Louis Burns
GISC Data Analysis Qualifier Exam Question, Spring 2018

I pledge that I will not discuss the particulars of this exam with any student who has not already taken the exam. Furthermore, I agree not to perform plagiarism.

Type your name here in agreement with these rules: Louis Burns

Analysis Tasks

[1] Specification of the Dependent Variable

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

For this calculation, I selected `TOTALVOT16` for the denominator rather than `REGVOT16` because the former is the number of registered voters who actually voted. My percentage calculation is:

```
Texas.shp$CLINTONVOT/Texas.shp$TOTALVOT16
```

[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.

Note: The R 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.

After reviewing histograms and log transformations, I see that the log transformation of `CLINTONVOT / TOTALVOT16` is the most normally distributed of the possible options. I will add 3.5 to the values to make them positive:

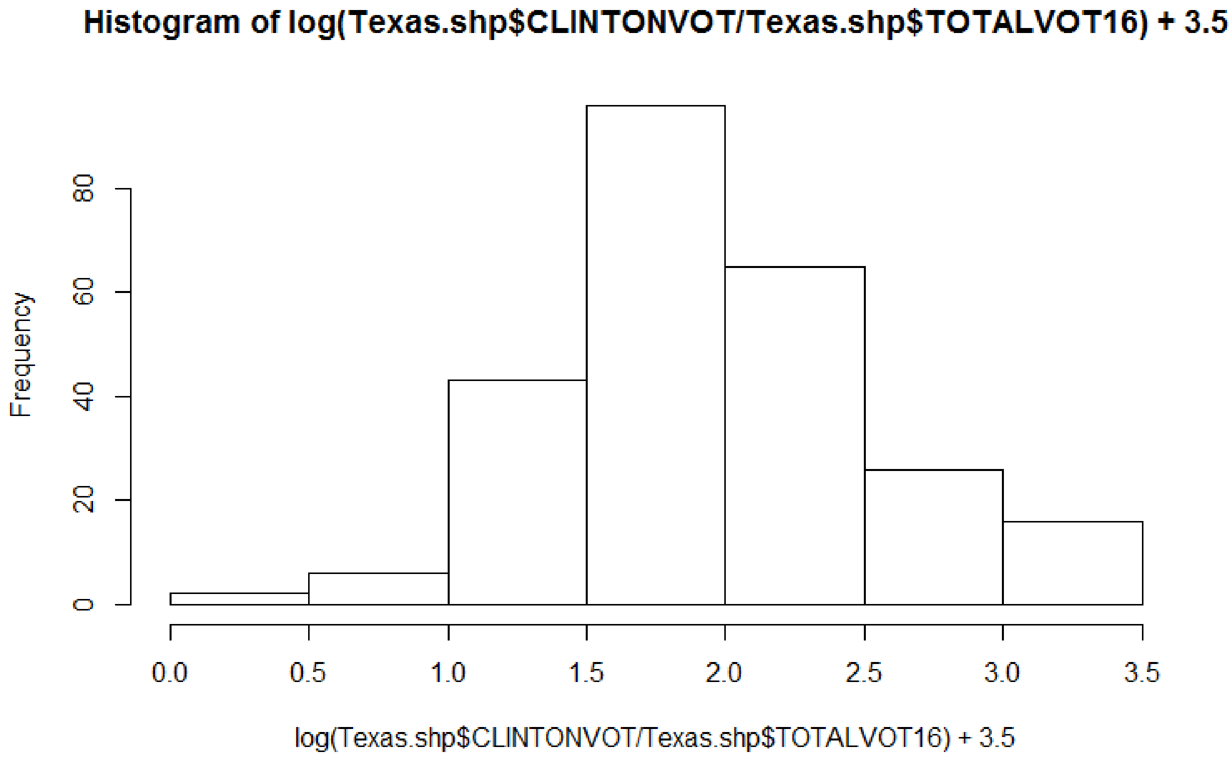
```
Texas.shp$CLINTONRATE <-
   log(Texas.shp$CLINTONVOT/Texas.shp$TOTALVOT16) + 3.5
hist(Texas.shp$CLINTONRATE)
```

[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.

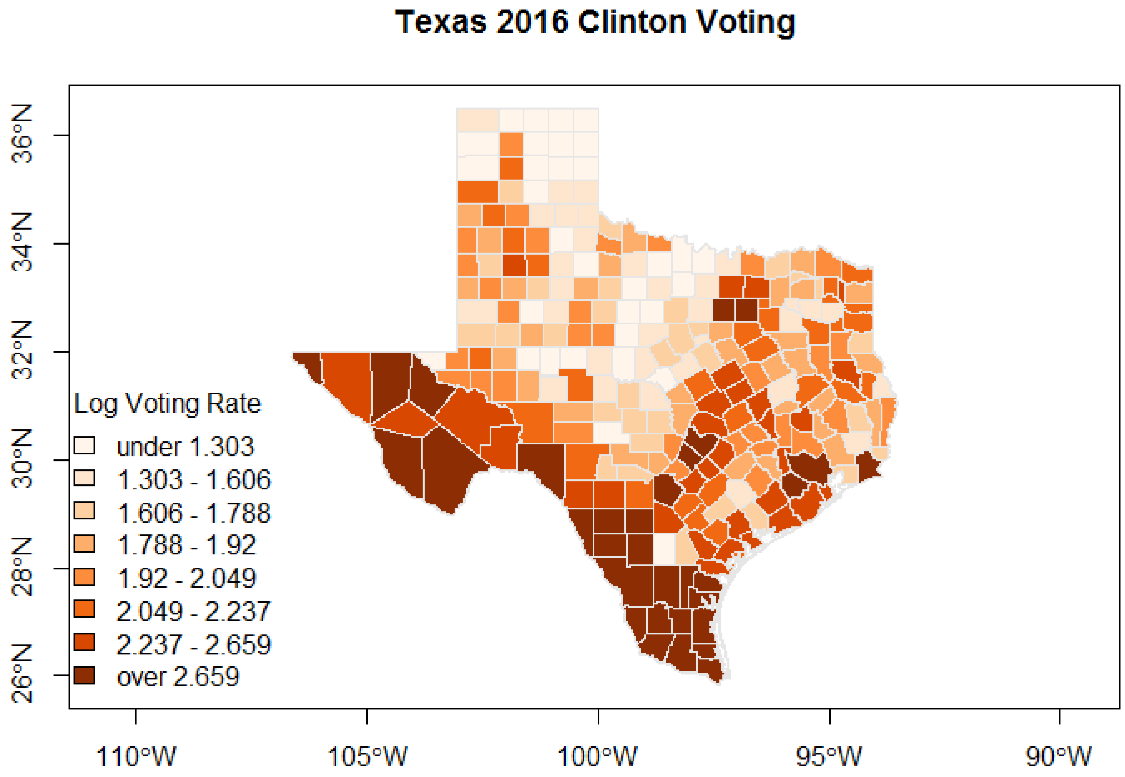
[2] Note that not all persons 18 years and older qualify to vote; for instance, because some are not U.S. citizens.

