Record Number: 21160
Author, Monographic: Girard, C.//Ouarda, T. B. M. J.//Bobée, B.
Author Role:
Title, Monographic: Étude du biais dans la prédiction depuis un modèle log-linéaire
Translated Title:
Reprint Status:
Edition:
Author, Subsidiary:
Author Role:
Place of Publication: Québec
Publisher Name: INRS-Eau
Date of Publication: 2000
Original Publication Date: Novembre 2000
Volume Identification:
Extent of Work: iii, 37
Packaging Method: pages incluant 4 annexes
Series Editor:
Series Editor Role:
Series Title: INRS-Eau, rapport de recherche
Series Volume ID: 575
Location/URL:
ISBN: 2-89146-443-5
Notes: Rapport annuel 2000-2001
Abstract:
Call Number: R000575
Keywords: rapport/ ok/ dl

Étude du biais dans la prédiction depuis un modèle log-linéaire

Rapport de recherche No R-575 *Novembre 2000*

Étude du biais dans la prédiction depuis un modèle log-linéaire

Rapport préparé par :

Claude Girard
Taha B.M.J. Ouarda
Bernard Bobée

Chaire industrielle en Hydrologie statistique
Institut national de la recherche scientifique, INRS-Eau
2800, rue Einstein, Case postale 7500, Sainte-Foy (Québec), G1V 4C7

Rapport de recherche No. R-575

Novembre 2000

ISBN : 2-89146-443-5

Table des matières

1. INTRODUCTION

En hydrologie, comme dans plusieurs domaines où la statistique est appliquée, l'utilisation d'un modèle log-linéaire est courante. Dans le cadre plus spécifique de l'estimation régionale des quantiles de crues, l'emploi d'un modèle log-linéaire liant un quantile de crue d'un bassin à certaines de ses caractéristiques physiographiques est très répandu. Un exemple d'application de ce modèle dans le cadre de l'estimation régionale est présenté par [Ribeiro-Corréa *et al.*1995] Rappelons que sous sa forme générale le modèle log-linéaire est décrit par :

$$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon \qquad (1.1)$$

où Y est la variable aléatoire d'intérêt, X est la variable (aléatoire) explicative, β_0 et β_1 sont des paramètres du modèle et ε est un terme d'erreur. Nous faisons l'hypothèse que le terme d'erreur, ε, est une variable aléatoire de densité normale $N(0,\sigma^2)$. Lorsque $X=x_0$, le terme d'erreur ε et la variable $\ln(Y)$ sont les seules variables aléatoires dans le modèle (1.1); il en résulte que $\ln(Y)$ admet une densité (conditionnelle à $X=x_0$) normale.

Le modèle log-linéaire est souvent utilisé pour établir une prédiction $\hat{\ln}(Y_0)$ de la valeur de la variable $\ln(Y)$ pour une valeur donnée x_0 de la variable X. À cette fin, la procédure généralement suivie consiste d'abord à obtenir les estimations par moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$ des paramètres β_0 et β_1, respectivement, pour établir l'équation de prédiction en fonction de X :

$$\hat{\ln}(Y) = \hat{\beta}_0 + \hat{\beta}_1 X \qquad (1.2)$$

L'équation (1.2) représente le modèle (1.1) ajusté aux données utilisées. Ensuite, en substituant à X dans (1.2) une valeur donnée x_0, une valeur correspondante est obtenue pour $\widehat{\ln}(Y_0)$; c'est la valeur prédite par le modèle (1.1) pour la valeur donnée x_0. Puisque Y est la véritable variable d'intérêt, et non $\ln(Y)$, il est courant d'obtenir de $\widehat{\ln}(Y_0)$ une estimation $\hat{Y}_0$ de Y en lui appliquant la transformation exponentielle.

Cette façon de procéder pour obtenir une prédiction $\hat{Y}_0$ pour Y (étant donné $X=x_0$), analogue à la façon de faire pour un modèle linéaire simple, fait l'objet de sérieuses critiques. La principale critique se rapporte au biais qui est introduit en passant d'une valeur prédite $\widehat{\ln}(Y_0)$ pour $\ln(Y)$ à la valeur prédite induite $\hat{Y}_0$ en lui appliquant la transformation exponentielle, l'inverse du logarithme. En effet, la valeur prédite $\widehat{\ln}(Y_0)$ obtenue par (1.2) est une estimation sans biais pour la moyenne de la distribution conditionnelle de la variable aléatoire $\ln(Y)$ étant donnée $X=x_0$. Puisque cette distribution est par hypothèse normale, et par conséquent symétrique, il s'ensuit que la valeur prédite $\widehat{\ln}(Y_0)$ est également une estimation sans biais pour la médiane de cette distribution. L'application de la transformation exponentielle à la valeur prédite $\widehat{\ln}(Y_0)$ obtenue de (1.2) fournit l'estimation suivante $\hat{Y}_0$ de Y pour la valeur donnée x_0 de X :

$$\hat{Y}_0 = \exp(\hat{\beta}_0) \times \exp(\hat{\beta}_1 x_0) \qquad (1.3)$$

Il est possible de montrer [Miller 1984] que cette estimation est biaisée pour la moyenne de la distribution de Y pour la valeur donnée x_0 de X, bien qu'elle ne le soit pas pour sa médiane. Puisque c'est la moyenne de la distribution conditionnelle de Y étant donnée $X=x_0$ qui nous intéresse généralement plutôt que sa médiane, l'importance de ce biais doit faire l'objet d'une étude approfondie.

