Lecture Notes: Multiple Linear Regression

Regression Analysis - Week 3

March 25 / April 1

Administrative Notes:

- **Recording:** Please note that the lecture recording started successfully after a brief initial issue. The camera now captures the entire board.
- **Questions:** Please feel free to ask questions at any point! Your understanding is the priority. Don't hesitate.
- **Midterm Exam:** We are planning the midterm exam. Tentatively, it will likely be scheduled around Week 8 of the course. We will confirm the exact date and details soon after coordinating with Yaniv. A formal announcement with plenty of notice will be made.
- **Office Hours:** Students needing to discuss specific matters (like the question raised about a project at the Central Bureau of Statistics) should please coordinate via email to schedule a meeting, as immediate post-lecture time may not always be available.

1 Introduction: From Simple to Multiple Regression

Good morning, everyone. Last week, we explored Simple Linear Regression, where we aimed to model a response variable Y using a single explanatory variable X. Our model took the form $Y \approx \beta_0 + \beta_1 X$. Today, we generalize this framework to **Multiple Linear Regression**, where we allow for multiple explanatory variables.

1.1 The Setup

Suppose we have n observations. For each observation i (where $i = 1, \ldots, n$), we measure:

- p explanatory variables (also called predictors, features, or independent variables): $x_{i1}, x_{i2}, \ldots, x_{ip}$.
- A response variable (or dependent variable): y_i.

We can collect the predictors for the i-th observation into a vector:

$$\mathbf{x}_i^{\text{raw}} = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \in \mathbb{R}^p$$

This is sometimes called the *feature vector* for observation i.

Our goal is to model the relationship between the response y_i and the predictors $x_{i1}, \ldots, x_{ip}$ using a linear equation. Generalizing the simple linear model, we propose the model:

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Here, β_0 is the intercept term, and $\beta_1, \ldots, \beta_p$ are the coefficients (or slopes) associated with each predictor. There are a total of $p + 1$ coefficients to determine.

1.2 Augmented Notation

It's convenient to handle the intercept term β_0 in the same way as the other coefficients. We can achieve this by introducing an artificial predictor x_{i0} that always takes the value 1 for every observation i. Then, the model equation can be written as a sum:

$$y_i \approx \sum_{j=0}^{p} \beta_j x_{ij} \quad \text{where } x_{i0} = 1$$

We can define the **augmented predictor vector** for observation i as:

$$\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T = (1, x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^{p+1}$$

And the vector of coefficients as:

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$$

With this notation, the model equation becomes a simple inner product:

$$y_i \approx \mathbf{x}_i^T \boldsymbol{\beta}$$

1.3 Predicted Values

Once we have estimated the coefficients (let's denote the estimated vector as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$), we can calculate the **predicted value** (or fitted value) for observation i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} = \sum_{j=0}^{p} \hat{\beta}_j x_{ij} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$$

This is the value our fitted model predicts for y_i given the predictor values $\mathbf{x}_i$.

2 The Least Squares Criterion

How do we find the "best" estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$? As in simple linear regression, we use the **method of least squares**. We aim to find the coefficient vector $\boldsymbol{\beta}$ that minimizes the sum of the squared differences between the observed responses y_i and the predicted responses $\hat{y}_i = \mathbf{x}_i^T \boldsymbol{\beta}$.

Definition 2.1 (Least Squares Objective Function)**.** The objective function $Q(\boldsymbol{\beta})$ is the sum of squared residuals (errors):

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 \\ &= \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \end{aligned}$$

Note that Q is a function of the $p+1$ candidate coefficients $\beta_0, \ldots, \beta_p$.

Remark 2.2. At this stage, we haven't assumed any statistical model for the data (e.g., error distributions). We are simply finding the linear combination of predictors that best fits the data in the least-squares sense. The β_j's in $Q(\boldsymbol{\beta})$ are variables we optimize over; the resulting minimizers $\hat{\beta}_j$ will be *estimators* (functions of the data). We reserve the term "parameter" for when we introduce a statistical model later.