[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.



```
plotColorRamp(Texas.shp$CLINTONRATE, Texas.shp, my.title="Texas 2016
        Clinton Voting", my.legend="Log Voting Rate")
```



[2] Selection of Independent Variables

[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.

Variable	Explanation	Hypothesis
CRIMERATE	I expect more Clinton support in the larger cities which may have an overall lower crime rate.	H_0: $\beta_k \geq 0$ H_1: $\beta_k < 0$
OBAVOT12	I expect former Obama voters to be Clinton supporters.	H_0: $\beta_k \leq 0$ H_1: $\beta_k > 0$
HISPORG	I expect more Hispanic voters to be Clinton supporters.	H_0: $\beta_k \leq 0$ H_1: $\beta_k > 0$
COLLEGEDEG	I expect more educated voters to be Clinton supporters.	H_0: $\beta_k \leq 0$ H_1: $\beta_k > 0$
POVERTY	I expect poorer areas to support Clinton's opponent.	H_0: $\beta_k \geq 0$ H_1: $\beta_k < 0$

URBRURAL (factor)	I expect more Clinton support in the more urban areas.	H_0: $\beta_k \le 0$ H_1: $\beta_k > 0$

[3] Exploration of Variables

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?

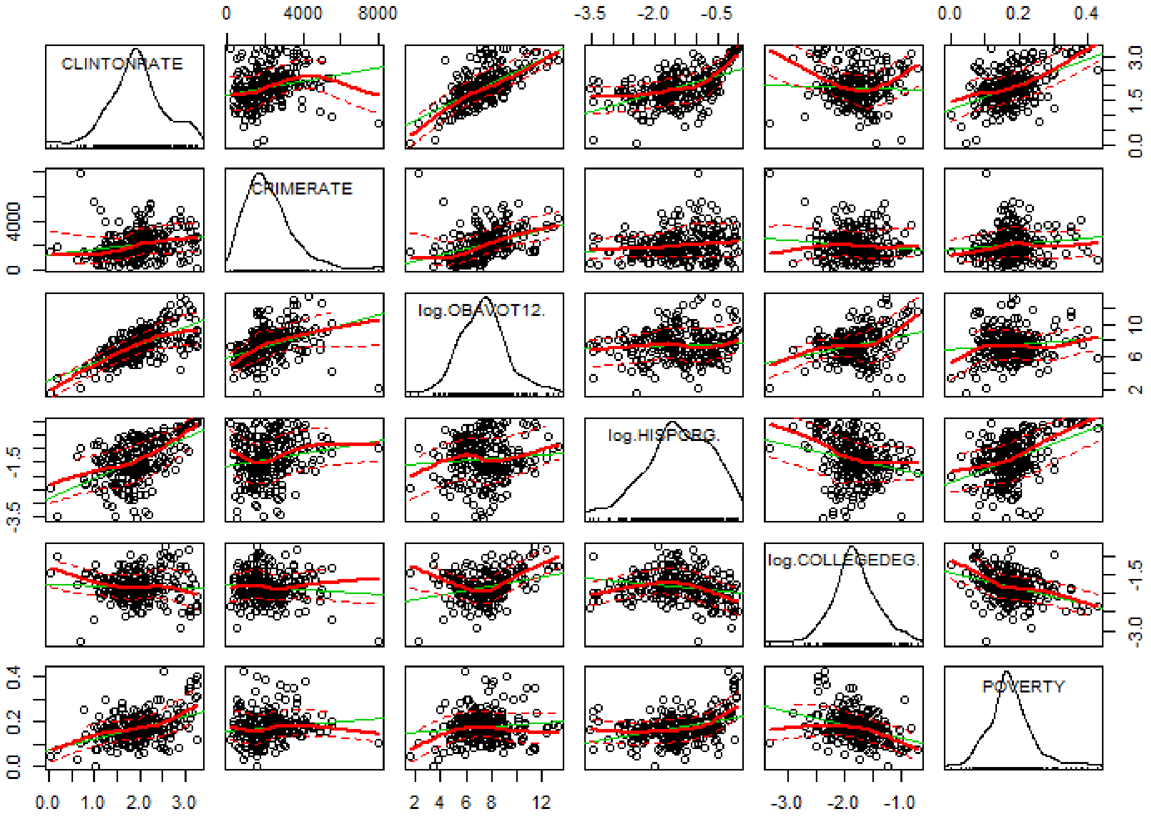
Note: Only consider transformations rounded to the closest 0.5 increments with a preference for the *log*-transformation, if required with an additional shift parameter. The log-transformation is preferred because it allows to interpret any relationships in percentage changes.

At this point redo the scatterplot matrix or box-plot with the any selected variable transformation. Remember that the scatterplot matrix function has an option to identify λ-values of the Box-Cox transformation.

Your initial trial model should already incorporate the identified transformation.