La présente étude a pour but d'investiguer plus à fond le problème de biais dans un modèle log-linéaire et de cerner l'impact que cette problématique a en pratique pour l'utilisateur d'un tel modèle.

2. APPROCHES DE RÉDUCTION/CORRECTION DU BIAIS

Il existe dans la littérature deux grandes approches pour apporter une solution considérée satisfaisante à ce problème de biais. La première approche consiste à substituer à l'estimateur défini par (1.3) un estimateur sans biais. Une possibilité consiste à introduire un facteur multiplicatif dans l'équation (1.3) afin de rendre l'estimation initiale $\hat{Y}_0$ non biaisée pour la moyenne de la distribution conditionnelle de Y étant donné $X = x_0$. Cette méthode de correction du biais est décrite par [Neyman et Scott 1960]. Cependant, [Miller 1984] décrit comme complexes les formulations qui y sont présentées pour obtenir des estimations sans biais pour la moyenne, un avis partagé par [Seber et Wild 1989].

La deuxième approche consiste à introduire un facteur multiplicatif dans l'équation (1.3) afin d'en atténuer le biais, et non plus de l'éliminer complètement. L'avantage premier de cette approche réside dans la simplicité du facteur multiplicatif employé. Un premier facteur de correction est présenté en détail par [Miller 1984], et mentionné par [Duan 1983]. C'est d'ailleurs celui que nous nous proposons d'étudier ici puisqu'il est possible d'en établir théoriquement les propriétés. [Duan 1983] présente également un facteur de correction dans un cadre non-paramétrique, c'est-à-dire qu'aucune densité n'est assignée au terme d'erreur. Puisque nous ne considérons ici que le cas où les erreurs sont normales, nous limitons cette étude au facteur de correction proposé par Duan-Miller. Il faut cependant noter que l'approche non-paramétrique s'avère très intéressante en pratique dans le cas où l'hypothèse de normalité des erreurs ne semble pas être adéquatement remplie. Mentionnons que [Hoos 1996] fait exclusivement usage de cette approche.

L'approche de réduction de biais de Duan-Miller consiste à remplacer l'équation (1.3) par l'équation de prédiction suivante, pour une valeur donnée x_0 de X :

$$\hat{Y} = \exp(\hat{\beta}_0) \times \exp(\hat{\beta}_1 x_0) \times \exp\left(\frac{\hat{\sigma}^2}{2}\right) \tag{2.1}$$

où $\hat{\sigma}^2$ est l'estimation sans biais de la variance du terme d'erreur du modèle, σ^2, induite de l'estimation des paramètres du modèle par la méthode des moindres carrés.

Pour justifier l'introduction de ce facteur, [Miller 1984] fait l'observation que la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$ est :

$$E(Y|X = x_0) = \exp(\beta_0) \times \exp(\beta_1 x_0) \times \exp\left(\frac{\sigma^2}{2}\right) \tag{2.2}$$

La démonstration de cette relation est présentée à l'annexe A. En comparant (1.3) et (2.2), il apparaît effectivement naturel d'introduire le facteur $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$ qui manque en quelque sorte dans (1.3) pour être la contrepartie estimée de (2.2).

Il faut cependant prendre note que l'estimation (2.1), tout comme (1.3), admet un biais pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$. Cela est essentiellement dû au fait que chacune des estimations $\exp(\hat{\beta}_0)$, $\exp(\hat{\beta}_1 x_0)$ et $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$ est biaisée. Toutefois, le biais est moins important dans le cas de (2.1) que dans (1.3) [Miller 1984].

3. CONSIDÉRATIONS GÉNÉRALES SUR LE BIAIS EN ESTIMATION

Utiliser en remplacement de (1.3) un estimateur qui présente un biais moindre, voire un biais nul, c'est admettre que la présence de biais est inacceptable dans la prédiction à partir d'un modèle log-linéaire. Cela semble être une attitude naturelle et justifiée en estimation. Mais qu'en est-il vraiment? Puisque la notion de biais est au centre de la présente étude, il convient d'en discuter de manière plus approfondie.

Depuis longtemps en statistique, du moins en statistique fréquentiste (par opposition à la statistique bayésienne), l'absence de biais constitue un principe clé sous-jacent à l'inférence. Il est pratique courante dans une problématique d'estimation d'un paramètre de limiter l'étude aux seuls estimateurs sans biais pour le paramètre en question. Par la suite, s'il faut choisir parmi plusieurs estimateurs sans biais, seulement alors étudie-t-on d'autres propriétés des estimateurs. C'est ainsi, par exemple, qu'on est amené à s'intéresser en statistique-mathématique (fréquentiste) à l'estimateur sans biais à variance minimale. Ce concept est cependant trompeur puisque cet estimateur, bien qu'à variance minimale parmi les estimateurs sans biais, peut en fait présenter une variance très grande lorsque comparée (adéquatement) à la "variance" de tous les autres estimateurs possibles.

