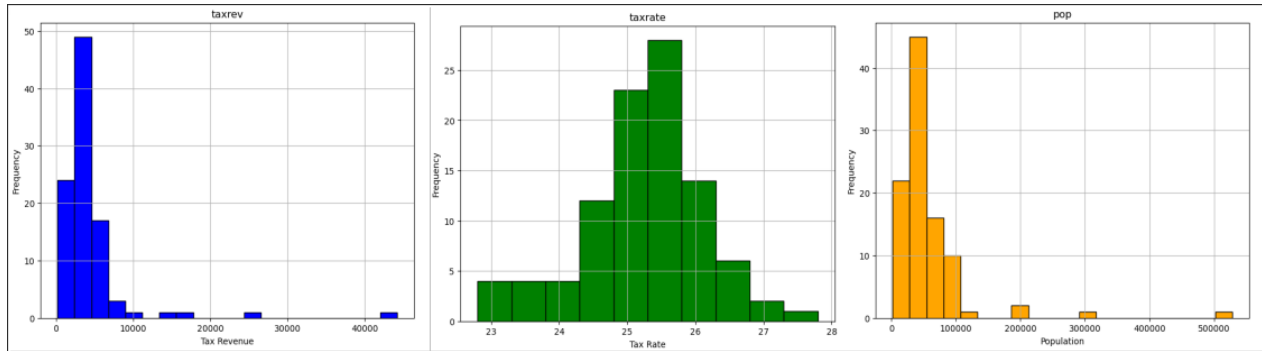


# Mandatory Assignment 1

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.

	count	mean	std	min	25%	50%	75%	max
nr	98.00	462.17	236.56	101.00	242.50	435.00	697.25	860.00
taxrev	98.00	4477.34	5251.18	211.23	2466.70	3317.85	4786.06	44170.34
taxrate	98.00	25.21	0.91	22.80	24.80	25.30	25.70	27.80
pop	98.00	56475.89	62925.30	1969.00	29997.75	43475.00	59733.00	528208.00
const	98.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
$\log_{\text{taxrev}}$	98.00	8.10	0.77	5.35	7.81	8.11	8.47	10.70
$\log_{\text{pop}}$	98.00	10.64	0.79	7.59	10.31	10.68	11.00	13.18
res1	98.00	0.00	0.89	-2.29	-0.37	0.11	0.49	2.33

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

### 1. What is the interpretation of $\delta_1$ in regression model 1?

In model 1,  $\delta_1$  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,  $\delta_1$  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  $\delta_1$ , 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  $\delta_1$  to be positive.

## 3. Estimate regression model (1) using OLS

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

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

<b>Omnibus:</b>	17.794	<b>Durbin-Watson:</b>	1.863
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	45.446
<b>Skew:</b>	-0.567	<b>Prob(JB):</b>	1.35e-10
<b>Kurtosis:</b>	6.138	<b>Cond. No.</b>	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 regression to explain tax revenue.

## 4. What is the interpretation of $\beta_1$ in regression model (2)?

$\beta_1$  represents the same as  $\delta_1$  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.  $\beta_1$  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

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

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

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

- Tax rate:

- From the table, 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 taxrate 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  $\delta_1$ .

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 (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,
- $\beta$  is a  $2 \times 1$  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  $\beta_1$ . First I will start by calculating  $X'X$ , 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  $2 \times 2$  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  $X'X$ :

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

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  $(X'X)^{-1}$ :

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

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

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

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:

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

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

## 2. Now consider two simple regressions:

Equation 2.66:  $\beta = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$

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:

$$u = \hat{r}_1 = x_1 - \frac{\sum_{i=1}^n x_{2i}x_{1i}}{\sum_{i=1}^n x_{2i}^2} x_2$$

**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), and then start to simplify the numerator, we get:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{j=1}^n \left( x_{j1} - \frac{\sum_{i=1}^n x_{2i} x_{1i}}{\sum_{i=1}^n x_{i2}^2} \cdot x_{j2} \right) y_j}{\sum_{j=1}^n \left( x_{j1} - \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} x_{j2} \right)^2} = \frac{\left( \sum_{j=1}^n x_{j1} - \frac{\sum_{i=1}^n x_{2i} x_{1i}}{\sum_{i=1}^n x_{i2}^2} \cdot \sum_{j=1}^n x_{j2} \right) \sum_{j=1}^n y_j}{\sum_{j=1}^n \left( x_{j1} - \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} x_{j2} \right)^2} = \\ &= \frac{\sum_{j=1}^n x_{j1} y_j - \frac{\sum_{i=1}^n x_{2i} x_{1i}}{\sum_{i=1}^n x_{i2}^2} \cdot \sum_{j=1}^n x_{j2} y_j}{\sum_{j=1}^n \left( x_{j1} - \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} x_{j2} \right)^2} = \frac{\sum_{j=1}^n x_{j1} y_j - \frac{\sum_{i=1}^n x_{2i} x_{1i}}{\sum_{i=1}^n x_{i2}^2} \cdot \sum_{j=1}^n x_{j2} y_j}{\sum_{j=1}^n \left( x_{j1} - \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} x_{j2} \right)^2} \end{aligned}$$

