Lecture 6: Regression Part 2

GGR376

Dr. Adams

Model Interpretation: Coefficients

```
model_mpg <- lm(cty~displ+cyl, data = mpg)
summary(model_mpg)$coefficients
```

```
             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 28.288512  0.6876399 41.138555 2.721700e-108
displ       -1.197882  0.3407738 -3.515181  5.287524e-04
cyl         -1.234654  0.2731967 -4.519285  9.908652e-06
```

Model Interpretation: \(R^2\)

```
summary(model_mpg)$adj.r.squared
```

```
[1] 0.6641936
```


Application

How do we use a linear regression model?

Explanatory

- Used to understand the relationships in existing data.
    - Coefficients, when x increases how does *Y* change

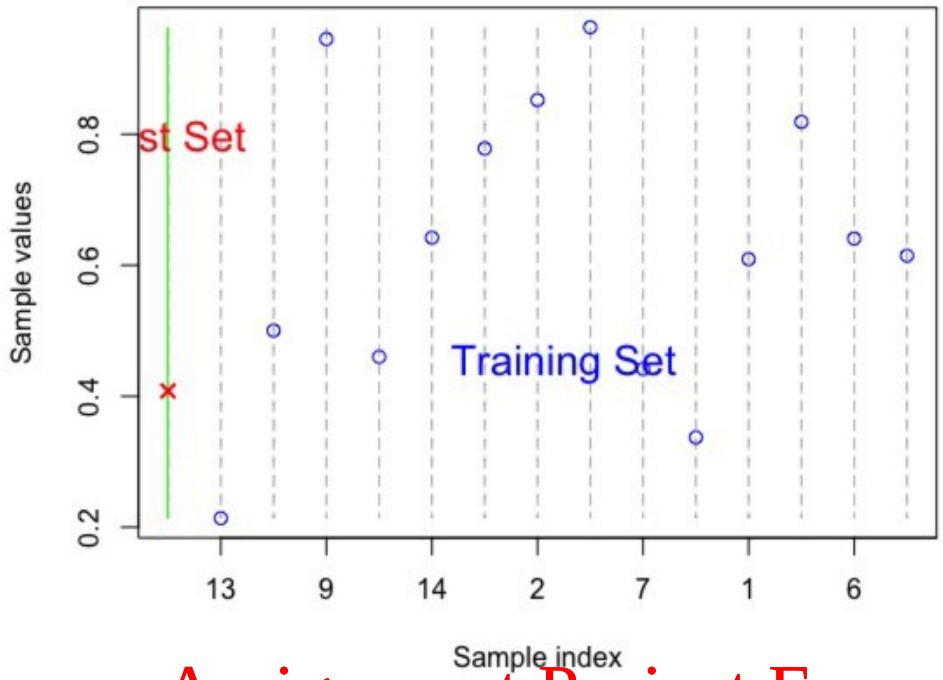
Predictive

- Predicting the known relationships in our data into the unknown.
    - Powerful, but requires more analysis steps.

