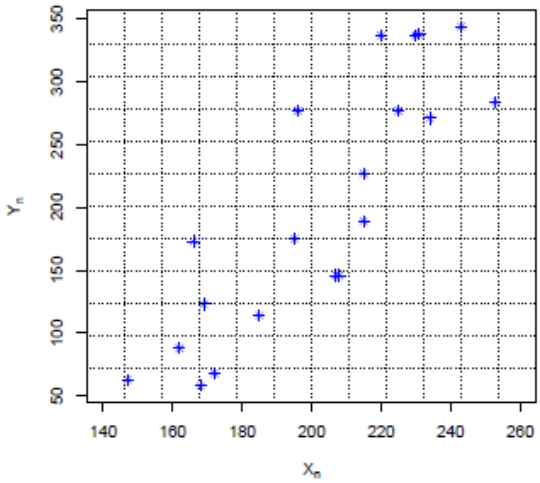
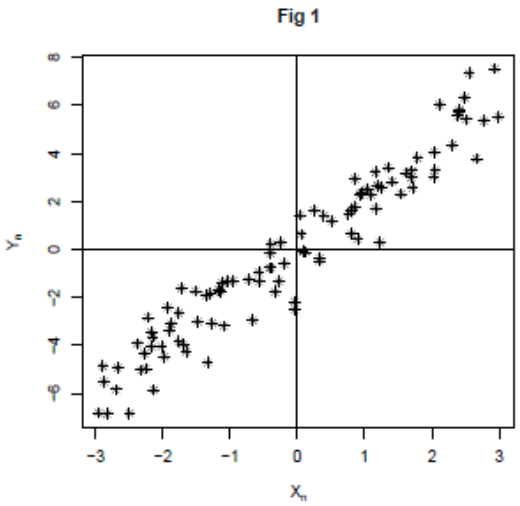
Modélisation Statistique

Régression Linéaire Simple

1- MONTRER QU'ON A UN MODELE LINEAIRE

- Graphique





Les points sont alignés

- Formules et hypothèses

$\boldsymbol{y_i = \beta_0 + \beta_1 + \varepsilon_i}$ ε_i terme d'erreur

Hypothèses :

- x_i observées non aléatoires
- y_i observées aléatoires
- ε_i non observé aléatoires
- $\mathbb{E}(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \sigma^2$ $\forall i \in \{1, \dots, n\}$ (homosédacité)
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ $i \neq j$
- Hypothèse supplémentaire ε_i $\mathcal{N}(0, \sigma^2)$ $\longrightarrow$ indépendance de ε_i (nécessaire pour les tests et intervalle de confiance)

2- DONNER/TRACER LA DROITE DE REGRESSION

$$\widehat{\beta_1} = \frac{Cov(x,y)}{s^2{}_x} = \frac{\sum_{i=1}^{n}(x_i - \overline{x_n})(y_i - \overline{y_n})}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x_n}\,\overline{y_n}}{\frac{1}{n}\sum_{i=1}^{n} x_i{}^2 - \overline{x_n}^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x_n}\,\overline{y_n}}{\sum_{i=1}^{n} x_i{}^2 - n\,\overline{x_n}^2}$$

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})(y_i - \overline{y_n}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x_n y_n}$$

$$s^2{}_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2 - \overline{x_n}^2 \qquad (Koening)$$

On peut faire cette opération s'il n'y a pas beaucoup de valeur, sinon, les sommes sont données

