Lecture 4: Multiple Linear Regression

Math 586

Trends/ Multiple Regression

"Response" = function of space, time etc.
e.g. Water level in space, Ore grade vs. position, porosity in 3 dimensions etc.

Let $\mathbf{x} = (x_1, x_2)$ - location. $y(\mathbf{x})$ = response (variable of interest).

Data: $(x_{1i}, x_{2i}), y_i, \quad i = 1, ..., N$

$$y_i = \texttt{Model mean} + \texttt{error}$$
$$\texttt{Model mean} = m(\mathbf{x}_i; \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ = parameter vector (typically unknown).
Goal: estimate "trend" $m(\mathbf{x}; \boldsymbol{\beta})$ or $\boldsymbol{\beta}$ based on observations.

Additive model

$$y_i = m(\mathbf{x}_i; \boldsymbol{\beta}) + \varepsilon_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$
$$i = 1, ..., N$$

That is, trend $m(\mathbf{x}; \boldsymbol{\beta})$ is a linear function.
"Errors" (residuals) ε_i are random:

- $\mathbb{E}(\varepsilon_i) = 0$
- may or may not be independent
- $Var(\varepsilon_i) = \sigma^2$ or maybe $Var(\varepsilon_i) = \sigma_i^2$.

Possible extension: $m(\mathbf{x}; \boldsymbol{\beta})$ is only linear in $\boldsymbol{\beta}$, i.e.

$$m(\mathbf{x}; \boldsymbol{\beta}) = \sum_{j=0}^{p} \beta_j f_j(\mathbf{x})$$

where f_j's are known.

Analog of Mean Square Error: Least Squares (OLS)

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} [y_i - m(\mathbf{x}_i; \boldsymbol{\beta})]^2$$

or Weighted Least Squares (WLS)

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} w_i [y_i - m(\mathbf{x}_i; \boldsymbol{\beta})]^2$$

with w_i's given set of weights. (Weights may be used to cope with non-stationarity of errors.)

Examples

1. $m_0(\mathbf{x}; \boldsymbol{\beta}) = \beta_0$ constant, $f_0 \equiv 1$ models intercept
2. $m_1(\mathbf{x}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ plane. Here $f_1((x_1, x_2)) = x_1$ and $f_2((x_1, x_2)) = x_2$
3. $m_2(\mathbf{x}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2$ quadratic surface
4. $m^\star(\mathbf{x}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \sin x_1 + \beta_4 \cos x_1$ periodic (e.g. to capture seasonal dependencies; note that an oscillatory function with a known period but unknown phase is a combination of sines and cosines.)

Vector-Matrix Form

$\mathbf{y} = (y_1, ..., y_N)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_N)'$ are column vectors.
The $N \times (p+1)$ matrix $(X_{ij}) = f_j(\mathbf{x}_i)$

$$\mathbf{X} = \begin{bmatrix} f_0(\mathbf{x}_1), & f_1(\mathbf{x}_1), & ... & f_p(\mathbf{x}_1) \\ f_0(\mathbf{x}_2), & f_1(\mathbf{x}_2), & ... & f_p(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(\mathbf{x}_N), & f_1(\mathbf{x}_N), & ... & f_p(\mathbf{x}_N) \end{bmatrix}$$

is called *design matrix.*

Model becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Least Squares minimizes

$$\texttt{SSE} = \sum_{i=1}^{N} [y_i - m(\mathbf{x}; \boldsymbol{\beta})]^2 = \qquad (1)$$

$$= \sum_{i=1}^{N} [y_i - \sum_{j=0}^{p} \beta_j f_j(\mathbf{x}_i)]^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Let

$$\mathbf{S} = \mathbf{X}'\mathbf{X}$$

it's a $p \times p$ symmetric matrix.

Assume full rank: $|\mathbf{S}| \neq 0$, therefore $\mathbf{S}^{-1}$ exists. Also, $\mathbf{S}$ is positive-definite since for any p-vector $\mathbf{b}$

$$\mathbf{b}'\mathbf{S}\mathbf{b} = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}\mathbf{b})'\mathbf{X}\mathbf{b} = \sum_{i=1}^{N} [(\mathbf{X}\mathbf{b})_i]^2 \geq 0$$

Let $\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{X}'\mathbf{y}$, then

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\,\mathbf{S}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

Thus, the min is attained when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

<Another approach to minimizing (1) would be to set partials with respect to β_j's equal to 0 and solve the resulting system. Results are equivalent to above.>

$$\boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

Example: planar case.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \qquad i = 1, ..., n$$

then

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

and

$$\mathbf{S} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} N & \sum_{i=1}^N x_{1i} & \sum_{i=1}^N x_{2i} \\ \sum_{i=1}^N x_{1i} & \sum_{i=1}^N x_{1i}^2 & \sum_{i=1}^N x_{1i}\, x_{2i} \\ \sum_{i=1}^N x_{2i} & \sum_{i=1}^N x_{1i}\, x_{2i} & \sum_{i=1}^N x_{2i}^2 \end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_{1i}\, y_i \\ \sum_{i=1}^N x_{2i}\, y_i \end{pmatrix}$$

Scaling problems: $\mathbf{S}$ may have a high condition number, computationally unstable. We can reparametrize: move all f_j to average 0 and rescale, that is, let

$$f_j^{**}(\mathbf{x}_i) = [f_j(\mathbf{x}_i) - \overline{f}_j]/s_j, \qquad j \geq 1,$$

$$f_0^{**} \equiv 1 \equiv f_0$$

This also has the advantage of making matrix $\mathbf{S}$ diagonal.

