1. Go to the website https://www.rug.nl/ggdc/productivity/pwt/ and download the PWT 10.0 dataset (either excel or state format). In the data file, you will find information on income, output, input, and productivity for 183 countries. Import the dataset into R. For this exercise, we are mainly interested in the following variables for the year **2019**:

   - *rgdpna*: Real GDP at constant prices of 2017 (in millions of USD of 2017)
   - *rnna*: Capital stock at constant prices of 2017 (in millions of USD of 2017)
   - *rtfpna*: Total factor productivity at constant prices of 2017
   - *pop*: Population (in millions)
   - *emp*: Employed population (in millions)
   - *avh*: Average hours worked per year by active people
   - *hc*: Human capital index, based on years of schooling and returns to education

   (a) Estimate a regression by OLS where the dependent variable is the real GDP (*rgdpna*) and the explanatory variable is the capital stock (*rnna*). Based on your intuition and knowledge of economics, which relevant variable(s) are we omitting in this equation?

```
> rgdpna <- data_2019$rgdpna
> rnna <- data_2019$rnna
>
> reg_gdp_capital <- lm(rgdpna ~ rnna)
> summary(reg_gdp_capital)

Call:
lm(formula = rgdpna ~ rnna)

Residuals:
     Min       1Q   Median       3Q      Max
-2437326   -11729    11494    36075  4942795

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.109e+04  3.826e+04   -0.29    0.772
rnna         2.264e-01  3.600e-03   62.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 490700 on 178 degrees of freedom
  (3 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.9569,    Adjusted R-squared:  0.9567
F-statistic:  3954 on 1 and 178 DF,  p-value: < 2.2e-16
```

The gross domestic product (GDP) of a country is not only determined by the country's accessible capital stock. Other factors affect GDP as well. Population together with employment is associated with the available labour force of a country. The unemployment rate has effect on the inflation rate. Human capital is a factor which determines development and growth of an economy. Together with technology, it provides which goods and services are being produced/offered. These factors as well as additional ones have an impact on the GDP. Therefore, the regression reg_gdp_capital suffers from omitted variable bias.

   (b) Estimate a second regression by OLS where the dependent variable is the real GDP (*rgdpna*) and the explanatory variables are the capital stock (*rnna*), the population (*pop*) and the total factor productivity (*rtfpna*). Interpret the coefficients and test the hypothesis that the coefficients of *pop* and *rnna* are jointly zero. Which are your conclusions?

```
> # linear regression model y=real GDP, x1=capital stock, x2=population, x3=productivity
> pop <- data_2019$pop
> rtfpna <- data_2019$rtfpna
>
> reg_gdp_capital_popul_product <- lm(rgdpna ~ rnna + pop + rtfpna)
> summary(reg_gdp_capital_popul_product)

Call:
lm(formula = rgdpna ~ rnna + pop + rtfpna)

Residuals:
     Min       1Q   Median       3Q      Max
-2324058   -34855    25523    75946  4624872

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.200e+05  9.762e+05  -0.635    0.527
rnna         2.338e-01  7.596e-03  30.780   <2e-16 ***
pop         -5.836e+02  5.073e+02  -1.150    0.252
rtfpna       5.943e+05  9.811e+05   0.606    0.546
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 602500 on 114 degrees of freedom
  (65 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.9569,    Adjusted R-squared:  0.9557
F-statistic: 843.2 on 3 and 114 DF,  p-value: < 2.2e-16
```

A 1 million USD increase in capital stock is associated with a 0.2338 million USD increase in real GDP, all other variables held constant. The effect is small but statistically significant at the 1% level as well as economically significant.