$$\widehat{\beta_0} = \overline{y_n} - \widehat{\beta_1}\,\overline{x_n}$$

3- ESTIMATION DE L'ERREUR DE LA COURBE

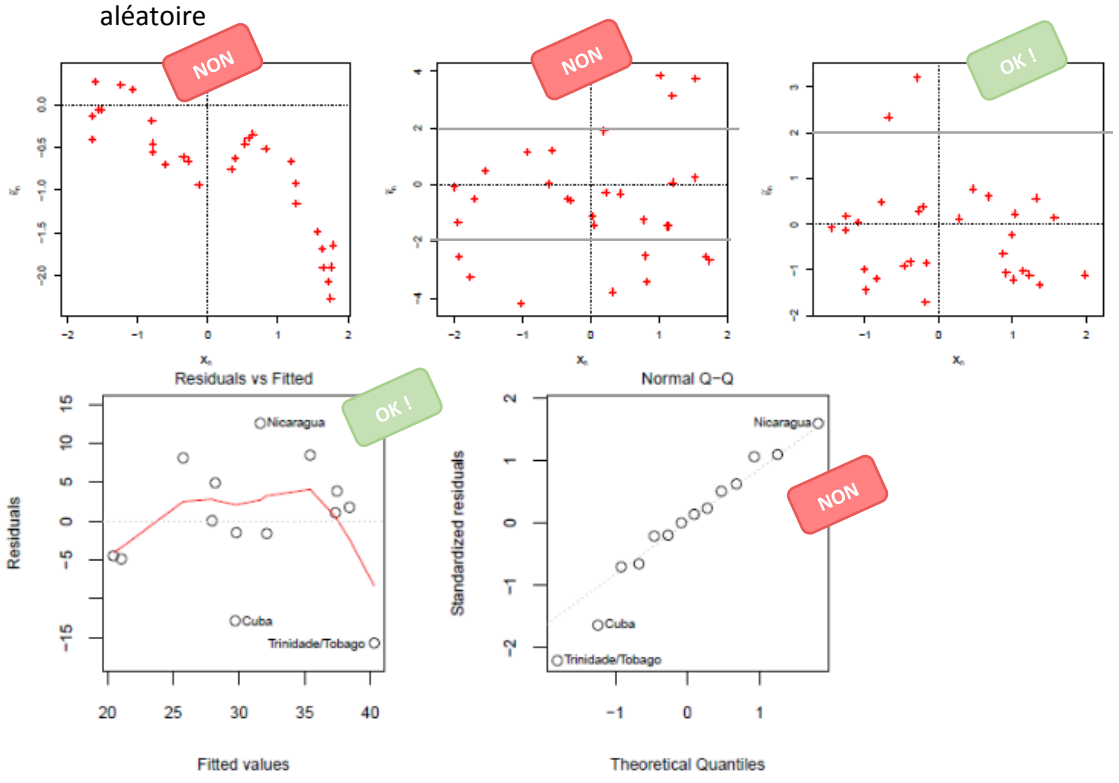
- Formules

$$\widehat{\varepsilon_i} = y_i - \widehat{y_i} \qquad \text{où} \quad \widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} + \varepsilon_i$$

- Graphique

On modélise graphiquement les erreurs :

- Les points doivent être centrés autour de 0 (faire attention au repère)
- Les points ne doivent pas avoir de forme significative, ils doivent etre répartis de manière aléatoire



4- SOMME DES CARRES DES RESIDUS

$$\sum_{i=1}^{n} \widehat{\varepsilon_i}^2 = SCR$$

Souvent donné car ça demande beaucoup de calculs

5- VARIANCE DES RESIDUS

$$\sigma^2 = s^2 = \frac{SCR}{n-2}$$

6- COEFFICIENT DE DETERMINATION R^2

$$\boldsymbol{R^2 = \frac{SCE}{SCT}} = \frac{SCT - SCR}{SCT} = 1 - \frac{SCR}{SCT} \qquad (SCT = SCE + SCR)$$

$SCT = \sum_{i=1}^{n} (y_i - \overline{y_n})^2$ données issues du graphique ou tableau

$SCE = \sum_{i=1}^{n} (\widehat{y_i} - \overline{y_n})^2$ données issues de l'estimation affine

$SCR = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$ représentation de l'erreur entre les données du tableau et celles estimées

Le résultat doit etre compris entre 0 et 1. C'est une bonne estimation quand il est proche de 1.
Quand R^2est compris entre 0,2 et 0,9 alors on a une dépendance linéaire partielle.