[Efron 1975] présente une critique très intéressante du principe d'absence de biais en estimation. La critique est essentiellement fondée sur l'approche des fonctions de risque, une ouverture en quelque sorte vers une vision bayésienne de la question. C'est d'ailleurs à l'aide des fonctions de risque que les premières perçées pour comprendre le paradoxe de Stein ont été obtenues [Efron et Morris 1977].

Les bayésiens, quant à eux, considèrent tout simplement non pertinent de considérer le biais d'un estimateur puisqu'il est basé sur un concept contradictoire avec le principe de vraisemblance. À ce sujet [Berger 1986] est très clair :

"[...] frequentist measures such as error probabilities, bias, coverage probability, and p-values, which involve averages over unobserved x, are not of this form, and can hence not be of basic interest from the conditional viewpoint."

En un mot, les bayésiens n'admettent pas les arguments qui font appel à des observations potentielles, c'est-à-dire à des résultats d'échantillon possibles autres que ceux obtenus. Par exemple, la notion de biais d'un estimateur consiste à faire la moyenne de tous les résultats potentiellement observables. Les bayésiens considèrent que toute l'information pertinente au niveau de l'échantillon, d'une réalisation, réside dans ce qui est réellement observé et non dans tout ce qui aurait pu être observé mais qui ne l'a pas été.

4. MESURE DE DISPERSION EN PRÉSENCE D'ESTIMATIONS BIAISÉES

L'estimation (2.1) de Duan-Miller, avec un facteur de correction du biais, présente un biais moins important pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$ que celui présenté par l'estimation directe (1.3). Sur la base de la réduction de biais, [Miller 1984] suggère que (2.1) soit préféré à (1.3), un avis repris par [Seber et Wild 1989]. Nous ne partageons cependant pas cet avis. En effet, il est fort possible, et nous montrerons dans les prochaines sections que c'est effectivement le cas ici, que les mesures prises pour corriger le biais des estimations en accroissent la dispersion. En d'autres termes, ce qui est gagné sur la précision moyenne des estimations est en quelque sorte perdu par l'étalement de ces dernières. Par conséquent, le fait que (2.1) admette un biais moins important que (1.3) ne constitue pas à lui seul, comme nous allons le montrer, un argument solide pour justifier l'emploi du premier au lieu du second pour fournir des estimations de qualité.

En présence de deux estimations biaisées, la mesure traditionnelle de dispersion qu'est la variance n'est pas un outil fiable pour juger de la qualité des estimations. Pour s'en convaincre, considérons une population de densité normale $N(\mu,1)$ de moyenne inconnue et de variance unité. Dans ce contexte considérons l'estimateur qui donne la valeur (arbitraire) 100, disons, comme estimation ponctuelle quelles que soient les données observées de cette population. Cet estimateur est clairement un choix d'estimateur absurde, et pourtant sa variance est nulle car l'estimation donnée est toujours la même valeur, à savoir ici 100.

Une mesure plus adéquate de dispersion en présence de biais est fournie par l'Écart Quadratique Moyen, l'EQM. Formellement, l'EQM d'un estimateur $\hat{\theta}$ pour le paramètre θ est défini par :

$$\mathrm{EQM}(\theta) = \mathrm{E}\left(\hat{\theta} - \theta\right)^2 \tag{4.1}$$

Une autre mesure possible est l'écart moyen absolu:

$$\mathrm{EQM}(\theta) = \mathrm{E}\left|\hat{\theta} - \theta\right| \tag{4.2}$$

Cependant, l'EQM lui est préféré en pratique pour deux raisons: 1) ses propriétés analytiques sont plus aisément établies; 2) l'EQM peut s'écrire facilement en fonction de la variance et du biais des estimations ; on a en effet:

$$\mathrm{EQM}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \left(\mathrm{biais}(\hat{\theta})\right)^2 \tag{4.3}$$

où $\mathrm{Var}(\hat{\theta})$ est la variance de l'estimateur et

$$\mathrm{biais}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta \tag{4.4}$$

Une forme dérivée de l'équation (4.3), qui nous sera utile par la suite, permet d'exprimer explicitement l'EQM en fonction de la variance et de l'espérance de l'estimateur de la façon suivante :

$$\mathrm{EQM}\left(\hat{\theta}\right) = \mathrm{Var}\left(\hat{\theta}\right) + \mathrm{E}^2\left(\hat{\theta}\right) - 2\theta\mathrm{E}\left(\hat{\theta}\right) + \theta^2 \tag{4.5}$$

Considérons $\hat{\theta}_1$ et $\hat{\theta}_2$, deux estimateurs d'un même paramètre θ de la distribution d'un phénomène aléatoire Y. En utilisant la forme (4.5) pour l'EQM, la différence des EQM, $\Delta\mathrm{EQM} = \mathrm{EQM}\left(\hat{\theta}_1\right) - \mathrm{EQM}\left(\hat{\theta}_2\right)$, pour les estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ s'écrit :

$$\begin{aligned}\Delta\mathrm{EQM} &= \mathrm{EQM}\left(\hat{\theta}_1\right) - \mathrm{EQM}\left(\hat{\theta}_2\right) \\ &= \mathrm{Var}\left(\hat{\theta}_1\right) - \mathrm{Var}\left(\hat{\theta}_2\right) + \mathrm{E}^2\left(\hat{\theta}_1\right) - \mathrm{E}^2\left(\hat{\theta}_2\right) + 2\theta\left(\mathrm{E}\left(\hat{\theta}_2\right) - \mathrm{E}\left(\hat{\theta}_1\right)\right)\end{aligned} \tag{4.6}$$

