Hypothesis testing and inference in Ordinary Least Squares: An application of the Gravitation Model of Trade

Juan David Rincón Mora

2023

1. Introduction.

The main objective of this notebook is to estimate a model of economic interest using the Ordinary Least Squares (OLS) method and to **make inferences about the estimated parameters**. This document will describe in detail the process of constructing statistics, variables, and hypothesis testing, as well as the interpretation of the results obtained.

It is important to emphasize that this notebook focuses on the development of the different estimators and statistics necessary for the estimation and inference of the econometric model in terms of the programming framework of R. The mathematical development behind them, beyond presenting the equations, formulas, and corresponding interpretations, will not be discussed in depth.

Some sections use functions that were introduced and programmed in a preceding notebook: *"An introductory guide to Ordinary Least Squares: Applying CAPM to Amazon stock"*; which are contained in the *OLS Function Pack #1*. Therefore, it is necessary to run that file or functions before running the rest of the code in this notebook.

Finally, it is relevant to mention that the mathematical, econometric, and statistical content presented in this notebook is largely based on the approach and methodology proposed by Judge, Hill, Griffiths, Lütkepohl, and Lee (1988).

2. Model.

In this notebook it will be estimated the **Gravitational Model of Trade**, which aims to explain the trade flow between two economies based on Newton's model of gravitation. According to this model, two objects attract each other with a force proportional to their masses and inversely proportional to the square of the distance between them. The mathematical equation describing this Newtonian model is as follows:

$$F = G \cdot \frac{m_1 \cdot m_2}{d^2}$$

Where F is the force of attraction between objects, G is the universal gravitational constant, m_1 and m_2 are the masses of the objects, and d is the distance between the objects. The gravitational model of trade, inspired by Newton's model, states that: *trade between two countries is influenced by the size of their economies and the distance that separates them.* The mathematical formulation of this model, based on Krugman and Obstfeld (2006), is as follows:

$$T_{ij} = A \cdot \frac{Y_i^a \cdot Y_j^b}{D_{ij}^c}$$

Where T_{ij} is the trade flow between two economies i and j measured as the sum of imports and exports, A is an unknown constant, Y_i and Y_j represent the GDP of economies i and j, respectively, and D_{ij} is the distance between the two economies. The parameters a, b, and c are expected to take the theoretical values of 1, 1, and 2, respectively. However, their indeterminacy allows for estimation to find the best fit for the data and real-world observations.

To facilitate the estimation using the ordinary least squares method, a logarithmic transformation of the mathematical model is performed. In addition, a perturbation term ε_{ij} is introduced to account for all unobservable factors affecting trade flow. The final model, in its linearized form, is expressed as follows:

$$log(T_{ij}) = log(A) + a \cdot log(Y_i) + b \cdot log(Y_j) - c \cdot log(D_{ij}) + \varepsilon_{ij}$$

To simplify the notation of the model, the following replacements are made:

$$\beta_0 = log(A), \quad \beta_1 = a, \quad \beta_2 = b, \quad \beta_3 = -c$$

On the other hand, Krugman and Obstfeld (2006) point out that the gravity model of trade may not capture certain particularities of international trade. Therefore, in this case it is introduced a dummy variable called frt_{ij}, which indicates whether the two economies share a common frontier (land border). This variable takes the value 1 if there is a shared frontier and 0 otherwise.

Taking into account the notation replacements and the new variable, the final model is expressed as follows:

$$log(C_{ij}) = \beta_0 + \beta_1 log(Y_i) + \beta_2 log(Y_j) + \beta_3 log(D_{ij}) + \beta_4 frt_{ij} + \varepsilon_{ij}$$

It is essential to note that this model is constructed using **cross-sectional data** since it considers information from economies i and j at a specific point in time.

2.1. Variables.

In this case, the gravitational model of trade is estimated on the trade interaction among the kingdoms of the fictional continent of Westeros from George R. R. Martin's "Song of Ice and Fire" and "Game of Thrones" universe (see Figure 1). Therefore, the variables are fully simulated. Specific details on how the simulations were performed can be found in Appendix 1.

Figure 1. Map of Westeros. Source: imgur.com

The most important information from the simulation is the following:

1. The Trade Flow variable T_{ij} and the respective GDPs, represented by Y_i and Y_j, are measured in millions of gold ingots hypothetically.
2. The variable frt_{ij}, which indicates whether or not the regions share a frontier, in this particular case, does not influence the model, so its coefficient is zero. However, it is included in the analysis to evaluate its statistical significance.
3. The population values of the parameters that effectively influence the model correspond to the theoretical values, and the unknown constant was set as $A = 3$.

Conducting this econometric exercise with simulated data is useful because it allows to compare the estimates obtained with the actual population values and determine whether they converge or are similar. Therefore, statistical and econometric inference techniques are employed.

The database used for this exercise is available at the following link: *Westeros Simulated Data*. The first values of the database are presented below:

```
#install.packages("readxl")
library(readxl)

# Import data.
DATA <- as.data.frame(read_excel(file.choose()))

# First values of data.
head(DATA)
```

```
##   Country A    Country B  Trade Flow GDP A  GDP B Distance Frontier
## 1 The North Vale of Aryn  301.054791  1093  5117      463        0
## 2 The North   Riverlands  136.904370  1093  2421      541        1
## 3 The North Iron Islands    8.979869  1093   267      531        0
## 4 The North  Westerlands  273.577482  1093  7697      704        0
## 5 The North   Crownlands 1179.569643  1093  4789      721        0
## 6 The North    The Reach   32.454623  1093  6820      917        0
```

2.1.1. Construction of model variables.

According to the econometric model, the variables are extracted from the database and a logarithmic transformation is applied to them, except in the case of the frontier dummy variable.

```
# Log. Trade Flow
log_tf <- log(DATA$`Trade Flow`)

# Log. Region A GDP (i)
log_yi <- log(DATA$`GDP A`)

# Log. Region B GDP (j)
log_yj <- log(DATA$`GDP B`)

# Log. Distance
log_dist <- log(DATA$Distance)

# Dummy Frontier.
front <- DATA$Frontier
```

3. OLS Estimation.

The following model is estimated using the Ordinary Least Squares (OLS) method in matrix form:

$$Y = X\beta + e$$

3.1. Design matrices.

```
# Y Matrix (Dependent variable).
Y <- as.matrix(log_tf)

# X Matrix (Independent variable).
X <- cbind(1, log_yi, log_yj, log_dist, front)

# Number of observations and variables.
N <- dim(X)[1]
K <- dim(X)[2]
```

3.2. Estimated coefficients.

```
# Estimated Betas.
B_ht <- Beta_OLS.f(X,Y)
B_ht
```

```
##                [,1]
##          -1.8784260
## log_yi    1.1227350
## log_yj    0.9804375
## log_dist -1.5288358
## front    -0.5560856
```

The interpretations of the estimated betas are as follows:

- $\hat{\beta}_1$: Given a 1% increase in the GDP of region i, the trade flow between regions i and j increases by approximately **1.12%**, holding everything else constant.
- $\hat{\beta}_2$: Given a 1% increase in the GDP of region j, the trade flow between regions i and j increases by approximately **0.98%**, holding everything else constant.
- $\hat{\beta}_3$: Given a 1% increase in the distance between regions i and j, the trade flow between them decreases by approximately **1.53%**, holding everything else constant.
- $\hat{\beta}_4$: If regions i and j share a frontier, the trade flow between them decreases by approximately **0.56%**, holding everything else constant.

