WS 2021/22 1048 Econometrics I
Case Study 1

Anja Kulagic (h12100543), Jovana Mileusnic (h12100542), Ema Vargova (h11914081)

1 Data Analysis

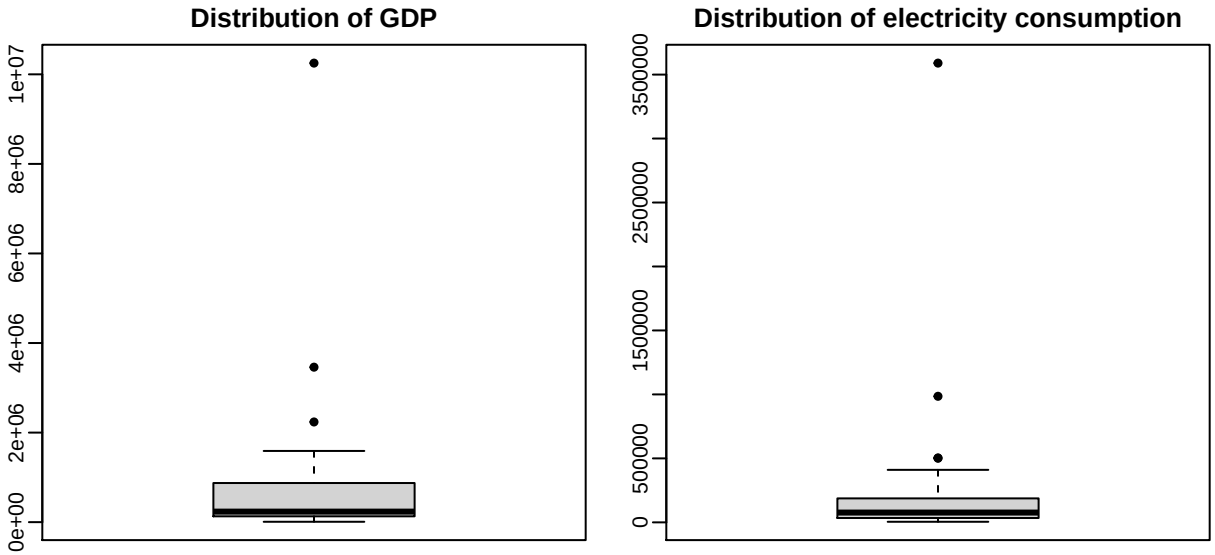
1.1

a)

```
data <- read.csv("ELCONS_GDP.csv",header=TRUE) #fetching the dataset
summary(data)
```

```
##    COUNTRY            TOTALCONS            GDP
##  Length:35          Min.   :   4484   Min.   :    8378
##  Class :character   1st Qu.:  33843   1st Qu.:  129894
##  Mode  :character   Median :  76468   Median :  235450
##                     Mean   : 245447   Mean   :  837337
##                     3rd Qu.: 187280   3rd Qu.:  873755
##                     Max.   :3589779   Max.   :10250952
```

```
par(mfrow = c(1, 2), cex = .7, mar = c(1,1.5, 2, 1.5), mgp = c(1.5, .5, 0), pch=21, lwd=1)
boxplot(data$GDP, xlab = "GDP", main="Distribution of GDP", xaxt="n", bg="black")
boxplot(data$TOTALCONS, xlab = "Electricity Consumption", bg="black", xaxt="n",
        main="Distribution of electricity consumption")
```



The interquartile range: $IQR = Q_3 - Q_1$

The interquartile range can be used to detect outliers. This is done using following steps:

(1) Calculate the interquartile range for the data.
(2) Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
(3) Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
(4) Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

```
boxplot.stats(data$GDP)$out
```

```
## [1]  2237046  3460607 10250952
```

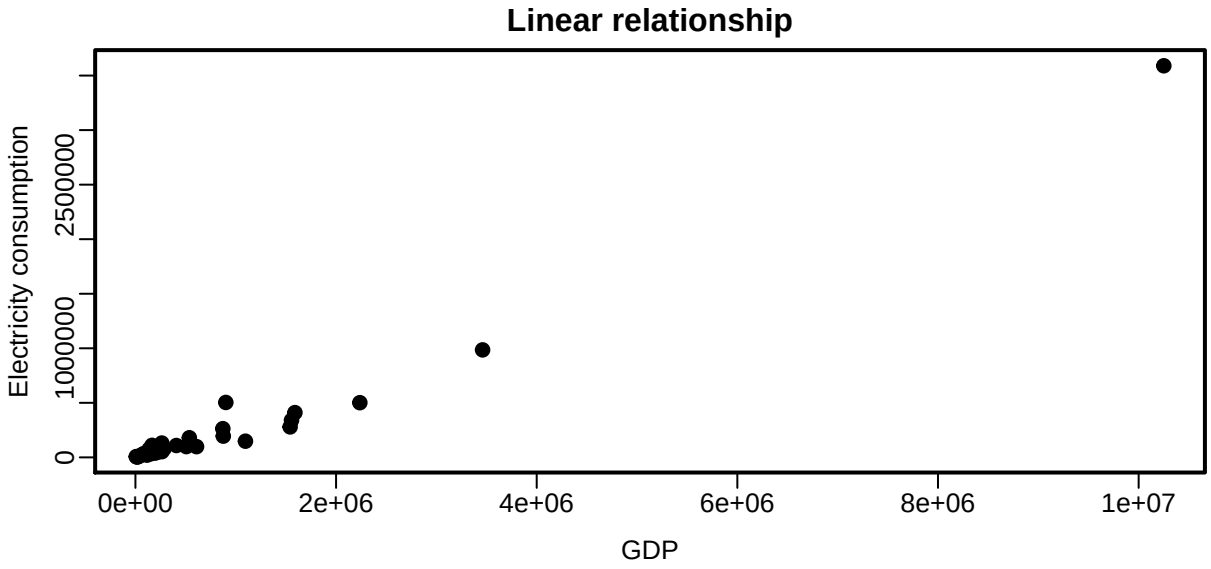
```
boxplot.stats(data$TOTALCONS)$out
```

```
## [1]  503403  501411  985360 3589779
```

