STAT 730 Chapter 6: Regression

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 730: Multivariate Analysis

Multivariate regression, simplest model

In STAT 704–705 we consider the model

$$y_i = \boldsymbol{\beta}' \mathbf{x}_i + u_i,$$

where $\mathbf{x}_i \in \mathbb{R}^q$ and $i = 1, \ldots, n$. Instead of having one response, we may have p correlated responses in $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$. Assume $V(\mathbf{y}_i) = \boldsymbol{\Sigma}$. Your book entertains a separate regression model for each element of $\mathbf{y}_i$. We will initially follow the book, but then discuss a few departures to this model that are in common use and easy to fit. Initially consider:

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{ip} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}'_{(1)} \mathbf{x}_i \\ \vdots \\ \boldsymbol{\beta}'_{(p)} \mathbf{x}_i \end{bmatrix} + \begin{bmatrix} u_{i1} \\ \vdots \\ u_{ip} \end{bmatrix},$$

or

$$\mathbf{y}_i = \mathbf{B}' \mathbf{x}_i + \mathbf{u}_i,$$

where $\mathbf{B} = [\boldsymbol{\beta}_{(1)} \cdots \boldsymbol{\beta}_{(p)}] \in \mathbb{R}^{q \times p}$.

Simplest regression model, continued

If we knock these over on their side, i.e. take the transpose, we have a $1 \times p$ row vector

$$\mathbf{y}_i' = \mathbf{x}_i'\mathbf{B} + \mathbf{u}_i'.$$

Stacking everything up we have

$$\underbrace{\mathbf{Y}}_{n\times p} = \underbrace{\mathbf{X}}_{n\times q}\underbrace{\mathbf{B}}_{q\times p} + \underbrace{\mathbf{U}}_{n\times p},$$

where $\mathbf{U}$ is d.m. from $N_p(\mathbf{0}, \mathbf{\Sigma})$.

Log-likelihood of multivariate normal data

Assume $\text{rank}(\mathbf{X}) = q$ and $n > p + q$.

$$L(\mathbf{Y}; \mathbf{B}, \mathbf{\Sigma}) = \prod_{i=1}^{n} |2\pi\mathbf{\Sigma}|^{-1/2} \exp\{-\tfrac{1}{2}(\mathbf{y}_i - \mathbf{B}'\mathbf{x}_i)'\mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{B}'\mathbf{x}_i)\},$$

implies

$$l(\mathbf{Y}; \mathbf{B}, \mathbf{\Sigma}) = -\tfrac{n}{2} \log |2\pi\mathbf{\Sigma}| - \tfrac{1}{2}\text{tr}\{(\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XB})'\}.$$

MLEs, simplest case

Let $\mathbf{P} = \mathcal{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathcal{I}_n - \mathbf{P}_\mathbf{X}$.

thm: MLEs are $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{Y}'\mathbf{P}\mathbf{Y}$.

Proof: Write

$$l(\mathbf{B}, \mathbf{\Sigma}) = -\tfrac{n}{2}\log|2\pi\mathbf{\Sigma}| - \tfrac{1}{2}\mathrm{tr}\{\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}\} - \tfrac{1}{2}\mathrm{tr}\{\mathbf{\Sigma}^{-1}\underbrace{(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{X}'\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B})}_{\text{indep. } \mathbf{\Sigma},\ \geq 0}\}.$$

This is maximized when $(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{X}'\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B})$ is minimized which is at $\mathbf{B} = \hat{\mathbf{B}}$. Plugging this into the log-likelihood we get

$$l(\hat{\mathbf{B}}, \mathbf{\Sigma}) = -\tfrac{np}{2}\log(2\pi) - \tfrac{n}{2}(\log|\mathbf{\Sigma}| + \mathrm{tr}\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}),$$

which is maximized at $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}$ from the result in Chapter 4. $\square$

Note that $(\hat{\mathbf{B}}, \hat{\mathbf{\Sigma}})$ are sufficient for $(\mathbf{B}, \mathbf{\Sigma})$.

MLEs, simplest case

thm: $\hat{\mathbf{B}}$ indep. $\hat{\mathbf{\Sigma}}$.

Proof: $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{Y}'\mathbf{P}\mathbf{Y} = \frac{1}{n}(\mathbf{Y} - \mathbf{B}\mathbf{X})'\mathbf{P}(\underbrace{\mathbf{Y} - \mathbf{B}\mathbf{X}}_{\text{d.m. } N_p(\mathbf{0}, \mathbf{\Sigma})})$ and

$\hat{\mathbf{B}} - \mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}(\underbrace{\mathbf{Y} - \mathbf{B}\mathbf{X}}_{\text{d.m. } N_p(\mathbf{0}, \mathbf{\Sigma})})$. Now note $\mathbf{P}\mathbf{Y}$ indep. $\hat{\mathbf{B}} - \mathbf{B}$ from Craig's theorem because $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{P} = \mathbf{0}$. $\square$

MLEs, sampling distributions

$n\hat{\mathbf{\Sigma}} = (\mathbf{Y} - \mathbf{BX})'\mathbf{P}(\underbrace{\mathbf{Y} - \mathbf{BX}}_{\text{d.m. } N_p(\mathbf{0}, \mathbf{\Sigma})})$, so Cochran's theorem gives us

