École d'actuariat

Université Laval

Modèles linéaires en actuariat

ACT-2003

Notes de cours de
Marie-Pier Côté

Rédigées en partie par
Christian Genest et Anne-Catherine Favre

Automne 2014

Remerciements

La première partie du document, sur la régression linéaire simple et multiple, **provient en grande partie des notes de cours STT-2100 rédigées par Christian Genest et Anne-Catherine Favre.** Je suis grandement reconnaissante à Christian qui m'a autorisé à utiliser ses notes et à les adapter pour ce cours d'actuariat.

La seconde partie du document est composée d'exemples pour les modèles linéaires généralisés, que j'ai trouvé dans des livres, ou que j'ai composé pour vous. Mes explications seront probablement teintées de celles des professeurs David Stephens (Département de mathématiques et de statistiques, Université McGill) et Jean-Philippe Boucher (Groupe de recherche en mathématiques financières et actuarielles, UQÀM) qui m'ont bien enseigné ces concepts.

Chapitre 1

Introduction

1.1 Un peu d'histoire

L'origine du mot *régression* vient de la génétique, plus précisément de Sir Francis Galton (1822–1911) qui était un homme de science britannique. Cousin de Charles Darwin, Galton était un touche-à-tout intuitif qui fut anthropologue, explorateur, géographe, inventeur, météorologue, généticien, psychométricien et statisticien [1]. En 1886, en travaillant sur l'hérédité, il s'est intéressé à la taille des enfants par rapport à leurs parents. Il constata que lorsque les parents étaient plus grands que la moyenne, leurs enfants avaient tendance à être plus petits qu'eux et, dans le cas contraire, lorsque les parents étaient plus courts que la moyenne, leurs enfants tendaient à être plus longs qu'eux. Il publia sa découverte dans l'article *Regression towards mediocrity in heriditary stature* **?** [2]. Il utilisa le terme régression dans le sens où les parents de taille extrême (petits et grands) voient en moyenne la taille de leurs enfants régresser vers la moyenne.

Cependant, bien que l'origine du mot régression remonte à F. Galton, l'analyse de causalité entre plusieurs variables est bien plus ancienne et remonte au milieu du XVIII$^{\text{ème}}$ siècle. En 1757, Ruder Josip Boščović (1711–1787), exposa une méthode minimisant la somme des valeurs absolues entre un modèle de causalité et les observations. Plus tard Adrien-Marie Legendre (1752–1833), célèbre mathématicien français, développa dans son mémoire de 1805, intitulé *Nouvelles méthodes pour la détermination des orbites des comètes*, la méthode d'estimation par moindres carrés des paramètres d'un modèle de causalité. Il donna ainsi le nom à la méthode. Parallèlement, Carl Friedrich Gauss (1777-1855) publia en 1809 dans le tome 2 de ses travaux sur la mécanique céleste, nommé *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*, le développement de la méthode des moindres carrés.

1. Plus d'informations sur la biographie de Sir Francis Galton sont disponibles à l'adresse `http://fr.wikipedia.org/wiki/Francis_Galton`
2. Disponible à l'adresse web suivante `http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf`

1.2 Buts de la régression

Dans plusieurs domaines, (incluant l'actuariat et la finance) on se pose la question de l'effet des variables les unes sur les autres. Par exemple, on peut citer :

- Quel est l'effet de la pollution par l'ozone (O_3) sur la santé ?
- Quel est l'effet de la température en hiver sur la consommation d'électricité à Québec ?
- Quel est le lien entre le revenu moyen d'un pays et l'espérance de vie de sa population ?

La figure 1.1 illustre par exemple l'effet de l'argent sur la santé en montrant le lien entre le revenu moyen domestique d'un pays et le nombre d'enfants survivants au-delà de leur cinquième année. Il est à souligner que cette figure est riche en informations car la couleur du cercle indique le continent d'appartenance du pays alors que sa taille montre le nombre d'habitants. Elle a été tracée dans la cadre du projet `Gapminder` développé en Suède dans le but de produire un logiciel interactif permettant de visualiser les liens entre diverses variables tels l'argent, la santé et le développement.

Le but d'une analyse de régression consiste à étudier les relations qui existent entre des variables (ou facteurs) mesurables à partir d'observations (données) prises sur ces variables.

