Spring 2009 Eric Slud

Stat 430 Handout on Partial Correlation

First we recall the definitions. We give formal definitions first for **sample** partial correlations, in a linear algebra framework.

Definition 1. Suppose that $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{z}$ are three n-dimensional vectors, e.g. three labelled columns in an n-observation `SAS` dataset. The *(sample) partial correlation of* $\mathbf{y}$ *and* $\mathbf{x}$ *with* $\mathbf{z}$*-effect removed* is the number $\hat{\rho}_{yx.z}$ given as follows. For a vector $\mathbf{w}$, let $\tilde{w} = \mathbf{w} - \bar{w}\,\mathbf{1}$ be the vector centered to have mean 0, where $\bar{w} = n^{-1}\sum_{i=1}^{n} w_i$. Then, using the notation

$$\hat{\rho}_{vw} \;=\; \mathrm{cor}(\mathbf{v},\mathbf{w}) \;=\; (\tilde{\mathbf{v}}'\tilde{\mathbf{w}})/\sqrt{(\tilde{\mathbf{v}}'\tilde{\mathbf{v}})\,(\tilde{\mathbf{w}}'\tilde{\mathbf{w}})}$$

we define

$$\hat{\rho}_{yx.z} \;=\; \mathrm{cor}\Big(\tilde{\mathbf{y}} - \frac{\tilde{\mathbf{y}}'\tilde{\mathbf{z}}}{\tilde{\mathbf{z}}'\tilde{\mathbf{z}}}\,\tilde{\mathbf{z}}\,,\; \tilde{\mathbf{x}} - \frac{\tilde{\mathbf{x}}'\tilde{\mathbf{z}}}{\tilde{\mathbf{z}}'\tilde{\mathbf{z}}}\,\tilde{\mathbf{z}}\Big) \tag{1}$$

As discussed in class, the two vectors whose correlation is calculated in (1) are respectively the residuals from the simple linear regression of $\mathbf{y}$ on $\mathbf{z}$ and of $\mathbf{x}$ on $\mathbf{z}$. □

These are the same as the definitions given in class, and are the quantities actually calculated in `SAS` by `PROC CORR` with a `partial z` command line.

Correlation and Partial Correlation are also concepts that relate `random variables`, that is, theoretical concepts defining numbers from probability distributions, numbers which are well estimated for large *iid* samples of data (y_i, x_i, z_i) from the theoretical probability distribution. In this case we use capital letters to denote the random variables, and expectations of products $E(VW)$ replace the former inner products $\mathbf{v}'\mathbf{w}$. The constant 'random variable' 1 replaces the former vector $\mathbf{1}$, and the centered random variable $\tilde{W} = W - E(W)$ replaces the former vector $\tilde{\mathbf{w}}$. This is the appropriate replacement because $E(\tilde{W}1) = 0$, just as formerly the centered vectors satisfied $\tilde{\mathbf{w}}'\mathbf{1} = 0$. Now the correlation becomes

$$\rho_{VW} \;=\; Cor(V,W) \;=\; E(\tilde{V}\tilde{W})/\sqrt{E(\tilde{V}^2)\,E(\tilde{W}^2)}$$

since

$$E(\tilde{W}^2) \;=\; E(W - E(W))^2 \;=\; \mathrm{Var}(W)$$

$$E(\tilde{V}\tilde{W}) \;=\; E\Big((V - E(V))(W - E(W))\Big) \;=\; \mathrm{Cov}(V,W)$$

Finally, we have the 'theoretical' partial correlation:

Definition 2. Partial Correlation of random variables Y, X after removing the linear effect of Z is

$$\rho_{YX.Z} \;=\; Cor\Big(\tilde{Y} - \frac{E(\tilde{Y}\tilde{Z})}{E(\tilde{Z}\tilde{Z})}\,\tilde{Z},\; \tilde{X} - \frac{E(\tilde{X}\tilde{Z})}{E(\tilde{Z}\tilde{Z})}\,\tilde{Z}\Big)$$

Worksheet on Partial Correlation

Do the problems on this worksheet and hand them in as Homework.

Problem 1. Suppose that $\mathbf{y}$, $\mathbf{x}$, $\mathbf{z}$ are each n-dimensional vectors with components y_i, x_i, z_i as in Definition 1, with $n > 3$, and assume that the n triplets $(y_i, x_i, z_i) \in \mathbf{R}^3$ are distinct. Suppose that Y, X, and Z are discrete random variables with joint probability mass function given by

$$P((Y, X, Z) = (y_i, x_i, z_i)) = 1/n \quad \text{for} \qquad i = 1, 2, \ldots, n$$

Then show that the quantities $\hat{\rho}_{yx.z}$ given in Definition 1 and $\rho_{YX.Z}$ given in Definition 2 are identical.

Problem 2. Suppose that for some large but fixed value of n, the n-vectors $\mathbf{x}, \mathbf{z}$ are fixed and have positive length ($\mathbf{x}'\mathbf{x} > 0,\ \mathbf{z}'\mathbf{z} > 0$) and means 0 (that is, $\mathbf{x}'\mathbf{1} = \mathbf{z}'\mathbf{1} = 0$) and that for some constants $a, b > 0$ $\mathbf{y} = b\mathbf{x} + a\mathbf{z}$. Show that $\rho_{yx.z} = 1$. The objective is to do this mathematically, but if you cannot prove this in symbols, show it instead with several choices of a, b using `SAS` with the columns $\mathbf{x} = \tilde{\mathbf{u}}$ where $\mathbf{u} =$ `PRICE` and $\mathbf{z} = \tilde{\mathbf{v}}$ for $\mathbf{v} =$ `SQFT` from the dataset `home` in the Data directory of the course web-pages.

Problem 3. Create a `SAS` dataset with $n = 200$ records with 3 columns y, x, z, with entries defined as follows:

$$x_k = x_{100+k} = k/100 \quad \text{for} \quad k = 1, \ldots, 100$$

$$z_k = 1 + \texttt{int}(\frac{k-1}{100}) \quad , \quad y_k = 1 + 20z_k - 3x_k + 0.5\sin(k/10) \ , \quad 1 \le k \le 200$$

Using `SAS` find the correlation between $\mathbf{y}$ and $\mathbf{x}$, and compare it to the partial correlation $\rho_{yx.z}$ after removing the linear effect of z. Try to interpret the result in terms of a plot of $\mathbf{y}$ versus $\mathbf{x}$ using z_k as a plotting character.