Microeconomics

Labour market economics: explaining individual wages

Programming projects 2020

Contents

1 Introduction

This documentation presents our study concerning Belgium individual wages by analyzing the provided dataset and by explaining data insight thanks to the **Ordinary Least Squares** method we defined.

Wages in Belgium[3] dataset consist of 1472 individuals, randomly sampled from the working population in Belgium for the year 1994. It contains 893 males and 579 females taken from the Belgian part of the European Community Household Panel.

The following sections will explain how to run the program, describe the script organization with a short description of the implemented functions. Section 4 provides a brief description of the economic problem followed by section 5 in which we present the main mathematical/computational methods applied. The last section present a showcase for our main function, considering wages in Belgium.

2 How to run the program

Project folder contains the main R script called *OLS.R* which contains the definition of the main `OLS(...)` function. This folder also contains another subfolder (*wages_in_Belgium*) where data file is stored as *bwages.dat*.
The program needs the following packages to run correctly:

- *rstudioapi* - used to automatically set the working directory in RStudio
- *tibble* - used to group output in a neat data structure
- *knitr* - used to format a table-like

These packages are automatically installed and loaded in the initial section of the program. Running the code should be straightforward and there is no need for user intervention.

3 User guide

After the initial loading of the required packages, the working directory is automatically set considering the path where the *OLS.R* script is located. The only constant value needed to be specified is the path where to find the dataset, starting from project root. This constant has been preset to load wages in Belgium and stored in the "*data*" variable.

Then our code includes the definition of the main `OLS(...)` function which is based on auxiliary functions to delegate sub-problem computation such as data validity test, regression estimates and statistics calculus.

The last part of the code is organized as a showcase for testing the main function, presenting regression results over different variables.

3.1 Functions

OLS(...) represent the main function, the only function intended to be used by the user. This method performs the *Ordinary Least Squares* estimation method in order to estimate the parameters of a linear regression model. Providing dependent and independent variables as a matrix-like object, it's possible to specify whether the function should *solve directly* the problem using the traditional formula to estimate coefficients $b = (X'X)^{-1} * X'y$, or use the *QR-decomposition* method and the appropriate formula $b = R^{-1} * Q' * y$. This can be done by setting the `useQRdecomposition` logical parameter. Also, this function allows to specify logical parameters such as `skipDataValidityTest` to enable/disable data validity check, `robustErrors` in order to request *heteroskedasticity-robust errors* computation, or `returnTable` to output results in the form of a table. The whole problem is divided into sub-problems solved the following auxiliary functions.

TestDataValidity(...) auxiliary function to check the validity of the data and give errors and warnings, including less observations than variables, special values check (NaN, Inf) and collinearity.

TestCorrelation(...) auxiliary function to test bivariate correlation between input regressors.

QRdecomposition(...) auxiliary function that performs QR decomposition over regressors data. This operation is performed when QR decomposition is requested, and returns an orthogonal matrix *Q* and an upper triangular matrix *R*.

CalculateStatistics(...) auxiliary functions dedicated to regression statistics computation, such as R-squared, adjusted R-squared, residuals standard error, coefficients standard errors, t-statistics, F-statistics and p-values.

CustomPlot(...) this simple function is just a wrapper to customize the process of plotting fitted-residuals.

4 Description of the economic problem

Prediction of wages using regression methods is a recurrent topic[2], in this paper we consider the wages in Belgium dataset, consisting of 1472 individuals (893 males and 579 females) and we want to find out how factor such as gender, education level and years of experience affect the wage rate distribution. The provided dataset is composed by four main variables:

wage	gross hourly wage rate in euro
male	dummy variable, 1 = male, 0 = female
educ	education level from 1 [low] to 5 [high]
exper	years of experience

