Les questions scientifiques en médecine ont besoin de modèles statistiques

Régression linéaire et régression logistique

Ulrike Held

Horten Zentrum, UniversitätsSpital, Zürich

Dans le domaine de la recherche, il est souvent intéressant de trouver une relation de cause à effet, à savoir d'examiner l'influence d'un paramètre (comme le poids) sur une valeur cible (comme la tension artérielle systolique). Ou de savoir si certaines habitudes alimentaires ont une influence sur l'apparition de certaines maladies. Pour répondre à cette question, ou à d'autres du même type, il est possible de recourir à des modèles de régression statistiques permettant d'étudier l'influence d'un ou plusieurs paramètres sur une variable cible. Un modèle statistique ad hoc est utilisé en fonction de la modalité de cette variable, à savoir si elle se déplace sur une échelle continue (métrique), si elle est dichotomique («oui/non») ou si elle dépend d'une durée (le temps écoulé jusqu'à la manifestation de tel ou tel incident). Le choix du modèle suppose la compétence clinique du médecin et un certain know-how biostatistique.

Dans cet article, nous décrivons les modèles statistiques peut-être les plus courants en recherche médicale: la régression linéaire et la régression logistique.

Régression linéaire

Par modèle linéaire, nous entendons que la relation entre le paramètre d'influence et la valeur cible peut typiquement se décrire de la manière suivante:

Valeur cible = a + b × paramètre d'influence

L'interprétation de cette équation est la suivante: étant donné un paramètre d'influence fixe, non aléatoire (par ex. le poids), la valeur cible (ici la tension systolique) est en principe donnée par la formule ci-dessus. La linéarité se rapporte ici à *la relation* entre les deux (une droite caractérisée par les paramètres a et b), et non pas au paramètre d'influence lui-même. Un modèle de régression «linéaire» permet aussi d'examiner la relation entre a + b × (paramètre d'influence)² et la valeur cible.

Le but de l'analyse de régression est d'estimer les paramètres inconnus a et b, correspondant à la section d'axe et à la pente de la droite dite de régression.

Soit les données fictives de 20 observations de poids, de tension artérielle systolique et d'âge chez des hommes, comme illustré au tableau 1 ⬅.

Les 18 paires de données complètes de poids et de tension artérielle peuvent être reportées dans un diagramme de dispersion ou «nuage de points», comme le montre la figure 1 📷. Une mesure du poids et une autre de tension artérielle doivent être exclues car la donnée correspondante est absente. Les estimations de la section d'axe a et de la pente b de la droite de régression donnent l'équation de régression suivante:

Tension artérielle systolique = 41,2 + 0,95 × poids

La droite de régression qui en résulte a également été tracée sur la figure 1.

La figure 1 montre une droite croissante: la relation entre la tension artérielle et le poids est positive, car la pente de la droite de régression (b = 0,95) est supérieure à zéro. Dans le cas contraire, on parlerait de relation négative.

Il y a naturellement au-delà de la relation linéaire une certaine déviation due au hasard pour chaque individu. La différence entre la valeur prédite par le modèle et l'observation effective est appelée terme d'écart. Il est en moyenne égal à zéro pour tous les patients mais a une certaine variation. A titre d'exemple, le terme d'écart pour la personne ayant le poids le plus bas a été reporté sur la figure, il s'agit de la distance verticale entre l'observation réelle et la droite de régression. Concrètement, dans ce cas, la valeur effectivement mesurée de la tension artérielle était de 17,6 mm Hg, à savoir plus élevée que celle estimée par le modèle de régression.

Pour savoir maintenant encore si le poids a une influence *significative* sur la valeur cible, c.-à-d. si cette influence est «statistiquement certainement» différente de zéro, il faut considérer le seuil de probabilité (appelé *valeur p*) de l'hypothèse faite sur la pente. La valeur p qui correspond est de 0,04. La relation entre les variables poids/tension est donc significative car le seuil de rejet de l'hypothèse sur la pente (5%) n'est pas atteint.

