Lecture 8: Point Level Models - Model Fitting

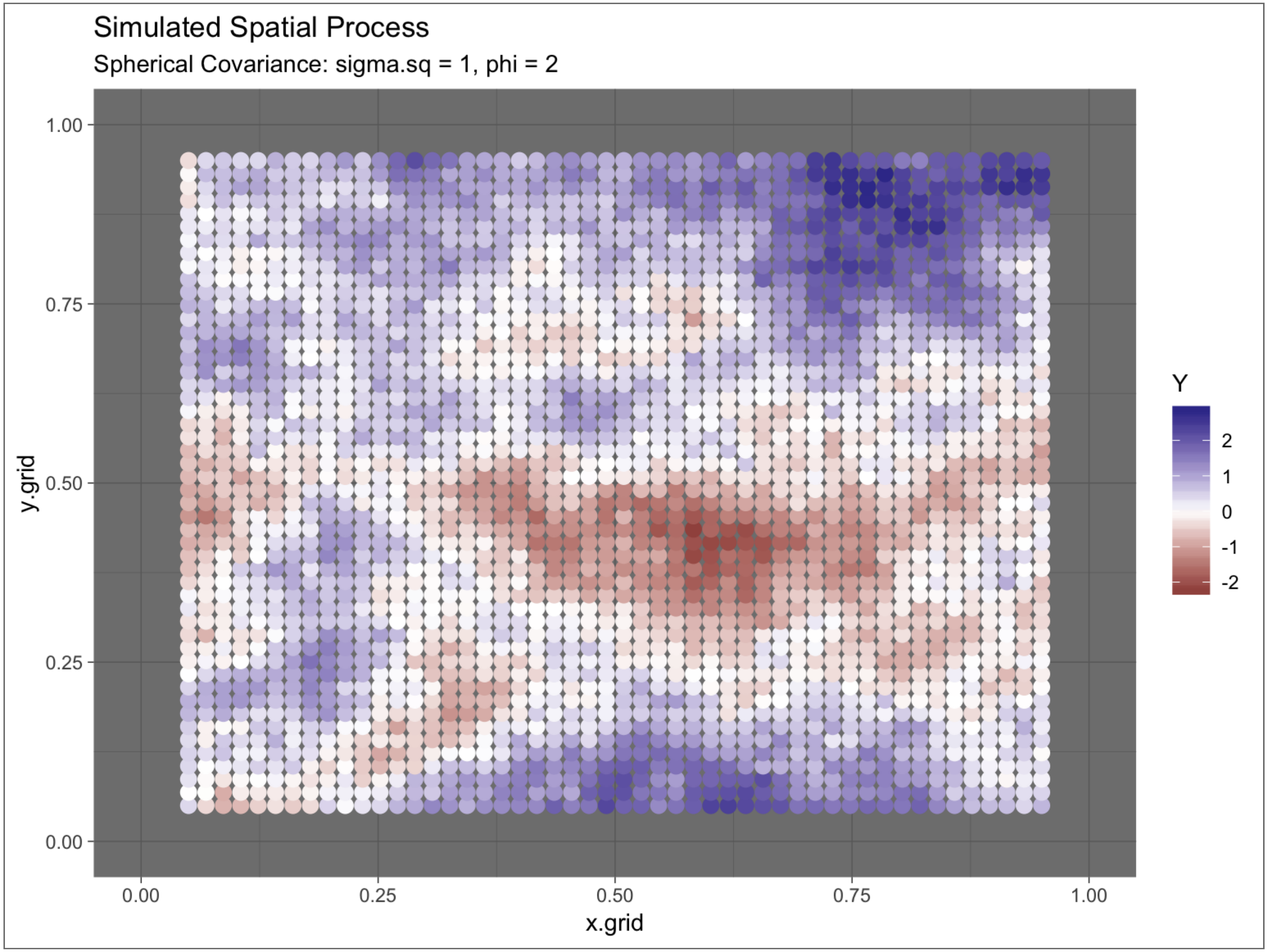
Class Intro

Intro Questions

- Why are we creating simulated spatial processes?
- For Today:
  - Model Fitting

Model Simulation

Simulating Spatial Process



Simulated Spatial Process: Exercise

How does the spatial process change with:

- another draw with same parameters?
- a different value of ϕ
- a different value of σ^2
- adding a nugget term, τ^2

