35. **Analyzing Medication Dosage Predictors Using Linear Regression and Stepwise Selection A Data Science Approach**

**Abstract:**

This study explores the relationship between several demographic and physiological variables, such as Age, Gender, BMI, and Cholesterol levels, in predicting the dosage of a medication. Utilizing multiple linear regression and stepwise selection models, we examine the significance of these variables in influencing medication dosage. The findings reveal no significant predictors, with low R-squared values across models, suggesting the limited explanatory power of these variables. The analysis utilizes extensive statistical methods in R, including Variance Inflation Factor (VIF) calculations to check multicollinearity and ANOVA tables to validate model assumptions.

**1. Introduction**

The primary goal of this research is to identify key demographic and physiological predictors influencing the dosage of a specific medication, represented by the variable Dosage\_Med1\_mg. We employed a multiple linear regression approach, incorporating demographic factors (Age, Gender, BMI) and physiological markers (Cholesterol levels). Furthermore, a stepwise selection model was applied to refine the model and improve interpretability. This study aims to determine whether these predictors significantly influence medication dosage and assess the model's overall fit.

**2. Methodology**

We used the R programming language for statistical analysis, focusing on multiple linear regression and stepwise selection models to identify the most relevant predictors for Dosage\_Med1\_mg.

* **Linear Regression Model:**

Dosage\_Med1\_mg∼Age+Gender+BMI+Cholesterol\_mg\_dLDosage\\_Med1\\_mg \sim Age + Gender + BMI + Cholesterol\\_mg\\_dLDosage\_Med1\_mg∼Age+Gender+BMI+Cholesterol\_mg\_dL

* **Stepwise Selection Model:** A combination of forward and backward selection was employed to refine the model. The variables were added or removed based on their significance, using the Akaike Information Criterion (AIC) as the selection metric.
* **Diagnostic Analysis:**
  + **Variance Inflation Factor (VIF):** To check for multicollinearity among predictors.
  + **Residual Analysis:** Residuals vs. fitted values plots to evaluate the assumptions of homoscedasticity and linearity.

**3. Results**

**3.1 Linear Regression Model Output:**

* **Coefficients Table:**
  + Age: Estimate = 0.111359, Std. Error = 0.089247, t-value = 1.248, p-value = 0.212
  + Gender (Male): Estimate = -4.968817, Std. Error = 3.765255, t-value = -1.320, p-value = 0.187
  + Gender (Other): Estimate = -12.116377, Std. Error = 9.546798, t-value = -1.271, p-value = 0.204
  + BMI: Estimate = -0.518637, Std. Error = 0.369638, t-value = -1.403, p-value = 0.161
  + Cholesterol (mg/dL): Estimate = -0.007314, Std. Error = 0.031929, t-value = -0.229, p-value = 0.819
* **Model Diagnostics:**
  + Residual Standard Error: 140.8 on 5828 degrees of freedom.
  + Multiple R-squared: 0.001211, Adjusted R-squared: 0.0002641.
  + F-statistic: 1.308 on 5 and 5828 DF, p-value: 0.2573.

**Interpretation:**

None of the predictor variables showed statistical significance (p > 0.05), indicating that Age, Gender, BMI, and Cholesterol levels do not significantly affect the dosage of the medication. The R-squared value is extremely low (0.001211), suggesting that the model explains only 0.12% of the variance in the dosage. This indicates poor model fit and limited predictive power.

**3.2 Residual Analysis:**

The residual plot (Residuals vs. Fitted Values) for the stepwise selection model reveals a random scatter around zero, suggesting that the assumptions of homoscedasticity and linearity are met. However, the residuals show substantial variability, reinforcing the model's limited predictive capability.

**3.3 Variance Inflation Factor (VIF) Analysis:**

* Age: VIF = 1.001216
* Gender: VIF = 1.000680
* BMI: VIF = 1.001216
* Cholesterol (mg/dL): VIF = 1.000552

All VIF values are approximately 1, indicating no multicollinearity among the predictors. This suggests that the predictors are independent, but they still fail to significantly explain the variation in dosage.

**3.4 Stepwise Selection Results:**

* The stepwise selection process, guided by AIC, retained all initial predictors. However, the model's adjusted R-squared value remains low (0.0002641), indicating that the refined model does not perform better than the full model.

**3.5 Boxplot Analysis by Ethnicity:**

The boxplot for medication dosage across different ethnic groups (African-American, Asian, and Caucasian) shows similar median dosages across all groups, with a considerable overlap in interquartile ranges. This suggests no apparent differences in dosage based on ethnicity.

**4. Discussion**

The findings indicate that Age, Gender, BMI, and Cholesterol levels do not significantly impact the dosage of medication, as reflected by the low R-squared values and lack of statistical significance in the coefficients. Despite the use of stepwise selection to optimize the model, the improvement in model fit was negligible.

These results suggest that other unexamined factors may better explain the variance in medication dosage. Future research could explore additional predictors, such as genetic factors, lifestyle variables, or comorbidities, to develop a more comprehensive model.

**5. Conclusion**

This study demonstrates that the selected demographic and physiological variables have limited predictive power for medication dosage. Further investigation is needed to identify more robust predictors to inform dosage decisions effectively. The analysis emphasizes the importance of rigorous statistical testing and highlights the challenges of predictive modeling in clinical contexts.