Multiple Linear Regression

Gioele La Manno

École Polytechnique Fédérale de Lausanne (EPFL)

School of Life Science (SV)

March 2025

EPFL - BMI - UPLAMANNO

Contents

Introduction to Multiple LinearRegression

Extending to Multiple Predictors

Simple linear regression models relationship between two variables:

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

Multiple linear regression captures multiple factors simultaneously:

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

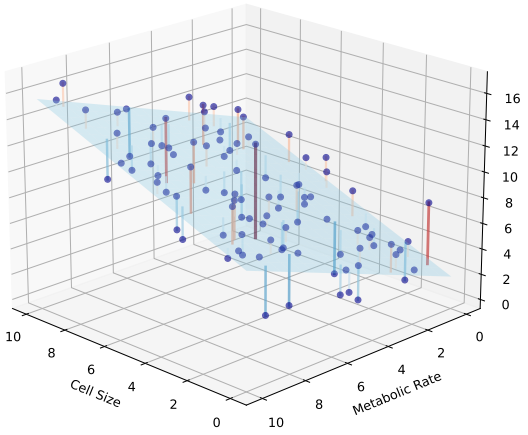
The general form with p predictors:

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

In multiple regression, each coefficient represents the effect of changing one predictor while holding all others constant.

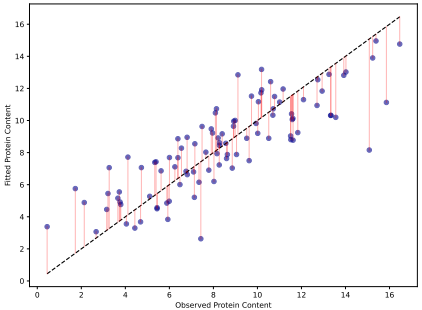
Geometric Representation of Multiple Regression

- In simple regression, we fit a line in 2D space.
- In multiple regression with two predictors, we fit a plane in 3D space.
- In general, with p predictors, we fit a hyperplane in $p + 1$ dimensions.



Residuals in Multiple Regression

Residuals represent the differences between observed values and those predicted by the model.



Matrix Formulation

Multiple regression with n observations and p predictors can be expressed concisely using matrices:

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

Where:

- $\mathbf{y}$ is an $n \times 1$ vector of responses
- $\mathbf{X}$ is an $n \times (p+1)$ design matrix
- $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of coefficients
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors

This compact notation reveals structural patterns and simplifies manipulations that would be cumbersome with individual equations.

The Design Matrix: Definition

The design matrix $\mathbf{X}$ is a fundamental concept in multiple regression that organizes all predictor variables for all observations:

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

Where:

- Each row represents one observation
- Each column represents one predictor variable
- The first column (all 1s) accounts for the intercept term β_0
- x_{ij} is the value of the j-th predictor for the i-th observation

The Design Matrix: Example

For our cell protein content example with 6 cells and two predictors (cell size and metabolic rate):

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

This matrix encodes:

- Each row: data from a single cell
- Column 1: intercept term (always 1)
- Column 2: cell size (in cubic micrometers)
- Column 3: metabolic rate (arbitrary units)

Parameter Estimation and the Normal Equations

To estimate coefficients, we use least squares to minimize the sum of squared residuals:

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

Taking derivatives and setting equal to zero leads to the normal equations:

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

The solution gives us the least squares estimator:

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

This formula generalizes simple linear regression to multiple predictors in an elegant way.

Two Geometric Perspectives on Regression

- **Predictor-Response Space:** Fitting a hyperplane where each axis is a variable
- **Observation Space:** Working in n-dimensional space where each axis is an observation
  - Response vector $\mathbf{y}$ and columns of $\mathbf{X}$ are points/vectors
  - Column space of $\mathbf{X}$ is a subspace of possible fitted values

The observation space perspective reveals deeper mathematical insights about regression.

Least Squares as Orthogonal Projection

$\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$
This geometric view explains why:

- Residuals are orthogonal to all predictor vectors
- Adding predictors never increases residual sum of squares
- Residuals sum to zero when model includes an intercept

The Projection Matrix

The "hat matrix" implements the orthogonal projection:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \quad \text{and} \quad \mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Properties: $\mathbf{H}$ is symmetric, idempotent ($\mathbf{H}^2 = \mathbf{H}$), with trace $= p + 1$
This perspective connects regression to broader concepts in linear algebra and provides the foundation for advanced techniques like ridge regression and principal components analysis, principal components regression and partial least squares

Interpreting Regression Coefficients

In multiple regression, each coefficient represents the effect of its predictor while holding all others constant.
For the model:

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

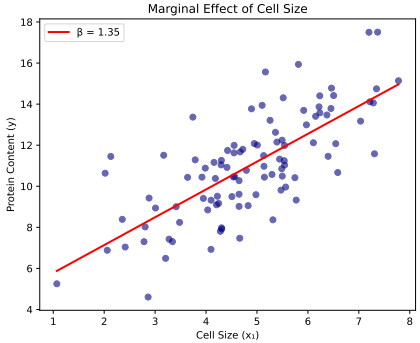
With estimates $\hat{\beta}_0 = 2.1$, $\hat{\beta}_1 = 0.08$, and $\hat{\beta}_2 = 2.5$:

- $\hat{\beta}_1 = 0.08$: 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.