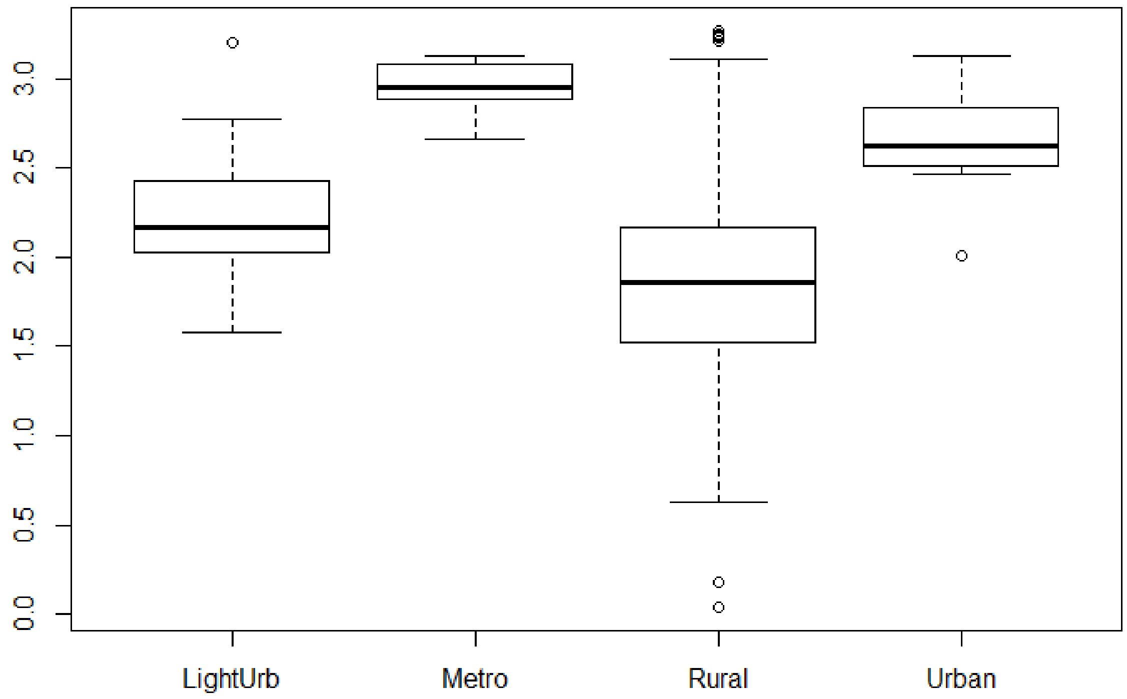
My initial scatterplot matrix suggested several log transformations which I show here:

```
scatterplotMatrix(~ CLINTONRATE + CRIMERATE + log(OBAVOT12) +
   log(HISPORG) + log(COLLEGEDEG) + POVERTY, Texas.shp)
```



I tried transforming `CRIMERATE` as well but it skewed it as much in the other direction so I will leave it as is. So far, all of my hypothesis appear to be correct. Visualizing the URBRURAL in a boxplot we have:

```
boxplot(CLINTONRATE ~ URBRURAL, data = Texas.shp)
```



We see that as the area becomes more urban, the median Clinton support increases.

[4] Initial Trial Regression Model

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?

[b] Is multicollinearity a problem?

[c] Are the model residuals approximately normally distributed?

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

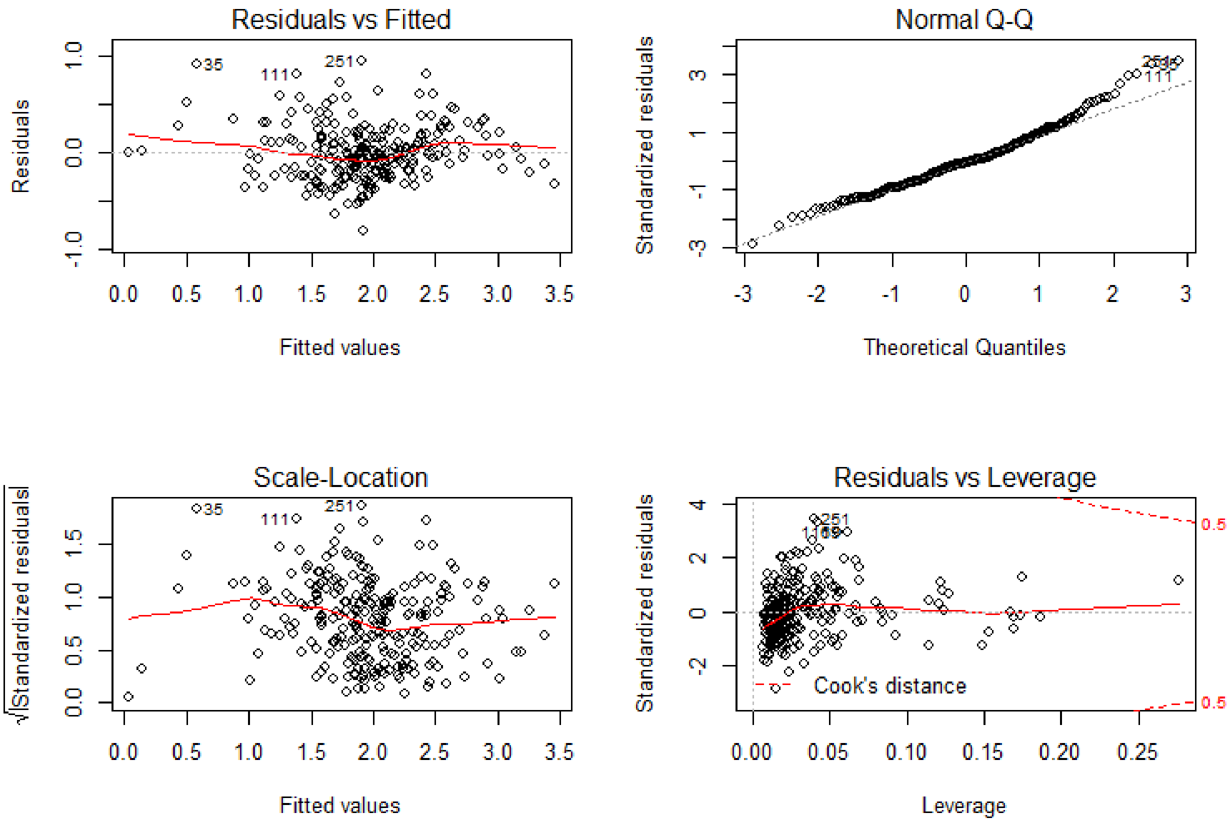
[e] Are there influential cases and outliers present in the model?

