Multivariate Regression

Advanced Methods for Multivariate Analysis

Juliho Castillo Colmenares

Tec de Monterrey

Today's Agenda

1. Logistic Regression Model
2. Inferences for Variances and Covariance Matrices
3. Inferences for a Vector of Means
4. MANOVA (Multivariate Analysis of Variance)
5. Canonical Correlation Analysis
6. Factor Analysis with Regression
7. Programming and Commercial Systems

Case Study: Healthcare Risk Assessment

Companion Example Throughout This Presentation

A hospital study analyzing cardiovascular disease (CVD) risk using multivariate methods to:

- Predict high-risk patients from lifestyle and physiological factors
- Compare health profiles between risk groups
- Evaluate the effectiveness of lifestyle interventions
- Understand complex relationships between multiple health variables

Real dataset with 1,000 patients demonstrating practical applications of each technique

The Research Question

Scenario: Hospital evaluating cardiovascular disease (CVD) risk

Dataset:

- 1,000 patients
- 13 predictor variables
- Multiple health outcomes
- Treatment intervention study

Variables in Our Study

Lifestyle Factors:

- **Age:** Years lived, affects baseline health risk
- **BMI:** Body Mass Index, ratio of weight to height indicating body composition
- **Exercise hours/week:** Physical activity level, linked to cardiovascular health
- **Smoking years:** Duration of tobacco use, major risk factor for multiple diseases
- **Alcohol consumption:** Frequency/amount of alcohol intake, impacts liver and heart health

- **Stress score:** Quantified psychological stress level, affects blood pressure and hormones
- **Sleep hours:** Nightly sleep duration, influences metabolic and cardiovascular function

Physiological Measurements:

- **Blood pressure (systolic/diastolic):** Force of blood against artery walls (peak/resting)
- **Cholesterol:** Blood lipid levels, excess increases arterial plaque buildup
- **Glucose:** Blood sugar level, indicator of diabetes risk and metabolic function

- **Triglycerides:** Fat molecules in blood, high levels increase heart disease risk
- **HDL:** "Good cholesterol" that removes harmful lipids from bloodstream

Research Objectives

1. **Predict** CVD risk from patient characteristics
2. **Compare** health profiles between risk groups
3. **Evaluate** lifestyle intervention effectiveness
4. **Understand** relationships between lifestyle and physiology
5. **Validate** assumptions for multivariate tests

Logistic Regression

Moving Beyond Linear Regression

What is Logistic Regression?

A statistical method for modeling binary outcomes (Yes/No, Success/Failure, 0/1) using predictor variables

Key Features:

- Predicts probabilities (0 to 1) rather than continuous values
- Uses the logistic (sigmoid) function to model the probability of an event
- Estimates coefficients via maximum likelihood (not least squares)

- Widely used in classification problems: medical diagnosis, credit scoring, marketing

When to Use: When your outcome variable is categorical (especially binary)

When Linear Regression Fails

Problem: Binary outcomes (Yes/No, Success/Failure, 0/1)

Using ordinary least squares (OLS) for binary responses creates fundamental problems:

Prediction Issues:

- Predicted probabilities can be negative or exceed 1 (nonsensical values)
- No mechanism to constrain predictions to [0, 1]

Assumption Violations:

- Response is not continuous (violates normality assumption)

- Errors follow Bernoulli distribution, not Normal distribution
- Heteroscedastic errors (variance depends on X: $\text{Var}(Y|X) = p(X)(1 - p(X)))$
- Non-constant variance violates homoscedasticity assumption

Linear regression is mathematically possible but statistically inappropriate for binary data.

Logistic Regression Solution

Our goal is to find a function that maps predictor values to probabilities while staying within $[0, 1]$. Linear regression fails here because predictions can fall outside this valid probability range.

Given a binary outcome variable Y (taking values 0 or 1 for failure or success) and a vector of predictor variables $X = \left(X_1, X_2, ..., X_p\right)$, we model the probability of success as:

$$P(Y = 1 \mid X) = p(X)$$

where $0 \leq p(X) \leq 1$ represents the probability that $Y = 1$ given the values of X. This approach ensures our predictions always remain valid probabilities.

The Logit Transformation

To connect probabilities (bounded between 0 and 1) to a linear combination of predictors (unbounded), we use the **logit transformation**, also known as the **log-odds**.

Recall that $p(X) = P(Y = 1 \mid X)$ is our probability of success. The odds of success are $\frac{p(X)}{1-p(X)}$, and taking the natural logarithm gives us:

$$\text{logit}(p(X)) = \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

where:

- $p(X)$ is the same probability function defined in the previous slide

- $\beta_0, \beta_1, ..., \beta_p$ are coefficients to be estimated
- The logit maps probabilities from $[0, 1]$ to $(-\infty, +\infty)$
- This transformation makes the model linear in the parameters

The Logit Transformation

Properties:

- Maps [0,1] to $(-\infty, +\infty)$
- Linear in parameters
- Interpretable as log-odds ratio

The Logistic Function

Inverse Logit: Solving for $p(X)$ from the logit equation gives the logistic function.

