Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #5 (régression sur une variable factorielle)

Régression sur une variable x factorielle

On suppose disposer d'observations individuelles $y_{i,j}$ appartenant à un groupe $j \in \{1, 2, \cdots, k\}$. On suppose que

$$y_{i,j} = \alpha_j + \varepsilon_{i,j} = \beta_0 + \beta_j + \varepsilon_{i,j}$$

où (classiquement) les $\varepsilon_{i,j}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

```
> with(Davis, t.test(height~sex))

t = -15.28, df = 174.29, p-value < 2.2e-16
alternative hypothesis: true difference in means is
    not equal to 0
95 percent confidence interval:
 -15.01467 -11.57949
sample estimates:
mean in group F mean in group M
       164.7143        178.0114
```

Régression sur une variable x factorielle

Ce modèle n'est pas identifiable

$$y_{i,j} = \beta_0 + \beta_j + \varepsilon_{i,j},\ \forall j = 1, \cdots, k$$

on a trois options possibles :

1. Supprimer la constante

$$y_{i,j} = \alpha_j + \varepsilon_{i,j}$$

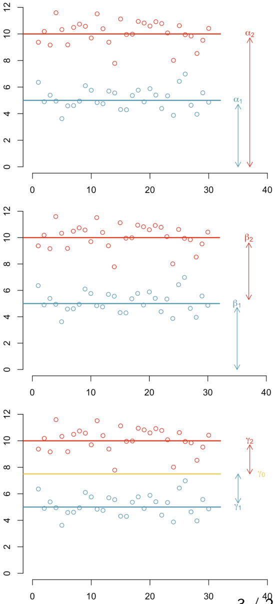
2. Supprimer une modalité, e.g. la première,

$$y_{i,j} = \beta_0 + \beta_j + \varepsilon_{i,j},\ \text{and } \beta_1 = 0$$

3. Rajouter une contrainte (linéaire)

on parlera de contrastes.

$$y_{i,j} = \gamma_0 + \gamma_j + \varepsilon_{i,j},\ \text{and } \sum_{j=1}^{k} \gamma_j = 0$$



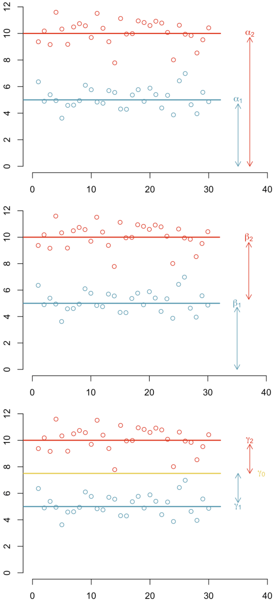
Régression sur une variable x factorielle

Les estimateurs du maximum de vraisemblance du modèle sont

$$\widehat{\alpha}_j = \frac{1}{n_j}\sum_{i\in\mathcal{I}_j} y_{i,j} = \overline{y}_{\cdot,j}$$

$$\widehat{\beta}_0 = \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} y_{i,j} = \overline{y}_{\cdot,1} \text{ et } \widehat{\beta}_j = \overline{y}_{\cdot,j} - \overline{y}_{\cdot,1}$$

$$\widehat{\gamma}_0 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j} y_{i,j} \text{ et } \widehat{\gamma}_j = \overline{y}_{\cdot,j} - \overline{y}$$



Régression sur une variable x factorielle

```
> with(Davis,mean(height[sex=="F"]))
[1] 164.7143
> with(Davis,mean(height[sex=="M"]))
[1] 178.0114

> summary(lm(height~0+sex,data=Davis))
Coefficients:
     Estimate Std. Error t value Pr(>t)
sexF 164.7143     0.5684   289.8   <2e-16 ***
sexM 178.0114     0.6412   277.6   <2e-16 ***

> summary(lm(height~1+sex,data=Davis))
Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) 164.7143     0.5684  289.80   <2e-16 ***
sexM         13.2971     0.8569   15.52   <2e-16 ***
```

