La régression linéaire simple

Yohann.Foucher@univ-nantes.fr

Equipe d'Accueil 4275 "Biostatistique, recherche clinique et mesures subjectives en santé", Université de Nantes

Master 2 - Bioinformatique, 23 Novembre 2011

Plan

Plan

Les acquis

- Variable catégorielle à expliquer en fonction d'une autre variable catégorielle explicative (comparaison de deux pourcentages).
  - Test de comparaison de deux fréquences (approximation normale).
  - Test du Chi-deux (plus de deux groupes)
  - Test du Chi-deux exact de Fisher (test non-paramétrique).
  - Test du Chi-deux de Mac Nehmar (données appariées).
- Variable continue à expliquer en fonction d'une autre variable catégorielle explicative à deux modalités (comparaison de deux moyennes).
  - Test de Student : comparaison de deux moyennes.
  - Test de Student sur différence (données appariées).
  - Test de Mann-Withney (test non-paramétrique).
  - Test de Wilcoxon (données appariées, test non-paramétrique).

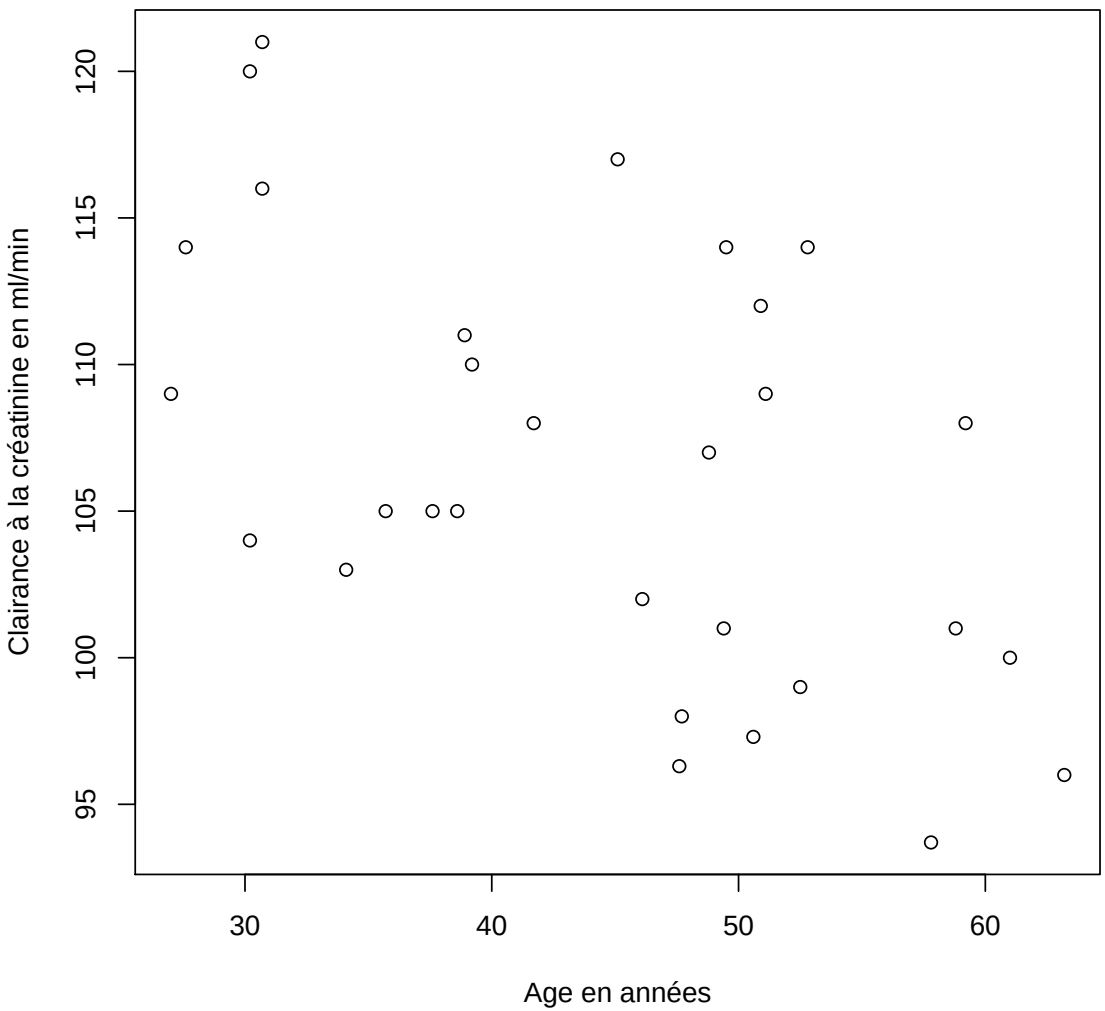
$\rightarrow$ Comment étudier le lien entre deux variables continues ?

Exemple

Etude de la fonction rénale (Y) selon l'âge (X).

Sujet	Y	X	Sujet	Y	X
1	114.0	27.6	16	96.0	63.2
2	110.0	39.2	17	99.0	52.5
3	114.0	52.8	18	103.0	34.1
4	102.0	46.1	19	98.0	47.7
5	101.0	49.4	20	109.0	27.0
6	107.0	48.8	21	120.0	30.2
7	109.0	51.1	22	111.0	38.9
8	117.0	45.1	23	105.0	38.6
9	108.0	41.7	24	116.0	30.7
10	97.3	50.6	25	96.3	47.6
11	101.0	58.8	26	112.0	50.9
12	104.0	30.2	27	105.0	35.7
13	105.0	37.6	28	93.7	57.8
14	121.0	30.7	29	108.0	59.2
15	114.0	49.5	30	100.0	61.0

Exemple


Clairance à la créatinine en ml/min
95
100
105
110
115
120
30
40
50
60
Age en années

Deux méthodes

1. Etude de la corrélation :
   - Interprétation limitée des résultats.
2. Régression linéaire simple :
   - Interprétation plus riche des résultats.
   - Généralisation à plusieurs facteurs explicatifs (variables continues ou catégorielles).

