# Økonometri I

## Efterår 2021

### Obligatorisk Opgave 1

Deadline: 14. oktober, 23:59

Lavet af:

Emil Møller Rasmussen 0311872893

# Problem 1

**Introduction** To give a proper overview of the data, we will provide some descriptive statistics, as well as some insight into the variables we are working with. The dataset consists of five variables, of the following types:

| variable name | type |
|---|---|
| nr | integer |
| kommune | string |
| taxrev | float |
| taxrate | float |
| pop | double |

Table 1: Variables and data types

In addition to the technical details above, the descriptive statistics are as follows (I have ommited the variables nr and kommune, as these dont contain information that's statistically relevant):

|  | taxrev | taxrate | pop |
|---|---|---|---|
| min | 211.2 | 22.80 | 1969 |
| max | 44170.3 | 27.80 | 528,208 |
| mean | 4477.3 | 25.21 | 56476 |
| median | 3317.8 | 25.30 | 43475 |
| std. dev. | 5251 | 0.9 | 56476 |

Table 2: Descriptive statistics about dataset kommune.dta

A total of 98 observations are available - as we have one observation of each variable per municipality. The variables are fairly spread out. This is a given, as Danish municipalities can vary significantly in geographic size and population size.

When analysing the taxrate variable the values have little variance around the mean. This intuitively makes sense, as even though the setting of tax rates is within the municipalities mandate, national legislation limits it significantly.

# Problem 2

We are given the following regression model:

$$log(taxrev_m) = \delta_0 + \delta_1 taxrate_m + \epsilon_m m = 1, ..., 98. \quad (1)$$

And it is assumed that MLR.1-5 are satisfied.

1. First off, the general intepretation of the regression coefficient $\delta_1$, is the expected change in $y$ when the coefficient changes. Here $\delta_1$ indicates the change in tax revenues of a municipality when the tax rate changes. Additionally the equation is of the log-level form, which means that each percentage-point increase of the taxrate will increase the tax revenue by a constant percentage - each 1 unit increase in taxrate multiplies the expected value of $taxrev_m$ by $e^{\delta_1}$

2. Intuitively the the expected sign of $\delta_1$ is positive. The causal interpretation is that when the taxrate increases we expect that the revenue of municipalities will rise, and they will fall when the taxrate decreases.

3. After the model has run, we are provided the following values:

| | Estimate | Std. error |
|---|---|---|
| $\delta_1$ | -0.14 | 0.85 |
| $\delta_0$ | 11.7 | 2.14 |

Table 3: Results of SLR of $log(taxrev_m)$ on $taxrate_m$

These values when inserted in (1) the following equation: $log(taxrev_m) = 11.7 - 0.14 taxrate_m + \epsilon_m$. This result is surprising, as it means that the taxrate has a negative influence on municipalities tax revenue. Looking at Wooldridge [2.19], we can infer that there must be a negative correlation between $y$ and $x$. If we had done some more math on the variables involved, we would have been able to predict this result.

The standard error is relatively high for $\delta_1$ and $\delta_0$, indicating that our estimates are fairly spread out from the mean.

We are now provided with the following equation, with the local population size as an additional explanatory variable:

$$log(taxrev_m) = \beta_0 + \beta_1 taxrate_m + \beta_2 log(pop_m) + u_m \quad m = 1, ..., 98. \quad (2)$$

4. The intepretation of $\beta_2$ in (2) is the **ceteris paribus** effect of population size on municipal tax revenue. The model is now also of a log-log

functional form, which means that $\beta_2$ also demonstrates a percentage change in tax revenue when there is a percentage change in population (or **elasticity**) .

Even though it can be tempting to directly compare $\delta_1$ and $\beta_1$, it must be remembered that we are dealing with the comparison of the ”same” variable between an SLR and MLR model. There are two cases where the variables will be the same, and the relationship between them can be described as $\delta_1 = \beta_1 + \beta_2\gamma$ where $\gamma$ is the slope coefficient from the simple regression of $log(pop_m)$ on $taxrate_m$ - described in Wooldridge [3.23].

5. After the model has run, we are provided the results found in Table 4.

| | Estimate | Std. error |
|---|---|---|
| $\beta_2$ | 0.97 | 0.014 |
| $\beta_1$ | 0.022 | 0.012 |
| $\beta_0$ | -2.80 | 0.3755 |

Table 4: Results of MLR of $log(taxrev_m)$ on $taxrate_m$ and $log(pop_m)$

The variable formerly known as $\hat{\delta}_1$, now $\hat{\beta}_1$ has changed sign, and the magnitude of the variable has decreased significantly. also the std. error on all the parameters have reduced significantly.

