Multivariate Regression

Lecture Notes

MA2003B - Application of Multivariate Methods in Data Science

1 Document Information

1.1 Credits and Roles

Instructional Architect: Juliho Castillo Colmenares (Course Instructor)

- Designed course structure and learning objectives
- Established pedagogical direction and content scope
- Defined assessment criteria and evaluation approach
- Reviewed content for academic accuracy and rigor
- Maintains ongoing responsibility for course materials

AI Assistant: Claude Sonnet 4.5 (Anthropic)

- Generated comprehensive lecture notes based on curriculum design
- Developed mathematical formulations and technical explanations
- Created practice questions and detailed answers
- Structured content according to pedagogical best practices
- Aligned material with E07 Multivariate Regression quiz topics

1.2 Contact Information

Course Instructor:

- **Name:** Juliho Castillo Colmenares
- **Email:** julihocc@tec
- **Office:** Tec de Monterrey CCM Office 1540
- **Office Hours:** Monday-Friday, 9:00 AM - 5:00 PM

Course Information:

- **Course Code:** MA2003B
- **Course Title:** Application of Multivariate Methods in Data Science
- **Institution:** Tec de Monterrey
- **Department:** Escuela de Ingeniería y Ciencias

1.3 Version Information

- **Document Version:** 1.0
- **Created:** 2025-11-12
- **Last Updated:** 2025-11-12
- **Status:** Active - Current Semester

1.4 Usage and Distribution

These lecture notes are intended for students enrolled in MA2003B. The material is designed to prepare students for the Multivariate Regression evaluation (E07) and provide comprehensive understanding of advanced regression methods in multivariate settings.

Academic Integrity: Students are expected to use these notes as a study resource in accordance with institutional academic integrity policies.

Contents

2 Logistic Regression Model

2.1 Introduction to Logistic Regression

Logistic regression is a statistical method for modeling binary response variables. Unlike linear regression, which models continuous outcomes, logistic regression predicts the probability that an observation belongs to one of two categories.

Info: The response variable in logistic regression follows a **Bernoulli distribution**, meaning it can take only two values: 0 (failure/absence) or 1 (success/presence).

2.2 The Logit Link Function

The core of logistic regression is the **logit function**, which transforms probabilities to the real line:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

Where:

- p is the probability of the event occurring
- $\frac{p}{1-p}$ is the odds ratio
- $\beta_0, \beta_1, ..., \beta_k$ are regression coefficients

The inverse transformation gives us the predicted probability:

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k))}$$

2.3 Key Differences from Linear Regression

Warning: Do not use ordinary least squares (OLS) for binary outcomes. The assumptions of linear regression are violated when the response is binary:

- Residuals are not normally distributed
- Variance is not constant (heteroscedasticity)
- Predicted values can fall outside 0,1 range

Logistic regression addresses these issues by:

1. Using a link function to bound predictions between 0 and 1
2. Modeling probabilities rather than raw values
3. Using maximum likelihood estimation instead of least squares

2.4 Maximum Likelihood Estimation

Since OLS assumptions are violated, we use **Maximum Likelihood Estimation (MLE)** to estimate coefficients. The likelihood function for binary data is:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

Where y_i is the observed outcome (0 or 1) and p_i is the predicted probability.

The log-likelihood is maximized using iterative numerical methods (typically Newton-Raphson or Fisher scoring).

2.5 Interpretation of Coefficients

In logistic regression, coefficients represent the change in log-odds for a one-unit increase in the predictor:

- $\beta_j > 0$: Increasing x_j increases the odds of the event
- $\beta_j < 0$: Increasing x_j decreases the odds of the event

Tip: To interpret effects on probability scale, compute the odds ratio: $\text{OR} = \exp(\beta_j)$. An OR greater than 1 indicates increased odds, while OR less than 1 indicates decreased odds.

