Chapitre 6 : Régression linéaire

- Le problème de régression
- Le modèle de régression linéaire simple
- Estimation par la méthode des moindres carrés
- Le modèle linéaire simple gaussien
- Etude complète d'un exemple en R

Le problème de régression

Introduction à l'analyse statistique des données multidimensionnelles :

- On mesure plusieurs variables sur chaque individu.
- Les observations sont des réalisations de vecteurs aléatoires, dont les composantes sont dépendantes.

Cas des données bidimensionnelles :

- On observe les réalisations de couples aléatoires $(X_1, Y_1), \ldots (X_n, Y_n)$, indépendants et de même loi.
- **Dépendance** entre X_i et Y_i ?

Problème de **régression** :

- On cherche une fonction f telle que $\forall i,\ Y_i \approx f(X_i)$.
- **Régression linéaire simple** : $\forall i,\ Y_i \approx \beta_1 X_i + \beta_0$.
- Pour estimer β_1 et β_0, on utilise la **méthode des moindres carrés**.

Exemple : vitesse et distance de freinage

- x : vitesse d'une voiture (en m/s) au moment où l'on freine.
- y : distance de freinage (en m).
- n expériences pour n vitesses différentes.

numéro de mesure i	1	2	3	4	5	6	7	8
vitesse x_i	5	10	15	20	25	30	35	40
distance de freinage y_i	3.42	5.96	31.14	41.76	74.54	94.92	133.78	169.16

- Quel modèle de dépendance entre la distance de freinage et la vitesse peut-on proposer ?
- Une dépendance affine est-elle raisonnable ?
- Peut-on estimer la distance de freinage d'une voiture lancée à 50 m/s ? Avec quelle précision ?

Modèle de régression de Y sur x

Hypothèses :

- Les x_i sont des constantes connues.
- Les y_i sont des réalisations de variables aléatoires Y_i.
- Les effets sur Y de x et des autres facteurs s'ajoutent.

Modèle de régression de Y sur x

$$Y = f(x) + \varepsilon$$

- Y est la **variable à expliquer** ou variable expliquée.
- x est la **variable explicative** ou prédicteur ou régresseur.
- ε est l'erreur de prévision de Y par $f(x)$ ou **résidu**.

Le modèle de régression linéaire simple

Hypothèses supplémentaires :

- f est une fonction affine de x.
- Les n mesures sont indépendantes, effectuées dans les mêmes conditions.
- Les effets des facteurs autres que le prédicteur s'équilibrent.

Modèle de régression linéaire simple

$$\forall i \in \{1, \dots, n\}, \;\; Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$$

- β_0 et β_1 sont des paramètres réels inconnus.
- Les résidus ε_i sont indépendants, de même loi, centrés et de variance σ^2.

$$\begin{aligned} E[Y_i] &= \beta_1 x_i + \beta_0 \\ Var[Y_i] &= \sigma^2 \end{aligned}$$

σ^2 mesure le bruit ou le poids des facteurs autres que x.

Plus σ^2 est élevé, plus Y_i fluctue autour de $\beta_1 x_i + \beta_0$.

Inversement, pour $\sigma^2 = 0$, les points (x_i, Y_i) sont parfaitement alignés : Y_i n'est plus aléatoire.

Problèmes statistiques :

- *Estimation* de β_1, β_0 et σ^2, ponctuelle et par intervalle de confiance.
- Construction de *tests d'hypothèses* portant sur β_1, β_0 et σ^2.
- *Prévision* de y connaissant x.
- *Validation* du modèle : la liaison entre la vitesse et la distance de freinage est-elle bien affine ?

Remarque fondamentale

Dans un modèle de régression linéaire simple, la liaison entre x et y n'est pas **linéaire**, mais **affine**.

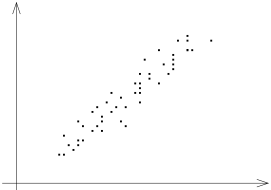
Mais en fait, ce qui est important, c'est que y dépende linéairement du couple (β_1, β_0).

Ainsi, contrairement aux apparences, les modèles suivants sont bien des modèles linéaires :

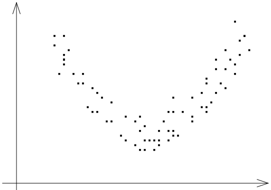
1. $Y_i = \beta_1 \ln x_i + \beta_0 + \varepsilon_i$.
2. $Y_i = \beta_2 x_i^2 + \beta_1 x_i + \beta_0 + \varepsilon_i$.
3. $Y_i = \beta_0 x_i^{\beta_1} \varepsilon_i$ (car $\ln Y_i = \beta_1 \ln x_i + \ln \beta_0 + \ln \varepsilon_i$).
4. $Y_i = \dfrac{e^{(\beta_1 x_i + \beta_0 + \varepsilon_i)}}{1 + e^{(\beta_1 x_i + \beta_0 + \varepsilon_i)}}$ (car $\ln \dfrac{Y_i}{1 - Y_i} = \beta_1 x_i + \beta_0 + \varepsilon_i$).

