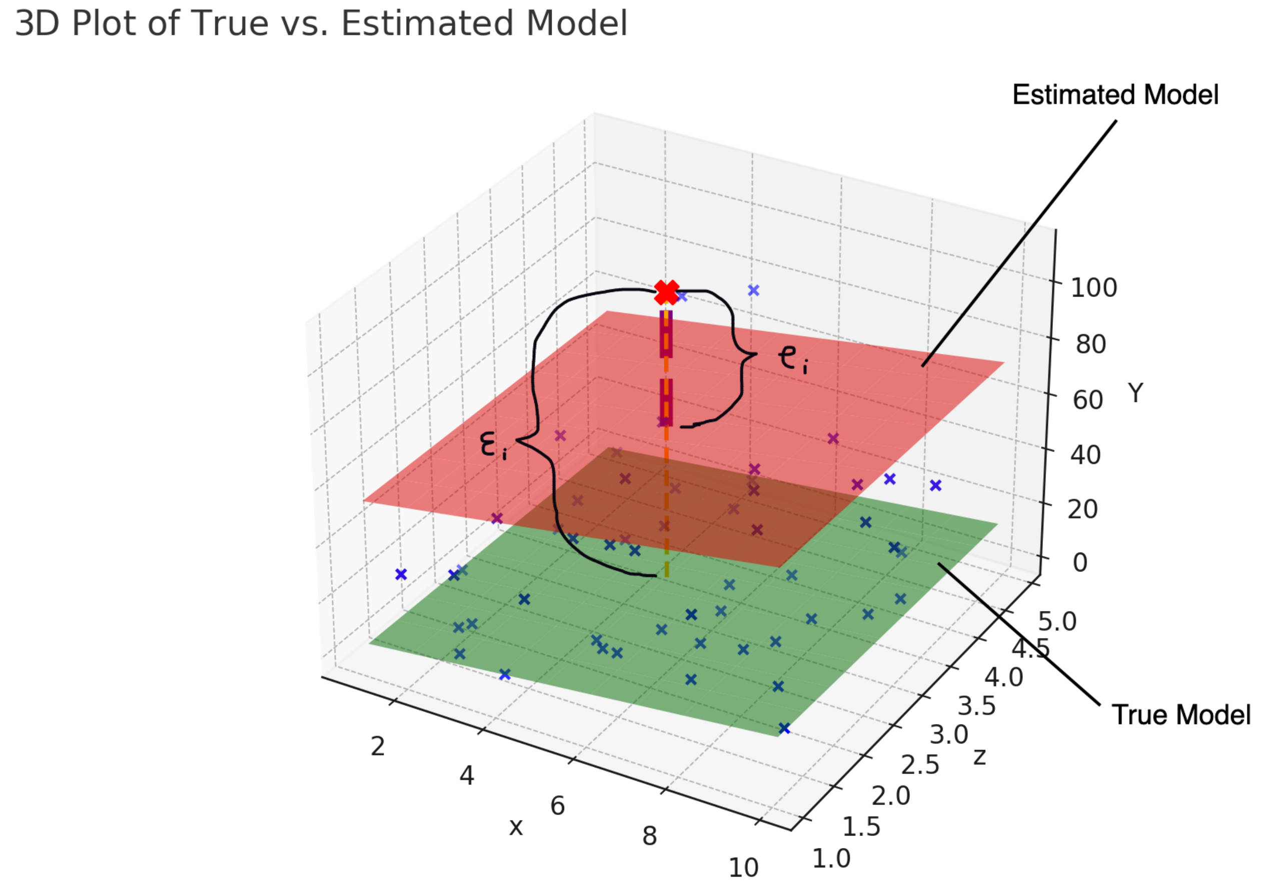
Coursework – Statistical Modelling 1 – Due at 13:00 on Wednesday, 19 March 2025

1. Consider a linear regression model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$, $i = 1, ..., n$. Assume the Full Rank (FR) assumption and the Second Order Assumption (SOA). Provide a graph in which you plot both the true model and the estimated one, and indicate in the graph the error and the residual of the i-th data. [Note: You can do it on paper or using a software] [3 Marks]
[UPDATE: What I am asking for here is just a rough sketch; there is no need to use real data. In class, all the sketches I have drawn were for the linear model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, ..., n$. So the question is: what mental picture should one have when there is one additional covariate?]
Solution:



Marks: 0.5 for attempt, additional 0.5 for having two planes in the picture, additional 0.5 for correct representation of the residual and additional 0.5 for correct representation of the error.

2. Consider the matrix

$$P = \begin{pmatrix} 0.8 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}.$$

Provide two different ways to demonstrate that P is not a projection matrix. [2 Marks]
[UPDATE: Before the matrix P was

$$P = \begin{pmatrix} 08 & 06 \\ 06 & 04 \end{pmatrix}.$$

As you can see, there was a typo because I forgot the dots after the zeros.]
Solution: For the first method, we use the fact that $P \neq P^2$. Indeed,

$$P^2 = \begin{pmatrix} 1 & 0.72 \\ 0.72 & 0.52 \end{pmatrix}.$$

For the second method, note that the eigenvalues of a projection matrix needs to be equal to 1 or 0, here we have that they are 0.8 and 0.4. Note that a third way is the check that $Px = x$ for an element of the column space of P. However, if you do the computations (even for a general vector x) you will see that $Px = x$ only for $x = \mathbf{0}$.
Marks: 1 point for attempt, additional 0.5 points for each correct method used (max total point is 2).