This ceteris paribus interpretation is central to multiple regression.

Marginal vs. Conditional Effects

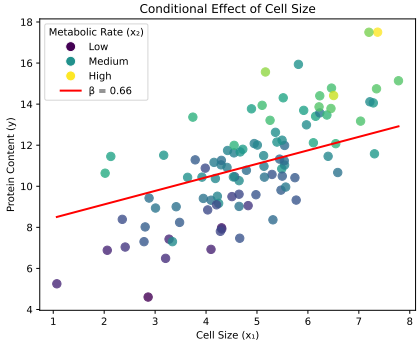
When predictors are correlated, marginal and conditional effects can differ substantially:



The marginal effect (simple regression) doesn't account for other variables that may influence the relationship.

Marginal vs. Conditional Effects

Conditional effects show the relationship after controlling for other variables:



Notice how the slope changes when we control for another predictor - this reveals the unique contribution of each variable.

Standardized Coefficients

Comparing coefficients can be challenging when predictors have different units of measurement.

Standardized coefficients convert all variables to a common scale:

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

These represent the expected change in the response (in standard deviation units) for a one standard deviation increase in the predictor.

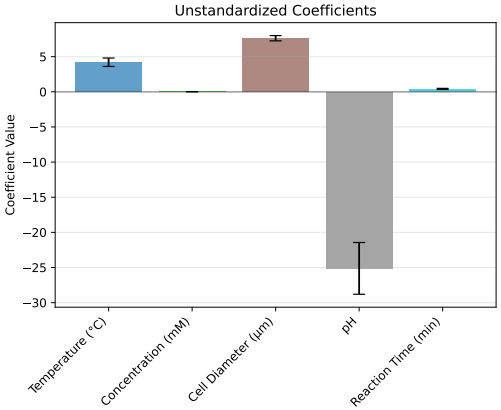
Benefits:

- Allow direct comparison of predictor influence regardless of original units
- Help identify which variables have the strongest relationship with the outcome
- Useful when predictors are measured on vastly different scales (e.g., age in years vs. concentration in nanomolar)

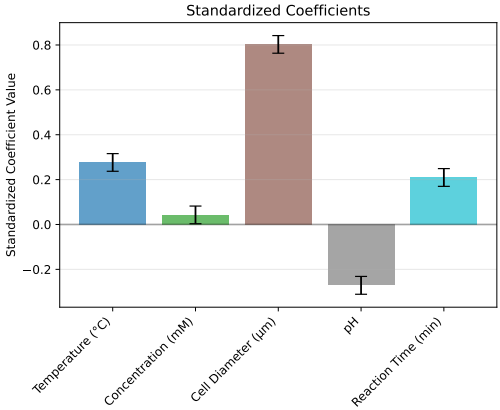
Limitation: Values become more abstract as they're expressed in standard deviation units rather than natural units.

Standardized vs. Unstandardized Coefficients: Visual Comparison

Unstandardized



Standardized



Notice how the relative importance of variables changes after standardization, revealing which predictors have the strongest relationship with the outcome variable.

Model Comparison in Multiple Regression

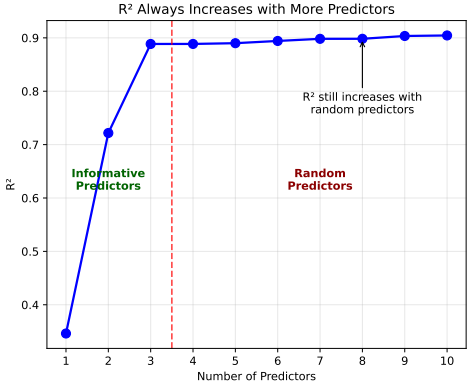
Assessing Model Fit: R^2

The coefficient of determination (R^2) quantifies the proportion of variance explained by the model:

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

R^2 ranges from 0 (model explains none of the variability) to 1 (model explains all variability).
However, R^2 has a critical flaw in multiple regression: it always increases (or at least never decreases) when additional predictors are added, even if they're just random noise.

The Problem with R^2 in Multiple Regression



R^2 continues to increase as predictors are added, even when they contribute only random noise (after predictor 3). This makes R^2 problematic for comparing models with different numbers of predictors.

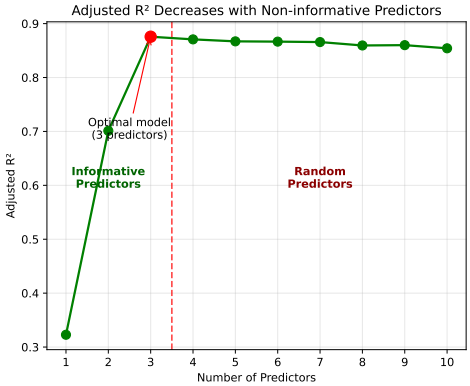
Adjusted R^2: Penalizing Model Complexity

The adjusted R^2 introduces a penalty for including additional predictors:

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

Unlike 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.

