Advanced Statistics
Assignment 4
Claude M. Schrader

Research Question:

This report will analyze whether or not broadband adoption in Kentucky is correlated with college education level, population density, and broadband availability, and measure the autocorrelation between counties in this data.

Measures:

adopt05 is the percentage of Kentucky households subscribing to broadband from a 2005 survey. This is the dependent variable.
college tracks the percent of college graduates in the population in 2004.
hhpct_0401 is the percentage of Kentucky households with broadband available.
popden2000 is the Population Density as of the 2000 census.

	adopt05	college	hhpct_0401	popden2000
Minimum	0.071000000	0.041000000	0.00400000	16.100000
Maximum	0.548000000	0.472000000	1.00000000	1336.700000
Mean	0.238783333	0.208175000	0.55985000	83.035833
Variance	0.010237717	0.006586062	0.08080245	23495.153075
Std. Dev.	0.101181606	0.081154554	0.28425771	153.281287

Methods:

We begin with an OLS model of our chosen variables. This will serve as a baseline to which we can compare the spatial model. We have chosen to regress adopt05 on college, hhpct_0401, and popden2000. We have already seen in Assignment 1 that all three independent variables have a statistically significant effect on adopt05, and that this combination explains a decent amount of variance in adopt05, as compared to a few other attempted models.

We have decided on using the Rook's case to calculate the neighbor list in this analysis. When calculating the list of a state's neighbors, it will include counties with which it shares a border, but not counties only adjacent where their corners meet.

Once this list is calculated, we will add spatial weights, which will give us a lagged means list, which is key to the rest of our analysis. This calculates our dependent variable (adopt05) compared to the mean of its neighbors as calculated previously. We will create a map of these lagged means, which will provide a good illustration of how strongly the broadband adoption is clustered (spatial autocorrelation). We will also create a Moran's Plot, which is a scatterplot visualizing the strength of the spatial lag in the adopt05 variable. This can provide a clear illustration of the strength and direction of any spatial autocorrelation.

For another quantitative measure, we will calculate the Moran's index of autocorrelation. This provides a quantitative measurement of the strength of spatial autocorrelation in the data, and provides a p-value to determine its significance.

Lastly, we will run our Spatial Autoregression model. This provides a model similar to the traditional linear OLS model which is aware of spatial dependency. This can help resolve errors and inconsistencies introduced by autocorrelation, and can provide a more accurate model.

Results:

Figures two and three show the relationships between counties. Figure three is a map created from the lagged means data. Darker values on this map indicate areas where a county's broadband adoption is affected more strongly by its neighbors. Figure two is a visualization of the Rook's Case relationships used to generate this information.

Figure four is a Moran's plot, or a scatterplot of the lagged broadband adoption values with a fit line. The Y axis indicates the lagged adopt05, and the x axis is the raw adopt05. Since the fit line has a positive slope, we see there may be some moderate positive autocorrelation present in our data. Our Moran's I calculation includes a possible range of -.77 to 1.02, and our Moran's I test has a value of 0.2561, and is statistically significant at the 99% level. The Moran's I value from a Monte Carlo simulation with 10,000 trials gives very similar values.

```
Call:lagsarlm(formula = kentucky$adopt05 ~ kentucky$college + kentucky$hhpct_0401 +
kentucky$popden2000, data = kentucky, listw = KY_Neigh_lw)

Residuals:
       Min         1Q     Median         3Q        Max
-0.1594178 -0.0520077 -0.0072749  0.0500588  0.1878281

Type: lag
Coefficients: (asymptotic standard errors)
                      Estimate Std. Error z value  Pr(>|z|)
(Intercept)         1.2856e-02 2.5373e-02  0.5067 0.6123623
kentucky$college    3.6944e-01 9.6832e-02  3.8153 0.0001360
kentucky$hhpct_0401 9.1700e-02 2.6068e-02  3.5177 0.0004353
kentucky$popden2000 1.4045e-04 4.8248e-05  2.9110 0.0036029

Rho: 0.3616, LR test value: 10.488, p-value: 0.0012012
Asymptotic standard error: 0.097578
    z-value: 3.7057, p-value: 0.00021078
Wald statistic: 13.732, p-value: 0.00021078

Log likelihood: 145.8057 for lag model
ML residual variance (sigma squared): 0.0050063, (sigma: 0.070755)
Number of observations: 120
Number of parameters estimated: 6
AIC: -279.61, (AIC for lm: -271.12)
LM test for residual autocorrelation
test value: 2.6443, p-value: 0.10392
```

