Modèle Linéaire

Régression linéaire à deux variables explicatives

Thierry Dhorne

Institut Universitaire de Technologie de Vannes
Université de Bretagne Sud

Année Universitaire 2014-2015

Plan du cours

1. Régression simple
2. Régression à deux variables explicatives
3. Régression multiple

Modèle Linéaire - Régression

Régression linéaire à deux variables explicatives

Thierry Dhorne

Institut Universitaire de Technologie de Vannes
Université de Bretagne Sud

Année Universitaire 2014-2015

Pourquoi (enseigner) la régression linéaire à deux variables explicatives ?

- nous avons vu l'intérêt et l'efficacité de la régression linéaire simple pour certains exemples particuliers (les moules)
- on peut avoir l'idée et l'envie de généraliser la prédiction d'une variable d'intérêt au cas où on dispose de plusieurs prédictrices

★ l'intuition n'est pas si facile que dans le cas d'une seule prédictrice

- la régression linéaire à deux variables explicatives est rarement présentée dans un cours

→ en général, on passe directement de la régression simple à la régression à plusieurs (p) variables explicatives

★ il nous paraît important de la présenter car bien que les calculs soient encore faisables, on se persuade de leur complexité relative

→ elle aide donc à la compréhension de la régression à plusieurs variables explicatives

Un exemple fictif
Pour comprendre la difficulté

- l'aptitude d'un individu au saut en hauteur dépend :
  - de son poids (plus l'on est léger plus on saute haut ?)
  - de sa taille (plus on a de grandes jambes plus on saute haut ?)

→ à la fois du poids et de la taille

★ d'où l'intérêt de prendre en compte ces deux variables dans un même modèle

La difficulté du problème

- l'influence du poids et de la taille s'effectue en sens contraire

★ mais pourtant le poids et la taille sont liés positivement (en moyenne, plus on est grand plus on pèse lourd ?)

→ l'analyse de l'effet conjoint des deux variables poids et taille sur l'aptitude au saut doit donc donner des informations différentes de celles des analyses séparées.

★ il est quasiment impossible d'intuiter la résultante de tous ces phénomènes

→ c'est l'objectif principal de la régression linéaire multivariable que de comprendre cette résultante

- on commence par deux variables car cela est suffisant pour essayer de comprendre (et réussir ?)

Contexte

- certaines notions présentées dans le cadre de l'étude de la régression à une variable explicative se généralisent sans difficultés au cas de plusieurs variables explicatives
- c'est le cas en particulier de l'analyse de la variance et des tests

→ mais des éléments nouveaux liés au caractère multivariable du problème apparaissent

- c'est le cas des conséquences du lien entre les variables explicatives

→ il faut en particulier aborder le problème de l'influence partielle ou conditionnelle d'une variable explicative par rapport à une autre

Un exemple réel

- l'exemple étudié est un exemple opérationnel au quotidien

➙ il s'agit de l'évaluation de la qualité des carcasses de porc dans les abattoirs

- ce travail est réalisé chaque jour dans tous les abattoirs européens

➙ en effet les éleveurs sont rétribués en fonction de la qualité de la carcasse qui doit être évaluée objectivement

- la qualité d'une carcasse est appréciée entre autres choses à travers sa teneur en viande (consommable) par opposition au gras (peu diététique)

Classement de carcasses

Une application pratique de la régression

- pour atteindre cet objectif une expérimentation très lourde doit être mise en place

➞ en France, c'est maintenant France-Agrimer qui a la responsabilité de cette évaluation

➞ et qui charge l'Institut Français du Porc de sa mise en œuvre

FranceAgriMer
ETABLISSEMENT NATIONAL DES PRODUITS DE L'AGRICULTURE ET DE LA MER

ifip institut du porc

Classification de carcasses

La variable d'intérêt

- une carcasse de porc est constituée grosso modo
  - de viande
  - de gras
  - d'éléments de structure (os, tendons, ...)
- la variabilité des éléments de structure est assez faible
- en revanche la teneur en viande dépend beaucoup de la race (type génétique)
- la teneur en gras de l'alimentation, de l'âge et du sexe (type sexuel)

→ ce qui intéresse le consommateur et donc le transformateur c'est la viande

★ d'où la variable stratégique : la teneur en viande

- la variable d'intérêt est très difficile à mesurer
  - on doit disséquer la carcasse
  - ou utiliser des méthodes physiques complexes : RMN, Rayons X
- en abattoir, il faut trouver des variables plus accessibles

→ l'idée consiste à prédire la teneur en viande à l'aide de variables faciles à mesurer en abattoir

Classification de carcasses

Le matériel

- les prédictrices les plus simples peuvent être des épaisseurs de gras et de muscles mesurées à différents endroits de la carcasse
- les mesures sont réalisées à l'aide de différents appareils