Plan

Calcul et interprétation

- Echantillon composé de n individus ($i = 1, ..., n$)
- Observation des couples (y_i, x_i)
- Indépendance des observations (les $Y_i|x_i$ sont indépendantes).

$$r = \frac{\sum_i y_i \sum_i x_i - n \sum_i x_i y_i}{\sqrt{\left((\sum_i x_i)^2 - n \sum_i x_i^2\right)\left((\sum_i y_i)^2 - n \sum_i y_i^2\right)}}$$

- Interprétation du coefficient de corrélation linéaire :
  - $r = 1$: lien linéaire parfait dans le même sens
  - $r = -1$: lien linéaire parfait dans le sens inverse
  - $|r| > 0.5$: lien linéaire fort
  - $0.3 < |r| < 0.5$: lien linéaire moyen
  - $0.1 < |r| < 0.3$: lien linéaire faible
  - $r = 0$: pas de liaison linéaire

Application à notre exemple

Etude de la fonction rénale (Y) en fonction de l'âge (X)

- $\sum_i y_i = 3196,3$; $\sum_i y_i^2 = 342125,7$; $\sum_i x_i = 1334,3$; $\sum_i x_i^2 = 62626,5$; $\sum_i x_i y_i = 140943,0$
- $r = -0,53$
- Forte corrélation : Il semble que la fonction du rein diminue avec l'âge du patient.

Application à notre exemple

Etude de la fonction rénale (Y) en fonction de l'âge (X)

- $\sum_i y_i = 3196,3$; $\sum_i y_i^2 = 342125,7$; $\sum_i x_i = 1334,3$; $\sum_i x_i^2 = 62626,5$; $\sum_i x_i y_i = 140943,0$
- $r = -0,53$
- Forte corrélation : Il semble que la fonction du rein diminue avec l'âge du patient.

Problème

Peut-on conclure que le coefficient de corrélation linéaire ρ de la population est significativement différent de 0 ?

Test de corrélation

- Définition des hypothèses :
  - $H_0 : \rho = 0$
  - $H_1 : \rho \neq 0$
- Statistique de test : $T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim \mathcal{T}_{n-2\ ddl}$
- $n = 30$; $ddl = 28$; $\alpha = 5\%$
- Région non-critique (test bilatéral) : $[-2,048; 2,048]$
- $t = \frac{0,53\sqrt{28}}{\sqrt{1-0,53^2}} = 3,21 \in$ Région critique
- On rejette l'hypothèse nulle selon laquelle le coefficient de régression linéaire est nul ($p < 5\%$). Il semble qu'il y ait un lien entre la clairance à la créatinine et l'âge.

Table de la loi de Student

LOI DE STUDENT

On connaît α et on cherche t vérifiant P(T>t)

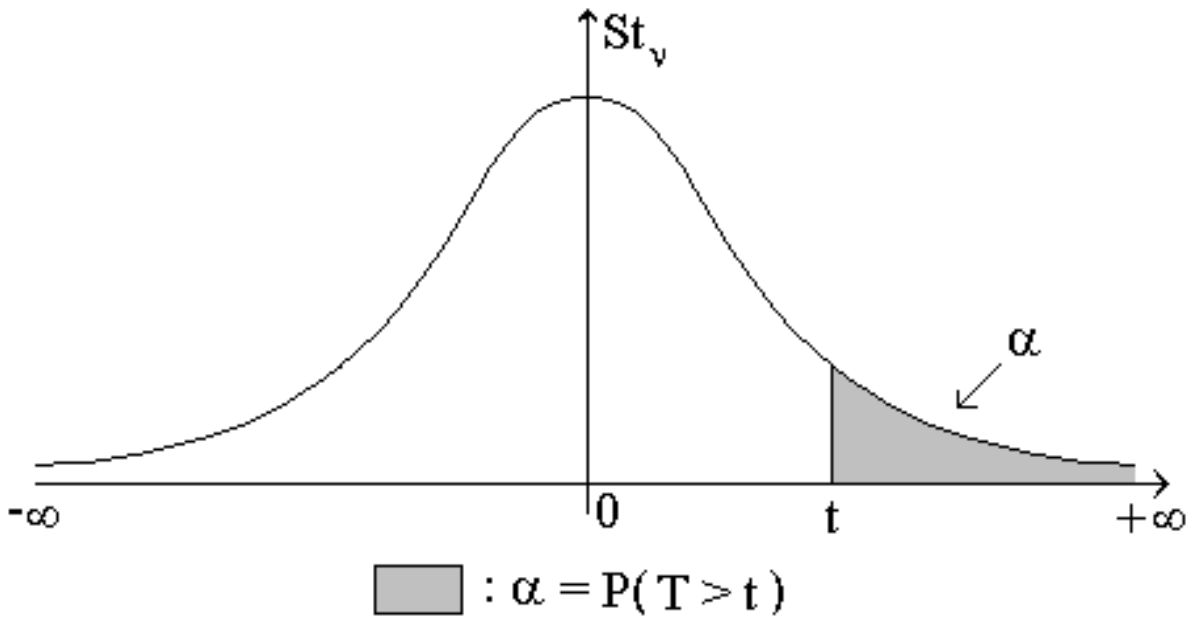
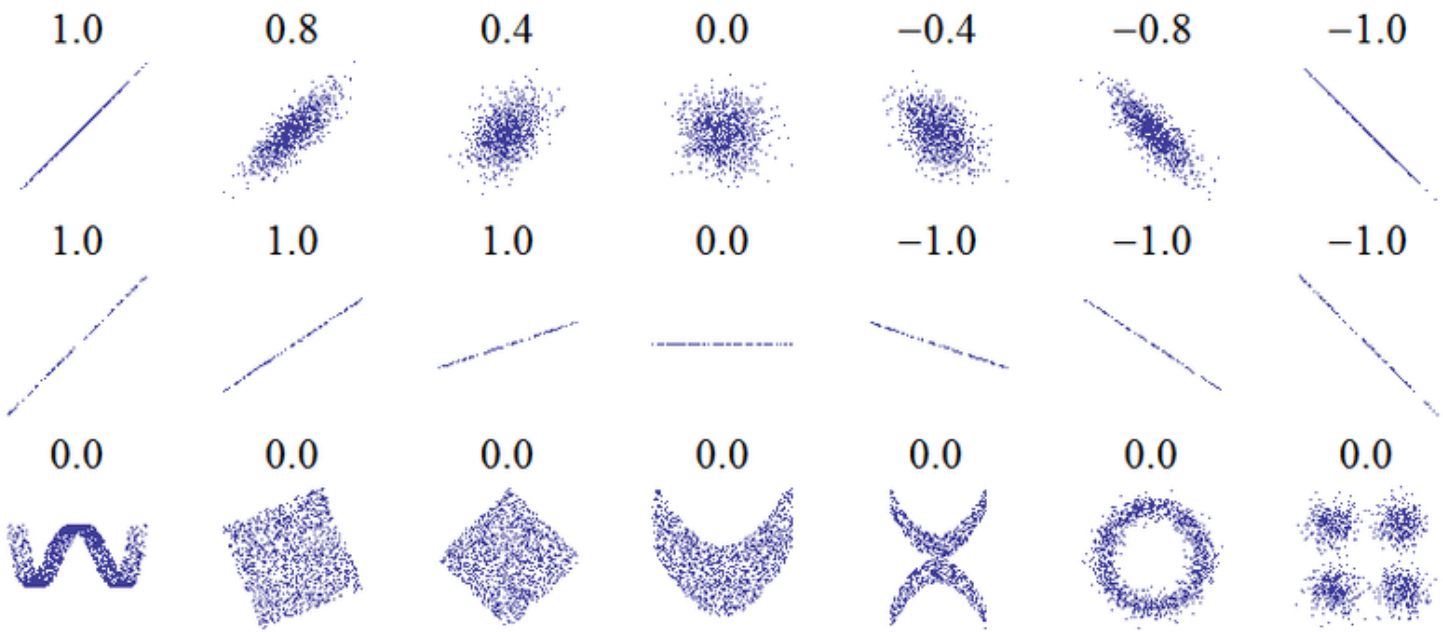