En recourant à un modèle statistique cherchant à décrire une relation importante en médecine, il faut toujours se demander si le choix du modèle a été «bon», ou quel est effectivement l'écart entre l'observation réelle et la valeur estimée par toutes les paires d'observations. Dans le cas d'un modèle de régression linéaire, il est donc judicieux de calculer le coefficient R^2 qui décrit la proportion d'explication de la variation de la valeur cible (tension artérielle) par la variation du paramètre d'influence (poids). Si nous voulons juger la qualité d'adaptation d'un modèle, nous pouvons dire qu'un R^2 proche de zéro est une adaptation non satisfaisante, alors qu'un R^2 proche de 1 signifie une bonne adaptation linéaire. Dans notre exemple, R^2 est de 0,23, soit 23%.

Ulrike Held

Si nous obtenons un R^2 très bas, cela pourrait provenir du fait que la relation entre le poids et la tension artérielle n'est pas vraiment linéaire. La représentation graphique de cette relation ne permet pas vraiment d'en juger. Il se pourrait aussi que d'autres importantes variables manquent encore pour expliquer la variable cible «tension artérielle systolique». Si nous voulons élargir le modèle de régression linéaire à plusieurs paramètres d'influence, nous parlons alors de régression multiple au lieu de simple.

Si, dans notre exemple, nous prenons l'âge des patients comme second paramètre d'influence, le modèle de régression donne:

Tension systolique = a + b × poids + c × âge

Le but consiste maintenant à estimer les paramètres a, b et c, c.-à-d. à quantifier l'influence à la fois du poids et de l'âge sur la tension systolique.

Le résultat donne une estimation de b = 0,95 (influence du poids), avec un p de 0,047 et de c = 0,38 (pour l'âge), avec un p de 0,50. Un modèle de régression multiple convient donc bien ici. Le paramètre de poids est pratiquement identique à celui donné par le modèle de régression simple et reste statistiquement significatif. Le paramètre supplémentaire âge n'est pas significatif (p supérieur à 5%). Il se peut dans quelques cas que l'influence de chaque coefficient de régression change nettement si d'autres variables sont inclues dans le modèle. Si nous reprenons la qualité d'adaptation, nous constatons que R^2 s'améliore à 26% en ajoutant l'âge.

Un autre aspect qui n'a pas encore été évoqué est la généralisation éventuelle des relations linéaires d'un modèle de régression. Dans nos données fictives sur 20 patients, la fourchette *(range)* ou l'intervalle des données sur le poids va de 73 à 115 kg. Les valeurs correspondantes de tension systolique se situent entre 90 et 168 mm Hg. Ces données ont fourni un modèle de régression linéaire mais les résultats ne peuvent être extrapolés sans autre à tous les autres intervalles de ces variables. Selon ce modèle, une personne pesant 60 kg aurait une valeur prédictive de tension de 98 mmHg, mais le type de relation dans les valeurs inférieures de poids pourrait également être tout autre que linéaire.

Tableau 1. Poids, taille et âge des 20 hommes.

Patient	Poids (kg)	Tension artérielle syst. (mm Hg)	Age (ans)
1	92,8	153,6	70
2	82,9	116,2	72
3	101,6	157,8	74
4	97,7	153,5	55
5	111,3	153,2	57
6	manque	123,3	59
7	73,3	128,4	61
8	87,2	manque	59
9	114,7	126,1	71
10	113,1	167,9	66
11	97,4	130,5	67
12	90,2	141,2	70
13	95,0	109,9	54
14	89,4	89,8	53
15	90,4	137,6	62
16	92,2	114,1	74
17	105,1	130,6	69
18	89,4	138,1	80
19	88,7	109,4	79
20	86,3	103,5	58

