Inférence statistique

Séance 9

Régression linéaire

Frédéric Sur

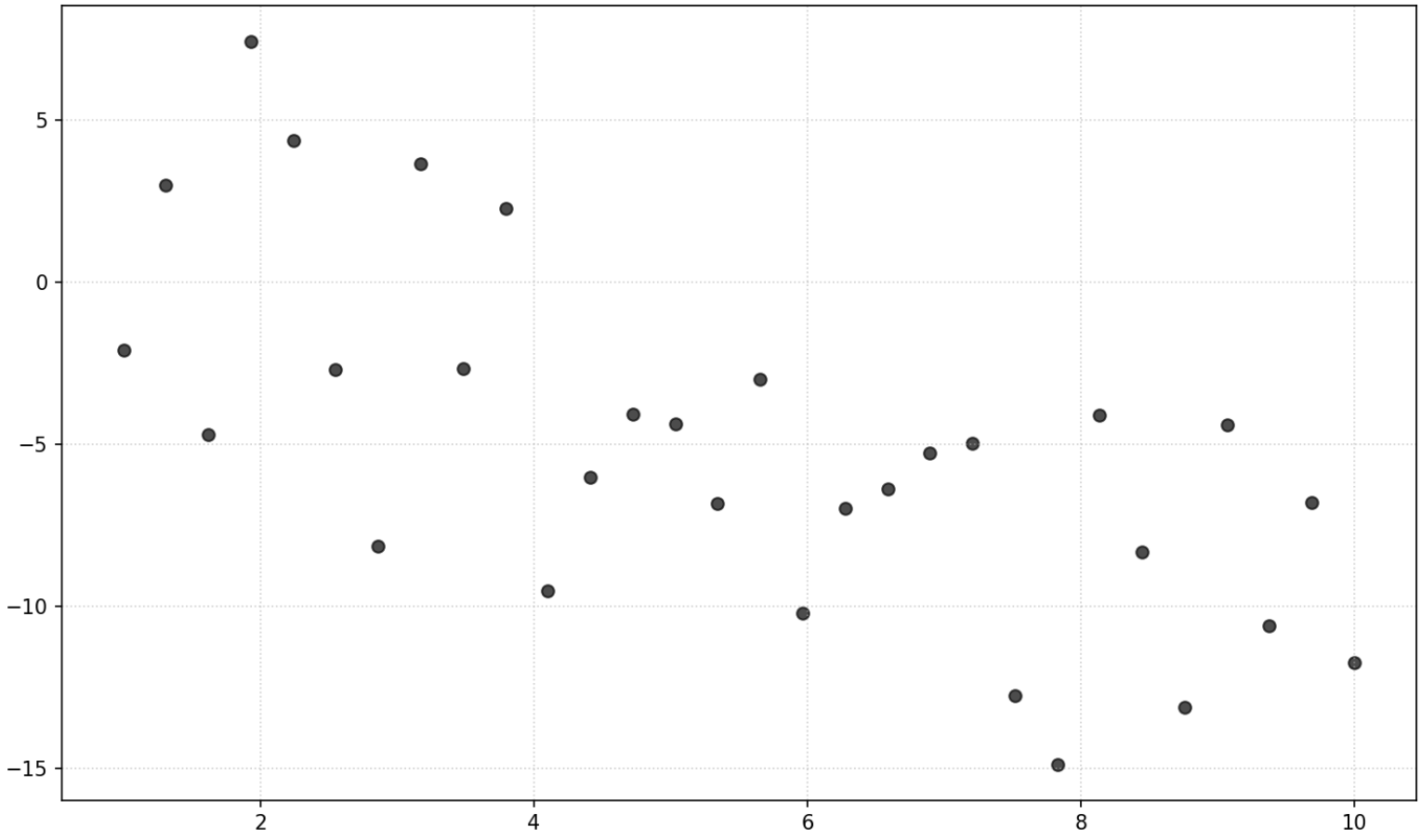
Mines Nancy

https://members.loria.fr/FSur/

Plan

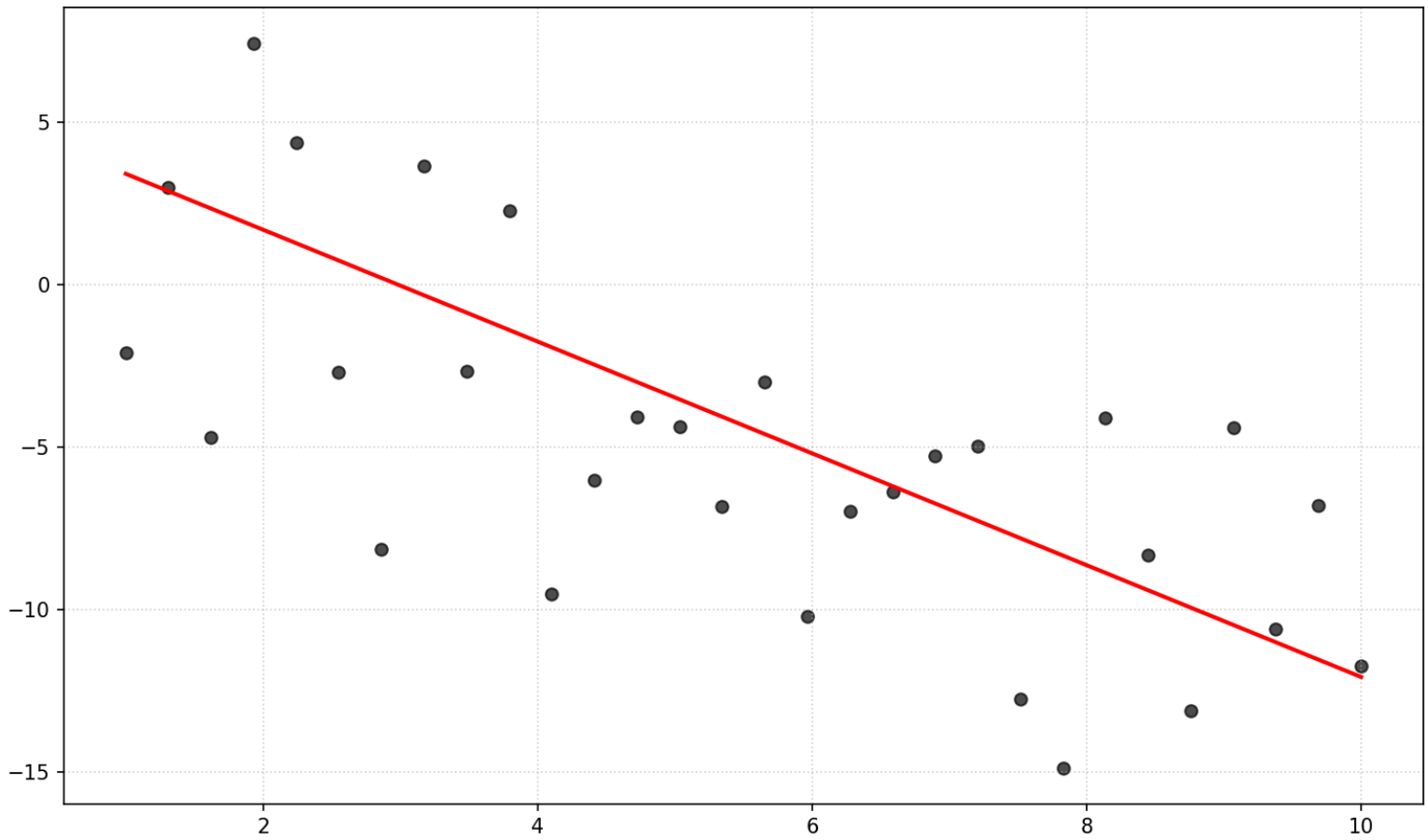
1. Introduction et formulation du modèle
2. Propriétés statistiques
3. Diagnostics
4. Conclusion

Exemple



Modélisation ?
Prédiction ?

Exemple



→ Relation linéaire en y et x ?

Est-elle *statistiquement significative* ?

→ y_0 pour x_0 ? Incertitude sur la prédiction ?

Modélisation mathématique

Objectif de la régression simple

Expliquer ou **prédire** une variable quantitative dépendante Y à l'aide d'une unique variable explicative x à travers une relation linéaire.

Pour chaque observation (x_i, Y_i), $i \in \{1, \ldots, n\}$:

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

- Y_i : variable à expliquer (endogène / dépendante / cible)
- x_i : variable explicative (exogène / indépendante / régresseur)
- β_0 : ordonnée à l'origine (*intercept*)
- β_1 : pente de la droite de régression (*slope*)
- ε_i : terme d'erreur ou perturbation **aléatoire**

Linéarité par rapport aux paramètres β_0 et β_1

Les hypothèses fondamentales (Gauss-Markov)

On suppose :

1. **H1** (erreur nulle en moyenne) :

$$\forall 1 \leqslant i \leqslant n,\ E(\varepsilon_i) = 0$$

2. **H2** (homoscédasticité) :

$$\forall 1 \leqslant i \leqslant n,\ \mathrm{Var}(\varepsilon_i) = \sigma^2$$

3. **H3** (absence d'autocorrélation) :

$$\forall 1 \leqslant i \neq j \leqslant n,\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$$

Estimation par la méthode des MCO

Principe des mondres carrés ordinaires (MCO)

Trouver les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ qui minimisent la somme des carrés des résidus (SCR).

Soit le résidu estimé : $e_i = Y_i - \hat{y}_i = Y_i - (\beta_0 + \beta_1 x_i)$.

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

Conditions du premier ordre : (nécessaires et suffisantes ici)

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (Y_i - \beta_0 - \beta_1 x_i) = 0$$

Solutions des MCO

En résolvant le système, on obtient les estimateurs (des v.a.) :

Estimateur de la pente $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \ll = \gg \quad \frac{\text{Cov}(x, Y)}{\text{Var}(x)}$$

Attention : x n'est pas aléatoire dans ce cours

Estimateur de l'ordonnée à l'origine $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

où $\bar{x}$ et $\bar{Y}$ sont les moyennes empiriques.