$n\hat{\mathbf{\Sigma}} \sim W_p(\mathbf{\Sigma}, \underbrace{\text{rank}(\mathbf{P})}_{n-q})$. Now note

$$\begin{bmatrix} \mathbf{y}_{(1)} \\ \mathbf{y}_{(2)} \\ \vdots \\ \mathbf{y}_{(p)} \end{bmatrix} \sim N_{np}\left(\begin{bmatrix} \mathbf{X} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)} \\ \vdots \\ \boldsymbol{\beta}_{(p)} \end{bmatrix}, \begin{bmatrix} \sigma_{11}\mathcal{I}_n & \sigma_{12}\mathcal{I}_n & \cdots & \sigma_{1p}\mathcal{I}_n \\ \sigma_{21}\mathcal{I}_n & \sigma_{22}\mathcal{I}_n & \cdots & \sigma_{2p}\mathcal{I}_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}\mathcal{I}_n & \sigma_{p2}\mathcal{I}_n & \cdots & \sigma_{pp}\mathcal{I}_n \end{bmatrix} \right).$$

In matrix terms $\mathbf{Y}^v \sim N_{np}([\mathcal{I}_p \otimes \mathbf{X}]\mathbf{B}^v, \mathbf{\Sigma} \otimes \mathcal{I}_n)$.

Sampling dist'n of $\hat{\mathbf{B}}^v$

From last slide, $\mathbf{Y}^v \sim N_{np}([\mathcal{I}_p \otimes \mathbf{X}]\mathbf{B}^v, \mathbf{\Sigma} \otimes \mathcal{I}_n)$. We can show $\hat{\mathbf{B}}^v = [\mathcal{I}_p \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}^v$. So

$$\hat{\mathbf{B}}^v \sim N_{qp}(\underbrace{[\mathcal{I}_p \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}][\mathcal{I}_p \otimes \mathbf{X}]\mathbf{B}^v}_{\mathbf{B}^v}, \underbrace{[\mathcal{I}_p \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}][\mathbf{\Sigma} \otimes \mathcal{I}_n][\mathcal{I}_p \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]'}_{\mathbf{\Sigma}\otimes(\mathbf{X}'\mathbf{X})^{-1}}).$$

That is, the sampling distribution of $\hat{\mathbf{B}}^v$ is

$$\hat{\mathbf{B}}^v \sim N_{qp}(\mathbf{B}^v, \mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}).$$

Note then, marginally,

$$\hat{\boldsymbol{\beta}}_{(j)} \sim N_q(\boldsymbol{\beta}_{(j)}, \sigma_{jj}[(\mathbf{X}'\mathbf{X})^{-1}]).$$

Since we also have $n\hat{\sigma}_{jj} \sim \sigma_{jj}\chi^2_{n-q}$ independent of $\hat{\boldsymbol{\beta}}_{(j)}$ we can make t-statistics and obtain confidence intervals for elements of $\boldsymbol{\beta}_{(j)}$. Let β_{ij} be the ith element of $\boldsymbol{\beta}_{(j)}$. Then

$$\frac{(\hat{\beta}_{ij} - \beta_{ij})/\sqrt{[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}}{\sqrt{n\hat{\sigma}_{jj}/(n-q)}} \sim t_{n-q}.$$

ANOVA decomposition

Instead of sums of squares, we now have matrices of sums of squares and cross products (SSCP). Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$.

$$\mathbf{T} = \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}' = \hat{\mathbf{U}}'\hat{\mathbf{U}} + [\hat{\mathbf{Y}}'\hat{\mathbf{Y}} - n\bar{\mathbf{y}}\bar{\mathbf{y}}'] = \mathbf{E} + \mathbf{R},$$

where $\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}$. Here, $\mathbf{E}$ is the SSCP for error, $\mathbf{R}$ is the SSCP for regression, and $\mathbf{T}$ is the total SSCP. Note also $\mathbf{T} = \mathbf{Y}'(\mathcal{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Y}$, $\mathbf{E} = \mathbf{Y}'(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}$ and $\mathbf{R} = \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{1}_n})\mathbf{Y}$ where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{P}_{\mathbf{1}_n} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$.

ANOVA decomposition

The **E** and **R** matrices cook up a test that nothing in the model is important beyond p intercepts, i.e. $H_0 : \mathbf{B} = [\beta_{10} \cdots \beta_{p0}]$. The test statistics are based on $\mathbf{ET}^{-1}$, as we shall see shortly.

This is a bit different than univariate regression where we focus on SSR/SSE. Instead we look at $SSE/SSTO = 1/(1 + \frac{SSR}{SSE})$. Recall that this has a beta distribution in the univariate case when H_0 is true. For the multivariate case we will have Wilks lambda from the LRT, the product of betas.

Hypotheses among elements of **B**

For the dental data of Pothoff and Roy (1964) there is only one covariate, gender. There are 16 boys and 11 girls.

$$\begin{bmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_{16}' \\ \mathbf{y}_{17}' \\ \vdots \\ \mathbf{y}_{27}' \end{bmatrix}_{27\times 4} = \begin{bmatrix} \mathbf{1}_{16} & \mathbf{0}_{16} \\ \mathbf{0}_{11} & \mathbf{1}_{11} \end{bmatrix}_{27\times 2} \begin{bmatrix} \mu_{m1} & \mu_{m2} & \mu_{m3} & \mu_{m4} \\ \mu_{f1} & \mu_{f2} & \mu_{f3} & \mu_{f4} \end{bmatrix}_{2\times 4} + \begin{bmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_{16}' \\ \mathbf{u}_{17}' \\ \vdots \\ \mathbf{u}_{27}' \end{bmatrix}_{27\times 4}.$$

Of interest is testing (1) growth is linear, and (2) there is no difference between boys and girls.

