Homework1

BIOS6643 Fall 2021

8/20/2021

Question 1 PCA

Consider the eNO data, and how we applied PCA to the data for graphical purposes (see Graphs slides). Determine the slope of the regression of Post (Y_2) on Pre (Y_1) values (i.e., a standard 'baseline as covariate' model), and compare this to the 'slope' of the $PC1$ axis. Compare the slopes numerically and superimpose the lines on a scatterplot of Post versus Pre values.

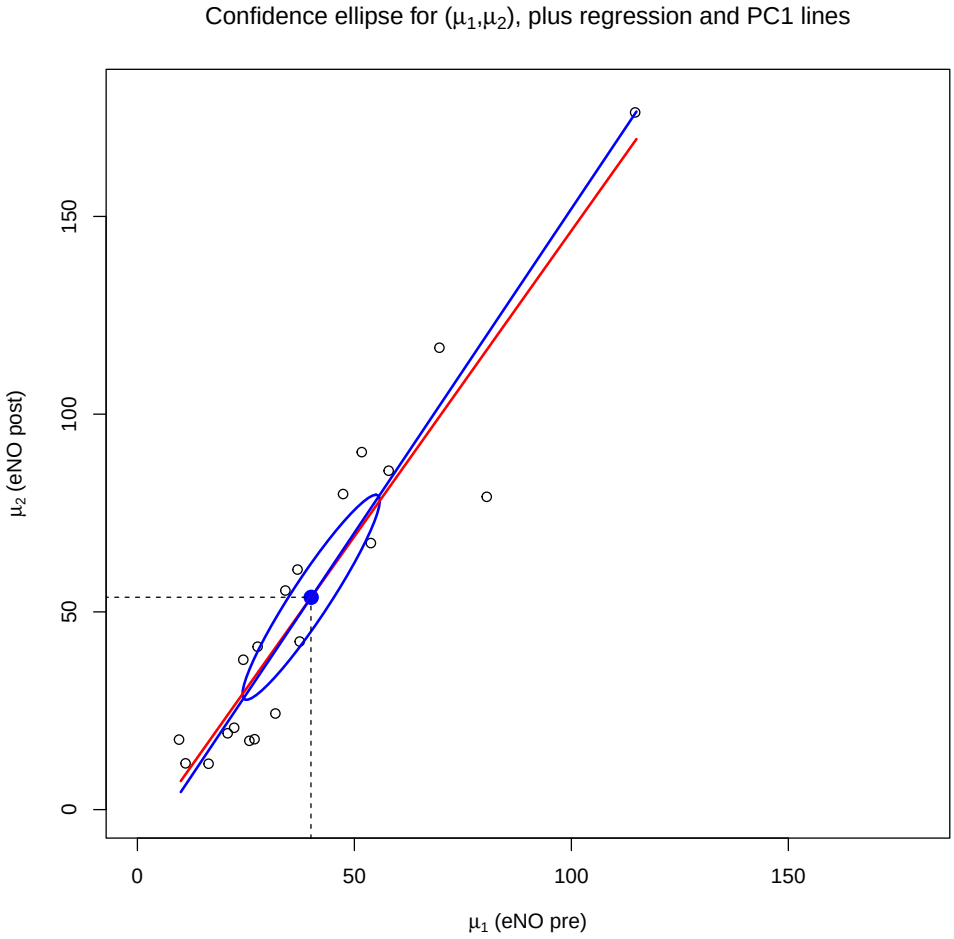
In order to do this, recall $PC1 = aY_1 + bY_2$, where a and b are chosen to maximize the variance of $PC1$ (recall $a = 0.51$, $b = 0.86$ for the data; see the slides).

Note: in terms of Y_2 versus Y_1, the 'slope' of the $PC1$ axis is simply b/a; to create a line to graph for $PC1$, you can have it go through the joint sample mean of Y_1 and Y_2. This exercise helps demonstrate the 'regression' principle in a regression line.