[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?

```
mod1 <- lm(CLINTONRATE ~ CRIMERATE + log(OBAVOT12) + log(HISPORG) +
             log(COLLEGEDEG) + POVERTY + URBRURAL, Texas.shp)

> summary(mod1)

Call:
lm(formula = CLINTONRATE ~ CRIMERATE + log(OBAVOT12) + log(HISPORG)
+ log(COLLEGEDEG) + POVERTY + URBRURAL, data = Texas.shp)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8124 -0.1935 -0.0206  0.1563  0.9583

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.843e-01  1.691e-01   1.090   0.2769
CRIMERATE       -7.812e-05  1.787e-05  -4.371 1.83e-05 ***
log(OBAVOT12)    2.368e-01  1.464e-02  16.176  < 2e-16 ***
log(HISPORG)     3.096e-01  2.559e-02  12.100  < 2e-16 ***
log(COLLEGEDEG) -1.028e-01  5.968e-02  -1.723   0.0862 .
POVERTY          1.900e+00  3.257e-01   5.834 1.71e-08 ***
URBRURALMetro   -1.422e-01  1.369e-01  -1.038   0.3001
URBRURALRural    1.151e-01  7.695e-02   1.496   0.1360
URBRURALUrban   -3.232e-02  1.139e-01  -0.284   0.7769
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2822 on 245 degrees of freedom
Multiple R-squared:  0.7679,  Adjusted R-squared:  0.7603
F-statistic: 101.3 on 8 and 245 DF,  p-value: < 2.2e-16
```

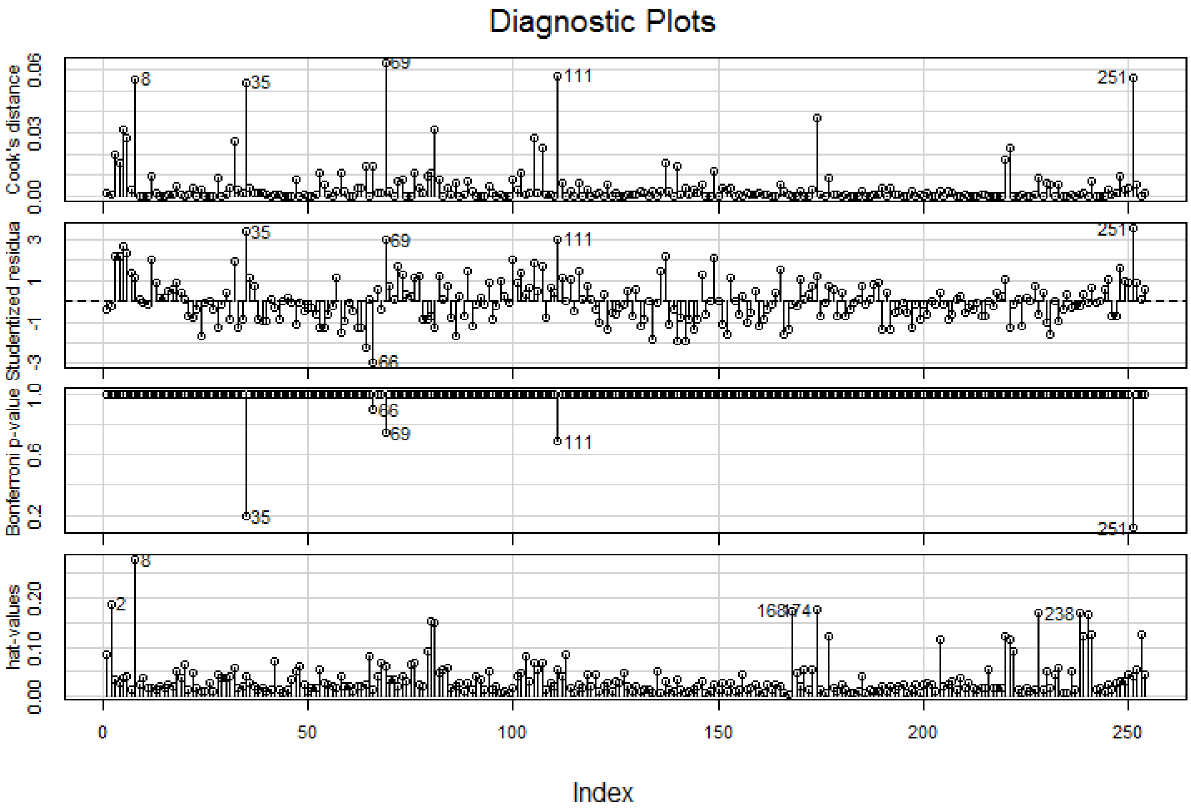
It appears that COLLEGEDEG and URBRURAL are not significant. The signs of the others are what we predicted although CRIMERATE carries a much lower influence (-0.000078) and POVERTY a much higher influence (1.9) than I anticipated. The adjusted R-squared value looks good at 0.7603 and significant. Checking for multicollinearity we see that it is not an issue here:

```
> vif(mod1)
                    GVIF Df GVIF^(1/(2*Df))
CRIMERATE       1.377770  1        1.173784
log(OBAVOT12)   2.369849  1        1.539431
log(HISPORG)    1.265293  1        1.124853
log(COLLEGEDEG) 1.662145  1        1.289242
POVERTY         1.376758  1        1.173353
URBRURAL        2.790384  3        1.186526

par(mfrow=c(2,2))
plot(mod1)
```



We would like to see a straighter line in the fitted values but this is okay and might improve after examining outliers, dropping insignificant variables or adjusting for spatial autocorrelation. From these plots, we can say the residuals look approximately normally distributed – no heteroscedasticity is immediately apparent. The log transformations performed previously appear to be appropriate.

```
influenceIndexPlot(mod1, id.n = 5)
```



Here we see potential outliers 35 and 251. They are Kent and Kenedy counties, the 3rd and 5th smallest estimated 2015 population counties. Kenedy is predominantly Hispanic at the north end of Padre Island on the coast and voted almost 50-50 in 2016. Kent is predominantly White and southeast of Lubbock and was a landslide against Clinton. Kenedy appears to have zero unemployment but I am not using that as a variable. Kent appears to have a very low poverty rate for its otherwise rural characteristics at 6% while Kenedy has a higher than average one at 28%. While they are on the edge of the dataset, I do not currently have a good reason to adjust or exclude them.