7- NULLITE DE LA PENTE

On a les hypothèses suivantes :

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

On utilise la loi de Student :

$$\boldsymbol{T_n = \frac{\widehat{\beta_1} - \beta_1}{\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}}}} \sim St(n-2)$$

Avec H_0 on a $\beta_1 = 0$ et donc :

On cherche son IC/IR

$$T_n = \frac{\widehat{\beta_1} - \beta_1}{\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}}} = \frac{\widehat{\beta_1} - 0}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}}} = \frac{\widehat{\beta_1}}{\frac{\sqrt{s^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}}} = \widehat{\boldsymbol{\beta_1}} \sqrt{\frac{\boldsymbol{s^2}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{x_i} - \overline{\boldsymbol{x_n}})^2}}$$

On cherche l'intervalle de confiance/de rejet :

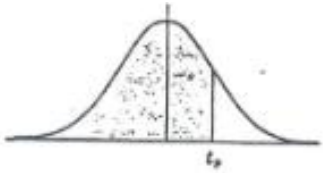
$$IR = \left[\widehat{\beta_1} - t_{n-2,1-\alpha} * \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}} \quad ; \quad \widehat{\beta_1} + t_{n-2,1-\alpha} * \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}}\right]$$

Rappel :

Test à … %
$1 - \alpha = 0, \ldots$
$\Leftrightarrow \alpha =$ le reste
On prend $\frac{\alpha}{2}$ car intervalle centré

Test à 95% → $\alpha = 0{,}05$
$\frac{\alpha}{2} = 0{,}025$
$1 - \frac{\alpha}{2} = 0{,}975$

On regarde le $\left(1 - \frac{\alpha}{2}\right)$



PERCENTILE VALUES (t_p)
for
STUDENT'S t DISTRIBUTION
with ν degrees of freedom
(shaded area = p)

On regarde le $(n-2)$

ν	$t_{.995}$	$t_{.99}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.55}$
1	63.66	31.82	12.71	6.31	3.08	1.376	1.000	.727	.325	.158
2	9.92	6.96	4.30	2.92	1.89	1.061	.816	.617	.289	.142
3	5.84	4.54	3.18	2.35	1.64	.978	.765	.584	.277	.137
4	4.60	3.75	2.78	2.13	1.53	.941	.741	.569	.271	.134
5	4.03	3.36	2.57	2.02	1.48	.920	.727	.559	.267	.132
6	3.71	3.14	2.45	1.94	1.44	.906	.718	.553	.265	.131
7	3.50	3.00	2.36	1.90	1.42	.896	.711	.549	.263	.130
8	3.36	2.90	2.31	1.86	1.40	.889	.706	.546	.262	.130
9	3.25	2.82	2.26	1.83	1.38	.883	.703	.543	.261	.129
10	3.17	2.76	2.23	1.81	1.37	.879	.700	.542	.260	.129
11	3.11	2.72	2.20	1.80	1.36	.876	.697	.540	.260	.129
12	3.06	2.68	2.18	1.78	1.36	.873	.695	.539	.259	.128
13	3.01	2.65	2.16	1.77	1.35	.870	.694	.538	.259	.128
14	2.98	2.62	2.14	1.76	1.34	.868	.692	.537	.258	.128
15	2.95	2.60	2.13	1.75	1.34	.866	.691	.536	.258	.128
16	2.92	2.58	2.12	1.75	1.34	.865	.690	.535	.258	.128
17	2.90	2.57	2.11	1.74	1.33	.863	.689	.534	.257	.128
18	2.88	2.55	2.10	1.73	1.33	.862	.688	.534	.257	.127
19	2.86	2.54	2.09	1.73	1.33	.861	.688	.533	.257	.127
20	2.84	2.53	2.09	1.72	1.32	.860	.687	.533	.257	.127
21	2.83	2.52	2.08	1.72	1.32	.859	.686	.532	.257	.127
22	2.82	2.51	2.07	1.72	1.32	.858	.686	.532	.256	.127
23	2.81	2.50	2.07	1.71	1.32	.858	.685	.532	.256	.127
24	2.80	2.49	2.06	1.71	1.32	.857	.685	.531	.256	.127
25	2.79	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
26	2.78	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
27	2.77	2.47	2.05	1.70	1.31	.855	.684	.531	.256	.127
28	2.76	2.47	2.05	1.70	1.31	.855	.683	.530	.256	.127
29	2.76	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
30	2.75	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
40	2.70	2.42	2.02	1.68	1.30	.851	.681	.529	.255	.126
60	2.66	2.39	2.00	1.67	1.30	.848	.679	.527	.254	.126
120	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
∞	2.58	2.33	1.96	1.645	1.28	.842	.674	.524	.253	.126

Mais on veut juste une partie de l'intervalle (?)