Among these results, the coefficient of the frontier variable stands out as not intuitive. According to theory, it would be logical to expect that when two regions share a frontier, the trade flow between them would increase due to the ease of transportation and communication between regions. However, in this case, the model indicates a relationship opposite to what is expected.

3.3. Dependent variable estimation.

$$\hat{Y} = X\hat{\beta}$$

```
# Estimated dependent variable.
Y_ht <- X%*%B_ht
head(Y_ht)
```

```
##          [,1]
## [1,] 4.966670
## [2,] 3.438809
## [3,] 1.861860
## [4,] 4.726285
## [5,] 4.224579
## [6,] 4.203564
```

3.4. Errors estimation.

$$\hat{e} = Y - \hat{Y}$$

```
# Estimated errors.
e_ht <- Y - Y_ht
head(e_ht)
```

```
##            [,1]
## [1,]  0.7406225
## [2,]  1.4804736
## [3,]  0.3331249
## [4,]  0.8853001
## [5,]  2.8483258
## [6,] -0.7237214
```

3.5. Estimated variance of errors.

$$\hat{\sigma}^2 = \frac{\hat{e}' \cdot \hat{e}}{T - K}$$

```
# Estimated Variance.
sigma2_ht <- sigma2_ht.f(e_ht, N, K)
sigma2_ht
```

```
## [1] 1.778371
```

```
# Estimated Standard Deviation.
sigma_ht <- sqrt(sigma2_ht)
sigma_ht
```

```
## [1] 1.333556
```

3.6. Variance-Covariance matrix of betas.

$$\sum_{\hat{\beta}} = \text{VarCov}(\hat{\beta}) = \hat{\sigma}^2 \left(X'X\right)^{-1}$$

```
# Variance-Covariance Matrix.
varcov_beta_ht <- sigma2_ht*solve(t(X)%*%X)
varcov_beta_ht
```

```
##                          log_yi       log_yj     log_dist       front
##           19.7009615 -0.454802100 -0.401899112 -2.121918917 -0.71619730
## log_yi   -0.4548021  0.047391016  0.008767665  0.006081651 -0.03428764
## log_yj   -0.4018991  0.008767665  0.040774594  0.005781605 -0.02331682
## log_dist -2.1219189  0.006081651  0.005781605  0.324886194  0.17069378
## front    -0.7161973 -0.034287636 -0.023316819  0.170693780  0.35141341
```

```
# Standard deviation of betas.
sd_beta_ht <- sqrt(diag(varcov_beta_ht))
sd_beta_ht
```

```
##              log_yi    log_yj  log_dist     front
## 4.4385765 0.2176948 0.2019272 0.5699879 0.5928013
```

3.7. Coefficient of determination.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{t=1}^{T} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2}$$

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{t=1}^{T} (\hat{e}_t)^2}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2}$$

```
# Coefficient of determination R2.
R2 <- R2.f(Y, Y_ht)
R2
```

```
## [1] 0.66804
```

This result indicates that approximately **66.8%** of the variability of the logarithm of the trade flow between two regions in Westeros is explained by the variability of the model regressors.

3.7.1. Adjusted coefficient of determination.

$$R^2_{adj} = 1 - \frac{T-1}{T-k-1}(1 - R^2)$$

```
# Adjusted R2 Coefficient of determination.
R2adj <- R2adj.f(R2, N, K)
R2adj
```

```
## [1] 0.6252064
```

This result indicates that, adjusted for the number of regressors, approximately **62.5%** of the variability of the logarithm of the trade flow between two regions in Westeros is explained by the variability of the model regressors.

4. Hypothesis Testing.

Once the results of the initial estimation have been obtained, it is important to conduct hypothesis testing about the estimated parameters and values to evaluate the global significance of the model, the individual significance of the variables, and to verify whether the estimates of the econometric model agree with those of the theoretical model.

4.1. Normality test on residuals.

Before carrying out hypothesis testing and model inference, it is crucial to verify if the model errors or residuals follow a normal distribution. This check is a prerequisite for proceeding with the inference process, since, according to theory, the underlying distribution of the errors is transferred to the dependent random variable and the parameters. This allows to make inferences about them using statistics constructed under the assumption of normality.

$$e \sim \mathcal{N}(0, \sigma^2 I_T)$$
$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_T)$$

Moreover, when the errors follow a normal distribution, the Maximum Likelihood estimators coincide with the Ordinary Least Squares estimators. Therefore, they share the properties of the Gauss-Markov Theorem.

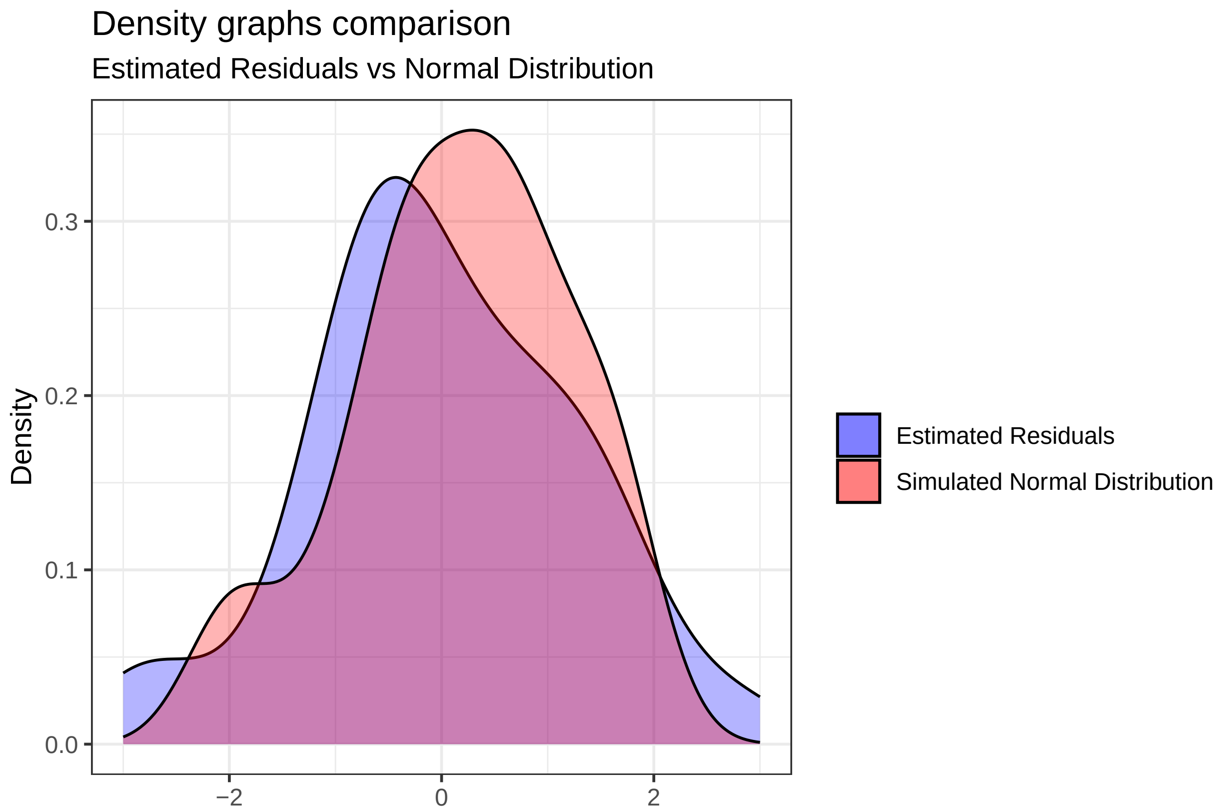
$$\hat{\beta} = \tilde{\beta}$$
$$\tilde{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$$