Droite de régression : $y = \beta_0' + \beta_1' x$ pour β_0' et β_1' des réalisations de $\hat{\beta}_0$ et $\hat{\beta}_1$

$\rightarrow$ la droite de régression passe toujours par le point moyen $(\bar{x}, \bar{y})$

Plan

1. Introduction et formulation du modèle
2. Propriétés statistiques
3. Diagnostics
4. Conclusion

Propriétés statistiques de $\hat{\beta}_1$ et $\hat{\beta}_0$

Sous les hypothèses de Gauss-Markov :

- **Absence de biais** : $E(\hat{\beta}_1) = \beta_1$ et $E(\hat{\beta}_0) = \beta_0$
- **Variances des estimateurs** :

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \text{ et } \text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]$$

(estimateurs convergents sauf cas pathologique)

- **Estimateur sans biais de σ^2** :

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\text{SCR}}{n-2}$$

Théorème de Gauss-Markov

$(\hat{\beta}_0, \hat{\beta}_1)$ est le **BLUE** (*Best Linear Unbiased Estimator*)
$\rightarrow$ estimateur linéaire sans biais de variance minimale

Hypothèse de normalité

H4 (normalité)

$$\forall\, 1 \leqslant i \leqslant n,\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$\rightarrow$ hypothèse qui permet de déterminer des *intervalles de confiance* et de faire des *tests statistiques*.

Conséquence immédiate :

$$\begin{aligned}
\hat{\beta}_0 &\sim \mathcal{N}\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right) \\
\hat{\beta}_1 &\sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)
\end{aligned}$$

Test sur la pente β_1

On cherche souvent à tester si x a un impact *significatif* sur Y.

$$\mathcal{H}_0 : \beta_1 = 0 \quad \text{contre} \quad \mathcal{H}_1 : \beta_1 \neq 0$$

Sous **H4**, la statistique de test est :

$$t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} \sim T(n-2) \quad \text{(loi de Student à } n-2 \text{ d.d.l.)}$$

où $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\text{Var}(\hat{\beta}_1)}$

Règle de décision (au risquel $\alpha = 5\%$)

Si $|t| > t_{\text{critique}}$ (test bilatéral), on rejette $\mathcal{H}_0$
$\rightarrow$ la relation est statistiquement significative

Intervalles de confiance des coefficients

Intervalles de confiance des coefficients

L'intervalle de confiance à $(1-\alpha)$ pour un paramètre β_k ($k \in \{0,1\}$) est donné par :

$$\text{IC}_{1-\alpha}(\beta_k) = \left[\hat{\beta}_k - t^{n-2}_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_k} \; ; \; \hat{\beta}_k + t^{n-2}_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_k}\right]$$

Interprétation : si cet intervalle contient la valeur 0, cela équivaut à ne pas rejeter l'hypothèse nulle $\beta_k = 0$ au risque α.

Intervalles de confiance des prédictions

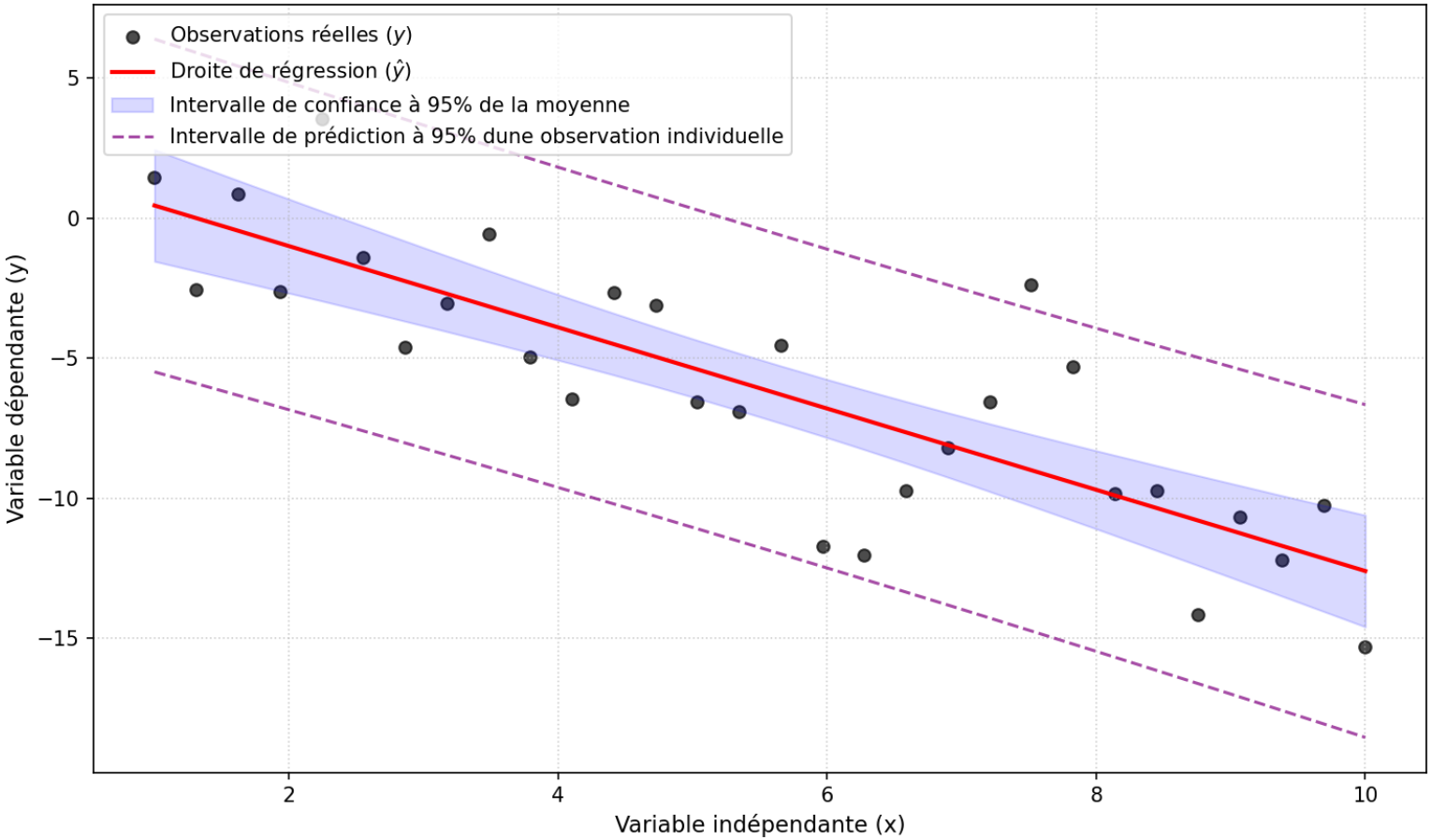
Intervalle de confiance d'un point de la « vraie » droite de régression