Table de la loi de Student

$\nu \setminus \alpha$	0.35	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.0125	0.01	0.005	0.0025	0.0005
1	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.71	25.45	31.82	63.66	127.3	636.6
2	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.205	6.965	9.925	14.09	31.60
3	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.177	4.541	5.841	7.453	12.93
4	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.495	3.747	4.604	5.598	8.610
5	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.163	3.365	4.032	4.773	6.869
6	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	2.969	3.143	3.707	4.317	5.959
7	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.841	2.998	3.499	4.029	5.408
8	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.752	2.896	3.355	3.833	5.041
9	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.685	2.821	3.250	3.690	4.781
10	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.634	2.764	3.169	3.581	4.587
11	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.593	2.718	3.106	3.497	4.437
12	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.560	2.681	3.055	3.428	4.318
13	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.533	2.650	3.012	3.373	4.221
14	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.510	2.624	2.977	3.326	4.140
15	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.490	2.602	2.947	3.286	4.073
16	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.473	2.583	2.921	3.252	4.015
17	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.458	2.567	2.898	3.223	3.965
18	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.445	2.552	2.878	3.197	3.922
19	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.433	2.539	2.861	3.174	3.883
20	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.423	2.528	2.845	3.153	3.850
21	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.414	2.518	2.831	3.135	3.819
22	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.406	2.508	2.819	3.119	3.792
23	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.398	2.500	2.807	3.104	3.767
24	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.391	2.492	2.797	3.091	3.745
25	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.385	2.485	2.787	3.078	3.725
26	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.379	2.479	2.779	3.067	3.707
27	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.373	2.473	2.771	3.057	3.690
28	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.369	2.467	2.763	3.047	3.674
29	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.364	2.462	2.756	3.038	3.659
30	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.360	2.457	2.750	3.030	3.646
40	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.329	2.423	2.704	2.971	3.551
50	0.388	0.528	0.679	0.849	1.047	1.299	1.676	2.009	2.311	2.403	2.678	2.937	3.497
60	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.298	2.390	2.660	2.921	3.460
120	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.267	2.358	2.617	2.873	3.373
∞	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.236	2.326	2.576	2.828	3.291

Limites



Limites

- Attention aux conclusions non-valides pour une relation non-linéaire.
- Plusieurs relations possibles pour un même coefficient de corrélation.
- Aucune quantification de la relation.

Plan

Définition du modèle

Le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_0 est l'ordonnée à l'origine (moyenne de Y_i quand $x_i = 0$).
- β_1 est la pente (changement moyen de Y_i quand x_i augmente d'une unité).
- ϵ_i est le résidu (différence entre la valeur prédite et celle observée).
- Les résidus sont distribués selon une loi normale de moyenne nulle et de variance σ^2 (variance résiduelle).

