Applications

4433LALG3: Linear Algebra

Week 3, Lecture 11, Valente Ramírez

Mathematical & Statistical Methods group — Biometris, Wageningen University & Research

Overview

- Analysis of variance
- Detecting collinearity
- Statistical distance

Section 1

Analysis of variance

Recap: Useful facts about the vector $\mathbf{1}_n$

Let $\mathbf{y}$ be a vector in $\mathbb{R}^n$.

- The dot product with $\mathbf{1}_n$ computes the sum of entries: $\mathbf{1}_n^\top \mathbf{y} = \sum y_i$.
- The projection $\text{proj}_{\mathbf{1}_n} \mathbf{y}$ computes the mean of $\mathbf{y}$:

$$\begin{aligned}
\text{proj}_{\mathbf{1}_n} \mathbf{y} &= \frac{\mathbf{1}_n^\top \mathbf{y}}{\|\mathbf{1}_n\|^2} \mathbf{1}_n \\
&= \frac{\sum y_i}{n} \mathbf{1}_n \\
&= \bar{y}\, \mathbf{1}_n
\end{aligned}$$

One-way ANOVA

Example

Consider n independent observations of a variable y, where the samples come from one of three possible treatment groups (i.e. a single factor with three levels).

We aim to decompose the i^{th} observation from the j^{th} group, y_{ij}, as:

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij},$$

where μ is the base response, τ_j is the treatment effect, and ε_{ij} is a residual.

The method of *analysis of variance* aims to account for the total sample variance

$$s^2 = \frac{1}{n-1} \sum_{i,j} (y_{ij} - \bar{y})^2,$$

by partitioning it into a component coming from the so-called *within-sample variability* and a component coming from the variability between different treatments.

How exactly is the variance partitioned? And what is the meaning of *"degrees of freedom"*?

One-way ANOVA

The model in the previous slide can be identified as a **linear model** by introducing dummy variables. For each $i = 1, 2, 3$ define:

$$x_i = \begin{cases} 1, & \text{for group } i \\ 0, & \text{otherwise} \end{cases}$$

The model becomes: $y = \mu + \tau_1 x_1 + \tau_2 x_2 + \tau_3 x_3 + \varepsilon$.

Our n–dimensional vectors of observations are:

$$\mathbf{1}_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \hline 1 \\ \vdots \\ 1 \\ \hline 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{g}_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \hline 0 \\ \vdots \\ 0 \\ \hline 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{g}_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \hline 1 \\ \vdots \\ 1 \\ \hline 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{g}_3 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \hline 0 \\ \vdots \\ 0 \\ \hline 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Not full rank

Warning

It should be obvious from the previous slide that the design matrix

$$X = \begin{bmatrix} \mathbf{1}_n & \mathbf{g}_1 & \mathbf{g}_2 & \mathbf{g}_3 \end{bmatrix}$$

does not have full rank: its columns are linearly dependent!

Indeed, $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$ is an independent set and $\mathbf{1}_n = \mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_3$.

Therefore $\mathrm{rank}(X) = 3$, one fewer than the number of columns.

We are dealing with an experimental design of **less than full rank**.

Not full rank: what to do about it?

Let us denote $V = \operatorname{span}\{\mathbf{1}_n, \mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$. As with a linear regression, we project the vector $\mathbf{y}$ onto the subspace V.

The projection $\hat{\mathbf{y}} = \operatorname{proj}_V \mathbf{y}$ is uniquely defined. However, the coefficients μ, τ_i such that $\hat{\mathbf{y}} = \mu\mathbf{1}_n + \tau_1\mathbf{g}_1 + \tau_2\mathbf{g}_2 + \tau_3\mathbf{g}_3$ are not: there are infinitely many ways to expand $\hat{\mathbf{y}}$ as a combination of these four linearly dependent vectors.

In order to identify the parameters uniquely, we impose an additional condition:

$$\tau_1 + \tau_2 + \tau_3 = 0.$$

This also guarantees that μ can be interpreted as the *general mean*.

