Why study multivariate data analysis?

- Multivariate data more common in research
- GLM approach: ANOVA, regression, etc. within a common framework: linear models
  \[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \]
- In matrix form (\( \mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{\varepsilon} \)), GLM extends to MANOVA, MMReg, etc.
- Idea of linear combinations extends readily to other methods: PCA, discriminant analysis, etc.
- Graphical methods, geometry → Insight

Sample problem: workers’ data

<table>
<thead>
<tr>
<th>Name</th>
<th>Income</th>
<th>Experience</th>
<th>Skill</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Abby</td>
<td>20</td>
<td>0</td>
<td>Female</td>
</tr>
<tr>
<td>2</td>
<td>Betty</td>
<td>35</td>
<td>5</td>
<td>Female</td>
</tr>
<tr>
<td>3</td>
<td>Charles</td>
<td>40</td>
<td>5</td>
<td>Male</td>
</tr>
<tr>
<td>4</td>
<td>Doreen</td>
<td>30</td>
<td>10</td>
<td>Female</td>
</tr>
<tr>
<td>5</td>
<td>Ethan</td>
<td>50</td>
<td>10</td>
<td>Male</td>
</tr>
<tr>
<td>6</td>
<td>Francie</td>
<td>50</td>
<td>15</td>
<td>Female</td>
</tr>
<tr>
<td>7</td>
<td>Georges</td>
<td>60</td>
<td>20</td>
<td>Male</td>
</tr>
<tr>
<td>8</td>
<td>Harry</td>
<td>50</td>
<td>25</td>
<td>Male</td>
</tr>
<tr>
<td>9</td>
<td>Isaac</td>
<td>70</td>
<td>30</td>
<td>Male</td>
</tr>
<tr>
<td>10</td>
<td>Juan</td>
<td>60</td>
<td>35</td>
<td>Male</td>
</tr>
</tbody>
</table>

In truly multivariate data, we may have several outcomes:
- Income
- Job satisfaction
- Manager ratings
- etc.

Linear models: Regression

Regression: understanding the relation of quantitative predictor(s) on a quantitative outcome.

Model: \( E(y \mid x) = \beta_0 + \beta_1 x \)

e.g., Income = 29 + 1.12 Experience

Parameters:
- \( \beta_0 = 29 = \text{Income at 0 years} \)
- \( \beta_1 = 1.12 = \text{Increase / year} = \frac{\Delta y}{\Delta x} \)

The regression line on the graph, and the fitted equation are just summaries. It is important to think about what they mean for a given problem!
Linear models: Regression

Regression: a “linear model” need only be linear in the parameters. It can have terms like $x^2$, $\log(x)$, etc.

Model: \[ E(y \mid x) = \beta_0 + \beta_1 x + \beta_2 x^2 \]
e.g., Income = $23 + 2.3 \text{ Exp} -0.33 \text{ Exp}^2$

Parameters:
- $\beta_0 = 23 = \text{Income at 0 years}$
- $\beta_1 = 2.3 = \text{Slope at 0 years}$
- $\beta_2 = -0.33 = \text{Decrease in slope/year}$

What does $\beta_1 = 2.3$ mean?
What does $\beta_2 < 0$ mean?

Linear models: Multiple regression

Regression models can have any number of linear predictors

Model: \[ E(y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \]
e.g., Income = $14.8 + 0.11 \text{ Exper} + 3.4 \text{ Skill}$

Parameters:
- $\beta_0 = 14.8 = \text{Income at 0 years, 0 skill}$
- $\beta_1 = 0.11 = \Delta \text{Income} / \Delta \text{Experience} | \text{Skill}$
- $\beta_2 = 3.4 = \Delta \text{Income} / \Delta \text{Skill} | \text{Experience}$

Control: The estimated effect for each predictor controls (adjusts) for all others in the model

Linear models: ANOVA

ANOVA: How does mean of quantitative response vary with a discrete factor?

Model: $E(Y) = \mu + \beta (G=\text{Male'})$
e.g., Income = $33.75 + 21.25 (G=\text{Male'})$

\[ (G = \text{`Male'}) \equiv \begin{bmatrix} 1 \\ 0 \end{bmatrix} \]

Parameters:
- $\mu = 33.75 = \text{Female mean Income}$
- $\beta = 21.25 = \text{Increment for Male}$

How would you describe this in words?

Linear models: Regression + Anova

ANCOVA: Is there a difference in a factor, controlling for a quantitative predictor?

Homogeneity of regression: Are the regression lines for two or more groups the same? Are they parallel?