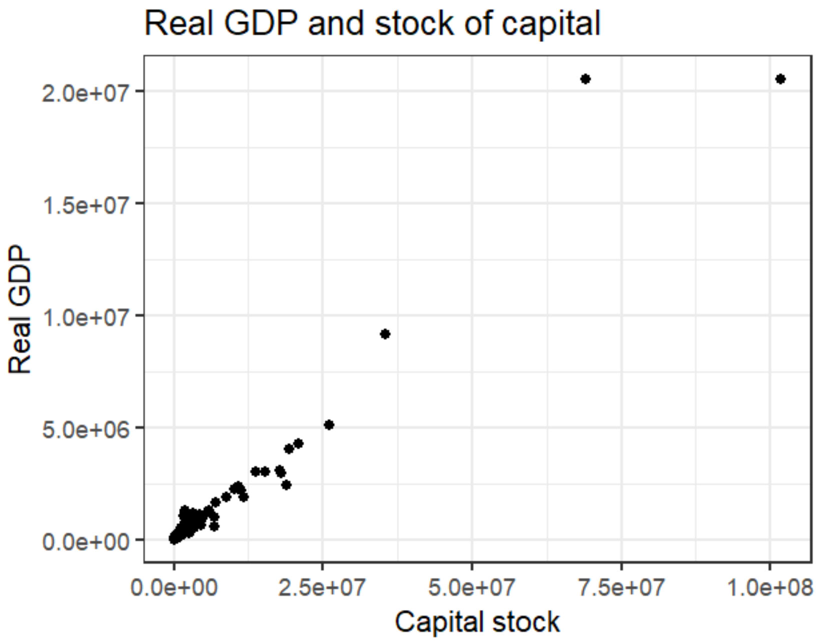
A 1 million increase in population is associated with a 583.6 million USD decrease in real GDP, all other variables held constant. The effect is economically not significant as we would expect a positive sign.

A 1 unit increase in total factor productivity is associated with a 594,300 million USD increase in real GDP, all other variables held constant. The effect is economically significant.

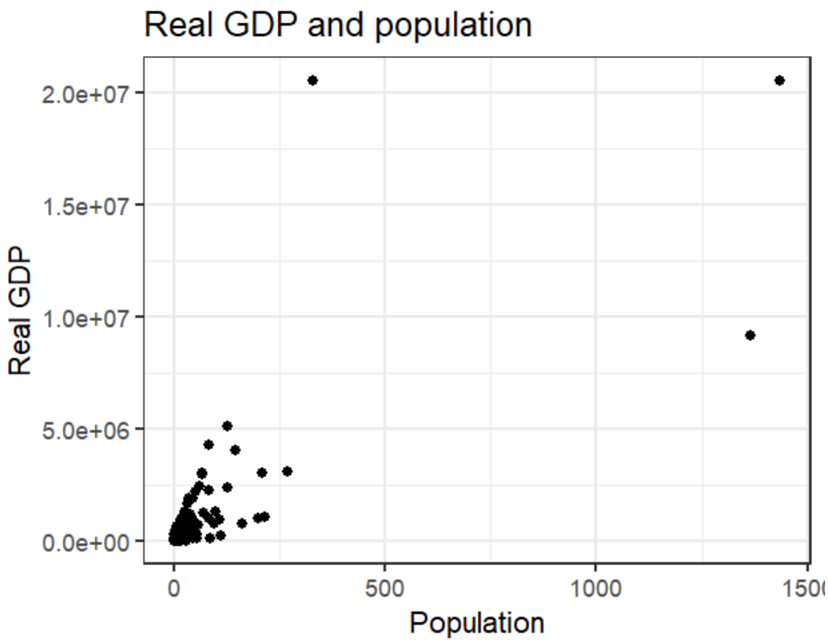
```
> # F-Test
> # H_0 = beta_pop + beta_rnna = 0
> # F = ((SSRr-SSRur)/q) / (SSRur/(n-k-1))
>
> SSRur <- sum(reg_gdp_capital_popul_product$residuals^2)
> reg_gdp_product <- lm(rgdpna ~ rtfpna)
> SSRr <- sum((reg_gdp_product$residuals^2))
>
> F_statistic <- ((SSRr - SSRur)/2)/(SSRur/(183-5-1))
> critical_value <- qf(0.95, 2, (183-5-1))
> F_statistic
[1] 1962.822
> critical_value
[1] 3.047012
>
> # for 5% significanc level
> if((abs(F_statistic) > critical_value)){
+   print("reject H_0")
+   } else
+   print("do not reject H_0")
[1] "reject H_0"
```

The F-statistic takes a very high value for the null hypotheses. Therefore, we reject H_0 at a significance level of 5%. That means with statistical significance the joint effect of population and the capital stock are not equal to 0.

(c) Create a graphical scatter plot of real GDP ($rgdpna$) and the stock of capital ($rnna$) and another scatter plot of real GDP ($rgdpna$) and population (pop). Are both relationships linear?