Additionally the model now more clearly interprets the intution of how taxrate influences the revenues of a municipality, i.e. an increase of the taxrate will **ceteris paribus** increase the tax revenue of a municipality.

The standard error is now significantly lower than in the previous model, giving us a more accurate estimate of the population.

6. - **Omitted variable bias** is when the regressor is correlated with an omitted variable. I.e there is something that implicitly is included in the error term $u$, that should be taken out and included as it's own independent variable. As it were, (1) most likely had an omitted variable in the form of an unspecified *pop* regressor. (And probably still has more).

     If we want to get an unbiased estimate of the $\beta_1$-estimator an assumption is that $E(u|x) = 0$, else $E(\beta_1) \neq \beta_1$.

   - According to Wooldridge [3.46] the bias of the independent variable of an underspecified model $\tilde{y}$ can be written as $Bias(\tilde{\beta}_1) = \beta_2\tilde{\delta}_1$, which in our case would be $Bias(\delta_1) = \beta_2\gamma$, where $\gamma$ is the slope coefficient from the regression of $log(pop_m)$ on $taxrate_m$. As $\gamma$ can be defined as the covariance between $log(pop_m)$ and $taxrate_m$ divded by the variance of $taxrate_m$ (Wooldridge [2.19]), or $\frac{covar(log(pop)_m, taxrate_m)}{log(pop_m)^2}$. we can evaluate that:

(a) The sign and magnitude of the bias is determined by a combination of the signs and sizes of $\beta_2$ and covariance.

(b) there is no bias when $\beta_2 || covariance = 0$ (bias might exist for $\delta_1 = 0$ as mentioned in Wooldridge p.85.

# Problem 3

The true statistical model is defined as:

$$y = \beta_1 x_1 + \beta_2 x_2 + u \tag{3}$$

1. The estimated statistical model is as follows:

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u} \tag{4}$$

Writing (4) using matrix notation we get the following:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{1,2} & x_{2,2} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{5}$$

Which can be written as:

$$Y = X\beta + \varepsilon \tag{6}$$

which gives us the following model in estimated form:

$$Y = X\hat{\beta} + \varepsilon \tag{7}$$

We now have to show that $\hat{\beta}_1$ can be written as (4) on page 4 of obligatory assignment 1. We are provided the follwing hint:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1}Xy \tag{8}$$

*Note: In the following proofs, for easy of reading I will substitute*

$$\sum_{i=1}^{n} = \Sigma \tag{9}$$

We have to prove that:

$$\hat{\beta}_1 = \frac{\Sigma x_{i1} y_i - \frac{\Sigma x_{i1} x_{i2}}{x_{i2}^2} x_{i2} y_i}{\Sigma x_{i1}^2 - \frac{(\Sigma x_{i1} x_{i2})^2}{\Sigma x_{i2}^2}} \tag{10}$$

Additionally it is assumed that $\overline{y}, \overline{x}_1, \overline{x}_1$ are equal to 0. We plug hint (8) hint into (7):

$$\hat{\beta} = (X'X)^{-1} X'y + \varepsilon \tag{11}$$

We now start multiplying first $(X'X)^{-1}$ and $X'y$. Multiplying vectors of dimensions $(n \times 2)$ and $(2 \times 1)$ gives an output matrix of size $(n \times 2)$ matrix

$$X'X = \begin{pmatrix} x_{1,1} & x_{2,1} & \dots & x_{n1} \\ x_{1,2} & x_{2,2} & \dots & x_{n2} \end{pmatrix} \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{1,2} & x_{2,2} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} \Sigma x_{i1} x_{i1} & \Sigma x_{i1} x_{i2} \\ \Sigma x_{i2} x_{i1} & \Sigma x_{i2} x_{i2} \end{pmatrix} \tag{12}$$

We then invert this matrix:

$$(X'X)^{-1} = \frac{1}{\Sigma x_{i1}^2 \Sigma x_{i2}^2 - (\Sigma x_{i1} x_{i2})^2} \begin{pmatrix} \Sigma x_{i2}^2 & -\Sigma x_{i1} x_{i2} \\ -\Sigma x_{i2} x_{i1} & \Sigma x_{i1}^2 \end{pmatrix} \tag{13}$$