A first approach to verify if the residuals follow a normal distribution is to examine their density plot and compare it with the density plot of a normal distribution with the same mean and standard deviation. In this way, it can be visually assessed if the two plots coincide and provide an indication of the normality of the residuals.

```r
#install.packages("ggplot2")
library(ggplot2)

# Provisional DataFrame.
df1 <- data.frame(e_ht = e_ht, norm_e = rnorm(N, 0, sigma_ht))

# Density graphs.
ggplot(data = df1) +
  geom_density(aes(x = e_ht, fill = "Estimated Residuals"), alpha = 0.3) +
  geom_density(aes(x = norm_e, fill = "Simulated Normal Distribution"), alpha = 0.3) +
  scale_fill_manual(values = c("blue", "red")) +
  guides(fill = guide_legend(title = "")) +
  ggtitle("Density graphs comparison",
          subtitle = "Estimated Residuals vs Normal Distribution") +
  xlab("") + ylab("Density") + theme_bw() + xlim(-3, 3)
```



In this case, it can be observed that the density plots do not coincide perfectly. However, it should be noted that, depending on the randomness of the normal distribution used for the contrast, the plots may be more or less similar. Therefore, to obtain a more accurate assessment of the normality of the residuals, a parametric test must be performed to provide a more robust contrast.

To perform the normality test, in this case, it is used the Jarque-Bera test. This goodness-of-fit test evaluates if a variable follows a normal distribution based on its skewness and kurtosis. The Jarque-Bera (JB) statistical test is defined as follows:

$$JB = \frac{N}{6}\left(\mathrm{Skw}^2 + \frac{(\mathrm{Krt}-3)^2}{4}\right) \sim \chi^2_{(2)}$$

Where Skw is the **asymmetry coefficient (skewness)**, calculated as follows:

$$\mathrm{Skw} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^{3/2}}$$

And Krt is the **kurtosis coefficient**, calculated as follows:

$$\mathrm{Krt} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^{2}}$$

Here N is the total number of data, and x is the random variable being evaluated. The Jarque-Bera statistic follows a **chi-squared distribution** with 2 degrees of freedom.

The hypothesis test for normality is presented as follows:

$$H_0 : e_i \sim \mathcal{N}(0,\sigma^2)$$
$$H_1 : e_i \nsim \mathcal{N}(0,\sigma^2)$$

The null hypothesis states that the errors follow a normal distribution, and the alternative hypothesis states the opposite. The rejection rule for such hypothesis testing is as follows:

$$H_0 \text{ is not rejected if } JB < \chi^2_{(2,\ 1-\alpha)}$$
$$H_0 \text{ is rejected if } JB \geq \chi^2_{(2,\ 1-\alpha)}$$

Where JB is the value of the Jarque-Bera statistic and $\chi^2_{(2,\ 1-\alpha)}$ is the critical value of a chi-squared distribution with 2 degrees of freedom at a given significance level α (1%, 5%, 10%).

On the other hand, the null hypothesis can also be rejected if the **p-value** of the statistic is less than or equal to the chosen level of significance to be evaluated. This approach through p-value applies to any statistical hypothesis test, not only in this particular case.

```r
# Jarque-Bera Test function.
Jarque_Bera.f <- function(data, alpha = 0.05){

  # Length of data.
  N <- length(data)

  # Mean of data.
  mean1 <- mean(data)

  # Skewness.
  Skw <- (sum((data-mean1)^3)/N)/(sum((data-mean1)^2)/N)^(3/2)

  # Kurtosis.
  Krt <- (sum((data-mean1)^4)/N)/(sum((data-mean1)^2)/N)^(4/2)

  # Formula.
  JB <- (N/6)*((Skw^2)+(((Krt-3)^2)/4))

  # p-value.
  pValue <- 1-pchisq(JB, 2)

  # Hypothesis testing.
  if(JB < qchisq(alpha, 2, lower.tail = FALSE)){
  hp <- "It is not rejected the hypothesis that residuals follow a normal distribution"
} else{
  hp <- "It is rejected the hypothesis that residuals follow a normal distribution"
}

  # Result.
  return(list(JB=JB, hp=hp, pValue=pValue))
}

# JB statistic.
Jarque_Bera.f(e_ht)$JB
```

```
## [1] 0.02695773
```

```r
# JB statistic p-value.
Jarque_Bera.f(e_ht)$pValue
```

```
## [1] 0.9866116
```

```r
# Hypothesis testing result (5%).
Jarque_Bera.f(e_ht)$hp
```

```
## [1] "It is not rejected the hypothesis that residuals follow a normal distribution"
```

Since the hypothesis of normality of the errors has not been rejected, it can proceed with the inference of the other parameters of the model. However, before going into the specific hypothesis tests, it is important to introduce and explain the test statistics that will be used throughout these assessments.

4.2. Test and inference statistics.

4.2.1. Restricted Maximum Likelihood.

The restricted maximum likelihood estimator, denoted as β^*, is obtained by including an adjustment to the model coefficients that are estimated through unrestricted maximum likelihood (or ordinary least squares if the normality hypothesis is not rejected). This adjustment is made by applying fixed linear restrictions, which can be known with total certainty or used to evaluate one or more hypotheses, either individually or jointly. The definition of the estimator is as follows:

$$\beta^* = \tilde{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\tilde{\beta})$$

In this formula, it is considered an R matrix of dimension $J \times K$, allowing to select the parameters subject to constraints, where J represents the total number of constraints evaluated. In addition, it is defined an r matrix of dimensions $J \times 1$ that contains the values corresponding to the constraints associated with each parameter or linear combination of parameters.

```r
# Restricted Beta function.
rBeta.f <- function(beta_ols, X, R, r){

  # Adjustment parts.
  partA <- solve(t(X)%*%X)%*%t(R)
  partB <- solve(R%*%solve(t(X)%*%X)%*%t(R))
  partC <- r - R%*%beta_ols

  # Formula with adjustment.
  rBeta <- beta_ols + partA%*%partB%*%partC

  # Result.
  return(rBeta)
}
```

