

Corrigé du chapitre 16

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 16 de *Comprendre les données sociales*. Le corrigé des exercices situés à l'intérieur de chaque section se trouve dans le livre lui-même, immédiatement après les exercices auxquels il répond.

Faites l'exercice avant de lire le corrigé. Un corrigé lu en premier n'est qu'un texte de plus ; un corrigé lu en second est une vérification de votre raisonnement, et seul le second enseigne quelque chose.

Le corrigé complet, pour tous les chapitres, se trouve à <https://www.markpara-dispolitics.com/books/comprendre-les-donnees-sociales/online/solutions/>.

Chapitre 16 La régression linéaire simple

Corrigé de l'exercice 16.E1. (a) La statistique est la pente sur son erreur type :

$$t = \frac{0,42}{0,15} = 2,80$$

avec $df = n - 2 = 1\,998$. Deux degrés de liberté sont dépensés, un pour l'ordonnée à l'origine et un pour la pente. Le résultat est significatif, $p = 0,005$.

(b) La valeur critique à ce nombre de degrés de liberté vaut essentiellement 1,960, donc

$$0,42 \pm 1,960 \times 0,15 = 0,42 \pm 0,29 = [0,13, 0,71]$$

(c) Chez ces 1 200 adultes, chaque année de scolarité supplémentaire va de pair avec 0,42 heures de bénévolat de plus par mois, avec un intervalle à 95 pour cent allant de 0,13 à 0,71 heures.

Les quatre parties sont la quantité (0,42 heures), le résultat (le bénévolat par mois), l'unité du prédicteur (une année de scolarité) et la population (ces 1 200 adultes). Laisser tomber la dernière invite la lectrice à appliquer le constat à n'importe qui.

(d) Sans le r^2 , vous ne pouvez pas dire quelle part de la variation du bénévolat le modèle explique, et vous ne pouvez pas retrouver l'erreur type de l'estimation — de sorte que vous ne pouvez pas dire de combien les prédictions du modèle se trompent habituellement. La pente et son intervalle décrivent la relation ; ni l'une ni l'autre ne décrit l'ajustement.

Corrigé de l'exercice 16.E2. (a) L'étude B. L'erreur type vaut $s_e / \sqrt{\sum (x_i - \bar{x})^2}$, et un seul quartier a bien moins de variation de densité qu'une province entière. À n égal et à dispersion semblable, le plus petit dénominateur donne la plus grande erreur type — l'effet mesuré à la §16.3.1, où resserrer l'étendue des âges a multiplié l'erreur type par 3,5.

(b) L'homogénéité aide contre la confusion et nuit à la précision, et il faut peser les deux plutôt que d'affirmer l'une. Les cas de l'étude B se ressemblent sur ce qu'un quartier détermine, ce qui écarte effectivement certaines explications rivales. Mais une pente s'estime à partir de la variation du prédicteur, et l'étude B en a retiré l'essentiel. Leur échantillon est mieux placé pour répondre à une question à laquelle il s'est lui-même rendu incapable de répondre avec précision.

(c) Cela confond le fait de ne pas détecter une relation avec celui d'établir qu'il n'y en a pas. Le devis de l'étude B garantit une grande erreur type, de sorte qu'un intervalle large autour de zéro était l'issue probable, qu'une relation existe ou non. Le constat rapportable est l'intervalle : s'il est étroit, l'étude a borné la relation ; s'il

est large, elle n'a rien établi dans un sens ni dans l'autre.

Corrigé de l'exercice 16.E3. (a)

$$s_e \approx 12,0 \times \sqrt{1 - 0,09} = 12,0 \times 0,954 = 11,45 \text{ points}$$

(b) Deviner la moyenne pour chaque élève se trompe d'environ 12,0 points ; employer la régression et le sommeil de l'élève se trompe de 11,45. Le gain est de 0,55 points, soit 4,6 pour cent.