Also, logaritmic version of some of these variables are provided as: *lnwage = ln(wage)*, *lnexper = ln(1+exper)* and *lneduc = ln(educ)*. Data is available in the project folder following the path `/wages_in_Belgium/bwages.dat`.
An initial data summary is provided in the Figure 1.

```
      WAGE            LNWAGE           EDUC           EXPER          LNEXPER         LNEDUC          MALE
 Min.   : 2.191   Min.   :0.7843   Min.   :1.000   Min.   : 0.00   Min.   :0.000   Min.   :0.000   Min.   :0.0000
 1st Qu.: 8.113   1st Qu.:2.0935   1st Qu.:3.000   1st Qu.: 9.00   1st Qu.:2.303   1st Qu.:1.099   1st Qu.:0.0000
 Median :10.127   Median :2.3152   Median :3.000   Median :16.50   Median :2.862   Median :1.099   Median :1.0000
 Mean   :11.051   Mean   :2.3344   Mean   :3.378   Mean   :17.22   Mean   :2.691   Mean   :1.137   Mean   :0.6067
 3rd Qu.:12.755   3rd Qu.:2.5460   3rd Qu.:4.000   3rd Qu.:24.00   3rd Qu.:3.219   3rd Qu.:1.386   3rd Qu.:1.0000
 Max.   :47.576   Max.   :3.8623   Max.   :5.000   Max.   :47.00   Max.   :3.871   Max.   :1.609   Max.   :1.0000
```

Figure 1: Data summary

And Figure 2 shows the histogram of wage values.

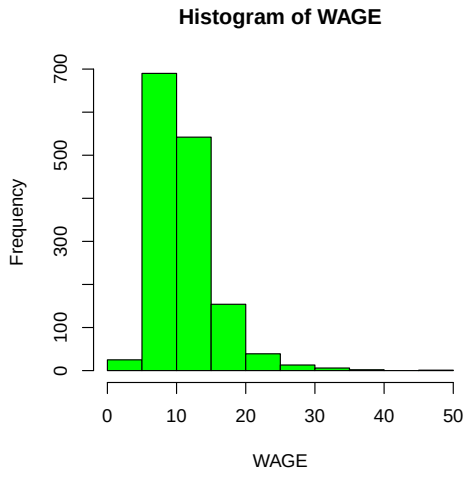

Figure 2: Wages histogram

In section 6 we'll use our OLS function to get some data insight and explain their relations.

5 Mathematical/computational methods

5.1 Ordinary Least Squares

In econometrics, Ordinary Least Squares method is widely used to estimate the parameter of a linear regression model. The aim of this method is to minimize the sum of the squared errors, that is the difference between observed values (ground truth) and predicted (fitted) values.

Geometrically, this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression surface. Obviously, small differences means better model fit capacity.

The OLS estimator is consistent when the regressors (independent variables) are exogenous, and—by the Gauss–Markov theorem—optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated. Under these conditions, the method of OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances.

The traditional matrix formulation for coefficients estimation is expressed as matrix multiplication between the inverse of the *normal matrix* $X'X$ and the *moment matrix* $X'y$ resulting in the following fomula: $b = (X'X)^{-1} * X'y$ where X represent regressors, y is the vector of observed values and $'$ is the *transpose* operator.

5.2 Regression statistics

Usually, linear regression methods include a series of values and statistics that describe the regression results. Once coefficents have been calculated, predicting data and calculating residuals is straightforward. Using all these informations and general data metrics such as observations number, variables number, mean and standard deviation, we can compute the following values.

Standard error represents the average distance that the observed values fall from the regression line. Smaller values are better because indicate that the observations are closer to the fitted line.

R squared is a popular measure for the goodness-of-fit, it states the proportion of the variance of dependent variable that is explained by the model. Sometimes the R2 is inter-

preted as a measure of quality of the statistical model, while in fact it measures nothing more than the quality of the linear approximation. Analytical R2 formula is the following: $R2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$

Adjusted R squared is a statistics that corrects R2 in order to consider explanatory power of additional variables in the model, making it possible to decrease if these regressors aren't meaningful. This is usually done by correcting the variance estimates for the degrees of freedom, the new formula looks like this: $1 - \frac{(1-R^2)*(n-1)}{n-k-1}, n = obs.number, k = vars.number$

T-statistic is the coefficient divided by its standard error. This statistic is used in a t-test to determine if you should support or reject the null hypothesis, it's also used to calculate p-values and determine whether a regressor has a significant impact on the dependent variable or not.

