REGRESSION ANALYSIS AND CAUSALITY WITH R

João Cerejeira[1] Miguel Portela[1,2,3]

[1]NIPE – UMinho
[2]IZA, Bonn
[3]Banco de Portugal

October 26, 2021

Instructors

João Cerejeira
[joao.cerejeira@eeg.uminho.pt]
https://sites.google.com/site/joaocerejeira/

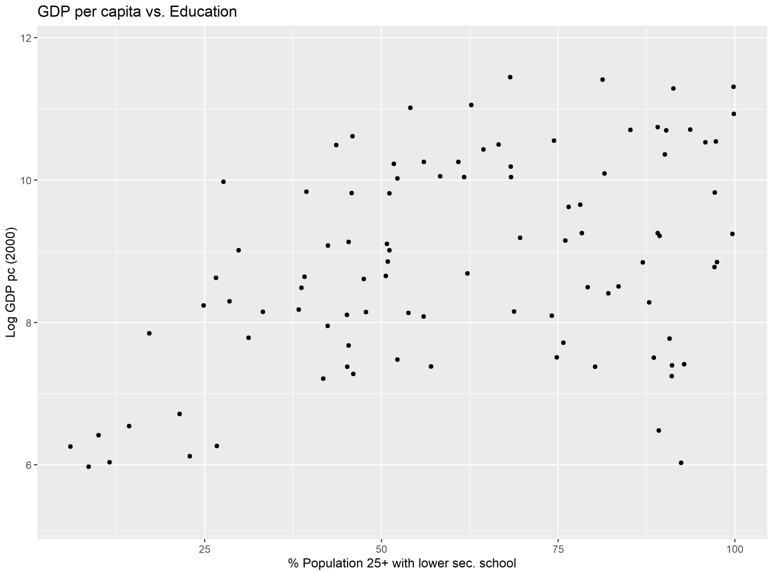
Miguel Portela
[miguel.portela@eeg.uminho.pt]
http://www1.eeg.uminho.pt/economia/mangelo/

Outline

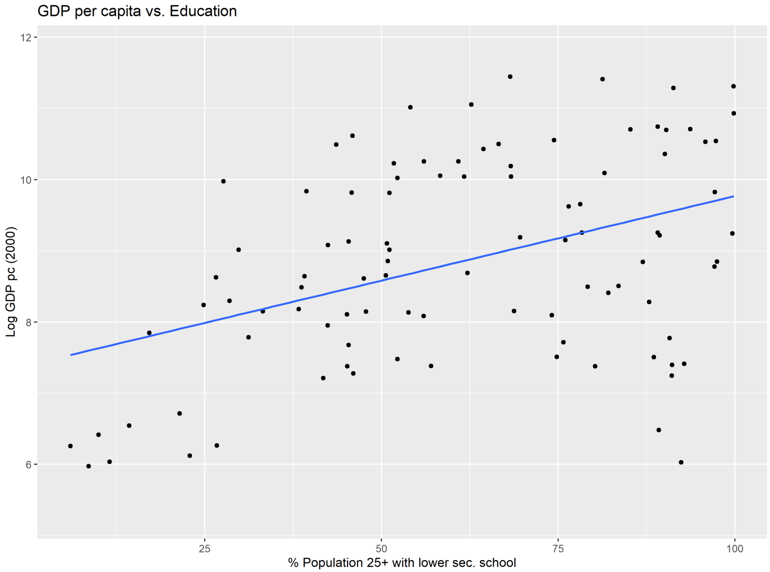
- Computing linear regression estimates
- Regression as a method-of-moments-estimator
- Regression residuals
- Sampling distribution of regression estimates
- Presenting regression estimates
- The ANOVA table
- Hypothesis tests, linear restrictions
- Computing residuals and predicted values
- Specification issues
- The generalized linear regression model
- Heteroskedasticity: causes and test
- Types of heteroskedasticity
- Robust estimation
- The GLS and FGLS estimator

The conditional mean of a response variable y as a linear function of k independent variables is:

$$E[y|x_1, x_2, ..., x_k] = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k. \tag{1}$$


GDP per capita vs. Education
Log GDP pc (2000)
12
10
8
6
25
50
75
100
% Population 25+ with lower sec. school

$$E[\text{Gdp per capita}|\text{education}] = \beta_1 + \beta_2[\text{education}]$$


GDP per capita vs. Education
12
10
8
6
Log GDP pc (2000)
25
50
75
100
% Population 25+ with lower sec. school

But we don't know the population values $\beta_{1,}\beta_2, ..., \beta_k$. We work with a sample of N observations of data from population.
Using this information, we must:
- obtain estimates of the coefficients $\beta_{1,}\beta_2, ..., \beta_k$;
- estimate their variance;
- test coefficients estimate;
- use estimated $\widehat{\beta}_{1,}\widehat{\beta}_2, ..., \widehat{\beta}_k$ to interpret the model.