→ en Bretagne, l'appareil utilisé, appelé CGM (capteur gras-muscle) a été conçu par la société SYDEL basée à Lorient

le cgm : capteur gras-muscle

Classification de carcasses

Les prédictrices

- au quotidien dans les abattoirs de l'Ouest, les trois prédictrices utilisées sont
  - une épaisseur de gras dorsal
  - une épaisseur de muscle dorsal
  - une épaisseur de gras lombaire
- la mise en œuvre quotidienne de cette méthode relève de l'organisme inter-professionnel Uniporc-Ouest qui classe plus de 15 millions de carcasses par an

la mesure des prédictrices

Classification de carcasses
Les données

- les données utilisées dans la suite forment un sous-ensemble des données obtenues lors de l'expérimentation de 1996
- il s'agit du sous-ensemble des carcasses femelles de type piétrain
- il y a quatre variables
  - le taux de muscle `tmus` en %
  - l'épaisseur de gras dorsal `grd` en mm
  - l'épaisseur de muscle dorsal `mud` en mm
  - l'épaisseur de gras lombaire `grl` en mm

	tmus	grd	mud	grl
1	59.43	18.24	56.98	19.97
2	61.11	13.87	50.91	15.90
3	56.68	16.91	54.14	19.40
4	62.62	13.68	51.86	14.81
5	63.22	16.53	58.69	17.70
6	59.57	17.10	52.43	22.02
7	59.99	18.43	56.04	19.54

	tmus	grd	mud	grl
8	58.93	16.53	49.58	16.25
9	63.45	12.92	54.90	15.82
10	62.98	18.62	60.21	21.13
11	58.80	19.75	57.74	22.33
12	58.09	17.67	54.52	19.26
13	65.12	12.92	65.15	17.62
14	58.98	17.48	56.98	20.70

Classification de carcasses

Les données - suite

	tmus	grd	mud	grl		tmus	grd	mud	grl
15	64.38	14.06	56.61	16.26	46	58.95	14.63	60.59	17.87
16	62.13	15.58	55.28	17.58	47	57.10	15.96	47.30	20.34
17	56.32	20.32	54.14	21.86	48	55.84	14.44	54.71	15.02
18	57.52	17.48	56.04	19.39	49	61.30	13.11	57.36	17.53
19	60.22	17.10	61.35	19.79	50	62.56	15.58	57.36	16.82
20	57.98	17.48	53.38	18.59	51	62.75	14.06	55.66	17.04
21	61.89	18.43	61.35	19.61	52	59.09	18.43	62.11	19.67
22	63.86	13.49	66.10	14.67	53	56.76	17.10	55.66	21.04
23	58.98	12.73	49.01	15.72	54	58.12	16.53	55.28	22.40
24	59.58	14.25	53.00	18.90	55	56.97	18.05	57.74	22.35
25	60.33	19.37	59.07	20.04	56	59.15	19.37	53.57	20.11
26	55.93	18.43	54.33	21.02	57	63.53	12.92	66.48	14.12
27	59.29	16.72	52.05	19.61	58	60.00	14.06	61.73	18.25
28	59.09	15.20	43.88	16.74	59	61.22	15.20	62.30	15.50
29	56.89	18.99	60.21	20.30	60	65.07	14.82	62.11	18.45
30	62.10	12.35	55.09	14.77	61	55.63	15.77	51.29	19.40
31	62.11	13.11	53.38	16.60	62	58.82	17.48	53.95	20.68
32	56.47	19.18	51.10	17.01	63	61.92	13.49	49.39	16.26
33	66.94	10.64	68.57	12.26	64	64.67	10.26	58.88	13.02
34	58.51	16.53	52.43	15.34	65	60.98	14.44	55.66	16.06
35	61.94	15.77	56.61	16.97	66	59.01	19.56	65.34	22.65
36	64.16	11.78	63.44	14.79	67	57.73	18.05	57.74	17.58
37	58.68	17.29	49.01	16.34	68	58.83	17.10	55.85	16.25
38	61.45	17.86	51.10	21.51	69	60.78	12.16	51.48	17.04
39	60.59	18.81	58.69	21.54	70	60.74	19.18	56.42	20.76
40	61.30	15.20	60.78	13.09	71	58.01	13.11	51.10	16.20
41	54.83	19.37	49.20	22.07	72	56.46	17.29	56.98	24.30
42	60.87	15.77	54.90	18.34	73	65.26	9.88	56.04	12.12
43	57.30	19.37	56.42	22.77	74	63.32	15.01	57.93	17.14
44	60.62	20.32	58.69	24.70	75	65.32	11.40	57.36	13.15
45	59.28	15.77	56.04	21.28					

Étude univariée

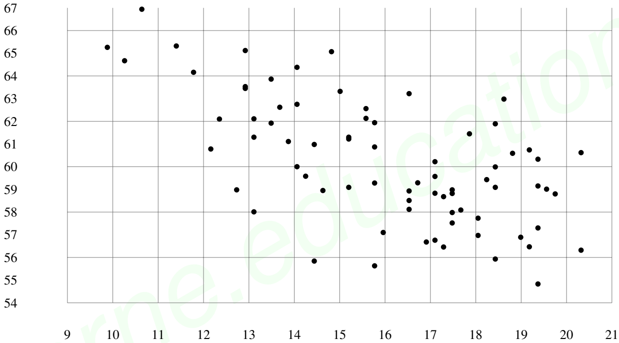
- on peut évidemment étudier le lien entre la variable d'intérêt : le taux de muscle
- et chacune des variables prédictrices
- indépendamment l'une des autres

