Cours 2
Régression Multiple

ISMIN ITS

Ariane Ferreira

Ecole Nationale
Supérieure des Mines
SAINT-ETIENNE

Centre Microélectronique de Provence

Sommaire

Cours 1 : Introduction à l' Analyse de Données
Cours 2 : Régression Multiple
Cours 3 : Analyse en Composantes principales
Cours 4 : Régression PLS (Partial Least Squares)

Références

[1] Lebart, L.; Piron, M.; Morineau, A. Statistique Exploratoire Multidimensionnelle, 4 ed. Dunod.
[2] Saporta, G. Probabilités et Statistique – analyse de données, 2 ed.Technip.
[3] Tenenhaus, M. Statistique Méthodes pour décrire, expliquer et prévoir, ed. Dunod.
[4] Tenenhaus, M. La Régression PLS théorie et pratique , ed. Technip.
[5] Montgomery, D.C.; Runger, G.C. Applied Statistics and Probability for Engineers, fourth edition, 2007, John Wiley & Sons, inc.

Lien intéressant :

http://faculty.chass.ncsu.edu/garson/PA765/regress.htm

Modèle Linéaire Simple

Le modèle simple

- X et Y deux variables continues.
- Les valeurs x_i de X sont contrôlées et sans erreur de mesure.
- On observe les valeurs correspondantes $y_1, \ldots, y_n$ de Y.

Exemples

- X peut être le temps et Y une grandeur mesurée à différentes dates.
- Y peut être la différence de potentiel mesurée aux bornes d'une résistance pour différentes valeurs de l'intensité du courant.

Hypothèse

- X et Y ne sont pas indépendantes et la connaissance de X permet d'améliorer la connaissance de Y.
- la valeur moyenne E(Y|X=x), l'espérance conditionnelle de Y sachant que X = x.

Fonction Linéaire

$$\mathrm{E}(\mathrm{Y}_\mathrm{i}) = \alpha + \beta x_i$$

$$\mathrm{Y}_\mathrm{i} = \alpha + \beta x_i + \varepsilon_i$$

Avec $\mathrm{E}(\varepsilon_i)=0$, pour tout $i=1,\ldots,n$

n = nb d'observations; ε_i =résidu de l'obs i

Les Données

Y = Variable à expliquer

- numérique
- (ou dépendante)

X_1, …, X_p = Variables explicatives

- numériques ou binaires
- (ou indépendantes)

Le tableau des données

	X_1	…	X_p	Y
1	x_{11}	…	X_{1p}	y_1
⋮				
i	x_{1i}	…	X_{ip}	y_i
⋮				
n	x_{1n}	…	X_{np}	y_n

où les x_{ji} sont fixes et ε_i est un terme aléatoire de loi $N(0,\sigma)$;
Les ε_i sont indépendants les uns des autres.

Le modèle linéaire multiple

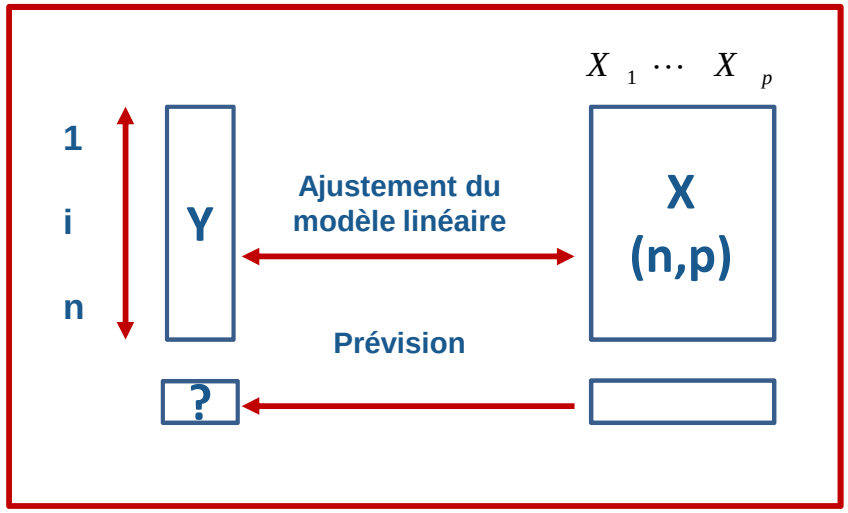
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Supposition : indépendance linéaire des X_i.

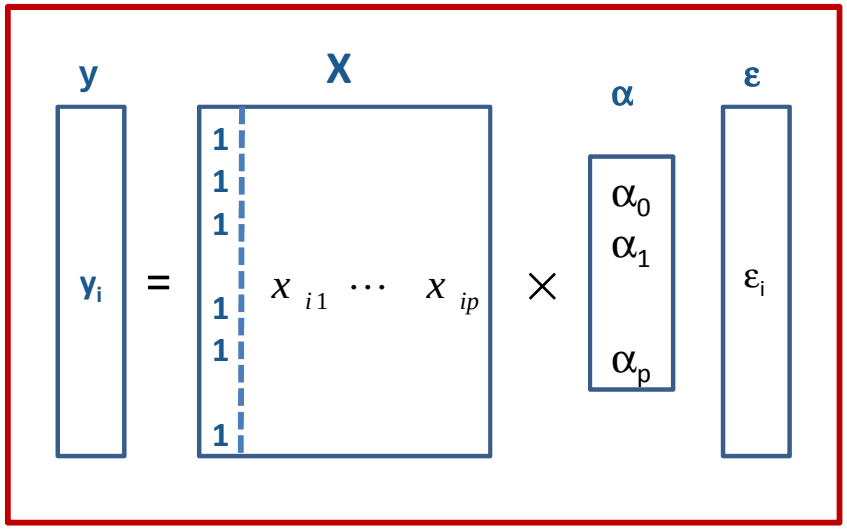
Modèle sous forme matricielle

$$y = \mathbf{X}\boldsymbol{\alpha} + \varepsilon$$

(n,1) (n,p+1) (p+1,1) (n,1)



Prévision Linéaire



Schématisation du modèle Linéaire

Hypothèses du modèle linéaire

Résidus

- La variance des résidus est la même pour toutes les valeurs de X
  - **Homoscédasticité** : $V(\varepsilon_i) = \sigma^2$
- Les résidus sont linéairement indépendants : $cov(\varepsilon_i,\varepsilon_j) = 0 \; \forall \; i \neq j$
- Les résidus sont normalement distribués : $\varepsilon_i \sim N(0,\sigma^2)$

L'existence de la composante stochastique (ε_i) correspond au fait que :

- **variation synchronique :** individus avec même valeur x_i peuvent avoir des réponses Y différentes;
- **variation diachronique** : un même individu mesuré à plusieurs reprises avec la même valeur x_i peut avoir des réponses Y différentes.