The relationship between *Real GDP at constant prices of 2017* and *Capital stock at constant prices of 2017* is approximately linear.



The relationship between *Real GDP at constant prices of 2017* and *Population* might not be linear as it looks like having a high variance (because of the spread of the data points).

(d) Create the logarithm of real GDP ($rgdpna$), the stock of capital ($rnna$), the population (pop) and the total factor productivity ($rtfpna$) and estimate by OLS a regression where the dependent variable is the logarithm of real GDP ($rgdpna$), and the explanatory variables are the logarithm of the stock of capital ($rnna$), the logarithm of the population (pop) and the logarithm of the total factor productivity ($rtfpna$). Interpret the estimated coefficients.

```
> log_rgdpna <- log(rgdpna)
> log_rnna <- log(rnna)
> log_pop <- log(pop)
> log_rtfpna <- log(rtfpna)
>
> reg_log <- lm(log_rgdpna ~ log_rnna + log_pop + log_rtfpna)
> summary(reg_log)

Call:
lm(formula = log_rgdpna ~ log_rnna + log_pop + log_rtfpna)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.15956 -0.16075  0.03155  0.22564  0.75449 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.83674    0.28970   2.888  0.00464 ** 
log_rnna     0.79909    0.02487  32.134  < 2e-16 ***
log_pop      0.19837    0.03001   6.610 1.29e-09 ***
log_rtfpna   2.65115    0.50272   5.274 6.43e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3554 on 114 degrees of freedom
  (65 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.9634,	Adjusted R-squared:  0.9624 
F-statistic:  1001 on 3 and 114 DF,  p-value: < 2.2e-16
```

A 1% increase in capital stock is associated with a 0.80% increase in real GDP, all other variables held constant.

A 1% increase in population is associated with a 0.20% increase in real GDP, all other variables held constant.

A 1% increase in total factor productivity is associated with a 2.65% increase in real GDP, all other variables held constant.

(e) Using the results of the previous question, test the hypothesis of constant returns to scale (that is, that the sum of the two estimated coefficients is equal to 1). Does the data verify this hypothesis? Rewrite the previous equations to test the same hypothesis but now based on a single estimated coefficient.

```
> reg_log_coefficiants <- as.matrix(reg_log$coefficients)
> reg_log_coefficiants
                 [,1]
(Intercept) 0.8367386
log_rnna    0.7990854
log_pop     0.1983711
log_rtfpna  2.6511541

> beta_1 <- reg_log_coefficiants[2]
> beta_2 <- reg_log_coefficiants[3]
>
> var_beta_1 <- vcov(reg_log)[2,2]
> var_beta_2 <- vcov(reg_log)[3,3]
> cov_beta_1_2 <- vcov(reg_log)[2,3]
>
> se_beta_1_2 <- sqrt(var_beta_1 + var_beta_2 + 2*cov_beta_1_2)
>
> t <- (beta_1 + beta_2 - 1)/se_beta_1_2
> t
[1] -0.1225907
```

```
> # für 5%iges Signifikanzniveau
> t_statistic <- abs(t)
> c.0025 <- 1.96
> if((t_statistic > c.0025)){
+   print("reject H_0")
+   } else
+     print("do not reject H_0")
[1] "do not reject H_0"
```

H_0 cannot be rejected at the 5% significance level. The sum of the two estimated coefficients *stock of capital* and *population* are not statistically significantly different from 1. Although we fail to reject the null hypotheses, we still cannot verify that we have constant returns to scale.

```
> # H_0 = beta_1+2 = 1
> beta_sum <- data_2019$rnna + data_2019$pop
> log_beta_sum <- log(beta_sum)
>
> reg_beta_sum <- lm(log_rgdpna ~ log_beta_sum + log_rtfpna)
> summary(reg_beta_sum)

Call:
lm(formula = log_rgdpna ~ log_beta_sum + log_rtfpna)

Residuals:
     Min       1Q   Median       3Q      Max
-1.19930 -0.24047  0.04364  0.24004  1.07785

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.29470    0.27370  -1.077 0.283864
log_beta_sum  0.91899    0.01992  46.133  < 2e-16 ***
log_rtfpna    2.13561    0.58150   3.673 0.000366 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4161 on 115 degrees of freedom
  (65 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.9494,	Adjusted R-squared:  0.9485
F-statistic:  1079 on 2 and 115 DF,  p-value: < 2.2e-16
```

```
> se_beta_sum <- 0.01992
>
> reg_beta_sum$coefficients
 (Intercept) log_beta_sum   log_rtfpna
  -0.2946980    0.9189879    2.1356118
> t_2 <- (coef(reg_beta_sum)[2] - 1)/se_beta_sum
> t_2
log_beta_sum
   -4.066873
>
> # für 5%iges Signifikanzniveau
> t_statistic_2 <- abs(t_2)
> if((t_statistic_2 > c.0025)){
+   print("reject H_0")
+ } else
+   print("do not reject H_0")
[1] "reject H_0"
```