→ on pourrait donc faire plusieurs régressions simples

- commençons par les analyses graphiques

Graphe croisé

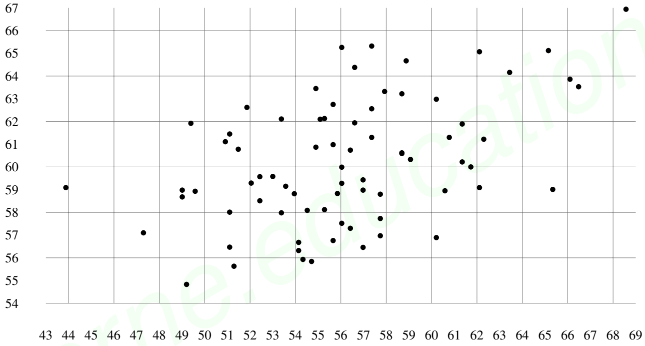
Taux de muscle en fonction d'épaisseur de gras dorsal



→ tendance linéaire décroissante assez lâche avec variance résiduelle constante ⇒ régression simple adaptée

Graphe croisé

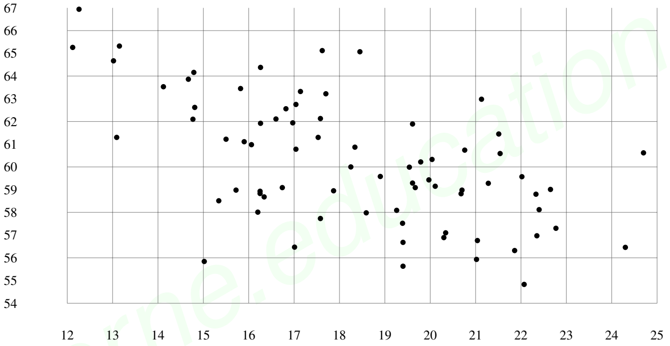
Taux de muscle en fonction d'épaisseur de muscle dorsal



➡ tendance linéaire croissante assez lâche avec variance résiduelle constante ⇒ régression simple adaptée

Graphe croisé

Taux de muscle en fonction d'épaisseur de gras lombaire



→ tendance linéaire décroissante assez lâche avec variance résiduelle constante ⇒ régression simple adaptée

Un modèle pour chaque variable explicative

- nous venons de voir que pour chacune des variables explicatives un modèle de régression linéaire simple serait adapté

→ pour le gras dorsal (x^1)

$$Y = \beta_0^1 + \beta_1^1 x^1 + E^1$$

→ pour le muscle dorsal (x^2)

$$Y = \beta_0^2 + \beta_1^2 x^2 + E^2$$

→ pour le gras lombaire (x^3)

$$Y = \beta_0^3 + \beta_1^3 x^1 + E^3$$

★ remarques

- l'indice haut représente la variable
- il n'y a pas d'indice r de répétition (pour simplifier)

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.92713    1.55441  45.630  < 2e-16 ***
grd         -0.67049    0.09609  -6.978 1.15e-09 ***

Residual standard error : 2.138 on 73 degrees of freedom
Multiple R-Squared : 0.4001,     Adjusted R-squared : 0.3919
F-statistic : 48.69 on 1 and 73 DF,  p-value : 1.146e-09
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.98395    3.40422  13.214  < 2e-16 ***
mud          0.27097    0.06034   4.491 2.60e-05 ***

Residual standard error : 2.443 on 73 degrees of freedom
Multiple R-Squared : 0.2165,     Adjusted R-squared : 0.2057
F-statistic : 20.17 on 1 and 73 DF,  p-value : 2.603e-05
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.1093     1.6730  41.905  < 2e-16 ***
grl          -0.5411     0.0904  -5.985 7.42e-08 ***