5. EXPRESSIONS THÉORIQUES POUR L'EQM DES ESTIMATIONS BIAISÉES ET NON BIAISÉES

Convenons de noter par $\hat{\theta}_1$ et $\hat{\theta}_2$ les estimations obtenues, respectivement, avec (équation (2.1)) et sans (équation (1.3)) l'ajout du facteur de correction de Duan-Miller. Pour exploiter la forme (4.6) pour comparer les EQM des estimations (1.3) et (2.1) pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$, nous devons disposer de leurs moyennes et de leurs variances, ainsi que d'une expression pour le paramètre à estimer.

À ce propos, il est possible de montrer les identités suivantes (voir Annexe A pour (5.1), Annexe C pour (5.3) et (5.5) et Annexe D pour (5.2) et (5.4)) :

- $$\theta = \mathrm{E}\left(\mathrm{Y}|\mathrm{X}=\mathrm{x}_0\right) = \exp\left(\beta_0 + \beta_1 \mathrm{x}_0 + \frac{\sigma^2}{2}\right) \tag{5.1}$$

- $$\mathrm{E}\left(\hat{\theta}_1\right) = \mathrm{E}\left(\hat{\mathrm{Y}}_{\mathrm{M}}|\mathrm{X}=\mathrm{x}_0\right) = \exp\left(\beta_0 + \beta_1 \mathrm{x}_0 + \frac{\sigma^2}{2}\left(1 + \frac{1}{n} + \frac{\left(\mathrm{x}_0 - \bar{\mathrm{x}}\right)^2}{\mathrm{S}_{\mathrm{xx}}} + \frac{\sigma^2}{2(n-2)}\right)\right) \tag{5.2}$$

- $$\mathrm{E}\left(\hat{\theta}_2\right) = \mathrm{E}\left(\hat{\mathrm{Y}}|\mathrm{X}=\mathrm{x}_0\right) = \exp\left(\beta_0 + \beta_1 \mathrm{x}_0 + \frac{\sigma^2}{2}\left(\frac{1}{n} + \frac{\left(\mathrm{x}_0 - \bar{\mathrm{x}}\right)^2}{\mathrm{S}_{\mathrm{xx}}}\right)\right) \tag{5.3}$$

- $$\mathrm{Var}\left(\hat{\theta}_1\right) = \mathrm{Var}\left(\hat{\mathrm{Y}}_{\mathrm{M}}|\mathrm{X}=\mathrm{x}_0\right) = \exp\left(2\beta_0 + 2\beta_1 \mathrm{x}_0 + \sigma^2\left(1 + \frac{1}{n} + \frac{\left(\mathrm{x}_0 - \bar{\mathrm{x}}\right)^2}{\mathrm{S}_{\mathrm{xx}}} + \frac{\sigma^2}{2(n-2)}\right)\right) \times$$
$$\left[\exp\left(\sigma^2\left(\frac{1}{n} + \frac{\left(\mathrm{x}_0 - \bar{\mathrm{x}}\right)^2}{\mathrm{S}_{\mathrm{xx}}} + \frac{\sigma^2}{2(n-2)}\right)\right) - 1\right] \tag{5.4}$$

- $$\mathrm{Var}\left(\hat{\theta}_2\right) = \mathrm{Var}\left(\hat{\mathrm{Y}}|\mathrm{X}=\mathrm{x}_0\right) = \exp\left(2\beta_0 + 2\beta_1 \mathrm{x}_0 + \sigma^2\left(\frac{1}{n} + \frac{\left(\mathrm{x}_0 - \bar{\mathrm{x}}\right)^2}{\mathrm{S}_{\mathrm{xx}}}\right)\right) \times$$

$$\left[\exp\left(\sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}\right)\right)-1\right] \tag{5.5}$$

où

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{5.6}$$

$$S_{XX}=\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{5.7}$$

Pour alléger la présentation qui suit, réécrivons les identités (5.1) à (5.5) avec :

- $\beta_0+\beta_1 x_0=A$ (5.8)
- $\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}=B$ (5.9)
- $\frac{\sigma^2}{2(n-2)}=C$ (5.10)

pour obtenir :

- $\theta=E(Y|X=x_0)=\exp\left(A+\frac{\sigma^2}{2}\right)$ (5.11)
- $E(\hat{\theta}_1)=E(\hat{Y}_M|X=x_0)=\exp\left(A+\frac{\sigma^2}{2}(1+B+C)\right)$ (5.12)
- $E(\hat{\theta}_2)=E(\hat{Y}|X=x_0)=\exp\left(A+\frac{\sigma^2}{2}B\right)$ (5.13)
- $\mathrm{Var}(\hat{\theta}_1)=\mathrm{Var}(\hat{Y}_M|X=x_0)=\exp(2A+\sigma^2(1+B+C))\times[\exp(\sigma^2(B+C))-1]$ (5.14)
- $\mathrm{Var}(\hat{\theta}_2)=\mathrm{Var}(\hat{Y}|X=x_0)=\exp(2A+\sigma^2B)\times[\exp(\sigma^2B)-1]$ (5.15)