Built on the IQR criteria, we can conclude that Germany, Japan, and the USA represent outliers within the defined data set on the basis of GPA and the total electricity consumption.

b)

```
par(cex = .8, mar = c(3,3, 2, 1.5), mgp = c(2, .5, 0), lwd=2, pch=19)
plot(data$GDP, data$TOTALCONS,
     xlab="GDP", ylab="Electricity consumption",
     main="Linear relationship")
```



There are several reasons why *Model 1* does not instill confidence in us, primarily, due to the high level of the SER, and presence of the extreme values.

```
cor.test(x = data$GDP, y = data$TOTALCONS)
```

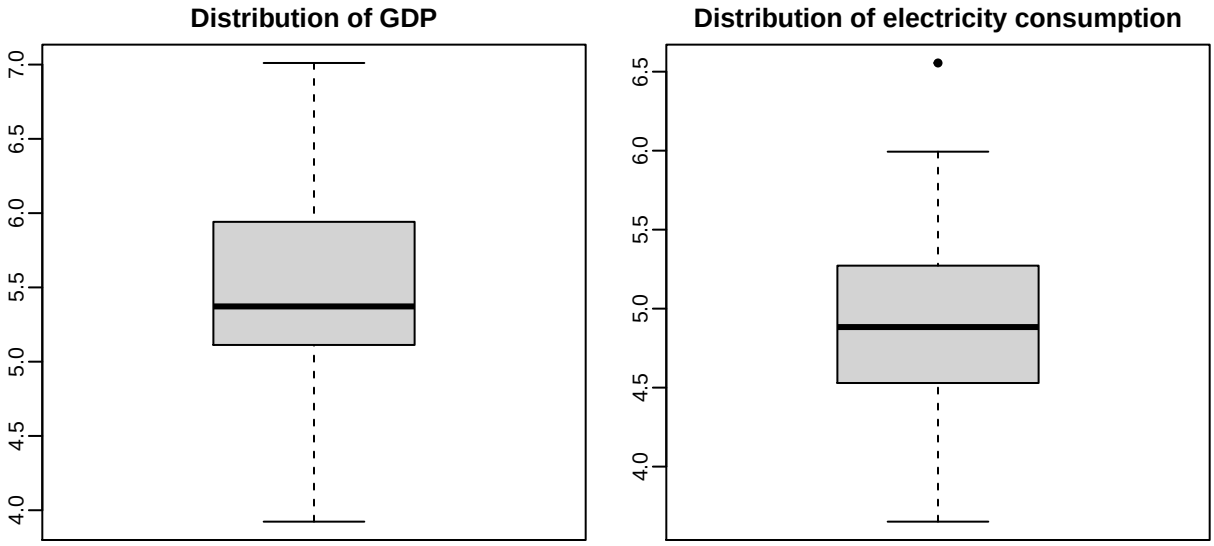
```
##
##  Pearson's product-moment correlation
##
## data:  data$GDP and data$TOTALCONS
## t = 37.001, df = 33, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9764672 0.9940621
## sample estimates:
##       cor
## 0.9881617
```

Despite extremely high value of Pearson's coefficient, the fact that the *Linear Relationship* plot does not show the presence of linear relationship between observed variables, the "significant" coefficient could be meaningless.

