

Module 11 - Régression linéaire multiple

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Introduction

Contexte

1. Plusieurs variables explicatives disponibles.
2. A-t-on au moins une variable significative ?
3. Quelle proportion de la variabilité de Y est expliquée par toutes ces variables globalement ?
4. Devrait-on utiliser toutes les variables disponibles pour prédire Y avec plus de précision ?
5. Lesquelles ont un impact significatif sur Y ?

Les données

- n unités expérimentales indépendantes;
- une variable réponse Y ;
- k variables explicatives $X_1, \dots, X_k$.

Les données sont de la forme

$$(Y_1, X_{11}, \dots, X_{1k})$$

$$(Y_2, X_{21}, \dots, X_{2k})$$

...

$$(Y_n, X_{n1}, \dots, X_{nk})$$

Le modèle

Modèle de régression linéaire multiple

Extension du modèle de régression linéaire simple.

Ici, la moyenne de Y dépend de plusieurs variables explicatives.

Modèle de régression linéaire multiple (RLM)

$$Y_i = \beta_0 + \underbrace{\beta_1 X_{i1} + \dots + \beta_k X_{ik}}_{E(Y)=\text{hyperplan}} + E_i,$$

où $E_1, \dots, E_n$ sont indépendantes de loi $N(0, \sigma^2)$.

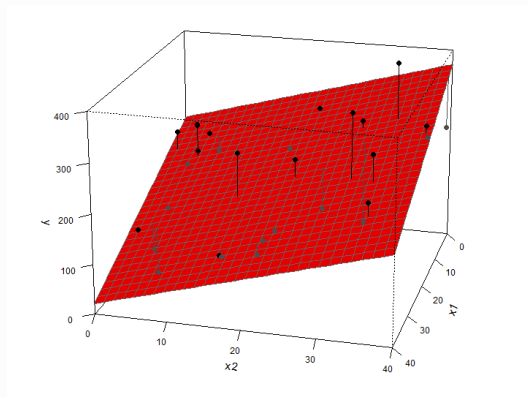
Il est possible d'écrire ce modèle de façon matricielle

$$Y = X\beta + \epsilon,$$

où Y est le vecteur des n réponses, X est la matrice des variables explicatives, β est le vecteur des $k + 1$ paramètres et ϵ est le vecteur des E_i

Interprétation géométrique

Avec $k = 2$ variables explicatives $\Rightarrow$ le nuage des points (x_{i1}, x_{i2}, y_i) sera distribué de façon aléatoire $N(0, \sigma^2)$ autour du plan $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.



Si $k > 2$, le modèle est un "hyperplan" $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$.

Postulats du modèle

Les mêmes 4 postulats qu'avec une variable explicative :

P1 **linéarité** : l'espérance de Y_i est linéaire en $X_{i1}, \dots, X_{ik}$;

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} \Leftrightarrow E(E_i) = 0$$

P2 **homoscédasticité** : la variance de Y_i autour du plan est homogène (elle ne dépend pas des X_{ij}) ;

$$V(Y_i) = \sigma^2 \Leftrightarrow V(E_i) = \sigma^2$$

P3 **indépendance** : les observations sont indépendantes ;

$$Y_i \text{ indépendants} \Leftrightarrow E_i \text{ indépendants}$$

P4 **normalité** : les Y_i sont de loi normale autour du plan.

$$\Leftrightarrow E_i \text{ de loi normale autour de } 0$$

Nous pouvons vérifier ces postulats à l'aide des résidus de manière similaire à ce qui a été fait à la fin du module précédent.

Interprétation des paramètres

Quand toutes les variables entrent linéairement dans le modèle,

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + E,$$

les coefficients de régression $\beta_0, \dots, \beta_k$ s'interprètent comme suit :

β_0 : Valeur **moyenne** de Y lorsque $X_1, \dots, X_k$ prennent **simultanément** la valeur 0.

β_j : Hausse **moyenne** de Y lorsque X_j augmente d'une unité et que les valeurs des autres variables sont prises en compte mais demeurent inchangées.