Estimation de la droite de régression

- Objectf : Trouver la meilleure droite pour un nuage de points.
- Minimisation des valeurs des résidus.
- Critère des Moindres Carrés :

$$CMC = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

- Calcul des dérivées partielles de CMC :

$$\begin{cases} \partial CMC/\partial \beta_0 = -1 \times 2 \times \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \partial CMC/\partial \beta_1 = -x_i \times 2 \times \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \end{cases}$$

$$\begin{cases} \partial CMC/\partial \beta_0 = -2 \sum_i y_i + 2n\beta_0 + 2\beta_1 \sum_i x_i \\ \partial CMC/\partial \beta_1 = -2 \sum_i y_i x_i + 2\beta_0 \sum_i x_i + 2\beta_1 \sum_i x_i^2 \end{cases}$$

Estimation de la droite de régression

- Les valeurs optimales, $\hat{\beta}_0$ et $\hat{\beta}_1$, minimisent le CMC :

$$\begin{cases} \sum_i y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i = 0 \\ \sum_i y_i x_i - \hat{\beta}_0 \sum_i x_i - \hat{\beta}_1 \sum_i x_i^2 = 0 \end{cases}$$

- Le CMC est minimum pour :

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)/n}{\sum_i x_i^2 - (\sum_i x_i)^2/n} \\ \hat{\beta}_0 = (\sum_i y_i)/n - \hat{\beta}_1 (\sum_i y_i)/n \end{cases}$$

Estimation des autres paramètres importants

- On peut alors simplement déduire la variance résiduelle des estimations précédentes.

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

- Si $\hat{\beta}_1$ représente la pente de Y en fonction de X et que $\hat{\beta}'_1$ représente la pente de X en fonction de Y, alors on montre que :

$$\hat{r}^2 = \hat{\beta}_1 \hat{\beta}'_1$$

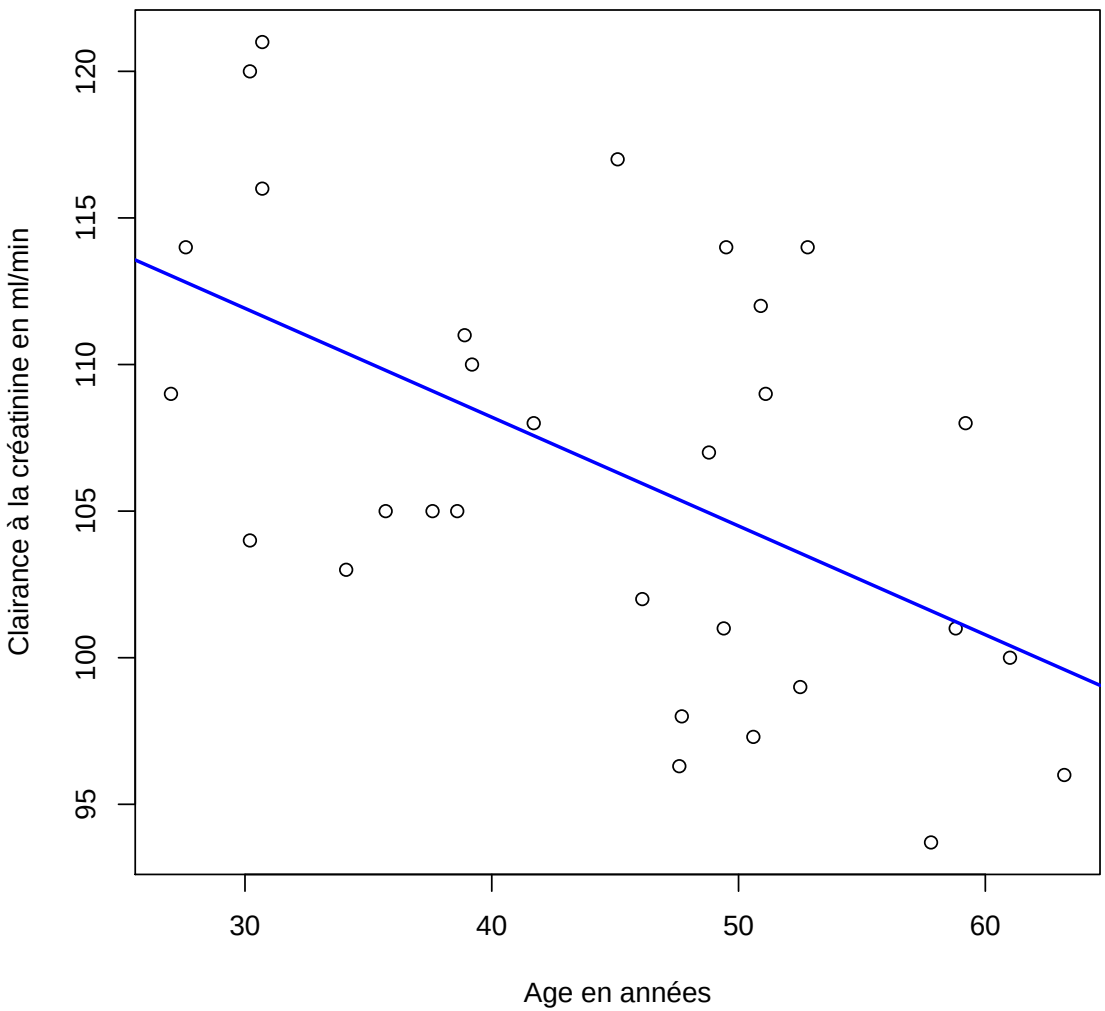
r^2 représente la proportion de variation de Y expliquée par X. Rappelons que r est le coefficient de corrélation linéaire.