3. In the proof of Lemma 17, it is written that "Thus $B\mathbf{X}$ and $L^T\mathbf{X}$ are independent (because they are jointly normally distributed)." Show that $B\mathbf{X}$ and $L^T\mathbf{X}$ are indeed jointly normally distributed. Moreover, what is the dimension of the matrix B? [3 Marks]
We know that $X \sim \mathcal{N}(\mu, I)$. Then, by Definition 23 from the lecture notes, we know that for a deterministic matrix D we have
$$DX \sim \mathcal{N}(D\mu, DD^T)$$
Then setting
$$D = \begin{bmatrix} B \\ L^T \end{bmatrix}$$
we get
$$\begin{bmatrix} B \\ L^T \end{bmatrix} X \sim \mathcal{N}\left(\begin{bmatrix} B \\ L^T \end{bmatrix} \mu, \begin{bmatrix} B \\ L^T \end{bmatrix} \begin{bmatrix} B \\ L^T \end{bmatrix}^T \right)$$
that is
$$\begin{bmatrix} BX \\ L^T X \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} B \\ L^T \end{bmatrix} \mu, \begin{bmatrix} B \\ L^T \end{bmatrix} \begin{bmatrix} B \\ L^T \end{bmatrix}^T \right)$$
Thus, $(BX, L^T X)$ are jointly normally distributed. Concerning the dimension of B, we need to make sure that BA can be computed. Since A is an $n \times n$ matrix, then B must be a $k \times n$ matrix, where $k \in \mathbb{N}^+$.
Marks: 1 for attempt, additional 0.5 point for showing joint normality, additional 0.5 point for correct dimension of B.

4. The dataset "World_Bank", which you can find on BB, is directly taken from the World Bank dataset and it contains the following variables for every country in the world[1]:
- Birth rate, crude (per 1,000 people)
- GDP per capita, PPP (current international $)
- Gini index[2]
- GDP per capita growth (annual %)
- Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)
- Inflation, consumer prices (annual %)
- Educational attainment, at least completed primary, population 25+ years, total (%) (cumulative)

We will consider *GDP per capita, PPP (current international $)* as our dependent variable.

(a) Consider the output of the R command `summary` applied to the linear model with the intercept and the covariate *Inflation, consumer prices (annual %)*. Describe how all the values in such output are computed giving also an explanation on the reason why they are computed in such way. [2 Marks]
Solutions: See the lecture on the Tooth growth example.
Marks: 1 for attempt, additional 0.5 for correct explanation of the p-value of the coefficient of the covariate *Inflation, consumer prices (annual %)*, additional 0.5 for explanation of the p-value of the F-statistic.

[1] I removed all countries that had missing values in at least one of the considered variables
[2] It is a measure of income inequality.

(b) Compare the linear model with covariates *Gini index*, *Inflation, consumer prices (annual %)*, *Birth rate, crude (per 1,000 people)* and *GDP per capita growth (annual %)* with the one with all the covariates. Which of these two models is to be preferred? Motivate your answer. [2 Marks]

Solutions: The model with 4 covariates is simpler and so more interpretable and has a higher adjusted R^2. Moreover, the additional two covariates in the full model are not significant, hence we cannot use them as (good) predictors for the GDP per capita. Therefore, the model with 4 covariate is to be preferred over the full model.

Marks: 1 for attempt, additional 1 point for correct solution.

(c) Compare the linear models in point (b) with the linear model with covariates *Gini index*, *Inflation, consumer prices (annual %)*, *Birth rate, crude (per 1,000 people)*, *GDP per capita growth (annual %)* and *Educational attainment, at least completed primary, population 25+ years, total (%) (cumulative)*. Which of these three models is to be preferred? Motivate your answer. [2 Marks]

Solutions: Since the model with 4 covariates is better than the the full model, we need just to compare this new model with the 4-covariate model. First, notice that this new 5-covariate model has a slightly higher adjusted R^2. Then, one can argue that this new model is better than the 4-covariate model. However, notice that the new covariate ($Educational\_attainment$) is not significant at all (p-value of 0.28). Thus, the 5-covariate model is not significantly better than the 4-covariate model and so we are not really improving our understanding of the underlying true model. Therefore, since the 4-covariate model is simpler than the 5-covariate model, the 4-covariate model is to be preferred to the 5-covariate model.

Marks: here there are 4 possible outcomes:

- 1.5 out of 2 points for saying that the 5-covariate model is better,
- 2 out of 2 points for saying that we cannot decide whether one model is better than the other,
- 2 out of 2 points for saying that the 4-covariate model is better.
- 1 out of 2 points if none of the above three outcomes occur.

(d) If you had to choose only one covariate to predict GDP per capita, which one would be the best? Which one would be the worst? Motivate your answer. [2 Marks]

Solutions: By looking at the values for R^2 the best is Birth rate and the worst is Labor force.

Marks: 1 for attempt, additional 0.5 for correct best covariate, and additional 0.5 for correct worst covariate.

5. Write down a statistical question you would like to answer (e.g. Does social media usage affect academic performance? What is a good predictor for the price of Bitcoin? Which economic factors drive up the market?). Next, select a relevant dataset of your choice. Explain why you selected it, describe its content, and specify its source. Then, conduct a statistical analysis using linear regression models and discuss the results. Be sure to include your code and the output in your solution. [Note: This question is designed to encourage exploration of a personally interesting topic without the pressure of achieving full marks. Hence, it will be graded generously, with a score of either 4 or 0.] [4 Marks]

Solutions: There is no general solution for this question.

Marks: 4 points if the student has done something and 0 otherwise. So I am expecting all students gets 4 points.