Growth is linear if $H_0 : \mu_{f2} - \mu_{f1} = \mu_{f3} - \mu_{f2} = \mu_{f4} - \mu_{f3}$ and $\mu_{m2} - \mu_{m1} = \mu_{m3} - \mu_{m2} = \mu_{m4} - \mu_{m3}$. This is written

$$\begin{bmatrix} \mu_{m1} & \mu_{m2} & \mu_{m3} & \mu_{m4} \\ \mu_{f1} & \mu_{f2} & \mu_{f3} & \mu_{f4} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 2 & -1 \\ -1 & 2 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hypotheses among covariate groups

No difference between boys and girls is written $H_0 : \mu_{m1} = \mu_{f1}$ and $\mu_{m2} = \mu_{f2}$ and $\mu_{m3} = \mu_{f3}$ and $\mu_{m4} = \mu_{f4}$. In matrix terms

$$\begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_{m1} & \mu_{m2} & \mu_{m3} & \mu_{m4} \\ \mu_{f1} & \mu_{f2} & \mu_{f3} & \mu_{f4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}.$$

Both hypotheses are written $H_0 : \mathbf{CBM} = \mathbf{D}$. The text considers both LRT and UIT of this (very general) hypothesis.

Let's derive LRT and UIT for a simpler hypothesis. Let $\mathbf{B} = \begin{bmatrix} \mathbf{B}_1((q-k) \times p) \\ \mathbf{B}_2(k \times p) \end{bmatrix}_{q \times p}$ and consider $H_0 : \mathbf{B}_2 = 0$. Similarly let $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$. H_0 tests whether we can drop the last k predictors in each of the p regressions. This test is the common "big model / little model" test for regression, where the first column of $\mathbf{X}$ is $\mathbf{x}_{(1)} = \mathbf{1}_n$,

LRT of $H_0 : \mathbf{B}_2 = \mathbf{0}$

Under $H_0 : \mathbf{B}_2 = \mathbf{0}$, $\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1$ d.m. $N_p(\mathbf{0}, \mathbf{\Sigma})$.

We can show $n\hat{\mathbf{\Sigma}}_r = \mathbf{Y}'(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}$ and $n\hat{\mathbf{\Sigma}}_f = \mathbf{Y}'(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}$ where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{P}_{\mathbf{X}_1} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Plugging into the log-likelihood (p. 159) gives

$$\lambda^{2/n} = \frac{L_0^*}{L_1^*} = \frac{|\mathbf{Y}'(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}|}{|\mathbf{Y}'(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}|} = \frac{|\mathbf{Y}'(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}|}{|\mathbf{Y}'(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y} + \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}|} = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|}.$$

Under H_0, $\mathbf{E} = (\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1)'(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})(\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1)$ and $\mathbf{H} = (\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1)'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_1})(\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1)$. Cochran's theorem tells us that $\mathbf{E} \sim W_p(\mathbf{\Sigma}, n - p)$ and $\mathbf{H} \sim W_p(\mathbf{\Sigma}, k)$; Craig's theorem tells us they are independent, as $(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_1}) = \mathbf{0}$. Therefore,

$$\lambda^{2/n} \sim \Lambda(p, n - p, k),$$

Wilk's lambda.

UIT of $H_0 : \mathbf{B}_2 = \mathbf{0}$

Since $\mathbf{y}_i \sim N_p(\mathbf{B}'\mathbf{x}_i, \mathbf{\Sigma})$ we have $\mathbf{a}'\mathbf{y}_i \sim N(\mathbf{a}'\mathbf{B}'\mathbf{x}_i, \mathbf{a}'\mathbf{\Sigma}\mathbf{a}) = N\left(\sum_{j=1}^{p} a_j \beta'_{(j)}\mathbf{x}_i, \mathbf{a}'\mathbf{\Sigma}\mathbf{a}\right)$. This is tested via (homework!)

$$\frac{(\mathbf{Ya})'(\mathbf{P_X} - \mathbf{P_{X_1}})(\mathbf{Ya})}{(\mathbf{Ya})'(\mathcal{I}_n - \mathbf{P_X})(\mathbf{Ya})} = \frac{\mathbf{a}'\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_{X_1}})\mathbf{Ya}}{\mathbf{a}'\mathbf{Y}'(\mathcal{I}_n - \mathbf{P_X})\mathbf{Ya}} \sim \frac{k}{n-p} F_{k,n-p},$$

under H_{0a} : last k elements of $\mathbf{B}'\mathbf{a}$ are zero. Note that this holds for all $\mathbf{a} \Leftrightarrow \mathbf{B}_2 = \mathbf{0}$. Maximizing over $\mathbf{a}$ gives λ_1, the largest e-value of $\mathbf{HE}^{-1}$. Page 84 implies that $\theta = \frac{\lambda_1}{1+\lambda_1}$, the largest e-value of $\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}$ is $\theta \sim \theta(p, k, n-p)$, Roy's greatest root.

