**KUNSKAPSKONTROLL (PART 2)**

**THEORITICAL QUESTIONS AND**

**GROUP WORK QUESTIONS (DATACOLLECTION)**



Aikaterini Antoniou

EC Utbildning

Course: R Programmering for dataanalysis

2024-04

**THEORITICAL QUESTIONS**

1. A Quantile-Quantile (QQ) plot is a visual (graphical) tool used to compare the distribution of a dataset against a theoretical distribution, typically to assess normality. It plots the quantiles of the sample data against the expected quantiles of the theoretical distribution, usually the normal distribution. If the dataset follows the expected distribution, the points will align closely with a straight line. Deviations from this line suggest discrepancies from the theoretical model, highlighting issues like skewness or outliers in the data. This tool is valuable for verifying assumptions in statistical tests that require normality.(Wikipedia,2024)

2. I would say to Karin that it is a great observation. In machine learning the primary goal is prediction. This field uses algorithms to process data, learn patterns from data and make predictions about future events based on past data which means predicted outcomes for new, unseen data. It's heavily focused on performance and often employs complex models that can handle large datasets and intricate patterns but might not provide much insight into the underlying causes or relationships. That means that the main goal is to optimize for accuracy and minimize prediction errors, often without a deep emphasis on understanding why or how decisions are made by the model.

In contrast, statistical regression analysis is used not only to predict outcomes but also to understand the underlying dynamics between variables. For instance, using regression, we can quantify how different factors like education level and working hours affect salary. This is what we mean by statistical inference. It allows us to test hypotheses and make conclusions about the data's structure and the significance of variables. Another example is when in a linear regression model predicting house prices we can infer how much increase in price can be expected for a unit increase in square footage or how the location affects the price, while controlling for other factors.

So, while machine learning models are typically used for their predictive accuracy without necessarily explaining why outcomes are predicted as they are, statistical regression models can provide explanations and quantify relationships, which is invaluable for understanding complex processes and making informed decisions.(Stewart, 2019)

3. A confidence interval primarily focuses on estimating the uncertainty of a parameter within a statistical model, such as the mean outcome from a set of predictors in regression analysis. It gives us a range within which we believe the true mean value lies for a given level of confidence (typically 95%). This interval reflects the accuracy of the model's parameters but does not include the variability observed in individual data points.

A prediction interval, on the other hand, provides a range around a single predicted response and is used to account for the uncertainty in predicting a single new observation. This interval is always wider than the confidence interval for the same predicted mean response because it includes both the uncertainty of the model estimate (like the confidence interval) and the variability of new data points around that estimated regression line. It tells us where a future observation will fall, at a specified level of confidence, considering both the model's prediction and the individual data point's potential deviation. So, while the confidence interval addresses the question: "Within what range do we expect the average outcome to fall", the prediction interval answers: "Within what range should we expect a new single observation to fall".(Parth,2020)

4. In a multiple linear regression model the beta coefficients (β) signify how much the dependent variable (Y) is expected to increase when the corresponding independent variable (X) increases by one unit, assuming that all other variables in the model are held constant. More specifically each component is interpreted as following:

𝛽₀ (Intercept): This coefficient provides the estimated value of Y when all the X variables are zero. It's the starting point of the regression line when it crosses the Y-axis.

𝛽₁, 𝛽₂, ... 𝛽p (Slope Coefficients): Each of these coefficients corresponds to a different predictor variable (X₁, X₂, ... Xp). The value of a beta coefficient represents the change in the outcome associated with a one-unit change in the corresponding predictor while holding other predictors in the model constant. For example, 𝛽₁ associated with X₁ indicates how much Y will change for each one-unit increase in X₁, assuming all other variables remain the same.

𝜀 (Error term): This represents the residual effect, not explained by the linear model. It captures the deviation of the observed values from the line defined by the model, due to factors not included in the model. (EasyMedStat,2023)

5. Hassan is not entirely correct to say that using Bayesian Information Criterion (BIC) in statistical regression modeling eliminates the need for dividing dataset into training, validation, and test sets. While BIC is useful for model selection, helping to choose a model that balances fit and complexity by penalizing free parameters, it doesn't substitute for the validation process.

The main function of BIC, (and similarly Akaike Information Criterion (AIC)), is to compare models by penalizing those with more parameters, which might otherwise appear overly accurate due to overfitting. These criteria help identify a model that's likely efficient at prediction while avoiding overfitting to the training data.