Model Fitting

Classical Model Fitting

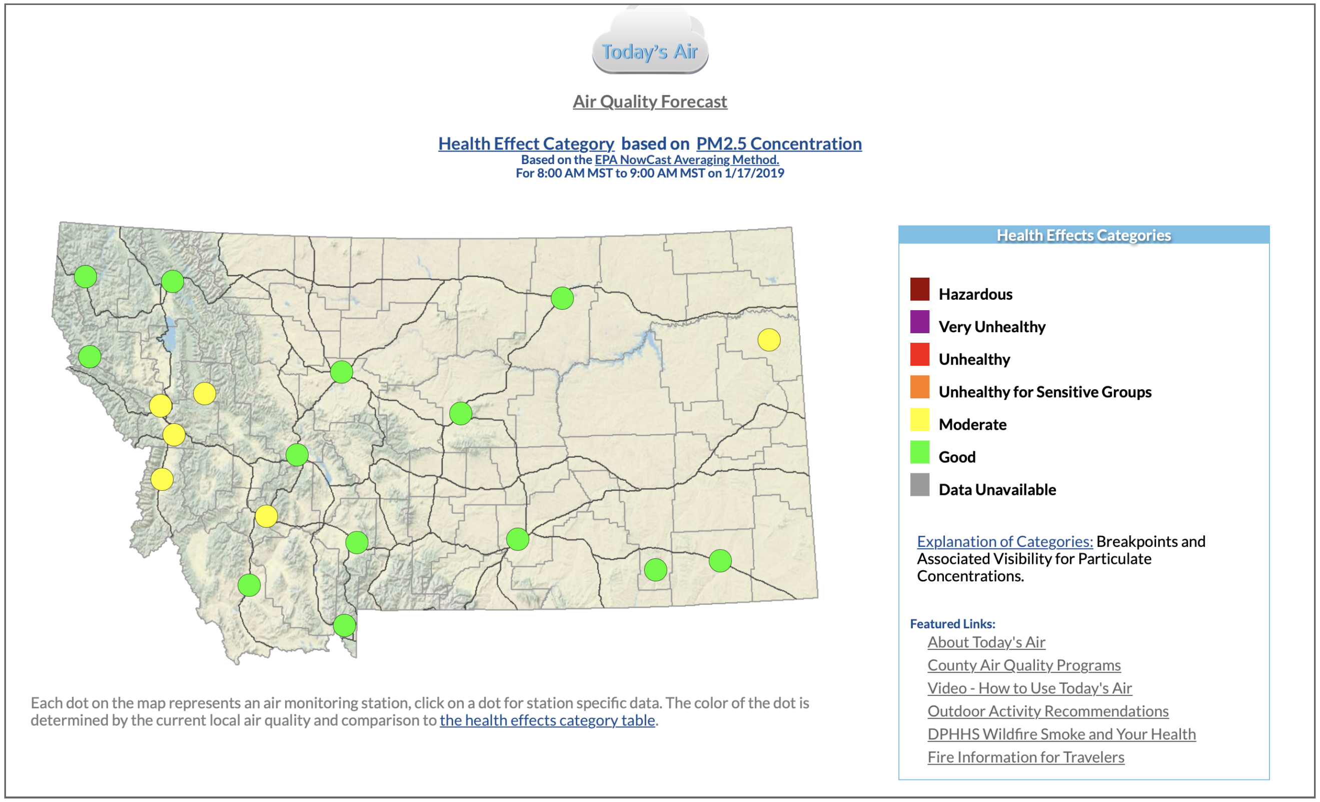
- The classical approach to spatial prediction is rooted in minimizing the mean-squared error.
- This approach is often referred to as *Kriging* in honor of D.G. Krige a South African mining engineer.
- As a result of Krige’s work (along with others), point-level spatial analysis and geostatistical analysis are used interchangeably.

Mathematical Motivation

- Let $\boldsymbol{Y} = \{Y(\boldsymbol{s_1}), \ldots, Y(\boldsymbol{s_n})\}$ be observations of a spatial process and n sites.
- Then $Y(\boldsymbol{s_0})$ is a site where the spatial process has not been observed.
- **The Goal:** What is the best predictor for $Y(\boldsymbol{s_0})$ given that $\boldsymbol{Y} = \{Y(\boldsymbol{s_1}), \ldots, Y(\boldsymbol{s_n})\}$ was observed?