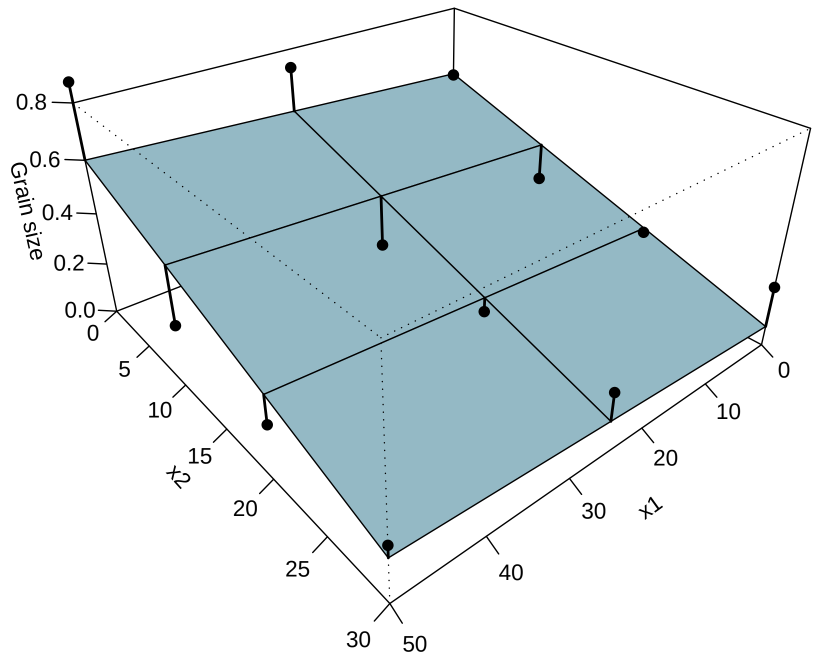

Figure 1: Linear fit

Example

Median grain size y: spatial data over a regular grid (Krumbein and Graybill, 1965)

	1	2	3	4	5	6	7	8	9	10	11	12
x_1	0	0	0	0	25	25	25	25	50	50	50	50
x_2	0	10	20	30	0	10	20	30	0	10	20	30
y	0.51	0.22	0.205	0.234	0.73	0.214	0.212	0.225	0.87	0.234	0.202	0.204

Initially, $\mathbf{S} = \mathbf{X}'\mathbf{X}$ is not diagonal. Let

$$x_1^* = (x_1 - \overline{x}_1)/25$$

$$x_2^* = (x_2 - \overline{x}_2)/15$$

Here, $\mathbf{S}^* = (\mathbf{X}^*)'\mathbf{X}^* = diag\{12, 8, 6.667\}$ and solving

$$\mathbf{S}^*\hat{\boldsymbol{\alpha}} = (\mathbf{X}^*)'\mathbf{y} = (4.06, 0.341, -1.463)'$$

yields

$$\hat{\boldsymbol{\alpha}} = (0.3383, 0.04265, -0.2195)'$$

and the fitted model

$$\hat{y}_i = 0.3383 + 0.04265\left(\frac{x_1 - 25}{25}\right) - 0.2195\left(\frac{x_2 - 15}{15}\right) = 0.5152 + 0.0017x_1 - 0.0146x_2$$

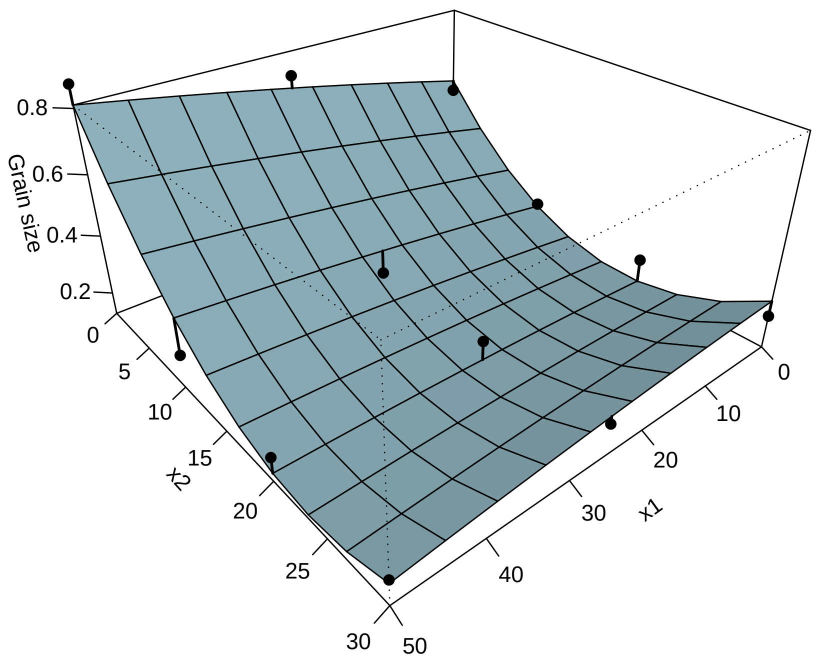

Figure 2: Quadratic fit $y = 0.547 + 0.0061x_1 - 0.046x_2 - 0.000016x_1^2 + 0.0012x_2^2 - 0.00024x_1x_2$

Measure of fit

Sums of squares:

$$\texttt{SS for Errors (SSE)} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$\texttt{Total(SST)} = \sum_{i=1}^{N} (y_i - \overline{y})^2$$

$$\texttt{SS(Regression)} = \texttt{SST} - \texttt{SSE}$$

Multiple squared correlation coefficient ("coefficient of determination") is

$$R^2 = SS(Regression)/SST = 1 - SSE/SST$$

Becomes the usual r^2 when using one predictor x.
E.g. $R^2 = 0.56$ for linear model fitted above and $R^2 = 0.93$ for quadratic model.