The least squares estimator $\hat{\boldsymbol{\beta}}$ is the vector that minimizes this objective function:

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} Q(\boldsymbol{\beta})$$

3 Finding the Least Squares Estimator: The Normal Equations

To find the minimum of $Q(\boldsymbol{\beta})$, we use calculus. Since Q is a differentiable function of the $p+1$ variables $\beta_0, \ldots, \beta_p$, a necessary condition for a minimum is that all partial derivatives are zero. We need to solve the system of $p+1$ equations:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_r} = 0 \quad \text{for } r = 0, 1, \ldots, p$$

Let's compute the partial derivative with respect to a specific coefficient β_r:

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_r} &= \frac{\partial}{\partial \beta_r} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_r} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 \\
&= \sum_{i=1}^{n} 2 \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right) \cdot \frac{\partial}{\partial \beta_r} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right) \qquad \text{(Chain Rule)}
\end{aligned}$$

Now, the inner derivative is:

$$\frac{\partial}{\partial \beta_r} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right) = 0 - \sum_{j=0}^{p} \frac{\partial(\beta_j x_{ij})}{\partial \beta_r}$$

The derivative $\frac{\partial(\beta_j x_{ij})}{\partial \beta_r}$ is x_{ij} if $j = r$, and 0 if $j \neq r$. So the inner derivative is simply $-x_{ir}$. Plugging this back in:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{i=1}^{n} 2 \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right) (-x_{ir})$$

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_r} = -2 \sum_{i=1}^{n} \left(y_i x_{ir} - \sum_{j=0}^{p} \beta_j x_{ij} x_{ir} \right)$$

Setting this derivative to zero for all $r = 0, \ldots, p$:

$$\sum_{i=1}^{n} \left(y_i x_{ir} - \sum_{j=0}^{p} \beta_j x_{ij} x_{ir} \right) = 0$$

Rearranging the terms, we move the sum involving β_j to the other side:

$$\sum_{i=1}^{n} \sum_{j=0}^{p} \beta_j x_{ij} x_{ir} = \sum_{i=1}^{n} y_i x_{ir}$$

We can switch the order of summation (since they are finite sums) and factor out β_j which doesn't depend on i:

$$\sum_{j=0}^{p} \left(\sum_{i=1}^{n} x_{ij} x_{ir} \right) \beta_j = \sum_{i=1}^{n} y_i x_{ir}$$

This gives us a system of $p+1$ linear equations in the $p+1$ unknown coefficients $\beta_0, \ldots, \beta_p$. These are known as the **Normal Equations**.

Definition 3.1 (Normal Equations (Scalar Form))**.** The least squares estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$ must satisfy the following system of $p+1$ linear equations:

$$\sum_{j=0}^{p} \left(\sum_{i=1}^{n} x_{ij} x_{ir} \right) \hat{\beta}_j = \sum_{i=1}^{n} y_i x_{ir} \quad \text{for } r = 0, 1, \ldots, p$$

4 Matrix Formulation of Least Squares

While the scalar form of the normal equations is correct, it becomes cumbersome for larger p. A much more compact and insightful formulation uses matrix algebra.

Notation 4.1 (Model Matrix and Response Vector).

- Let $\mathbf{y}$ be the $n \times 1$ vector of observed responses:

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$$

- Let X be the $n \times (p+1)$ **model matrix** (or design matrix), whose rows are the augmented predictor vectors $\mathbf{x}_i^T$:

$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

- Recall the coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, which is $(p+1) \times 1$.

Notice the structure of X:

- Each *row* corresponds to an observation.
- Each *column* corresponds to a predictor variable (with the first column being all 1s for the intercept). Let $\mathbf{X}_j$ denote the j-th column ($j = 0, \ldots, p$), representing the n values of the j-th predictor across all observations. So $X = [\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_p]$. Note $\mathbf{X}_0$ is a vector of ones.

Now, let's rewrite the vector of predicted values $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$. The i-th predicted value is $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$. This is precisely the i-th element of the matrix-vector product $X\hat{\boldsymbol{\beta}}$.

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

The vector of residuals is $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$.

The objective function $Q(\boldsymbol{\beta})$ is the sum of squared residuals, which is the squared Euclidean norm of the residual vector:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 = \|\mathbf{y} - X\boldsymbol{\beta}\|^2 = (\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta})$$

Now, let's re-derive the normal equations using matrix calculus (or by careful inspection of the scalar form). The system we found was:

$$\sum_{j=0}^{p} \left(\sum_{i=1}^{n} x_{ir} x_{ij} \right) \beta_j = \sum_{i=1}^{n} x_{ir} y_i \quad \text{for } r = 0, 1, \ldots, p$$

Let's examine the terms:

- The term $\sum_{i=1}^{n} x_{ir} x_{ij}$ is the inner product of the r-th column and the j-th column of X. This is exactly the (r, j)-th element of the matrix product $X^T X$. Note that X^T is $(p+1) \times n$ and X is $n \times (p+1)$, so $X^T X$ is a $(p+1) \times (p+1)$ matrix.
- The term $\sum_{i=1}^{n} x_{ir} y_i$ is the inner product of the r-th column of X and the vector $\mathbf{y}$. This is exactly the r-th element of the vector product $X^T \mathbf{y}$. Note that $X^T \mathbf{y}$ is a $(p+1) \times 1$ vector.

