

Yalin Yang 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.*

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.

Analysis and Modelling Tasks

Analysis Tasks

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

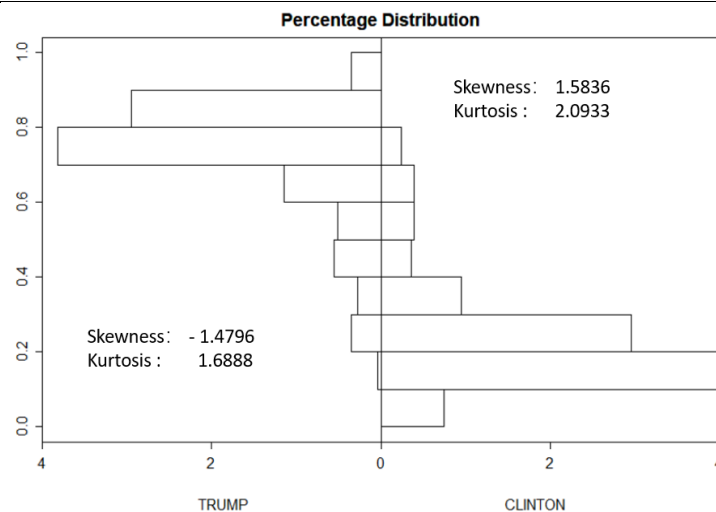
You are given the *absolute counts* of votes for Trump (**TRUMPVOT16**), Clinton (**CLINTONVOT**) and others¹ (**OTHERVOT16**), as well as the number of persons 18 years and older² (**POP18PLUS**), number of registered voters³ (**REGVOT16**) and the turnout rate⁴ (**TURNOUT16**).

[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.*

```
ct.shp <- rgdal::readOGR(dsn=getwd(), layer="TXCnty",
integer64="warn.loss")
ct.data <- as.data.frame(ct.shp)
ct.data$Voters <- ct.data$REGVOT16 * ct.data$TURNOUT16
ct.data$TRUMPRate <- ct.data$TRUMPVOT16 / ct.data$Voters
ct.data$CLINTONRate <- ct.data$CLINTONVOT / ct.data$Voters
```

For evaluating the percentage of voters who voted for either Clinton or trump, the research population should be all voters who actually participate in an election. That is why I select the total number of registered voters multiple turnout percentage as my denominator.

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



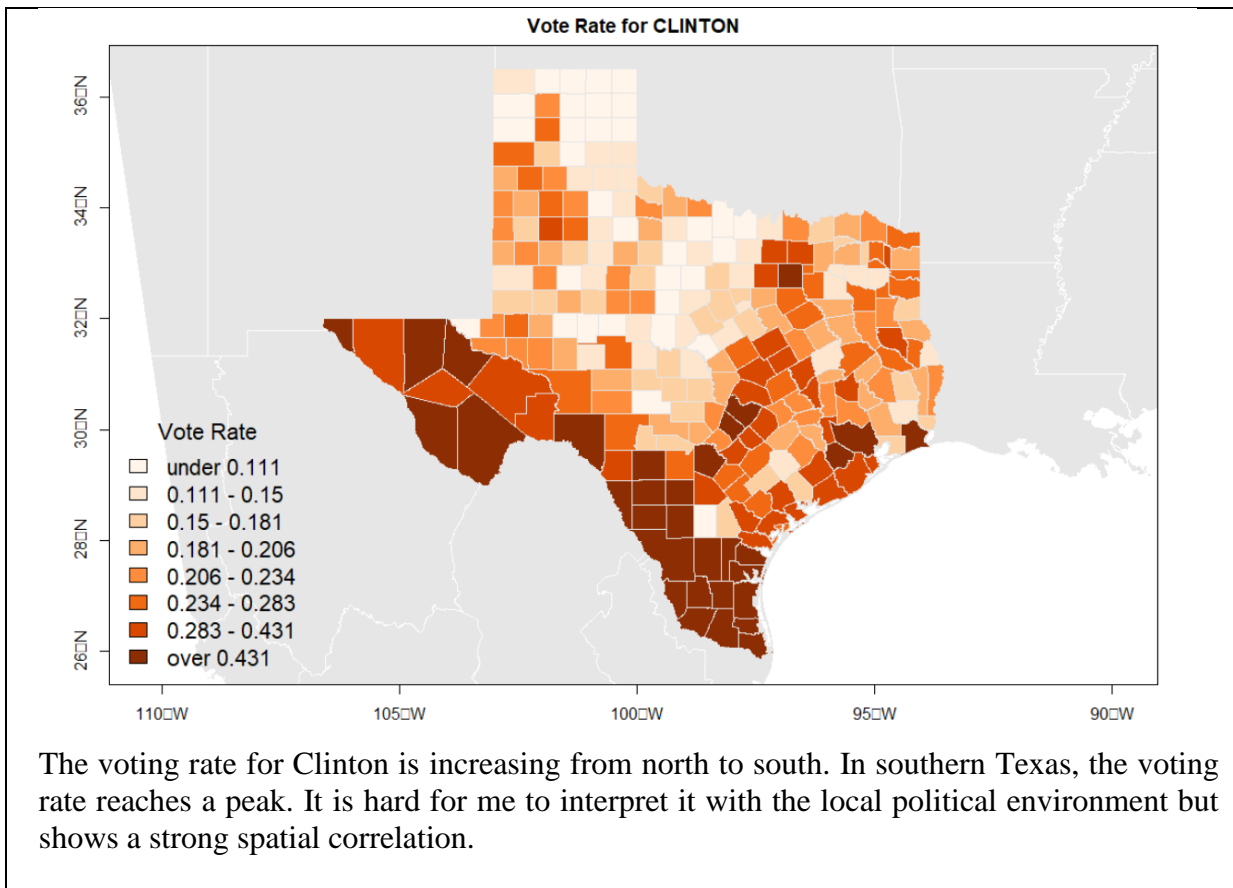
From both skewness and kurtosis metric, both two distributions are not normally and evenly distributed. But for later transforming and interpreting purposes, I would choose Clinton distribution since it positively skews, could perform log transformation on it.

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.

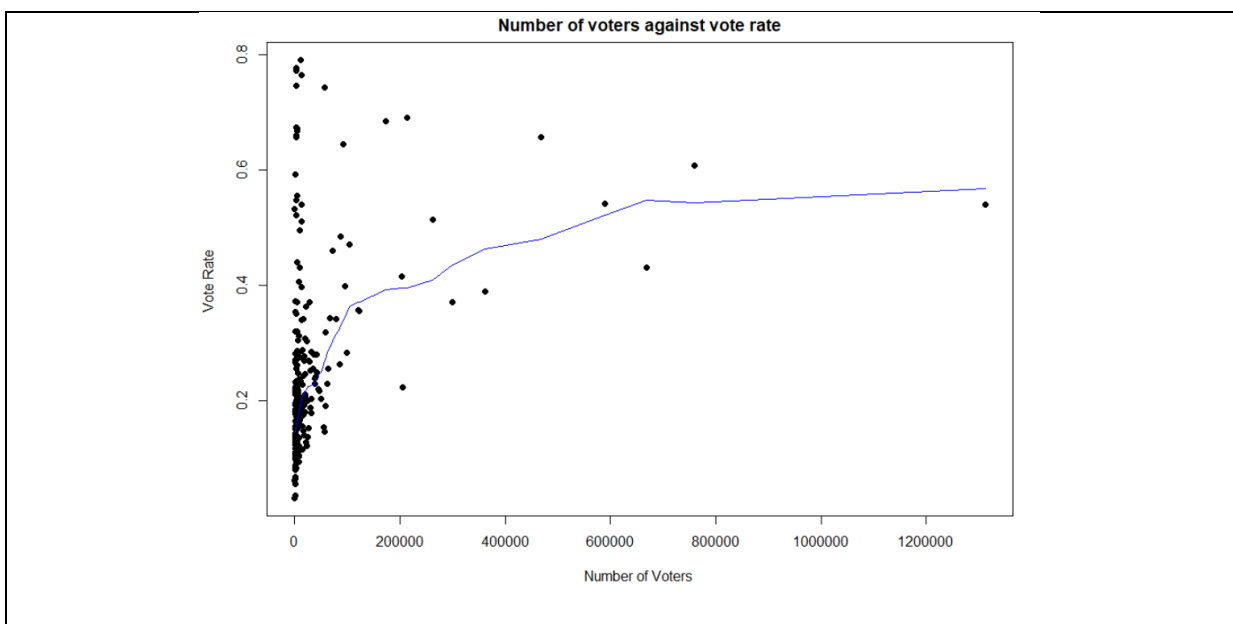
² Note that not all persons 18 years and older qualify to vote; for instance, because some are not U.S. citizens.

