Coursework – Statistical Modelling 1 – Due at 13:00 on Friday, 22 March 2024

Ensure that your submission is clearly legible and well-organised.

1. Consider a linear regression model $\mathbf{Y} = \mathsf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Explain the difference between errors and residuals. In your answer you should also include whether or not the errors and the residuals are observable random variables. [2 Marks]
   Solution: The errors are $\boldsymbol{\epsilon} = \mathbf{Y} - \mathsf{X}\boldsymbol{\beta}$ while the residuals are $\boldsymbol{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathsf{X}\boldsymbol{\beta} - \mathsf{X}\hat{\boldsymbol{\beta}}$. The errors are not observable random variables because $\boldsymbol{\beta}$ (the true parameter) is not observed, while the residuals are observable because $\mathbf{Y}$, X and $\hat{\boldsymbol{\beta}}$ are all observable – in particular we observe $\mathbf{Y}$ and X and so we observe $\hat{\boldsymbol{\beta}}$ since $\hat{\boldsymbol{\beta}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\mathbf{Y}$)

2. Consider a linear regression model $\mathbf{Y} = \mathsf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Assume the Full Rank assumption and the Normal Theory assumption. Compute the maximum likelihood estimator (MLE) for the parameter $\theta = (\boldsymbol{\beta}, \sigma^2)$. Is this estimator unbiased for θ? Motivate your answer. [2 Marks]
   Solution: It is possible to notice that the likelihood is

$$L(\theta; \mathbf{Y}) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\Big(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathsf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathsf{X}\boldsymbol{\beta})\Big)$$

   and the log-likelihood is

$$l(\theta; \mathbf{Y}) = -\frac{n}{2}\log(\sigma^2 2\pi) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathsf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathsf{X}\boldsymbol{\beta})$$

   Fix σ. Note that maximizing $l(\theta; \mathbf{Y})$ with respect to β is equivalent to minimizing $(\mathbf{Y} - \mathsf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathsf{X}\boldsymbol{\beta})$ which gives you that $\hat{\boldsymbol{\beta}}_{MLE} = \hat{\boldsymbol{\beta}}_{LSE}$. Now, fix β. We have

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{Y} - \mathsf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathsf{X}\boldsymbol{\beta})$$

   by equaling it to zero and letting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{LSE}$ we obtain

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathsf{X}\hat{\boldsymbol{\beta}}_{LSE})^T(\mathbf{Y} - \mathsf{X}\hat{\boldsymbol{\beta}}_{LSE}) = \frac{RSS}{n}$$

   This is indeed a maximum because

$$\frac{\partial^2 l}{(\partial\sigma^2)^2}|_{\sigma^2=\hat{\sigma}} = \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}(\mathbf{Y} - \mathsf{X}\hat{\boldsymbol{\beta}}_{LSE})^T(\mathbf{Y} - \mathsf{X}\hat{\boldsymbol{\beta}}_{LSE}) = \frac{n}{2\hat{\sigma}^4} - \frac{n}{\hat{\sigma}^4} = -\frac{n}{2\hat{\sigma}^4} < 0$$

   Therefore, $\hat{\theta} = (\hat{\boldsymbol{\beta}}_{LSE}, RSS/n)$. It is possible to notice that $E[RSS/n] = \frac{n-p}{n}\sigma^2 \neq \sigma^2$. Hence, the estimator is biased.

3. Consider a linear regression model $\mathbf{Y} = \mathsf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Which are the minimal assumptions needed to define uniquely the least square estimator? Motivate your answer. [2 Marks]
   Solution: The minimal assumption needed to define uniquely the least square estimator is the Full Rank assumption, as mentioned in the lecture notes in page 68.

4. Let f be the pdf of a standard Normal distribution and let g_n be the pdf of a student's t-distribution with n degrees of freedom. For what value of n we have that $\int_{\mathbb{R}} |f(x) - g_n(x)|dx < 0.01$? (You can use a program to compute it, if you do Please add your codes and the output of your codes to the solutions) [2 Marks]
   Solution: $n = 64$.

5. In the Higgs Boson announcement (check https://www.youtube.com/watch?v=0CugLD9HF94), the speaker affirmed that they obtained a "significance of 5σ". Why does this result support the existence of the Higgs Boson? Can you guess the hypothesis tests they used? *(Note: their pivotal quantity follows a Normal distribution).* [2 Marks]
Solution: We have that H_0 : Higgs Boson does not exists vs H_1 : Higgs Boson exists. Suppose that X is the pivotal quantity and $X \sim N(\mu, \sigma^2)$. Without loss of generality we can consider $\mu = 0$ (otherwise we would just consider $\mu + 5\sigma$ as our observed value). Since they obtained 5σ, the p-value is $P(|X| \geq 5\sigma) = 2P(X \geq 5\sigma) = 0.00000058$. Thus, we can reject H_0 to the level $\alpha = 0.00000058$.

