REGRESSION LINEAIRE SIMPLE

Veillez à avoir lu les sections relatives à cette séance dans le document disponible sur le site « Introduction aux séances de TP de biostatistique » avant de vous présenter au TP.

Exercice 1

La régression linéaire permet de décrire un ensemble de données par une relation linéaire entre la variable dépendante Y et la ou les variables indépendantes X_i. Y est donc une fonction du (régression linéaire simple) ou des (régression linéaire multiple) X_i. R et Excel nous permettent de calculer les paramètres de la régression, d'en tester la signification et de faire une représentation graphique des résultats. Nous étudierons ici la situation simple avec une seule variable Xi (régression linéaire simple). Pour cela, nous allons générer 21 couples de données (X_i, Y_i) de manière aléatoire, puis étudier la relation entre les deux variables Y et X.

a) Génération des données

Si la relation entre Y et X est linéaire, elle peut s'écrire :

$$Y=\beta * X+\beta_0 \text{ où :}$$

- β est la pente de la droite décrivant la relation (quand X augmente de 1 unité, Y augmente de β unités),
- β_0 est l'intercept, c'est-à-dire la valeur de Y quand X = 0.

Par exemple, on peut émettre l'hypothèse que le poids des porcs est une fonction linéaire du périmètre thoracique (PT). En pratique, on nepeut cependant pas affirmer que deux porcs de PT identiques ont nécessairement exactement le même poids : en fait, on supposera que les moyennes des poids des porcs de PT = X varient de manière linéaire avec Y :

$$E[Y(X)] = \beta * X +\beta_0 \text{ où :}$$

- E[Y(X)] est l'espérance mathématique, c'est-à-dire la moyenne, de Y pour un X donné.

Pour un porc particulier, le poids ne sera pas forcément égal à la moyenne des poids des porcs de même PT : il y aura donc un écart aléatoire, en moyenne nul, entre le poids de ce porc et le poids moyen des porcs de même PT, écart noté e_i pour le porc pesant Y_i :

$$Yi = \beta *X_i + \beta_0 + e_i$$

Dans les situations pratiques, β et β_0 sont des paramètres inconnus, qu'on essaiera d'estimer sur base des données récoltées. Dans notre problème, nous supposerons dans un premier temps ces paramètres connus et nous les utiliserons pour générer aléatoirement nos 21 couples de données. On supposera donc que β = 0,7 et que β_0 = 25. On supposera en outre que les valeurs de PT varient de 100 à 120 cm par incréments de 1 cm (soit, 21 valeurs de X_i) et que les écarts e_i proviennent d'une distribution normale de moyenne nulle (voir plus haut) et de déviation standard 15.

Dans Excel : utilisez la fonction *LOI.NORMALE.INVERSE()* pour générer les Y en utilisant les paramètres donnés ci-dessus.

Dans R : créez le vecteur de périmètres thoraciques X et ensuite utilisez la fonction *rnorm(n,m,s)->Y* pour générer les poids. Pour rappels : n est le nombre de poids à générer, m est la moyenne de la distribution des poids (qui, vu ce qui précède, devra donc dépendre de la valeur de X correspondante...) et s est la déviation standard de cette distribution.

b) Graphique

A partir de votre échantillon de 21 couples (X,Y), il s'agit à présent d'essayer d'estimer les paramètres de la droite supposée afin de relier ces Y aux X. Dans cet exemple, les « vraies » valeurs de β et β_o sont connues, mais, à partir d'ici, nous supposerons qu'elles sont inconnues et que vous devez donc les estimer. Vous commencerez par une représentation graphique de vos données, sous forme d'un nuage de points non reliés, pour voir si la tendance attendue se confirme…

Dans Excel : après avoir sélectionné vos données, menu Insertion > Graphique > Nuage de poids (XY)

Dans R : fonction *plot(X,Y)*. Essayez de formater au mieux le graphique. Cela peut se faire en ajoutant des arguments à la fonction graphique (pour cela n'hésitez pas à utiliser l'aide : *?plot*).

c) Régression linéaire

Pour pouvoir tracer la droite de tendance (droite de régression) et prédire les valeurs attendues en utilisant cette droite, vous devez au préalable calculer les estimateurs des paramètres de cette droite, c'est-à-dire les estimateurs b et b_0 des coefficients de régression β et β_0. Les formules sont données dans le cours théorique : $b = \Sigma x*y / \Sigma x^2$ et $b_0 = Y_m - b*X_m$ (Y_m et X_m représentent la moyenne des Y et des X, respectivement. x représente l'écart entre X et X_m, et y l'écart entre Y et Y_m).

