PE I: Multivariable Regression

General Modelling Difficulties
(Chapter 4.9)

Andrius Buteikis, andrius.buteikis@mif.vu.lt
http://web.vu.lt/mif/a.buteikis/

Multiple Regression: Model Assumptions

Much like in the case of the univariate regression with one independent variable, the multiple regression model has a number of required assumptions:

(MR.1): Linear Model The Data Generating Process (**DGP**), or in other words, the population, is described by a linear (*in terms of the coefficients*) model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon \tag{MR.1}$$

(MR.2): Strict Exogeneity Conditional expectation of ε, given all observations of the explanatory variable matrix $\boldsymbol{X}$, is zero:

$$\mathbb{E}\left(\varepsilon|\boldsymbol{X}\right) = \mathbf{0} \tag{MR.2}$$

This assumption also implies that $\mathbb{E}(\varepsilon) = \mathbb{E}\left(\mathbb{E}(\varepsilon|\boldsymbol{X})\right) = \mathbf{0}$, $\mathbb{E}(\varepsilon\boldsymbol{X}) = \mathbf{0}$ and $\mathbb{C}\mathrm{ov}(\varepsilon, \boldsymbol{X}) = \mathbf{0}$. Furthermore, this property implies that: $\mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{X}\right) = \boldsymbol{X}\beta$

(MR.3): Conditional Homoskedasticity The variance-covariance matrix of the error term, conditional on $\boldsymbol{X}$ is constant:

$$\mathbb{V}\text{ar}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \begin{bmatrix} \mathbb{V}\text{ar}(\epsilon_1) & \mathbb{C}\text{ov}(\epsilon_1, \epsilon_2) & ... & \mathbb{C}\text{ov}(\epsilon_1, \epsilon_N) \\ \mathbb{C}\text{ov}(\epsilon_2, \epsilon_1) & \mathbb{V}\text{ar}(\epsilon_2) & ... & \mathbb{C}\text{ov}(\epsilon_2, \epsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}(\epsilon_N, \epsilon_1) & \mathbb{C}\text{ov}(\epsilon_N, \epsilon_2) & ... & \mathbb{V}\text{ar}(\epsilon_N) \end{bmatrix} = \sigma_\epsilon^2 \boldsymbol{I} \qquad \text{(MR.3)}$$

(MR.4): Conditionally Uncorrelated Errors The covariance between different error term pairs, conditional on $\boldsymbol{X}$, is zero:

$$\mathbb{C}\text{ov}\left(\epsilon_i, \epsilon_j|\boldsymbol{X}\right) = 0, \quad i \neq j \qquad \text{(MR.4)}$$

This assumption implies that all error pairs are uncorrelated. For cross-sectional data, this assumption implies that there is no spatial correlation between errors.

(MR.5) There exists no exact linear relationship between the explanatory variables. This means that:

$$c_1 X_{i1} + c_2 X_{i2} + ... + c_k X_{ik} = 0, \ \forall i = 1, ..., N \iff c_1 = c_2 = ... = c_k = 0 \quad \text{(MR.5)}$$

This assumption is violated if there exists some $c_j \neq 0$.
Alternatively, this requirement means that:

$$\text{rank}\,(\boldsymbol{X}) = k + 1$$

or, alternatively, that:

$$\det\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \neq 0$$

This assumption is important, because a linear relationship between independent variables means that we cannot separately estimate the effects of changes in each variable separately.

(MR.6) (optional) The residuals are normally distributed:

$$\varepsilon | \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}\right) \quad \text{(MR.6)}$$

General Modelling Difficulties

- Up until now, we have assumed that the multiple regression equation we are estimating includes **all the relevant explanatory variables from the underlying DGP**.

In practice, this is rarely the case:

- sometimes some variables may not be included due to various factors like oversight, ignorance, or even lack of observations.
- sometimes irrelevant variables are included in the model itself.

Finally, if we attempt to take this into account, we may find ourselves with more than one model, that may be acceptable. Yet, we may want to select one model to present coefficient interpretations, predict values, test results, etc.

Causality vs. Prediction

When specifying a model, an important consideration should be taken with regards to the **purpose of creating the model**. Do we want to:

a) Use the model for prediction?
b) Use the model for causal analysis?

With **causal inference** we are interested in the **effect of a one unit (or one percent, depending on the variable) change in a regressor on the conditional mean of the dependent variable, other factors held constant (i.e. ceteris paribus)**. This includes:

- whether the effect is statistically significant;
- the direction (positive or negative) of the effect;
- the magnitude of the effect;

This type of analysis is important for measuring the effects of policies, laws, taxes, firm expenditure, and so on, on some performance measure, like income, education quality, revenue, etc. In order to measure these effects, we need to be able to separate them from all other possible effects, which may also have an effect - we want to be able to hold all these other effects constant to be able to isolate and measure the effects that we are interested in.

When analysing causality, we may be tempted to *primarily* rely on the correlation between variables, however this is known as the *cum hoc ergo propter hoc* ("with this, therefore because of this") problem, in which two events occur simultaneously.

This is also usually stated as **correlation does not imply causation**.