- **On a équivalence de** $\varepsilon_i \sim N(0,\sigma^2)$ et $Y/X=x_i \sim N(\alpha + \beta x_i,\sigma^2)$

Les problèmes

1. **Estimation des coefficients de régression**

   $\beta_0, \beta_1, \ldots, \beta_p$.
2. **Estimation de l'écart-type σ du terme résiduel ε_i**
3. **Analyse des résidus**
4. **Mesurer la force de la liaison entre Y et les variables $X_1,\ldots,X_p$**

   Coefficients : de corrélation multiple (R), de détermination (R^2)
5. **La liaison globale entre Y et $X_1,\ldots,X_p$ est-elle significative ?**
6. **L'apport marginal de chaque variable X_j (en plus des autres) à l'explication de Y est-il significatif ?**
7. **Sélection automatiques des « bonnes » variables X_j.**
8. **Comparaison de modèles.**
9. **Intervalle de prévision à 95% de y.**
10. **Intervalle de confiance à 95% de E(Y).**

Estimation des coefficients de régression β_j

Notations :

y_i = valeur observée

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}$$

= valeur calculée

= prévision de y_i

- $e_i = y_i - \hat{y}_i$ = erreur

Méthode des moindres carrés :

On recherche $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ minimisant $\sum_{i=1}^{n} e_i^2$.

Des modèles sur des échantillons différents donneront des estimateurs différents.
D'où une variance des estimateurs.

Estimation des coefficients de régression β_j

Afin de simplifier la notation nous utiliserons a à la place de β
Système de *n* équations et *p* inconnus :

$$y_1 = a_0 + a_1 x_{1,1} + a_2 x_{1,2} + \cdots + a_{p-1} x_{1,p-1} + \varepsilon_1$$
$$\vdots = \vdots$$
$$y_n = a_0 + a_1 x_{n,1} + a_2 x_{n,2} + \cdots + a_{p-1} x_{n,p-1} + \varepsilon_n$$

Sous forme matricielle :

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} \qquad \mathbf{X} \qquad \mathbf{a} \qquad \boldsymbol{\varepsilon}$$

Estimation des coefficients de régression

Les coefficients a_0, ..., a_{p-1} sont obtenus par la minimisation des moindres carrées :

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = (\mathbf{y} - \mathbf{X}\mathbf{a})^T (\mathbf{y} - \mathbf{X}\mathbf{a})$$

La solution est donnée par:

$$\hat{\mathbf{a}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y} = \mathbf{C}\mathbf{X}^\mathbf{T}\mathbf{y}$$

$$\mathbf{C} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1} \text{ est une matrice symétrique de taille } (p,p)$$

Propriétés statistiques de $\hat{\mathbf{a}}$:

$$E(\hat{\mathbf{a}}) = \mathbf{a}$$

$$V(\hat{\mathbf{a}}) = \sigma^2 \mathbf{C}$$

Les valeurs prédites par le modèle:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{a}} = \mathbf{X}\mathbf{C}\mathbf{X}^\mathbf{T}\mathbf{y} = \mathbf{H}\mathbf{y}$$

$$\mathbf{H} = (\mathbf{X}\mathbf{C}\mathbf{X}^\mathbf{T}) \text{ est une matrice symétrique qui vérifie } \mathbf{H}^2 = \mathbf{H}$$

Vecteur des résidus

Le vecteur des résidus du modèle :

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$
$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Propriétés orthogonales des résidus :

$$\mathbf{1}^T \mathbf{e} = 0$$
$$\hat{\mathbf{y}}^T \mathbf{e} = 0$$
$$\mathbf{X}^T \mathbf{e} = 0$$

Estimation de l'écart-type σ du terme résiduel :

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2 \quad \Rightarrow \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Sommes des carrés

Décomposition de la somme des carrés totale :

$$\underbrace{\sum(y_i - \bar{y})^2}_{\text{Somme des carrés totale}} = \underbrace{\sum(\hat{y}_i - \bar{y})^2}_{\text{Somme des carrés expliquée Régression}} + \underbrace{\sum e_i^2}_{\text{Somme des carrés résiduelle (erreurs)}}$$

Valeur moyenne de la variable réponse y : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\mathbf{1}^T\mathbf{y}}{n}$

Somme des Carrés Totale : $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y}^T\mathbf{y} - \frac{(\mathbf{1}^T\mathbf{y})^2}{n}$

Somme des Carrés Régression : $SCR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\mathbf{a}}\mathbf{X}^T\mathbf{y} - \frac{(\mathbf{1}^T\mathbf{y})^2}{n}$

Somme des Carrés Erreurs : $SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}^T\mathbf{y} - \hat{\mathbf{a}}^T\mathbf{X}^T\mathbf{y}$

