PE I: Multivariable Regression

Introduction: Model Specification (Chapter 4.1)

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

Multiple Regression Model Specification

Estimating a univariate regression and transforming the dependent and/or the independent variable still leaves a dependence structure in the residuals. A common reason is that it is rarely the case, that economic relationships involve just two variables.

Multiple Regression With k Independent Variables

Generally, the dependent variable Y may depend on k different independent variables. Thus in practice we are often faced with a **multiple regression model**:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i, \quad i = 1, ..., N$$

we can think of the term $\beta_0 = \beta_0 \cdot 1$, i.e. as the coefficient of a **constant term**.

This is not the same as a **multivariate** regression, where $\boldsymbol{Y}$ is an $N \times q$ **matrix**, as opposed to the $N \times 1$ vector, which is the case in the univariate and multivariable regressions.
For example, an *equation system*, where $Y_{1,i}$ is the supply and $Y_{2,i}$ is the demand and $X_{1,i}, ..., X_{k,N}$ a vector of explanatory variables - is a form of a multivariate regression.

A single coefficient β_j measures the effect of a unit change in X_j on the expected value of Y, **ceteris paribus** (**holding all else constant**):

$$\beta_j = \left.\frac{\Delta \mathbb{E}(Y|X_1, ..., X_k)}{\Delta X_j}\right|_{\text{other X's held constant}} = \frac{\partial \mathbb{E}(Y|X_1, ..., X_k)}{\partial X_j}$$

Alternatively, if X_j increases by 1, then Y value changes from $\widetilde{Y}$ to $\widetilde{\widetilde{Y}}$. The total change is:

$$\widetilde{\widetilde{Y}} - \widetilde{Y} = \beta_0 + \beta_1 X_1 + ... + \beta_j (X_j + 1) + ... + \beta_k X_k - (\beta_0 + \beta_1 X_1 + ... + \beta_j X_j + ... + \beta_k X_k) = \beta_j$$

Regarding β_0 - the **intercept** parameter - is the expected value of the dependent variable Y, when **all** of the independent variables $X_1, ..., X_k$ are zero.

However, just like in the univariate regression, β_0 usually does not have a clear economic interpretation and can be considered more of a garbage collector.

Generally, the same kind of variable transformations and interpretations of their coefficients carry over from univariate regression to the multiple regression models.

Polynomial Regression

Previously, we constrained the quadratic model in the univariate regression case by including only one variable. Now, working with the multiple regression model allows us to consider unconstrained polynomials with all their terms. So, a p-order polynomial regression would take the following form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + ... + \beta_p X_{1i}^p + \epsilon_i$$

To investigate the slopes, and to interpret the parameters, we take the partial derivative:

$$\frac{d\mathbb{E}(Y|X_1)}{dX_1} = \beta_1 + 2\beta_2 X_1 + ... + p\beta_p X_1^{p-1}$$

this is the slope of the average value of Y, which changes for every value of X_1 and depends on the parameters $\beta_1, ..., \beta_p$.

For $p = 2$ we have a **quadratic model**, where:

- If $\beta_1 > 0$ and $\beta_2 < 0$, this would indicate that an increase of X, when X is large, has a decreasing effect on Y. For example, if $X =$ age, then $\beta_1 > 0$ and $\beta_2 < 0$ would indicate that as people get older, the effect of age on Y is lessened (diminishing returns for extra years of age).
- If $\beta_1 > 0$ and $\beta_2 > 0$ - this would mean that as people get older, the effect of age on Y is stronger.

However, when interpreting the **ceteris paribus** effect of a change in X_1 on Y, we have to look at the equation:

$$\Delta \mathbb{E}(Y|X_1) = \left(\beta_1 + 2\beta_2 X_1 + ... + p\beta_p X_1^{p-1}\right) \Delta X_1$$

That is, we **cannot** interpret β_1 separately once we add higher order polynomial terms of X_1, like $X_1^2, X_1^3, ..., X_1^p$.

Furthermore, the inclusion of polynomial terms does not complicate least squares estimation. Nevertheless, in some cases having a variable and its square, or cube, in the same model causes *collinearity* problems.

When including higher order polynomial variables, it is generally recommended to include the lower order polynomials as well. So, if we wanted to include a X^p variable, we would also need to include the lower order polynomials $X_1^2, X_1^3, ..., X_1^{p-1}$ as well.

