# Spatial Analysis and Modeling (GIST 4302/5302)

Guofeng Cao
Department of Geosciences
Texas Tech University

# Outline of This Week

- Last week, we learned:
  - Spatial autocorrelation of areal data
    - Moran’s I, Geary’s C, Getis-Ord General G
    - Anselin’s LISA
- This week, we will learn:
  - Regression
  - Spatial regression

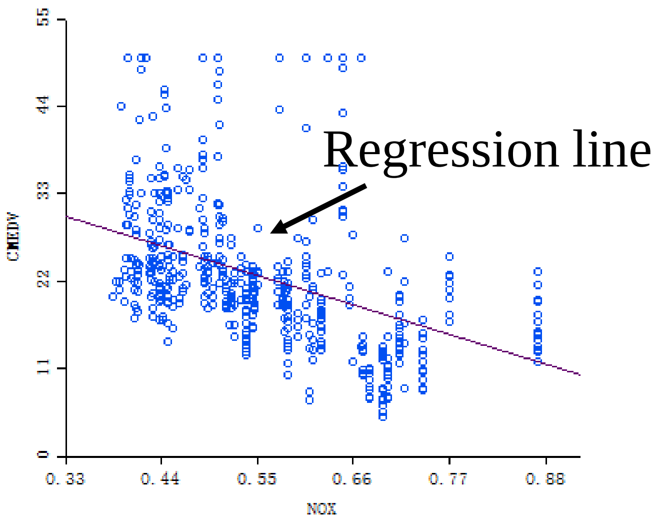
# From Correlation to Regression

## Correlation

- Co-variation
- Relationship or association
- No direction or causation is implied

## Regression

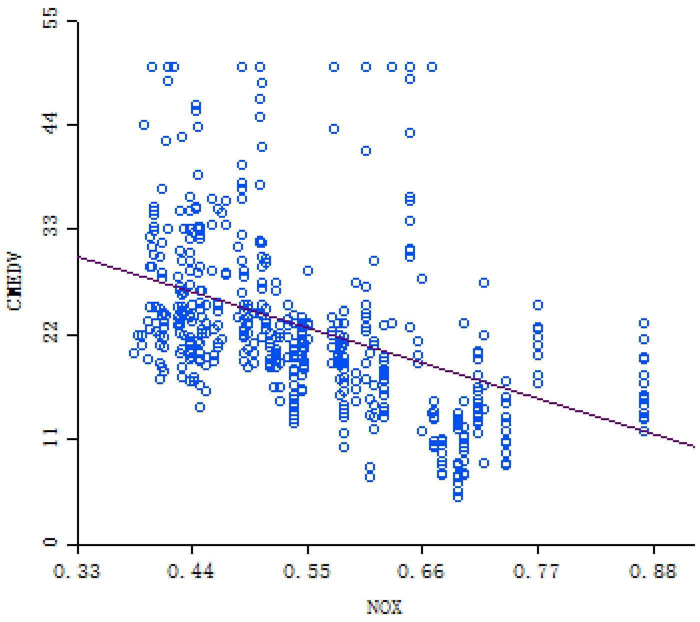
- Prediction of Y from X
- Implies, but does not prove, causation
- X (independent variable)
- Y (dependent variable)



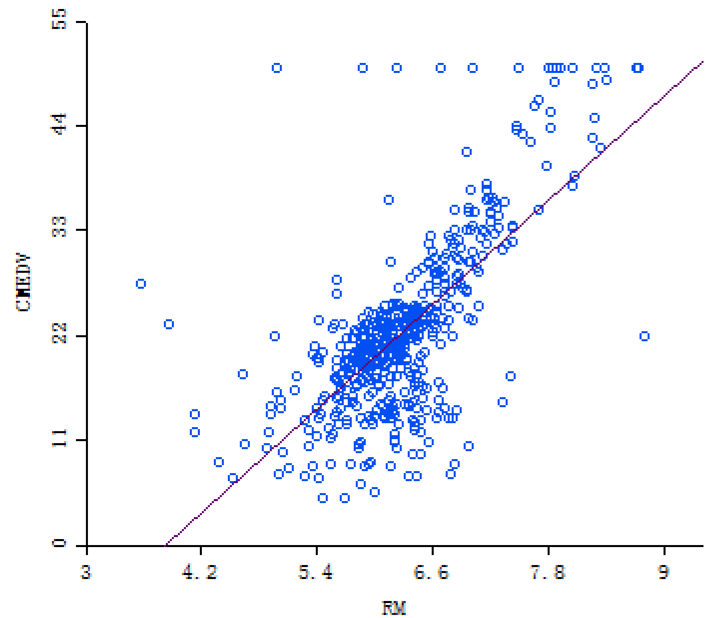
# Correlation Coefficient (r)

- One of the most common statistic
- measures the <u>strength of the relationship</u> (or "association") between <u>two</u> variables e.g. income and education
- Full name<u>: Pearson product-moment correlation coefficient</u>
- Varies on a scale from –1 thru 0 to +1

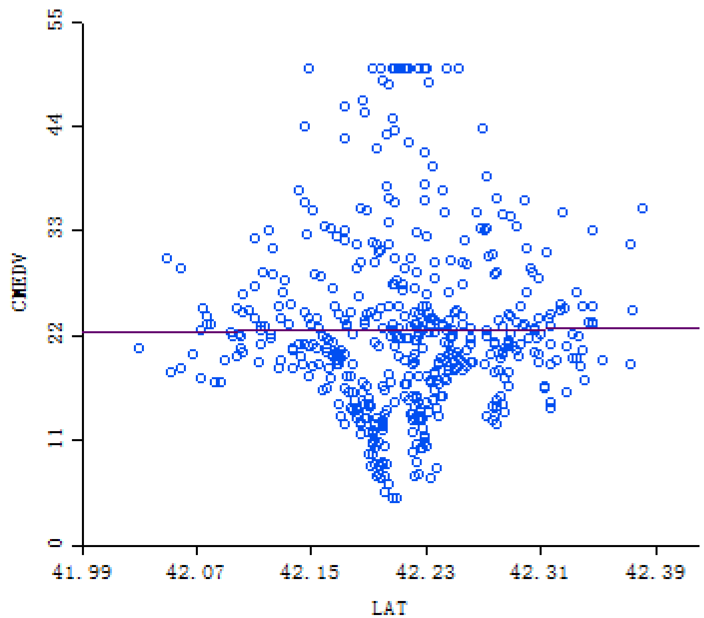
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$



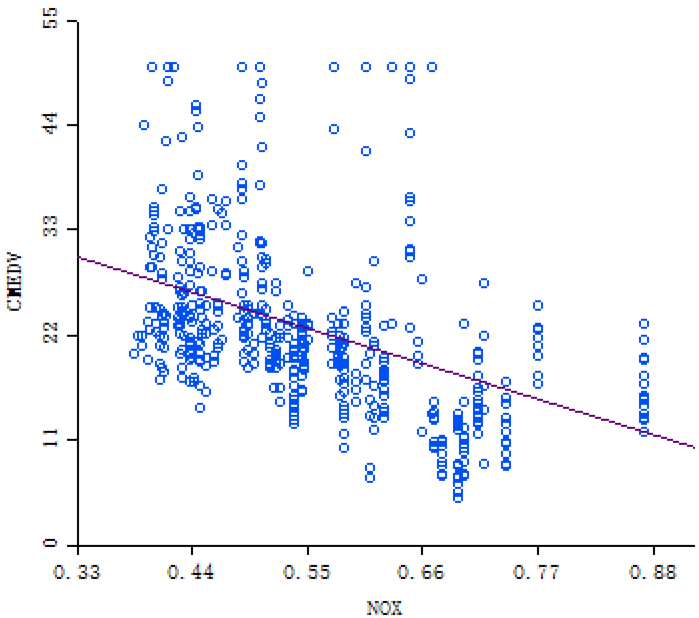
55
44
33
22
11
0
CMEDV
0.33
0.44
0.55
0.66
0.77
0.88
NOX

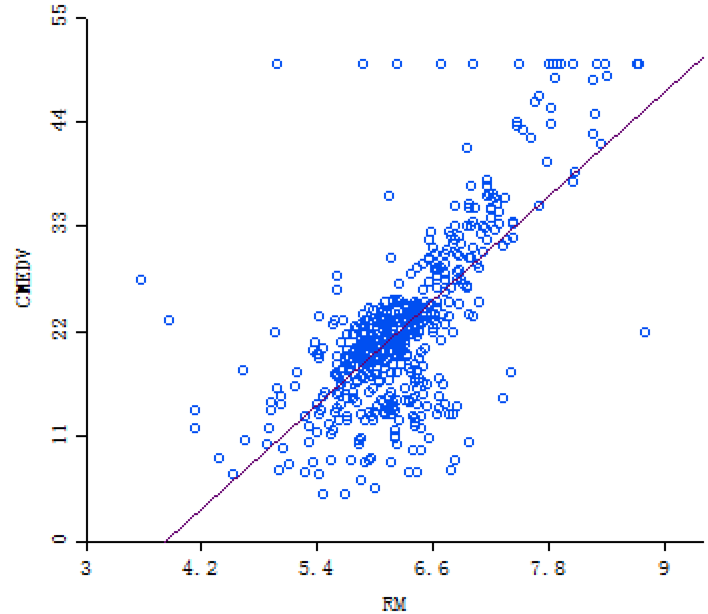



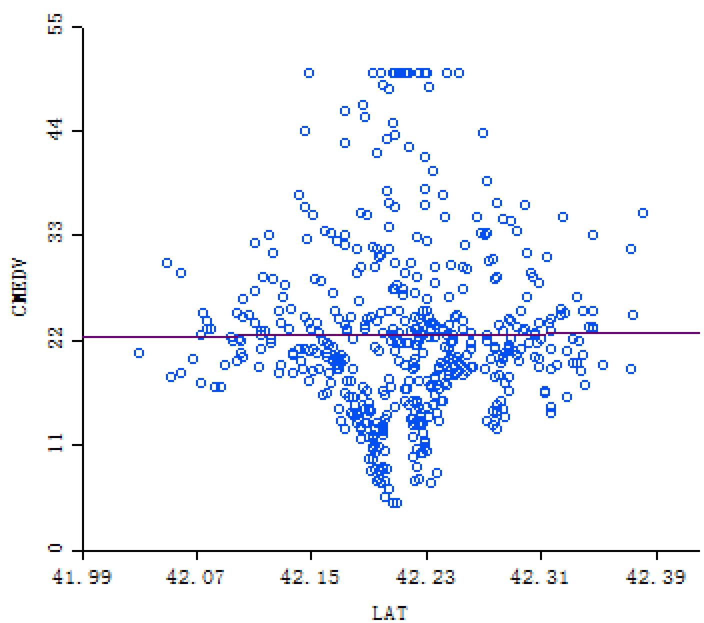
55
44
33
22
11
0
CMEDV
3
4.2
5.4
6.6
7.8
9
RM