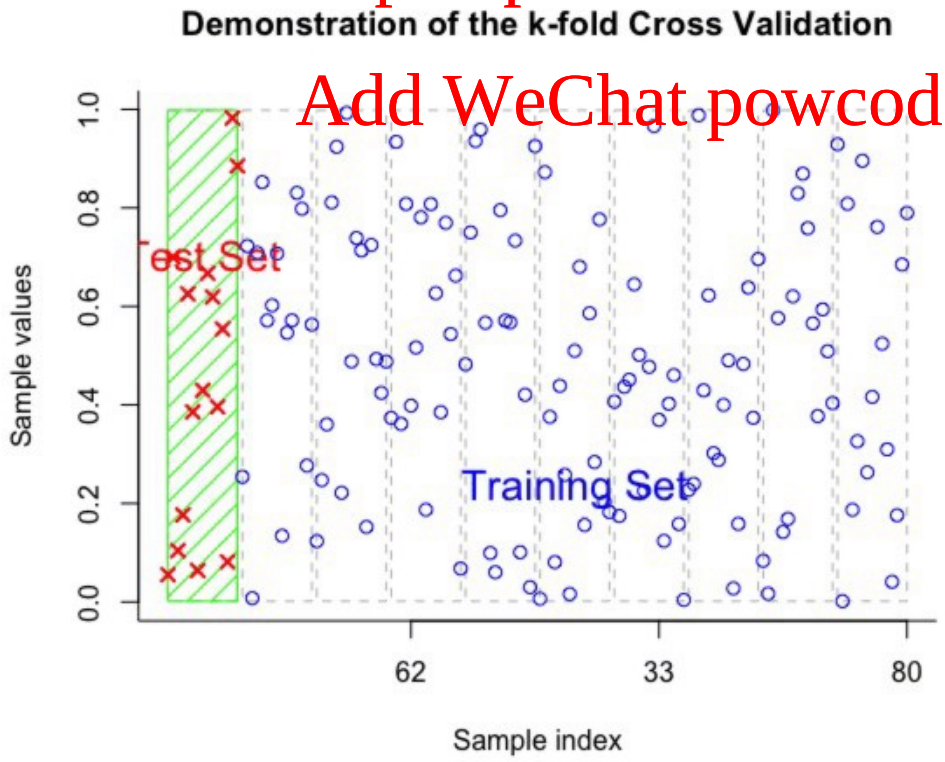
Cross-Validation

- Leave One Out (LOO)
    - Useful for smaller data samples
- Sub-setting
    - Training Data
    - Testing Data
- **Required for Predictive Models!**

Cross-Validation LOO



Cross-Validation Subsetting



Predictive Modelling

1. Split the data
    - Training Data ~80%
    - Testing Data, remaining

2. Fit the model to the training data.
3. predict() the testing data using the model.
4. Compare predicted vs. actual of testing data.
5. Repeat

Predictive Modelling Demo

- In Class Demo
    - lm(cty~displ+cyl, data = mpg)
    - dplyr::slice

Variable Selection

How do we determine how and which variables are included in the final model.

- Manual
- Step-wise
- All subsets

Manual Selection

- Requires some expert knowledge
- Typically begins by including strongest predictor
- Strategically add and remove variables

Step-wise

`MASS::stepAIC()`

- Forward selection, begin with no variables
    - Add a variable
    - Test if improves model
    - Repeat
- Backward elimination, begin with all candidate variables
    - Test loss in model by removal of each variable
    - Delete variable from model if no significant difference
- Bidirectional elimination, a combination of the above
    - Testing at each step for variables to be included or excluded.

All Subsets

- Test all combinations
- Useful for smaller sets of data

```
library(caret)
leaps<-train(y ~ .,
             data=mydata,
             method = "lm")
```

All Subsets Example I

```
library(caret)
data(swiss)
```

Swiss Fertility and Socioeconomic Indicators

- Fertility, *lg*, ‘common standardized fertility measure’
- Agriculture, % of males involved in agriculture as occupation
- Examination, % draftees receiving highest mark on army examination
- Education, % education beyond primary school for draftees.
- Catholic, % ‘catholic’ (as opposed to ‘protestant’).
- Infant.Mortality, live births who live less than 1 year.

All Subsets Example II

```
all <- train(Fertility ~ ., data = swiss, method = "lm")
all$finalModel
```

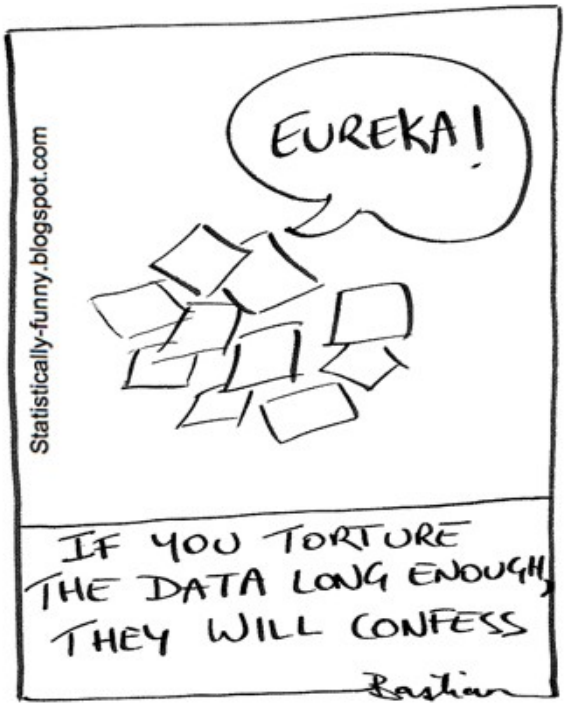
```

Call:
lm(formula = .outcome ~ ., data = dat)

Coefficients:
     (Intercept)       Agriculture       Examination         Education
         66.9152           -0.1721           -0.2580           -0.8709
        Catholic  Infant.Mortality
          0.1041            1.0770
```

All Subsets Example III


```
options(scipen = 999)
summary(all$finalModel)$coefficients[,c(1,3,4)]
```

```
                   Estimate   t value     Pr(>|t|)
(Intercept)      66.9151817  6.250229 0.0000001906051
Agriculture      -0.1721140 -2.448142 0.0187271543852
Examination      -0.2580082 -1.0162683 0.3154617231437
Education        -0.8709401 -4.7758492 0.0000243060459
Catholic          0.1041153  2.952969 0.0051900785452
Infant.Mortality  1.0770481  2.821568 0.0073357153206
```

P-hacking



Prediction Activity

- Five Dice
- Roll *n* dice and sum values.
- For *n* = 1, 2, 3, 4, 5.
- Predict the value if you were to roll 6, 10, and 20 dice?