Régression sur une variable x factorielle

In the second casde, we considered $y_i = \beta_0 + \beta_1 \mathbf{1}(x_i = \text{'M'})$, and

$$\widehat{\beta}_0 = \frac{1}{n_F} \sum_{i=1}^{n} y_i \mathbf{1}(x_i = \text{F})$$

while

$$\widehat{\beta}_1 = \widehat{\beta}_0 = \frac{1}{n_M} \sum_{i=1}^{n} y_i \mathbf{1}(x_i = \text{M}) - \widehat{\beta}_0$$

Régression sur une variable x factorielle

On va alors chercher à tester

$$\begin{cases} H_0 : \alpha_1 = \cdots = \alpha_k \\ H_0 : \beta_2 = \cdots = \beta_k = 0 \\ H_0 : \gamma_1 = \cdots = \gamma_k = 0 \end{cases}$$

Techniquement, on a ici un test multiple.
Notons pour commencer que

$$SCR_{\text{total}} = SCR_{\text{facteur}} + SCR_{\text{résidus}}$$

où

$$\underbrace{\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}(y_{i,j}-\overline{y})^2}_{SCR_{\text{total}}} = \underbrace{\sum_{j=1}^{k}(\overline{y}_{\cdot,j}-\overline{y})^2}_{SCR_{\text{facteur}}} + \underbrace{\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}(y_{i,j}-\overline{y}_{\cdot,j})^2}_{SCR_{\text{résidus}}}$$

Régression sur une variable x factorielle

Sous une hypothèse de normalité des résidus $\varepsilon_{i,j}$, on peut montrer que si H_0 est vraie

$$SCR_{\text{facteur}} = \sum_{j=1}^{k} n_j(\overline{y_{\cdot,j}} - \overline{y})^2 \sim \chi^2(DL_{\text{facteur}}) \text{ avec } DL_{\text{facteur}} = k - 1$$

$$SCR_{\text{résidus}} = \sum_{j=1}^{k} \sum_{i \in \mathcal{I}_j} (y_{i,j} - \overline{y_{\cdot,j}})^2 \sim \chi^2(DL_{\text{résidus}})$$

avec

$$DL_{\text{résidus}} = \sum_{j=1}^{k} (n_j - 1) = (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) = n - k$$

Alors

$$F = \frac{\dfrac{SCR_{\text{facteur}}}{k-1}}{\dfrac{SCR_{\text{résidus}}}{n-k}} \sim \mathcal{F}(k-1, n-k)$$

Régression sur une variable x factorielle

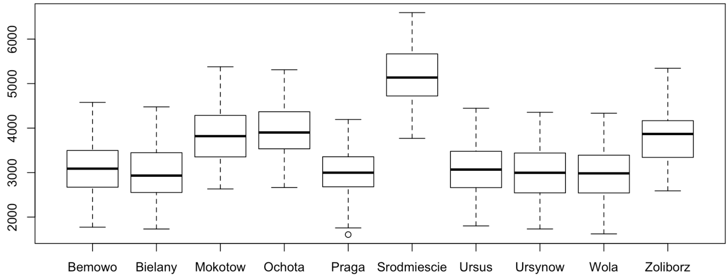
```
> summary(aov(height ~ sex, data = Davis))
             Df Sum Sq Mean Sq F value Pr(>F)
sex           1   8713    8713   240.8 <2e-16 ***
Residuals   198   7164      36
> anova(lm(height~sex,data=Davis))
Analysis of Variance Table

Response: height
           Df Sum Sq Mean Sq F value    Pr(>F)
sex         1 8713.3  8713.3  240.83 < 2.2e-16 ***
Residuals 198 7163.8    36.2
```

$$F = \frac{\dfrac{8713.3}{1}}{\dfrac{7163.8}{198}} = 240.83 \sim \mathcal{F}(1,198)$$

```
> qf(.95,1,198)
[1] 3.888853
```

