_0 + \beta_1 X_i + e_i$ (1.1) where e is the error, X is the explanatory variable, Y is the dependent variable and betas denotes the population coefficients. This is called the Population Regression Equation.

What we have is only a sample (sample observations are denoted by lowercase letters):

$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + u_i$ (1.2) where u is the residual which is our estimate of the error. This is the Sample Regression Equation. y and x are hence a sample drawn from the population Y and X and the beta hats are our estimates of the betas (population parameters) from the sample.

We arrived at the OLS estimates of the beta coefficients by the least squares principle (SSR – sum of squares residuals):

$SSR = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$ (1.3). The first order conditions for a minimum requires that:

$$\frac{\sum_{i=1}^{n} u_i^2}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = -2\sum_{i=1}^{n} u_i = 0 \rightarrow \sum_{i=1}^{n} u_i = 0 \rightarrow \frac{1}{n}\sum_{i=1}^{n} u_i = 0 \rightarrow E(u) = 0$$ (1.4) in other words, the residual of the OLS will always have zero mean, provided we include an intercept or constant term.

$$\frac{\partial \sum_{i=1}^{n} u_i^2}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = -2\sum_{i=1}^{n} u_i x_i = 0 \rightarrow \sum_{i=1}^{n} u_i x_i = 0 \rightarrow \frac{1}{n}\sum_{i=1}^{n} u_i x_i = E(ux) = 0$$ (1.5) This is the sample version of the orthogonality or exogeneity condition. The OLS will always assume that the residuals and the explanatory variables are uncorrelated. If this is not true, the OLS is biased.

From the above conditions:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{y} - \hat{\beta}_1 \overline{x}$$ (1.6)

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = \sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = \frac{E(x \cdot y) - \overline{y} \cdot \overline{x}}{E(x^2) - \overline{x} \cdot \overline{x}}$$ (1.7)

We can now substitute our estimate for the intercept into the above expression to arrive at the OLS estimator for the slope coefficient.

$$\hat{\beta}_1 = \frac{E(x \cdot y) - \bar{y} \cdot \bar{x}}{E(x^2) - \bar{x} \cdot \bar{x}} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{y} \cdot \bar{x}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x} \cdot \bar{x}} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (1.8)$$

These are all equivalent and we can use any of these as we please.

You should be able to reproduce above derivation without any difficulty before advancing further. If you do not understand a step, you will find plenty of help on the internet. The idea is not that you memorize the derivations but that you arrive at valid results starting out from the same assumptions, in other words you can reproduce the derivations. Let me share my experience with you: you completely understand something only if you can derive it. Knowing the big picture only, will help you to advance faster initially, but it will hold you back from some point on.

Let us look at the properties of the OLS now using the two-variate case!

1. **Linearity:** First of all we will show why the regression as in (1.2) is linear. If the parameters can be expressed as a linear combination or weighted average of the observations of the dependent variable, we call the regression linear. We take one version of the estimator:

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sum_{i=1}^{n} k_i y_i$ (1.9), where $k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ (1.10). So every single observation of the dependent variable is going to affect our estimate of the slope parameter β_1 by a unique weight. The weight depends on the variance of the explanatory variable and the deviation of the explanatory variable at observation i from the mean.

Properties of k_i

$$\sum_{i=1}^{n} k_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0, \sum_{i=1}^{n} k_i^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{k_i}{(x_i - \bar{x})},$$

$$\sum_{i=1}^{n} k_i x_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 1, \sum_{i=1}^{n} k_i (x_i - \bar{x}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 1$$

2. **Unbiasedness:** An estimator is unbiased if its expected value equals the population parameter. That is $E(\hat{\beta}) = \beta$. If there is a difference then that difference is called the bias.