Donc
$\overline{t_{n-2}}\left(\frac{\alpha}{2}\right) = t_{n-2}\,; 1 - \frac{\alpha}{2}$
A B

A et B à regarder dans la table

Si la valeur trouvée T_n appartient à l'intervalle de rejet IR alors, on rejette l'hypothèse.

Régression Linéaire Multiple

1- MONTRER QU'ON A UN MODELE LINEAIRE

- Graphique

C'est une surface, on ne la représente pas graphiquement (la plupart du temps)

- Formules et hypothèses

$\boldsymbol{y_i = \beta_0 + \beta_1 l_i + \beta_2 k_i + \varepsilon_i}$ $\quad \forall i \in \{1, \dots, n\}$

Hypothèses :

- ε_i non observé aléatoires
- y_i observées aléatoires
- l_i et k_i observées non aléatoires
- (A1) $\mathbb{E}(\varepsilon_i) = 0 \quad \forall i \in \{1, \dots, n\} \quad \Leftrightarrow \mathbb{E}(y_i) = \beta_0 + \beta_1 l_i + \beta_2 k_i$
- (A2) $Var(\varepsilon_i) = \sigma^2$ (homosédacité) $\Leftrightarrow Var\ y_i = \sigma^2$
- (A3) $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j \quad \Leftrightarrow Cov(y_i, y_j) = 0$
- (A4) $\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i \in \{1, \dots, n\} \quad \Leftrightarrow y_i \sim \mathcal{N}(\beta_0 + \beta_1 l_i + \beta_2 k_i, \sigma^2) \longrightarrow$ indépendance de $\varepsilon_i \Rightarrow$ indépendance de y_i

2- DONNER L'ECRITURE SOUS FORME MATRICIELLE // DONNER LA DROITE DE REGRESSION

On a :

$y_i = \beta_0 + \beta_1 l_i + \beta_2 k_i + \varepsilon_i$ avec

$$\mathrm{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 & \dots & z_1 \\ 1 & x_2 & \dots & z_2 \\ \dots & \dots & \dots & \dots \\ 1 & x_n & \dots & z_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

D'où $\mathrm{Y} = \mathrm{X}\beta + \varepsilon$ X matrice des inconnues

3- DONNER $\widehat{\boldsymbol{\beta}}$ // ESTIMATION DE L'ERREUR DE LA COURBE

On a : $\mathrm{Y} = \mathrm{X}\beta + \varepsilon$ d'où $\mathrm{Y} = \mathrm{X}\hat{\beta} + \varepsilon$

$\Leftrightarrow Y = X\hat{\beta} \quad \Leftrightarrow X'Y = X'X\hat{\beta}$

$\Leftrightarrow (X'X)^{-1} * (X'Y) = (X'X)^{-1} * (X'X)\hat{\beta}$

$\Leftrightarrow \hat{\beta} = (X'X)^{-1} * (X'Y)$

4- SOMME DES CARRES DES RESIDUS

Truc compliqué à retenir TD2, exo 1, 5) a completer

5- VARIANCE DES RESIDUS

$$s^2 = \frac{SCR}{n-p-1}$$

p = nombre de variables explicites

$p+1$ = nombre de paramètres

6- COEFFICIENT DE DETERMINATION R^2

7- NULLITE DE LA PENTE

On a les hypothèses suivantes :

$H_0 : \beta_1 = \beta_2 = 0$

$H_1 : \beta_1 \neq 0$ ou $\beta_2 \neq 0$

On utilise la loi de Fisher :

$$\boldsymbol{F = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}}} \quad \sim \quad F(p\,, n-p-1)$$

