Lecture 6 - **Multiple Linear Regression**

BIOENG-210 Course Notes
Prof. Gioele La Manno

March 2025

1 Introduction to Multiple Linear Regression

1.1 Extending to Multiple Predictors

While simple linear regression provides a powerful framework for modeling the relationship between two variables, many biological phenomena are influenced by multiple factors simultaneously. For instance, the total protein content of a cell is not determined solely by its size but also by its metabolic rate, cell cycle stage, and various other factors.

Consider a scenario where we are studying protein synthesis in cells. In a simple linear regression model, we might use only the cell size (in μm^3) as our predictor:

$$\text{Protein Content} = \beta_0 + \beta_1 \times \text{Cell Size} + \varepsilon$$

But this model captures only part of the biological story. What if we also measure each cell's metabolic rate, quantified by oxygen consumption? We could incorporate this additional predictor:

$$\text{Protein Content} = \beta_0 + \beta_1 \times \text{Cell Size} + \beta_2 \times \text{Metabolic Rate} + \varepsilon$$

This extended model is an example of multiple linear regression, where we use more than one predictor variable to explain the response. The general form for a multiple linear regression model with p predictors is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

Where:

- Y is the response variable we are trying to predict (protein content)
- $x_1, x_2, \ldots, x_p$ are the predictor variables (cell size, metabolic rate, etc.)
- β_0 is the intercept term (the expected protein content when all predictors are zero)
- $\beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients
- ε is the error term, assumed to be normally distributed with mean 0 and constant variance σ^2

The interpretation of the regression coefficients in multiple regression differs subtly but importantly from simple regression. Each coefficient represents the expected change in the response variable for a one-unit increase in the corresponding predictor, holding all other predictors constant. This "holding other variables constant" qualification is crucial - it means we are measuring the unique contribution of each predictor after accounting for the effects of all others.

For instance, in our cell protein content example:

- β_1 represents the expected change in protein content (in picograms) for a one-μm^3 increase in cell size, for cells with the same metabolic rate
- β_2 represents the expected change in protein content for a one-unit increase in metabolic rate, for cells of the same size

This ability to isolate the effect of individual biological factors while controlling for others makes multiple regression an exceptionally valuable tool for understanding complex relationships in multivariate biological data. It allows researchers to disentangle the various factors that simultaneously influence cellular processes like protein synthesis, gene expression, or metabolic flux.

1.2 Geometric Representation

To develop intuition for multiple linear regression, it is valuable to understand its geometric interpretation. In simple linear regression, our model corresponds to a line in a two-dimensional space: one dimension for the predictor variable and one for the response variable. The regression coefficients β_0 and β_1 determine the line's intercept and slope, respectively.

When we add a second predictor, our geometric representation extends to three dimensions: one for each predictor and one for the response. In this case, the regression model corresponds to a plane in this three-dimensional space.

For instance, in our cellular protein content example with two predictors (cell size and metabolic rate), the regression model:

$$\text{Protein Content} = \beta_0 + \beta_1 \times \text{Cell Size} + \beta_2 \times \text{Metabolic Rate} + \varepsilon$$

represents a plane in 3D space where:

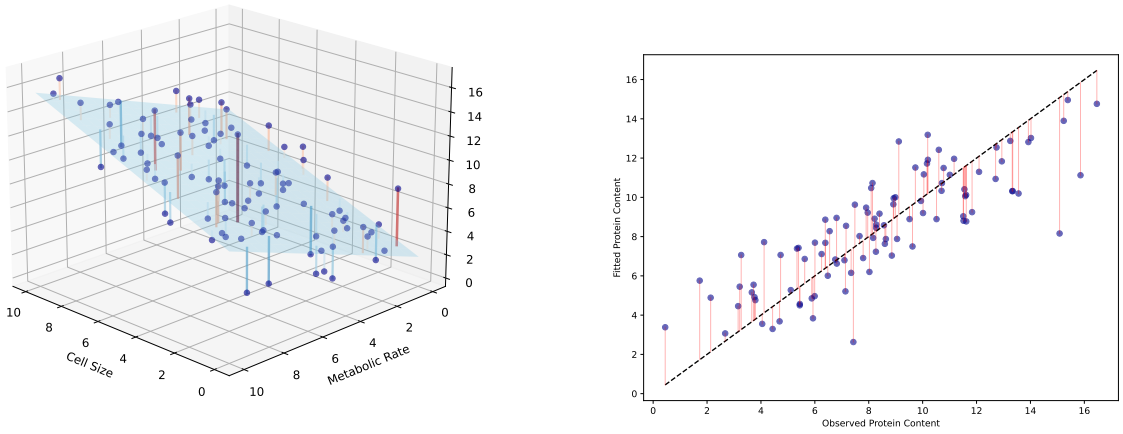
- The z-axis represents protein content
- The x-axis represents cell size
- The y-axis represents metabolic rate
- β_0 is the z-intercept (height of the plane at the origin)
- β_1 is the slope of the plane in the x-direction
- β_2 is the slope of the plane in the y-direction

The coefficients tell us how the response variable changes as we move along the coordinate axes. Moving one unit along the x-axis (increasing cell size by one unit while keeping metabolic rate constant) changes the height of the plane by β_1 units. Similarly, moving one unit along the y-axis (increasing metabolic rate by one unit while keeping cell size constant) changes the height by β_2 units.

For every point (x_1, x_2) in the predictor space, the model predicts a response value that corresponds to the height of the plane at that point. The residuals represent the vertical distances between the observed data points and this plane.

When we extend to more than two predictors, we move beyond what we can directly visualize. With p predictors, our model exists in a $(p+1)$-dimensional space and corresponds to a hyperplane – the higher-dimensional analog of a plane. While we cannot visualize hyperplanes in more than three dimensions, the mathematical principles remain the same:

- β_0 is the intercept (value of Y when all predictors are zero)
- Each β_j represents the change in Y for a one-unit increase in x_j, holding all other predictors constant



(a) Multiple linear regression as a plane in 3D space

(b) Residuals in multiple linear regression

Figure 1: **Geometric visualization of multiple linear regression with two predictors**

- Residuals are the differences between observed values and the hyperplane

This geometric perspective provides insight into the fitting process as well. When we estimate the regression coefficients using least squares, we are essentially finding the hyperplane that minimizes the sum of squared vertical distances between the observed data points and the hyperplane – a direct extension of the two-dimensional case.