However, using training, validation and test sets is essential for evaluating how a model performs on new data, which BIC alone cannot ascertain. This split ensures that the evaluation of the model’s performance reflects its ability to generalize beyond the specific data on which it was trained. Thus, while BIC is a valuable tool for model comparison, it does not obviate the need for traditional train-test splits in assessing model generalizability and robustness.

6. The algorithm for "Best subset selection" is a method used in regression analysis to choose a subset of predictors that results in the best fitting model. The algorithm is characterized by the following steps:

1.Null Model: The procedure initiates with a null model that contains no predictors. This model predicts the sample mean for each observation.

2.For each number of predictors k, ranging from 1 to the total number p:

-Models containing exactly k predictors are fitted.

-The best among these models is selected, defined by the smallest Residual Sum of Squares (RSS) or the largest R-squared (R²) value.

3.The optimal model is selected from among the best models for each number of predictors. This selection is based on the prediction error on a validation set, considering criteria such as Cp (AIC), BIC, or adjusted R², or by employing the cross-validation method.

In this algorithm a comprehensive search is conducted across all possible combinations of variables to identify the model that balances simplicity with a high degree of fit to the data.

7. The quote by George Box “All models are wrong, some are useful” emphasizes the concept that no statistical model can capture the full complexity of reality. Models, by necessity, are simplifications that abstract away certain details to focus on the main features of a phenomenon. They are "wrong" in the sense that they are incomplete representations but they can still be "useful" if they capture enough of the essential details to be applicable in practical situations.

The statement suggests a pragmatic approach to modeling: the goal is not to find a perfect model but one that serves a particular purpose well enough. For instance, a model of climate change might not be able to predict every variable in the Earth's climate system but if it can reliably predict trends it is still valuable.

In this light, the value of a model is not measured by its perfection but by its utility. A model that offers actionable insights or a reasonable approximation of the truth can be a powerful tool for analysis and decision-making. Thus, when using models, one should be aware of their limitations, continuously test their predictions against real-world data and refine them as needed, always considering their utility in the context they are applied.

**REFERENCES**

1.EasyMedStat(2023).The Beta (ß) Coefficient in Multivariate Linear Regression <https://help.easymedstat.com/support/solutions/articles/77000538788-the-beta-%C3%9F-coefficient-in-multivariate-linear-regression#:~:text=What%20is%20the%20beta%20coefficient,holding%20all%20other%20predictors%20constant>. (2023, Feb 01)

2. Parth, S. (2020). How Confidence and Prediction intervals work. Towards Data Science. Retrieved from https://towardsdatascience.com/how-confidence-and-prediction-intervals-work-4592019576d8 . (2020, Feb 04)

3. Stewart, M. (2019). The actual difference between statistics and machine learning. Towards Data Science. Retrieved from <https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3> . (2019, March 25)

4. Wikipedia. (2024).Q-Q plot. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Q%E2%80%93Q\_plot (2024, April 05)

**GROUP WORK QUESTIONS (DATA COLLECTION)**

1.For the group work portion of the project, collaboration was undertaken with Anders, Jacob, Manna, Lidiia, Girlie, and Nil.

2. In the group, the collaboration was structured and systematic. Each member was assigned specific tasks and regular meetings were held to discuss progress, brainstorm ideas and address any challenges while creating and using shared files everyone could see the progress in real time. Open communication was maintained throughout the project through communication channels (Teams), allowing for an exchange of knowledge and feedback, which contributed to a cohesive final outcome. The group's dynamics were positive, fostering a productive work environment that was conducive to creativity and mutual learning.

3. The group work was marked by a collaborative spirit and a collective effort to meet the first part´s goals. The division of tasks (gathering one year´s data for each member) and the creation of shared files resulted in a well-rounded approach to the work. There was a fluid exchange of ideas and when challenges arose, the group demonstrated commendable flexibility, problem-solving skills and mainly willingness to solve problems.

What could possibly be improved so that there is maximum development is the consequence in the delivery times of the responsibilities by all members of the team so that there is no delay and time limitation in workflow for the rest.

4.When working in a group, one of my strengths is my ability to listen actively and synthesize diverse viewpoints. This helps in building consensus and fostering a collaborative environment. In group settings, I'm good at organizing and ensuring tasks are completed accurately, which contributes to meeting our objectives effectively. However i want to enhance my presentation skills to more effectively communicate our results, which will help in advocating for our group's work and ideas more persuasively. This improvement will make my contributions even more valuable to the team.

5. If I were to change something it would be the presence of all the members of the group from the beginning, to start the group work and discussions at the same time, all together, because when a member comes later and with limited time there is no possibility to go back from the beginning, something that makes it difficult for both the new member and the team together.