Après quelques manipulations algébriques, l'introduction des quantités (5.11) à (5.15) dans l'équation (4.6) permet de la récrire sous la forme suivante :

$$\Delta EQM = EQM\left(\hat{Y}_M | X = x_0\right) - EQM\left(\hat{Y} | X = x_0\right)$$
$$= \exp(2A) \times \left[\exp\left(\sigma^2(2B + 2C + 1)\right) - \exp\left(2B\sigma^2\right) + 2\exp\left(\frac{\sigma^2}{2}(1 + B)\right) - 2\exp\left(\sigma^2\left(1 + \frac{B + C}{2}\right)\right)\right] \quad (5.16)$$

Ce qu'il importe de connaître au sujet de ΔEQM donné par (5.16) c'est son signe. En effet, si (5.16) est toujours de signe négatif, par exemple, alors nous pourrons en conclure que l'estimateur (2.1) est meilleur que (1.3) en ce sens qu'il donne lieu à des estimations moins dispersées. Puisque le terme exp(2A) de (5.16) est toujours positif, le signe de l'expression (5.16) est le signe que prend l'expression entre crochets dans (5.16) qui dépend des quantités B et C seulement. Le prochain chapitre donne des conditions suffisantes fonction des quantités B et C afin que (5.16) soit positive.

6. COMPARAISON DES EQM

Nous avons vu à la fin du chapitre 5 que le de l'expression (5.16) dépend seulement des quantités B et C. Or, des expressions (5.9) et (5.10) pour B et C, respectivement, il est clair qu'une fois le modèle ajusté, la quantité B est l'unique quantité dans (5.16) qui peut varier dépendant du x_0 qui est fixé. Pour que l'expression (5.16) soit de signe positif il suffit d'avoir simultanément :

$$\exp\left(\sigma^2(2B+2C+1)\right)-2\exp\left(\sigma^2\left(1+\frac{B+C}{2}\right)\right)>0 \tag{6.1}$$

et

$$2\exp\left(\frac{\sigma^2}{2}(1+B)\right)-\exp\left(2B\sigma^2\right)>0 \tag{6.2}$$

La condition (6.1) est équivalente à :

$$\exp\left(\sigma^2(2B+2C+1)\right)>\exp\left(\sigma^2\left(1+\frac{B+C}{2}\right)+\ln(2)\right) \tag{6.3}$$

$$\Leftrightarrow 2B+2C+1>1+\frac{B+C}{2}+\frac{\ln(2)}{\sigma^2} \tag{6.4}$$

$$\Leftrightarrow B>\frac{2}{3}\frac{\ln(2)}{\sigma^2}-C \tag{6.5}$$

La condition (6.2), quant à elle, est équivalente à :

$$\exp\left(\frac{\sigma^2}{2}(1+B)+\ln(2)\right)>\exp\left(2B\sigma^2\right) \tag{6.6}$$

$$\Leftrightarrow 1+B+\frac{2}{\sigma^2}\ln(2)>4B \tag{6.7}$$

$$\Leftrightarrow B < \frac{1}{3} + \frac{2}{3}\frac{\ln(2)}{\sigma^2} \tag{6.8}$$

En combinant (6.5) et (6.8), nous avons alors que l'expression (5.16) est de signe positif si la double condition suivante sur B est satisfaite :

$$\frac{2}{3}\frac{\ln(2)}{\sigma^2} - C < B < \frac{1}{3} + \frac{2}{3}\frac{\ln(2)}{\sigma^2} \tag{6.9}$$

ou, de façon équivalente, si la double condition suivante sur x est satisfaite :

$$\frac{2}{3}\frac{\ln(2)}{\sigma^2} - C < \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} < \frac{1}{3} + \frac{2}{3}\frac{\ln(2)}{\sigma^2} \tag{6.10}$$

Puisque $C = \frac{\sigma^2}{2(n-2)} > 0$, la condition (6.10) peut toujours être satisfaite, pourvu qu'une valeur appropriée pour x_0 soit choisie. Par conséquent, nous avons montré qu'il existe des situations où l'emploi du facteur de réduction de biais proposé par [Miller 1984] peut mener à des estimations de moindre qualité (au sens de l'EQM) que celles obtenues sans qu'une réduction du biais n'ait été appliquée.

7. DISCUSSION ET CONCLUSION

Nous avons montré que l'utilisation du facteur de correction de Duan-Miller pour atténuer le biais dans l'estimation de la moyenne de la distribution conditionnelle étant donné $X=x_0$ ne mène pas nécessairement à des estimations de meilleure qualité au sens de l'EQM. En effet, nous avons montré que pour certaines valeurs de x_0, données par l'équation (6.10), l'EQM pour l'estimation sans facteur de réduction de biais est inférieur à l'EQM pour les estimations rectifiées.