Under this convention, the other parameters also have a meaningful interpretation:

$\mu + \tau_i$	Mean of group i
τ_i	Difference between general mean and group mean

Fitting parameters

Set $\boldsymbol{\alpha} = \begin{bmatrix} \mu & \tau_1 & \tau_2 & \tau_3 \end{bmatrix}^\top$. We can now fit the parameters by solving the following system of equations for $\boldsymbol{\alpha}$:

$$X^\top X \boldsymbol{\alpha} = X^\top \mathbf{y} \qquad \text{Normal equations}$$

$$\begin{bmatrix} 0 & \mathbf{1}_3^\top \end{bmatrix} \boldsymbol{\alpha} = 0 \qquad \text{Zero-sum constraint}$$

Warning

Because X does not have full rank, it is no longer true that $X^\top X$ is invertible!

Therefore, we cannot use the formula $\boldsymbol{\alpha} = (X^\top X)^{-1} X^\top \mathbf{y}$.

Sum-of-squares decomposition

Define $M = \text{span}\{\mathbf{1}_n\}$. We've seen that $\text{proj}_{\mathbf{1}_n} \mathbf{y} = \bar{y}\,\mathbf{1}_n$.
Thus M represents the subspace of all possibilities for: $\bar{y}\,\mathbf{1}_n$.

Now define $T = \left\{ t_1\mathbf{g}_1 + t_2\mathbf{g}_2 + t_3\mathbf{g}_3 \,\middle|\, t_1 + t_2 + t_3 = 0 \right\}$. This represents the space where we could find the vector: $\hat{\mathbf{y}} - \bar{y}\mathbf{1}_n$.

Finally, define $E = V^{\perp}$. It represents the space of possible residuals: $\mathbf{y} - \hat{\mathbf{y}}$.

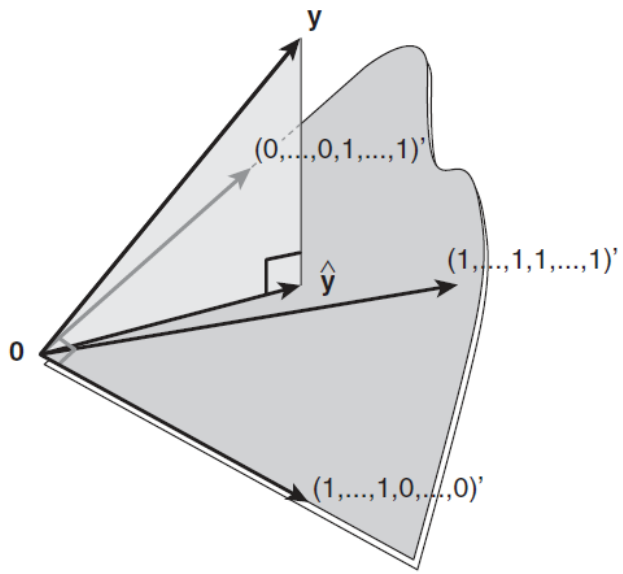

Figure 10.12 The vector geometry of least-squares fit for the overparametrized one-way ANOVA model when there are two groups. The $m+1 = 3$ columns of the model matrix are collinear and span a subspace of dimension $m = 2$.

Source: *John Fox. Applied Regression Analysis and Generalized Linear Models, 3rd ed.*

Sum-of-squares decomposition

Define $M = \text{span}\{\mathbf{1}_n\}$. Recall that $\text{proj}_{\mathbf{1}_n} \mathbf{y} = \bar{y}\mathbf{1}_n$.
Thus M represents the subspace of all possibilities for: $\bar{y}\mathbf{1}_n$.

Now define $T = \left\{ t_1\mathbf{g}_1 + t_2\mathbf{g}_2 + t_3\mathbf{g}_3 \,\middle|\, t_1 + t_2 + t_3 = 0 \right\}$. This represents the space where we could find the vector: $\hat{\mathbf{y}} - \bar{y}\mathbf{1}_n$.

Finally, define $E = V^{\perp}$. It represents the space of possible residuals: $\mathbf{y} - \hat{\mathbf{y}}$.