Spatial Correlation

“everything is related to everything else, but near things are more related than distant things.”

- Waldo Tobler

Temporal Correlation

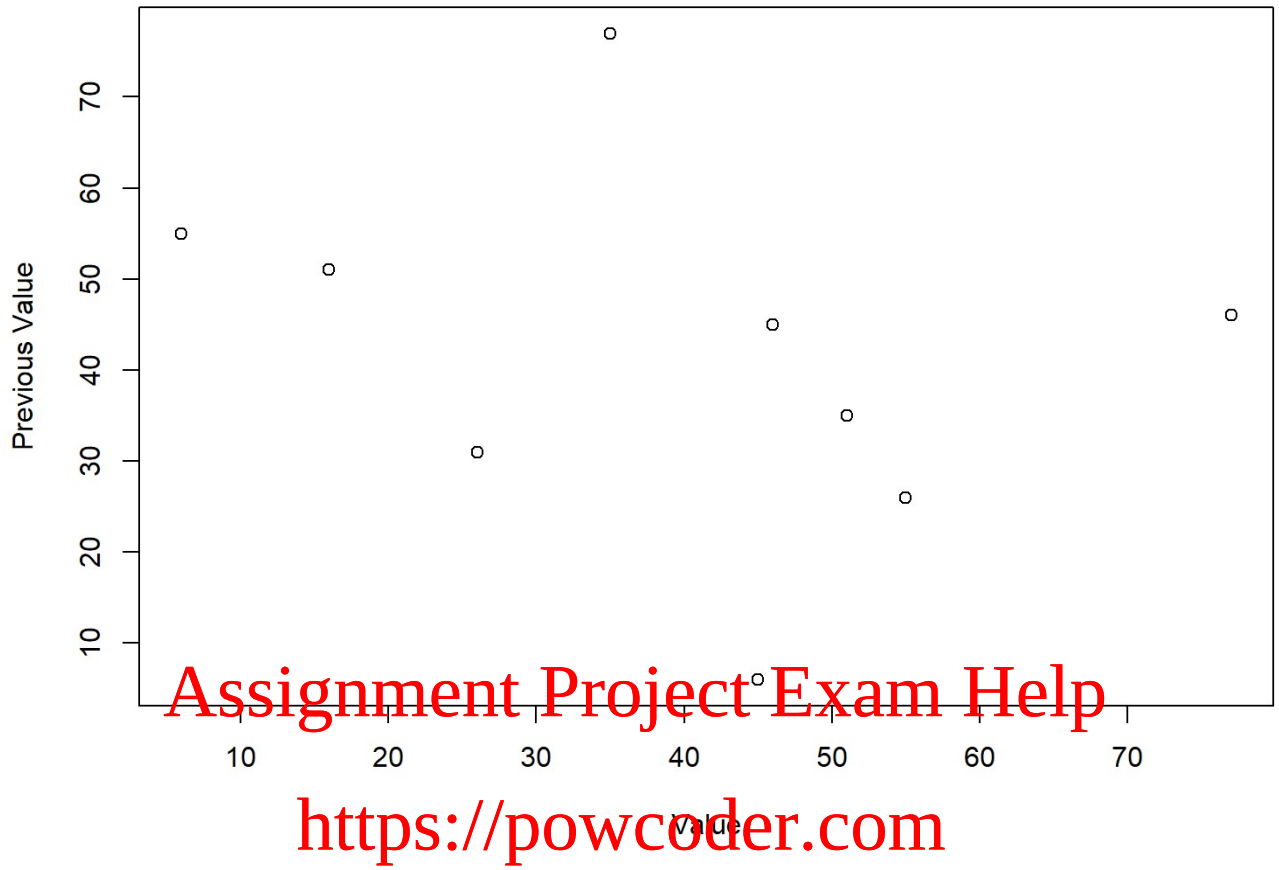
```
set.seed(100)

# Generate a random sequence of numbers
t <- sample(100, 10)

# Vector with last value removed
t_reg <- t[-length(t)]
t_reg[1:5]
```

```
[1] 31 26 55  6 45
```

```
# Vector of lags
t_lag <- t[-1]
t_lag[1:5]
```