[5] Revised Regression Model

[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 1?

```
mod2 <- lm(CLINTONRATE ~ CRIMERATE + log(OBAVOT12) + log(HISPORG) +
POVERTY, Texas.shp)
> summary(mod2)

Call:
lm(formula = CLINTONRATE ~ CRIMERATE + log(OBAVOT12) + log(HISPORG)
   + POVERTY, data = Texas.shp)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8418 -0.1827 -0.0138  0.1528  0.9107 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.061e-01  1.031e-01   5.881 1.30e-08 ***
CRIMERATE     -7.337e-05  1.711e-05  -4.289 2.57e-05 ***
log(OBAVOT12)  2.066e-01  1.058e-02  19.529  < 2e-16 ***
log(HISPORG)   3.063e-01  2.529e-02  12.114  < 2e-16 ***
POVERTY        2.280e+00  3.069e-01   7.428 1.76e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2861 on 249 degrees of freedom
Multiple R-squared:  0.7574,	Adjusted R-squared:  0.7535 
F-statistic: 194.3 on 4 and 249 DF,  p-value: < 2.2e-16
```

All of the variables are now significant. The coefficients are close to what they were before except for `POVERTY` which went from 1.9 to 2.28 which may have taken up some of the influence lost by dropping `COLLEGEDEG` and `URBRURAL`. The adjusted R-squared value is about the same but we are using 4 degrees of freedom less so that's better. Overall, this model still supports the hypothesis that crime, poverty work against Clinton support and prior Obama support and Hispanic origin work in favor of it. A more thorough investigation would be needed to fully explain the role of urban/rural variation.

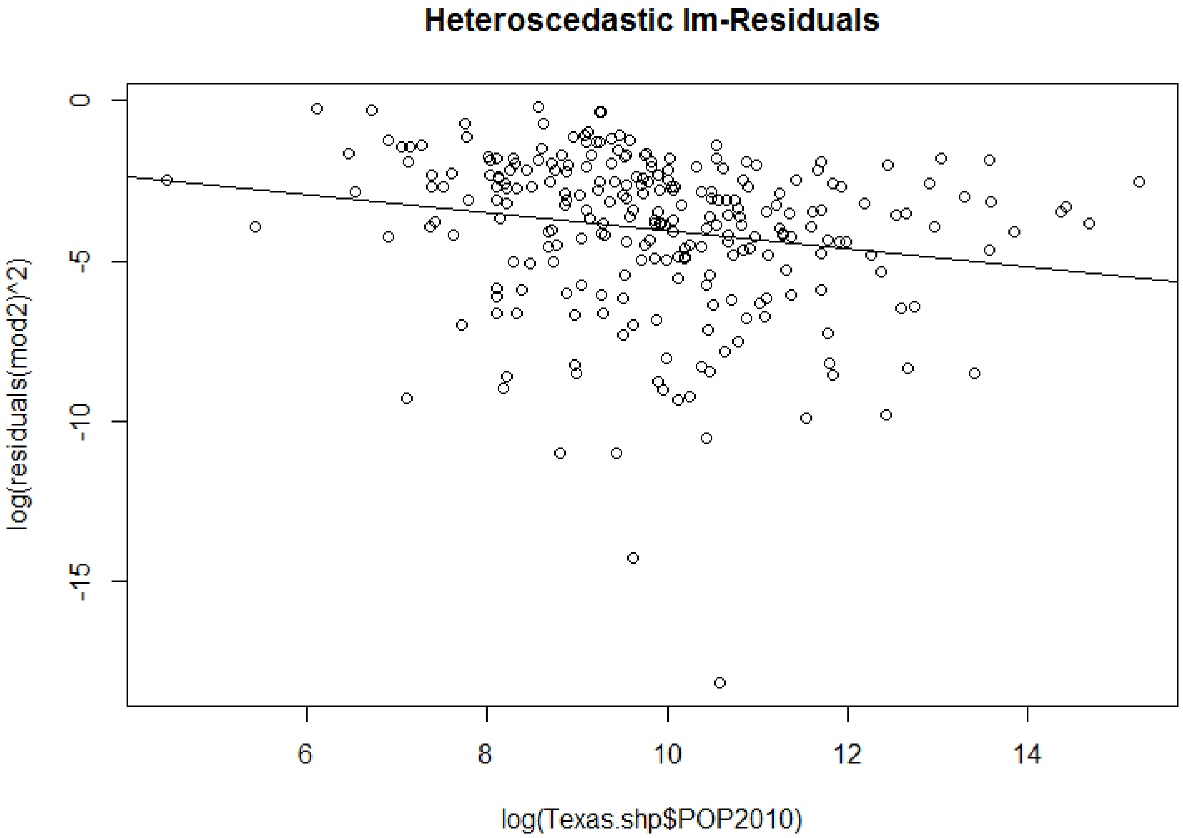
[6] Heteroscedasticity Investigation

Note: The size of the reference population underlying the voters' percentages for selected candidate varies widely from county to county.

[a] Estimate and interpret the parameters $\{\gamma_0, \gamma_1\}$ of the multiplicative heteroscedasticity model $\sigma_i^2 = \exp(\gamma_0 \cdot 1 + \gamma_1 \cdot \log{(refpop_i)}.)$.
[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.

```
> auxreg1 <- lm(log(residuals(mod2)^2) ~ log(Texas.shp$POP2010), 
Texas.shp)
> summary(auxreg1)

Call:
lm(formula = log(residuals(mod2)^2) ~ log(Texas.shp$POP2010), 
    data = Texas.shp)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.9332  -0.8729   0.5973   1.6377   3.5204 
```

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.22806    0.91955  -1.336  0.18292
log(Texas.shp$POP2010) -0.28197    0.09187  -3.069  0.00238 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.444 on 252 degrees of freedom
Multiple R-squared:  0.03604, Adjusted R-squared:  0.03221
F-statistic: 9.421 on 1 and 252 DF,  p-value: 0.00238

plot(log(residuals(mod2)^2) ~ log(Texas.shp$POP2010))
abline(auxreg1)
title("Heteroscedastic lm-Residuals")
```



The parameters are both negative and plot as a slightly negatively sloping line suggesting more variance of the residuals as the population increases.