This yields the decomposition:

$$\mathbb{R}^n = M \oplus T \oplus E,$$

matching the decomposition

$$\mathbf{y} = \bar{y}\mathbf{1}_n + (\hat{\mathbf{y}} - \bar{y}\mathbf{1}_n) + (\mathbf{y} - \hat{\mathbf{y}}).$$

Comparing the squared-lengths:

$$\|\mathbf{y}\|^2 = \|\bar{y}\mathbf{1}_n\|^2 + \|\hat{\mathbf{y}} - \bar{y}\mathbf{1}_n\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2 .$$

This decomposition partitions the total (e.g. uncorrected) sum of squares as:

$$\sum y_{ij}^2 = \sum \bar{y}^2 + \sum (\bar{y}_j - \bar{y})^2 + \sum (y_{ij} - \bar{y}_j)^2.$$

Orthogonality

Warning

When a vector decomposes as

$$\mathbf{c} = \mathbf{a} + \mathbf{b},$$

it is not always true that the *squared-lengths* also decompose as

$$\|\mathbf{c}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2.$$

This is only true when the vectors $\mathbf{a}, \mathbf{b}$ are orthogonal (e.g. Pythagoras' Theorem).

Orthogonality

Recall V was defined as $V = \text{span}\{\mathbf{1}_n, \mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$.

We defined $\hat{\mathbf{y}}$ as the projection onto V, and $\mathbf{e}$ the residual:

$$\hat{\mathbf{y}} = \text{proj}_V \mathbf{y}, \quad \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}.$$

Because this is an orthogonal projection, V and E are always orthogonal.

Question

When we decompose $V = M + T$, why do we get an orthogonal decomposition?

Let $\mathbf{t} = t_1\mathbf{g}_1 + t_2\mathbf{g}_2 + t_3\mathbf{g}_3$ be any vector in T (thus $t_1 + t_2 + t_3 = 0$).

$$\begin{aligned}
\mathbf{1}_n^\top \mathbf{t} &= \mathbf{1}_n^\top (t_1\mathbf{g}_1 + t_2\mathbf{g}_2 + t_3\mathbf{g}_3) \\
&= t_1\mathbf{1}_n^\top \mathbf{g}_1 + t_2\mathbf{1}_n^\top \mathbf{g}_3 + t_3\mathbf{1}_n^\top \mathbf{g}_3 \\
&= t_1 n_1 + t_2 n_2 + t_3 n_3 \\
&= n_1(t_1 + t_2 + t_3) && \text{assuming } n_1 = n_2 = n_3 \\
&= 0.
\end{aligned}$$

Orthogonality

Warning

Throughout this application I have made the implicit assumption that the data set was balanced. This is also called the **orthogonal design**.

Thus a balanced design guarantees orthogonality, which in turn provides a meaningful/interpretable sum-of-squares decomposition.

Degrees of freedom

Degrees of freedom of a sum of squares $\sum w_i^2 = \|\mathbf{w}\|^2$ refers to the **dimension** of the space of all possibilities for the vector $\mathbf{w}$.

If there is one factor with p levels, we have the following:

Subspace	**Dimension**	**Vector**	**Sum-of-squares**	**df**
$\mathbb{R}^n$	n	$\mathbf{y}$	$\sum y_{ij}^2$	n
M	1	$\bar{y}\mathbf{1}_n$	$\sum \bar{y}^2$	1
T	$p-1$	$\hat{\mathbf{y}} - \bar{y}\mathbf{1}_n$	$\sum(\bar{y}_j - \bar{y})^2$	$p-1$
E	$n-p$	$\mathbf{y} - \hat{\mathbf{y}}$	$\sum(y_{ij} - \bar{y}_j)^2$	$n-p$

We can also consider the *corrected* sum-of-squares:

$M^\perp$	$n-1$	$\mathbf{y} - \bar{y}\mathbf{1}_n$	$\sum(y_{ij} - \bar{y})^2$	$n-1$

Section 2

Detecting collinearity

Collinearity example

Example

Consider a data set consisting of grades for a mathematics course. Suppose there are three partial grades and one final grade.

What is the rank of this data set?

Is this data set suitable for a regression analysis?