General linear hypothesis for simplest model

LRT and UIT for the general hypothesis $H_0 : \mathbf{CBM} = \mathbf{D}$ are derived on pp. 161–163. Here, $\mathbf{C} \in \mathbb{R}^{g \times q}$ and $\mathbf{M} \in \mathbb{R}^{p \times r}$. Let $\mathbf{H} = (\mathbf{C}\hat{\mathbf{B}}\mathbf{M} - \mathbf{D})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\hat{\mathbf{B}}\mathbf{M} - \mathbf{D})$ and $\tilde{\mathbf{E}} = \mathbf{M}'\mathbf{E}\mathbf{M}$. Then

$$(LRT) : \lambda^{2/n} = \frac{|\tilde{\mathbf{E}}|}{|\tilde{\mathbf{E}} + \mathbf{H}|} \sim \Lambda(r, n - q, g),$$

and

$$(UIT) : \theta \sim \theta(r, n - q, g), \ \theta \text{ is greatest e-value of } \mathbf{H}(\tilde{\mathbf{E}} + \mathbf{H})^{-1}.$$

Note that this test is for any design $\mathbf{X}$ in the model $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, including a MANOVA design where $\mathbf{X} = \text{block-diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_k})$ as in the dental data a few slides ago.

Dental data hypothesis tests

```
library(reshape) # need to turn $\by$ into $\bY$.
library(car)     # allows for multivariate linear hypotheses
library(heavy)   # has dental data
data(dental)
d2=cast(melt(dental,id=c("Subject","age","Sex")),Subject+Sex~age)
names(d2)[3:6]=c("d8","d10","d12","d14")
r=lm(cbind(d8,d10,d12,d14)~0+Sex,data=d2) # no intercept!
model.matrix(cbind(d8,d10,d12,d14)~0+Sex,data=d2)
M=matrix(c(1,-2,1,0,0,1,-2,1),4,2)
C=matrix(c(1,-1),1,2)
linearHypothesis(r,hypothesis.matrix=C) # accept linear trends!
linearHypothesis(r,hypothesis.matrix=diag(2),P=M) # sexes different
```

The output includes Wilk's lambda (LRT) and Roy's greatest root (UIT), as well as two additional tests: Pillai-Bartlett trace and Hotelling-Lawley Trace. All four are based on $\mathbf{HE}^{-1}$.

Uh oh...

We accept linear trends for both boys and girls. Such a model would help with interpretation and we could also investigate whether growth *rate* is different in boys and girls.

Seems like we need a more general model...

$$y_{ij} = [\beta_0 + \tau_0 g_i] + [\beta_1 + \tau_1 g_i] t_j + u_{ij},$$

where $\mathbf{u}_i \overset{iid}{\sim} N_p(\mathbf{0}, \mathbf{\Sigma})$, $t_j = 8 + 2(j-1)$, and g_i is gender coded 0/1. This allows different linear growth for boys and girls. We may want to then test $H_0 : \tau_1 = 0$, i.e. the growth rate is the same.

Different predictors for each outcome

We first generalize the simplest model by allowing different predictors for each of the p outcomes.

Different predictors may be important for each measurement in $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$. This results in

$$\mathbf{y}_{(j)} = \mathbf{X}_{(j)}\boldsymbol{\beta}_{(j)} + \mathbf{u}_{(j)},$$

a different regression for each outcome $\mathbf{y}_{(j)}$. Let $\boldsymbol{\beta}_{(j)} \in \mathbb{R}^{q_j}$ and take $\mathbf{Y}^v \sim N_{np}(\mathbf{X}^*\mathbf{B}^v, \boldsymbol{\Sigma} \otimes \mathcal{I}_n)$ where

$$\mathbf{X}^* = \begin{bmatrix} \mathbf{X}_{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_{(p)} \end{bmatrix}_{np \times q+}. \text{ Then}$$

$$\hat{\mathbf{B}}^v = [\mathbf{X}^{*\prime}(\boldsymbol{\Sigma}^{-1} \otimes \mathcal{I}_n)\mathbf{X}^*]^{-1}\mathbf{X}^{*\prime}(\boldsymbol{\Sigma}^{-1} \otimes \mathcal{I}_n)\mathbf{Y}^v$$

is the MLE of $\mathbf{B}^v$.

Different predictors for each outcome, con't

To estimate $\mathbf{\Sigma}$, we need $\hat{y}_{ij} = E(y_{ij})$ under the model. Let $\hat{\mathbf{Y}}^v = \mathbf{X}^* \hat{\mathbf{B}}^v$ be the fitted values, the estimates of $E(y_{ij})$. We can get the fitted vector $\hat{\mathbf{y}}_i$ as $\hat{\mathbf{y}}_i = [\mathcal{I}_p \otimes \mathbf{e}_i'] \hat{\mathbf{Y}}^v$ where $\mathbf{e}_i = \mathbf{0}_n$ with a one at element i.

Then $\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)(\mathbf{y}_i - \hat{\mathbf{y}}_i)'$ is the MLE of $\mathbf{\Sigma}$.

This model is more general than the simplest model, but still does not allow parameters to be shared across the p responses in $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$. Such models are important in fitting longitudinal or spatial data models, commonly allowed in SAS' `proc mixed` or in R's `gls` and `lme` functions from the `nlme` package.

This model is also a special case of the general linear model coming up obtained by zeroing out some predictors.