From here, we can switch the  $j$  index to  $i$ , as the values are the same sum expressions:

$$\frac{\sum_{i=1}^n x_{i1} y_i - \frac{\sum_{i=1}^n x_{2i} x_{1i}}{\sum_{i=1}^n x_{i2}^2} \cdot \sum_{i=1}^n x_{i2} y_i}{\sum_{j=1}^n \left( x_{j1} - \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} x_{j2} \right)^2} = \frac{\sum_{i=1}^n x_{i1} y_i - \left( \sum_{i=1}^n x_{2i} x_{1i} \right) \frac{\sum_{i=1}^n x_{i2} y_i}{\sum_{i=1}^n x_{i2}^2}}{\sum_{j=1}^n \left( x_{j1} - \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} x_{j2} \right)^2}$$

Now we have found the numerator, and from here, we will now simplify the denominator:

$$\sum_{j=1}^n \hat{r}_1^2 = \sum_{j=1}^n \left( x_{j1} - \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} x_{j2} \right)^2 = \sum_{j=1}^n x_{j1}^2 + \sum_{j=1}^n \left( \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} \cdot x_{j2} \right)^2 - 2 \cdot \sum_{j=1}^n x_{j1} \cdot x_{j2} \cdot \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2}$$

This simplifies to:

$$= \sum_{j=1}^n x_{j1}^2 + \sum_{j=1}^n \frac{\left( \sum_{i=1}^n x_{i2} x_{i1} \right)^2}{\left( \sum_{i=1}^n x_{i2}^2 \right)^2} \cdot x_{j2}^2 - 2 \cdot \sum_{j=1}^n x_{j1} \cdot x_{j2} \cdot \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2}$$

Now we factor out the second and third terms:

$$= \sum_{j=1}^n x_{j1}^2 + \frac{\left( \sum_{i=1}^n x_{i2} x_{i1} \right)^2}{\left( \sum_{i=1}^n x_{i2}^2 \right)^2} \sum_{j=1}^n x_{j2}^2 - 2 \cdot \frac{\sum_{i=1}^n x_{i2} x_{i1}}{\sum_{i=1}^n x_{i2}^2} \cdot \sum_{j=1}^n x_{j1} \cdot x_{j2}$$

Now we can switch the  $j$  index to  $i$ , as the values are the same sum expressions:

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

Here we can now combine the second and third terms:

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

And now we have the denominator:

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

Now we can show the final expression:

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

Hereby we have showed that the estimated slope parameter from step b) is identical to  $\hat{\beta}_1$  as defined in equation (4) and found that the two expressions are equal to each other.

### 3. Verify your theoretical result from Problem 3.2

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

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

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

From the table 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  $\log(\text{taxrev}_m)$  on  $\text{res1}_m$ .

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

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

<b>Omnibus:</b>	15.581	<b>Durbin-Watson:</b>	0.015
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	46.400
<b>Skew:</b>	-0.383	<b>Prob(JB):</b>	8.40e-11
<b>Kurtosis:</b>	6.283	<b>Cond. No.</b>	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  $\beta_1$  in the second step when  $\log(\text{pop}_m)$  is an omitted variable.**

In the two-step procedure, we first regress  $\text{taxrate}$  on  $\log(\text{pop}_m)$ , and the residuals  $\text{res1}$  represent the variation in  $\text{taxrate}$  that is uncorrelated with  $\log(\text{pop}_m)$ . Essentially,  $\text{res1}$  is the “purified” variation in  $\text{taxrate}$ , free from any influence of  $\log(\text{pop}_m)$ .

When we regress  $\log(\text{taxrev})$  on  $\text{res1}$ , the regression no longer suffers from omitted variable bias because  $\text{res1}$  is orthogonal (uncorrelated) to  $\log(\text{pop}_m)$ .

The residuals  $\text{res1}$  contain no information about  $\log(\text{pop}_m)$ , meaning any variation in  $\log(\text{taxrev})$  explained by  $\text{res1}$  is independent of the variation in  $\log(\text{pop}_m)$ .

Thus, the second step does not suffer from omitted variable bias, and this explains why the estimate from this regression can be identical to  $\beta_1$ . The first step “separates out” the influence of  $\log(\text{pop}_m)$  on  $\text{taxrate}$ , and the second step works with the residual variation in  $\text{taxrate}$ , which is uncorrelated with the omitted variable.

When you omit  $\log(\text{pop}_m)$  in the second step, you’re effectively isolating the variation in  $\text{taxrate}$  that is independent of  $\log(\text{pop}_m)$  (which is captured by  $\text{res1}$ ). This is similar to what happens in a multiple regression where you would control for  $\log(\text{pop}_m)$ .

#### **Problem 4: Conclusion**