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.

⁴ The turnout percentage is that proportion of registered voters who participate in an election.

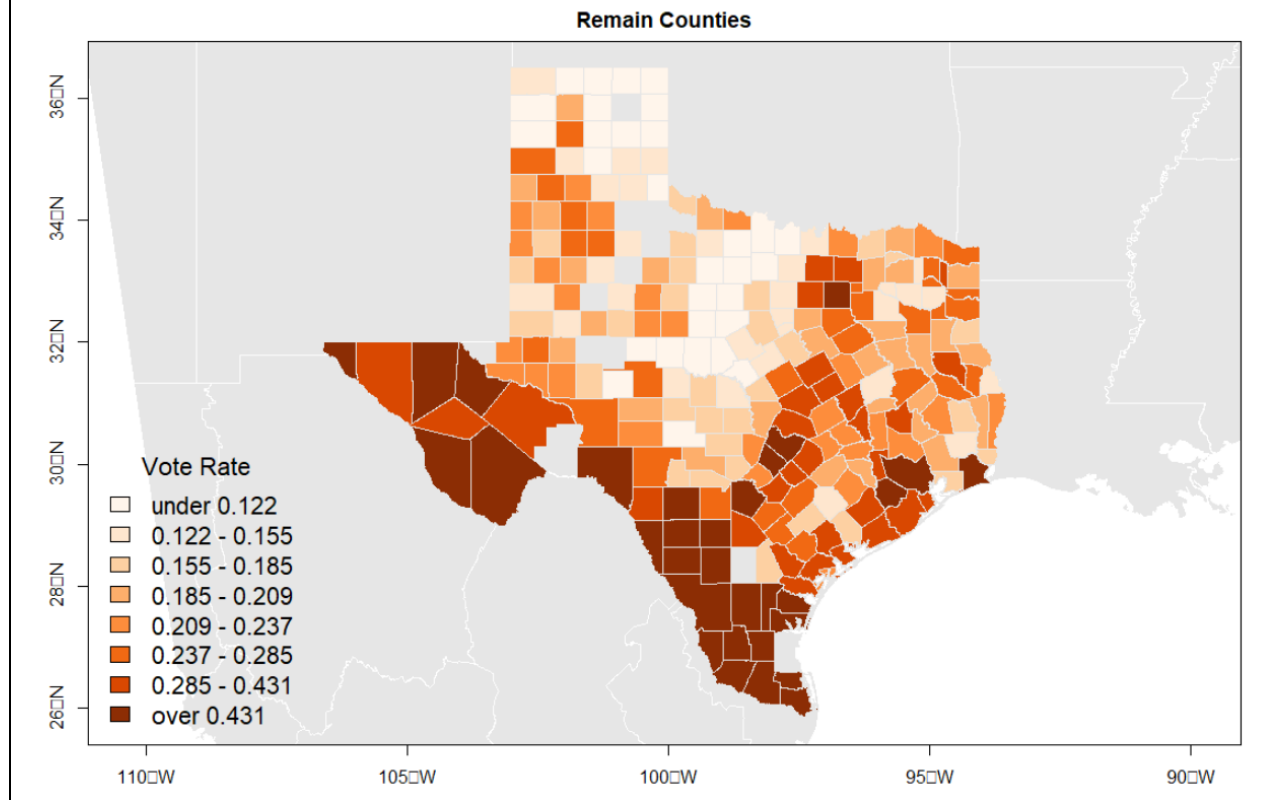



[c] Can all 254 counties be used in the analysis or do a few counties have a too small denominator, thus leading to instable percentage estimates.



From this graph, we could clearly notice that when the number of voters goes lower, the fluctuation of vote rate increased substantially. From getting a more accurate regression result, we should eliminate those counties to keep variance stable.

In here, I drop the lowest 5% counties from my dataset.



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.

All assumptions below based on my very limited knowledge of Clinton and the US election.

Independent Variables	Common-sense hypotheses	Hypothesis
-----------------------	-------------------------	------------

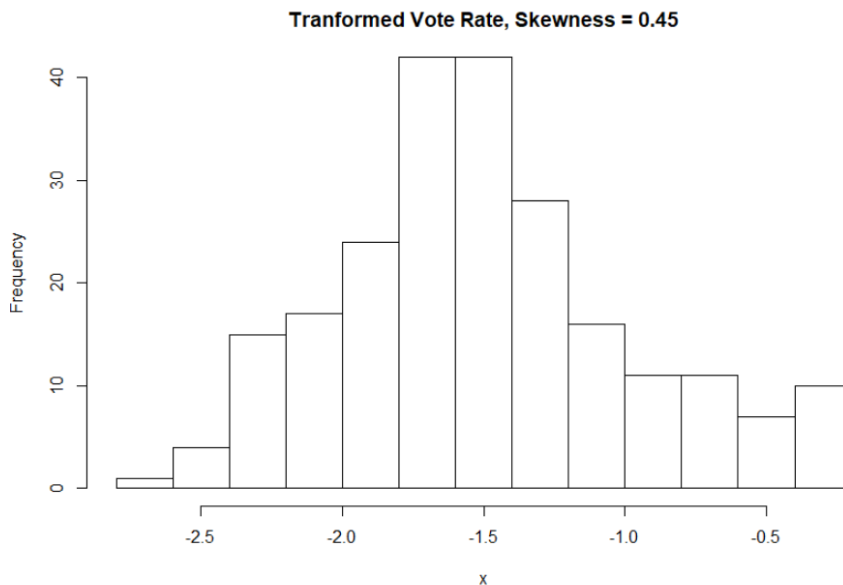
REGION	From the map shown above, the southern areas may have a higher vote rate for Clinton.	$H_0: \beta_0 = 0$ $H_1: \beta_0 \neq 0$
INCOME	I was told that a lot of rich people support Trump at that time. So, I guess with higher income, the voting rate should be lower when income go higher.	$H_0: \beta_0 \geq 0$ $H_1: \beta_0 < 0$
COLLEGEDEG	Just guess that higher education would have higher probability to support Clinton. The regression slope should be positive.	$H_0: \beta_0 \leq 0$ $H_1: \beta_0 > 0$
MEDAGE	Just guess younger people like to support Clinton more. The regression slope should be negative when age goes larger.	$H_0: \beta_0 \geq 0$ $H_1: \beta_0 < 0$
CRIMERATE	I just guess Clinton may propose some solutions for high crime rate areas. It may have a positive impact.	$H_0: \beta_0 \leq 0$ $H_1: \beta_0 > 0$
SINGLEMOM	Just guess single mom like to support Clinton more since she might be the first female president. It may have a positive impact.	$H_0: \beta_0 \leq 0$ $H_1: \beta_0 > 0$
LANEMILES	longer mile of the highway means more well-establish within the county (although it should be normalized using the area). So, it should have a positive impact since Clinton propose plans for urban establishing.	$H_0: \beta_0 \leq 0$ $H_1: \beta_0 > 0$

[3] Exploration of Variables (3 points)

In a scatter plot matrix or, where appropriate, boxplot:

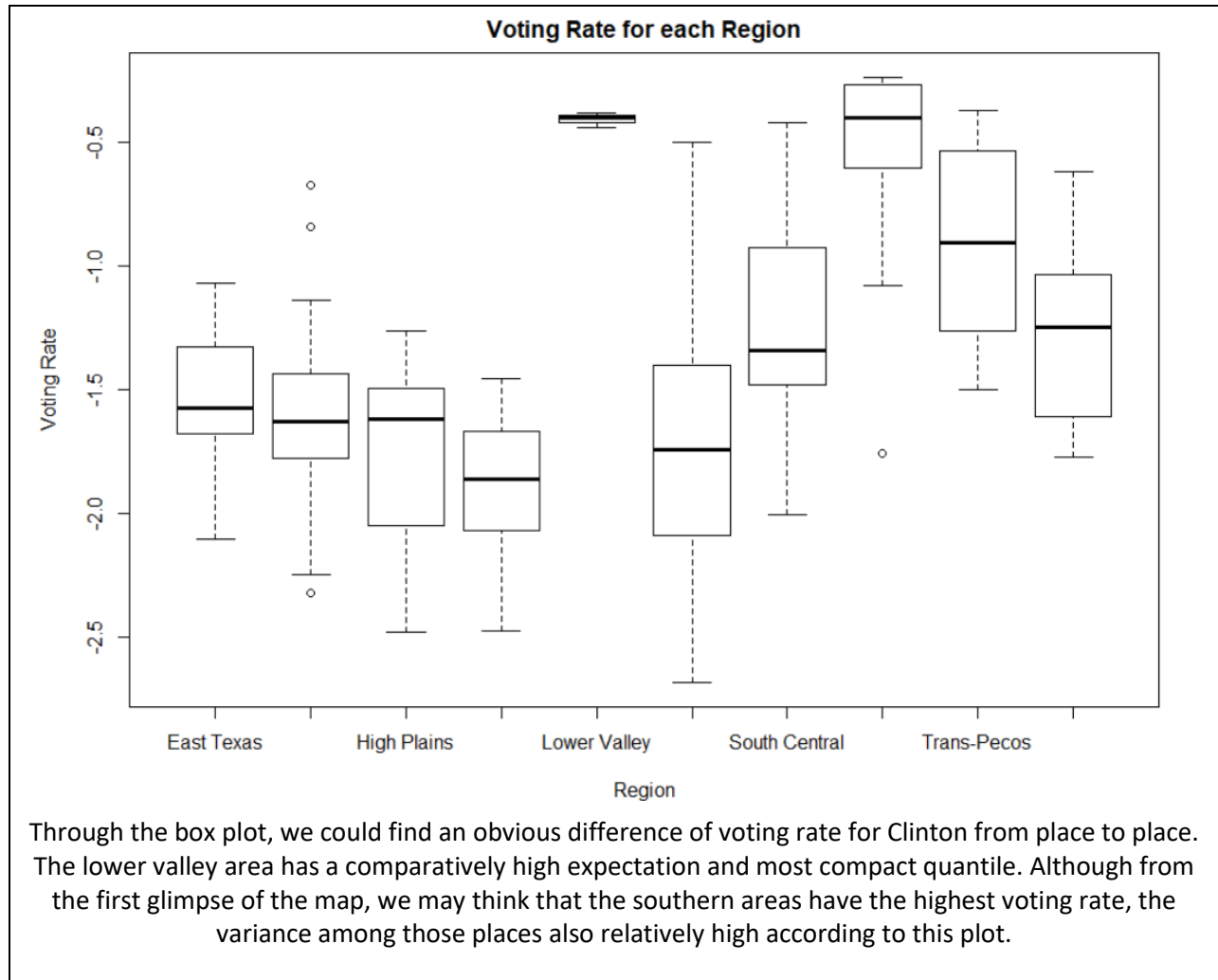
[a] Explore the univariate distribution of the dependent variable.

```
ct.data.remain$CLINTONRate.log <- log(ct.data.remain$CLINTONRate)
hist(ct.data.remain$CLINTONRate.log,breaks = 12,main =
paste('Tranformed Vote Rate, Skewness
=',round(e1071::skewness(ct.data.remain$CLINTONRate.log),2)),xlab
= 'x')
```



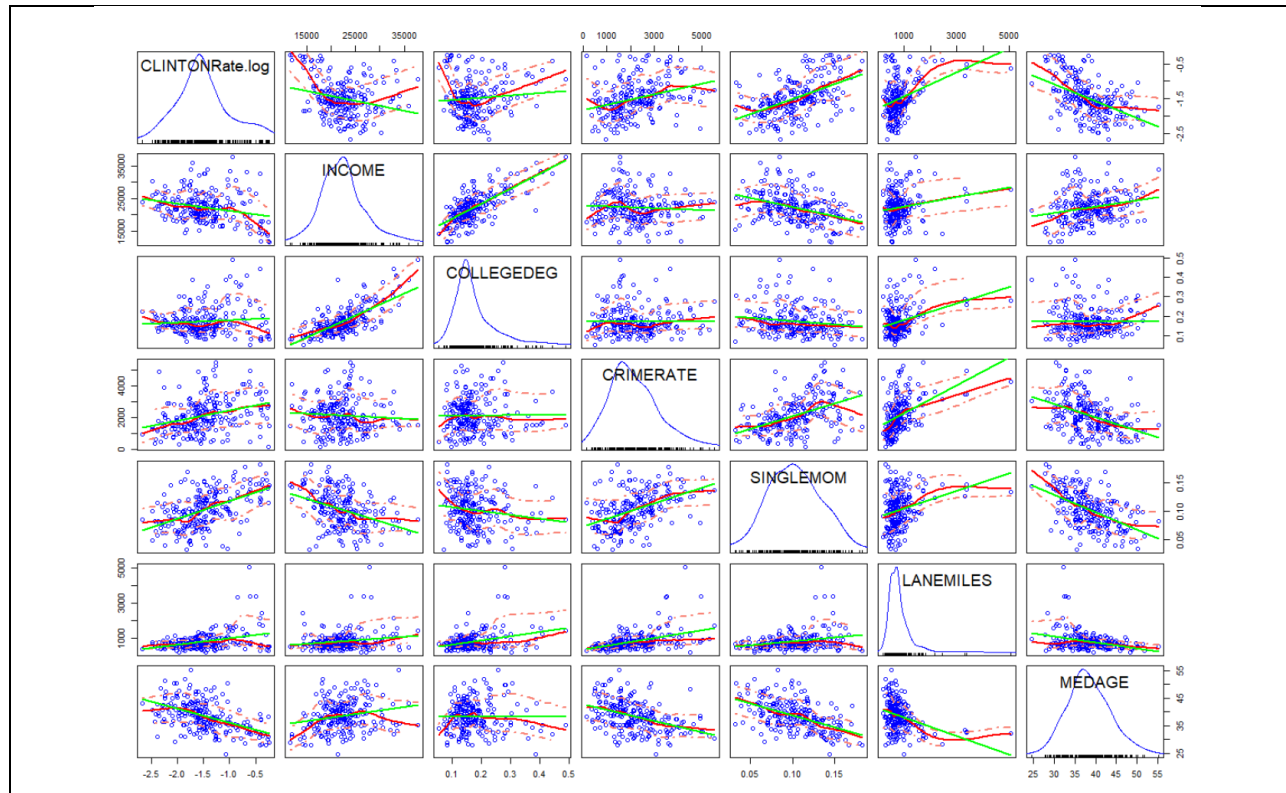
As discussed before, the original voting rate is positively skewed. Therefore, we applied the log transformation to calibrate the dependent variable.

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



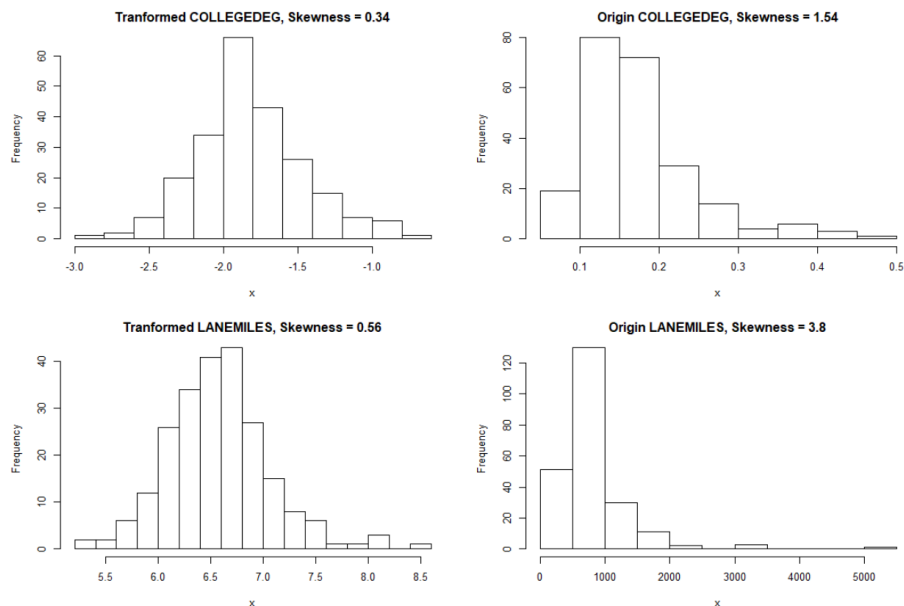
[c] Explore the univariate and bivariate distributions of the independent metric variables.

COLLEGE DEGREE and LANE MILES are highly positively skewed, so I would apply the logistic transformation to them. Additionally, Income and COLLEGE DEGREE are correlated with each other. For avoiding multicollinearity, we may consider dropping one or two of them later.

