

The OLS Estimator (1)

STA 211: The Mathematics of Regression

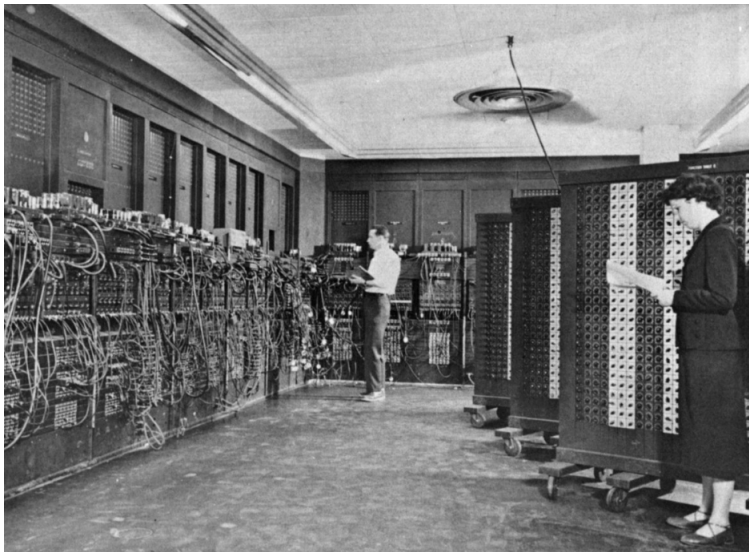
Yue Jiang

January 17, 2023

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Why care about linear models?



Review: vector operations

Let \mathbf{x} be a k -vector (this is to say, with dimensions $k \times 1$):

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

Review: vector operations

Suppose we have some function of \mathbf{x} , $f(\mathbf{x})$. Then the gradient ∇f (with respect to \mathbf{x}) is the k -vector of partial derivatives:

$$\nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_k} \end{bmatrix}$$

Review: vector operations

Similarly, the Hessian $\nabla^2 f$ is the $k \times k$ matrix of second partial derivatives:

$$\nabla_x^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_k} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_k \partial x_1} & \frac{\partial^2 f}{\partial x_k \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_k^2} \end{bmatrix}$$

Review: vector operations

The gradient has some convenient properties. For instance, suppose we wish to differentiate the linear form $\mathbf{x}^T \mathbf{z}$, where \mathbf{z} is also a k -vector (and not a function of \mathbf{x}):

$$\mathbf{x}^T \mathbf{z} = \begin{bmatrix} x_1 & \cdots & x_k \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix} = x_1 z_1 + \cdots + x_k z_k$$

Review: vector operations

The gradient has some convenient properties. For instance, suppose we wish to differentiate the linear form $\mathbf{x}^T \mathbf{z}$, where \mathbf{z} is also a k -vector (and not a function of \mathbf{x}):

$$\mathbf{x}^T \mathbf{z} = \begin{bmatrix} x_1 & \cdots & x_k \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix} = x_1 z_1 + \cdots + x_k z_k$$

Then the gradient with respect to \mathbf{x} is

$$\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{z} = \begin{bmatrix} \frac{\partial \mathbf{x}^T \mathbf{z}}{\partial x_1} \\ \vdots \\ \frac{\partial \mathbf{x}^T \mathbf{z}}{\partial x_k} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} (x_1 z_1 + \cdots + x_k z_k) \\ \vdots \\ \frac{\partial}{\partial x_k} (x_1 z_1 + \cdots + x_k z_k) \end{bmatrix} = \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix} = \mathbf{z}$$

Review: vector operations

Similarly, if \mathbf{A} is a $k \times k$ matrix that is not a function of \mathbf{x} , then the gradient of the quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$ (again, with respect to \mathbf{x}) is

$$\nabla \left(\mathbf{x}^T \mathbf{A} \mathbf{x} \right) = \left(\mathbf{A} + \mathbf{A}^T \right) \mathbf{x},$$

which is $2\mathbf{A}\mathbf{x}$ if \mathbf{A} is symmetric.

The linear model in matrix form

We have n observations of a response variable Y and each predictor X_1, X_2, \dots, X_p . As well, each observation has some unobserved error ϵ (which may have certain properties, which we'll talk about later this semester).

The linear model in matrix form

We have n observations of a response variable Y and each predictor X_1, X_2, \dots, X_p . As well, each observation has some unobserved error ϵ (which may have certain properties, which we'll talk about later this semester).