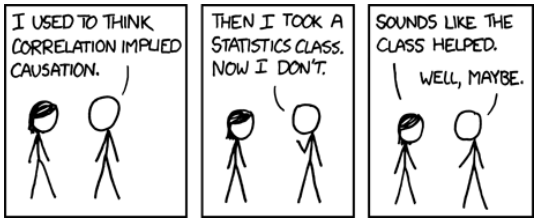


Figure 1: https://xkcd.com/552/

If **the purpose of our model is to predict** the value of the dependent variable, then **we want to choose the regressors, that are highly correlated with the dependent variable** (and subsequently lead to a high value of R^2, or small value of AIC/BIC, or overall a small variance of the residuals).
Whether or not these variables have a direct effect on the dependent variable, or even if we have omitted some relevant variables - **is less important**.

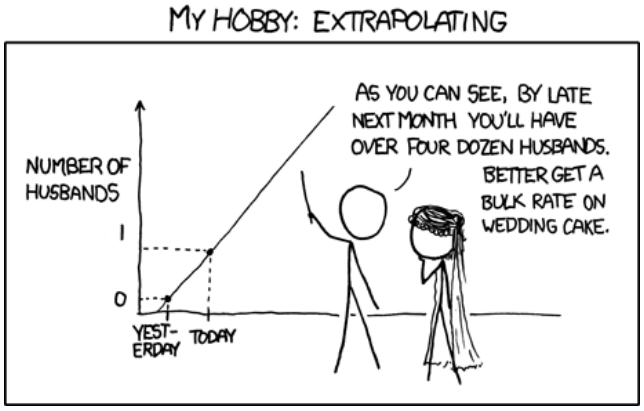


Figure 2: https://xkcd.com/605/

When carrying out regression analysis it is assumed that only the correct and important explanatory variables are included in the model and the correct functional form of the relationship is specified. If any of these requirements are violated, then we are making a **specification error**.

In practical applications, we seldom know the true functional form of the model, so we have to infer it based on the data and relationships between variables, as well as economic theory.

Furthermore, having selected a functional form, we usually have a limited pool of variables to choose from, which often influence not only the subset of variables included in our model, but also the functional form.

We will look at how the usually encountered problems like:

- correlation between explanatory variables;
- omitting a relevant variable;
- including an irrelevant variable;

affect our estimated model and our results.

Correlation Between Explanatory Variables

Assuming $\mathbb{E}(\epsilon_i|X_{1,i}, X_{2,i}) = 0$, looking at the conditional expectation:

$$\mathbb{E}\left(Y_i|X_{1,i}, X_{2,i}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}$$

allows us to have **causal** interpretations for the coefficients:

$$\beta_j = \frac{\partial \mathbb{E}\left(Y_i|X_{1,i}, X_{2,i}\right)}{\partial X_j, i}, j = 1, 2$$

where β_j represents the change in the mean of Y, resulting from a change in X_i, **other factors held constant**.

Assumption $\mathbb{E}(\epsilon_i|X_{1,i}, X_{2,i}) = 0$ means that changes in X_1 and X_2 have no effects on the error term.

Now, suppose that the explanatory regressors are correlated:

$$\mathbb{C}\text{orr}(X_{1,i}, X_{2,i}) \neq 0$$

This is a **common occurrence among explanatory variables in practical applications**.

For simplicity, assume that the correlation can be described as:

$$\mathbb{E}\left(X_{2,i}|X_{1,i}\right) = \gamma_0 + \gamma_1 X_{1,i}$$

If we plug this expression into our conditional expectation of the dependent variable, but this time only condition on $X_{1,i}$, we get:

$$\begin{aligned}
\mathbb{E}\left(Y_i|X_{1,i}\right) &= \beta_0 + \beta_1 X_{1,i} + \beta_2 \mathbb{E}\left(X_{2,i}|X_{1,i}\right) + \mathbb{E}\left(\epsilon_i|X_{1,i}\right) \\
&= \beta_0 + \beta_1 X_{1,i} + \beta_2 \left[\gamma_0 + \gamma_1 X_{1,i}\right] \\
&= \left(\beta_0 + \beta_2\gamma_0\right) + \left(\beta_1 + \beta_2\gamma_1\right) X_{1,i} \\
&= \alpha_0 + \alpha_1 X_{1,i}
\end{aligned}$$

From this specification, we may believe that we only need to estimate

$$Y_i = \alpha_0 + \alpha_1 X_{1,i} + u_i$$

via OLS. In such a case, the OLS estimates of α_0 and α_1 would be unbiased. But this may not be as straightforward as it seems.

If our goal is to use $X_{1,i}$ to predict Y_i, we can use the model with only $X_{1,i}$ and not worry about excluding $X_{2,i}$.

On the other hand, since $X_{2,i}$ is not held constant, $\alpha_1 = \beta_1 + \beta_2\gamma_1$ does not measure the causal effect of $X_{1,i}$ on Y_i, as it includes an indirect effects of the correlation between $X_{1,i}$ on $X_{2,i}$, as γ_1, and the effect of a unit change in $X_{2,i}$, as β_2. The true causal effect of a unit change in $X_{1,i}$, given in β_1, is not directly estimated, unless $\beta_2 = 0$, i.e. if $X_{2,i}$ does not affect Y_i.

Consequently, **to estimate the causal effect of X_1 using OLS**, we would need to include **ALL** explanatory variables, which:

- **are correlated with X_1**
- **are correlated with Y**

On the other hand, we need to be mindful of the possibility of **multicollinearity**.

Because of this, we may need to replace the correlated variables with **control variables**, which act as a proxy for the omitted correlated explanatory variable.