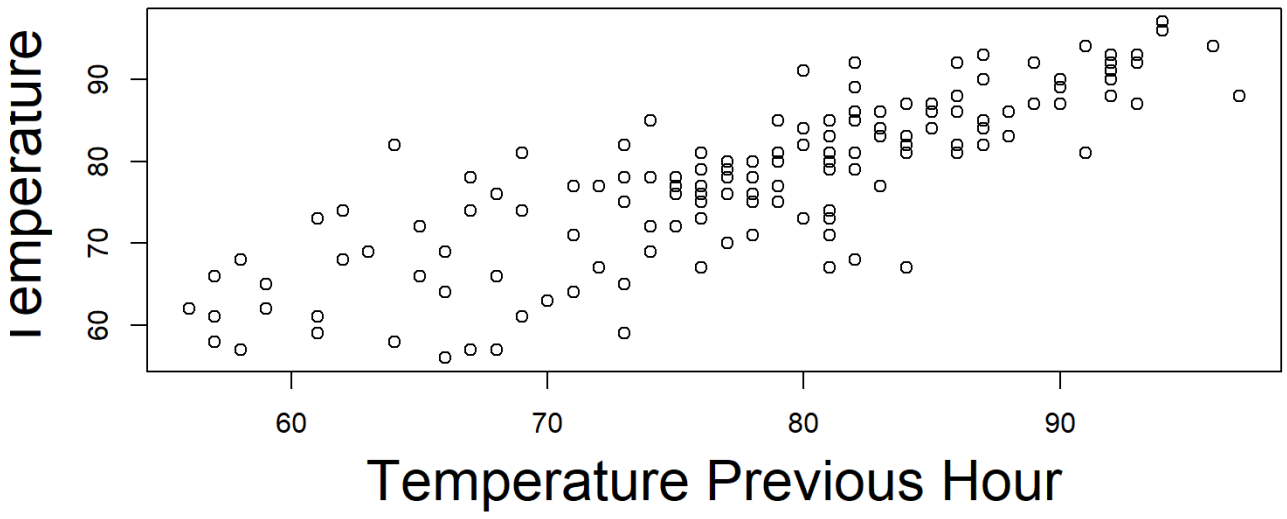
```
[1] 26 55  6 45 46
```

Random Values Test



Temperature data

```
temp <- airquality$Temp
temp_reg <- temp[-length(temp)]
temp_lag <- temp[-1]
```