F-statistic is a statistic commonly used to compare statistical models that have been fitted to a dataset, in order to identify the model that best fits the population from which the data were sampled. When comparing a regression model with a restricted model containing only an intercept term the f-statistic can be written as: $F = \frac{R^2/(k-1}{(1-R^2)/(n-k)}$

5.3 QR decomposition

A problem with the traditional approach is that the matrix inverse is both computationally expensive and numerically unstable. An alternative approach is to use a matrix decomposition to avoid this operation.

The *QR matrix decomposition* allows to express a matrix as a product of two separate matrices: the orthogonal matrix Q (meaning that $Q'Q = I$) and an upper triangular matrix R.

The approach still involves a matrix inversion, but in this case only on the simpler R matrix. Using this method a new formula for coefficients estimation need to be defined: $b = R^{-1} * Q' * y$

5.4 Heteroskedasticity

Heteroscedasticity is a systematic change in the spread of the residuals over the range of measured values. This is a problem because Ordinary Least Squares regression assumes

homoscedasticity, that is all residuals are drawn from a population that has a constant variance.

Heteroscedasticity[1] does not cause problems for estimating the coefficients, it only causes problems for getting the "correct" standard errors.

Thus to mitigate this problem the traditional solution is provided by the White estimator which computes the variance-covariance matrix of the coefficient vector as:
$(X'X)^{-1}X'\hat{\Sigma}X(X'X)^{-1}$.

6 Wages in Belgium showcase

This section provide a showcase for testing our function, we initially used a few simple data points (2D) found on the internet in order to show pratically that the method is working as intended, that is, it's performing a coherent linear regression. A plot of these results is presented in the Figure 3.

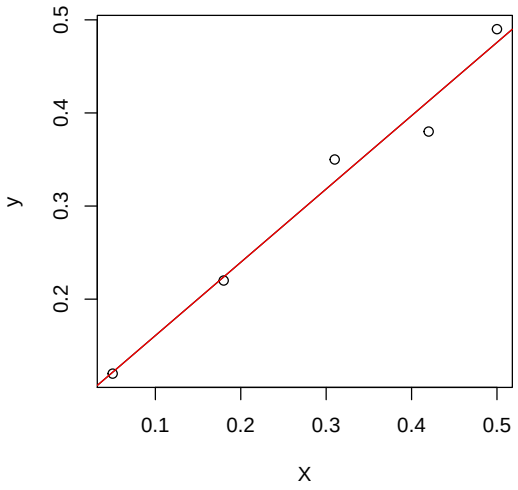

Figure 3: Initial functionality test, simple data

Afterwards we tested that the function returned errors when providing independent variables having less observations than variables and correlated variables. For demonstration purposes these errors have been caught and a simple character string has been printed in the console.

WAGE regression on MALE After this initial testing phase we considered Belgium individual wages data and run a first regression of WAGE on MALE dummy variable (variable that can only assume binary 0-1 values).

term	coefficients	standardError	tStatistic	pValue
INTERCEPT	10.261544	0.1831234	56.036229	0.00e+00
MALE	1.300687	0.2351105	5.532238	3.74e-08

ResidualsStandardError	R2	R2adj	fStatistic
4.404894	0.0203955	0.0190618	30.60566

Figure 4: Regression of WAGE on MALE results

Figure 4 shows that the estimated gross hourly wage rate for males is 10.26+1.3 = 11.56€ and 10.26€ for females. Looking at the R2 value we see that the 20% of the variation in individual wages can be attributed to gender differences.

WAGE regression on MALE, FEMALE Computing the FEMALE variable, which can be seen as the logical negation of the MALE variable, and adding it to the previous model as a regressor results in an error. This is obvious because the two variables are in an *exact linear combination* relation, that is, one of the two variables is redundant and doesn't bring more information to the model. Also, checking with the built-in R function *lm()* we can see that the FEMALE variable gets ignored.

