Visualizing Multivariate Data/Models in R

This document is the working outline for the book

## Introduction

Not quite sure what should go here …

### Preliminaries

Maybe not a separate chapter, but list the main packages used here and datasets for examples.

* R packages
* Datasets

## Getting Started

File: getting\_started.qmd

* Why plot your data?
  + Anscombe data
  + Davis data
* data plots
* model plots
* diagnostic plots

## Plots of Multivariate Data

This chapter introduces a toolbox of basic graphical methods for visualizing multivariate datasets. It starts with some some simple techniques to enhance a basic scatterplot with annotations to summarize the relation between two variables. To visualize more than two variables, we can view all pairs of variables in a scatterplot matrix for all pairs, or shift gears entirely to show multiple variables along a set of parallel axes.

* Bivariate summaries
  + smoothers
  + data ellipses
* Quantitative data:
  + scatterplot matrices
  + corrplots
  + parallel coordinate plots
* Categorical data:
  + mosaic plots
* Generalized pair plots
* Heatmaps

## Dimension reduction techniques: PCA and Biplots

Beyond a few variables, the limitations of 2D (or even 3D) graphs become quickly apparent. In such cases, it is often profitable to view the data in a low-D space that extracts the most “juice” – the important information within the data. This chapter describes the simplest multivariate juicer, principal components analysis and its visualization in the related biplot, which shows the data and the original variables projected into a space of small dimension.

* PCA, the multivariate juicer
* Biplot, a low-dimensional view

## Overview of Linear Models

In this chapter, I review the standard statistical methods for explaining or predicting a quantitative response using a linear model composed of quantitative and/or categorical predictors.

* Regression
* ANOVA
* ANCOVA
* Discriminant analysis
* Regression trees

## Plots for univariate response models

For a univariate linear model fit using lm(), glm() and similar functions, the standard plot() method gives basic versions of *diagnostic* plots of residuals and other calculated quantities for assessing possible violations of the model assumptions. Some of these can be considerably enhanced using other packages.

* the “regression quartet”
* coefficient plots
* marginal (effect) plots
* diagnostic plots

## Collinearity & Ridge Regression

This chapter focuses on the problems associated with high correlations among predictors in linear models, which can lead to numerical instability and paradoxical findings that, while a linear model can be highly predictive, few or none of the independent variables appear to be significant.

* visualizing multicollinearity
  + collinearity diagnostics
  + tableplots
  + collinearity biplots
* ridge regression – generalized ridge trace plots

## Hotelling’s T^2

Just as the one- and two- sample univariate -test is the gateway drug for understanding analysis of variance, so too Hotelling’s test provides an entry point to multivariate analysis of variance. This simple case provides an entry point to understanding the collection of methods I call the **HE plot framework** for visualizing effects in multivariate linear models, which are a main focus of this book.

* as a generalized -test
* properties
* HE plot and discriminant axis

## Brief review of the MLM

The general multivariate linear model (MLM) can be understood as a simple extension of the univariate linear model, with the main difference being that there are multiple response variables instead of just one.

* ANOVA -> MANOVA
* MRA -> MMRA
* ANCOVA -> MANCOVA
* Repeated measures designs

## Visualizing Multivariate Models

* HEplot framework
* Toy example
* HE plot details
  + evidence vs. effect scaling
* Canonical discriminant analysis

## Case studies

* NeuroCog and SocialCog

### MANOVA Examples

This is a collection of examples of multivariate analysis of variance (MANOVA), listed here with the sources from other publications I may draw on as case studies or use in earlier chapters

* iris data [maybe use this earlier?]
* parenting data [from TQMP tutorial]
* diabetes data [from candisc vignette]
* Neuro-Cog & Social-Cog [from TQMP tutorial]

### MMRA Examples

* Rohwer data from [HE-plot-examples vignette & TQMP tutorial]

## Multivariate influence

This material should go earlier …

* univariate influence
* multivariate influence

## Visualizing equality of covariance matrices

* Homogeneity of Variance in Univariate ANOVA
* Homogeneity of variance in MANOVA
* Box’s M test