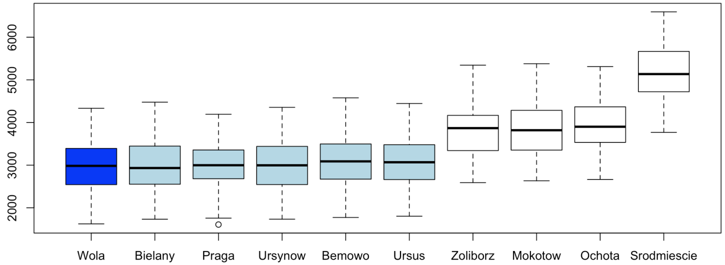
Variable x factorielle $\{A, B, \cdots, J\}$

```
> A = with(data = apartments, aggregate(m2.price,by=
    list(district),FUN=mean))
> A = A[order(A$x),]
> L = as.character(A$Group.1)
> apartments$district = factor(apartments$district,
    level=L)
> with(data=apartments, boxplot(m2.price ~ district))
```



Variable x factorielle $\{A, B, \cdots, J\}$

```
> summary(lm(m2.price ~ district, data=apartments))
Coefficients:
                    Estimate Std. Error t value Pr(>t)
(Intercept)          2968.36    58.02  51.160   <2e-16 ***
districtBielany        17.38    84.16   0.207    0.836
districtPraga          26.45    85.12   0.311    0.756
districtUrsynow        42.01    82.65   0.508    0.611
districtBemowo         80.10    83.71   0.957    0.339
districtUrsus         102.01    82.25   1.240    0.215
districtZoliborz.     829.59    83.94   9.884   <2e-16 ***
```



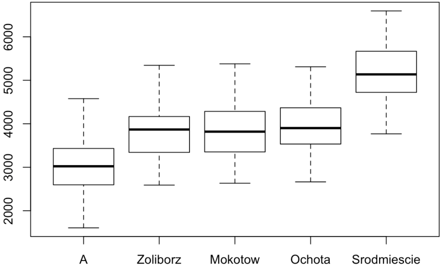
Variable x factorielle $\{A, B, \cdots, J\}$ (1)

Fisher (multiple test): $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

```
> library(car)
> linearHypothesis(reg, c("districtBielany = 0", "
    districtPraga = 0", "districtUrsynow = 0", "
    districtBemowo = 0","districtUrsus = 0"))

Model 1: restricted model
Model 2: m2.price ~ district

  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    995 354051715
2    990 353269202  5    782513 0.4386 0.8217
```

Variable x factorielle $\{A, B, \cdots, J\}$ (1)

```
> levels(apartments$district) = c(rep("A",6),levels(
    apartments$district)[7:11])
> with(data=apartments, boxplot(m2.price ~ district))
> apartments$district = relevel(apartments$district, "
    Zoliborz")
```

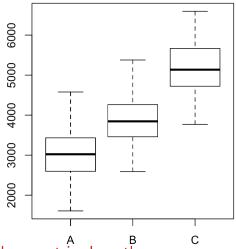


Variable x factorielle $\{A, B, \cdots, J\}$ (1)

```
> reg = lm(m2.price ~ district, data=apartments)
> summary(reg)
Coefficients:
                    Estimate Std. Error t value Pr(>t)
(Intercept)          3797.95    60.57   62.707 <2e-16 ***
districtA            -784.61    65.28  -12.019 <2e-16 ***
districtMokotow        57.51    83.63    0.688 0.4918
districtOchota        158.34    85.88    1.844 0.0655 .
districtSrodmiescie  1384.80    85.01   16.290 <2e-16 ***
```

Variable x factorielle $\{A, B, \cdots, J\}$ (1)