Rewriting the simplest model

Recall $\mathbf{Y}^v \sim N_{np}([\mathcal{I}_p \otimes \mathbf{X}]\mathbf{B}^v, \mathbf{\Sigma} \otimes \mathcal{I}_n)$. Instead we can write

$$\mathbf{y}_i \overset{ind.}{\sim} N_p(\mathcal{I}_p \otimes \mathbf{x}_i\mathbf{B}^v, \mathbf{\Sigma}).$$

If we stack these into one vector $\mathbf{y}' = (\mathbf{y}_1', \ldots, \mathbf{y}_n')$ we get

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} \sim N_{np}\left(\begin{bmatrix} \mathcal{I}_p \otimes \mathbf{x}_1' \\ \vdots \\ \mathcal{I}_p \otimes \mathbf{x}_n' \end{bmatrix} \mathbf{B}^v, \mathcal{I}_n \otimes \mathbf{\Sigma}\right).$$

We can define $\mathbf{X}_i = \mathcal{I}_p \otimes \mathbf{x}_i$ and stack these as well into $\mathbf{X}' = [\mathbf{X}_1' \cdots \mathbf{X}_n']_{np \times pq}$. Then we have

$$\mathbf{y} \sim N_{np}(\mathbf{X}\beta, \mathcal{I}_n \otimes \mathbf{\Sigma}),$$

where $\beta = \mathbf{B}^v$, in the form of a general linear model.

Rewriting model w/ different predictors for each $\boldsymbol{\beta}_{(j)}$

Recall $\boldsymbol{\beta}_{(j)} \in \mathbb{R}^{q_j}$. Let

$$\mathbf{X}_i = \text{block-diag}(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})_{p \times q+},$$

where $\mathbf{x}_{ij}$ is the set of q_j predictors from the i subject for response j and $q+ = \sum_{j=1}^{p} q_j$. Then

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} \sim N_{np} \left(\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \mathbf{B}^v, \mathcal{I}_n \otimes \boldsymbol{\Sigma} \right),$$

or simply

$$\mathbf{y} \sim N_{np}(\mathbf{X}\boldsymbol{\beta}, \mathcal{I}_n \otimes \boldsymbol{\Sigma}),$$

where $\boldsymbol{\beta} = \mathbf{B}^v$, in the form of a general linear model.

General linear model

In general, the different elements of $\mathbf{y}_i$ can share parameters in $\boldsymbol{\beta}$. The model is then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim N_{np}(\mathbf{0}, \underbrace{\mathcal{I}_n \otimes \boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}).$$

Here, $\mathbf{X}$ does not necessarily have any special structure. $\boldsymbol{\Sigma}$ can be unconstrained or have special structure, e.g. compound symmetry, AR(1), spatial correlation etc.

Fitting proceeds by first ridding ourselves of the fixed effects. One can show that maximizing $l(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\Sigma})$ over $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is equivalent to first maximizing $l(\mathbf{X}; \boldsymbol{\Sigma})$ (via Newton-Raphson), where

$$l(\mathbf{X}; \boldsymbol{\Sigma}) = -\tfrac{1}{2} \log |\boldsymbol{\Gamma}| - \tfrac{n}{2} \log(2\pi) - \tfrac{1}{2} \mathbf{y}' [\mathcal{I}_{np} - \mathbf{X}(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Gamma}^{-1}]' \boldsymbol{\Gamma}^{-1} [\mathcal{I}_{np} - \mathbf{X}(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Gamma}^{-1}]\mathbf{y}$$

yielding $\hat{\boldsymbol{\Sigma}}$, then $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{X})^{-1}\mathbf{X}\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{y}$ is the generalized least squares estimate (and MLE) of $\boldsymbol{\beta}$.

Wald tests in the general linear model

Say $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\mathbf{M} \in \mathbb{R}^{r \times q}$ where $r < q$. The linear hypothesis $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{0}$ is tested via

$$F^* = \tfrac{1}{r}[\mathbf{M}\hat{\boldsymbol{\beta}}]'[\mathbf{M}(\mathbf{X}'\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{X})^{-1}\mathbf{M}']^{-1}\mathbf{M}\hat{\boldsymbol{\beta}}.$$

Since $\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_q(\boldsymbol{\beta}, (\mathbf{X}'\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{X})^{-1})$, $[\mathbf{M}\hat{\boldsymbol{\beta}}]'[\mathbf{M}(\mathbf{X}'\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{X})^{-1}\mathbf{M}']^{-1}\mathbf{M}\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} \chi^2_r$ under H_0. `gls` and `lme` rather uses $F^* \overset{\bullet}{\sim} F_{r,\hat{\nu}}$ where $\hat{\nu}$ is an estimate of the denominator degrees of freedom; for `gls` this is $n - q$.

Aside: $\frac{1}{n-q}\chi^2_{n-q} \xrightarrow{P} 1$ by LLN, so the sampling distribution converges to the scaled χ^2_r anyway.

gls and lme parameterization

For these functions $\mathbf{\Sigma} = \mathbf{VRV}$ where $\mathbf{V}$ is diagonal and $\mathbf{R}$ is a correlation matrix. The default is $\mathbf{V} = \sigma\mathcal{I}_p$, but this can be relaxed to $\mathbf{V} = \text{diag}(\sigma_1, \ldots, \sigma_p)$ by adding

```
weights=varIdent(form=~1|time)
```

where `time` is $j \in \{1, \ldots, p\}$ in y_{ij}. The correlation structure $\mathbf{R}$ is specified by, e.g.

```
cor=corSymm(form=~time|Subject)
```

where `Subject` is $i \in \{1, \ldots, n\}$. A general unstructured $\mathbf{\Sigma}$ requires both of these.