Visual Motivation

- **The Goal:** What is the best predictor for $Y(\boldsymbol{s_0})$ given that $\boldsymbol{Y} = \{Y(\boldsymbol{s_1}), \ldots, Y(\boldsymbol{s_n})\}$ was observed?



source: airnow.gov

Mathematical Notation

- A linear predictor for $Y(\boldsymbol{s_0})$, given $\boldsymbol{Y}$ takes the form $\sum_i l_i Y(\boldsymbol{s_i}) + \delta_0$
- Using squared error loss, we'd seek to minimize
  $$E\left[Y(\boldsymbol{s_0}) - \left(\sum_i l_i Y(\boldsymbol{s_i}) + \delta_0\right)\right]^2$$
  as a function of l_i and δ_0.
- Describe and interpret l_i and δ_0

Connection to variogram

- Recall the intrinsic stationarity assumption
  $E[Y(\boldsymbol{s}+\boldsymbol{h})-Y(\boldsymbol{s})]=0$,
  thus $\sum_i l_i = 1$ such that
  $$E[Y(\boldsymbol{s_0})-\sum_i l_i Y(\boldsymbol{s_i})]=0$$
- Following this logic, we would now minimize
  $$E[Y(\boldsymbol{s_0})-\sum_i l_i Y(\boldsymbol{s_i})]^2+\delta_0^2,$$
  thus $\delta_0=0$.

Connection to variogram: part 2

- Define $a_0 = 1$ and $a_i = -l_i$, then we can rewrite
$$E[Y(\boldsymbol{s_0}) - \sum_i l_i Y(\boldsymbol{s_i})]^2 \quad \text{as} \quad E[\sum_{i=0}^{n} a_i Y(\boldsymbol{s_i})]^2$$
- It turns out that
$$E[\sum_{i=0}^{n} a_i Y(\boldsymbol{s_i})]^2 = -\sum_i \sum_j a_i a_j \gamma(\boldsymbol{s_i} - \boldsymbol{s_j})$$
- In other words, minimizing the squared error, under assumptions, justifies the variogram.
- This is a contrained optimization of a quadratic form that is typically handled with a Lagrange multiplier. **Khan Academy Refresher Video**

Lagrange multipliers

To rewrite the constrained optimization in terms of l_i we get

$$-\sum_{i=0}^{n}\sum_{j=0}^{n} a_i a_j \gamma(\boldsymbol{s_i} - \boldsymbol{s_j}) = -\sum_{i=1}^{n}\sum_{j=1}^{n} l_i l_j \gamma_{ij} + 2\sum_{i=1}^{n} l_i \gamma_{0i},$$

where $\gamma_{ij} = \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j)$ and hence $\gamma_{0j} = \gamma(\boldsymbol{s}_0 - \boldsymbol{s}_j)$

