HOMEWORK 1 MATH 463 – DUE FRIDAY APRIL 14, 2017

Here for a vector $\boldsymbol{y}$ and a subspace V, we denote by $\pi(\boldsymbol{y} \mid V)$ the projection of $\boldsymbol{y}$ on V. Also Π_V denotes the projection operator onto V.

Problem A. (A simulation.) The notation $N(\mu,\sigma^2)$ means the Normal distribution with mean μ and variance σ^2.

Generate values $z_1,\ldots,z_{100},e_1,\ldots,e_{100},d_1,\ldots,d_{100}$ all from a $N(0,1)$ distribution. (Use the R function `rnorm`. Prior to generating these values, use `set.seed` with an argument you select, so that your data can be reproduced. Each student must use a different seed value.) Create values, for $i=1,\ldots,100$,

$$y_i = 10z_i + e_i\,, \quad x_i = 8z_i + d_i\,. \tag{1}$$

(a) Create a scatterplot of y_i against x_i and superimpose the least-squares line.

(b) Suppose that x_{50} is changed to the value 5. Is the right-hand equation in (1) still true for $i=50$? Does the equation in (b) still hold for $i=50$? Does the value y_{50} change?

(c) The random vector (X,Y) has a multivariate Normal distribution if there is an $2\times r$ matrix $\boldsymbol{A}$ and a vector $(W_1,W_2,\ldots,W_r)$ of independent standard Normal random variables such that that

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} W_1 \\ W_1 \\ \vdots \\ W_r \end{bmatrix}$$

Suppose that Z,ε,δ are i.i.d. each with a standard Normal distribution, and

$$Y = 10Z + \varepsilon, \quad X = 8Z + \delta\,. \tag{2}$$

Show that (X,Y) has a multivariate Normal distribution.

(d) If (X,Y) has a multivariate Normal distribution with $\mu_X=\mu_Y=0$, the conditional distribution of Y given $X=x$ is $N(\rho\frac{\sigma_X}{\sigma_Y}x,(1-\rho^2)\sigma_y^2)$, where $\rho=\mathrm{corr}(X,Y)$. Find the conditional distribution of Y given $X=x$ for the pair defined in (2).

(e) Find a fixed number b and Normal random variable γ so that, for any x the random variable

$$Y'_x = bx + \gamma$$

has the same distribution as Y given $X=x$.

(f) Generate $g_1,\ldots,g_{100}$, each with the same distribution as γ found above, and set

$$y'_i = bx_i + g_i\,.$$

Plot $\{y'_i\}$ against $\{x_i\}$. Can you distinguish this plot from the plot made in (a)? What is the least-squares line for this data, and how does it compare to the least-squares line for the data $\{(x_i,y_i)\}$?

(g) What does this exercise say about the ability to infer, based on observational data, the effect of an intervention to change the value of a single variable?

Problem B. Load the R object `gala` by

```
load(url("http://pages.uoregon.edu/dlevin/DATA/gala.R"))
```

The variable $\boldsymbol{y}$ is given in the first column "Species", and the variables $\boldsymbol{x}_i$ for $i = 1,2,3,4,5$ are given by the last four columns. This data records the number of species on islands in the Galapagos chain, along with other geographical and topological variables.

(a) Find the coefficients $b_0, b_1, b_2, b_3, b_4, b_5$ of the least-squares fit

$$\boldsymbol{y} = b_0\mathbf{1} + b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + b_3\boldsymbol{x}_3 + b_4\boldsymbol{x}_4 + b_5\boldsymbol{x}_5 + \boldsymbol{e}$$

(where $\boldsymbol{e} \perp \mathscr{L}\{\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_5\}$) using the function `lm` in R.

(b) Plot $\boldsymbol{e}$ against $\hat{\boldsymbol{y}}$. What does this plot say about the fit of the least-squares linear function?

(c) Compute the least-squares coefficients in R using the matrix multiplication $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Note to get a matrix product in R, use `A %*% B`. The function `solve` can be used to invert a matrix.

