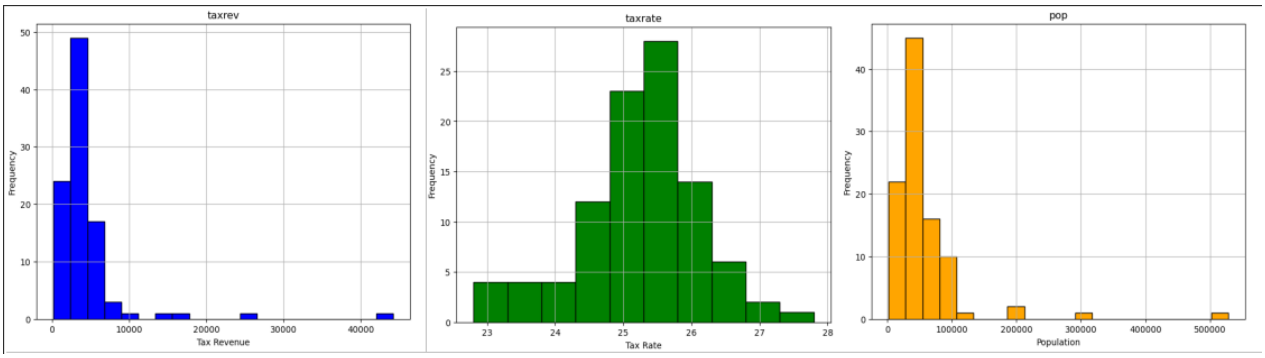
Mandatory Assignment 1

Max Marius Toft (vsx222) & Sofus Holm (gsk440)

September 18, 2024

Problem 1. Descriptive analysis



- Tax Revenue (taxrev): The distribution of tax revenues among the municipalities is highly skewed, with most municipalities generating revenues below 10,000, and a few outliers generating significantly higher tax revenues, especially Københavns Kommune.
- Tax Rate (taxrate): The tax rate has a relatively narrow range, centered around 25%, with few municipalities below or above this central value. Most municipalities fall between 24.8% and 25.7%.
- Population (pop): The population distribution is heavily skewed to the right, with the majority of municipalities having populations below 100,000, while Københavns Kommune has a significantly larger population, acting as an outlier.

	nr	taxrev	taxrate	pop
count	98.000000	98.000000	98.000000	98.000000
mean	462.173469	4477.341309	25.208162	56475.887755
std	236.555437	5251.175293	0.908003	62925.301713
min	101.000000	211.228409	22.799999	1969.000000
255075max	860.000000	44170.335938	27.799999	528208.000000

Problem 2. Empirical analysis of tax revenues and municipal tax rates

1. What is the interpretation of δ_1 in regression model 1?

In model 1, δ_1 is first of all the slope of the linear model, but it also represents the elasticity of total tax revenues with respect to the municipal tax rate. Since we are using a log-linear model, δ_1 can be interpreted as the percentage change in tax revenue when there is a 1% change in the tax rate in the municipality.

2. What is the expected sign of $\delta 1$?

In general, we would expect a positive sign for δ_1, as municipalities increase their income tax rate, they are likely to collect more tax revenue, at least in the short term. This is because the tax rate is directly applied to the income base of the residents, and higher rates would typically yield higher revenues.
However, if the tax-rate keep rising, we do not expect the same relation. For instance, higher tax rates could make people move from the municipality, or work less, which result in lowered tax revenues. But for purpose of the analysis, we expect δ_1 to be positive.

3. Estimate regression model (1) using OLS

Dep. Variable:	log_taxrev	**R-squared:**	0.029
Model:	OLS	**Adj. R-squared:**	0.018
Method:	Least Squares	**F-statistic:**	2.818
Date:	Mon, 16 Sep 2024	**Prob (F-statistic):**	0.0965
Time:	19:44:56	**Log-Likelihood:**	-111.12
No. Observations:	98	**AIC:**	226.2
Df Residuals:	96	**BIC:**	231.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	11.6982	2.143	5.459	0.000	7.444	15.952
taxrate	-0.1426	0.085	-1.679	0.096	-0.311	0.026

Omnibus:	17.794	**Durbin-Watson:**	1.863
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	45.446
Skew:	-0.567	**Prob(JB):**	1.35e-10
Kurtosis:	6.138	**Cond. No.**	705.

