Chapitre 4 – Estimation par le quotient, par la régression, et par la différence

30. La caractéristique de la population à laquelle on s'intéresse dans une enquête est $\alpha = 1/\mu$, où μ est la moyenne de la population. Dans un EAS de taille $n = 105$, on obtient $\overline{y} = 5.25$ et $s = 0.37$. Dans ce qui suit, nous considérons $\hat{\alpha} = \overline{y}^{-1}$ comme estimateur de α.

    (a) Utiliser un développement en série de Taylor (de deuxième ordre) de $\hat{\alpha}$ autour de $\overline{y} = \mu$ afin d'obtenir une expression approximative du biais de $\hat{\alpha}$ en tant qu'estimateur de α.

    (b) Utiliser un développement en série de Taylor (du premier ordre) de $\hat{\alpha}$ autour de $\overline{y} = \mu$ afin d'obtenir une expression approximative du biais de $\hat{\alpha}$ en tant qu'estimateur de α.

    (c) En supposant que la distribution de $\hat{\alpha}$ suit approximativement une loi normale pour des valeurs de n suffisamment élevées, utiliser le résultat de (b) afin d'obtenir un I.C. de α à environ 95%. [Ignorer le biais de $\hat{\alpha}$ et le facteur de correction de la population finie, en supposant dans ce dernier cas que N est très grand.]

    (d) Trouver un I.C. de α à environ 95% en trouvant d'abord un intervalle analogue pour μ, puis en inversant les bornes. Comparer avec le résultat obtenu en (c).

    Solution: Soient un EAS $\{y_i\}$ de taille n provenant d'une population $\{u_j\}$ de taille N, et l'estimateur

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Dans le contexte EAS, nous savons que $E(\overline{y}) = \mu$ et $V(\overline{y}) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$, où σ^2 represente la variance de la population.

(a) La série de Taylor de deuxième ordre de $f(x) = \frac{1}{x}$ autour de $x = \mu$ est

$$\frac{1}{x} \approx \frac{1}{\mu} - \frac{1}{\mu^2}(x-\mu) + \frac{1}{\mu^3}(x-\mu)^2.$$

Ainsi, $\overline{y}^{-1} \approx \frac{3}{\mu} - \frac{3\overline{y}}{\mu^2} + \frac{\overline{y}^2}{\mu^3}$ et

$$\begin{aligned}\text{Biais}\,(\hat{\alpha}) = E(\hat{\alpha}-\alpha) = E(\hat{\alpha}) - \alpha = E\left(\overline{y}^{-1}\right) - \frac{1}{\mu} &\approx E\left(\frac{3}{\mu} - \frac{3\overline{y}}{\mu^2} + \frac{\overline{y}^2}{\mu^3}\right) - \frac{1}{\mu} \\ &= \frac{3}{\mu} - \frac{3}{\mu^2}E(\overline{y}) + \frac{1}{\mu^3}E\left(\overline{y}^2\right) - \frac{1}{\mu} \\ &= \frac{3}{\mu} - \frac{3}{\mu^2}\mu + \frac{1}{\mu^3}\left[V(\overline{y}) + (E(\overline{y}))^2\right] - \frac{1}{\mu} = \frac{1}{\mu^3}\left[V(\overline{y}) + \mu^2\right] - \frac{1}{\mu} = \frac{1}{\mu^3}V(\overline{y})\end{aligned}$$

L'approximation de deuxième ordre du Biais($\hat{\alpha}$) dans un EAS est alors $\frac{1}{\mu^3}\cdot\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$.

(b) L'approximation de premier ordre de $f(x) = \frac{1}{x}$ autour de $x = \mu$ est

$$\frac{1}{x} \approx \frac{1}{\mu} - \frac{1}{\mu^2}(x-\mu).$$

Ainsi, $\overline{y}^{-1} \approx \frac{2}{\mu} - \frac{\overline{y}}{\mu^2}$ et

$$\begin{aligned}\text{Biais}\,(\hat{\alpha}) = E(\hat{\alpha}-\alpha) = E(\hat{\alpha}) - \alpha = E\left(\overline{y}^{-1}\right) - \frac{1}{\mu} &\approx E\left(\frac{2}{\mu} - \frac{\overline{y}}{\mu^2}\right) - \frac{1}{\mu} \\ &= \frac{2}{\mu} - \frac{1}{\mu^2}E(\overline{y}) - \frac{1}{\mu} = \frac{2}{\mu} - \frac{1}{\mu} - \frac{1}{\mu} = 0\end{aligned}$$

L'approximation de premier ordre du Biais($\hat{\alpha}$) dans un EAS est donc nulle.

(c) Si $\hat{\alpha}$ suit approximativement une loie normale lorsque la taille n est élevée, et si Biais$(\hat{\alpha}) \approx 0$, alors $\hat{\alpha} \pm 2\sqrt{\hat{V}(\hat{\alpha})}$ représente un I.C. de α à environ 95%. On se sert de l'expansion de premier ordre de la partie (b) et l'on obtient

$$V(\hat{\alpha}) = V\left(\overline{y}^{-1}\right) \approx V\left(\frac{2}{\mu} - \frac{\overline{y}}{\mu^2}\right) = V\left(\frac{\overline{y}}{\mu^2}\right) = \frac{1}{\mu^4} V(\overline{y}) = \frac{1}{\mu^4} \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

On constate alors, en ignorant le facteur de correction en population finie, que

$$\hat{V}(\hat{\alpha}) \approx \overline{y}^{-4} \cdot \frac{s^2}{n}\left(1 - \frac{n}{N}\right) \approx \frac{\overline{y}^{-4} s^2}{n}.$$

Conséquemment,

$$\hat{\alpha} \pm 2\sqrt{\hat{V}(\hat{\alpha})} \equiv \overline{y}^{-1} \pm 2 \cdot \frac{\overline{y}^{-2} s}{\sqrt{n}} \equiv \frac{1}{5.25} \pm 2 \cdot \frac{0.37}{(5.25)^2\sqrt{105}} \equiv (0.1878561, 0.1930963)$$

est l'I.C. recherché.

(d) Dans un EAS, si l'on suppose que le FCPF est ≈ 0, on obtient un I.C. de μ à environ 95% à l'aide de

$$\overline{y} \pm 2\sqrt{\hat{V}\left(\overline{y}\right)} \approx \overline{y} \pm 2 \cdot \frac{s}{\sqrt{n}} \equiv 5.25 \pm 2 \cdot \frac{0.37}{\sqrt{105}} \equiv (5.177783, 5.322217).$$

En inversant les bornes, on obtient un autre I.C. de α à environ 95%:

$$\left(\frac{1}{5.322217}, \frac{1}{5.177783}\right) = (0.1878916, 0.1931329).$$

■

31. Notre ami forestier souhaite maintenant donner un estimé du volume total des arbres d'une vente de bois ($N = 250$). Il prélève un EAS (de taille $n = 12$) de ces arbres et enregistre le volume de chaque arbre dans l'échantillon. En outre, il mesure la superficie de la base de chaque arbre marqué pour la vente. Il utilise ensuite un estimateur par le quotient pour le volume total. Soit X la superficie de la base et Y le volume en pieds cubes d'un arbre. Le total de la superficie de la base des 250 arbres est de $\tau_X = 75$ pieds carrés. Il recueille les données suivantes:

Arbre	**Superficie de la base**	**Volume**	**Arbre**	**Superficie de la base**	**Volume**
1	0.3	6	7	0.6	12
2	0.5	9	8	0.5	9
3	0.4	7	9	0.8	20
4	0.9	19	10	0.4	9
5	0.7	15	11	0.8	18
6	0.2	5	12	0.6	13

(a) Obtenir les moyennes et les écarts types de l'échantillon pour la superficie de la base et pour le volume, ainsi qu'une estimé de la corrélation entre les deux variables.