En revanche, le modèle $Y_i = \beta_1 + e^{\beta_0 x_i} + \varepsilon_i$ n'est pas linéaire.

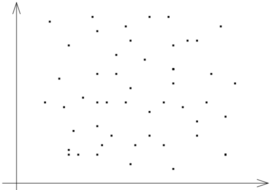
Liaisons possibles entre x et y nuage des points $(x_i, y_i), \forall i \in \{1, \ldots, n\}$



liaison affine



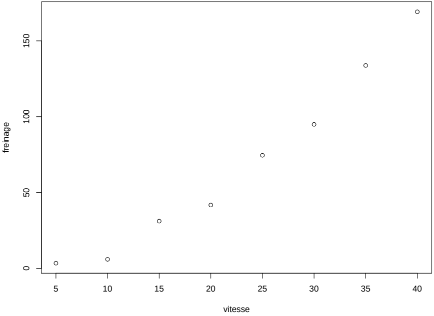
liaison non affine en x



pas de liaison apparente

Vitesse/distance de freinage : nuage des points (x_i, y_i)

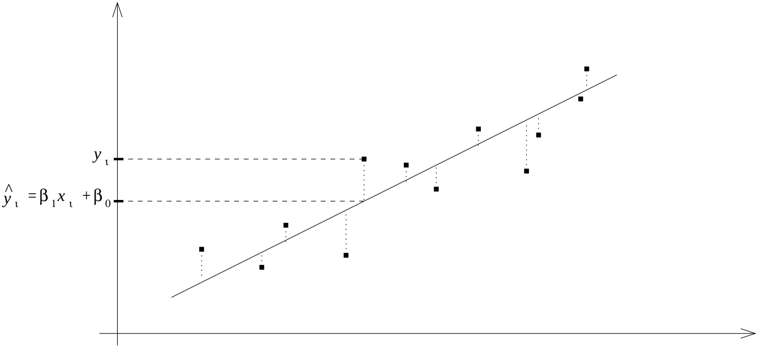
```
>  vitesse<-c(5,10,15,20,25,30,35,40)
>  freinage<-c(3.42,5.96,31.14,41.76,74.54,94.92,133.78,169.16)
>  plot(vitesse,freinage)
```



Méthode des moindres carrés

Problème : trouver la droite “la plus proche” du nuage de points, en un certain sens.

Méthode des moindres carrés : prendre la droite pour laquelle la somme des carrés des distances verticales des points à la droite est minimale.



Notations

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$ moyenne empirique des x_i

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$ moyenne empirique des y_i

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$$ variance empirique des x_i

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}_n^2$$ variance empirique des y_i

$$c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n\bar{y}_n$$ covariance empirique entre les x_i et les y_i

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$ coefficient de corrélation linéaire empirique entre les x_i et les y_i

Coefficient de corrélation linéaire empirique

r_{xy} est la version empirique du coefficient de corrélation linéaire $\rho(X, Y)$ et possède des propriétés équivalentes :

- $r_{xy} \in [-1, 1]$.
- $r_{xy} = +1 \iff$ les points (x_i, y_i) sont alignés sur une droite de pente positive.
- $r_{xy} = -1 \iff$ les points (x_i, y_i) sont alignés sur une droite de pente négative.
- Si y ne dépend pas de x, r_{xy} doit être proche de 0. Réciproquement, si r_{xy} est proche de 0, alors il n'y a pas de dépendance affine entre x et y, mais il est possible qu'il existe une dépendance non affine.

Estimateurs des moindres carrés

On minimise l'**erreur quadratique moyenne** :

$$\delta^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

Estimateurs des moindres carrés

On minimise l'**erreur quadratique moyenne** :

$$\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$$

$\dfrac{\partial\delta^2}{\partial\beta_0} = 0 \Rightarrow \bar{y}_n - \beta_1\bar{x}_n - \beta_0 = 0$. Par conséquent, la droite des moindres carrés passe par le barycentre du nuage $(\bar{x}_n, \bar{y}_n)$.