4.2.2. Lagrange Multipliers Test.

The Lagrange Multipliers Test, denoted as λ_1, is the first of the three classical statistics used to test joint hypothesis tests. In particular, this statistic quantifies the difference between the sample data and the restrictions to be applied, allowing to compare jointly all the available information. The formula of the statistic is as follows:

$$\lambda_1 = \frac{(R\tilde{\beta} - r)' \cdot [R(X'X)^{-1}R']^{-1} \cdot (R\tilde{\beta} - r)}{J\hat{\sigma}^2} \sim F_{(J,\, T-K)}$$

The statistic follows an **F-distribution** with J and $(T-K)$ degrees of freedom. It is important to note that the formula may have a slight variation. Instead of placing $\hat{\sigma}^2$ in the denominator, it is possible to place it in the numerator and combine it with the expression $(X'X)^{-1}$, which jointly represents the variance-covariance matrix of the coefficients.

The general hypothesis that applies to the test, no matter what the specific linear constraints are, is stated as follows:

$$H_0 : R\beta = r$$
$$H_1 : R\beta \neq r$$

The null hypothesis states that all linear constraints are satisfied jointly, while the alternative hypothesis states that at least one constraint is not satisfied together with the others. In either case, the rule for rejecting the hypothesis is as follows:

$$H_0 \text{ is not rejected if } \lambda_1 < F_{(J,\, T-K,\, 1-\alpha)}$$
$$H_0 \text{ is rejected if } \lambda_1 \geq F_{(J,\, T-K,\, 1-\alpha)}$$

Alternatively, when there is only one constraint ($J = 1$), the statistic may have a different transformation. Because only one constraint is evaluated, products between matrices are reduced to operations between real numbers. Moreover, statistically it is satisfied that $F_{(1,\, T-K)} = t^2_{(T-K)}$. This allows that, in the case of a single constraint, the statistic λ_1 can be rewritten as follows:

$$\sqrt{\lambda_1} = \frac{R_1\tilde{\beta} - r_1}{\sqrt{R_1 \cdot \hat{\sigma}^2 (X'X)^{-1} \cdot R_1'}} \sim t_{(T-K)}$$

In this case, R_1 and r_1 represent the selection vector and the value of the single hypothesis constraint, respectively. Now it is evaluated the statistic $\sqrt{\lambda_1}$, which follows a **t-distribution** with $(T - K)$ degrees of freedom. Because of this, the rule for rejecting the general null hypothesis, which remains the same, is equivalent to a two-tailed t-test.

$$H_0 \text{ is not rejected if } \left|\sqrt{\lambda_1}\right| < t_{(T-K,\, 1-\alpha/2)}$$
$$H_0 \text{ is rejected if } \left|\sqrt{\lambda_1}\right| \geq t_{(T-K,\, 1-\alpha/2)}$$

```r
# Lagrange Multipliers Test function.
lambda1_LM.f <- function(R, r, beta, varcov_beta, T_obs, alpha = 0.05){

  # Number of constraints.
  J <- length(r)

  # Numerator parts.
  partA <- t(R%*%beta - r)
  partB <- solve(R%*%varcov_beta%*%t(R))
  partC <- t(partA)

  # Formula.
  lambda1 <- as.numeric((partA%*%partB%*%partC)/J)

  # p-value.
  pValue <- 1-pf(lambda1, J, T_obs - dim(beta)[1])

  # Hypothesis testing.
  if (lambda1 < qf(1-alpha, J, T_obs - dim(beta)[1])){
    hp <- "The null hypothesis is not rejected"
  } else{
    hp <- "The null hypothesis is rejected"
  }
```

```r
  # Result.
  return(list(lambda1 = lambda1, hp = hp, pValue=pValue))
}
```

4.2.3. Likelihood Ratio Test.

The Likelihood Ratio Test, denoted as λ_2, is the second of the three classical statistics used to test joint hypothesis tests. In particular, this statistic quantifies the difference between the residuals of the unrestricted OLS model and the residuals restricted to the hypotheses to be evaluated. The formula of the statistic is as follows:

$$\lambda_2 = \frac{\text{SSEr} - \text{SSEnr}}{J\hat{\sigma}^2} \sim F_{(J,\, T-K)}$$

Where SSEr and SSEnr represent the sum of squared errors for the restricted and unrestricted model, respectively. The unrestricted errors are obtained in the same way as in the OLS method, but using the previously defined restricted maximum likelihood estimator β^*.

It is important to note that the values of λ_1 and λ_2 are the same when evaluating the same hypothesis, i.e., the same constraints. Moreover, the rule for rejecting the null hypothesis is the same as in the Lagrange Multipliers Test.

$$H_0 \text{ is not rejected if } \lambda_2 < F_{(J,\, T-K,\, 1-\alpha)}$$
$$H_0 \text{ is rejected if } \lambda_2 \geq F_{(J,\, T-K,\, 1-\alpha)}$$

```r
# Likelihood Ratio Test function.
lambda2_LR.f <- function(X, SSEr, SSEnr, var_ht, J, alpha = 0.05){

  # Formula.
  lambda2 <- as.numeric((SSEr - SSEnr)/(var_ht*J))

  # p-value.
  pValue <- 1-pf(lambda2, J, dim(X)[1] - dim(X)[2])

  # Hypothesis testing.
  if (lambda2 < qf(1-alpha, J, dim(X)[1] - dim(X)[2])){
    hp <- "The null hypothesis is not rejected"
  } else{
    hp <- "The null hypothesis is rejected"
  }

  # Result.
  return(list(lambda2 = lambda2, hp = hp, pValue=pValue))
}
```

4.2.4. Wald Test.

The Wald test, denoted as λ_3, is the third of the three classical statistics used to test joint hypothesis tests. This statistic quantifies the difference between the coefficients obtained from the initial OLS estimation and the restricted coefficients according to the hypothesis to be tested. The formula of the statistic is as follows:

$$\lambda_3 = \frac{(\beta^* - \tilde{\beta})' \cdot (X'X) \cdot (\beta^* - \tilde{\beta})}{J\hat{\sigma}^2} \sim F_{(J,\, T-K)}$$

Again, it is important to mention that the values of λ_1, λ_2, and λ_3 are identical when evaluating the same hypothesis. Also, the rule for rejecting the null hypothesis is the same as in the previous tests.

$$H_0 \text{ is not rejected if } \lambda_3 < F_{(J,\, T-K,\, 1-\alpha)}$$
$$H_0 \text{ is rejected if } \lambda_3 \geq F_{(J,\, T-K,\, 1-\alpha)}$$

```r
# Wald test function.
lambda3_WT.f <- function(rBeta, beta_ols, X, var_ht, J, alpha = 0.05){

  # Formula.
  lambda3 <- (t(rBeta - beta_ols)%*%(t(X)%*%X)%*%(rBeta - beta_ols))/(var_ht*J)
  lambda3 <- as.numeric(lambda3)

  # p-value.
  pValue <- 1-pf(lambda3, J, dim(X)[1] - dim(X)[2])

  # Hypothesis testing.
  if (lambda3 < qf(1-alpha, J, dim(X)[1] - dim(X)[2])){
    hp <- "The null hypothesis is not rejected"
  } else{
    hp <- "The null hypothesis is rejected"
  }

  # Result.
  return(list(lambda3 = lambda3, hp = hp, pValue=pValue))
}
```

Once presented and explained the statistics necessary to perform joint hypothesis testing in the econometric context, it is possible to continue with the process of inference on the model.