A few comments: First, in terms of the graph, $PC1$ is an axis rather than a line, just like Y_1 and Y_2. This is why we need to anchor it through something; it makes sense to have it go through the joint sample means of Y_1 and Y_2, just like the regression line does. This will allows us to determine an intercept for $PC1$ in addition to the slope, which we already know. See the code in the Appendix that walks through the calculations. Note in the graph below I added the 95Note that the slope of the regression line is $(SD_{post}/SD_{pre}) \times r$ and the slope of the $PC1$ line is SD_{post}/SD_{pre}; since r is close to 1, we do not see much difference between the two.

```
library(car)
library(tidyverse)
library(grDevices)

eno <- here::here("data", "eno_data.txt") %>%
  read.table(header = T,  sep = " ", skip = 0)

# compute radius
N <- length(eno$eno_pre); n <- 2
f <- qf(0.95, n, N - n)
r <- sqrt((n * (N - 1) * f) / ((N - n) * N))
```

```r
# covariance matrix
sigma <- mat.or.vec(2, 2)
sigma[1, 2] <- cov(eno$eno_pre, eno$eno_post); sigma[2, 1] <- sigma[1, 2]
sigma[1, 1] <- var(eno$eno_pre); sigma[2, 2] <- var(eno$eno_post)

# ellipse center (means)
mny1 <- mean(eno$eno_pre); mny2 <- mean(eno$eno_post)
# plot the data
matplot(eno$eno_pre, eno$eno_post,
  xlim = c(0, 180), ylim = c(0, 180),
  xlab = expression(mu[1] * " (eNO pre)"),
  ylab = expression(mu[2] * " (eNO post)"),
  main = expression("Confidence ellipse for (" * mu[1] * "," *  mu[2] *
                    "), plus regression and PC1 lines"), pch = 1)

# add the ellipse
ellipse(center = c(mny1, mny2), shape = sigma, radius = r)

# indicate marginal sample means
segments(40, -10, 40, 53.7, lty = 2); segments(-10, 53.7, 40, 53.7, lty = 2)

# Other Confidence ellipse info
eig <- eigen(sigma); corr <- cov2cor(sigma)

# Parts to answer the HW question
linreg <- lm(eno$eno_post ~ eno$eno_pre)
x <- c(10:115)
linregy <- -8.230 + 1.546 * x
lines(x, linregy, col = "red", lwd = 2)
slope <- sqrt(sigma[2, 2]) / sqrt(sigma[1, 1])
yint <- mean(eno$eno_post) - mean(eno$eno_pre) * (slope)
pcy <- yint + slope * x
lines(x, pcy, col = "blue", lwd = 2)
```


Confidence ellipse for (μ1,μ2), plus regression and PC1 lines
μ2 (eNO post)
μ1 (eNO pre)
0
50
100
150

Question 2 GLM, GzLM, and LMM

In a paragraph, explain the difference between a general linear model (GLM; not a generalized linear model, which I denote with GzLM and which will be discussed more later) and a linear mixed model (LMM).

Basically, a general linear model (GLM) is for independent (e.g., cross-sectional) data, and a linear mixed model (LMM) accounts for correlated data. The GLM is a special case of the LMM when there are no random effects and the error covariance matrix is simple ($\sigma^2\boldsymbol{I}$). Both modeling approaches are regression-type models, where we are trying to understand the relationship between an outcome and several.

Question 3 Profiled likelihood, restricted likelihood, and Likelihood functions

In a short paragraph, explain the difference between a profiled likelihood and a restricted likelihood for a linear mixed model, and how and why they are used. Which one is a re-expression of the standard likelihood?

A profiled likelihood is a re-expression of the standard likelihood. The common profiled likelihood for a linear mixed model is expressed completely in terms of the covariance parameters. This is accomplished by maximizing the likelihood conditioned on the covariance parameters, and then solving for the fixed effects. This leads to an algebraic form for $\hat{\boldsymbol{\beta}}$, expressed as a function of the covariance parameters. This form can then be substituted back in for $\boldsymbol{\beta}$, so that the likelihood is completely expressed in terms of covariance parameters, but it is intrinsically the same likelihood. The restricted likelihood considers a linear form of the original $\boldsymbol{Y}$ that eliminates the fixed effects completely, so it is a different likelihood. The purpose is to get unbiased (or at least less biased) estimators of covariance parameters. The difficulty is there is no true mechanism to estimate the fixed effect parameters with the restricted likelihood, so what is typically done is that the ML algebraic form for $\hat{\boldsymbol{\beta}}$ is employed.

Question 4 Variance in LMM

Derive $Var[\hat{\boldsymbol{\beta}}]$ in a full-rank linear mixed model, given the algebraic form of $\hat{\boldsymbol{\beta}}$ that is obtained via ML estimation.

NOTE: there are two types of variance, model-based and empirical (or sandwich estimator). The difference is whether the middle $\boldsymbol{V}$ is determined via the model or using squared residual quantities; derive *the model-based form*. To answer this question, work with the 'complete data' form of $\hat{\boldsymbol{\beta}}$.

The ML estimator has form $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^t\boldsymbol{V}^-\boldsymbol{Y}$, which is a linear form of $\boldsymbol{Y}$. Since we are dealing with a model with full rank $\boldsymbol{X}$, then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{Y}$. The linear form result says $Var[\boldsymbol{AY}] = \boldsymbol{A}Var[\boldsymbol{Y}]\boldsymbol{A}^t$; so let $\boldsymbol{A} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}$ and

$$
\begin{aligned}
Var[\hat{\beta}] &= \boldsymbol{A}Var[\boldsymbol{Y}]\boldsymbol{A}^t \\
&= \left[(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\right]Var[\boldsymbol{Y}]\left[(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\right]^t \\
&= \left[(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\right]\boldsymbol{V}\left[(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\right]^t \\
&= (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{V}(\boldsymbol{V}^{-1})^t\boldsymbol{X}\left[(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\right]^t \\
&= (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{X}\left[(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\right]^t \\
&= (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}
\end{aligned}
$$

. Another good practice question is to derive $Var[\boldsymbol{L}\hat{\boldsymbol{\beta}}]$ for an estimable $\boldsymbol{L}\boldsymbol{\beta}$, for a less-than-full-rank model.