Generating log values of GDP and Electricity consumption in order to **investigate non-linear relationship** between the two variables:

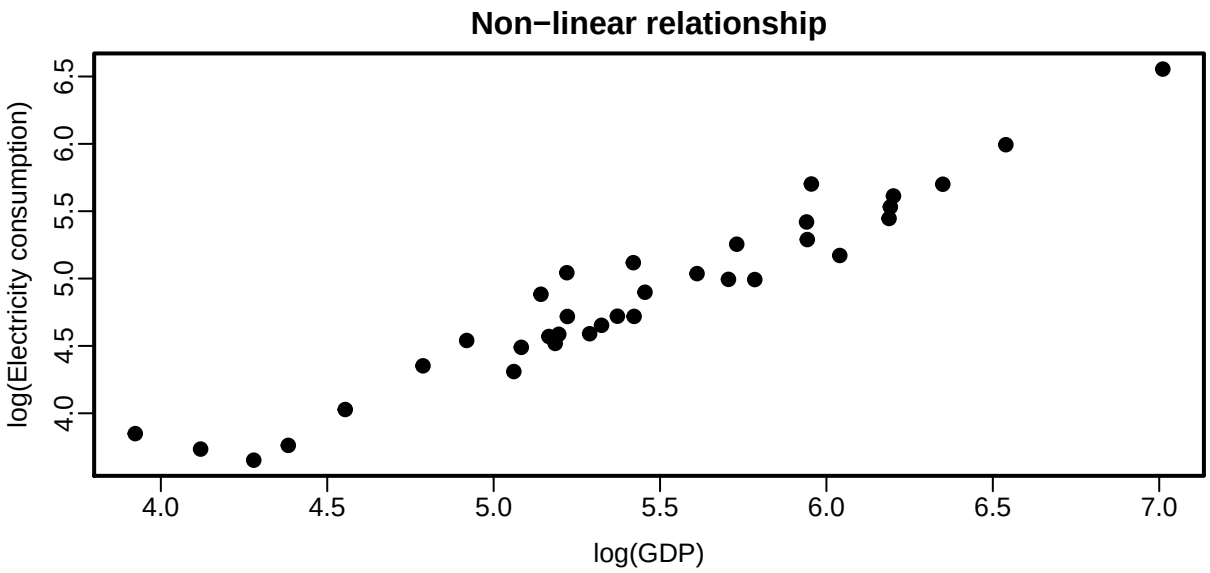
```
data$GDP_log <- log(data$GDP,base=10)
data$TOTALCONS_log <- log(data$TOTALCONS,base=10)
```

```
par(mfrow = c(1, 2), cex = .7, mar = c(1,1.5, 2, 1.5), mgp = c(1.5, .5, 0), pch=21, lwd=1)
boxplot(data$GDP_log, xlab = "GDP", main="Distribution of GDP", xaxt="n", bg="black")
boxplot(data$TOTALCONS_log, xlab = "Electricity Consumption", bg="black", xaxt="n",
        main="Distribution of electricity consumption")
```



After the log transformation, we can see that the outliers are not as extreme anymore with the distributions not being as skewed is it is the case with the original data.

```
par(cex = .8, mar = c(3,3, 2, 1.5), mgp = c(2, .5, 0), lwd=2, pch=19)
plot(data$GDP_log, data$TOTALCONS_log,
     xlab="log(GDP)", ylab="log(Electricity consumption)", main="Non-linear relationship")
```



As we can see the plots above, the **non-linear relationship** after transformation of both variables to a logarithmic scale seems to have a better fit than the **linear relationship** as the variance in the *electricity consumption* seems to be increasing as the value of *GDP* increases which might be a violation of some assumptions for using the linear model.

1.2

```
model1 <- lm(TOTALCONS ~ GDP, data=data)
summary(model1)
```

```
##
## Call:
## lm(formula = TOTALCONS ~ GDP, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217087  -17446   20405   37307  236430
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.754e+04  1.791e+04  -2.096   0.0438 *
## GDP          3.380e-01  9.134e-03  37.001   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 95810 on 33 degrees of freedom
## Multiple R-squared:  0.9765, Adjusted R-squared:  0.9758
## F-statistic:  1369 on 1 and 33 DF,  p-value: < 2.2e-16
```

For Model 1, the estimate of $\hat{\beta_0}$ is -3.7542325×10^4 and the estimate of $\hat{\beta_1}$ is 0.338 meaning that when the value of GDP increases by 1, the value of electricity consumption increases by 0.338, **on average**.

The R^2 in the output is called **Multiple R-squared** and has a value of 0.9765. Hence, 97.65% of the variance of the dependent variable is explained by the predictor.