Residual standard error : 2.261 on 73 degrees of freedom
Multiple R-Squared : 0.3292,     Adjusted R-squared :  0.32
F-statistic : 35.82 on 1 and 73 DF,  p-value : 7.419e-08
```

Conclusions et Suggestion

- chacune des variables prise indépendamment est une bonne prédictrice pour une régression linéaire simple
- mais on a envie de « combiner » l'information apportée par chacune d'entre elle

➞ en un seul modèle

★ comment combiner ?

$$Y = \beta_0^1 + \beta_1^1 x^1 + E^1$$

et

$$Y = \beta_0^2 + \beta_1^2 x^2 + E^2$$

➞ en accordant dans le même modèle un coefficient pour chaque variable

Modèle

- le modèle de régression linéaire simple peut être généralisé au cas de deux variables explicatives

Modèle de régression linéaire à deux variables explicatives

$$Y_r = \beta_0 + \beta_1 x_r^1 + \beta_2 x_r^2 + E_r, \ r = 1, \dots, n$$

- Y est la variable aléatoire à expliquer
- x^1, x^2 sont les deux variables explicatives
- β_0, β_1, β_2 sont les trois paramètres de l'espérance
- les E_r représentent les écarts au modèle et sont considérées comme des variables aléatoires
- n est la taille de l'échantillon ou nombre d'observations

Postulats

- le modèle qui vient d'être présenté

$$Y_r = \beta_0 + \beta_1 x_r^1 + \beta_2 x_r^2 + E_r,\ r = 1, \ldots, n$$

s'accompagne des mêmes postulats que ceux du modèle de régression linéaire simple

- les E_r sont des variables aléatoires telles que
  - $E(E_r) = 0 \Rightarrow E(Y_r) = \beta_0 + \beta_1 x_r^1 + \beta_2 x_r^2$
  - $V(E_r) = V(Y_r) = \sigma_E^2$
  - E_r et $E_{r' \neq r}$ sont indépendants

Estimation des paramètres
Méthode

- l'estimation des paramètres est obtenue, comme pour la régression simple, par la méthode des moindres carrés
- ou, ce qui est équivalent, du maximum de vraisemblance gaussien

➞ la quantité à minimiser est donc

$$\sum_{r=1}^{n} (Y_r - \beta_0 - \beta_1 x_r^1 - \beta_2 x_r^2)^2$$

➞ et les solutions sont

$$(B_0, B_1, B_2) = \arg \min_{(\beta_0, \beta_1, \beta_2)} \sum_{r=1}^{n} (Y_r - \beta_0 - \beta_1 x_r^1 - \beta_2 x_r^2)^2$$

Estimation des paramètres

Dérivation

→ en dérivant par rapport à chacun des paramètres, on obtient le système linéaire suivant

$$
\begin{aligned}
\sum_{r=1}^{n}(Y_r - B_0 - B_1 x_r^1 - B_2 x_r^2) &= 0,\\
\sum_{r=1}^{n} x_r^1 (Y_r - B_0 - B_1 x_r^1 - B_2 x_r^2) &= 0,\\
\sum_{r=1}^{n} x_r^2 (Y_r - B0 - B_1 x_r^1 - B_2 x_r^2) &= 0.
\end{aligned}
$$

Estimation des paramètres

Analogie avec la régression simple

- en divisant la première équation par n, on retrouve un résultat analogue à celui de la régression linéaire simple

$$\bar{Y} = B_0 + B_1 \overline{x^1} + B_2 \overline{x^2},$$

➜ ce qui montre que le plan de régression des moindres carrés passe par le point moyen

- on peut réécrire les deux dernières équations en éliminant l'inconnue B_0 :

$$\begin{aligned}
\sum_{r=1}^{n}(x_r^1 - \bar{x}^1)[(Y_r - \bar{Y}) - B_1(x_r^1 - \bar{x}^1) - B_2(x_r^2 - \bar{x}^2)] &= 0, \\
\sum_{r=1}^{n}(x_r^2 - \bar{x}^2)[(Y_r - \bar{Y}) - B_1(x_r^1 - \bar{x}^1) - B_2(x_r^2 - \bar{x}^2)] &= 0.
\end{aligned}$$

- ceci conduit finalement au système de deux équations à deux inconnues

$$\begin{array}{ccccccc} B_1 & SCE(x^1) & + & B_2 & SPE(x^1,x^2) & = & SPE(x^1,Y), \\ B_1 & SPE(x^1,x^2) & + & B_2 & SCE(x^2) & = & SPE(x^2,Y), \end{array}$$

où

- $SCE(x^1) = \sum_{=1}^{n}(x_r^1 - \overline{x^1})^2$
- $SCE(x^2) = \sum_{=1}^{n}(x_r^2 - \overline{x^2})^2$
- $SPE(x^1,x^2) = \sum_{=1}^{n}(x_r^1 - \overline{x^1})(x_r^2 - \overline{x^2})$