55
44
33
22
11
0
CMEDV
41.99
42.07
42.15
42.23
42.31
42.39
LAT


CMEDV
55
44
33
22
11
0
0.33
0.44
0.55
0.66
0.77
0.88
NOX

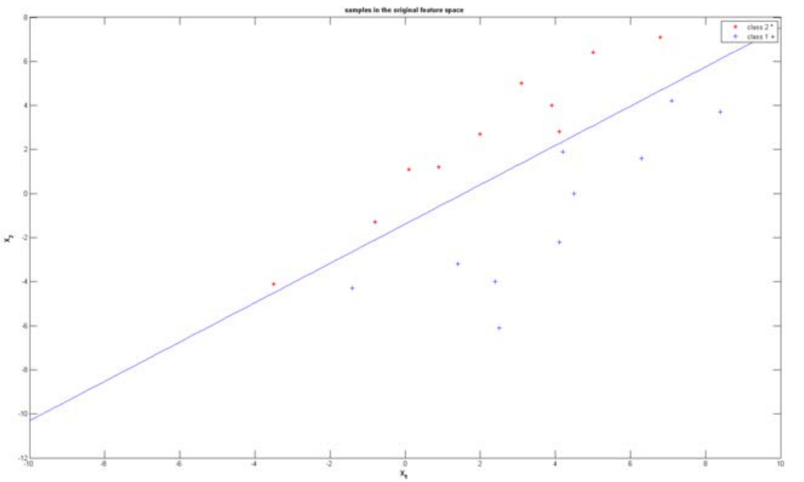

CMEDV
55
44
33
22
11
0
3
4.2
5.4
6.6
7.8
9
RM


CMEDV
55
44
33
22
11
0
41.99
42.07
42.15
42.23
42.31
42.39
LAT

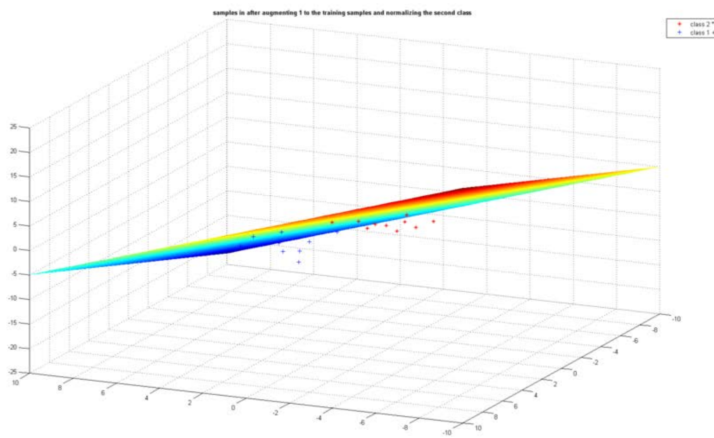
# Regression

- Simple regression
  - Between two variables
    - One dependent variable (Y)
    - One independent variable (X)

$$Y = aX + b + \varepsilon$$




- Multiple Regression
  - Between three or more variables
    - One dependent variable (Y)
    - Two or independent variable ($X_1$ ,$X_{2\ldots}$)




$$Y = b_1 X_1 + b_2 X_2 + \cdots + b_0 + \varepsilon$$

# Simple Linear Regression

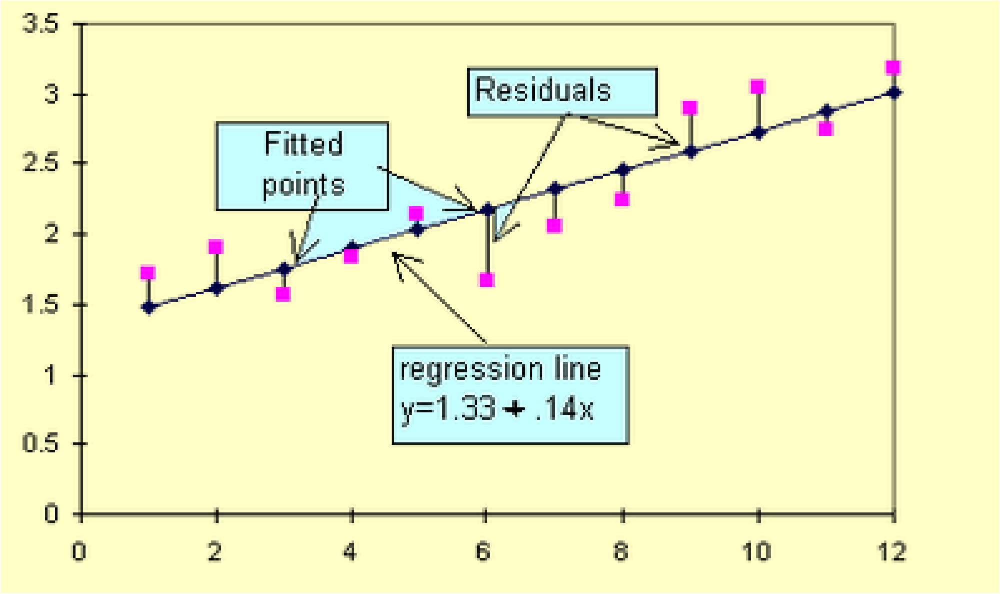
- Concerned with "predicting" one variable (Y - the dependent variable) from another variable (X - the independent variable)

$$Y = aX + b + \varepsilon$$

$$Y_i \sim \text{observations}$$

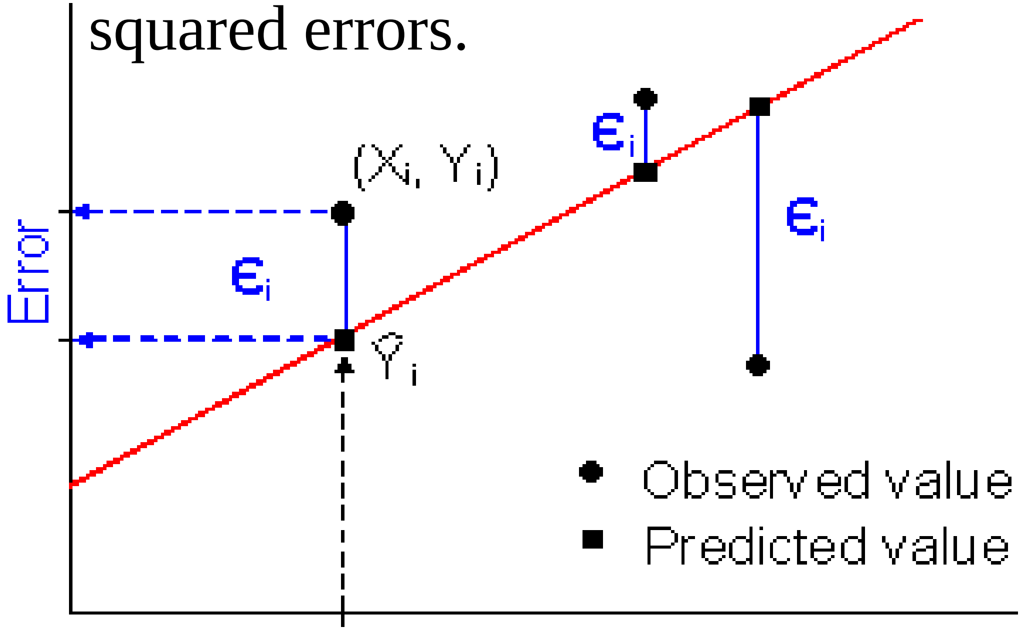
$$\hat{Y}_i \sim \textit{predictions}$$

$$\varepsilon_i = Y_i - \hat{Y}_i$$



# How to Find the line of the Best Fit

- **Ordinary Least Square** (OLS) is the mostly common used procedure
- This procedure evaluates the difference (or error) between each observed value and the corresponding value on the line of best fit.
- This procedure finds a line that minimizes the sum of the squared errors.



