Multivariate Regression

September 3, 2024

Multivariate Regression Review (S&S 5.7)

Many methods for multivariate time series analysis build on **multivariate linear regression**, also known as **general linear regression** (not to be confused with generalized linear regression!). When we perform multivariate linear regression, we jointly model r $n \times 1$ response vectors $\boldsymbol{y}_1, \dots, \boldsymbol{y}_r$ arranged as an $n \times r$ matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_r]$ as a linear function of the same $n \times 1$ covariate vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_q$ arranged as an $n \times q$ matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_q]$. We want to find an $q \times r$ matrix of regression coefficients $\boldsymbol{B}$ such that $\boldsymbol{Y} \approx \boldsymbol{X}\boldsymbol{B}$ by solving:

$$\min_{\boldsymbol{\beta}} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}||_2^2, \tag{1}$$

where $||\boldsymbol{Y}||_2^2 = \sum_{i=1}^n \sum_{j=1}^r y_{ij}^2$ gives the sum of squared elements of the matrix $\boldsymbol{Y}$.

We still refer to the quantity $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}||_2^2$ as the **residual sum of squares (RSS)**, as it measures how much of the variability of $\boldsymbol{Y}$ remains after subtracting off a linear function of the covariates. We can also still minimize (1) by differentiating; the minimizing value $\hat{\boldsymbol{B}}$ will satisfy:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{B}} - \boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{0} \implies \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{X}'\boldsymbol{Y}.$$

If the matrix $\boldsymbol{X}$ is full rank with rank q, then the minimizing value is

$$\hat{\boldsymbol{B}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Y}. \tag{2}$$

If we want to say more about $\hat{\boldsymbol{B}}$, we need to make some more assumptions. First, note that we can always decompose the observed response $\boldsymbol{Y}$ into a linear part $\boldsymbol{XB}$ and a remainder $\boldsymbol{W}$:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{W}. \tag{3}$$

If we assume:

- $\mathbb{E}[\boldsymbol{W}] = \boldsymbol{0}$, then $\hat{\boldsymbol{B}}$ is **unbiased**, i.e. $\mathbb{E}\left[\hat{\boldsymbol{B}}\right] = \boldsymbol{B}$.
- $\boldsymbol{w}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}_w)$, where $\boldsymbol{w}_i$ are columns of the remainder $\boldsymbol{W}$, then:
  - ($\star$) $\hat{\boldsymbol{B}}$ is the maximum likelihood estimator of $\boldsymbol{B}$;
  - ($*$) Elements of $\hat{\boldsymbol{B}}$ are normally distributed, with $\mathbb{V}\left[\hat{\boldsymbol{b}}_i\right] = \sigma_{ii}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and $\mathrm{Cov}\left[\hat{\boldsymbol{b}}_i, \hat{\boldsymbol{b}}_j\right] = \sigma_{ij}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ where $\hat{\boldsymbol{b}}_i$ be the i-th column of $\hat{\boldsymbol{B}}$;
  - ($\dagger$) The residuals $\boldsymbol{R} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}$ are normally distributed, with $\mathbb{E}[\boldsymbol{R}] = \boldsymbol{0}$, $\mathbb{V}[\boldsymbol{r}_i] = \sigma_{ii}\left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)$, and $\mathrm{Cov}[\boldsymbol{r}_i, \boldsymbol{r}_j] = \sigma_{ij}\left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)$ where $\boldsymbol{r}_i$ be the i-th column of $\boldsymbol{R}$;
  - ($\circ$) $\hat{\boldsymbol{B}}$ and $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}$ are independent.

We're not going to derive ($\star$) this time around. Standard practice for constructing standard errors and confidence intervals is to use ($*$), plugging in an unbiased estimator of the variance-covariance matrix $\boldsymbol{\Sigma}_w$:

$$\boldsymbol{S}_w = \frac{\boldsymbol{R}'\boldsymbol{R}}{n-q}. \tag{4}$$

Note that this is *not* the maximum likelihood estimate of $\boldsymbol{\Sigma}_w$ - the maximum likelihood estimator $\hat{\sigma}_w^2 = \boldsymbol{R}'\boldsymbol{R}/n$ is biased.

It follows from ($*$), ($\dagger$), and ($\circ$) that

$$t_{n-q} = \frac{\hat{b}_{ij} - b_{ij}}{\sqrt{s_{w,jj}}\sqrt{\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{ii}}} \sim \mathcal{T}_{n-q}. \tag{5}$$

This gives us a way of testing the null hypothesis that b_{ij} is exactly equal to a specific value because it tells us the approximate distribution of $\hat{b}_{ij}$ for specific values of b_{ij}. We call such tests ***t*-tests**.

***F*-tests** are a bit trickier to derive for multivariate linear models, so we'll just talk about performing model selection (choosing the covariates or columns of $\boldsymbol{X}$ to use) using AIC, AICc and SIC. Letting $\boldsymbol{X}_k$ refer to a matrix containing k covariates and $\boldsymbol{B}_k$ and $\hat{\boldsymbol{B}}_k$ the corresponding regression coefficients and their linear regression estimates, several popular methods for performing model selection are:

(⋆) Compute **Akaike's Information Criterion (AIC)**

$$AIC = \ln\left(\left|\frac{\left(\boldsymbol{Y} - \boldsymbol{X}_k\hat{\boldsymbol{B}}_k\right)'\left(\boldsymbol{Y} - \boldsymbol{X}_k\hat{\boldsymbol{B}}_k\right)}{n}\right|\right) + \frac{2}{n}\left(rk + \frac{r\left(r+1\right)}{2}\right) \tag{6}$$

for models with k and k' covariates, and choose the model with the lower AIC value.

(⋆) Compute **AIC, Bias Corrected (AICc)**

$$AICc = \ln\left(\left|\frac{\left(\boldsymbol{Y} - \boldsymbol{X}_k\hat{\boldsymbol{B}}_k\right)'\left(\boldsymbol{Y} - \boldsymbol{X}_k\hat{\boldsymbol{B}}_k\right)}{n}\right|\right) + \frac{r\left(n+q\right)}{n-r-q-1} \tag{7}$$

for models with k and k' covariates, and choose the model with the lower $AICc$ value.

(⋆) Compute **Schwarz's/Bayesian Information Criterion (SIC/BIC)**

$$SIC = \ln\left(\left|\frac{\left(\boldsymbol{Y} - \boldsymbol{X}_k\hat{\boldsymbol{B}}_k\right)'\left(\boldsymbol{Y} - \boldsymbol{X}_k\hat{\boldsymbol{B}}_k\right)}{n}\right|\right) + \left(kr + r\left(r+1\right)/2\right)\left(\frac{\log\left(n\right)}{n}\right) \tag{8}$$

for models with k and k' covariates, and choose the model with the lower SIC value.

Recall that whether AIC, AICc, or BIC is most appropriate for a given problem is problem-specific; AICc can perform better than AIC when n is relatively small, and SIC/BIC can perform better than AIC when the number of covariates k is relatively large. Because including one additional covariate (column of $\boldsymbol{X}$) yields r additional regression coefficients when we are performing multivariate linear regression, we may tend to prefer SIC/BIC.