Lecture 5
Prediction and Kriging

Dennis Sun
Stanford University
Stats 253

July 1, 2015

1. Prediction

2. Kriging

1. Prediction

2. Kriging

How do you predict?

- In the usual linear regression model where $\mathrm{Var}[\boldsymbol{\epsilon}|X] = \sigma^2 I$, how do we predict y_0 for a new set of covariates?
- Everybody "knows" that the answer is $x_0^T \hat{\boldsymbol{\beta}}^{OLS}$.
- So how do we predict y_0 in the correlated model where $\mathrm{Var}[\boldsymbol{\epsilon}|X] = \Sigma$? Is it just $x_0^T \hat{\boldsymbol{\beta}}^{GLS}$?
- No! This is why it's important to think carefully about optimality of estimators.

Best Linear Unbiased Prediction

Remember that the model for the data is

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \mathrm{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}.$$

The same model holds at the point we are trying to predict:

$$y_0 = \mathbf{x}_0^T\boldsymbol{\beta} + \epsilon_0.$$

Let's try to find the **best linear unbiased predictor**. That is, we would like a predictor of the form $\hat{y}_0 = \mathbf{w}^T\mathbf{y}$ satisfying $\mathrm{E}[\hat{y}_0] = \mathrm{E}[y_0]$. This means that $\mathbf{w}^T X\boldsymbol{\beta} = \mathbf{x}_0^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$.

Now we can write down an optimization problem:

$$\underset{\mathbf{w}}{\text{minimize}}\ \mathrm{E}(y_0 - \mathbf{w}^T\mathbf{y})^2 \ \text{ subject to }\ \mathbf{w}^T X = \mathbf{x}_0^T$$

Solving the Optimization Problem

$$\underset{\mathbf{w}}{\text{minimize}}\ \mathrm{E}(y_0 - \mathbf{w}^T\mathbf{y})^2 \ \text{ subject to } \ \mathbf{w}^T X = \mathbf{x}_0^T.$$

Let's first rewrite the objective function by adding and subtracting $\mathrm{E}[y_0] = \mathbf{x}_0^T\boldsymbol{\beta}$ and $\mathrm{E}[\mathbf{w}^T\mathbf{y}] = \mathbf{w}^T X\boldsymbol{\beta}$:

$$\begin{aligned}
\mathrm{E}(y_0 - \mathbf{w}^T\mathbf{y})^2 &= \mathrm{E}(\underbrace{y_0 - \mathbf{x}_0^T\boldsymbol{\beta}}_{\epsilon_0} + \underbrace{\mathbf{x}_0^T\boldsymbol{\beta} - \mathbf{w}^T X\boldsymbol{\beta}}_{0} - \mathbf{w}^T(\underbrace{\mathbf{y} - X\boldsymbol{\beta}}_{\boldsymbol{\epsilon}}))^2 \\
&= \mathrm{E}(\epsilon_0 - \mathbf{w}^T\boldsymbol{\epsilon})^2
\end{aligned}$$

So our optimization problem becomes

$$\underset{\mathbf{w}}{\text{minimize}}\ \mathrm{E}(\epsilon_0 - \mathbf{w}^T\boldsymbol{\epsilon})^2 \ \text{ subject to } \ \mathbf{w}^T X = \mathbf{x}_0^T.$$

Solve by Lagrange multipliers! The Lagrangian is:

$$\mathrm{E}(\epsilon_0 - \mathbf{w}^T\boldsymbol{\epsilon})^2 + (\mathbf{x}_0^T - \mathbf{w}^T X)\boldsymbol{\lambda}.$$

Solving the Optimization Problem

$$\mathrm{E}(\epsilon_0 - \mathbf{w}^T \boldsymbol{\epsilon})^2 + (\mathbf{x}_0^T - \mathbf{w}^T X)\boldsymbol{\lambda}.$$