Dans Excel : faire un tableau avec les sommes des x, x^2, y, y^2 et xy, après avoir naturellement calculé les moyennes des X et des Y. En utilisant les formules vues plus haut, vous obtiendrez vos b et b_0. Vous pourrez ainsi calculer les Y prédits par la droite de régression à partir des mêmes valeurs de X. Votre droite pourra ensuite être superposée au premier graphique réalisé (*clic droit sur le graphique > sélectionner des données > ajouter*).

Dans R : pour obtenir les valeurs prédites par le modèle afin de tracer la droite de régression, il est plus facile de passer directement par la fonction *lm()* expliquée ci-dessous. L'accès aux coefficients et aux valeurs prédites est alors assez facile. Sinon, il faut passer par les mêmes calculs qu'avec Excel afin d'obtenir b et b_0 …

d) Test de la régression

Vous avez trouvé un coefficient de régression correspondant à vos données. L'hypothèse nulle H_0 (qui est "il n'y a pas de régression linéaire", ou encore, "Y ne dépend pas (linéairement) de X") prédisait $\beta = 0$. La valeur obtenue (b) s'écarte-t-elle significativement de cette valeur, ou bien est-elle simplement le résultat d'une fluctuation statistique n'impliquant pas une vraie dépendance ? Pour répondre à cette question, il faut calculer la probabilité d'obtenir une telle valeur du coefficient de régression si l'hypothèse nulle d'indépendance linéaire entre X et Y est vraie. Pour cela, deux possibilités équivalentes (c'est-à-dire qui conduiront à la même conclusion) existent :

- ***la statistique t***
- ***l'analyse de la variance***

Dans Excel : pour tester la signification, vous aurez besoin des valeurs attendues de Y (Poids) d'après la droite de régression trouvée au point précédent (aussi appelées valeurs prédites ou Y prédits). La différence entre ces Y prédits et les Y effectivement fournis est appelée « erreur du modèle ». La « somme des carrés des erreurs », appelée SCE, est donc simplement la somme de ces différences élevées au carré. Cette somme constitue en fait une mesure de la différence entre ce qu'on observe réellement et ce qui est attendu par la régression. Cette valeur va servir de base au calcul de la variance erreur et à la conclusion de notre test.

Comme d'habitude (?), la statistique t est une mesure standardisée de l'écart entre ce qui était prévu et ce qui est effectivement observé. Elle se calcule par :

$$t = (X - \mu_X)/s_X$$

où X est un estimateur, μ_X est la valeur attendue de X (c'est-à-dire la moyenne de X) et s_X est sa déviation standard. Dans le cas qui nous occupe ici, cette formule devient :

$$t = b/S_b$$

puisque E(b) = β et, si H0 est vraie, β = 0.La déviation standard de l'estimateur S_b se calcule en divisant la variance erreur par la somme des x^2 puis en prenant la racine carrée du résultat obtenu (voir théorie si nécessaire). La variance erreur est la variance liée aux fluctuations: nous l'avions choisie égale à 15^2 dans l'énoncé, mais nous la supposerons inconnue, et nous devons l'estimer. Pour cela, nous diviserons notre SCE évoquée plus haut par le nombre de degrés de liberté lié à l'erreur (c'est-à-dire le nombre de couples de données moins 2). On obtient alors une valeur de t calculée sous l'hypothèse nulle H0: β = 0. On peut ensuite tester s'il est probable ou non d'obtenir une telle valeur de t. Si c'est improbable (p < 0,05), on rejettera l'hypothèse nulle, indiquant qu'il y a bien régression linéaire de Y sur X (la régression est qualifiée alors de significative). Vous pourrez obtenir cette probabilité en utilisant la fonction *LOI.STUDENT(t ;ddl ;2)*, où le dernier paramètre signifie qu'on souhaite effectuer un calcul bilatéral de probabilité (on se demande s'il est probable d'avoir une valeur qui s'écarte plus fort de 0 que la valeur de t calculée : soit en étant plus négative si t est négatif, soit en étant plus positive si t est positif).