Adjusted R^2 for Model Selection



The adjusted R^2 decreases when non-informative predictors are added, making it more suitable for model selection than the regular R^2.

Comparing Nested Models

Two models are considered **nested** when one model contains a subset of the predictors in the other model.

Example: A study examining factors affecting enzyme activity

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

Research question: Do temperature and pH significantly affect enzyme activity after accounting for substrate concentration and ionic strength?

Model comparison:

$$\text{Restricted: } Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$$

$$\text{Unrestricted: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Likelihood Ratio Test for Nested Models

The likelihood ratio test (LRT) compares the maximum likelihood achieved by each model:

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

For linear regression with normally distributed errors, this relates to the residual sum of squares:

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

Under the null hypothesis, this LR statistic follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

Applying the Likelihood Ratio Test

For our enzyme activity example with 50 experimental runs:

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:

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

With 2 degrees of freedom, the critical value at $\alpha = 0.05$ is 5.99.
Since $16.8 > 5.99$, we reject the null hypothesis and conclude that temperature and pH significantly improve the model's fit.

ANOVA in the Regression Framework

Analysis of Variance: Comparing Multiple Groups

Analysis of Variance (ANOVA) extends hypothesis testing beyond two groups. Instead of performing multiple pairwise comparisons, ANOVA tests a single null hypothesis:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

Traditional ANOVA decomposes variation into:

- Between-group variation: Differences among group means
- Within-group variation: Differences within each group

The test statistic is the F-ratio:

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

ANOVA as Multiple Regression

ANOVA can be reformulated as linear regression with categorical predictors.
For comparing bacterial strains (A, B, C, D) on growth rates:

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

Where:

- $X_{Ji} = 1$ if observation i is strain J, 0 otherwise (for $J \in \{B, C, D\}$)

Interpretation:

- β_0 = mean for strain A (reference group)
- β_j = difference between strain j and strain A

The ANOVA null hypothesis becomes: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Testing in ANOVA Using Model Comparison

In the regression framework, ANOVA becomes a comparison between two nested models:

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

We can test this using the F-statistic:

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

This is identical to the traditional ANOVA F-ratio:

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

Connecting Traditional ANOVA and Regression: Sums of Squares

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:

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

Connecting Traditional ANOVA and Regression: Test Statistics

The relationship between likelihood ratio (LR) and F-statistics can be shown through Taylor approximation:

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

Therefore:

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

Sample Question 1

In multiple linear regression, what does the coefficient β_j represent?

A) The change in all predictors together
B) The marginal effect of predictor x_j
C) The effect of x_j holding all other predictors constant
D) The average residual error across all observations
E) The change in the response if all predictors change by one unit

Sample Question 2

Which of the following statements is true about the design matrix X in multiple regression?

A) Each column corresponds to an observation
B) The last column always contains the response variable
C) The first column is always a vector of zeros
D) Each row corresponds to one observation and the first column contains 1s for the intercept
E) It contains only the residual values

Sample Question 3

Which statement best describes the difference between **marginal** and **conditional** effects in regression?

A) Marginal effects control for other variables, conditional do not
B) Marginal effects only exist in univariate regression
C) Conditional effects represent the effect of a variable while holding others constant
D) Marginal effects are more accurate than conditional ones
E) Conditional effects are only used in ANOVA

Sample Question 4

Why is **adjusted** R^2 preferred over R^2 for comparing multiple regression models?

A) It provides the exact same value as R^2 for all models

B) It always increases when more predictors are added

C) It penalizes model complexity and can decrease if non-informative variables are added

D) It is easier to calculate

E) It only applies to nested models

Sample Question 5

In the context of comparing nested models using a likelihood ratio test (LRT), what does a **large** test statistic indicate?

A) The null model fits the data significantly better
B) The additional predictors in the full model do not improve fit
C) The restricted model should be preferred
D) The full model fits significantly better than the restricted model
E) The models are not nested

Sample Question 6

A biologist models protein content (Y) using two predictors: cell size (X_1) and metabolic rate (X_2). For cells with $X_1 = 100\ \mu m^3$ and $X_2 = 3.5$ AU, the 95% **confidence interval** for the **mean protein content** is [8.2, 9.6], while the 95% **prediction interval** for **a new observation** is [6.1, 11.8].

Which of the following statements is most accurate?

A) The confidence interval is wider because it includes both model and individual-level uncertainty
B) If the experiment is repeated, 95% of individual values will fall within [8.2, 9.6]
C) The prediction interval reflects greater uncertainty because it includes random variation across individual cells
D) The true mean protein content for $X_1 = 100$ and $X_2 = 3.5$ is between 6.1 and 11.8 with 95% probability
E) The intervals would be identical if the model had only one predictor instead of two