(c) Cela dépend de l'affirmation. Si « appréciable » veut dire que la relation est réelle et mérite une phrase, un r^2 de 0,09 à une taille d'échantillon convenable peut l'appuyer. Si cela veut dire que le modèle prédit utilement le résultat d'un élève, non : la prédiction typique se trompe encore de 11,45 points.

Ce qui trancherait la question, c'est la pente dans les unités des deux variables, avec son intervalle. Combien de points vaut une heure de sommeil ? C'est la question vers laquelle « appréciable » fait signe, et le r^2 ne peut pas y répondre.

Corrigé de l'exercice 16.E4. (a) *La linéarité.* La vraie relation est courbe et on y a ajusté une droite. Cela endommage l'*estimation* : la pente ne décrit ni l'une ni l'autre extrémité de l'étendue. Remède : transformer le prédicteur (§16.5.3) ou ajuster une courbe.

(b) *La variance constante.* La dispersion croît avec la valeur ajustée. Cela endommage l'*erreur type*, qui emploie un seul nombre pour une quantité qui varie. La pente elle-même n'est pas touchée. Remède : des erreurs types robustes, ou une transformation du résultat.

(c) *L'indépendance.* Des cas consécutifs qui se ressemblent signifient que les erreurs sont liées, ce qui arrive habituellement quand les données sont ordonnées dans le temps ou regroupées en grappes. Cela endommage l'*erreur type*, et presque toujours vers le bas, de sorte que l'analyse déclare plus de confiance qu'elle n'en a méritée. Remède : une méthode qui modélise le regroupement — au-delà de ce livre, et la raison pour laquelle `cma_panel` est absent de ce chapitre.

(d) *Rien n'est menacé.* Une asymétrie à droite des résidus viole la normalité, et à $n = 900$ le théorème central limite rend la distribution d'échantillonnage de b_1 proche de la normale de toute façon. Ni l'estimation ni l'erreur type ne sont endommagées. Remède : aucun n'est nécessaire. L'asymétrie peut tout de même mériter une phrase, parce qu'elle dit quelque chose du résultat — un plancher et pas de plafond — même si elle ne dit rien de l'arithmétique.

Corrigé de l'exercice 16.E5. (a) Une hausse de un dans un logarithme naturel est une multiplication par e , soit environ 2,7 — donc une ville environ deux fois et trois quarts plus grande, et non une ville comptant une personne de plus.

(b)

$$3,8 \times \ln 2 = 3,8 \times 0,693 = 2,63$$

Doubler la population d'une ville va de pair avec environ 2,63 lieux culturels de plus par cent mille habitants.

(c) Une ville de un million contre une ville de dix mille, c'est une différence d'un facteur cent, soit 2 paliers d'un facteur dix :

$$3,8 \times \ln 100 = 2 \times 8,75 = 17,50$$

lieux par cent mille habitants.

(d) Parce que la population des villes a un plancher, pas de plafond et une très longue queue à droite, de sorte qu'une poignée de grandes villes porterait autrement tout le levier (§16.5.2) et que la relation a peu de chances d'être droite sur l'échelle brute. La raison de fond est plus forte que l'une ou l'autre : ce qui compte vraisemblablement pour l'offre culturelle, c'est qu'une ville soit *deux fois* plus grande qu'une autre, et non qu'elle compte dix mille personnes de plus, et le modèle logarithmique est celui qui pose cette question-là.

Corrigé de l'exercice 16.E6. (a) Elle doit être appréciable. Un contrôle ne peut prendre à une pente que ce qu'il partage avec le prédicteur (§16.6.1), de sorte qu'une baisse de 66 pour cent signifie que le temps d'écran et le revenu familial se recouvrent fortement — les ménages à faible revenu déclarent plus de temps d'écran.

(b) Le temps d'écran et le revenu familial sont deux indicateurs d'à peu près la même situation, et c'est le revenu qui fait le travail. Ce qui ressemblait à un effet du temps d'écran était le reflet du revenu du ménage depuis le début, et le contrôle l'a révélé.