6. The dataset "btc_2015_2024", which you can find on BB, offers a detailed examination of Bitcoin's price behavior over the last eight years, featuring a range of technical indicators for analyzing its trends. It captures the daily opening, highest, lowest, and closing prices, along with trading volume. The dataset includes momentum indicators such as the 7-day and 14-day Relative Strength Index (RSI) to determine if the asset is overbought or oversold. It also contains the 7-day and 14-day Commodity Channel Index (CCI), which compares the current price to the historical average to spot short-and medium-term trends. Additionally, it encompasses moving averages like the 50-day and 100-day Simple Moving Average (SMA) and Exponential Moving Average (EMA), which shed light on the asset's trend direction. Other essential indicators in the dataset are the Moving Average Convergence Divergence (MACD), Bollinger Bands for assessing price volatility, the True Range, and the 7-day and 14-day Average True Range (ATR) that provide insights into market volatility. We want to predict the price of the bitcoin, hence we consider the variable $open$ as our dependent random variable.

   (a) Consider the output of the R command `summary` applied to the linear model. Describe how all the values in such output are computed giving also an explanation on the reason why they are computed in such way. [1 Mark]
   Solution: see Tooth growth example (in the annotated slides of Lecture 19).

   (b) Compare the linear model with covariates ema_50 and sma_50 and the one with covariates ema_100, sma_100, rsi_14, rsi_7, $volume$, $TrueRange$, atr_14, and atr_7. Which of these two models is to be preferred? Motivate your answer. [2 Marks]
   Solution: The first model has higher R^2 and lower number of covariates than the second model. Hence, it should be preferred.

   (c) Compare the linear models in point (b) with the linear model with covariates sma_100, rsi_14, $volume$, $TrueRange$, atr_14, atr_7, cci_14, and $macd$. Which of these three models is to be preferred? Motivate your answer.
   Solution: This third model has slightly higher R^2 than the other two models of point (b). However, it has more covariates than the first model of point (b). Hence, here there is no clear choice of which model is better. The first model is more interpretable while the third one explain the data slightly better. Surely, the second model of point (b) is the worst one.
   That said, solutions that mention the third model as the best one (due to its higher R^2) are also accepted as correct. [2 Marks]

7. Take a real dataset of your choice (possibly related to an area you are passionate about). Explain the dataset, perform a statistical analysis using linear regression models and explain the outcome of your statistical analysis. Please add your codes and the output of your codes to the solutions. (Note: this question has only two points because you get 2 points if you provide a good (but necessarily very good) statistical analysis – for example you get two points if you discuss the outcome of a linear model and you compare it with at least another linear model with different covariates. Thus, I want that you explore something you are interested in

without the pressure of not getting the total marks for this question) [2 Marks]
Solution: There is no unique solution for this question. But if you do what is asked in the Note of this question, you should get 2 points.

8. A group of economists want to study the productivity of firms. To fit their model, they have collected $n = 100$ firms and recorded the annual profit y_i of each firm measured in millions of pound. The following characteristics were also recorded for each firm: number of workers $x_{1,i}$, average level of education of the workers x_{2i} measured in years spent studying (in school and university), having a female CEO or not $x_{3,i}$ (in particular $x_{3,i} = 1$ if firm i has a female CEO and $x_{3,i} = 0$ otherwise). The economists will use the model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i, \tag{1}$$

where the errors are i.i.d. and $\epsilon_i \sim N(0, \sigma^2)$, where σ is unknown. Assume that the full rank assumption is satisfied.

The following information is obtained after using least squares to fit the linear model:

Coefficient	Point estimate ($\hat{\beta}$)	p-value*
β_0	0.679	0.0405
β_1	0.102	0.0010
β_2	1.886	0.0545
β_3	2.127	0.0647

$R^2 = 0.74$ and $n = 100$.
*All p-values were computed as $p = P(|T| > |t|)$ for $T \sim t_{96}$.

(a) From this statistical analysis, what would be the predicted difference in profit between a firm with a male CEO and a firm with a female CEO? Is this significant? [1 Mark]
Solution: The predicted difference is 2.127 million pounds. In particular, firms with female CEOs have 2.127 in profit more than firms with male CEOs. This values is significantly different from zero at level of significance 0.0647. Thus, for $\alpha = 0.1$ we reject the null hypothesis that this effect is significant, however we do not reject it at level $\alpha = 0.05$.

(b) Add realistic covariates of your choice to the model (1) so that the full rank (FR) assumption is violated. Motivate your answer, that is motivate the choice of the covariates and explain the reason why the FR assumption is violated. [2 Marks]
Solution: There is no unique solution for this question. One example is given by the introduction of $x_{4,i}$ and $x_{5,i}$ in the model, where $x_{4,i}$ numbers of female workers and $x_{4,i}$ numbers of male workers of firm i (for every $i = 1, .., n$). In this case we have $x_{4,i} + x_{5,i} = x_{1,i}$ for every i and so the FR would be violated.