(b) En utilisant les résultats de (a), donner un estimé du volume total des arbres marqués pour la vente en utilisant l'estimation par le quotient, et une marge d'erreur sur l'estimation.

Solution: Soit X la superficie de la base, et Y le volume. On utilise le code `R` suivant:

```
> x=1/10*c(3,5,4,9,7,2,6,5,8,4,8,6)
> y=c(6,9,7,19,15,5,12,9,20,9,18,13)

> mean(x)
> sd(x)

> mean(y)
> sd(y)

> cor(x,y)
```

(a) Ainsi, $\mu_X = 0.55833$, $s_X = 0.21515$, $\overline{y} = 11.83333$, $s_Y = 5.18448$, et $\hat{\rho} = \frac{s_{XY}}{s_X s_Y} = 0.97123$.

(b) L'estimateur par le quotient du total du volume est

$$\hat{\tau}_Y = r\tau_X = \frac{\overline{y}}{\mu_X}\tau_X = \frac{11.83333}{0.55833}(75) = (21.19)(75) = 1589.561281,$$

tandis que la marge d'erreur sur l'estimation est

$$\begin{aligned} B &= 2\sqrt{\hat{V}\left(\hat{\tau}_Y\right)} = 2\sqrt{\hat{V}\left(r\tau_X\right)} = 2\tau_X\sqrt{\hat{V}\left(r\right)} \approx 2\tau_X\sqrt{\frac{s_W^2}{n\mu_X^2}\left(1-\frac{n}{N}\right)} = 2N\sqrt{\frac{s_W^2}{n}\left(1-\frac{n}{N}\right)} \\ &= 2N\sqrt{\frac{s_Y^2 + r^2 s_X^2 - 2r\hat{\rho}s_X s_Y}{n}\left(1-\frac{n}{N}\right)} \\ &= 2(250)\sqrt{\frac{(5.2)^2 + (21.2)^2(0.2)^2 - 2(21.2)(0.97)(0.2)(5.2)}{12}\left(1-\frac{12}{250}\right)} \\ &\approx 186.321 \end{aligned}$$

■

32. On souhaite donner un estimé de la moyenne μ_Y d'une population donnée. Un EAS contient les observations y_i et l'information auxiliaire x_i, $i = 1, \ldots, n$, (la moyenne μ_X de la population est connue). Discuter des mérites relatifs de l'utilisation de:

(a) La moyenne de l'échantillon $\overline{y}$.

(b) L'estimateur par le quotient $\hat{\mu}_{Y;R}$.

(c) L'estimateur par la régression $\hat{\mu}_{Y;L}$.

(d) L'estimateur par la différence $\hat{\mu}_{Y;D}$.

Solution:

(a) La moyenne de l'échantillon $\overline{y}$ représente l'estimateur EAS.

Bénéfices: Facile à calculer. Pas besoin d'informations auxiliaires x. S'il y a des informations auxiliaires x mais que la corrélation avec y est faible, l'estimateur EAS fournit un estimateur plus efficace.

Inconvénients: S'il y a une forte corrélation entre x et y, l'estimateur EAS perd en précision en n'utilisant pas les informations auxiliaires.

(b) L'estimateur par le quotient $\hat{\mu}_{Y;R} = \frac{\overline{y}}{\mu_X}\mu_X$ présume l'existence d'une relation linéaire forte entre x et y, et que la droite de régression passe par l'origine.

Bénéfices: Si les hypothèses se réalisent, l'estimateur du ratio est plus efficace que l'estimateur EAS.

Inconvénients: Si ces hypothèses ne sont pas vérifiées, l'estimateur du ratio peut être inefficace.

(c) L'estimateur par la régression $\hat{\mu}_{Y;L} = a + b\mu_X$ présume l'existence d'une rlation linéaire forte entre x et y, mais la droite de régression ne passe pas nécessairement pas l'origine.

Bénéfices: L'estimateur par la régression est plus efficace que l'estimateur par le quotient, sauf si $b = \frac{\overline{y}}{\mu_X}$, auquel cas ils sont équivalents.

Inconvénients: Si $a \approx 0$, l'estimateur par le quotient devrait être utilisé car il est plus facile à mettre en œuvre. L'estimateur par la régression peut présenter un biais important si la corrélation entre x et y est faible.

(d) L'estimateur par la différence est $\hat{\mu}_{Y;D} = \mu_X + d = \mu_X + (\overline{y} - \mu_X)$.

Bénéfices: Plus facile à calculer que l'estimateur par la régression.

Inconvénients: Peut être moins efficace que l'estimateur par la régression lorsque la corrélation entre x et y n'est pas forte. ■

33. Une société souhaite donner un estimé du revenu total des ventes d'un produit durant une période de trois mois. Pour chacun des $N = 123$ bureaux de district, le total des revenus est disponible durant la période de trois mois correspondante de l'année précédente: $\tau_X = 128,200$. Un EAS de 13 bureaux de district est prélevé parmi les 123 bureaux de la société. Les données résultantes sont présentées dans le tableau ci-dessous.

Bureau i	1	2	3	4	5	6	7	8	9	10	11	12	13
Précédent x_i	550	720	1500	1020	620	980	928	1200	1350	1750	670	729	1530
Actuel y_i	610	780	1600	1030	600	1050	977	1440	1570	2210	980	865	2020

(a) Tracer un graphique de dispersion de y_i en fonction de x_i et appliquer un modèle linéaire simple. Quel estimateur le modèle suggère-t-il? Expliquer.

(b) Utiliser un estimateur par le quotient afin de donner un estimé de la moyenne des revenus actuels μ_Y (par bureau) et donner une marge d'erreur sur l'estimation.

(c) Utiliser un estimateur par le quotient afin de donner un estimé du total τ_Y des revenus actuels (société) et donner une marge d'erreur sur l'estimation.