Understanding this geometric representation helps clarify both the meaning of the regression coefficients and the overall modeling approach. However, it is important to remember that the true power of multiple regression lies in its ability to model complex relationships involving many predictors, even when those relationships cannot be directly visualized.

1.3 Formulation

Moving from specific examples to a general formulation, we can express multiple linear regression in a more compact and powerful form using linear algebra. This shift from scalar to matrix notation not only simplifies our representation but also facilitates theoretical analysis and computational implementation.

Let's begin with our standard multiple regression model for the i-th observation:

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

Here, y_i is the observed response value, x_{ij} is the value of the j-th predictor for the i-th observation, and ε_i is the error term.

Using index notation, we can rewrite this as:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

Now, consider a dataset with n observations. We can organize our data as follows:

- $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ - an $n \times 1$ vector of response values
- $\mathbf{x_j} = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$ - an $n \times 1$ vector of values for the j-th predictor
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ - a $p \times 1$ vector of regression coefficients

- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ - an $n \times 1$ vector of error terms

To handle the intercept term β_0, we can introduce a predictor that has the value 1 for all observations. Let's define $x_{i0} = 1$ for all i, and include β_0 in our vector of coefficients. Now our model becomes:

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \varepsilon_i$$

We can organize all predictor values into a matrix $\mathbf{X}$ of size $n \times (p+1)$:

$$\mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

With these definitions, we can express our multiple regression model succinctly using matrix notation:

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

This single equation encapsulates the entire system of n linear equations for our n observations. If we expand it, we get:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The matrix formulation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is remarkably elegant and concise. It allows us to express a complex system of equations in a single line and facilitates the application of powerful linear algebra techniques. However, this elegance comes with a shift in mathematical tools - we must now work with matrix operations rather than scalar algebra.

This brings both advantages and challenges. On one hand, matrix notation reveals structural patterns and simplifies manipulations that would be cumbersome with individual equations. On the other hand, understanding the model requires familiarity with matrix operations. For example, the product $\mathbf{X}\boldsymbol{\beta}$ involves multiplying each row of $\mathbf{X}$ with the vector $\boldsymbol{\beta}$, producing a vector of fitted values.

The assumptions of multiple linear regression can also be expressed concisely using matrix notation. For instance, the assumption that the errors have constant variance and are uncorrelated can be written as:

$$\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

This matrix formulation provides a foundation for extending regression to more complex models and for developing efficient computational methods for parameter estimation and inference.

1.4 Meet the Design Matrix

The matrix $\mathbf{X}$ that we introduced in the previous section plays a central role in regression analysis and is known as the design matrix or model matrix. This matrix organizes all predictor variables

for all observations, with each row corresponding to an observation and each column corresponding to a predictor variable (including the intercept).

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

The first column of ones represents the intercept term. Each subsequent column contains the values of a particular predictor variable across all observations. The design matrix serves as the bridge between our data and the statistical model, encoding the structure of the predictor variables that we believe influence the response.

Let's consider some concrete examples of design matrices in biological contexts:

1.4.1 Example 1: Cell Protein Content Study

In our previous example investigating how cell size and metabolic rate affect protein content, we might have data from 6 cells:

Cell	Protein Content (pg)	Size (μm^3)	Metabolic Rate
1	15.2	120	3.2
2	10.7	95	2.8
3	18.3	135	3.5
4	12.9	105	3.0
5	21.1	150	3.9
6	14.5	110	3.1

The design matrix for this dataset would be:

$$\mathbf{X} = \begin{pmatrix} 1 & 120 & 3.2 \\ 1 & 95 & 2.8 \\ 1 & 135 & 3.5 \\ 1 & 105 & 3.0 \\ 1 & 150 & 3.9 \\ 1 & 110 & 3.1 \end{pmatrix}$$

This 6×3 matrix contains the predictor values for each of our 6 observations. The response vector would be:

$$\mathbf{y} = \begin{pmatrix} 15.2 \\ 10.7 \\ 18.3 \\ 12.9 \\ 21.1 \\ 14.5 \end{pmatrix}$$

1.4.2 Example 2: Gene Expression Study

Consider a study examining how gene expression (measured by mRNA levels) is affected by temperature and exposure time in a cell culture. The researchers collect data under various conditions:

Sample	Expression Level	Temperature (°C)	Exposure Time (h)
1	25.3	37.0	6
2	18.7	37.0	3
3	32.6	39.5	6
4	22.4	39.5	3
5	15.8	34.5	6
6	10.2	34.5	3
7	28.9	38.0	9
8	20.1	36.0	9

The design matrix would be:

$$\mathbf{X} = \begin{pmatrix} 1 & 37.0 & 6 \\ 1 & 37.0 & 3 \\ 1 & 39.5 & 6 \\ 1 & 39.5 & 3 \\ 1 & 34.5 & 6 \\ 1 & 34.5 & 3 \\ 1 & 38.0 & 9 \\ 1 & 36.0 & 9 \end{pmatrix}$$

These examples illustrate design matrices with continuous predictors. In practice, design matrices can become more complex when we include:

- Transformations of variables (e.g., log transformations or polynomials)
- Interaction terms between predictors
- Categorical predictors (requiring dummy variable encoding)

The careful construction of the design matrix is a critical step in regression modeling, as it determines the form of the relationship we are assuming between the predictors and the response. By examining the structure of $\mathbf{X}$, we can understand the model being fitted and its underlying assumptions.

Additionally, the properties of the design matrix have important implications for estimation. For example:

- The columns of $\mathbf{X}$ must be linearly independent for the model to be identifiable
- The condition number of $\mathbf{X}^T\mathbf{X}$ affects the stability of parameter estimation
- The rank of $\mathbf{X}$ determines the degrees of freedom available for estimation

In subsequent sections, we'll see how the design matrix is used in parameter estimation and inference, serving as the cornerstone of the regression analysis framework.