The linear regression model has the form:

$$y_i = \beta_1 x_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i \tag{2}$$

with $i = 1, 2, .., N$.

In matrix form:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{3}$$

where $\mathbf{X}$ is an $N \times k$ matrix of sample values.

Regression as a method-of-moments-estimator

The key assumption in the linear regression model is:

$$E[u|\mathbf{x}] = 0. \tag{4}$$

The unobserved factors involved in the regression function are not related systematically to observed factors. For linear relationships, the later assumption implies:

$$E[\mathbf{x}'u] = \mathbf{0}$$

$$E[\mathbf{x}'(\mathbf{y} - \mathbf{x}\beta)] = \mathbf{0}. \tag{5}$$

Substituting calculated moments from our sample into the expression, yields:

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{0} \tag{6}$$
$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \tag{7}$$

Regression residuals

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\beta}. \tag{8}$$

The estimator of population variance of the stochastic disturbance is:

$$s^2 = \frac{\sum_{i=1}^{N} \widehat{u}_i^2}{N-k} = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{N-k}. \tag{9}$$

$\sqrt{s^2}$—standard error of regression or root mean squared error.

Sampling distribution of regression estimates

The OLS estimator $\widehat{\beta}$ is a vector of random variables because it is a function of the random variable y, which in turn is a function of the stochastic disturbance u.

The OLS estimator takes on different values for each sample of N observations drawn from the population.

Assume that u_i are independent draws from a identical distribution (i.i.d).

Large sample theory shows that the sampling distribution of the OLS estimator is approximately normal.

OLS estimator $\widehat{\beta}$ has a large sample normal distribution with expected value β and variance $\sigma^2\mathbf{Q}^{-1}$, where $\mathbf{Q}^{-1}$ is the variance-covariance matrix of $\mathbf{X}$ in the population.

Because $\sigma^2\mathbf{Q}^{-1}$ is unknown, a consistent estimator of $\sigma^2\mathbf{Q}^{-1}$ is $s^2(\mathbf{X}'\mathbf{X})^{-1}$.

Example 1

$$
\begin{aligned}
\log(Grow\_GDPper\ capita_c) &= \beta_1 + \\
&+\beta_2[\log GDPpc2000_c] + \\
&+\beta_2[\%\ Educ\_Sec_c] + \\
&+\beta_3[Invest.Grow_c] + \\
&+\beta_4[Trade2000_c] + \\
&+\beta_5[Gov2000_c] + u_c
\end{aligned}
$$

Presenting regression estimates

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.2333 -0.4785 -0.0029  0.6294  3.5823 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5.557012   1.843849   3.014  0.00368 ** 
logGDPpc2000  -0.726476   0.216851  -3.350  0.00135 ** 
educ_sec       0.032574   0.006704   4.859 7.79e-06 ***
invest_growth  0.232956   0.043563   5.348 1.23e-06 ***
trade2000      0.005198   0.002567   2.025  0.04696 *  
gov2000       -0.096830   0.313265  -0.309  0.75823    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.224 on 65 degrees of freedom
  (147 observations deleted due to missingness)
Multiple R-squared:  0.5909,	Adjusted R-squared:  0.5594 
F-statistic: 18.78 on 5 and 65 DF,  p-value: 1.674e-11
```

Presenting regression estimates

The ANOVA table