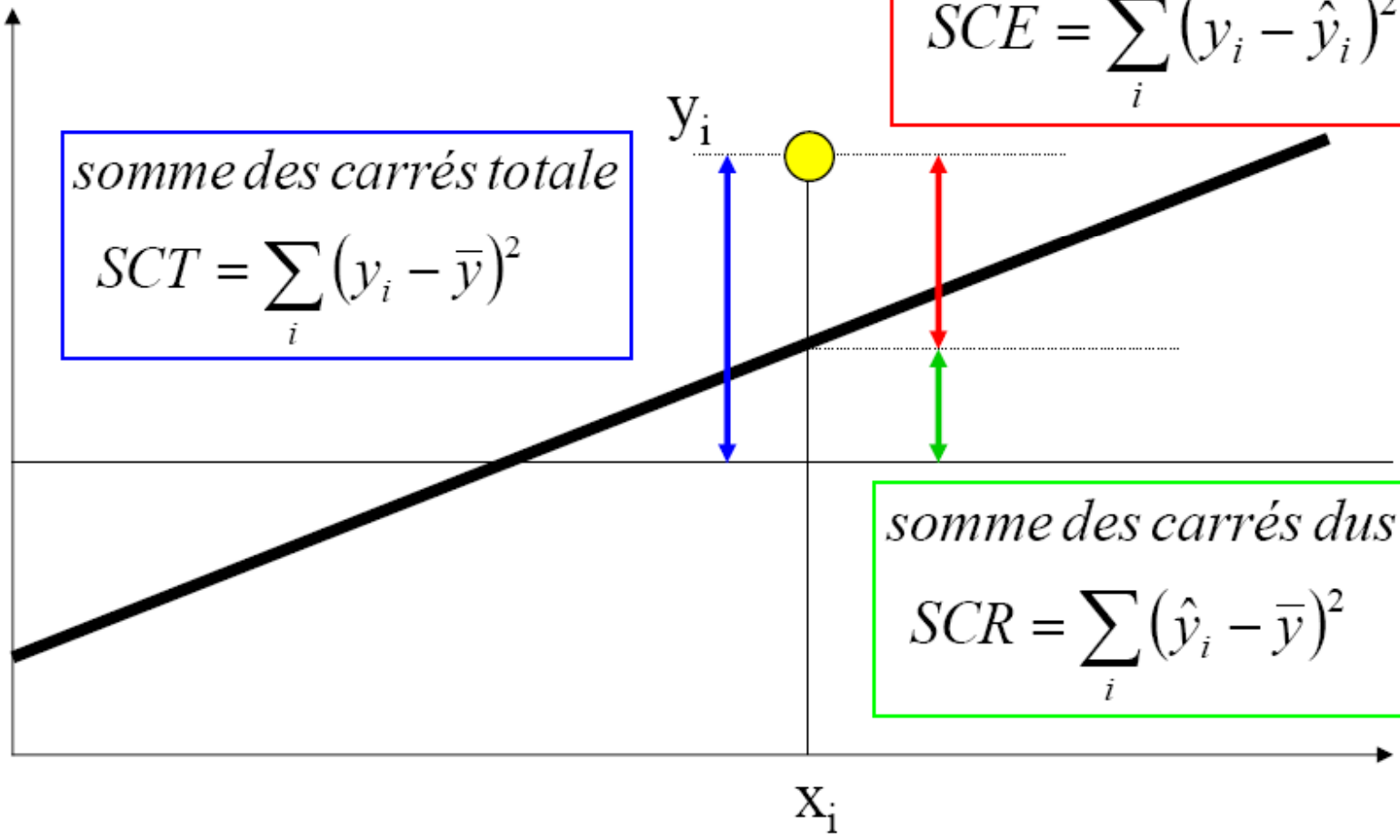
Sommes des carrés

somme des carrés résiduels

$$SCE = \sum_i (y_i - \hat{y}_i)^2$$

somme des carrés totale

$$SCT = \sum_i (y_i - \bar{y})^2$$



somme des carrés dus à la rég.

$$SCR = \sum_i (\hat{y}_i - \bar{y})^2$$

Carrés Moyens

Somme des Carrés : SCT=SCR+SCE

Carré Moyen de la Régression :

$$CMR = \frac{SCR}{p-1}$$

Carré Moyen Résiduel (Erreurs) :

$$CME = \frac{SCE}{n-p}$$

SCR = somme des carrés Régression
SCE = somme des carrés Erreurs
p = nombre de variables
n = nombre d'observations

Coefficient de détermination multiple $R^2 \in (0,1)$

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Coefficient de détermination Ajusté $R_a^2 \in (0,1)$

R^2 augment toujours avec l'addition de variables explicatives au modèle.

Comment comparer les R^2 de deux modèles construits avec des nombres d'observations et des variables différents ?

- Le R_a^2 permet de tenir compte du nombre d'observations et du nombre de variables explicatives.
- On modifie le coefficient R^2 en tenant compte du nombre de degrés de liberté
  - de la somme des carrés totale (n-1) et
  - de la somme des carrés résiduelle (n-p-1)

$$R_a^2 = 1 - \frac{SCE/(n-p)}{SCT/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)\left(1-R^2\right)$$

- **Grâce au R_a^2 on peut comparer les pouvoir explicatifs de différents modèles.**

Mesure de la multi-colinéarité : Tolérance et VIF

Tolérance

- **Tolérance** (Xj) = 1 - R^2(Xj ; Autres X)
- Il est préférable d'observer une tolérance supérieure à 0.33.

VIF

- **Variance Inflation Factor** = 1 / Tolérance
- Il est préférable d'observer un VIF inférieur à 3.

La multi-colinéarité

$S(X_1,\ldots,X_p)$ est la somme des carrés expliquée par les variables $X_1,\ldots,X_p$.

F partiel

$$F_j = t_j^2 = \frac{\hat{a}_j^2}{s_j^2} = \frac{S(X_1,\cdots,X_p) - S(X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_p)}{\hat{\sigma}^2_{\text{modèle complet}}}$$

On obtient un $|t_j|$ petit si :

|cor(Y,Xj)| est petite ou bien

$R^2(X_j$; Autres variables X) est grande.

Le Test d'hypothèse Globale

La liaison globale entre Y et les variables explicatives $X_1,\ldots,X_p$ est-elle significative?

- **Test d'hypothèse :**

$$H_0 : \forall j,\ a_j = 0$$

$$H_1 : \exists j,\ a_j \neq 0$$ au moins un coefficient différent non nul.