Problem C.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)', \boldsymbol{x} = (x_1, \ldots, x_n)', \mathbf{1} = (1, \ldots, 1)'$ and $V = \mathscr{L}(\mathbf{1}, \boldsymbol{x})$.

(a) Use Gram-Schmidt orthogonalization on the vectors $\mathbf{1}, \boldsymbol{x}$ (in this order) to find orthogonal vectors $\mathbf{1}, \boldsymbol{x}^\star$ spanning V. Express $\boldsymbol{x}^\star$ in terms of $\mathbf{1}$ and $\boldsymbol{x}$, and find b_0, b_1 such that

$$\hat{\boldsymbol{y}} = b_0\mathbf{1} + b_1\boldsymbol{x}.$$

To simplify the notation let

$$\boldsymbol{y}^\star = \boldsymbol{y} - \pi(\boldsymbol{y} \mid \mathbf{1}) = \boldsymbol{y} - \bar{y}\mathbf{1},$$

$$\begin{aligned}
S_{xy} &= \langle \boldsymbol{x}^\star, \boldsymbol{y}^\star \rangle = \langle \boldsymbol{x}^\star, \boldsymbol{y} \rangle = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i = \sum_i x_i y_i - \bar{x}\bar{y}n, \\
S_{xx} &= \langle \boldsymbol{x}^\star, \boldsymbol{x}^\star \rangle = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - \bar{x}^2 n, \\
S_{yy} &= \langle \boldsymbol{y}^\star, \boldsymbol{y}^\star \rangle = \sum_i (y_i - \bar{y})^2.
\end{aligned}$$

(b) Suppose

$$\hat{\boldsymbol{y}} = \pi(\boldsymbol{y} \mid V) = a_0\mathbf{1} + a_1\boldsymbol{x}^\star.$$

Find formulas for a_1 and a_0 in terms of $\bar{y}, S_{xy}, S_{xx}$.

(c) Express $\boldsymbol{x}^\star$ in terms of $\mathbf{1}$ and $\boldsymbol{x}$, and use this to determine formulas for b_1 and b_0 so that

$$\hat{\boldsymbol{y}} = b_0\mathbf{1} + b_1\boldsymbol{x}.$$

(d) Express $\|\hat{\boldsymbol{y}}\|^2$ and $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$ in terms of S_{xy}, S_{xx} and S_{yy}.

(e) Use the formula $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ for $\boldsymbol{b} = (b_0, b_1)'$ and verify that they are the same as those found in (c).

(f) For

$$\boldsymbol{y} = \begin{bmatrix} 2 \\ 6 \\ 7 \\ 8 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix},$$

find $a_0, a_1, \hat{\boldsymbol{y}}, b_0, b_1, \|\boldsymbol{y}\|^2, \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$. Verify that

$$\|\hat{\boldsymbol{y}}\| = b_0\langle \boldsymbol{y}, \mathbf{1} \rangle + b_1\langle \boldsymbol{y}, \boldsymbol{x} \rangle,$$

and that $\boldsymbol{y} - \hat{\boldsymbol{y}} \perp V$.

Problem D. Let $\Omega = \mathbb{R}^4$, and

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix},$$

and let $V_0 = \mathscr{L}(\boldsymbol{x}_4)$ for $\boldsymbol{x}_4 = 3\boldsymbol{x}_3 - 2\boldsymbol{x}_2$, $V = \mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$. Find Π_{V_0}, Π_V and Π_{V_1} for $V_1 = V_0^{\perp} \cap V$. For $\boldsymbol{y} = (0, 2, 14, 1)'$ find $\pi(\boldsymbol{y} \mid V_0), \pi(\boldsymbol{y} \mid V_1), \pi(\boldsymbol{y} \mid V)$.

Problem E. For an $n \times k$ matrix $\boldsymbol{X}$ of rank k, what are the eigenvalues and vectors for $\boldsymbol{\Pi} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$? What is $\mathsf{trace}(\boldsymbol{\Pi})$? What is $\mathsf{det}(\boldsymbol{\Pi})$ if $n > k$? If $n = k$?