Notation: $X = (X_1, X_2, ..., X_p)$ is the vector of p predictor variables. We write $p(X)$ to mean the probability depends on all predictors.

Define the linear combination: $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$

Then the logistic function is:

$$p(X) = \frac{e^{\eta}}{1 + e^{\eta}} = \frac{1}{1 + e^{-\eta}}$$

This is the **logistic** or **sigmoid** function, guaranteeing $p(X) \in [0, 1]$ for any predictor values.

Maximum Likelihood Estimation

Since we cannot use ordinary least squares, we estimate coefficients using **Maximum Likelihood Estimation (MLE)**.

Our Data: We have collected n observations. Each observation consists of:

- An outcome: $y_i \in \{0, 1\}$ (e.g., $y_i = 1$ means "has disease", $y_i = 0$ means "no disease")
- Predictor values: $x_{i1}, x_{i2}, \ldots, x_{ip}$ (the values of p predictors for person i)

Unknown Parameters: We need to estimate $(p + 1)$ coefficients

- β_0 = intercept coefficient

- $\beta_1, \beta_2, ..., \beta_p$ = coefficients for the p predictors

We write $\boldsymbol{\beta} = \left(\beta_0, \beta_1, ..., \beta_p\right)^T$ for the entire vector of unknown parameters.

Maximum Likelihood Estimation

The Probability Model: For each observation i, the probability of success depends on:

1. The predictor values for that observation: $x_{i1}, x_{i2}, ..., x_{ip}$
2. The unknown parameters: $\beta_0, \beta_1, ..., \beta_p$

Define the linear combination: $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$

The probability is computed using the logistic function:

$$P\left(Y_i = 1 \mid x_{i1}, ..., x_{ip}; \beta_0, ..., \beta_p\right) = \frac{1}{1 + e^{-\eta_i}}$$

We write this as $\pi_i(\beta)$ for brevity, emphasizing it depends on the parameters β.

Important: Each $\pi_i(\beta)$ depends on BOTH the observed predictors for observation i AND the unknown parameters β.

Maximum Likelihood Estimation

Probability of Observing Outcome y_i: Given parameters β, the probability of observing the actual outcome y_i for observation i is formally written as $P(Y_i = y_i \mid ...)$, where Y_i is the random variable for the outcome. The following is a common shorthand:

$$P(Y_i = y_i \mid \boldsymbol{x}_i, \beta) = [\pi_i(\beta)]^{y_i} [1 - \pi_i(\beta)]^{1-y_i}$$

This formula evaluates to:

- $\pi_i(\beta)$ when the observed outcome is $y_i = 1$ (success)
- $1 - \pi_i(\beta)$ when the observed outcome is $y_i = 0$ (failure)

Key Idea of MLE: Find the parameter values β that make the observed data $(y_1, y_2, \ldots, y_n)$ most probable.

Maximum Likelihood Estimation

Likelihood Function: Assuming observations are independent, the probability of observing ALL our data is:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y_i = y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = \prod_{i=1}^{n} [\pi_i(\boldsymbol{\beta})]^{y_i} [1 - \pi_i(\boldsymbol{\beta})]^{1-y_i}$$

Log-Likelihood Function: Taking natural logarithm (easier to maximize):

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(\pi_i(\boldsymbol{\beta})) + (1 - y_i) \log(1 - \pi_i(\boldsymbol{\beta}))]$$

Goal: Find β^* that maximizes $\ell(\beta)$. This requires numerical optimization (Newton-Raphson, gradient descent, etc.) since no closed-form solution exists.

Interpreting Coefficients: Log-Odds vs. Odds Ratios

Recall the logit equation:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

1. The Raw Coefficient (β_j)

- A one-unit increase in X_j corresponds to a β_j change in the **log-odds** of the outcome.
- This is mathematically precise but not intuitive. A "0.2 change in log-odds" is hard to grasp.

2. The Odds Ratio (e^{β_j}) – A Better Way

- By exponentiating the coefficient, we get the **Odds Ratio** (OR).
- The OR tells us the **multiplicative change** in the **odds** of the outcome for a one-unit increase in X_j.

Example: Stress Score ($\beta_{\text{stress}} = 0.223$)

The odds ratio is $e^{0.223} \approx 1.25$.

Interpretation: For each one-point increase in the stress score, the **odds** of having high CVD risk are multiplied by 1.25 (i.e., increase by 25%), holding other variables constant.