This hypothesis captures the effects of population and capital stock in one variable. In this test we reject the null hypotheses. Meaning we do not observe constant returns to scale.

(f) Estimate by OLS a regression where the dependent variable is the logarithm of real GDP (*rgdpna*) and the explanatory variables are the logarithm of the stock of capital (*rnna*) and the logarithm of the population (*pop*). Then, estimate another regression where you use the logarithm of the employed population (*emp*) instead of the logarithm of the population (*pop*). Does the estimated coefficient change much?

```
> reg_log_rnna_pop <- lm(log_rgdpna ~ log_rnna + log_pop)
> summary(reg_log_rnna_pop)

Call:
lm(formula = log_rgdpna ~ log_rnna + log_pop)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.48967 -0.16685  0.04009  0.23792  0.77760 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.67835    0.24632   2.754   0.0065 ** 
log_rnna     0.80242    0.02175  36.892   <2e-16 ***
log_pop      0.22527    0.02266   9.943   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.394 on 177 degrees of freedom
  (3 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.9696,	Adjusted R-squared:  0.9692 
F-statistic:  2818 on 2 and 177 DF,  p-value: < 2.2e-16

> reg_log_emp_pop <- lm(log_rgdpna ~ log_rnna + log_emp)
> summary(reg_log_emp_pop)

Call:
lm(formula = log_rgdpna ~ log_rnna + log_emp)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.52126 -0.14926  0.02792  0.23357  0.87780 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.14054    0.27427   4.158 5.05e-05 ***
log_rnna     0.78130    0.02294  34.054  < 2e-16 ***
log_emp      0.23735    0.02482   9.562  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3881 on 172 degrees of freedom
  (8 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.9668,	Adjusted R-squared:  0.9665 
F-statistic:  2508 on 2 and 172 DF,  p-value: < 2.2e-16
```

When using the *logarithm of the employed population* instead of the *logarithm of the population*, the estimated coefficient does not change much. This indicates a strong correlation between both estimates.

2. In empirical studies, the regression results for different models will often be reported as in the following table. Below the coefficients for each independent variable is its standard error. I estimated three models. The dependent variables are the log of CEO salary for 177 companies. The independent variable *lsales* is the log of sales, *lmktval* is the log of the market value of the firm, *profmarg* is profit as a percentage of sales, *ceoten* is years as CEO with the current company, and *comten* is total years with the company.

Table 1: Regression on factors for CEO salaries

	Dependent variable: log(salary)		
	(1)	(2)	(3)
lsales	0.224	0.158	0.188
	(0.027)	(0.040)	(0.040)
lmktval		0.112	0.100
		(0.050)	(0.049)
profmarg		−0.002	−0.002
		(0.002)	(0.002)
ceoten			0.017
			(0.006)
comten			−0.009
			(0.003)
Constant	4.961	4.621	4.572
	(0.200)	(0.254)	(0.253)
Observations	177	177	177
R^2	0.281	0.303	0.353
Adjusted R^2	0.277	0.291	0.334
Residual Std. Error	0.515 (df = 175)	0.510 (df = 173)	0.495 (df = 171)
F Statistic	68.345 (df = 1; 175)	25.128 (df = 3; 173)	18.622 (df = 5; 171)

(a) Explain the coefficient on *lmktval* and answer whether the company's market value significantly affects the CEO salary.

<u>specification (2):</u> A 1% increase in market value of the firm is associated with a 0.112% increase in CEO salary, all other variables held constant.