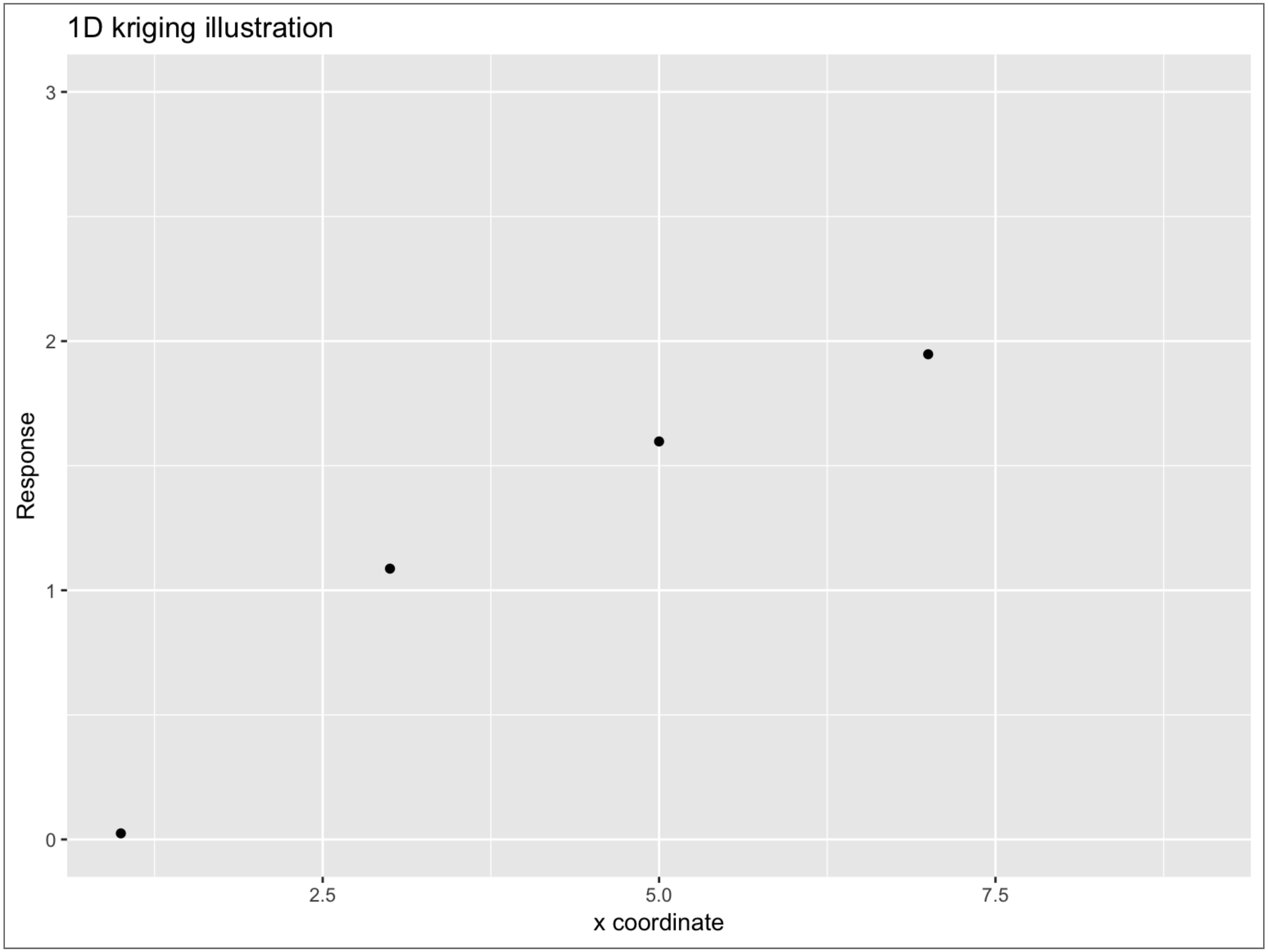
- Solving equations with this constraint requires partial derivatives and the use of Lagrange multipliers, typically denoted as λ.

BLUP

- It turns out that the solution for the vector $\boldsymbol{l}$ is
  $\boldsymbol{l} = \Gamma^{-1}\left(\boldsymbol{\gamma_0} + \frac{(1 - \mathbf{1}^T\Gamma^{-1}\boldsymbol{\gamma_0})}{\mathbf{1}^T\Gamma^{-1}\mathbf{1}}\mathbf{1}\right),$
  where Γ is an $n \times n$ matrix with entries $\Gamma_{ij} - \gamma_{ij}$ and $\boldsymbol{\gamma_0}$ is the vector of γ_{0i} values.
- Then the Best Linear Unbiased Predictor is $\boldsymbol{l}^T\boldsymbol{Y}$
- This BLUP also requires an estimate of $\gamma(\boldsymbol{h})$