2.6 Model Diagnostics

Common diagnostics for logistic regression include:

- **Deviance**: Measures goodness of fit
- **AIC/BIC**: Model selection criteria
- **ROC Curve and AUC**: Classification performance
- **Hosmer-Lemeshow test**: Calibration assessment

Example: Predicting customer churn: $p(\text{churn}) = \frac{1}{1+\exp(-(2.5-0.8\times \text{ tenure } +1.2\times \text{ complaints}))}$

- Coefficient for tenure: -0.8 (longer tenure reduces churn probability)
- Coefficient for complaints: 1.2 (more complaints increase churn probability)

3 Inferences for Variances and Covariances Matrices

3.1 Multivariate Normal Distribution

Before discussing inference for covariance matrices, we must establish that the data comes from a multivariate normal distribution:

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Where:

- $\boldsymbol{X}$ is a $p \times 1$ random vector
- $\boldsymbol{\mu}$ is the mean vector
- $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix

3.2 Sample Covariance Matrix

For a sample of n observations, the unbiased estimate of the covariance matrix is:

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)'$$

3.3 The Wishart Distribution

The Wishart distribution is the multivariate generalization of the chi-square distribution. When $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are independent $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$(n-1)\boldsymbol{S} \sim W_p(n-1, \boldsymbol{\Sigma})$$

This is read as "$(n-1)S$ follows a Wishart distribution with $n-1$ degrees of freedom and scale matrix Σ."

Info: Just as the chi-square distribution is used for inference about a single variance, the Wishart distribution is used for inference about covariance matrices.

3.4 Testing Equality of Covariance Matrices

A common question is whether two or more groups have the same covariance structure.

3.4.1 Box's M Test

For testing $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_g$, Box's M statistic is:

$$M = (n-g)\log|\boldsymbol{S}_{\text{pooled}}| - \sum_{i=1}^{g}(n_i - 1)\log|\boldsymbol{S}_i|$$

Where:

- $\boldsymbol{S}_{\text{pooled}}$ is the pooled covariance matrix
- $\boldsymbol{S}_i$ is the covariance matrix for group i
- n_i is the sample size for group i

Under H_0, M follows approximately a chi-square distribution for large samples.

Warning: Box's M test is sensitive to departures from multivariate normality. Use with caution and verify normality assumptions first.

3.4.2 Bartlett's Test for Univariate Variances

For the special case of testing equality of variances in univariate data:

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

Bartlett's test statistic is:

$$\chi^2 = \frac{(n-g)\log\left(s^2_{\text{pooled}}\right) - \sum_{i=1}^{g}(n_i - 1)\log(s_i^2)}{C}$$

Where C is a correction factor for small samples.

3.5 Assumptions for Valid Inference

For valid inference about covariance matrices:

1. Data must come from a multivariate normal distribution
2. Observations must be independent
3. Sample sizes should be adequate (general rule: $n > 5p$)

Tip: Always test for multivariate normality before conducting inference on covariance matrices. Methods include Mardia's test or examining Q-Q plots for each variable.

4 Inferences for a Vector of Means

4.1 Hotelling's T-Squared Test

Hotelling's T^2 statistic is the multivariate generalization of the univariate t-test. It tests hypotheses about mean vectors.

4.1.1 One-Sample Test

For testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$:

$$T^2 = n\left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right)' \boldsymbol{S}^{-1} \left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right)$$

Under H_0, the test statistic can be transformed to an F-distribution:

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

Where:

- n is sample size
- p is number of variables
- $\overline{\boldsymbol{X}}$ is the sample mean vector
- $\boldsymbol{S}$ is the sample covariance matrix

4.1.2 Two-Sample Test

For comparing two independent groups, testing $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2\right)' \boldsymbol{S}_{\text{pooled}}^{-1} \left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2\right)$$

The pooled covariance matrix is:

$$\boldsymbol{S}_{\text{pooled}} = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2}$$

The F-transformation is:

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

Info: The key advantage of Hotelling's T^2 over multiple univariate t-tests is that it controls the overall Type I error rate when testing multiple variables simultaneously.

4.2 Paired Samples

For paired observations, test $H_0 : \boldsymbol{\mu}_D = \boldsymbol{0}$ where $\boldsymbol{D} = \boldsymbol{X}_1 - \boldsymbol{X}_2$:

$$T^2 = n\overline{\boldsymbol{D}}' \boldsymbol{S}_D^{-1} \overline{\boldsymbol{D}}$$

This reduces to the one-sample case applied to difference vectors.

4.3 Confidence Regions

Simultaneous confidence intervals for mean vectors form ellipsoidal regions in p-dimensional space:

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

Example: Comparing nutritional content (protein, fat, carbohydrates) between two diets using two-sample Hotelling's T^2 test. This accounts for correlations between nutritional components and provides a single omnibus test.

4.4 Rejection Criteria

Reject H_0 when:

- T^2 exceeds the critical value from the T^2 distribution, or equivalently
- F exceeds the critical value $F_{p,n-p,\alpha}$

Tip: After rejecting the null hypothesis with Hotelling's T^2, follow up with univariate tests or discriminant analysis to identify which variables contribute most to the difference.

