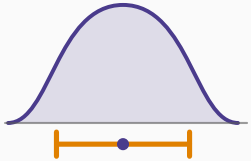
Statistique inférentielle

Chapitre 20 : La régression devient un modèle



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Plan du chapitre

Le changement de statut

Ce que la statistique descriptive en faisait

Un résumé, et rien de plus

Le chapitre 14 du cours de descriptive calculait la droite qui minimise la somme des carrés des écarts :

$$\hat{y} = 4\,347{,}23 + 277{,}17\,x$$

C'était un **résumé du nuage** — une façon commode de décrire cinq cents points par deux nombres. Aucune population n'était en jeu.

Une relation vraie, et son estimation

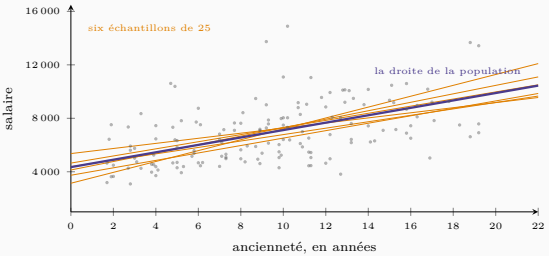
On pose qu'il existe une relation **vraie**, dans la population, dont le nuage observé n'est qu'une réalisation :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0 et β_1 sont des **paramètres** : fixes, inconnus. La droite calculée n'en est qu'une **estimation**.

La droite change avec l'échantillon

la droite d'ajustement change avec l'échantillon — la « vraie » droite, elle, ne bouge pas



Le dernier pas du cursus

La statistique descriptive s'arrêtait à la droite d'ajustement : un résumé du nuage, rien de plus.

L'inférence change son statut. On pose qu'il existe une relation vraie

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

et que la droite calculée n'en est qu'une *estimation*, tirée d'un échantillon.

$\widehat{\beta}_1$ est donc une variable aléatoire, avec une espérance, un écart-type et une loi — exactement comme $\overline{X}$ au chapitre 3.

Tout ce qui suit — intervalle de confiance, test de significativité — n'est que l'application des chapitres 9 à 16 à ce nouveau paramètre.

Le terme d'erreur

Ce que le terme d'erreur représente

Tout ce qui influence le salaire sans figurer dans le modèle : le poste, la performance, la négociation, la chance.

C'est l'écart entre la vraie droite et l'observation individuelle — **pas** l'écart à la droite estimée, qui s'appelle le *résidu*.

Deux mots à ne pas confondre

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i) \qquad e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

L'**erreur** est inconnue, comme les paramètres. Le **résidu** est calculé, comme la droite. La même distinction que m et $\overline{x}$, appliquée à chaque point.

Les hypothèses du modèle

Ce qu'il faut supposer pour aller plus loin

Les quatre hypothèses

1. **Linéarité** : la relation vraie est bien une droite.
2. **Espérance nulle** : $E(\varepsilon_i) = 0$ — les erreurs ne penchent pas d'un côté.
3. **Homoscédasticité** : $V(\varepsilon_i) = \sigma_\varepsilon^2$, la même pour tous.
4. **Indépendance** : les erreurs ne sont pas liées entre elles.

Elles ne sont pas décoratives

Sans la deuxième, l'estimateur est biaisé. Sans la troisième et la quatrième, il reste sans biais mais son **écart-type est faux** — donc tous les tests et intervalles qui suivront sont faux.

C'est le point de départ de tout un cours : l'économétrie consiste largement à détecter la violation de ces hypothèses et à y remédier.

Une cinquième, pour les petits échantillons

La normalité des erreurs

Si les ε_i sont normales, tous les résultats du chapitre 21 sont **exacts**, quel que soit n.

Sinon, ils restent valables **approximativement** pour n grand — par le théorème central limite, une fois de plus.

Ce que cela signifie en pratique

Sur cinq cents observations, la normalité des erreurs n'a aucune importance.

Sur quinze, elle est décisive — et il faut regarder l'histogramme des résidus avant de publier le moindre test.

Les estimateurs

Les formules, inchangées

Les moindres carrés ordinaires

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{V(x)} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \, \overline{x}$$

Ce sont exactement les formules du cours de descriptive. **Le calcul ne change pas ; c'est son interprétation qui change.**

Sur nos données

$$\text{cov} = 9\,859{,}9 \qquad V(x) = 35{,}574$$

$$\hat{\beta}_1 = \frac{9\,859{,}9}{35{,}574} = 277{,}17 \qquad \hat{\beta}_0 = 7\,158{,}26 - 277{,}17 \times 10{,}142 = 4\,347{,}23$$

Le théorème de Gauss-Markov

Sous les quatre hypothèses, les estimateurs des moindres carrés sont **sans biais** et de **variance minimale** parmi tous les estimateurs linéaires sans biais.

$$E(\hat{\beta}_1) = \beta_1 \qquad V(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \overline{x})^2}$$

C'est le vocabulaire du chapitre 7

Sans biais, efficace : les deux qualités demandées à un estimateur, démontrées ici pour la pente.

Gauss-Markov n'introduit rien de neuf. Il applique à $\hat{\beta}_1$ les critères que le chapitre 7 avait posés pour $\overline{X}$.

Lire la formule de la variance

Trois enseignements pratiques

$$V(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \overline{x})^2}$$

- Moins de **bruit** (σ_ε petit) : pente mieux estimée.
- Plus d'**observations** : le dénominateur grandit, la variance baisse.
- Des x plus **étalés** : le dénominateur grandit aussi.

Le troisième point est le moins connu et le plus utile

Pour estimer une pente, mieux vaut vingt observations bien réparties de 1 à 20 ans d'ancienneté que cent observations toutes comprises entre 9 et 11 ans.

On mesure une pente d'autant mieux que le levier est long. C'est le principe des plans d'expérience — et une raison de plus de ne pas se contenter d'un échantillon de convenance.

Ce qu’il faut retenir

Le résumé du chapitre

Six points

1. La descriptive **résumait** un nuage ; l'inférence pose une relation **vraie** dans la population.
2. $Y = \beta_0 + \beta_1 X + \varepsilon$: les β sont des paramètres, la droite calculée est une estimation.
3. **Erreur** ε : inconnue. **Résidu** e : calculé. Comme m et $\overline{x}$.
4. Quatre hypothèses : linéarité, espérance nulle, variance constante, indépendance.
5. Les formules des moindres carrés sont **inchangées** : $\widehat{\beta}_1 = \text{cov}(x, y)/V(x) = 277{,}17$.
6. **Gauss-Markov** : sans biais et de variance minimale — les critères du chapitre 7, appliqués à la pente.

La suite

$\widehat{\beta}_1$ est une variable aléatoire dotée d'une loi. Tout ce que le cours a construit — intervalle, test, p-value — s'y applique donc sans une formule nouvelle.

Le chapitre 21 le fait, et l'on y verra apparaître, dans l'ordre, tous les nombres qu'affiche un logiciel d'économétrie.