Loi de Fisher-Snedecor

Valeurs de $f(n_1, n_2; 0{,}05)$

Tableau du F de Fisher Snedecor

n1 / n2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.8	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.35	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.89	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.09	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.88	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.21	4.21	4.15	4.10	4.05	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.79	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.50	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.29	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.14	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.01	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	2.91	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.83	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.76	2.7	2.70	2.65	2.62	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.71	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.66	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.61	2.61	2.55	2.46	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.58	2.58	2.51	2.49	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.54	2.54	2.48	2.46	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.51	2.51	2.45	2.42	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.49	2.49	2.42	2.39	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.46	2.46	2.40	2.37	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.44	2.44	2.37	2.34	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.42	2.42	2.36	2.32	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.40	2.40	2.34	2.30	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.39	2.39	2.32	2.28	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.37	2.37	2.31	2.27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.36	2.36	2.29	2.25	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.35	2.35	2.28	2.24	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.33	2.33	2.27	2.22	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.25	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.17	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.09	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.01	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

8- RESUMER DANS UN TABLEAU ANOVA

On résume les informations calculées dans un tableau ANOVA de cette forme :

Source de variation	ddl	Somme des carrés	Carrés moyens	F
Régression	p	$SCE = \sum_{i=1}^{n} (\widehat{y_i} - \overline{y_n})^2$	$\frac{SCE}{p} = \frac{\sum_{i=1}^{n} (\widehat{y_i} - \overline{y_n})^2}{p}$	$\frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{\frac{\sum_{i=1}^{n} (\widehat{y_i} - \overline{y_n})^2}{p}}{\frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n-p-1}}$
Résiduelle	$n-p-1$	$SCR = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$	$\frac{SCR}{n-p-1} = \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n-p-1} = \boldsymbol{s^2}$	
Totale	$n-1$	$SCT = \sum_{i=1}^{n} (y_i - \overline{y_n})^2$		

Analyse de la variance (ANOVA)

1 - INTRO

Le but est d'étudier les différences entre les résultats selon l'influence du facteur. Même si $A \neq B$ on veut voir si $resulat\ A\ \approx resultat\ B$.

On cherche à voir si le résultat dépend du facteur.

Variable quantitative : x (résultat) échantillon

Facteur : élément dont est issu le résultat (souvent le titre des colonnes dans le tableau)

Niveau : nombre de facteurs

2 - MODELE

On a le même nombre de modèles que le nombre de niveaux.

a **1er modèle**

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

Modèle plutôt théorique (« grain grossier » vérifie que c'est grosso modo ok)

j = nombre de niveaux
i = nombre de résultats par niveaux
μ_j = effet du niveau j
ε_{ij} = erreurs

b **2eme modèle**

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

Modèle plus pratique (version plus précise)

μ = effet qui dépend des facteurs
α = effet sur les variables quantitatives

a Hypothèses

- $\mathbb{E}(\varepsilon_{ij}) = 0$
- $Var(\varepsilon_{ij}) = \sigma^2$ la variance est constante
- $Cov(\varepsilon_{ij}, \varepsilon_{kl}) = 0$ les erreurs ne sont pas corrélées (il n'y a pas de lien entre les erreurs)
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \rightarrow y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$ les erreurs sont indépendantes

Donc, conclusion des hypothèses : $\varepsilon_{ij} = 0$

On pose alors $\overline{y_{ij}} = \widehat{\mu_j}$

On applique ça aux nombre de niveaux qu'on a :

$\widehat{\mu_1} = \overline{y_1}$
$\widehat{\mu_2} = \overline{y_2}$
...
$\widehat{\mu_n} = \overline{y_n}$

} n niveaux

Rappel
■ = normal (le vrai)
$\overline{■}$ = moyenne
$\widehat{■}$ = estimateur

Exemple :

b Le modèle 1 est un peu trop « grossier » donc on ajoute un paramètre extérieur qui dépend de la situation (sans lien avec le facteur) : $\boldsymbol{\mu}$
Il faut qu'il influence tous les i, pas seulement l'un d'entre eux (ex : le vent, un attentat, le stress, …)
On ajoute aussi un paramètre lié au facteur : $\boldsymbol{\alpha_j}$
Il faut que ça soit quelque chose qui influence l'état du facteur (ex : humeur du prof, sucre dans une boisson, produits toxiques dans l'engrais)