Possible $\mathbf{R}$ structures include `corAR1` autoregressive process of order 1, `corARMA` autoregressive moving average process, `corCAR1` AR(1) process for a continuous time covariate, `corCompSymm` compound symmetry, `corExp` exponential spatial correlation, `corGaus` Gaussian spatial correlation, `corSymm` general correlation matrix, with no additional structure.

Some require additional covariate(s), e.g. latitude/longitude.

Unbalanced data

The most general version of the model allows $\mathbf{y}_i \in \mathbb{R}^{p_i}$ and assumes

$$\mathbf{y} \sim N_{p+}(\mathbf{X}\beta, \text{block-diag}(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_n).$$

For example $\mathbf{\Sigma}_i = \sigma_i^2[(1-\rho_i)\mathcal{I}_{p_i} + \rho_i \mathbf{1}_{p_i}\mathbf{1}'_{p_i}]$, compound symmetry within each i.

When subjects are seen irregularly and/or at differing time points, `gls`, `lme`, and SAS `proc mixed` still provide correct inference. An alternative to the general linear model with unbalanced data is a mixed model, discussed shortly.

A related problem is missing data. If data are MCAR (y_{ij} missing does not depend on y_{ij} or $\mathbf{x}_i$) or MAR (y_{ij} missing does not depend on y_{ij}), likelihood-based methods are superior to multiple imputation. Proceed as usual.

General model in R

The function `gls` requires i (`Subject`, below) and j (`time`).

```
library(nlme)
dental$time=dental$age/2-3 # consecutive integers for each element of y_i
f1=gls(distance~factor(Sex)*age,data=dental,cor=corSymm(form=~time|Subject),
 weights=varIdent(form=~1|time),method="ML") # unstructured correlation
f2=gls(distance~factor(Sex)*age,data=dental,cor=corSymm(form=~time|Subject),
 method="ML") # same variances across time points
anova(f1,f2) # are different variances necessary across time points? LRT test for nested models
```

The first `gls` function fits $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i$, where

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix} = \begin{bmatrix} 1 & g_i & 8 & 8g_i \\ 1 & g_i & 10 & 10g_i \\ 1 & g_i & 12 & 12g_i \\ 1 & g_i & 14 & 14g_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_{i1} \\ u_{i2} \\ u_{i3} \\ u_{i4} \end{bmatrix},$$

and $V(\mathbf{u}_i) = \boldsymbol{\Sigma}$. Here, $\boldsymbol{\Sigma} = \mathbf{VRV}$ where $\mathbf{R}$ is a correlation matrix and $\mathbf{V} = \sigma\text{diag}(1, p_2, p_3, p_4)$. The 2nd fits the model where $\mathbf{V} = \sigma\mathcal{I}_4$. Are the growth *rates* the same across gender?

Simple model as general model

We can rewrite $\mathbf{Y}$ in the simple model as $\mathbf{y}$ and fit the simple model using `gls`:

```
f3=gls(distance~factor(time)*factor(Sex),data=dental,cor=corSymm(form=~time|Subject),
 weights=varIdent(form=~1|time),method="ML") # equivalent to simple regression model
summary(f3) # gives order of regression effects to figure out contrast matrix
model.matrix(distance~factor(time)*factor(Sex),data=dental)
```

Fits the model (j is time)

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 I\{j=2\} + \beta_2 I\{j=3\} + \beta_3 I\{j=4\} + \beta_4 I\{g_i = F\} \\ &\quad + \beta_5 I\{j=2\} I\{g_i = F\} + \beta_6 I\{j=3\} I\{g_i = F\} + \beta_7 I\{j=4\} I\{g_i = F\} + u_{ij}, \end{aligned}$$

where $\mathbf{u}_i \overset{iid}{\sim} N_4(\mathbf{0}, \mathbf{\Sigma})$ for $i = 1, \ldots, 27$. In terms of the model $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$ we have

$$\begin{array}{ll} \mu_{m1} = \beta_0 & \mu_{f1} = \beta_0 + \beta_4 \\ \mu_{m1} = \beta_0 + \beta_1 & \mu_{f2} = \beta_0 + \beta_1 + \beta_4 + \beta_5 \\ \mu_{m2} = \beta_0 + \beta_2 & \mu_{f3} = \beta_0 + \beta_2 + \beta_4 + \beta_6 \\ \mu_{m4} = \beta_0 + \beta_3 & \mu_{f4} = \beta_0 + \beta_3 + \beta_4 + \beta_7 \end{array}.$$

P-values are a bit different than using simpler model, but conclusions are the same.