4.3. Global significance.

The first test to be performed refers to the global significance of the model, that is, to verify whether the independent variables collectively are statistically significant in explaining the dependent variable. The hypotheses proposed for this test are the following:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_1 : \text{At least one is not satisfied}$$

In the case that the null hypothesis is rejected, it means that at least one variable in the model has a significant effect on the dependent variable. On the other hand, if the null hypothesis is not rejected, this indicates that there is no evidence in the data to affirm that any of the variables evaluated are significant.

For pedagogical purposes, all three lambda tests will be used to perform the global significance test. However, **using only one test is sufficient**, because of what was previously mentioned: all three tests lead to the same conclusion and yield the same result.

Lagrange Multipliers Test.

To carry out the hypothesis test using the Lagrange Multipliers Test λ_1, it is necessary to initially design the selection matrix R and the constraint vector r. Considering the null and alternative hypothesis established for the global significance test, the matrices are defined as follows:

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{4\times 5} \qquad r = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{4\times 1}$$

```r
# Selection matrix of constraints.
R_gs <- cbind(0, diag(1, nrow=K-1))

# Constraints value vector.
r_gs <- as.matrix(rep(0, K-1))
```

Once these matrices are available, the other necessary inputs have already been calculated. Therefore, it is simply a matter of calculating the corresponding statistic.

```r
# Lambda 1 value.
lambda1_LM.f(R_gs, r_gs, B_ht, varcov_beta_ht, N)$lambda1
```

```
## [1] 15.59618
```

```r
# Lambda 1 p-value.
lambda1_LM.f(R_gs, r_gs, B_ht, varcov_beta_ht, N)$pValue
```

```
## [1] 4.285508e-07
```

```r
# Hypothesis testing.
lambda1_LM.f(R_gs, r_gs, B_ht, varcov_beta_ht, N)$hp
```

```
## [1] "The null hypothesis is rejected"
```

Likelihood Ratio Test.

To carry out the hypothesis test using the Likelihood Ratio Test λ_2, it is required to initially calculate the restricted maximum likelihood estimator β^* using the previously designed R and r matrices.

```r
# Restricted estimator.
rBeta_gs <- rBeta.f(B_ht, X, R_gs, r_gs)
rBeta_gs
```

```
##                  [,1]
##          4.637129e+00
## log_yi   0.000000e+00
## log_yj   0.000000e+00
## log_dist 0.000000e+00
## front    3.330669e-16
```

Subsequently, the restricted residuals are estimated using these coefficients and the sum of these residuals squared, as well as the unrestricted residuals.

```r
# Sum of restricted squared errors.
SSEr_gs <- sum((Y - X%*%rBeta_gs)^2)

# Sum of unrestricted squared errors.
SSEnr_gs <- sum(e_ht^2)
```

Once all the necessary inputs are available, it is possible to proceed to calculate the corresponding statistic.

```r
# Lambda 2 value.
lambda2_LR.f(X, SSEr_gs, SSEnr_gs, sigma2_ht, K-1)$lambda2
```

```
## [1] 15.59618
```

```r
# Lambda 2 p-value.
lambda2_LR.f(X, SSEr_gs, SSEnr_gs, sigma2_ht, K-1)$pValue
```

```
## [1] 4.285508e-07
```

```r
# Hypothesis testing.
lambda2_LR.f(X, SSEr_gs, SSEnr_gs, sigma2_ht, K-1)$hp
```

```
## [1] "The null hypothesis is rejected"
```

Wald test.

In this instance, all the necessary inputs for the statistic were already calculated previously, therefore, it is only necessary to apply the statistic.

```r
# Lambda 3 value.
lambda3_WT.f(rBeta_gs, B_ht, X, sigma2_ht, K-1)$lambda3
```

```
## [1] 15.59618
```

```r
# Lambda 3 p-value.
lambda3_WT.f(rBeta_gs, B_ht, X, sigma2_ht, K-1)$pValue
```

```
## [1] 4.285508e-07
```

```r
# Hypothesis testing.
lambda3_WT.f(rBeta_gs, B_ht, X, sigma2_ht, K-1)$hp
```

```
## [1] "The null hypothesis is rejected"
```

After performing the hypothesis test using the three classic tests, it is confirmed that, as mentioned above, they all coincide in the conclusion and provide the same result.

As for the global significance test, the three statistics indicate that, **with a confidence level of 95%, the null hypothesis of global non-significance is rejected**. Furthermore, the p-value is very close to 0, indicating the rejection of the null hypothesis up to a 99% confidence level. Therefore, it is concluded that among the independent variables of the model, there is at least one that is significant in explaining the dependent variable, in this case, trade flow.

Taking this into account, the next step is to perform the corresponding individual significance tests for each variable and determine if there is any variable that is not significant for the model.

4.4. Individual significance.

Two approaches will be employed for conducting individual significance tests: first, through individual hypothesis tests using t-statistics; and second, by constructing the respective confidence intervals for each coefficient.

4.4.1. Individual hypothesis testing.

For the individual significance test using hypothesis testing, the procedure begins by calculating the following statistic:

$$t(\beta_k) = \frac{\tilde{\beta}_k - \beta_k}{\sqrt{Var(\tilde{\beta}_k)}} \sim t_{(T-K)}$$

Where $\tilde{\beta}_k$ is the k-th estimated coefficient to be evaluated, β_k is the population value of the k-th coefficient to be tested in the hypothesis, and $Var(\tilde{\beta}_k)$ is the estimated variance of the k-th coefficient. The statistic follows a t-distribution with $(T-K)$ degrees of freedom.

It is worth mentioning that this statistic is identical to the already mentioned case of the Lagrange Multipliers Test when a single constraint is evaluated. Therefore, the individual significance test can also be performed under that method.