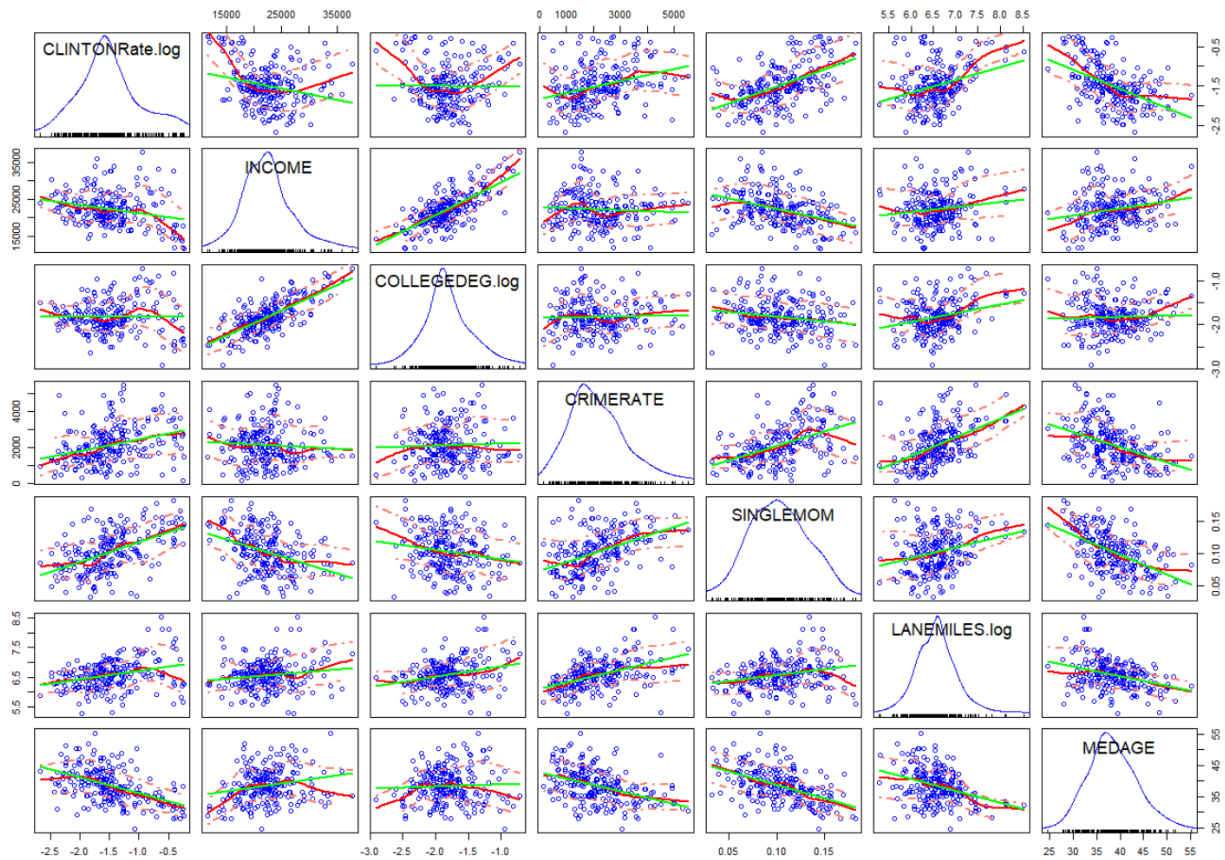


[d] Does this exploration point at any variable transformations for your initial regression model?

Yes, since COLLEGE DEGREE and LANE MILES are highly positively skewed, I applied the logistic transformation on them.



At this point redo the scatterplot matrix or boxplot 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:

[a] Are all selected variables and factors relevant and do their regression coefficients exhibit the expected sign?

```
lm(formula = CLINTONRate.log ~ REGION + INCOME + COLLEGEDEG.log +
  CRIMRATE + SINGLEMOM + LANEMILES.log + MEDAGE, data = ct.data.remain)
```

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept)	-4.070e-01	5.348e-01	-0.761	0.447573	
REGIONEdwards Plateau	8.408e-02	7.952e-02	1.057	0.291526	
REGIONHigh Plains	-3.197e-01	7.590e-02	-4.213	3.73e-05	***
REGIONLow Rolling Plains	-3.342e-01	8.089e-02	-4.132	5.19e-05	***
REGIONLower Valley	5.840e-01	1.874e-01	3.116	0.002084	**


```

REGIONNorth Central    -1.784e-01  6.724e-02  -2.653  0.008580  **
REGIONSouth Central    2.955e-01  7.648e-02   3.864  0.000148  ***
REGIONSouth Texas      8.233e-01  1.008e-01   8.168  2.80e-14   ***
REGIONTrans-Pecos      5.025e-01  1.156e-01   4.345  2.16e-05   ***
REGIONUpper Coast      2.205e-01  9.889e-02   2.229  0.026832   *
INCOME                  -2.385e-05  7.160e-06  -3.331  0.001021   **
COLLEGEDEG.log         3.684e-01  8.379e-02   4.397  1.74e-05   ***
CRIMERATE              1.158e-05  2.388e-05   0.485  0.628081
SINGLEMEMOM              2.662e+00  8.738e-01   3.046  0.002614   **
LANEMILES.log          1.250e-01  5.210e-02   2.399  0.017306   *
MEDAGE                  -2.659e-02  5.271e-03  -5.044  9.78e-07   ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2966 on 212 degrees of freedom

Multiple R-squared: 0.7029, Adjusted R-squared: **0.6819**

F-statistic: 33.43 on 15 and 212 DF, p-value: < 2.2e-16

Since the relationship between the crime rate and the dependent variable shares a similar pattern with the relationship between the crime rate and single mom. We may assume the "single mom" variable accounts for the influence of the crime rate. Based on the t-test, we should drop crime rate and do regression again.

```

lm(formula = CLINTONRate.log ~ REGION + INCOME + COLLEGEDEG.log +
    SINGLEMEMOM + LANEMILES.log + MEDAGE, data = ct.data.remain)

```

Residuals:

```

      Min       1Q   Median       3Q      Max
-0.80712 -0.17571  0.01182  0.18108  0.71703

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.499e-01  5.265e-01  -0.855  0.393753
REGIONEdwards Plateau  8.230e-02  7.929e-02   1.038  0.300493
REGIONHigh Plains    -3.145e-01  7.499e-02  -4.194  4.03e-05 ***
REGIONLow Rolling Plains -3.340e-01  8.075e-02  -4.137  5.07e-05 ***
REGIONLower Valley    5.908e-01  1.865e-01   3.167  0.001764 **
REGIONNorth Central  -1.799e-01  6.705e-02  -2.683  0.007878 **
REGIONSouth Central   2.983e-01  7.613e-02   3.918  0.000120 ***
REGIONSouth Texas     8.221e-01  1.006e-01   8.173  2.67e-14 ***
REGIONTrans-Pecos     4.979e-01  1.150e-01   4.328  2.32e-05 ***
REGIONUpper Coast     2.261e-01  9.803e-02   2.307  0.022038 *
INCOME           -2.381e-05  7.147e-06  -3.331  0.001020 **
COLLEGEDEG.log       3.689e-01  8.363e-02   4.411  1.64e-05 ***
SINGLEMEMOM           2.801e+00  8.240e-01   3.399  0.000808 ***
LANEMILES.log        1.337e-01  4.885e-02   2.736  0.006736 **
MEDAGE             -2.670e-02  5.257e-03  -5.078  8.30e-07 ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.296 on 213 degrees of freedom

