Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #13 (multicolinérité et interprétations)

- **Définition**: on parle de phénomène de colinéarité lorsque 2 variables explicatives (ou plus) sont fortement liées linéairement.
- **Conséquences**:
  - Les valeurs/signes des coefficients sont contradictoires, elles ne concordent pas avec les connaissances du domaine;
  - Les variances des estimateurs sont exagérées;
  - Les coefficients ne paraissent pas significatifs; Risque de passer à côté d'un régresseur important pour l'explication de y.

- Par l'hypothèse $\mathcal{H}_1$, $\mathbf{X}^\top\mathbf{X}$ est inversible, donc les colonnes de $\mathbf{X}$ sont linéairement indépendantes (il n'existe pas de relation linéaire parfaite entre les variables).
- Phénomène de colinéarité: quand il existe une relation presque linéaire entre les variables.
- $(\mathbf{X}^\top\mathbf{X})^{-1}$ invervient dans la définition de $\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$. Or, $(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{C}/\det(\mathbf{X}^\top\mathbf{X})$ (où $\mathbf{C}$ est la matrice des cofacteurs) et

  $$\text{colinéarité } \Rightarrow \det(\mathbf{X}^\top\mathbf{X}) \simeq 0 \Rightarrow (\mathbf{X}^\top\mathbf{X})^{-1} \text{ aura de grandes valeurs}$$

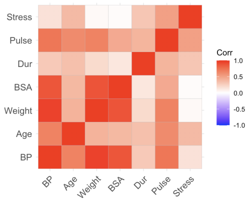
Colinéarité avec R

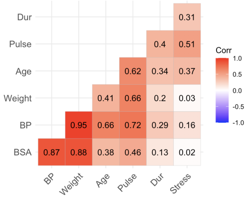
```
> loc = "https://online.stat.psu.edu/stat462/sites/
    onlinecourses.science.psu.edu.stat462/files/data/
    bloodpress/index.txt"
> base = read.table(loc,header=TRUE)
> base = base[,-1]
```

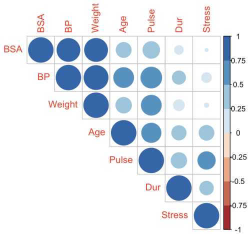
- blood pressure (x_1 = BP, in mm Hg)
- age (x_2 = Age, in years)
- weight (x_3 = Weight, in kg)
- body surface area (x_4 = BSA, in sq m)
- duration of hypertension (x_5 = Dur, in years)
- basal pulse (x_6 = Pulse, in beats per minute)
- stress index (y = Stress)

Colinéarité avec R

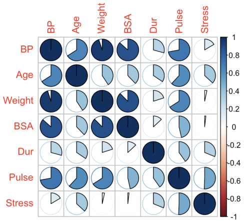
```
> r = cor(base)
> library(ggplot2)
> library(ggcorrplot)
> ggcorrplot(r)
> ggcorrplot(r, hc.order = TRUE, type = "lower", lab =
    TRUE)
```

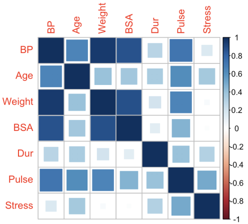


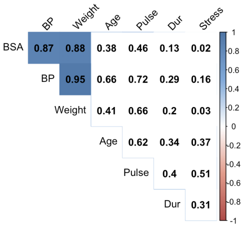




```
> library(corrplot)
> corrplot(r, method="circle")
> corrplot(r, method="pie")
> library(RColorBrewer)
> corrplot(r, type="upper", order="hclust", col=brewer
    .pal(n=8, name="RdBu"))
```







```
> eigen(r)
eigen() decomposition
$values
[1] 3.908 1.470 0.709 0.522 0.308 0.081 0.002
> solve(r)
              BP   Age Weight   BSA  Dur Pulse Stress
BP         259.8 -84.1 -199.3 -24.7 -7.0  15.4   -9.9
Age        -84.1  29.0   65.3   7.2  2.1  -6.1    3.2
Weight    -199.3  65.3  161.4  13.2  5.3 -16.0    9.3
BSA        -24.7   7.2   13.2   7.7  0.8   0.6    0.2
Dur         -7.0   2.1    5.3   0.8  1.4  -0.7    0.1
Pulse       15.4  -6.1  -16.0   0.6 -0.7   5.3   -2.2
Stress      -9.9   3.2    9.3   0.2  0.1  -2.2    2.2
```