For easier calculation (and overview), we temporarily call the fraction $\frac{1}{\Sigma x_{i1}^2 \Sigma x_{i2}^2 - (\Sigma x_{i1} x_{i2})^2} = \Gamma$

$$(X'X)^{-1} = \Gamma \begin{pmatrix} \Sigma x_{i2}^2 & -\Sigma x_{i1} x_{i2} \\ -\Sigma x_{i2} x_{i1} & \Sigma x_{i1}^2 \end{pmatrix} \tag{14}$$

We then do the calculation $X'Y$:

$$(X'Y) = \begin{pmatrix} x_{1,1} & x_{2,1} & ... & x_{n1} \\ x_{1,2} & x_{2,2} & ... & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \Sigma x_{i1} y_i \\ \Sigma x_{i2} y_i \end{pmatrix} \tag{15}$$

Then we multiply $(X'X)^{-1}$ with $X'Y$:

$$(X'X)^{-1} X'Y = \Gamma \begin{pmatrix} \Sigma x_{i2}^2 & -\Sigma x_{i1} x_{i2} \\ -\Sigma x_{i2} x_{i1} & \Sigma x_{i1}^2 \end{pmatrix} \begin{pmatrix} \Sigma x_{i1} y_i \\ \Sigma x_{i2} y_i \end{pmatrix} = \Gamma \begin{pmatrix} \Sigma x_{i2}^2 \Sigma x_{i1} y_i - \Sigma x_{i1} x_{i2} \Sigma x_{i2} y_i \\ -\Sigma x_{i2} x_{i1} \Sigma x_{i1} y_i + \Sigma x_{i1}^2 \Sigma x_{i2} y_i \end{pmatrix} \tag{16}$$

We now replace $\Gamma$ with the scalar it represents:

$$\frac{1}{\Sigma x_{i1}^2 \Sigma x_{i2}^2 - (\Sigma x_{i1} x_{i2})^2} \begin{pmatrix} \Sigma x_{i2}^2 \Sigma x_{i1} y_i - \Sigma x_{i1} x_{i2} \Sigma x_{i2} y_i \\ -\Sigma x_{i2} x_{i1} \Sigma x_{i1} y_i + \Sigma x_{i1}^2 \Sigma x_{i2} y_i \end{pmatrix} \tag{17}$$

And we multiply it out:

$$\hat{\beta} = \begin{pmatrix} \frac{\Sigma x_{i2}^2 \Sigma x_{i1} y_i - \Sigma x_{i1} x_{i2} \Sigma x_{i2} y_i}{\Sigma x_{i1}^2 \Sigma x_{i2}^2 - (\Sigma x_{i1} x_{i2})^2} \\ \frac{-\Sigma x_{i2} x_{i1} \Sigma x_{i1} y_i + \Sigma x_{i1}^2 \Sigma x_{i2} y_i}{\Sigma x_{i1}^2 \Sigma x_{i2}^2 - (\Sigma x_{i1} x_{i2})^2} \end{pmatrix} \tag{18}$$

As $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$, $\hat{\beta}_1 = \frac{\Sigma x_{i2}^2 \Sigma x_{i1} y_i - \Sigma x_{i1} x_{i2} \Sigma x_{i2} y_i}{\Sigma x_{i1}^2 \Sigma x_{i2}^2 - (\Sigma x_{i1} x_{i2})^2}$

And if we simplify, by dividing by $\Sigma x_{i2}^2$

$$\hat{\beta}_1 = \frac{\Sigma x_{i1} y_i - \frac{\Sigma x_{i1} x_{i2} \Sigma x_{i2} y_i}{x_{i2}^2}}{\Sigma x_{i1}^2 - \frac{(\Sigma x_{i1} x_{i2})^2}{\Sigma x_{i2}^2}} \tag{19}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_{i1} y_i - \frac{\sum_{i=1}^n x_{i1} x_{i2}}{x_{i2}^2} \sum_{i=1}^n x_{i2} y_i}{\sum_{i=1}^n x_{i1}^2 - \frac{(\sum_{i=1}^n x_{i1} x_{i2})^2}{\sum_{i=1}^n x_{i2}^2}} \tag{20}$$

And it is thus shown.