# Evaluating the Goodness of Fit: Coefficient of Determination ($r^2$)

- The coefficient of determination ($r^2$) measures the proportion of the variance in Y (the dependent variable) which can be predicted or "explained by" X (the independent variable). Varies from 1 to 0.
- It equals the correlation coefficient (r) squared.

$$r^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$\sum (\hat{Y}_i - \bar{Y})^2 \leftarrow$ SS Regression or Explained Sum of Squares

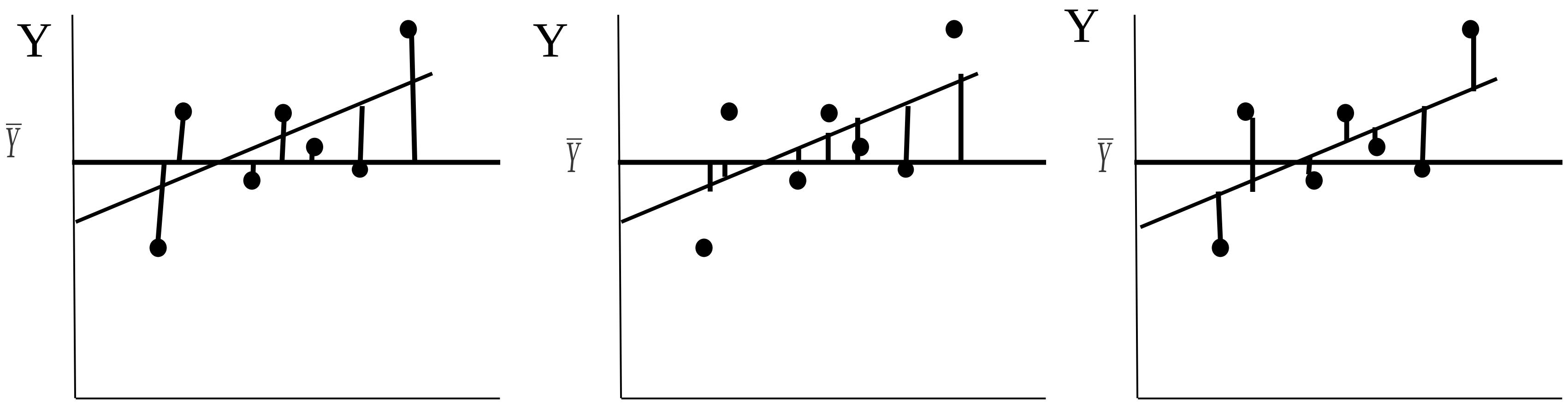
$\sum (Y_i - \bar{Y})^2 \leftarrow$ SS Total or Total Sum of Squares

Note:

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{\text{SS Total or Total Sum of Squares}} = \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{\text{SS Regression or Explained Sum of Squares}} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{\text{SS Residual or Error Sum of Squares}}$$

SS Total or Total Sum of Squares = SS Regression or Explained Sum of Squares + SS Residual or Error Sum of Squares

# Partitioning the Variance on Y




$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

SS Total
or Total Sum of
Squares

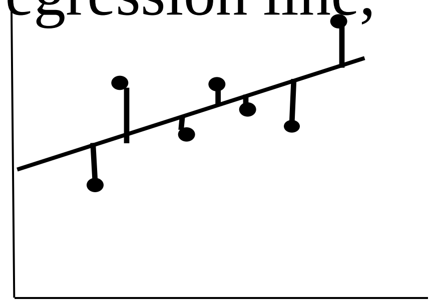
SS Regression
or Explained Sum of
Squares

SS Residual
or Error Sum of Squares

$$r^2 = \frac{\sum (\hat{Y}_i - \bar{Y}_i)^2}{\sum (Y_i - \bar{Y}_i)^2}$$

# Standard Error of the Estimate

- Measures *predictive accuracy*: the bigger the standard error, the greater the spread of the observations about the regression line, thus the predictions are less accurate



- $\sigma =$ error mean square, or average squared residual
  = variance of the estimate, variance about regression

$$\sigma = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-k}}$$

Sum of squared residuals

Number of observations minus *degrees of freedom*
(for simple regression, degrees of freedom = 2)

# Sample Statistics, Population Parameters and Statistical Significance tests

$$Y = aX + b + \varepsilon \qquad Y = \alpha + \beta X + \varepsilon$$

- a and b are *sample statistics* which are estimates of *population parameters* α and β
- β (and b) measure the change in Y for a one unit change in X. If β = 0 then X has no effect on Y
- **Significant test**
  - Test whether X has a statistically significant affect on Y
  - **Null Hypothesis** ($H_0$)**:** in the population $\beta = 0$
  - **Alternative Hypothesis** ($H_1$)**:** in the population $\beta \neq 0$

# Test Statistics in Simple Regression

- Student's t test, similar to normal, but with heavier tails

$$t = \frac{b}{\mathrm{SE(b)}} = \frac{b}{\sqrt{\dfrac{\sigma_e^2}{\sum_i (X - \overline{X})^2}}}$$

where $\sigma_{\mathrm{e}}^2$ is the variance of the estimate, with degrees of freedom = n – 2

- F-test, A test can also be conducted on the *coefficient of determination* ($r^2$) to test if it is significantly greater than zero, using the *F* frequency distribution.

$$F = \frac{\text{Regression S.S./d.f.}}{\text{Residual S.S./d.f.}} = \frac{\sum (\hat{\mathrm{Y}}_{\mathrm{i}} - \overline{\mathrm{Y}})^2 / 1}{\sum (\mathrm{Y}_{\mathrm{i}} - \hat{\mathrm{Y}}_{\mathrm{i}})^2 / n - 2}$$

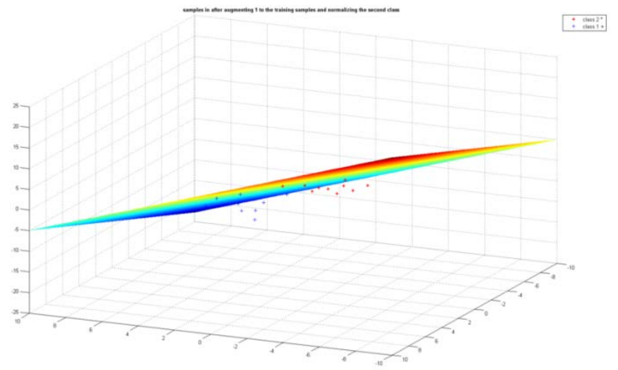
- Mathematically identical to each other

# Multiple regression

- Multiple regression: Y is predicted from 2 or more independent variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

- $\beta_0$ is the *intercept* —the value of Y when values of <u>all</u> $X_j = 0$
- $\beta_1 \ldots \beta_m$ are *<u>partial</u> regression coefficients* which give the change in Y for a one unit change in $X_j$, all other X variables held constant
- *m* is the number of independent variables



# How to Decide the Best Multiple Regression Hyperplane?

- Least square - Same as in the simple regression case

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

$$\text{or } \; Y_i = \sum_{j=0}^{m} X_{ij}\beta_j + \varepsilon_i \quad (\text{actual} \quad \mathrm{Y_i}).$$

$$\hat{Y}_i = \sum_{j=0}^{m} X_{ij} b_j \text{ predicted values for Y } (\text{regression hyperplane})$$

$$) = \text{ residuals } ¿$$

$$e_i = Y_i - \sum_{j=0}^{m} X_{ij} b_j = (Y_i - \hat{Y}_i) = (\text{Actual } Y_i \text{ - Predicted } \{ \hat{Y} \quad i$$

$$\text{Min } \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

# Evaluating the Goodness of Fit: Coefficient of Multiple Determination ($R^2$)

- Similar to simple regression, the coefficient of multiple determination ($R^2$) measures the proportion of the variance in Y (the dependent variable) which can be predicted or "explained by" all of X variables in combination.

Varies from 0 to 1.

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

SS Regression or Explained Sum of Squares

*Formulae identical to simple regression*

SS Total or Total Sum of Squares

---

As with simple regression

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

SS Total or Total Sum of Squares = SS Regression or Explained Sum of Squares + SS Residual or Error Sum of Squares

# Reduced or Adjusted $\bar{R}^2$

- $R^2$ will <u>always</u> increase each time another independent variable is included
  - an additional dimension is available for fitting the regression *hyperplane* (the multiple regression equivalent of the regression line)
- Adjusted $\bar{R}^2$ is normally used instead of $R^2$ in multiple regression

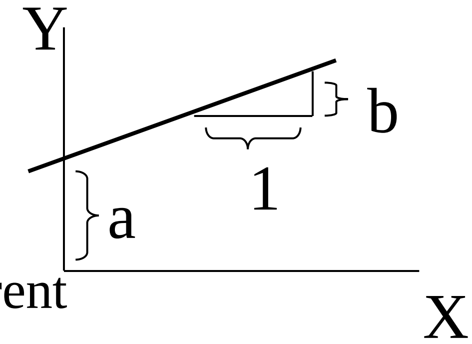
$$\bar{R}^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k}\right)$$

***k*** is the number of coefficients in the regression equation, normally equal to the number of independent variables plus 1 for the intercept.