Omitting Relevant Variables

- In order to keep our model easy to interpret, we may sometimes choose to exclude some explanatory variables (including interaction and polynomial terms), which may be significant in the true (unobserved) model.
- Another reason for omitting variables, is that we may not be able to quantify them (e.g. taste, responsibility, etc.), or simply because we do not have them in our dataset (e.g. it may be very expensive to create/acquire a large dataset, compared to a simplified one with less variables).
- On the other hand, omission of relevant variables is always a possibility when having a large selection of various explanatory variables. Finally, we may sometimes simply overlook a particular variable.

Assume that our initial model with k regressors and N observations can be written as:

$$\underset{N\times 1}{\mathbf{Y}} = \underset{N\times k}{\mathbf{X}}\ \underset{k\times 1}{\beta} + \underset{N\times 1}{\varepsilon}, \quad \mathbb{E}(\varepsilon|\mathbf{X}) = \mathbf{0}, \quad \mathbb{V}\mathrm{ar}(\varepsilon|\mathbf{X}) = \sigma_\epsilon^2\mathbf{I}$$

For simplicity, assume that of the k variables, **we decide to include only r variables**. Then the matrix and parameter vector can be **partitioned** as follows:

$$\underset{N\times k}{\mathbf{X}} = \begin{bmatrix} \underset{N\times r}{\mathbf{X}_1} & \underset{N\times(k-r)}{\mathbf{X}_2} \end{bmatrix}, \quad \underset{k\times 1}{\beta} = \begin{bmatrix} \underset{r\times 1}{\beta_1} & \underset{(k-r)\times 1}{\beta_2} \end{bmatrix}$$

So, the **full** (or **true**) model can be written as:

$$\underset{N\times 1}{\mathbf{Y}} = \underset{N\times r}{\mathbf{X}_1}\ \underset{r\times 1}{\beta_1} + \underset{N\times(k-r)}{\mathbf{X}_2}\ \underset{(k-r)\times 1}{\beta_2} + \underset{N\times 1}{\varepsilon}$$

But we decided to specify a model with $k - r$ excluded variables:

$$\underset{N\times 1}{\mathbf{Y}} = \underset{N\times r}{\mathbf{X}_1}\ \underset{r\times 1}{\beta_1} + \underset{N\times 1}{\zeta}, \text{ where } \zeta = \mathbf{X}_2\beta_2 + \varepsilon$$

If we were to estimate the parameters of this **misspecified** model via OLS and use the linear model expression of the true $\mathbf{Y}$:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_1 &= \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{Y} = \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \left(\mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}\right) \\ &= \boldsymbol{\beta}_1 + \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \boldsymbol{\varepsilon} + \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \boldsymbol{\beta}_2\end{aligned}$$

Then, the expected value is calculated similarly to how it was done in the univariate OLS case:

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_1\right] &= \boldsymbol{\beta}_1 + \mathbb{E}\left[\left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \boldsymbol{\varepsilon}\right] + \mathbb{E}\left[\left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \boldsymbol{\beta}_2\right] \\ &= \boldsymbol{\beta}_1 + \mathbb{E}\left[\mathbb{E}\left(\left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \boldsymbol{\varepsilon} \middle| \mathbf{X}_1\right)\right] + \mathbb{E}\left[\mathbb{E}\left(\left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \boldsymbol{\beta}_2 \middle| \mathbf{X}_1\right)\right] \\ &= \boldsymbol{\beta}_1 + \mathbb{E}\left[\left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbb{E}\left(\boldsymbol{\varepsilon} | \mathbf{X}_1\right)\right] + \mathbb{E}\left[\left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbb{E}\left(\mathbf{X}_1^\top \mathbf{X}_2 | \mathbf{X}_1\right) \boldsymbol{\beta}_2\right] \\ &= \boldsymbol{\beta}_1 + \mathbb{E}\left[\left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbb{E}\left(\mathbf{X}_1^\top \mathbf{X}_2 | \mathbf{X}_1\right) \boldsymbol{\beta}_2\right]\end{aligned}$$

- $\widehat{\boldsymbol{\beta}}_1$ is generally biased, since $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_1\right] \neq \boldsymbol{\beta}_1$.
- The bias vanishes if $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$ - every element of $\mathbf{X}_1^\top \mathbf{X}_2$ is a linear combination, which sums to zero - that is, if the omitted variables in $\mathbf{X}_2$ are **not correlated** with any of the included variables in $\mathbf{X}_1$.

Similarly, the variance can be expressed as:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}_1|\mathbf{X}_1) &= \mathbb{E}\left[(\widehat{\boldsymbol{\beta}}_1 - \mathbb{E}(\widehat{\boldsymbol{\beta}}_1))(\widehat{\boldsymbol{\beta}}_1 - \mathbb{E}(\widehat{\boldsymbol{\beta}}_1))^\top|\mathbf{X}_1\right] \\
&= \mathbb{E}\left[\left(\left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\top\boldsymbol{\varepsilon} + \left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\top\mathbf{X}_2\boldsymbol{\beta}_2\right)\left(\left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\top\boldsymbol{\varepsilon} + \left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\top\mathbf{X}_2\boldsymbol{\beta}_2\right)^\top\middle|\mathbf{X}_1\right] \\
&= \ldots \\
&= \sigma^2\left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1} + \left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\top\mathbb{E}\left[\mathbf{X}_2\boldsymbol{\beta}_2\boldsymbol{\beta}_2^\top\mathbf{X}_2^\top\middle|\mathbf{X}_1\right]\mathbf{X}_1\left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1} \geq \sigma^2\left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}
\end{aligned}$$

The variance of the OLS estimates is not the smallest in the misspecified model. This means that the OLS estimates are inefficient.