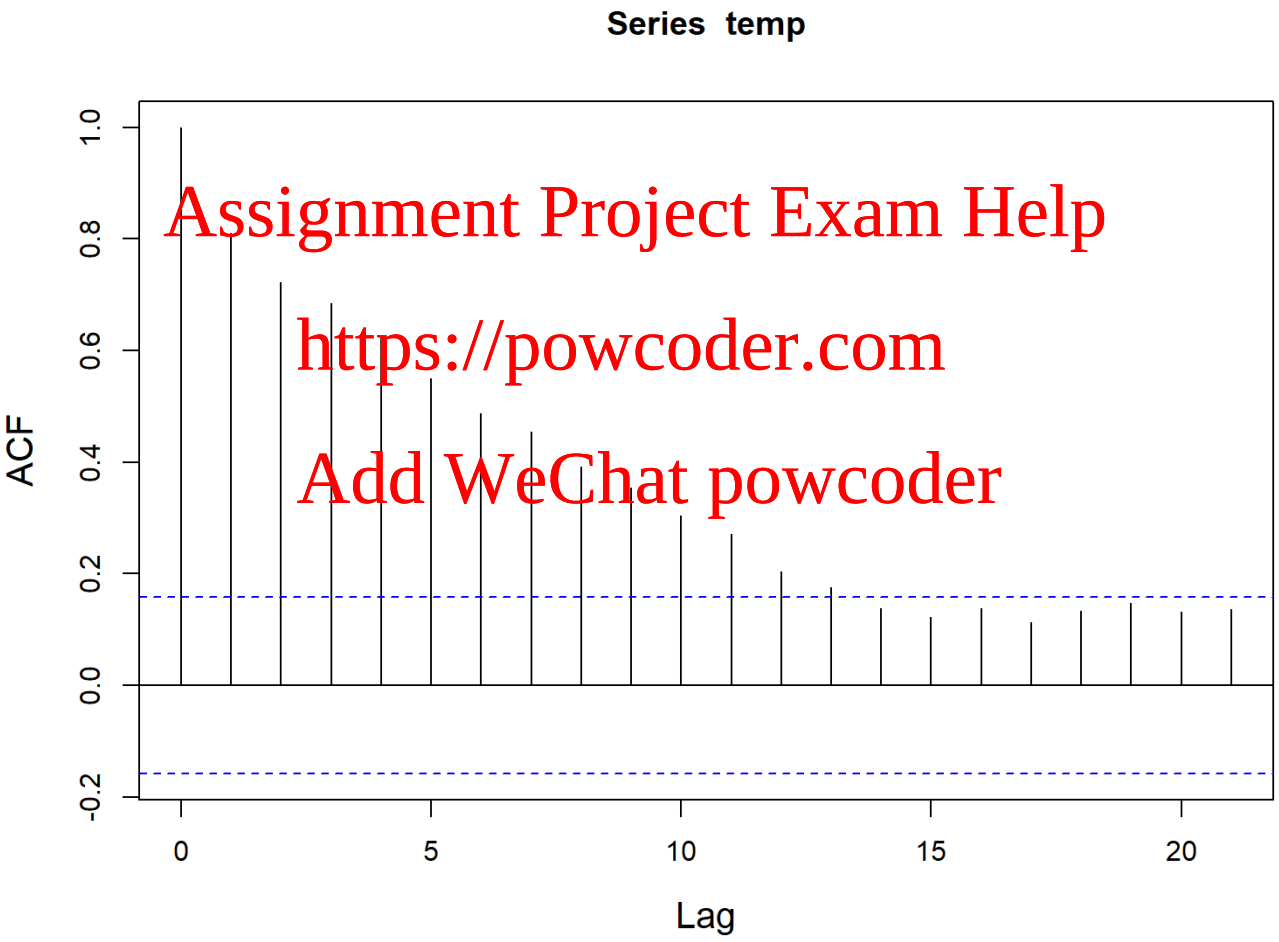


Correlation

```
cor(t_reg,t_lag)

[1] -0.2921794

cor.test(t_reg,t_lag)$p.value

[1] 0.4455116

cor(temp_reg, temp_lag)

[1] 0.8154956
```

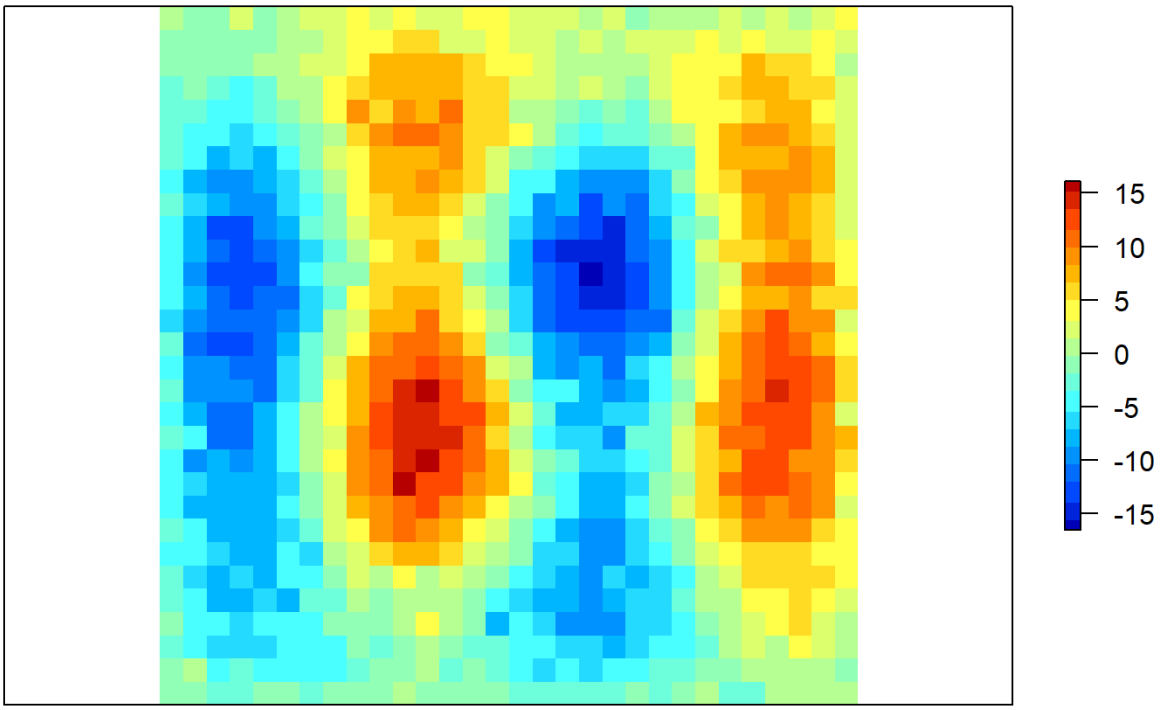
Temporal Lag Plot

```
acf(temp, cex.lab = 1.3)
```