# Interpreting *partial regression coefficients*

- The regression coefficients ($b_j$) tell us the change in Y for a <u>1 unit</u> change in $X_j$, all other X variables "held constant"
- Can we compare these $b_j$ values to tell us the relative importance of the independent variables in affecting the dependent variable?
  - If $b_1 = 2$ and $b_2 = 4$, is the affect of $X_2$ twice as big as the affect of $X_1$ ?
  - NO!
- The size of $b_j$ depends on the <u>measurement scale</u> used for each independent variable
  - if $X_1$ is income, then a 1 unit change is $1
  - but if $X_2$ is rmb or Euro(€) or even cents (₵)

    1 unit is <u>not</u> the same!
  - And if $X_2$ is *% population urban*, 1 unit is <u>very</u> different




- Regression coefficients are <u>only</u> directly comparable if the <u>units are all the same</u>: all $ for example

# Standardized partial regression coefficients

- How do we compare the relative importance of independent variables?
- We know we cannot use partial regression coefficients to directly compare independent variables unless they are all measured on the same scale
- However, we can use *standardized partial regression coefficients* (also called *beta weights*, *beta coefficients*, or *path coefficients*).
- They tell us the number of standard deviation (SD) unit changes in Y for a one SD change in X)
- They are the partial regression coefficients if we had measured every variable in *standardized form*

$$\beta_{YX_j}^{std} = b_j \left( \frac{s_{X_j}}{s_Y} \right)$$

# Test Statistics in Multiple Regression:

- Similar as in the simple regression case, but for each independent variable
- The student's test can be conducted for <u>each</u> partial regression coefficient $b_j$ to test if the associated independent variable influences the dependent variable.
  - Null Hypothesis $H_o : b_j = 0$

$$t = \frac{b_j}{\mathrm{SE}(b_j)}$$

with degrees of freedom = n – k, where ***k*** is the number of coefficients in the regression equation, normally equal to the number of independent variables plus 1 for the intercept (m+1).

- The formula for calculating the standard error (SE) of $b_j$ is more complex than for simple regression , so it is not shown here.

# Test Statistics in Mutiple Regression

## *testing the overall model*

- We test the *coefficient of multiple determination* ($R^2$ ) to see if it is significantly greater than zero, using the *F* frequency distribution.
- It is an overall test to see if at least one independent variable, or two or more in combination, affect the dependent variable.
- Does not test if each and every independent variable has an effect

$$F = \frac{\text{Regression S.S./d.f.}}{\text{Residual S.S./d.f.}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2 / k-1}{\sum (Y_i - \hat{Y}_i)^2 / n-k}$$

Again, k is the number of coefficients in the regression equation, normally equal to the number of variables (m) plus 1.

- Similar to the F test in simple regression.
  - But unlike simple regression, it is not identical to the t tests.
- It is possible (but unusual) for the F test to be significant but all t tests *not significant.*

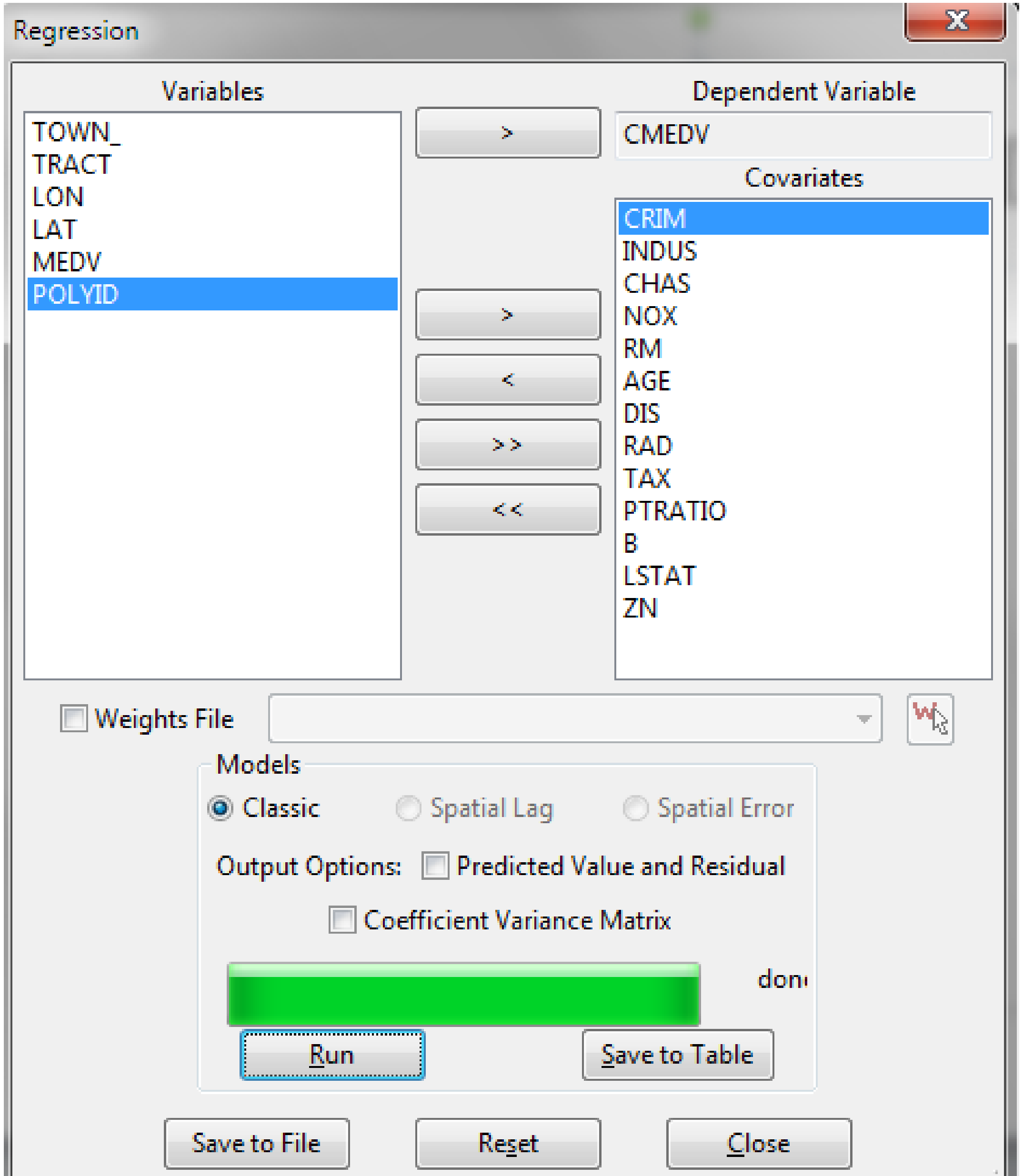
# Model/Variable Selection

- Model selection is usually an iterative process
- $R^2$ nor Adjusted $\bar{R}^2$
- P-value of coefficient
- Maximum likelihood
- *Akaike Information Criteria* (AIC)
  - the <u>smaller</u> the AIC value the <u>better</u> the model

$$AIC = 2\mathrm{k} + n\left[\ln\left(\text{Residual Sum of Squares}\right)\right]$$

***k*** is the number of coefficients in the regression equation, normally equal to the number of independent variables plus 1 for the intercept term.

# Regression in GeoDa



Regression Report

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION

Data set : boston
Dependent Variable : CMEDV
Number of Observations: 506
Mean dependent var : 22.5289
Number of Variables : 14
S.D. dependent var : 9.1731
Degrees of Freedom : 492

R-squared : 0.744464
F-statistic : 110.259
Adjusted R-squared : 0.737712
Prob(F-statistic) : 0
Sum squared residual: 10880.2
Log likelihood : -1494.23
Sigma-square : 22.1141
Akaike info criterion : 3016.45
S.E. of regression : 4.70257
Schwarz criterion : 3075.63
Sigma-square ML : 21.5023
S.E of regression ML: 4.63706

| Variable | Coefficient | Std.Error | t-Statistic | Probability |
|---|---|---|---|---|
| CONSTANT | 36.38279 | 5.057427 | 7.193933 | 0.0000000 |
| CRIM | -0.1062316 | 0.03256946 | -3.261692 | 0.0011844 |
| INDUS | 0.02330444 | 0.0609425 | 0.3824005 | 0.7023286 |
| CHAS | 2.691086 | 0.8538237 | 3.151805 | 0.0017216 |
| NOX | -17.74832 | 3.785282 | -4.688769 | 0.0000036 |
| RM | 3.788596 | 0.4141828 | 9.147159 | 0.0000000 |
| AGE | 0.00059854 | 0.01309137 | 0.04572021 | 0.9636235 |
| DIS | -1.501691 | 0.1976006 | -7.599627 | 0.0000000 |
| RAD | 0.3038247 | 0.06574986 | 4.620918 | 0.0000049 |
| TAX | -0.01270635 | 0.003726515 | -3.409715 | 0.0007038 |
| PTRATIO | -0.9242695 | 0.129599 | -7.131766 | 0.0000000 |
| B | 0.009230156 | 0.002661772 | 3.467674 | 0.0005710 |
| LSTAT | -0.53059 | 0.0502573 | -10.55747 | 0.0000000 |
| ZN | 0.04778387 | 0.01360136 | 3.513167 | 0.0004836 |

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER 87.315931
TEST ON NORMALITY OF ERRORS

| TEST | DF | VALUE | PROB |
|---|---|---|---|
| Jarque-Bera | 2 | 842.5171 | 0.0000000 |

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS

| TEST | DF | VALUE | PROB |
|---|---|---|---|
| Breusch-Pagan test | 13 | 181.575 | 0.0000000 |
| Koenker-Bassett test | 13 | 48.55038 | 0.0000053 |

