Preamble:

- Regression modeling is an iterative process in which, based on outcomes of the model diagnostics, the structure of the regression model needs to be revised.
- It requires additional knowledge and experiences beyond GISC6301 and will be trained in the sequel GISC7310.

Data Preparations

- Add row names to the spatial dataframe by: **`row.names(map.shp@data) <- 1:length(map.shp)`**. Notice the syntax **`map.shp@data`** to directly address the attributed table of the spatial dataframe.

Hypotheses Formulation

- State explicit hypotheses about the direction that an independent variable is influencing the dependent variable. If you are uncertain about the direction speculate which argument could lead to a positive and/or negative influence.

Data Exploration

- Evaluate the spatial pattern of the dependent variable and perhaps the independent variables. Are there spatial patterns and outlier (values that don't match the surrounding environment)? Do spatial patterns co-vary?
- Generate a scatterplot matrix. Make sure that the dependent variable is the first variable in the right-hand side formula.

- Interpret the univariate distributions on the diagonal. Are variable transformation advisable to achieve approximate symmetry.
- Evaluate the relationships between the dependent variable and the independent variables. In which direction point the relationships and are there potential non-linearities.
- Evaluate the relationships among the independent variables.
- Repeat the analysis of the scatterplot matrix after variables have been transformed.

Regression Analysis

- Run a regression with the full set of independent variables, which potentially are transformed to approximate symmetry. Evaluate the model and remove irrelevant variables one at a time.
- Interpret the final regression model.
  - Have the regression coefficient the expected sign?
  - Are the relevant. Look at the stars and drop variables without a star.
  - Discuss the adjusted R^2.

Regression Diagnostics

- With the **`vif( )`** function evaluate the degree of multicollinearity. Highly redundant variables have a value of 10 or larger. If the associated variables are relevant and have the expected sign don't do anything. Otherwise, drop one offending variable at a time and rerun the regression model.
- With the **`qqPlot( )`** function evaluate whether the regression residuals are approximately normally distributed. Ideally all observations are on the diagonal line, but random variations within the confidence bands are acceptable. Observation outside the bands, usually in the tails, indicate deviation from the normal distribution and are potential outliers. The most extreme observations are labelled by their row index.

Remember: the residuals measure the difference between the observed and the predicted values. Thus, how far does a particular observation deviate from the general model structure.

- With the **`residualPlots( )`** function evaluate whether individual variables should be added as quadratic term to the regression model. Only consider adding quadratic relationships **`y~x+I(x^2)`** to your model if the quadratic effect is strong.
- With the **`influenceIndexPlot( )`** function evaluate the data for potential outliers and influential observations.
  - The Cook distance is a measure of how the set of regression coefficients would change if the i^{th} observation is dropped. Large Cook distances relative to the other Cook distances indicate that the i^{th} observation is influential.
  - The studentized residuals indicate the presence of potential outliers. Values outside the interval $[-2,2]$ could be potential outliers and extreme relative to the distribution of the residuals.
  - The Bonferroni p-value measures the extremity of the regression residuals. Observations with a value less than 0.05 require careful investigation. Are they members of the population that we want to analyze? Why are these observations deviating from the general model structure?
    Note: the scale of the Bonferroni p-value may not stretch down to 0.05 or lower. In this case there are no relevant outliers.
  - The hat-values measures how far the independent variables for the i^{th} observation deviate from the ensemble of the independent variables of the other variables.

Adding Factors

- It is possible to add factors to a regression model. They modify the intercepts for each factor level.
- Due to technical issues one factor needs to be suppressed. The general intercept b_0 is associated with this suppressed factor level.

- The remaining intercepts are $b_0 + b_{level_g}$.
- It is advisable to only add factors to a modal if they improve the model fit substantially.
- Note: Due to multicollinearity of the factor with the metric independent variables some regression coefficients may no longer differ substantially from zero.

Spatial Pattern of Residuals

- Remember: The neutral value of regression residuals is zero.
- Make sure that the class intervals in the lower and upper branch are approximately equally populated with observations.
- Scan the pattern in the regression residuals for clusters in negative and positive values as well as spatial outliers, i.e., positive residual surrounded by negative residuals and vice versa.
- Spatial patterns among the residuals hint at an underlying spatial process which ties the observations spatially together. This can be further investigated with spatial autocorrelation analysis techniques.