5 Multivariate Analysis of Variance (MANOVA)

5.1 Introduction to MANOVA

MANOVA (Multivariate Analysis of Variance) extends ANOVA to situations with multiple dependent variables. It tests whether group means on a combination of dependent variables differ across levels of categorical independent variables.

5.2 Why Use MANOVA Instead of Multiple ANOVAs?

Warning: Conducting separate ANOVAs for each dependent variable increases Type I error through multiple testing. If you test 5 variables at alpha equals 0.05, your overall error rate could be as high as 1 minus 0.95 raised to 5 equals 0.226.

Advantages of MANOVA:

1. **Controls Type I error** by testing all variables simultaneously
2. **Accounts for correlations** between dependent variables
3. **Increases statistical power** by combining information across variables
4. **Detects patterns** that univariate tests might miss

5.3 MANOVA Model

The multivariate linear model:

$$\boldsymbol{Y}_{n\times p} = \boldsymbol{X}_{n\times k}\boldsymbol{B}_{k\times p} + \boldsymbol{E}_{n\times p}$$

Where:

- $\boldsymbol{Y}$ is the matrix of responses (n observations, p variables)
- $\boldsymbol{X}$ is the design matrix (group indicators)
- $\boldsymbol{B}$ is the matrix of coefficients
- $\boldsymbol{E}$ is the matrix of errors

The null hypothesis is $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_g$ for all p variables.

5.4 Test Statistics

5.4.1 Wilks' Lambda

The most commonly used MANOVA test statistic is **Wilks' Lambda**:

$$\Lambda = |\boldsymbol{W}\frac{|}{|}\boldsymbol{T}| = |\boldsymbol{W}\frac{|}{|\boldsymbol{W}| + |\boldsymbol{B}|}$$

Where:

- $\boldsymbol{W}$ is the within-groups sum of squares and cross-products matrix
- $\boldsymbol{B}$ is the between-groups sum of squares and cross-products matrix
- $\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}$ is the total sum of squares and cross-products matrix

Info: Wilks' Lambda represents the proportion of total variance not explained by group differences. Values range from 0 to 1, with smaller values indicating greater group separation.

5.4.2 Other Test Statistics

Alternative MANOVA test statistics include:

- **Pillai's Trace**: More robust to violations of assumptions
- **Hotelling-Lawley Trace**: Most powerful when assumptions are met
- **Roy's Greatest Root**: Tests for difference on the first discriminant function only

5.5 Assumptions

For valid MANOVA inference:

1. **Independence**: Observations must be independent
2. **Multivariate Normality**: Errors follow multivariate normal distribution
3. **Homogeneity of Covariance Matrices**: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_g$
4. **Adequate Sample Size**: At least $n > p + g$ (preferably much larger)

Tip: Test homogeneity of covariance matrices using Box's M test before conducting MANOVA. If violated, consider robust alternatives or data transformations.

5.6 Interpreting MANOVA Results

After a significant MANOVA result:

1. **Examine univariate F-tests** to see which individual variables differ across groups
2. **Apply Bonferroni correction** or other multiple comparison adjustments
3. **Conduct discriminant analysis** to identify linear combinations that best separate groups
4. **Examine effect sizes** (e.g., partial eta-squared) for practical significance

Example: Comparing teaching methods (lecture, discussion, online) on student outcomes measured by exam score, engagement rating, and satisfaction score. MANOVA tests whether the three methods produce different patterns across all three outcomes simultaneously.

5.7 Follow-Up Analyses

After rejecting the null hypothesis in MANOVA:

- **Univariate ANOVAs** with adjusted alpha levels
- **Discriminant Function Analysis** to identify separation patterns
- **Post-hoc pairwise comparisons** with appropriate corrections
- **Contrast analysis** for planned comparisons

Warning: Always report which assumptions were tested and any violations found. MANOVA can be sensitive to assumption violations, particularly with unequal sample sizes.

6 Canonical Correlation Analysis

6.1 Introduction to Canonical Correlation

Canonical Correlation Analysis (CCA) examines relationships between **two sets of variables** simultaneously. It finds linear combinations of variables in each set that maximize their correlation.

6.2 Motivation and Applications

CCA answers questions like:

- How do personality traits (set 1) relate to job performance measures (set 2)?
- What is the relationship between physical characteristics and athletic performance?
- How do economic indicators relate to social welfare measures?