$$IP_{1-\alpha}(\bar{Y}_0) = \left[\hat{Y}_0 \pm t^{n-2}_{1-\alpha/2} \cdot s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

Intervalle de confiance d'une observation

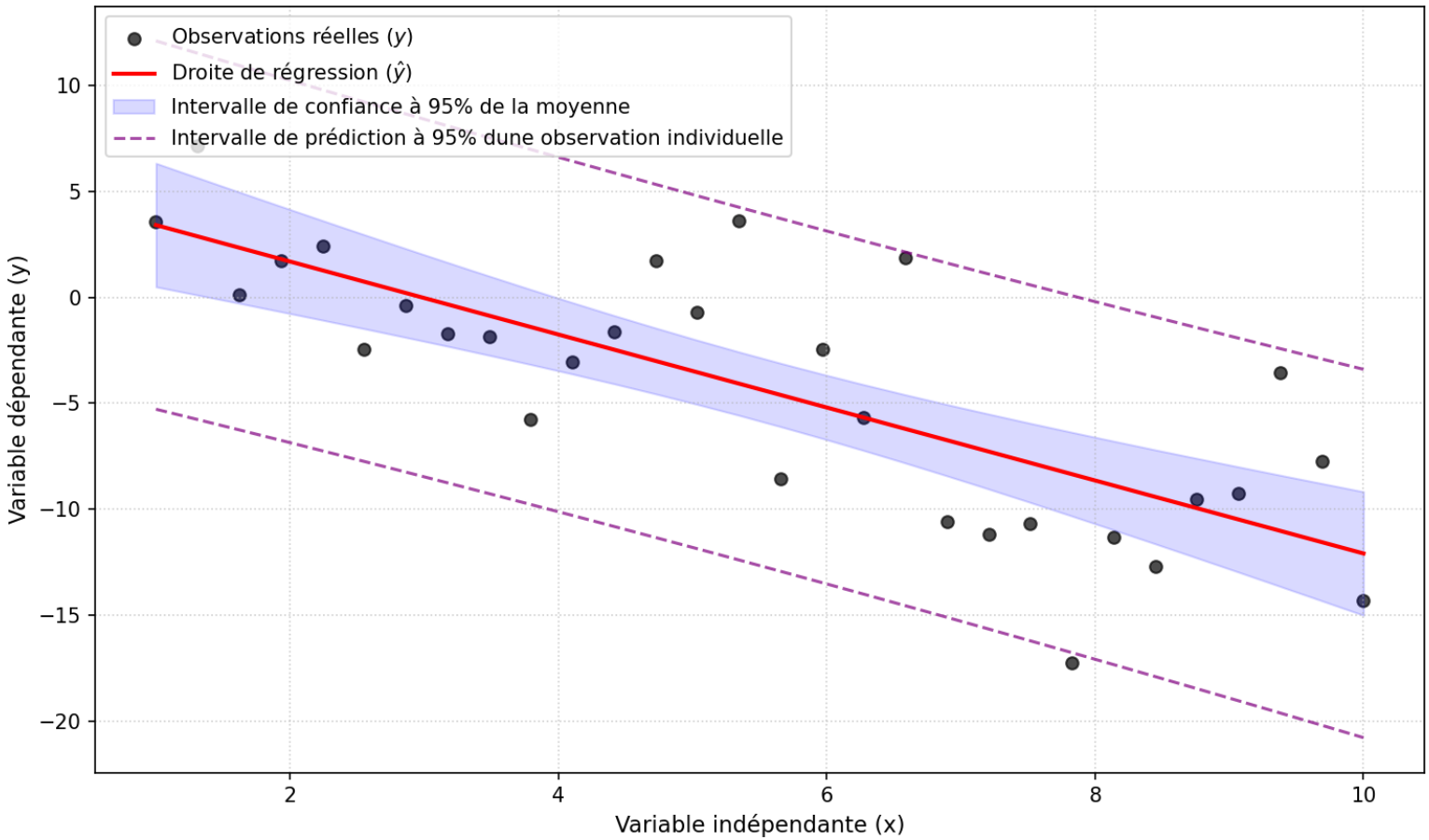
$$IP_{1-\alpha}(Y_0) = \left[\hat{Y}_0 \pm t^{n-2}_{1-\alpha/2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