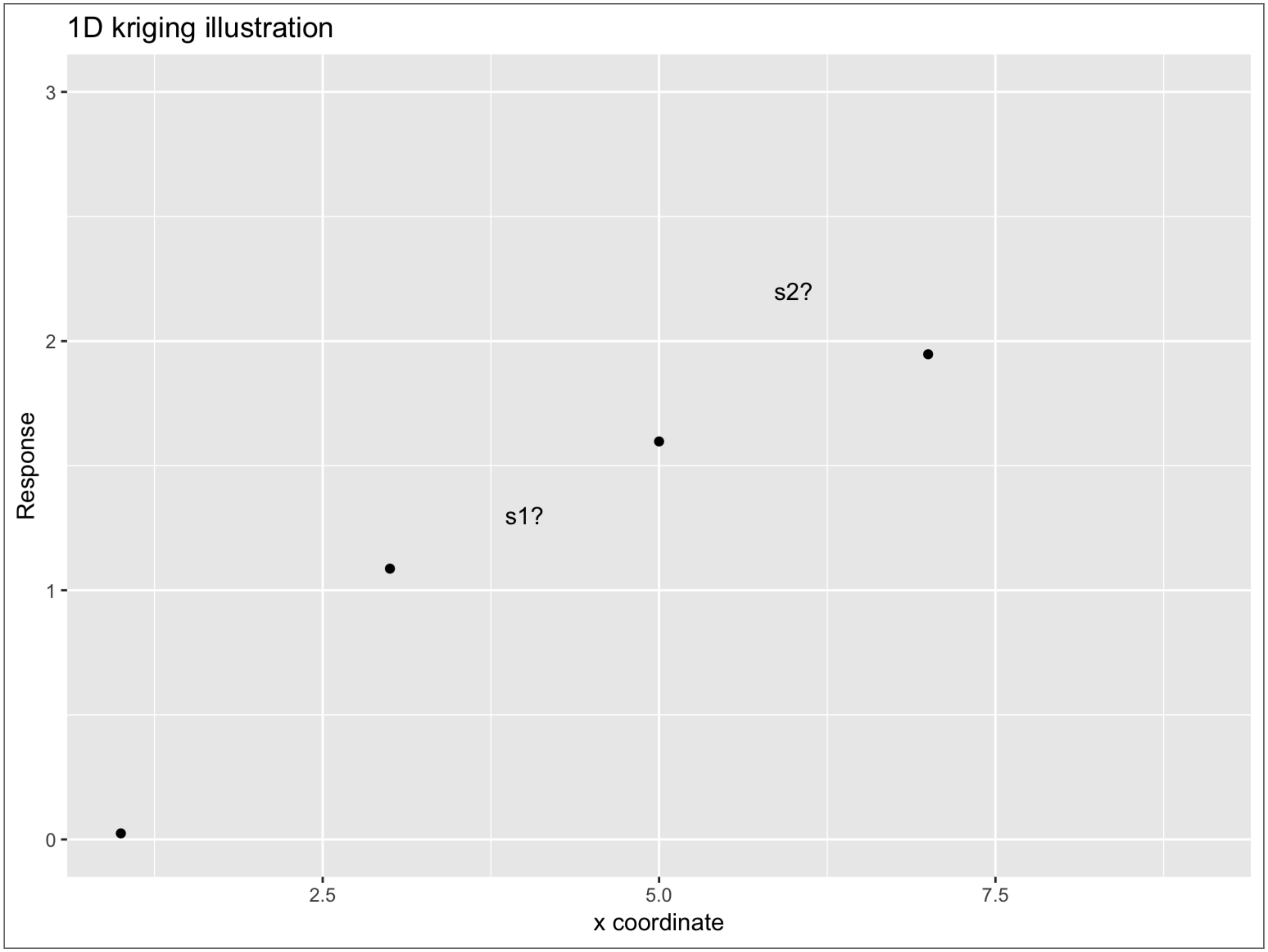
So what does this all mean …

Consider a small example on 1-dimension.