Multiple R-squared: **0.7026**, Adjusted R-squared: **0.683**

F-statistic: 35.94 on 14 and 213 DF, p-value: < 2.2e-16

[b] Is multicollinearity a problem?

```

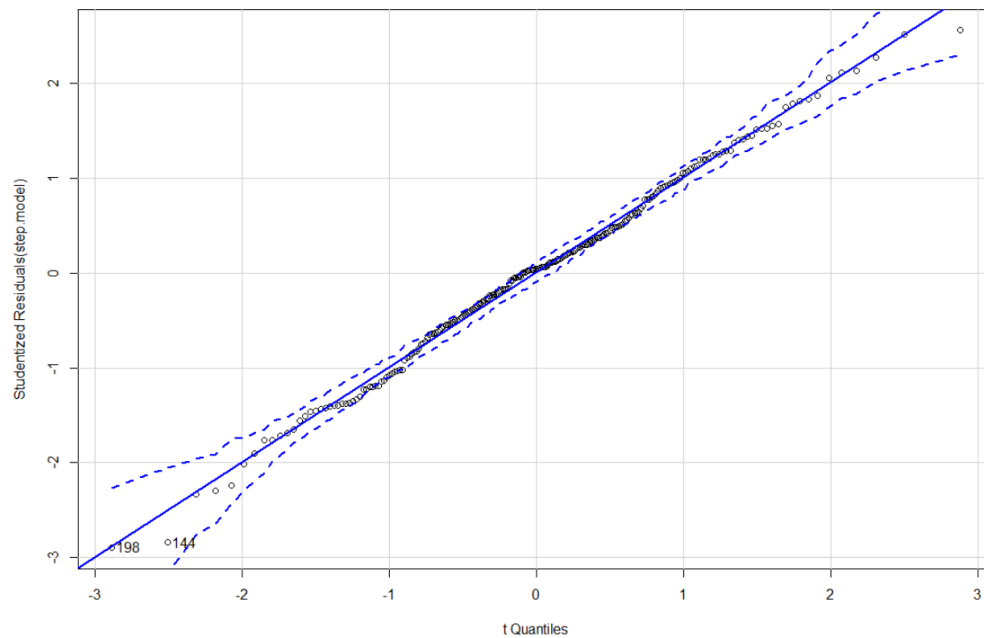
vif(step.model)
              GVIF Df GVIF^(1/(2*Df))
REGION          2.607425  9      1.054685
INCOME          2.793170  1      1.671278
COLLEGEDEG.log  2.459038  1      1.568132

```

SINGLEMEM	1.749766	1	1.322787
LANEMILES.log	1.471179	1	1.212922
MEDAGE	2.054428	1	1.433327

Vif values are all smaller than 10, so there is no multicollinearity issue.

[c] Are the model residuals approximately normally distributed?



```
shapiro.test(residuals(step.model))
```

```
data: residuals(step.model)
W = 0.99553, p-value = 0.7501
```

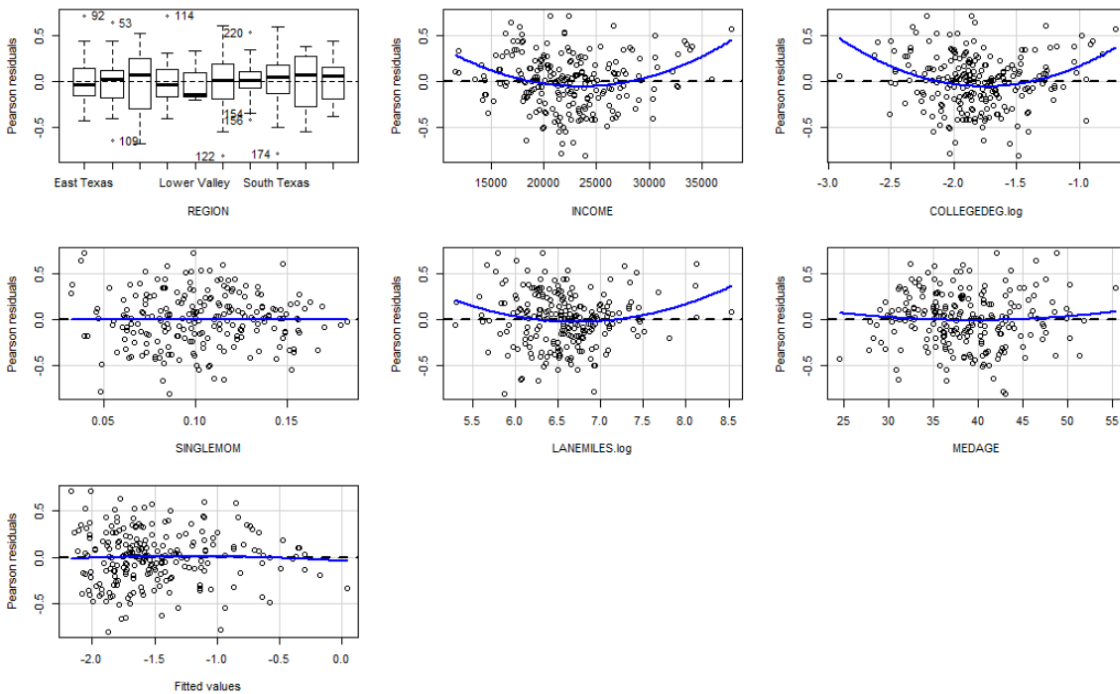
Results from the Shapiro test indicates that we failed to reject the none hypothesis, which means our residuals are normally distributed.

[d] Do you need to refine the variable transformations or add quadratic terms?

```
residualPlots(step.model)
```

REGION	Test stat	Pr(> Test stat)
INCOME	4.8501	2.387e-06 ***
COLLEGEDEG.log	4.8615	2.266e-06 ***
SINGLEMEM	-0.0579	0.953872
LANEMILES.log	2.7377	0.006712 **
MEDAGE	0.8197	0.413286
Tukey test	-0.5195	0.603383

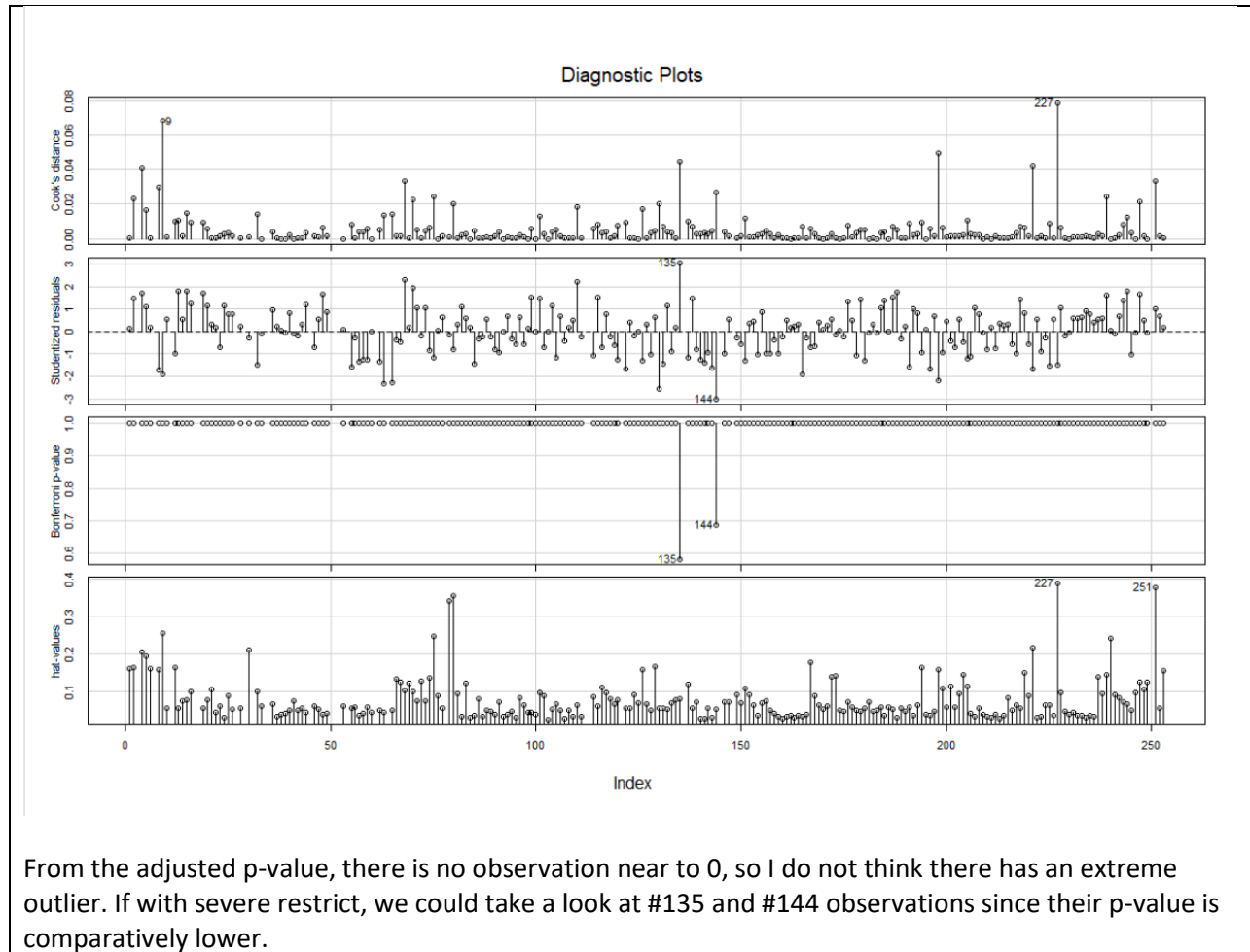
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Results from the Tukey test indicates that INCOME, COLLEGEDEG.log, and LANEMILES.log has a significant quadratic impact on the dependent variable

[e] Are there influential cases and outliers present in the model?

```
car::influenceIndexPlot(updated.model)
```



- [f] 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?