Looking at the results of the SAR model, all three of our independent variables are all statistically significant at the 99% level, which tells us that they all are useful additions to the model. The LM test for residual autocorrelation is not statistically significant, which tells us there is no strong autocorrelation remaining in the residuals. This is a sign that our model explains a good amount of the variation in adopt05.

Looking at Figure One, we can detect some clustering in the data. Because Rho is significant with a value of .3616, that tells us there is some level of spatial autocorrelation in our data, although it is moderate. If Rho was not significant, that would mean that the clustering on our map is by chance, and that a county's level of broadband adoption is not strongly affected by its neighbors.

```
> anova(model01.lag, model01)
            Model df     AIC logLik Test L.Ratio   p-value
model01.lag     1  6 -279.61 145.81    1
model01         2  5 -271.12 140.56    2  10.489 0.0012012
```

This ANOVA table compares our original OLS model with the spatially lagged model. We see that for the lagged model, AIC is lower and the log likelihoods test is higher, which indicates a small preference for the SAR model. However, the two tests have the same p-value, so any preference might not actually affect the statistical significance very strongly.

Conclusion:

Looking at the data and charts generated by this analysis, our data does have some spatial autocorrelation, so a spatial model should be appropriate. From the outputs from the SAR model and ANOVA table, this model is statistically significant, and may offer a small advantage over the linear model.