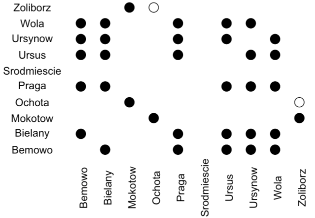
```
> linearHypothesis(reg, c("districtMokotow = 0", "districtOchota = 0"))

Model 1: restricted model
Model 2: m2.price ~ district

  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    997 355292524
2    995 354051715  2   1240809 1.7435 0.1754
> levels(apartments$district) = c("B","A","B","B","C")
> apartments$district = relevel(apartments$distr,"A")
```

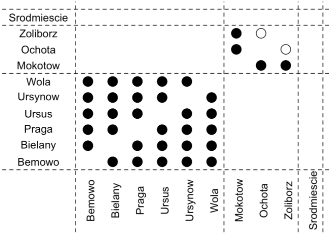


Variable x factorielle $\{A, B, \cdots, J\}$ (2)

Let us try all possible reference categories,

```
> LvL=levels(apartments$district)
> P=matrix(NA,10,10)
> for(i in 1:10){
  apartments$district=relevel(apartments$district,Lvl[i])
  p=summary(lm(m2.price~district,data=apartments))$coefficients[-1,4]
  names(p)=substr(names(p),9,nchar(names(p)))
  P[LvL[i],names(p)]=p
}
```





ANOVA - Analysis of Variance

Consider k groups, of size $n_1, \cdots, n_k$ respectively.

Sample mean in group j: $\overline{y}_{\cdot j} = \dfrac{1}{n_j}\displaystyle\sum_{i=1}^{n_j} y_{i,j}$

Sample standard deviation in group j: $s^2_{\cdot j} = \dfrac{1}{n_j - 1}\displaystyle\sum_{i=1}^{n}(y_{i,j} - \overline{y}_{\cdot j})^2$

grand mean is $\overline{y} = \dfrac{1}{n}\displaystyle\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{i,j} = \sum_{j=1}^{k}\left(\frac{n_j}{n}\right)\overline{y}_{\cdot j}$

(weighted average of the sample means, weighted by sample size)

Modeling assumptions:

$\mathbb{E}[Y_{i,j}] = \mu_j$ and $\text{Var}[Y_{i,j}] = \sigma^2$ (identical), $Y_{i,j}$ are independent.

Set $\mu = \displaystyle\sum_{j=1}^{k}\left(\frac{n_j}{n}\right)\mu_j$

ANOVA - Analysis of Variance

Then $\mathbb{E}\big(\overline{Y}_j\big) = \mu_j$ and $\text{Var}\big(\overline{Y}_j\big) = \dfrac{\sigma^2}{n_j}$

If we assume $Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$, $\overline{Y}_{\cdot j} \sim \mathcal{N}(\mu_j, \dfrac{\sigma^2}{n})$

Let $GSS = \displaystyle\sum_{j=1}^{k} n_j\big(\overline{y}_j - \overline{y}\big)^2$

$$GSS = (k-1)\sigma^2 + \sum_{j=1}^{k} n_j\big(\mu_j - \mu\big)^2 =^{*} (k-1)\sigma^2$$

$*$ if $\mu_1 = \cdots = \mu_k = \mu$.
GSS is the group sum of squares., related to variation among samples

ANOVA - Analysis of Variance

Similarly, observe that

$$\mathbb{E}\left(\sum_{j=1}^{n}(n_j - 1)s_j^2\right) = (n-k)\sigma^2$$

Define group mean square

$$MS_{group} = \frac{1}{k-1}\sum_{j=1}^{k} n_j\left(\overline{y}_j - \overline{y}\right)^2$$

and the mean square error

$$MS_{error} = \frac{1}{n-k}\sum_{j=1}^{n}(n_j - 1)s_j^2$$

ANOVA - Analysis of Variance

Assume $H_0 : \mu_1 = \cdots = \mu_k = \mu$ and $Y_{ij} \sim \mathcal{N}(\mu, \sigma^2)$
Under H_0, $MS_{group} \perp\!\!\!\perp MS_{error}$ and