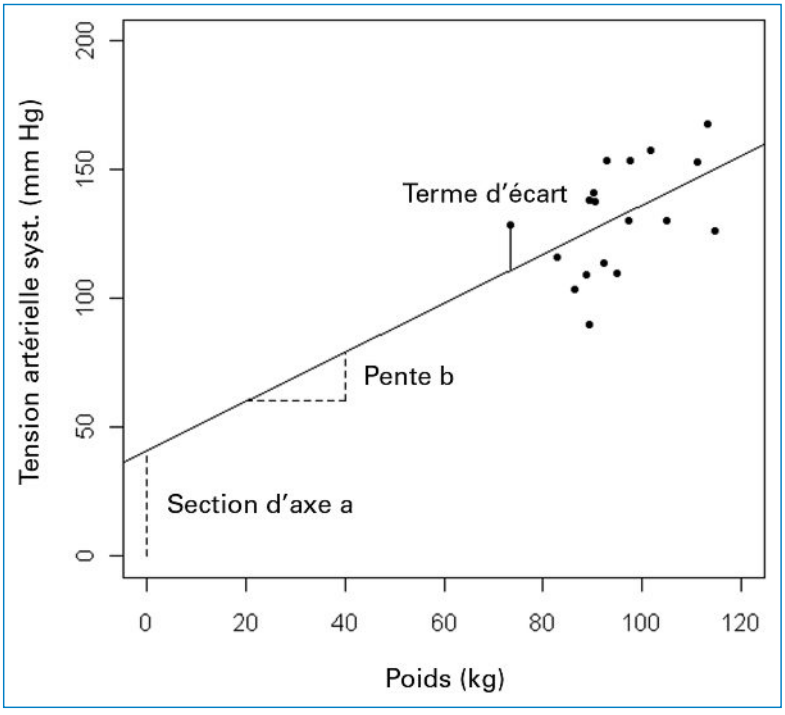

Figure 1
Diagramme de dispersion des 18 paires observées avec droite de régression linéaire.

Modèle de régression logistique

Contrairement au modèle de régression linéaire, un modèle de régression logistique est utilisé si la variable cible n'est pas continue, mais peut n'avoir que deux caractéristiques (par ex. oui/non ou présent/absent). La théorie sur la régression logistique est tout aussi complète que celle sur la régression linéaire mais les différences ne sont pas particulièrement importantes pour l'utilisateur et résident principalement dans l'interprétation des résultats: dans la régression linéaire, l'effet des paramètres d'influence sur la variable cible peut se déduire directement du coefficient estimé. Dans notre exemple, l'effet du poids a été de 0,95, c.-à-d. que la tension systolique augmente de 9,5 mm Hg par tranche de 10 kilos. Dans la régression logistique par contre, la variable cible (avec ses deux caractéristiques possibles) n'est pas directement modélisée mais est fonction de la probabilité que l'événement se produise dans des conditions (de risque) données. Les effets estimés du/des paramètre/s d'influence doivent alors être interprétés selon un risque relatif (ou Odds ratio), que nous expliquerons séparément dans un futur article.

En plus des régressions linéaire et logistique, nous présenterons d'autres modèles tels que la régression de Cox pour l'analyse des survies, des modèles d'analyse de variance et de séquence chronologique.

Glossaire

p

Après avoir utilisé un test statistique d'examen d'une hypothèse nulle (à savoir pas de relation entre les variables) sur la base des données, nous obtenons la valeur de la statistique du test. Le p indique à quel point cette valeur est «extrême» et correspond à la probabilité d'obtenir la valeur calculée ou une valeur encore plus extrême si en réalité c'est l'hypothèse nulle qui est la bonne.

R^2

La mesure (multiple) de définition R^2 donne la proportion de variabilité dans les variables cibles pouvant être expliquée par le modèle (donc par les paramètres d'influence). Plus la variabilité peut être expliquée, plus la variable cible est adaptée par les paramètres d'influence.

Range

Le terme anglais «*range*» décrit la dispersion effective des données, du minimum au maximum.

Signification statistique

Les différences entre statistiques sont significatives si la probabilité qu'elles soient le fruit du hasard, donc sans qu'il n'y ait de différence effective, est très faible. On souhaite typiquement obtenir que la probabilité d'erreur soit inférieure à 5% ou 1%, et on parle alors d'un résultat significatif. Le p (voir ci-dessus) est alors <5% resp. à <1% (soit p <0,05 ou <0,01).

Correspondance:
Dr rer. nat. Ulrike Held
Horten Zentrum
UniversitätsSpital Zürich
Postfach Nord
CH-8091 Zürich
ulrike.held@usz.ch

Références recommandées

– Held L, Rufibach C, Seifert B. Einführung in die Biostatistik. 4. Auflage. Zürich: Abteilung Biostatistik, Institut für Sozial- und Präventivmedizin der Universität Zürich; Juli 2009. http://www.biostat.uzh.ch.
– Hüsler J, Zimmermann H. Statistische Prinzipien für medizinische Projekte. 4. Auflage. Bern: Huber-Verlag; 2006.
– Kreienbrock L, Schach S. Epidemiologische Methoden. 4. Auflage. München: Elsevier-Verlag; 2005.
– R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN 3-900051-07-0, URL http://www.R-project.org.