Pour faire le tableau d'analyse de la variance, il faut se rappeler que la variation dans les données peut être scindée en deux parties : une partie due à la régression et une autre partie due à la déviation par rapport à la droite de régression (qu'on a appelée l'erreur ci-dessus).

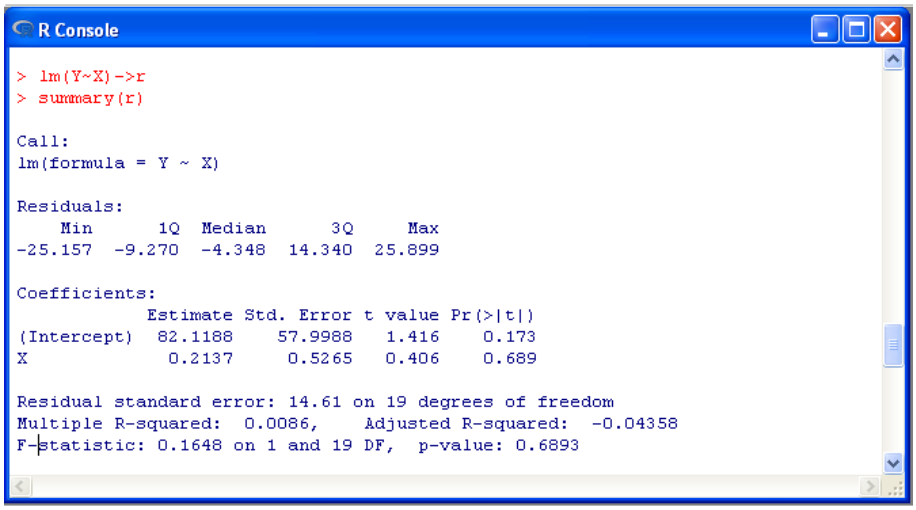
Mathématiquement: $(Y_i - Y_m) = (Y_i - Yp_i) + (Yp_i - Y_m) = d_i + yp_i$ où Yp_i désigne la valeur prédite par la régression en $X = X_i$ et $yp_i = Yp_i - Y_m$.

En élevant cette expression au carré, et en sommant sur toutes les données, le double produit s'annule (cf. cours théorique) et on obtient :

$\Sigma y_i^2 = \Sigma d_i^2 + \Sigma yp_i^2$, souvent exprimé sous la forme SCTotaux = SCErreur + SC Régression.

Après avoir calculé ces trois composantes, le reste de la table d'analyse de la variance de la régression se construit aisément (cf. ANOVA1). La probabilité calculée via cette table est-elle identique à celle obtenue par le test de t ?

Dans R : on peut naturellement faire les calculs de manière tout à fait similaire à Excel comme évoqué plus haut, mais ce programme offre des possibilités statistiques plus importantes. Nous pouvons donc tester notre régression linéaire directement via la fonction *lm(Y~X)->r*. La fonction *lm()* est une fonction qui permet de tester des modèles linéaires (linear models). On lui spécifie le modèle de régression linéaire comme indiqué dans la commande ci-dessus : "~" indique "dépend de". Notre fonction est enregistrée dans une variable nommée, par exemple ici, r. La commande *summary(r)* fournit directement le tableau des coefficients de régression ainsi que leur signification.



Dans l'exemple représenté ci-dessus, le coefficient b vaut 0,2137, la statistique F pour 1 et 19 degrés de liberté vaut 0,1648 et la p-value (probabilité sous l'hypothèse nulle d'observer une telle valeur ou une valeur plus élevée de F) vaut 0,6893.

Notre seuil de signification arbitraire étant 0,05, la probabilité observée est largement plus grande et conduit donc à accepter l'hypothèse nulle. Ma régression n'est donc pas significative…

NB: - summary() est une fonction générique qui peut être appliquée sur une multitude de variables ou résultats d'autres fonctions. Les informations de sortie dépendront de la variable sur laquelle on applique summary()…

- pour obtenir uniquement les coefficients de la régression, on peut utiliser l'instruction r$coeff, qui fournit un vecteur avec les deux coefficients b_0 et b.