1.5 Parameter Estimation

Just as in simple linear regression, our goal in multiple regression is to estimate the unknown parameters that best describe the relationship between our predictors and the response variable. The method of least squares provides a principled approach to this estimation problem.

When working with a single predictor, we found values of β_0 and β_1 that minimized the sum of squared residuals:

$$\text{RSS} = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

For multiple regression, we extend this approach by minimizing:

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2$$

Using our matrix notation, we can express this objective function more elegantly. Since the vector of residuals is $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, the sum of squared residuals becomes:

$$\text{RSS}(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

This formulation generalizes the least squares approach from simple to multiple regression in a natural way. In both cases, we seek to minimize the squared Euclidean distance between the observed responses and the predictions from our model.

To find the values of $\boldsymbol{\beta}$ that minimize this expression, we need to use calculus - specifically, we must take the derivative of the RSS with respect to the vector $\boldsymbol{\beta}$ and set it equal to zero. This will lead us to the normal equations, which we'll explore in the next section.

1.6 Normal Equations

To find the minimum of the RSS function, we need to differentiate it with respect to the vector of parameters $\boldsymbol{\beta}$ and set the result equal to zero. This requires us to work with matrix calculus, which extends the familiar rules of scalar calculus to matrices and vectors.

Starting with our RSS function:

$$\text{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

We can expand this expression:

$$\begin{aligned}\text{RSS}(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Note that $\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}$ is a scalar (specifically, a 1×1 matrix), so it equals its transpose $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$. This allows us to simplify:

$$\text{RSS}(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

Now, we take the partial derivative with respect to $\boldsymbol{\beta}$. Using the rules of matrix differentiation: 1. $\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{y}^T\mathbf{y} = \mathbf{0}$ (constant term) 2. $\frac{\partial}{\partial\boldsymbol{\beta}}(2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}) = 2\mathbf{X}^T\mathbf{y}$ 3. $\frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$
Therefore:

$$\frac{\partial\text{RSS}}{\partial\boldsymbol{\beta}} = \mathbf{0} - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

Setting this equal to zero:

$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

Dividing by 2 and rearranging:

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

These are the normal equations for multiple linear regression. They represent a system of $p+1$ linear equations in $p+1$ unknowns (the elements of $\boldsymbol{\beta}$).

The normal equations have several important properties:

- They provide a necessary condition for the minimizer of the RSS function.
- If the columns of $\mathbf{X}$ are linearly independent (i.e., $\mathbf{X}$ has full column rank), then $\mathbf{X}^T\mathbf{X}$ is positive definite and invertible, ensuring a unique solution.
- If $\mathbf{X}$ does not have full column rank (e.g., when we have perfect multicollinearity among predictors), then the normal equations have infinitely many solutions, and additional constraints are needed to obtain a unique solution.

The normal equations have a clear geometric interpretation as well: they ensure that the residual vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is orthogonal to the column space of $\mathbf{X}$. In other words, the residuals cannot be further reduced by adjusting the coefficients in any direction within the model's capabilities.

1.7 Formula for the Solution

When the design matrix $\mathbf{X}$ has full column rank, we can solve the normal equations $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$ by multiplying both sides by $(\mathbf{X}^T\mathbf{X})^{-1}$:

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

This formula gives the least squares estimator for the regression coefficients. The expression $\mathbf{X}^T\mathbf{X}$ is a $(p+1) \times (p+1)$ matrix of sums of squares and cross-products of the predictor variables, while $\mathbf{X}^T\mathbf{y}$ is a $(p+1) \times 1$ vector of sums of products between predictors and the response.

The least squares estimator $\hat{\boldsymbol{\beta}}$ possesses several important statistical properties when the standard regression assumptions hold:

- **Unbiasedness**: $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, meaning the estimator is correct on average.
- **Consistency**: As the sample size increases, $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$.
- **Efficiency**: Among all unbiased linear estimators, $\hat{\boldsymbol{\beta}}$ has the smallest variance (Gauss-Markov theorem).
- **Normality**: If the errors ε_i are normally distributed, then $\hat{\boldsymbol{\beta}}$ follows a multivariate normal distribution:
$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

This expression generalizes what we saw in simple linear regression. For the simplest case with one predictor and an intercept, where:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

The formula $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ yields exactly the familiar expressions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

This demonstrates how the matrix approach provides a unified framework that encompasses both simple and multiple regression as special cases.

The geometric interpretation of this solution is that $\hat{\boldsymbol{\beta}}$ gives the coefficients of the hyperplane that minimizes the sum of squared vertical distances between the observed data points and the hyperplane. In the n-dimensional space of observations, $\mathbf{X}\hat{\boldsymbol{\beta}}$ represents the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$ – the space of all possible fitted values that can be achieved with our model.

In the next sections, we will explore how to make inference about these estimated coefficients and how to assess the overall quality of the regression model.

1.8 Linear Regression in the Observation Space: A Linear Algebra Perspective

While we've discussed regression as fitting a hyperplane in the predictor-response space, there's another powerful geometric perspective: viewing regression in the n-dimensional observation space. This approach reveals fundamental properties of the least squares solution.

1.8.1 Two Different Geometric Spaces

it is important to distinguish between two geometric views of regression:

- **Predictor-Response Space:** Our earlier visualization where we fit a (p)-dimensional hyperplane in a $(p+1)$-dimensional space (with p predictors and 1 response dimension). Each axis represents a variable.
- **Observation Space:** The n-dimensional space where each axis represents an observation. In this space, vectors have n components, one for each data point.

The observation space perspective provides deeper insights into the mathematical structure of regression.

1.8.2 Vectors in Observation Space

In the n-dimensional observation space:

- The response vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is a single point
- Each column of $\mathbf{X}$ is a vector in this space
- The fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ represent another point

The column space of $\mathbf{X}$ is the subspace spanned by its columns - all possible linear combinations of the predictor vectors. This is a $(p+1)$-dimensional subspace within the n-dimensional observation space.