Here we get $\hat{\delta_1} = -0.1426$, which is not what we expected, as it shows a negative correlation between tax rates and tax revenue. This could likely be due to the fact that the population, a major factor in looking at tax revenue, is not included. As we can see above, the 95% confidence interval includes zero, meaning that the coefficient is not statistically significant, and with an $R^2$of 0.03, we can only explain 3% of the variance from the model (which is not very much). Hereby it suggest we cannot use this use this regression to explain tax revenue.

4. What is the interpretation of β_1 in regression model (2)?

β_1 represents the same as δ_1 as before, as the elasticity of tax revenue with respect to the tax rate. However, in this regression, population is now included in the model. β_1 measures the percentage change in tax revenue for a one-unit increase in the tax rate, while holding population constant.
If $\beta_1 > 0$, this portrays a positive correlation between tax rate and tax revenue. On the other hand, if $\beta_1 < 0$, it shows a negative correlation, meaning higher tax rates reduce revenue. In general we expect that the inclusion of population will improve the estimate since population size significantly affects tax revenue.

5. Estimate regression model (2) by OLS

Dep. Variable:	taxrev	**R-squared:**	0.980
Model:	OLS	**Adj. R-squared:**	0.980
Method:	Least Squares	**F-statistic:**	2344.
Date:	Wed, 18 Sep 2024	**Prob (F-statistic):**	1.42e-81
Time:	11:52:45	**Log-Likelihood:**	79.497
No. Observations:	98	**AIC:**	-153.0
Df Residuals:	95	**BIC:**	-145.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
taxrate	0.0226	0.012	1.816	0.072	-0.002	0.047
const	-2.8022	0.376	-7.461	0.000	-3.548	-2.057
logpop	0.9711	0.014	67.471	0.000	0.943	1.000

Omnibus:	15.730	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	20.068
Skew:	0.810	**Prob(JB):**	4.39e-05
Kurtosis:	4.514	**Cond. No.**	933.

- Tax rate:
  - From the table (Husk at indsætte tabellen), we observe that the coefficient for the tax rate in this regression is 0.0226. A one-unit increase in the tax rate is thus associated with a 2.26% increase in tax revenue. This result aligns more closely with our expected \beta, as predicted in problem 2.2. Additionally, the p-value is around 0.72, indicating that our coefficient is near the 10% significance level. This suggests a positive but weak relationship, as it does not reach the 5% significance level.
- Log population (logpop):
  - We also observe that the coefficient for log(population) is 0.9711, meaning that a 1% increase in population correlates with an increase in tax revenue of approximately 0.97%.

6. Explain what omitted variable bias is and describe how the covariance between taxratem and log(popm) influences the sign and size of the bias in the OLS estimate $\delta\beta_1$.

Omitted Variable Bias (OVB) occurs when a regression model fails to include all relevant variables that influence either the dependent or independent variables, leading to biased and inconsistent coefficient estimates. The bias is calculated as:

$$Bias(\hat{\beta_1}) = \beta_2 \cdot Cov(x_1, x_2)$$

If $Cov(x_1, x_2)$ equals zero, there will be no bias in the estimation.

If $Cov(taxrate_m, log(pop_m)) > 0$, meaning that the tax rate and $log(population)$ are positively correlated, the bias will reflect the potential effect of the omitted variable. Therefore, if $log(pop)$ impacts the dependent variable, omitting it will increase the bias, as seen with δ_1.

Conversely, if $Cov(taxrate_m, log(pop_m)) < 0$, the bias will move in the opposite direction.