Illustration sur différentes réalisations



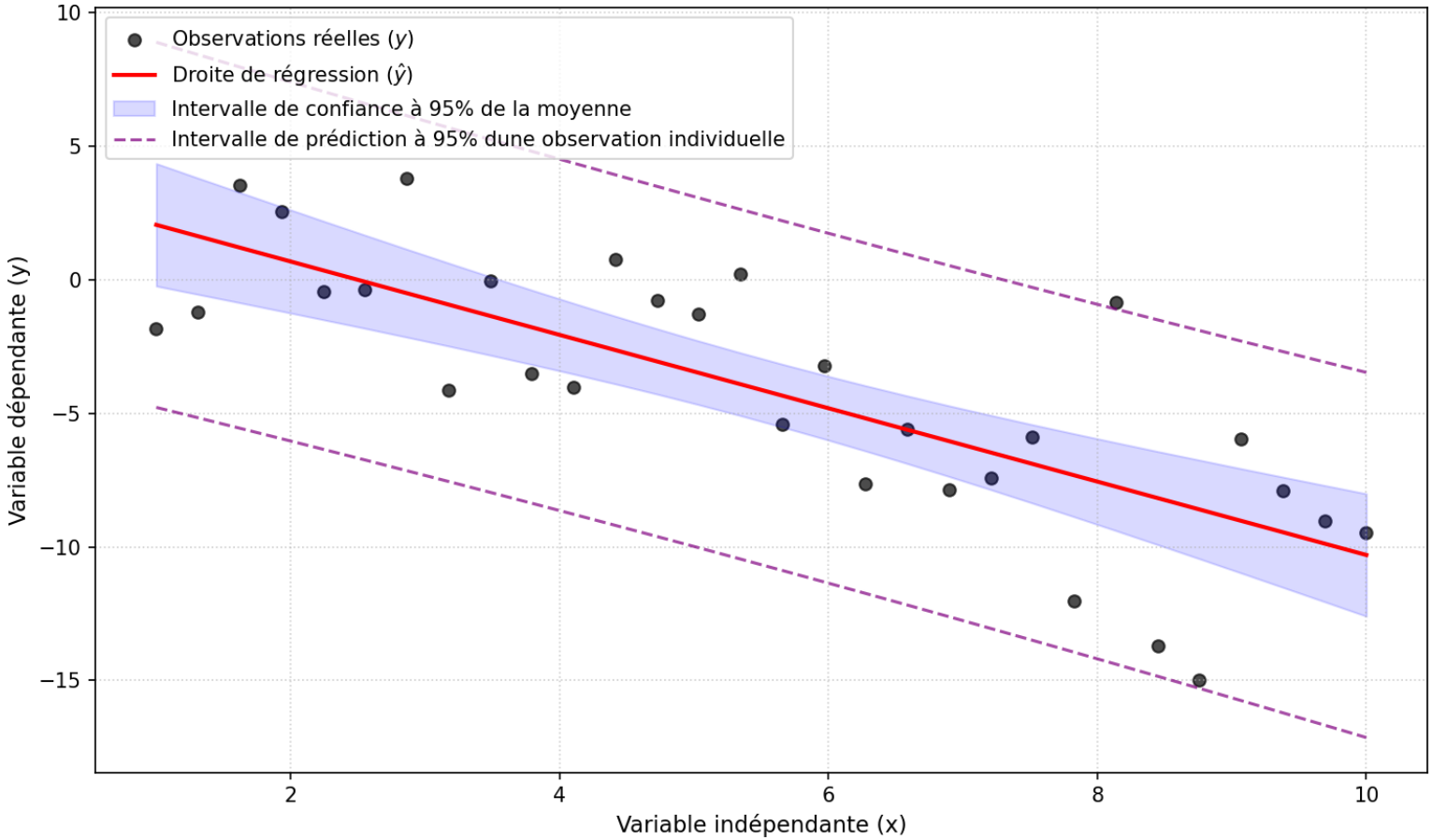
y=1,91 - 1,14 x

Illustration sur différentes réalisations



y=5,13 - 1,72 x

Illustration sur différentes réalisations



y=2,94 - 1,41 x

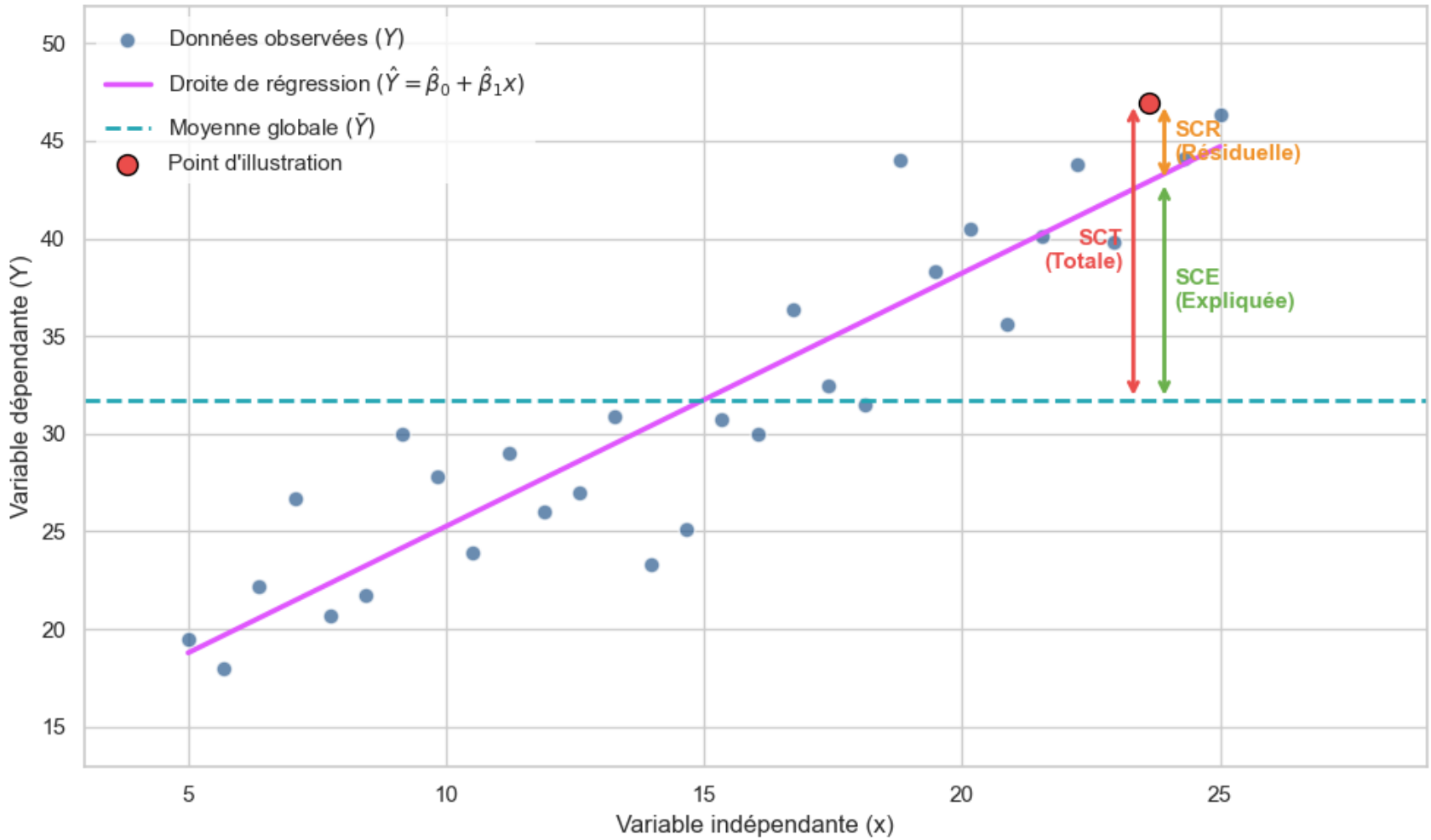
Plan

1. Introduction et formulation du modèle
2. Propriétés statistiques
3. Diagnostics
4. Conclusion

Décomposition de la variance

La variabilité totale de Y se décompose en deux Sommes des Carrés :

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{\text{SCT (SC totale)}} = \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{\text{SCE (SC expliquée)}} + \underbrace{\sum e_i^2}_{\text{SCR (SC résiduelle)}}$$



Test de Fisher-Snedecor sur l'analyse de la variance

Sous l'hypothèse $\beta_1 = 0$:

$$F = \frac{SCE}{s^2} = \frac{SCE/1}{SCR/(n-2)} \sim F(1, n-2)$$

On démontre (simple manipulation algébrique) que $F = t^2_{\beta_1}$
Or (d'après la définition), si $t \sim T(n-2)$, alors $t^2 \sim F(1, n-2)$