Solution:

(a) Les sommes et les données qui seront nécessaires pour calculer les paramètres et les coefficients sont les suivantes :

$$\sum_{i=1}^{13} x_i = 13547 \qquad \sum_{i=1}^{13} y_i = 15732 \qquad \sum_{i=1}^{13} x_i y_i = 18748141$$
$$\sum_{i=1}^{13} x_i^2 = 15963525 \qquad \sum_{i=1}^{13} y_i^2 = 22230054.$$

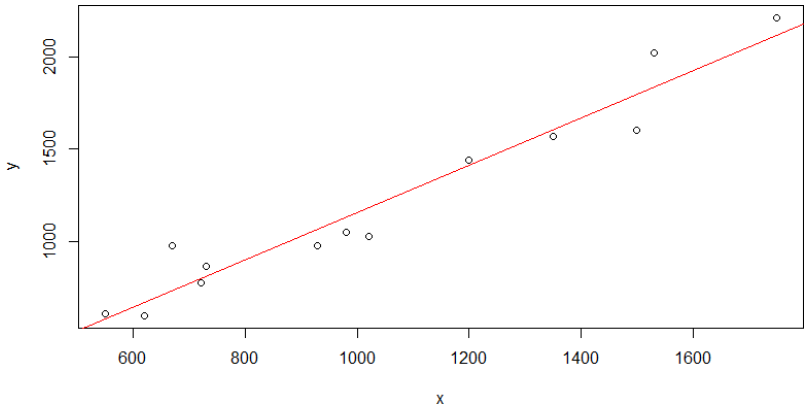
Les moyennes sont alors $\overline{x} = 1042.077$ et $\overline{y} = 1210.154$. Selon les formules, les paramètres de régression de la droite sont les suivants :

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - 13\overline{xy}}{\sum x_i^2 - 13\overline{x}^2} = \frac{18748141 - 13(1042.077)(1210.154)}{15963525 - 13(1042.0769)^2} = 1.2749$$
$$a = \overline{y} - b\overline{x} = 1210.154 - 1.2749(1042.0769) = -118.390,$$

et la droite de régression est tout simplement $\hat{y} = -118.390 + 1.275x$. Le coefficient de corrélation entre y et x est de

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.9697.$$

Le nuage de points et la droite des moindres carrés sont présentés ci-dessous.



L'utilisation de l'estimateur par le quotient est acceptable lorsque $1 \geq \rho \gg 0$ et $a = 0$. Ici, $1 \geq \rho = 0.9697 \gg 0$. À première vue, il semblerait que $a = -118.390 \neq 0$. On peut tester

$$\begin{cases} H_0 : a = 0 \\ H_a : a \neq 0 \end{cases}$$

à l'aide d'un test t bilatéral. D'une part, note that

$$s\{a\} = \sqrt{\text{EQM}}\sqrt{\frac{1}{13} + \frac{\overline{x}}{S_{xx}}} = 107.30798.$$

Selon la théorie, la statistique $t^* = \frac{a}{s\{a\}}$ suit une loi $t(n-2)$ lorsque H_0 est valide. Pour $\alpha = 0.05$, la valeur critique $t(1-\alpha/2; 11) = t(0.975; 11) = 2.200985$ est plus élevée que $|t^*| = \frac{118.390}{107.30798} = 1.10$. Conséquemment, il n'y a pas assez d'évidence pour rejeter H_0 et on considère que l'estimateur par le quotient est acceptable.

(b) L'estimateur de μ_Y par le quotient est

$$\hat{\mu}_{Y;R} = \frac{\overline{y}}{\overline{x}}\mu_X = \frac{1210.154}{1042.0769} \cdot \frac{128200}{123} = 1210.386.$$

L'approximation de la variance de l'estimateur par le quotient est :

$$\begin{aligned} \hat{V}\left(\hat{\mu}_{Y;R}\right) &= \frac{N-n}{Nn}s_r^2 = \frac{N-n}{Nn}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{\overline{y}}{\overline{x}}x_i\right)^2\right] \\ &= \frac{123-13}{123(13)} \cdot \frac{1}{12}214317.9834 = 1228.6313. \end{aligned}$$

On peut alors construire un intervalle de confiance pour μ_Y à environ 95%:

$$1210.386 \pm 2\sqrt{1228.6313} = 1210.386 \pm 70.1.$$

(c) L'estimateur de τ_Y par le quotient est

$$\hat{\tau}_{Y;R} = \frac{\overline{y}}{\overline{x}}\tau_X = \frac{1210.154}{1042.0769} \cdot 128200 \approx 148877.44.$$

L'approximation de la variance de l'estimateur par le quotient est :

$$\hat{V}\left(\hat{\tau}_{Y;R}\right) = N^2\hat{V}\left(\hat{\mu}_{Y;R}\right) = 123^2 \cdot 1228.6313 \approx 18587962.9377.$$

On peut alors construire un intervalle de confiance pour τ_Y à environ 95%:

$$148877.44 \pm 2\sqrt{18587962.9377} = 148877.44 \pm 8622.7520.$$

■

34. Une gestionnaire de ressources forestières souhaite donner un estimé du nombre de sapins morts dans une zone de 400 acres. À l'aide d'une photo aérienne, elle divise la zone en 200 parcelles de 2 acres. Soit x le compte des sapins morts sur la photo et y le compte réel au sol pour un EAS de $n = 10$ parcelles. Le nombre total de sapins morts obtenu à partir du compte photographique est $X = 4300$. Les données résultantes sont présentées dans le tableau ci-dessous.

Parcelle i	1	2	3	4	5	6	7	8	9	10
Compte photo x_i	12	30	24	24	18	30	12	6	36	42
Compte réel y_i	18	42	24	36	24	36	14	10	48	54

(a) Tracer un graphique de dispersion de y_i en fonction de x_i et appliquer un modèle linéaire simple. Quel estimateur le modèle suggère-t-il? Expliquer.

(b) Utiliser un estimateur par le quotient afin de donner un estimé du nombre total τ_Y de sapins mort dans la zone de 400 acres et donner une marge d'erreur sur l'estimation.

(c) Utiliser un estimateur par la régression afin de donner un estimé du nombre total τ_Y de sapins mort dans la zone de 400 acres et donner une marge d'erreur sur l'estimation.

(d) Utiliser un estimateur par la différence afin de donner un estimé du nombre total τ_Y de sapins mort dans la zone de 400 acres et donner une marge d'erreur sur l'estimation.

(e) Quel estimateur est préférable pour ce problème? Expliquer.

Solution:

(a) On utilise le code suivant:

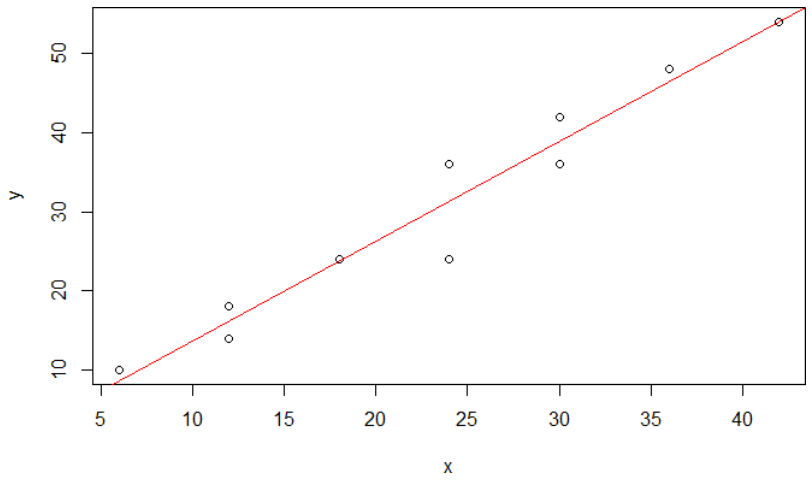
```
> x=c(12,30,24,24,18,30,12,6,36,42)
> y=c(18,42,24,36,24,36,14,10,48,54)
> plot(x,y)
> abline(lm(y ~ x), col = "red")
> summary(lm(y~x))

Residuals:
    Min      1Q  Median      3Q     Max
-7.3556 -1.6884  0.7568  1.7006  4.6444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1307     2.7286   0.414    0.689
x             1.2594     0.1057  11.911 2.27e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.639 on 8 degrees of freedom
Multiple R-squared:  0.9466,Adjusted R-squared:   0.94
F-statistic: 141.9 on 1 and 8 DF,  p-value: 2.269e-06
```