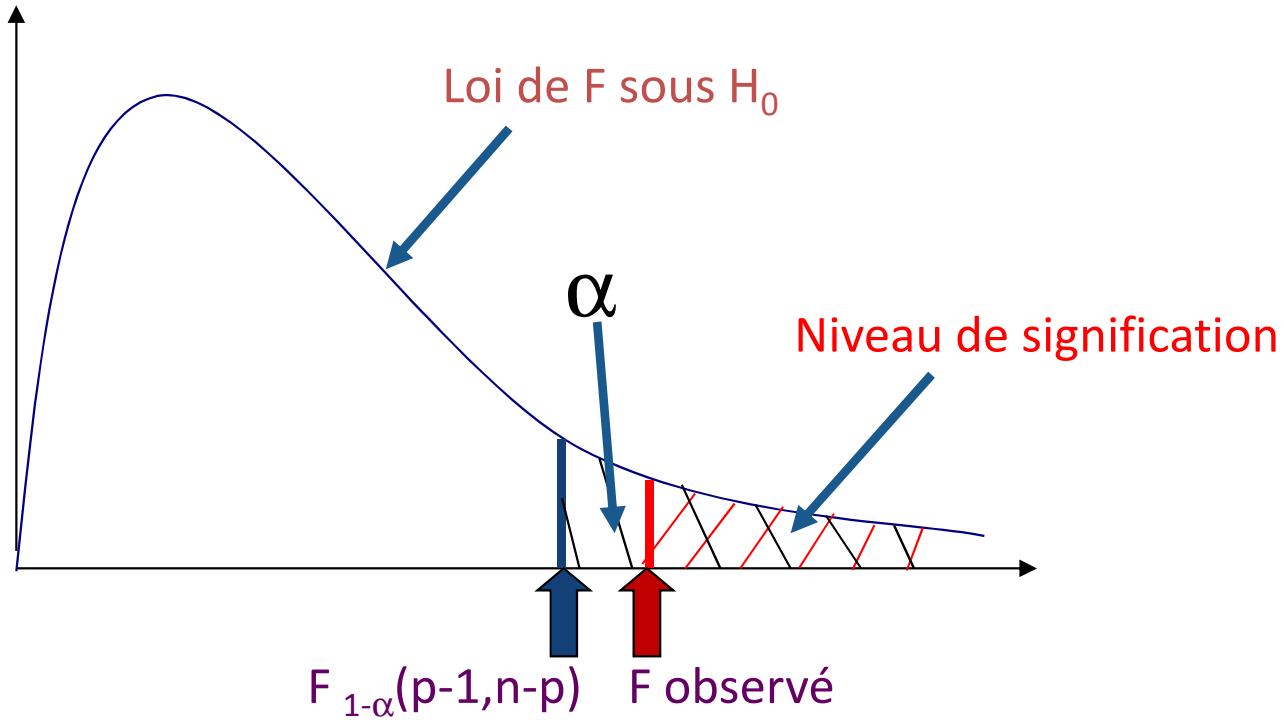
- Si l'hypothèse H_0 est acceptée : la variable Y ne dépend pas du tout des variables $X_1,\ldots,X_p$.
- l'hypothèse H_0 est rejetée si : $1 - F_F\left(\frac{CMR}{CME}, p-1, n-p\right) \leq \alpha$

- **Tableau ANOVA** :
  - Décision de rejeter H0 au risque α de se tromper :
  - Rejet de H_0 si $F \geq F_{1-\alpha}(p-1, n-p)$

Fractile d'une loi de Fisher-Snedecor

Le Test d'hypothèse Globale Niveau de signification

Plus petite valeur de α conduisant au rejet de H_0



On rejette H_0 au risque α de se tromper si $NS \leq \alpha$

Les Tests d'hypothèse Marginaux

Lorsque le test global conduit au rejet de l'hypothèse nulle, il faut rechercher quels sont les coefficients de régression $\hat{a}_j$ significatifs (différent de zéro) :

- **Test d'hypothèse :**

$$H_0 : a_j = 0$$

$$H_1 : a_j \neq 0$$

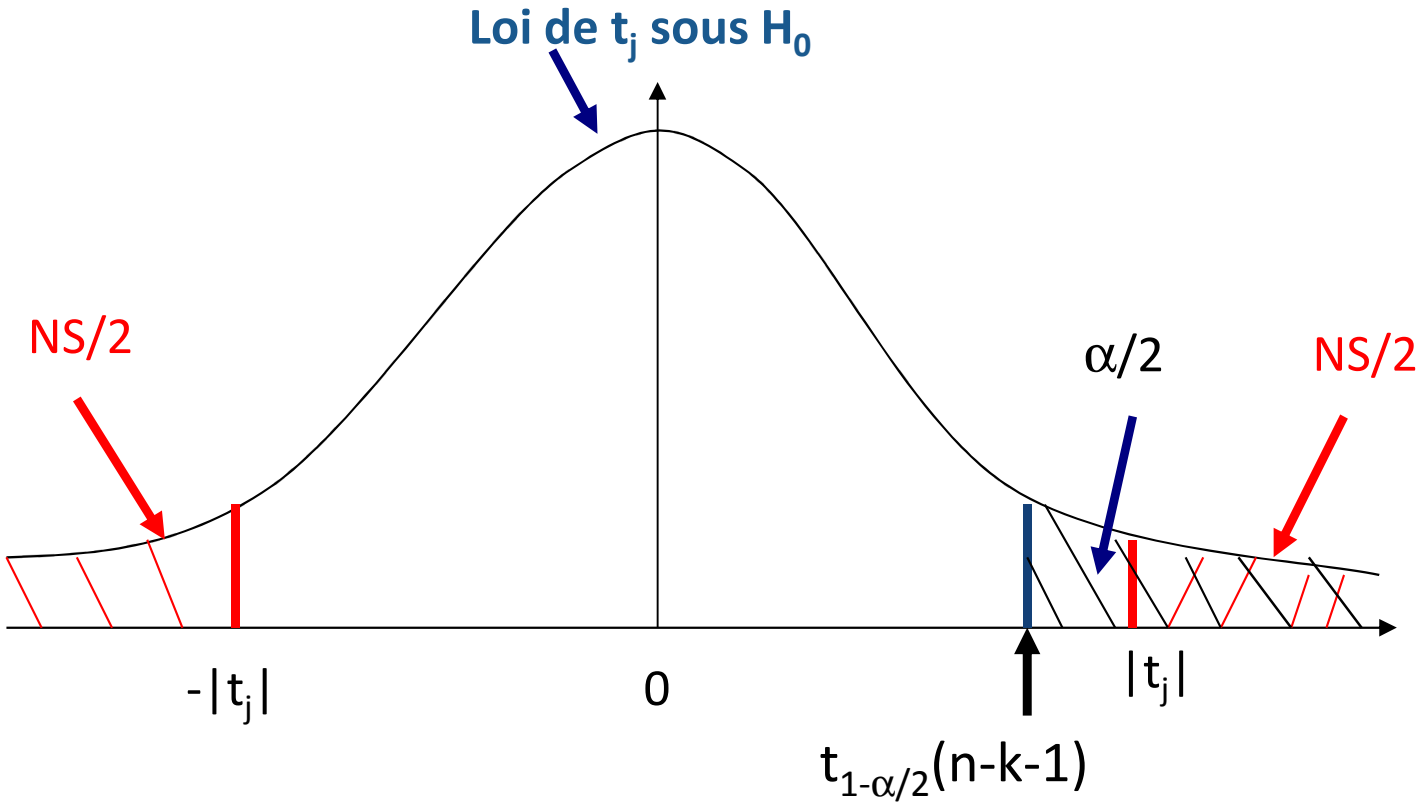
- l'hypothèse H_0 est rejetée si : $2\left\{1 - F_T\left(\left|\frac{\hat{a}_j}{\sqrt{c_{jj} CME}}\right|, n-p\right)\right\} \leq \alpha$