We show the unbiasedness first. The trick is that since the population regression function (or the Data Generating Process) is $Y_i = \beta_0 + \beta_1 X_i + e_i$, hence $y_i = \beta_0 + \beta_1 x_i + e_i$. This can be substituted into the estimator to derive the relationship between the population parameter and our estimate.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i + e_i - \beta_0 - \beta_1 \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})e_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \beta_1 + \sum_{i=1}^{n} k_i e_i$$

(1.11)

$E(\hat{\beta}_1) = \beta_1 + \frac{\sigma_{xe}}{\sigma_x^2}$ hence, if the error and the explanatory variables are uncorrelated, the OLS estimator is unbiased (orthogonality condition again). Then: $E(\hat{\beta}_1) = \beta_1$

Similarly, $\hat{\beta}_o = \overline{y} - \hat{\beta}_1 \overline{x} = \beta_0 - \left(\hat{\beta}_1 - \beta_1\right)\overline{x}$ hence $E(\hat{\beta}_o) = \beta_0 - E\left(\hat{\beta}_1 - \beta_1\right)\overline{x}$, so if $\hat{\beta}_1$ is unbiased, $\hat{\beta}_0$ is unbiased too.

3. **Efficiency:** This concept can only be understood in a relative sense. An estimator is going to have a standard deviation (called standard error) since with any new sample drawn from the same population, you are going to obtain different estimates for the population parameter. If you have two alternative estimators, the one with lower standard error is called more efficient. There is a lower limit of standard errors, given by the Cramér–Rao lower bound (in other words, no estimator can have less variance than this limit). If the estimator's variance is at the lower bound, then we call it efficient.

We can express the variance of the OLS estimator for β_1 as follows:

$$\sigma_{\hat{\beta}_1}^2 = E\left[\left(\hat{\beta}_1 - E(\hat{\beta}_1)\right)^2\right] = E\left(\sum_{i=1}^{n} k_i e_i\right)^2 \quad (1.12)$$

If the orthogonality condition holds, $E\left(\sum_{i=1}^{n} k_i e_i\right)^2 = E\left(\sum_{i=1}^{n} k_i^2 e_i^2\right)$ since the cross products are all zero. If the error is homoscedastic then $E\left(e^2\right) = \sigma_e^2$ and we can treat it as constant and bring it in front of the summa sign.

$$\sigma_{\hat{\beta}_1}^2 = E\left(\sum_{i=1}^{n} k_i^2 e_i^2\right) = E(e^2)E\left(\sum_{i=1}^{n} k_i^2\right) = \frac{\sigma_e^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \rightarrow \sigma_{\hat{\beta}_1} = \frac{\sigma_e}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \quad (1.13)$$

The problem is that we do not know the standard deviation of the error. But we can use the residual variance (or mean sum of squares residual) as estimator of the error variance. $\sigma_e^2 = \frac{\sigma_u^2}{n-2}$.

But why does our OLS estimator for the population error variance equals $\frac{\sigma_u^2}{n-2}$ in a two-variate regression? Everyone seems to accept this, yet only rarely is it derived algebraically (it is quite simple to do with matrix algebra, though). Let us see the derivation!

First, you will need some uncomfortable algebra to express the residual as function of the error.

$$u_i = y_i - \hat{\beta}_o - \hat{\beta}_1 x_i = \beta_o + \beta_1 x_i + e_i - (\overline{y} - \hat{\beta}_1 \overline{x}) - \hat{\beta}_1 x_i = \beta_o + \beta_1 x_i + e_i - (\beta_o + \beta_1 \overline{x} + \overline{e}) + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 x_i =$$
$$= e_i - \overline{e} - \left(\hat{\beta}_1 - \beta_1\right)\left(x_i - \overline{x}\right)$$

No we take the square residual:

$u_i^2 = \left(e_i - \overline{e}\right)^2 + \left(\hat{\beta}_1 - \beta_1\right)^2 \left(x_i - \overline{x}\right)^2 - 2\left(\hat{\beta}_1 - \beta_1\right)\left(x_i - \overline{x}\right)\left(e_i - \overline{e}\right)$ and the sum of squared residuals is then:

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n}\left(e_i - \overline{e}\right)^2 + \left(\hat{\beta}_1 - \beta_1\right)^2 \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 - 2\left(\hat{\beta}_1 - \beta_1\right)\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(e_i - \overline{e}\right)$$

Now we need the expectation of the sum of squared residuals.