Le nuage de points et la droite de régression se retrouvent plus bas:



La droite de régression est $\hat{y} = 1.1307 + 1.2594x$ et la correlation de x et y dans l'échantillon est

```
> cor(x,y)
[1] 0.9729
```

L'estimateur par le quotient est approprié lorsque $1 \geq \rho \gg 0$ et $a = 0$.
Ici, nous avons $1 \geq \rho = 0.9729 \gg 0$ et $a = 1.1307$ et l'erreur-type de l'ordonnée à l'origine est

$$s\{a\} = 2.7286;$$

la statistique observée $t^* = \frac{a}{s\{a\}}$ suit une distribution $t(n-2) = t(8)$ si la valeur réelle de a est nulle. Si $\alpha = 0.05$, la valeur critique $t(1-\alpha/2; 8) = t(0.975; 8) = 2.306$ est plus élevée que la valeur observée $|t^*| = 1.1307/2.7286 = 0.4144$. Nous n'avons donc pas assez d'évidence afin de rejeter l'hypothèse $a = 0$. Conséquemment, l'estimateur par le quotient est préférable.

(b) Nous calculons aisément que $\overline{x} = 23.4$ et $\overline{y} = 30.6$. L'estimateur par le quotient de τ_Y est

$$\hat{\tau}_{Y;R} = \frac{\overline{y}}{\overline{x}}\tau_X = \frac{30.6}{23.4}4300 = 5623.077,$$

et sa variance approximative est

$$\hat{V}\left(\hat{\tau}_{Y;R}\right) = \hat{V}\left(N\hat{\mu}_{Y;R}\right) = N^2\left(\frac{N-n}{Nn}\right)s_r^2 = N^2\left(\frac{N-n}{Nn}\right)\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{\overline{y}}{\overline{x}}x_i\right)^2\right]$$
$$= (200)^2\frac{200-10}{200(10)}\cdot\frac{1}{9}(12.08) \approx 45890.$$

On forme alors un intervalle de confiance à environ 95% de τ_Y à l'aide de $5623.077 \pm 2\sqrt{45890} \approx 5623.01 \pm 428.44$.

(c) L'estimateur par le régression est à propos lorsqu'il y a une forte corrélation linéaire entre x et y: puisque le coefficient de corrélation est $\rho = 0.9727$, c'est une hypothèse adéquate.

L'estimateur du total τ_Y par la régression est

$$\hat{\tau}_{Y;L} = N(a + b\mu_X) = Na + b\tau_X = 200(1.1307) + 1.2594(4300) = 5641.56,$$

et sa variance approximative est

$$\hat{V}\left(\hat{\tau}_{Y;L}\right) = \hat{V}\left(N\hat{\mu}_{Y;L}\right) = N^2\left(\frac{N-n}{Nn}\right)\text{EQM} = N^2\left(\frac{N-n}{Nn}\right)\left[\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2 - b\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right]$$
$$= (200)^2\frac{200-10}{200(10)}\cdot\frac{1}{8}\left(13.24\right) \approx 50312.$$

On forme alors un intervalle de confiance à environ 95% de τ_Y à l'aide de $5641.56 \pm 2\sqrt{50312} \approx 5641.56 \pm 448.61$.

(d) Pas de solution.

(e) Pas de solution.

35. Un contrôle traditionnel exprime les ventes au détail comme étant l'inventaire d'ouverture plus les achats du magasin, duquel on retranche l'inventaire de fermeture, sur une période de 6 semaines afin de rapporter les ventes totales. De telles données, provenant de plusieurs magasins et recueillies pour une variété de marques concurrentes, permettent de donner un estimé dus parts de marché. Mais les méthodes de vérification des ventes de la fin de semaine et des achats en magasin offrent des méthodes plus rapides pour donner un estimé dus parts de marché. La première élimine les achats en magasin, car les achats sont minimes le fin de semaine, mais utilise une période plus courte et est sujette à des irrégularités dues aux promotions de fin de semaine. La seconde utilise uniquement l'information sur les achats pour calculer la part de marché et n'implique aucune vérification des stocks. Pour une certaine marque de bière, les données sur les parts de marché calculées par les trois méthodes [traditionnelle (T), fin de semaine (F), achats (A)] sont présentées dans le tableau ci-dessous [les observations ont été effectuées à six périodes différentes au cours de l'année].

Traditionnelle (T)	**Fin de semaine** (F)	**Achats** (A)
15	16	12
18	17	14
16	17	20
14	16	11
13	12	8
16	18	15

(a) Présenter un estimé du quotient de la part de marché moyenne calculée avec la méthode F par celle calculée avec la méthode T, et donner une marge d'erreur sur l'estimation.

(b) Présenter un estimé du quotient de la part de marché moyenne calculée avec la méthode A par celle calculée avec la méthode T, et donner une marge d'erreur sur l'estimation.

(c) Quelle méthode se compare le plus favorablement à la méthode traditionnelle?

(d) Y a-t-il des obstacles qui se manifestent dans les divers diagrammes de dispersion?

Solution:

(a) Les sommes et les données qui seront nécessaires pour calculer les paramètres et les coefficients sont les suivantes :

$$\sum_{i=1}^{6} t_i = 92 \qquad \sum_{i=1}^{6} f_i = 96 \qquad \sum_{i=1}^{6} t_i f_i = 1486$$
$$\sum_{i=1}^{6} t_i^2 = 1426 \qquad \sum_{i=1}^{6} f_i^2 = 1558$$

Les moyennes sont $\overline{t} = 15.3333$, $\overline{f} = 16$. Les paramètres de régression sont ainsi :

$$b_F = \frac{1486 - 6(15.3333)(16)}{1426 - 6(15.3333)^2} = 0.9130$$
$$a_F = 16 - 0.9130(15.3333) = 2$$

et la droite de régression est $\hat{f} = 2 + 0.9130t$. De plus, $s\{a_F\} = 5.9764$ et $\rho_{T,F} = 0.7623$. Alors $r_F = \frac{\overline{f}}{\overline{t}} = 1.043$ et

$$\hat{V}(r_F) = \left(\frac{N-n}{Nn}\right) \frac{1}{\overline{T}^2} s_{r_F}^2.$$

Puisque $\overline{T}$ et N sont inconnus, nous utilisons $\bar{t} \approx \overline{T}$ et $\frac{N-n}{N} \approx 1$. Alors

$$\hat{V}(r_F) = \frac{1}{n\overline{T}^2} s^2_{r_F} = \frac{1}{5(15.3333)^2} 1.895 = 0.00134,$$

et on obtient un I.C. de R_F à environ 95% à l'aide de $1.043 \pm 2\sqrt{0.00134} = 1.043 \pm 0.0733$.