Figure 1: Kentucky Broadband Adoption

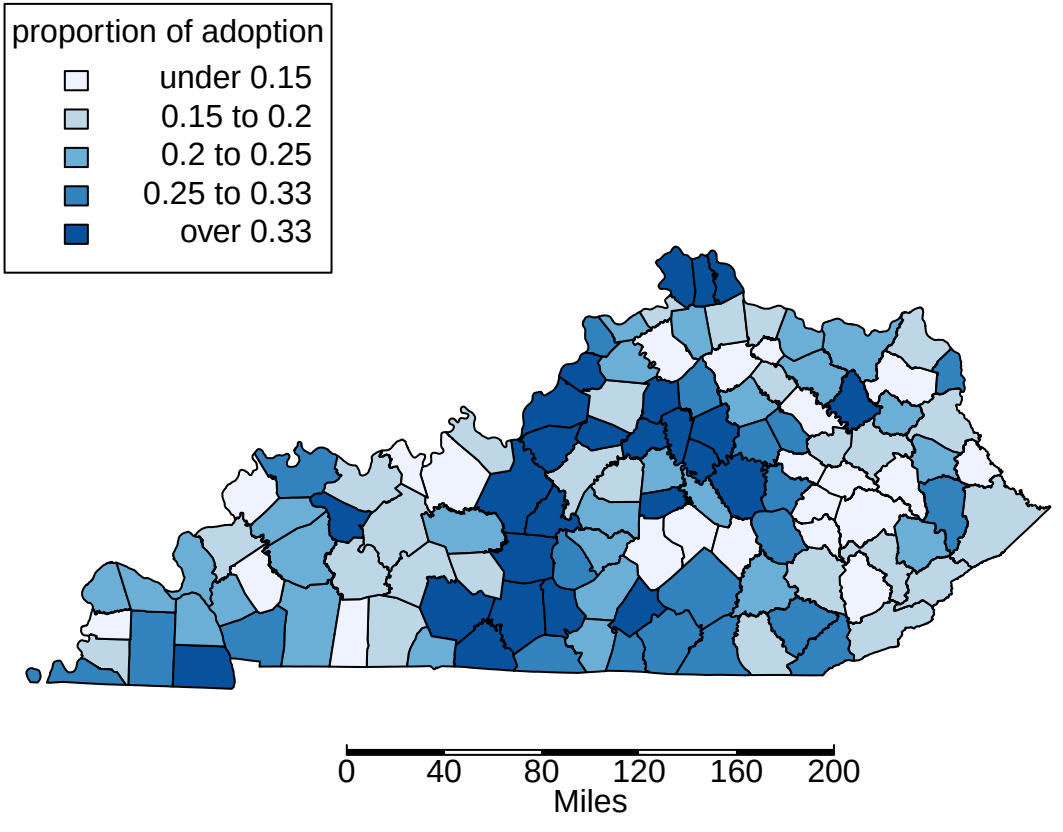


Figure 2: Weighted Neighbor plot of Kentucky Broadband Adoption