Interpreting Coefficients: The Odds Ratio (cont.)

Interpreting Different Odds Ratios:

If β_j is...	Then e^{β_j} is...	Meaning for a 1-unit increase in X_j...
Positive (> 0)	Greater than 1	The odds of the outcome **increase**. (e.g., OR=1.25 means 25% increase in odds)

Negative (< 0)	Less than 1	The odds of the outcome **decrease**. (e.g., OR=0.80 means 20% decrease in odds)
Zero	Exactly 1	There is no change in the odds of the outcome.

Example: Exercise ($\beta_{\text{exercise}} = -0.328$)

The odds ratio is $e^{-0.328} \approx 0.72$.

Interpretation: For each additional hour of exercise per week, the **odds** of having high CVD risk are multiplied by 0.72 (i.e., decrease by 28%), holding other variables constant.

Model Fit and Diagnostics

Assessing how well our logistic regression model explains the data is crucial. Unlike linear regression, we use different metrics.

Deviance:

- A key metric for models where ordinary least squares doesn't apply (like logistic regression). It's a measure of how much your model's predictions deviate from the actual data.
- Think of it as the equivalent of the **Sum of Squared Errors** you see in linear regression.
- A lower deviance means your model fits the data more closely.

- For a deeper dive into deviance, see this video: Deviance Residuals Explained

Pseudo R-squared:

- While linear regression has a standard R-squared to measure goodness-of-fit, logistic and other generalized linear models use "Pseudo" R-squared.
- It tells you how much better your model is than a "null" model that only includes an intercept (essentially, a model that just guesses the average outcome).
- **McFadden's R^2** is a popular version:

$$R^2_{\text{McF}} = 1 - \frac{\log(\mathcal{L}_{\text{model}})}{\log(\mathcal{L}_{\text{null}})}$$

.

- A higher value (closer to 1) indicates a better fit, but the scale is not directly comparable to the R-squared from linear regression.

Classification Performance

Confusion Matrix:

	Predicted 0	**Predicted 1**
Actual 0	True Negative (TN)	False Positive (FP)
Actual 1	False Negative (FN)	True Positive (TP)

Classification Metrics

Accuracy: $\frac{\text{TP} + \text{TN}}{n}$

Sensitivity (Recall): $\frac{\text{TP}}{\text{TP} + \text{FN}}$

Specificity: $\frac{\text{TN}}{\text{TN} + \text{FP}}$

Precision: $\frac{\text{TP}}{\text{TP} + \text{FP}}$

Case Study: Predicting CVD Risk

Application of Logistic Regression

Objective: Predict high CVD risk (0/1) from 13 patient characteristics

CVD Prediction: Model Setup

Predictors (13 variables):

- Demographics: age, BMI
- Lifestyle: exercise, smoking, alcohol, stress, sleep
- Physiology: BP, cholesterol, glucose, triglycerides, HDL

Outcome: CVD risk high (binary: 0 = low risk, 1 = high risk)

Data split: 70% training (n=700), 30% testing (n=300)

Top Risk Factors: Odds Ratios

Predictor	Odds Ratio	Interpretation
Exercise hours/week	0.72	28% lower odds per hour
Stress score	1.25	25% higher odds per point
Sleep hours	0.80	20% lower odds per hour
BMI	1.19	19% higher odds per unit

All significant predictors contribute to risk assessment

CVD Prediction: Model Performance

Confusion Matrix (Test Set):

	Pred Low	Pred High
Actual Low	106	44
Actual High	42	108

Metrics:

- Accuracy: 71%
- AUC-ROC: 0.77
- Balanced precision/recall

Key Insights: CVD Prediction

1. Exercise is the strongest protective factor (OR = 0.72)
2. Stress significantly increases risk (OR = 1.25)
3. Model achieves good discrimination (AUC = 0.77)
4. Can identify high-risk patients for intervention

Clinical Value: Early identification enables preventive care

Inferences for Covariance Matrices

Testing Variability Structure

Why Test Covariance Matrices?

Applications:

- Homogeneity assumptions in MANOVA
- Comparing variability between groups
- Validating models
- Quality control

The Wishart Distribution

Multivariate Generalization of Chi-Square

If $X_1, ..., X_n \sim N_{p(\mu,\Sigma)}$, then:

$$S \sim W_{p(n-1,\Sigma)}$$

where $S = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T$

Testing Single Covariance Matrix

Null Hypothesis:

$$H_0 : \Sigma = \Sigma_0$$

Test Statistic: Based on likelihood ratio

$$\Lambda = |S\Sigma_0^{-1}|$$

Box's M Test

Testing Equality of Covariance Matrices

$$H_0 : \Sigma_1 = \Sigma_2 = ... = \Sigma_g$$

Box's M Test Statistic

$$M = (n - g) \log|S_{\text{pooled}}| - \sum_{i=1}^{g} (n_i - 1) \log|S_i|$$

where:

- S_i = covariance matrix for group i
- S_{pooled} = pooled covariance matrix

Box's M Test Properties

Asymptotic Distribution: Chi-square for large samples

Limitation: Very sensitive to normality violations

Alternatives: Permutation tests, robust methods

Bartlett's Test for Univariate Data

Special Case: Testing equality of variances (p=1)

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_g^2$$

Test Statistic: Chi-square distributed

Case Study: Validating MANOVA Assumptions

Application of Box's M Test

Question: Are covariance matrices equal between treatment groups (MANOVA assumption)?

Box's M Test: Setup

Testing Homogeneity of Covariances:

$$H_0 : \Sigma_{\text{Control}} = \Sigma_{\text{Intervention}}$$

Variables (p = 4):

- Systolic BP, Diastolic BP
- Cholesterol, Glucose

Groups:

- Control: n = 479
- Intervention: n = 521

Why test? MANOVA assumes equal covariance matrices across groups

Box's M Test: Results

Test Statistic:

$$M = 8.49$$

Degrees of Freedom: 10

Interpretation: $M < 30$ (rule of thumb)

Conclusion: Covariance matrices are approximately equal. MANOVA assumption satisfied.

Implication: Our MANOVA results are valid and trustworthy

Key Insights: Assumption Testing

1. Box's M test validates MANOVA assumptions
2. Equal covariances ensure valid inference
3. Small M statistic (8.49) indicates homogeneity
4. Treatment groups have similar variability patterns

Methodological importance: Always check assumptions before interpreting results

Inferences for a Vector of Means

Multivariate Hypothesis Testing

From t-test to Hotelling's T-squared

Univariate: t-test for single mean

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Multivariate: Hotelling's T^2 for mean vector

Hotelling's T-squared Test

One-Sample Test:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

Test Statistic:

$$T^2 = n\big(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0\big)^T S^{-1}\big(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0\big)$$

Distribution of T-squared

Transform to F Distribution:

$$F = \frac{(n-p)T^2}{(n-1)p} \sim F_{p,n-p}$$

where:

- p = number of variables
- n = sample size

Two-Sample Hotelling's T-squared

Testing Difference Between Groups:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

Two-Sample T-squared Statistic

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2} \right) (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^T S_{\text{pooled}}^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$$

F Transformation:

$$F = \frac{(n_1 + n_2 - p - 1) T^2}{(n_1 + n_2 - 2) p} \sim F_{p, n_1 + n_2 - p - 1}$$

Confidence Region for Mean Vector

Multivariate Confidence Region:

Ellipsoid centered at $\bar{\boldsymbol{X}}$

$$n\left(\boldsymbol{\mu}-\bar{\boldsymbol{X}}\right)^{T} S^{-1}\left(\boldsymbol{\mu}-\bar{\boldsymbol{X}}\right) \leq \frac{(n-1)p}{n-p} F_{\alpha;p,n-p}$$

Simultaneous Confidence Intervals

Bonferroni Correction:

For p variables, use $\frac{\alpha}{p}$ for each interval

T-squared Intervals: More efficient but wider than individual intervals

Case Study: Comparing Risk Groups

Application of Hotelling's T-squared

Question: Do high-risk and low-risk CVD patients differ in their multivariate health profile?

Health Profile Comparison: Setup

Two Groups:

- Low risk: n = 500
- High risk: n = 500

Variables (p = 6):

- Systolic BP, Diastolic BP
- Cholesterol, Glucose
- Triglycerides, HDL

Goal: Single omnibus test for all 6 variables simultaneously

Mean Differences by Risk Group

Variable	Low Risk	High Risk	Difference
Systolic BP	123.8	131.1	+7.3
Diastolic BP	78.1	82.9	+4.7
Cholesterol	184.1	196.4	+12.3
Glucose	110.0	116.8	+6.9
Triglycerides	133.6	145.5	+11.9
HDL	44.5	41.2	−3.3

Hotelling's T-squared Results

Test Statistic:

$$T^2 = 228.65$$

F Transformation:

$$F = 37.92, \quad \text{df} = (6, 993)$$

P-value: < 0.0001

Conclusion: Strong evidence that high-risk and low-risk patients have significantly different health profiles

Key Insights: Risk Group Differences

1. High-risk patients show higher values across all adverse markers
2. Largest differences: cholesterol (+12.3 mg/dL) and triglycerides (+11.9 mg/dL)
3. HDL (protective) is lower in high-risk group (-3.3 mg/dL)
4. Multivariate test accounts for correlations among measurements

Clinical significance: Pattern of differences suggests metabolic syndrome

MANOVA

Multivariate Analysis of Variance

What is MANOVA?