Finally, we know that the variance of the OLS residuals can be calculated as:

$$\widehat{\sigma}^2 = \frac{\widehat{\epsilon}^\top \widehat{\epsilon}}{N - r}$$

where we can write the residuals as:

$$\begin{aligned}
\widehat{\epsilon} &= \mathbf{Y} - \mathbf{X}_1 \widehat{\beta}_1 \\
&= \mathbf{Y} - \mathbf{X}_1 \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{Y} \\
&= \left[\mathbf{I} - \mathbf{X}_1 \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top\right] \mathbf{Y} \\
&= \left[\mathbf{I} - \mathbf{X}_1 \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top\right] \left[\mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \varepsilon\right] \\
&= \left[\mathbf{I} - \mathbf{X}_1 \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top\right] \left[\mathbf{X}_2 \beta_2 + \varepsilon\right] \\
&= \mathbf{H} \left[\mathbf{X}_2 \beta_2 + \varepsilon\right],
\end{aligned}$$

where$\mathbf{H} = \mathbf{I} - \mathbf{X}_1 \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top$. We also have that $\mathbf{H}$ and $\mathbf{X}_1 \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1}$ are symmetric and idempotent, i.e. $\mathbf{H}^\top \mathbf{H} = \mathbf{H}\mathbf{H}^\top = \mathbf{H}$.

A **trace** of a matrix can be defined as the sum of the elements on the main diagonal:

$$\begin{aligned}
\text{tr}(\mathbf{X}) &= \sum_{i=1}^{N} x_{ii} \\
\text{tr}(\mathbf{ABC}) &= \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CAB}) \\
\text{tr}(\mathbf{ABC}) &\neq \text{tr}(\mathbf{CBA}) \\
\text{tr}(\mathbf{X}^\top\mathbf{Y}) = \text{tr}(\mathbf{X}\mathbf{Y}^\top) &= \text{tr}(\mathbf{Y}^\top\mathbf{X}) = \text{tr}(\mathbf{Y}\mathbf{X}^\top) = \sum_{i,j=1}^{N} X_{i,j}Y_{i,j}
\end{aligned}$$

then:

$$\text{tr}\,(\mathbf{H}) = \text{tr}\left(\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\top\right) = \text{tr}\,(\mathbf{I}) - \text{tr}\left(\mathbf{X}_1\left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\top\right) = N - r$$

Furthermore, since $\varepsilon^\top\mathbf{H}\varepsilon$ is a scalar (i.e. a number), it holds that:

$$\varepsilon^\top\mathbf{H}\varepsilon = \text{tr}\left(\varepsilon^\top\mathbf{H}\varepsilon\right) = \text{tr}\left(\mathbf{H}\varepsilon\varepsilon^\top\right)$$

and:

$$\mathbb{E}(\varepsilon^\top\mathbf{H}\varepsilon) = \text{tr}\left(\mathbf{H}\mathbb{E}(\varepsilon\varepsilon^\top)\right) = \sigma^2\text{tr}\,(\mathbf{H})$$

Which means that the expected value of the residual variance estimate is:

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}^2\right] &= \frac{1}{N-r}\mathbb{E}\left[\widehat{\epsilon}^\top\widehat{\epsilon}\right] \\
&= \frac{1}{N-r}\mathbb{E}\left[\left(\mathbf{H}\mathbf{X}_2\beta_2 + \mathbf{H}\varepsilon\right)^\top\left(\mathbf{H}\mathbf{X}_2\beta_2 + \mathbf{H}\varepsilon\right)\right] \\
&= \frac{1}{N-r}\mathbb{E}\left[\left(\mathbf{H}\mathbf{X}_2\beta_2\right)^\top\left(\mathbf{H}\mathbf{X}_2\beta_2\right) + \left(\mathbf{H}\mathbf{X}_2\beta_2\right)^\top\varepsilon + \varepsilon^\top\mathbf{H}^\top\left(\mathbf{H}\mathbf{X}_2\beta_2\right) + \varepsilon^\top\mathbf{H}^\top\mathbf{H}\varepsilon\right] \\
&= \frac{1}{N-r}\mathbb{E}\left[\left(\mathbf{H}\mathbf{X}_2\beta_2\right)^\top\left(\mathbf{H}\mathbf{X}_2\beta_2\right) + \varepsilon^\top\mathbf{H}\varepsilon\right] \\
&= \frac{1}{N-r}\mathbb{E}\left[\left(\mathbf{H}\mathbf{X}_2\beta_2\right)^\top\left(\mathbf{H}\mathbf{X}_2\beta_2\right)\right] + \frac{1}{N-r}\mathrm{tr}\left(\mathbf{H}\mathbb{E}\left[\varepsilon\varepsilon^\top\right]\right) \\
&= \frac{1}{N-r}\sigma^2\mathrm{tr}\left(\mathbf{H}\right) + \frac{1}{N-r}\mathbb{E}\left[\left(\mathbf{H}\mathbf{X}_2\beta_2\right)^\top\left(\mathbf{H}\mathbf{X}_2\beta_2\right)\right] \\
&= \sigma^2 + \frac{1}{N-r}\mathbb{E}\left[\left(\mathbf{H}\mathbf{X}_2\beta_2\right)^\top\left(\mathbf{H}\mathbf{X}_2\beta_2\right)\right] > \sigma^2
\end{aligned}
$$

If we omit relevant variables from our regression model:

- The OLS estimator is **biased** and inefficient;
- The residual OLS variance estimator is **biased**;
- $\widehat{\sigma}^2$ is, on average, overestimated, since $\mathbb{E}\left[\widehat{\sigma}^2\right] > \sigma^2$;
- This overestimation of $\widehat{\sigma}^2$ holds true, even if $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$;

As a result, statistical inferences on the coefficients based on t-tests and confidence intervals are **invalid**.