Application à notre exemple

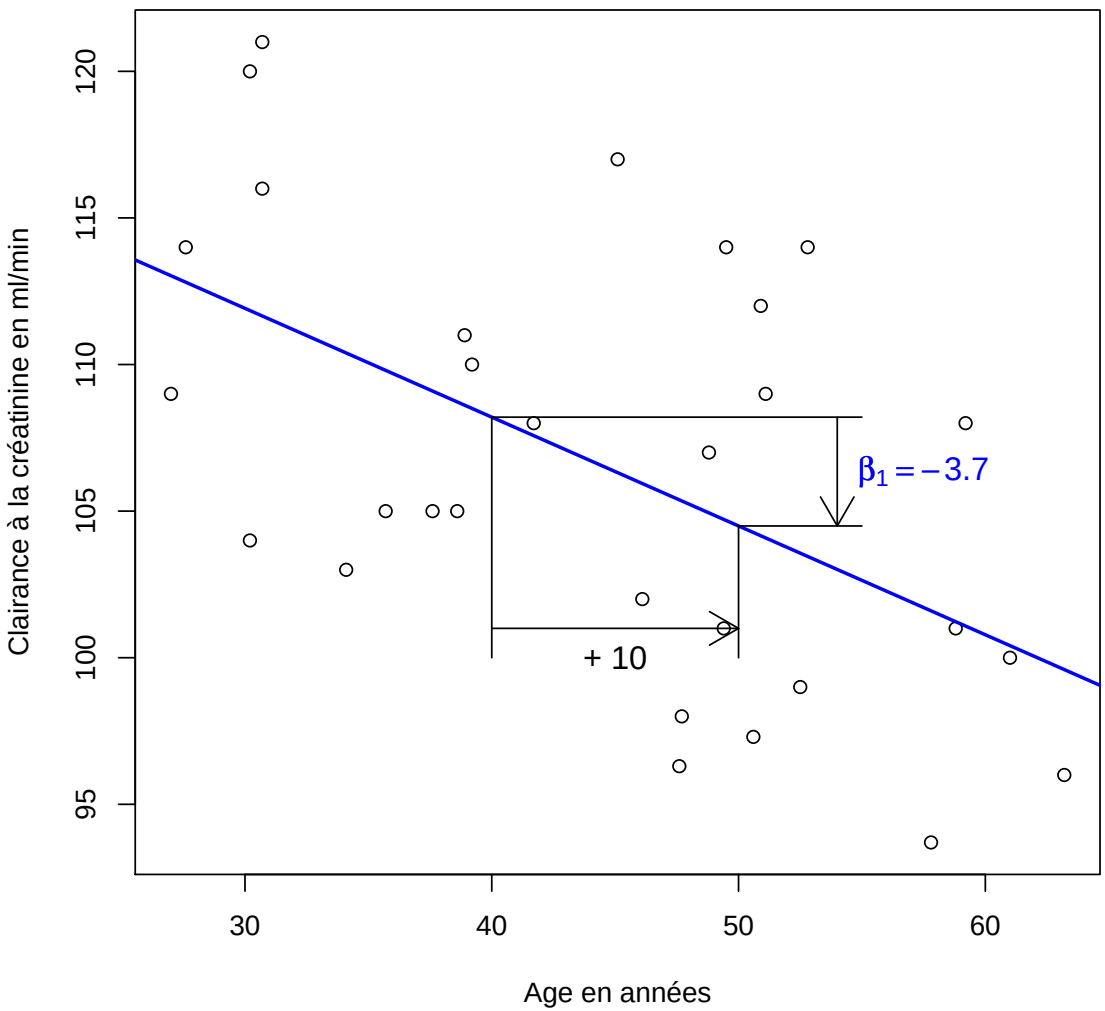
Etude de la fonction rénale (Y) en fonction de l'âge (X)

- $n = 30$; $\sum_i y_i = 3196,3$; $\sum_i y_i^2 = 342125,7$; $\sum_i x_i = 1334,3$; $\sum_i x_i^2 = 62626,5$; $\sum_i x_i y_i = 140943,0$;
- $\hat{\beta}_0 = 123,0$: La fonction rénale d'un nouveau né ($x = 0$) est estimée à 123,0 ml/min en moyenne. Attention : cette valeur n'est pas fiable (aucun enfant dans l'étude).
- $\hat{\beta}_1 = -0,37$: La fonction rénale chute en moyenne de 3,7 ml/min tous les 10 ans.
- $\hat{\sigma}^2 = 6,35$.
- $\hat{r}^2 = 0,29$: 29% de la variation de Y est expliquée par X.