(b) Les sommes et les données qui seront nécessaires pour calculer les paramètres et les coefficients sont les suivantes :

$$\sum_{i=1}^{6} t_i = 92 \qquad \sum_{i=1}^{6} a_i = 80 \qquad \sum_{i=1}^{6} t_i a_i = 1250$$
$$\sum_{i=1}^{6} t_i^2 = 1426 \qquad \sum_{i=1}^{6} a_i^2 = 1150$$

Les moyennes sont $\bar{t} = 15.3333$, $\bar{a} = 13.3333$. Les paramètres de régression sont ainsi :

$$b_A = \frac{1250 - 6(15.3333)(13.3333)}{1426 - 6(15.3333)^2} = 1.5217$$
$$a_A = 13.3333 - 1.5217(15.3333) = -10$$

et la droite de régression est $\hat{a} = -10 + 1.5217t$. De plus, $s\{a_A\} = 13.6135$ et $\rho_{T,A} = 0.6528$. Alors $r_A = \frac{\bar{a}}{\bar{t}} = 0.8696$ et

$$\hat{V}(r_A) = \left(\frac{N-n}{Nn}\right) \frac{1}{\overline{T}^2} s^2_{r_A}.$$

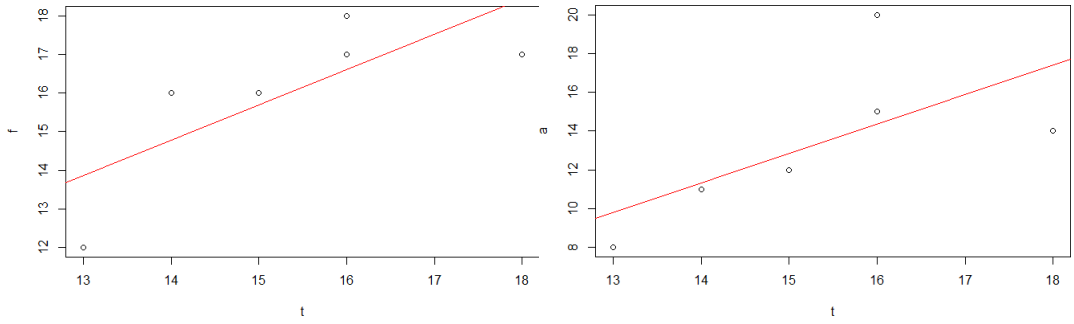
Puisque $\overline{T}$ et N sont inconnus, nous utilisons $\bar{t} \approx \overline{T}$ et $\frac{N-n}{N} \approx 1$. Alors

$$\hat{V}(r_A) = \frac{1}{n\overline{T}^2} s^2_{r_A} = \frac{1}{5(15.3333)^2} 10.868 = 0.0077,$$

et on obtient un I.C. de R_A à environ 95% à l'aide de $0.868 \pm 2\sqrt{0.0077} = 0.868 \pm 0.1755$.

(c) La méthode F est favorable à la méthode A par le quotient puisque $\hat{V}(r_F) < \hat{V}(r_A)$, et r_F est plus près de 1 que r_A ne l'est.

(d) Les méthodes par le quotient utilisées en (a) et (b) nécessitent $a_A, a_P \approx 0$ et $1 \geq \rho_{T,F}, \rho_{T,P} \gg 0$. On peut montrer à l'aide de tests t qu'il n'est pas impossible que $a_A, a_F \approx 0$ et les coefficients de corrélation $1 \geq \rho_{T,F}, \rho_{T,A} \gg 0$ sont relativement élevés comme on peut le voir dans les nuages. ■



36. Une population est composée de $N = 5$ unités dont les valeurs de X et Y sont les suivantes:

$(X_1, Y_1) = (3, 2), \quad (X_2, Y_2) = (5, 3), \quad (X_3, Y_3) = (3, 3), \quad (X_4, Y_4) = (4, 2), \quad (X_5, Y_5) = (6, 5).$

(a) Déterminer le quotient R dans cette population.

(b) Pour chaque échantillon possible de taille $n = 3$, déterminer le quotient r. Calculer ensuite le biais d'échantillonnage de r, à savoir $E[r - R]$.

(c) Nous avons développé, en classe, l'approximation théorique de l'erreur systématique:

$$E[r - R] \approx \frac{1}{n\mu_X^2} \left(\frac{N-n}{N-1} \right) (R\sigma_X^2 - \rho\sigma_X\sigma_Y).$$

Calculer la valeur de l'approximation théorique de l'erreur systématique pour cette population, et comparer avec la valeur réelle.

(d) Calculer les deux estimations de la moyenne de la population, $\overline{y}_{\text{EAS}}$ et $\hat{\mu}_{Y;R}$, pour chaque échantillon. À partir de ces résultats, calculer $V(\overline{y}_{\text{EAS}})$ et $E\left[(\hat{\mu}_{Y;R} - \mu_Y)^2\right]$. Discutez des avantages et des inconvénients de l'utilisation respective de y_{EAS} et de $\hat{\mu}_{Y;R}$ en tant qu'estimateurs de μ_Y.

Solution:

(a) Le quotient est

$$R = \frac{\mu_Y}{\mu_X} = \frac{\sum Y_i}{\sum X_i} = \frac{2+3+3+2+5}{3+5+3+4+6} = \frac{15}{21} = \frac{5}{7}.$$

(b) Il y a $\binom{5}{3} = 10$ échantillons de taille $n = 3$:

$(x_1, y_1), (x_2, y_2), (x_3, y_3)$	**Échantillon**	r
$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$	$(3,2), (5,3), (3,3)$	8/11
$(X_1, Y_1), (X_2, Y_2), (X_4, Y_4)$	$(3,2), (5,3), (4,2)$	7/12
$(X_1, Y_1), (X_2, Y_2), (X_5, Y_5)$	$(3,2), (5,3), (6,5)$	5/7
$(X_1, Y_1), (X_3, Y_3), (X_4, Y_4)$	$(3,2), (3,3), (4,2)$	7/10
$(X_1, Y_1), (X_3, Y_3), (X_5, Y_5)$	$(3,2), (3,3), (6,5)$	5/6
$(X_1, Y_1), (X_4, Y_4), (X_5, Y_5)$	$(3,2), (4,2), (6,5)$	9/13
$(X_2, Y_2), (X_3, Y_3), (X_4, Y_4)$	$(5,3), (3,3), (4,2)$	2/3
$(X_2, Y_2), (X_3, Y_3), (X_5, Y_5)$	$(5,3), (3,3), (6,5)$	11/14
$(X_2, Y_2), (X_4, Y_4), (X_5, Y_5)$	$(5,3), (4,2), (6,5)$	2/3
$(X_3, Y_3), (X_4, Y_4), (X_5, Y_5)$	$(3,3), (4,2), (6,5)$	10/13

Le biais d'échantillonnage est ainsi

$$E(r - R) = E(r) - R = \frac{1}{10}\left(\frac{8}{11} + \frac{7}{12} + \frac{5}{7} + \frac{7}{10} + \frac{5}{6} + \frac{9}{13} + \frac{2}{3} + \frac{11}{14} + \frac{2}{3} + \frac{10}{13}\right) - \frac{5}{7} = -0.0004045954.$$

(c) Les valeurs importantes sont :

$$\mu_X = \frac{1}{5}(3+5+3+4+6) = 4.2$$
$$\sigma_X^2 = \frac{1}{5}\left(3^2+5^2+3^2+4^2+6^2\right) - (4.2)^2 = 1.36$$
$$\mu_Y = \frac{1}{5}(2+3+3+2+5) = 3$$
$$\sigma_Y^2 = \frac{1}{5}\left(2^2+3^2+3^2+2^2+5^2\right) - 3^2 = 1.2$$
$$\text{Cov}(X,Y) = E(XY) - \mu_X\mu_Y = \frac{1}{5}((3\cdot 2)+(5\cdot 3)+(3\cdot 3)+(4\cdot 2)+(6\cdot 5)) - (4.2)(3) = 1.$$