Info: Unlike multiple regression (many predictors, one response), CCA treats both sets symmetrically and finds the strongest relationship between linear combinations of each set.

6.3 Mathematical Framework

Given two sets of variables:

- Set 1: $\boldsymbol{X}$ with p variables
- Set 2: $\boldsymbol{Y}$ with q variables

CCA finds coefficients $\boldsymbol{a}$ and $\boldsymbol{b}$ such that:

$$\text{Cor}(\boldsymbol{a}'\boldsymbol{X}, \boldsymbol{b}'\boldsymbol{Y})$$

is maximized, subject to:

- $\text{Var}(\boldsymbol{a}'\boldsymbol{X}) = 1$
- $\text{Var}(\boldsymbol{b}'\boldsymbol{Y}) = 1$

The resulting linear combinations are called **canonical variates**:

- First canonical variate for Set 1: $U_1 = \boldsymbol{a}_{1'}\boldsymbol{X}$
- First canonical variate for Set 2: $V_1 = \boldsymbol{b}_{1'}\boldsymbol{Y}$

6.4 Number of Canonical Correlations

The number of canonical correlation pairs equals:

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

Where p and q are the number of variables in sets 1 and 2.

Each successive pair of canonical variates:

- Is uncorrelated with previous pairs
- Maximizes correlation subject to this constraint
- Has correlation less than or equal to the previous pair

Example: If you have 5 academic performance measures and 3 personality traits, you can extract at most min(5,3) equals 3 canonical correlation pairs.

6.5 Canonical Loadings

Canonical loadings (also called **structure coefficients**) measure the correlation between original variables and canonical variates:

- Loading of X_i on U_j: r_{X_i,U_j}
- Loading of Y_i on V_j: r_{Y_i,V_j}

These loadings help interpret what each canonical variate represents.

Tip: Canonical loadings are generally more interpretable than canonical weights (coefficients) because they are not affected by multicollinearity among predictors.

6.6 Testing Significance

Test whether there are any significant canonical correlations using Wilks' Lambda:

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

Where r_i are the canonical correlations.

The test statistic:

$$\chi^2 = -\left(n - 1 - \frac{p + q + 1}{2}\right) \log(\Lambda)$$

follows approximately a chi-square distribution with pq degrees of freedom.

6.7 Interpretation Guidelines

When interpreting canonical correlations:

1. Test statistical significance of each canonical correlation
2. Examine the proportion of variance explained
3. Interpret canonical loadings to understand variable contributions
4. Name canonical variates based on which variables load highly
5. Consider practical significance, not just statistical significance

Warning: Canonical correlation requires large sample sizes for stable results. A general guideline is at least 10 observations per variable, preferably 20.

6.8 Relationship to Other Methods

Canonical correlation generalizes several statistical techniques:

- Multiple regression: Special case where Set 2 has one variable
- Discriminant analysis: When Set 2 is group membership
- Principal components: When Set 1 and Set 2 are the same

Example: Examining the relationship between socioeconomic factors (income, education, housing quality) and health outcomes (life expectancy, disease prevalence, healthcare access). CCA identifies which combinations of socioeconomic factors most strongly relate to which patterns of health outcomes.

7 Analysis by Factors and Regression

7.1 Combining Factor Analysis with Regression

Factor analysis can be integrated with regression to address **multicollinearity** and **dimensionality reduction** in predictive modeling.

7.2 The Problem: Multicollinearity

When predictor variables are highly correlated:

- Regression coefficients become unstable
- Standard errors inflate
- Interpretation becomes difficult
- Prediction may remain good, but inference suffers

Warning: High correlation among predictors (typically |r| greater than 0.8) can lead to:

- Nonsensical coefficient estimates
- Coefficients with unexpected signs
- High variance inflation factors (VIF greater than 10)

7.3 Factor Scores as Predictors

The general approach:

1. **Extract factors** from the predictor variables using factor analysis
2. **Compute factor scores** for each observation
3. **Use factor scores** as predictors in regression instead of original variables

This addresses multicollinearity because:

- Factor scores are orthogonal (uncorrelated) when using orthogonal rotation
- Dimensionality is reduced (fewer predictors than original variables)
- Factors represent underlying constructs rather than individual variables

7.4 Mathematical Framework

Given original predictors $\boldsymbol{X}$ with p variables, extract m factors (where $m < p$):

$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{U}$$

Where:

- $\boldsymbol{L}$ is the loading matrix
- $\boldsymbol{F}$ are factor scores
- $\boldsymbol{U}$ are unique factors

Then fit regression:

$$Y = \beta_0 + \beta_1 F_1 + \beta_2 F_2 + ... + \beta_m F_m + \varepsilon$$

7.5 Methods for Computing Factor Scores

7.5.1 Regression Method

$$\hat{\boldsymbol{F}} = (\boldsymbol{L}'\boldsymbol{\Sigma}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$$

Advantages:
- Minimizes least squares error
- Most commonly used

7.5.2 Bartlett Method

$$\hat{\boldsymbol{F}} = \left(\boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{L}\right)^{-1}\boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}$$

Where $\boldsymbol{\Psi}$ is the diagonal matrix of unique variances.