If errors are uncorrelated, then this is

$$\mathrm{E}[\epsilon_0^2] + \mathrm{E}(\mathbf{w}^T \boldsymbol{\epsilon})^2 + (\mathbf{x}_0^T - \mathbf{w}^T X)\boldsymbol{\lambda} = \sigma^2 + \sigma^2 \mathbf{w}^T \mathbf{w} + (\mathbf{x}_0^T - \mathbf{w}^T X)\boldsymbol{\lambda}.$$

Setting the derivatives with respect to $\mathbf{w}$ and $\boldsymbol{\lambda}$ equal to zero, we obtain the first-order conditions:

$$2\sigma^2 \mathbf{w} = X\boldsymbol{\lambda} \qquad\qquad X^T \mathbf{w} = \mathbf{x}_0.$$

Multiply the first equation by X^T. Then, by the second equation, we can replace $X^T\mathbf{w}$ by $\mathbf{x}_0$ to obtain $2\sigma^2\mathbf{x}_0 = X^T X \boldsymbol{\lambda}$, so the Lagrange multiplier is

$$\boldsymbol{\lambda} = 2\sigma^2 (X^T X)^{-1} \mathbf{x}_0.$$

Substituting this into the first equation, we obtain

$$\boxed{\mathbf{w} = X(X^T X)^{-1} \mathbf{x}_0}.$$

Does the solution make sense?

$$\mathbf{w} = X(X^T X)^{-1}\mathbf{x}_0.$$

This is correct because it says that when the errors are uncorrelated, the optimal predictor of y_0 is

$$\hat{y}_0 = \mathbf{w}^T\mathbf{y} = \mathbf{x}_0^T(X^T X)^{-1}X^T\mathbf{y} = \mathbf{x}_0^T\hat{\boldsymbol{\beta}}^{OLS}.$$

Let's try to do the same calculation when the errors are correlated. Call the covariances:

$$\begin{aligned} \Sigma_{00} &= \mathrm{Var}[\epsilon_0] \\ &= \mathrm{E}[\epsilon_0^2] \end{aligned} \qquad \begin{aligned} \Sigma_{01} &= \mathrm{Cov}[\epsilon_0, \boldsymbol{\epsilon}] \\ &= \mathrm{E}[\epsilon_0\boldsymbol{\epsilon}] \end{aligned} \qquad \begin{aligned} \Sigma_{11} &= \mathrm{Var}[\boldsymbol{\epsilon}] \\ &= \mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \end{aligned}$$

The Correlated Case

The objective we are trying to solve is

$$\mathrm{E}(\epsilon_0 - \mathbf{w}^T\boldsymbol{\epsilon})^2 + (\mathbf{x}_0^T - \mathbf{w}^T X)\boldsymbol{\lambda}.$$

Expanding the expectation, we obtain:

$$\Sigma_{00} - 2\Sigma_{01}\mathbf{w} + \mathbf{w}^T\Sigma_{11}\mathbf{w} + (\mathbf{x}_0^T - \mathbf{w}^T X)\boldsymbol{\lambda}.$$

Setting the derivatives with respect to $\mathbf{w}$ and $\boldsymbol{\lambda}$ equal to zero, we obtain the first-order conditions:

$$2\Sigma_{11}\mathbf{w} - X\boldsymbol{\lambda} = 2\Sigma_{10} \qquad\qquad X^T\mathbf{w} = \mathbf{x}_0$$

To solve for $\mathbf{w}$, we multiply the first equation by $X^T\Sigma_{11}^{-1}$ to obtain

$$2\underbrace{X^T\mathbf{w}}_{\mathbf{x}_0} - X^T\Sigma_{11}^{-1}X\boldsymbol{\lambda} = 2X^T\Sigma_{11}^{-1}\Sigma_{10}$$

Now we can substitute the second equation $X^T\mathbf{w} = \mathbf{x}_0$ into this equation and solve for $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = 2(X^T\Sigma_{11}^{-1}X)^{-1}(\mathbf{x}_0 - X^T\Sigma_{11}^{-1}\Sigma_{10}).$$

