Lesson 9: Geographical Weighted Regression Concepts and Methods

Dr. Kam Tin Seong
Assoc. Professor of Information Systems

School of Information Systems,
Singapore Management University

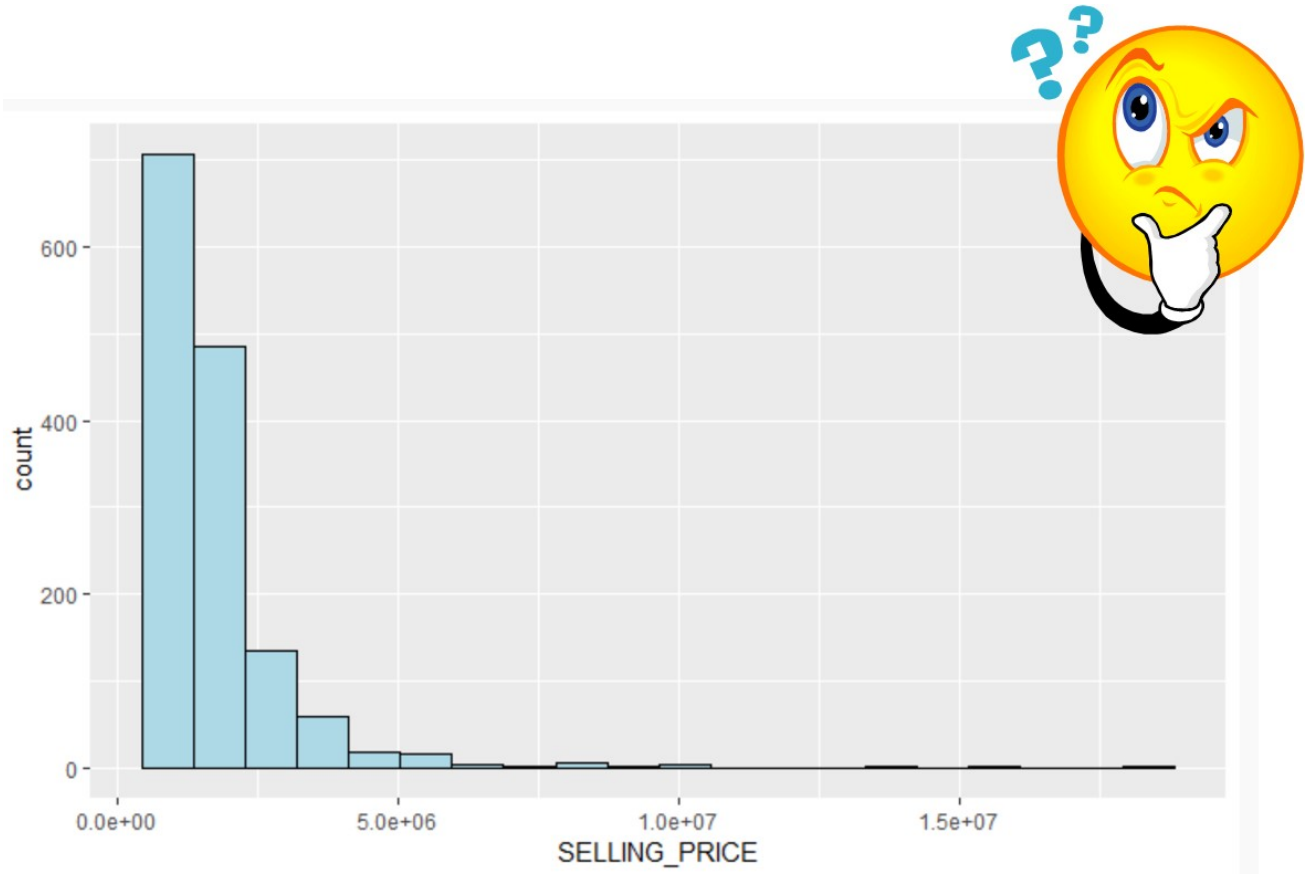
2020-5-16 (updated: 2021-10-12)

Content

- Introducing Regression Modelling
  - Simple Linear Regression
  - Multiple Linear Regression
- What is Spatial Non-stationary
- Introducing Geographically Weighted Regression
  - Weighting functions (kernel)
  - Weighting schemes
  - Bandwidth
- Interpreting and Visualising

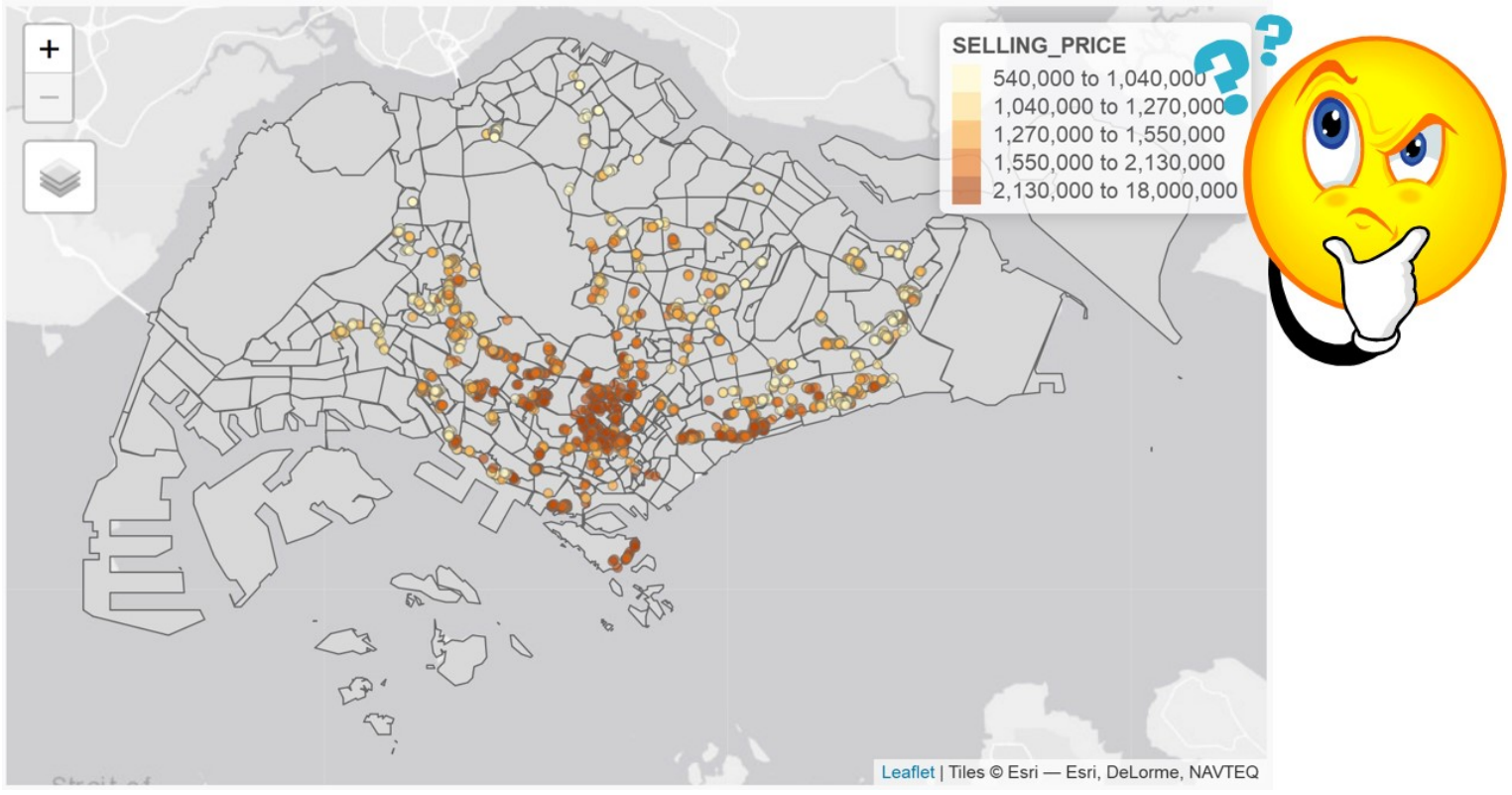
The WHY Questions

- Why some condominium units were transacted at relatively higher prices than others?



The WHY Questions

Why condominium units located at the central part of Singapore were transacted at relatively higher prices than others?