	CLINTON Rate	REGION	INCOME	COLLEGE DEG	MEDAGE	SINGLE OM	LANEMIL ES
135	0.1028665	North Central	24667	0.150396	42.6	0.080748	852
144	0.2289878	South Central	32736	0.322579	42.7	0.068572	703

By comparing the distribution of independent variables, I do not find the unreasonable or extreme value those two counties have. So, I would not drop them from my dataset.

[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.

```

updated.model <- update(step.model, .~.+I(COLLEGEDEG.log^2) + I(LANEMILES.log^2) +
I(INCOME^2))
summary(updated.model)

lm(formula = CLINTONRate.log ~ REGION + INCOME + COLLEGEDEG.log +
  SINGLEMOM + LANEMILES.log + MEDAGE + I(COLLEGEDEG.log^2) +
  I(LANEMILES.log^2) + I(INCOME^2), data = ct.data.remain)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.168e+00  2.086e+00   3.437 0.000710 ***
REGIONEdwards Plateau  7.836e-02  7.329e-02   1.069 0.286218
REGIONHigh Plains    -2.726e-01  7.028e-02  -3.878 0.000141 ***
REGIONLow Rolling Plains -3.345e-01  7.476e-02  -4.475 1.25e-05 ***
REGIONLower Valley    2.638e-01  1.878e-01   1.405 0.161513
REGIONNorth Central  -1.740e-01  6.228e-02  -2.794 0.005681 **
REGIONSouth Central   3.068e-01  7.028e-02   4.366 1.99e-05 ***
REGIONSouth Texas     6.001e-01  9.983e-02   6.011 8.07e-09 ***
REGIONTrans-Pecos     4.325e-01  1.072e-01   4.034 7.67e-05 ***
REGIONUpper Coast     2.373e-01  9.083e-02   2.613 0.009628 **
INCOME             -1.457e-04  3.789e-05  -3.845 0.000160 ***
COLLEGEDEG.log      1.315e+00  4.586e-01   2.866 0.004574 **
SINGLEMOM            2.626e+00  7.706e-01   3.407 0.000787 ***
LANEMILES.log       -1.583e+00  6.139e-01  -2.579 0.010583 *
MEDAGE              -2.089e-02  4.997e-03  -4.181 4.26e-05 ***
I(COLLEGEDEG.log^2)  2.835e-01  1.239e-01   2.288 0.023155 *
I(LANEMILES.log^2)   1.309e-01  4.616e-02   2.835 0.005036 **
I(INCOME^2)          2.417e-09  7.983e-10   3.028 0.002769 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.273 on 210 degrees of freedom
Multiple R-squared:  0.7506,    Adjusted R-squared:  0.7304
F-statistic: 37.18 on 17 and 210 DF,  p-value: < 2.2e-16

anova(step.model,updated.model)

  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     213 18.665
2     210 15.650   3     3.0144 13.483 4.432e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

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

Region: Based on the voting rate in the east area, south Texas, Trans Pecos, upper coast and south-central areas have a significant positive influence, which fitted with our observation, southern areas with a higher rate. In contrast, low rolling plains, high plains, and north-central areas have a significantly lower voting rate. The rest areas do not have a big difference with the east area.

Income: The regression slope of the income is nested but much larger than else, plus the variation among incomes spans tremendously, which makes income becomes a dominant variable. It means lower-income loves Trump more, extremely rich people more likely to support Clinton.

Single Mom: The slope is positive, which fits our expectation.

Median Age: The slope is negative, which also fits our assumption.

highway miles: It has a quadratic impact on the dependent variable. but hard to interpret it since it needs to be normalized.

education: both higher and lower education loves Clinton more, except medium.

Overall: Those independent variables could account for 75% variation among voting rates. Results from the t-test show that all regression coefficients significantly differ from 0. And F test confirms our model is significant.

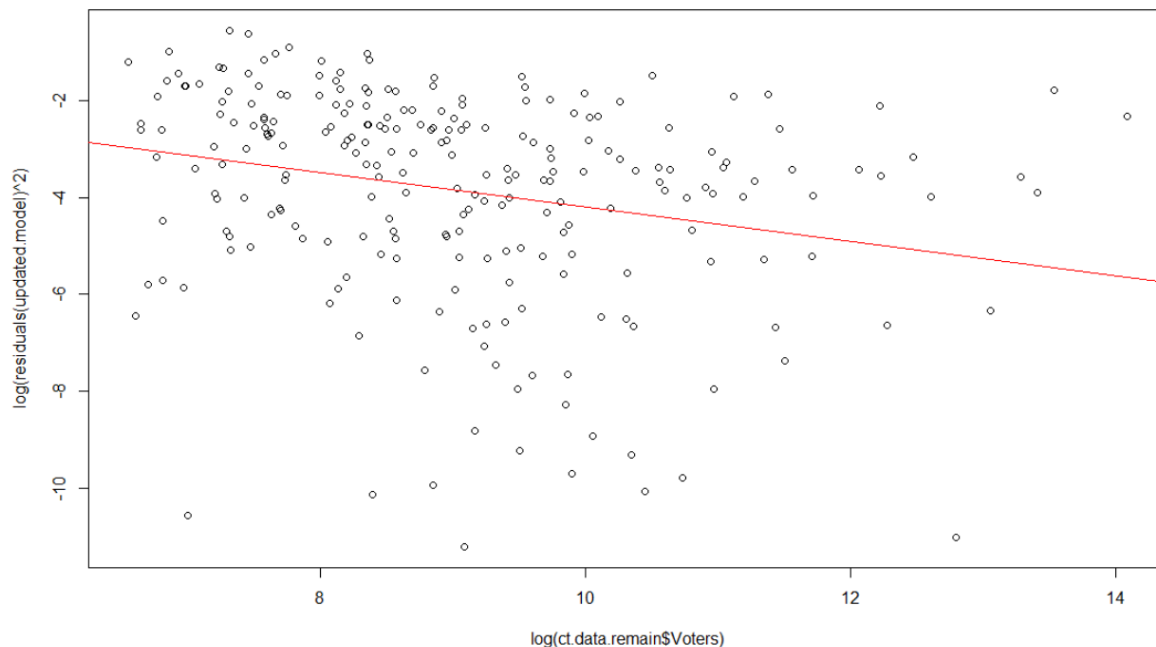
[6] Heteroscedasticity Investigation (2 points)

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 $\{\gamma_0, \gamma_1\}$ of the multiplicative heteroscedasticity model

$$\sigma_i^2 = \exp(\gamma_0 + \gamma_1 \cdot \log(\text{refpop}_i)).$$

```
auxreg<- lm(log(residuals(updated.model)^2)~log(ct.data.remain$Voters))
plot(log(residuals(updated.model)^2)~log(ct.data.remain$Voters)); abline(auxreg,
col="red")
```



it is obvious that the variance of residual is not consistent, it varies when the voters' number changes. And the slope is negative, which means variance should be smaller when the voters' number goes larger.

[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.