Spatial Autocorrelation

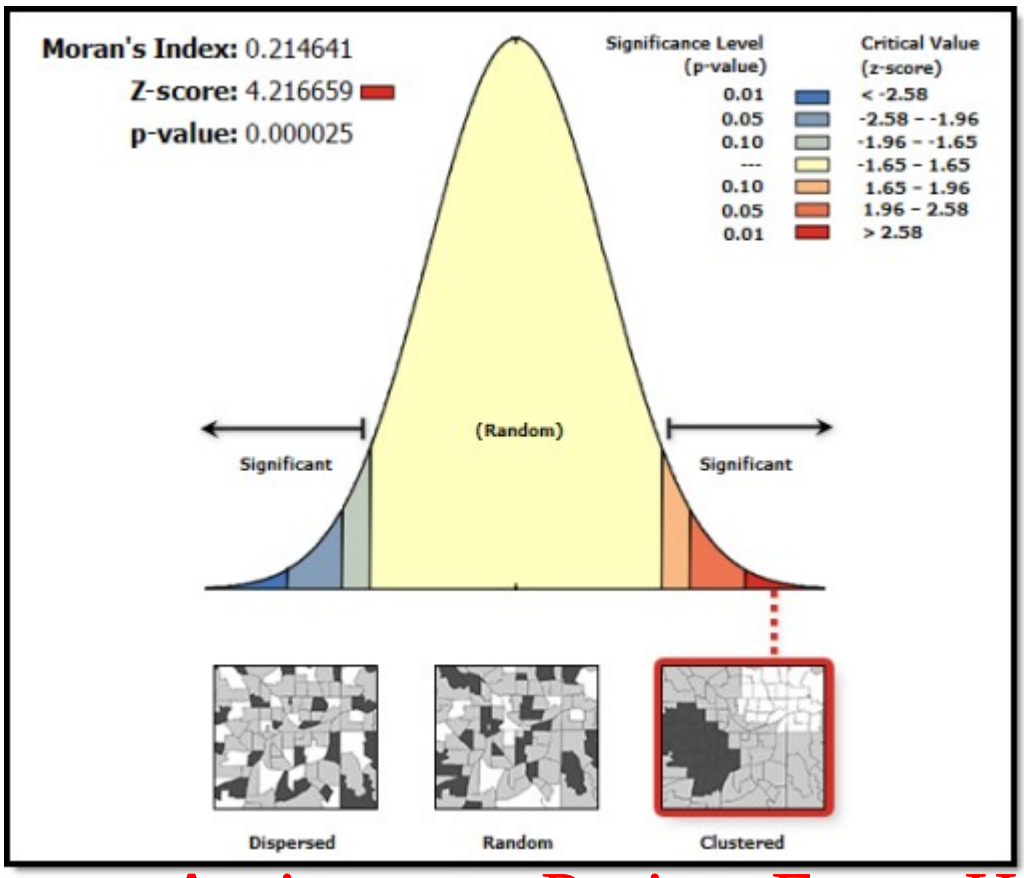
- Time is in one dimension
- Space dealing with, at least, two dimensions
    - Less clear how to measure “near”



Measure of Spatial Autocorrelation

A measure of SA describes the degree to which values are similair to other nearby objects.

- Moran's I
  - Global test statistic
    - Overall test for spatial autocorrelation

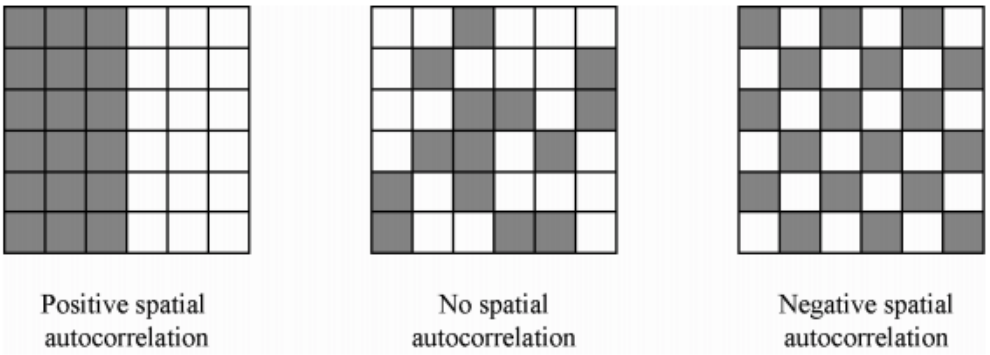


Moran's I

- Ranges from -1 to +1
- Negative 1
  - Dissimilar values are near each other
- Positive 1
  - Similar values are near each other
- Zero, no spatial autocorrelation

Moran's I & Spatial Correlation

Figure 2.1: Spatial data may demonstrate a pattern of positive spatial autocorrelation (left), negative spatial autocorrelation (right), or a pattern that is not spatially autocorrelated (center). Statistical tests, such as Moran's *I*, should always be used to evaluate the presence of spatial autocorrelation.