Model: $E(Y) = \mu + \beta_1 x_1 + \beta_2 (G=\text{Male'})$
e.g.,
- Inc = $27.27 + 0.86 \text{ Exp} + 9.73 (G=\text{Male'})$

The coefficient, $\beta_2$ for G=Male’ allows the intercepts (or means) to differ. Slopes are forced to be equal.
Linear models: Regression + Anova

Homogeneity of regression: Test equal slopes by allowing a different slope for each group [X * Group interaction]

Model: \( E(Y) = \mu + \beta_1 X_1 + \beta_2 (G='Male') + \beta_3 X_1 * (G='Male') \)

e.g.,
Inc = 21.0 + 1.70 Exp + 19.25 (G='Male') - 1.0 Exp * (G='Male')

Thus, we have two separate models:

- Females: Inc = 21.0 + 1.7 Exp
- Males: Inc = (21+19.25) + (1.7-1.0) Exp = 40.25 + 0.7 Exp

A more complete description, but maybe overly complex!

General Linear Model (GLM)

All of these are special cases of the General Linear Model:

\[ y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \]

data = fitted (explained part) + residual (unexplained)

where,
- **x** Quantitative predictor (experience, skill)
- **\( \beta \)** Effect of predictor (\( \Delta y/\Delta x \))

Regression ANOVA

- Dependent (response) Quantitative Quantitative
- Independent (predictors) Quantitative Discrete factors
- Concepts, statistics Terms: \( X_1, X_2 \) Interactions: \( A_B \) Linear hypotheses Main effects: A, B
  Contrasts F stats, factor effects

They all become unified when cast in matrix terms:

\[
\begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{pmatrix} =
\begin{pmatrix}
1 & x_{11} & \ldots & \beta_0 \\
1 & x_{21} & \ldots & \beta_1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & \ldots & \beta_p
\end{pmatrix}
\begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{pmatrix}
\]

or,

\[
y_{n \times 1} = X_{n \times (p+1)} \beta_{(p+1) \times 1} + \epsilon_{n \times 1}
\]

For all cases:
- parameter estimates, std. errors, etc. have the same form
- all hypothesis tests are special cases of \( H_0: C \beta = 0 \)
- methods extend directly to: multivariate \( Y \), non-normal errors, etc.
Linear models & linear combinations

• All methods of multivariate statistics involve linear combinations of variables, with weights (coefficients) chosen to optimize some criterion (measure of fit)

• Methods differ according to:
  - 1 set of variables (PCA, FA) vs. 2+ sets (GLM, canonical correlation, discrim. analysis)
  - Nature of variables (2 sets):
    • Xs: discrete / continuous
    • Ys: discrete / continuous

Linear combinations: 1 set of variables

With $p$ variables, $p$ components account for 100% of variance, and correspond to a rotation of the variable space to uncorrelated components.

Goal in PCA is to account for most variance with $k<<p$ components.

Factor analysis: Latent variables

FA: find weights for latent (unobserved) factors to account for correlations among observed variables

$$x_1 = \lambda_{11} F_1 + \lambda_{12} F_2 + \epsilon_1$$
$$x_2 = \lambda_{21} F_1 + \epsilon_2$$
$$x_3 = \lambda_{32} F_2 + \epsilon_3$$
$$x_4 = \lambda_{42} F_2 + \epsilon_4$$

Differs from PCA in that error variance is taken into account.

FA can often give a simpler account with fewer factors or non-zero weights.
**Linear combinations: 2 sets of variables**

Univariate response:

**MRA**: find weights to maximize correlation (R) between y and predicted y,

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \]

**2 sets, multivariate response: MMRA**

Multivariate MRA: find weights to maximize correlation between each y and predicted y,

\[ \hat{y}_1 = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \]
\[ \hat{y}_2 = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3 \]

- Coefficients for each response are the same as in separate MRAs
- But: Multivariate tests take correlations among the y’s into account. Can be more powerful, by “pooling strength.”

**2 sets, multivariate response: CanCorr**

Canonical correlation:

Find linear combinations of the x’s that best predicts linear combination of the y’s

- Choose weights to maximize \( r^2 (v_1, w_1) \)
- Up to \( s = \min(p, q) \) additional pairs of canonical variables: \((v_2, w_2), \ldots (v_s, w_s)\)
- All correlations between the Ys and Xs are explained thru the correlation of each \( v_i \) with \( w_i \).

**Discrete predictors: 2 groups**

**t-test**

\[ \begin{align*}
1 & \quad \text{Male} \\
0 & \quad \text{Female}
\end{align*} \]

**Hotelling’s T^2**

\[ \beta = \bar{y}_M - \bar{y}_F \]

**Multivariate generalization**: find lin. comb. of y’s → max. univariate \( t^2 \). (Wts are discriminant coefficients.)
Discrete predictors: 1 factor

1-way ANOVA

\[ y \sim \mu \beta_1 \beta_2 \]

1-way MANOVA

\[ y_1 y_2 y_3 \sim \mu \beta_1 \beta_2 \]

\begin{array}{ccc}
L & M & H \\
0 & 0 & 0 \\
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{array}

\begin{array}{ccc}
Gp1 & Gp2 & Gp3 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 0 \\
\end{array}