All in all, we want to avoid excluding important variables from our model.

An alternative (and possibly more intuitive) way of thinking is that omitting a relevant variable X_j is equivalent to using restricted least squares (RLS), where the restriction $\beta_j = 0$ is **incorrect**.

As was mentioned for RLS, this results in **biased** and **inconsistent** parameter estimators (however, they are still efficient). However, if the *excluded* explanatory variable is **not** correlated with the *included* explanatory variables - then there will be no omitted variable bias.

Including Irrelevant Variables

- Assume that wanting to avoid omitting an important variable, or simply being too enthusiastic, we may include explanatory variables, which are not relevant to the model.
- These variables contribute very little to the explanatory power of the model and reduce the degrees of freedom $N - k$.
- This may effect any inference we may draw on the model, for example, by arbitrary increasing the R^2 of the model.

Assume that our **true** model includes k explanatory variables:

$$\underset{N\times 1}{\mathbf{Y}} = \underset{N\times k}{\mathbf{X}}\underset{k\times 1}{\beta} + \underset{N\times 1}{\varepsilon}, \quad \mathbb{E}(\varepsilon|\mathbf{X}) = \mathbf{0}, \quad \mathbb{V}\mathrm{ar}(\varepsilon|\mathbf{X}) = \sigma_\epsilon^2\mathbf{I}$$

However, we have included r additional variables, which we will collect in a separate matrix $\mathbf{Z}$. Then our **misspecified** model is:

$$\underset{N\times 1}{\mathbf{Y}} = \underset{N\times k}{\mathbf{X}}\underset{k\times 1}{\beta} + \underset{N\times r}{\mathbf{Z}}\underset{r\times 1}{\gamma} + \underset{N\times 1}{\boldsymbol{u}}$$

or, more compactly:

$$\mathbf{Y} = \begin{bmatrix}\mathbf{X} & \mathbf{Z}\end{bmatrix}\begin{bmatrix}\beta \\ \gamma\end{bmatrix} + \boldsymbol{u}$$

If we were to estimate our misspecified model via OLS, we would get:

$$\begin{bmatrix}\widehat{\beta} \\ \widehat{\gamma}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{X} & \mathbf{Z}^\top\mathbf{Z}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}^\top\mathbf{Y} \\ \mathbf{Z}^\top\mathbf{Y}\end{bmatrix}$$

which we can also write as:

$$\begin{bmatrix}\mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{X} & \mathbf{Z}^\top\mathbf{Z}\end{bmatrix}\begin{bmatrix}\widehat{\beta} \\ \widehat{\gamma}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^\top\mathbf{Y} \\ \mathbf{Z}^\top\mathbf{Y}\end{bmatrix}$$

We are interested how well OLS estimates β.

Notice that the first and second rows of the equality can be written as separate equalities:

$$\mathbf{X}^\top \mathbf{X}\widehat{\beta} + \mathbf{X}^\top \mathbf{Z}\widehat{\gamma} = \mathbf{X}^\top \mathbf{Y}$$
$$\mathbf{Z}^\top \mathbf{X}\widehat{\beta} + \mathbf{Z}^\top \mathbf{Z}\widehat{\gamma} = \mathbf{Z}^\top \mathbf{Y}$$

Multiplying both sides of the *second row* by $\mathbf{X}^\top \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1}$ yields:

$$\mathbf{X}^\top \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X}\widehat{\beta} + \mathbf{X}^\top \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{Z}\widehat{\gamma} = \mathbf{X}^\top \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{Y}$$
$$\Downarrow$$
$$\mathbf{X}^\top \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X}\widehat{\beta} + \mathbf{X}^\top \mathbf{Z}\widehat{\gamma} = \mathbf{X}^\top \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{Y}$$

Then, subtracting it from the first row gives us:

$$\left(\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X}\right) \widehat{\beta} = \mathbf{X}^\top \mathbf{Y} - \mathbf{X}^\top \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{Y}$$
$$\Downarrow$$
$$\mathbf{X}^\top \left(\mathbf{I} - \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top\right) \mathbf{X}\widehat{\beta} = \mathbf{X}^\top \left(\mathbf{I} - \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top\right) \mathbf{Y}$$

Setting $\mathbf{H} = \mathbf{I} - \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1}$ allows to to get the OLS estimate of β:

$$\begin{aligned}
\widehat{\beta} &= \left(\mathbf{X}^\top \mathbf{H}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{H}\mathbf{Y} \\
&= \left(\mathbf{X}^\top \mathbf{H}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{H}(\mathbf{X}\beta + \varepsilon) \\
&= \beta + \left(\mathbf{X}^\top \mathbf{H}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{H}\varepsilon
\end{aligned}$$

Then, the expected value:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}\right] &= \beta + \mathbb{E}\left[\left(\mathbf{X}^\top \mathbf{H}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{H}\varepsilon\right] \\
&= \beta + \mathbb{E}\left[\mathbb{E}\left(\left(\mathbf{X}^\top \mathbf{H}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{H}\varepsilon \middle| \mathbf{X}\right)\right] \\
&= \beta + \mathbb{E}\left[\left(\mathbf{X}^\top \mathbf{H}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{H}\mathbb{E}(\varepsilon)\right] \\
&= \beta
\end{aligned}$$

$\widehat{\beta}$ is an unbiased estimate, even when some irrelevant variables are included in the model.