1D kriging illustration
Response
x coordinate
0
1
2
3
2.5
5.0
7.5

So what does this all mean …

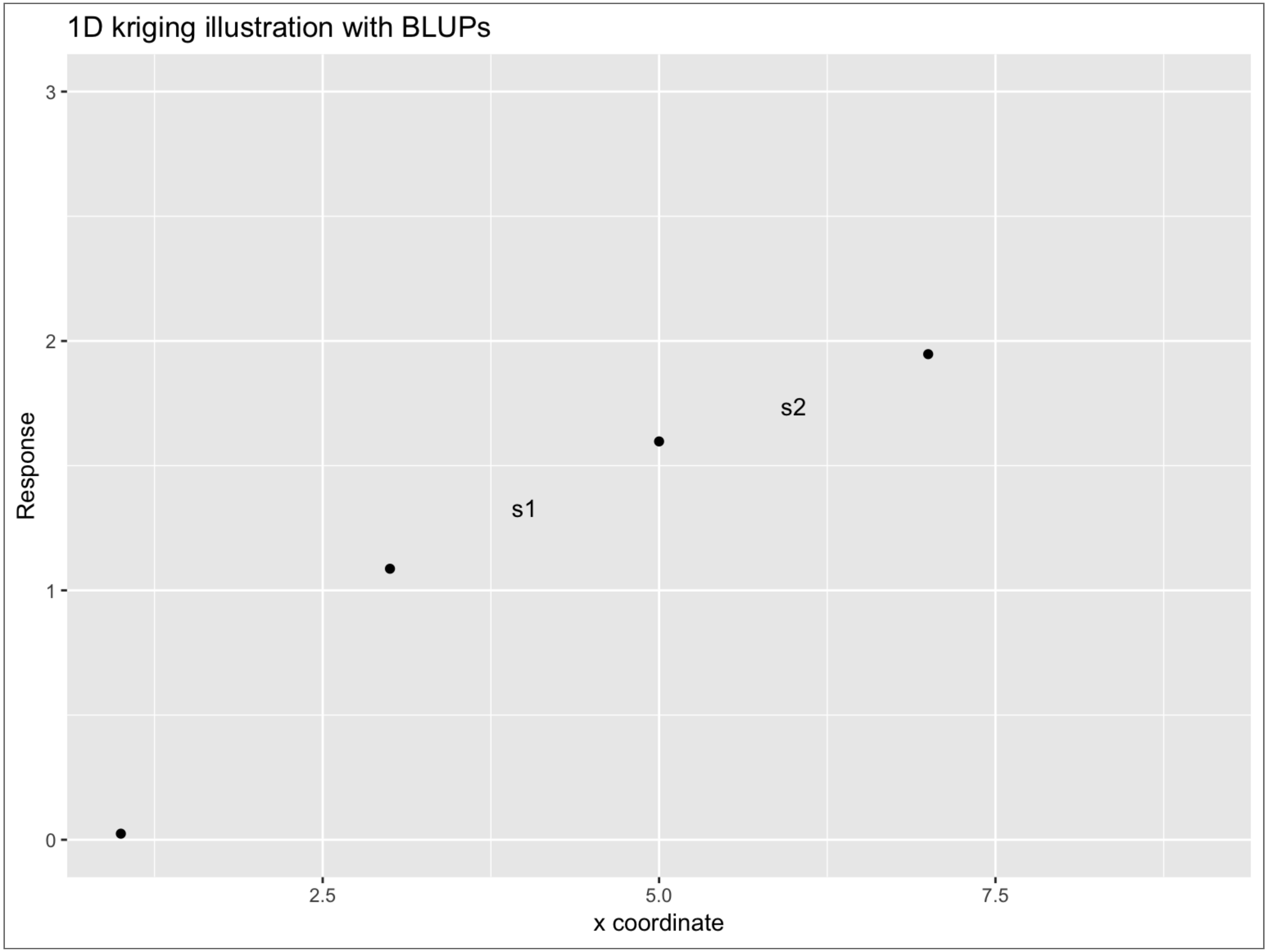
What should the predictions be at $\boldsymbol{s}_1^* = 4$ and $\boldsymbol{s}_2^* = 6$


1D kriging illustration
Response
x coordinate
3
2
1
0
2.5
5.0
7.5
s1?
s2?

Kriging Exercise:

- Recall
$$\boldsymbol{l} = \Gamma^{-1}\left(\boldsymbol{\gamma_0} + \frac{(1 - \mathbf{1}^T\Gamma^{-1}\boldsymbol{\gamma_0})}{\mathbf{1}^T\Gamma^{-1}\mathbf{1}}\mathbf{1}\right),$$
- Define $\gamma(h) = 1 - \exp(-\frac{h}{3})$ and compute the BLUPs for $\boldsymbol{s}_1^*$ and $\boldsymbol{s}_2^*$
- Interpret and explain $\boldsymbol{l}$ for each sample point.
- If you have time, fill in the line (rather than the surface) from (0.5, 7.5)

Kriging Solution



Kriging with Gaussian Processes

A Gaussian Process

- The BLUP does not contain a distributional assumptions, but rather comes from an optimization framework.
- Now assume that
  $\boldsymbol{Y} = \mu\mathbf{1} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma)$
- With no nugget, let $\Sigma = \sigma^2 H(\phi)$, where $(H(\phi))_{ij} = \rho(\phi; d_{ij})$, where d_{ij} is the distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$.
- A nugget can be included by modifying Σ to be $\Sigma = \sigma^2 H(\phi) + \tau^2 I$