We wish to fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

The linear model in matrix form

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\boldsymbol{\epsilon}}$$

- What are the dimensions of each of the pieces above?

The mean squared error

Note the error term, ϵ ; suppose we are interested in the mean squared error (MSE), given by

$$\frac{1}{n} \epsilon^T \epsilon$$

- ▶ What is the dimension of the MSE?
- ▶ How would you express the MSE in terms of \mathbf{y} , \mathbf{X} , and β ?

Minimizing the mean squared error

$$\frac{1}{n} \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = \frac{1}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Minimizing the mean squared error

$$\begin{aligned}\frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} &= \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{1}{n}(\mathbf{y}^T - \boldsymbol{\beta}^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

Minimizing the mean squared error

$$\begin{aligned}\frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} &= \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{1}{n}(\mathbf{y}^T - \boldsymbol{\beta}^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{1}{n}(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})\end{aligned}$$

Minimizing the mean squared error

$$\begin{aligned}\frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} &= \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{1}{n}(\mathbf{y}^T - \boldsymbol{\beta}^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{1}{n}(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})\end{aligned}$$

Suppose we wanted to minimize the MSE with respect to $\boldsymbol{\beta}$.

- What is the gradient of the MSE with respect to $\boldsymbol{\beta}$?

The mean squared error

$$\nabla_{\beta} \frac{1}{n} \epsilon^T \epsilon = \frac{1}{n} \left(\nabla \mathbf{y}^T \mathbf{y} - 2 \nabla \beta^T \mathbf{X}^T \mathbf{y} + \nabla \beta^T \mathbf{X}^T \mathbf{X} \beta \right)$$

The mean squared error

$$\begin{aligned}\nabla_{\beta} \frac{1}{n} \epsilon^T \epsilon &= \frac{1}{n} \left(\nabla \mathbf{y}^T \mathbf{y} - 2 \nabla \beta^T \mathbf{X}^T \mathbf{y} + \nabla \beta^T \mathbf{X}^T \mathbf{X} \beta \right) \\ &= \frac{1}{n} \left(0 - 2 \mathbf{X}^T \mathbf{y} + 2 \mathbf{X}^T \mathbf{X} \beta \right)\end{aligned}$$

The mean squared error

$$\begin{aligned}\nabla_{\beta} \frac{1}{n} \epsilon^T \epsilon &= \frac{1}{n} \left(\nabla \mathbf{y}^T \mathbf{y} - 2 \nabla \beta^T \mathbf{X}^T \mathbf{y} + \nabla \beta^T \mathbf{X}^T \mathbf{X} \beta \right) \\ &= \frac{1}{n} \left(0 - 2 \mathbf{X}^T \mathbf{y} + 2 \mathbf{X}^T \mathbf{X} \beta \right) \\ &\propto \mathbf{X}^T \mathbf{X} \beta - \mathbf{X}^T \mathbf{y}\end{aligned}$$

The mean squared error

$$\begin{aligned}\nabla_{\beta} \frac{1}{n} \epsilon^T \epsilon &= \frac{1}{n} \left(\nabla \mathbf{y}^T \mathbf{y} - 2 \nabla \beta^T \mathbf{X}^T \mathbf{y} + \nabla \beta^T \mathbf{X}^T \mathbf{X} \beta \right) \\ &= \frac{1}{n} \left(0 - 2 \mathbf{X}^T \mathbf{y} + 2 \mathbf{X}^T \mathbf{X} \beta \right) \\ &\propto \mathbf{X}^T \mathbf{X} \beta - \mathbf{X}^T \mathbf{y}\end{aligned}$$

If a solution $\hat{\beta}$ exists, it occurs when this quantity is zero:

$$\mathbf{0} \stackrel{\text{set}}{=} \mathbf{X}^T \mathbf{X} \hat{\beta} - \mathbf{X}^T \mathbf{y}$$

- What is the (candidate) solution?

Minimizing the mean squared error

$$\mathbf{0} \stackrel{\text{set}}{=} \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} - \mathbf{X}^T \mathbf{y}$$

Minimizing the mean squared error

$$\mathbf{0} \stackrel{\text{set}}{=} \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} - \mathbf{X}^T \mathbf{y}$$
$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}$$

Minimizing the mean squared error

$$\mathbf{0} \stackrel{\text{set}}{=} \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} - \mathbf{X}^T \mathbf{y}$$

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}$$

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

- Is this proposed solution a minimum?