Illustration — Génie du bois

Soit le modèle linéaire :

$$Y = 35 + 1,8 X_1 - 2,4 X_2 + 0,9 X_3 + E$$

Y : résistance à la flexion (en MPa)

X_1 : densité du panneau (kg/m^3)

X_2 : teneur en humidité (%)

X_3 : proportion de colle structurale utilisée (%)

- Si la teneur en humidité augmente de 1%, toutes choses égales par ailleurs, la résistance moyenne **diminue de 2,4 MPa**.
- Si la densité du panneau augmente de 1 kg/m^3 , la résistance augmente en moyenne de **1,8 MPa**.

Estimation ponctuelle des coefficients β_j

Les $\beta_0, \dots, \beta_k$ sont estimés en minimisant la somme des carrés des erreurs :

$$\begin{aligned} SS_E &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_k X_{ik}])^2 \end{aligned}$$

En version matricielle, le vecteur des paramètres estimés est

$$\hat{\beta} = (X'X)^{-1}X'Y$$

On en déduit que les $\hat{\beta}_j$ sont des combinaisons linéaires des Y_i , et donc qu'ils suivent une loi normale.

(Nous ne faisons pas ces calculs à la main.)

Exemple 1 : consommation d'essence per capita

Weisberg étudie la consommation d'essence des 48 états continentaux des États-Unis en 1971. Soit

Y_i : la consommation d'essence (gallons par habitant) dans l'état i ;

X_{i1} : les taxes sur l'essence en vigueur dans l'état i (en cents/gallon);

X_{i2} : le pourcentage de la population de l'état i possédant un permis de conduire;

X_{i3} : le revenu annuel per capita (en milliers \$) dans l'état i ;

X_{i4} : la longueur des routes fédérales (milliers de miles) dans l'état i .

Exemple 1

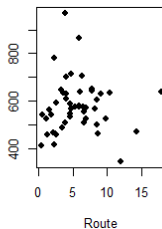
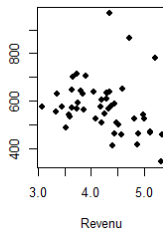
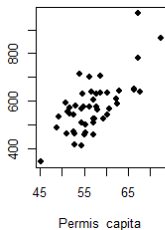
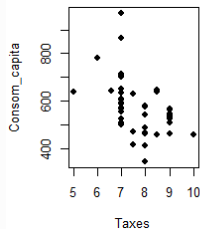
	A	B	C	D	E	F	G	H	I	J
1	Obs	Etat	Population	Taxes	Permis	Revenu	Route	Consommation	Permis_capita	Consom_capita
2	1	ME	1029	9.00	540	3.571	1.976	557	52.5	541
3	2	NH	771	9.00	441	4.092	1.250	404	57.2	524
4	3	VT	462	9.00	268	3.865	1.586	259	58.0	561
5	4	MA	5787	7.50	3060	4.870	2.351	2396	52.9	414
6	5	RI	968	8.00	527	4.399	.431	397	54.4	410
7	6	CN	3082	10.00	1760	5.342	1.333	1408	57.1	457
8	7	NY	18366	8.00	8278	5.319	11.868	6312	45.1	344
9	8	NJ	7367	8.00	4074	5.126	2.138	3439	55.3	467
10	9	PA	11926	8.00	6312	4.447	8.577	5528	52.9	464
11	10	OH	10783	7.00	5948	4.512	8.507	5375	55.2	498
12	11	IN	5291	8.00	2804	4.391	5.939	3068	53.0	580
13	12	IL	11251	7.50	5903	5.126	14.186	5301	52.5	471
14	13	MI	9082	7.00	5213	4.817	6.930	4768	57.4	525
15	14	WI	4520	7.00	2465	4.207	6.580	2294	54.5	508
16	15	MN	3896	7.00	2368	4.332	8.159	2204	60.8	566
17	16	IA	2883	7.00	1689	4.318	10.340	1830	58.6	635
18	17	MO	4753	7.00	2719	4.206	8.508	2865	57.2	603
19	18	ND	632	7.00	341	3.718	4.725	451	54.0	714
20	19	SD	579	7.00	419	4.716	5.915	501	72.4	865
21	20	NF	1525	8.50	1033	4.341	6.010	976	67.7	640

