Многомерный дисперсионный анализ

Пусть вектор $Y_i = (y_{i1}, \ldots, y_{in})'$ соответствует n независимым измерениям i-й переменной, $i = 1, \ldots, p$. Для каждого Y_i можно рассмотреть одномерную линейную модель

$$Y_i = X\beta_i + e_i, \quad \operatorname{cov}(Y_i) = \sigma_{ii} I, \tag{1}$$

где X – матрица плана ранга $r \le m < n$, $\beta_i = (\beta_{i1}, \ldots, \beta_{im})'$ – вектор параметров, специфичных для каждой переменной. Через σ_{ij} обозначена ковариация между переменными Y_i и Y_j, $p \le n - r$, $r < n$. Взятые вместе p линейных моделей (1) составляют многомерную обобщенную линейную модель: $Y = X\beta + e$, где

$$Y = \begin{bmatrix} y_{11} & y_{21} & \ldots & y_{p1} \\ y_{12} & y_{22} & \ldots & y_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ y_{1n} & y_{2n} & \ldots & y_{pn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{11} & \beta_{21} & \ldots & \beta_{p1} \\ \beta_{12} & \beta_{22} & \ldots & \beta_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ \beta_{1m} & \beta_{2m} & \ldots & \beta_{pm} \end{bmatrix}$$

X – матрица плана размерности n на m ранга r, e – матрица n на p, строки которой составляют случайную выборку из $\mathcal{N}(0, \Sigma)$.
Оценкой параметров β_i, $i = 1, \ldots, p$, служит МНК-оценка по одной переменной Y_i вида

$$\hat{\beta}_i = (X'X)^{-}(X'Y_i), \quad i = 1, \ldots, p\,.$$

Многомерный дисперсионный анализ

Для вычисления несмещенных оценок для σ_{ii} и σ_{ij} используются остаточные суммы квадратов

$$R_0^2(i,i) = Y_i'Y_i - Y_i'X\hat{\beta}_i, \quad \hat{\sigma}_{ii} = \frac{R_0^2(i,i)}{n-r}$$

и остаточные суммы произведений

$$R_0^2(i,j) = Y_i'Y_j - Y_i'X\hat{\beta}_j, \quad \hat{\sigma}_{ij} = \frac{R_0^2(i,j)}{n-r}\,.$$

Матрица остаточных сумм и произведений имеет вид

$$\mathbb{R}_0 = \begin{bmatrix} R_0^2(1,1) & \dots & R_0^2(1,p) \\ \dots & \dots & \dots \\ R_0^2(p,1) & \dots & R_0^2(p,p) \end{bmatrix}$$

Требуется одновременно проверить гипотезы о том, что

$$H'\beta_i = \theta_i, \ i = 1, \dots, p\,, \tag{2}$$

где матрица H' размерности s на m имеет ранг $s \le r$. Оценки модели с ограничениями β_i^* и матрица остаточных сумм и произведений $\mathbb{R}_1$. Матрица $\mathbb{R}_1 - \mathbb{R}_1$ называется матрицей квадратов и произведений, обусловленных отклонением от гипотезы.

Многомерный дисперсионный анализ

Сведем задачу (2) к одномерному случаю, рассмотрев

$$\beta = \mathrm{L}_1\beta_1 + \ldots + \mathrm{L}_\mathrm{p}\beta_\mathrm{p}, \quad \xi = \mathrm{L}_1\xi_1 + \ldots + \mathrm{L}_\mathrm{p}\xi_\mathrm{p},$$

гипотезу $\mathrm{H}'\beta = \xi$, множество случайных величин $\mathrm{Y} = \mathrm{L}_1\mathrm{Y}_1 + \ldots + \mathrm{L}_\mathrm{p}\mathrm{Y}_\mathrm{p}$. Подходящий критерий основывается на двух минимальных суммах квадратов:

$$\mathrm{R}_0^2 = \min(\mathrm{Y} - \mathrm{X}\beta)'(\mathrm{Y} - \mathrm{X}\beta), \quad \mathrm{R}_1^2 = \min_{\mathrm{H}'\beta=\xi}(\mathrm{Y} - \mathrm{X}\beta)'(\mathrm{Y} - \mathrm{X}\beta).$$