The magnitude of the bias depends on the strength of the correlation between tax rate and $log(population)$. A strong correlation results in a larger bias, while a weak correlation leads to a smaller bias.

Problem 3: Multiple linear regression as two simple linear regressions

1. Write up model model (3) using matrix notation. Show the expression of the OLS estimator, $\hat{\beta}$.

We will estimate the regression model using OLS.
We start by expressing the model in matrix form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Or

$$y = X\beta + u$$

Here:
- y is an $n \times 1$ vector of the dependent variable,
- X is an $n \times 2$ matrix of independent variables,
- β is a 2×1 vector of coefficients,
- u is an $n \times 1$ vector of error terms.

The OLS estimator for the coefficients $\hat{\beta}$ is given by:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1}X'y$$

Now we can calculate the estimates of β_1. First I will start by calculationg $X'X$, the product of the transpose of X, and X itself:

$$X'X = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix}$$

Now we will find the inverse of $X'X$:
To invert a 2×2 matrix with the form $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ we use:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Here we can do that to $X'X$:

$$(X'X)^{-1} = \frac{1}{\sum_{i=1}^n x_{i1}^2 \cdot \sum_{i=1}^n x_{i2}^2 - \left(\sum_{i=1}^n x_{i1}x_{i2}\right)^2} \begin{pmatrix} \sum_{i=1}^n x_{i2}^2 & -\sum_{i=1}^n x_{i1}x_{i2} \\ -\sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i1}^2 \end{pmatrix}$$

Next up, we can compute $X'y$, the product of X' and y:

$$X'y = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \end{pmatrix}$$

To estimate $\hat{\beta}_1$, we know from the hint, that we only need the first row of the matrix product $(X'X)^{-1}$:

$$\frac{1}{\sum_{i=1}^n x_{i1}^2 \cdot \sum_{i=1}^n x_{i2}^2 - \left(\sum_{i=1}^n x_{i1}x_{i2}\right)^2} \begin{pmatrix} \sum_{i=1}^n x_{i2}^2 & -\sum_{i=1}^n x_{i1}x_{i2} \end{pmatrix}$$

Now we can multiply this by $X'y$, and hereby get the estimator for $\hat{\beta}_1$:

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

This can be simplified to:

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

Which equals:

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

Hereby we have shown that the OLS estimator, $\hat{\beta}_1$ can be written as $\frac{\sum_{i=1}^n x_{i1}y_i - \left(\sum_{i=1}^n x_{i1}x_{i2}\right)\frac{\sum_{i=1}^n x_{i2}y_i}{\sum_{i=1}^n x_{i2}^2}}{\sum_{i=1}^n x_{i1}^2 - \left(\frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i2}^2}\right)^2}$.

2. Now consider two simple regressions:

a. A simple regression of x_1 on x_2 (without an intercept). Define the residuals from this regression as $\hat{r_1}$

b. A simple regression of y on $\hat{r}_1$ (without an intercept).

Show that the estimated slope parameter from step b) is identical to β1 as defined in the previous equation.

[Hint: Use equation (2.66) in Wooldridge to obtain an expression for the slope parameter in step a). Calculate the residuals based on this expression and use these in step b)]

equation 2.66=$\beta \frac{\sum_{i=1}^n \tilde{x}_i y_i}{\sum_{i=1}^n x_i^2}$

To show that the estimated slope parameter from step b) is identical to $\hat{\beta}_1$ as defined in the previous equation, we will use the provided hint and apply equation (2.66) from Wooldridge.