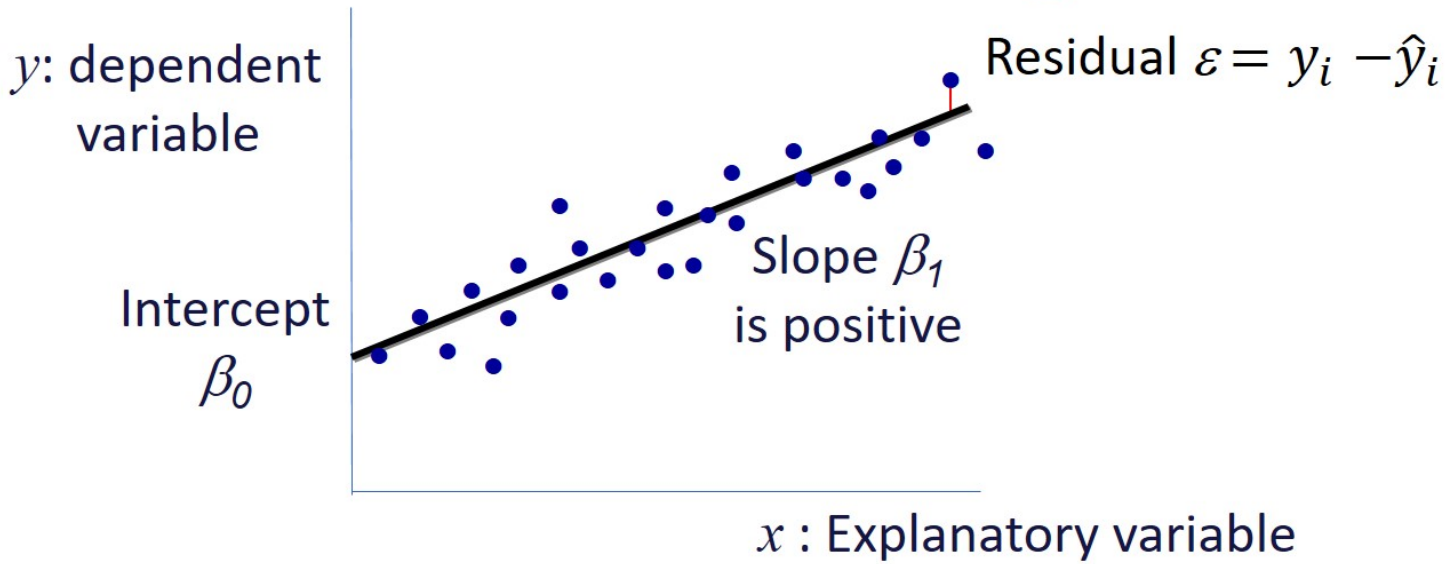
What is regression analysis?

- A set of statistical processes for explaining the relationships among variables.
- The focus is on the relationship between a dependent variable (y) and one or more independent variables (x)
  - Does X affect Y? If so, how?
  - What is the change in Y given a one unit change in X?
- Estimate outcomes based on the relationships modelled.

A Simple Linear Regression Model

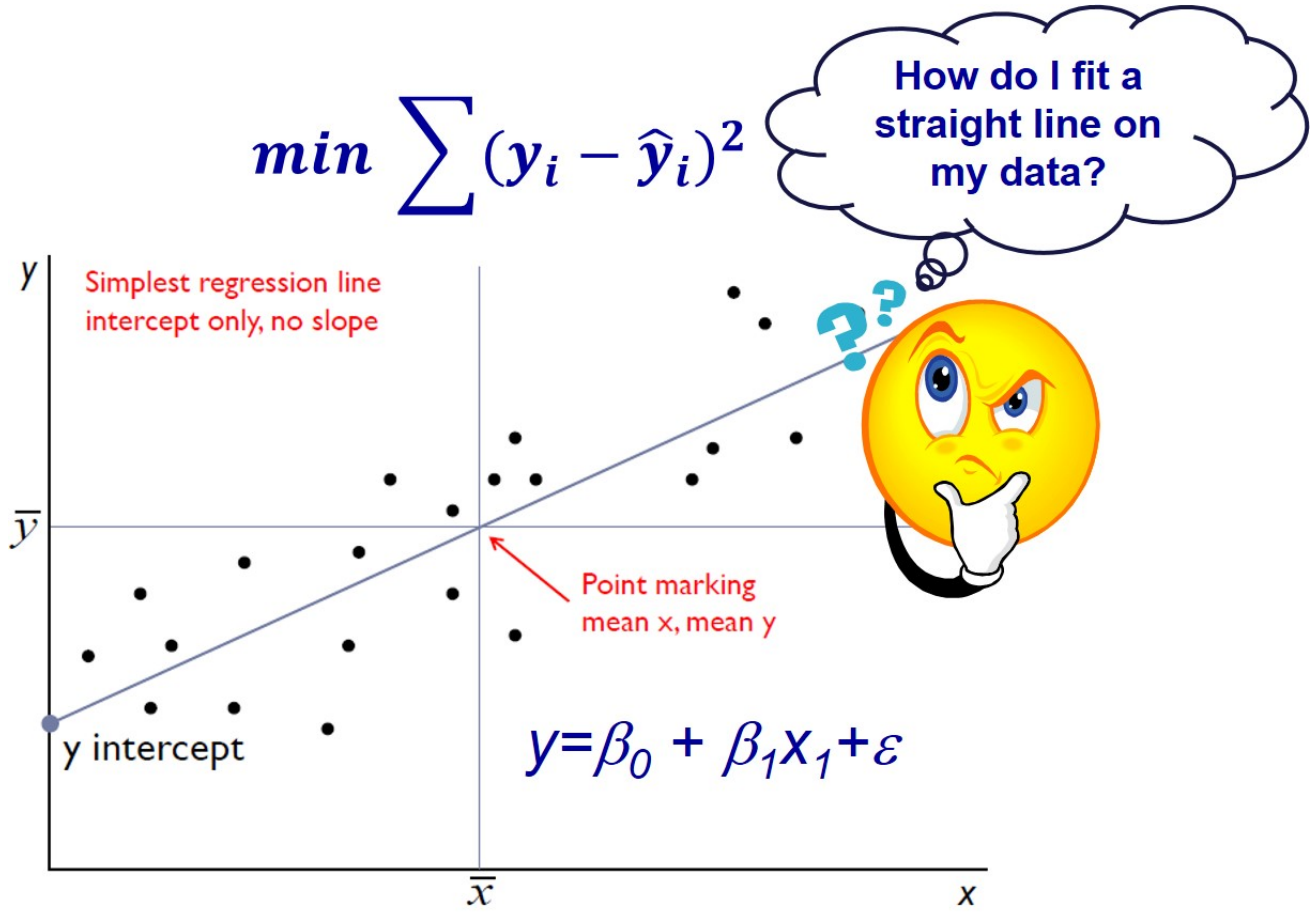
The formula

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$



The Least Squares Method

- The sum of the vertical deviations (y axis) of the points from the line is minimal



Multiple Linear Regression

- Regression establishes relationship among a dependent variable and a set of independent variable(s)
- A typical linear regression model looks like:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_n X_{ni}) + \varepsilon_i$$

- With Y_i the dependent variable, X_{ji} (j from 1 to n) the set of independent variables, β_0.. β_n are the regression coefficients and ε_i the residual.

