Lab04: Extended Topics of Regression Analysis

**Handed out:** Monday, March 30, 2020

**Return date:** Friday, April 17, 2020

**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 the percentage of votes for either Trump or Clinton in the presidential election in 2016 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 data are documented in the file **TXCntyVoteVars2016.pdf** which can be found in the zipped file **TXCnty2018.zip** together with the necessary data.

Three ESRI shape files are packed into **TXCnty2018.zip**. All are in the ***long/lat*** format and must be imported in  with the option **proj4string=CRS("+proj=longlat")** to map properly projected. Since these files were digitized for high resolution maps it may take  a several seconds to draw the maps. Be a little patient and try to get right the first time!

The file names are:

* **TXCnty.shp**: *Area layer* with the 254 counties of Texas. Its associate **dBase** file holds the attribute information for this analysis.
* **TXNeighbors.shp**: *Area layer* with the neighboring states of Mexico and the United States of America. You *may* use this shape file as reference frame for the Texas counties.
* **InterStateHwy.shp**: *Line layer* of the main highways in and around Texas. You *may* use this shape file as spatial reference frame to locate the Texas counties.

## Variable in TXCnty.dbf

All variable for this project were retrieved from [a] the American Community Survey by the U.S. Census Bureau (2012 release of the 5 years summary estimate), [b] the presidential election results for 2012 are documented in [www.data.gov](http://www.data.gov), and the 2016 presidential election results can be found at <http://www.sos.state.tx.us/index.html>, [c] the Texas Department of Public Safety for 2015 (<https://www.dps.texas.gov/>), the Texas Department of Transportation (<http://www.txdot.gov/inside-txdot/division/finance/discos.html>) for 2015/2016 and [d] calculated by Caliper Inc.

# Analysis and Modelling Tasks

## Analysis Tasks

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

You are given the *absolute counts* of votes for Trump (**TRUMPVOT16**), Clinton (**CLINTONVOT**) and others[[1]](#footnote-1) (**OTHERVOT16**), as well as the number of persons 18 years and older[[2]](#footnote-2) (**POP18PLUS**), number of registered voters[[3]](#footnote-3) (**REGVOT16**) and the turnout rate[[4]](#footnote-4) (**TURNOUT16**).

[a] Calculate the ***percentage of voters*** who voted for either candidate. Be careful to select the proper reference population in the denominator. *Justify your calculation*.

[b] Evaluate the ***distribution*** of both percentages and chose that candidate those percentage distribution is easier to transform to symmetry. Map the percentage of voters of your candidate and interpret its spatial distribution.

[c] Can all 254 counties be used in the analysis or do a few counties have a too small denominator, thus leading to instable percentage estimates.

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 in 10% increments.

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

[a] Identify 4 to 6 potential independent *metric* variables plus at least one *factor* that you expect to influence the proportion of voters.

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

**[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)**

Even though the dependent variable is a rate and therefore technically follows a binomial distribution, proceed in your analysis with ordinary least squares, which is approximately valid. Based on the selected variables build an *initial trial* ordinary least squares regression 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] Build a *revised* regression model and re-check its properties. Are all identified problems from item 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] 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 underlying the voters’ percentages for selected candidate varies widely from county to county. Use the model structure 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 item 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 item 5.

[b] Interpret the model. What is the spatial autocorrelation coefficient? Are the estimated regression coefficients of the autoregressive model and their significances substantially different from the refined OLS model in item 5?

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

1. Besides the two main candidates, the electorate also has had a choice to vote for independent candidates and Libertarians. Only a very small number of voters in each county has chosen these alternatives. [↑](#footnote-ref-1)
2. Note that not all persons 18 years and older qualify to vote; for instance, because some are not U.S. citizens. [↑](#footnote-ref-2)
3. In Texas, voters need to register in order to be eligible to vote. This does not imply that all registered voters will participate in an election. [↑](#footnote-ref-3)
4. The turnout percentage is that proportion of registered voters who participate in an election. [↑](#footnote-ref-4)