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

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

$$\hat{\rho}_{yx.z} = \text{cor}(\tilde{y} - \hat{\tilde{y}}\tilde{z}, \tilde{x} - \hat{\tilde{x}}\tilde{z})$$

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

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 $v'w$. The constant ‘random variable’ 1 replaces the former vector 1, and the centered random variable $\tilde{W} = W - E(W)$ replaces the former vector $\tilde{w}$. This is the appropriate replacement because $E(\tilde{W}1) = 0$, just as formerly the centered vectors satisfied $\tilde{w}'1 = 0$. Now the correlation becomes

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

since

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

$$E(\tilde{V}\tilde{W}) = E\left(\left((V - E(V))(W - E(W))\right)\right) = \text{Cov}(V, W)$$

Finally, we have the ‘theoretical’ partial correlation:

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

$$\rho_{YX.Z} = \text{Cor}\left(\tilde{Y} - \frac{E(\tilde{Y}\tilde{Z})}{E(\tilde{Z}\tilde{Z})}\tilde{Z}, \tilde{X} - \frac{E(\tilde{X}\tilde{Z})}{E(\tilde{Z}\tilde{Z})}\tilde{Z}\right)$$
Worksheet on Partial Correlation

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

**Problem 1.** Suppose that \( y, x, z \) 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 \( (y_i, x_i, z_i) \in \mathbb{R}^3 \) are distinct. Suppose that \( Y, X, \) and \( Z \) are discrete random variables with joint probability mass function given by

\[
P((Y, X, Z) = (y_i, x_i, z_i)) = \frac{1}{n} \quad \text{for} \quad i = 1, 2, \ldots, n
\]

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 \( x, z \) are fixed and have positive length \( (x'x > 0, z'z > 0) \) and means 0 (that is, \( x'1 = z'1 = 0 \)) and that for some constants \( a, b > 0 \) \( y = bx + az \). 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 \( x = \tilde{u} \) where \( u = \text{PRICE} \) and \( z = \tilde{v} \) for \( v = \text{SQFT} \) 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:

\[
x_k = x_{100+k} = \frac{k}{100} \quad \text{for} \quad k = 1, \ldots, 100
\]

\[
z_k = 1 + \text{int}(\frac{k-1}{100}) \quad , \quad y_k = 1 + 20z_k - 3x_k + 0.5\sin(k/10) \quad , \quad 1 \leq k \leq 200
\]

Using SAS find the correlation between \( y \) and \( x \), 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 \( y \) versus \( x \) using \( z_k \) as a plotting character.