Conséquence : les deux tests sont équivalents
$\rightarrow$ l'ANOVA n'apporte pas d'information
(voir exercices Python : on ne parlera pas de la statistique F affichée)

Remarque : vrai dans le cas de la régression linéaire **simple**

Coefficient de détermination R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{SCR}{SCT}$$

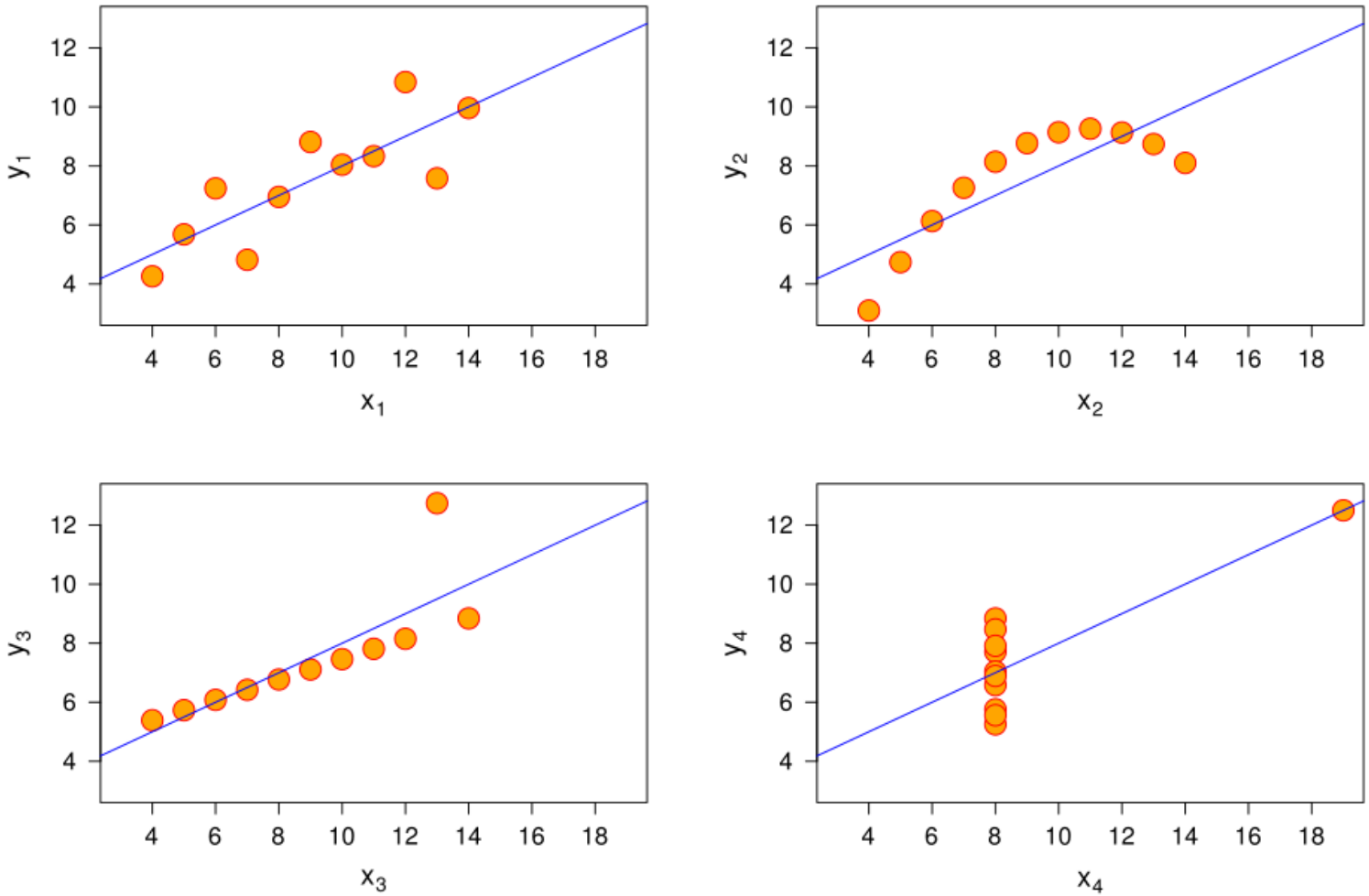
Dans le cas de la **régression linéaire** :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{SCE}{SCT}$$

$\rightarrow$ dans ce cas, R^2 compris entre 0 et 1.