Extension of ANOVA to Multiple Dependent Variables

- ANOVA: One response variable
- MANOVA: Multiple response variables simultaneously

Why Use MANOVA?

Instead of Multiple ANOVAs:

1. Controls Type I error rate
2. Accounts for correlations among responses
3. More powerful when responses related
4. Tests overall group effect

MANOVA Model

One-Way MANOVA:

$$\boldsymbol{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij}$$

where:

- $\boldsymbol{Y}_{ij}$ = response vector for observation j in group i
- $\boldsymbol{\mu}$ = overall mean vector
- $\boldsymbol{\alpha}_i$ = group effect vector
- $\boldsymbol{\varepsilon}_{ij}$ = error vector

MANOVA Assumptions

1. **Multivariate Normality:** Errors follow multivariate normal
2. **Independence:** Observations independent
3. **Homogeneity of Covariance:** Equal covariance matrices across groups

Testing Assumptions

Multivariate Normality:

- Mardia's test
- Q-Q plots for each variable

Homogeneity: Box's M test

MANOVA Matrices

Between-Groups Matrix (H):

$$\boldsymbol{H} = \sum_{i=1}^{g} n_i \left(\bar{\boldsymbol{Y}}_i - \bar{\boldsymbol{Y}}\right)\left(\bar{\boldsymbol{Y}}_i - \bar{\boldsymbol{Y}}\right)^T$$

Within-Groups Matrix (E):

$$\boldsymbol{E} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_i\right)\left(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_i\right)^T$$

Wilks' Lambda

Most Common Test Statistic:

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|}$$

Wilks' Lambda Properties

Interpretation:

- Range: [0, 1]
- Small values: Strong group differences
- Lambda = 1: No group differences

Represents: Proportion of total variance not explained by groups

Other MANOVA Test Statistics

Pillai's Trace:

$$V = \text{tr}\big(\boldsymbol{H}(\boldsymbol{H}+\boldsymbol{E})^{-1}\big)$$

Hotelling-Lawley Trace:

$$U = \text{tr}\big(\boldsymbol{H}\boldsymbol{E}^{-1}\big)$$

Roy's Largest Root: Largest eigenvalue of $\boldsymbol{H}\boldsymbol{E}^{-1}$

Choosing Test Statistic

Statistic	Best When
Wilks' Lambda	General use (most common)
Pillai's Trace	Robust to violations
Hotelling-Lawley	Equal group sizes
Roy's Root	Group difference on one dimension

Post-Hoc Tests in MANOVA

After Significant MANOVA:

1. Univariate ANOVAs (with correction)
2. Discriminant analysis
3. Contrast tests for specific hypotheses

Case Study: Treatment Intervention

Application of MANOVA

Question: Does a lifestyle intervention improve multiple health outcomes simultaneously?

Treatment Intervention Study: Setup

Groups:

- Control: n = 479 (standard care)
- Intervention: n = 521 (lifestyle program)

Outcomes (p = 4):

- Systolic BP
- Diastolic BP
- Cholesterol
- Glucose

Why MANOVA? Controls Type I error while testing all outcomes together

MANOVA Results: Treatment Effect

Wilks' Lambda: $\Lambda = 0.889$

F Statistic: $F = 31.05$, df = (4, 995)

P-value: < 0.0001

Conclusion: The intervention significantly improves health outcomes across the multivariate profile

Also significant: Pillai's trace, Hotelling-Lawley, Roy's root (all p < 0.0001)

Mean Improvements by Treatment Group

Outcome	Control	Intervention	Difference
Systolic BP	130.8	124.3	−6.5***
Diastolic BP	82.7	78.5	−4.2***
Cholesterol	194.4	186.5	−7.9***
Glucose	116.3	110.7	−5.6***

*** All differences significant at $p < 0.001$ in follow-up ANOVAs

Key Insights: Intervention Effects

1. Intervention reduces all cardiovascular risk markers
2. Largest effect on blood pressure (-6.5 / −4.2 mmHg)
3. Clinically meaningful reductions in cholesterol and glucose
4. MANOVA provides single omnibus test (no Type I error inflation)

Clinical significance: Comprehensive lifestyle changes yield broad health benefits

Canonical Correlation Analysis

Relating Two Sets of Variables

What is Canonical Correlation?