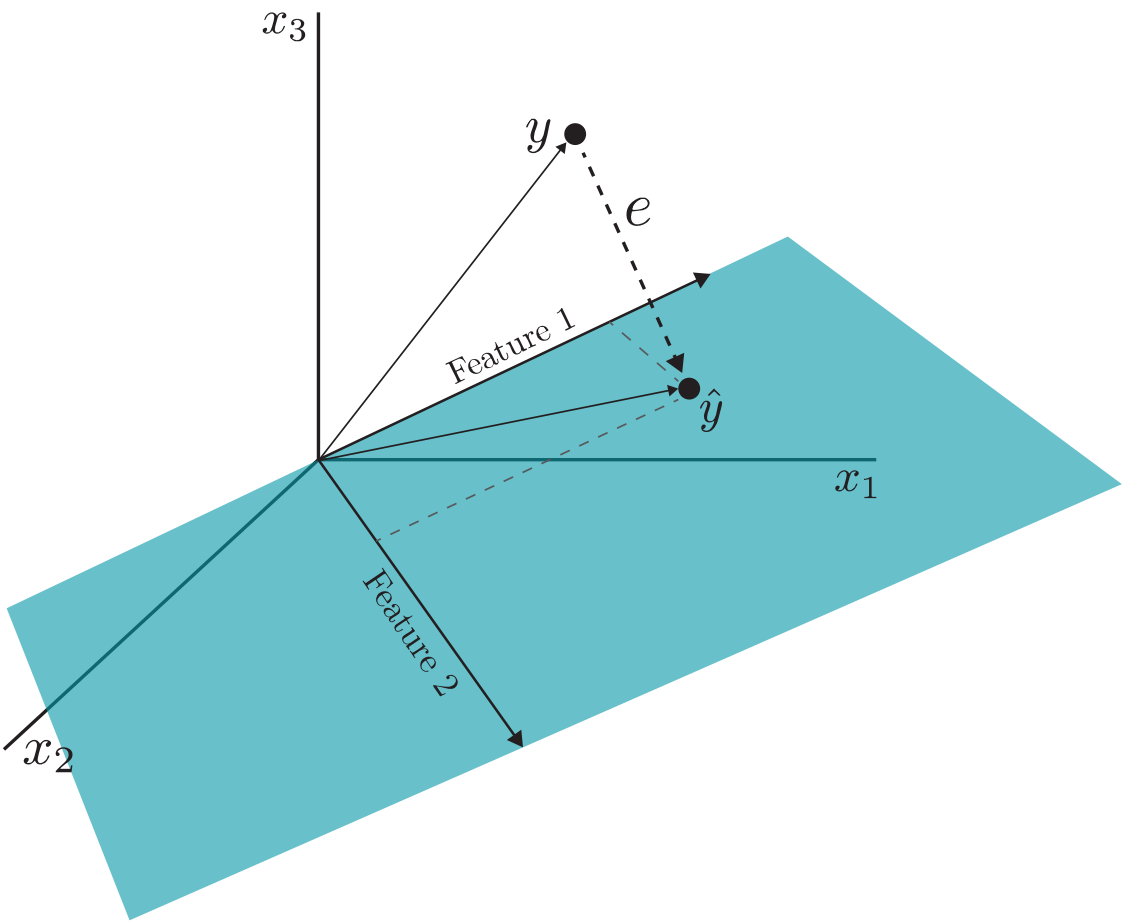


Figure 2: Orthogonal projection in observation space for a case with 3 samples and 2 features. The response vector $\mathbf{y}$ is projected onto the column space of $\mathbf{X}$, yielding the fitted values $\hat{\mathbf{y}}$. The residual vector $\mathbf{e}$ is perpendicular to the column space. In this case, $\mathbf{X}$ has two columns, corresponding to "Feature 1" and "Feature 2".

1.8.3 Least Squares as Orthogonal Projection

The least squares solution has a remarkable interpretation: $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$.

This means:

- $\hat{\mathbf{y}}$ is the closest point in the column space to $\mathbf{y}$ (minimizing Euclidean distance)
- The residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to every vector in the column space
- This orthogonality gives us $\mathbf{X}^T\mathbf{e} = \mathbf{0}$, which leads directly to the normal equations

1.8.4 The Projection Matrix

The hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ mathematically implements this projection:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

As a projection matrix, $\mathbf{H}$ has several important properties:

- It is idempotent: $\mathbf{H}^2 = \mathbf{H}$ (applying the projection twice is the same as applying it once)
- It is symmetric: $\mathbf{H}^T = \mathbf{H}$
- Its eigenvalues are either 0 or 1
- Its trace equals $p + 1$, the dimension of the column space

The complementary projection matrix $\mathbf{M} = \mathbf{I} - \mathbf{H}$ projects onto the orthogonal complement of the column space:

$$\mathbf{e} = \mathbf{M}\mathbf{y}$$

1.8.5 Why This Perspective Matters

This geometric understanding reveals why least squares has certain properties:

- The residuals sum to zero when the model includes an intercept (because the residual vector is orthogonal to the constant vector)
- The residuals are uncorrelated with all predictors (due to orthogonality with the column space)
- Adding more predictors never increases the residual sum of squares (as the column space only grows larger)

Moreover, this perspective connects regression to broader concepts in linear algebra and provides the foundation for advanced topics.

1.9 Interpreting Regression Coefficients

In multiple regression, the interpretation of coefficients requires careful attention to the presence of other variables in the model. Each coefficient represents the expected change in the response variable for a one-unit increase in the corresponding predictor, while holding all other predictors constant.

1.9.1 The Ceteris Paribus Interpretation

The phrase "holding all other variables constant" (or *ceteris paribus* in Latin) is crucial in multiple regression. It distinguishes the interpretation from simple regression, where no other predictors are involved.