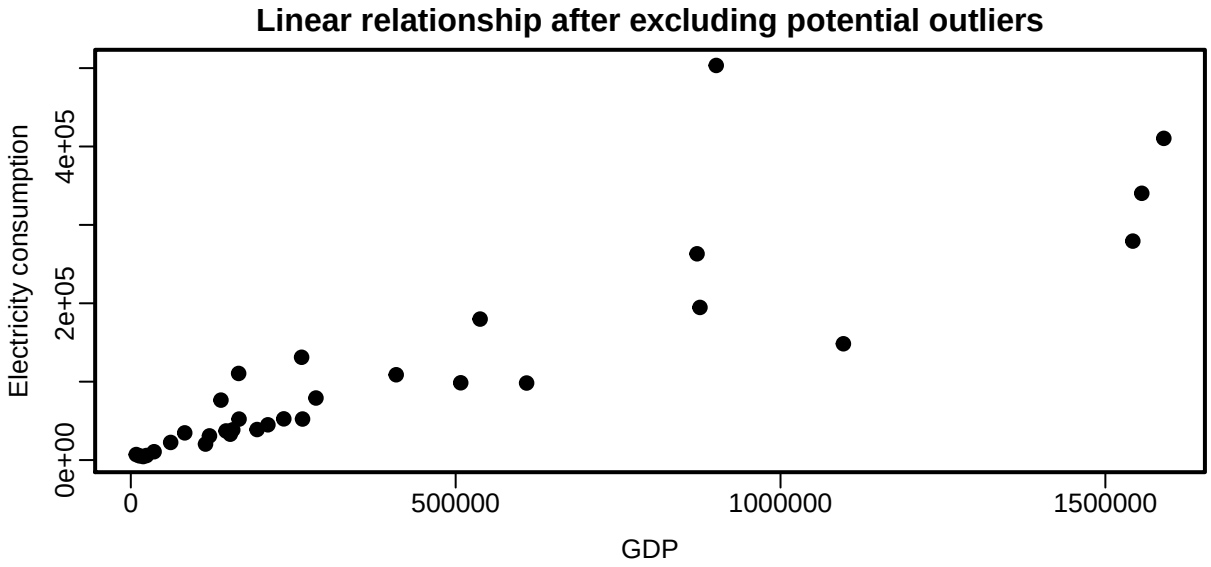
The SER is called **Residual standard error** and equals 95810. The unit of the SER is the same as the unit of the dependent variable. That is, on average the deviation of the actual achieved test score and the regression line is 95,810 points (GWh).As an estimator of the standard deviation of the regression residual U_i, it provides a measure of the spread of the observations around the regression line. The higher the SER is, the larger the spread of the observation is. Consequently, when the SER is high, the prediction of the dependent variable that has been made from a particular predictor can be wrong by a large amount. Contrary, if the SER is close to 0, the prediction of the independent variable could be considered as a good one.

The Least Square Assumption: Large outliers can make OLS regression results misleading.

$$\Downarrow$$

Idea - Excluding USA, Japan and Germany as the outliers:

```
data1 <- subset(data, data$COUNTRY!="USA" & data$COUNTRY!="DEU" & data$COUNTRY!="JPN")
par(cex = .8, mar = c(3,3, 2, 1.5), mgp = c(2, .5, 0), lwd=2, pch=19)
plot(data1$GDP, data1$TOTALCONS, xlab="GDP", ylab="Electricity consumption",
     main="Linear relationship after excluding potential outliers",)
```




```
model1_mod <- lm(data1$TOTALCONS ~ data1$GDP, data = data1)
summary(model1_mod)
```

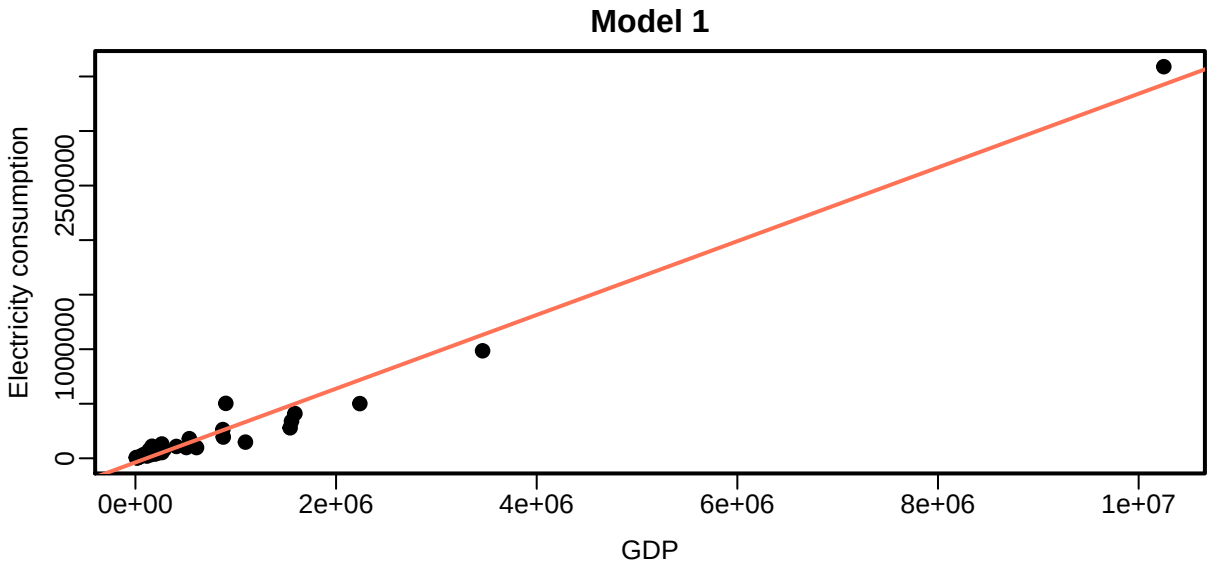
```
##
## Call:
## lm(formula = data1$TOTALCONS ~ data1$GDP, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -116439  -19777  -11958    1078  283278
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.459e+04  1.521e+04   0.960    0.345
## data1$GDP   2.281e-01  2.433e-02   9.374 2.02e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64020 on 30 degrees of freedom
## Multiple R-squared:  0.7455, Adjusted R-squared:  0.737
## F-statistic: 87.86 on 1 and 30 DF,  p-value: 2.019e-10
```