```
lmH1 <- lmHetero(mod2, hetero= ~log(Texas.shp$POP2010),
   data = Texas.shp)
> summary(lmH1)

===============================================================
Multiplicatively Weighted Heteroscedasticity ML-Regrssion Model
===============================================================

Regression Coefficients:
                 Estimate     Std.Err z-value  Pr(>|z|)
(Intercept)    6.0611e-01  1.0205e-01  5.9393 2.863e-09 ***
CRIMERATE     -7.3372e-05  1.6939e-05 -4.3314 1.481e-05 ***
log(OBAVOT12)  2.0662e-01  1.0477e-02 19.7205 < 2.2e-16 ***
log(HISPORG)   3.0634e-01  2.5042e-02 12.2329 < 2.2e-16 ***
POVERTY        2.2795e+00  3.0391e-01  7.5006 6.350e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Gamma Coefficients:
                Gamma   Std.Err z-value  Pr(>|z|)
(Intercept) -2.522419  0.088736 -28.426 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

log-likelihood = -40.06316
```

When I tried to do a likelihood ratio test, the `lmHetero` output was not recognized as a linear model for some reason. I ran out of time to resolve the issue. I have posted the error below. The `anova` function did not work either.

```
> logLik(lmH1)
Error in UseMethod("logLik") :
  no applicable method for 'logLik' applied to an object of class
"lmHetero"
```

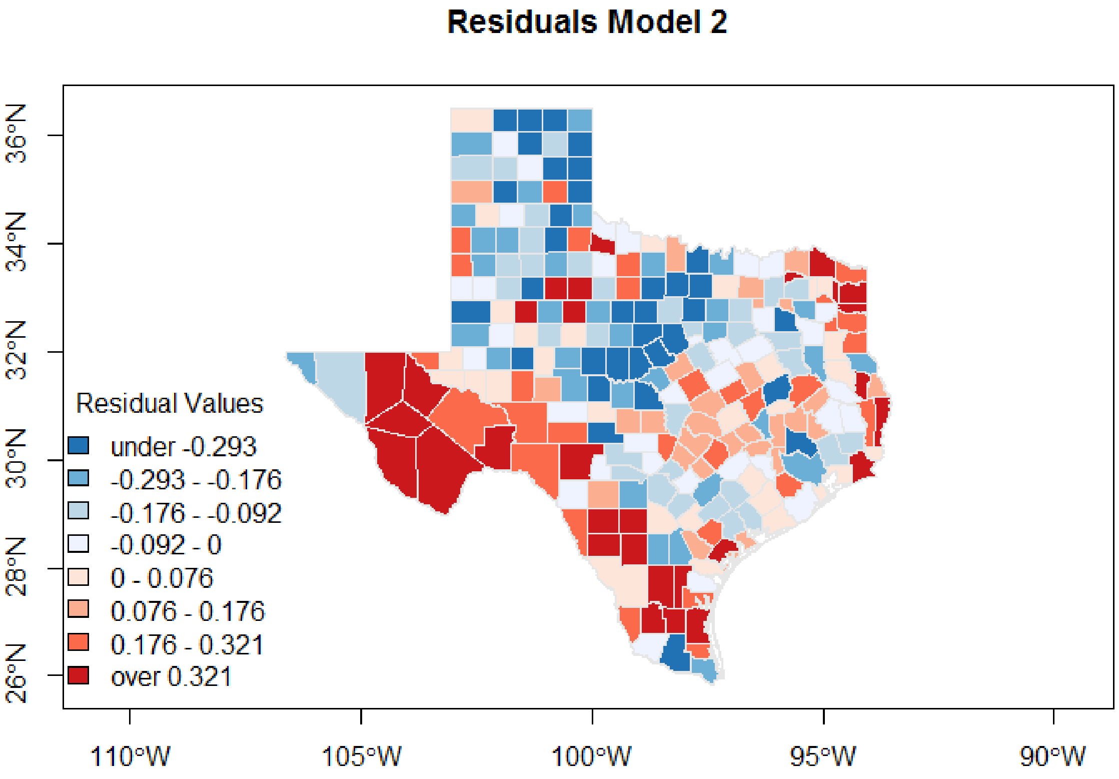
I do not see a substantial amount of heteroscedasticity from this output. The coefficients are similar to the prior model.

[7] Spatial Residual Analysis

For the spatial residual analysis tasks, you ***do not need to account for any potential heteroscedasticity*** in your model and you can proceed with the refined OLS model in item 5.