Assessing the goodness of fit

- Sums of squares, R and R^2

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

Total sum of squares (TSS)

Regression (explained) sum of squares (ESS)

Residual (unexplained) sum of squares (RSS)

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

The proportion of the total explained variation in y is called the coefficient of determination (R^2)

Significance testing in regression

- Test hypothesis: That the variation explained by the models is not due to chance (F-test)

$$F = \frac{MS_M}{MS_R}$$

where

$$MS_M = \frac{\sum(\hat{y}_i - \bar{y})^2}{1}$$

$$MS_R = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2}$$

Goodness of fit test

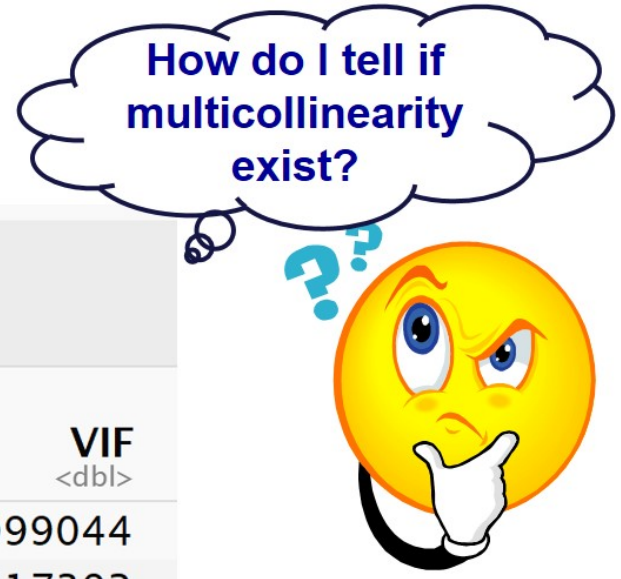
```
Residual standard error: 866500 on 1426 degrees of freedom
Multiple R-squared:  0.5394,    Adjusted R-squared:  0.5365
F-statistic: 185.6 on 9 and 1426 DF,  p-value: < 2.2e-16
```

Assessing individual parameters

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         6425.06   96357.57   0.067   0.9468
AREA_SQM           14106.74     412.56  34.193  < 2e-16 ***
AGE               -22712.35    3270.91  -6.944 5.78e-12 ***
PROX_CHILDCARE    -86666.14  118508.00  -0.731   0.4647
PROX_PRIMARY_SCH  367856.82   65457.91   5.620 2.30e-08 ***
PROX_SUPERMARKET -531151.08   75969.25  -6.992 4.16e-12 ***
PROX_BUS_STOP     129723.16  145395.72   0.892   0.3724
NO_Of_UNITS         -174.12      95.87  -1.816   0.0695 .
CARPARK            -2034.67   54772.69  -0.037   0.9704
FREEHOLD          517795.14   52704.94   9.824  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Are there redundant explanatory variables?

- Variance Inflation Factors (VIF)
- *ols_vif_tol()* of **olsrr** package



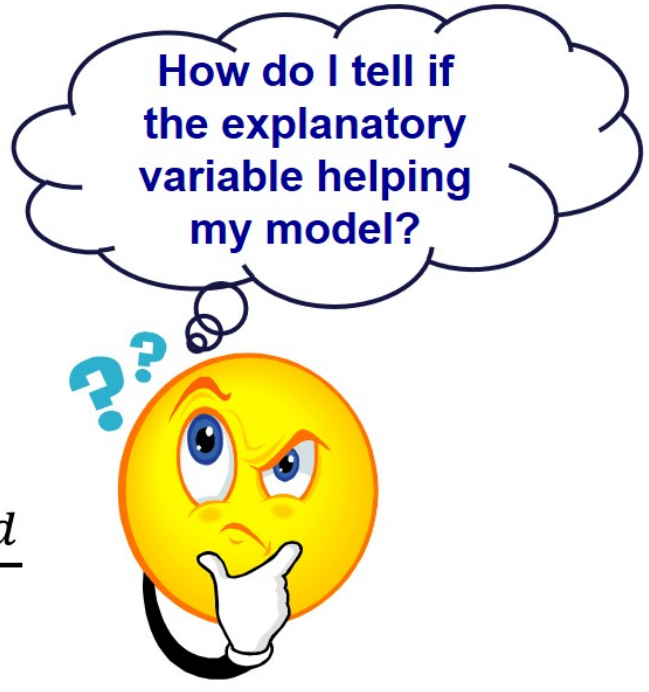

```{r}
ols_vif_tol(mlr)
```

Variables <chr>	Tolerance <dbl>	VIF <dbl>
AREA_SQM	0.9098817	1.099044
AGE	0.6590257	1.517392
PROX_CHILDCARE	0.3343099	2.991237
PROX_PRIMARY_SCH	0.5033832	1.986558
PROX_SUPERMARKET	0.8324691	1.201246
PROX_BUS_STOP	0.3951765	2.530515
NO_Of_UNITS	0.7630640	1.310506
CARPARK	0.7059033	1.416625
FREEHOLD	0.7713492	1.296430

Individual parameter testing

- Null hypothesis: $b = 0$
- Test statistics, t

$$\frac{b_{observed} - b_{expected}}{SE_b}$$



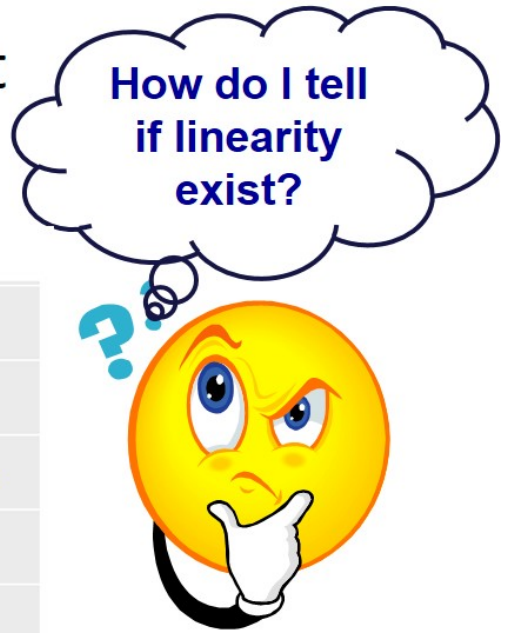
- The degrees of freedom are N - p - 1, where N is the total sample size and p is the number of predictors.

Assumptions of linear regression models

- **Linearity assumption**. The relationship between the dependent variable and independent variables is (approximately) linear.
- **Normality assumption**. The residual errors are assumed to be normally distributed.
- **Homogeneity of residuals variance**. The residuals are assumed to have a constant variance (homoscedasticity).
- The residuals are uncorrelated with each other.
  - serial correlation, as with time series
- (Optional) The errors (residuals) are normally distributed and have a 0 population mean.

The linearity assumption

- Actual values vs Fitted values scatterplot

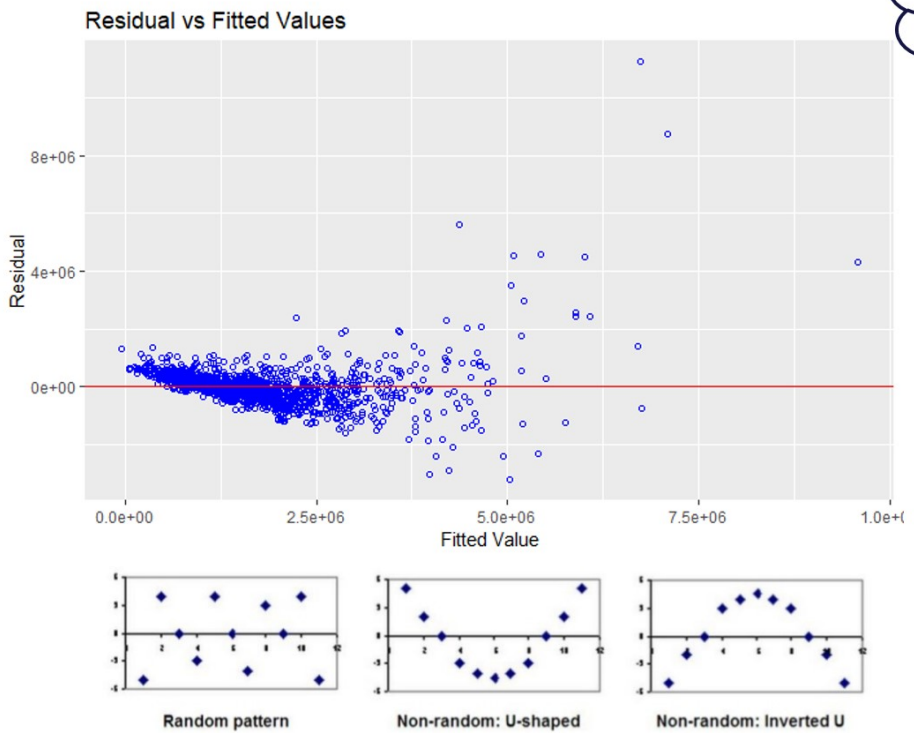


The linearity assumption

Residuals vs Fitted plot

- Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, what is good.