Plus R^2 est proche de 1, meilleur est l'ajustement global du modèle aux données.
Cas où $R^2 = 1$: ajustement linéaire parfait

Le quartet d'Anscombe

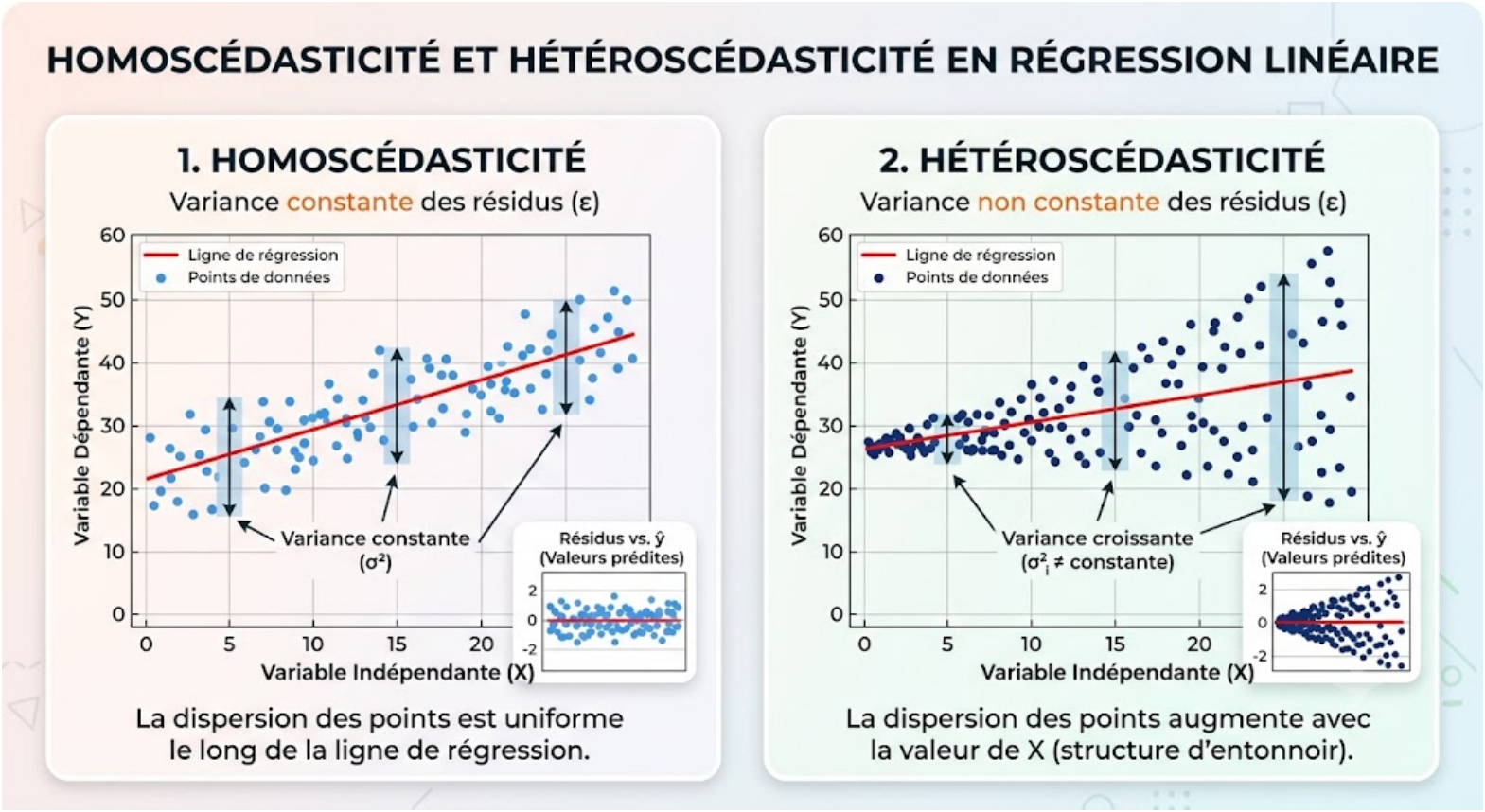


$R^2 = (0,816)^2 \ldots$

Conséquence : toujours faire des graphiques

Source : `https://en.wikipedia.org/wiki/Anscombe's_quartet`

Validation des hypothèses de Gauss-Markov



Source : Nano Banana 2

(entonnoir ou autre forme)

Normalité des résidus : graphiques (QQ plot, histogramme) et tests statistiques

Voir exercices Python

Plan

1. Introduction et formulation du modèle
2. Propriétés statistiques
3. Diagnostics
4. Conclusion

Conclusion

Modèle de la régression linéaire :

- **Avantages**
  - interprétation géométrique claire
  - validité statistique adossée aux hypothèses de Gauss-Markov
- **Limites et perspectives**
  - risque d'omission de variables pertinentes
    $\rightarrow$ nécessité du modèle de **régression multiple**
  - hypothèse de linéarité parfois restrictive
    $\rightarrow$ **modèles non-linéaires** ou **transformations des variables**
  - danger de la confusion entre **corrélation** et **causalité**
    $\rightarrow$ `https://tylervigen.com/spurious-correlations`

Examen

Lundi 8 juin matin

→ durée : 2h30

→ examen ≪ papier ≫

→ horaires et répartition dans les salles à venir sur Arche

Autorisé

→ une feuille A4 recto-verso manuscrite (taille d'écriture raisonnable)

→ fourni dans le sujet : un formulaire contenant les éventuelles formules dont vous pourriez avoir besoin

Interdit

→ tout le reste

→ tout appareil électronique