Regression Analysis

Regression Analysis in Practice

Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Emergency Department Healthcare Costs : Model Fit and Assessment

Georgia Tech

About This Lesson

Georgia Tech

Multiple Linear Regression Model

```
## Exclude GEOID, scaling factor PMPM, and confounding factors EDCost and ED
## Exclude OtherPop & ComplexPop because of linear dependence
dataAdult.red = dataAdult[, -c(1, 3, 4, 5, 10, 13)]

fullmodel = lm(log(EDCost.pmpm) ~ ., data=dataAdult.red)
summary(fullmodel)
```

Georgia Tech

Multiple Linear Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.208e+00	1.175e-01	18.788	< 2e-16 ***
StateAR	9.235e-01	1.610e-02	57.353	< 2e-16 ***
StateLA	9.081e-01	1.358e-02	66.853	< 2e-16 ***
StateNC	1.418e+00	1.650e-02	85.909	< 2e-16 ***
HO	1.168e+01	7.587e-01	15.401	< 2e-16 ***
PO	1.378e-01	4.114e-02	3.350	0.000815 ***
WhitePop	4.416e-03	5.800e-04	7.614	3.16e-14 ***
BlackPop	4.894e-03	5.824e-04	8.403	< 2e-16 ***
HealthyPop	-9.044e-04	8.160e-04	-1.108	0.267751
ChronicPop	-5.949e-03	2.052e-03	-2.899	0.003760 **
Unemployment	4.390e-04	7.377e-04	0.595	0.551797
Income	-2.556e-07	2.774e-07	-0.922	0.356769
Poverty	-3.306e-04	4.460e-04	-0.741	0.458529
Education	-1.447e-03	3.296e-04	-4.390	1.16e-05 ***
UrbanicitySuburban	-4.565e-04	1.369e-02	-0.033	0.973406
UrbanicityUrban	2.067e-02	1.269e-02	1.629	0.103356
Accessibility	-1.965e-03	7.094e-04	-2.770	0.005623 **
Availability	8.037e-02	1.975e-02	4.068	4.81e-05 ***
RankingsPCP	7.596e-04	1.819e-04	4.175	3.03e-05 ***
RankingsFood	6.586e-03	5.203e-03	1.266	0.205642
RankingsHousing	-4.642e-03	1.562e-03	-2.973	0.002967 **
RankingsExercise	3.993e-04	2.332e-04	1.712	0.086907 .
RankingsSocial	-3.895e-04	1.347e-03	-0.289	0.772497
ProvDensity	6.042e-02	1.573e-02	3.841	0.000124 ***

Socioeconomic predicting variables *Unemployment*, *Income*, *Poverty* and *RankingsSocial* are **not** statistically significant given other predicting variables in the model.

Access to primary care variables *Accessibility* and *Availability* are statistically significant.

85% of the variability in the ED cost is explained.

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

Residual standard error: 0.2321 on 4995 degrees of freedom
Multiple R-squared: 0.8486, Adjusted R-squared: 0.8479
F-statistic: 1218 on 23 and 4995 DF, p-value: < 2.2e-16

Georgia Tech

Residual Analysis: Outliers & Normality

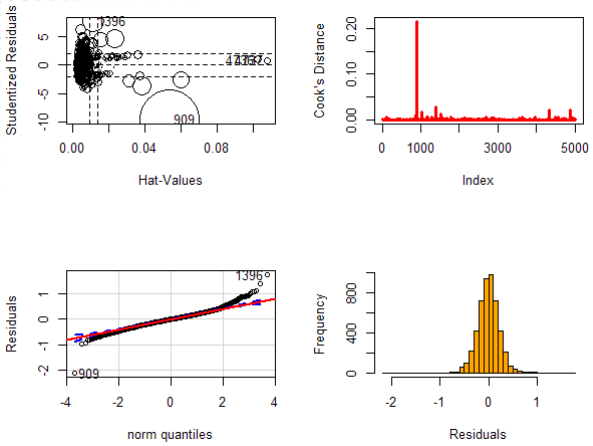
```
## Residuals versus individual predicting variables
full.resid = residuals(fullmodel)
cook = cooks.distance(fullmodel)

## Check outliers
influencePlot(fullmodel)
plot(cook, type="h", lwd=3, col="red", ylab="Cook's Distance")

## Check Normality
qqPlot(full.resid, ylab="Residuals", main = "")
qqline(full.resid, col="red", lwd=2)
hist(full.resid, xlab="Residuals", main = "", nclass=30, col="orange")
```

Georgia Tech

Residual Analysis: Outliers & Normality



Outliers

Observation 909 stands out.