(Radil 2011)

Moran's I and Spatial Weights

Moran's I Formula: $$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

- Similar to correlation coefficient
- Spatial Weights Matrix \(w_{ij}\)

Spatial Weights

The measure of how "near" are objects in space.

- Points
  - Calculate a distance
- Polygons
  - Could use distance, centroid?
  - Based on contiguity

Contiguity

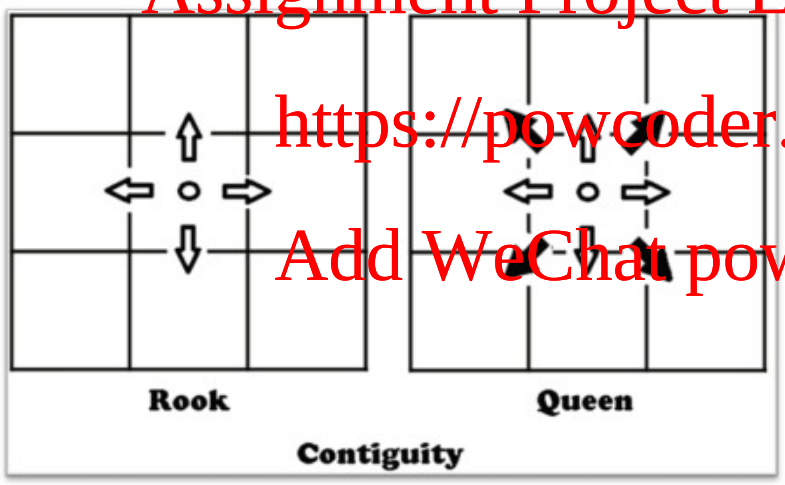
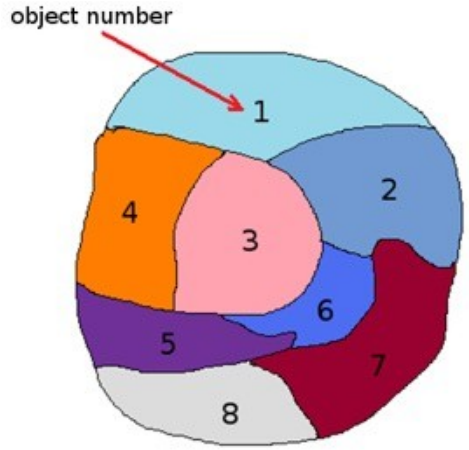


Figure 14 Rook's vs. Queen's Contiguity

(Tenney 2013)

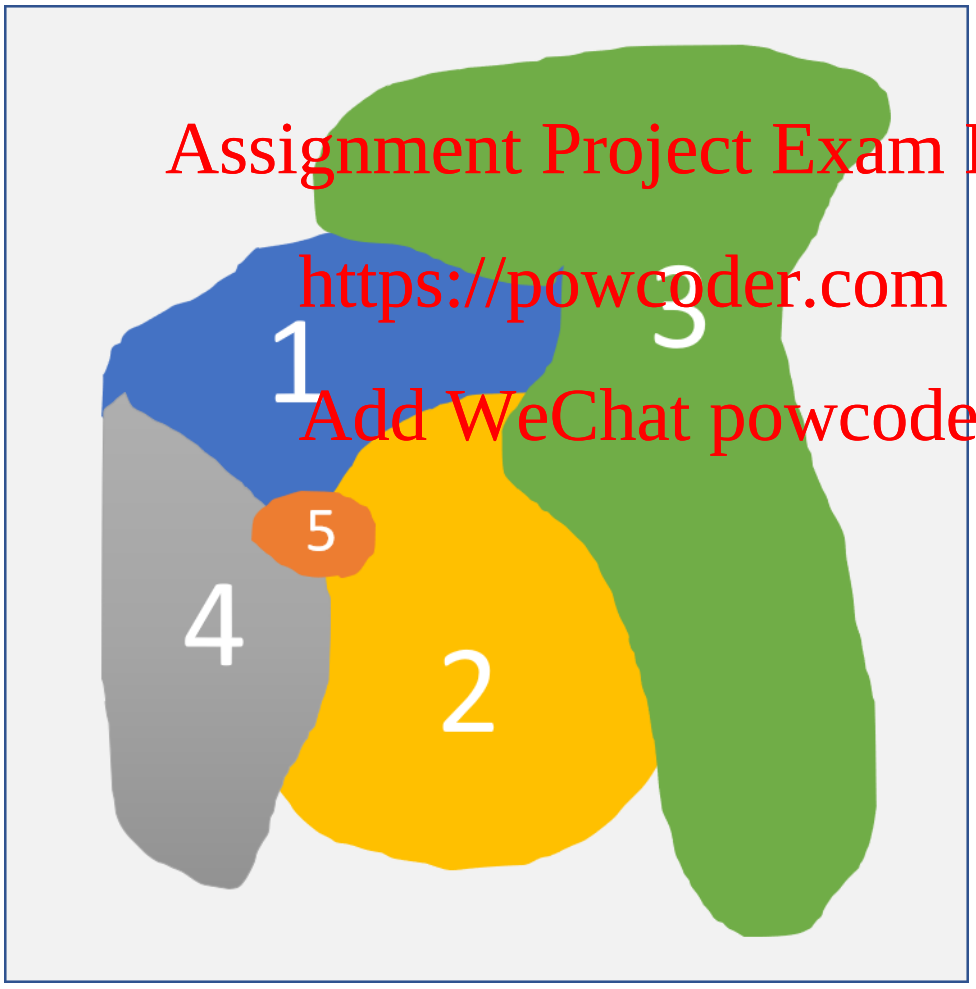
Weights Matrix (Row Standardized)