The Correlated Case

$$\boldsymbol{\lambda} = 2(X^T\Sigma_{11}^{-1}X)^{-1}(\mathbf{x}_0 - X^T\Sigma_{11}^{-1}\Sigma_{10}).$$

Now substitute this value of $\boldsymbol{\lambda}$ into the original first-order condition $2\Sigma_{11}\mathbf{w} - X\boldsymbol{\lambda} = 2\Sigma_{10}$ to solve for $\mathbf{w}$:

$$\mathbf{w} = \Sigma_{11}^{-1}(\Sigma_{10} + X(X^T\Sigma_{11}^{-1}X)^{-1}(\mathbf{x}_0 - X^T\Sigma_{11}^{-1}\Sigma_{10}))$$

So what is $\hat{y}_0 = \mathbf{w}^T\mathbf{y}$, ultimately?

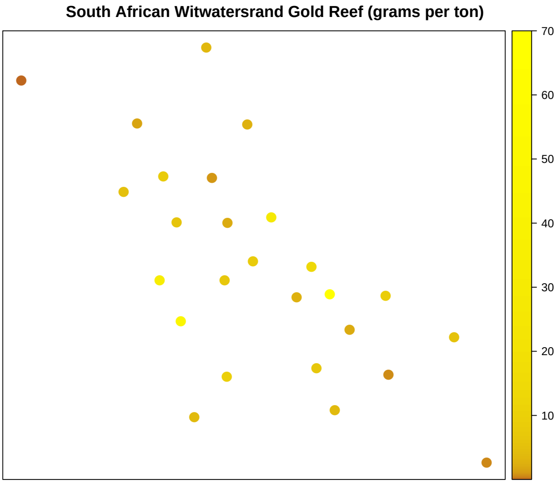
$$\begin{aligned}
\mathbf{w}^T\mathbf{y} &= \Sigma_{01}\Sigma_{11}^{-1}\mathbf{y} + (\mathbf{x}_0 - X^T\Sigma_{11}^{-1}\Sigma_{10})^T \underbrace{(X^T\Sigma_{11}^{-1}X)^{-1}X^T\Sigma_{11}^{-1}\mathbf{y}}_{\hat{\boldsymbol{\beta}}^{GLS}} \\
&= \mathbf{x}_0^T\hat{\boldsymbol{\beta}}^{GLS} + \Sigma_{01}\Sigma_{11}^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}}^{GLS}).
\end{aligned}$$

1. Prediction
2. Kriging

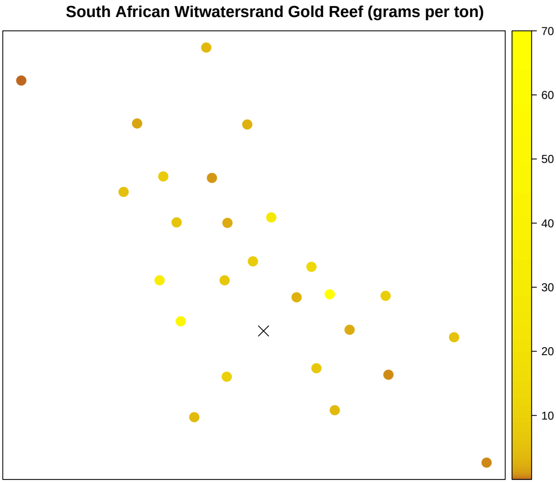
A Brief History of Kriging

- Kriging is named for Danie Krige (1919-2013), a South African mining engineer.
- He was trying to predict gold grades at the Witwatersrand reef complex.
- The prediction method that he used was the one just discussed.
- For these historical reasons, spatial prediction is often called **geostatistics**.