Multivariate generalization: find lin. comb. of y’s \rightarrow max. univariate F

Discrete responses

Continuous predictors: 1 factor

1-way ANOVA

\[ y \sim \mu \beta_1 \beta_2 \]

1-way MANOVA

\[ y_1 y_2 y_3 \sim \mu \beta_1 \beta_2 \]

\begin{array}{ccc}
L & M & H \\
0 & 0 & 0 \\
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{array}

\begin{array}{ccc}
Gp1 & Gp2 & Gp3 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 0 \\
\end{array}

Discrete responses & predictors

Job Satisfac | Education
---|---
L | M | H | L | M | H
1 | 0 | 0 | 1 | 0 | 0
0 | 1 | 0 | 1 | 0 | 0
1 | 0 | 0 | 0 | 1 | 0
0 | 1 | 0 | 0 | 1 | 0
0 | 0 | 1 | 0 | 0 | 1
0 | 1 | 0 | 0 | 0 | 1
0 | 0 | 1 | 0 | 0 | 1
0 | 0 | 1 | 0 | 0 | 1

Education (x)

\begin{array}{ccc}
Lo & M & Hi \\
L & 23 & 10 & 5 \\
M & 12 & 37 & 9 \\
H & 4 & 9 & 43 \\
\end{array}

Satisfaction (y)

\begin{array}{ccc}
L & M & H \\
0 & 0 & 0 \\
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{array}

Simplest example: x2 for 2-way table

Multi-way frequency tables: loglinear models account for associations among discrete factors

\[ \log(f) = X\beta \]

Techniques, by variable type

<table>
<thead>
<tr>
<th>Response variables: y_1, \ldots, y_q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
</tr>
<tr>
<td>q=1</td>
</tr>
<tr>
<td>p=1</td>
</tr>
</tbody>
</table>

Simplest example: 2-way table

Logistic regression as a generalized linear model:

\[ \log \text{odds} = \log \left( \frac{p}{1-p} \right) = X\beta \]

Full generalized linear model for non-normal data:

\[ g(y) = X\beta \]

Loglinear models

<table>
<thead>
<tr>
<th>Predictor variables: x_1, \ldots, x_p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
</tr>
<tr>
<td>p=1</td>
</tr>
<tr>
<td>t-test</td>
</tr>
<tr>
<td>1-way ANOVA</td>
</tr>
<tr>
<td>Hotelling T^2</td>
</tr>
<tr>
<td>Simple \chi^2</td>
</tr>
<tr>
<td>Loglinear models</td>
</tr>
</tbody>
</table>
Graphical methods + Geometry = Insight

- **Graphical methods**: major theme of this course
  - No data analysis is well-begun or well-completed without extensive, well-chosen data displays
  - **Data analysis = Summarization + Exposure** (statistical model) (graphs)
  - **Visual statistics**: Let your data tell you what they seem to say – graphs speak more clearly than a $p$-value.
  - **Visual diagnostics**: graphical methods for diagnosing violations of model assumptions & suggesting corrective actions.

Visual statistics: Why plot your data?

Three data sets with exactly the same bivariate summary statistics:
- Same correlations, linear regression lines, etc
- Indistinguishable from standard printed output

![Visual statistics examples](image)

Graphical methods + Geometry = Insight

- **Geometry**: visual understanding of statistical concepts
  - Regression: fitting lines, planes, hyperplanes
  - Fitting by least squares: projection of $y$ on $X$
  - df: # of dimensions of a vector space
  - SS: lengths of vectors
  - Ellipses: visual summaries of data (data ellipses) and models (confidence ellipses)
  - Helps to use 2D (& 3D) to understand high-D data

Geometry: Data ellipse

Looking at scatterplots:
- What is SD of x? of y?
- What is correlation?
- What is regression line?
- Is relationship linear?
- Are there unusual pts?
Geometry: Data ellipse

Data ellipse:
- Encloses (1-\(\alpha\))% in bivariate normal dist
- 40% = univariate std interval = mean ± 1 SD
- 68% = bivariate std interval

Regression & correlation:
- Regression of y on x goes thru pts of vertical tangency
- correlation is the ratio of height of regression line to height of data ellipse
- visual estimates:
  \[ \text{Inc} = 29 + 1.1 \text{ Exp} \]
  \[ r \approx 0.85 \]

Summary

- Multivariate analysis unifies all traditional linear models within the GLM framework
- Concepts, statistics, and tests apply equally for regression & ANOVA
- All methods involve linear combinations, optimizing some criterion
- Easy generalizations:
  - Multivariate models: \( y = X \beta + \varepsilon \rightarrow Y = X \beta + E \)
  - Non-normal data: models for \( g(y) \)
    - Logistic/logit models: \( \log \left[ \frac{p}{1-p} \right] = X \beta \)
    - Loglinear models: \( \log(f) = X \beta \)
- Graphical methods + Geometry = Insight!