SPECIFICATION ROBUST TEST

| TEST | DF | VALUE | PROB |
|---|---|---|---|
| White | 104 | N/A | N/A |

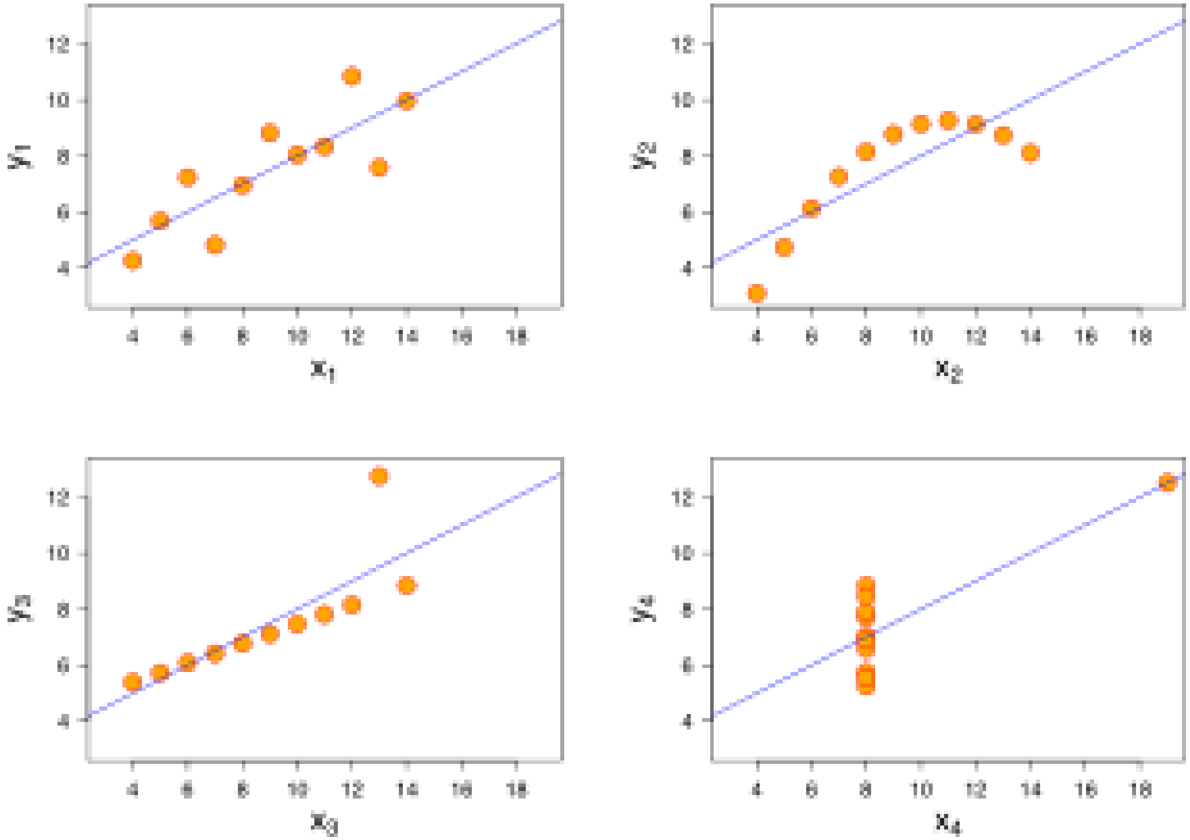
======================= END OF REPORT =============================

# Procedures for Regression

- Diagnostic
  - Outlier
  - Constant variance
  - Normality
- Transformation
  - Transforming the response
  - Transforming the predictors
- Scale Change, principal component and collinearity, and auto/cross-correlation
- Variable selection
  - Step-wise procedures
- Model fit and analysis

# Always look at your data

## *Statistics might lie*



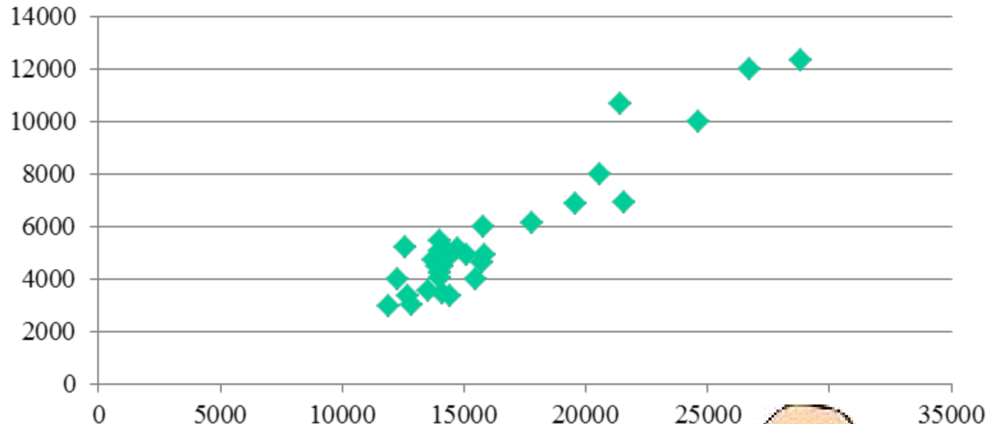

Anscombe, Francis J. (1973). "Graphs in statistical analysis". *The American Statistician* **27**: 17–21.

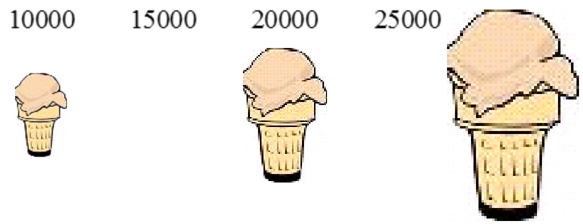
Source: Brigs UT Dallas

# Spurious relationships

## Ice Cream sales related to Drownings

Help!



- *Omitted variable* problem

  --both are related to a <u>third variable</u> not included in the analysis

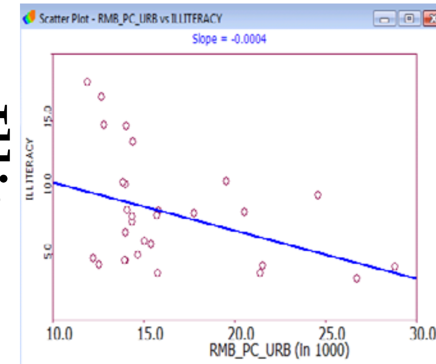


Source: Briggs UT Dallas

# Linear Regression does not prove causal effects!

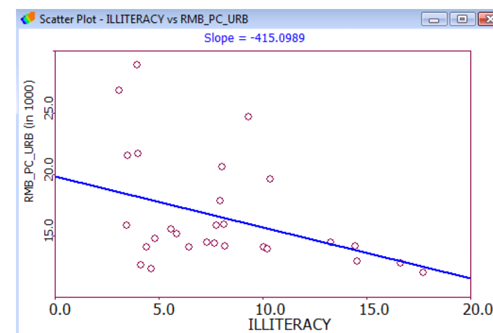
- States with higher incomes can afford to spend more on education, so illiteracy is lower
  - – Higher Income -> Less Illiteracy
- The higher the level of literacy (and thus the lower the level of <u>il</u>literacy) the more high income jobs.
  - – Less Illiteracy -> Higher Income
- Regression <u>will not</u> decide!