Dependent variables

Suppose first that the **final grade** is the average of the partial grades.

```
# Data frame contains 3 partial grades
> colnames(grades1)
[1] "test1" "test2" "test3"

# Compute final grade
> grades1$final <- rowMeans(grades1)
> grades1
  test1 test2 test3    final
1   8.2   7.4   5.8   7.1333
2   6.5   7.8   7.1   7.1333
3   6.6   7.5   5.0   6.3667
4   5.3   6.8   6.2   6.1000
5   9.1   8.9   7.4   8.4667
.
> X1 <- as.matrix(grades1)
> R(X1)  # rank
[1] 3

> gaussianElimination(X1)
     test1 test2 test3     final
[1,]     1     0     0   0.33333
[2,]     0     1     0   0.33333
[3,]     0     0     1   0.33333
[4,]     0     0     0   0.00000
[5,]     0     0     0   0.00000
```

Almost-dependent variables

Suppose that the **final grade** is rounded to one decimal.

```
# Copy matrix X1 and recompute 'final'
> X2 <- X1
> X2[,4] <- round(grades1$final, 1)
> X2
     test1 test2 test3 final
[1,]   8.2   7.4   5.8   7.1
[2,]   6.5   7.8   7.1   7.1
[3,]   6.6   7.5   5.0   6.4
[4,]   5.3   6.8   6.2   6.1
[5,]   9.1   8.9   7.4   8.5

> R(X2)
[1] 4
```

Morally, 'final' is the average of the other variables. But numerically, it is not exactly true. Therefore the columns of X_2 are linearly independent, but they **almost** satisfy the linear relationship $\mathbf{x}_4 = \frac{1}{3}(\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3)$.

This could cause numerical instability.

```
> kappa(X2)
[1] 610.3643
> kappa(X2[,1:3])
[1] 25.5681
```

Collinear variables

Suppose the **final grade** is given by $\mathbf{x}_4 = 1 + 0.9\left(\frac{\mathbf{x}_1+\mathbf{x}_2+\mathbf{x}_3}{3}\right)$.

```
# Copy matrix X1 and recompute 'final'
> X3 <- X1
> X3[,4] <- 1 + 0.9*rowMeans(grades1[,1:3]); X3
     test1 test2 test3  final
[1,]   8.2   7.4   5.8   7.42
[2,]   6.5   7.8   7.1   7.42
[3,]   6.6   7.5   5.0   6.73
[4,]   5.3   6.8   6.2   6.49
[5,]   9.1   8.9   7.4   8.62

> R(X3)  # rank
[1] 4
```

Notice how the matrix $\begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_4\end{bmatrix}$ has full rank, but the (potential) design matrix $\begin{bmatrix}\mathbf{1}_n & \mathbf{x}_1 & \dots & \mathbf{x}_4\end{bmatrix}$ does not.

In this example, it is obvious that the variables are collinear, but in other examples it might be hard to identify collinearity.

Collinear variables

Definition

A set of (population) variables $X_1, \ldots, X_k$ is called **collinear** if there exists a perfect linear combination of them that results in a constant variable:

$$a_1 X_1 + \ldots + a_k X_k = c.$$

A sample $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is (almost) collinear if there exists an (almost) perfect linear combination:

$$a_1 \mathbf{x}_1 + \ldots + a_k \mathbf{x}_k = c\,\mathbf{1}_n.$$

Collinear variables

If a set of variables is collinear, bad things could happen.

Suppose $a_1 X_1 + \ldots + a_k X_k = c$, and let $\mathrm{Cov}(\mathbf{X}) = \Sigma$.

Then the variable $W = \mathbf{a}^\top \mathbf{X}$ is constant and so has zero variance.

But $\mathrm{Var}(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \Sigma \, \mathbf{a} = 0$.

This is only possible if Σ is **singular** (i.e. not invertible).

Remark

In general, the eigenvalues of a covariance matrix are non-negative.

- They are **positive** in the absence of collinarity.
- There is a **zero-eigenvalue** if and only if a subset of the variables is collinear.