The variance of the OLS estimate is:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \mathbb{E}(\widehat{\boldsymbol{\beta}}))(\widehat{\boldsymbol{\beta}} - \mathbb{E}(\widehat{\boldsymbol{\beta}}))^\top | \mathbf{X}\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{H}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{H}\boldsymbol{\varepsilon}\left(\left(\mathbf{X}^\top\mathbf{H}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{H}\boldsymbol{\varepsilon}\right)^\top \middle| \mathbf{X}\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{H}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{H}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{H}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{H}\mathbf{X}\right)^{-1} \middle| \mathbf{X}\right] \\
&= \left(\mathbf{X}^\top\mathbf{H}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{H}\left(\sigma^2\mathbf{I}\right)\mathbf{H}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{H}\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}^\top\mathbf{H}\mathbf{X}\right)^{-1} \geq \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}
\end{aligned}$$

If $\mathbf{X}$ and $\mathbf{Z}$ are orthogonal (i.e. **uncorrelated**), then the OLS estimates are efficient, despite inclusion of irrelevant variables. Otherwise, they are inefficient.

Finally, it can be shown, that:

$$\widehat{\sigma}^2 = \frac{\widehat{\epsilon}^\top\widehat{\epsilon}}{N-k-r}$$

is an unbiased estimator of σ^2.

Incorrect Functional Form

Specifying an incorrect functional form may result in the following:

- Heteroskedastic residuals;
- Autocorrelated residuals;

Hence, do not believe that a significant heteroskedasticity or autocorrelation can be adequately fixed by using robust standard errors. Examine the model for any specification problems.
Many economic applications use either a **log-log**, or **log-linear** model, since logarithmic transformations **stabilize the variance** of the transformed variable.

Example

Assume that we specify a linear, instead of a log-linear model.
Specifically, assume that the true model is:

$$\log(Y_i) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$

But assume that we fit the following model:

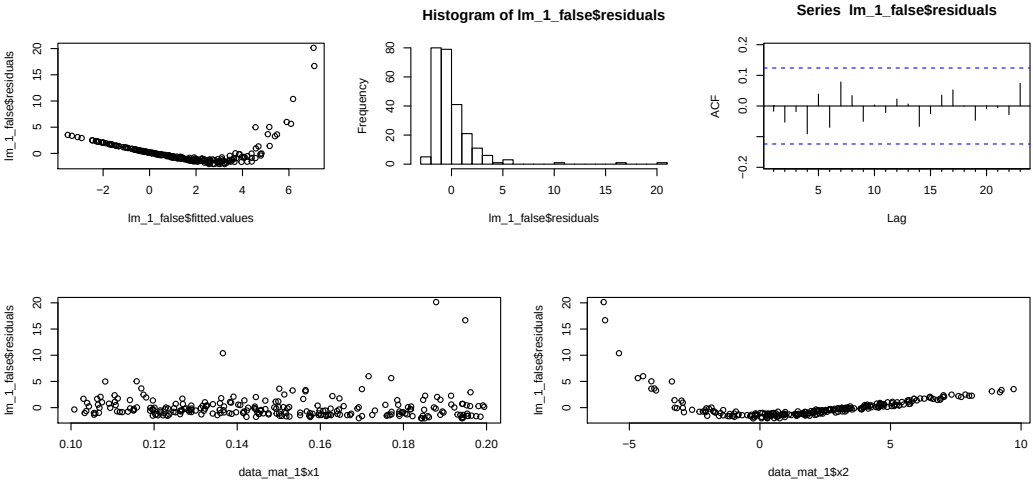
$$Y_i = \alpha_0 + \alpha_1 X_{1,i} + \alpha_2 X_{2,i} + \zeta_i$$

```r
set.seed(123)
#
N <- 250
beta_vec <- c(0.1, 0.5, -0.5)
#
x1 <- sample(seq(from = 0.1, to = 0.2, length.out = 5000), size = N, replace = TRUE)
x2 <- rnorm(mean = 2, sd = 3, n = N)
e  <- rnorm(mean = 0, sd = 0.2, n = N)
#
y  <- exp(cbind(1, x1, x2) %*% beta_vec + e)
#
data_mat_1 <- data.frame(y, x1, x2)
```

```r
lm_1_false <- lm(y ~ 1 + x1 + x2, data = data_mat_1)
print(round(coef(summary(lm_1_false)), 5))
```

```
##             Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  1.28436    0.80315   1.59915  0.11107
## x1           9.62872    5.25331   1.83289  0.06802
## x2          -0.66323    0.04810 -13.78934  0.00000
```

```r
layout(matrix(c(1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5), ncol = 6, byrow = TRUE))
#
plot(lm_1_false$fitted.values, lm_1_false$residuals)
hist(lm_1_false$residuals, breaks = 25)
forecast::Acf(lm_1_false$residuals)
plot(data_mat_1$x1, lm_1_false$residuals)
plot(data_mat_1$x2, lm_1_false$residuals)
```



```r
GQ_t <- lmtest::gqtest(lm_1_false, alternative = "two.sided", order.by = ~ x1)
print(GQ_t)
```

```
##
##  Goldfeld-Quandt test
##
## data:  lm_1_false
## GQ = 3.4473, df1 = 122, df2 = 122, p-value = 3.579e-11
## alternative hypothesis: variance changes from segment 1 to 2
```

```r
GQ_t <- lmtest::gqtest(lm_1_false, alternative = "two.sided", order.by = ~ x2)
print(GQ_t)
```

```
##
##  Goldfeld-Quandt test
##
## data:  lm_1_false
## GQ = 0.00097945, df1 = 122, df2 = 122, p-value < 2.2e-16
## alternative hypothesis: variance changes from segment 1 to 2
```

```r
W_t <- lmtest::bptest(lm_1_false, ~ x1 + I(x1^2) + x2 + I(x2^2) + x1:x2)
print(W_t)
```