$$E\left(\sum_{i=1}^{n} u_i^2\right) = E\left(\sum_{i=1}^{n}\left(e_i - \overline{e}\right)^2\right) + E\left(\left(\hat{\beta}_1 - \beta_1\right)^2\right)\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 - 2E\left(\hat{\beta}_1 - \beta_1\right)\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(e_i - \overline{e}\right) =$$

$$= (n-1)\sigma_e^2 + \sigma_e^2 - 2\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)e_i}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(e_i - \overline{e}\right) = n\sigma_e^2 - 2\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = n\sigma_e^2 - 2\sigma_e^2 =$$
$$= (n-2)\sigma_e^2 \quad (1.14)$$

where we made us of the following:

$$E\left(\sum_{i=1}^{n}\left(e_i - \overline{e}\right)^2\right) = E\left(\sum_{i=1}^{n} e_i^2 - n\overline{e}^2\right) = E\left(\sum_{i=1}^{n}\left(e_i^2\right) - n\left(\frac{\sum_{i=1}^{n} e_i}{n}\right)^2\right) =$$

$$= \sum_{i=1}^{n} E\left(e_i^2\right) - \frac{1}{n}E\left(\sum_{i=1}^{n} e_i\right)^2 = n\sigma_e^2 - \frac{1}{n}\left(\sum_{i=1}^{n} E(e_i^2)\right) = \left(n-1\right)\sigma_e^2$$

where the conversion: $\left(\sum_{i=1}^{n} e_i\right)^2 = \sum_{i=1}^{n} e_i^2$ is true under the assumption of no autocorrelation.

The standard error of the constant term can be derived as follows:

$\hat{\beta}_o - \beta_0 = \left(\hat{\beta}_1 - \beta_1\right)\overline{x} + \overline{e} = \overline{e} - \overline{x}\sum_{i=1}^{n} k_i e_i$ hence

$\left(\hat{\beta}_o - \beta_0\right)^2 = \overline{e}^2 + \overline{x}^2\left(\sum_{i=1}^{n} k_i e_i\right)^2 - 2\overline{ex}\sum_{i=1}^{n} k_i e_i = \frac{\sum_{i=1}^{n} e_i^2}{n^2} + \overline{x}^2\left(\sum_{i=1}^{n} k_i e_i\right)^2 - 2\overline{ex}\sum_{i=1}^{n} k_i e_i$ and

$E\left(\left(\hat{\beta}_o - \beta_0\right)^2\right) = \frac{\sigma_e^2}{n} + \frac{\overline{x}^2\sigma_e^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$ Let us remember: $\overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \sigma_x^2$ substituting this into the previous equation gives us:

$$E\left(\left(\hat{\beta}_o - \beta_0\right)^2\right) = \frac{\sigma_e^2}{n} + \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2\sigma_e^2 - \sigma_x^2\sigma_e^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \frac{\sigma_e^2\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \quad (1.15)$$

The Cramér-Rao lower bound

The Cramér-Rao lower bound for the variance of an estimator ($\hat{\theta}$) is expressed as: $\sigma_{\hat{\theta}}^2 \geq \frac{1}{I(\theta)}$, where θ denotes the population parameter to be estimated. I(θ) is the Fischer information which is defined as:

$I(\theta) = E\left(\left(\frac{\partial \ell(x,\theta)}{\partial\theta}\right)^2\right) = -E\left(\frac{\partial^2 \ell(x,\theta)}{\partial\theta\partial\theta}\right)$, where $\ell(\mathrm{x},\theta)$ is the log-likelihood function, and we have a single parameter to estimate.

For example if the dependent variable Y follows a normal distribution and we estimate its population mean only (μ_Y).[1] Then the log-likelihood function is:

$$\ell(X, \beta_0, \beta_1) = -\frac{n}{2}\ln(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum e_i^2 = -\frac{n}{2}\ln(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum (Y_i - \mu_Y)^2$$

$\frac{\partial \ell(X, \mu_Y)}{\partial \mu_Y} = \frac{1}{\sigma_e^2}\sum (\mathrm{Y}_i - \mu_Y) = \frac{\sum e_i}{\sigma_e^2}$, also known as the score function.