```
lm.hetero <- lmHetero(CLINTONRate.log ~ REGION + INCOME + COLLEGEDEG.log + SINGLEMOM
+ LANEMILES.log + MEDAGE, hetero=~log(Voters), data=ct.data.remain )
summary(lm.hetero)
```

Call:

```
lmHetero(formula = CLINTONRate.log ~ REGION + INCOME + COLLEGEDEG.log +
SINGLEMOM + LANEMILES.log + MEDAGE | log(Voters), data = ct.data.remain)
```

Regression Coefficients:

	Estimate	Std.Err	z-value	Pr(> z)
(Intercept)	-9.8932e-01	4.5753e-01	-2.1623	0.0305924 *
REGIONEdwards Plateau	1.1409e-01	7.4190e-02	1.5378	0.1241089
REGIONHigh Plains	-3.1257e-01	6.7718e-02	-4.6158	3.917e-06 ***
REGIONLow Rolling Plains	-2.7342e-01	8.0359e-02	-3.4025	0.0006676 ***
REGIONLower Valley	5.6368e-01	1.3713e-01	4.1105	3.947e-05 ***
REGIONNorth Central	-7.7827e-02	5.6566e-02	-1.3759	0.1688610
REGIONSouth Central	3.0783e-01	6.3684e-02	4.8338	1.340e-06 ***
REGIONSouth Texas	8.3341e-01	9.5351e-02	8.7405	< 2.2e-16 ***
REGIONTrans-Pecos	5.4044e-01	1.0544e-01	5.1257	2.964e-07 ***
REGIONUpper Coast	2.3560e-01	7.4297e-02	3.1711	0.0015185 **
INCOME	-1.0456e-05	6.4970e-06	-1.6094	0.1075292
COLLEGEDEG.log	3.3577e-01	7.6301e-02	4.4006	1.080e-05 ***
SINGLEMEM	3.6132e+00	7.7999e-01	4.6324	3.614e-06 ***
LANEMILES.log	1.5634e-01	4.0268e-02	3.8824	0.0001034 ***
MEDAGE	-2.8758e-02	4.8405e-03	-5.9412	2.830e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Gamma Coefficients:

	Gamma	Std.Err	z-value	Pr(> z)
(Intercept)	0.259984	0.552649	0.4704	0.638
log(Voters)	-0.309259	0.060374	-5.1224	3.017e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

log-likelihood = -36.83892

Heteroscedasticity likelihood ratio test:

LR	df	Pr(Chi > LR)
14.61004	1	0.0001322081

The likelihood ratio test indicates the p-value is smaller than 0.05, so we can reject the null hypothesis and tentatively conclude that there has heterogeneity, so it is necessary to use the population as the weighted index. The Gamma Coefficients is -0.309259 and the p-value is pretty small, which means when the population goes larger, the variance of residual decrease significantly.

The previous dominant variable income becomes insignificant in this model. It may have a high correlation with the population. The rest estimator remains a similar range as before.

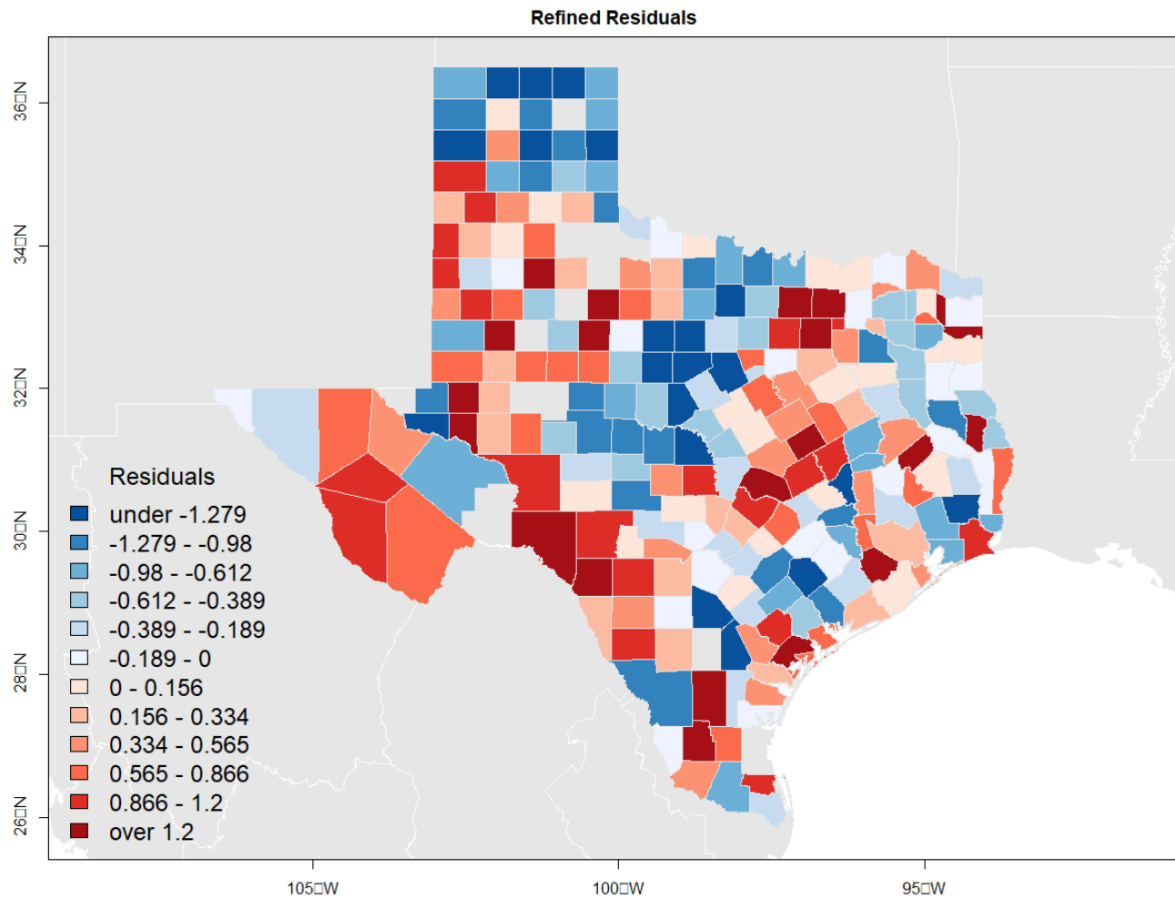
[7] Spatial Residual Analysis (3 points)

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.

