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, 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)
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 (X₁, X₂, ...)

\[ Y = bₙXₙ + b₂X₂ + L + b₀ + \varepsilon \]
Simple Linear Regression

• Concerned with “predicting” one variable (Y - the dependent variable) from another variable (X - the independent variable)

\[ Y = aX + b + \varepsilon \]

\( Y_i \sim \) observations

\( \hat{Y}_i \sim \) 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.

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

Note:

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

\[ \sigma = \text{error mean square, or average squared residual} = \text{variance of the estimate, variance about regression} \]

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

- Number of observations minus *degrees of freedom* (for simple regression, degrees of freedom = 2)
Sample Statistics, Population Parameters and Statistical Significance tests

\[ Y = aX + b + \varepsilon \quad Y = \alpha + \beta X + \varepsilon \]

- \(a\) and \(b\) are sample statistics which are estimates of population parameters \(\alpha\) and \(\beta\).
- \(\beta\) (and \(b\)) measure the change in \(Y\) for a one unit change in \(X\). If \(\beta = 0\) 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

\[
t = \frac{b}{\text{SE}(b)} = \frac{b}{\sigma_e^2 \sqrt{\sum_i (X - \bar{X})^2}}
\]

where \( \sigma_e^2 \) 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.

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

- Mathematically identical to each other
Multiple regression

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

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

• $\beta_0$ is the intercept — the value of $Y$ when values of all $X_j = 0$
• $\beta_1 \ldots \beta_m$ are partial 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

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

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

\[
\hat{Y}_i = \sum_{j=0}^{m} X_{ij} \beta_j \quad \text{predicted values for } Y \quad \text{(regression hyperplane)}
\]

\[
e_i = Y_i - \sum_{j=0}^{m} X_{ij} \beta_j = (Y_i - \hat{Y}_i) = (\text{Actual } Y_i \text{ - Predicted } \hat{Y}_i)
\]

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

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

\( R^2 \) = SS Regression or Explained Sum of Squares

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

Formulae identical to simple regression

As with simple regression

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

\( \uparrow \) SS Total or Total Sum of Squares

\( \uparrow \) SS Regression or Explained Sum of Squares

\( \uparrow \) SS Residual or Error Sum of Squares
Reduced or Adjusted $\overline{R^2}$

- $R^2$ will **always** 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 $\overline{R^2}$ is normally used instead of $R^2$ in multiple regression

\[
\overline{R^2} = 1 - (1 - R^2) \left( \frac{n-1}{n-k} \right)
\]

$k$ 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 1 unit 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 measurement scale 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 not the same!
  - And if \(X_2\) is % *population urban*, 1 unit is very different

- Regression coefficients are only directly comparable if the units are all the same: 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 each partial regression coefficient $b_j$ to test if the associated independent variable influences the dependent variable.

- Null Hypothesis $H_0 : b_j = 0$

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

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 Multiple 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 smaller the AIC value the better the model

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

$k$ 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 interface in GeoDa](image)

- **Variables**: TOWN, TRACT, LON, LAT, MEDV, POLYID
- **Dependent Variable**: CMEDV
- **Covariates**: CRIM, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, ZN

- **Models**:
  - Classic
  - Spatial Lag
  - Spatial Error

- **Output Options**:
  - Predicted Value and Residual
  - Coefficient Variance Matrix

- **Buttons**:
  - Run
  - Save to Table
  - Save to File
  - Reset
  - Close
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION

Data set : boston
Dependent Variable : CMDEV 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 : 0.737124 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std.Error</th>
<th>t-Statistic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>36.38279</td>
<td>5.057427</td>
<td>7.193933</td>
<td>0.00000000</td>
</tr>
<tr>
<td>CRIM</td>
<td>-0.1062316</td>
<td>0.03256946</td>
<td>-3.261692</td>
<td>0.0011844</td>
</tr>
<tr>
<td>INDUS</td>
<td>0.02330444</td>
<td>0.0609425</td>
<td>0.3824005</td>
<td>0.7023286</td>
</tr>
<tr>
<td>CHAS</td>
<td>2.691086</td>
<td>0.8538237</td>
<td>3.151805</td>
<td>0.0017216</td>
</tr>
<tr>
<td>NOX</td>
<td>-17.74832</td>
<td>3.785282</td>
<td>-4.688769</td>
<td>0.000036</td>
</tr>
<tr>
<td>RM</td>
<td>3.788596</td>
<td>0.4141828</td>
<td>9.147195</td>
<td>0.000000</td>
</tr>
<tr>
<td>AGE</td>
<td>0.00059854</td>
<td>0.01309137</td>
<td>0.04572021</td>
<td>0.9636235</td>
</tr>
<tr>
<td>DIS</td>
<td>-1.501691</td>
<td>0.1976006</td>
<td>-7.599627</td>
<td>0.000000</td>
</tr>
<tr>
<td>RAD</td>
<td>0.3038247</td>
<td>0.06574986</td>
<td>4.620918</td>
<td>0.000049</td>
</tr>
<tr>
<td>TAX</td>
<td>-0.01270635</td>
<td>0.00372651</td>
<td>-3.409715</td>
<td>0.0007083</td>
</tr>
<tr>
<td>PTRATIO</td>
<td>-0.9242695</td>
<td>0.129599</td>
<td>-7.131766</td>
<td>0.000000</td>
</tr>
<tr>
<td>B</td>
<td>0.009230156</td>
<td>0.00265177</td>
<td>3.467674</td>
<td>0.0005710</td>
</tr>
<tr>
<td>LSTAT</td>
<td>-0.53059</td>
<td>0.0502573</td>
<td>-10.55747</td>
<td>0.000000</td>
</tr>
<tr>
<td>ZN</td>
<td>0.04778387</td>
<td>0.01360136</td>
<td>-3.513167</td>
<td>0.0004836</td>
</tr>
</tbody>
</table>

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


Source: Brigs UT Dallas
Spurious relationships

Ice Cream sales related to Drownings

- Omitted variable problem
  --both are related to a third variable 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 illiteracy), the more high income jobs.
  - Less Illiteracy -> Higher Income

- Regression will not decide!

Source: Briggs UT Dallas
Spatial Regression
Spatial Autocorrelation: shows the association or relationship between the same variable in “nearby” areas.

Standard Correlation shows the association or relationship between two different variables.
Consequences of Ignoring Spatial Autocorrelation

• correlation coefficients and coefficients of determination appear **bigger** 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 **smaller** 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?
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**

- \( X_j \)
- \( X_i \)
- \( y_j \)
- \( y_i \)
- \( \varepsilon_j \)
- \( \varepsilon_i \)

No influence from neighbors

**SPATIAL LAG**

- \( X_i \)
- \( X_i \)
- \( y_j \) (dependent variable)
- \( y_i \)
- \( \varepsilon_j \)
- \( \varepsilon_i \)

Dependent variable influenced by neighbors

**SPATIAL ERROR**

- \( X_j \)
- \( X_i \)
- \( y_j \)
- \( y_i \)
- \( \varepsilon_j \)
- \( \varepsilon_i \)

Residuals influenced by neighbors


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’ $W_y$ 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 dependent variable.
- **Error** model controls spatial autocorrelation in the residuals, thus it controls autocorrelation in both the dependent and the independent variables.

**Conclusion**: the error model is more robust and generally the better choice.

**Statistical tests** called the *LM Robust* test can also be used to select
  - Will not 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std.Error</th>
<th>t-Statistic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>41.39839</td>
<td>1.806375</td>
<td>22.91793</td>
<td>0.00000000</td>
</tr>
<tr>
<td>NOX</td>
<td>-34.01786</td>
<td>3.187837</td>
<td>-10.67114</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

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 HETROSKEDEASTICITY
RANDOM COEFFICIENTS
TEST          DF    VALUE   PROB
Breusch-Pagan test 1   1.131862 0.287385
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.gvt
(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
  \[ \varepsilon = Y - (\beta_0 + \lambda WY + X\beta) \]

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

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

$k$ 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*

• Occam's Razor
• End of this topic