Régression Linéaire

- Rappelons (si besoin) que $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$; $\sigma^2_{\hat{\beta}_j} = \text{Var}(\hat{\beta}_j)$.
- Colinéarité $\Rightarrow$ tendance à augmenter de façon abusive $\sigma^2_{\hat{\beta}_j} = \text{Var}(\hat{\beta}_j)$ et donc $\hat{\sigma}_{\hat{\beta}_j}$.
- Et puisque $T_j = \beta_j/\hat{\sigma}_{\hat{\beta}_j}$, $t_{j,obs}$ aura tendance à être davantage proche zéro, et donc on concluera peut-être à tort que $\mathbf{x}_j$ n'est pas significative.

- idée simple: mesurer le degré d'influence de chaque régresseur sur les $p-1$ restants;
- influence est mesurée par le R_j^2 coefficient de détermination multiple du modèle: $\mathbf{x}_j = \sum_{k=1,k\neq j}^{p} \eta_k \mathbf{x}_k + \boldsymbol{\varepsilon}_j$.

On appelle facteur d'inflation de la variance (VIF) la quantité:

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}, \quad j = 1,\ldots,p.$$

On parle d'inflation de la variance car

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{S_j^2}\,\mathrm{VIF}_j, \quad \text{où } S_j^2 = \sum_i (x_{ij}-\bar{x}_j)^2.$$

- Plus VIF_j sera élévé, plus $\mathrm{Var}(\hat{\beta}_j)$ sera élevée (tout comme son estimation); $\hat{\beta}_j$ sera beaucoup plus instable et il aura moins de chances d'être détecté significatif dans le test $H_0 : \beta_j = 0$.
- Si $\mathbf{x}_j \in \mathcal{V}(\mathbf{X}^{(j)})^{\perp}$ (c-a-d $\mathbf{x}_j$ linéairement indépendant des autres régresseurs), alors $\mathrm{VIF}_j = 1$.
- Plus $R_j^2 \to 1$ et plus $\mathrm{VIF}_j \to \infty$.
- Règle usuelle: problème de colinéarité si la dépendance entre régresseurs implique une augmentation de l'écart-type par 2, c-a-d si $\mathrm{VIF}_j \geq 4$. (certains auteurs préfèrent une valeur moins contraignant comme 10).

Soit $\mathbf{C}_{\mathbf{X}}$ la matrice de corrélation des régresseurs (sans $\mathbf{1}$ le cas échéant) alors

$$\mathrm{VIF}_j = ((\mathbf{C}_{\mathbf{X}})^{-1})_{jj}.$$

```
> model = lm(Stress~., data=base)
> summary(model)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 354.9763   309.8312   1.146  0.27257
BP           30.5437    18.7067   1.633  0.12649
Age         -21.2579    13.6854  -1.553  0.14434
Weight      -36.1523    17.8328  -2.027  0.06365 .
BSA         -27.8159   140.1533  -0.198  0.84575
Dur          -0.8911     3.8438  -0.232  0.82029
Pulse         9.7162     3.2171   3.020  0.00985 **

Residual standard error: 30.15 on 13 deg of freedom
Multiple R-squared:  0.5477
F-statistic: 2.624 on 6 and 13 DF,  p-value: 0.06833

> VIF(model)
     BP      Age  Weight     BSA     Dur   Pulse
215.553   24.474 122.596   7.647   1.421   3.128
```

Colinéarité & Indépendance (Frish-Waugh)

Under-identification is obtained when the true model is $y=\beta_0+\mathbf{x}_1^\top\boldsymbol{\beta}_1+\mathbf{x}_2^\top\boldsymbol{\beta}_2+\varepsilon$, but we estimate $y=b_0+\mathbf{x}_1^\top\mathbf{b}_1+\eta$.
Maximum likelihood estimator for $\mathbf{b}_1$ is

$$\begin{aligned}
\widehat{\mathbf{b}_1} &= (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{y}\\
&= (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top[\mathbf{X}_{1,i}\boldsymbol{\beta}_1+\mathbf{X}_{2,i}\boldsymbol{\beta}_2+\varepsilon]\\
&= \boldsymbol{\beta}_1+\underbrace{(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2\boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}}+\underbrace{(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\varepsilon}_{\nu_i}
\end{aligned}$$

so that $\mathbb{E}[\widehat{\mathbf{b}_1}]=\boldsymbol{\beta}_1+\boldsymbol{\beta}_{12}$,
and the bias is null when $\mathbf{X}_1^\top\mathbf{X}_2=\mathbf{0}$ i.e. $\mathbf{X}_1\perp\mathbf{X}_2$.