C'est pourquoi on a :

$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$

Hypothèses

- $\mathbb{E}(\varepsilon_{ij}) = 0$
- $Var(\varepsilon_{ij}) = \sigma^2$ la variance est constante
- $Cov(\varepsilon_{ij}, \varepsilon_{kl}) = 0$ les erreurs ne sont pas corrélées (il n'y a pas de lien entre les erreurs)
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \rightarrow y_{ij} \sim \mathcal{N}(\mu + \alpha_j, \sigma^2)$ les erreurs sont indépendantes

Donc, conclusion des hypothèses : $\varepsilon_{ij} = 0$

3 - ESTIMER LES PARAMETRES α ET μ

On part du 2eme modèle $y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ (b)

On pose alors $y_{ij} = \mu + \alpha_j$ (1) ou $y_{ij} = \mu_k + \alpha_j$ (2)

(1) $y_{ij} = \mu + \alpha_j$	(2) $y_{ij} = \mu_k + \alpha_j$
μ = effet global : moyenne pondérée des effets des niveaux (moyenne globale de tout)	μ_k = effet extérieur qui **influence le facteur** mais n'as de lien direct avec le résultat. (ex : place d'une plante sur l'étagère, examinateur dans un lycée différent, couche de pommade appliquée, …)
$\widehat{\mu_0} \approx \frac{\sum x_i}{n}$	
$y_{ij} = \mu + \alpha_j$ découle du modèle 1 Or dans le modèle 1, on a $\overline{y_{ij}} = \widehat{\mu_j}$ D'où $\widehat{\mu_j} = \mu + \alpha_j$ → ce qu'on cherche → moyenne globale $\widehat{\mu_0}$ → Valeurs du 1er modèle calculées en (a)	$y_{ij} = \mu_k + \alpha_j$ découle du modèle 1 Or dans le modèle 1, on a $\overline{y_{ij}} = \widehat{\mu_j}$ D'où $\widehat{\mu_j} = \mu_k + \alpha_j$ → ce qu'on cherche → on choisit un niveau de référence un peu freestyle, en général $k = 1$ et reste constant. D'où $\widehat{\mu_k} = \widehat{\mu_1}$ tout le temps → Du coup $\widehat{\mu_j} = \widehat{\mu_1} + \alpha_j$
D'où $\alpha_j = \widehat{\mu_j} - \widehat{\mu_0}$ On calcule : $\alpha_1 = \widehat{\mu_1} - \widehat{\mu_0}$ $\alpha_2 = \widehat{\mu_2} - \widehat{\mu_0}$ $\alpha_3 = \widehat{\mu_3} - \widehat{\mu_0}$ … $\alpha_n = \widehat{\mu_n} - \widehat{\mu_0}$ } n niveaux	D'où $\alpha_j = \widehat{\mu_j} - \widehat{\mu_1}$ On calcule : $\alpha_1 = \widehat{\mu_1} - \widehat{\mu_1} = 0$ $\alpha_2 = \widehat{\mu_2} - \widehat{\mu_1}$ $\alpha_3 = \widehat{\mu_3} - \widehat{\mu_1}$ … $\alpha_n = \widehat{\mu_n} - \widehat{\mu_1}$ } n niveaux
Paramètres estimés	Paramètres estimés

4 - TABLEAU ANOVA

Seuls les ddl indices changent car ici on a i (variable) et j (facteur) au lieu de juste i

Source de variation	ddl	Somme des carrés	Carrés moyens	F
Régression	$k-1$	$SCE = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}_{..})^2$	$\frac{SCE}{k-1} = \frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}_{..})^2}{k-1}$	$\frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}} = \frac{\frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}_{..})^2}{k-1}}{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{n-k}}$
Résiduelle	$n-k$	$SCR = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$	$\frac{SCR}{n-k} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{n-k}$	
Totale	$n-1$	$SCT = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{..})^2$		

4 – TEST D'HYPOTHESE

Student ?