Exemple 1

D'après cette matrice de corrélation entre les variables, quelle serait la meilleure variable prise seule pour prédire Y?

```
> cor(essence[,c("Consom_capita", "Taxes", "Permis_capita", "Revenu", "Route")])
```

	Consom_capita	Taxes	Permis_capita	Revenu	Route
Consom_capita	1.000	-0.451	0.699	-0.245	0.019
Taxes	-0.451	1.000	-0.288	0.013	-0.522
Permis_capita	0.699	-0.288	1.000	0.157	-0.064
Revenu	-0.245	0.013	0.157	1.000	0.050
Route	0.019	-0.522	-0.064	0.050	1.000



Peut-on en déduire les 2 meilleures variables?
Le meilleur modèle à plusieurs variables?

Exemple 1

Si on inclut toutes les variables dans le modèle :

```
> LinearModel.5 <- lm(Consom_capita ~ Taxes + Permis_capita + Revenu + Route,data=essence)
> summary(LinearModel.5)
Call:
lm(formula = Consom_capita ~ Taxes + Permis_capita + Revenu +
    Route, data = essence)

Residuals:
    Min       1Q   Median       3Q      Max
-122.03  -45.57  -10.66   31.53  234.95

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    377.291    185.541   2.033 0.048207 *
Taxes          -34.790     12.970  -2.682 0.010332 *
Permis_capita   13.364      1.923   6.950 1.52e-08 ***
Revenu         -66.589     17.222  -3.867 0.000368 ***
Route           -2.426      3.389  -0.716 0.477999

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.31 on 43 degrees of freedom
Multiple R-squared:  0.6787, Adjusted R-squared:  0.6488
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

Exemple 1 :

L'équation de l'hyperplan estimé est

$$\hat{Y} = 377,29 - 34,79X_1 + 13,36X_2 - 66,59X_3 - 2,43X_4.$$

Interprétation de $\hat{\beta}_1$:

Si on augmente la taxe sur l'essence de 1 cent par gallon, mais qu'**on ne change pas la valeur des autres variables**, alors la consommation d'essence annuelle moyenne devrait diminuer de 34,79 gallons par habitant.

REMARQUE

- Si on retire une variable du modèle (p.ex. X_4), les estimations des coefficients de régression pour les autres variables (p.ex. β_1 , β_2 , β_3) sont susceptibles de changer.
- Si on garde le même échantillon, l'estimation de l'ordonnée à l'origine β_0 va changer selon les variables que l'on inclut dans le modèle.

Inférence

1) Test simultané de tous les coefficients

Première question importante en régression linéaire multiple :

“Est-ce qu’au moins une des variables explicatives est associée à la variable réponse?”

1) Test global de la régression.

$$H_0 : \beta_1 = \dots = \beta_k = 0$$

H_1 : au moins un paramètre parmi $\beta_1, \dots, \beta_k$ ne vaut pas zéro.

Analyse de la variance

Ce test s'effectue à l'aide d'une table d'analyse de la variance.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_k X_{ik} = \text{la } i^{\text{e}} \text{ valeur ajustée (ou prédite),}$$

$$e_i = Y_i - \hat{Y}_i = \text{le } i^{\text{e}} \text{ résidu.}$$

On définit les sommes de carrés

$$\left. \begin{aligned} SS_R &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ SS_E &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2 \\ SS_T &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \end{aligned} \right\} \begin{aligned} &\text{On peut montrer que} \\ &SS_T = SS_R + SS_E. \end{aligned}$$

Analyse de la variance

Si H_0 est vraie, les β_j sont tous nuls, l'hyperplan est horizontal et