Colinéarité & Indépendance (Frish-Waugh)

Asumme that the true model is

$$\mathbf{y} = \beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \varepsilon$$

If $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$, $\widehat{\boldsymbol{\beta}}_2$ can be estimated using

$$(b_0, \widehat{\boldsymbol{\beta}}_2) = (\mathbf{X}_2'^\top \mathbf{X}_2')^{-1}\mathbf{X}_2'^\top \mathbf{y} \text{ where } \mathbf{X}_2' = [\mathbf{1}|\mathbf{X}_2]$$

Otherwise, let $\mathbf{y}_2^\star = \Pi_{\mathcal{X}_1^\perp}\mathbf{y}$ and $\mathbf{X}_2^\star = \Pi_{\mathcal{X}_1^\perp}\mathbf{X}_2$, then

$$\widehat{\boldsymbol{\beta}}_2 = [\mathbf{X}_2^{\star\top}\mathbf{X}_2^\star]^{-1}\mathbf{X}_2^{\star-1}\mathbf{y}_2^\star$$

$\mathbf{X}_2^\star = \mathbf{X}_2$ if $\mathbf{X}_1 \perp \mathbf{X}_2$,

Régression sur des variables orthogonales

```
> n=100
> set.seed(1)
> Z=princomp(cbind(rnorm(n),rnorm(n)))$scores
> cor(Z)
              Comp.1        Comp.2
Comp.1  1.000000e+00 -4.738026e-18
Comp.2 -4.738026e-18  1.000000e+00
```

We generated variables x_1 and x_2 such that $x_1 \perp x_2$,

```
> Y=1+Z[,1]-Z[,2]+rnorm(n)
> summary(lm(Y~Z[,1]+Z[,2]))

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)   1.0297     0.1043   9.868 2.58e-16 ***
Z[, 1]        1.0536     0.1095   9.624 8.69e-16 ***
Z[, 2]       -0.9793     0.1168  -8.388 4.00e-13 ***
```

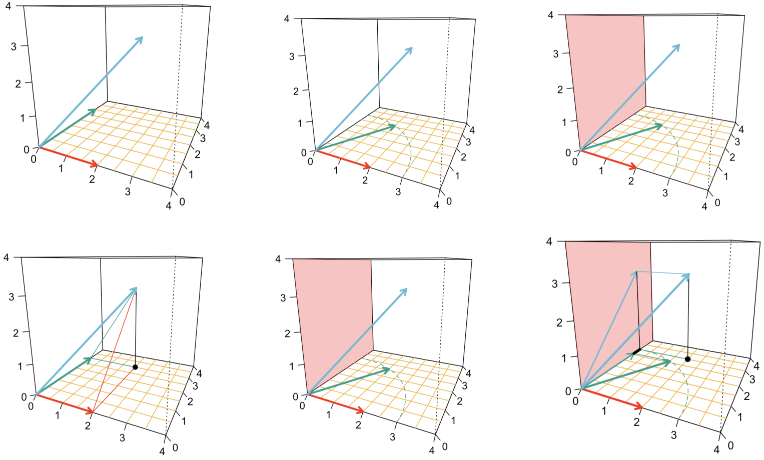
Régression sur des variables orthogonales

```
> summary(lm(Y~Z[,1]))

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)   1.0297     0.1364   7.552 2.27e-11 ***
Z[, 1]        1.0536     0.1431   7.365 5.60e-11 ***
```

```
> summary(lm(Y~Z[,2]))

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)   1.0297     0.1451   7.094 2.05e-10 ***
Z[, 2]       -0.9793     0.1624  -6.030 2.90e-08 ***
```

Note: here constants are equal since $\overline{x}_1 = \overline{x}_2 = 0$ (and $\widehat{\beta}_0 = \overline{y}$)