WAGE regression on MALE, EDUC, EXPER, EXPER2 The next step in our showcase is regressing WAGE on MALE, EDUC, EXPER, EXPER2 (to be considered as EXPER^2) in order to show how regressors affect hourly wage rate. Figure 5 shows regression results.

term	coefficients	standardError	tStatistic	pValue
INTERCEPT	-0.8924849	0.4329127	-2.061582	3.94e-02
MALE	1.3336935	0.1908668	6.987562	4.23e-12
EDUC	1.9881267	0.0798526	24.897461	2.02e-114
EXPER	0.3579993	0.0316566	11.308845	1.72e-28
EXPER2	-0.0043692	0.0007962	-5.487409	4.80e-08

ResidualsStandardError	R2	R2adj	fStatistic
3.509021	0.3783418	0.3762215	223.2044

Figure 5: Regression of WAGE on EDUC, EXPER, EXPER2 results

Results show that EDUC and EXPER are statistically highly significant as their t-statistics have high values (respectively 24,9 and 11,31). The introduction of the EXPER2 term can be interpreted as modeling the fact that experience affects an individual's wage nonlinearly,

decreasing after many years of experience. In fact, results show that EXPER2 has a negative coefficient confirming our statement. Thus including this variable significantly improves the model. Also looking at the R2 value we can see that this model is explain wages differential much better that the first simple regression model, as the new R2 value amounts to 0.378.

Next we plotted residuals versus the fitted values to check if this model suffer from heteroskedasticity, invalidating the homoskedasticity assumption.

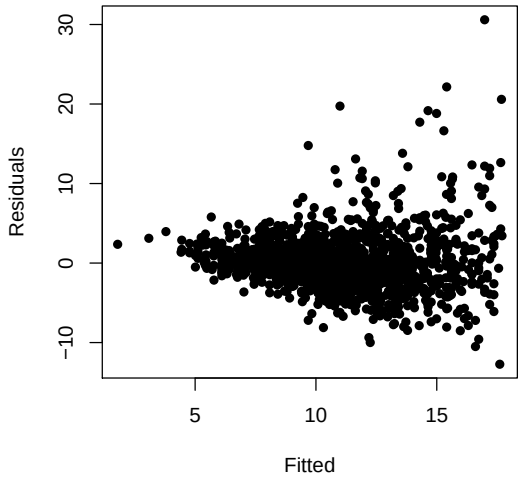

Figure 6: Residuals versus the fitted values - linear model

Indeed as shown in Figure 6 variation in the residuals increase for higher fitted values, proving a clear symptom of heteroskedasticity and thus standard errors are biased and not appropriate.

LNWAGE regression on MALE, LNEDUC, LNEXPER, LNEXPER2 to mitigate heteroskedasticity problem we regress a log-linear model of LNWAGE on MALE, LNEDUC, LNEXPER, LNEXPER2 (there is no need to apply logarithm to MALE dummy variable, as it would be inconsistent). Results of this log-linear model are presented in Figure 7.

term	coefficients	standardError	tStatistic	pValue
INTERCEPT	1.2627056	0.0663418	19.033337	2.32e-72
MALE	0.1179433	0.0155711	7.574488	6.35e-14
LNEDUC	0.4421764	0.0181921	24.305988	6.13e-110
LNEXPER	0.1098205	0.0543838	2.019360	4.36e-02
LNEXPER2	0.0260073	0.0114762	2.266193	2.36e-02

ResidualsStandardError	R2	R2adj	fStatistic
0.2858605	0.3782603	0.3761398	223.1271

Figure 7: Regression of LNWAGE on MALE, LNEDUC, LNEXPER, LNEXPER2 results

These results aren't directly comparable with the previous ones because the nature of the model is different, in fact now the coefficient of MALE measures the relative difference in expected wages for males and females. Let's plot residuals versus fitted values to check the heteroskedasticity question.

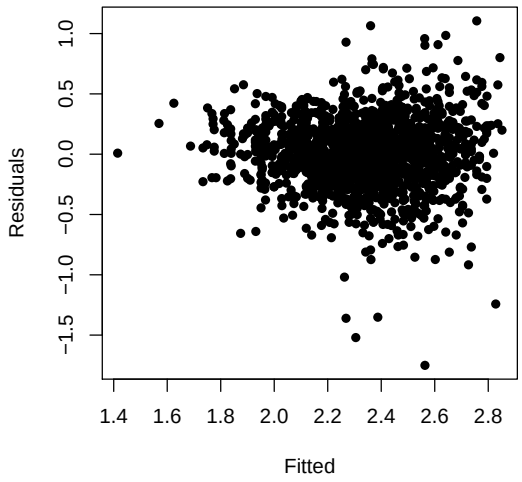

Figure 8: Residuals versus the fitted values - log-linear model

Looking at the Figure 8, we can see that the variation in the residuals for higher fitted values is much less pronounced, that is, its magnitude is much smaller in the log-linear model than the previous linear model, suggesting a better and more correct calculus of standard errors.