[Miller 1984] fait la remarque que le facteur de réduction de biais $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$ devient de plus en plus important à mesure que la variance de Y s'accroît. Cela semble suggérer que le besoin d'introduire un facteur de réduction de biais est d'autant plus important que la variance de Y est grande, puisqu'alors le biais dans l'estimation est plus considérable. À ce sujet, il est intéressant de noter que le domaine de valeurs de x décrit par (6.10) pour lesquelles l'expression ΔEQM, donnée par (5.16), est positive ne rétrécit pas de façon appréciable à mesure que la variance de Y augmente. Cela signifie que l'utilité d'introduire le facteur de réduction $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$ en raison d'un biais grandissant est annihilée par la forte dispersion dans les estimations que provoque son insertion dans le modèle.

En pratique, il ne semble pas facile de déterminer pour un x_0 donné laquelle des prédictions (1.3) ou (2.1) il est préférable d'utiliser sur la base de (5.16). En effet, les quantités A et C qui apparaissent dans (5.16) dépendent des paramètres β_0, β_1 et σ^2 du modèle considéré et qui sont généralement inconnus. Certes, nous pouvons remplacer A et C en substituant aux paramètres que ces expressions contiennent leurs estimations sans biais correspondantes. Cependant, cette façon de faire n'apparaît pas très sûre. En effet, nous

croyons que les estimations des paramètres β_0, β_1 et σ^2 ne seront pas suffisamment précises pour assurer que le signe de (5.16) établi à partir des estimations soit vraiment le signe que nous aurions obtenu avec les véritables valeurs des paramètres.

Nous croyons avoir montré clairement que la seule considération du biais pour départager deux estimateurs ne donne pas un aperçu fiable de la qualité des estimations à être produites. Même si dans l'expression de l'EQM le biais apparaît au carré, cela ne signifie pas pour autant qu'il soit un facteur prépondérant quant à la qualité des estimations. En fait, la formule suggère plutôt, et c'est ce que nous avons montré dans cette étude, que la qualité des estimations dépend d'au moins deux facteurs: le biais et la variance. Et plus encore, là où le biais parvient à être diminué, la variance s'en trouve accrue.

8. RÉFÉRENCES

Berger, J.O. (1986), *Bayesian Inference and Decision Techniques, edited by P. Goel and A. Zellner,* Elsevier Science Publishers, pp. 473-487.

Casella G., R.L. Berger (1990), Statistical Inference, Duxbury Press.

Duan, N. (1983), Smearing estimate : a non parametric retransformation method, *Journal of the American Statistical Association,* 78 : 605-610.

Dudewicz E.J. (1988), Modern Mathematical Statistics, John Wiley & Sons.

Efron, B. (1975), Biased Versus Unbiased Estimation, *Advances in Mathematics*, 16 pp:259-277.

Efron, B. et C. Morris, 1977, Le paradoxe de Stein, *Pour La Science,* pp. 28-37.

Hoos, A.B. (1996), Improving regional-model estimates of urban-runoff quality using local data, *Water Resources Bulletin,* 32 : 855-863.

Miller, D. M., 1984, Reducing transformation bias in curve fitting, *The American Statistician*, vol.38, no.2, 124-126.

Neyman, J., E. L. Scott, Correction for bias introduced by a transformation of variables, Ann. Math. Statist., 1960, 31, pp. 643-655.

Seber, G.A.F. et C.J. Wild, 1989, *Nonlinear regression*, John Wiley & Sons.

ANNEXE A :

Moyenne de la distribution conditionnelle de Y étant donné

$X=x_0$

Cette annexe fournit les détails qui justifient l'équation (5.1).

Rappelons (1) qui décrit la forme générale du modèle log-linéaire :

$$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon \tag{A1}$$

Par hypothèse, ln(Y) est distribuée normalement et a pour moyenne :

$$E(\ln(Y)) = E(\beta_0 + \beta_1 X) + E(\varepsilon)$$

$$= \beta_0 + \beta_1 X \tag{A2}$$

puisque $E(\varepsilon) = 0$. De plus,

$$\mathrm{Var}(\ln(Y)) = \mathrm{Var}(\varepsilon) = \sigma^2 \tag{A3}$$

Par conséquent, la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$, $E(Y|X = x_0)$, peut être obtenue de (B4) de l'annexe B en utilisant (A2) et (A3):

$$E(Y|X = x_0) = \exp\left(\beta_0 + \beta_1 x_0 + \frac{\sigma^2}{2}\right) \tag{A4}$$

ANNEXE B :

Expressions pour la moyenne et la variance d'une variable log-normale

Si X est une variable aléatoire de densité normale $N(\mu,\sigma^2)$, alors la variable aléatoire $Y = e^X$ est dite log-normale. Nous sommes intéressés à exprimer la moyenne et la variance de Y en fonction de μ et σ^2, respectivement la moyenne et la variance de X.