In particular, the hypothesis test being performed is as follows:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

The null hypothesis states that the k-th coefficient is not significant, while the alternative hypothesis states that it has some non-zero magnitude. The rejection rule for the null hypothesis is the same as a two-tailed t-test:

$$H_0 \text{ is not rejected if } |t(\beta_k)| < t_{(T-K,\, 1-\alpha/2)}$$
$$H_0 \text{ is rejected if } |t(\beta_k)| \geq t_{(T-K,\, 1-\alpha/2)}$$

In this case, the significance test will be performed under a confidence level of 95% (significance α of 5%), however, the p-value is calculated to be able to contrast under any significance level.

```r
# Individual Significance function.
ind.sig.f <- function(betas_ols, sd_betas, T_obs, alpha = 0.05){

  # Betas significance data frame.
  sigDF <- data.frame(Beta = row.names(betas_ols)[-1],
                      tValue = NA, pValue = NA, Hypothesis = NA)

  # Betas iteration.
  K <- length(betas_ols)
  for(k in 2:K){

    # t-statistic.
    tValue_beta <- betas_ols[k]/sd_betas[k]

    # p-value.
    pValue_beta <- 2*(1-pt(abs(tValue_beta), T_obs-K))

    # Hypothesis rejection rule.
    if (abs(tValue_beta) < qt(1-(alpha/2), T_obs-K)){
      hp <- paste("Beta", k-1, "is not statistically significative")
    } else{
      hp <- paste("Beta", k-1, "is statistically significative")
    }

    # Fill data frame.
    sigDF[k-1, 2:4] <- c(round(tValue_beta, 3),
                         round(pValue_beta, 3), hp)
```

```r
  }

  # Hypothesis testing results.
  return(sigDF)
}

# Individual significance results.
ind.sig.f(B_ht, sd_beta_ht, N)
```

```
##        Beta tValue pValue                              Hypothesis
## 1    log_yi  5.157      0     Beta 1 is statistically significative
## 2    log_yj  4.855      0     Beta 2 is statistically significative
## 3 log_dist -2.682  0.012     Beta 3 is statistically significative
## 4     front -0.938  0.355 Beta 4 is not statistically significative
```

The results of the individual hypothesis tests indicate that, at a 95% confidence level, all variables in the model, except the frontier dummy variable, are significant in explaining the trade flow between the different regions of Westeros.

4.4.2. Confidence Intervals.

The way to construct the confidence interval at a certain significance level, for any estimated coefficient, is based on the following:

$$Pr\left[\tilde{\beta}_k - t_{(T-K,\,1-\alpha/2)} \cdot \hat{\sigma}(\tilde{\beta}_k) \leq \beta_k \leq \tilde{\beta}_k + t_{(T-K,\,1-\alpha/2)} \cdot \hat{\sigma}(\tilde{\beta}_k)\right] = 1-\alpha$$

Where $t_{(T-K,\,1-\alpha/2)}$ is the critical value that accumulates $(1-\alpha/2)$ probability in a t-distribution with $(T-K)$ degrees of freedom, and $\hat{\sigma}(\tilde{\beta}_k)$ is the estimated standard deviation of the k-th coefficient.

The way by which it is possible to identify through this method if the coefficient (associated with a variable) is significant or not, lies in observing if within the calculated interval the number zero is found. If this is the case, it means that the null value is among the possible values that the coefficient can take at that confidence level.

The confidence intervals will be constructed at both 95% and 99% confidence levels.

```r
# Confidence intervals function.
conf.int.f <- function(betas_ols, sd_betas, T_obs, alpha = 0.05){

  # Confidence intervals data frames.
  CIdf <- data.frame(Beta = row.names(betas_ols)[-1],
                     Left.Limit = NA, Right.Limit = NA,
                     Conclusion = NA)

  # Betas iteration.
  K <- length(betas_ols)
  for(k in 2:K){

    # Interval limits.
    left <- betas_ols[k] - qt(1-(alpha/2), T_obs-K)*sd_betas[k]
    right <- betas_ols[k] + qt(1-(alpha/2), T_obs-K)*sd_betas[k]

    # Conclusion.
```

```
    if(left <= 0 & 0 <= right){
      conc <- paste("Beta", k-1, "is not statistically significative")
    } else{
      conc <- paste("Beta", k-1, "is statistically significative")
    }

    # Fill data frames.
    CIdf[k-1, 2:4] <- c(round(left, 3), round(right, 3), conc)
  }

  # Result.
  return(CIdf)
}

# Confidence intervals (95%).
conf.int.f(B_ht, sd_beta_ht, N, alpha = 0.05)
```

```
##       Beta Left.Limit Right.Limit                              Conclusion
## 1   log_yi      0.679       1.567     Beta 1 is statistically significative
## 2   log_yj      0.569       1.392     Beta 2 is statistically significative
## 3 log_dist     -2.691      -0.366     Beta 3 is statistically significative
## 4    front     -1.765       0.653 Beta 4 is not statistically significative
```

```
# Confidence intervals (99%).
conf.int.f(B_ht, sd_beta_ht, N, alpha = 0.01)
```

```
##       Beta Left.Limit Right.Limit                              Conclusion
## 1   log_yi      0.525        1.72     Beta 1 is statistically significative
## 2   log_yj      0.426       1.535     Beta 2 is statistically significative
## 3 log_dist     -3.093       0.035 Beta 3 is not statistically significative
## 4    front     -2.183       1.071 Beta 4 is not statistically significative
```

Based on the confidence intervals, the following conclusions can be drawn:

- Both variables $log(y_i)$ and $log(y_j)$ demonstrate statistical significance in explaining trade flow at both 95% and 99% confidence levels.
- The variable $log(d_{ij})$ appears significant in explaining trade flow at a 95% confidence level, but it marginally loses significance at the 99% confidence level.
- The variable frt_{ij}, associated with the shared frontier between regions, lacks statistical significance at both the 95% and 99% confidence levels.

Considering these conclusions, firstly, the variable $log(d_{ij})$ will remain in the model. It is one of the original theoretical variables, and its lack of significance occurs marginally only at the 99% confidence level. However, the variable frt_{ij} will be removed from the model. Its lack of significance was rejected at both the 95% and 99% confidence intervals. Furthermore, the individual hypothesis test also rejected its significance, with a relatively high p-value indicating a strong inclination toward non-significance.

Consequently, the gravitational model, after conducting both global and individual significance tests, is defined as follows:

$$log(T_{ij}) = \beta_0 + \beta_1 log(Y_i) + \beta_2 log(Y_j) + \beta_3 log(D_{ij}) + \varepsilon_{ij}$$

Now it is necessary to re-estimate the model, this time including only the remaining variables.