Illiteracy



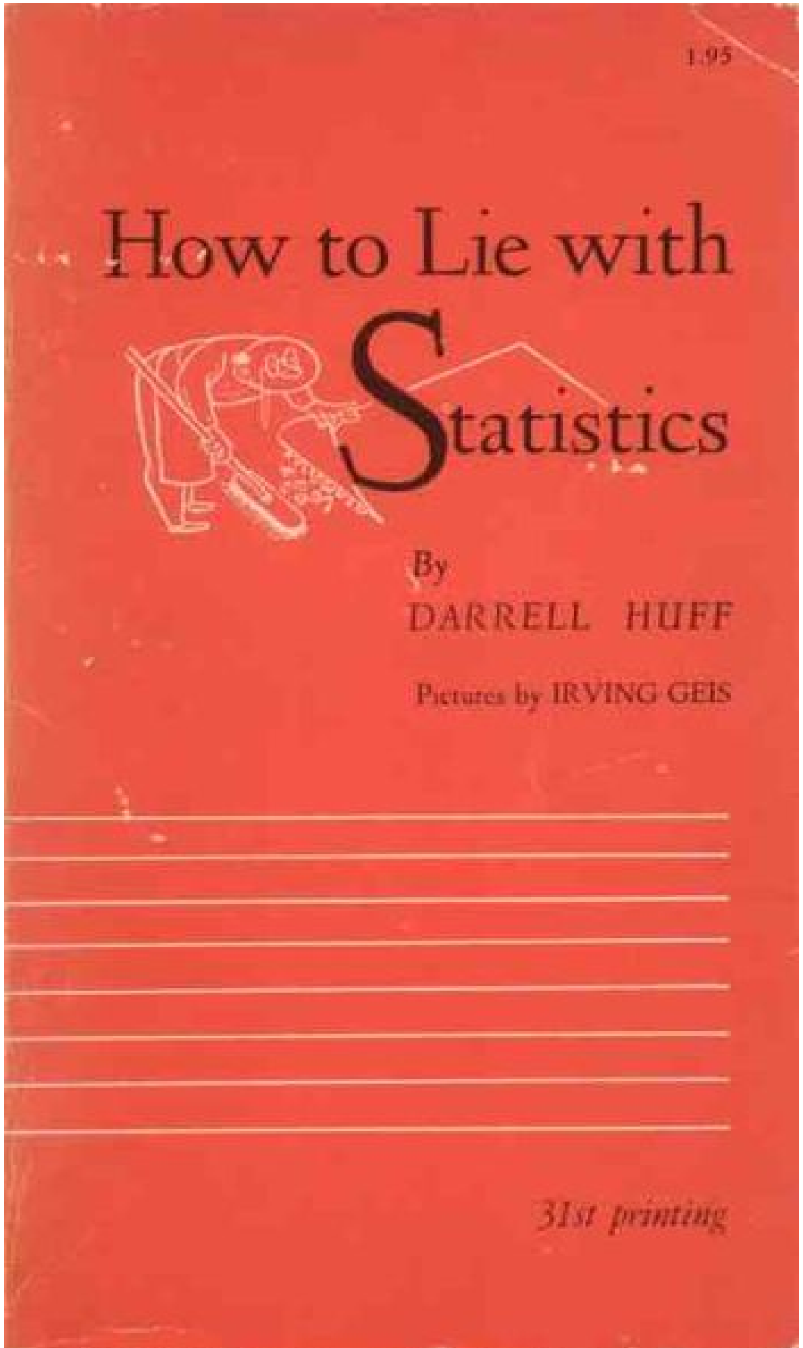

Income

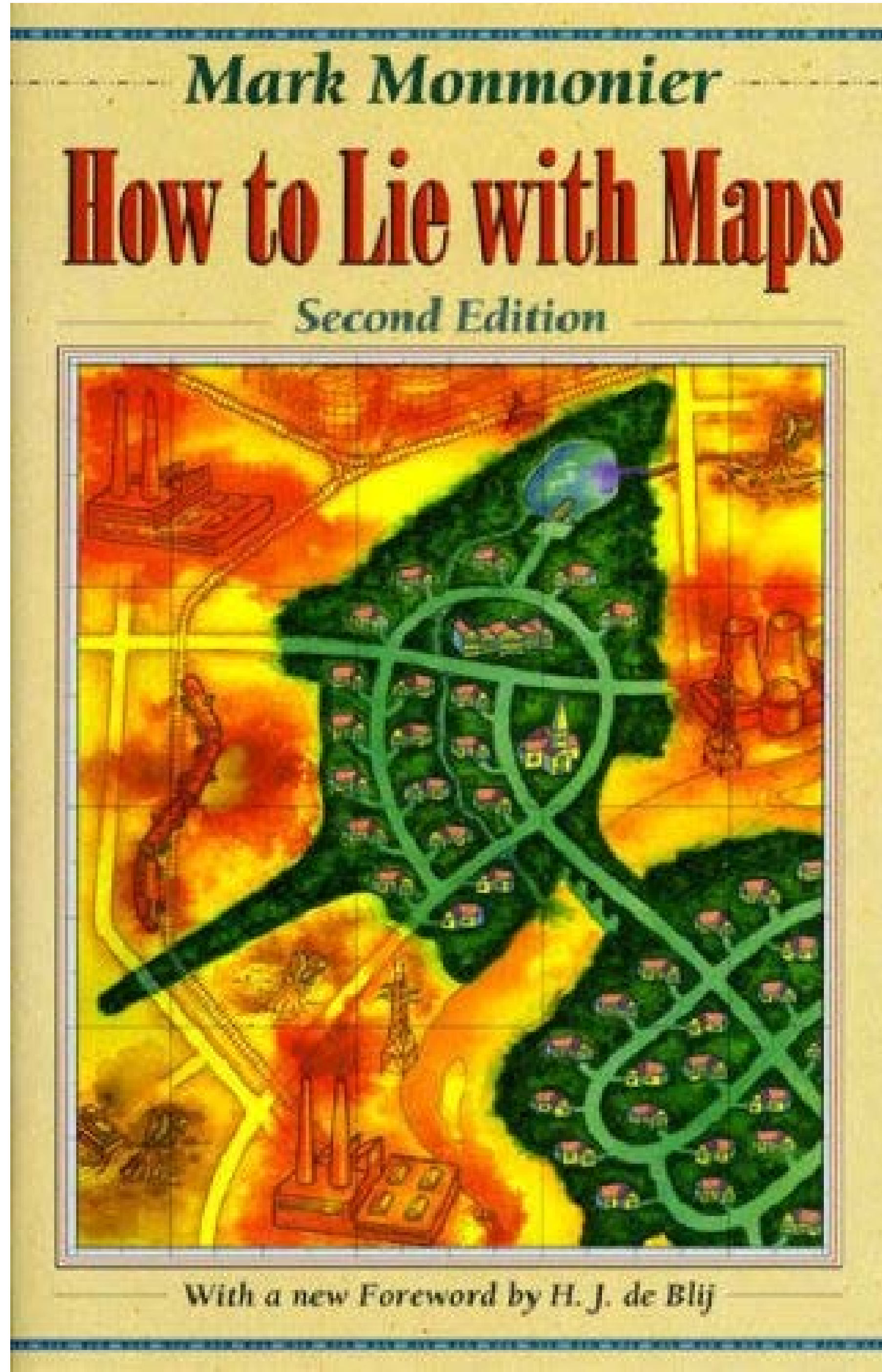
Income




Illiteracy

Source: Briggs UT Dallas


1.95
How to Lie with Statistics
By
DARRELL HUFF
Pictures by IRVING GEIS
31st printing

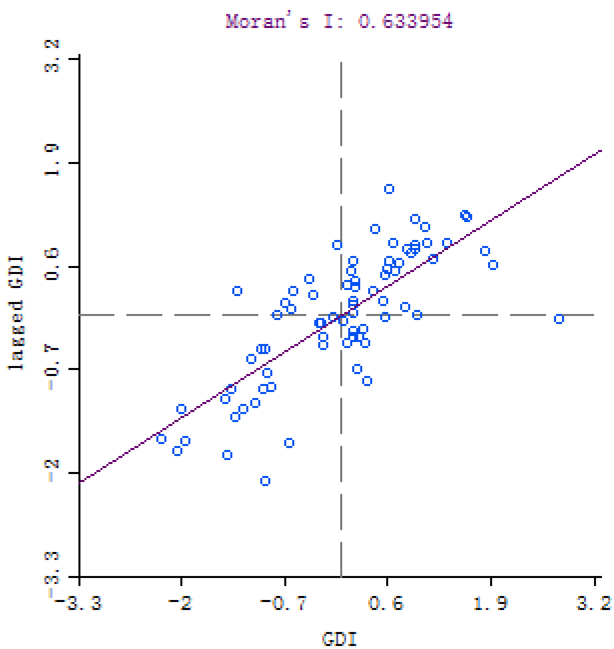

Mark Monmonier
How to Lie with Maps
Second Edition
With a new Foreword by H. J. de Blij

# Spatial Regression

# Spatial Autocorrelation vs Correlation

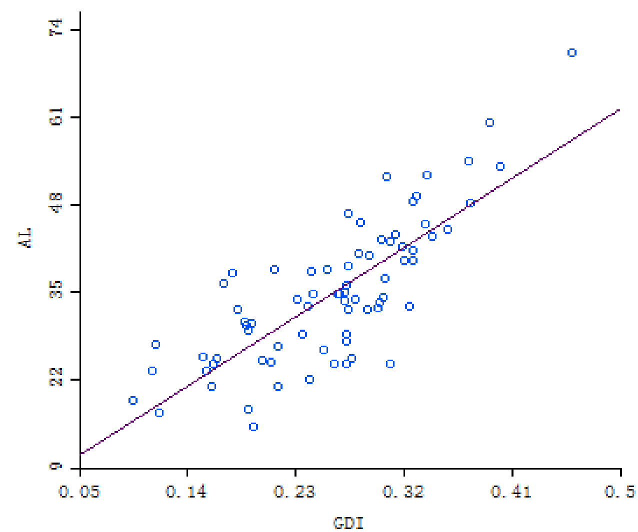
**Spatial Autocorrelation:**

shows the association or relationship between the <u>same</u> variable in "near-by" areas.




**Standard Correlation**

shows the association or relationship between two <u>different</u> variables



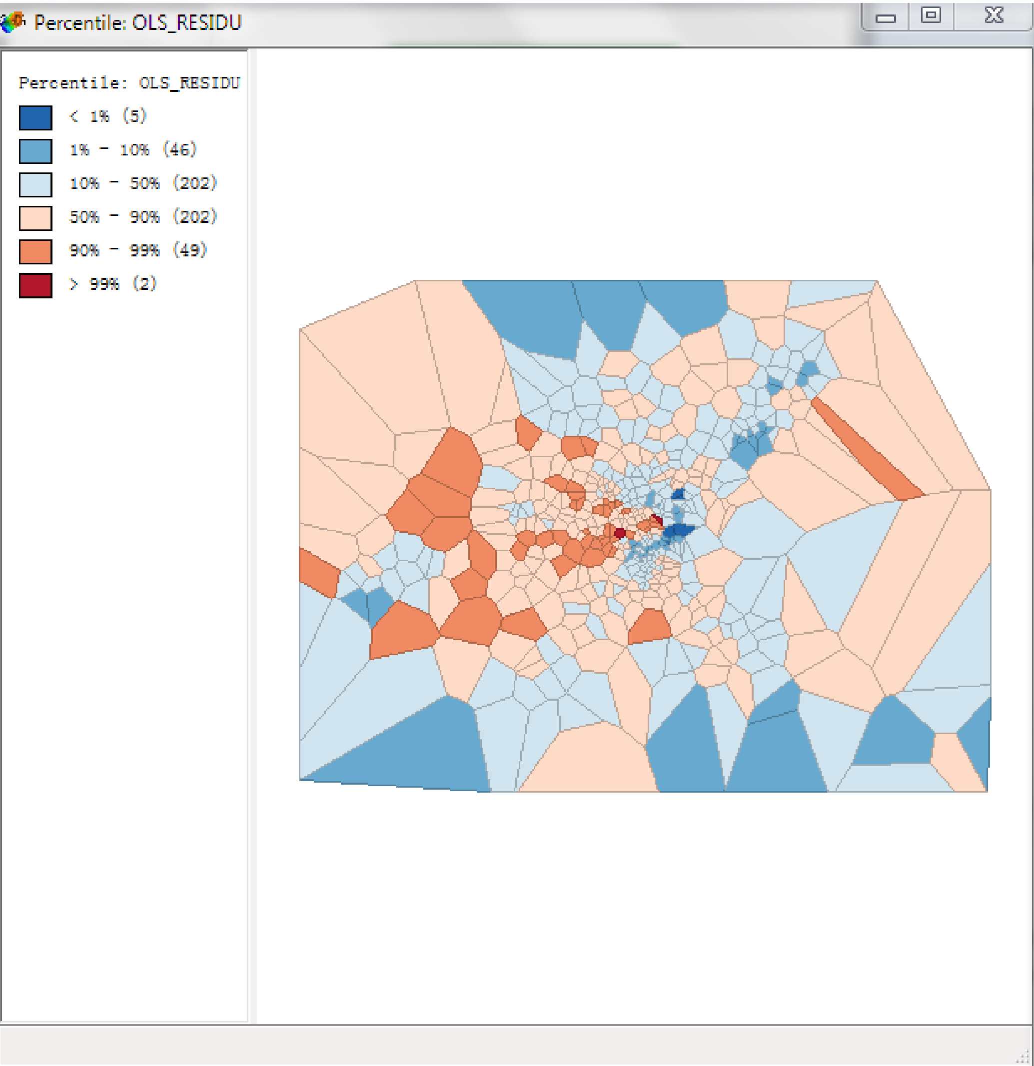
# Consequences of Ignoring Spatial Autocorrelation