- Residual vs Fitted Values plot

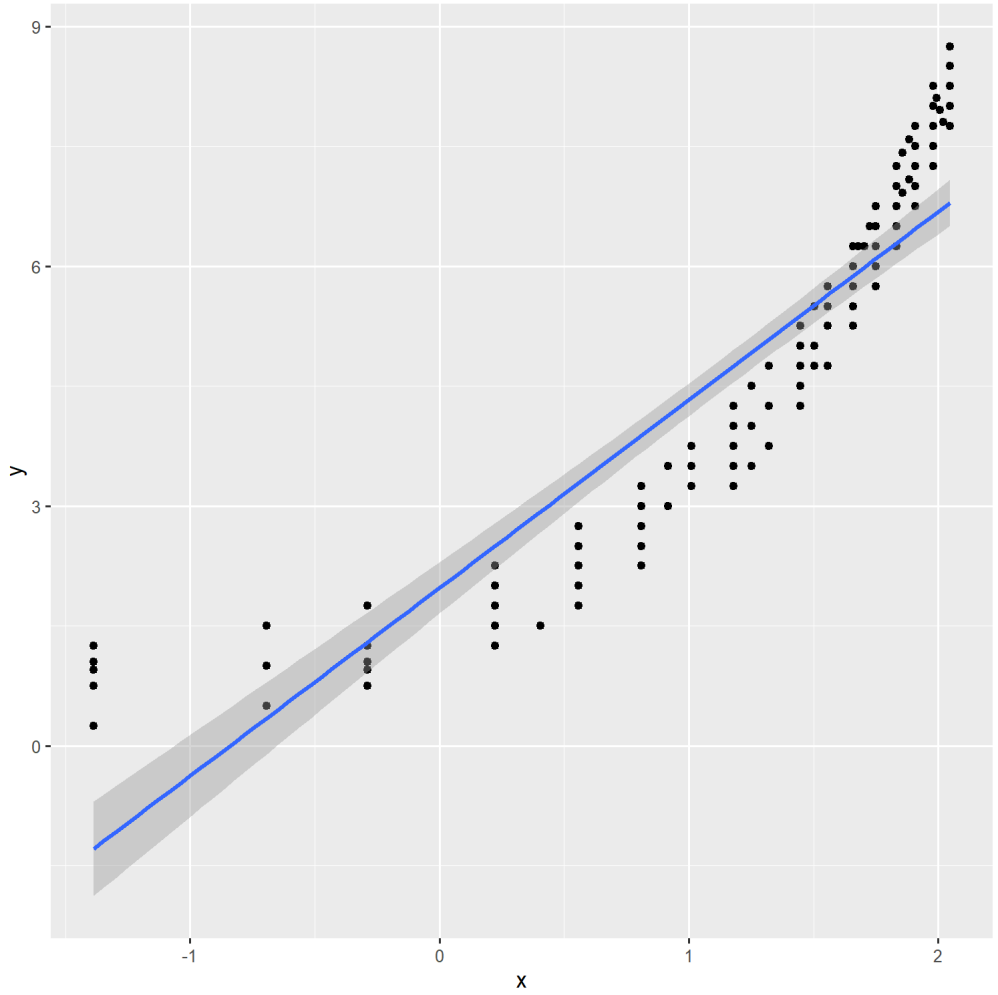


Demystifying the linearity assumption myth

The myth:

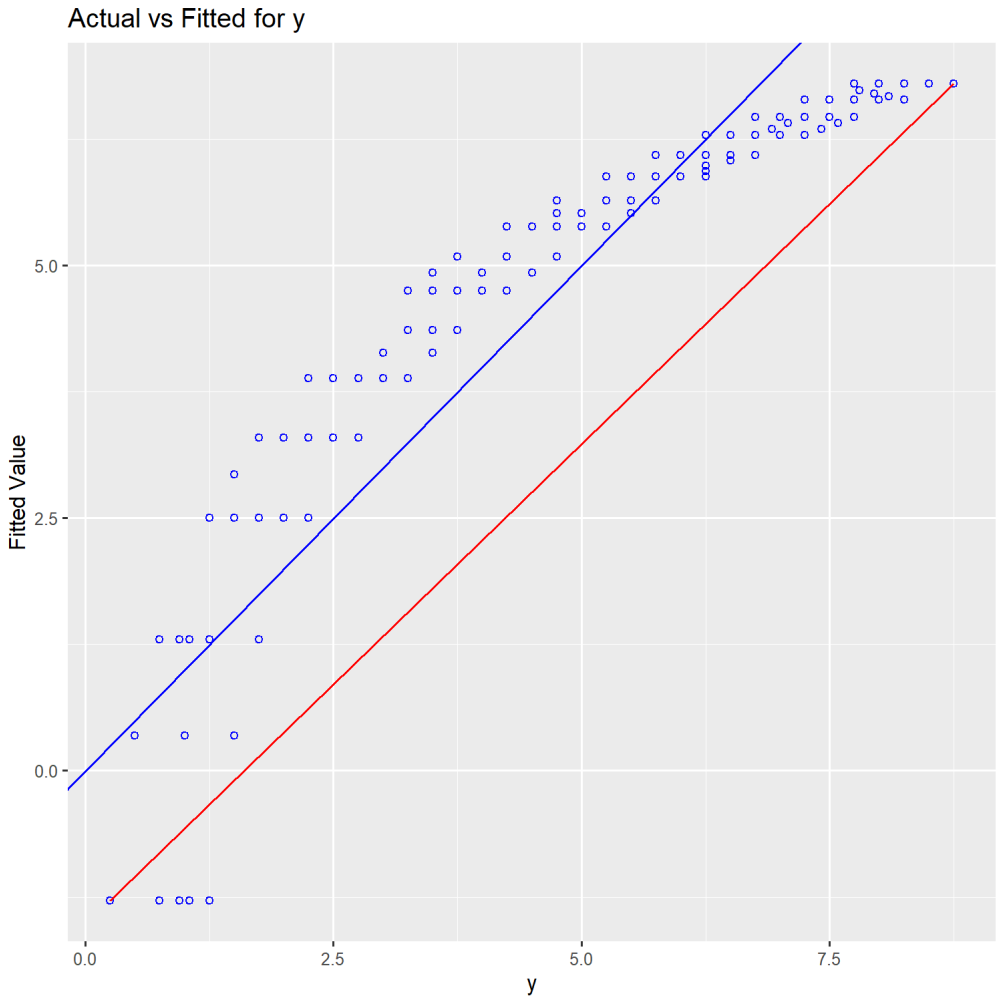
- We should transform the values of the y variable when they are large.

	y		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.98	1.66 – 2.30	**<0.001**
x	2.35	2.13 – 2.57	**<0.001**
Observations	100		
R^2 / R^2 adjusted	0.819 / 0.817		