Discussion of EDUC Looking at the EDUC variable we can see that it represents an individual's level of education starting from 1, which means low, to 5, which means high. This kind of variables are called "categorical" as they're representing unique factor values,

as opposed to continuous values. Figure 10 shows how EDUC values are distributed.

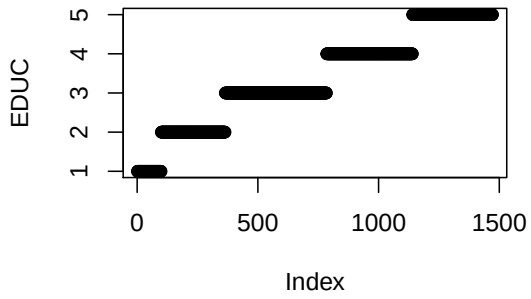

Figure 9: EDUC data distribution

Using EDUC categorical variable in our OLS regression as a single variable restrict the estimation power, thus to unrestrictedly estimate coefficients for each level of education we have to encode our single EDUC variable in n different dummy variables (binary 0-1), where n is equal to the numerosity of unique values in the original categorical variable. This process is called *One Hot Encoding* and it's performed by our function `OneHotEncode(...)`.

Applying this procedure to EDUC, we get 5 new dummy variables ready to be used in a linear regression model. Recall that to avoid *exact multicollinearity*, it's sufficient to include only four of the five variables in the model. Omitting EDUC=1 (identified in the program as EDUC_1) sets this level of education as reference for the economic interpretation.

This final introduction of *cross terms* (variables obtained by the multiplication of two o more initial variables) allows to model differences between males and females having different levels of education, thus we are not assuming anymore that the influence of gender is constant. Figure 10 shows the results of our final regression.

Looking at the coefficients estimate and t-statistic values we can state that each of the four included dummy variables (EDUC_2, EDUC_3, EDUC_4, EDUC_5) is individually highly significant. The coefficients of the last set of regressor (cross terms) measure to what extent the effect is different for males. Note that the R2 value is increased to 0.38.

term	coefficients	standardError	tStatistic	pValue
INTERCEPT	1.0836187	0.4509711	2.4028560	1.64e-02
EDUC_2	1.7494126	0.4767023	3.6698222	2.51e-04
EDUC_3	4.2504483	0.4628325	9.1835554	1.39e-19
EDUC_4	5.8560589	0.4268404	13.7195511	2.21e-40
EDUC_5	7.9373315	0.5302877	14.9679733	3.13e-47
MALE	1.7981066	0.4250078	4.2307619	2.47e-05
EXPER	0.3602362	0.0348115	10.3482005	2.87e-24
EXPER2	-0.0044629	0.0010006	-4.4601245	8.82e-06
EDUC_2_MALE	-0.0709015	0.5557163	-0.1275858	8.98e-01
EDUC_3_MALE	-1.0455045	0.5255907	-1.9891991	4.69e-02
EDUC_4_MALE	-0.7121700	0.5313768	-1.3402353	1.80e-01
EDUC_5_MALE	-0.0326556	0.6842381	-0.0477254	9.62e-01

ResidualsStandardError	R2	R2adj	fStatistic
3.497503	0.382416	0.3773365	82.18645

Figure 10: Regression of WAGE on MALE, EXPER, EXPER2, EDUC=2, EDUC=3, EDUC=4, EDUC=5 and cross terms results

References

[1] *OLS in Matrix Form*. URL: https://web.stanford.edu/~mrosenfe/soc_meth_proj3/matrix_OLS_NYU_notes.pdf.

[2] Marno Verbeek. *A Guide to Modern Econometrics*. URL: https://thenigerianprofessionalaccountant.files.wordpress.com/2013/04/modern-econometrics.pdf.

[3] *Wages in Belgium dataset*. URL: https://www.wiley.com/legacy/wileychi/verbeek2ed/datasets.html.