After excluding the outliers, the *Linear relationship after excluding potential outliers* plot is proof of the presence non-linear relationship. As we can notice, although the *SER* is still quite high, it is important to emphasize that it is significantly lower than in *Model 1*, thus to conclude that the presence of extremes can violate one of the main Linear regression assumptions.

1.3

```
par(cex = .8, mar = c(3,3, 2, 1.5), mgp = c(2, .5, 0), lwd=2, pch=19)
plot(data$GDP, data$TOTALCONS,
     xlab="GDP", ylab="Electricity consumption", main="Model 1")
abline(reg=model1, col="coral1")
```

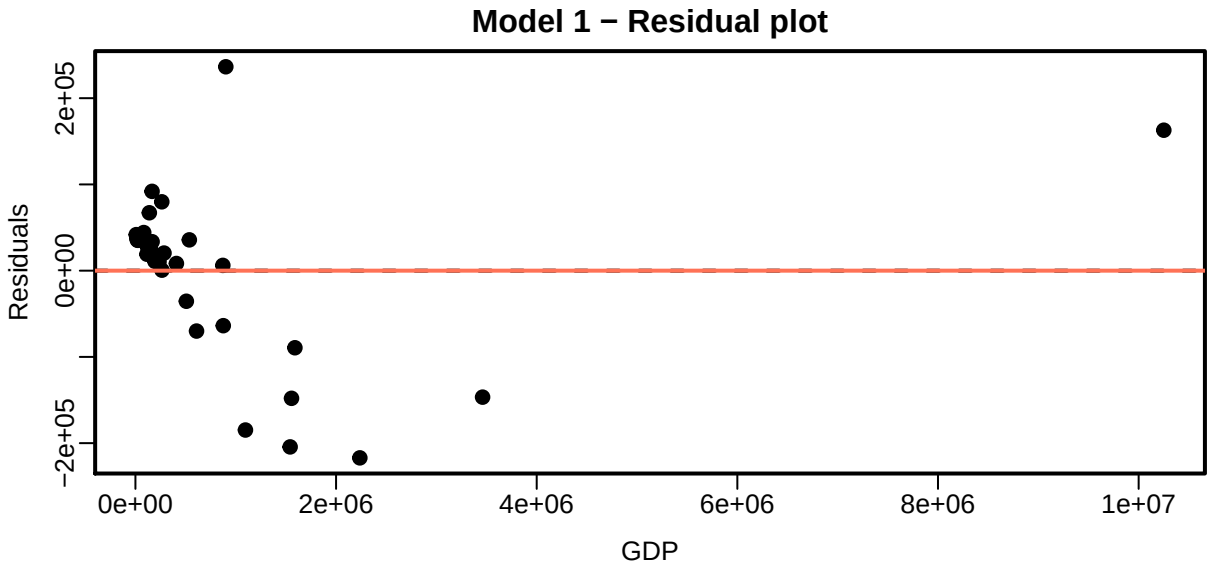


1.4

Residual Diagnosis:
(1) The residuals spread randomly around the 0 line indicating that the relationship is linear.
(2) The residuals form an approximate horizontal band around the 0 line indicating homogeneity of error variance.
(3) In other words, the residuals scatterplot should be roughly rectangular-shaped.

```
par(cex = .8, mar = c(3,3, 2, 1.5), mgp = c(2, .5, 0), lwd=2, pch=19)
plot(data$GDP, model1$residuals,
     xlab="GDP", ylab="Residuals",
     main="Model 1 - Residual plot")
```

```r
abline(h=0, col="black",lty=2)
abline(h=mean(model1$residuals), col="coral1")
```



The residuals spread randomly around the 0 line indicating that the model is **unbiased** and relationship is linear. The residuals form an approximate horizontal band around the 0 line indicating homogeneity of error variance. However, the residuals do not seem to be normally distributed meaning that the model **violates homoskedacity assumption**.