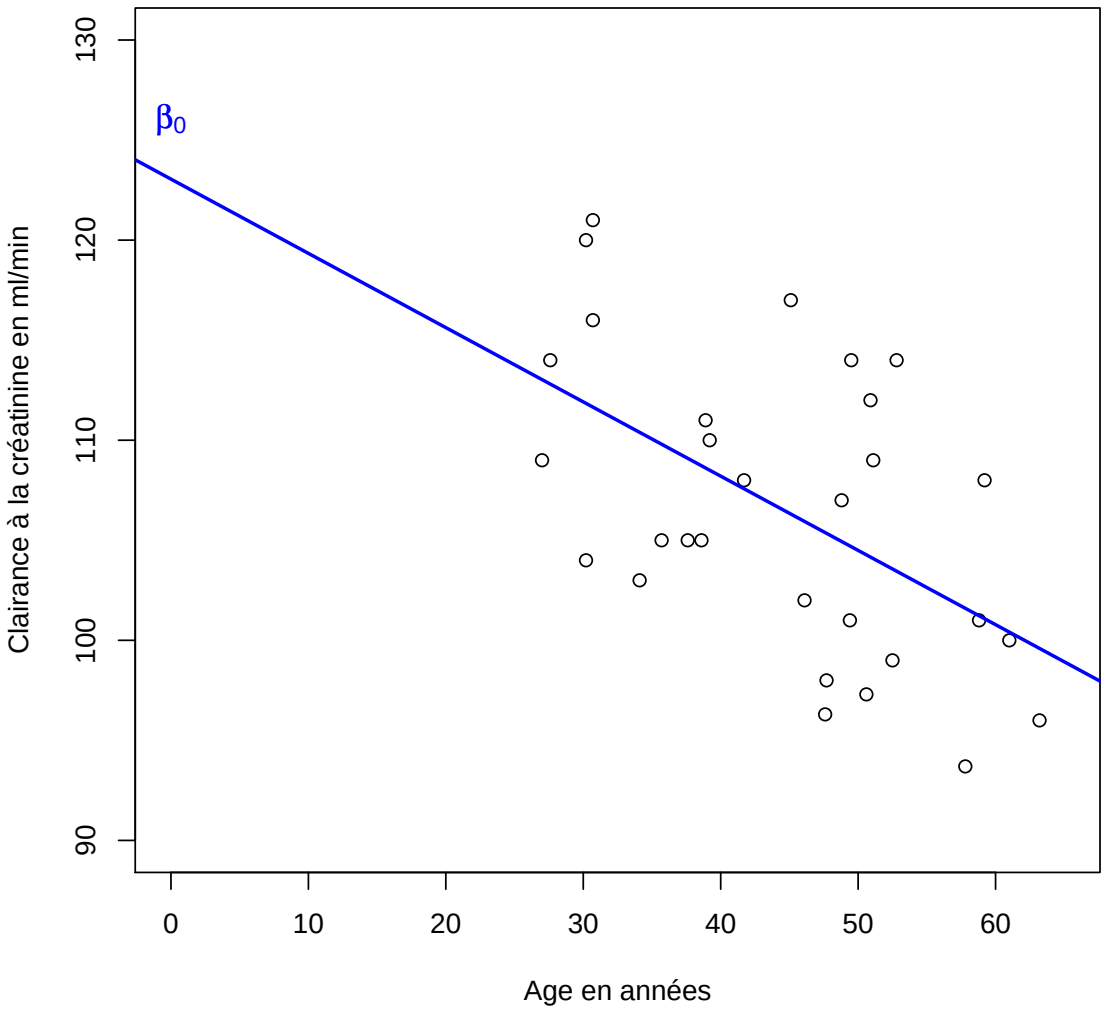
Application à notre exemple



Application à notre exemple



Application à notre exemple



Intervalle de confiance de β_1

$$IC_{(1-\alpha)} = \left[\hat{\beta}_1 \pm t_{\alpha,n-2} s(\hat{\beta}_1)\right]$$

- $t_{\alpha,n-2}$: fractile de la loi de Student à $n-2$ ddl (lue dans la table).
- $s(\hat{\beta}_1)$: écart-type estimé de la pente
  - Remarque : $s(\hat{\beta}_1) = \hat{\sigma}/(\hat{\sigma}_x\sqrt{n-1})$, où $\hat{\sigma}_x$ est l'écart-type de X.
- Si l'intervalle de confiance comprend la valeur 0, on conclura que la pente n'est pas significativement différente de 0.
- Si l'intervalle de confiance ne comprend pas la valeur 0, on conclura que la pente est significativement différente de 0.

Test de β_1 (méthode 1 : Test de Student)

- Définition des htypothèses
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$
- Statistique de test : $T = \beta_1 / s(\beta_1) \sim \mathcal{T}_{n-2 \ ddl}$
- Définition de la région critique.
- Si t appartient à la région critique, on rejette H_0, sinon on ne peut pas rejeter H_0.

Application à notre exemple

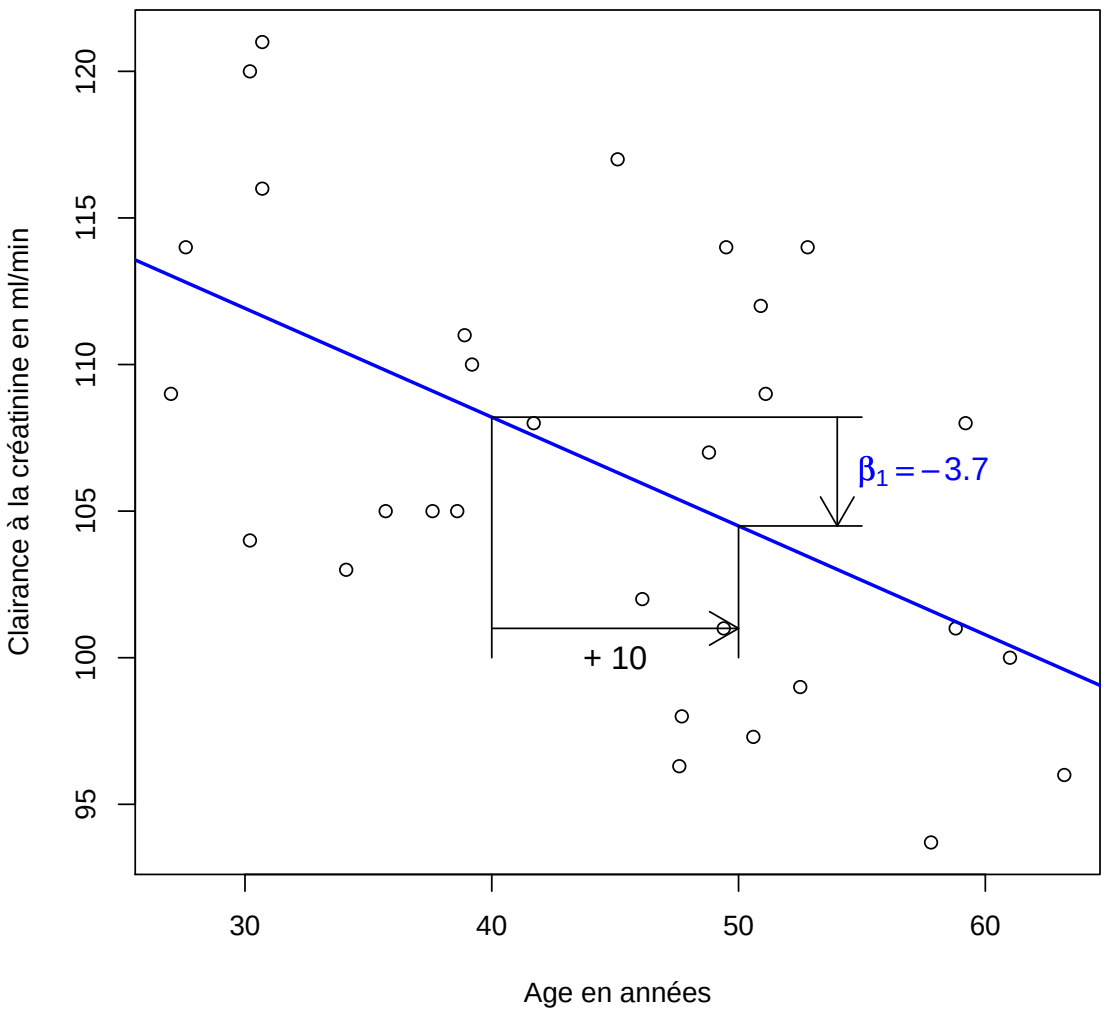
Etude de la fonction rénale (Y) en fonction de l'âge (X)