$\dfrac{\partial\delta^2}{\partial\beta_1} = 0 \Rightarrow c_{xy} - \beta_1 s_x^2 = 0.$

Estimateurs des moindres carrés de β_1 et β_0

$$\widehat{\beta}_1 = \frac{C_{xY}}{s_x^2} \qquad \text{et} \qquad \widehat{\beta}_0 = \bar{Y}_n - \frac{C_{xY}}{s_x^2}\bar{x}_n$$

Droite des moindres carrés

Equation de la droite des moindres carrés

$$y = \widehat{\beta}_1 x + \widehat{\beta}_0 = \bar{y}_n + \frac{c_{xy}}{s_x^2}(x - \bar{x}_n)$$

Erreur quadratique moyenne minimale :

$$\delta^2_{min} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0\right)^2 = s_y^2(1 - r_{xy}^2)$$

$\Rightarrow r_{xy} \in [-1, +1]$

et $\delta^2_{min} = 0 \iff r_{xy} = \pm 1 \iff$ les points du nuage sont alignés.

Qualité des estimateurs des moindres carrés de β_1 et β_0

$\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont des estimateurs sans biais et convergents de β_1 et β_0.

$$Var[\widehat{\beta}_1] = \frac{\sigma^2}{ns_x^2}. \qquad Var[\widehat{\beta}_0] = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}_n^2}{s_x^2}\right).$$

$$Cov(\widehat{\beta}_1, \widehat{\beta}_0) = -\frac{\sigma^2 \bar{x}_n}{ns_x^2}. \qquad Cov(\widehat{\beta}_1, \bar{Y}_n) = 0.$$

Théorème de Gauss-Markov

$\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont les estimateurs de β_1 et β_0 sans biais et de variance minimale parmi tous les estimateurs sans biais linéaires (qui s'écrivent comme des combinaisons linéaires des Y_i) .

Estimation de la variance des résidus σ^2

Puisque, $\forall i$, $\sigma^2 = Var[\varepsilon_i] = Var[Y_i - \beta_1 x_i - \beta_0]$, il est naturel d'estimer σ^2 par la variance empirique des $Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0$.

- **Résidus empiriques** : $E_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0$.
- **Variance résiduelle** : $S^2_{Y|x}$ = variance empirique des résidus empiriques.

$$S^2_{Y|x} = \frac{1}{n}\sum_{i=1}^{n} E_i^2 - \bar{E}_n^2 = \delta^2_{min} = S^2_Y(1 - R^2_{xY})$$

Estimation de σ^2

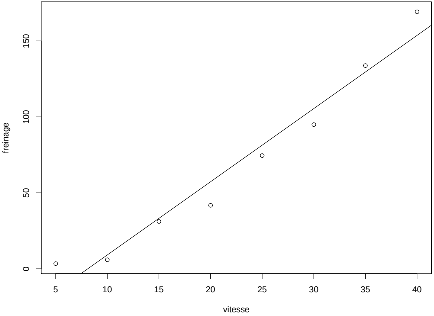
$$\widehat{\sigma}^2 = \frac{n}{n-2} S^2_{Y|x} = \frac{n}{n-2} S^2_Y(1 - R^2_{xY}) = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0)^2$$

est un estimateur sans biais de σ^2.

Exemple : vitesse/distance de freinage

$\widehat{\beta}_1 = 4.82, \quad \widehat{\beta}_0 = -39.06, \quad \widehat{\sigma}^2 = 168.4, \quad r_{xy} = 0.9799.$

Droite des moindres carrés : $y = 4.82x - 39.06$.



Prévision de la distance de freinage d'une voiture lancée à 50 m/s :

$$4.82 * 50 - 39.06 = 201.9 \text{ m}.$$

Le modèle de régression linéaire simple gaussien

Modèle de régression linéaire simple

$$\forall i \in \{1, \dots, n\}, \;\; Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$$

- β_0 et β_1 sont des paramètres réels inconnus.
- Les résidus ε_i sont indépendants, de même loi, centrés et de variance σ^2.

Le modèle de régression linéaire simple gaussien

Modèle de régression linéaire simple

$$\forall i \in \{1, \dots, n\}, \;\; Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$$

- β_0 et β_1 sont des paramètres réels inconnus.
- Les résidus ε_i sont indépendants, de même loi, centrés et de variance σ^2.

Modèle de régression linéaire simple **gaussien**

$$\forall i \in \{1, \dots, n\}, \;\; Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$$

- β_0 et β_1 sont des paramètres réels inconnus.
- Les résidus ε_i sont indépendants, de même loi **normale** $\mathcal{N}(0, \sigma^2)$.