Advantages:
- Unbiased estimates
- Accounts for measurement error

Info: Most statistical software (R, Python, SPSS) provides options for different factor score estimation methods.

7.6 Benefits of Factor-Based Regression

1. **Dimensionality Reduction**: m factors instead of p variables (where $m \ll p$)
2. **Elimination of Multicollinearity**: Orthogonal factors have zero correlation
3. **Interpretability**: Factors represent meaningful constructs
4. **Robustness**: Less sensitive to measurement error in individual variables

7.7 Interpretation Considerations

When interpreting factor-based regression:

- **Factor coefficients** indicate importance of underlying constructs
- **Factor loadings** show which original variables contribute to each factor
- **Indirect effects** can be traced from original variables through factors to outcome

Example: Predicting customer satisfaction from 20 service quality items. Instead of using all 20 items (with high multicollinearity), extract 4 factors: Responsiveness, Reliability, Tangibles, and Empathy. Regress satisfaction on these 4 factors, which are uncorrelated and more interpretable.

7.8 Limitations and Considerations

Warning: Factor-based regression has limitations:
- Factors must be meaningful and interpretable
- Factor extraction is somewhat subjective (number of factors, rotation method)
- Results depend on the specific sample used for factor analysis
- Prediction in new samples requires computing factor scores using the same loadings

7.9 Alternative Approaches

Other methods for addressing multicollinearity:
- **Ridge Regression**: Shrinks coefficients toward zero
- **Principal Components Regression**: Similar to factor regression but uses PCA
- **Partial Least Squares**: Finds components that predict Y well

- **Variable Selection**: Remove redundant predictors

Tip: Consider factor-based regression when you have many correlated predictors representing a smaller number of underlying constructs. Use PCA-based regression when dimensionality reduction is the primary goal rather than construct interpretation.

8 Programming and Commercial Systems

8.1 Statistical Software for Multivariate Analysis

Modern multivariate analysis relies heavily on specialized statistical software. This section covers the primary tools used in practice and research.

8.2 Python Ecosystem

8.2.1 Core Libraries

NumPy and Pandas: Foundation for data manipulation

- Matrix operations and linear algebra
- Data structures for multivariate data
- Missing data handling

Statsmodels: Statistical modeling

```
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Logistic regression
model = smf.logit('y ~ x1 + x2 + x3', data=df)
result = model.fit()

# MANOVA
from statsmodels.multivariate.manova import MANOVA
manova = MANOVA.from_formula('y1 + y2 + y3 ~ group', data=df)
print(manova.mv_test())
```

Scikit-learn: Machine learning and predictive modeling

```
from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

# Logistic regression with regularization
logreg = LogisticRegression(penalty='l2', C=1.0)
logreg.fit(X_train, y_train)

# LDA for classification
lda = LinearDiscriminantAnalysis()
lda.fit(X_train, y_train)
```

8.2.2 Specialized Packages

Factor Analyzer: Factor analysis

```
from factor_analyzer import FactorAnalyzer, calculate_kmo

# KMO test
kmo_all, kmo_model = calculate_kmo(df)

# Factor analysis
fa = FactorAnalyzer(n_factors=3, rotation='varimax')
fa.fit(df)
loadings = fa.loadings_
```

Scipy: Scientific computing

```
from scipy.stats import wishart
from scipy.spatial.distance import mahalanobis

# Wishart distribution sampling
W = wishart.rvs(df=10, scale=Sigma, size=1)

# Mahalanobis distance
d = mahalanobis(x, mean, cov_inv)
```

Tip: Python's ecosystem is ideal for integrating multivariate analysis with machine learning pipelines and production systems. Use Jupyter notebooks for exploratory analysis and Python scripts for production code.

8.3 R Programming

R provides extensive support for multivariate statistics through base functions and specialized packages.

