Spatial Statistics

Assignment No 5

Multivariate Regression and Spatial Lag Model

Muhammad Bilal
Matriculation: - 12214473

Q.No.1

In Assignment 4 you have investigate the effect of the explanatory variable Median Age 2013 (MdA_2013) on the dependent variable Average GCSE Score (AGc_2).

A. Include at least one more predictor / independent variable in your model that improves the overall model fit.
B. Visually inspect and interpret linearity of relationships between predictors and criterion variable as well as homoscedasticity. Transform variables if needed.
Ans: -

To start by loading required libraries like sf, ggplot2, dplyr and then Visualize Ward data based on Average GCSE Score

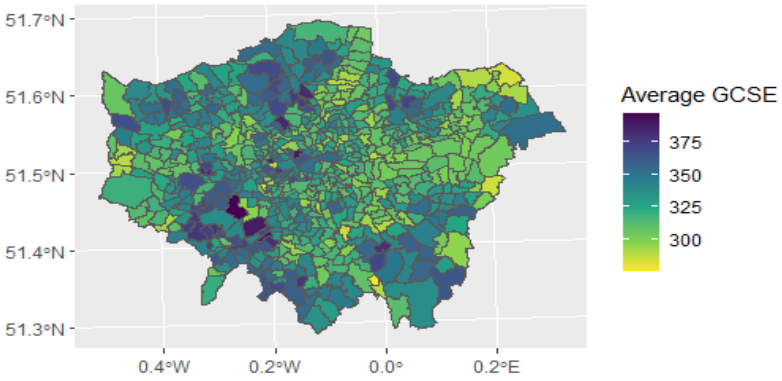


Figure 1

Employment Rate another variable is chosen to plot the values against the Average GCSE Score

```
#Ploting Another Independent Variable Employment rate (16-64) - 2011 and Average GCSE
ggplot(data = wards, aes(x = E_16_, y = AGc_2)) +
  geom_point() +
  xlab("Employment rate (16-64)") +
  ylab("Average GCSE - 2014") +
  theme_minimal()
```

Implementing Linear Regression Model to see correlation between Employment rate (16-64) and Average GCSE score.

```
#Apply Linear Regression Method to see correlation between Employment and GSCE Score
ggplot(data = wards, aes(x = E_16_, y = AGc_2)) +
  geom_point() +xlab("Employment rate (16-64)") +
  ylab("Average GCSE - 2014") +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=FALSE) +
  theme_minimal()
```

Output of the LM Model as shown below in figure 2. From Visually you can see there is positive correlation between employment rate and Average GCSE Score as one variable increases the other variable also increases.

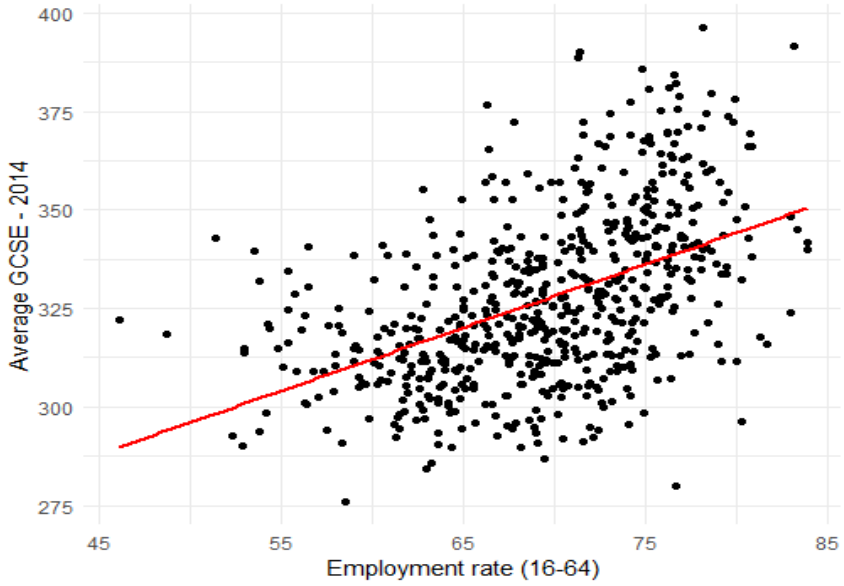


Figure 2

Multivariate Regression Model

Now I take both independent variables that are median age and Employment rate into account to see their correlation with dependent variables that is Average GCSE Score.

```
# Summary of the Multivariate model for two Independent variable with GCSE Score
multivariate <- lm(AGc_2 ~ MdA_2013+E_16_, data=wards)
summary(multivariate)
```

```
> # Summary of the Multivariate model for two Independent variable with GCSE Score
> multivariate <- lm(AGc_2 ~ MdA_2013+E_16_, data=wards)
> summary(multivariate)

Call:
lm(formula = AGc_2 ~ MdA_2013 + E_16_, data = wards)

Residuals:
   Min     1Q Median     3Q    Max
-50.11 -12.24  -1.32  10.26  61.87

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 200.7477     8.0670  24.885  < 2e-16 ***
MdA_2013      1.7210     0.2314   7.437 3.42e-13 ***
E_16_         0.9673     0.1409   6.864 1.62e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.14 on 623 degrees of freedom
Multiple R-squared:  0.2952,    Adjusted R-squared:  0.293
F-statistic: 130.5 on 2 and 623 DF,  p-value: < 2.2e-16
```

Figure 3

Interpretation: -

R-Squared

R squares show the percentage of variation in dependent variable explained by the two independent variables so in our case 29% of variation in Average GCSE Score can be explained by the variables Median Age and employment rate and about 71% can be attributed to other factors. Although the model does not show significant improvement while considering two independent variables but still it improves from the previous one.