$$\frac{(k-1)}{\sigma^2} MS_{group} \sim \chi^2(k-1)$$

$$\frac{(n-k)}{\sigma^2} MS_{error} \sim \chi^2(k-k)$$

so it follows that

$$F = \frac{MS_{group}}{MS_{error}} \sim \mathcal{F}(k-1, n-k)$$

source	df	sum of squares	mean	F	p-value
groups	$k-1$	SS_{group}	MS_{group}	F	p
error	$n-k$	SS_{error}	MS_{error}		
total	$n-1$	SS_{total}			

ANOVA - Analysis of Variance

The proportion of variability explained by the groups is

$$R^2 = \frac{SS_{group}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$$

ANOVA - Analysis of Variance

Recall that $\text{Var}[\overline{y}_j] = \dfrac{\sigma^2}{n_j}$ where $\hat{\sigma}^2 = MS_{error}$. One can derive 95% confidence interval for μ_j

$$\mu_j \in \left[\overline{y}_j \pm 1.96\frac{\hat{\sigma}^2}{n_j}\right]$$

and since $\text{Var}[\overline{y}_{j_1} - \overline{y}_{j_2}] = \sigma^2\left(\dfrac{1}{n_{j_1}} + \dfrac{1}{n_{j_2}}\right)$

$$\mu_{j_1} - \mu_{j_2} \in \left[\overline{y}_{j_1} - \overline{y}_{j_2} \pm 1.96\hat{\sigma}\sqrt{\frac{1}{n_{j_1}} + \frac{1}{n_{j_2}}}\right]$$

but that's not simultaneous confidence intervals

ANOVA - Analysis of Variance

```
                        diff       lwr      upr
Bielany-Bemowo        -62.72   -317.91   192.47
Mokotow-Bemowo        807.00    558.52  1055.48
Ochota-Bemowo         907.83    652.64  1163.02
Praga-Bemowo          -53.65   -311.63   204.32
Srodmiescie-Bemowo   2134.29   1881.69  2386.89
Ursus-Bemowo           21.91   -227.69   271.52
Ursynow-Bemowo        -38.09   -288.86   212.68
Wola-Bemowo           -80.10   -329.14   168.94
Zoliborz-Bemowo       869.72    619.89  1119.54
Mokotow-Bielany       970.55    714.05  1227.06
Ochota-Bielany          9.06   -250.21   268.34
Praga-Bielany        2197.01   1943.08  2450.94
Srodmiescie-Bielany    84.63   -166.32   335.58
Ursus-Bielany          24.63   -227.48   276.74
Ursynow-Bielany       -17.38   -267.76   233.00
Wola-Bielany          100.83   -148.99   350.66
Zoliborz-Bielany     -860.65  -1113.33  -607.98
```

ANOVA - Analysis of Variance

See Tukey's honestly significant difference (HSD)

If we want to be 95% confident that **all** population mean differences are contained in their intervals, we need to increase the size of the multipler.

```
> TukeyHSD(aov(formula = m2.price ~ 0 + district, data
    = apartments))
  Tukey multiple comparisons of means
    95% family-wise confidence level

                         diff      lwr      upr p adj
Bielany-Bemowo         -62.72  -334.72   209.29  1.00
Mokotow-Bemowo         807.00   542.15  1071.85  0.00
Ochota-Bemowo          907.83   635.83  1179.84  0.00
Praga-Bemowo           -53.65  -328.63   221.32  1.00
Srodmiescie-Bemowo    2134.29  1865.05  2403.53  0.00
Ursus-Bemowo            21.91  -244.14   287.96  1.00
Ursynow-Bemowo         -38.09  -305.39   229.21  1.00
Wola-Bemowo            -80.10  -345.55   185.34  0.99
Zoliborz-Bemowo        749.49   478.19  1020.79  0.00
```