8.3.1 Base R Functions

```
# Hotelling's T-squared test
library(Hotelling)
hotel.test <- hotelling.test(cbind(y1, y2) ~ group, data=df)

# MANOVA
manova_result <- manova(cbind(y1, y2, y3) ~ group, data=df)
summary(manova_result, test="Wilks")

# Canonical correlation
library(CCA)
cc_result <- cc(X, Y)
```

8.3.2 Specialized Packages

- `mvtnorm`: Multivariate normal and t distributions
- `MVN`: Multivariate normality tests
- `car`: Companion to Applied Regression (includes MANOVA)
- `FactoMineR`: Multivariate exploratory analysis
- `psych`: Psychological methods including factor analysis

Info: R excels at statistical analysis and has more specialized multivariate packages than Python. It is widely used in academic research and has excellent documentation.

8.4 Commercial Software Packages

8.4.1 SPSS (IBM)

Graphical User Interface:

- Point-and-click menus for common analyses
- Integrated data editor and output viewer
- Extensive built-in help and tutorials

Syntax Programming:

```
* Logistic regression
LOGISTIC REGRESSION VAR=outcome
  /METHOD=ENTER predictor1 predictor2 predictor3
  /PRINT=GOODFIT CI(95).

* MANOVA
MANOVA y1 y2 y3 BY group(1,3)
  /PRINT=CELLINFO(MEANS)
  /DESIGN.

* Factor analysis
FACTOR /VARIABLES var1 TO var10
  /EXTRACTION PC
  /ROTATION VARIMAX.
```

Strengths:

- User-friendly for non-programmers
- Comprehensive documentation
- Wide adoption in social sciences

8.4.2 SAS (Statistical Analysis System)

Procedures for Multivariate Analysis:

```
/* Logistic regression */
PROC LOGISTIC DATA=mydata;
  MODEL outcome(EVENT='1') = x1 x2 x3 / LINK=LOGIT;
  OUTPUT OUT=pred PREDICTED=prob;
RUN;

/* MANOVA */
PROC GLM DATA=mydata;
  CLASS group;
  MODEL y1 y2 y3 = group;
  MANOVA H=group / PRINTH PRINTE;
RUN;

/* Canonical correlation */
PROC CANCORR DATA=mydata;
  VAR x1 x2 x3;
  WITH y1 y2 y3;
RUN;

/* Factor analysis */
PROC FACTOR DATA=mydata METHOD=PRINCIPAL
  ROTATE=VARIMAX NFACTORS=3;
  VAR var1-var10;
RUN;
```

Strengths:

- Enterprise-level data handling
- Excellent performance with large datasets
- Strong in clinical trials and regulatory environments

8.4.3 MATLAB

MATLAB's Statistics and Machine Learning Toolbox includes:

- `fitglm`: Generalized linear models (logistic regression)
- `manova1`: One-way MANOVA
- `canoncorr`: Canonical correlation analysis
- `factoran`: Factor analysis

Strengths:

- Matrix operations are native
- Integration with engineering applications
- Excellent for simulation studies

Warning: Commercial software (SPSS, SAS, MATLAB) requires licenses that can be expensive. Academic institutions often provide access. Consider open-source alternatives (R, Python) for cost-effective solutions with comparable functionality.

8.5 Software Selection Guidelines

Choose software based on:

1. **Task Requirements**:
   - Exploratory analysis: R or SPSS
   - Production systems: Python
   - Large-scale data: SAS
   - Engineering integration: MATLAB
2. **Team Expertise**:
   - Programmers: Python, R
   - Non-programmers: SPSS
   - Mixed teams: R + RStudio
3. **Budget Constraints**:
   - Free: R, Python
   - Institutional license: SPSS, SAS, MATLAB
   - Cloud-based: Python (widely available on cloud platforms)
4. **Reproducibility Needs**:
   - Code-based: Python, R, SAS, MATLAB
   - Documentation: All have good capabilities

Example: A pharmaceutical company might use:

- SAS for regulatory submissions (required by FDA)
- R for exploratory analysis and visualization
- Python for machine learning and predictive modeling
- SPSS for surveys and marketing research

8.6 Best Practices

Regardless of software choice:

1. **Document your analysis**: Use scripts/syntax, not just point-and-click
2. **Version control**: Track changes to code and data
3. **Reproducibility**: Include random seeds, package versions
4. **Validation**: Cross-check critical results across software when possible
5. **Backup**: Maintain copies of data, code, and results

Tip: Learn one statistical language well (Python or R) and become familiar with at least one commercial package. This combination provides flexibility for different work environments and requirements.