$\frac{\partial \ell(X, \mu_Y)}{\partial \mu_Y \partial \mu_Y} = -\frac{1}{\sigma_e^2}\sum 1 = -\frac{n}{\sigma_e^2}$, hence $\mathrm{var}(\hat{\mu}_Y) \geq \frac{\sigma_e^2}{n}$. You should remember that when we the sample mean is used as an estimator for the population mean, its standard error was: $\sigma_{\bar{y}}^2 = \frac{\sigma_e^2}{n}$ hence the sample mean is at the Cramér-Rao lower bound and is an efficient estimator of the population mean.

[1] Actually, the standard deviation is also a parameter to estimate, but this is independent of the mean, so I disregard it now.

What if, as usually the case, we have a vector of parameters to estimate (i.e. multiple parameters)? Then we have the Fischer information matrix. The i,j[th] element of which is:

$$\mathbf{I}(\boldsymbol{\theta})_{i,j} = E\left(\left(\frac{\partial \ell(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_i}\right)\left(\frac{\partial \ell(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_j}\right)\right) = -E\left(\frac{\partial^2 \ell(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right).$$

For example if we have the PRF (1.1) and e is assumed to be normally distributed, then

$$\ell(X,\beta_0,\beta_1) = -\frac{n}{2}\ln(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum e_i^2 = -\frac{n}{2}\ln(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum (\mathrm{Y}_i - \beta_0 - \beta_1 \mathrm{X}_i)^2$$

$$\frac{\partial \ell(X,\beta_0,\beta_1)}{\partial \beta_1} = \frac{1}{\sigma_e^2}\sum (\mathrm{Y}_i - \beta_0 - \beta_1 \mathrm{X}_i)\,\mathrm{X}_i$$

$$\frac{\partial^2 \ell(X,\beta_0,\beta_1)}{\partial \beta_1 \partial \beta_1} = -\frac{\sum \mathrm{X}_i^2}{\sigma_e^2}, \frac{\partial^2 \ell(X,\beta_0,\beta_1)}{\partial \beta_1 \partial \beta_0} = -\frac{\sum \mathrm{X}_i}{\sigma_e^2}$$

$$\frac{\partial \ell(X,\beta_0,\beta_1)}{\partial \beta_0} = \frac{1}{\sigma_e^2}\sum (\mathrm{Y}_i - \beta_0 - \beta_1 \mathrm{X}_i) = \frac{\sum e_i}{\sigma_e^2}, \frac{\partial \ell(X,\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_0} = -\frac{n}{\sigma_e^2}, \frac{\partial \ell(X,\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} = -\frac{\sum \mathrm{X}_i}{\sigma_e^2}$$

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -\frac{n}{\sigma_e^2} & -\frac{\sum \mathrm{X}_i}{\sigma_e^2} \\ -\frac{\sum \mathrm{X}_i}{\sigma_e^2} & -\frac{\sum \mathrm{X}_i^2}{\sigma_e^2} \end{bmatrix} \quad \mathbf{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} -\frac{\sigma_e^2 \sum X_i^2}{n\sum\left(X_i - \bar{X}\right)^2} & \frac{\bar{X}\sigma_e^2}{\sum\left(X_i - \bar{X}\right)^2} \\ \frac{\bar{X}\sigma_e^2}{\sum\left(X_i - \bar{X}\right)^2} & -\frac{\sigma_e^2}{\sum\left(X_i - \bar{X}\right)^2} \end{bmatrix}$$

Hence: $\sigma_{\hat{\beta}_0}^2 \geq \frac{\sigma_e^2 \sum X_i^2}{n\sum\left(X_i - \bar{X}\right)^2}$ and $\sigma_{\hat{\beta}_1}^2 \geq \frac{\sigma_e^2}{\sum\left(X_i - \bar{X}\right)^2}$, which equal the variances (1.13), (1.14) of the OLS estimates under exogeneity, homoscedasticity and no autocorrelation assumptions.

The Gauss Markov theorem

If the following conditions are met:

1. $Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{j,i} + e_i$ (the model is linear)
2. $E(e) = 0$
3. $Var(e) = \sigma_e^2 < \infty$ (homoscedasticity)
4. $Cov(e_i, \mathrm{e}_j) = 0, i \neq j$ (no autocorrelation)
5. $Cov(\mathrm{X}_j, \mathrm{e}) = 0$ for any X_j (exogeneity)

Then the OLS is the best linear unbiased estimator or BLUE. Best refers to the fact that its standard errors are on the Cramér-Rao lower bound, hence we cannot have any estimator with a lower standard error. This is core result in statistics. Observe that the normality of the

error term is not required, even though it is customary to list among the assumptions of the Classical Linear Model, but we used it to derive the Cramér-Rao Lower Bound. Yet, since the coefficients are calculated as the weighted sum of observations drawn from the same probability distribution (y), their distribution should converge to the normal distribution according to the Central Limit Theorem (CLT).