Regression with Transformed Variables

As in the simple univariate regression case, we may need to transform some of the variables in our regression:

$$f(Y_i) = \beta_0 + \beta_1 g_1(X_{1i}) + ... + \beta_k g_k(X_{ki}) + \epsilon_i, \quad i = 1, ..., N$$

where $f(\cdot), g_1(\cdot), ..., g_k(\cdot)$ are some kind of transformations of our variables. For example, the transformations could yield the following regression:

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_{1i}) + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{3i}^2 + \epsilon_i$$

The interpretation of its coefficients would be similar to the univariate regression case:

- a **one percent** increase in X_1 yields a β_1 percentage change (i.e. $\%\beta_1$) in Y, **ceteris paribus**;
- a **one unit** increase in X_2 yields a (approximately) $100 \cdot \beta_2$ percentage change (i.e. $\%(100 \cdot \beta_2)$) in Y, **ceteris paribus**;
- etc.

Regression with Interaction Variables

Another model generalization is to include a **cross-product**, or **interaction** terms in the model. Such terms are the products, or the multiplication, or different independent variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \left(X_{1i} X_{2i} \right) + \epsilon_i$$

where:

- β_1 and β_2 are the coefficients of the **main effects** X_1 and X_2, which interaction we also want to include;
- β_3 is the coefficient of the interaction term between X_1 and X_2;

We can also rewrite this equation to better understand how to interpret the coefficients. We can write it either as:

$$Y_i = \beta_0 + (\beta_1 + \beta_3 X_{2i}) X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

or, alternatively, as;

$$Y_i = \beta_0 + \beta_1 X_{1i} + (\beta_2 + \beta_3 X_{1i}) X_{2i} + \epsilon_i$$

Interpretation of β_3 requires holding other variables fixed - so a unit increase in X_1 would result in a $\beta_1 + \beta_3 X_2$ unit change in Y, ceteris paribus. In other words, $\beta_1 + \beta_3 X_2$ changes depending on the value of X_2, therefore, **the effect of a unit increase in X_1 is no longer constant**.

Furthermore, a unit increase in X_2 would result in a $\beta_2 + \beta_3 X_1$ unit change in Y, ceteris paribus, which leads to similar conclusions for the effect of a unit change in X_2 (i.e. it is not constant as well). Depending on the values of X_1 and X_2, this may help drawing some conclusions about variable effects on Y.

It may sometimes be the case that the *p*-value of an interaction effect is very small (i.e. less than α), but the *p*-values of the main effects are large (i.e. greater than α). If we include an interaction term in a model, then we should also include its main effects, **even if the the main effects are insignificant** (i.e. even if their associated *p*-values $> \alpha$).

Furthermore, polynomial variables can be thought of as interaction terms. Consequently, if we include higher order polynomials, then we should not remove the lower order ones, even if they are insignificant. For example, if X^2 is significant, but X is not, we should leave *both* in the regression.

Removing the main effects, but leaving their interaction terms makes the interpretation of the interaction coefficients more difficult.

Furthermore, if some of the values in the interaction can attain **zero** values, then, **if at least one variable in the interaction is zero, then any changes in other variables in the interaction do not have an effect on the dependent variable**.

For example, if $X_1 = 0$, then regardless of the value of X_2, the interaction term would be $\beta_3 \cdot 0 = 0$ - including the main effects separately controls for these cases.

Regression With Indicator Variables

An **indicator variable** is a binary variable that takes values of either zero, or one. It is often used to represent a non-quantitative characteristic - gender, race, location, etc. Indicator variables are often called **dummy**, **binary** or **dichotomous** variables:

$$D_j = \begin{cases} 1, & \text{if characteristic is present in obsetvation } j \\ 0, & \text{if characteristic is not present in obsetvation } j \end{cases}$$

Indicator variables can be used to capture changes in the model intercept, or slopes, or both.