(c) Un faible revenu familial cause un temps d'écran plus élevé — moins de solutions de rechange, moins d'activités encadrées, des logements plus petits — et le temps d'écran agit ensuite sur le bien-être. Le revenu est une étape sur le chemin qui va de la situation au résultat, de sorte que contrôler pour lui retire une partie du mécanisme par lequel le temps d'écran opère. Le temps d'écran a compté, et l'analyse a soustrait la raison pour laquelle il comptait.

(d) Rien dans les données. Les deux lectures produisent des coefficients identiques, et un échantillon plus grand produirait les mêmes coefficients identiques avec plus de précision. Les départager exige un argument sur l'ordre causal — idéalement un devis dans lequel le temps d'écran varie pour des raisons sans lien avec le revenu (chapitre 3).

Corrigé de l'exercice 16.E7. (a) Les réponses dépendront du téléchargement, mais la régression de `infant_mort` sur `educ_years` sur les 52 pays de la vague 1 donne

une pente nettement négative : plus de scolarité va de pair avec moins de décès infantiles. Rapportez-la en décès par millier de naissances vivantes par année de scolarité, avec son intervalle, le r^2 et l'erreur type de l'estimation en décès par millier.

(b) Oui. La mortalité infantile se comporte comme le PIB par habitant à l'envers — elle chute abruptement chez les pays les plus pauvres et s'aplatit chez les plus riches — de sorte qu'une droite tracée à travers l'une ou l'autre montre le même arc dans ses résidus : la droite au-dessus des données aux deux extrémités et en dessous au milieu.

(c) D'après le graphique des résidus et la forme de la variable, non d'après le r^2 . Si les résidus du modèle logarithmique sont informes là où ceux du modèle brut décrivent un arc, c'est une donnée probante sur la spécification plutôt qu'un nombre plus grand. L'argument de fond est le plus fort : la mortalité infantile répond vraisemblablement à des différences proportionnelles de ressources plutôt qu'à des différences absolues, ce qu'encode un prédicteur logarithmique.

(d) Le pays au levier le plus élevé se trouvera à une extrémité de l'étendue du prédicteur — le plus pauvre ou le plus scolarisé, selon le modèle. Savoir s'il est *influent* est une question distincte, et la façon d'y répondre est de réajuster sans lui et de comparer les pentes. Dans ce jeu de données, la réponse est presque certainement non : la §16.5.2 donne la plus grande distance de Cook sur l'ensemble des 52 pays comme valant 0,11.

Corrigé de l'exercice 16.E8. (a)–(b) Les réponses dépendent de l'année de recensement. Une régression du revenu médian après impôt sur la part de diplômés, sur les régions métropolitaines de recensement, donne une pente positive de quelques centaines de dollars par point de pourcentage, avec un intervalle assez étroit : les régions sont assez peu nombreuses pour que l'échantillon soit petit, mais la relation est assez forte pour que l'intervalle exclue confortablement zéro.

(c) Votre r^2 sera bien plus élevé que 0,017 — des valeurs supérieures à 0,5 sont habituelles pour ce couple au niveau des régions.

La raison est l'agrégation. Le revenu médian d'une région métropolitaine est une moyenne sur des centaines de milliers de personnes, et faire une moyenne annule la variation individuelle qui rend 0,017 petit. Ce qui reste est la part systématique, de sorte que les relations au niveau des régions paraissent bien plus nettes que les relations individuelles qui les sous-tendent. Rien dans le monde n'est devenu plus fort ; c'est le bruit qui a été moyenné.

(d) Qu'un r^2 au niveau des régions ne dit à peu près rien de la qualité avec laquelle les mêmes variables prédisent une personne. Conclure d'une relation forte au niveau des régions que les individus scolarisés gagnent proportionnellement plus est le sophisme écologique (§10.7.5), et ce couple de variables en est une des

illustrations classiques.