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

Logistic Regression

Moving Beyond Linear Regression

When Linear Regression Fails

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

Linear regression assumptions violated:

- Response not continuous
- Errors not normal
- Predictions can exceed [0,1]

Logistic Regression Solution

Key Idea: Model the probability of success

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

where $0 \leq p(X) \leq 1$

The Logit Transformation

Logit (Log-Odds):

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

The Logit Transformation

Properties:

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

The Logistic Function

Inverse Logit:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

Also written as:

$$p(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_p X_p)}}$$

Why Not Ordinary Least Squares?

Problems with OLS for Binary Response:

- Predicted probabilities can be negative or exceed 1
- Errors follow Bernoulli distribution, not Normal
- Heteroscedastic errors
- Violates fundamental assumptions

Maximum Likelihood Estimation

Bernoulli Distribution:

$$P(Y_i = y_i \mid X_i) = p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$$

Maximum Likelihood Estimation

Log-Likelihood Function:

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

Goal: Find β that maximizes $\ell(\beta)$

Interpreting Coefficients

Coefficient β_j:

- One unit increase in X_j changes log-odds by β_j
- Odds ratio: e^{β_j}

Example: If $\beta_1 = 0.5$, then $e^{0.5} = 1.65$ means 65% increase in odds

Model Fit and Diagnostics

Deviance: Measures goodness of fit

$$D = -2\log(\mathcal{L})$$

Pseudo R-squared: McFadden's R^2, Nagelkerke R^2

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: $\dfrac{\text{TP} + \text{TN}}{n}$

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

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

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

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

$$\Lambda = |S \frac{|}{|} \Sigma_0|$$

Box's M Test

Testing Equality of Covariance Matrices

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

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)

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

Test Statistic: Chi-square distributed

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:

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

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 $\bar{\boldsymbol{X}}$

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

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

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:

$$V = \mathrm{tr}\left(\boldsymbol{H}(\boldsymbol{H} + \boldsymbol{E})^{-1}\right)$$

Hotelling-Lawley Trace:

$$U = \mathrm{tr}\left(\boldsymbol{H}\boldsymbol{E}^{-1}\right)$$

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

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: $X_1, X_2, \ldots, X_p$
- Set 2: $Y_1, Y_2, \ldots, Y_q$

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:

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

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

such that $\mathrm{cor}(U_1, V_1)$ 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:

$$\rho_1 \geq \rho_2 \geq \dots \geq \rho_k \geq 0$$

where ρ_i is the i-th canonical correlation

Testing Significance

Test All Correlations:

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

Test Remaining Correlations:

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

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?

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

```python
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

```python
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

Key Takeaways: Models

Logistic Regression:

- Use for binary outcomes
- Maximum likelihood estimation
- Interpret via odds ratios

Key Takeaways: Inference

Covariance Matrix Tests:

- Box's M test for equality
- Wishart distribution foundation

Mean Vector Tests:

- Hotelling's T-squared generalizes t-test
- Confidence regions are ellipsoids

Key Takeaways: Advanced Methods

MANOVA:

- Multiple response variables simultaneously
- Wilks' Lambda most common test
- Controls Type I error

Canonical Correlation:

- Relates two variable sets
- Multiple correlation pairs

Key Takeaways: Applications

Factor-Based Regression:

- Addresses multicollinearity
- Dimension reduction
- Interpretable factors

Software:

- Python: scikit-learn, statsmodels
- R: stats, MASS
- Commercial: SPSS, SAS

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) $\longrightarrow$ Factor-based regression
- Discriminant Analysis (L05) $\longrightarrow$ MANOVA post-hoc
- PCA principles $\longrightarrow$ Principal components regression

Comprehensive Framework: All methods part of the multivariate toolkit