L'approximation théorique du biais is

$$E(r-R) \approx \frac{1}{n\mu_X^2}\left(\frac{N-n}{N-1}\right)\left(R\sigma_X^2 - \mathrm{Cov}(X,Y)\right)$$
$$= \frac{1}{3(4.2)^2}\left(\frac{5-3}{5-1}\right)\left(\frac{5}{7}(1.36)-1\right) = -0.0002699492.$$

(d) Nous avons la table suivante :

$(x_1,y_1),(x_2,y_2),(x_3,y_3)$	**Échantillon**	r	$\overline{y}$	$\hat{\mu}_{Y;R} = r\mu_X$
$(X_1,Y_1),(X_2,Y_2),(X_3,Y_3)$	$(3,2),(5,3),(3,3)$	8/11	8/3	3.054545
$(X_1,Y_1),(X_2,Y_2),(X_4,Y_4)$	$(3,2),(5,3),(4,2)$	7/12	7/3	2.45
$(X_1,Y_1),(X_2,Y_2),(X_5,Y_5)$	$(3,2),(5,3),(6,5)$	5/7	10/3	3
$(X_1,Y_1),(X_3,Y_3),(X_4,Y_4)$	$(3,2),(3,3),(4,2)$	7/10	7/3	2.94
$(X_1,Y_1),(X_3,Y_3),(X_5,Y_5)$	$(3,2),(3,3),(6,5)$	5/6	10/3	3.5
$(X_1,Y_1),(X_4,Y_4),(X_5,Y_5)$	$(3,2),(4,2),(6,5)$	9/13	3	2.907692
$(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)$	$(5,3),(3,3),(4,2)$	2/3	8/3	2.8
$(X_2,Y_2),(X_3,Y_3),(X_5,Y_5)$	$(5,3),(3,3),(6,5)$	11/14	11/3	3.3
$(X_2,Y_2),(X_4,Y_4),(X_5,Y_5)$	$(5,3),(4,2),(6,5)$	2/3	10/3	2.8
$(X_3,Y_3),(X_4,Y_4),(X_5,Y_5)$	$(3,3),(4,2),(6,5)$	10/13	10/3	3.230769

Puisque $\overline{y}$ est un estimateur sans biais de μ_Y, nous avons ainsi

$$\mathrm{EQM}(\overline{y}) = V(\overline{y}) = \frac{1}{10}\sum_{\overline{y}} \overline{y}^2 - \mu_Y^2 = \frac{1}{10}((8/3)^2 + \cdots + (10/3)^2) - 3^2 = 9.2 - 9 = 0.2$$

$$\mathrm{EQM}(\hat{\mu}_{Y;R}) = \frac{1}{10}\left[(3.0... - 3)^2 + \cdots + (3.2... - 3)^2\right] + (E(\hat{\mu}_{Y;R} - \mu_Y))^2 = 0.079 + (-0.017)^2 = 0.079.$$

Ceci suggère que l'estimateur par le quotient, quoique biaisé, est plus précis et sans doute préférable dans ce cas. En général, $\hat{\mu}_Y$ est un bon choix d'estimateur pour μ_Y si :

i. la relation entre Y et X est linéaire et passe par l'origine, et si

ii. si la variance de Y le long de la droite de régression est proportionnelle à la valeur prise par X.

Dans ce cas, cependant, il n'y a pas vraiment assez de points dans la population pour déterminer si les hypothèses sont réelement valides. ■

37. Les données relatives à la taille de la famille x_i et aux dépenses alimentaires y_i au cours de la semaine d'enquête sont enregistrées pour chaque famille d'un échantillon de 33 familles provenant d'une grande population de familles.

(a) Exprimer les dépenses alimentaires (pour cette semaine) par personne dans la population sous forme de quotient de populations.

(b) En utilisant les données de l'échantillon, nous obtenons

$$\sum_{i=1}^{33} x_i = 123, \quad \sum_{i=1}^{33} x_i^2 = 533, \quad \sum_{i=1}^{33} y_i = 2721.30, \quad \sum_{i=1}^{33} y_i^2 = 254196, \quad \sum_{i=1}^{33} x_i y_i = 10786.5$$

donner un estimé et un I.C. (à environ 95%) des dépenses alimentaires par capita dans la population.

Solution:

(a) Le quotient recherché est $R = \frac{\sum_{j=1}^{N} y_j}{\sum_{j=1}^{N} x_j}$.

(b) Avec les données du problème, l'estimé ponctuel pour le quotient est $r = \frac{\sum_{i=1}^{33} y_i}{\sum_{i=1}^{33} x_i} = \frac{2721.3}{123} = 22.12$. Si on néglige le facteur de correction en population finie, la variance de l'estimateur est environ

$$\hat{V}(r) = \frac{s_r^2}{33\overline{x}^2} = \frac{\sum_{i=1}^{33}(y_i - rx_i)^2}{32 \cdot 33 \cdot \overline{x}^2} = \frac{1}{123^2 \cdot 32/33}\left[\sum_{i=1}^{33} y_i^2 - 2r\sum_{i=1}^{33} x_i y_i + r^2 \sum_{i=1}^{33} x_i^2\right]$$
$$= \frac{254196 - 2(22.12)10786.5 + (22.12)^2 533}{123^2 \cdot 32/33} = 2.58.$$

La marge d'erreur sur l'estimateur r est ainsi $B_r = 2\sqrt{\hat{V}(r)} = 3.21$, d'où l'intervalle de confiance recherché est 22.12 ± 3.21. ■

38. Donner un estimé du volume total (en pieds cube) des arbres marqués pour la vente (cf. donnés de la question 31) en utilisant l'estimation par la régression et l'estimation par EAS, et placer une limite sur l'erreur d'estimation dans les deux cas.

Solution: On commence par l'EAS, et on termine avec l'estimateur de régression.

(a) Puisque $\overline{y} = \frac{142}{12} = 11.83$, l'estimateur du total dans le contexte EAS est

$$\tau = N\overline{y} = 250(11.83) = 2958.33.$$

La variance de l'estimateur est environ

$$\begin{aligned}\hat{V}(\hat{\tau}) &= N^2 \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = 250^2 \frac{s^2}{12}\left(1 - \frac{12}{250}\right) = \frac{14875}{3}s^2 \\ &= \frac{14875}{3} \cdot \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right] = \frac{14875}{3} \cdot \frac{1}{11}\left[1976 - 12(11.83)^2\right] = 133700.6\end{aligned}$$

et la marge d'erreur sur l'estimation est $2\sqrt{\hat{V}(\tau)} = 731.302$.

(b) Nous obtenons :

i. $\sum_{i=1}^{12} X_i = 6.7000$

ii. $\sum_{i=1}^{12} Y_i = 142.0000$

iii. $\sum_{i=1}^{12} X_i Y_i = 91.2000$

iv. $\sum_{i=1}^{12} X_i^2 = 4.2500$

v. $\sum_{i=1}^{12} Y_i^2 = 1976.0000$

vi. $n = 12$.