```
               Df Sum Sq Mean Sq F value    Pr(>F)    
logGDPpc2000    1  61.16   61.16  40.825 2.05e-08 ***
educ_sec        1  28.02   28.02  18.704 5.36e-05 ***
invest_growth   1  45.29   45.29  30.229 6.91e-07 ***
trade2000       1   6.04    6.04   4.033   0.0488 *  
gov2000         1   0.14    0.14   0.096   0.7582    
Residuals      65  97.38    1.50                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
147 observations deleted due to missingness
```

`Sum Sq`=$\widehat{\mathbf{y}}'\widehat{\mathbf{y}} = \sum_{i=1}^{N} \left(\widehat{y}_i - \overline{y}\right)^2$, is the sum of the squares of the deviations of the predicted values of y from the mean value of y.
`Residual SS`=$\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = \sum_{i=1}^{N} \widehat{u}_i^2 = \sum_{i=1}^{N} \left(y_i - \widehat{y}_i\right)^2$.
`Total SS`= $\widetilde{\mathbf{y}}'\widetilde{\mathbf{y}} = \sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2$, where $\widetilde{\mathbf{y}} = y - \overline{y}$.

`R-squared=Model SS / Total SS=1-(Residual SS / Total SS)=`R^2
`Adj R-squared=`$1 - \left(1 - R^2\right)\frac{N-1}{N-k}$

The other measures to compare competing regression models are the Akaike information criterion (AIC) and Bayesian information criterion (BIC, or Schwarz criterion).
These measures account for both the goodness of fit and its parsimony by rewarding improvements in the goodness of fit and penalizing the additional degrees of freedom.
The preferred model is the one with the minimum AIC or BIC value.The AIC penalizes the number of parameters less strongly than does the Bayesian information criterion.

F statistic=Model MS / Residual MS

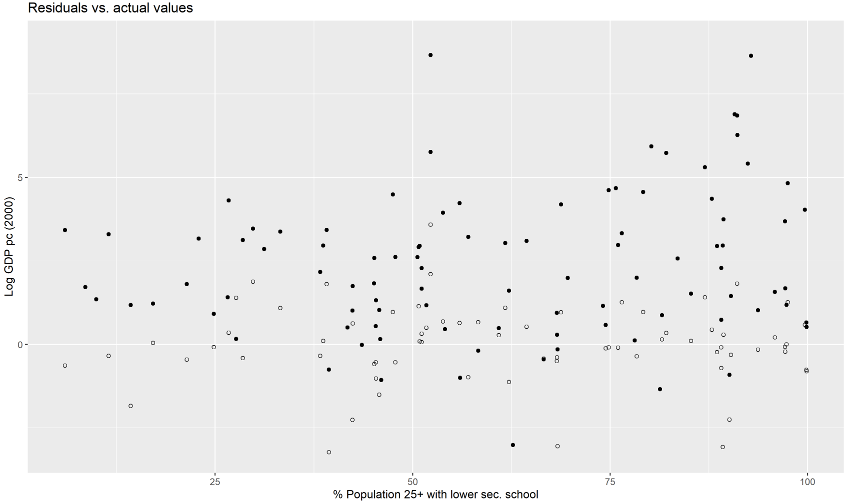
Root MSE=$\sqrt{\text{Residual MS}}$

Estimated variance covariance (VCE) matrix $s^2(\mathbf{X}'\mathbf{X})^{-1}$.

	(Intercept)	logGDPpc2000	educ_sec
(Intercept)	3.3997778995	-3.838101e-01	2.808327e-04
logGDPpc2000	-0.3838100668	4.702440e-02	-3.548833e-04
educ_sec	0.0002808327	-3.548833e-04	4.494372e-05
invest_growth	-0.0238176748	1.597596e-03	2.344525e-05
trade2000	-0.0004613550	-5.556980e-06	7.067824e-07
gov2000	0.4269619849	-5.055677e-02	-1.074546e-04

	invest_growth	trade2000	gov2000
(Intercept)	-2.381767e-02	-4.613550e-04	0.4269619849
logGDPpc2000	1.597596e-03	-5.556980e-06	-0.0505567717
educ_sec	2.344525e-05	7.067824e-07	-0.0001074546
invest_growth	1.897755e-03	-7.114476e-06	0.0015162991
trade2000	-7.114476e-06	6.587727e-06	-0.0001871097
gov2000	1.516299e-03	-1.871097e-04	0.0981352360

Predicted residuals



Hypothesis tests, linear restrictions

Three tests are commonly used in econometrics: Wald tests, Lagrange multiplier (LM) tests and likelihood-ratio (LR) tests.
Here I present the Wald tests.
Given the population regression equation:

$$y = x\beta + u$$

any set of linear restrictions on the coefficient vector may be expressed as

$$\mathbf{R}\beta = \mathbf{r}.$$

Example: Wald test

Given:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

we want to test $H_0 : \beta_2 = 0$. The restriction is:

$$\mathbf{R} = \begin{Bmatrix} 0 & 1 & 0 \end{Bmatrix}$$

$$\mathbf{r} = (0)$$