- c_{jj} est l'élément diagonal de la matrice **C** correspondant à $\hat{a}_j$.
- **Tableau ANOVA** :
  - Décision de rejeter H_0 au risque α de se tromper :
  - Rejet de H_0 si $|t_j| \geq t_{1-\alpha/2}(n-p)$

Fractile d'une loi de Student

Le Tests d'hypothèses Marginaux Niveau de signification

Plus petite valeur de α conduisant au rejet de H_0



On rejette « $H_0 : \beta_j = 0$ » au risque α de se tromper si $NS \leq \alpha$

Intervalles de confiance

Intervalle de Confiance des coefficients de régression $\hat{a}_j$

L'intervalle de confiance du coefficient $\hat{a}_j$ au niveau 1-α est donnée par la formule

$$(a_j)_L = \hat{a}_j - F_T^{-1}(1-\alpha/2, n-p)\sqrt{c_{jj}CME}$$

$$(a_j)_U = \hat{a}_j + F_T^{-1}(1-\alpha/2, n-p)\sqrt{c_{jj}CME}$$

•Intervalle de Confiance des valeurs prédits par le modèle

$$(\mathbf{x}^T\mathbf{a})_L = \hat{y} - F_T^{-1}(1-\alpha/2, n-p)\sqrt{CME(\mathbf{x}^T\mathbf{C}\mathbf{x})}$$

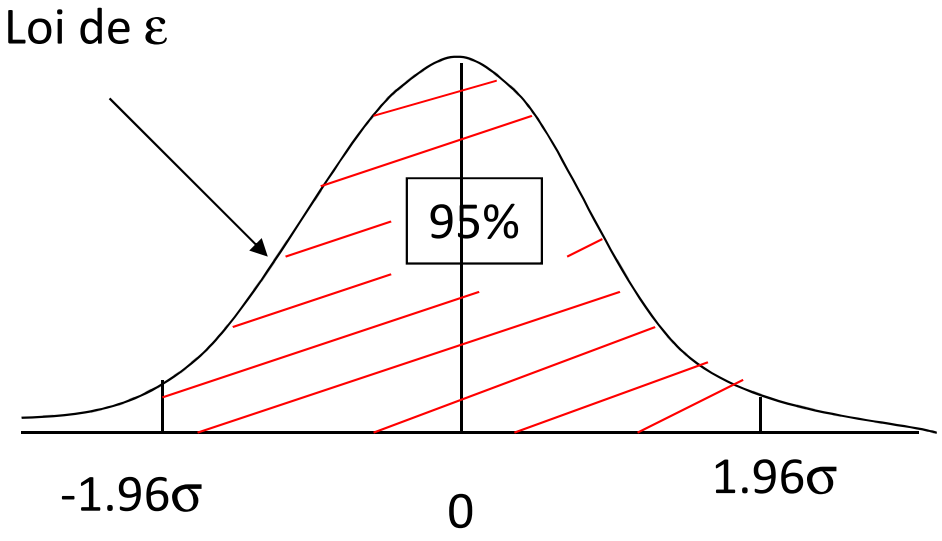
$$(\mathbf{x}^T\mathbf{a})_U = \hat{y} + F_T^{-1}(1-\alpha/2, n-p)\sqrt{CME(\mathbf{x}^T\mathbf{C}\mathbf{x})}$$

Analyse des résidus

Modèle de régression :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

avec $\varepsilon \sim N(0, \sigma)$



Un résidu e_i est considéré comme trop important si

$$|e_i| > 2\hat{\sigma}$$

ou

Résidu standardisé

$$\left|\frac{e_i}{\hat{\sigma}}\right| > 2$$

Analyse des résidus et des observations

Résidus et Observations

•L'ensemble des individus étudiés forme-t-il un ensemble homogène au niveau des variables explicatives, ou bien existe-t-il des points atypiques éloignés des autres?

•Il peut avoir des observations mal reconstituées par le modèle. À partir de quel niveau peut-on considérer une erreur comme trop importante?

•Certaines observations peuvent avoir une influence importante sur la construction du modèle.

•L'estimation des paramètres du modèle réalisée en utilisant toutes les observations peut être assez différente des résultats de l'estimation sans utilisation de ces observations.

•Comment identifier les observations influentes?

