Lab04: Extended Topics of Regression Analysis

**Handed out:** Monday, April 5, 2021

**Return date:** Friday, April 16, 2021

**Grading:** This lab counts 18 % towards your final grade.

**Objectives:** *You will build, analyze and interpret an election regression model that explains the aspatial and spatial variation of median home values 2012 in Texas using its 254 counties.*

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as Word file at the submit tab in the Lab04 folder in eLearning. Add a running title with the following information: Lab04, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. Trial and error answers will lead to a deduction of points. Label each answer properly with the bold task headings. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code and output. Use mathematical type-setting when equations are required. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

# Data

## Data Files

The zipped file **TXCnty2020.zip** contains the shape file for the Texas counties with their attribute information (**TXCnty2020.shp**), a shape file outlining neighboring US and Mexican states (**TXNeighbors.shp**) and a shape file of the major highway system (**InterStateHwy.shp**). It also includes the document **TXCntyVars2020.pdf** with the data dictionary and basic descriptive statistics.

# Analysis and Modelling Tasks

## Analysis Tasks

**[1] Specification of the Dependent Variable (2 points)**

The dependent variable is the median home value in dollars (**MEDVALHOME**) of owner-occupied housing units for the 254 counties of Texas.

[a] Identity the ***underlying statistical universe*** (a.k.a ***population at risk***) from which the median home value was estimated. This variable will be use later as weights variable to investigate any potential heteroscedasticity in your regression model.

[b] Carefully evaluate the ***distribution*** of the median home value. Are there any missing observations that you need to deal with?

[c] Map the spatial distribution using quantiles and interpret the map pattern.

Note: The  mapping function uses quantiles; therefore, your map pattern will look slightly different from that shown in the back of your handout, which uses fixed intervals.

**[2] Selection of Independent Variables (2 points)**

[a] Identify 4 to 6 potential independent *metric* variables plus at least one *factor* that you expect will influence the median home values.

[b] Formulate common-sense hypotheses why and which direction these potential independent variables will influence the median home values.   
Document items 2 [a] and [b] in a table.

|  |  |  |
| --- | --- | --- |
| Variable | Rationale | Statistical Hypotheses |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**[3] Exploration of Variables (3 points)**

In a scatter plot matrix or, where appropriate, box-plot:

[a] Explore the univariate distribution of the dependent variable.

[b] Explore the relationship of the independent variables and factor(s) with the dependent variable.

[c] Explore the univariate and bivariate distributions of the independent metric variables.

[d] Does this exploration point at any variable transformations for your initial regression model?

Note: If required only consider transformations rounded to the closest 0.5 increments with a preference for the *log*-transformation, if required with an additional shift parameter. The log-transformation is preferred because it allows to interpret any relationships in percentage changes.

At this point redo the scatterplot matrix or box-plot with the any selected variable transformation.

***Your initial trial model should already incorporate these transformations.***

**[4] Initial Trial Regression Model (4 points)**

Proceed in your initial analysis with ordinary least squares. Based on the selected variables build an *initial trial* model and perform a thorough aspatial model diagnostics. Provide supportive plots and statistics.

Guiding questions are:

1. Are all selected variables and factors relevant and do their regression coefficients exhibit the expected sign?
2. Is multicollinearity a problem?
3. Are the model residuals approximately normally distributed?
4. Do you need to refine the variable transformations or add quadratic terms?
5. Are there influential cases and outliers present in the model?
6. Speculate why some observations appear to be “extreme” and decide what to do with these observations: Do you need to drop the associated counties from the analysis because they are not representative of the underlying population or have “unstable” variable values?

**[5] Revised Regression Model (2 points)**

[a] Based on what you have learned about your model in Task 4 build a *revised* regression model and re-check its properties. Are all identified problems from Task 4 ─ at least to some degree ─ addressed? Make sure to work with at least 4 meaningful metric variables and if the selected factor remains relevant, then keep it.

[b] Fully interpret your final model. Does it support the hypotheses that you have formulated in Task 2?

**[6] Heteroscedasticity Investigation (2 points)**

Note: The size of the reference population varies widely from county to county. Use the revised model from Task 5.

[a] Estimate and interpret the parameters of the multiplicative heteroscedasticity model .   
[b] Interpret the likelihood ratio test whether it is necessary to account for heteroscedasticity.  
[c] Interpret the regression parameters of your independent variables with regards to whether they or their significances are substantially different from those of your revised OLS model in Task 5.

**[7] Spatial Residual Analysis (3 points)**

For the spatial residual analysis, you can proceed either with the refined OLS model from Task 5 or, if there is significant heteroscedasticity, with heteroscedasticity model from Task 6.

[a] Map the regression residuals of your refined OLS model in a choropleth map with a bi-polar map theme broken around the neutral zero value.   
*Interpret* the observed map pattern of positive and negative residuals.

[b] Generate the spatial links and plot its graph onto a map of the Texas Counties. Check whether this graph is connecting all counties properly.

[c] Generate a Moran scatterplot of the regression residuals and interpret it.

[d] Test with the Moran’s *I* statistic whether the regression residuals of your final model are spatially independent or exhibit spatial autocorrelation.

**[8] Estimate a Spatial Autoregressive Model (2 points)**

For the SAR model you can proceed either with the refined OLS model from Task 5 or, if there is significant heteroscedasticity, with heteroscedasticity model from Task 6.

[a] Estimate a spatial autoregressive regression model and test with a likelihood ratio test whether the spatial autoregressive model improves significantly over your refined OLS model in Task 5 (or in Task 6 if heteroscedasticity is present).

[b] Interpret the model. What is the spatial autocorrelation coefficient telling you? Are the estimated regression coefficients of the autoregressive model and their significances substantially different from the refined OLS model in Task 5 (or in Task 6)?

[c] Test the residuals of the autoregressive model for spatial autocorrelation and comment on the result.