При нулевой гипотезе $\mathrm{H}'\beta = \xi$ статистики $\mathrm{R}_1^2 - \mathrm{R}_0^2$ и R_0^2 распределены независимо как центральные хи-квадрат распределения $\sigma_\mathrm{L}^2\chi^2(\mathrm{s})$ и $\sigma_\mathrm{L}^2\chi^2(\mathrm{n}-\mathrm{r})$. Можно непосредственно убедиться в том, что

$$\mathrm{L}'\mathbb{R}_0\mathrm{L} = \mathrm{R}_0^2, \quad \mathrm{L}'\mathbb{R}_1\mathrm{L} = \mathrm{R}_1^2, \quad \mathrm{L}'(\mathbb{R}_1 - \mathbb{R}_0)\mathrm{L} = \mathrm{R}_1^2 - \mathrm{R}_0^2,$$

например,

$$\mathrm{R}_0^2 = (\mathrm{Y} - \mathrm{X}\beta)'(\mathrm{Y} - \mathrm{X}\beta) = \Big[\sum \mathrm{L}_\mathrm{i}(\mathrm{Y}_\mathrm{i} - \mathrm{X}\beta_\mathrm{i})\Big]' \Big[\sum \mathrm{L}_\mathrm{i}(\mathrm{Y}_\mathrm{i} - \mathrm{X}\beta_\mathrm{i})\Big] =$$

$$= \sum \mathrm{L}_\mathrm{i}\mathrm{L}_\mathrm{j}(\mathrm{Y}_\mathrm{i} - \mathrm{X}\beta_\mathrm{i})'(\mathrm{Y}_\mathrm{i} - \mathrm{X}\beta_\mathrm{i}) = \sum \mathrm{L}_\mathrm{i}\mathrm{L}_\mathrm{j}\mathrm{R}_0^2(\mathrm{i},\mathrm{j}) = \mathrm{L}'\mathbb{R}_0\mathrm{L}\,.$$

Следовательно, матрицы $\mathbb{R}_1 - \mathbb{R}_0$ и $\mathbb{R}_0$ имеют независимые распределения Уишарта

$$\mathbb{R}_0 \;\sim\; \mathrm{W}_\mathrm{p}(\mathrm{n}-\mathrm{r}), \quad \mathbb{R}_1 - \mathbb{R}_0 \;\sim\; \mathrm{W}_\mathrm{p}(\mathrm{s}).$$

Многомерный дисперсионный анализ

При справедливости нулевой гипотезы (2) для каждого L верна гипотеза $H'\beta = \xi$ и подходящими являются статистики

$$F = \frac{n - r}{s} \cdot \frac{R_1^2 - R_0^2}{R_0^2} \quad \text{или} \quad B = \frac{R_0^2}{R_1^2} = \frac{1}{1 + \frac{sF}{n-r}} = \frac{L^T \mathbb{R}_0 L}{L^T \mathbb{R}_1 L},$$

последняя имеет бета-распределение. Чем больше F и значимость, тем меньше B. Выберем L так, чтобы статистика B имела наименьшее значение (F наибольшее), тем самым пытаясь опровергнуть гипотезу. Тогда критерий сводится к виду:

$$\lambda = \min_L B = \min_L \frac{L^T \mathbb{R}_0 L}{L^T \mathbb{R}_1 L},$$

или λ является наименьшим корнем характеристического уравнения

$$|\mathbb{R}_0 - \lambda \mathbb{R}_1| = 0\,, \quad \text{так как} \quad \mathbb{R}_0 L = \lambda \mathbb{R}_1 L \iff L^T \mathbb{R}_0 L = \\ = \lambda L^T \mathbb{R}_1 L \iff \lambda = \frac{L^T \mathbb{R}_0 L}{L^T \mathbb{R}_1 L}\,.$$

Один из возможных критериев на собственных числах - Λ-критерий Уилкса:

$$\Lambda = \lambda_1 \cdot \ldots \cdot \lambda_p = \frac{|\mathbb{R}_0|}{|\mathbb{R}_1|}\,.$$

Пример из фтизиопульмонологии и код в R

Выборка – больные туберкулезом без ХОБЛ (хроническая обструкция легких)
Зависимые переменные

- ММП9 - протеолитический фермент, повышение которого свидетельствует о процессе, сязанном с расщеплением коллагена,
- индекс Генслера — чувствительный индекс к наличию ухудшения проходимости дыхательных путей.