Intercept Indicator Variables

The most common use of indicator variables is to modify the regression model intercept parameter:

$$Y_j = \beta_0 + \alpha D_j + \beta_1 X_1 + \epsilon_j$$

which leads to the following conditional expected value of Y:

$$\mathbb{E}(Y|X) = \begin{cases} (\beta_0 + \alpha) + \beta_1 X_1, & \text{if } D = 1 \\ \\ \beta_0 + \beta_1 X_1, & \text{if } D = 0 \end{cases}$$

Adding an indicator variable to the regression causes a parallel shift in the relationship by amount α (i.e. the regression shifts either up, or down). An indicator variable, which is used to capture the shift in the intercept as a result of some qualitative factor is called an **intercept indicator (or intercept dummy) variable**.

The least squares estimators properties are not affected by the value range of D - we can construct an interval estimate for α and test whether it is statistically significant.

On the other hand, assume that we define an indicator variable as $LD = 1 - D$ and include it *alongside* D:

$$Y_j = \beta_0 + \alpha D_j + \lambda LD_j + \beta_1 X_1 + \epsilon_j$$

In this model the variables $D + LD = 1$ and $\beta_0 = \beta_0 \cdot 1$ - this is a case of **exact collinearity**, which we will expand in a later section.

For now, what you should know is that **the least squares estimator is not defined in such cases**. Consequently , this is sometimes described as the **dummy variable trap**. In other to avoid this problem, we need to choose to include only one of the two indicator variables in our model - either LD, or D.

The regression with $D = 0$ defines a **reference** (or **base**) **group**. Therefore α coefficient indicates the magnitude that observations with a specific characteristic (when $D = 1$) differ from the base group without that characteristic (when $D = 0$).

If α is not statistically significantly different from zero, then there is no difference in Y values between the two groups.

Slope Indicator Variables

As mentioned before r - we can also specify **interaction** variable terms. Since generally, there are no restrictions on what types of data we can create interactions from (as long as we can provide a clear economic interpretation), we can specify the following model, where we interact the indicator variable along with an independent variable X_1:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \gamma(X_{1j} \times D_j) + \epsilon_j$$

where $(X_{1j} \times D_j$ is the interaction variable (the product of X_j and D_j). Alternatively, it can be called the **slope indicator**, or the **slope dummy variable**. A slope indicator variable is treated just like any other variable when estimating the parameters.

This leads to the following conditional expected value of Y:

$$\mathbb{E}(Y|X) = \begin{cases} \beta_0 + (\beta_1 + \gamma)X_1, & \text{if } D = 1 \\ \\ \beta_0 + \beta_1 X_1, & \text{if } D = 0 \end{cases}$$

The interaction term allows to capture the effect of a unit change on X_1 for observations with a specific characteristic (when $D = 1$) - in this case Y changes by $\beta_1 + \gamma$ - and compare it to a unit change in X_1 for observation in the base group (when $D = 0$) - in this case Y changes by β_1.

If γ is not statistically significantly different from zero, then a unit change in X_1 has an identical effect on Y in both groups.

Intercept and Slope Indicator Variable

If we assume that a characteristic D affects **both** the intercept (i.e. the average value of Y when all the other explanatory variables are zero), and the slope (i.e. when a unit increase in X_1 results in a different change in Y, depending on the value of D), then we can specify the multiple regression as:

$$Y_j = \beta_0 + \alpha D_j + \beta_1 X_{1j} + \gamma (X_{1j} \times D_j) + \epsilon_j$$

which leads to the following expected value:

$$\mathbb{E}(Y|X) = \begin{cases} (\beta_0 + \alpha) + (\beta_1 + \gamma) X_1, & \text{if } D = 1 \\ \\ \beta_0 + \beta_1 X_1, & \text{if } D = 0 \end{cases}$$

When including interaction terms, regardless whether they are indicator variables or not, it is generally recommended to include both the variables separately in the equation as well.

Qualitative Factors With Several categories (Categorical Variables)

Many qualitative factors have more than two categories. For example, a variable, which describes m different regions:

$$REGION_{i,j} = \begin{cases} 1, & \text{if observation } j \text{ is from region } i \\ 0, & \text{if observation } j \text{ is not from region } i \end{cases}$$

Furthermore, the indicator variables are such that $\sum_{i=1}^{m} REGION_{i,j} = 1$. Just like before, since we do not want to fall in the **dummy variable trap**, we need to omit one indicator variable. The omitted indicator variable will then define a **reference group** - a group that remains when the remaining regional indicator variables are set to zero.

Assume that we omit the ith region indicator variable.