- les prédictions $\hat{Y}_i$ sont toutes “proches” de la moyenne $\bar{Y}$;
- SS_R tend à prendre de faibles valeurs;

- $F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$ suit une loi $F_{k,n-k-1}$.

Analyse de la variance

Si H_0 est vraie, les β_j sont tous nuls, l'hyperplan est horizontal et

- les prédictions $\hat{Y}_i$ sont toutes “proches” de la moyenne $\bar{Y}$;
- SS_R tend à prendre de faibles valeurs;
- $F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$ suit une loi $F_{k,n-k-1}$.

Si H_1 est vraie, certains β_j sont non nuls, et

- les prédictions $\hat{Y}_i$ changent en fonction des valeurs des X_{ij} ;
- SS_R tend à prendre de grandes valeurs;
- $F_0 = \frac{MS_R}{MS_E}$ sera beaucoup plus grand que 1.

On rejettera donc H_0 au seuil α si $f_0 > F_{\alpha;k,n-k-1}$.

Table d'analyse de la variance

Source de variabilité	Somme de carrés	Deg. de liberté	Carrés moyens	F_0	Valeur p
Régression	SS_R	k	$MS_R = \frac{SS_R}{k}$	$f_0 = \frac{MS_R}{MS_E}$	$P(F_{k,n-k-1} > f_0)$
Erreur	SS_E	$n - k - 1$	$MS_E = \frac{SS_E}{n - k - 1}$		
Totale	SS_T	$n - 1$			

Table d'analyse de la variance

Source de variabilité	Somme de carrés	Deg. de liberté	Carrés moyens	F_0	Valeur p
Régression	SS_R	k	$MS_R = \frac{SS_R}{k}$	$f_0 = \frac{MS_R}{MS_E}$	$P(F_{k,n-k-1} > f_0)$
Erreur	SS_E	$n - k - 1$	$MS_E = \frac{SS_E}{n - k - 1}$		
Totale	SS_T	$n - 1$			

Si le test F rejette H_0 , on sait qu'au moins une des variables explicatives est associée à la valeur de Y. Laquelle ou lesquelles des k variables le sont ?

⇒ sélection de variables (plus loin).

Exemple 2 sur la consommation d'essence

Residual standard error: 66.31 on 43 degrees of freedom
Multiple R-squared: 0.6787, Adjusted R-squared: 0.6488
F-statistic: 22.71 on 4 and 43 DF, p-value: 3.907e-10

Source de variabilité	Somme de carrés	Degrés de liberté	Carré moyen	F_0	Valeur P
Régression	399 425	4	99 856,32	22,71	$3,907 \times 10^{-10}$
Erreur	189 072	43	$66,31^2 = 4 397,02$		
Totale	588 497	47			

Expliquez pourquoi les données supportent l'hypothèse qu'*au moins une des quatre variables explicatives est associée à la consommation d'essence* (pour tous les seuils usuels).

Test partiel sur un coefficient β_j

Objectif : tester la force de l'effet de la variable X_j sur Y ,
en tenant compte des autres variables du modèle.

$H_0 : \beta_j = c$ (souvent $c = 0$)

$\Rightarrow$ Compte tenu de la présence des autres variables, une hausse de 1 unité de X_j fait varier Y de c unités en moyenne

$H_1 : \beta_j \neq c$

$\Rightarrow$ En présence des autres variables, β_j diffère de c .

Statistique de test et IC

Statistique de test :

$$T_0 = \frac{\hat{\beta}_j - c}{\text{se}(\hat{\beta}_j)} \sim t_{n-k-1} \quad \text{si } H_0 \text{ est vraie}$$

Intervalle de confiance pour β_j au niveau $1 - \alpha$:

$$\hat{\beta}_j \pm t_{\alpha/2; n-k-1} \cdot \text{se}(\hat{\beta}_j)$$

Décision :

- On rejette H_0 si $|T_0| > t_{\alpha/2; n-k-1}$
- ou si c ne se trouve pas dans l'intervalle de confiance

Exemple 3 sur l'essence

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	377.291	185.541	2.033	0.048207	*
Taxes	-34.790	12.970	-2.682	0.010332	*
Permis_capita	13.364	1.923	6.950	1.52e-08	***
Revenu	-66.589	17.222	-3.867	0.000368	***
Route	-2.426	3.389	-0.716	0.477999	