- $SPE(x^1,Y) = \sum_{=1}^{n}(x_r^1 - \overline{x^1})(Y_r - \overline{Y})$
- $SPE(x^2,Y) = \sum_{=1}^{n}(x_r^2 - \overline{x^2})(Y_r - \overline{Y})$

Estimation des paramètres

Écriture simplifiée

- on peut donner une forme plus synthétique au système précédent :

$$\begin{array}{cccccccc} B_1 & S_1^2 & + & B_2 & S_{12} & = & S_{1Y}, \\ B_1 & S_{12} & + & B_2 & S_2^2 & = & S_{2Y}, \end{array}$$

où

- $S_1^2 = \sum_{=1}^{n}(x_r^1 - \bar{x}^1)^2$
- $S_2^2 = \sum_{=1}^{n}(x_r^2 - \bar{x}^2)^2$
- $S_{12} = \sum_{=1}^{n}(x_r^1 - \bar{x}^1)(x_r^2 - \bar{x}^2)$
- $S_{1Y} = \sum_{=1}^{n}(x_r^1 - \bar{x}^1)(Y_r - \bar{Y})$
- $S_{2Y} = \sum_{=1}^{n}(x_r^2 - \bar{x}^2)(Y_r - \bar{Y})$

- la résolution du système précédent conduit aux solutions suivantes

$$\begin{aligned} B_1 &= \frac{S_2^2 S_{1Y} - S_{12} S_{2Y}}{S_1^2 S_2^2 - (S_{12})^2}, \\ B_2 &= \frac{S_1^2 S_{2Y} - S_{12} S_{1Y}}{S_1^2 S_2^2 - (S_{12})^2}. \end{aligned}$$

qui donnent donc la forme analytique des estimateurs des moindres carrés

- en remplaçant dans l'équation au point moyen, on obtient

$$B_0 = \overline{Y} - B_1 \overline{x^1} - B_2 \overline{x^2}$$

Estimateur de la variance résiduelle

Naturellement, l'estimateur non biaisé de σ^2 construit sur l'estimateur du maximum de vraisemblance est donné par

$$S^2 = \frac{1}{n-3}\sum_{i=1}^{n}(Y_r - B_0 - B_1 x_r^1 - B_2 x_r^2)^2$$

Comparaison avec la régression simple

Cas de prédictrices non liées

- l'intérêt de la régression multiple consiste à améliorer la qualité de l'ajustement en introduisant plusieurs variables

→ en quoi l'introduction d'une nouvelle variable modifie la régression simple initiale ?

★ si la quantité S_{12} est nulle, les estimateurs deviennent :

$$
\begin{aligned}
B_1 &= \frac{S_{1Y}}{S_{11}}, \\
B_2 &= \frac{S_{2Y}}{S_{22}}.
\end{aligned}
$$

c'est-à-dire sont identiques à ceux de la régression simple

→ il y a donc «additivité» dans ce cas

Influence du lien entre les variables

- S_{12} est d'autant plus faible que les deux variables varient indépendamment

➡ les coefficients de la régression à deux variables seront donc d'autant plus proches de ceux des deux régressions simples que la liaison entre les deux variables explicatives sera faible

➡ il y a quasi-additivité

★ lorsque ce n'est pas le cas, les explicatives «empiètent» en quelque sorte l'une sur l'autre

- Décomposition de $B_{1/2}$

$$
\begin{aligned}
B_{1/2} &= \frac{S_{22}S_{1Y} - S_{12}S_{2Y}}{S_{11}S_{22} - (S_{12})^2} \\
&= \frac{S_{22}S_{1Y} - \frac{(S_{12})^2}{S_{11}}S_{1Y} + \frac{(S_{12})^2}{S_{11}}S_{1Y} - S_{12}S_{2Y}}{S_{11}S_{22} - (S_{12})^2} \\
&= \frac{(S_{22} - \frac{(S_{12})^2}{S_{11}})S_{1Y}}{S_{11}S_{22} - (S_{12})^2} + \frac{S_{12}}{S_{11}(S_{22} - \frac{(S_{12})^2}{S_{11}})}(\frac{S_{12}}{S_{11}}S_{1Y} - S_{2Y}) \\
&= \frac{S_{1Y}}{S_{11}} + \frac{S_{12}}{S_{11}} \frac{\frac{S_{12}}{S_{11}}S_{1Y} - S_{2Y}}{S_{22} - \frac{(S_{12})^2}{S_{11}}}
\end{aligned}
$$

Etude analytique - II

- le terme $\frac{S_{1Y}}{S_{11}}$ correspond au coefficient de régression de Y sur x^1
- le terme $\frac{S_{12}}{S_{11}} S_{1Y} - S_{2Y}$ peut être vu comme la somme des produits des écarts

$$\begin{aligned} SPE\left(Y - \frac{S_{12}}{S_{11}} x^1, x^2 - \frac{S_{12}}{S_{11}} x^1\right) &= S_{2Y} - \frac{S_{12}}{S_{11}} S_{12} - \frac{S_{12}}{S_{11}} S_{1Y} + \frac{S_{12}^2}{S_{11}} \\ &= S_{2Y} - \frac{S_{12}}{S_{11}} S_{1Y} \end{aligned}$$

- de même le terme $S_{22} - \frac{(S_{12})^2}{S_{11}}$ peut être vu comme la somme des carrés des écarts