Then the regression:

$$Y_j = \beta_0 + \alpha_1 REGION_{1,j} + ... + \alpha_{i-1} REGION_{i-1,j} + \alpha_{i+1} REGION_{i+1,j} + ... + \alpha_m REGION_{m,j} + \beta_1 X_{1j} + \epsilon_j$$

its expected value:

$$\mathbb{E}(Y|X) = \begin{cases} (\beta_0 + \alpha_1) + \beta_1 X_1, & \text{if } REGION_1 = 1 \\ (\beta_0 + \alpha_2) + \beta_1 X_1, & \text{if } REGION_2 = 1 \\ & \vdots \\ (\beta_0 + \alpha_{i-1}) + \beta_1 X_1, & \text{if } REGION_{i-1} = 1 \\ \beta_0 + \beta_1 X_1, & \text{if } REGION_i = 1 \\ (\beta_0 + \alpha_{i+1}) + \beta_1 X_1, & \text{if } REGION_{i+1} = 1 \\ & \vdots \\ (\beta_0 + \alpha_m) + \beta_1 X_1, & \text{if } REGION_m = 1 \end{cases}$$

In this case β_1 measures the expected value of Y in $REGION_i$ if $X_1 = 0$. Furthermore, the parameter α_1 measures the expected Y differential between $REGION_1$ and $REGION_i$; α_2 measures the expected Y differential between $REGION_2$ and $REGION_i$, etc.

Alternatively, we may have a single variable REGION, which is coded as follows:

$$\text{REGION}_i = \begin{cases} 0, & \text{if observation j is in the 1-st region} \\ 1, & \text{if observation j is in the 2-nd region} \\ \vdots \\ m-1, & \text{if observation j is in the m-th region} \end{cases}$$

Unfortunately, if we were to include REGION$_i$ as a single variable - it would be difficult to interpret - what does a "unit increase in REGION" mean? As such, a much better approach would be to create the aforementioned regional dummy variables for each case:

$$\text{REGION}_{i,j} = \begin{cases} 1, & \text{if observation j is in region } i \\ 0, & \text{if observation j is not in region } i \end{cases} \quad , \quad i = 1, ..., m-1$$

Note that doing it this way, we create $m-1$ dummy variables, and leave one region (in this case the last m-th region) as the base group.

In most econometric software instead of multiple dummy variables for each subgroup, we have a single **categorical variable**, which defines possible values. Examples include:

- **region** (south, east, north, west),
- **state** (Texas, Alabama, Florida, etc.),
- **country** (USA, UK, France, Germany, China, India, etc.)

and many more, which are often provided either as text strings, or integer codes.

In such cases, when carrying out OLS estimation, these categorical variables are split into dummy variables for each subcategory, with one subcategory automatically selected as the base group.

The Matrix Notation of a Multiple Linear Regression

In general, we can write the multiple regression model as:

$$\begin{cases} Y_1 &= \beta_0 + \beta_1 X_{11} + ... + \beta_k X_{k1} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{12} + ... + \beta_k X_{k21} + \epsilon_2 \\ \vdots & \\ Y_N &= \beta_0 + \beta_1 X_{1N} + ... + \beta_k X_{kN} + \epsilon_N \end{cases}$$

where $X_{j\cdot}$ may be any type of independent variables - logarithms, polynomial, indicator, multiple variable interaction, or non-transformed data. Y may be non-transformed, or may be logarithms of the original dependent variable, or some other transformation, as long as it still results in a linear relationship model.

Then, we can re-write the model in matrix notation:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{k1} \\ 1 & X_{12} & \dots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & \dots & X_{kN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

or in a more compact form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Note that sometimes the explanatory variable matrix $\mathbf{X}$ is called the **design matrix**, or the model matrix, or regressor matrix.

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:

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

This requirement is similar to **(UR.1)** - regardless of how we transform the dependent and/or independent variables, the model must be linear in parameters.

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

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

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

$$\mathbb{E}\left(\mathbf{Y}|\mathbf{X}\right) = \mathbf{X}\beta$$

(MR.3): Conditional Homoskedasticity The variance-covariance matrix of the error term, conditional on **X** is constant:

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

The variance is constant throughout observations. Consequently, the fact that we require the variance-covariance matrix to be diagonal, leads to another condition.