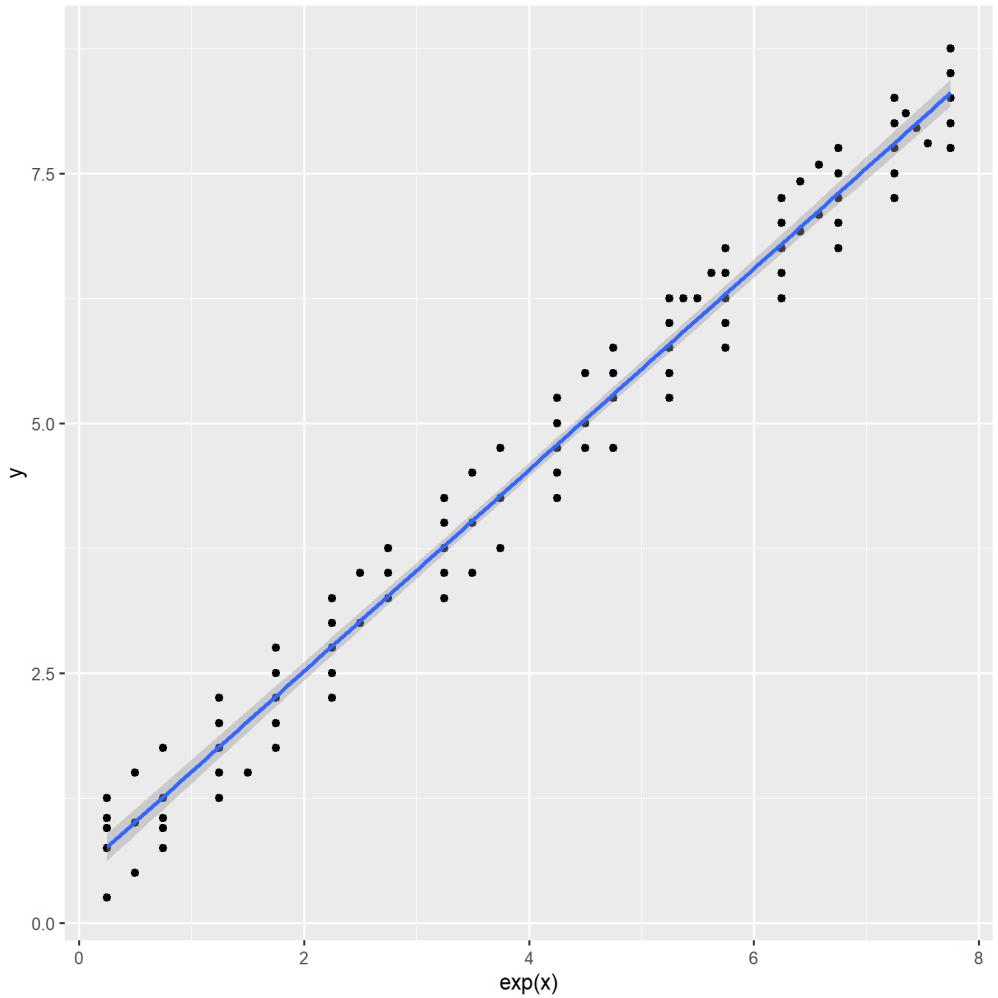
The linearity assumption

Despite the values of the dependent variable is rather similar to the values of the independent variable, the diagnostic plot shows that the linearity assumption has been violated.



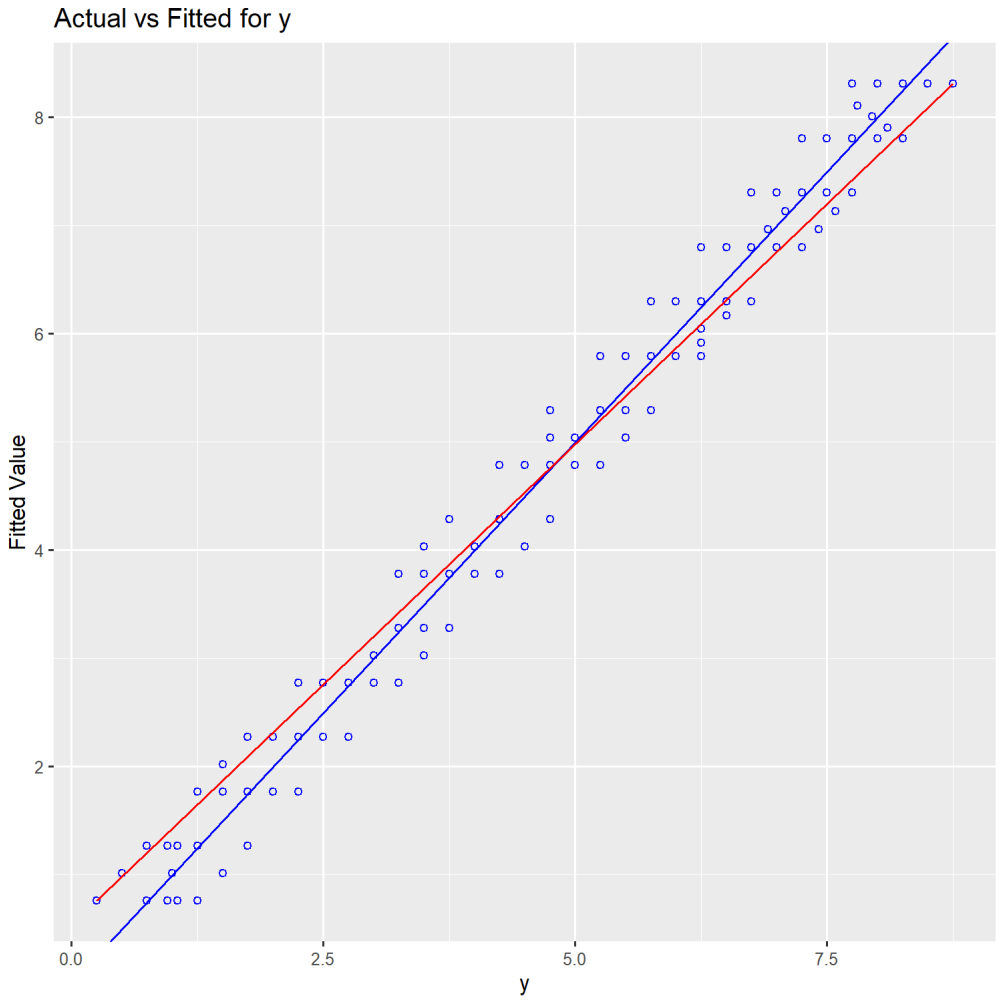
Data transformation come to rescue

	y		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.51	0.36 – 0.65	**<0.001**
x [exp]	1.01	0.98 – 1.04	**<0.001**
Observations	100		
R^2 / R^2 adjusted	0.977 / 0.977		



The linearity assumption

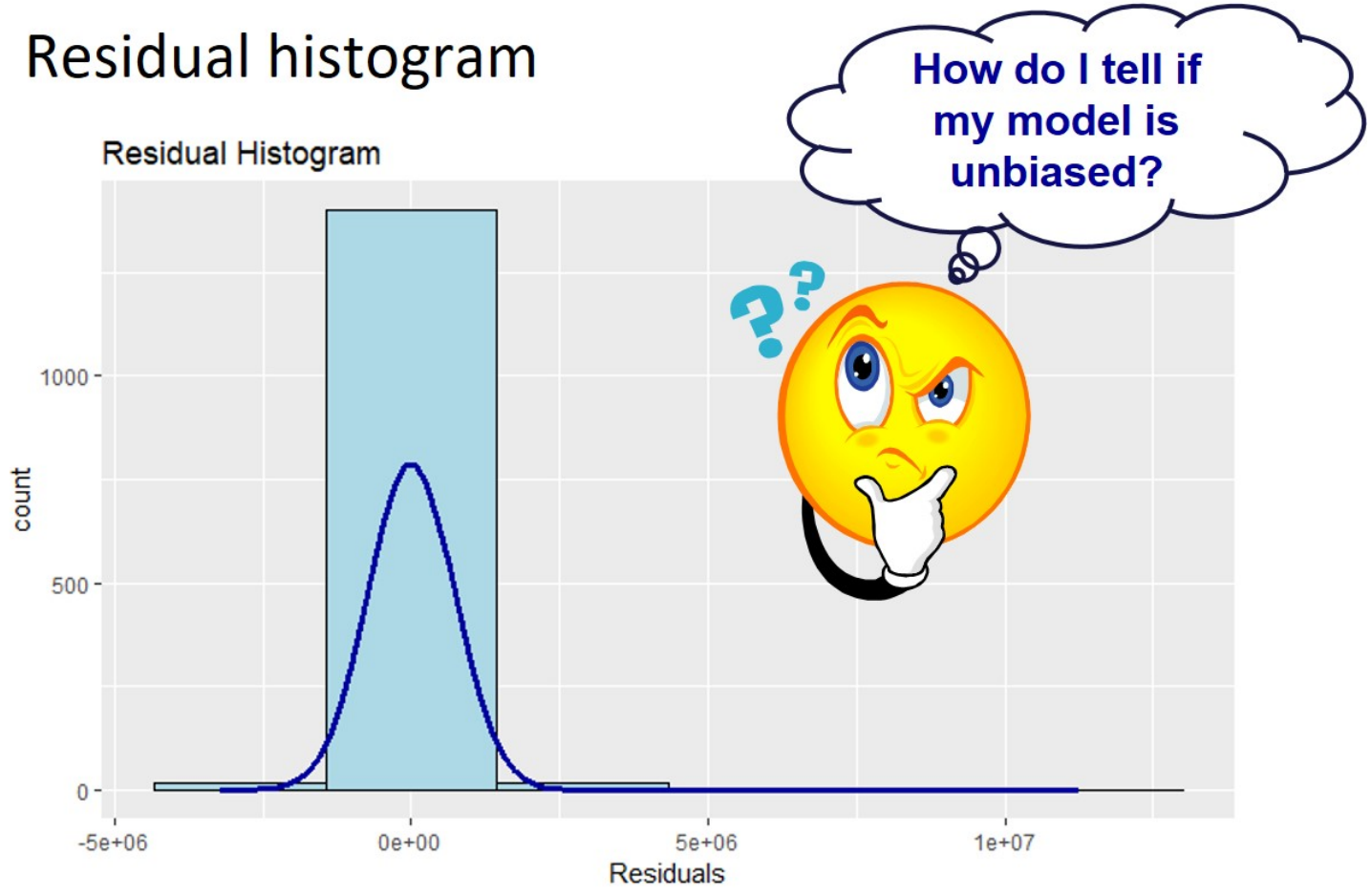
The diagnostic plot on the right shows that the linearity assumption has been conformed.