$$\begin{aligned} SCE\left(x^2 - \frac{S_{12}}{S_{11}} x^1\right) &= S_{22} + \frac{(S_{12})^2}{(S_{11})^2} S_{11} - 2 \frac{S_{12}}{S_{11}} S_{12} \\ &= S_{22} - \frac{(S_{12})^2}{S_{11}} \end{aligned}$$

Etude analytique - III

- on obtient donc l'égalité :

$$\text{coef}(Y, x^1/x^2) = \text{coef}(Y, x^1) - \text{coef}(x^2, x^1)\text{coef}(Y - x^1, x^2 - x^1).$$

- on a évidemment la forme symétrique pour x^2 :

$$\text{coef}(Y, x^2/x^1) = \text{coef}(Y, x^2) - \text{coef}(x^1, x^2)\text{coef}(Y - x^2, x^1 - x^2).$$

→ la régression à deux variables explicatives peut donc être vue comme

- une régression à une variable explicative
- corrigée d'une autre régression du résidu de la variable à prédire sur le résidu de la variable prédictrice
- pondérée par la régression de la première prédictrice sur la seconde

Variables explicatives non covariantes

- on vérifie ici que si les deux variables explicatives ne sont pas liées linéairement, c'est-à-dire ne s'expliquent pas l'une l'autre alors $\text{coef}(x^2, x^1) = \text{coef}(x^1, x^2) = 0$ et donc :

$$\begin{aligned} \text{coef}(Y, x^1/x^2) &= \text{coef}(Y, x^1), \\ \text{coef}(Y, x^2/x^1) &= \text{coef}(Y, x^2). \end{aligned}$$

On peut calculer l'espérance de l'estimateur B_1 :

$$
\begin{aligned}
E(B_1) &= \frac{1}{S_{11}S_{22}-(S_{12})^2}[S_{22}E(S_{1Y}) - S_{12}E(S_{2Y})], \\
&= \frac{1}{S_{11}S_{22}-(S_{12})^2}\{S_{22}\sum_{r=1}^{n}(x^1-\overline{x^1})[\beta_1(x_r^1-\overline{x^1})+\beta_2(x_r^2-\overline{x^2})] \\
&\quad \ldots - S_{12}\sum_{r=1}^{n}(x^2-\overline{x^2})[\beta_1(x_r^1-\overline{x^1})+\beta_2(x_r^2-\overline{x^2})]\}, \\
&= \frac{1}{S_{11}S_{22}-(S_{12})^2}[S_{22}(\beta_1 S_{11}+\beta_2 S_{12})] - [S_{12}(\beta_1 S_{12}+\beta_2 S_{22})], \\
&= \frac{1}{S_{11}S_{22}-(S_{12})^2}\beta_1[S_{11}S_{22}-(S_{12})^2], \\
&= \beta_1.
\end{aligned}
$$

Propriétés des estimateurs
Espérance II

- De manière symétrique, on démontre que B_2 est non biaisé :

$$E(B_2) = \beta_2.$$

- La relation :

$$B_0 = \bar{Y} - B_1\overline{x^1} - B_2\overline{x^2}.$$

donne en espérance :

$$\begin{aligned} E(B_0) &= E(\bar{Y}) - E(B_1)\overline{x^1} - E(B_2)\overline{x^2}, \\ &= \beta_0 + \beta_1\overline{x^1} + \beta_2\overline{x^2} - \beta_1\overline{x^1} - \beta_2\overline{x^2}, \\ &= \beta_0. \end{aligned}$$

Propriétés des estimateurs
Variance I

- Calcul de la variance de B_1, de celle de B_2 ainsi que de leur covariance

$$
\begin{aligned}
V(B_1) &= \frac{1}{(S_{11}S_{22}-(S_{12})^2)^2}\left[(S_{22})^2V(S_{1Y})+(S_{12})^2V(S_{2Y})\right.\\
&\qquad \left. \ldots - 2S_{22}S_{12}C(S_{1Y},S_{2Y})\right]\\
V(B_2) &= \frac{1}{(S_{11}S_{22}-(S_{12})^2)^2}\left[(S_{11})^2V(S_{2Y})+(S_{12})^2V(S_{1Y})\right.\\
&\qquad \left. \ldots - 2S_{11}S_{12}C(S_{1Y},S_{2Y})\right]\\
C(B_1,B_2) &= \frac{1}{(S_{11}S_{22}-(S_{12})^2)^2}\left[(S_{11}S_{22}+(S_{12})^2)C(S_{1Y},S_{2Y})\right.\\
&\qquad \left. \ldots - S_{22}S_{12}V(S_{1Y}) - S_{11}S_{12}V(S_{2Y})\right]
\end{aligned}
$$

→ Il faut donc calculer les expressions :

- $V(S_{1Y})$, $V(S_{2Y})$ et $C(S_{1Y},S_{2Y})$

$$
\begin{aligned}
V(S_{1Y}) &= \sum_{r=1}^{n} (x_r^1 - \bar{x^1})^2 V(Y_r), \\
&= S_{11}\sigma^2, \\
V(S_{2Y}) &= \sum_{r=1}^{n} (x_r^2 - \bar{x^2})^2 V(Y_r), \\
&= S_{22}\sigma^2. \\
C(S_{1Y}, S_{2Y}) &= \sum_{r=1}^{n} (x_r^1 - \bar{x^1})(x_r^2 - \bar{x^2}) V(Y_r), \\
&= S_{12}\sigma^2.
\end{aligned}
$$

D'où

$$\begin{aligned}
V(B_1) &= \frac{\sigma^2}{(S_{11}S_{22}-(S_{12})^2)^2}\left[(S_{22})^2 S_{11} + (S_{12})^2 V(S_{22}) - 2S_{11}S_{12}^2)\right] \\