2. Regressing $x_1$ on $x_2$:

$$x_1 = \gamma x_2 + \zeta, \quad where \quad \gamma = \frac{\Sigma x_{i2} x_{i1}}{\Sigma x_{i2}^2} \tag{21}$$

The residuals of this regression are written as:

$$residuals = x_{i1} - \hat{x}_{i1} \tag{22}$$

As $\hat{x}_{i1} = \Sigma\gamma x_{i2}$ we can rewrite $\hat{r}$ as:

$$\hat{r} = \Sigma x_{i1} - \Sigma\gamma x_{i2} \tag{23}$$

Now we will regress $y$ on $\hat{r}$:

$$y = \lambda\hat{r} + \epsilon \tag{24}$$

In this situation it will give us the lambda estimator as:

$$\hat{\lambda} = \frac{\Sigma\hat{r}_i y_i}{\Sigma\hat{r}_i^2} \tag{25}$$

as $\hat{r} = \Sigma x_{i1} - \Sigma\gamma x_{i1}$ and $\gamma = \frac{\Sigma x_{i1}x_{i2}}{\Sigma x_{i1}^2}$ we can rewrite $\hat{r}$ as $\hat{r} = \Sigma x_{i1} - \frac{\Sigma x_{i2}x_{i1}}{\Sigma x_{i2}^2} x_{i2}$ insert this into the formula and we get:

$$\hat{\lambda} = \frac{(\Sigma x_{i1} - \frac{\Sigma x_{i2}x_{i1}}{\Sigma x_{i2}^2}\Sigma x_{i2})\Sigma y_i}{(\Sigma x_{i1} - \frac{\Sigma x_{i2}x_{i1}}{\Sigma x_{i2}^2}\Sigma x_{i2})^2} \tag{26}$$

We simplify this further:

$$\hat{\lambda} = \frac{\Sigma x_{i1}y_i - \frac{\Sigma x_{i2}x_{i1}}{\Sigma x_{i2}^2}\Sigma x_{i2}y_i}{\Sigma x_{i1}^2 - \frac{(\Sigma x_{i2}x_{i1})^2}{\Sigma x_{i2}^2}} \tag{27}$$

And we have now shown that $\hat{\lambda} = \hat{\beta}_1$

3. The parameter estimates from the two processes are exactly similar. This must be true as the residuals of $taxrate_m$ on $log(pop_m)$ are the part of $taxrate_m$ that is uncorellated with $log(pop_{m_i})$, and thus shows the relationship between $log(taxrev_m)$ and $taxrate_m$ after $log(pop_m)$ has been partialled out. In other words, the part of $log(pop_m)$ that correlates with $taxrate_m$ has been "taken" out and placed into the error term of the equation $log(\tilde{taxrate}_m) = \tilde{\delta}_1 + \varepsilon$. This also means that there is **omitted variable bias** present in the tilde formula - as the technical definition of omitted variable bias is that it is present if $x_1$ is correlated with $x_2$.

# Problem 4: Conclusion

In problem 1 we gave some descriptive statistics of the variables involved with the KOMMUNE.dta dataset.

We went on to look at the simple regression model of the relationship between municipal tax revenue ($log(taxrev_m)$) and the municipal taxrate ($taxrate_m$). In the simple regression model it was, contrary to expectations, shown that an increase in municipal taxrate, would reduce the municipal revenue. There was some uncertainty in the accuracy of these estimates, by analysing the std. errors.

By including an additional explanatory variable, the population of the municipality $log(pop_m)$, a more intuitive model was proposed - with seemingly stronger explanatory power. Here tax revenue was positively influenced by both the taxrate, but to a larger extent the relative change of population. This gave rise to the notion that our original model for modelling the tax revenue was suffering from omitted variable bias - i.e. the variable $log(pop_m)$ was included in the error term.

In problem 3 we demonstrated different ways of defining the slope parameter for the $\beta_1$ parameter in the provided multiple linear regression model - one using the **Frisch Waugh Theorem**, and the other by first regressing the $\beta_2$ on $\beta_1$ and then regressing $y$ on the residuals of this. This demonstration effectively partialed out the correlating part of the variables $log(pop_m)$ and $taxrate_m$ out of $\beta_2$. We proved that these methods provided the same result - and indirectly also showed the link between the SLR and MLR models.

Throughout this assignment we have shown how an underspecified model lacks explanatory power - and even can end up providing the ”wrong” or misleading conclusions based on our available data and variables.

It has been shown that tax rate alone has poor explanatory power when it comes to the tax revenue of municipalities. More independent variables were needed to create a proper model describing municipal tax revenues. Specifically population provides a strong candidate in combination with tax rate for describing tax rate. This also makes intuitive sense, as tax rate combined with population size seems a strong decider of tax revenues.

**Further work** could be done to extract further explanatory variables - such as a variable describing income brackets, as a higher percentage of high-income earners intuitively will increase the tax revenue of a municipality. This hypothesis could be explored both by adding more independent variables, but also by delving further into the specific municipalities and their population distributions.