1.5

Using logs for variables on both sides of your econometric specification is called **a log-log model**. It represents a widespread way to handle situations where a non-linear relationship exists between the independent and dependent variables. This model is practical because the log transformation generates the desired linearity in parameters. It is also important, having in mind that **linearity in parameters is one of the OLS assumptions.**

$$\log X]Y = \tilde{\beta}_0 \cdot X^{\beta_1} \cdot \tilde{u}, \quad \tilde{\beta}_0, \tilde{u} > 0$$
$$\Downarrow$$
$$\log Y = \beta_0 + \beta_1 \log X + u, \quad \mathbb{E}[u \mid \log X] = 0$$

```r
model2 <- lm(TOTALCONS_log ~ GDP_log, data=data)
summary(model2)
```

```
##
## Call:
## lm(formula = TOTALCONS_log ~ GDP_log, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26774 -0.13193 -0.04587  0.08451  0.35785
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1103     0.2333  -0.473    0.639
## GDP_log       0.9187     0.0427  21.517   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1725 on 33 degrees of freedom
## Multiple R-squared:  0.9335, Adjusted R-squared:  0.9314
## F-statistic:   463 on 1 and 33 DF,  p-value: < 2.2e-16
```

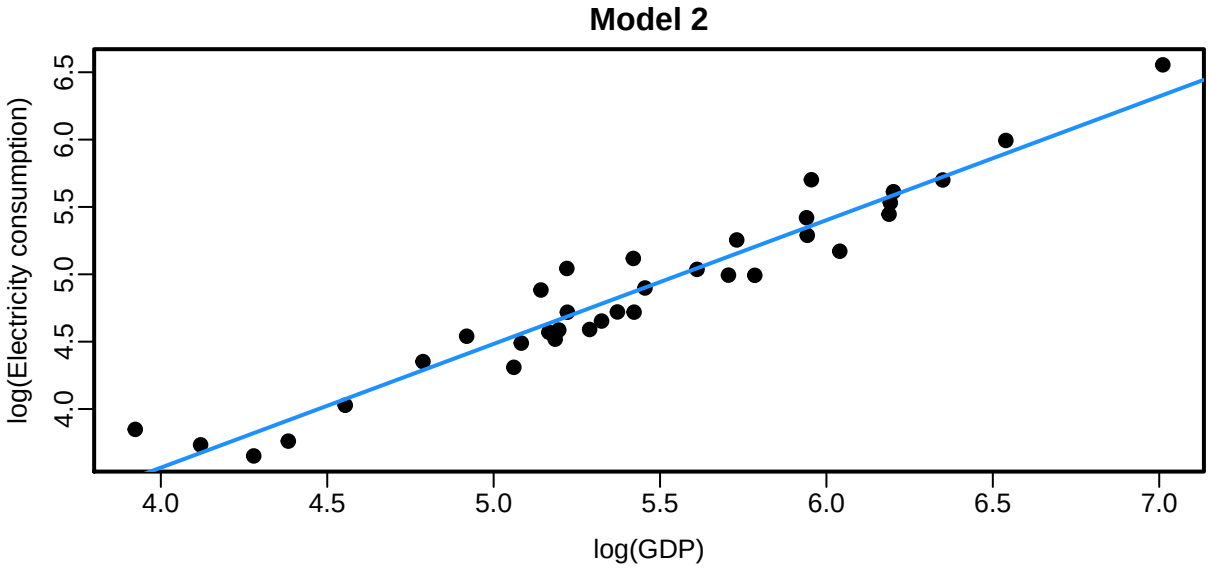
For Model 2, the estimate of $\hat{\beta}_0$ is -0.1103 and the estimate of $\hat{\beta}_1$ is 0.9187 meaning that when the value of GDP increases by 1%, the value of electricity consumption increases by 0.9187%, **on average**.

The R^2 in the output is called **Multiple R-squared** and has a value of 0.9335. Hence, 93.35% of the variance of the dependent variable is explained by the predictor.

The SER is called **Residual standard error** and equals 95810. The unit of the SER is the same as the unit of the dependent variable. That is, on average, the deviation of the actual achieved test score and the regression line is 0.1725 points (GWh).

1.6

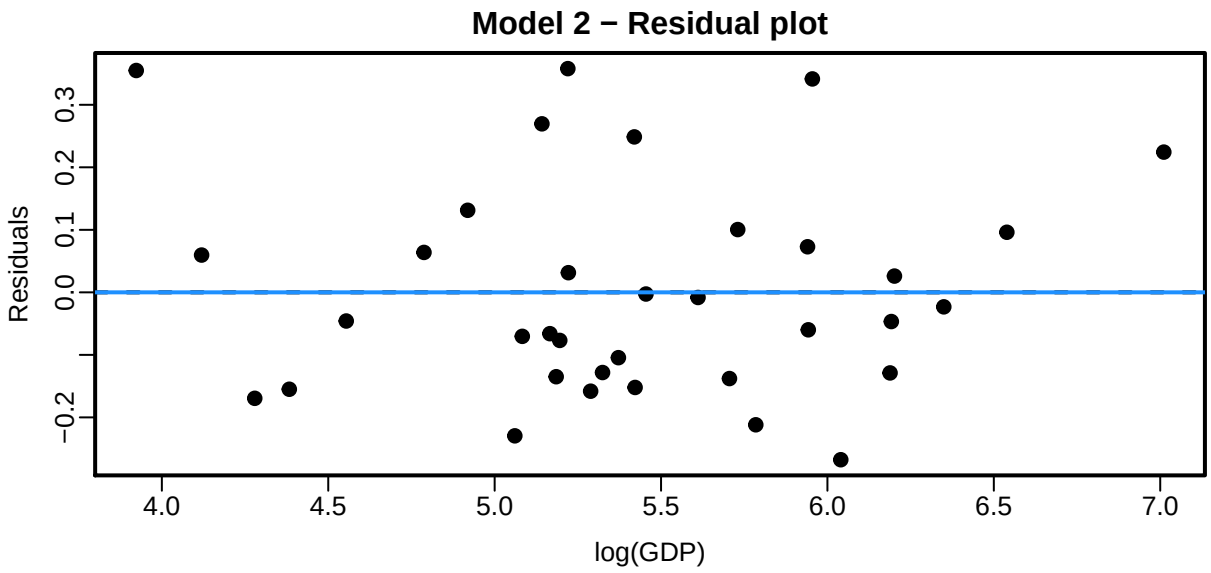
```
par(cex = .8, mar = c(3,3, 2, 1.5), mgp = c(2, .5, 0), lwd=2, pch=19)
plot(data$GDP_log, data$TOTALCONS_log,
     xlab="log(GDP)", ylab="log(Electricity consumption)", main="Model 2")
abline(reg=model2, col="dodgerblue")
```