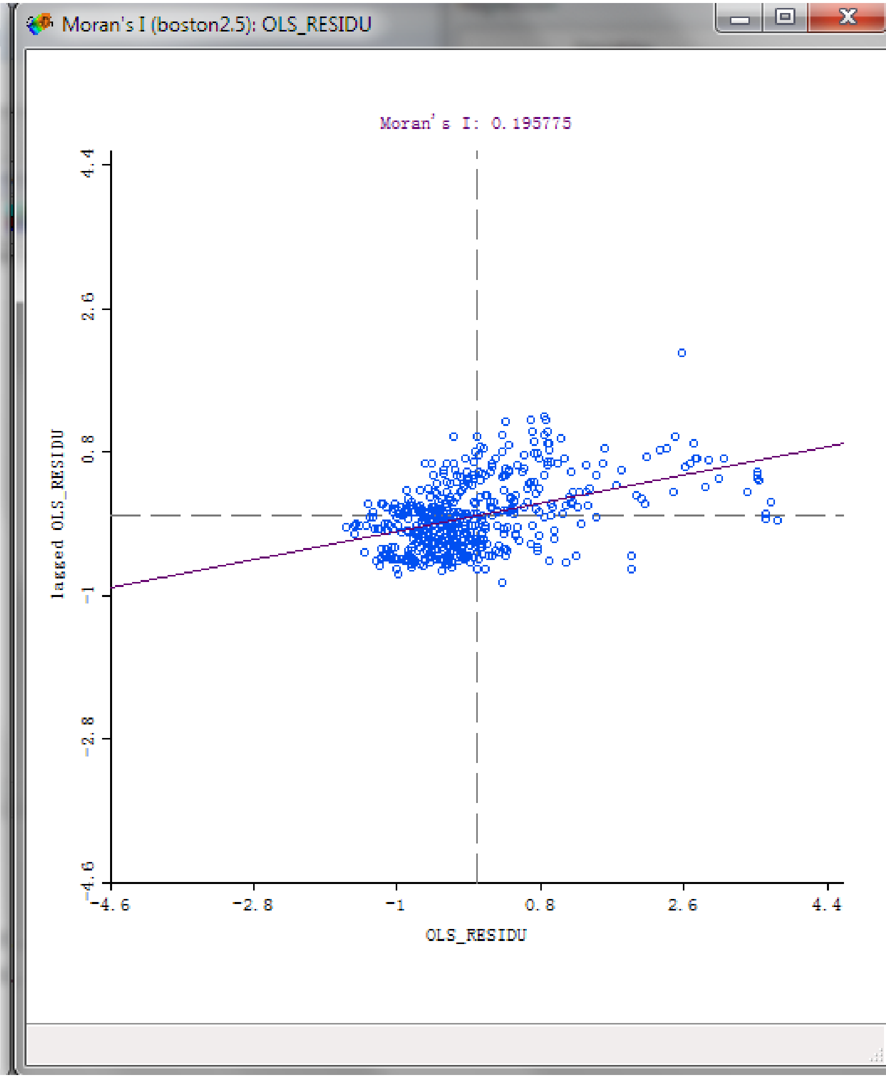
- correlation coefficients and coefficients of determination appear <u>bigger</u> than they really are
  - You think the relationship is stronger than it really is
  - the variables in nearby areas affect each other
- Standard errors appear <u>smaller</u> than they really are
  - *exaggerated precision*
  - You think your predictions are better than they really are since standard errors measure *predictive accuracy*
  - More likely to conclude relationship is *statistically significant*.

# Diagnostic of Spatial Dependence

- **For correlation**
  - calculate Moran's I for each variable and test its statistical significance
  - If Moran's I is significant, you may have a problem!
- **For regression**
  - calculate the residuals

    map the residuals: do you see any spatial patterns?
  - Calculate Moran's I for the residuals: is it statistically significant?


Percentile: OLS_RESIDU
Percentile: OLS_RESIDU
< 1% (5)
1% - 10% (46)
10% - 50% (202)
50% - 90% (202)
90% - 99% (49)
> 99% (2)


Moran's I (boston2.5): OLS_RESIDU
Moran's I: 0.195775
lagged OLS_RESIDU
OLS_RESIDU
4.4
2.6
0.8
-1
-2.8
-4.6
-4.6
-2.8
-1
0.8
2.6
4.4

# When (spatial) correlation happens

- Try to think of omitted variables and include them in a multiple regression.
  - Missing (omitted) variables may cause spatial autocorrelation
- Regression assumes all relevant variables influencing the dependent variable are included
  - If relevant variables are missing, model is *misspecified*

# Spatial Regression Methods

- Spatial Econometrics Approaches
  - Lag model
  - Error model
- Spatial Statistics Approaches
  - Simultaneous Autoregressive Models (SAR)
    - A more general case of Spatial Econometrics
  - Conditional Autoregressive Models (CAR)
- Other methods:
  - Generalized linear model with mixed effects
  - Generalized additive model
  - Generalized Estimating Equations

Source: Briggs UT Dallas

# Spatial Econometrics Approaches

- **Spatial lag model**

$$Y = \beta_0 + \lambda WY + X\beta + \varepsilon$$

values of the dependent variable in neighboring locations (*WY)* are included as an extra explanatory variable

  - these are the "spatial lag" of Y

- **Spatial error model**

$$Y = \beta_0 + X\beta + \rho W\varepsilon + \xi$$

$\xi$ is "white noise"

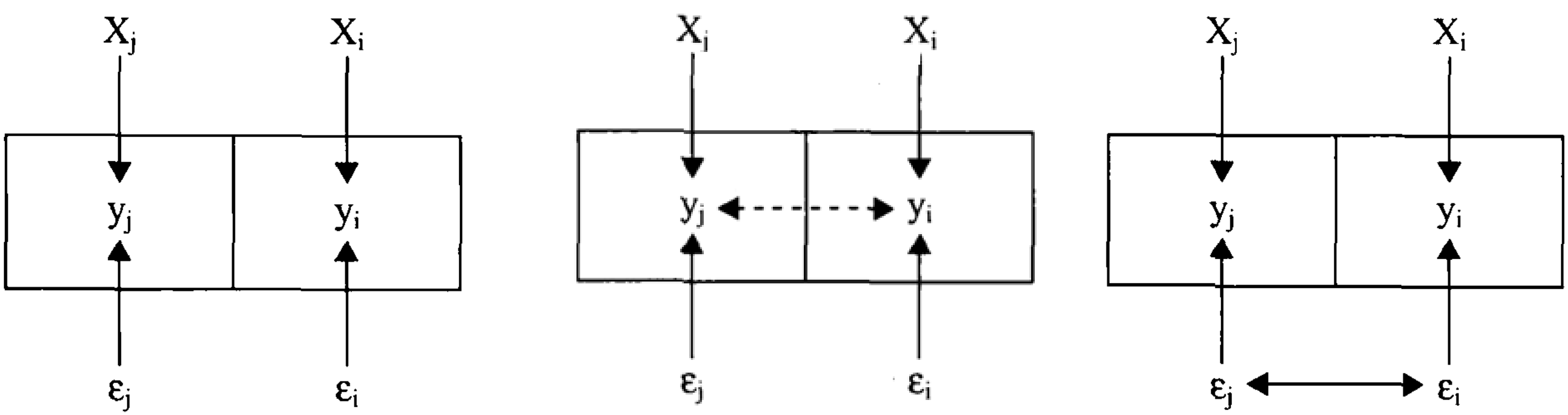
values of the residuals in neighboring locations ($W\varepsilon$) are included as an extra term in the equation;

  - these are "spatial error"

# Spatial Lag and Spatial Error Models: *conceptual comparison*

**Ordinary Least Squares**

**OLS** | **SPATIAL LAG** | **SPATIAL ERROR**




No influence from neighbors

Dependent variable influenced by neighbors

Residuals influenced by neighbors

Baller, R., L. Anselin, S. Messner, G. Deane and D. Hawkins. 2001. *Structural covariates of US County homicide rates: incorporating spatial effects*,. Criminology , 39, 561-590

Source: Briggs UT Dallas

# Spatial Lag Model

- Incorporates spatial effects by including a spatially lagged dependent variable as an additional predictor
- Outcome is dependent on the outcome for neighbors
- The ‘spatially lagged’ or ‘average neighbouring’ Wy is correlated with the unobserved error term, thus the model leads to biased and inefficient coefficients if using OLS