&= \frac{S_{22}(S_{22}S_{11}-S_{12}^2)}{(S_{11}S_{22}-(S_{12})^2)^2}\sigma^2 \\
&= \frac{S_{22}}{S_{11}S_{22}-S_{12}^2}\sigma^2 \\
V(B_2) &= \frac{S_{11}}{S_{11}S_{22}-S_{12}^2}\sigma^2
\end{aligned}$$

Propriétés des estimateurs
Variance IV

$$
\begin{aligned}
C(B_1, B_2) &= \frac{\sigma^2}{(S_{11}S_{22} - (S_{12})^2)^2}[(S_{11}S_{22} + (S_{12})^2)S_{12} - 2S_{11}S_{22}S_{12}] \\
&= \frac{-S_{12}(S_{22}S_{11} - S_{12}^2)}{(S_{11}S_{22} - (S_{12})^2)^2}\sigma^2 \\
&= -\frac{S_{12}}{S_{11}S_{22} - S_{12}^2}\sigma^2
\end{aligned}
$$

Résumé théorique

Forme des estimateurs

$$B_1 = \frac{S_2^2 S_{1Y} - S_{12} S_{2,Y}}{S_1^2 S_2^2 - (S_{12})^2} \qquad B_2 = \frac{S_1^2 S_{2Y} - S_{12} S_{1Y}}{S_1^2 S_2^2 - (S_{12})^2}$$

$$B_0 = \bar{Y} - B_1 \overline{x^1} - B_2$$

Variances et covariance des estimateurs

$$V(B_1) = \frac{S_{22}}{S_{11} S_{22} - S_{12}^2} \sigma^2 \qquad V(B_2) = \frac{S_{11}}{S_{11} S_{22} - S_{12}^2} \sigma^2$$

$$C(B_1, B_2) = -\frac{S_{12}}{S_{11} S_{22} - S_{12}^2} \sigma^2$$

- dans le cas de la régression simple, on a construit le test de l'effet de la variable explicative.

→ ceci a été fait à partir de l'équation d'analyse de la variance

$$
\begin{array}{ccccc}
\sum_{r=1}^{n}(Y_r - \bar{Y})^2 & = & \sum_{r=1}^{n}(\hat{Y}_r - \bar{Y})^2 & + & \sum_{r=1}^{n}(Y_r - \hat{Y}_r)^2, \\
SCET & = & SCEM & + & SCER.
\end{array}
$$

- cette écriture se généralise au cas de deux variables explicatives

→ on peut, dans un premier temps, écrire une équation d'analyse de la variance globale pour les deux variables

$$\begin{array}{ccccc} \sum_{r=1}^{n}(Y_r - \bar{Y})^2 & = & \sum_{r=1}^{n}(\hat{Y}_{r(1,2)} - \bar{Y})^2 & + & \sum_{r=1}^{n}(Y_r - \hat{Y}_{r(1,2)})^2 \\ SCET & = & SCEM_{(1,2)} & + & SCER_{(1,2)} \end{array}$$

où l'indice $_{(1,2)}$ indique que les estimateurs ou les sommes de carrés sont obtenus dans le modèle à 2 régresseurs.

Validation du modèle global

- hypothèses

$$H_0 : \beta_1 = \beta_2 = 0 \qquad H1 : \beta_1 \neq 0 \text{ ou } \beta_2 \neq 0$$

→ le test de validation de Fisher est basé sur la statistique

$$F_{(1,2)} = \frac{SCEM_{(1,2)}/2}{SCER_{1,2)}/(n-3)}$$

→ il compare 2 sources de variabilité : celle du modèle et celle des erreurs

- la statistique $F_{(1,2)}$ suit, sous H_0, une loi de Fisher $\mathcal{F}(2, n-3)$

→ la région de rejet est au seuil α

$$F_{(1,2)} > \mathcal{F}(2, n-3, 1-\alpha)$$

Test de l'effet d'une variable additionnelle - I

- On peut aussi écrire les équations d'analyse de la variance pour les deux modèles de régression simples envisageables :

$$\begin{array}{cccccc} \sum_{r=1}^{n}(Y_r - \bar{Y})^2 & = & \sum_{r=1}^{n}(Y_r - \hat{Y}_{r(1)})^2 & + & \sum_{r=1}^{n}(\hat{Y}_{r(1)} - \bar{Y})^2 \\ SCET & = & SCER_{(1)} & + & SCEM_{(1)} \\ & = & \sum_{r=1}^{n}(Y_r - \hat{Y}_{r(2)})^2 & + & \sum_{r=1}^{n}(\hat{Y}_{r(2)} - \bar{Y})^2 \\ SCET & = & SCER_{(2)} & + & SCEM_{(2)} \end{array}$$

Test de l'effet d'une variable additionnelle - II

→ Ces deux décompositions relatives à des sous-modèles se combinent à la décomposition globale pour donner :

$$\begin{aligned} SCET &= SCEM_{(1)} + SCER_{(1)} - SCER_{(1,2)} + SCER_{(1,2)}, \\ &= SCEM_{(2)} + SCER_{(2)} - SCER_{(1,2)} + SCER_{(1,2)}. \end{aligned}$$

- $SCER_{(1,2)}$ représente la somme de carré résiduelle du modèle complet (à deux variables),
- $SCEM_{(1)}$ ou $SCEM_{(2)}$ représente la sommes de carrés associée à un sous-modèle (à une variable explicative),