The normality assumption

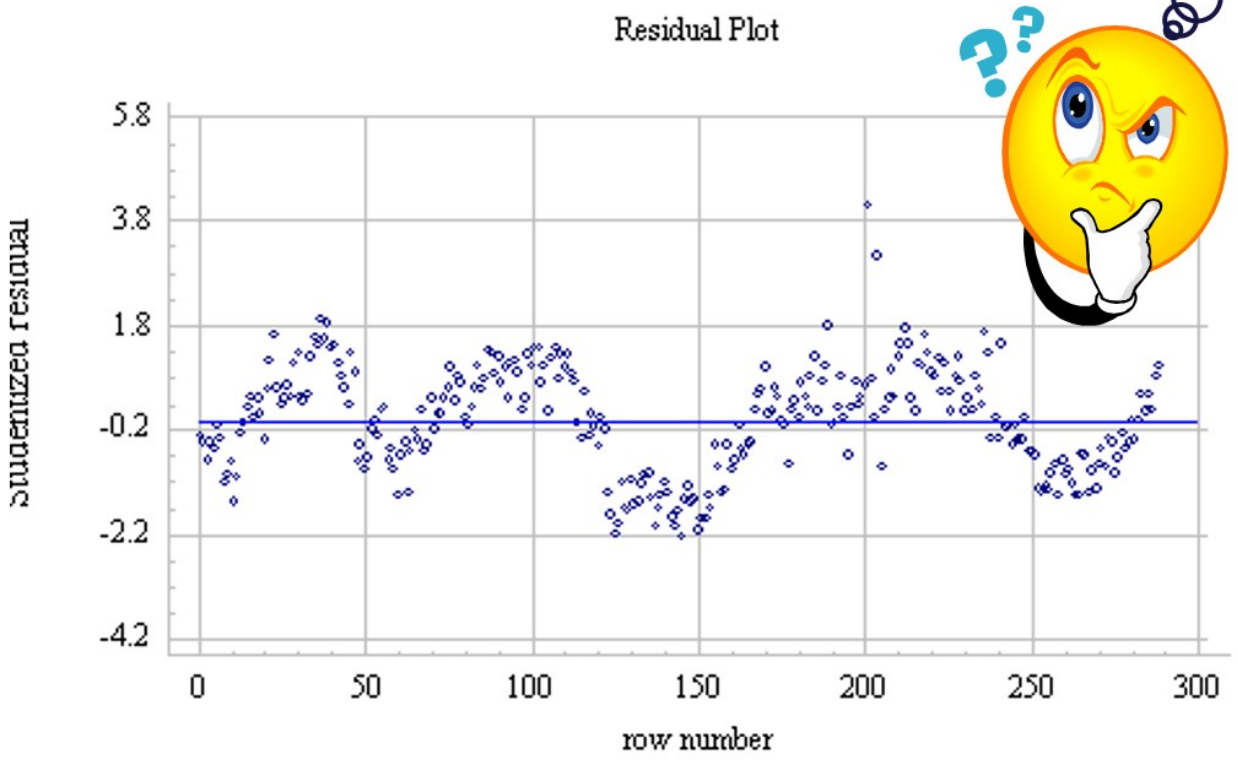
Warning: This is the test on the residual and not on the dependent variable.

- Residual histogram



Checking for serial correlation

- Residual vs Row Number plot



Spatial Non-stationary

- When applied to spatial data, as can be seen, it assumes a stationary spatial process
  - The same stimulus provokes the same response in all parts of the study region.
  - Highly untenable for spatial process

Why do relationships vary spatially?

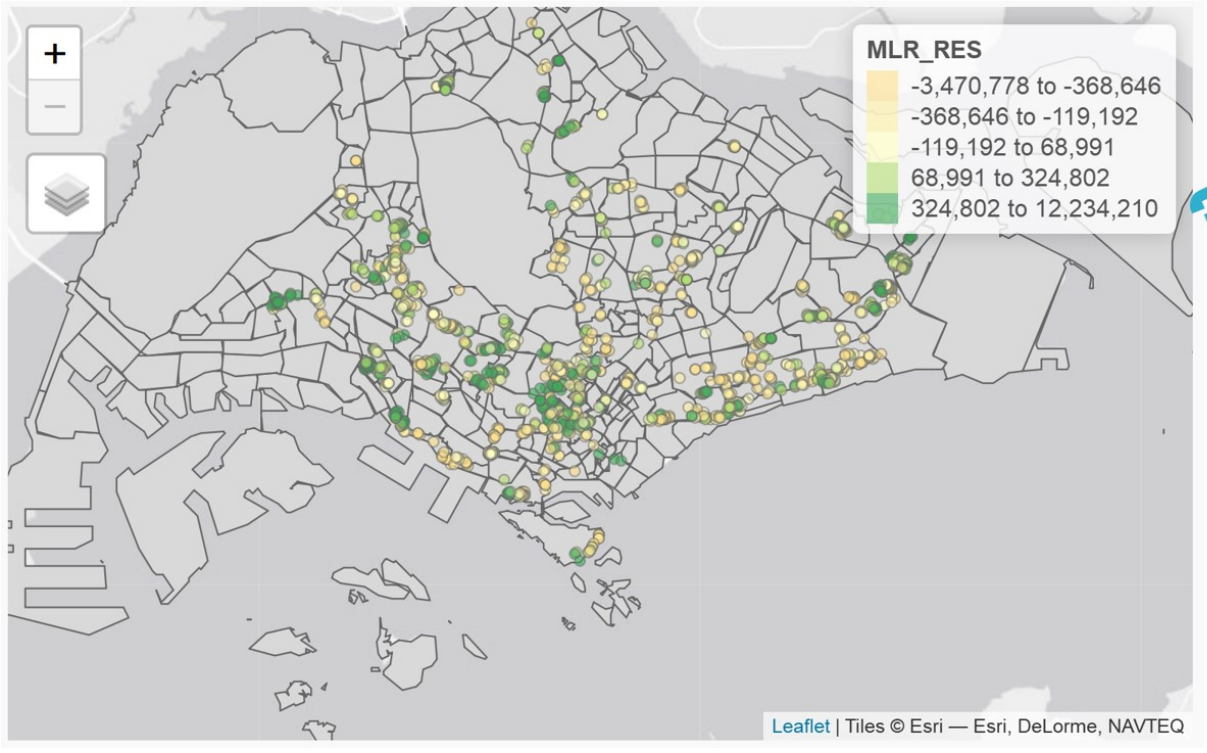
- Sampling variation
  - Nuisance variation, not real spatial non-stationarity
- Relationships intrinsically different across space
  - Real spatial non-stationarity
- Model misspecification
  - Can significant local variations be removed?

Some definitions

- Spatial non-stationarity: the same stimulus provokes a different response in different parts of the study region.
- Global models: statements about processes which are assumed to be stationary and as such are *location independent*.
- Local models: spatial decompositions of global models, the results of local models are *location dependent* – a characteristic we usually anticipate from geographic (spatial) data.

Spatial Autocorrelation assumption

The **residuals** are assumed to be distributed at random over geographical space.



Test of spatial autocorrelation

To test if the relationships in the model are **non-stationary**.

- *lm.morantest()* of spdep package will be used.

```
        Global Moran I for regression residuals

data:
model: lm(formula = SELLING_PRICE ~ AREA_SQM + AGE +
PROX_CBD + PROX_CHILDCARE + PROX_ELDERLYCARE +
PROX_URA_GROWTH_AREA + PROX_MRT + PROX_PARK +
PROX_PRIMARY_SCH + PROX_SHOPPING_MALL + PROX_BUS_STOP +
NO_Of_UNITS + FAMILY_FRIENDLY + FREEHOLD, data =
condo_resale.sf)
weights: nb_lw

Moran I statistic standard deviate = 35.586, p-value <
2.2e-16
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
    1.576993e-01    -4.343880e-03     2.073533e-05
```

Geographically Weighted Regression (GWR)

- Local statistical technique to analyze spatial variations in relationships.
- Spatial non-stationarity is assumed and will be tested.
- Based on the “First Law of Geography”: everything is related with everything else, but closer things are more related.