Given the hypothesis $H_0 = \mathbf{R}\beta = 0$, the Wald statistic is:

$$W = (\mathbf{R}\widehat{\beta} - \mathbf{r})'\{\mathbf{R}(\widehat{\mathbf{VCE}})\mathbf{R}'\}^{-1}(\mathbf{R}\widehat{\beta} - \mathbf{r}) \tag{10}$$

w has a large-sample χ^2 distribution when H_0 is true. In small samples w/q is better approximated by an F distribution with q (the number of restrictions) and $(N-k)$ degress of freedom. If $q = 1$, $\sqrt{w}$ can be approximated by a Student t distribution with $(N-k)$ d.f.

Since we know the distribution of w when H_0 is true, the standard hypothesis test is:

$$\Pr(\text{Reject } H_0 \mid H_0 \text{ is true}) = \alpha \tag{11}$$

where α is the significance level of the test.
Stata presents $p-$values, which measure the evidence against H_0 - the largest significance level at which a test can be conducted without rejecting H_0.

Example:

Coefficient of `education` `.032574`, with s.d. `.006704`.
t statistic of the the null

$$H_0 : \beta_{[Educ\_Sec]} = 0$$

is

$$= \widehat{\beta}_{[Educ\_Sec]} / s.d. \left(\widehat{\beta}_{[Educ\_Sec]} \right) = .0325741 / .006704 = 4.859$$

Confidence intervals:

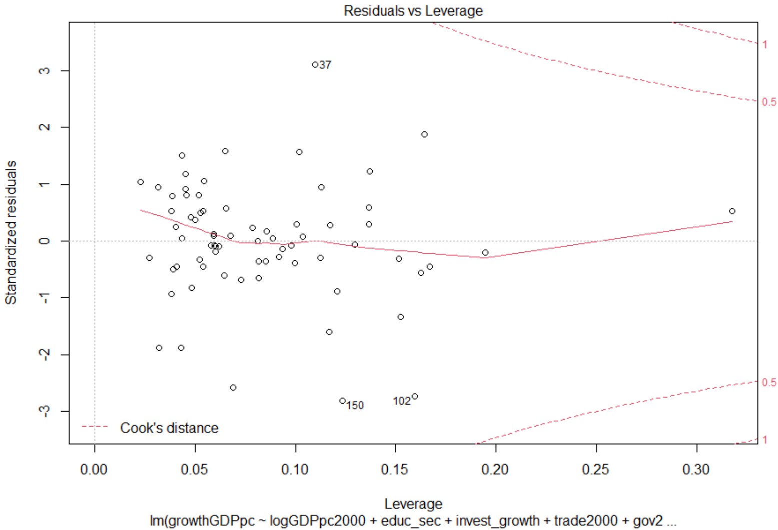
$$\widehat{\beta}_{[Educ\_Sec]} - s.d.(\widehat{\beta}_{[Educ\_Sec]}) \times t_{crit,\ 5\%} \leq \beta_{[Educ\_Sec.]} \leq \widehat{\beta}_{[Educ\_Sec]} + s.d.(\widehat{\beta}_{[Ed$$

$$.03257 - .00670 \times 1.99714 \leq \beta_{[Educ\_Sec]} \leq ..03257 + .00670 \times 1.99714$$

$$.01919 \leq \beta_{[Educ\_Sec]} \leq .04597$$

Detecting Outliers

An outlier is a data point with an unusual value (observed or residual). Evidence that the model's coefficients are strongly influenced by a few data points casts doubt on the fitted model's worth in a broder context. A data point has a high degree of leverage on the estimates if including it in the sample alters considerably the estimated coefficients The leverage values are computed from the diagonal elements of the matrix $h_i = x_i(\mathbf{X}'\mathbf{X})^{-1}x_i'$.


Residuals vs Leverage
Standardized residuals
3
2
1
0
-1
-2
-3
37
150
102
Cook's distance
1
0.5
0.5
1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(growthGDPpc ~ logGDPpc2000 + educ_sec + invest_growth + trade2000 + gov2 ...

The generalized linear regression model

Suppose that $\Sigma_u \neq \sigma^2 I_N$. The OLS estimator is unbiased, consistent, but is no longer efficient as demonstrated by:

$$\begin{aligned}
\widehat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}'\beta + \mathbf{u}) \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\
& E[\widehat{\beta} - \beta] = 0.
\end{aligned} \tag{12}$$

$$\begin{aligned}
Var[\widehat{\beta}|\mathbf{X}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Sigma_u\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned} \tag{13}$$

The VCE computed by regress is $s_u^2(\mathbf{X}'\mathbf{X})^{-1}$. When $\Sigma_u \neq \sigma^2 I_N$ this estimator of the VCE is not consistent and the usual inference procedures are inappropriate.

Heteroskedasticity: causes

Potential causes of heteroskedasticity:
- disturbances are often related to some measure of scale (e.g. income);