# Spatial Error Model

- Incorporates spatial effects through error term
- Unobserved factors in neighboring locations are correlated
- With spatial error violate the assumption that error terms are uncorrelated and coefficients are inefficient if using OLS

# Lag or Error Model: *Which to use?*

- **Lag** model primarily controls spatial autocorrelation in the <u>dependent</u> variable
- **Error** model controls spatial autocorrelation in the <u>residuals</u>, thus it controls autocorrelation in <u>both</u> the dependent <u>and</u> the independent variables
- **Conclusion:** the <u>error model</u> is more robust and generally the better choice.
- **Statistical tests** called the *LM Robust* test can also be used to select
  - Will <u>not</u> discuss these

Regression Report

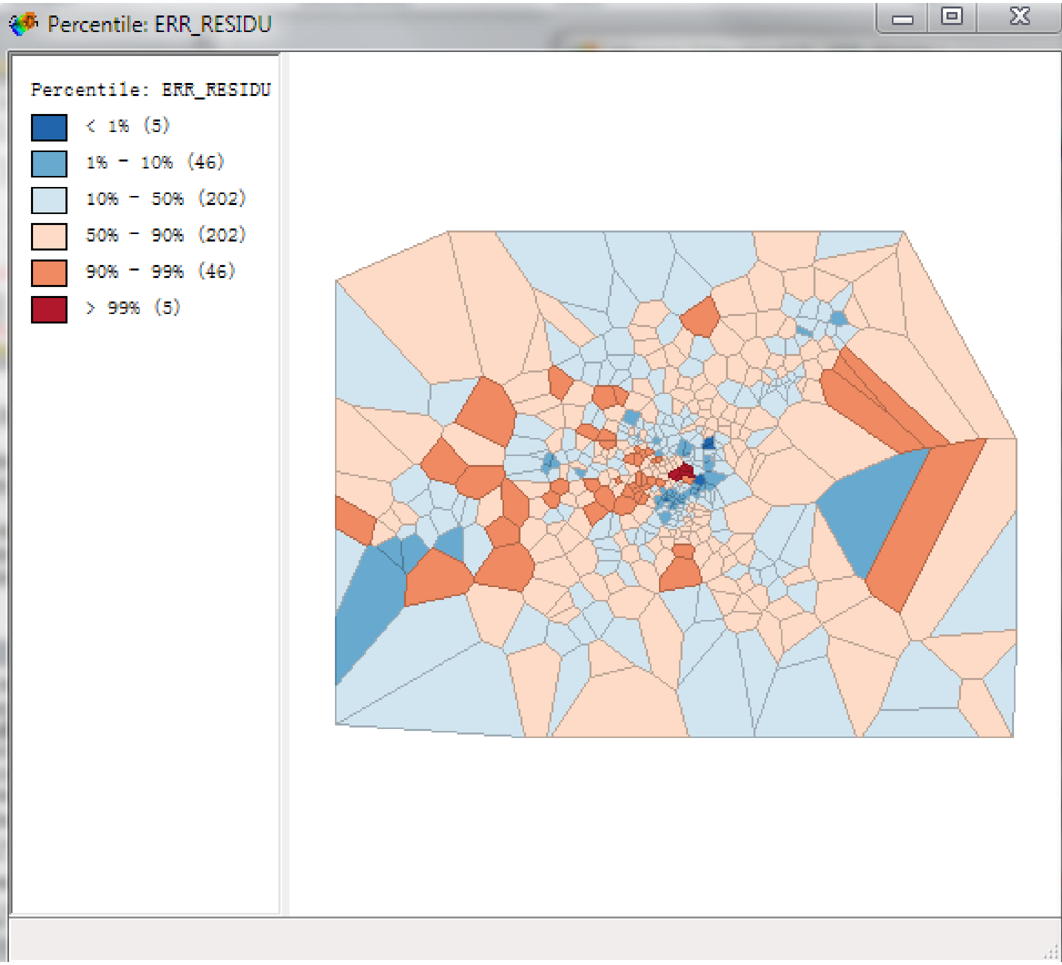
```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :   bostonpolygon
Dependent Variable  :        CMEDV  Number of Observations:  506
Mean dependent var  :      22.5289  Number of Variables   :    2
S.D. dependent var  :       9.1731  Degrees of Freedom    :  504

R-squared           :     0.184299  F-statistic           :     113.873
Adjusted R-squared  :     0.182680  Prob(F-statistic)     :4.16755e-024
Sum squared residual:      34730.7  Log likelihood        :    -1787.88
Sigma-square        :      68.9102  Akaike info criterion :     3579.76
S.E. of regression  :      8.30121  Schwarz criterion     :     3588.21
Sigma-square ML     :      68.6378
S.E of regression ML:      8.28479

-----------------------------------------------------------------------
    Variable   Coefficient      Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT      41.39839       1.806375        22.91793     0.0000000
         NOX     -34.01786       3.187837       -10.67114     0.0000000
-----------------------------------------------------------------------


REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   9.686514
TEST ON NORMALITY OF ERRORS
TEST                  DF           VALUE            PROB
Jarque-Bera            2            443.2973        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF           VALUE            PROB
Breusch-Pagan test     1            1.131862        0.2873785
Koenker-Bassett test   1           0.4377741        0.5081988
SPECIFICATION ROBUST TEST
TEST                  DF           VALUE            PROB
White                  2            6.069546        0.0480856

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : boston2.5.gwt
   (row-standardized weights)
TEST                          MI/DF       VALUE           PROB
Moran's I (error)           0.195775     15.2444755      0.0000000
Lagrange Multiplier (lag)       1       127.4022649      0.0000000
Robust LM (lag)                 1         1.7548967      0.1852623
Lagrange Multiplier (error)     1       207.8469315      0.0000000
Robust LM (error)               1        82.1995633      0.0000000
Lagrange Multiplier (SARMA)     2       209.6018282      0.0000000
```

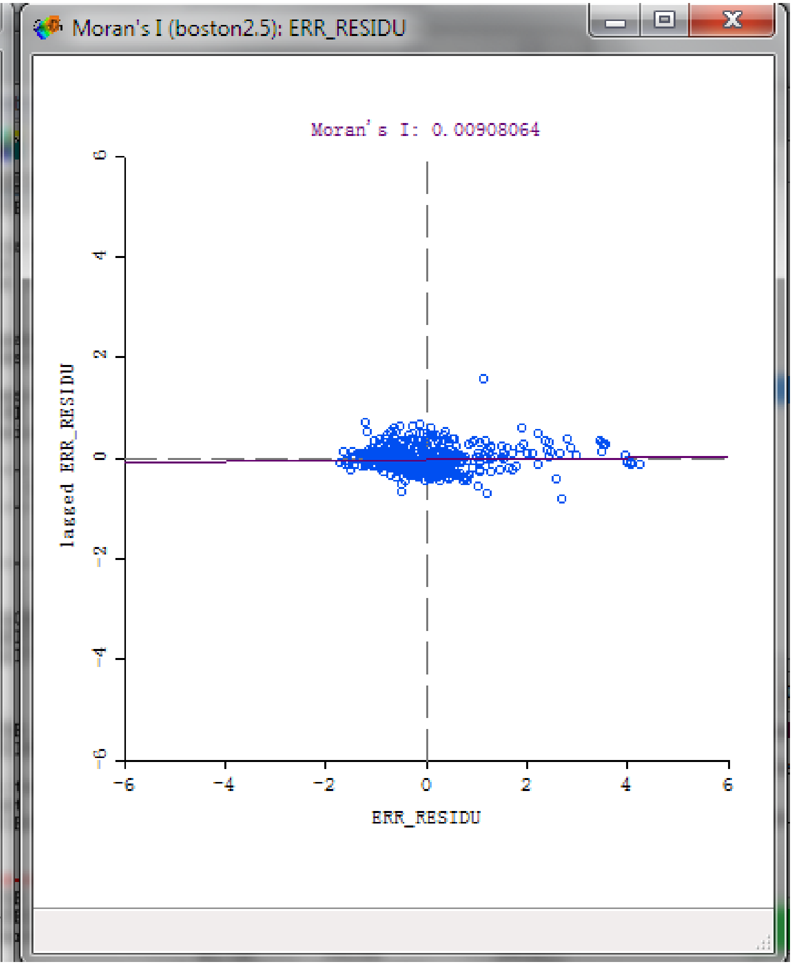
# Model Fitting

- Maximum likelihood estimation

$$\varepsilon = Y\text{-}(\beta_0 + \lambda WY + X\beta)$$

- $\varepsilon$ are assumed to be normally distributed
- Likelihood distribution of ε can be derived
- *$I\text{-}\lambda W$ must be invertible matrix (non-singular)*


Percentile: ERR_RESIDU
Percentile: ERR_RESIDU
< 1% (5)
1% - 10% (46)
10% - 50% (202)
50% - 90% (202)
90% - 99% (46)
> 99% (5)


Moran's I (boston2.5): ERR_RESIDU
Moran's I: 0.00908064
lagged ERR_RESIDU
ERR_RESIDU
-6
-4
-2
0
2
4
6

# Model/Variable Selection

- Which model best predicts the dependent variable?
- Neither $R^2$ nor Adjusted $\bar{R}^2$ can be used to compare different spatial regression models
- We use *Akaike Information Criteria* (AIC)
  - the smaller the AIC value the better the model

$$AIC = 2\mathrm{k} + n\left[\ln\left(\text{Residual Sum of Squares}\right)\right]$$

***k*** is the number of coefficients in the regression equation, normally equal to the number of independent variables plus 1 for the intercept term.

Note: can only be used to compare models with the same dependent variable

- End of this topic