Therefore, the r-th equation in the scalar system can be written as:

$$(X^T X\boldsymbol{\beta})_r = (X^T\mathbf{y})_r$$

where $(\cdot)_r$ denotes the r-th element (using 0-based indexing for rows/elements, consistent with our β_r indexing). Since this must hold for all $r = 0, \ldots, p$, the entire system is equivalent to the matrix equation:

Proposition 4.2 (Normal Equations (Matrix Form))**.** *The least squares estimator $\hat{\boldsymbol{\beta}}$ is any solution to the system of linear equations:*

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

This is a remarkably compact representation of the $p+1$ normal equations.

5 The Solution for $\hat{\boldsymbol{\beta}}$ and Conditions for Uniqueness

The normal equations $(X^T X)\hat{\boldsymbol{\beta}} = X^T\mathbf{y}$ form a system of linear equations for the unknown vector $\hat{\boldsymbol{\beta}}$. The matrix $X^T X$ is always square, with dimensions $(p+1) \times (p+1)$.

When does this system have a unique solution? From linear algebra, we know that a system $A\mathbf{x} = \mathbf{b}$ has a unique solution if and only if the matrix A is invertible. In our case, $A = X^T X$.

Theorem 5.1 (Condition for Unique Least Squares Solution)**.** *The matrix $X^T X$ is invertible if and only if the columns of the model matrix X are linearly independent.*

Proof. (Sketch - relies on properties of null spaces) The columns of X are linearly independent iff the only solution to $X\mathbf{c} = \mathbf{0}$ is $\mathbf{c} = \mathbf{0}$. Suppose $X^T X$ is not invertible. Then there exists $\mathbf{c} \neq \mathbf{0}$ such that $X^T X\mathbf{c} = \mathbf{0}$. Multiplying by $\mathbf{c}^T$, we get $\mathbf{c}^T X^T X\mathbf{c} = (X\mathbf{c})^T(X\mathbf{c}) = \|X\mathbf{c}\|^2 = 0$. This implies $X\mathbf{c} = \mathbf{0}$ for $\mathbf{c} \neq \mathbf{0}$, so the columns of X are linearly dependent. Conversely, suppose the columns of X are linearly dependent. Then there exists $\mathbf{c} \neq \mathbf{0}$ such that $X\mathbf{c} = \mathbf{0}$. Multiplying by X^T, we get $X^T X\mathbf{c} = X^T\mathbf{0} = \mathbf{0}$. Since we found a non-zero vector $\mathbf{c}$ in the null space of $X^T X$, the matrix $X^T X$ is not invertible. □

If $X^T X$ is invertible, we can simply multiply both sides of the normal equations by its inverse to find the unique solution:

Corollary 5.2 (Unique Least Squares Estimator)**.** *If the columns of the model matrix X are linearly independent, then the unique least squares estimator $\hat{\boldsymbol{\beta}}$ is given by:*

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

Remark 5.3 (Dimensions Check)**.** Let's check the dimensions: X is $n \times (p+1)$. X^T is $(p+1) \times n$. $\mathbf{y}$ is $n \times 1$. $X^T X$ is $(p+1) \times (p+1)$. $(X^T X)^{-1}$ is $(p+1) \times (p+1)$. $X^T\mathbf{y}$ is $(p+1) \times 1$. So, $(X^T X)^{-1}(X^T\mathbf{y})$ is $((p+1) \times (p+1)) \times ((p+1) \times 1)$, resulting in a $(p+1) \times 1$ vector, which matches the dimension of $\hat{\boldsymbol{\beta}}$.

5.1 When are the Columns of X Linearly Independent?

The matrix X has $p+1$ columns and n rows. For its columns to be linearly independent, the rank of the matrix must be equal to the number of columns, i.e., $\text{rank}(X) = p+1$. However, we also know that the rank of a matrix cannot exceed the minimum of its number of rows and number of columns: $\text{rank}(X) \leq \min(n, p+1)$. Therefore, a necessary condition for the columns of X to be linearly independent is that $\min(n, p+1) \geq p+1$, which implies $n \geq p+1$.

Proposition 5.4. *A necessary condition for the uniqueness of the least squares solution $\hat{\boldsymbol{\beta}}$ is that the number of observations n must be greater than or equal to the number of coefficients $p+1$ (i.e., $n \geq p+1$).*

If $n < p+1$, then $\text{rank}(X) \leq n < p+1$. Since the rank is strictly less than the number of columns, the columns of X *must* be linearly dependent. In this case, $X^T X$ is singular (not invertible), and the normal equations $(X^T X)\hat{\boldsymbol{\beta}} = X^T \mathbf{y}$ will have infinitely many solutions. Each of these solutions minimizes the sum of squares $Q(\boldsymbol{\beta})$, but we don't have a single, unique estimator.

