Plan of work for the fourth session, November 16

1. Project no 1 – comments (orginality od data sets, ChatGPT, defense).
2. Multiple regression. Fitting the model. Selection of variables. VIF. Collinearity.

Transformations of variables. Square root transform, Quadratic transformation, Log transformation.

1. Selection of the best subset of variables. Model building flowchart.
2. Sensitivity analysis. DFFITS and DFBETAs. Cross-validation.
3. Multiple regression. Selection of variables.
4. Introduction to classification. Practical exercises.

Class notes. Class of Nov 16, 7:30 – 10:00.

Equation of the multiple regression regression model in matrix term

**Y = Xbeta + epsilon, X – design matrix**

Estimate of the vectorial parameter beta is obtained in the following way

**(beta)hat = ( Xtr X)^(-1)XtrY**

**( Xtr X)^(-1)** - inverse matrix. Collinearity – linear dependence (in the sense of algebra) of several columns of the design matrix **X**

When we have collinearity – calculation of **( Xtr X)^(-1)** is numerically unstable.

Stepwise regression.

* **Forward method**. Start – select 1st variable that has the highest correlation with Y. Then continue with the 2nd highest correlated until you reach no significant increase of adjusted R square. Goal – reduce the number of variables without compromising the quality of the regression model.
* **Backward method**. Start – full model. Reduce 1st variable that has the smallest correlation with Y. Check for some drop in R square adjusted. If no significant drop – reduce 2nd variable that has the smallest correlation with Y. Stop when significant drop of R square adjusted is obtained.

Use also Cp Mallow’s coefficient. Cp can be calculated when n – no of observations is MUCH larger than the number of variables. Usually it is recommended that n >= 15\*p, p – no of variables.