Multiple Linear Regression Extension

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Multiple predictors model

Simple vs Multiple Regression

Aspect	Simple
Predictors	1 variable
Model	$Y = \beta_0 + \beta_1 X$
Aspect	**Multiple**
Predictors	p variables
Model	$Y = X\beta$

Matrix Notation

$$Y = X\beta + \varepsilon$$

$$X \in \mathbb{R}^{n \times (p+1)}$$

Same Solution

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Normal equation applies

Model Fit Metrics

R^2 (R-squared)

$$R^2 = 1 - SSE/SST$$

Proportion of variance explained

Adjusted R^2

Penalizes model complexity

⚠ Multicollinearity

Correlated predictors
cause instability in estimates

Challenges

- Increased complexity
- More parameters to estimate
- Risk of overfitting
- Interpretation becomes harder

Feature Selection

Choosing relevant predictors
to improve model performance

Coefficient Interpretation

β_j is the effect of X_j
holding all other predictors constant

Understanding Multicollinearity in Detail

📊 Definition

High correlation between two or more predictor variables

$$Corr(X_i, X_j) \approx \pm 1$$

Makes it difficult to isolate individual effects

⚠️ Problems

1. Unstable estimates:
$\hat{\beta}$ coefficients become highly sensitive to small data changes

2. Large standard errors:
Var($\hat{\beta}$) increases dramatically

3. Poor interpretability:
Cannot determine which variable is truly important

$(X^TX)^{-1}$ becomes ill-conditioned

🔍 Detection Methods

1. Correlation Matrix:
Check pairwise correlations

$$|corr| > 0.8 \rightarrow \text{Warning}$$

2. VIF (Variance Inflation Factor):

$$VIF = 1/(1 - R^2_j)$$

R^2_j: R^2 when X_j is regressed on other predictors

VIF > 10 → Serious multicollinearity
VIF > 5 → Moderate concern

💡 Solutions

1. Remove variables:
Drop one of the highly correlated predictors

2. Combine variables:
Create composite index (e.g., PCA)

3. Regularization:
Ridge regression (L2) or Lasso (L1)

$$\text{Ridge: } \min ||Y - X\beta||^2 + \lambda||\beta||^2$$

📝 Example Scenario

Predicting house price with:

- Square footage
- Number of rooms

These are highly correlated!
Larger houses → more rooms

Solution: Use only one,
or create "size index"

🎯 Key Insights

Important: Multicollinearity affects *interpretation* and *stability*, but NOT prediction accuracy

The model can still predict well, but we cannot trust individual coefficient values

Overall model fit (R^2) remains good
Individual t-tests become unreliable