Minimizing Mean-Square Prediction Error

- **Goal:** find $h(\boldsymbol{y})$ that minimizes $E[(\boldsymbol{Y}(\boldsymbol{s_0}) - h(\boldsymbol{y}))^2|\boldsymbol{y}]$
- $E[(\boldsymbol{Y}(\boldsymbol{s_0}) - h(\boldsymbol{y}))^2|\boldsymbol{y}]$
- $= E[(\boldsymbol{Y}(\boldsymbol{s_0}) - h(\boldsymbol{y}) \pm E[(\boldsymbol{Y}(\boldsymbol{s_0}|\boldsymbol{y}))])^2|\boldsymbol{y}]$
- $= E\{(\boldsymbol{Y}(\boldsymbol{s_0}) - E[(\boldsymbol{Y}(\boldsymbol{s_0})|\boldsymbol{y}])^2|\boldsymbol{y}\} + \{E[(\boldsymbol{Y}(\boldsymbol{s_0})|y] - h(\boldsymbol{y})\}^2$

Minimizing Mean-Square Prediction Error: Part 2

- As $\{E[(\boldsymbol{Y}(\boldsymbol{s_0})|y] - h(\boldsymbol{y})\}^2 \geq 0$
- we have $E[(\boldsymbol{Y}(\boldsymbol{s_0}) - h(\boldsymbol{y}))^2|\boldsymbol{y}] \geq E\{(\boldsymbol{Y}(\boldsymbol{s_0}) - E[(\boldsymbol{Y}(\boldsymbol{s_0})|\boldsymbol{y}])^2|\boldsymbol{y}\}$
- Hence to minimize $E[(\boldsymbol{Y}(\boldsymbol{s_0}) - h(\boldsymbol{y}))^2|\boldsymbol{y}]$, we set ...
- $h(\boldsymbol{y}) = E[(\boldsymbol{Y}(\boldsymbol{s_0})|\boldsymbol{y}]$
- Hence, $h(\boldsymbol{y})$ that minimizes the error is the conditional expectation of $\boldsymbol{Y}(\boldsymbol{s_0})$
- Note this is also the *posterior mean* of $\boldsymbol{Y}(\boldsymbol{s_0})$

Multivariate Normal Theory

- For consider partioning a multivariate normal distribution into two parts
  $$\begin{pmatrix} \boldsymbol{Y_1} \\ \boldsymbol{Y_2} \end{pmatrix} = N\left(\begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}\right),$$
  where $\Omega_{12} = \Omega_{21}^T$

Conditional Multivariate Normal Theory

- The conditional distribution, $p(\boldsymbol{Y_1}|\boldsymbol{Y_2})$ is normal with:
- $E[\boldsymbol{Y_1}|\boldsymbol{Y_2}] = \boldsymbol{\mu_1} + \Omega_{12}\Omega_{22}^{-1}(\boldsymbol{Y_2} - \mu_2)$
- $Var[\boldsymbol{Y_1}|\boldsymbol{Y_2}] = \Omega_{11} - \Omega_{12}\Omega_{[}22]^{-1}\Omega_{21}$
- Thus with $\boldsymbol{Y_1} = Y(\boldsymbol{s_0})$ and $\boldsymbol{Y_2} = \boldsymbol{y}$ \
  $\Omega_{11} = \sigma^2 + \tau^2, \quad \Omega_{12} = (\sigma^2 p(\phi; d_{01})), \ldots, p(\phi; d_{0n}))), \quad \Sigma_{22} = \sigma^2 H(\phi) + \tau^2$

Universal Kriging

- When covariate information is available for inclusion in the analysis, this is often referred to as *universal kriging*
- Now we have
  $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma)$
- The conditional distributions are very similar to what we have derived above, watch for HW question.
- In each case, kriging or universal kriging, it is still necessary to estimate the following parameters: σ^2, τ^2, ϕ, and μ or β.
- This can be done with least-squares methods or in a Bayesian framework.

Gaussian Process Exercise:

Overview

- Similar to the previous exercise, we will simulate data from a 1D process and make predictions at unobserved locations.
- In this situation, please plot the mean of the distribution as well as some uncertainty metric.
- You do not need to estimate σ^2, τ^2, and ϕ but can use the known values in the R code.

Data Sampling