For our model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$:

- β_0 (intercept): The expected value of Y when all predictors equal zero. This may not always have a meaningful interpretation, especially if zero values for predictors are outside their realistic range.
- β_j (slope): The expected change in Y for a one-unit increase in X_j, while holding all other predictors constant. Mathematically, this can be expressed as:

$$\beta_j = \frac{\partial E[Y|X_1, X_2, \ldots, X_p]}{\partial X_j}$$

1.9.2 Concrete Examples

To illustrate these interpretations, let's revisit our cellular protein content example with the model:

$$\text{Protein Content} = \beta_0 + \beta_1 \times \text{Cell Size} + \beta_2 \times \text{Metabolic Rate} + \varepsilon$$

If we estimate $\hat{\beta}_0 = 2.1$, $\hat{\beta}_1 = 0.08$, and $\hat{\beta}_2 = 2.5$, we interpret:

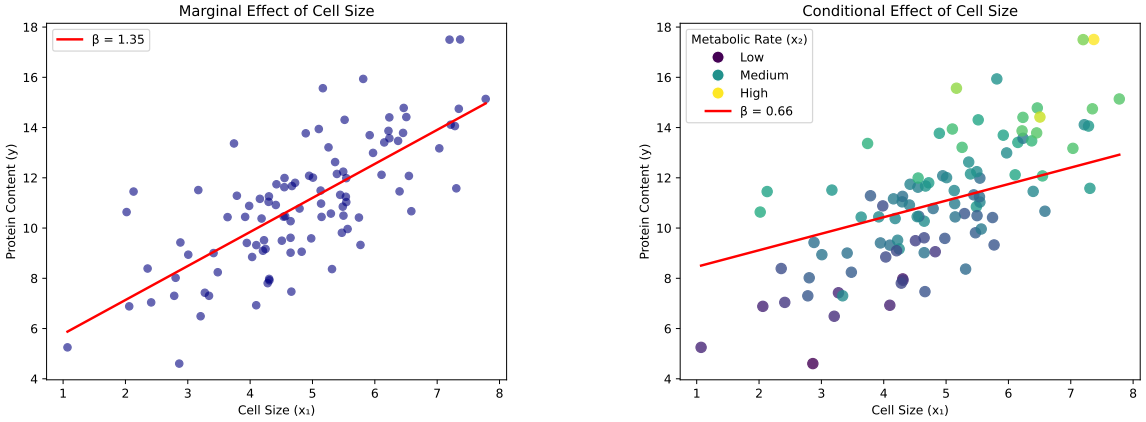
- $\hat{\beta}_0 = 2.1$: The expected protein content is 2.1 picograms for a cell with zero size and zero metabolic rate. This is not biologically meaningful since cells cannot have zero size.
- $\hat{\beta}_1 = 0.08$: For cells with the same metabolic rate, each additional cubic micrometer of cell size is associated with an expected increase of 0.08 picograms in protein content.
- $\hat{\beta}_2 = 2.5$: For cells of the same size, each additional unit of metabolic rate is associated with an expected increase of 2.5 picograms in protein content.

1.9.3 Conditional vs. Marginal Effects

A key distinction in multiple regression is between conditional and marginal effects:

- **Conditional effect** (captured by β_j): The effect of X_j on Y after controlling for other predictors in the model.
- **Marginal effect** (from simple regression of Y on X_j alone): The overall association between X_j and Y without controlling for other variables.

These effects can differ substantially, especially when predictors are correlated. For example, if cell size and metabolic rate are positively correlated, the marginal effect of cell size on protein content (without controlling for metabolic rate) would likely be larger than its conditional effect (controlling for metabolic rate).



(a) Marginal effect (simple regression) of x_1 on y

(b) Conditional effect of x_1 after controlling for x_2

Figure 3: **Comparison of marginal effects (simple regression) and conditional effects (multiple regression)**

This figure illustrates how the slope of the relationship between X_1 and Y changes when we control for another predictor X_2. The steeper line represents the marginal effect (simple regression), while the flatter line represents the conditional effect (multiple regression).

1.9.4 Standardized Coefficients

In multiple regression, predictor variables often have different units of measurement, making direct comparison of their coefficients challenging. For instance, in a model predicting enzyme activity, temperature might be measured in degrees Celsius while substrate concentration is in millimolar units. The coefficient for temperature might be 0.25 activity units per degree, while the coefficient for concentration might be 2.5 activity units per millimolar. Which variable has a "stronger" effect?

Standardized coefficients (also called beta coefficients) address this issue by converting all variables to a common scale, allowing for direct comparison of their relative influences on the response variable.

Definition and Calculation To compute standardized coefficients, we standardize both the predictors and the response variable by subtracting their means and dividing by their standard deviations:

$$Z_Y = \frac{Y - \bar{Y}}{s_Y}, \quad Z_{X_j} = \frac{X_j - \bar{X}_j}{s_{X_j}}$$

We then perform the regression using these standardized variables:

$$Z_Y = \beta_0^* + \beta_1^* Z_{X_1} + \beta_2^* Z_{X_2} + \cdots + \beta_p^* Z_{X_p} + \varepsilon^*$$

The resulting coefficients β_j^* are the standardized regression coefficients. They can also be calculated directly from the unstandardized coefficients using:

$$\beta_j^* = \beta_j \frac{s_{X_j}}{s_Y}$$

where β_j is the unstandardized coefficient, s_{X_j} is the standard deviation of predictor X_j, and s_Y is the standard deviation of the response variable.

Interpretation Standardized coefficients represent the expected change in the response variable (in standard deviation units) for a one standard deviation increase in the predictor, holding all other predictors constant. For example, a standardized coefficient of 0.5 for a predictor means that a one standard deviation increase in that predictor is associated with a 0.5 standard deviation increase in the response variable, controlling for other predictors.

This standardization allows for direct comparison of the "strength" or "importance" of different predictors in the model, regardless of their original units of measurement.

This figure illustrates how standardized coefficients can reveal different patterns of variable importance compared to unstandardized coefficients. In this example, the variable pH has the largest unstandardized coefficient (in absolute value) but not the largest standardized coefficient, indicating that its apparent importance was partly due to its scale of measurement.