Les moyennes sont ainsi $\overline{X} = 0.5583$ et $\overline{Y} = 11.8333$, et conséquemment, les coefficients de la régression sont

$$b_1 = \frac{91.2000 - 12(0.5583)(11.8333)}{4.2500 - 12(0.5583)^2} = 23.4043 \quad \text{et} \quad b_0 = 11.8333 - 23.4043(0.5583) = -1.2340$$

et la droite de régression est $\hat{Y} = -1.2340 + 23.4043X$. La corrélation entre X et Y est forte ($\rho = .9712$). L'estimateur ponctuel du total est donc

$$\hat{\tau}_{Y;L} = N\hat{\mu}_{Y;L} = 250(-1.2340 + 23.4043 \cdot \frac{75}{250}) = 1446.822,$$

et la variance d'échantillonnage est

$$\hat{V}(\hat{\tau}_{Y;L}) = N^2 \cdot \frac{s_Y^2(1-\rho)}{n}\left(\frac{N-n}{N-1}\right) = 250^2 \cdot \frac{26.96(1-0.9712)}{12}\left(\frac{250-12}{250-1}\right) = 3865.349,$$

et la marge d'erreur sur l'estimation est $2\sqrt{\hat{V}(\hat{\tau}_{Y;L})} = 2\sqrt{3865.349} = 124.3439$, ce qui est de loin meilleur. ■

39. Une agence de publicité s'inquiète de l'effet que peut avoir une nouvelle campagne promotionnelle régionale sur les ventes totales en dollars d'un produit particulier. Un EAS de 20 magasins a été constitué à partir de la population de 452 magasins dans lesquels le produit est vendu. Les données trimestrielles sur les ventes ont été obtenues pour la période de trois mois en cours et la période de trois mois précédant la nouvelle campagne et sont présentées dans le tableau ci-dessous. On sait également que les ventes totales pour l'ensemble des 452 magasins au cours de la période de trois mois précédant la nouvelle campagne étaient de 216,256.

Magasin	**Antérieures** x	**Actuelles** y	**Magasin**	**Antérieures** x	**Actuelles** y
1	208	239	11	599	626
2	400	428	12	510	538
3	440	472	13	828	888
4	259	276	14	473	510
5	351	363	15	924	998
6	880	942	16	110	171
7	273	294	17	829	889
8	487	514	18	257	265
9	183	195	19	388	419
10	863	897	20	244	257

(a) Tracer un diagramme de dispersion des valeurs des ventes actuelles par rapport aux valeurs des ventes antérieures. Quelle méthode d'estimation semble plus appropriée? Expliquer.

(b) Déterminer un I.C. pour les ventes totales actuelles à environ 95% en utilisant l'estimation par le quotient.

(c) Répéter l'étape (b) en utilisant l'estimation par la régression.

(d) Comparer les marges d'erreur sur l'estimation pour les intervalles de confiance obtenus aux étapes (b) et (c). Laquelle est la plus élevée? Est-ce conforme aux attentes? Expliquer.

(e) Répéter l'étape (b) en utilisant l'estimation par la différence.

(f) L'estimation par la différence est-elle une approche raisonnable pour donner un estimé du total des ventes en cours? Expliquer.

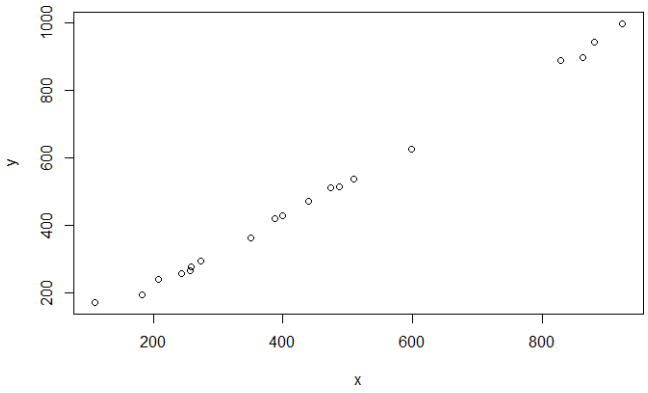
(g) Comparer les marges d'erreur sur l'estimation pour les intervalles de confiance obtenus aux étapes (c) et (f). Laquelle est la plus élevée? Est-ce conforme aux attentes? Expliquer.

(h) Combien de magasins faudrait-il échantillonner afin de donner un estimé du total des ventes actuelles en préservant une marge d'erreur sur l'estimation de 2500$ si l'on utilise l'estimation par le quotient?

(i) Répéter l'étape (h) en utilisant l'estimation par la régression et l'estimation par la différence.

Solution:

(a) Voici le nuage de points:



La droite de régression est $y = 10.11 + 1.05x$; la relation entre les deux variables est clairement fortement linéaire (et la droite de régression croise l'axe des Y très près de l'origine, à 10.1052), mais la variance ne semble pas être proportionnelle à X, et l'estimateur de régression est probablement plus efficace que l'estimateur du quotient.

(b) On obtient 231611.63 ± 3073.83 selon la méthode du quotient.

(c) On obtient 231581.43 ± 2950.85 selon la méthode de la régression.

(d) L'estimateur du quotient est effectivement moins "serré" que l'estimateur de régression, mais la différence n'est pas si importante que cela.

(e) La méthode de la différence est sans doute adéquate, puisque la relation entre la variable auxiliaire X et la variable réponse Y est fortement linéaire, et puisque la pente de la droite de régression s'approche de 1.

(f) On obtient 231510.78 ± 3849.01 selon la méthode de la différence.

(g) L'estimateur de régression est plus "serré" que celui de la différence, ce qui n'est pas surprenant puisque c'est toujours le cas sauf si $\rho\frac{\sigma_Y}{\sigma_X} = 1$.

(h) Selon la méthode du quotient, nous devons avoir

$$n = \frac{N\sigma_W^2}{(N-1)D + \sigma_W^2} \approx \frac{452(241.940)}{\frac{451(2500)^2}{4(452)^2} + 241.940} = 29.63;$$

on devrait donc prélever au moins 30 unités.

(i) Selon la méthode de l'estimateur de régression, nous avons

$$V(\hat{\mu}_{Y;L}) \approx \frac{\text{EQM}}{n}\left(\frac{N-n}{N-1}\right).$$

On résoud $B = 2N\sqrt{\frac{\text{EQM}}{n}\left(\frac{N-n}{N-1}\right)}$ pour n, ce qui donne $n \approx \frac{N\cdot\text{EQM}}{(N-1)D+\text{EQM}}$, où $D = \frac{B^2}{4N^2}$. Avec la marge requise, cela nous donne

$$n \approx \frac{452(222.968)}{\frac{451(2500)^2}{4(452)^2} + 222.968} = 27.44,$$

d'où il nous faudrait un échantillon d'au moins 28 unités. Finalement, pour l'estimateur de la différence, nous avons

$$V(\hat{\mu}_{Y;D}) \approx \frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right).$$

On résoud, $B = 2N\sqrt{\frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right)}$ pour n, ce qui donne $n \approx \frac{N\cdot\sigma_D^2}{(N-1)D+\sigma_D^2}$, où $D = \frac{B^2}{4N^2}$. Avec la marge requise, cela nous donne

$$n \approx \frac{452(379.355)}{\frac{451(2500)^2}{4(452)^2} + 379.355} = 44.78,$$

d'où il nous faudrait prélever au moins 45 unités. ■

41. Le modèle théorique utile $Y_i = \beta X_i + D_i$, où D_i représente l'écart par rapport à la droite, peut être utilisé afin de comparer divers estimateurs de quotients. Pour une valeur donnée de $X = x$, supposons que les valeurs de Y soient éparpillées autour de la droite, de sorte que l'espérance et la variance des écarts soient

$$E[D \mid X = x] = 0 \quad \text{et} \quad V[D \mid X = x] = \sigma^2 x^{2a}.$$

Considérons un estimateur général de β ayant la forme $b = \sum_{i=1}^{n} c_i y_i$, où c_i peut dépendre de x_i.