- $SCER_{(1)} - SCER_{(1,2)}$ ou $SCER_{(2)} - SCER_{(1,2)}$ correspond à la sommes de carrés de l'effet additionnel du modèle complet par rapport au sous-modèle.

Test de l'effet d'une variable additionnelle - III

- on prendra donc l'habitude d'écrire les trois formes de l'équation de l'analyse de la variance d'un modèle de régression à deux variables explicatives :

$$
\begin{array}{rcccccc}
SCET & = & & & SCEM_{(1,2)} & + & SCER_{(1,2)} \\
\\
 & = & SCEM_{(1)} & + & SCEM_{(1,2)} - SCEM_{(1)} & + & SCER_{(1,2)} \\
 & = & SCEM_{(1)} & + & SCER_{(1)} - SCER_{(1,2)} & + & SCER_{(1,2)} \\
 & = & SCEM_{(1)} & + & SCEM_{(2/1)} & + & SCER_{(1,2)} \\
\\
 & = & SCEM_{(2)} & + & SCEM_{(1,2)} - SCEM_{(2)} & + & SCER_{(1,2)} \\
 & = & SCEM_{(2)} & + & SCER_{(2)} - SCER_{(1,2)} & + & SCER_{(1,2)} \\
 & = & SCEM_{(2)} & + & SCEM_{(1/2)} & + & SCER_{(1,2)}
\end{array}
$$

Interprétation

- Ces trois décompositions correspondent à trois manières différentes de percevoir le même modèle

➙ La première considère l'effet conjoint des deux variables explicatives

➙ les deux autres l'effet brut (ou non ajusté) d'une variable et l'effet additionnel (ou ajusté) de l'autre

★ Il est clair que pour tester l'intérêt d'un modèle à deux variables par rapport à un modèle à une variable, la somme de carrés ajustée va être d'un intérêt fondamental

Degrés de liberté

- Comme dans le cas du modèle à une variable explicative, à ces différentes sommes de carrés sont associés les degrés de liberté correspondant aux variables aléatoires indépendantes qui leur ont donné naissance.

→ On obtient naturellement :

SC	$SCET$	=	$SCEM_{(1,2)}$			+	$SCER$
Ddl	$n-1$		2				$n-3$
SC		=	$SCEM_{(1)}$	+	$SCEM_{(2/1)}$	+	$SCER$
Ddl	$n-1$		1		1		$n-3$
SC		=	$SCEM_{(2)}$	+	$SCEM_{(1/2)}$	+	$SCER$
Ddl	$n-1$		1		1		$n-3$

Test sur l'ajout d'une variable

- on veut tester l'intérêt d'ajouter la variable X^2

$$H_0 : \beta_2 = 0 \, qd \, \beta_1 \neq 0 \quad H_1 : \beta_2 \neq 0 \, qd \, \beta_1 \neq 0$$

→ le test de validation de Fisher est basé sur la statistique

$$F_{2/(12)} = \frac{SCEM_{(1,2)} - SCEM_1/1}{SCER_{1,2)}/(n-3)}$$

→ $F_{2/(12)}$ significativement grand conduit à ajouter la variable X^2

- la statistique $F_{2/(1,2)}$ suit, sous H_0 une loi de Fisher $\mathcal{F}(1, n-3)$

→ a région de rejet est au seuil α

$$F_{2/(12)} > \mathcal{F}(1, n-3, 1-\alpha)$$

★ si le est est significatif, on retient le modèle à 2 variables

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.5908     1.6444  43.536  < 2e-16 ***
gd           -0.5139     0.1614  -3.184  0.00214 **
gl           -0.1731     0.1436  -1.206  0.23181

Residual standard error : 2.131 on 72 degrees of freedom
Multiple R-squared : 0.412, Adjusted R-squared : 0.3956
F-statistic : 25.22 on 2 and 72 DF,  p-value : 4.991e-09
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.81207    2.97522  19.095  < 2e-16 ***
gd          -0.63147    0.08239  -7.665 6.47e-11 ***
md           0.23997    0.04527   5.301 1.21e-06 ***

Residual standard error : 1.825 on 72 degrees of freedom
Multiple R-squared : 0.5685, Adjusted R-squared : 0.5565
F-statistic : 47.43 on 2 and 72 DF,  p-value : 7.227e-14
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.49175    3.00286  19.146  < 2e-16 ***
gd          -0.48405    0.13761  -3.517 0.000764 ***
md           0.23904    0.04503   5.308 1.20e-06 ***
gl          -0.16313    0.12234  -1.333 0.186656

Residual standard error : 1.816 on 71 degrees of freedom
Multiple R-squared : 0.5791, Adjusted R-squared : 0.5613
F-statistic : 32.56 on 3 and 71 DF,  p-value : 2.387e-13
```