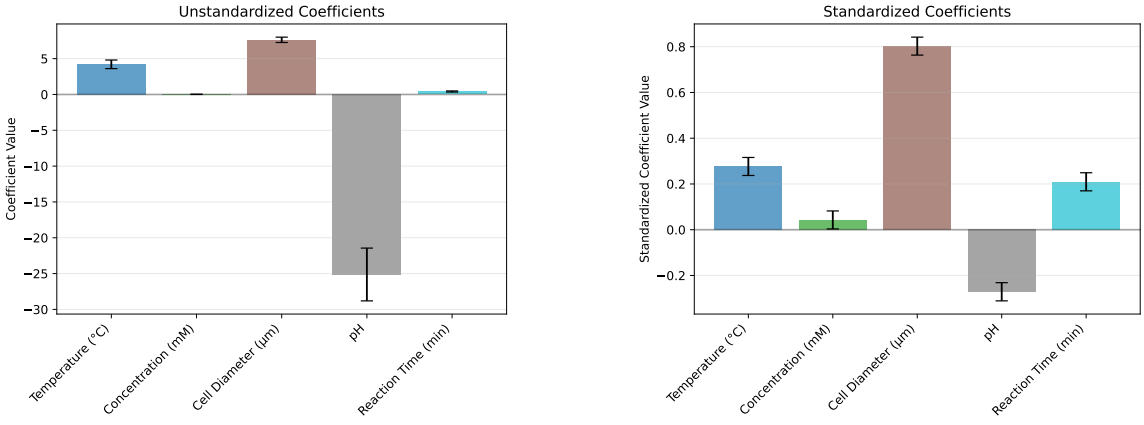
Advantages and Limitations Standardized coefficients make it easier to compare predictors measured on different scales, helping researchers identify which variables have the strongest statistical relationships with the outcome. This is particularly valuable when comparing results across different studies that may use different measurement units or methods.

However, these coefficients have important limitations. Their interpretation becomes more abstract since they're expressed in standard deviation units rather than natural units. They also depend on the sample's variability, meaning they can change across different datasets even when the underlying relationship remains constant. Perhaps most importantly, standardized coefficients reflect statistical associations rather than biological significance - a variable might have a large standardized coefficient simply because it is measured with less error, not because it is more biologically

Variable	Unstandardized Coefficient (± SE)	Standardized Coefficient (± SE)
Temperature (°C)	4.207 ± 0.599	0.276 ± 0.039
Concentration (mM)	0.016 ± 0.015	0.043 ± 0.039
Cell Diameter (µm)	7.625 ± 0.373	0.803 ± 0.039
pH	-25.125 ± 3.682	-0.271 ± 0.040
Reaction Time (min)	0.414 ± 0.078	0.210 ± 0.040

(a) Comparison table of both coefficient types

Figure 4: **Comparison table of unstandardized and standardized coefficients**



(a) Unstandardized regression coefficients

(b) Standardized regression coefficients

Figure 5: **Comparison of unstandardized (original) and standardized regression coefficients**

meaningful. They also don't account for non-linearities or correct for multicollinearity between predictors.

Application in Biological Research Standardized coefficients are valuable in biological research when comparing variables with different units or when synthesizing results across studies with different measurement approaches. They reveal which factors have the strongest statistical associations with outcomes - for example, showing that age has a stronger relationship with bone mineral density than calcium intake. However, these statistical relationships don't necessarily reflect biological importance; a variable might show a strong association simply because it is measured more precisely. Meaningful interpretation always requires integrating these statistical results with domain knowledge about the biological system under study.

2 Model Comparison in Multiple Regression

2.1 Assessing Model Fit: R^2 and Adjusted R^2

When evaluating regression models, we need measures to assess how well our model fits the data. The coefficient of determination, R^2, is a widely used metric, but it has limitations in the context of multiple regression. This leads us to the adjusted R^2, which addresses some of these limitations.

2.1.1 Coefficient of Determination (R^2)

The coefficient of determination, R^2, quantifies the proportion of variance in the response variable that is explained by the regression model:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where:

- RSS is the residual sum of squares (unexplained variation)
- TSS is the total sum of squares (total variation in Y)
- $\hat{y}_i$ are the fitted values from the regression model
- $\bar{y}$ is the mean of the response variable

R^2 ranges from 0 to 1, with:

- $R^2 = 0$: The model explains none of the variability in the response
- $R^2 = 1$: The model explains all the variability in the response

In simple linear regression, R^2 equals the square of the Pearson correlation coefficient between X and Y. In multiple regression, it can be interpreted as the square of the correlation between the observed values and the fitted values.

While R^2 provides a seemingly intuitive measure of fit, it has a critical flaw in the multiple regression context: it always increases (or at least never decreases) when additional predictors are added to the model, regardless of whether these predictors are actually related to the response.

The figure illustrates how R^2 tends to increase as more predictors are added to the model, even when those predictors are random noise. This property makes R^2 problematic for comparing models with different numbers of predictors, as it will always favor more complex models.

2.1.2 Adjusted R^2

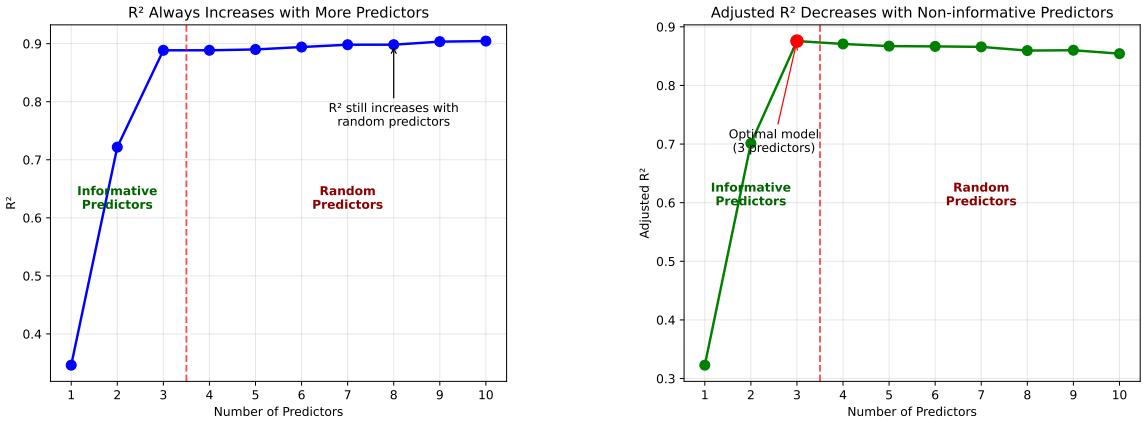
To address this limitation, the adjusted R^2 introduces a penalty for including additional predictors:

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

where:

- n is the number of observations
- p is the number of predictors (excluding the intercept)

The adjusted R^2 balances the goodness of fit against model complexity. Unlike the regular R^2, the adjusted R^2 can decrease when a predictor is added that doesn't improve the model enough to justify the loss of a degree of freedom.