Step a) Regress x_1 on x_2 without an intercept:

Part (a):

The OLS estimate of $\hat{\beta}_1$ is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_{2i}x_{1i}}{\sum_{i=1}^n x_{2i}^2}$$

Also we can express x_1 from x_2 as:

$$x_1 = \beta_1 x_2 + u$$

Here the residual u for the model is:

$$u = x_1 - \hat{\beta}_1 x_2$$

Which also can be written as:

$$u = \hat{r_1} = x_1 - \frac{\sum_{i=1}^n x_{2i}x_{1i}}{\sum_{i=1}^n x_{2i}^2} x_{2i}$$

b)

$y = \beta_1 \cdot \hat{r_1} + u_2$
$y = \frac{\sum_{i=1}^{n} r_1 y_i}{\sum_{i=1}^{n} r_1^2} \cdot \hat{r_1} + u_2$
$\hat{\beta_1} = \frac{\sum_{i=1}^{n} r_{i1} y_i}{\sum_{i=1}^{n} r_{i1}^2} = \frac{\sum_{i=1}^{n} (x_{i1} - \frac{\sum_{i=1}^{n} x_{i2} x_{i1}}{\sum_{i=1}^{n} x_{i2}^2} \cdot x_{i2}) y_i}{(\sum_{i=1}^{n} x_{i1} - \frac{\sum_{i=1}^{n} x_{i2} x_{i1}}{\sum_{i=1}^{n} x_{i2}^2} \cdot x_{i2})^2}$

Part (b):
For the second regression model:

$$y = \beta_1 \cdot \hat{r_1} + u_2$$

The OLS estimate of $\hat{\beta_1}$ is:

$$\hat{\beta_1} = \frac{\sum_{j=1}^{n} \hat{r_1} y_j}{\sum_{j=1}^{n} \hat{r_1}^2}$$

Here we can substitute $\hat{r_1}$ from part (a), we get:

$$\hat{\beta_1} = \frac{\sum_{j=1}^{n} \left(\frac{\sum_{i=1}^{n} x_{2i} x_{1i}}{\sum_{i=1}^{n} x_{2i}^2} \cdot x_{i2} \right) y_j}{\sum_{j=1}^{n} \hat{r_1}^2}$$

Now I will simplify the denominator:

$$\hat{r_1}^2 = \left(x_{i1} - \frac{\sum_{i=1}^{n} x_{i2} x_{i1}}{\sum_{i=1}^{n} x_{i2}^2} x_{i2} \right)^2 = x_{i1}^2 + \left(\frac{\sum_{i=1}^{n} x_{i2} x_{i1}}{\sum_{i=1}^{n} x_{i2}^2} \cdot x_{i2} \right)^2 - 2 \cdot x_{i1} \cdot \frac{\sum_{i=1}^{n} x_{i2} x_{i1}}{\sum_{i=1}^{n} x_{i2}^2} \cdot x_{i2}$$

This simplifies to:

$$= x_{i1}^2 + (\frac{\sum_{i=1}^{n} x_{i1}}{\sum_{i=1}^{n} x_{i2}})^2 - \sum_{i=1}^{n} 2x_{i1}^2 = \sum_{i=1}^{n} x_{i1}^2 + \frac{\sum_{i=1}^{n} x_{i1}^2}{\sum_{i=1}^{n} x_{i2}^2} - \sum_{i=1}^{n} 2x_{i1}^2$$

Further simplifying:

$$= -\sum_{i=1}^{n} x_{i1}^2 + \frac{\sum_{i=1}^{n} x_{i1}^2}{\sum_{i=1}^{n} x_{i2}^2} = \left(-1 + \frac{1}{\sum_{i=1}^{n} x_2^2} \right) \sum_{i=1}^{n} x_{i1}^2 = \left(\frac{1 - \sum_{i=1}^{n} x_{i2}^2}{\sum_{i=1}^{n} x_{i2}^2} \right) \sum_{i=1}^{n} x_{i1}^2$$

$$= \frac{(1 - x_2^2) x_1^2}{x_2^2} = \frac{x_1^2 - x_1^2 x_2^2}{x_2^2}$$

Hereby we have found that the complete expression is now simplified to
Final Result:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n} \left(\frac{\sum_{i=1}^{n} x_{2i} x_{1i}}{\sum_{i=1}^{n} x_{2i}^2} \cdot x_{i2} \right) y_i}{(\sum_{i=1}^{n} x_{i1}^2 - \left(\frac{\sum_{i=1}^{n} x_{i2} x_{i1}}{x_{i2}^2} \cdot x_{i2} \right)^2)}$$