- Intervalle de confiance à 95% de β_1 ($t_{5\%;28} = 2,048$) :

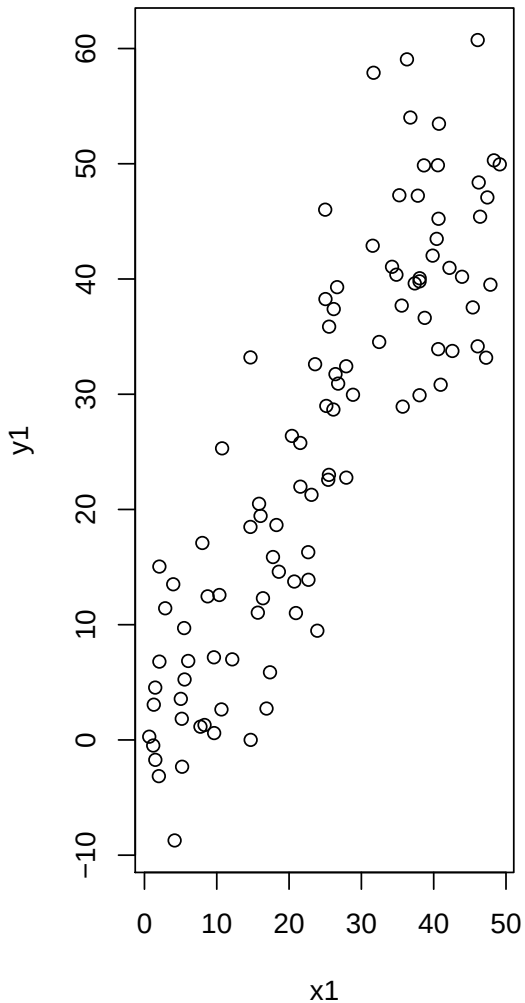
$$IC_{95\%} = [-0,37 \pm 2,048 \times 0,11] = [-0,48; -0,26]$$

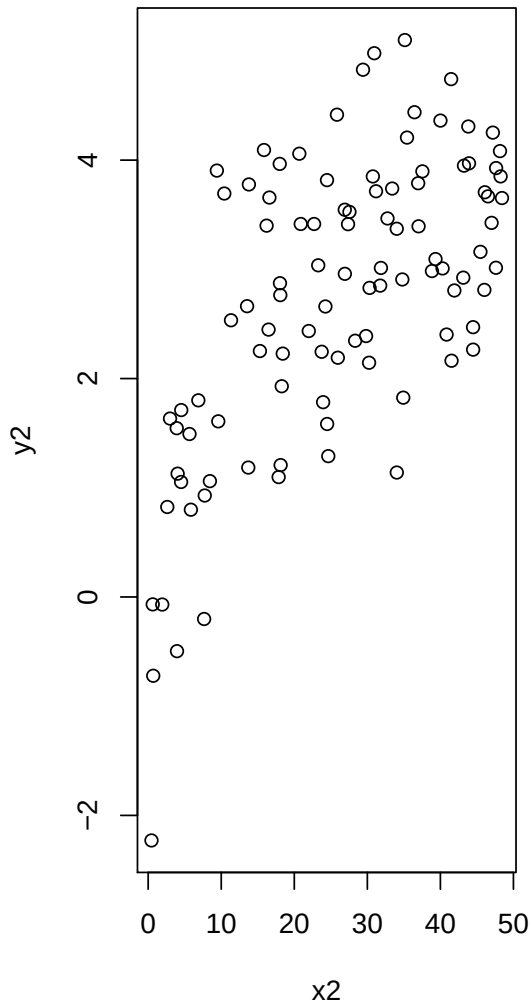
- L'intervalle de confiance ne comprend pas la valeur 0, il semble donc que la pente soit significativement différente de zéro.
- Test $\rightarrow$ $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$
- $\alpha = 0,05$, $t_{5\%;28} = 2,048$
- $t = -0,37/0,11 = -3,35$
- $|t| > 2,048$, on rejette H_0.