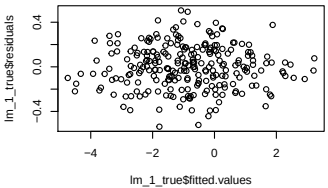
```
##
##  studentized Breusch-Pagan test
##
## data:  lm_1_false
## BP = 108.6, df = 5, p-value < 2.2e-16
```

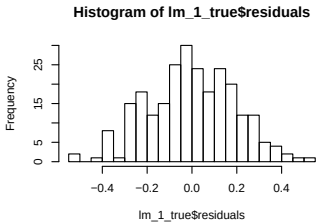
So, we would reject the null hypothesis of homoskedasticity, which means that the residuals would appear to be heteroskedastic.

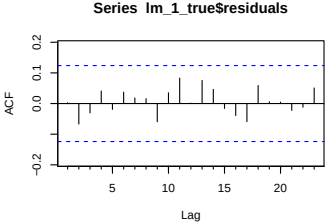
However, if instead of opting to use HCE, we were to specify a log-linear model:

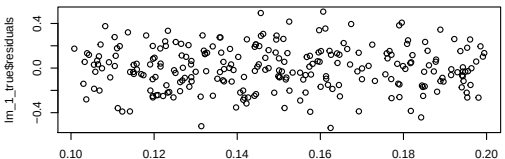
```
lm_1_true <- lm(log(y) ~ 1 + x1 + x2, data = data_mat_1)
print(round(coef(summary(lm_1_true)), 5))
```

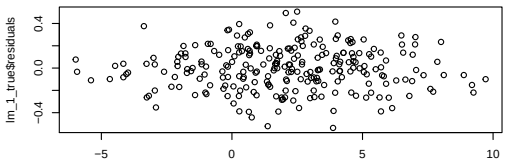
```
##             Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  0.14242    0.06768    2.10452  0.03634
## x1           0.26047    0.44266    0.58844  0.55678
## x2          -0.50738    0.00405 -125.19452  0.00000
```











```r
GQ_t <- lmtest::gqtest(lm_1_true, alternative = "two.sided", order.by = ~ x1)
print(GQ_t)
```

```
##
##  Goldfeld-Quandt test
##
## data:  lm_1_true
## GQ = 1, df1 = 122, df2 = 122, p-value = 1
## alternative hypothesis: variance changes from segment 1 to 2
```

```r
GQ_t <- lmtest::gqtest(lm_1_true, alternative = "two.sided", order.by = ~ x2)
print(GQ_t)
```

```
##
##  Goldfeld-Quandt test
##
## data:  lm_1_true
## GQ = 0.97116, df1 = 122, df2 = 122, p-value = 0.8719
## alternative hypothesis: variance changes from segment 1 to 2
```

```r
W_t <- lmtest::bptest(lm_1_true, ~ x1 + I(x1^2) + x2 + I(x2^2) + x1:x2)
print(W_t)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_1_true
## BP = 8.5609, df = 5, p-value = 0.1279
```

We would not reject the null hypothesis that the residuals are homoskedastic.

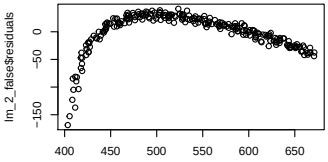
Example

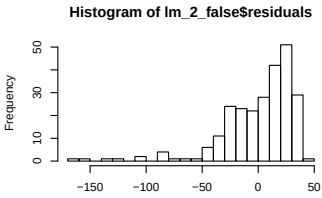
Assume that we specify a correct form for the dependent variable, but an incorrect form for the **independent** variable.