(a) R^2 increases with additional predictors, even random ones

(b) Adjusted R^2 decreases with non-informative predictors

Figure 6: **Comparison of R^2 and adjusted R^2 for model selection**

This figure shows how the adjusted R^2 can actually decrease when non-informative predictors are added to the model, making it a more reliable metric for model selection.

2.1.3 Interpretation and Limitations

While the adjusted R^2 is useful for comparing models with different numbers of predictors, it has several limitations:

- Like R^2, it assumes the model is correctly specified (e.g., linear relationship, homoscedasticity).
- It doesn't provide a formal hypothesis test for model comparison.
- It doesn't account for multicollinearity, transformations, or interaction terms.
- Adding a variable that is only slightly correlated with the dependent variable can make the adjusted R^2 go up, even if it is not materially important.

In biological research, adjusted R^2 is best used as one of several tools for model assessment, alongside other metrics, residual analysis, and domain knowledge. Different fields have different standards for what constitutes a "good" R^2 or adjusted R^2 value.

2.2 Comparing Nested Models

When working with multiple predictors, researchers often need to determine which variables contribute meaningfully to explaining the response. A fundamental approach to this question is nested model comparison, where we test whether adding certain variables significantly improves model fit.

Two models are considered "nested" when one model (the restricted model) contains a subset of the predictors in the other model (the unrestricted model). The statistical question becomes: Does the addition of extra predictors in the unrestricted model provide a significantly better fit to the data?

Consider a study examining factors affecting enzyme activity in a biochemical reaction. Researchers measure:

- Y: Enzyme activity (nmol/min)
- X_1: Substrate concentration (mM)
- X_2: Temperature (°C)
- X_3: pH
- X_4: Ionic strength (mM)

The researchers want to determine if temperature and pH (variables X_2 and X_3) significantly affect enzyme activity after accounting for substrate concentration and ionic strength. This leads to a comparison between:

- Restricted model: $Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$
- Unrestricted model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$

The null hypothesis for this comparison is:

$$H_0 : \beta_2 = \beta_3 = 0$$

This hypothesis states that temperature and pH have no effect on enzyme activity after controlling for substrate concentration and ionic strength.

2.3 Likelihood Ratio Test for Nested Models

To test this hypothesis, we can use the likelihood ratio test (LRT), which compares the maximum likelihood achieved by each model. The LRT is especially valuable because it generalizes beyond ordinary least squares to any maximum likelihood framework, including generalized linear models and other statistical approaches.

The test statistic for the likelihood ratio test is:

$$LR = 2(\ell_{unrestricted} - \ell_{restricted})$$

Where $\ell_{unrestricted}$ and $\ell_{restricted}$ are the log-likelihoods of the unrestricted and restricted models, respectively. Under the null hypothesis, this LR statistic asymptotically follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

In the context of linear regression with normally distributed errors, the log-likelihood is directly related to the residual sum of squares (RSS). The relationship is:

$$\ell = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{RSS}{2\sigma^2}$$

Where σ^2 is the error variance. When comparing nested models, the LR statistic can be expressed as:

$$LR = n \cdot \ln\left(\frac{RSS_{restricted}}{RSS_{unrestricted}}\right)$$

This formulation reveals how model comparison inherently balances improvement in fit against model complexity.

2.4 Applying the Likelihood Ratio Test

Returning to our enzyme activity example, let's examine how the likelihood ratio test works in practice.

Suppose researchers collected data from 50 experimental runs with the following results:

Model	Parameters	RSS	df	Log-likelihood
M1: X_1, X_4	3	2450	47	-173.6
M2: X_1, X_2, X_3, X_4	5	1820	45	-165.2

The likelihood ratio test statistic can be calculated as:

$$LR = 2((-165.2) - (-173.6)) = 2(8.4) = 16.8$$

Alternatively, using the RSS formulation:

$$LR = 50 \cdot \ln\left(\frac{2450}{1820}\right) = 50 \cdot \ln(1.346) \approx 14.85$$

With 2 degrees of freedom (the difference in parameters between models), we compare this to a chi-square distribution. The critical value at $\alpha = 0.05$ with 2 df is 5.99. Since our calculated LR value exceeds this threshold, we reject the null hypothesis and conclude that temperature and pH significantly improve the model's fit.

This approach can be incorporated into model-building strategies:

1. Begin with predictors supported by prior knowledge
2. Add variables in meaningful groups, using LR tests to evaluate their contribution
3. Examine individual coefficients within significant groups

The LR test provides a principled framework for comparing nested models and extends naturally to other maximum likelihood frameworks, including generalized linear models and non-linear models often used in biological research.

3 ANOVA in the Regression Framework

3.1 Introduction

In the lecture on statistical tests, we learned how to compare two groups using methods like the t-test. But what if we have more than two groups? How do we compare multiple means simultaneously?

One approach could be to test all possible pairs of groups, but this would lead to multiple hypothesis tests and a higher risk of false positives. A more meaningful question might be whether all the group means are equal. This is the fundamental idea behind Analysis of Variance (ANOVA), a statistical method used to compare means across multiple groups.