Il est facile d'établir que la densité de la variable aléatoire Y, $f_Y(y)$ est donnée par :

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma}}\frac{1}{y}\exp\left(\frac{-(\ln(y)-\mu)}{2\sigma^2}\right),\ y>0 \qquad \text{(B1)}$$

Pour obtenir l'espérance et la variance de Y il est utile de rappeler que la fonction génératrice des moments pour la variable X, $M_X(t)$, est :

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)dx = \exp\left(\mu t + \frac{\sigma^2}{2}\right) \qquad \text{(B2)}$$

En effet,

$$E(Y) = \int_0^{\infty}\frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{-(\ln(y)-\mu)^2}{2\sigma^2}\right)dy \qquad \text{(B3)}$$

devient, en posant x=ln(y),

$$E(Y) = \int_{-\infty}^{\infty} e^x\frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)dx$$
$$= M_X(1)$$

$$= \exp\left(\mu + \frac{\sigma^2}{2}\right) \qquad \text{(B4)}$$

Pour établir l'expression pour la variance de Y, nous avons besoin d'établir $E(Y^2)$ puisque

$$Var(Y) = E(Y^2) - E^2(Y) \qquad \text{(B5)}$$

Nous avons,

$$E(Y^2) = \int_0^{\infty} \frac{y}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(\ln(y) - \mu)^2}{2\sigma^2} \right) dy \qquad \text{(B6)}$$

En posant à nouveau x=ln(y), nous obtenons

$$E(Y^2) = \int_{-\infty}^{\infty} \frac{e^{2x}}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x - \mu)^2}{2\sigma^2} \right) dx$$
$$= M_X(2)$$

$$= \exp(2\mu + 2\sigma^2) \qquad \text{(B7)}$$

Par (B5) nous obtenons :

$$Var(Y) = e^{2\mu}\left(e^{2\sigma^2} - e^{\sigma^2}\right) \qquad \text{(B8)}$$

ANNEXE C :
Espérance et variance de la prédiction

Cette annexe présente le détail des calculs pour l'obtention des expressions pour $E(\hat{Y})$ et $Var(\hat{Y})$ présentées en (5.3) et (5.5), respectivement.

Rappelons que le modèle log-linéaire ajusté fournit, pour x_0 donné, l'estimation suivante pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$:

$$\ln(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 x_0 \qquad \text{(C1)}$$

Puisque $\hat{\beta}_0 = \overline{\ln}(Y) - \hat{\beta}_1 \bar{x}$, où $\overline{\ln}(Y) = \frac{\sum_{i=1}^{n} \ln(y_i)}{n}$, et $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$ sont des estimations sans biais pour β_0 et β_1, respectivement, il s'ensuit que :

$$E(\ln(\hat{Y})) = \beta_0 + \beta_1 x_0 \qquad \text{(C2)}$$

Pour obtenir la variance $Var(\hat{Y})$ il vaut mieux récrire l'équation (C1) sous la forme équivalente suivante :

$$\ln(\hat{Y}) = \overline{\ln}(Y) + \hat{\beta}_1 (x_0 - \bar{x}) \qquad \text{(C3)}$$

puisque sous cette forme, contrairement à la forme (C1), l'ordonnée à l'origine et la pente, en raison de l'hypothèse de normalité sur le modèle, sont indépendantes. En effet,

$$\begin{aligned} Cov(\overline{\ln}(Y), \hat{\beta}_1) &= \frac{\sum_i \sum_j Cov(\ln(y_i), \ln(y_j)(x_j - \bar{x}))}{nS_{XX}} \\ &= \frac{\sum_i \sum_{j \neq i} Cov(\ln(y_i), \ln(y_j)(x_j - \bar{x})) + \sum_i (x_i - \bar{x})\sigma^2}{nS_{XX}} \\ &= 0 \end{aligned} \qquad \text{(C4)}$$

puisque $Cov(\ln(y_i), \ln(y_j)) = 0$ (indépendance de $\ln(y_i)$et $\ln(y_j)$pour $i \neq j$) et $\sum_i (x_i - \bar{x}) = 0$.

Par conséquent,

$$Var(\ln(\hat{Y})) = Var(\overline{\ln}(Y)) + (x - \bar{x})^2 Var(\hat{\beta}_1)$$

$$= \frac{1}{n^2}\sum_i \mathrm{Var}(\ln(y_i)) + \frac{(x_0 - \bar{x})^2}{S_{XX}^2}\sum_i (x_i - \bar{x})^2 \mathrm{Var}(\ln(y_i))$$

$$= \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2 \sigma^2}{S_{XX}} \qquad \text{(C5)}$$

Nous avons donc que $\ln(\hat{Y})$ est normalement distribuée dont la moyenne et la variance sont donnés par (C2) et (C5), respectivement. Il s'ensuit que $\hat{Y}$ est de densité log-normale. La moyenne de $\hat{Y}$, $E(\hat{Y})$, est alors, d'après (B4) :

$$E(\hat{Y}) = \exp\left(\beta_0 + \beta_1 X + \frac{\sigma^2}{2}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)\right) \qquad \text{(C6)}$$

et la variance de $\hat{Y}$, $\mathrm{Var}(\hat{Y})$, est donnée par (B8) :

$$\mathrm{Var}(\hat{Y}) = \exp\left(2\beta_0 + 2\beta_1 x_0 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)\right) \times \left[\exp\left(\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)\right) - 1\right] \qquad \text{(C7)}$$

ANNEXE D :

Espérance et variance de la prédiction corrigée pour le biais

Cette annexe présente le détail des calculs pour l'obtention des expressions pour $E(\hat{Y}_M|X = x)$ et $Var(\hat{Y}_M|X = x)$ présentées en (5.2) et (5.4), respectivement.