Propriétés

- Les Y_i sont indépendantes et de lois de probabilité respectives $\mathcal{N}(\beta_1 x_i + \beta_0, \sigma^2)$.
- $\widehat{\beta}_1$ est de loi $\mathcal{N}\left(\beta_1, \dfrac{\sigma^2}{n s_x^2}\right)$.
- $\widehat{\beta}_0$ est de loi $\mathcal{N}\left(\beta_0, \dfrac{\sigma^2}{n}(1 + \dfrac{\overline{x}_n^2}{s_x^2})\right)$.
- $\dfrac{(n-2)\widehat{\sigma}^2}{\sigma^2} = \dfrac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0\right)^2$ est de loi $\mathcal{X}_{n-2}^2$.
- $\widehat{\sigma}^2$ est indépendant de $\bar{Y}_n, \widehat{\beta}_1$ et $\widehat{\beta}_0$.
- $\widehat{\beta}_1$, $\widehat{\beta}_0$ et $\widehat{\sigma}^2$ sont les ESBVM de β_1, β_0 et σ^2.

Maximum de vraisemblance

Les estimateurs de maximum de vraisemblance de β_1, β_0 et σ^2 sont $\widehat{\beta}_1$, $\widehat{\beta}_0$ et $\dfrac{n-2}{n}\widehat{\sigma}^2$.

Intervalles de confiance

- Un intervalle de confiance de seuil α pour β_1 est :

$$\left[\widehat{\beta}_1 - \frac{\widehat{\sigma}\, t_{n-2,\alpha}}{s_x\sqrt{n}}, \widehat{\beta}_1 + \frac{\widehat{\sigma}\, t_{n-2,\alpha}}{s_x\sqrt{n}}\right]$$

- Un intervalle de confiance de seuil α pour β_0 est :

$$\left[\widehat{\beta}_0 - \frac{t_{n-2,\alpha}\widehat{\sigma}\sqrt{s_x^2 + \bar{x}_n^2}}{s_x\sqrt{n}}, \widehat{\beta}_0 + \frac{t_{n-2,\alpha}\widehat{\sigma}\sqrt{s_x^2 + \bar{x}_n^2}}{s_x\sqrt{n}}\right]$$

- Un intervalle de confiance de seuil α pour σ^2 est :

$$\left[\frac{(n-2)\widehat{\sigma}^2}{z_{n-2,\frac{\alpha}{2}}}, \frac{(n-2)\widehat{\sigma}^2}{z_{n-2,1-\frac{\alpha}{2}}}\right]$$

Exemple vitesse/distance de freinage

$\widehat{\beta}_1 = 4.82, \quad \widehat{\beta}_0 = -39.06, \quad \widehat{\sigma}^2 = 168.4, \quad r_{xy} = 0.9799.$

$n = 8$ et $\alpha = 10\% \Rightarrow t_{6,0.1} = 1.943$, $z_{6,0.05} = 12.59$ et $z_{6,0.95} = 1.64$.

$$
\begin{aligned}
IC(\beta_1) &= [4.04, 5.60] \\
IC(\beta_0) &= [-58.71, -19.41] \\
IC(\sigma^2) &= [80.2, 617.8]
\end{aligned}
$$

Tests d'hypothèses

- Test de seuil α de "$\beta_1 = b$" contre "$\beta_1 \neq b$" :

$$W = \left\{ \left| \frac{\widehat{\beta}_1 - b}{\widehat{\sigma}} \right| s_x \sqrt{n} > t_{n-2,\alpha} \right\}$$

- Test de seuil α de "$\beta_0 = b$" contre "$\beta_0 \neq b$" :

$$W = \left\{ \left| \frac{\widehat{\beta}_0 - b}{\widehat{\sigma}} \right| \frac{s_x \sqrt{n}}{\sqrt{s_x^2 + \bar{x}_n^2}} > t_{n-2,\alpha} \right\}$$

- Test de seuil α de "$\sigma = \sigma_0$" contre "$\sigma \neq \sigma_0$" :

$$W = \left\{ \frac{(n-2)\widehat{\sigma}^2}{\sigma_0^2} < z_{n-2,1-\frac{\alpha}{2}} \quad \text{ou} \quad \frac{(n-2)\widehat{\sigma}^2}{\sigma_0^2} > z_{n-2,\frac{\alpha}{2}} \right\}$$

Test de pertinence du modèle de régression linéaire

Si les points (x_i, y_i) sont parfaitement alignés, alors r_{xy} est égal à ± 1.

Inversement, si r_{xy} est proche de 0, on peut rejeter l'hypothèse de liaison affine entre x et y.