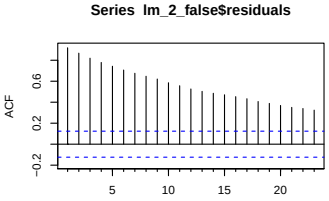
```
set.seed(123)
#
N <- 250
beta_vec <- c(2, 100, 3)
#
x1 <- seq(from = 10, to = 500, length.out = N)
x2 <- rnorm(mean = 2, sd = 1, n = N)
e  <- rnorm(mean = 0, sd = 5, n = N)
#
y  <- cbind(1, log(x1), x2) %*% beta_vec + e
#
data_mat_2 <- data.frame(y, x1, x2)
```

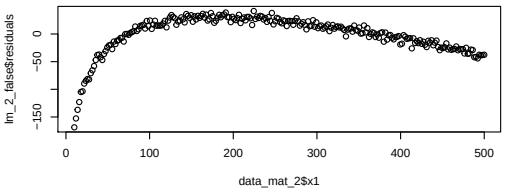
```
lm_2_false <- lm(y ~ 1 + x1 + x2, data = data_mat_2)
print(round(coef(summary(lm_2_false)), 5))
```

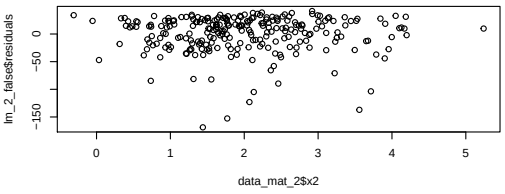
```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 393.97792    6.32438 62.29509  0.00000
## x1            0.54056    0.01485 36.39325  0.00000
## x2            2.79541    2.24355  1.24597  0.21395
```


lm_2_false$residuals
0
−50
−150
400
450
500
550
600
650
lm_2_false$fitted.values


Histogram of lm_2_false$residuals
Frequency
0
10
30
50
−150
−100
−50
0
50
lm_2_false$residuals


Series lm_2_false$residuals
ACF
−0.2
0.2
0.6
5
10
15
20
Lag


lm_2_false$residuals
0
−50
−150
0
100
200
300
400
500
data_mat_2$x1


lm_2_false$residuals
0
−50
−150
0
1
2
3
4
5
data_mat_2$x2

If we test for autocorrelation:

```
DW_t <- lmtest::dwtest(lm_2_false, alternative = "two.sided")
print(DW_t)
```

```
##
##  Durbin-Watson test
##
## data:  lm_2_false
## DW = 0.053879, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is not 0
```

```
BG_T <- lmtest::bgtest(lm_2_false, order = 2)
print(BG_T)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 2
##
## data:  lm_2_false
## LM test = 213.53, df = 2, p-value < 2.2e-16
```

We would not reject the null hypothesis of no serial correlation.

We would also reject the null hypothesis of homoskedasticity:

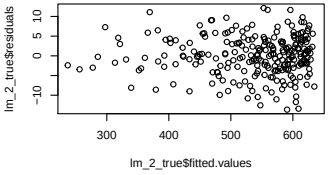
```
W_t <- lmtest::bptest(lm_2_false, ~ x1 + I(x1^2) + x2 + I(x2^2) + x1:x2)
print(W_t)
```

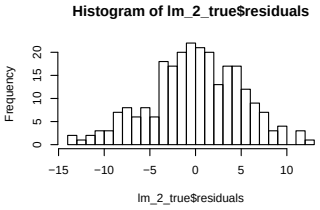
```
##
##  studentized Breusch-Pagan test
##
## data:  lm_2_false
## BP = 70.538, df = 5, p-value = 7.92e-14
```

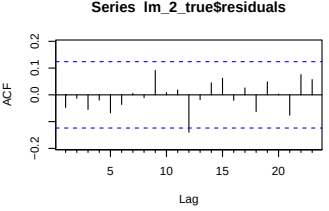
Specifying the true model:

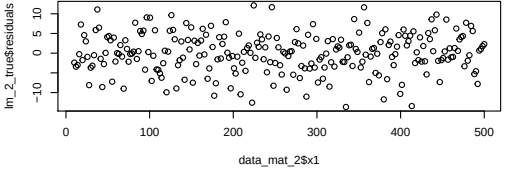
```
lm_2_true <- lm(y ~ 1 + log(x1) + x2, data = data_mat_2)
print(round(coef(summary(lm_2_true)), 5))
```

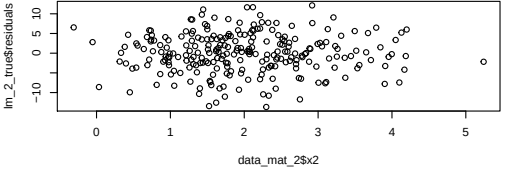
```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept)   2.23478    2.16937   1.03015  0.30395
## log(x1)      99.88577    0.38009 262.79397  0.00000
## x2            3.38075    0.33788  10.00584  0.00000
```











```r
DW_t <- lmtest::dwtest(lm_2_true, alternative = "two.sided")
print(DW_t)
```

```
##
##  Durbin-Watson test
##
## data:  lm_2_true
## DW = 2.0926, p-value = 0.5
## alternative hypothesis: true autocorrelation is not 0
```

```r
BG_T <- lmtest::bgtest(lm_2_true, order = 2)
print(BG_T)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 2
##
## data:  lm_2_true
## LM test = 0.61856, df = 2, p-value = 0.734
```

```r
W_t <- lmtest::bptest(lm_2_true, ~ log(x1) + I(log(x1)^2) + x2 + I(x2^2) + log(x1):x2)
print(W_t)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_2_true
## BP = 4.7901, df = 5, p-value = 0.442
```

We would not reject the null hypothesis of no autocorrelation, and we would not reject the null hypothesis of homoskedasticity.

Remaining modelling difficulties will be presented directly from the lecture notes (from section 4.9.6)