$H_0: \beta_2 = 0 \qquad H_1: \beta_2 \neq 0$

$$t = \frac{\widehat{\beta_2}}{\mathrm{se}(\widehat{\beta_2})} = \frac{.112}{.050} = 2.24$$

```
> ## specification (2)
> ## p-value for a two-sided hypothesis test
> 2*(1-pt(2.24, df=173))
[1] 0.02636421
```

As $0.026 < 0.05$ we can reject H_0 at the 5% significance level. The market value of the firm has a statistically significant effect on CEO salary.

<u>specification (3):</u> A 1% increase in market value of the firm is associated with a 0.100% increase in CEO salary, all other variables held constant.

$H_0: \beta_2 = 0 \qquad H_1: \beta_2 \neq 0$

$$t = \frac{\widehat{\beta_2}}{\mathrm{se}(\widehat{\beta_2})} = \frac{.100}{.049} \approx 2.04$$

```
> ## specification (3)
> ## p-value for a two-sided hypothesis test
> 2*(1-pt((.100/.049), df=171))
[1] 0.04280647
```

As $0.043 < 0.05$ we can reject H_0 at the 5% significance level. The market value of the firm has a statistically significant effect on CEO salary.

(b) Interpret the coefficients on *ceoten* and *comten*. Are they statistically significant individually?

ceoten: A one year increase in being CEO with the current company is associated with a 1.7% increase in CEO salary, all other variables held constant.

$H_0: \beta_4 = 0 \qquad H_1: \beta_4 \neq 0$

$$t = \frac{\widehat{\beta_4}}{\operatorname{se}(\widehat{\beta_4})} = \frac{.017}{.006} \approx 2.83$$

```
> ## ceoten
> ## p-value for a two-sided hypothesis test
> 2*(1-pt((.017/.006), df=171))
[1] 0.005160772
```

As $0.005 < 0.01$ we can reject H_0 at the 1% significance level. The years as CEO with the current company has a statistically significant effect on CEO salary.

comten: A one year increase in total years of being with the company is associated with a 0.9% decrease in CEO salary, all other variables held constant.

$H_0: \beta_5 = 0 \qquad H_1: \beta_5 \neq 0$

$$t = \frac{\widehat{\beta_5}}{\operatorname{se}(\widehat{\beta_5})} = \frac{-.009}{.003} = -3$$

```
> ## comten
> ## p-value for a two-sided hypothesis test
> 2*(1-pt(3, df=171))
[1] 0.003103408
```

As $0.003 < 0.01$ we can reject H_0 at the 1% significance level. The total years with the company has a statistically significant effect on CEO salary.

(c) Are the coefficients on *ceoten* and *comten* jointly significant from zero? (Hint: use R^2)

$H_0: \beta_4 = 0, \beta_5 = 0$

$H_1: \beta_4 \neq 0, \beta_5 \neq 0$

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n-k-1)} = \frac{(.353 - .303)/2}{(1 - .353)/(177 - 5 - 1)} \approx 6.607$$

The critical value is 3.049 for a two-sided F test at the 10% significance level.

The critical value is 3.770 for a two-sided F test at the 5% significance level.

The critical value is 5.466 for a two-sided F test at the 1% significance level.

```
> cv10 <- qf(0.95, 2, 171)
> cv10
[1] 3.048833
>
> ## 5% significance niveau
>
> cv5 <- qf(0.975, 2, 171)
> cv5
[1] 3.769614
>
> ## 1% significance niveau
>
> cv1 <- qf(0.995, 2, 171)
> cv1
[1] 5.465926
```

As $F > 5.466$ we can reject H_0 at the 1% significance level. The years as CEO with the current company (ceoten) and total years of being with the company (comten) have a statistically significant effect on CEO salary.

(d) What do you make of the fact that longer tenure with the company, holding the other factors fixed, is associated with a lower salary?

There could be many reasons why employees have a long tenure with the company. When employees are satisfied with their profession, they have no incentives to leave the company. Other may have high responsibilities (also towards others) so they feel like they cannot change their job. When employees are risk-averse, they may do not want to ask for higher salary as they fear getting replaced. This could result in lower salary.

Moreover, there may be differences in urban and rural area. In cities it is easier for persons to change the company as they have more possibilities to find a profession that fits to their skills and know-how.

Dedication, loyalty as well as limited mobility may result in lower salary as the employers do not have incentives to increase their wages because the employees stay anyway.