Le tableau d'analyse de la variance sera obtenu par l'instruction *anova(r)*. Le résultat affiché est un peu moins complet que *summary(r)* mais la valeur de F et la p-value sont naturellement identiques.

Une fois l'analyse de régression effectuée, il est assez simple d'obtenir les valeurs attendues (ou prédites) et de tracer la droite de régression. Notre objet s'appelant *r*, il suffit de taper *r$fitted.values* pour obtenir les valeurs, qui peuvent ensuite être enregistrées dans un vecteur (par exemple, *P*). Ainsi la droite de régression peut être tracée par l'instruction : *lines(X,P,col= « red »)*.

Autre solution : utiliser la fonction *predict(),* qui fournit les valeurs prédites. Par défaut, elle donnera les valeurs prédites pour nos X : *predict(r,newdata=list("X"=X))* ou plus simplement *predict(r)*. On peut directement utiliser ses valeurs dans la fonction graphique secondaire *lines(…)* évoquée ci-dessus.

NB: une fonction graphique secondaire dans R est une fonction qu'on utilise après l'emploi préalable d'une fonction graphique principale (par exemple plot()). Une telle fonction permet, par exemple, de tracer la droite de régression en superposition sur le graphique principal (en utilisant les mêmes échelles…).

Exercice 2

Le coefficient de régression b est un estimateur ponctuel de β (le vrai coefficient de régression de la population que l'on ne connaît naturellement pas). Comme pour d'autres estimateurs, un intervalle de confiance de l'estimateur peut être établi. Nous aurons donc, par exemple, 95% de chances de trouver le vrai paramètre de la population dans cet intervalle. Les limites de cet intervalle correspondent donc à deux valeurs de b entre lesquelles β a 95% de chance de tomber. Ainsi, à partir de ces deux coefficients extrêmes, deux nouvelles droites de régression "limites" peuvent être tracées en remplaçant b dans l'équation de la régression par b_i (limite inférieure de b) et b_s (limite supérieure de b). Et donc, b limites = b +/- t (0,05; ddl erreur)*S_b. La valeur de t théorique s'obtient soit via la table de t, soit via Excel ou R… (avec quelles fonctions ?) S_b est la racine carrée de la variance associée au coefficient de régression, calculée plus haut lors du test de t.

Insérez les deux droites limites sur le graphique de départ. Les trois droites (la droite de régression et les 2 droites « limites ») se coupent en un point unique, de coordonnées (X_m; Y_m).

Exercice 3

Corrélation : calculez le coefficient de corrélation de vos données (coefficient de corrélation de Pearson). Ce coefficient complète l'analyse de la régression linéaire en donnant une information sur le degré de liaison entre les deux variables, indépendamment du sens de la relation les unissant. Cela est surtout intéressant quand le sens de la relation ne peut pas être clairement défini : le poids dépend-il de la taille ou bien la taille dépend-elle du poids ?

$r = \Sigma xy$ / racine carrée ($\Sigma x^2 * \Sigma y^2$)

Quelle est la plage théorique de variation du coefficient de corrélation ?

Comme r est un estimateur, basé sur votre échantillon, de la « vraie » corrélation ρ existant entre X et Y, il peut être utile de tester l'hypothèse nulle $H_0 : \rho = 0$, pour savoir si la corrélation obtenue r s'écarte de manière significative de cette corrélation nulle, qui indiquerait qu'il n'y a aucun lien (linéaire) entre X et Y. Dans R, vous pouvez directement obtenir le coefficient de corrélation de vos données et en avoir l'analyse statistique grâce à la fonction *cor.test(Y,X)*.

Exercice facultatif pour le brainstorming

Dans R, à l'aide d'une boucle, générez 1000 valeurs échantillonnées dans la distribution de F avec 1 et 18 degrés de liberté. Triez ensuite les valeurs et extrayez la 950^e, qui correspondra à la limite supérieure du seuil 5% unilatéral.

Comme les autres distributions, la distribution de F peut être déclinée dans R selon 4 méthodes (rf, df, pf, qf). En utilisant l'une d'entre elles, écrivez l'instruction qui permettra de trouver la valeur seuil théorique de la distribution de F pour 1 et 18 degrés de liberté au seuil 5% unilatéral droit.

Comparez-la à la valeur empirique trouvée précédemment.