Analyse des résidus

Vecteur des résidus standardisés d : $\mathbf{d} = \dfrac{\mathbf{e}}{\sqrt{CME}}$

Vecteur des résidus studentisés r : $r_j = \dfrac{e_j}{\sqrt{CME(1-h_{ii})}}$

h_{jj} est le j[ième] élément diagonal de la matrice **H**

$\mathbf{H} = \left(\mathbf{XCX}^{\mathbf{T}}\right)$ est une matrice symétrique qui vérifie $\mathbf{H}^2 = \mathbf{H}$

Mesure de l'influence des observations :

•une observation a une influence significative sur le modèle si :

•**Le Levier h_{jj} :** $h_{jj} > \dfrac{2p}{n}$

•Ou la **Distance de Cook's :** $D_j = \dfrac{\left(\hat{\mathbf{a}}_{(\mathbf{j})} - \hat{\mathbf{a}}\right)^{\mathbf{T}} \mathbf{X}^{\mathbf{T}} \mathbf{X} \left(\hat{\mathbf{a}}_{(\mathbf{j})} - \hat{\mathbf{a}}\right)}{p\hat{\sigma}^2} = \dfrac{r_j^2 h_{jj}}{p(1-h_{jj})} > 1$

Utilisation d'observations répétés

Exemple: Plan d'expériences factoriels avec répétition

- **n** expériences avec **m** répétitions

Total de réponses observés : $n=n_1+...+n_m$

- La décomposition de la **Somme des Carrés Erreurs** :

$$SCE = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{i,j}-\hat{y}_i\right)^2 = \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{i,j}-\bar{y}_i\right)^2}_{\text{Somme des carrés des erreurs}\ SCPE} + \underbrace{\sum_{i=1}^{m} n_i\left(\bar{y}_i-\hat{y}_i\right)^2}_{\text{Somme des carrés Lack-of-fit}\ SCLOF}$$

- Les **Carrés Moyens des Erreurs** :

$$CMLOF = \frac{SCLOF}{m-p}$$

$$CMPE = \frac{SCPE}{n-m}$$

Teste d'hypothèse : le modèle de régression pourra être rejeté si :

$$1 - F_F\left(\frac{CMLOF}{CMPE}, m-p, n-m\right) \le \alpha$$

Régression pas à pas descendante (Backward)

On part du modèle complet.

A chaque étape on enlève la variable X_j ayant l'apport marginal le plus faible :

$|t_j|$ minimum ou $NS(t_j)$ maximum

à condition que cet apport soit non significatif

($NS(t_j) \geq 0.1$ = valeur par défaut).

Exemple réf.[3]

Données de Wheelwright et Makridakis

- Il s'agit de relier les ventes semestrielles **y** d'un produit à huit variables potentiellement explicatives **X**:
  - X_1 = marché total de la branche (MT);
  - X_2 = remise aux grossistes (RG);
  - X_3 = prix (PRIX);
  - X_4 = budget de recherche (BR);
  - X_5 = Investissements (INV);
  - X_6 = Publicité (PUB);
  - X_7 = Frais de ventes (FV);
  - X_8 = Total du budget publicité de la branche (TPUB).

Variable à expliquer :

- Y = Ventes semestrielles (k€)