4.5. Corrected OLS Estimation.

Design matrix.

```
# New X matrix.
X <- cbind(1, log_yi, log_yj, log_dist)
head(X)
```

```
##          log_yi   log_yj log_dist
## [1,] 1 6.996681 8.540324 6.137727
## [2,] 1 6.996681 7.791936 6.293419
## [3,] 1 6.996681 5.587249 6.274762
## [4,] 1 6.996681 8.948586 6.556778
## [5,] 1 6.996681 8.474077 6.580639
## [6,] 1 6.996681 8.827615 6.821107
```

```
# New number of variables.
K <- dim(X)[2]
```

OLS Betas.

```
# New betas.
B_ht <- Beta_OLS.f(X, Y)
B_ht
```

```
##                [,1]
##          -3.0117550
## log_yi    1.0684774
## log_yj    0.9435403
## log_dist -1.2587256
```

Dependent variable estimation.

```
# Estimated dependent variable.
Y_ht <- X%*%B_ht
head(Y_ht)
```

```
##          [,1]
## [1,] 4.796467
## [2,] 3.894359
## [3,] 1.837632
## [4,] 4.654208
## [5,] 4.176456
## [6,] 4.207349
```

Errors estimation.

```
# Estimated errors.
e_ht <- Y - Y_ht
head(e_ht)
```

```
##             [,1]
## [1,]  0.9108255
## [2,]  1.0249236
## [3,]  0.3573534
## [4,]  0.9573768
## [5,]  2.8964494
## [6,] -0.7275062
```

Estimated variance.

```r
# Estimated Variance.
sigma2_ht <- sigma2_ht.f(e_ht, N, K)
sigma2_ht
```

```
## [1] 1.7717
```

```r
# Estimated Standard Deviation.
sigma_ht <- sqrt(sigma2_ht)
sigma_ht
```

```
## [1] 1.331052
```

Variance-Covariance Matrix of betas.

```r
# Variance-Covariance Matrix.
varcov_beta_ht <- sigma2_ht*solve(t(X)%*%X)
varcov_beta_ht
```

```
##                           log_yi       log_yj    log_dist
##          18.1728919 -0.522713794 -0.447734075 -1.76738222
## log_yi   -0.5227138  0.043880329  0.006468273  0.02265107
## log_yj   -0.4477341  0.006468273  0.039080340  0.01704323
## log_dist -1.7673822  0.022651067  0.017043228  0.24106658
```

```r
# Standard deviation of betas.
sd_beta_ht <- sqrt(diag(varcov_beta_ht))
sd_beta_ht
```

```
##              log_yi    log_yj  log_dist
## 4.2629675 0.2094763 0.1976875 0.4909853
```

Coefficient of determination.

```r
# Coefficient of determination R2.
R2 <- R2.f(Y, Y_ht)
R2
```

```
## [1] 0.658617
```

This result indicates that approximately **65.9%** of the variability of the logarithm of the trade flow between two regions in Westeros is explained by the variability of the regressors.

```r
# Adjusted Coefficient of determination.
R2adj <- R2adj.f(R2, N, K)
R2adj
```

```
## [1] 0.6266123
```

This result indicates that, adjusted for the number of regressors, approximately **62.7%** of the variability of the logarithm of the trade flow between two regions in Westeros is explained by the variability of the regressors.

4.5.1 Hypothesis testing over corrected model.

Test of normality over residuals.

```r
# JB statistic.
Jarque_Bera.f(e_ht)$JB
```

```
## [1] 0.6963568
```

```r
# JB statistic p-value.
Jarque_Bera.f(e_ht)$pValue
```

```
## [1] 0.7059729
```

```r
# Hypothesis testing result (5%).
Jarque_Bera.f(e_ht)$hp
```

```
## [1] "It is not rejected the hypothesis that residuals follow a normal distribution"
```

Since the hypothesis of normality of the residuals has not been rejected, it is possible to proceed with the inference over the other parameters of the model.

Global significance.

```r
# Selection matrix of constraints.
R_gs <- cbind(0, diag(1, nrow=K-1))

# Constraints value vector.
r_gs <- as.matrix(rep(0, K-1))

# Lambda 1 value.
lambda1_LM.f(R_gs, r_gs, B_ht, varcov_beta_ht, N)$lambda1
```

```
## [1] 20.57878
```

```r
# Lambda 1 p-value.
lambda1_LM.f(R_gs, r_gs, B_ht, varcov_beta_ht, N)$pValue
```

```
## [1] 1.294656e-07
```

```r
# Hypothesis testing.
lambda1_LM.f(R_gs, r_gs, B_ht, varcov_beta_ht, N)$hp
```

```
## [1] "The null hypothesis is rejected"
```

At a confidence level of 95%, the null hypothesis of global non-significance is rejected, i.e., at least one variable is statistically significant in the model.

Individual significance.

```r
# Hypothesis testing (95% confidence).
ind.sig.f(B_ht, sd_beta_ht, N)
```

```
##       Beta tValue pValue                            Hypothesis
## 1   log_yi  5.101      0 Beta 1 is statistically significative
## 2   log_yj  4.773      0 Beta 2 is statistically significative
## 3 log_dist -2.564  0.015 Beta 3 is statistically significative
```

The results of the individual hypothesis tests indicate that all the variables in the model, at a 95% level of confidence, are significant in explaining the trade flow between the different regions of Westeros.

Confidence intervals.

```r
# Confidence intervals (95% confidence).
conf.int.f(B_ht, sd_beta_ht, N, alpha = 0.05)
```

```
##       Beta Left.Limit Right.Limit                            Conclusion
## 1   log_yi      0.642       1.495 Beta 1 is statistically significative
## 2   log_yj      0.541       1.346 Beta 2 is statistically significative
## 3 log_dist     -2.259      -0.259 Beta 3 is statistically significative
```

```r
# Confidence intervals (99% confidence).
conf.int.f(B_ht, sd_beta_ht, N, alpha = 0.01)
```

```
##       Beta Left.Limit Right.Limit                                Conclusion
## 1   log_yi      0.495       1.642     Beta 1 is statistically significative
## 2   log_yj      0.402       1.485     Beta 2 is statistically significative
## 3 log_dist     -2.603       0.086 Beta 3 is not statistically significative
```

These results indicate that at a 95% level of confidence, all variables demostrated statistical significance. However, at a 99% level of confidence, the variable $log(d_{ij})$ appears to lack statistical significance once again. Despite this observation, it has been decided to keep this variable in the model.

5. Inference over the theoretical model.

Once the definitive model has been obtained from the evaluation of global and individual significance, it is pertinent to verify whether what was initially proposed in the theoretical model can be valid in the population. This verification is carried out by conducting a hypothesis test, similar to the global significance test, but with restrictions on the parameters to align with the theoretical values. In consideration of this, the hypotheses proposed for this analysis are as follows:

$$H_0 : \beta_1 = 1, \quad \beta_2 = 1, \quad \beta_3 = -2$$
$$H_1 : \text{At least one is not satisfied}$$

This hypothesis test will be carried out using the Lambda 1 test (Lagrange Multipliers). First, it will be necessary to construct the selection matrix and the vector of constraints.

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{3\times4} \qquad r = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}_{3\times1}$$

```
# Selection matrix of constraints.
R_tv <- cbind(0, diag(1, nrow=3))

# Constraints value vector.
r_tv <- as.matrix(c(1, 1, -2))

# Lambda 1 value.
lambda1_LM.f(R_tv, r_tv, B_ht, varcov_beta_ht, N)$lambda1
```

```
## [1] 0.8654394
```

```
# Lambda 1 p-value.
lambda1_LM.f(R_tv, r_tv, B_ht, varcov_beta_ht, N)$pValue
```

```
## [1] 0.4690677
```

```
# Hypothesis testing.
lambda1_LM.f(R_tv, r_tv, B_ht, varcov_beta_ht, N)$hp
```

```
## [1] "The null hypothesis is not rejected"
```

The results indicate, with a 95% confidence level, that the null hypothesis stating the estimated model aligns with the theoretical model is not rejected. Consequently, it is permissible to proceed by adopting the theoretical values as the model parameters. Therefore, the new coefficient vector will be constructed using the Restricted Maximum Likelihood Estimator.

```
# Beta of theoretical values.
B_ht <- rBeta.f(B_ht, X, R_tv, r_tv)
B_ht
```

```
##               [,1]
##           1.593325
## log_yi    1.000000
## log_yj    1.000000
## log_dist -2.000000
```

Now, it is necessary to estimate the rest of variables based on these parameters.