2. OLS with matrix algebra

Let us define the following linear model in the population:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{e} \text{ or } \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_n \end{bmatrix} = \begin{bmatrix} \tilde{X}_{11} & \cdots & \tilde{X}_{1k} \\ \tilde{X}_{21} & \cdots & \tilde{X}_{1k} \\ \vdots & \ddots & \vdots \\ \tilde{X}_{n1} & \cdots & \tilde{X}_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad (2.1)$$

we can estimate the vector of coefficients (β) using the least squares principle. The vector **e** denotes the vector of errors: $\mathbf{e} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}$.

Hence the sum of squares residual (SSE) is

$$SSE = \mathbf{u}^{\mathbf{T}}\mathbf{u} = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^T \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{y}^{\mathbf{T}}\mathbf{y} - \mathbf{y}^{\mathbf{T}}\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y} + \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}^{\mathbf{T}}\mathbf{y} - 2\hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y} + \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\hat{\boldsymbol{\beta}} \quad (2.2)$$

, which is a scalar. Here we made us of the fact that $\boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{y}^{\mathbf{T}}\mathbf{X}\boldsymbol{\beta}$, since they are scalars (their dimension is 1x1).

The First Order Condition of an extremum requires that:

$$\frac{\partial \mathbf{u}^{\mathbf{T}}\mathbf{u}}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}^{\mathbf{T}}\mathbf{y} + 2\mathbf{X}^{\mathbf{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \quad (2.3)$$

or $\mathbf{X}^{\mathbf{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathbf{T}}\mathbf{y}$, which is called the normal eqaution

Where I made use of the following rules:

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}, \frac{\partial \mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^{\mathbf{T}}(\mathbf{A} + \mathbf{A}^{\mathbf{T}}) \text{ and if } \mathbf{A} \text{ is symmetric : } \frac{\partial \mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^{\mathbf{T}}\mathbf{A}$$

The vector of betas is hence:

$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$ (2.4) we can further differentiate (2.5) by $\hat{\boldsymbol{\beta}}$ in order to check the Second Order Condition and obtain: $\frac{\partial^2 \mathbf{u}^{\mathbf{T}}\mathbf{u}}{\partial \hat{\boldsymbol{\beta}}^2} = 2\mathbf{X}^{\mathbf{T}}\mathbf{X} > 0$, that is, we indeed have a minimum. From this point it follows that $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ must be invertible, hence it must be of full rank. This is only possible if the matrix X has full

column rank, i.e., our explanatory variables are linearly independent. (this is the condition of **no multicollinearity**).

It will make our life much easier if we introduce two important matrices. The first is the projection matrix (sometimes referred to as the "hat" matrix) (**P**): $\mathbf{P} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}$ (2.6), the second is the annihilator matrix (**M**): $\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}$ (2.7). These matrices are square matrices (nxn), symmetric, that is $\mathbf{P} = \mathbf{P}^{\mathbf{T}}$ and $\mathbf{M} = \mathbf{M}^{\mathbf{T}}$ and idempotent, i.e., $\mathbf{PP} = \mathbf{P}$ and $\mathbf{MM} = \mathbf{M}$.

Proof: $\mathbf{PP} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}} = \mathbf{P}$

$\mathbf{MM} = (\mathbf{I}_n - \mathbf{P})(\mathbf{I}_n - \mathbf{P}) = \mathbf{I}_n - 2\mathbf{P} + \mathbf{PP} = \mathbf{I}_n - 2\mathbf{P} + \mathbf{P} = \mathbf{I}_n - \mathbf{P} = \mathbf{M}$

The projection matrix projects **y** onto a column vector space defined by the explanatory variables **X**. That is:

$\mathbf{Py} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$ (2.8) Hence the projection matrix contains the weights and plays the same role as the weights in (1.10).