```
> Confint(LinearModel.5, level=0.95)
```

	Estimate	2.5 %	97.5 %
(Intercept)	377.291146	3.111674	751.470618
Taxes	-34.790149	-60.947054	-8.633245
Permis_capita	13.364494	9.486432	17.242555
Revenu	-66.588752	-101.319716	-31.857787
Route	-2.425889	-9.260810	4.409033

- L'effet d'une hausse de taxes est-il significatif, une fois qu'on tient compte des permis, du revenu et des routes? ($\alpha = 0,05$)
- L'effet d'une hausse de 1% des permis diffère-t-il de 10 gallons de plus per capita en moyenne? ($\alpha = 0,05$)

QUIZ!

Vrai ou Faux?

Les coefficients estimés, $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ ont tous la même marge d'erreur dans un intervalle de niveau 0,95.

Multicolinéarité

- Lorsque les $X_1, \dots, X_k$ sont corrélées entre elles, elles contiennent la même information $\Rightarrow$ pas nécessaire de toutes les inclure dans le modèle.
- Si X_1 et X_2 sont très corrélées entre elles, et corrélées avec Y , et qu'on les inclut toutes les deux dans le modèle, on pourrait voir :
 - un test global significatif;
 - deux tests partiels non significatifs.

Prévision

Prévision de la MOYENNE de Y : $E[Y|X_1, \dots, X_k]$

On veut prédire la **valeur moyenne** de Y pour tous les individus de la population ayant les mêmes valeurs pour les variables explicatives, disons $X_1 = x_{01}, \dots, X_k = x_{0k}$. La prévision ponctuelle sera

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_k x_{0k}.$$

L'erreur-type associée à $\hat{Y}_0$ sera donnée par

$$se(\hat{Y}_0) = \frac{\sqrt{\widehat{Var}(\hat{\beta}_0) + x_{01}^2 \widehat{Var}(\hat{\beta}_1) + \dots + x_{0k}^2 \widehat{Var}(\hat{\beta}_k) + 2x_{01} \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \dots + 2x_{0k} \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_k) + 2x_{01}x_{02} \widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2) + \dots + 2x_{0(k-1)}x_{0k} \widehat{Cov}(\hat{\beta}_{k-1}, \hat{\beta}_k)}}{1}$$

L'intervalle de confiance à $100(1 - \alpha)\%$ pour $E[Y|X_1, \dots, X_k]$ est donné par

$$\hat{Y}_0 \pm t_{\alpha/2; n-k-1} se(\hat{Y}_0)$$

Prévision d'une observation INDIVIDUELLE de Y

Si l'on cherche plutôt à prédire **une seule observation** de Y, pour certaines valeurs de $(X_1, \dots, X_k)$, alors la prévision ponctuelle demeure la même $(\hat{Y}_0)$. Cependant l'erreur-type sera plus grande :

$$\begin{aligned} se(\hat{Y}_0 + E_0) &= \sqrt{\widehat{Var}(\hat{Y}_0) + \widehat{Var}(E_0)} \\ &= \sqrt{[se(\hat{Y}_0)]^2 + MS_E} \end{aligned}$$

L'intervalle de confiance correspondant, souvent appelé intervalle de prévision, est

$$\hat{Y}_0 \pm t_{\alpha/2; n-k-1} se(\hat{Y}_0 + E_0)$$

Exemple 5 sur la consommation d'essence

On retire la variable Route de notre modèle. On obtient :

```
lm(formula = Consom_capita ~ Taxes + Permis_capita + Revenu,  
    data = essence)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	307.328	156.831	1.960	0.05639	.
Taxes	-29.484	10.584	-2.786	0.00785	**
Permis_capita	13.748	1.837	7.485	2.24e-09	***
Revenu	-68.023	17.010	-3.999	0.00024	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.94 on 44 degrees of freedom

Multiple R-squared: 0.6749, Adjusted R-squared: 0.6527

F-statistic: 30.44 on 3 and 44 DF, p-value: 8.235e-11

Exemple sur la consommation d'essence

On s'intéresse aux prédictions du modèle linéaire dans les conditions suivantes :

- taxe sur l'essence de 7 cents par gallon,
 - pourcentage de la population avec permis de conduire de 54%
 - revenu per capita de 4000\$
- (i) Donnez un intervalle de confiance à 95% pour la consommation d'essence par habitant en moyenne dans des états où les conditions ci-dessus sont rencontrées.
- (ii) Que devient votre prédiction et sa marge d'erreur si on remplace “en moyenne dans des états” dans la phrase précédente par “dans un état en particulier”?