5.1. Theoretical model estimations.

Dependent variable estimation.

```
# Estimated dependent variable.
Y_ht <- X%*%B_ht
head(Y_ht)
```

```
##          [,1]
## [1,] 4.854876
## [2,] 3.795104
## [3,] 1.627731
## [4,] 4.425036
## [5,] 3.902805
## [6,] 3.775406
```

Errors estimation.

```
# Estimated errors.
e_ht <- Y - Y_ht
head(e_ht)
```

```
##            [,1]
## [1,]  0.8524162
## [2,]  1.1241787
## [3,]  0.5672541
## [4,]  1.1865491
## [5,]  3.1700997
## [6,] -0.2955635
```

Estimated variance.

```
# Estimated Variance.
sigma2_ht <- sigma2_ht.f(e_ht, N, K)
sigma2_ht
```

```
## [1] 1.915446
```

```
# Estimated Standard Deviation.
sigma_ht <- sqrt(sigma2_ht)
sigma_ht
```

```
## [1] 1.383997
```

Variance-Covariance Matrix of betas.

```
# Variance-Covariance Matrix.
varcov_beta_ht <- sigma2_ht*solve(t(X)%*%X)
varcov_beta_ht
```

```
##                           log_yi       log_yj    log_dist
##          19.6473485 -0.565124148 -0.484060953 -1.91077868
## log_yi   -0.5651241  0.047440557  0.006993076  0.02448886
## log_yj   -0.4840610  0.006993076  0.042251121  0.01842603
## log_dist -1.9107787  0.024488861  0.018426029  0.26062551
```

```
# Standard deviation of betas.
sd_beta_ht <- sqrt(diag(varcov_beta_ht))
sd_beta_ht
```

```
##              log_yi    log_yj  log_dist
## 4.4325330 0.2178085 0.2055508 0.5105149
```

Coefficient of determination.

```
# Coefficient of determination R2.
R2 <- R2.f(Y, Y_ht)
R2
```

```
## [1] 0.8252146
```

This result indicates that approximately **82.5%** of the variability of the logarithm of the trade flow between two regions in Westeros is explained by the variability of the regressors.

```
# Adjusted Coefficient of determination.
R2adj <- R2adj.f(R2, N, K)
R2adj
```

```
## [1] 0.8088284
```

This result indicates that, adjusted for the number of regressors, approximately **80.9%** of the variability of the logarithm of the trade flow between two regions in Westeros is explained by the variability of the regressors.

5.2. Final interpretations.

Firstly, as the theoretical parameter values were not statistically rejected, the interpretations of the coefficients are as follows:

- $\hat{\beta}_1$: Given a 1% increase in the GDP of region i, the trade flow between regions i and j increases by approximately **1%**, holding everything else constant.
- $\hat{\beta}_2$: Given a 1% increase in the GDP of region j, the trade flow between regions i and j increases by approximately **1%**, holding everything else constant.
- $\hat{\beta}_3$: Given a 1% increase in the distance between regions i and j, the trade flow between them decreases by **2%**, holding everything else constant.

On the other hand, it is interesting to note that despite the variances of the residuals and coefficients are considerably higher in the restricted theoretical model compared to the unrestricted OLS estimations, the R^2 and R^2_{adj} are much higher. This suggests that the theoretical model fits the data better than the unrestricted model.

6. Conclusions.

The objective of this document was to estimate a Gravitational Model of Trade applied to the fictional world of Westeros, using simulated data. The exercise was carried out for pedagogical reasons to review the process of hypothesis testing and inferences over the estimated parameters of the model, and demonstrate its practical application through programming in R.

Given the non-real nature of the data, the model and estimations lack direct applicability in the real world. However, conducting the exercise with simulated data allows for a comparison of how estimations and hypothesis testing could approximate population parameters and the actual data generation process. In this case, the iterative process of eliminating non-significant variables and evaluating the theoretical model led to identifying a model that aligns well with the data. In this case, it was correctly inferred that the theoretical model was the real one.

Nevertheless, despite the efficacy of the performed hypothesis testing in approximating the true model, given the limited dataset and the inherent randomness of the exercise, the final model still estimates the standard deviation of the residuals and the value of the intercept coefficient quite inaccurately. This highlights that while inference evaluation techniques are effective in approximating real parameters, they are not entirely precise. Another perspective is that even when it is known that the variable $log(d_{ij})$ does influence the model, at a 99% confidence level, the tests indicated that it was not significant. Therefore, it is advisable to not completely rely on this test, as it has the random possibility to reject something true or accept something false.

7. References.

Judge, G., Hill R., Griffiths W., Lütkepohl H. & Lee T. (1988). Introduction to the Theory and Practice of Econometrics. John Wiley & Sons.

Krugman, P. & Obstfeld, M. (2006). Comercio mundial: una visión general. *Economía Internacional. Teoría y política* (pp. 13-26). PEARSON.

Appendix 1. Simulation of data.

As it was explained at the beginning of the notebook, all data used in the model were simulated and do not represent real-world measurements or official data of the fictional work and the economic context under analysis. In this appendix, it is explained how the variables and population parameters of the model were simulated.

- **GDP**: The "GDPs" of the regions were simulated based on a uniform distribution ranging from 100 to 10.000. The nine simulated values were assigned to each region according to a ranking of wealth created for this purpose, derived from a brief review of followers' opinions on the internet. The regions are ordered in descending wealth as follows: Westerlands, The Reach, Vale of Aryn, Crownlands, Riverlands, The North, Dorne, Stormlands, and Iron Islands. Finally, the GDP is measured in millions of gold ingots.

- **Distances**: The distances between regions were measured as the distance between their capital cities, manually extracted from the map in Figure 1. The distance is measured in pixels, hence its measurement is not presented explicitly in the exercise.

- **Frontiers (land borders)**: The binary variables indicating shared frontiers were directly obtained from the map in Figure 1.

- **Errors or residuals**: The unobservable errors or perturbances were simulated from a normal distribution with a mean of 0 and a standard deviation of 1.2. The value of the standard deviation was chosen arbitrarily.
- **Model parameters**: The main model parameters were defined according to the theoretical model The elasticity of the GDP, of both regions, over the trade flow (a and b) was defined exactly as 1. The elasticity of distance over the trade flow (c) was defined as -2. Finally, the constant term A was arbitrarily defined as 3, then the natural logarithm is aproximated as 1.0986.