[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.

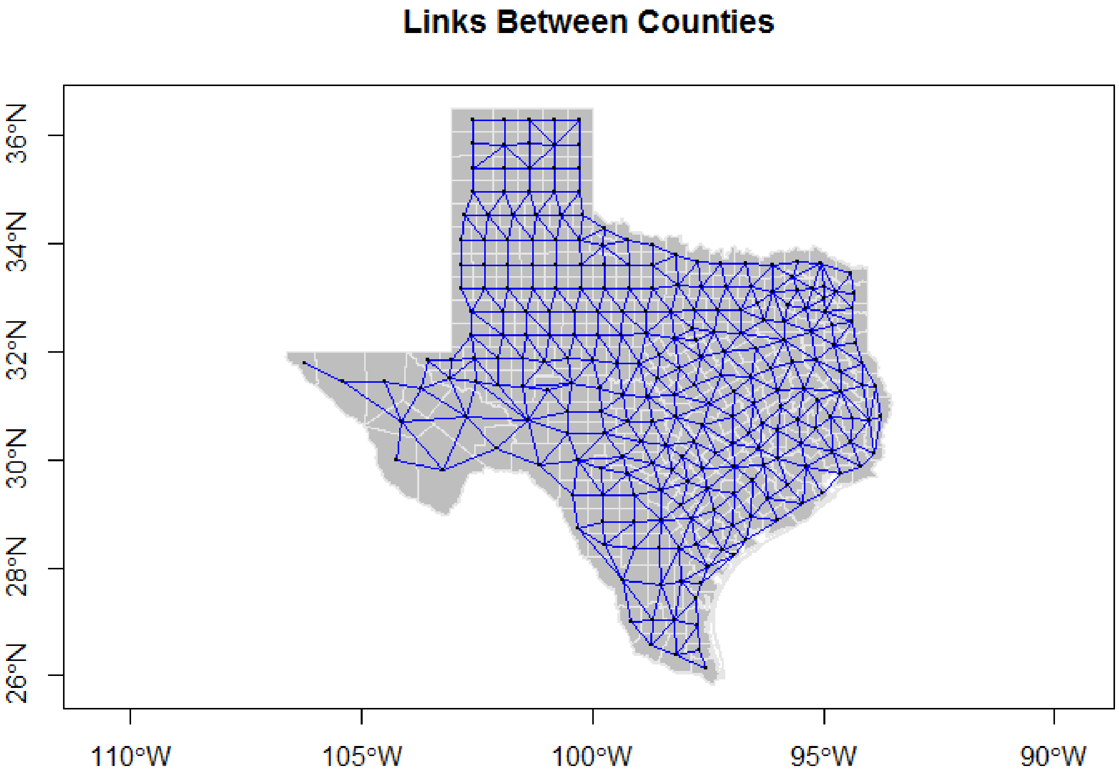
```
plotBiPolar(mod2$residuals, Texas.shp, my.title = "Residuals Model
   2", my.legend = "Residual Values")
```



There appear to be clustered residuals with strong negative clusters in central north Texas and the panhandle and moderate to strong positive clusters in central Texas as well as the southern and eastern borders of the state.

[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.

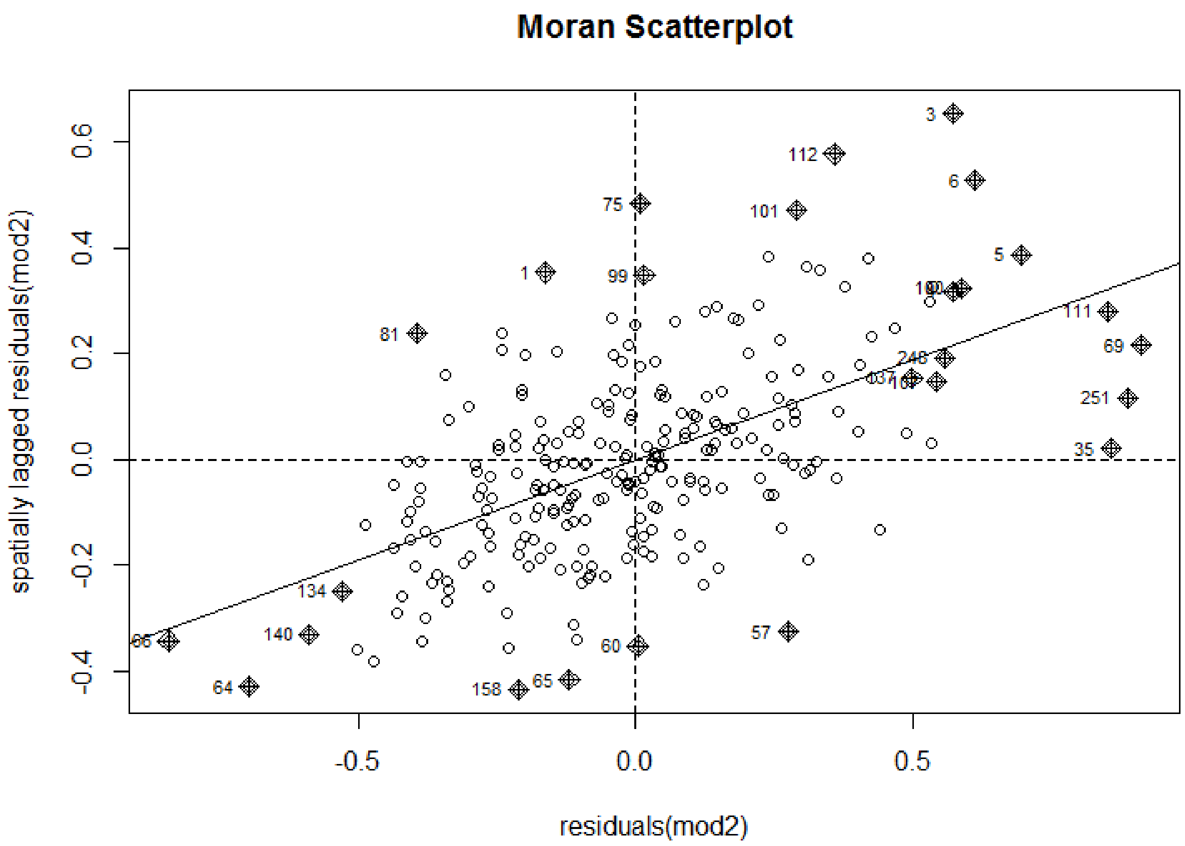
```
centroids <- coordinates(Texas.shp)
links <- poly2nb(Texas.shp, queen=F)
plot(Texas.shp, col="grey", border=grey(0.9), axes=T)
plot(links, coords=centroids, pch=19, cex=0.1,
     col="blue", add=T)
title("Links Between Counties")
```



The counties appear to be connected appropriately.

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

```
linkW <- nb2listw(links, style="W")
moran.plot(residuals(mod2), linkW)
```



There appears to be positive spatial autocorrelation give the positive slope in the line.

[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(mod2, linkW)

   Global Moran I for regression residuals

data:
model: lm(formula = CLINTONRATE ~ CRIMERATE + log(OBAVOT12) +
log(HISPORG) + POVERTY, data =
Texas.shp)
weights: linkW

Moran I statistic standard deviate = 9.9802, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
     0.377654333     -0.010627041      0.001513608
```

The observed Moran's I statistic is much larger than the expectation and we can therefore conclude there is significant positive spatial autocorrelation.