Simple model as general model

To test linearity in the dental data we need $\mathbf{M} = \begin{bmatrix} 0 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & 0 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & -1 & 2 & -1 \end{bmatrix}$. To test no difference between girls and boys $\mathbf{M} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$ R fills matrices by columns first – be careful!

```
M1=t(matrix(c( 0, 2,-1, 0, 0, 0, 0, 0,
               0,-1, 2,-1, 0, 0, 0, 0,
               0, 2,-1, 0, 0, 2,-1, 0,
               0,-1, 2,-1, 0,-1, 2,-1),8,4))
M2=t(matrix(c( 0, 0, 0, 0, 1, 0, 0, 0,
               0, 0, 0, 0, 1, 1, 0, 0,
               0, 0, 0, 0, 1, 0, 1, 0,
               0, 0, 0, 0, 1, 0, 0, 1),8,4))
anova(f3,L=M1) # linear? different p-value than Wilk's lambda or Roy's root
anova(f3,L=M2) # boys vs. girls, again different p-value
```

There are cleverer ways to do this, e.g. using `contrast`, but this "brute force" approach makes things transparent.

Mixed effects model

Each child's trajectory is approximately linear. We can postulate a separate regression model for each child instead. We'll fit a simpler model with gender as a simple additive effect rather than separate slopes.

$$y_{ij} = \theta_{0i} + \theta_{1i} a_j + \beta_2 g_i + u_{ij}, \quad u_{ij} \overset{iid}{\sim} N(0, \sigma^2),$$

for $j = 1, \ldots, 4$. We further assume that the intercepts and slopes come from a *population*

$$\begin{bmatrix} \theta_{i0} \\ \theta_{i1} \end{bmatrix} \overset{iid}{\sim} N_2 \left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \underbrace{\begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}}_{\boldsymbol{\Omega}} \right).$$

This is a mixed effects model, or mixed model for short. Note that $E(y_{ij}) = \beta_0 + \beta_1 a_j + \beta_2 g_i$, so the fixed effects are $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$. The variance components are $(\omega_{11}, \omega_{12}, \omega_{22}, \sigma^2)$.

Mixed effects model

The mixed effects model is also called the Laird-Ware model, after the groundbreaking 1982 *Biometrics* paper. It is common to reparameterize the model so the random effects are mean-zero

$$y_{ij} = \underbrace{\beta_0 + \gamma_{i0}}_{\boldsymbol{\theta}_{i0}} + \underbrace{\beta_1 a_j + \gamma_{1i} a_j}_{\boldsymbol{\theta}_{i1} a_j} + \beta_2 g_i + u_{ij},$$

where $\boldsymbol{\gamma}_i \overset{iid}{\sim} N_2(\mathbf{0}, \boldsymbol{\Omega})$ are indep. $\{u_{ij}\}$. Then

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix} = \begin{bmatrix} 1 & 8 & g_i \\ 1 & 10 & g_i \\ 1 & 12 & g_i \\ 1 & 14 & g_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{bmatrix} \begin{bmatrix} \gamma_{i0} \\ \gamma_{i1} \end{bmatrix} + \begin{bmatrix} u_{i1} \\ u_{i2} \\ u_{i3} \\ u_{i4} \end{bmatrix},$$

or succinctly $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{u}_i$. Note then that $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ and $V(\mathbf{y}_i) = \mathbf{Z}_i\boldsymbol{\Omega}\mathbf{Z}_i' + \sigma^2\mathcal{I}_4$. The model *induces* correlation among the elements of $\mathbf{y}_i$ through the random effects.

Laird-Ware model in R

```
f4=lme(distance~factor(Sex)+age,data=dental,random=~1+age|Subject,
 method="ML")
f5=lme(distance~factor(Sex)+age,data=dental,random=~1+age|Subject,
 weights=varExp(form=~age),method="ML")
anova(f5,f4) # test for homogeneity of variance akin to Breusch-Pagan
```

The first `lme` fits the model on the previous slide. The 2nd allows the variability of the u_{ij} to change with j as a function of age. Specifically, $V(u_{ij}) = \sigma^2 e^{\tau a_j}$ yielding $V(\mathbf{u}_i) = \sigma^2 \text{diag}(e^{\tau 8}, e^{\tau 10}, e^{\tau 12}, e^{\tau 14})$.

This type of model is also called a "random coefficient" model. `lme` allows for $V(\mathbf{u}_i) = \mathbf{\Sigma}$ as before (rather than $V(\mathbf{u}_i) = \sigma^2 \mathcal{I}_4$) in which case $V(\mathbf{y}_i) = \mathbf{Z}_i \mathbf{\Omega} \mathbf{Z}_i' + \mathbf{\Sigma}$. You need to be careful here; if $\mathbf{\Sigma}$ is unstructured the model is not identifiable.

AIC for model selection

The AIC (Akaike, 1974) is a widely accepted statistic for choosing among models, both mean and covariance models. The AIC is asymptotically justified as attempting to minimize the estimated Kullback-Liebler distance between the true probability model and several candidate models. As the true model is often more complex than our simple statistical models, the AIC will tend to pick larger, more complex models as more data are collected and more is known about the true data generating mechanism.

The AIC is

$$\text{AIC} = 2p - 2l(\mathbf{X}; \hat{\boldsymbol{\theta}}),$$

where p is the number of parameters in $\boldsymbol{\theta}$; for these models p includes both mean parameters and variance components (but *not* random effects).

AIC vs. BIC

The BIC (Schwarz, 1978) will pick the *correct* model as the sample size increases (it is consistent) *as long as the correct model is among those under consideration*. Since we do not know whether the true model is among those we are considering, I tend to use AIC and possibly err on the side of a more complex model, but one that better predicts the actual data that we saw.