(MR.4): Conditionally Uncorrelated Errors The covariance between different error term pairs, conditional on **X**, is zero:

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

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:

$$\text{rank}\left(\mathbf{X}\right) = k + 1$$

or, alternatively, that:

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

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. Note that in case of a linear relationship between explanatory variables, $\det\left(\mathbf{X}^\top \mathbf{X}\right) = 0$, which means that we cannot calculate $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$, and as such, we cannot carry out an OLS estimation.

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

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}\right) \tag{MR.6}$$

The normality assumption implies that **Y** is also normally distributed (as in the univariate regression case):

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}\left(\mathbf{X}\beta,\ \sigma^2\mathbf{I}\right)$$

- This assumption is useful for hypothesis testing and interval estimation **when the sample size is relatively small**. - However, it is not a necessary requirement, since many of the OLS estimator properties hold regardless.
- Furthermore, if the sample size is relatively large, this assumption is no longer necessary for hypothesis testing and interval estimation.

Note that (MR.5) assumption also holds for the polynomial model, as we are not creating linear transformations of the same variable, but rather polynomial ones.

Example

Assume that we have a design matrix of the following form:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} \\ \vdots & \vdots & \vdots \\ 1 & X_{1N} & X_{2N} \end{bmatrix}$$

```r
set.seed(123)
#
N = 100
#
x1 <- rnorm(mean = 5, sd = 2, n = N)
x2 <- rnorm(mean = 10, sd = 1.5, n = N)
x_mat <- cbind(1, x1, x2)
```

Then, we can easily check the rank of $\mathbf{X}$ and the determinant of $\mathbf{X}^\top \mathbf{X}$:

```r
print(Matrix::rankMatrix(x_mat)[1])
```

```
## [1] 3
```

```r
print(det(t(x_mat) %*% x_mat))
```

```
## [1] 6855725
```

Example

Assume that we have a design matrix where the second independent variable is the square of the first independent variable:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{11}^2 \\ \vdots & \vdots & \vdots \\ 1 & X_{1N} & X_{1N}^2 \end{bmatrix}$$

```r
set.seed(123)
#
N = 100
#
x <- rnorm(mean = 5, sd = 2, n = N)
x_mat <- cbind(1, x, x^2)
```

Then, we can easily check the rank of $\mathbf{X}$ and the determinant of $\mathbf{X}^\top \mathbf{X}$:

```r
print(Matrix::rankMatrix(x_mat)[1])
```

```
## [1] 3
```

```r
print(det(t(x_mat) %*% x_mat))
```

```
## [1] 65930843
```

Example

Assume that we have a design matrix where the second independent variable is a simple linear transformation of the first independent variable:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{11} + 3 \\ \vdots & \vdots & \vdots \\ 1 & X_{1N} & X_{1N} + 3 \end{bmatrix}$$

```r
set.seed(123)
#
N = 100
#
x <- rnorm(mean = 5, sd = 2, n = N)
x_mat <- cbind(1, x, x + 3)
```

Then, we can easily check the rank of $\mathbf{X}$ and the determinant of $\mathbf{X}^\top \mathbf{X}$:

```r
print(Matrix::rankMatrix(x_mat)[1])
```

```
## [1] 2
```

```r
print(det(t(x_mat) %*% x_mat))
```

```
## [1] -1.2379e-07
```

As we can clearly see, the design matrix has linearly-dependent variables, so (MR.5) does not hold.

Example

Assume that we have a design matrix of the following form:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{11} + X_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1N} & X_{2N} & X_{1N} + X_{2N} \end{bmatrix}$$

```r
set.seed(123)
#
N = 100
#
x1 <- rnorm(mean = 5, sd = 2, n = N)
x2 <- rnorm(mean = 10, sd = 1.5, n = N)
x3 <- x1 + x2
x_mat <- cbind(1, x1, x2, x3)
```

Then, we can verify that the design matrix has linearly-dependent variables, so (MR.5) does not hold.

```r
print(Matrix::rankMatrix(x_mat)[1])
```

```
## [1] 3
```

```r
print(det(t(x_mat) %*% x_mat))
```

```
## [1] 4.666691e-05
```

For example, one variable could be `education`, another could be `work_experience` and the last one could be `knowledge = education + work_experience`, which would be linearly dependent, if included with the previous variables.

To be continued after the midterm. . .