# Geostatistical models - Part I

Applied Spatial Statistics

# Motivating example

- $Y_i$ is the microbiome species richness (SR) of sample $i$
- $\mathbf{s}_i = (s_{i1}, s_{i2})$ is the lat/lon of sample $i$
- $X_i$ is the net primary production (NPP) in the vicinity of sample $i$
- Link to maps of the microbiome data

# Objectives

1. Estimate the effect of NPP on SR

2. Determine if there is spatial correlation

3. Predict SR where it has not been measured

# Simple methods

1. Estimate the effect of NPP on SR – linear regression
2. Determine if there is spatial correlation – plot sample correlations
3. Predict SR where it has not been measured – take an average of nearby points

We will review these methods, discuss their limitations and introduce geostatistical alternatives

# Review of linear regression

First we review non-spatial linear regression

- Least squares
- Linear regression in matrix notation
- Maximum likelihood analysis

# Review of linear regression

- Response: $Y_i$ for $i \in \{1, ..., n\}$
- Covariates: the $p$ covariates are $X_{i1}, ..., X_{ip}$
- The model is

$$Y_i = \beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p + \varepsilon_i$$

- The mean $\mathrm{E}(Y_i) = \mu_i = \beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p$ is a linear combination of the covariates
- The errors/residuals $\varepsilon_i = Y_i - \mu_i$ are assumed to be independent and identically distributed

# Review of linear regression - least squares

- The slope $\beta_j$ is interpreted as the increase in the mean if $X_{ij}$ increases by one with all other variables held fixed
- Let $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$ be the collection of all slopes put in a column vector
- We measure how well a candidate $\boldsymbol{\beta}$ fits the data using the sum of squared errors

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \mu_i)^2$$

where $\mu_i = \beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p$

# Review of linear regression - least squares

- We use as the estimate of $\boldsymbol{\beta}$ the value that minimizes the sum of squared errors
- Denote this estimate as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, ..., \hat{\beta}_p)^T$
- In math notation

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ SSE(\boldsymbol{\beta})$$

- The estimated mean and residuals are

$$\hat{\mu}_i = \hat{\beta}_0 + X_{i1}\hat{\beta}_1 + ... + X_{ip}\hat{\beta}_p$$

and $\hat{\varepsilon}_i = Y_i - \hat{\mu}_i$

# Review of linear regression - matrix notation

- The notation and least squares solution have tidy expressions when written using matrices
- The response vector is the $n \times 1$ matrix

$$\mathbf{Y} = (Y_1, ..., Y_n)^T$$

- The covariate matrix is the $n \times (p+1)$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & ... & X_{1p} \\ 1 & X_{21} & ... & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & ... & X_{np} \end{pmatrix}$$

- Note that matrices and vectors are written in bold face

# Review of linear regression - matrix notation

- Using this notation the model for all $n$ observations is simply written
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$
where $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ is the vector of errors

- The least squares solution is
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- This is a famous expression, wait for the chorus

# Review of linear regression - MLE

- Least squares is a great way to estimate parameters, but it only applies to a few problems
- Maximum likelihood estimation (MLE) is more general
- The likelihood function is the distribution of the data ($\mathbf{Y}$) given the parameters ($\beta$)
- This requires picking a distribution for the errors.
- The most common assumption is $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$, independent over $i$

# Review of linear regression - MLE

- The Gaussian linear regression model

$$Y_i \sim \text{Normal}(\mu_i, \sigma^2),$$

  indepenent over $i$

- Since the observations are independent, the distribution of $(Y_1, ..., Y_n)$ is the product of $n$ Gaussian distributions

- The likelihood is

$$L(\beta) = \prod_{i=1}^{n} \phi(Y_i; \mu_i, \sigma^2),$$

  where $\phi(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$ is the normal PDF

# Review of linear regression - MLE

- Putting this together gives

$$L(\boldsymbol{\beta}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2\right\}$$

- The likelihood is related to the sum of squared errors

$$L(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2\sigma^2} SSE(\boldsymbol{\beta})\right\}$$

- The MLE is the $\boldsymbol{\beta}$ that maximizes the likelihood function

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\ L(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\text{argmin}}\ SSE(\boldsymbol{\beta})$$

- Therefore, for linear regression assuming normality, the least squares solution is also the MLE

# Review of linear regression - R code

- The R function that performs linear regression is `lm`
- You can enter the variables one at a time

  fit <- lm(Y∼X1+X2+X3)

  or where $X$ is an $n \times p$ matrix

  fit <- lm(Y∼X)
- This stores the output in an object called fit, which can be accessed via

  summary(fit)
- Regression for the microbiome data

# Linear regression for spatially-correlated data

- Can we apply least squares to spatial data such as the microbiome data?
- Well, this is not the worst idea ever
- Correlated $\varepsilon_i$ violates a model assumption, but **the least squares estimator remains unbiased**
- However, the least squares estimator is suboptimal
- Also, uncertainty estimates (standard errors, confidence intervals, p-values) are invalid
- Ignoring correlation generally leads to standard errors that are too small

# Spatial covariance model

- To improve efficiency and have valid uncertainty quantification, we model/estimate the spatial covariance
- Estimating the covariance function also leads to optimal prediction at unmeasured locations (Kriging)
- How to estimate the correlation between $Y_1$ and $Y_2$?
- How about the sample correlation?
- The sample correlation is undefined with only one observation at each spatial location
- The sample correlation would be valid if we have replications, say data each day for a year at all locations

# Spatial covariance models

- For the canonical example without replication, we need assumptions about the spatial correlation
- These simplifying assumption give “spatial replications”
- For example, assume the correlation is the same for all pairs of sites separated by 20 miles
- If our dataset includes dozens of pairs of sites separated by 20 miles, then we collect all such pairs and compute the sample correlation estimator

# Spatial linear models

- Below we introduce a standard spatial regression model
- We assume the responses are Gaussian, which is an assumption that needs to be verified
- We will also introduce simplifying assumptions about the spatial covariance (isotropy, stationarity, etc)
- In this lecture we will introduce the model and discuss the role/interpretation of each component
- In future lectures we will discuss how to use data to estimate the parameters of the model

# Spatial linear models

The standard model is model is $Y_i = \mu_i + Z_i + \varepsilon_i$

- The mean is the same as linear regression

$$\mu_i = \beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p$$

- There are two error terms:
  - $Z_i$ is spatially-correlated
  - $\varepsilon_i$ are independent across $i$
- If the $Z_i = 0$, then this reduces to non-spatial linear regression

# Spatial linear models - mean structure

$$\mu_i = \beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p$$

- The covariates included in the model can be spatial variables: elevation, distance to a highway, latitude, etc
- They can also be non-spatial: time of day, visibility at the time of measurement, etc
- Covariate often explain the spatial pattern in the data and the residuals are uncorrelated
- For this reason, it is usually a good idea to include latitude, longitude and maybe their squares as covariates
- We will usually plot the least-squares residuals to inspect spatial correlation

# Spatial linear models - mean structure

Which variables might we include in the air pollution example?

- 
- 
- 
- 

# Spatial covariance models - nugget effect

- The independent error $\varepsilon_i$ are called the **nugget** term

- They are distributed $\varepsilon_i \sim \text{Normal}(0, \tau^2)$, independent over $i$

- Sources: measurement error, small-scale variation that cannot be explained

# Spatial linear models - nugget effect

Which factors might contribute to nugget error in the air pollution example?

- 
- 
- 
- 

# Spatial covariance models

- The $Z_i$ capture spatial correlation not explained by the covariates
- Modeling these spatial terms is the heart of spatial statistics
- In Part II we will discuss several models and their properties