Example 5.5 (The $n < p+1$ Case: Degeneracy)**.** Consider a situation in modern biostatistics or genomics. Suppose we are studying a very rare disease, and we only have data for $n = 1000$ patients. However, for each patient, we measure a vast number of potential predictors (genes, demographic variables, clinical measurements), say $p = 10000$. Here, the number of predictors (plus intercept) is $p + 1 = 10001$, which is much larger than the number of observations $n = 1000$. The model matrix X would be 1000×10001. Such a matrix is sometimes called "fat". Since $n < p+1$, the columns of X must be linearly dependent. The matrix $X^T X$ (which is 10001×10001) will be singular. There is no unique least squares solution $\hat{\boldsymbol{\beta}}$. Intuitively, we have far more parameters to estimate than we have data points to constrain them. It's like trying to solve $a + b = 10$ for a unique value of a – there are infinite solutions (a, b) depending on what we choose for b. The parameters are not identifiable from the data using least squares alone in this setup. This scenario requires different techniques (like regularization, which we might discuss later).

Convention 5.6. For the standard development of multiple linear regression using ordinary least squares, we will generally assume that $n \geq p+1$ and that the columns of X are linearly independent, ensuring $X^T X$ is invertible and $\hat{\boldsymbol{\beta}}$ is unique.

6 Towards a Geometric Interpretation

The algebraic solution $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ is powerful, but a geometric perspective can provide deeper intuition about what least squares is actually doing. To build this perspective, let's recall some fundamental concepts from linear algebra.

6.1 Review of Linear Algebra Concepts

Definition 6.1 (Image of a Matrix (Column Space))**.** Let A be an $n \times m$ matrix. The **image** of A, denoted $\text{Im}(A)$ (or sometimes $Col(A)$ for Column Space), is the set of all possible outputs when A multiplies vectors in $\mathbb{R}^m$:

$$\text{Im}(A) = \{A\mathbf{v} \mid \mathbf{v} \in \mathbb{R}^m\} \subseteq \mathbb{R}^n$$

Proposition 6.2 (Image as Span of Columns)**.** *Let A be an $n \times m$ matrix with columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m \in \mathbb{R}^n$. Then the product $A\mathbf{v}$ for $\mathbf{v} = (v_1, \ldots, v_m)^T$ is a linear combination of the columns of A:*

$$A\mathbf{v} = v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + \cdots + v_m\mathbf{a}_m = \sum_{j=1}^{m} v_j \mathbf{a}_j$$

Consequently, the image of A is the set of all possible linear combinations of its columns, which is the subspace spanned by the columns:

$$\text{Im}(A) = \text{span}\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$$

Proof. The first equality follows directly from the definition of matrix-vector multiplication. The second equality follows because any vector in $\mathrm{Im}(A)$ is of the form $A\mathbf{v}$ for some $\mathbf{v}$, which is a linear combination of the columns. Conversely, any linear combination $\sum v_j \mathbf{a}_j$ can be written as $A\mathbf{v}$ where $\mathbf{v} = (v_1, \ldots, v_m)^T$, so it belongs to $\mathrm{Im}(A)$. $\square$

This is a crucial insight: multiplying a matrix A by all possible vectors $\mathbf{v}$ sweeps out the subspace spanned by the columns of A.

Definition 6.3 (Standard Euclidean Inner Product)**.** For two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, their standard Euclidean inner product (or dot product) is:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v} = \sum_{i=1}^{n} u_i v_i$$

Definition 6.4 (Euclidean Norm)**.** The Euclidean norm (or length) of a vector $\mathbf{v} \in \mathbb{R}^n$ is:

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{\sum_{i=1}^{n} v_i^2}$$

Note that $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$.

Definition 6.5 (Euclidean Distance)**.** The Euclidean distance between two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is the norm of their difference:

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}$$

With these tools, we can rephrase the least squares problem: Find $\hat{\boldsymbol{\beta}}$ such that the vector of predicted values $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ is the vector in the column space of X (i.e., $\mathrm{Im}(X)$) that is closest to the observed response vector $\mathbf{y}$ in terms of Euclidean distance. That is,

$$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2 = \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 = \min_{\tilde{\mathbf{y}} \in \mathrm{Im}(X)} \|\mathbf{y} - \tilde{\mathbf{y}}\|^2$$

We will explore the geometric consequences of this next time.