Факторы

- курение
- диагноз: ИТЛ, туберкулема, ФКТ
- бронхит

Модель трехфакторного дисперсионного анализа с главными эффектами

ММП9	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sm	1	4.36	4.36	8.20	0.0050
fbs	1	2.06	2.06	3.87	0.0517
diag	2	3.98	1.99	3.74	0.0267
Residuals	115	61.24	0.53		

ИГ	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sm	1	9.89	9.89	18.44	0.0000
fbs	1	1.89	1.89	3.52	0.0632
diag	2	1.25	0.62	1.16	0.3159
Residuals	115	61.65	0.54		

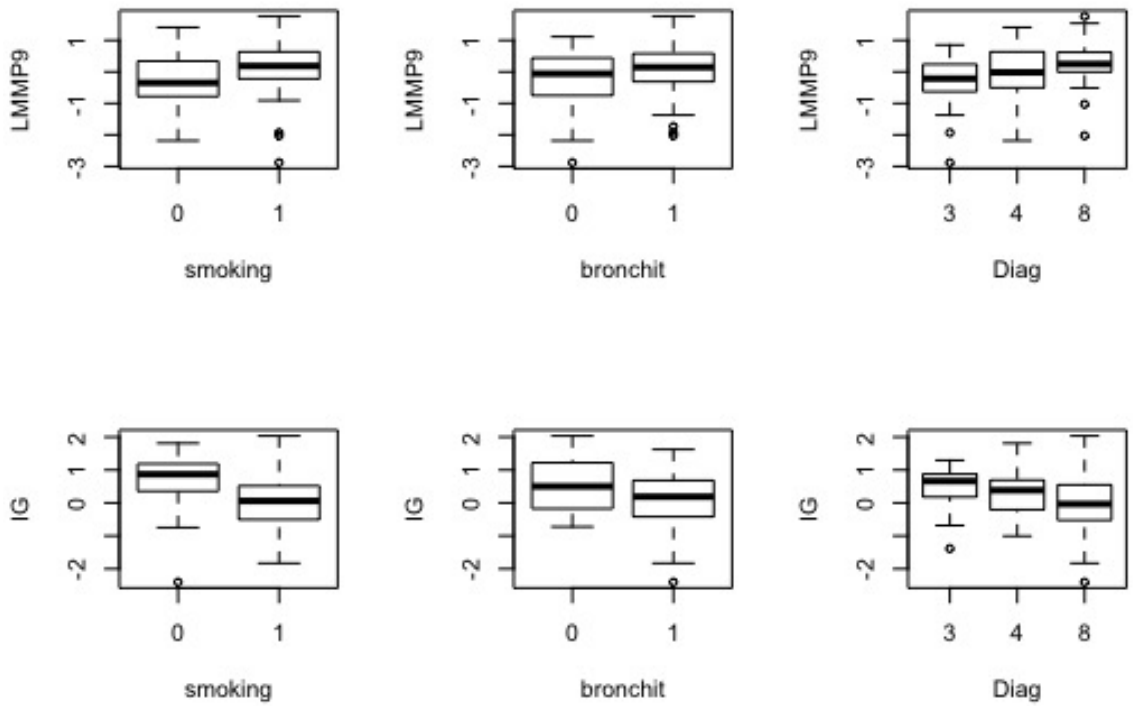

Figure: Зависимости ММП9 и ИГ от факторов курения, бронхита и клинической формы

Матрица частичного плана по матрице X с данными трех факторов

```
Matrix.Plan.N<-function(X)
{
a1<-rep(1,nrow(X))

tab1<-table(X$sm);tab1
a2<-ifelse(X$sm==1,1,-tab1[1]/tab1[2] )

tab2<-table(X$fbs);tab2

a3<-ifelse(X$fbs==1,1,-tab2[1]/tab2[2] )
tab3<-table(X$diag);tab3

a4<-ifelse(X$diag==3,1,ifelse(X$diag==4,0,-tab3["3"]/tab3["8"]))
a5<-ifelse(X$diag==3,0,ifelse(X$diag==4,1,-tab3["4"]/tab3["8"]))

cbind(a1,a2,a3,a4,a5) }
```