9 Practice Questions and Answers

9.1 Question 1: Logistic Regression

Question: Why is the logit link function necessary in logistic regression?

Answer: The logit link function $\log\left(\frac{p}{1-p}\right)$ transforms probabilities (bounded between 0 and 1) to the real line (negative infinity to positive infinity). This allows us to model probabilities using a linear combination of predictors. Without this transformation, linear regression would produce predicted probabilities outside the valid range of 0,1, leading to nonsensical results. The inverse logit (logistic function) then maps predictions back to valid probability scale.

9.2 Question 2: Wishart Distribution

Question: What is the relationship between the Wishart distribution and the chi-square distribution?

Answer: The Wishart distribution is the multivariate generalization of the chi-square distribution. Just as the chi-square distribution describes the distribution of a sum of squared standard normal variables, the Wishart distribution describes the distribution of sample covariance matrices from multivariate normal data. When $p = 1$ (univariate case), the Wishart distribution reduces to the chi-square distribution. This relationship is fundamental for conducting inference about covariance matrices in the multivariate setting.

9.3 Question 3: Hotelling's T-Squared

Question: When would you use Hotelling's T^2 test instead of multiple univariate t-tests?

Answer: Use Hotelling's T^2 when comparing mean vectors across groups on multiple correlated variables simultaneously. Conducting separate t-tests inflates Type I error (the family-wise error rate increases with each test). Hotelling's T^2 provides a single omnibus test that controls overall Type I error while accounting for correlations between variables. After rejecting with Hotelling's T^2, you can conduct follow-up univariate tests with appropriate corrections to identify which specific variables differ.

9.4 Question 4: MANOVA vs Multiple ANOVAs

Question: Explain why MANOVA is preferred over conducting separate ANOVAs on each dependent variable.

Answer: MANOVA is preferred for three main reasons:

1. **Type I Error Control**: Multiple ANOVAs increase the probability of false positives through multiple testing. MANOVA maintains the overall Type I error rate at the specified alpha level.
2. **Accounting for Correlations**: MANOVA incorporates the correlation structure among dependent variables, potentially increasing statistical power.
3. **Detecting Patterns**: MANOVA can detect differences in linear combinations of variables that might be missed by individual ANOVAs, revealing multivariate patterns of group differences.

9.5 Question 5: Wilks' Lambda Interpretation

Question: What does Wilks' Lambda measure and how is it interpreted in MANOVA?

Answer: Wilks' Lambda measures the proportion of total variance in the dependent variables that is not explained by differences between groups. It is calculated as $\Lambda = |W \mid T|$ where W is the within-groups matrix and T is the total matrix. Values range from 0 to 1:

- Λ near 0 indicates strong group separation (most variance is between-groups)

- Λ near 1 indicates weak group separation (most variance is within-groups)

Smaller values of Wilks' Lambda lead to rejection of the null hypothesis that group means are equal.

9.6 Question 6: Canonical Correlation

Question: How many canonical correlation pairs can be extracted when analyzing the relationship between 7 socioeconomic variables and 4 health outcomes?

Answer: The number of canonical correlation pairs equals $\min(p, q)$ where p and q are the number of variables in each set. In this case, $\min(7, 4) = 4$, so exactly 4 canonical correlation pairs can be extracted. The first pair has the maximum possible correlation, and each successive pair has decreasing correlation while being uncorrelated with previous pairs.

9.7 Question 7: Canonical Loadings

Question: What are canonical loadings and why are they important for interpretation?

Answer: Canonical loadings (structure coefficients) represent the correlation between original variables and canonical variates. For example, the loading of variable X_3 on canonical variate U_1 measures how strongly X_3 contributes to U_1. These loadings are crucial for interpretation because they help identify what each canonical variate represents. Variables with high loadings (typically |r| greater than 0.3 or 0.4) on a variate are considered important contributors to that dimension. Canonical loadings are generally more interpretable than canonical weights because they are not affected by multicollinearity.

9.8 Question 8: Factor Scores in Regression

Question: What are the main benefits of using factor scores as predictors in regression analysis?

Answer: Using factor scores as predictors provides several benefits:

1. **Multicollinearity Reduction**: Factor scores are orthogonal (uncorrelated) when using orthogonal rotation, eliminating multicollinearity problems
2. **Dimensionality Reduction**: Fewer predictors than original variables (m factors instead of p variables where $m \ll p$)
3. **Construct Interpretation**: Factors represent underlying latent constructs rather than individual observed variables
4. **Robustness**: Less sensitive to measurement error in individual variables since factors capture shared variance

This approach is particularly useful when you have many correlated predictors representing a smaller number of theoretical constructs.