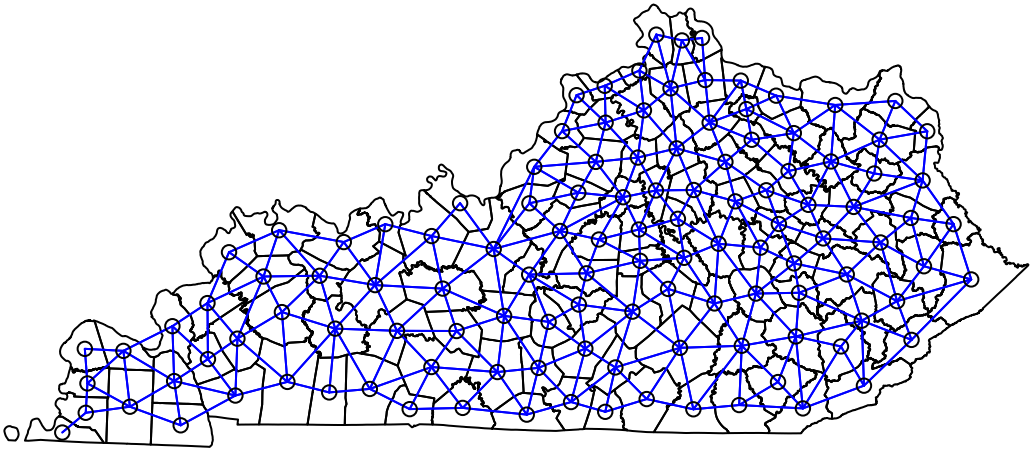


Figure 3: Lagged Means plot of Kentucky Broadband Adoption

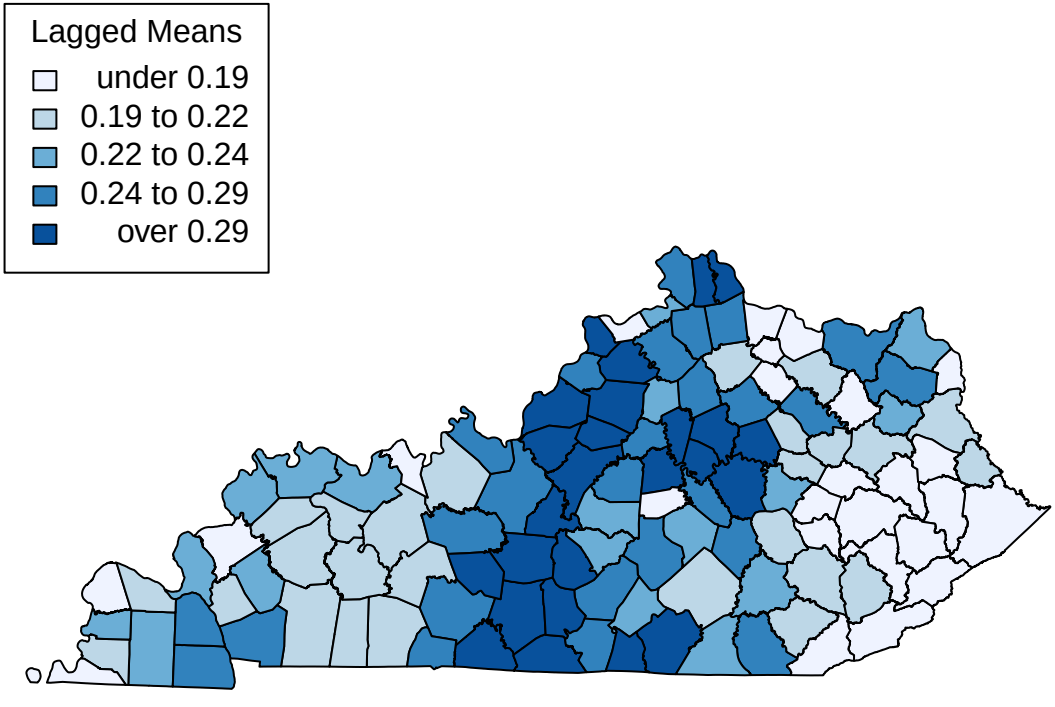
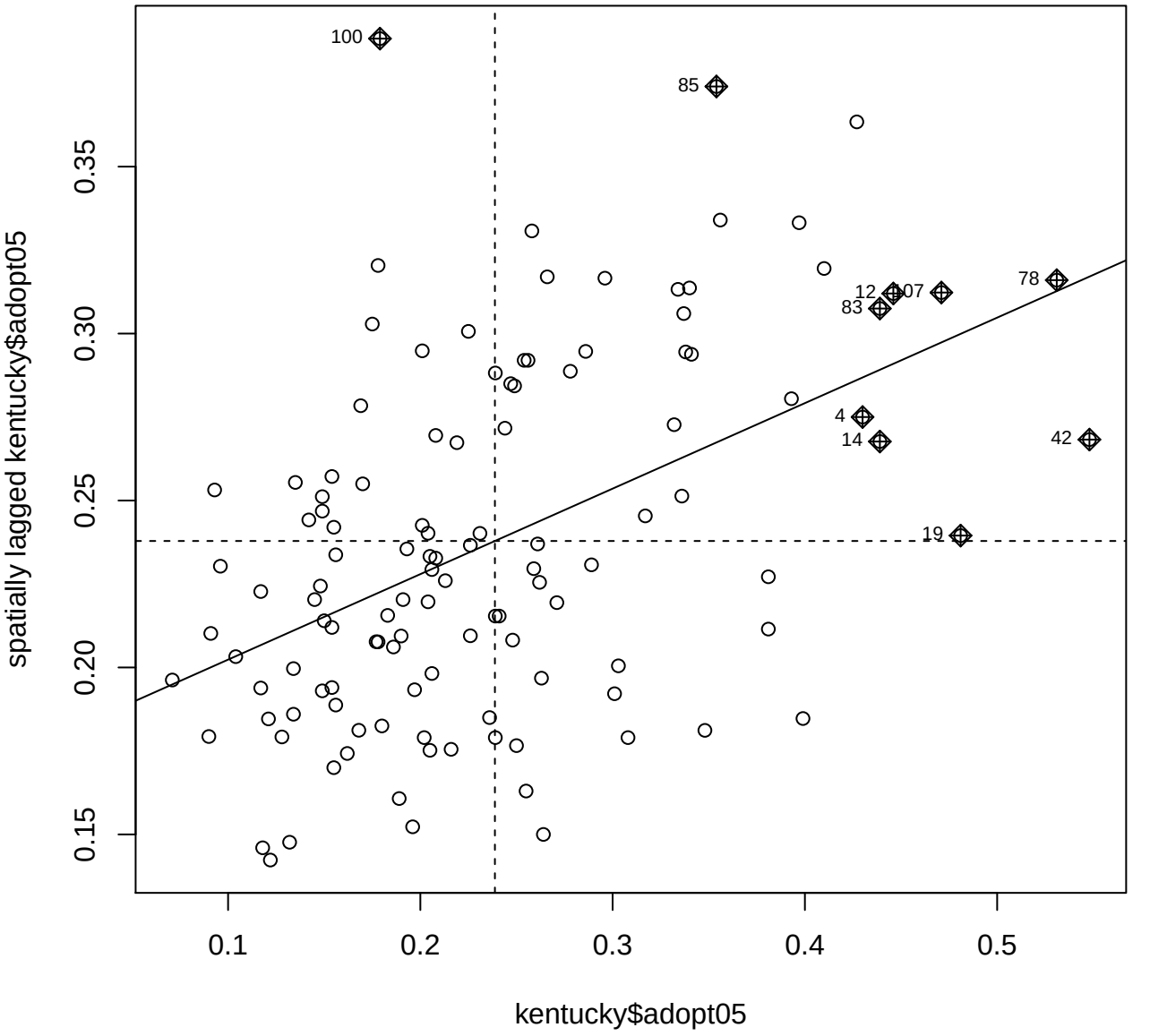