Exemple (réf. [3]) Cas Ventes : les données

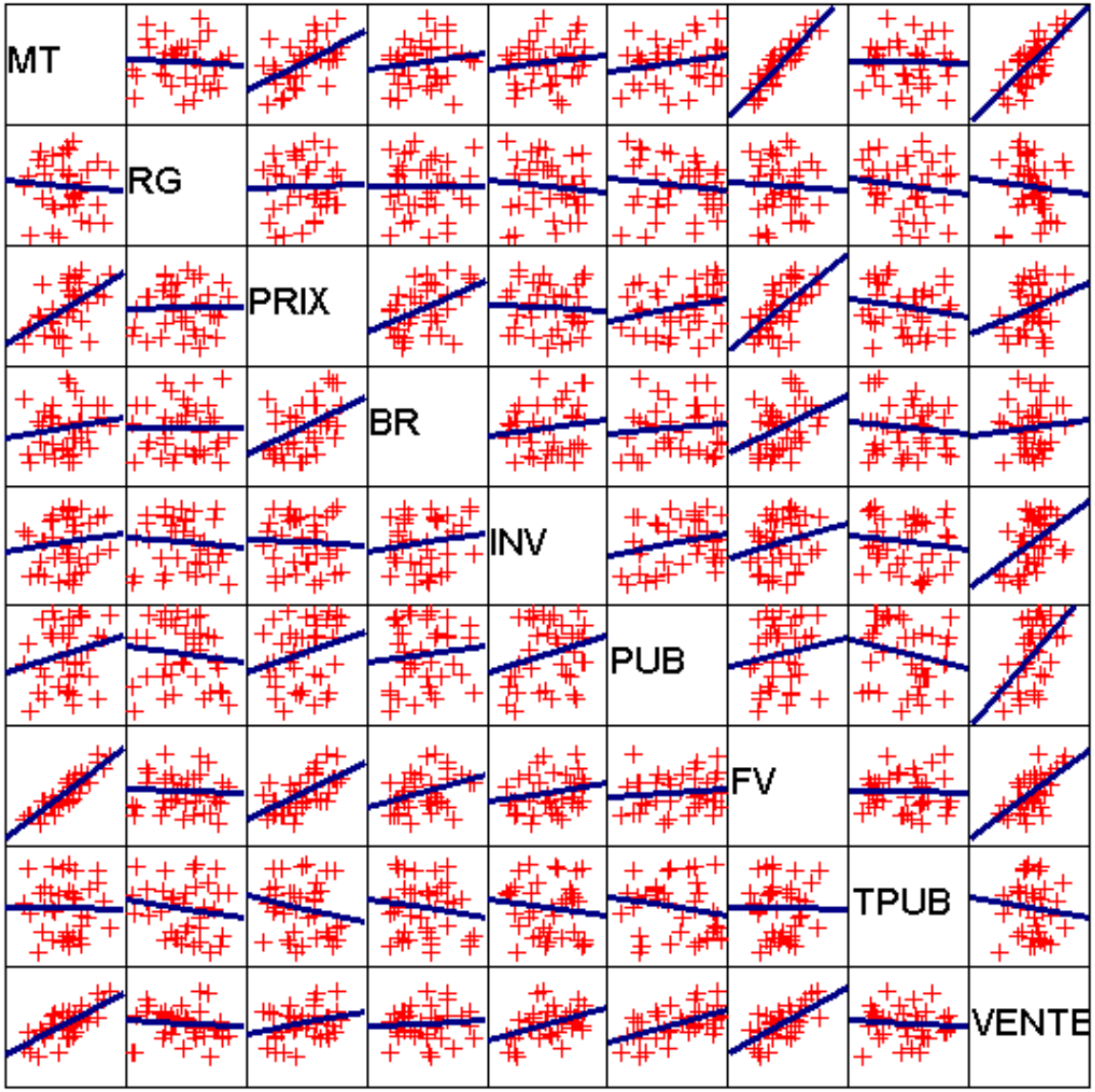
SEMESTRE	X_1 Marché total	X_2 Remises aux grossistes	X_3 Prix	X_4 Budget de recherche	X_5 Investissements	X_6 Publicité	X_7 Frais de ventes	X_8 Total publicité de la branche	Y Ventes
1	398	138	56	12	50	77	229	98	5540
2	369	118	59	9	17	89	177	225	5439
3	268	129	57	29	89	51	166	263	4290
4	484	111	58	13	107	40	258	321	5502
5	394	146	59	13	143	52	209	407	4872
6	332	140	60	11	61	21	180	247	4708
7	336	136	60	25	-30	40	213	328	4627
8	383	104	60	21	-45	32	201	298	4110
9	285	105	63	8	-28	12	176	218	4123
10	277	135	62	11	76	68	175	410	4842
11	456	128	65	22	144	52	253	93	5741
12	355	131	65	24	113	77	208	307	5094
13	364	120	64	14	128	96	195	107	5383
14	320	147	66	15	10	48	154	305	4888
15	311	143	67	22	-25	27	181	60	4033
16	362	145	67	23	117	73	220	239	4942
17	408	131	66	13	120	62	235	141	5313
18	433	124	68	8	122	25	258	291	5140
19	359	106	69	27	71	74	196	414	5397
20	476	138	71	18	4	63	279	206	5149
21	415	148	69	8	47	29	207	80	5151
22	420	136	70	10	8	91	213	429	4989
23	536	111	73	27	128	74	296	273	5927
24	432	152	73	16	-50	16	245	309	4704
25	436	123	73	32	100	43	276	280	5366
26	415	119	75	20	-40	41	211	315	4630
27	462	112	73	15	68	93	283	212	5712
28	429	125	74	11	88	83	218	118	5095
29	517	142	74	27	27	75	307	345	6124
30	328	123	77	20	59	88	211	141	4787
31	418	135	79	35	142	74	270	83	5036
32	515	120	77	23	126	21	328	398	5288
33	412	149	78	36	30	26	258	124	4647
34	455	126	78	22	18	95	233	118	5316
35	554	138	81	20	42	93	324	161	6180
36	441	120	80	16	-22	50	267	405	4801
37	417	120	81	35	148	83	257	111	5512
38	461	132	82	27	-18	91	267	170	5272
39	500	100	83	30	50	90	300	200	.

Problème de prévision des ventes

Prévoir les ventes semestrielles (en K €) y du 39e semestre sous le scénario suivant :

Marché total (MF) = 500
Remises aux grossistes (KF) = 100
Prix (F) = 83
Budget de Recherche (KF) = 30
Investissement (KF) = 50
Publicité (KF) = 90
Frais de ventes (KF) = 300
Total budget publicité de la branche (KF) = 200

Exemple (réf. [3]) Graphiques des liaisons deux à deux



Mesure de la multi-colinéarité : Tolérance et VIF

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3129.231	641.355		4.879	.000		
	MT	4.423	1.588	.605	2.785	.009	.142	7.051
	RG	1.676	3.291	.043	.509	.614	.946	1.057
	PRIX	-13.526	8.305	-.201	-1.629	.114	.439	2.276
	BR	-3.410	6.569	-.054	-.519	.608	.630	1.587
	INV	1.924	.778	.234	2.474	.019	.752	1.330
	PUB	8.547	1.826	.434	4.679	.000	.778	1.285
	FV	1.497	2.771	.130	.540	.593	.116	8.621
	TPUB	-2.15E-02	.401	-.005	-.054	.958	.882	1.133

a. Dependent Variable: VENTES

Prix d'un appartement

Il s'agit d'étudier le prix Y d'un appartement en fonction de sa surface X.

Nous avons relevé quelques annonces d'appartements à vendre dans le Figaro.

Les données sont reproduites dans le tableau 1.

Pour avoir une vision d'ensemble des données, nous avons construit le tableau 2 en associant à chaque appartement son prix, sa surface et son prix au m^2, puis le graphique prix / surface de la figure 1.