9.9 Question 9: Box's M Test

Question: What does Box's M test assess and why is it important for MANOVA?

Answer: Box's M test assesses the assumption of homogeneity of covariance matrices across groups, testing $H_0 : \Sigma_1 = \Sigma_2 = ... = \Sigma_g$. This is crucial for MANOVA because the method assumes groups have equal covariance structures. Violation of this assumption can affect the validity of MANOVA results, particularly with unequal sample sizes. However, Box's M is sensitive to departures from multivariate normality, so it should be used cautiously. If the test is significant, consider robust alternatives or data transformations.

9.10 Question 10: Software Selection

Question: What factors should guide the choice between Python's statsmodels and SPSS for conducting multivariate regression analysis in a corporate setting?

Answer: Key factors include:

1. **Team Skills**: SPSS is more suitable for teams without programming background (GUI-based), while Python requires coding skills
2. **Integration Needs**: Python integrates easily with databases, web applications, and machine learning pipelines; SPSS is more standalone
3. **Cost**: Python is free and open-source; SPSS requires expensive licenses
4. **Reproducibility**: Both support scripting (Python natively, SPSS through syntax), crucial for reproducible research
5. **Advanced Features**: Python offers more flexibility for custom analyses and newer methods; SPSS has more built-in standard procedures

In practice, many organizations use both: SPSS for standard analyses by non-programmers, Python for custom work and production systems.

10 Summary and Key Takeaways

10.1 Logistic Regression

- Models binary outcomes using Bernoulli distribution
- Logit link function: $\log\left(\frac{p}{1-p}\right)$
- Maximum likelihood estimation (not OLS)
- Coefficients represent change in log-odds

10.2 Covariance Matrix Inference

- Based on multivariate normal distribution
- Wishart distribution generalizes chi-square
- Box's M test for equality of covariances
- Bartlett's test for univariate variances

10.3 Mean Vector Inference

- Hotelling's T^2 generalizes t-test
- Can be transformed to F distribution
- Controls Type I error for multiple variables
- Applicable to one-sample, two-sample, and paired designs

10.4 MANOVA

- Tests multiple dependent variables simultaneously
- Controls Type I error inflation
- Wilks' Lambda most common test statistic
- Requires multivariate normality and homogeneous covariances

10.5 Canonical Correlation

- Examines relationships between two variable sets
- Maximizes correlation between linear combinations
- Number of pairs equals $\min(p, q)$
- Canonical loadings aid interpretation

10.6 Factor-Based Regression

- Addresses multicollinearity through dimensionality reduction
- Factor scores used as predictors
- Orthogonal factors have zero correlation
- Balances interpretability and parsimony

10.7 Statistical Software

- **Python**: statsmodels, scikit-learn, flexible integration
- **R**: Comprehensive multivariate packages, research-oriented
- **SPSS**: User-friendly GUI, social science standard
- **SAS**: Enterprise data handling, regulatory compliance
- Choose based on team skills, requirements, and budget

11 References and Further Reading

11.1 Recommended Textbooks

1. Johnson, R. A., & Wichern, D. W. (2007). **Applied Multivariate Statistical Analysis** (6th ed.). Pearson.
2. Rencher, A. C., & Christensen, W. F. (2012). **Methods of Multivariate Analysis** (3rd ed.). Wiley.
3. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). **Multivariate Data Analysis** (8th ed.). Cengage.
4. Tabachnick, B. G., & Fidell, L. S. (2018). **Using Multivariate Statistics** (7th ed.). Pearson.

11.2 Online Resources

- Python statsmodels documentation: https://www.statsmodels.org/
- Scikit-learn user guide: https://scikit-learn.org/stable/user_guide.html
- R CRAN Task View Multivariate: https://cran.r-project.org/web/views/Multivariate.html
- Penn State STAT 505 (Online course on multivariate methods)

11.3 Software Documentation

- **Python**: pandas, numpy, scipy, statsmodels, scikit-learn
- **R**: Base R stats package, car, MVN, FactoMineR, psych
- **SPSS**: IBM SPSS Statistics documentation
- **SAS**: SAS/STAT User's Guide

End of Lecture Notes

MA2003B - Multivariate Regression

Tec de Monterrey