Collinear variables

We can detect almost-collinearity by looking at the **smallest** eigenvalue of the covariance matrix and checking how close to zero it is.

```
> X3
     test1 test2 test3 final
[1,]   8.2   7.4   5.8  7.42
[2,]   6.5   7.8   7.1  7.42
[3,]   6.6   7.5   5.0  6.73
[4,]   5.3   6.8   6.2  6.49
[5,]   9.1   8.9   7.4  8.62

> eigen(cov(X3))$values
[1] 3.5630e+00 8.3078e-01 1.0261e-01 1.1938e-18

> eigen(cov(X3))$vectors[,4]
[1] -0.26620 -0.26620 -0.26620  0.88735
```

The **last** principal component tells how to combine the variables to obtain an almost-constant variable.

Section 3

Statistical distance

Motivation

Question

Suppose we have a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$.

We know that the mean is $\mu = 10$.

We sample from this distribution once and obtain $x = 60$.

Is this result surprising?

Motivation

The question on the previous slide is only relevant if we know the **standard deviation**.

- If $\sigma = 10$, the result is extremely surprising.
- If $\sigma = 100$, the result is not at all surprising.

Saying that x is 50 units above the mean is not informative (without more context).

Saying that x is 5 standard deviations above the mean is always informative.

How can we measure the distance from the mean for a **multivariate** distribution?

Case 1: Uncorrelated variables: $\sigma_{12} = 0$

Suppose we have a random vector of length 2, $\mathbf{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^\top$, with: $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$, and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$.

Given an observation $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$, we compute the Z–score for each variable to obtain a vector $\mathbf{z} = \begin{bmatrix} z_1 & z_2 \end{bmatrix}^\top$.

The natural way to summarize the size of $\mathbf{z}$ into a single number is to consider $d = \|\mathbf{z}\|$.

Case 1: Uncorrelated variables: $\sigma_{12} = 0$

Let's unpack the formula $d = \|\mathbf{z}\|$. It is easier to first analyze d^2.

$$\begin{aligned}
\|\mathbf{z}\|^2 &= \begin{bmatrix} z_1 & z_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^{-1}(x_1 - \mu_1) & \sigma_2^{-1}(x_2 - \mu_2) \end{bmatrix} \begin{bmatrix} \sigma_1^{-1}(x_1 - \mu_1) \\ \sigma_2^{-1}(x_2 - \mu_2) \end{bmatrix} \\
&= (\mathbf{x} - \boldsymbol{\mu})^\top \begin{bmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{bmatrix} (\mathbf{x} - \boldsymbol{\mu}) \\
&= (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})
\end{aligned}$$

The **statistical distance** between the observation $\mathbf{x}$ and the mean $\boldsymbol{\mu}$ is:

$$d = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}.$$

Case 2: Correlated variables: $\sigma_{12} \neq 0$

How to proceed?

Work in term of the **principal components**, which are **uncorrelated**!

Consider the spectral decomposition: $\Sigma = PDP^\top$.

Let W_1 and W_2 be the principal components.

We will use the following facts:

- $\mathrm{Cov}(\mathbf{W}) = D$,
- Change of basis from the stndard basis to the eigenbasis is given by $P^\top$,
- $\Sigma^{-1} = PD^{-1}P^\top$.

Case 2: Correlated variables: $\sigma_{12} \neq 0$

We begin with the observation $\mathbf{x}$, and do:

- Subtract the mean: $\mathbf{x} - \boldsymbol{\mu}$
- Write in terms of principal components: $\mathbf{w} = P^\top(\mathbf{x} - \boldsymbol{\mu})$
- Compute statistical distance squared: $d^2 = \mathbf{w}^\top D^{-1}\,\mathbf{w}$

But notice that:

$$\begin{aligned} d^2 &= \left(P^\top(\mathbf{x} - \boldsymbol{\mu})\right)^\top D^{-1} \left(P^\top(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= (\mathbf{x} - \boldsymbol{\mu})^\top P D^{-1} P^\top (\mathbf{x} - \boldsymbol{\mu}) \\ &= (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \end{aligned}$$

Statistical distance

Definition

Let $\mathbf{X}$ be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

The **statistical distance**, also called the **Mahalanobis distance** between an observation $\mathbf{x}$ and the mean $\boldsymbol{\mu}$ is:

$$d = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}.$$

The statistical distance is *the correct way* to measure distance taking into account the variance of, and covariance between, the variables.