Colinéarité & Indépendance (Frish-Waugh)



cf https://freakonometrics.hypotheses.org/61177

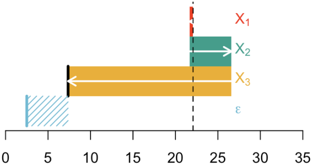
Colinéarité avec R

```
> chicago = read.table("http://freakonometrics.free.fr
    /chicago.txt", header=TRUE,sep=";")
> model = lm(Fire~.,data=chicago)
> summary(model)

Coefficients:
             Estimate Std. Error t value  Pr(|t|)
(Intercept) 22.07525    6.19447   3.564 0.000910 ***
X_1         -0.62764    5.28130  -0.119 0.905953
X_2          0.22378    0.06161   3.632 0.000744 ***
X_3         -1.55059    0.38195  -4.060 0.000204 ***
```

i.e. $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3$, soit ici

$$\widehat{y} = \underset{(6.2)}{22.06} - \underset{(5.3)}{0.63} x_1 + \underset{(0.06)}{0.22}\, x_2 - \underset{(0.38)}{1.55}\, x_3$$

```
> chicago[12,]
   Fire   X_1 X_2    X_3
12  2.2 0.402  14 13.722
```



Régressions en cascade

On peut faire des régressions en cascade, en commençant par x_1,

$$\widehat{y_i} = \underbrace{\widehat{b_0} + \widehat{b_1} x_{1,i}}_{(1)} + \underbrace{\widehat{b_{0,2}} + \widehat{b_2} \tilde{x}_{2,i}}_{(2)} + \underbrace{\widehat{b_{0,3}} + \widehat{b_3} \tilde{x}_{3,i}}_{(3)}$$

où

- (1) est obtenu en régression simplement y sur x_1 $y_i = b_0 + b_1 x_{1,i} + \eta_i$, puis on apporte une correction, pour tenir compte de ce que nous apprend x_2 une fois la première régression effectuée

```
> reg1=lm(Fire~X_1,data=chicago)
> summary(reg1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.448      3.272   1.054  0.29751
X_1            14.028      5.132   2.733  0.00893 **
```

Régressions en cascade

- (2) on va projeter y sur $x_1^\perp$ (ce que x_1 n'explique pas), et x_2 sur $x_1^\perp$ et faire la régression de ces deux projections

$$\Pi_{x_1^\perp} y_i = b_{0,2} + b_2 \Pi_{x_1^\perp} x_{2,i} + \eta_{2,i}$$

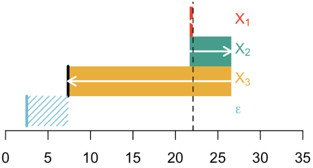
```
> reg2=lm(residuals(lm(Fire~X_1,data=chicago))~
    residuals(lm(X_2~X_1,data=chicago)))
> summary(reg2)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.378e-16  1.095e+00   0.000   1.0000
residuals()   1.511e-01  6.778e-02   2.229   0.0308 *
```

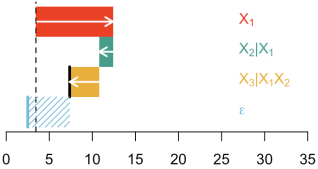
Régressions en cascade

- (3) on va projeter y et x_3 sur $\mathcal{V}(x_1, x_2)^\perp$

$$\Pi_{(x_1,x_2)^\perp} y_i = b_{0,3} + b_3 \Pi_{(x_1,x_2)^\perp} x_{3,i} + \eta_{2,i}$$

```
> reg3=lm(residuals(lm(Fire~X_1+X_2,data=chicago))~
   residuals(lm(X_3~X_1+X_2,data=chicago)))
> summary(reg3)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.124e-16  9.306e-01   0.000 1.000000
residuals() -1.551e+00  3.734e-01  -4.153 0.000144 ***
```





Régressions en cascade

ou en commençant par x_2 puis x_1 puis x_3

$$\widehat{y}_i = \underbrace{\widehat{b}_0 + \widehat{b}_2 x_{2,i}}_{(2)} + \underbrace{\widehat{b}_{0,1} + \widehat{b}_1 \tilde{x}_{1,i}}_{(1)} + \underbrace{\widehat{b}_{0,3} + \widehat{b}_3 \tilde{x}_{3,i}}_{(3)}$$