Exemple (suite)

$\hat{y}_0$: Estimation ponctuelle de 571,22 gallons par habitant dans les deux cas

- i) intervalle pour la moyenne : (542,69; 599,76) gallons
- ii) intervalle pour une valeur individuelle : (435,31; 707,14) gallons

```
> predict(LinearModel.2, newdata=essencePLUS, interval="confidence")
      fit      lwr      upr
1 571.2247 542.6901 599.7593

> predict(LinearModel.2, newdata=essencePLUS, interval="prediction")
      fit      lwr      upr
1 571.2247 435.307 707.1425
```

Critères de qualité d'ajustement

Sélection de variables

- Trop de variables dans le modèle ?

La variance des prévisions est trop grande.

- Manque de variables dans le modèle ?

Les prévisions sont biaisées.

- Il existe une multitude de *critères servant à évaluer la qualité d'un modèle de régression.*

Deux critères se calculent à partir du tableau d'analyse de la variance : R^2 et R^2_{ajuste} .

Deux critères d'ajustement

- Coefficient de détermination, R^2 , proportion de la variation de Y expliquée par le modèle :

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Problème pour la comparaison de modèles : le SS_R augmente toujours quand on ajoute des variables explicatives au modèle, même si elles sont inutiles...

Deux critères d'ajustement

- Coefficient de détermination, R^2 , proportion de la variation de Y expliquée par le modèle :

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Problème pour la comparaison de modèles : le SS_R augmente toujours quand on ajoute des variables explicatives au modèle, même si elles sont inutiles...

- Coefficient de détermination ajusté, R_{ajuste}^2 :

$$R_{ajuste}^2 = 1 - \frac{SS_E/(n - k - 1)}{SS_T/(n - 1)}.$$

Plus approprié pour sélectionner les variables.

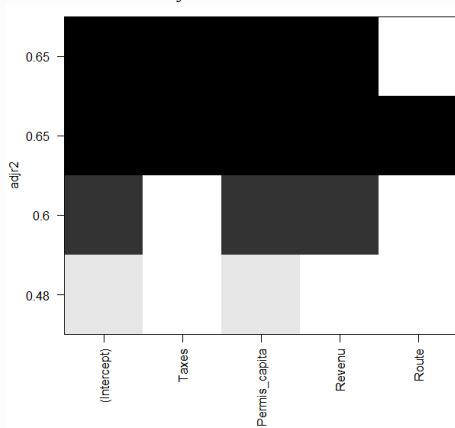
Procédure de sélection de modèle

Pour choisir les variables explicatives à retenir dans un modèle de régression :

1. ajuster tous les sous-modèles possibles et calculer R_{ajuste}^2 pour chacun d'entre eux;
2. conserver le modèle qui a le meilleur (plus grand) R_{ajuste}^2 .

Exemple 4 sur la consommation d'essence

D'après le R^2_{ajuste} , le meilleur sous-modèle est ...



QUIZ!

Vrai ou Faux?

- La valeur du MS_E est toujours plus petite pour un modèle contenant plus de variables.
- La valeur du MS_E diminue généralement quand on augmente la taille d'échantillon.

Une autre approche

Le critère que nous venons de voir nous donne de l'information sur la *qualité* du modèle.

Et si nous voulions plutôt avoir le meilleur modèle **prédictif**?

Surajustement et généralisation

Quand on utilise plusieurs variables explicatives, on risque de tomber dans le **surajustement**.

- Le modèle prédit très bien les données qui ont servi à estimer celui-ci, mais **échoue à bien prédire** de nouvelles données.
- On aimerait plutôt construire un modèle qui prédit bien à de nouvelles observations.

⇒ On a besoin d'**outils pour estimer la performance prédictive réelle du modèle**.

⇒ Nous n'avons qu'un seul jeu de données, que peut-on faire ?

Validation croisée

La **validation croisée** est une méthode pour évaluer la performance prédictive d'un modèle à partir des données disponibles.