Rappelons que le modèle log-linéaire ajusté, puis corrigé pour la réduction du biais, fournit, pour x donné, l'estimation suivante pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$:

$$\ln(\hat{Y}_M|X = x) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \exp\left(\frac{\hat{\sigma}^2}{2}\right) \qquad (D1)$$

Puisque $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$ et $\hat{\sigma}^2$ sont des estimations sans biais pour β_0, β_1 et σ^2, respectivement, il s'ensuit que :

$$E(\ln(\hat{Y}_M|X = x_0)) = \beta_0 + \beta_1 x_0 + \frac{\sigma^2}{2} \qquad (D2)$$

Pour obtenir la variance $Var(\hat{Y}_M|X = x_0)$ il est préférable de récrire l'équation (D1) sous la forme équivalente suivante :

$$\ln(\hat{Y}_M|X = x_0) = \overline{\ln}(Y) + \hat{\beta}_1(x_0 - \bar{x}) + \frac{\hat{\sigma}^2}{2} \qquad (D3)$$

puisque sous cette forme, contrairement à la forme (D1), l'ordonnée à l'origine et la pente, en raison de l'hypothèse de normalité sur le modèle, sont indépendantes. La démonstration de ce fait est présentée dans l'annexe C après l'équation (C3). De plus, il est possible de montrer que $\hat{\beta}_1$ et $\hat{\sigma}^2$ sont indépendants [Casella et Berger, 1990, p. 569], comme le sont également $\overline{\ln}(Y)$ et $\hat{\sigma}^2$. L'indépendance $\overline{\ln}(Y)$ et $\hat{\sigma}^2$ est un fait bien connu dont une démonstration élémentaire est présentée dans [Dudewicz 1988].

L'indépendance mutuelle des termes de (D3) justifie alors :

$$\mathrm{Var}\left(\ln\left(\hat{Y}_M|X=x_0\right)\right)=\mathrm{Var}\left(\overline{\ln}(Y)\right)+\left(x_0-\bar{x}\right)^2\mathrm{Var}\left(\hat{\beta}_1\right)+\frac{\mathrm{Var}\left(\hat{\sigma}^2\right)}{4} \qquad \text{(D4)}$$

$$=\frac{1}{n^2}\sum_i \mathrm{Var}\left(\ln\left(y_i\right)\right)+\frac{\left(x_0-\bar{x}\right)^2}{S_{XX}^2}\sum_i\left(x_i-\bar{x}\right)^2\mathrm{Var}\left(\ln\left(y_i\right)\right)+\frac{\mathrm{Var}\left(\hat{\sigma}^2\right)}{4} \qquad \text{(D5)}$$

La variance de l'estimation $\hat{\sigma}^2$, $\mathrm{Var}\left(\hat{\sigma}^2\right)$, s'obtient en établissant la distribution de $\hat{\sigma}^2$. En effet, de [Casella et Berger 1990 p.569], nous avons que la variable aléatoire $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ est distribuée selon un khi-deux χ^2_{n-2} à (n-2) degrés de liberté. Par conséquent,

$$\mathrm{Var}\left(\hat{\sigma}^2\right)=\mathrm{Var}\left(\frac{\sigma^2}{n-2}\frac{(n-2)\hat{\sigma}^2}{\sigma^2}\right)=\frac{\sigma^4}{(n-2)^2}\mathrm{Var}\left(\chi^2_{n-2}\right)=\frac{\sigma^4 2(n-2)}{(n-2)^2}=\frac{2\sigma^4}{n-2} \qquad \text{(D6)}$$

Puisque $\ln\left(\hat{Y}_M\right)$ est normalement distribuée dont la moyenne et la variance sont donnés par (D2) et (D6), respectivement. Il s'ensuit que $\hat{Y}_M$ est de densité log-normale. La moyenne de $\hat{Y}_M$, $E\left(\hat{Y}_M|X=x_0\right)$, est alors, d'après (A4) :

$$E\left(\hat{\theta}_1\right)=E\left(\hat{Y}_M|X=x_0\right)=\exp\left(\beta_0+\beta_1 x_0+\frac{\sigma^2}{2}\left(1+\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{S_{XX}}+\frac{\sigma^2}{2(n-2)}\right)\right) \qquad \text{(D7)}$$

et la variance de $\hat{Y}_M$, $\mathrm{Var}\left(\hat{Y}_M|X=x_0\right)$, est donnée par (A8) :

$$\mathrm{Var}\left(\hat{Y}_{corr}\right)=\exp\left(2\beta_0+2\beta_1 x_0+\sigma^2\left(1+\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{S_{XX}}+\frac{\sigma^2}{2(n-2)}\right)\right)\times$$

$$\left[\exp\left(\sigma^2\left(\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{S_{XX}}+\frac{\sigma^2}{2(n-2)}\right)\right)-1\right] \qquad \text{(D8)}$$