Geographically Weighted Regression (GWR): The method

- Mathematically, the GWR model is written as

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i$$

where $\beta_0(u_i, v_i)$ denotes the coordinates of the i-th point in space, and $\beta_k(u_i, v_i)$ is a realization of the continuous function at point i (Fotheringham et al. 1998)

Calibration of GWR

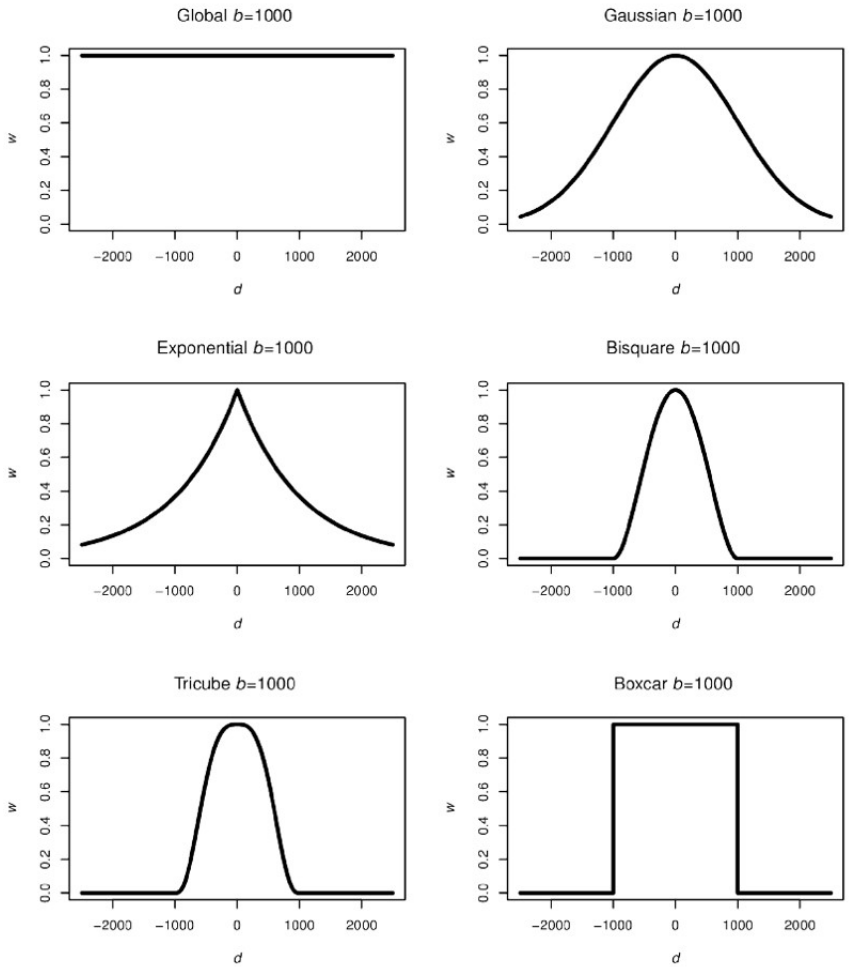
- Local weighted least squares
  - Weights are attached with locations
  - Based on the “First Law of Geography”: everything is related with everything else, but closer things are more related than remote ones

Calibration - Weighting functions

Global Model	$w_{ij} = 1$
Gaussian	$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right)$
Exponential	$w_{ij} = \exp\left(-\frac{\lvert d_{ij}\rvert}{b}\right)$
Box-car	$w_{ij} = \begin{cases} 1 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$
Bi-square	$w_{ij} = \begin{cases} (1-(d_{ij}/b)^2)^2 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$
Tri-cube	$w_{ij} = \begin{cases} (1-(\lvert d_{ij}\rvert/b)^3)^3 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$

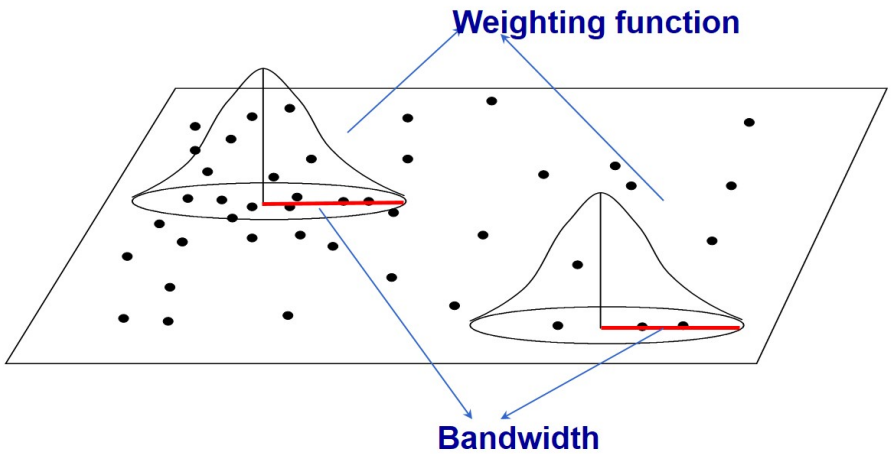
$\boldsymbol{w_{ij}}$ is the $\boldsymbol{j}$-th element of the diagonal of the matrix of geographical weights $W(u_i, v_i)$, and d_{ij} is the distance between observations i and j, and b is the bandwidth.

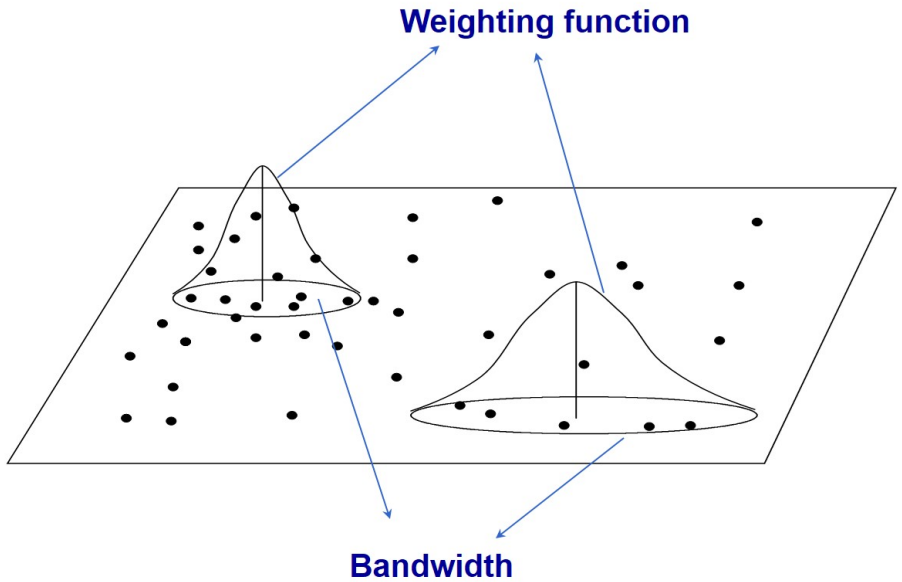
Calibration - Weighting functions



Calibration - Weighting schemes

- Determines weights
  - Most schemes tend to be Gaussian or Gaussian-like reflecting the type of dependency found in most spatial processes.
  - It can be either **Fixed** or **Adaptive**.