The BIC is

$$\text{BIC} = p\log(n) - 2l(\mathbf{X};\hat{\theta}).$$

This penalizes for adding predictors moreso than AIC when $n \geq 8$, and so tends to give simpler models. Section 9.5 in Marden (2012) (on the course webpage) gives more details if you are interested.

One troublesome aspect of the BIC is the sample size n. It is unclear what to use for n when data are missing, censored, or where data are highly dependent.

Comments

- SAS `proc mixed` fits all these models; `repeated` statement specifies $\mathbf{\Sigma}$ & `random` statement specifies $\mathbf{Z}_i\gamma_i$.
- We can generalize to non-normal data, e.g. y_{ij} is Poisson or Bernoulli, using conditional generalized linear mixed models (GLMM) or else marginal models.
- Marginal models typically use GEE to allow $V(\mathbf{y}_i) = \mathbf{\Sigma}$ where $Y_{ij} \sim Pois(e^{\eta_{ij}})$ or $y_{ij} \sim Bern(\frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}})$ and $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta}$.
- GLMM assumes $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\gamma_i$.
- R has many packages to do both. `geepack` fits marginal models via GEE and `lme4` fits GLMMs including crossed and nested random effects.
- `nlme` (original R package for mixed models) also fits nonlinear models but not GLMM.
- Marginal models can be fit in SAS `proc genmod` and GLMM can be fit in SAS `proc nlmixed` and `proc glimmix`.

Some notes on correlation

Multiple correlation coefficient (pp. 168–169)

$R_{y\cdot\mathbf{x}} = \max_{\mathbf{a}\in\mathbb{R}^q} \text{corr}(\mathbf{y}, \mathbf{X}\mathbf{a})$, attained by $\mathbf{a} = \hat{\boldsymbol{\beta}} = \mathbf{P}_{\mathbf{X}}\mathbf{y}$. R^2 in univariate regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ is square of this. This idea is generalized to $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}$ on pp. 170–171.

Dependence analysis (pp. 175–176). Want to keep $k < q$ predictors, can use subset $\{i_1, \ldots, i_k\} \subset \{1, 2, \ldots, q\}$ that maximizes $R_{y\cdot i_1 \ldots i_k}$, i.e. find subset most correlated with $\mathbf{y}$.

Partial correlation in $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}$. On p. 169 your book defines

$$r_{ij\cdot\mathbf{x}} = \frac{[(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{y}_{(i)}]'[(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{y}_{(j)}]}{||(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{y}_{(i)}||\ ||(\mathcal{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{y}_{(j)}||} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}.$$

This is the *partial correlation* between measurements $\mathbf{y}_{(i)}$ and $\mathbf{y}_{(j)}$ after adjusting for the predictors in $\mathbf{X}$. It's simply the correlation of the residuals from each regression.

Interdependency analysis (pp. 179–180) is related to variance inflation factors.

Regression diagnostics

Consider the simple model we started with where each variable has the same predictors. Recall that each of the p $\hat{\boldsymbol{\beta}}_{(j)}$ has sampling distribution

$$\hat{\boldsymbol{\beta}}_{(j)} \sim N_q(\boldsymbol{\beta}_{(j)}, \sigma_{jj}[(\mathbf{X}'\mathbf{X})^{-1}]).$$

The estimate of σ_{jj} used in F-tests is $\frac{n}{n-q}\hat{\sigma}_{jj} = \frac{1}{n-q}SSE_j$ where SSE_j is the usual sum of squared errors from considering variable j only. Tests on $\boldsymbol{\beta}_{(j)}$ are carried out as if only the univariate regression was considered.

Similarly, testing the assumptions of linear mean and constant variance rely on univariate diagnostics (e.g. residual and influence plots) for each of the p regressions. Multivariate normality on the $\hat{\mathbf{u}}_i$ can be checked using, e.g. `mardiaTest`.

Last example, from Johnson & Wichern

```
# Col. 1: z1 = orders    Col. 2: z2 = add-delete items
# Col. 3: Y1 = CPU time  Col. 4: Y2 = disk input/output capacity

d=t(matrix(c(
 123.5, 2.108, 141.5, 301.8,
 146.1, 9.213, 168.9, 396.1,
 133.9, 1.905, 154.8, 328.2,
 128.5, 0.815, 146.5, 307.4,
 151.5, 1.061, 172.8, 362.4,
 136.2, 8.603, 160.1, 369.5,
  92.0, 1.125, 108.5, 229.1),4,7))
colnames(d)=c("orders","add-delete","CPU time","input/output")

library(car)
library(MVN)
f=lm(d[,3:4]~d[,1:2]); fm=Manova(f)
mardiaTest(f$residuals,qqplot=T) # multivariate normality
```

Example from Johnson & Wichern

```
# http://www.statmethods.net/stats/rdiagnostics.html
f1=lm(d[,3]~d[,1:2])
f2=lm(d[,4]~d[,1:2])
avPlots(f1) # added variable plots
avPlots(f2)
influencePlot(f1,id.method="identify")
influencePlot(f2,id.method="identify")
ncvTest(f1) # non-constant variance
ncvTest(f2)
vif(f1); vif(f2)  # etc...
# partial correlation for CPU & I/O adjusting for orders/add-delete
cov2cor(fm$SSPE)
# original unadjusted correlation
cor(d[,3],d[,4])
```