3. Verify your theoretical result from Problem 3.2

Dep. Variable:	taxrate	**R-squared:**	0.039
Model:	OLS	**Adj. R-squared:**	0.029
Method:	Least Squares	**F-statistic:**	3.862
Date:	Wed, 18 Sep 2024	**Prob (F-statistic):**	0.0523
Time:	11:54:00	**Log-Likelihood:**	-127.16
No. Observations:	98	**AIC:**	258.3
Df Residuals:	96	**BIC:**	263.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	27.6268	1.234	22.386	0.000	25.177	30.077
log_pop	-0.2273	0.116	-1.965	0.052	-0.457	0.002

Omnibus:	3.367	**Durbin-Watson:**	1.570
Prob(Omnibus):	0.186	**Jarque-Bera (JB):**	2.699
Skew:	-0.328	**Prob(JB):**	0.259
Kurtosis:	3.480	**Cond. No.**	147.

From the table (indsæt tabellen) we get the following:

R-squared: 0.039, indicating that the log of population explains about 3.9% of the variation in the tax rate.

Coefficient for log(pop): -0.2273, with a p-value of 0.052, which is close to the 0.05 significance level. This suggests a weak negative relationship between the population size (in logarithmic terms) and the tax rate.

Intercept (constant): 27.63, meaning that when log(pop) is zero, the expected tax rate is around 27.63.

Next, run a simple regression of $log(taxrevm)$ on $res1_m$.

Dep. Variable:	log_taxrev	**R-squared (uncentered):**	0.000
Model:	OLS	**Adj. R-squared (uncentered):**	-0.010
Method:	Least Squares	**F-statistic:**	0.0005828
Date:	Wed, 18 Sep 2024	**Prob (F-statistic):**	0.981
Time:	11:54:45	**Log-Likelihood:**	-344.53
No. Observations:	98	**AIC:**	691.1
Df Residuals:	97	**BIC:**	693.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
res1	0.0226	0.933	0.024	0.981	-1.829	1.874

Omnibus:	15.581	**Durbin-Watson:**	0.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	46.400
Skew:	-0.383	**Prob(JB):**	8.40e-11
Kurtosis:	6.283	**Cond. No.**	1.00

R-squared (uncentered): 0.000, indicating that the residuals (res1) explain virtually none of the variation in log(taxrev).

Coefficient for res1 : 0.0226, with a very high p-value of 0.981. This suggests that the relationship between the residuals from the first regression and log(taxrev) is statistically insignificant.

F-statistic: 0.0005879, reinforcing that the model does not significantly explain the variation in the dependent variable.

Discuss why it is possible to obtain an estimate identical to β_1 in the second step when $log(pop_m)$ is an omitted variable.

In the two-step procedure, we first regress taxrate on log(popm) , and the residuals res1 represent the variation in taxrate that is uncorrelated with log(popm) . Essentially, $\textres1$ is the "purified" variation in taxrate , free from any influence of log(popm) .

When we regress log(taxrev) on res1 , the regression no longer suffers from omitted variable bias because res1 is orthogonal (uncorrelated) to log(popm) .

The residuals res1 contain no information about log(popm) , meaning any variation in log(taxrev) explained by res1 is independent of the variation in log(popm) .

Thus, the second step does not suffer from omitted variable bias, and this explains why the estimate from this regression can be identical to β_1 . The first step "separates out" the influence of log(popm) on taxrate , and the second step works with the residual variation in taxrate , which is uncorrelated with the omitted variable.

When you omit log(popm) in the second step, you're effectively isolating the variation in taxrate that is independent of log(popm) (which is captured by res1). This is similar to what happens in a multiple regression where you would control for log(popm)

Problem 4: Conclusion