Calibration - Determining Bandwidth

- Least cross-validation (CV) score

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}_{\neq i}(b))^2$$

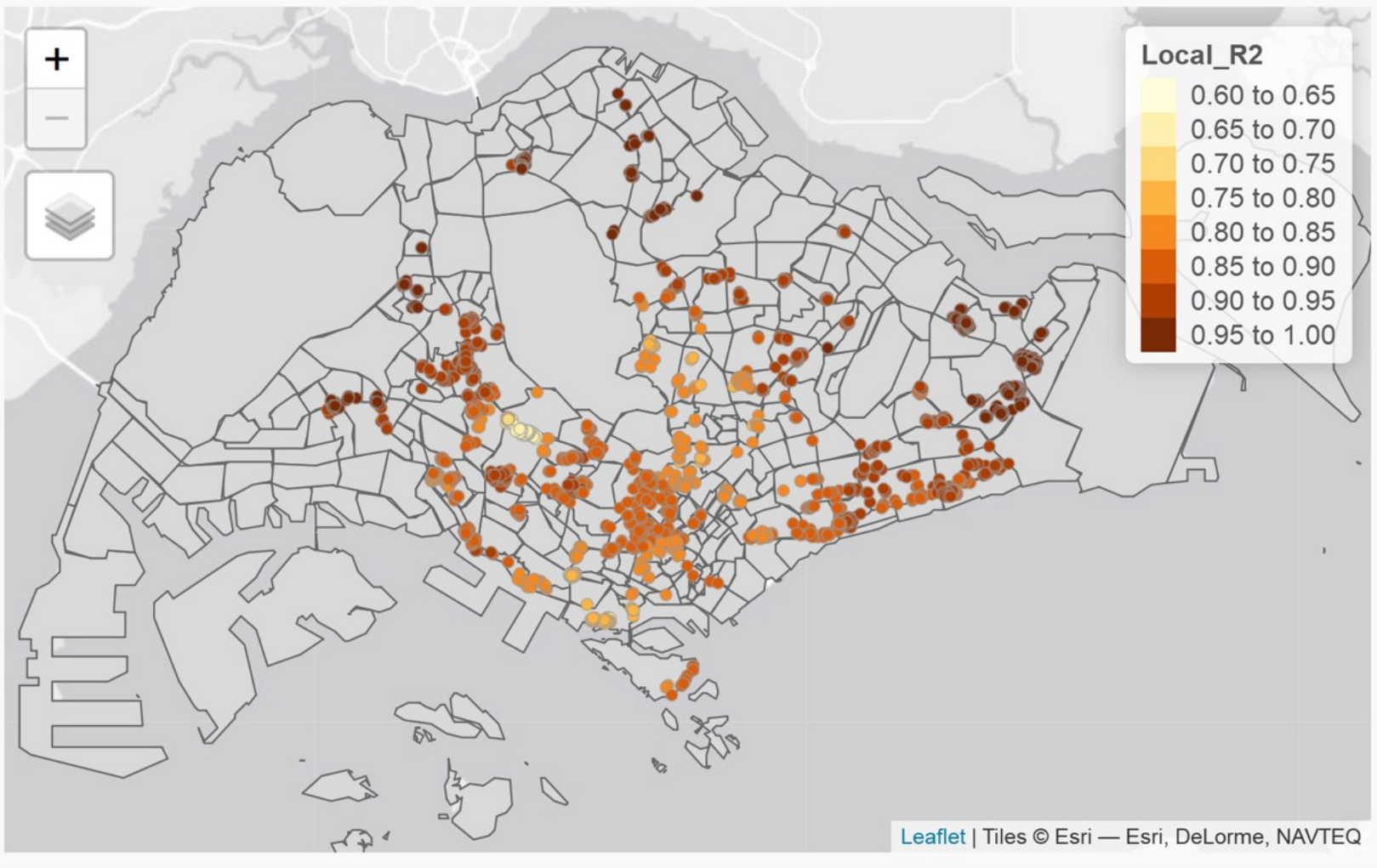
- Least Akaike Information Criterion (AIC)

$$\mathrm{AIC} = 2n\log_e(\hat{\sigma}) + n\log_e(2\pi) + n\left\{\frac{n + \mathrm{tr}(S)}{n - 2 - \mathrm{tr}(S)}\right\}$$

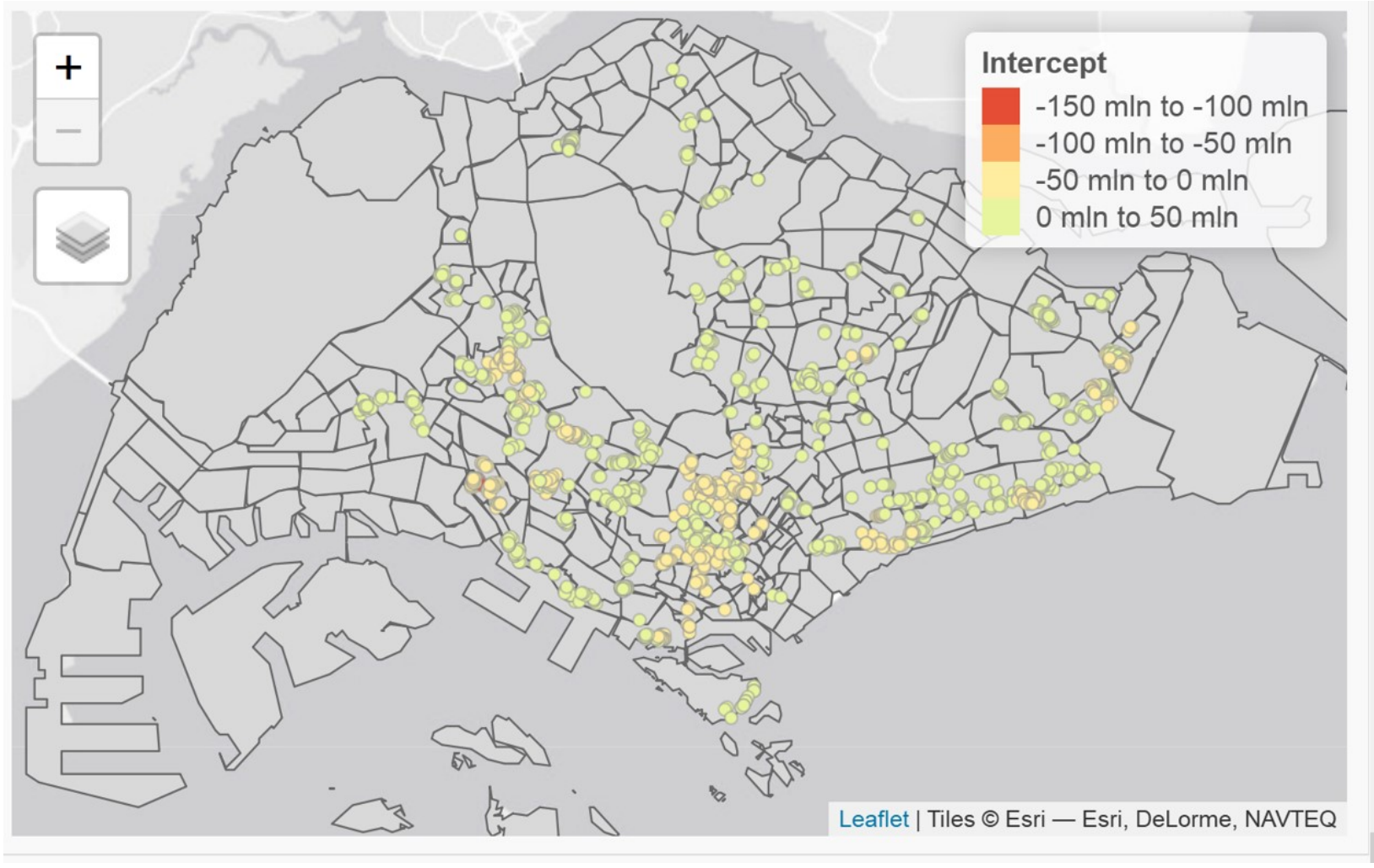
GWR Report

- Package Model
- Results of Global Regression
- Results of Geographically Weighted Regression
- SDF: A SpatialPointDataFrame

gwr: local R2


Local_R2
0.60 to 0.65
0.65 to 0.70
0.70 to 0.75
0.75 to 0.80
0.80 to 0.85
0.85 to 0.90
0.90 to 0.95
0.95 to 1.00
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

gwr: intercept



References

Brunsdon, C., Fotheringham, A.S., and Charlton, M. (2002) “Geographically weighted regression: A method for exploring spatial nonstationarity”. *Geographical Analysis*, 28: 281-289.

Brunsdon, C., Fotheringham, A.S. and Charlton, M., (1999) “Some Notes on Parametric Significance Tests for Geographically Weighted Regression”, 497-524.

Mennis, Jeremy (2006) "Mapping the Results of Geographically Weighted Regression", *The Cartographic Journal*, Vol.43 (2), p.171-179.

Stephen A. Matthews ; Tse-Chuan Yang (2012) "Mapping the results of local statistics: Using geographically weighted regression", *Demographic Research*, Vol.26, p.151-166.