Les objectifs d'une telle analyse sont multiples et comprennent par exemple :

1. **La spécification de modèle**

   Elle consiste à décire le lien entre les variables. Exemple : Comment la température de fonctionnement d'un procédé influence-t-elle le rendement du produit ?

2. **L'estimation de paramètres**

   Exemple : La relation de Faber-Jackson [3] est une loi empirique qui relie la luminosité L à la dispersion des vitesses σ des étoiles centrales des galaxies elliptiques. Mathématiquement elle s'exprime comme $L = k\sigma^{\gamma}$, où γ est un indice approximativement égal à 4, selon la catégorie de galaxies considérées et k est un paramètre du modèle. Le but consiste à estimer k et γ puis à tester si γ est non significativement différent de 4 avec nos données. Il est à noter qu'il est facile de linéariser la relation à prenant le logarithme à gauche et à droite de l'égalité. Nous obtenons ainsi $\log L = \gamma \log \sigma + \log k$.

3. **La sélection de variables**

   Exemple : Parmi les données disponibles sur un assuré tels que le sexe, l'âge, la profession, etc., quelles variables ont une influence significative sur les réclamations en assurance automobile ?

4. **La prévision**

   Exemple : Étant donné les caractéristiques d'un assuré et de sa maison, quel sera le coût des réclamations en assurance habitation pour l'année 2014 ?

3. Voir `http://fr.wikipedia.org/wiki/Relation_de_Faber-Jackson`

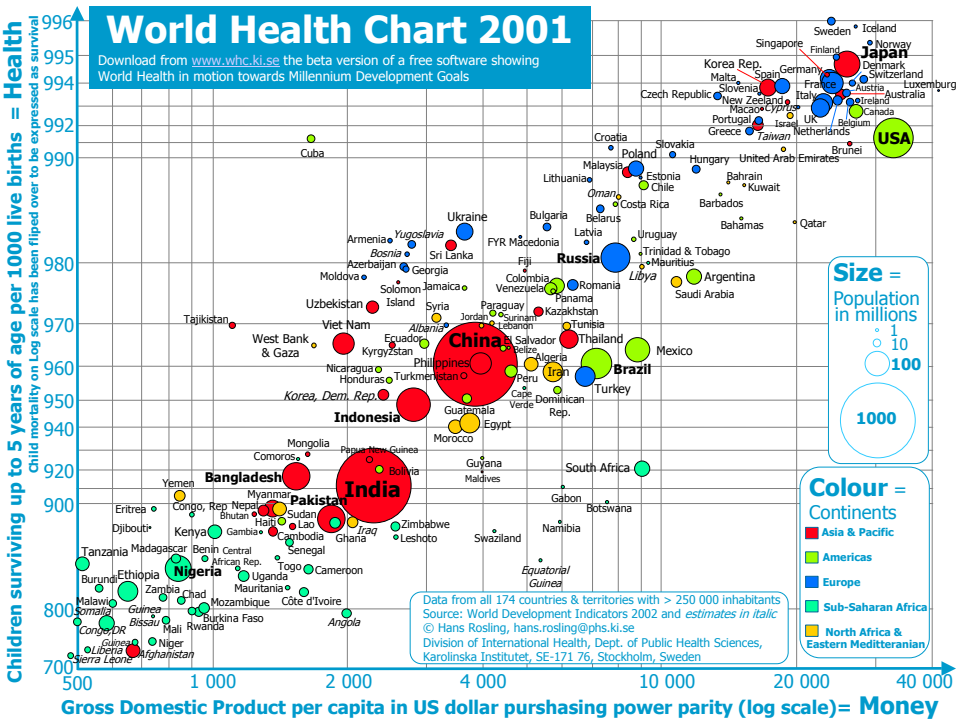

FIGURE 1.1: Graphique en nuage de points montrant le lien entre le revenu domestique moyen d'un pays et le nombre d'enfants survivants au-delà de leur cinquième année. La couleur du cercle indique le continent d'appartenance du pays et la taille caractérise le nombre d'habitants. Source :http ://www.math.yorku.ca/SCS/Gallery/

1.3 Modèle de régression

Dans ce cours toutes nos analyses et conclusions seront fondées sur un *modèle de régression*.