$\mathbf{My} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}})\mathbf{y} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{u}$ (2.9)

$\mathbf{MX} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}})\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X} = 0$ (2.10) and $\mathbf{My} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \mathbf{Me} = \mathbf{u}$ (2.11).

Unbiasedness: We can prove the unbiasedness of the OLS estimator as follows:

$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}\right) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{e}$ (2.12)

$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}E\left(\mathbf{X}^{\mathbf{T}}\mathbf{e}\right)$ (2.13). That is, if $E\left(\mathbf{X}^{\mathbf{T}}\mathbf{e}\right) = 0$ (**exogeneity**) the OLS estimates are unbiased.

Efficiency:

First we need the variance of the OLS estimator with unibiasedness assummed.

The variance of the estimator is then:

$\sigma^2_{\hat{\boldsymbol{\beta}}} = E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2\right) = E\left(\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{e}\mathbf{e}^{\mathbf{T}}\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\right) = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}E(\mathbf{e}\mathbf{e}^{\mathbf{T}})\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}$ (2.14)

If the error is **homoscedastic**, and **not autocorrealted** (this is a weak version of the condition of identically and independently distributed errors) then $E(\mathbf{e}\mathbf{e}^{\mathbf{T}}) = \begin{pmatrix} \sigma_e^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_e^2 \end{pmatrix} = \sigma_e^2\mathbf{I}_n$.

Hence (2.13) can be written in a much simpler form: $\sigma^2_{\hat{\boldsymbol{\beta}}} = \sigma_e^2\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}$.

Yet, we do not know the error variance, only the residual varince. We canhowever establish the relationship easily. Using (2.10):

$\mathbf{u}^{\mathbf{T}}\mathbf{u} = \mathbf{e}^{\mathbf{T}}\mathbf{M}^{\mathbf{T}}\mathbf{M}\mathbf{e} = \mathbf{e}^{\mathbf{T}}\mathbf{M}\mathbf{e}$ (2.15) Since this is a scalar, its value will be equal to its trace:
$tr(\mathbf{u}^{\mathbf{T}}\mathbf{u}) = tr(\mathbf{e}^{\mathbf{T}}\mathbf{M}\mathbf{e})$ for the trace there exists a rule regarding cyclic permutations, namely that
$tr(\mathbf{ABCD}) = \mathrm{tr}(\mathbf{DABC}) = \mathrm{tr}(\mathbf{CDAB}) = ...$ (2.16) using this rules we obtain that:

$tr(\mathbf{u}^{\mathbf{T}}\mathbf{u}) = tr\left(\mathbf{e}^{\mathbf{T}}\mathbf{e}\mathbf{M}\right) = \mathbf{e}^{\mathbf{T}}\mathbf{e}tr(\mathbf{M}) = \sigma_e^2 tr(\mathbf{M})$ (2.17)

But what is the trace of the annihilator matrix? The trace of an nxn identity matrix is n, and the trace of the projection matrix equals the rank of the matrix X), which is k.
$tr(\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}) = \mathrm{tr}((\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}) = \mathrm{tr}(\mathbf{I}_k) = \mathrm{k}$. Hence: $tr(\mathbf{M}) = tr(\mathbf{I}_n) - tr(\mathbf{P}) = n - k$ (2.18).

$\sigma_e^2 = \frac{\sigma_u^2}{n-k}$ (2.19) Here we received the same result for k<n parameters to be estimated as in (1.14) for k=2.

The effect of additional explanatory variables on the coefficient

Let assume that we have two sets of regressors, X_1 and X_2. If we regress y on both sets of variables:

$\mathbf{y} = \mathbf{X}_1\hat{\boldsymbol{\beta}} + \mathbf{X}_2\hat{\boldsymbol{\gamma}} + \mathbf{u}$ The residual will be: $\mathbf{u} = \mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}} - \mathbf{X}_2\hat{\boldsymbol{\gamma}}$ The sum of square residuals is:

$$\mathbf{u}^{\mathbf{T}}\mathbf{u} = \left(\mathbf{y}^{\mathbf{T}} - \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X}_1^{\mathbf{T}} - \hat{\boldsymbol{\gamma}}^{\mathbf{T}}\mathbf{X}_2^{\mathbf{T}}\right)\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}} - \mathbf{X}_2\hat{\boldsymbol{\gamma}}\right) =$$
$$= \mathbf{y}^{\mathbf{T}}\mathbf{y} - \mathbf{y}^{\mathbf{T}}\mathbf{X}_1\hat{\boldsymbol{\beta}} - \mathbf{y}^{\mathbf{T}}\mathbf{X}_2\boldsymbol{\gamma} - \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X}_1^{\mathbf{T}}\mathbf{y} + \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_1\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^{\mathbf{T}}\mathbf{X}_2^{\mathbf{T}}\mathbf{y} + \hat{\boldsymbol{\gamma}}^{\mathbf{T}}\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_1\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}^{\mathbf{T}}\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}}$$

which we seek to minimize by choosing the coefficient vectors:

$\frac{\partial \mathbf{u}^{\mathbf{T}}\mathbf{u}}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}_1^{\mathbf{T}}\mathbf{y} + 2\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}} + 2\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_1\hat{\boldsymbol{\beta}} = 0 \rightarrow \hat{\boldsymbol{\beta}} = \left(\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathbf{T}}\mathbf{y} - \left(\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}}$ (2.20) and

$\frac{\partial \mathbf{u}^{\mathbf{T}}\mathbf{u}}{\partial \hat{\boldsymbol{\gamma}}} = -2\mathbf{X}_2^{\mathbf{T}}\mathbf{y} + 2\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_1\hat{\boldsymbol{\beta}} + 2\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}} = 0 \rightarrow \hat{\boldsymbol{\gamma}} = \left(\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathbf{T}}\mathbf{y} - \left(\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_1\hat{\boldsymbol{\beta}}$ (2.21)

Or $\begin{pmatrix} \mathbf{X}_1^{\mathbf{T}}\mathbf{X}_1 & \mathbf{X}_1^{\mathbf{T}}\mathbf{X}_2 \\ \mathbf{X}_2^{\mathbf{T}}\mathbf{X}_1 & \mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2 \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1^{\mathbf{T}}\mathbf{y} \\ \mathbf{X}_2^{\mathbf{T}}\mathbf{y} \end{pmatrix}$ (2.22) which is the set of normal equations.

Hence we can see that the coefficients in a multivariate regression will reflect the effect of the correlation among the different regressors. If, and only if the two sets of regressors were uncorrelated, that is, $\left(\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_1 = \left(\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathbf{T}}\mathbf{X}_2 = 0$ could we expect that the coefficient from a regression of y on $\mathbf{X_1}$ would yield the same beta coefficients as in (1).

The Frisch-Waugh Theorem (also known as the Frisch-Waugh-Lovell Theorem)

Let us substitute (2.20) into (2.22)!

$$\begin{pmatrix} \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2 \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2 \end{pmatrix}\begin{pmatrix} \left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} - \left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}}\mathbf{y} \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{y} \end{pmatrix}$$

$$\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} - \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}} + \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}} = \mathbf{X}_2^{\mathrm{T}}\mathbf{y}$$

Let us define the projection matrix for the column vector space spanned by $\mathbf{X_1}$: $\mathbf{P}_1 = \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}$ and an annihilator matrix: $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$

$$\mathbf{X}_2^{\mathrm{T}}\mathbf{P}_1\mathbf{y} - \mathbf{X}_2^{\mathrm{T}}\mathbf{P}_1\mathbf{X}_2\hat{\boldsymbol{\gamma}} + \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2\hat{\boldsymbol{\gamma}} = \mathbf{X}_2^{\mathrm{T}}\mathbf{y}$$

$\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\gamma}} = \mathbf{X}_2^{\mathrm{T}}\mathbf{M}_1\mathbf{y} \rightarrow \hat{\boldsymbol{\gamma}} = \left(\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_1\mathbf{y}$ or, due to idempotence and symmetry

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_1^{\mathrm{T}}\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_1^{\mathrm{T}}\mathbf{M}_1\mathbf{y}$$

What is $\mathbf{M}_1\mathbf{y}$? It is the residual from a regression of y on X_1 only. Similarly, $\mathbf{M}_1\mathbf{X}_2$ is the set of residuals from the regressions of all columns of X_2 on X_1. The effect of X_1 on the coefficient vector γ is netted out or partialed out.