- disturbances are homoskedastic within groups but heteroskedastic between groups;
- grouped data, in which each observation is the average of microdata.

Types of heteroskedasticity

In the identically distributed assumption:

$$\Sigma_u = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 I_N. \tag{14}$$

If the diagonal elements differ:

$$\Sigma_u = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \sigma_N^2 \end{pmatrix} \tag{15}$$

If errors are correlated within clusters (m clusters) of observations, we have:

$$\Sigma_u = \begin{pmatrix} \Sigma_1 & 0 & ... & 0 \\ 0 & \Sigma_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & \Sigma_M \end{pmatrix} \tag{16}$$

Serial correlation in time-series regression models:

$$\Sigma_u = \sigma_u^2 \begin{pmatrix} 1 & \rho_1 & ... & \rho_{N-1} \\ \rho_1 & 1 & ... & \rho_{2N-3} \\ ... & ... & ... & ... \\ \rho_{N-1} & \rho_{2N-3} & 0 & 1 \end{pmatrix} \tag{17}$$

The robust estimator of the VCE

The term we must estimate $\{\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\} = \{\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X}\}$ is sandwiched between the $(\mathbf{X}'\mathbf{X})^{-1}$ terms. Huber (1967) and White (1980) showed that:

$$\widehat{S}_0 = \frac{1}{N}\sum_{i=1}^{N} \widehat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i, \tag{18}$$

consistently estimates $\{\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X}\}$ when the u_i are conditionally heteroskedastic. The robust estimator of the VCE is:

$$Var[\widehat{\beta}|\mathbf{X}] = \frac{N}{N-k}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N} \widehat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i\right)(\mathbf{X}'\mathbf{X})^{-1}. \tag{19}$$

The GLS and FGLS estimator

With a known Σ_u matrix, we can premultiply the model by $\mathbf{P}'=\Sigma_u^{-1}$:

$$\begin{aligned} \mathbf{P}'\mathbf{y} &= \mathbf{P}'\mathbf{X}\beta+\mathbf{P}'\mathbf{u} && (20) \\ \mathbf{y}^* &= \mathbf{X}^*\beta+\mathbf{u}^* && (21) \end{aligned}$$

where

$$Var[\mathbf{u}^*] = E[\mathbf{u}^*\mathbf{u}^{*\prime}] = \mathbf{P}'\Sigma_u\mathbf{P}' = \mathbf{I}_N. \qquad (22)$$

Then,

$$\widehat{\beta}_{GLS} = (\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}(\mathbf{X}^{*\prime}\mathbf{y}^{*}) \tag{23}$$

and

$$Var[\widehat{\beta}_{GLS}|\mathbf{X}] = (\mathbf{X}'\Sigma_u^{-1}\mathbf{X})^{-1}. \tag{24}$$

The FGLS estimator is applied when Σ_u is not known and if we have a consistent estimator of Σ_u, denoted $\widehat{\Sigma}_u$, replacing $\mathbf{P}'$ with $\widehat{\mathbf{P}}'$.
In grouped data we can estimate FGLS models multiplying original data with proper weights.

References

Hill, R. Carter, William E. Griffiths, Guay C. Lim, 2018. Principles of Econometrics, 5th Ed., Wiley. [Companion with R: Colonescu, Constantin (2016). Principles of Econometrics with R. Available at:https://bookdown.org/ccolonescu/RPoE4/]

Wooldridge, J. (2020) Introductory Econometrics: A Modern Approach, 7th Ed., Cengage Learning. [Companion with R: Heiss, Florian (2020). Using R for Introductory Econometrics. Available at: http://www.urfie.net/.]