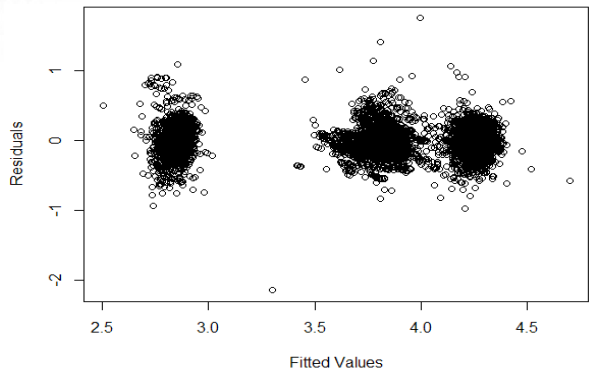
Normality

Symmetric, but with heavy tails.

Georgia Tech

Residual Analysis: Constant Variance and Uncorrelated Errors

```
## Check Constant Variance & Uncorrelated Errors
full.fitted = fitted(fullmodel)
par(mfrow=c(1,1))
plot(full.fitted, full.resid, xlab="Fitted Values", ylab="Residuals")
```



Constant Variance Assumption

No pattern

Uncorrelated Errors Assumption

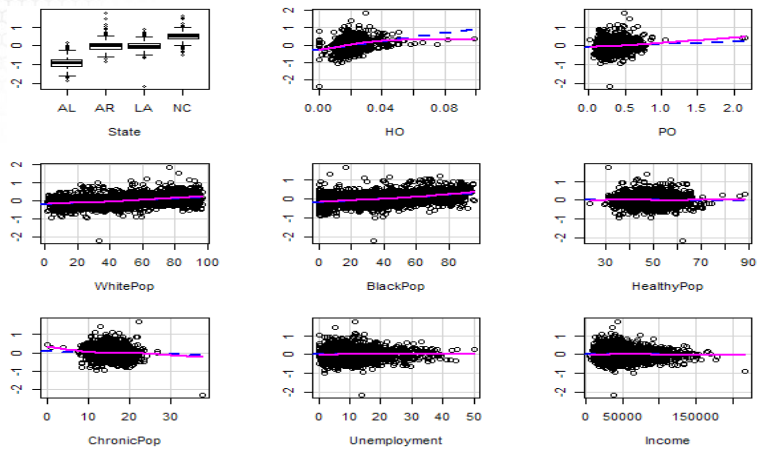
Three well-defined clusters

Spatial Dependence

Georgia Tech

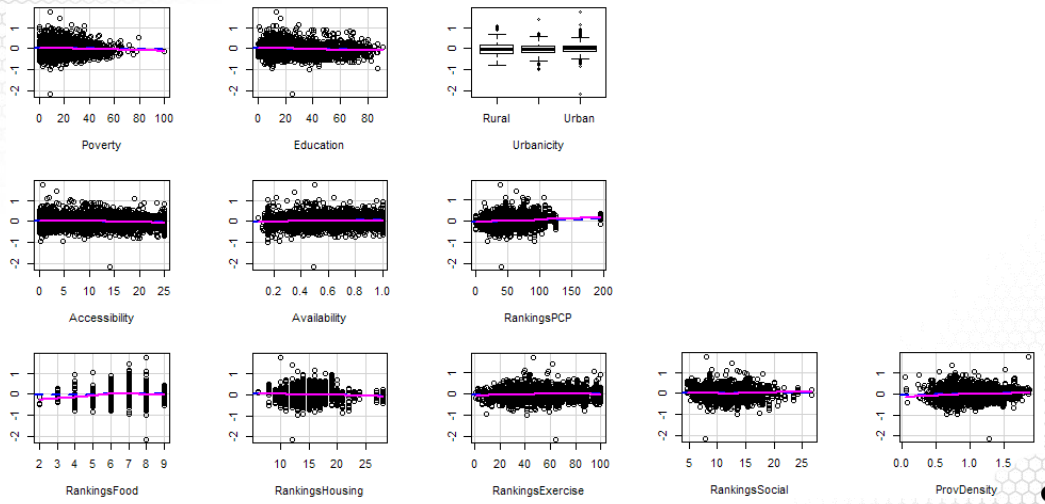
Residual Analysis: Linearity

```
## Check Linearity
crPlots(fullmodel, ylab="")
```



Georgia Tech

Residual Analysis: Linearity (cont'd)



Georgia Tech

Summary

Georgia Tech