[8] Estimate a Spatial Autoregressive Model

[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.

```
> mod3 <- spautolm(mod2, data = Texas.shp, listw=linkW,
        family="SAR")
> summary(mod3)

Call: spautolm(formula = mod2, data = Texas.shp, listw = linkW,
family = "SAR")

Residuals:
      Min        1Q    Median        3Q       Max
-0.560364 -0.143193 -0.019194  0.137291  0.730361

Coefficients:
                 Estimate  Std. Error z value  Pr(>|z|)
(Intercept)    8.6870e-01  1.2813e-01  6.7801 1.201e-11
CRIMERATE     -6.5073e-05  1.3155e-05 -4.9466 7.552e-07
log(OBAVOT12)  1.8429e-01  1.1415e-02 16.1448 < 2.2e-16
log(HISPORG)   2.6217e-01  3.3757e-02  7.7664 7.994e-15
POVERTY        1.2827e+00  2.8822e-01  4.4503 8.576e-06

Lambda: 0.70216 LR test value: 91.649 p-value: < 2.22e-16
Numerical Hessian standard error of lambda: 0.05205

Log likelihood: 5.761099
ML residual variance (sigma squared): 0.049182, (sigma: 0.22177)
Number of observations: 254
Number of parameters estimated: 7
AIC: 2.4778

> AIC(mod2)
[1] 92.12631
```

Since I cannot get R to recognize linear models at the moment with `logLik` or `anova`, I will compare the models by their AIC values. The SAR model has a value of 2.4778 while the prior unadjusted model has an AIC value of 92.12631. Therefore we can conclude the SAR model is a better fit.

[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?

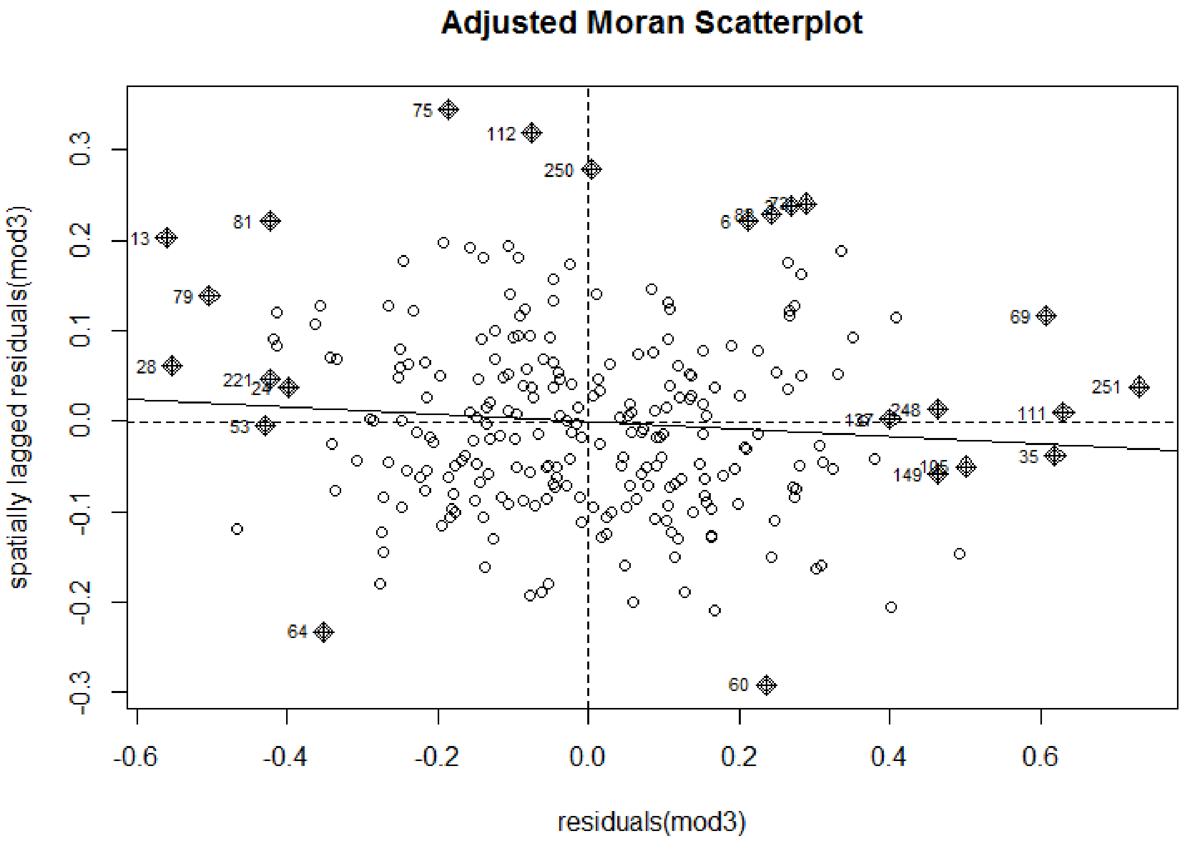
I'm having the same issue now with the Moran's I statistic:

```
> lm.morantest(mod3, linkW)
Error in lm.morantest(mod3, linkW) : mod3 not an lm object
```

As for the model coefficients, the intercept has changed from 0.61 to 0.87. The `CRIMERATE` influence was so small already that the change is negligible. `OBAVOT12` changed from 0.21 to 0.18. `HISPORG` changed from 0.31 to 0.26. And `POVERTY` changed from 2.28 to 1.28. They are all still highly significant. These differences suggest that prior support for Obama, Hispanic origin, and especially poverty are all spatially autocorrelated in Texas.

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

```
moran.plot(residuals(mod3), linkW)
title("Adjusted Moran Scatterplot")
```



From this scatterplot, we see the slope is now only slightly negative rather than the substantially positive prior plot slope. In this plot, the slope appears to be about -0.05 while the prior plot's slope was almost 0.4.