P-value

There were two hypotheses considered in this scenario first there is no relationship between the dependent and independent variables (H0) and second is there is relationship between them (HA). According to our model result from standard significance level 0.05, the p-value is quite small so we will reject the null hypothesis, and this shows that there is significant relationship between average GCSE score and median age + employment rate.

Heteroscedasticity

The residual is not constant across all levels of predictor variable therefore this residual distribution is Heteroscedasticity.

Part C: - Multicollinearity Test

This happens when two or more predictors strongly correlate with each other that affect the model output therefor to check this I applied Pearson method. The estimated correlation coefficient between the two variables is 0.6080128. This value indicates a moderate linear relationship between median age and employment rate.

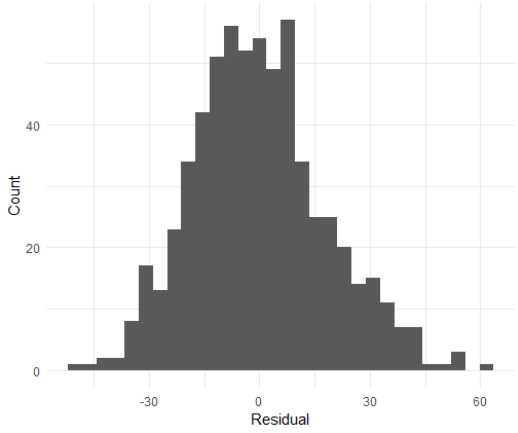
```
#Multicollinearity Test
cor.test(wards$MdA_2013, wards$E_16_, method = "pearson")
```

```
        Pearson's product-moment correlation

data:  wards$MdA_2013 and wards$E_16_
t = 19.13, df = 624, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5561477 0.6551604
sample estimates:
      cor
0.6080128
```

Part D: - Modelling and Visualization Residuals of Multivariate Model

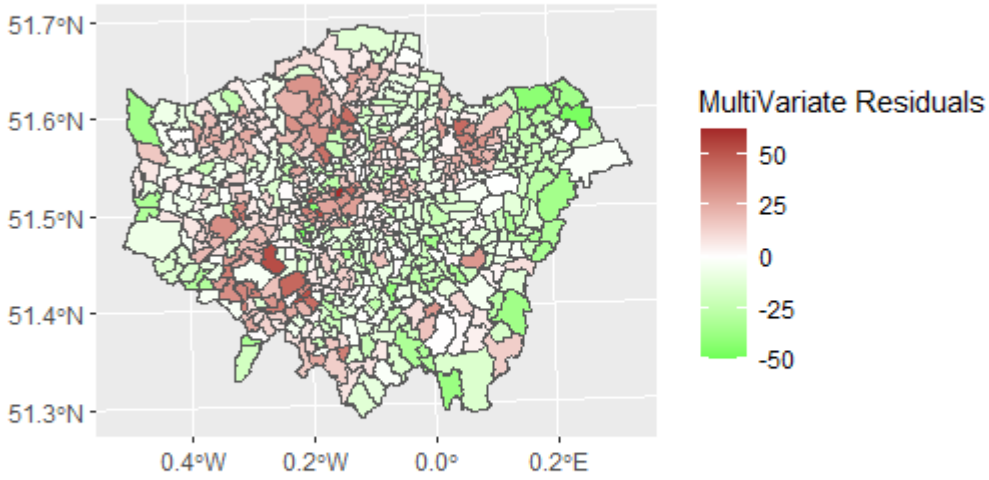
This shows that residuals are symmetrically distributed and represent good multivariate model.



Visualizing Residuals on Map

```
#Visualize residuals in map
wards <- wards %>%
  mutate(resids = res.df$res)

ggplot(data = wards) +
  geom_sf(aes(fill = resids)) +
  scale_colour_gradient2( low = "green", mid = "white",
                          high = "brown", midpoint = 0,
                          aesthetics = "fill") +
  labs(fill='MultiVariate Residuals')
```



Interpretation: -
This map of residuals shows that they are spatially autocorrelated. Positive values cluster together, and negative values are also cluster together means the wards with positive value have also positive in its nearby wards and vice versa.

Part E: - Spatial Lag Regression – Nearest Neighbors Matrix

```
#Spatial Lag Regression
coordsW <- st_geometry(wards1) %>%
  st_centroid()
LWard_nb2 <-coordsW %>%
  knearneigh(., k=4)

LWard_knn <- LWard_nb2 %>%
  knn2nb()
# run spatial lag model
slag_model_knn <- lagsarlm(AGc_2 ~MdA_2013+E_16_,
                           data = wards,
                           nb2listw(LWard_knn, style="C"))

tidy(slag_model_knn)
```

```
# A tibble: 4 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 rho            0.500    0.0401     12.5  0
2 (Intercept)   78.9     11.3         6.95 3.54e-12
3 MdA_2013       1.09     0.213       5.14 2.78e- 7
4 E_16_          0.675    0.126       5.34 9.37e- 8
>
```

The rho is a spatial lag coefficient shows 0.500 means there is a positive spatial autocorrelation in the data. GCSE scores in each ward are spatially autocorrelated, meaning that high scores in one ward are also associated with high scores in the neighboring wards. The median age and employment rate have a positive impact on Average GCSE Scores. This model is useful because it also considers the spatial distribution.