Figure 4: Moran's Plot



Appendix: R code

```
rm(list = ls())

setwd("C:/Users/tuj53509/Dropbox/docs/Temple/Advanced Statistics for Urban Applications/Assignment4")

library(GISTools)

library(maptools)

library(lm.beta)

library(spdep)



kentucky = readShapeSpatial("ky_counties/ky_counties", proj4string=CRS("+proj=lcc"))

kentucky_census = read.csv("KY.bband.csv")


# Join census file to shapefile

kentucky@data <- data.frame(kentucky@data, kentucky_census[match(kentucky@data[,"NAME10"],kentucky_census[,"Name"]),])


#Linear Model

model01 <- lm(kentucky$adopt05 ~ kentucky$college + kentucky$hhpct_0401 + kentucky$popden2000)

#add standardized betas

model01.std <- lm.beta(model01)

summary(model01.std)

summary(model01$residuals)


pdf(file="f1.adopt05choropleth.pdf")

adopt05.shades <- auto.shading(kentucky$adopt05, cols=brewer.pal(5,"Blues"))

choropleth(kentucky, kentucky$adopt05, shading = adopt05.shades)

title("Figure 1: Kentucky Broadband Adoption")

choro.legend(3751596,4833529, adopt05.shades, title = "proportion of adoption")

map.scale(5021404, 3205804, miles2ft(200), "Miles", 5, 40)

dev.off()
```

```
#Lagged Means

#plot queen's case just for comparison's sake

# kentucky_neighbors_queen <- poly2nb(kentucky, queen = TRUE)

# plot(kentucky, main = "DIAG: Queen's Case Weighted Neighbor plot")

# plot(kentucky_neighbors_queen, coordinates(kentucky), add=T, col='blue')



kentucky_neighbors <- poly2nb(kentucky, queen=FALSE)

pdf(file = "f2.weightedneighborplot.pdf")

plot(kentucky, main = "Figure 2: Weighted Neighbor plot of Kentucky Broadband
Adoption ")

plot(kentucky_neighbors, coordinates(kentucky), add=T, col='blue')

dev.off()

KY_Neigh_lw <- nb2listw(kentucky_neighbors)

bband <- kentucky$adopt05

bband_lagged_mean <- lag.listw(KY_Neigh_lw, bband)

#create a choropleth map from the lagged means result

par(mar = c(2,1,2,1))

pdf(file="f3.lagged_means_choropleth.pdf")

laggedmeanshades = auto.shading(bband_lagged_mean,
cols=brewer.pal(5,"Blues"))

choropleth(kentucky, bband_lagged_mean, shading =laggedmeanshades)

title("Figure 3: Lagged Means plot of Kentucky Broadband Adoption")

choro.legend(3751596,4833529, laggedmeanshades, title = "Lagged Means")

dev.off()


#create a Lagged Means Plot (Moran's Plot)

pdf(file = "f4.moransplot.pdf")

moran.plot(kentucky$adopt05, KY_Neigh_lw)

title("Figure 4: Moran's Plot")

#this closes the file handle

dev.off()
```

```
#Run a Moran's I test
moran.range <- function(lw) {
  wmat <- listw2mat(lw)
  return(range(eigen((wmat + t(wmat))/2)$values))
}
moran.range(KY_Neigh_lw)
#approximate test statistic using normal distribution
moran.test(kentucky$adopt05, KY_Neigh_lw)
#calculate the test statistic using 10,000 random trials
moran.mc(kentucky$adopt05, KY_Neigh_lw, 10000)


#run the SAR model
model01.lag <- lagsarlm(kentucky$adopt05 ~ kentucky$college +
kentucky$hhpct_0401 + kentucky$popden2000, data = kentucky, KY_Neigh_lw)
summary(model01.lag)
anova(model01.lag, model01)
```