- On **divise les données** en un ensemble d'entraînement et un ensemble de test (validation).
- On **ajuste le modèle sur l'ensemble d'entraînement**, puis on **évalue l'erreur de prédiction** sur l'ensemble test.
- Pour limiter l'effet du découpage, on répète cette opération plusieurs fois (ex : **validation croisée en K plis**).

⇒ Fournit une estimation plus réaliste de la performance prédictive.

Implémentation de la validation croisée en K plis

Pour calculer l'erreur de prédiction pour un modèle donné à l'aide de la validation croisée en K plis

1. Partitionner aléatoirement les n observations du jeu de données en K groupes.
2. Pour **chacun** des groupes 1 à K :
 - (a) Utiliser les données du groupe comme **jeu de test** et les données des $K - 1$ autres groupes comme **jeu d'entraînement**.
 - (b) Estimer les coefficients β à partir du **jeu d'entraînement**.
 - (c) Utiliser les $\hat{\beta}$ obtenus en (b) et les X des données du **jeu de test** pour calculer les $\hat{Y}$ pour les données du **jeu de test**.
 - (d) Calculer $SSE_{\text{pred}}^{(k)} = \sum_{i \in \text{jeu de test}} (\hat{Y}_i - Y_i)^2$.
3. Calculer $MSE_{\text{pred}} = \frac{1}{n} \sum_{k=1}^K SSE_{\text{pred}}^{(k)}$.

Les livres classiques de science des données (par exemple, *Elements of Statistical Learning*) recommandent souvent une valeur de $K = 10$.

Exemple : sélection de variables avec validation croisée

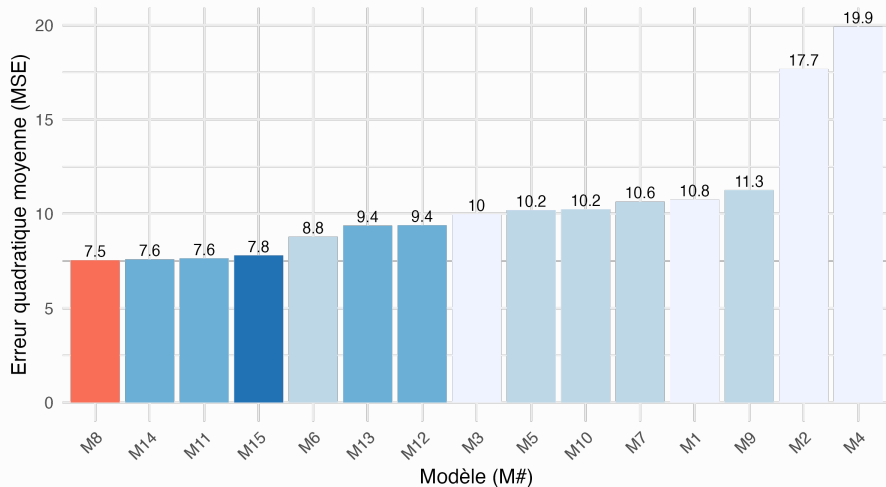
Supposons qu'on dispose de 4 variables explicatives $X_1, \dots, X_4$. On veut déterminer lesquelles conserver dans le modèle pour maximiser la capacité de prédiction.

- On considère différents sous-modèles (ex : toutes les combinaisons possibles de variables).
- Pour chaque modèle, on effectue une **validation croisée en K plis** et on calcule l'erreur de prédiction moyenne (p.ex., MSE) [voir page suivante pour les détails].
- On choisit le modèle avec la plus faible erreur de validation.

Exemple

Validation croisée (K = 10) — Modèles à 1 à 4 variables

Meilleur modèle : X2, X3



Ouverture : vers l'apprentissage machine

La régression linéaire multiple est un modèle simple mais fondamental.

Dans le contexte de grandes bases de données ou de relations complexes, d'autres méthodes sont souvent utilisées :

- arbres de décision, forêts aléatoires (random forests)
- modèles à base de voisins (k-NN)
- régularisation (ridge, lasso)
- réseaux de neurones et autres méthodes d'apprentissage profond

Ces approches s'appuient toutes sur des techniques de validation croisée pour éviter le surajustement.