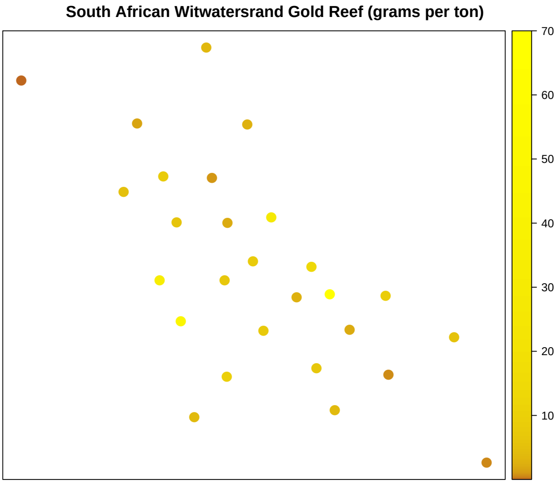
Witwatersrand Gold Data



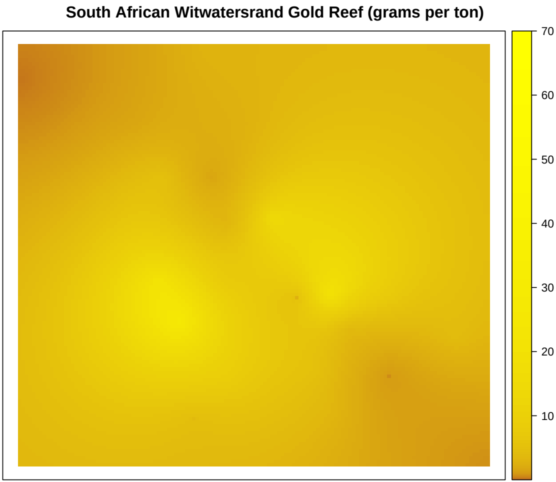
Witwatersrand Gold Data



Witwatersrand Gold Data



Witwatersrand Gold Data



Types of Kriging

Assume a covariance function $\Sigma(\mathbf{s}, \mathbf{s}')$ on the space.

- **Simple kriging:** $y_i = \epsilon_i$.
- **Ordinary kriging:** $y_i = \mu + \epsilon_i$.
- **Universal kriging:** $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$.

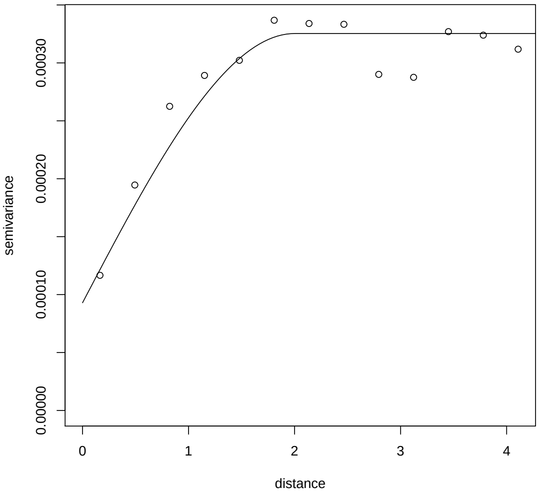
The Variogram

Instead of the covariance, they use the **variogram**.

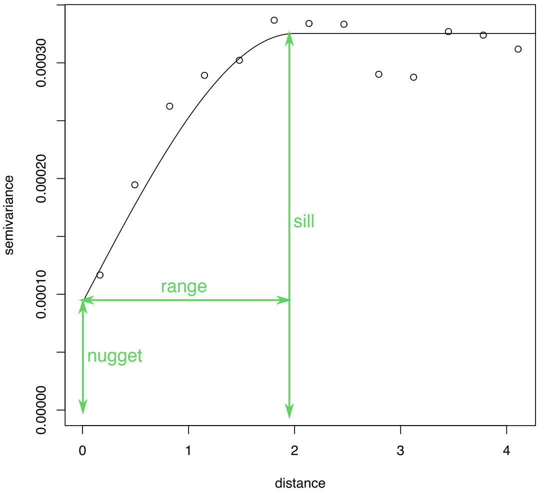
$$2\gamma(\mathbf{s}_i, \mathbf{s}_j) = \mathrm{Var}[y_i - y_j]$$

In the case of ordinary kriging, estimating the variogram doesn't require an estimate of the mean, unlike estimating the covariance.

The Variogram



The Variogram



The Variogram