Lecture 6: Multiple linear regression

©Christopher S. Bretherton

Winter 2015

A natural extension is to regress a predictand y on multiple predictor variables x_m. **Assuming all variables have been de-meaned**, we fit the linear model:

$$\hat{y} = a_1 x_1 + a_2 x_2 \ldots + a_M x_M. \tag{6.0.1}$$

6.1 Mathematical derivation

Least squares minimization of the mean square residual

$$Q(\mathbf{a}) = N^{-1} \sum_{j=1}^{N} [y_j - \hat{y}_j]^2 = N^{-1} \sum_{j=1}^{N} [y_j - (a_1 x_1 \ldots + a_M x_M)_j]^2$$

with respect to each coefficient a_k gives the **normal equations**:

$$\begin{aligned}
a_1 \overline{x_1 x_1} + a_2 \overline{x_1 x_2} \ldots + a_M \overline{x_1 x_M} &= \overline{x_1 y} \\
a_2 \overline{x_2 x_1} + a_2 \overline{x_2 x_2} \ldots + a_M \overline{x_2 x_M} &= \overline{x_2 y} \\
&\vdots \\
a_M \overline{x_M x_1} + a_2 \overline{x_M x_2} \ldots + a_M \overline{x_M x_M} &= \overline{x_M y}
\end{aligned}$$

The coefficients $\overline{x_i x_p}$ are just the elements C_{ip} of the $M \times M$ **covariance matrix** $\mathbf{C}_{xx}$ between the predictor variables, and the right-hand side is a column vector or $M \times 1$ matrix of covariances between the predictors and the predictand, so the normal equation takes the matrix form:

$$\mathbf{C}_{xx} \mathbf{a} = \mathbf{C}_{xy} \tag{6.1.1}$$

The covariance matrices can be written directly in terms of the $M \times N$ predictor matrix $\mathbf{X}$ (whose rows are the time series of the predictors) and the $1 \times N$ predictand matrix $\mathbf{Y}$ (a row vector):

$$\begin{aligned}
\mathbf{C}_{xx} &= \frac{1}{N-1} \mathbf{X}\mathbf{X}^T \\
\mathbf{C}_{xy} &= \frac{1}{N-1} \mathbf{X}\mathbf{Y}^T
\end{aligned} \tag{6.1.2}$$

As with simple linear regression, it is straightforward to apply multiple regression to a whole array of predictands. since the regression is computed separately for each predictand variable.

6.2 Matlab example

The Matlab script **regression_example.m** was introduced in the previous lecture. It continues with an example of multiple regression of MPG on $M = 2$ predictor variables, car weight and horsepower. Scatterplots show that MPG is highly correlated with horsepower as well as weight, but also that they are even more highly correlated with each other (r = -0.87). Thus, the incremental benefit of adding horsepower as a predictor is not a priori clear.

Using the statistics toolbox function **regress**, the two-predictor linear model is found to have a marginally larger correlation coefficient (0.87 vs. -0.86) with the MPG than does the linear fit with weight alone.

6.3 Overfitting and stepwise linear regression

A concern with multiple regression is **overfitting**; with a lot of predictors and a limited number of samples, random sampling fluctuations will allow some linear combination of the predictors to match the predictand perfectly over the limited samples we have, but the correlations will fall apart for a different set of samples. In particular, if the number m of predictors exceeds the number n of samples, the matrix $\mathbf{C}_{xx}$ will have rank no more than n, so will have at least $m - n$ zero eigenvalues. Then the normal equations will not even be mathematically solvable. Overfitting is often a serious problem for much smaller m than this.

Given independent data samples, a common approach that controls overfitting by keeping the number of predictands to a minimum is **stepwise multiple regression**. Start with the one predictor that explains the most predictand variance (i. e. has the highest correlation coefficient with the predictand). If this correlation coefficient passes an appropriate statistical significance test (i.e. compared to the null hypothesis of regressing a predictor on an equal number of predictors that are all uncorrelated with it), both the predictand and the other predictors are regressed on predictor 1. The linear regression fits are removed to create a modified predictand $\tilde{y}$ and modified predictors $\tilde{x}_{2,3,\ldots}$, all of which are uncorrelated with predictor 1 (this is analogous to creating an orthogonal basis out of an arbitrary set of linearly independent vectors). The modified predictand is then regressed on the modified predictor with which it has the highest correlation coefficient. If this is statistically significant, this predictor is added and a new modified predictor and predict ands are created. The process is repeated until no remaining modified predictor has a statistically significant correlation coefficients with the modified predictand. This process is manually implemented for the second predictor (x_2 = horsepower) in the Matlab script **regression_example.m**. The correlation coefficient of $\tilde{y}$ and $\tilde{x}_2$ is 0.23, which is just barely significant at the 95% level due to the large sample size ($N = 93$, so $2N^{-1/2} = 0.22$). Thus we keep the second predictor despite it only adding marginally to the fitting skill.

The last line of the Matlab script calls the statistics toolbox function **stepwisefit**, which implements this same process. As we did above, it concludes that the second predictor is just barely worth retaining, and calculates the same multiple regression coefficients as did **regress**.

6.4 Cross-validation as a test of overfitting

If the data is serially correlated, one must estimate an effective sample size for use in the stepwise regression. If the data has complicated correlation structure, there may not be a justifiable way to do this. In that case, it is sensible to test the obtained linear fit for overfitting using **cross-validation**; that is, calculating the linear fit based on a subset of the sample and testing whether it also predicting a similar fraction of the variance in the withheld subset of the sample. One way to do this ('leave-one-out') is to successively leave each sample out, compute the regression using the remaining samples, and use the left-out sample for cross-validation. Another approach ('split the data') is to divide the data in halves, and use each half to cross-validate the regression derived from the other half of the data.