The fit of Model 2 seems to be better than the fit of Model 1 as the *real values are closer to the fitted values* of Model 2 meaning that the **RSS** is smaller for Model 2.

```
par(cex = .8, mar = c(3,3, 2, 1.5), mgp = c(2, .5, 0), lwd=2, pch=19)
plot(data$GDP_log, model2$residuals,
     xlab="log(GDP)", ylab="Residuals", main="Model 2 - Residual plot")
```

```
abline(h=0, col="black",lty=2)
abline(h=mean(model2$residuals), col="dodgerblue")
```



Residual Diagnosis:

The error term of Model 2 is **homoskedastic** as there does not seem to be any correlation between the log value of GDP and the residuals. Moreover, the model specification is **unbiased** as the expected value of residuals given the log values of GDP is 0. This means that both of the standard assumptions for using the OLS model seem to be satisfied which is not the case for Model 1 as it violates homoskedacity assumption.

1.7

1) In Model 1, the expected **decrease** in electricity consumption is 8,449 GWh, **on average**.

```
-25*1000*model1$coefficients[2]
```

```
##       GDP
## -8449.092
```

2) The electricity consumption is expected to increase by 2.77%, **on average**.

```
3*model2$coefficients[2]
```

```
##  GDP_log
## 2.756209
```

3) According to Model 1, the expected electricity consumption is 300,421 GWh while based on Model 2, the expected electricity consumption is 252,428 GWh, **on average**.

```
predict(model1,newdata=data.frame(GDP=1E6))
```

```
##        1
## 300421.4
```

```
exp(predict(model2,newdata=data.frame(GDP_log=log(1E6,base=10)))*log(10))
```

```
##        1
## 252427.7
```

1.8

```
range(data$GDP)
```

```
## [1]     8377.672 10250952.000
```

```
range(data$TOTALCONS)
```

```
## [1]    4484 3589779
```

Considering the range of explanatory variable GDP, these data alone are not a reliable basis for predicting the effect of such a extremely high level of explanatory variable such as the GDP of China in 2020. Moreover, the defined linear model takes into account the relationship between the dependent and the independent variable that is characteristic of a given moment. Nevertheless, the characteristics of this connection are subject to change and we cannot say with certainty that the relationship we defined based on data from 2000 is valid assumption in 2021 as well.

However, *Model 2* and its low level of the SER, the R^2, and no presence of extreme values lead us to conclusion that both of the standard assumptions for using the OLS model seem to be satisfied, thus it could be a decent base for predicting given values, with a certain dose of uncertainty.

China's Gross Domestic Product in 2000: 1.211 trillion USD

```
exp(predict(model2,newdata=data.frame(GDP_log=log(1.211E6,base=10)))*log(10))
```

```
##        1
## 300970.9
```

According to Model 2, the expected electricity consumption is 3,009,709 GWh.

Austria's Gross Domestic Product in 2021: GDP in Austria is expected to reach 435.00 USD Billion by the end of 2021, according to Trading Economics global macro models and analysts expectations.

```
exp(predict(model2,newdata=data.frame(GDP_log=log(435E3,base=10)))*log(10))
```

```
##        1
## 117490.8
```

2 Theory

Model 3:

$$\log X = \gamma_0 + \gamma_1 \log x + u, \quad \mathbb{E}[u \mid \log Y] = 0$$

Basic properties:

(1) *The variance of a constant is zero:* $Var(\beta_0) = 0 \;\Rightarrow\; Var(\gamma_0) = 0$

(2) *If all values are scaled by a constant, the variance is scaled by the square of that constant:* $\text{Var(aX)} = a^2 Var(X)$

(3) *Symmetric:* $Cov(X, Y) = Cov(Y, X)$

Question of the interest: Is it true that $\gamma_1 = \frac{1}{\beta_1}$ where γ_1 is the slope coefficient of **Model 3** and γ_1 is that of **Model 2**?