Flexibilité du modèle

Flexibilité du modèle

Modèle linéaire : lien linéaire par rapport aux paramètres β_0 à β_k . Sa formulation permet d'exprimer des relations très générales entre Y et les X_j .

- relations polynomiales
- modèle avec interaction entre les X_j
- analyse de la variance
- etc.

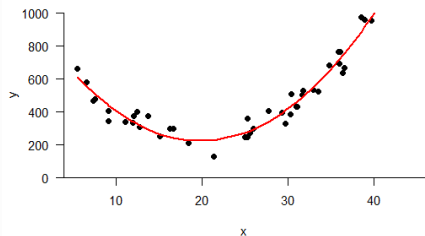
Exemple 6 : Relation quadratique

En posant $X_1 = X$ et $X_2 = X^2$, on a

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E \\ &= \beta_0 + \beta_1 X + \beta_2 X^2 + E, \end{aligned}$$

soit une **relation quadratique** entre Y et X .

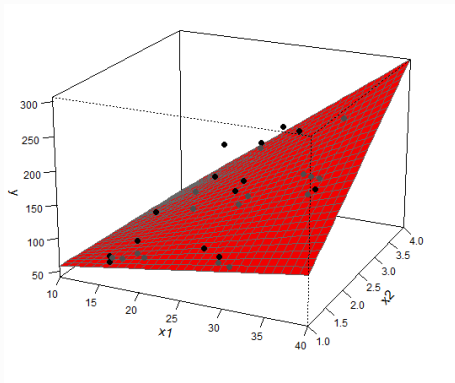
Quelle valeur de X minimise la valeur moyenne de Y ?



Exemple 7 : interactions entre deux variables explicatives

On modélise l'interaction en posant $X_3 = X_1X_2$ et

$$\begin{aligned} Y &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + E \\ &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + E. \end{aligned}$$



Attention !

En présence d'interaction ou de termes de degré supérieur, les coefficients ne s'interprètent pas de la même façon !

Avec interaction,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + E,$$

si X_2 passe de 3 à 4 unités, l'effet sur $E[Y]$ est :

$$\begin{aligned} E[Y|X_2 = 4] - E[Y|X_2 = 3] &= \beta_0 + \beta_1 X_1 + \beta_2(4) + \beta_3 X_1(4) \\ &\quad - [\beta_0 + \beta_1 X_1 + \beta_2(3) + \beta_3 X_1(3)] \\ &= \beta_2 + \beta_3 X_1. \end{aligned}$$

Ainsi l'effet d'un changement dans la valeur de X_2 *dépend* de la valeur de X_1 , d'où le nom de “terme d'interaction” pour le terme $X_1 X_2$.

Exemple 8 : Lien avec l'analyse de la variance

Supposons une expérience à un facteur à 3 niveaux (A, B et C). Posons

$$X_1 = \begin{cases} 1 & \text{au niveau A} \\ 0 & \text{sinon} \end{cases} \quad X_2 = \begin{cases} 1 & \text{au niveau B} \\ 0 & \text{sinon} \end{cases}$$

Si on utilise le modèle linéaire avec $E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, on retrouve le modèle d'analyse de la variance à un facteur.

Réponses aux exemples

1. Permis_capita est la plus corrélée avec Y = consommation d'essence. On ne peut pas déduire le meilleur modèle à plusieurs variables, car les coefficients dépendent de la présence des autres variables.
2. H_0 rejetée. Au moins une des quatre variables explicatives est associée à la consommation d'essence.
3. Valeur-P pour Taxes = 0,0103 < 0,05. On rejette $H_0 : \beta_1 = 0$, l'effet est significatif.
Puisque $10 \in IC(\beta_2)$, on ne rejette pas $H_0 : \beta_2 = 10$.
4. La ligne du haut représente le R_{ajuste}^2 maximal. On retient tout sauf Routes.
5. $\hat{Y}_0 = 307,328 - 29,484(7) + 13,748(54) - 68,023(4) = 571,22$
6. Le minimum de $E(Y)$ est atteint en $x = -\frac{\beta_1}{2\beta_2}$.