The ordinary least squares estimate

$$\nabla_{\beta} \left(\mathbf{X}^T \mathbf{X} \beta - \mathbf{X}^T \mathbf{y} \right) = \mathbf{X}^T \mathbf{X}$$

If \mathbf{X} has full column rank, then we have indeed found a *minimizing* solution, since for any non-zero \mathbf{z} $\mathbf{z}^T (\mathbf{X}^T \mathbf{X}) \mathbf{z} > 0$ (that is, the Hessian is positive definite). In fact, this is the unique global solution.

Fitted values

Suppose we use our OLS estimates $\hat{\beta}$ to try and predict \mathbf{y} from \mathbf{X} : $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$. This gives us a vector of fitted values:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} \\ &= \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\mathbf{H}}\mathbf{y}\end{aligned}$$

where \mathbf{H} is known as the "hat matrix" (it puts the hat on the \mathbf{y}). Note that this is only a function of \mathbf{X} , NOT of \mathbf{y} . From this matrix, we can see how our predictions change as \mathbf{X} varies.

Residuals

The *difference* between the observed outcome \mathbf{y} and the predicted outcome $\hat{\mathbf{y}}$ is known as the residual:

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

We can equivalently use the hat matrix here:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

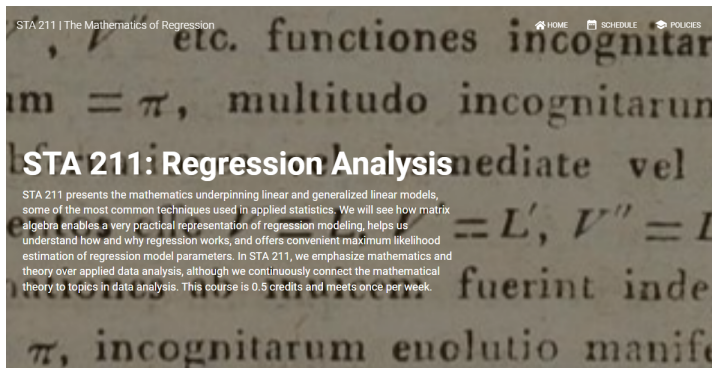
Homework 1: Due Jan. 24

1. Show that $\nabla (\mathbf{x}^T \mathbf{A} \mathbf{x}) = 2\mathbf{A} \mathbf{x}$ for a k -vector \mathbf{x} and symmetric $k \times k$ matrix \mathbf{A} .
2. Show that \mathbf{H} and $\mathbf{I} - \mathbf{H}$ are symmetric (i.e., $\mathbf{H}^T = \mathbf{H}$, etc.) and idempotent (i.e., $\mathbf{H}^2 = \mathbf{H}$, etc.).
3. Instead of the MSE, suppose we wanted to minimize the following function with respect to β , for some scalar $\lambda > 0$ (assuming full rank \mathbf{X}):

$$(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta.$$

Is there an analytical solution to this objective function? If so, provide the solution and demonstrate that it indeed minimizes the objective function. Otherwise, explain why not.

Some housekeeping

The image is a screenshot of the STA 211 course website. The background is a dark, textured image with faint Latin text, including words like "funciones incognitar", "multitudo incognitarum", "mediate vel", "fuerint inde", and "incognitarum evolutio manife". At the top left, the text "STA 211 | The Mathematics of Regression" is visible. At the top right, there are navigation links: "HOME", "SCHEDULE", and "POLICIES". The main heading "STA 211: Regression Analysis" is prominently displayed in the center. Below it, a paragraph describes the course content, mentioning linear and generalized linear models, matrix algebra, regression modeling, maximum likelihood estimation, and the emphasis on mathematics and theory over applied data analysis. It also states the course is 0.5 credits and meets once per week.

STA 211 | The Mathematics of Regression

HOME SCHEDULE POLICIES

STA 211: Regression Analysis

STA 211 presents the mathematics underpinning linear and generalized linear models, some of the most common techniques used in applied statistics. We will see how matrix algebra enables a very practical representation of regression modeling, helps us understand how and why regression works, and offers convenient maximum likelihood estimation of regression model parameters. In STA 211, we emphasize mathematics and theory over applied data analysis, although we continuously connect the mathematical theory to topics in data analysis. This course is 0.5 credits and meets once per week.