Frisch-Waugh theorem states that the coefficients from a multivariate regression are identical from a two-variate regression where the effect of all other variables is netted out. The coefficients from a multivariate regression hence can be interpreted as the partial effect of the variable in question on the dependent variable, that is, with all other effects removed.

But what is the practical importance of this another core result in statistics?

1. The idea of ceteris paribus is central in the methodology of economics, for example in comparative statics. In comparative statics we analyze the effect of a single variable or parameter on the outcome variable with all other factors fixed. Hence multivariate regressions are obvious ways to directly measure such relationships.
2. Have you ever considered what the right way is to regress y on x when you know that seasonal effects are present? Should you regress y on x with seasonal dummies included, or rather should you first deseasonalize y and x individually, and regress the deseasonalized y on the deseasonalized x? Frisch and Waugh have a good news to you. It is the same.

Practical example

We have data on the salary of employees and their education (years of education) and experience (years) (Ramanthan data6-4.gdt in Gretl). First we estimate the effect of both education and age on the logarithm of salary in a three-variate regression.

Model 2: OLS, using observations 1-49

Dependent variable: l_WAGE

```
Model 2: OLS, using observations 1-49
Dependent variable: l_WAGE

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       6,85060      0,135079     50,72     5,03e-042 ***
  EDUC        0,0645046    0,0165612     3,895    0,0003    ***
  EXPER       0,0229541    0,00628452    3,652    0,0007    ***

Mean dependent var   7,454952   S.D. dependent var   0,312741
Sum squared resid    3,157276   S.E. of regression   0,261986
R-squared            0,327484   Adjusted R-squared   0,298245
F(2, 46)             11,19995   P-value(F)           0,000109
Log-likelihood      -2,346281   Akaike criterion     10,69256
Schwarz criterion    16,36802   Hannan-Quinn         12,84582

Log-likelihood for WAGE = -367,639
```

Now we are going to partial out the effect of experience on education.

First we regress the log wage on experience and save the residual (res1).

```
Model 3: OLS, using observations 1-49
Dependent variable: l_WAGE

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       7,31134      0,0744048    98,26     4,64e-056 ***
  EXPER       0,0162518    0,00689559    2,357    0,0226    **

Mean dependent var   7,454952   S.D. dependent var   0,312741
Sum squared resid    4,198521   S.E. of regression   0,298882
R-squared            0,105694   Adjusted R-squared   0,086666
F(1, 47)             5,554725   P-value(F)           0,022649
Log-likelihood      -9,329333   Akaike criterion     22,65867
Schwarz criterion    26,44231   Hannan-Quinn         24,09417

Log-likelihood for WAGE = -374,622
```

Then we regress the education on experience and save the residual (res2).

```
Model 4: OLS, using observations 1-49
Dependent variable: EDUC

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        7,14266     0,574431     12,43     1,81e-016 ***
  EXPER       -0,103904    0,0532364    -1,952    0,0569    *

Mean dependent var   6,224490   S.D. dependent var   2,374038
Sum squared resid    250,2481   S.E. of regression   2,307472
R-squared            0,074973   Adjusted R-squared   0,055292
F(1, 47)             3,809332   P-value(F)           0,056941
Log-likelihood      -109,4785   Akaike criterion     222,9570
Schwarz criterion    226,7406   Hannan-Quinn         224,3925
```

```
Model 5: OLS, using observations 1-49
Dependent variable: res1

             coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------
  const       0,000000     0,0370262    0,0000    1,0000
  res2        0,0645046    0,0163841    3,937     0,0003  ***

Mean dependent var   0,000000   S.D. dependent var   0,295752
Sum squared resid    3,157276   S.E. of regression   0,259184
R-squared            0,248003   Adjusted R-squared   0,232003
F(1, 47)             15,50022   P-value(F)           0,000271
Log-likelihood      -2,346281   Akaike criterion     8,692561
Schwarz criterion    12,47620   Hannan-Quinn         10,12807
```

Which indeed yield the same coefficient as the education has in the multivariate regression.