By definition: $\beta_1 = \frac{Cov(\log X, \log Y)}{Var(X)} \;\Rightarrow\; \gamma_1 = \frac{Cov(\log X, \log Y)}{Var(Y)}$

$$Var(\log Y) = Var(\beta_0 + \beta_1 \log X) = Var\beta_0 + \beta_1^2 Var(\log X) = \beta_1^2 Var(\log X)$$

$$\Downarrow$$

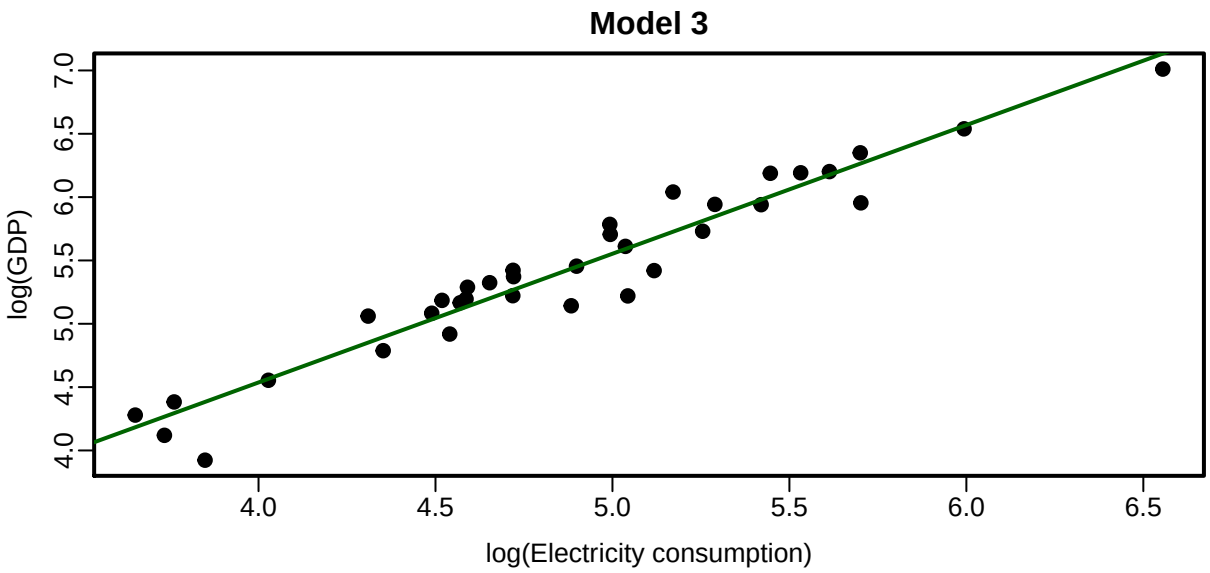
$$\gamma_1 = \frac{Cov(\log X, \log Y)}{Var(\log Y)} = \frac{Cov(\log X, \log Y)}{\beta_1^2 Var(\log X)} = \frac{1}{\beta_1}$$

Hence, γ_1 should be equal to $\frac{1}{0.9187} = 1.088$. By compering the features, we would notice that it is not a case in our example.

```
model3 <- lm(GDP_log ~ TOTALCONS_log, data=data)
summary(model3)
```

```
##
## Call:
## lm(formula = GDP_log ~ TOTALCONS_log, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46014 -0.09418  0.04905  0.11220  0.31327
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.47269    0.23197   2.038   0.0497 *
## TOTALCONS_log  1.01603    0.04722  21.517   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1814 on 33 degrees of freedom
## Multiple R-squared:  0.9335, Adjusted R-squared:  0.9314
## F-statistic:   463 on 1 and 33 DF,  p-value: < 2.2e-16
```

```
par(cex = .8, mar = c(3,3, 2, 1.5), mgp = c(2, .5, 0), lwd=2, pch=19)
plot(data$TOTALCONS_log, data$GDP_log,
     xlab="log(Electricity consumption)", ylab="log(GDP)", main="Model 3")
abline(reg=model3, col="darkgreen")
```



For Model 3, the estimate of $\hat{\beta}_0$ is 0.47269 and the estimate of $\hat{\beta}_1$ is 1.01603 meaning that when the value of electricity consumption increases by 1%, the value of GDP increases by 1.01603%, **on average**.

This leads us to conclude that the equation $\gamma_1 = \frac{1}{\beta_1}$ is valid only when a perfect linear relationship exists between observed variables.

The proof of its restriction:

$$\frac{s_{\log x}}{s_{\log y}} r_{\log x \log y} = \frac{s_{\log x}}{s_{\log y} r_{\log x \log y}}$$

$$r^2_{\log x \log y} = 1$$

which is valid only if

$$r_{\log x \log y} = \pm 1$$