$$\mathbb{R}_0 = \begin{bmatrix} 61.238 & 1.240 \\ 1.240 & 61.654 \end{bmatrix} \quad \mathbb{R}_1 = \begin{bmatrix} 71.646 & -9.471 \\ -9.471 & 74.674 \end{bmatrix}$$

$$|\mathbb{R}_0 - \lambda\mathbb{R}_1| = 0 \iff |\mathbb{R}_0\mathbb{R}_1^{-1} - \lambda\mathbb{I}| = 0\,.$$

Матрица $\mathbb{R}_0\mathbb{R}_1^{-1}$

- Собственные числа 0.9854889 и 0.7280059
- Собственные вектора
  0.7447 0.6631
  0.6673 −0.7485

$$\mathrm{Z} = 0.6631 \cdot \mathrm{X}_1 - 0.7485 \cdot \mathrm{X}_2$$

Интерпретация - фактор тяжести, связанный с повышением протеолитической активности (распад коллагена) и снижением функции внешнего дыхания

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X.$sm	1	13.98	13.98	26.69	0.0000
X.$fbs	1	3.92	3.92	7.48	0.0072
X.$diag	2	4.61	2.30	4.40	0.0144
Residuals	115	60.24	0.52		

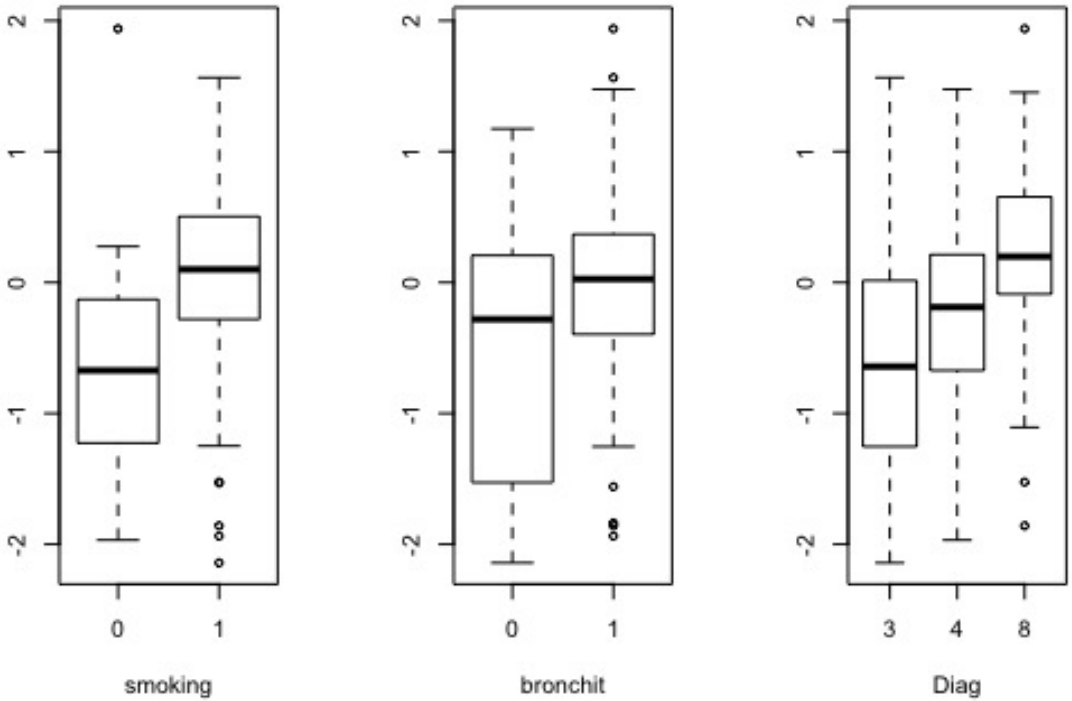

Figure: Зависимость фактора тяжести от факторов курения, клинической формы и бронхита