La régression est une technique de modélisation qui vise à expliquer une variable Y en fonction d'une ou de plusieurs variables $x_1, \ldots, x_p$. Cela peut s'exprimer mathématiquement par

$$Y = f(x_1, \ldots, x_p). \tag{1.1}$$

La variable Y est appelée *variable endogène*[4] (ou également variable réponse, variable dépendante). Les

4. Selon le Larousse, le terme *endogène* signifie : Qui est produit par la structure elle-même en dehors de tout apport extérieur, par opposition à exogène

variables $x_1, \ldots, x_p$ sont nommées *variables exogènes*[5]. Dans la littérature les termes « variables explicatives », « facteurs », « covariables », et « variables indépendantes »constituent des synonymes de variables exogènes.

Dans un monde idéal, la relation serait exacte et donc la valeur de Y serait dans ce cas uniquement déterminée par les valeurs de $x_1, \ldots, x_p$. Le modèle de régression serait alors un modèle dit *déterministe*.

Exemple 1.1. *Un grand nombre de lois physiques sont déterministes. Par exemple, la modèle d'atmosphère standard de l'Organisation de l'aviation civile internationale (OACI), couvrant les altitudes jusqu'à 80 [km] et ne prenant pas en compte la vapeur d'eau, exprime la relation entre la pression et l'altitude de la manière suivante :*

$$p = p_0(1 - 2,26 \times 10^{-5} h),$$

où p_0 désigne la pression au niveau de la mer en hPa ($p_0 = 1013,25\, hPa$) et h note l'altitude en m.

Dans le cas contraire, il est possible que les données soient entachées d'une erreur de mesure (expérimentale) ou encore, que certains facteurs contributifs de moindre importance aient été négligés. Dans un cas général, les valeurs de $x_1, \ldots, x_p$ peuvent expliquer une partie de Y, alors que l'autre partie demeure inexpliquée.

Exemple 1.2. *Soit Y, le poids d'un bébé fille à la naissance en kg et soit x sa taille en cm. Dans ce cas, on peut écrire $Y = f(x)$+fluctuation aléatoire, étant donné que deux bébés de même taille, par exemple 50 cm, peuvent avoir des poids très différents.*

1.4 Régression linéaire

Dans la première partie de ce cours nous nous intéressons d'abord aux modèles de régression linéaire. Dans un tel modèle la valeur de la variable endogène est une fonction linéaire des paramètres :

$$Y = \beta_0 + \beta_1 f(x_1) + \beta_2 f_2(x_2) + \cdots + \beta_p f_p(x_p) + \text{fluctuation}, \tag{1.2}$$

où Y dénote la variable endogène,
$x_1, \ldots, x_p$ représentent les variables exogènes,
$f_1, \ldots, f_p$ sont des transformations connues des variables exogènes et
$\beta_0, \ldots, \beta_p$ sont des paramètres de valeur inconnue qu'il s'agira d'estimer à l'aide des données.

Dans le cas où une seule variable explicative (exogène) est considérée, le modèle est appelé *régression linéaire simple*. Ce modèle largement utilisé fait l'objet du chapitre **??** de ces notes de cours. Dans le cas inverse où plusieurs variables explicatives sont nécessaires pour modéliser la variable réponse, le modèle est appelé *régression linéaire multiple*. Ce modèle est exposé dans le chapitre **??**.

5. Selon le Larousse, le terme *exogène* signifie : Qui provient du dehors, de l'extérieur du phénomène, par opposition à endogène

1.5 Analyse de régression

Supposons que l'on se place, par souci de simplicité, dans le cadre de la régression linéaire simple. Plusieurs questions potentielles se posent lors de la mise en œuvre de l'analyse de régression, dont

- Est-ce raisonnable de supposer qu'il existe une relation linéaire entre Y et x ? Un modèle du type $Y = \beta_0 + \beta_1 x_1 +$ fluctuation est-il approprié ? Il s'agit pour cela de vérifier le respect d'un certain nombre de postulats.
- Comment choisir les paramètres β_0 et β_1 ? Il s'agira de les estimer par exemple à l'aide de la méthode des moindres carrés ou du maximum de vraisemblance.
- A-t-on vraiment besoin de β_0 et β_1 dans le modèle ? Il est possible de développer des tests d'hypothèse permettant de tester la significativité des paramètres. Si, par exemple, β_0 est non significativement différent de zéro, il est alors possible de simplifier le modèle.
- Comment mettre en œuvre le modèle de régression pour effectuer une prévision ? Comment calculer l'erreur autour de cette prévision ?
- Comment mesurer l'effet d'une variable dans le modèle ? Il s'agira pour cela d'interpréter les paramètres estimés.