```
(length(Resid.weight[Resid.weight < 0]) ) 118
(length(Resid.weight[Resid.weight > 0]) ) 123
plot(ng.shp, axes=T, col=grey(0.9), border="white", xlim=ct.bbox[1,], ylim=ct.bbox[2,])
mapBiPolar(Resid.weight, ct.shp.remain, neg.breaks=6, pos.breaks=6, break.value=0.0,
map.title="Refined Residuals", legend.title="Residuals", legend.cex=1.5, add.to.map=T)
```



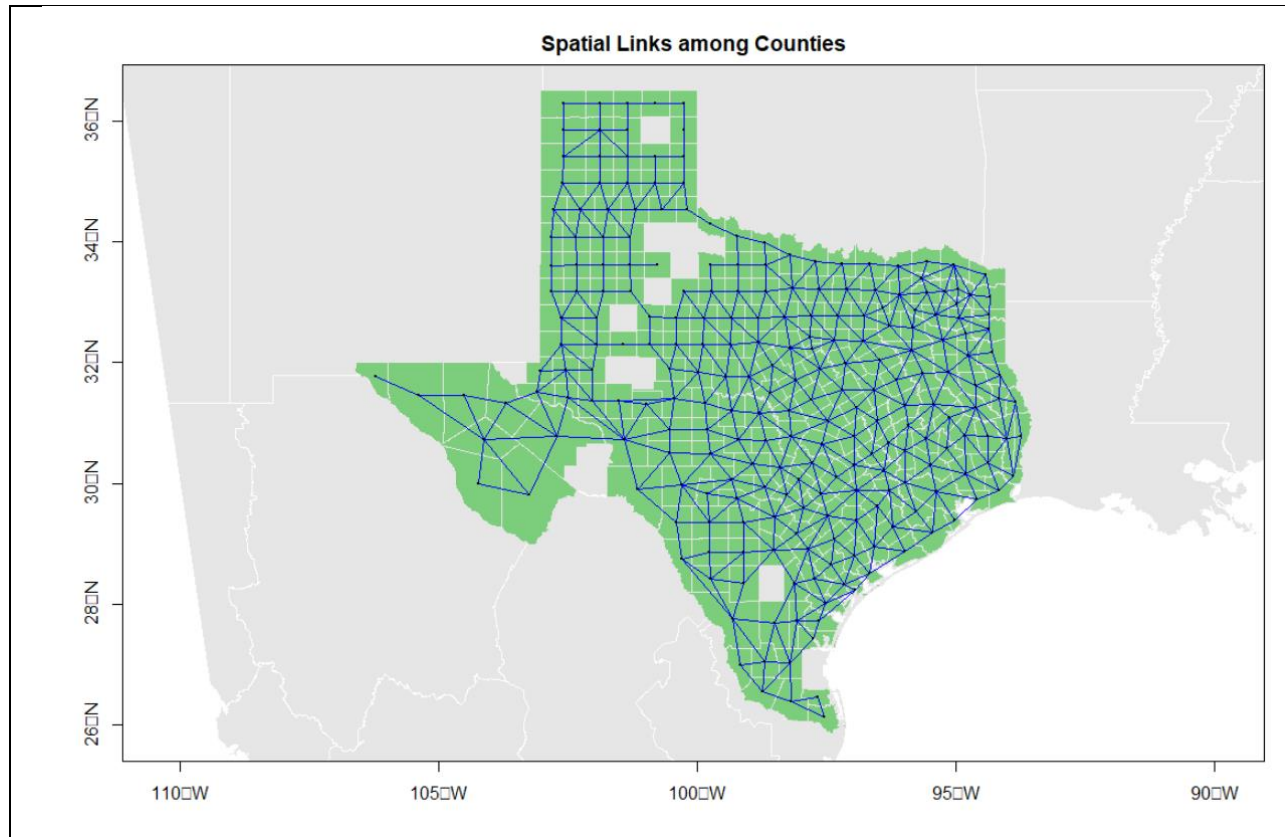
In the map, we could find that the northern areas are generally overestimated by our model, and Southern and east central areas are generally underestimated. From their observed pattern, it means our predicted surface are over smoother.

The overall pattern exhibited several clusters, which means positive spatial autocorrelation may exist in our dataset.

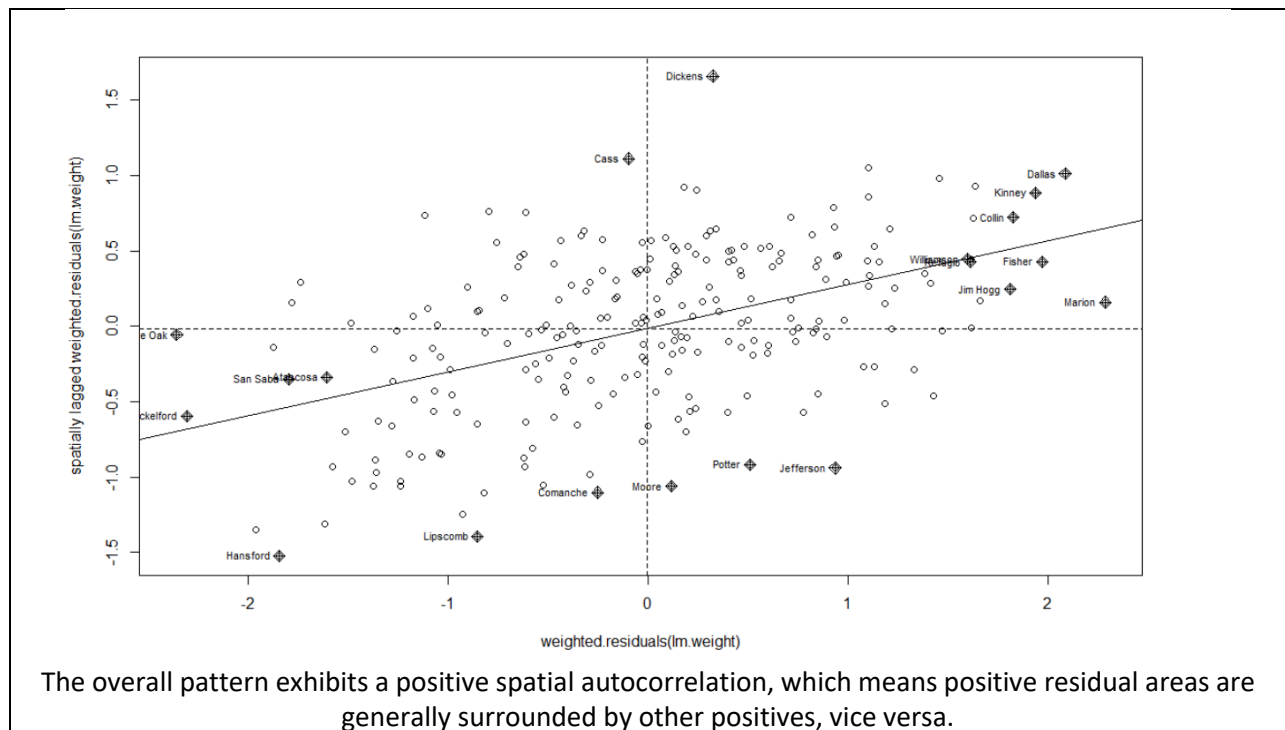
[b] Generate the spatial links and plot its graph onto a map of the Texas Counties. Check whether this graph is connecting all counties properly.

```
ct.link <- poly2nb(ct.shp.remain, queen=F)
ct.centroid <- coordinates(ct.shp.remain)
plot(ng.shp, axes=T, col=grey(0.9), border="white", xlim=ct.bbox[1,], ylim=ct.bbox[2,])
plot(ct.shp.remain, col="palegreen3", border=grey(0.9), axes=T, add=T)
plot(ct.link, coords=ct.centroid, pch=19, cex=0.1, col="blue", add=T)
title("Spatial Links among Counties")
```

All counties in my dataset are properly linked.



[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.

```
lm.morantest(lm.weight, ct.linkW)

Global Moran I for regression residuals
weights: ct.linkW

Moran I statistic standard deviate = 8.2401, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation      Variance
0.290126517          -0.038553096      0.001591053
```

The p-value confirmed that the observed Moran's Index is significant. Since it's 0.29, which also means positive spatial autocorrelation exists in our dataset.

[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?

```
rate.SAR <- spautolm(lm.weight, na.action="na.omit", listw=ct.linkW, family="SAR")
summary(rate.SAR)

Residuals:
      Min       1Q   Median       3Q      Max
-0.7271115 -0.1565658 -0.0094384  0.1572598  0.5775110

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.0191e+00  4.3342e-01 -2.3513 0.0187094
REGIONEdwards Plateau -2.4660e-02  1.2667e-01 -0.1947 0.8456390
REGIONHigh Plains   -1.3867e-01  1.3013e-01 -1.0656 0.2866002
REGIONLow Rolling Plains -1.7012e-01  1.2698e-01 -1.3397 0.1803299
REGIONLower Valley    2.9379e-01  2.9166e-01  1.0073 0.3137884
REGIONNorth Central  -1.1108e-01  1.0026e-01 -1.1080 0.2678708
REGIONSouth Central   2.7630e-01  1.0983e-01  2.5157 0.0118785
REGIONSouth Texas     5.2680e-01  1.5839e-01  3.3261 0.0008808
REGIONTrans-Pecos     1.3769e-01  2.0693e-01  0.6654 0.5057995
REGIONUpper Coast     2.3160e-01  1.1843e-01  1.9555 0.0505210
INCOME          -1.3143e-05  6.4917e-06 -2.0245 0.0429168
COLLEGEDEG.log      2.1761e-01  7.2802e-02  2.9890 0.0027989
SINGLEMEM           3.1980e+00  6.4725e-01  4.9409 7.777e-07
LANEMILES.log       1.0662e-01  3.8670e-02  2.7571 0.0058315
```

```
MEDAGE                -2.1370e-02  4.3717e-03 -4.8882 1.018e-06
```

Lambda: 0.70783 LR test value: 55.293 p-value: 1.0381e-13

Numerical Hessian standard error of lambda: 0.076002

Log likelihood: -11.93358

ML residual variance (sigma squared): 0.50053, (sigma: 0.70748)

Number of observations: 241

Number of parameters estimated: 17

AIC: 57.867

```
likeH0 <- lm.hetero$logLikeH1      # unrestricted model
likeH1 <- logLik(rate.SAR)
cat("chi-square value: ", chi <- -2*(likeH0[1]-likeH1[1]))
cat("\nerror-probability: ", pchisq(chi, df=1, lower.tail=F))
```

chi-square value: 49.81068

error-probability: 1.693194e-12

The error probability is much lower than 0.05, which means the calibration of spatial autocorrelation improves our model significantly. And the lambda from the SAR model is 0.7, which indicates positive spatial autocorrelation exists in our dataset. P-value also confirms it is significant.

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

```
moran.mc(residuals(rate.SAR), ct.linkW, nsim=9999)
```

Monte-Carlo simulation of Moran I

```
data: residuals(rate.SAR)
```

```
weights: ct.linkW
```

```
number of simulations + 1: 10000
```

```
statistic = -0.031287, observed rank = 2635, p-value = 0.7365
```

```
alternative hypothesis: greater
```

Results from a total of 10 thousand testings show that there is no significant spatial autocorrelation in the residual of the SAR model, which means the spatial autocorrelation in the original datasets already been eliminated by our SAR model.