Hypothèses du modèle : linéarité

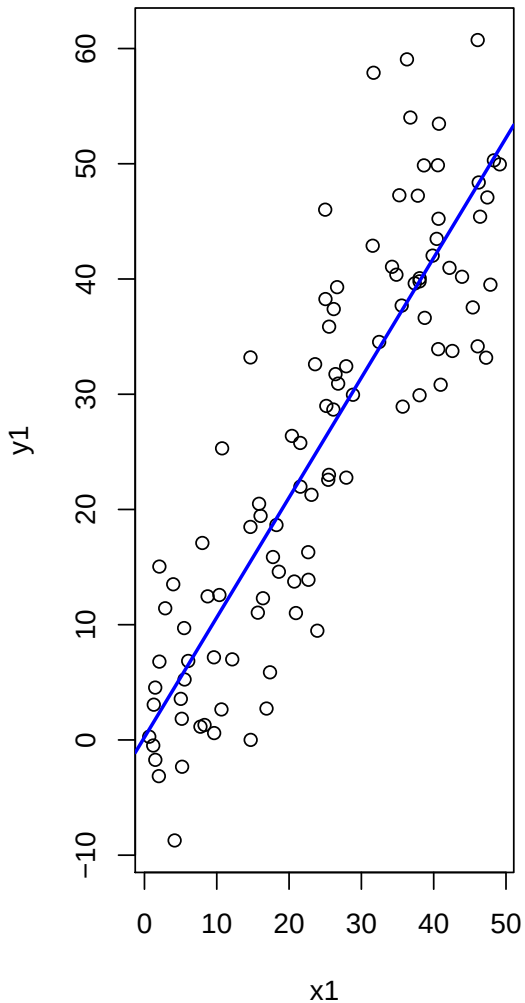


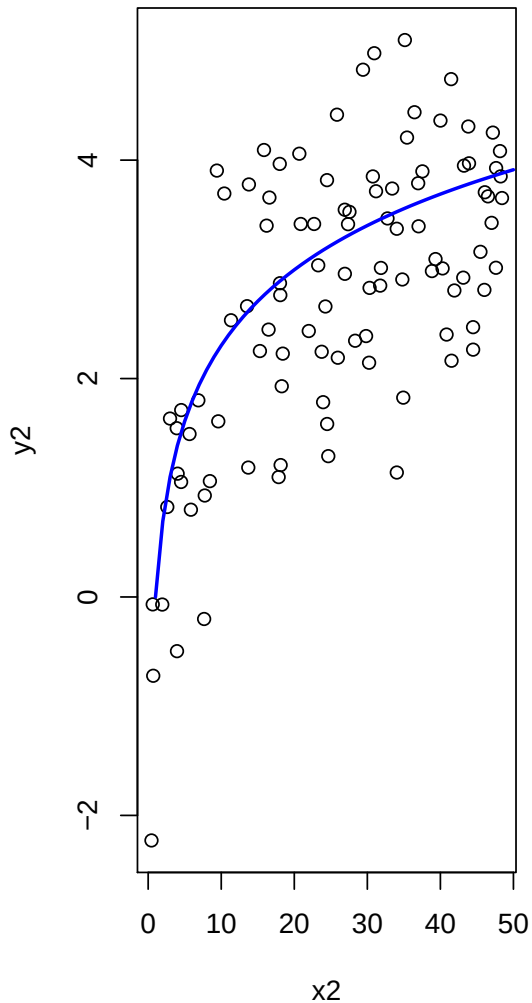
Hypothèses du modèle : linéarité



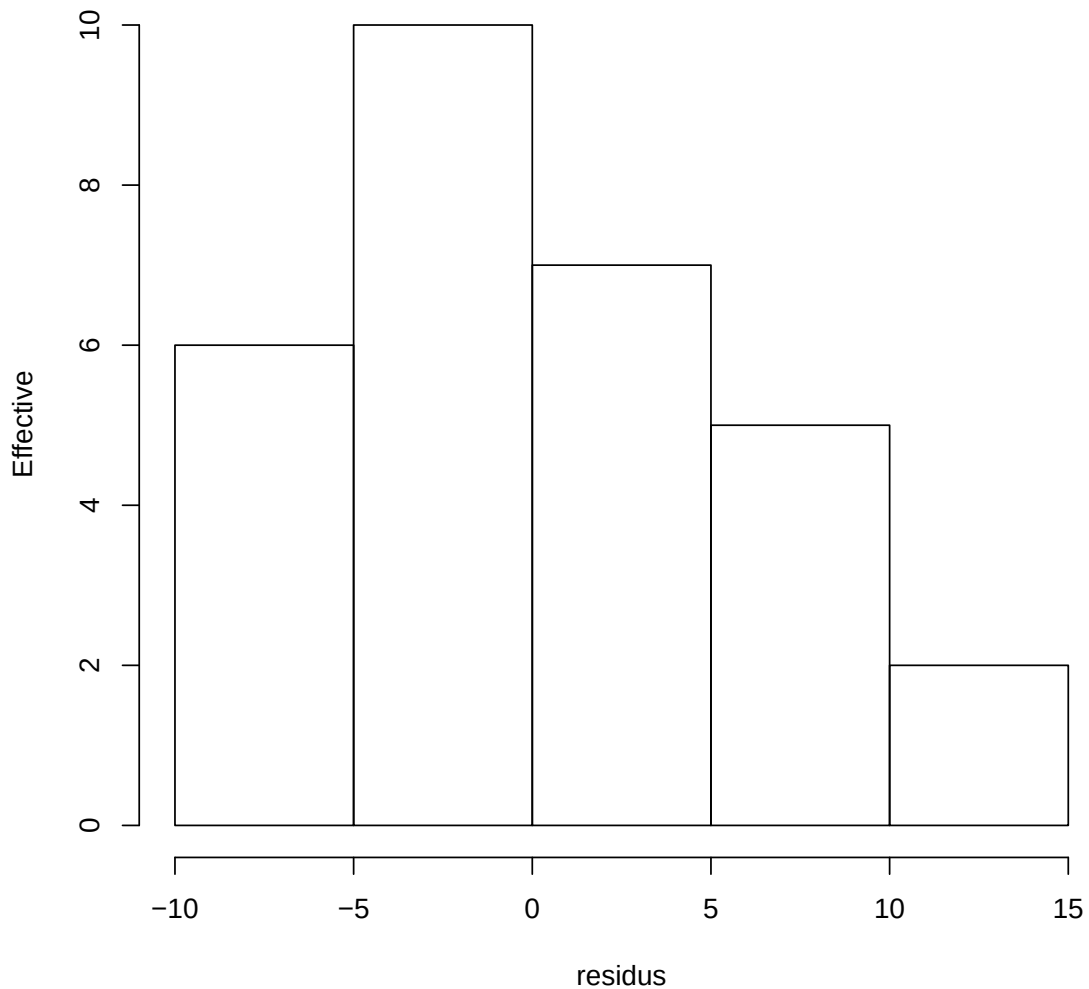


Hypothèses du modèle : linéarité





Hypothèses du modèle : normalité des résidus



Hypothèses du modèle : homoscédasticité des résidus