(a) Trouver une condition sur les coefficients afin de garantir que b est un estimateur non biaisé de β, étant donné les x observés.

(b) Déterminer une expression pour la variance de b en fonction de a, conditionnellement aux $x > 0$ observés.

(c) Pour une valeur donnée de a, trouver l'estimateur non biaisé de la classe ci-dessus avec une variance conditionnelle minimale.

(d) Si $a = 0$, quel estimateur présente la plus petite variance conditionnelle? Et si $a = 0.5$? Et pour $a = 1$?

(e) Discuter des conséquences de cette analyse pour l'estimation de μ_Y par le quotient et par la régression.

Solution:

(a) On doit avoir $E(b) = \beta$. Puisque $E(D_i|X = x_i) = 0$,

$$E(y_i|X = x_i) = E(\beta x_i|X = x_i) + E(D_i|X = x_i) = \beta x_i + 0 = \beta x_i \quad \text{for all } i = 1, \ldots, n.$$

Mais

$$\begin{aligned} E(b) &= E\left(c_1 y_1 + \cdots + c_n y_n\right) = c_1 E(y_1|X = x_1) + \cdots + c_n E(y_n|X = x_n) = c_1 \beta x_1 + \cdots + c_n \beta x_n \\ &= \beta(c_1 x_1 + \cdots + c_n x_n) \end{aligned}$$

Pour que $E(b) = \beta$, on doit avoir $c_1 x_1 + \cdots + c_n x_n = 1$, à moins, bien sûr, que $\beta = 0$, auquel cas b est nécessairement un estimateur sans biais de β.

(b) Puisque $V(D_i|X = x_i) = \sigma^2 x_i^{2a}$,

$$V(y_i|X = x_i) = V(\beta x_i + D_i|X = x_i) = V(D_i|X = x_i) = \sigma^2 x_i^{2a} \quad \text{pour tout } i = 1, \ldots, n.$$

Si les x_i sont indépendants, les y_i le sont également et

$$\begin{aligned} V(b) &= V\left(c_1 y_1 + \cdots + c_n y_n\right) = c_1^2 V(y_1|X = x_1) + \cdots + c_n^2 V(y_n|X = x_n) \\ &= c_1^2 \sigma^2 x_1^{2a} + \cdots + c_n^2 \sigma^2 x_n^{2a} = \sigma^2 \left(c_1^2 x_1^{2a} + \cdots + c_n^2 x_n^{2a}\right) \end{aligned}$$

(c) Supposons que $\beta \neq 0$. On cherche à minimiser $V(b) = f(x) = \sigma^2 \left(c_1^2 x_1^{2a} + \cdots + c_n^2 x_n^{2a}\right)$, sujet à la contrainte $g(x) = c_1 x_1 + \cdots + c_n x_n - 1 = 0$. Selon la méthode des multiplicateurs de Lagrange, on cherche les points x tels que

$$\begin{aligned} \nabla f(x) &= \lambda \nabla g(x) \\ g(x) &= 0 \end{aligned}$$

c'est-à-dire que

$$2a\sigma^2 c_i^2 x_i^{2a-1} = \lambda c_i, \quad i = 1, \ldots, n$$
$$c_1 x_1 + \cdots + c_n x_n = 1$$

On résoud pour c_i: la première ligne nous donne soit $c_i = 0$ ou soit $c_i = \frac{\lambda}{2a\sigma^2 x_i^{2a-1}}$ pour tout $i = 1, \ldots, n$. Ainsi, $c_i x_i = 0$ ou $c_i x_i = \frac{\lambda}{2a\sigma^2 x_i^{2a-2}}$ pour tout $i = 1, \ldots, n$.
Mais $c_i \neq 0$ pour au moins un i, sinon $g(x) = -1 \neq 0$. Soit $C = \{i : c_i \neq 0\} \neq \varnothing$; nous devons avoir

$$1 = \sum_{i \in C} \frac{\lambda}{2a\sigma^2 x_i^{2a-2}} = \frac{\lambda}{2a\sigma^2} \sum_{i \in C} x_i^{2-2a},$$

d'où $\lambda = \frac{2a\sigma^2}{\sum_{i \in C} x_i^{2-2a}}$. Ainsi,

$$c_i = \frac{\lambda}{2a\sigma^2 x_i^{2a-1}} = \left(x_i^{2a-1} \sum_{j \in C} x_j^{2-2a} \right)^{-1}, \quad \text{pour } i \in C, \quad \text{et} \quad c_i = 0, \quad \text{pour } i \notin C.$$

Si $c_i = 0$, on aurait pu tout aussi bien ne pas choisir le $i-$ème point dans le modèle. Conséquemment, on suppose que $c_i \neq 0$ pour tout i, de sorte que

$$c_i = \frac{x_i^{1-2a}}{x_1^{2-2a} + \cdots + x_n^{2-2a}}, \quad i = 1, \ldots, n.$$

(d) Lorsque $a = 0$, nous avons $c_i = \frac{x_i}{x_1^2 + \cdots + x_n^2}$, $i = 1, \ldots, n$, d'où l'estimateur sans biais ayant la plus faible variance est

$$b = c_1 y_1 + \cdots + c_n y_n = \frac{1}{\sum x_i^2} (x_1 y_1 + \cdots + x_n y_n) = \frac{\sum x_i y_i}{\sum x_i^2};$$

c'est l'estimateur des moindres carrés de la pente de la droite de régression à travers l'origine.
Lorsque $a = 0.5$, nous avons $c_i = \frac{1}{x_1 + \cdots + x_n}$, $i = 1, \ldots, n$, d'où l'estimateur sans biais ayant la plus faible variance est

$$b = c_1 y_1 + \cdots + c_n y_n = \frac{1}{\sum x_i} (y_1 + \cdots + y_n) = \frac{\sum y_i}{\sum x_i};$$

c'est l'estimateur du quotient.
Lorsque $a = 1$, nous avons $c_i = \frac{1}{x_i}$, $i = 1, \ldots, n$, , d'où l'estimateur sans biais ayant la plus faible variance est

$$b = c_1 y_1 + \cdots + c_n y_n = \frac{y_1}{x_1} + \cdots + \frac{y_n}{x_n} = \sum \frac{y_i}{x_i}.$$

(e) Si nous pouvons montrer que le modèle de la ligne (non-horizontale) passant par l'origine est approprié et que l'hypothèse de variance constante se vérifie, l'estimateur sans biais le plus efficace de μ_Y est basé sur l'estimateur de régression. Si, par contre, il est démontré que la variance est proportionnelle au niveau des x, l'estimateur sans biais le plus efficace de μ_Y est basé sur l'estimateur du quotient. ■