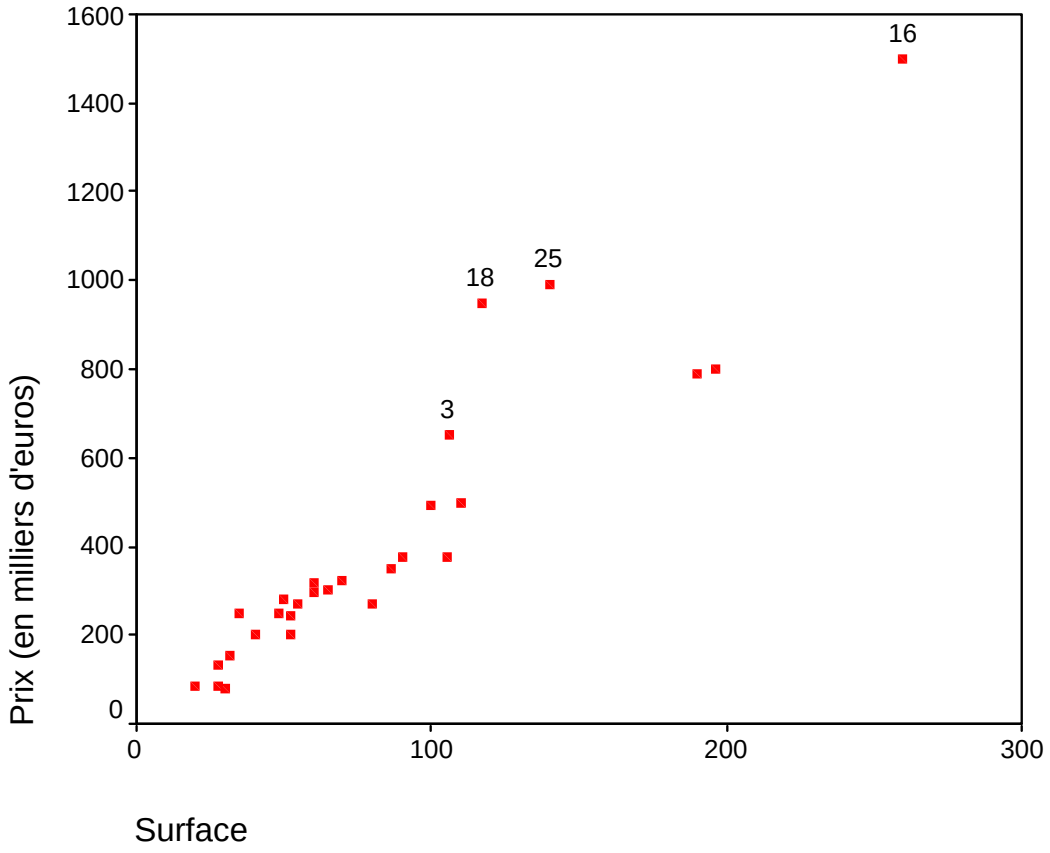
Il peut avoir des observations mal reconstituées par le modèle.

À partir de quel niveau peut-on considérer une erreur comme trop importante?

Prix, surface et prix au m^2 des 28 appartements

Numéro	Localisation	Surface (m^2)	Prix (x 1000€)	Prix au m^2 (en €)
1	censier	28	130	4640
2	contrescarpe	50	280	5600
3	rue saint-simon	106	650	6130
4	rapp	196	800	4080
5	saint-andré des arts	55	268	4870
6	5-ième, près quais	190	790	4160
7	gobelins	110	500	4550
8	gobelins	60	320	5330
9	censier	48	250	5210
10	panthéon	35	250	7140
11	rue madame	86	350	4070
12	rue de seine	65	300	4620
13	panthéon	32	155	4840
14	sèvres-babylone	52	245	4710
15	montparnasse	40	200	5000
16	rue d'assas	260	1500	5770
17	saint-germain	70	325	4640
18	ile saint-louis	117	950	8120
19	jussieu	90	378	4200
20	quartier-latin	30	78	2600
21	montparnasse	105	375	3570
22	rue mazarine	52	200	3850
23	censier	80	270	3380
24	assas luxembourg	60	295	4920
25	jardins de l'observatoire	140	990	7070
26	rue de savoie	20	85	4250
27	près luxembourg	100	495	4950
28	gobelins	28	85	3040

Graphique Prix / Surface des 28 appartements



Exercice 2 Régression multiple

From William E. Becker, Statistics for Business and Economics, S.R.B. Publishing, 1997, p. 509-511.

An article in Newsweek (January 22, 1996) stated that "Compared with real wine-drinking countries, the United States is practically dry. That may be a reason, scientists say, that our rate of heart disease is higher." Dr. Charles Fuchs is also quoted saying that drinking beer versus wine may produce extraordinary differences in life expectancy. Questions: 1) Does wine, beer or liquor consumption provide an explanation of death rates from heart disease? 2) Does wine consumption significantly lower death rates from heart disease? The following data on average country death rates and average country alcoholic beverage consumption, in liters per capita, were provided in the Newsweek article. "Heart Disease" is defined as the death rate per 100,000.

Exercice 2 Régression multiple

	Death Rate from Heart Disease	Wine*	Beer*	Liquor*
France	61.1	63.5	40.1	2.5
Italy	94.1	58.0	25.1	0.9
Switzerland	106.4	46.0	65.0	1.7
Australia	173.0	15.7	102.1	1.2
Britain	199.7	12.2	100.0	1.5
U.S.A.	176.0	8.9	87.8	2.0
Russia	373.6	2.7	17.1	3.8
Czech Republic	283.7	1.7	140.0	1.0
Japan	34.7	1.0	55.0	2.1
Mexico	36.4	0.2	50.4	0.8

*Consumption Per Capita

Semiconductor Manufacturing Plant
Wire bond pull strength data

The table contains data on three variables that were collected in an observational study in a semiconductor manufacturing plant. In this plant, the finished semiconductor is wire bonded to a framer. The variable reported are pull strength (a measure of the amount of force required to break the bond), the wire length, and the height of the die.

We would like to find a model relating pull strength to wire length and die height. Unfortunately, there is no physical mechanism that we can easily apply here, so it doesn't seem likely that a mechanistic modeling approach will be successful.

Semiconductor Manufacturing Plant
Wire bond pull strength data

Obs number	Y: Pull Strength	x1: Wire Length	x2: Die Height
1	9,95	2	50
2	24,45	8	110
3	31,75	11	120
4	35,00	10	550
5	25,02	8	295
6	16,86	4	200
7	14,38	2	375
8	9,60	2	52
9	24,35	9	100
10	27,50	8	300
11	17,08	4	412
12	37,00	11	400
13	41,95	12	500
14	11,66	2	360
15	21,65	4	205
16	17,89	4	400
17	69,00	20	600
18	10,30	1	585
19	34,93	10	540
20	46,59	15	250
21	44,88	15	290
22	54,12	16	510
23	56,63	17	590
24	22,13	6	100
25	21,15	5	400