nr	1	2	3	4	5	6	7	8
1	0	1/3	1/3	1/3	0	0	0	0
2	1/4	0	1/4	0	0	1/4	1/4	0
3	1/5	1/5	0	1/5	1/5	1/5	0	0
4	1/3	0	1/3	0	1/3	0	0	0
5	0	0	1/5	1/5	0	1/5	1/5	1/5
6	0	1/5	1/5	0	1/5	0	1/5	1/5
7	0	1/4	0	0	1/4	1/4	0	1/4
8	0	0	0	0	1/3	1/3	1/3	0

Modified from https://pqstat.com/?mod_f=macwag

Spatial Weights Exercise



Calculate Moran's I in R

```
# Spatial Dependence Library
library(spdep)

# Moran's I Test - Analytical
moran.test()

# Monte Carlo Simulation
moran.mc()
```

Monte Carlo

1. Assign values to random polygons and calculate *I*
2. Repeat several time to form a distribution
3. Calculate *I* for observed data
4. Is it likely the observed is a random draw

Autocorrelation: Residuals

The linear regression model requires the residuals to be independent.

- Auto-correlation violates this assumptions

1. Temporal Autocorrelation
2. Spatial Autocorrelation

Spatial Autocorrelation

- Model residuals need to be tested with Moran's I for spatial autocorrelation.

What to do after?

- Additional Variable
- Spatial Autoregressive Models
    - Spatial Lag Model
    - Spatial Error Model

Spatial Autoregresion Models

For this course you need to be aware of these two models.

```
- Their interpretation is challenging.
- When to use either model is at times unclear.
- Models are estimated with maximum liklihood
```

Spatial Error Model

- Captures the influence of unmeasured independent variables.
    - Examines the clustering in unexplained portion of the response variable with clustering of the error terms.

Spatial Lag Model

- Implies an influence from neighbouring variables
    - Not an artifact of unmeasured variables

The value of an outcome variable in one location affects the outcome variable in neighbouring locations.

Choosing a model

- Lagrange Multiplier diagnostics for spatial dependence in linear models

```
spdep::lm.LMtests(model, # Linear Model
                  listw, # Spatial Weights
                  test = "all")

summary(lm.LMtests())
```

Lagrange Multiplier Output

```
    Lagrange multiplier diagnostics for spatial dependence
data:
model: lm(formula = CRIME ~ HOVAL + INC, data = COL.OLD)
weights: nb2listw(COL.nb)

       statistic parameter  p.value
LMerr   5.723131         1 0.016743 *
LMlag   9.363684         1 0.002213 **
RLMerr  0.079495         1 0.777983
RLMlag  3.720048         1 0.053763 .
SARMA   9.443178         2 0.008901 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- May need an underlying theory to support your ideas.

References

Radil, Steven M. 2011. “Spatializing Social Networks: Making Space for Theory in Spatial Analysis.” *Dissertation*.

Tenney, Matthew. 2013. “A conceptual model of exploration wayfinding: An integrated theoretical framework and computational methodology.” *ProQuest Dissertations and Theses*, no. April 2013: 172. http://prx.library.gatech.edu/login?url=http://search.proquest.com/docview/1353676596?accountid=11107{\%}5Cnhttp://primo-pmtna03.hosted.exlibrisgroup.com/openurl/01GALI{_}GIT/01GALI{_}GIT{_}SERVICES??url{_}ver=Z39.88-2004{\&}rft{_}val{_}fmt=info:ofi/fmt:kev:mtx:dissertatio.