Linear regression

The theoretical regression model:

$$\underbrace{y_i}_{\text{Dependent Variable}} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}}_{\text{Independent Variables}} + \underbrace{\epsilon_i}_{\text{Random Error}}$$

An estimated regression equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_j x_{ij}$$

The loss function minimized with the OLS procedure:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_j x_{ij})^2$$

2026 Zehui Yin

Trend surface

General form:

$$y_i = f(lat_i, long_i) + \epsilon_i$$

First degree:

$$y_i = \beta_0 + \beta_1 lat_i + \beta_2 long_i + \epsilon_i$$

Second degree:

$$y_i = \beta_0 + \beta_1 lat_i^2 + \beta_2 lat_i +$$
$$\beta_3 lat_i \cdot long_i + \beta_4 long_i + \beta_5 long_i^2 + \epsilon_i$$

2026 Zehui Yin

Spatially-varying Coefficients

Expansion Method:

Use the trend surface function in place of a constant coefficient:

$$y_i = f^1_{(lat_i, long_i)} + f^2_{(lat_i, long_i)} x_{i1} + \epsilon_i$$

$$f^1_{(lat_i, long_i)} = \beta_{01} + \beta_{02} lat_i + \beta_{03} long_i$$
$$f^2_{(lat_i, long_i)} = \beta_{11} + \beta_{12} lat_i + \beta_{13} long_i$$

Geographically Weighted Regression:

Instead of fitting a regression model to all the observations, it fits to a subset of them each time based on the bandwidth defined.

It is like a spatial moving average; instead of calculating an average, we are fitting a regression model to each window.

Spatial error model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \underbrace{\epsilon_i}_{\text{Random Error}}$$

The error term is no longer independent of each other but follows an autoregressive form:

$$\epsilon_i = \lambda \sum_k w_{ik} \epsilon_k + u_i$$

Noting that the u_i is the pure random component now.

Capstone activity expectation

For lab activity submission, you need to at least conduct **descriptive statistics**.

Load the data and use the `summary` method or provide some descriptive plots like histograms.

```
1 summary(df)
```

```
      mpg              cyl             disp             hp        
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
```

Activities for today

- We will work on the following chapter from the textbook:
  - Chapter 28: Activity 13: Area Data V
  - Chapter 30: Activity 14: Area Data VI
- The hard deadline is **Thursday, Novermber 12**.