3.2 Traditional ANOVA Approach

ANOVA is traditionally presented as a method that partitions the total variation in the data into components attributed to different sources. For instance, when comparing multiple groups, we decompose the total sum of squares (SST) into:

- Between-group sum of squares (SSB): Variation due to differences among group means
- Within-group sum of squares (SSW): Variation due to differences within each group

This gives us the fundamental ANOVA identity: SST = SSB + SSW.
The test statistic in ANOVA is the F-ratio:

$$F = \frac{\text{Between-group variation}/(k-1)}{\text{Within-group variation}/(n-k)} = \frac{\text{MSB}}{\text{MSW}}$$

Where:

- MSB is the mean square between groups (SSB divided by its degrees of freedom)
- MSW is the mean square within groups (SSW divided by its degrees of freedom)
- k is the number of groups
- n is the total number of observations

Under the null hypothesis that all group means are equal, this F-statistic follows an F-distribution with $(k-1, n-k)$ degrees of freedom. Large values of F suggest that between-group variation exceeds what we would expect by chance, leading us to reject the null hypothesis.

3.3 ANOVA as a Special Case of Regression

Although commonly presented as a separate technique, ANOVA is actually a special case of linear regression with categorical predictors.

Consider a researcher studying the effects of different strains of a bacterial species on growth rates when stimulated with a particular nutrient. The researcher has four different strains (A, B, C, and D) and measures bacterial growth rate after 24 hours for multiple samples in each group.

The fundamental question is: Does the strain matter? In other words: Do any of these strains exhibit different growth responses when exposed to this nutrient?

Rather than performing six separate pairwise comparisons between all possible pairs of strains, ANOVA tests the single null hypothesis:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

where μ_i represents the population mean growth rate for strain i.

We can express this scenario as a regression model with categorical predictors using dummy variables. If we designate strain A as our reference category, the model becomes:

$$Y_i = \beta_0 + \beta_1 X_{Bi} + \beta_2 X_{Ci} + \beta_3 X_{Di} + \varepsilon_i$$

where:

- Y_i is the bacterial growth rate for the ith observation
- X_{Bi} equals 1 if the ith sample was strain B, 0 otherwise
- X_{Ci} equals 1 if the ith sample was strain C, 0 otherwise

- X_{Di} equals 1 if the ith sample was strain D, 0 otherwise

In this formulation:

- β_0 represents the mean growth rate for strain A
- β_1 represents the difference between mean growth rates of strains B and A
- β_2 represents the difference between mean growth rates of strains C and A
- β_3 represents the difference between mean growth rates of strains D and A

Our null hypothesis can now be rewritten as $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, which we can test using the principles of multiple linear regression.

3.4 Testing in ANOVA Using Model Comparison

In the regression framework, the ANOVA null hypothesis becomes a question of whether the model with group indicators fits significantly better than a model with just an intercept. This can be approached as a nested model comparison.

We can formulate this as a comparison between two nested models:

- Restricted model: $Y_i = \beta_0 + \varepsilon_i$ (all group means equal)
- Unrestricted model: $Y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j X_{ji} + \varepsilon_i$ (potentially different means)

Initially, we can evaluate this using a chi-squared test. Under certain assumptions, the difference in deviance between nested models follows a chi-squared distribution:

$$\chi^2 = -2(\ell_{restricted} - \ell_{unrestricted})$$

Where ℓ represents the log-likelihood of each model. With normally distributed errors, this can be expressed in terms of the residual sum of squares:

$$\chi^2 = n \cdot \ln\left(\frac{\text{RSS}_{restricted}}{\text{RSS}_{unrestricted}}\right)$$

This chi-squared statistic has $k-1$ degrees of freedom, representing the difference in parameters between the models.

Interestingly, this test is equivalent to the traditional F-test used in ANOVA. The F-statistic for this model comparison is:

$$F = \frac{\text{RSS}_{restricted} - \text{RSS}_{unrestricted}}{\text{RSS}_{unrestricted}} \cdot \frac{n-k}{k-1}$$

Where:

- $\text{RSS}_{restricted}$ is the residual sum of squares from the intercept-only model
- $\text{RSS}_{unrestricted}$ is the residual sum of squares from the model with group indicators
- n is the total number of observations
- k is the number of groups

3.5 Connecting the Two Frameworks - (Optional!)

The equivalence between traditional ANOVA and the regression framework can be demonstrated mathematically:

- $\text{RSS}_{restricted} = \text{SST}$ (Total sum of squares)
- $\text{RSS}_{unrestricted} = \text{SSW}$ (Within-group sum of squares)
- $\text{RSS}_{restricted} - \text{RSS}_{unrestricted} = \text{SSB}$ (Between-group sum of squares)

The F-statistic in regression model comparison:

$$F = \frac{(\text{RSS}_{restricted} - \text{RSS}_{unrestricted})/(k-1)}{\text{RSS}_{unrestricted}/(n-k)} = \frac{\text{SSB}/(k-1)}{\text{SSW}/(n-k)} = \frac{\text{MSB}}{\text{MSW}}$$

This is identical to the traditional ANOVA F-ratio, confirming that both approaches test the same hypothesis.

For the likelihood ratio test:

$$LR = 2(\ell_{unrestricted} - \ell_{restricted}) = n \cdot \ln\left(\frac{\text{RSS}_{restricted}}{\text{RSS}_{unrestricted}}\right)$$

The relationship between the LR statistic and F-statistic can be established through Taylor approximation:

$$\ln\left(\frac{\text{RSS}_{restricted}}{\text{RSS}_{unrestricted}}\right) = \ln\left(1 + \frac{\text{SSB}}{\text{SSW}}\right) \approx \frac{\text{SSB}}{\text{SSW}}$$

Therefore:

$$LR \approx n \cdot \frac{\text{SSB}}{\text{SSW}} = n \cdot \frac{k-1}{n-k} \cdot F$$

This demonstrates that under the null hypothesis:

- F follows an F-distribution with $(k-1, n-k)$ degrees of freedom
- LR follows a chi-square distribution with $k-1$ degrees of freedom
- As n increases, $\frac{k-1}{n-k} \cdot F$ approximates a chi-square distribution with $k-1$ degrees of freedom

These relationships confirm that ANOVA is mathematically equivalent to testing for the significance of categorical predictors in a regression model.