$\Rightarrow$ on va admettre la validité la liaison affine si r_{xy} est "suffisamment proche" de ± 1, ou si r_{xy} est "suffisamment éloigné" de 0 :

$$W = \left\{ r_{xy}^2 > l_\alpha \right\}$$

Test de pertinence de la régression

Sous H_0 : "$\beta_1 = 0$",

- $\dfrac{R_{xY}}{\sqrt{1 - R_{xY}^2}}\sqrt{n-2}$ est de loi $St(n-2)$.
- $\dfrac{(n-2)R_{xY}^2}{1 - R_{xY}^2}$ est de loi $F(1, n-2)$.

Test de pertinence de la régression

Test de seuil α de H_0 : "$\beta_1 = 0$" contre H_1 : "$\beta_1 \neq 0$".

$$W = \left\{ \frac{(n-2)r_{xy}^2}{1 - r_{xy}^2} > f_{1,n-2,\alpha} \right\} = \left\{ r_{xy}^2 > \frac{f_{1,n-2,\alpha}}{n - 2 + f_{1,n-2,\alpha}} \right\}$$

Exemple vitesse/distance de freinage

$$\frac{(n-2)r_{xy}^2}{1-r_{xy}^2} = 144.7.$$

$f_{1,6,0.05} = 5.99$ et $f_{1,6,0.01} = 13.8$.

$\Rightarrow$ même au seuil 1%, on est très largement dans la région critique, donc on conclut que la régression linéaire est ici très pertinente.

Exemple vitesse/distance de freinage

$$\frac{(n-2)r_{xy}^2}{1-r_{xy}^2} = 144.7.$$

$f_{1,6,0.05} = 5.99$ et $f_{1,6,0.01} = 13.8$.

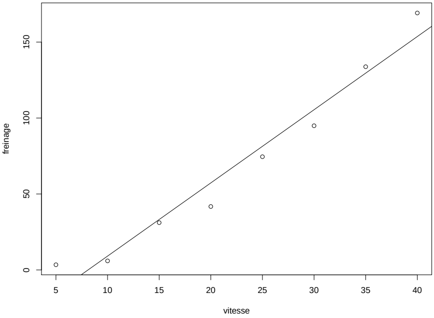
$\Rightarrow$ même au seuil 1%, on est très largement dans la région critique, donc on conclut que la régression linéaire est ici très pertinente.

Remarque : Le nom de "test de pertinence de la régression" est abusif : on teste en fait si, parmi toutes les droites $y = \beta_1 x + \beta_0$, la droite constante $y = \beta_0$ est plausible ou pas.

Etude complète de l'exemple en R

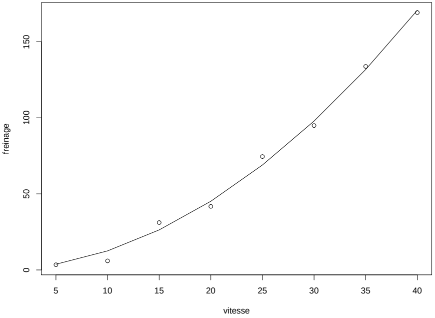
```
> regvf<-lm(freinage~vitesse)
> summary(regvf)

Call:
lm(formula = freinage~vitesse)

Residuals:
     Min       1Q   Median      3Q      Max
 -15.531   -7.766   -2.609   7.048   18.393

Coefficients:

               Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)   -39.0614      10.1113    -3.863    0.00833   **
 vitesse         4.8176       0.4005    12.030      2e-05   ***
--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.98 on 6 degrees of freedom
Multiple R-Squared: 0.9602, Adjusted R-squared: 0.9536
F-statistic: 144.7 on 1 and 6 DF, p-value: 2.002e-05
```

```
> plot(vitesse,freinage)
> lines(vitesse,fitted.values(regvf))
```



Etude du modèle $Y_i = \beta_1 x_i^2 + \beta_0 x_i + \varepsilon_i$

```
> v2<-vitesse^2
> regvf2<-lm(freinage~v2+vitesse-1)
> summary(regvf2)

Call:
lm(formula = freinage~v2+vitesse-1)

Residuals:
     Min        1Q    Median        3Q       Max
 -6.5569   -3.0400   -0.9151   2.7337   5.5614

Coefficients:
          Estimate   Std. Error   t value   Pr(>|t|)
 v2       0.100497     0.007826    12.842   1.37e-05   ***
 vitesse  0.246712     0.256589     0.962      0.373
--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.54 on 6 degrees of freedom
Multiple R-Squared: 0.9981, Adjusted R-squared: 0.9974
F-statistic: 1545 on 2 and 6 DF, p-value: 7.275e-09
```


freinage
0
50
100
150
5
10
15
20
25
30
35
40
vitesse

Exemple de régression polynomiale d'ordre 3