Purpose: Find maximum correlation between linear combinations of two sets of variables

- Set 1: $X_1, X_2, ..., X_p$
- Set 2: $Y_1, Y_2, ..., Y_q$

Canonical Correlation vs. Other Methods

Method	Set 1	Set 2
Correlation	1 variable	1 variable
Multiple Regression	Multiple	1 variable
Canonical Correlation	Multiple	Multiple

Canonical Variates

First Canonical Variate Pair:

$$U_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

$$V_1 = b_{11}Y_1 + b_{12}Y_2 + \ldots + b_{1q}Y_q$$

such that $\text{cor}(U_1, V_1)$ is maximized

Number of Canonical Correlations

How Many Pairs?

$$k = \min(p, q)$$

Each subsequent pair:

- Uncorrelated with previous pairs
- Maximizes remaining correlation

Canonical Correlation Coefficients

Ordering:

$$\rho_1 \geq \rho_2 \geq \ldots \geq \rho_k \geq 0$$

where ρ_i is the i-th canonical correlation

Testing Significance

Test All Correlations:

$$H_0 : \rho_1 = \rho_2 = ... = \rho_k = 0$$

Test Remaining Correlations:

$$H_0 : \rho_{m+1} = ... = \rho_k = 0$$

Wilks' Lambda for Canonical Correlation

$$\Lambda = \prod_{i=1}^{k} (1 - \rho_i^2)$$

Approximate chi-square distribution for testing

Canonical Loadings

Structure Coefficients:

Correlation between original variables and canonical variates

- Help interpret meaning of canonical variates
- More stable than canonical weights

Redundancy Analysis

Proportion of Variance Explained:

How much variance in one set is explained by the other set through canonical variates

$$\text{Redundancy} = \left(\frac{1}{p}\right) \sum_{j=1}^{p} R^2_{X_j, V_1}$$

Interpreting Canonical Correlations

1. **Examine significance:** Are correlations statistically significant?
2. **Check magnitude:** Are correlations practically meaningful?
3. **Interpret loadings:** What do canonical variates represent?
4. **Assess redundancy:** How much variance explained?

Case Study: Lifestyle vs. Physiology

Application of Canonical Correlation

Question: How do lifestyle factors relate to physiological health markers?

Lifestyle-Physiology Relationship: Setup

Set 1 - Lifestyle Factors (p = 5):

- Exercise hours/week
- Smoking years
- Alcohol units/week
- Stress score
- Sleep hours

Set 2 - Physiological Markers (q = 6):

- Systolic BP, Diastolic BP
- Cholesterol, Glucose
- Triglycerides, HDL

Maximum pairs: $\min(5, 6) = 5$

Canonical Correlations: Results

Pair	Correlation	Interpretation
1	0.639	Strong relationship
2	0.244	Moderate relationship
3-5	< 0.12	Weak relationships

Focus on first canonical correlation (r = 0.639)

First Canonical Variate: Lifestyle

Canonical Loadings (Structure Coefficients):

Variable	Loading
Exercise hours	+0.65
Stress score	−0.53
Alcohol units	−0.37
Sleep hours	+0.33
Smoking years	−0.27

Interpretation: Healthy lifestyle pattern (more exercise, less stress)

First Canonical Variate: Physiology

Canonical Loadings:

Variable	Loading
Diastolic BP	−0.70
Systolic BP	−0.68
Cholesterol	−0.65
HDL	+0.65
Glucose	−0.61
Triglycerides	−0.59

Interpretation: Favorable health profile (lower BP, higher HDL)

Key Insights: Lifestyle-Physiology Link

1. Strong canonical correlation (r = 0.639) between lifestyle and health
2. Healthy lifestyle pattern $\longrightarrow$ Favorable physiological profile
3. Exercise and low stress most important lifestyle factors
4. Blood pressure and cholesterol most related physiological markers

Clinical significance: Lifestyle interventions can meaningfully improve multiple health markers

Factor Analysis with Regression

Combining Dimension Reduction and Prediction

The Multicollinearity Problem

Issue: Highly correlated predictors in regression

Consequences:

- Unstable coefficient estimates
- Large standard errors
- Difficult interpretation
- Poor prediction in new samples

Factor-Based Regression Solution

Strategy:

1. Extract factors from correlated predictors
2. Use factor scores as predictors
3. Fit regression with orthogonal factors

Factor-Based Regression Workflow

1. **Factor Analysis:** Extract factors from X variables
2. **Compute Factor Scores:** For each observation
3. **Regression:** Predict Y using factor scores
4. **Interpretation:** Results in terms of factors

Benefits of Factor-Based Regression

Advantages:

- Reduces multicollinearity (orthogonal factors)
- Dimensionality reduction (fewer predictors)
- Conceptual interpretation (latent constructs)
- More stable estimates

Comparing Approaches

Aspect	Direct Regression	Factor Regression
Multicollinearity	Problem	Eliminated
Interpretation	Original variables	Latent factors
Predictors	Many	Few
Variance explained	Higher	May be lower

Principal Components Regression

Alternative Approach:

Use PCA instead of factor analysis

Difference:

- PCA: Explains total variance
- FA: Explains common variance (removes unique variance)

Other Methods for Multicollinearity

Ridge Regression: Shrinks coefficients toward zero

Lasso: Variable selection via L1 penalty

Partial Least Squares: Finds components that predict Y well

Cautions and Limitations

Factor-Based Regression Limitations:

- Factor extraction somewhat subjective
- Results depend on specific sample
- Prediction requires computing factor scores with same loadings
- May lose some predictive information

Programming and Commercial Systems

Implementing Multivariate Methods

Python for Multivariate Analysis

Key Libraries:

- `statsmodels`: Statistical models and tests
- `scikit-learn`: Machine learning algorithms
- `numpy` / `scipy`: Numerical computations
- `pandas`: Data manipulation

Python: Logistic Regression

```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
probabilities = model.predict_proba(X_test)
```

Python: Hotelling's T-squared

```
from scipy.stats import chi2
import numpy as np

# Compute T-squared statistic
diff = mean1 - mean2
S_pooled_inv = np.linalg.inv(S_pooled)
T2 = (n1 * n2) / (n1 + n2) * diff.T @ S_pooled_inv @ diff

# Transform to F
p = len(mean1)
F_stat = ((n1 + n2 - p - 1) * T2) / ((n1 + n2 - 2) * p)
```

Python: MANOVA

```
from statsmodels.multivariate.manova import MANOVA

# Fit MANOVA model
manova = MANOVA.from_formula(
    'Y1 + Y2 + Y3 ~ Group',
    data=df
)

# Test results
print(manova.mv_test())
```

Python: Canonical Correlation

```
from sklearn.cross_decomposition import CCA

# Canonical correlation analysis
cca = CCA(n_components=2)
cca.fit(X_set, Y_set)

# Transform to canonical variates
X_c, Y_c = cca.transform(X_set, Y_set)

# Canonical correlations
correlations = [np.corrcoef(X_c[:, i], Y_c[:, i])[0, 1]
                for i in range(2)]
```

R for Multivariate Analysis

Key Packages:

- `stats`: Base statistical functions
- `MASS`: Advanced statistical methods
- `car`: Companion to Applied Regression
- `vegan`: Multivariate analysis

R: MANOVA Example

```r
# Fit MANOVA
model <- manova(cbind(Y1, Y2, Y3) ~ Group, data = df)

# Test results
summary(model, test = "Wilks")
summary(model, test = "Pillai")

# Follow-up univariate tests
summary.aov(model)
```

Commercial Software: SPSS

GUI-Based Analysis:

- Analyze > General Linear Model > Multivariate
- Analyze > Regression > Binary Logistic
- Analyze > Correlate > Canonical Correlation

Syntax: Also supports command syntax for reproducibility

Commercial Software: SAS

Key Procedures:

- `PROC LOGISTIC`: Logistic regression
- `PROC GLM`: General linear models (MANOVA)
- `PROC CANCORR`: Canonical correlation
- `PROC FACTOR`: Factor analysis

Software Comparison

Software	Strengths	Limitations
Python	Free, flexible, ML integration	Statistical testing less developed
R	Free, comprehensive stats	Steeper learning curve
SPSS	GUI, easy to learn	Expensive, less flexible
SAS	Enterprise, comprehensive	Very expensive, complex

Choosing Software

Considerations:

- Cost (free vs. commercial)
- Learning curve
- Specific methods needed
- Integration with workflow
- Reproducibility requirements
- Team expertise

Best Practices: Code Documentation

Essential Elements:

- Comment your code clearly
- Document data preprocessing steps
- Record package versions
- Save random seeds for reproducibility
- Version control (Git)

Best Practices: Workflow

1. **Data Cleaning:** Handle missing values, outliers
2. **Exploratory Analysis:** Visualize distributions
3. **Check Assumptions:** Test before analysis
4. **Run Analysis:** Use appropriate methods
5. **Validate Results:** Cross-validation, diagnostics
6. **Document:** Clear reporting

Case Study Summary

Healthcare Risk Assessment: What We Learned

Key Findings: Prediction and Classification

Logistic Regression Results:

- 71% accuracy predicting CVD risk (AUC = 0.77)
- Exercise strongest protective factor (OR = 0.72)
- Stress increases risk 25% per point (OR = 1.25)
- Model identifies high-risk patients for early intervention

Clinical Value: Enables targeted prevention strategies

Key Findings: Group Comparisons

Hotelling's T-squared:

- High-risk patients differ significantly across 6 health markers ($T^2 = 228.65$, $p < 0.0001$)
- Largest differences: cholesterol (+12.3) and triglycerides (+11.9)
- Pattern suggests metabolic syndrome

Box's M Test:

- Covariance matrices equal between groups ($M = 8.49$)
- MANOVA assumptions validated

Key Findings: Treatment Effectiveness

MANOVA Results:

- Intervention improves all health outcomes ($\Lambda = 0.889$, $p < 0.0001$)
- Blood pressure: −6.5 / −4.2 mmHg
- Cholesterol: −7.9 mg/dL
- Glucose: −5.6 mg/dL

Clinical Impact: Comprehensive lifestyle changes yield broad benefits

Key Findings: Lifestyle-Health Relationships

Canonical Correlation:

- Strong link between lifestyle and physiology (r = 0.639)
- Healthy lifestyle pattern: ↑ exercise, ↓ stress
- Favorable health profile: ↓ BP, ↑ HDL
- 40.8% shared variance between domains

Clinical Insight: Lifestyle interventions affect multiple health markers simultaneously

Methodological Insights

1. **Multivariate methods reveal patterns** missed by univariate tests
2. **Type I error control** critical with multiple outcomes
3. **Assumption testing** (Box's M) validates results
4. **Effect sizes matter** beyond statistical significance
5. **Clinical context** guides interpretation

All methods demonstrated with real healthcare data

Key Takeaways: Models

Logistic Regression:

- Use for binary outcomes
- Maximum likelihood estimation
- Interpret via odds ratios
- **Case Study:** 71% accuracy predicting CVD risk

Key Takeaways: Inference

Covariance Matrix Tests:

- Box's M test for equality
- Wishart distribution foundation
- **Case Study:** M = 8.49 (assumption satisfied)

Mean Vector Tests:

- Hotelling's T-squared generalizes t-test
- Confidence regions are ellipsoids
- **Case Study:** $T^2 = 228.65$ (strong group differences)

Key Takeaways: Advanced Methods

MANOVA:

- Multiple response variables simultaneously
- Wilks' Lambda most common test
- Controls Type I error
- **Case Study:** $\Lambda = 0.889$ (intervention effective)

Canonical Correlation:

- Relates two variable sets
- Multiple correlation pairs
- **Case Study:** r = 0.639 (lifestyle-health link)

Key Takeaways: Applications

Factor-Based Regression:

- Addresses multicollinearity
- Dimension reduction
- Interpretable factors

Software:

- Python: scikit-learn, statsmodels
- R: stats, MASS
- Commercial: SPSS, SAS
- **Case Study:** All analyses implemented in Python

Common Pitfalls to Avoid

1. Using logistic regression without checking convergence
2. Ignoring multicollinearity in regression
3. Not checking MANOVA assumptions (Box's M)
4. Over-interpreting weak canonical correlations
5. Using too many factors in factor-based regression

Method Selection Guide

Situation	Method
Binary outcome	Logistic regression
Multiple groups, multiple responses	MANOVA
Relate two variable sets	Canonical correlation
Multicollinear predictors	Factor/PCA regression

Recommended Resources: Books

Textbooks:

- Agresti (2018) - Introduction to Categorical Data Analysis
- Johnson & Wichern (2007) - Applied Multivariate Statistical Analysis
- Rencher & Christensen (2012) - Methods of Multivariate Analysis

Recommended Resources: Online

StatQuest YouTube Channel:

1. **Logistic Regression:**
   https://www.youtube.com/watch?v=yIYKR4sgzI8
2. **MANOVA Concepts:** Search "MANOVA StatQuest"
3. **PCA (for PCR):**
   https://www.youtube.com/watch?v=FgakZw6K1QQ

Recommended Resources: Software

Documentation:

- Scikit-learn: https://scikit-learn.org
- Statsmodels: https://www.statsmodels.org
- R Documentation: https://www.rdocumentation.org

Questions?

Thank you for your attention!

Juliho Castillo Colmenares
julihocc@tec

Office: Tec de Monterrey CCM Office 1540
Office Hours: Monday-Friday, 9:00 AM - 5:00 PM

Next Steps: This Week

For This Week:

- Review lecture notes thoroughly
- Practice with provided examples
- Complete practice questions
- Prepare for E07 quiz

Next Steps: Preparation for Evaluation

Key Topics to Master:

- Logistic regression: logit transformation, MLE, interpretation
- Hotelling's T-squared: computation and F transformation
- MANOVA: assumptions, Wilks' Lambda, interpretation
- Canonical correlation: number of pairs, loadings, significance
- Factor-based regression: workflow, benefits, limitations
- Software implementation: Python and R basics

Integration with Previous Topics

Building on Earlier Concepts:

- Factor Analysis (L04) ⟶ Factor-based regression
- Discriminant Analysis (L05) ⟶ MANOVA post-hoc
- PCA principles ⟶ Principal components regression

Comprehensive Framework: All methods part of the multivariate toolkit