Modélisation Statistique
Master 1 Psychologie

Année 2017-18

www.univ-nantes.fr

UNIVERSITÉ DE NANTES

Régression linéaire

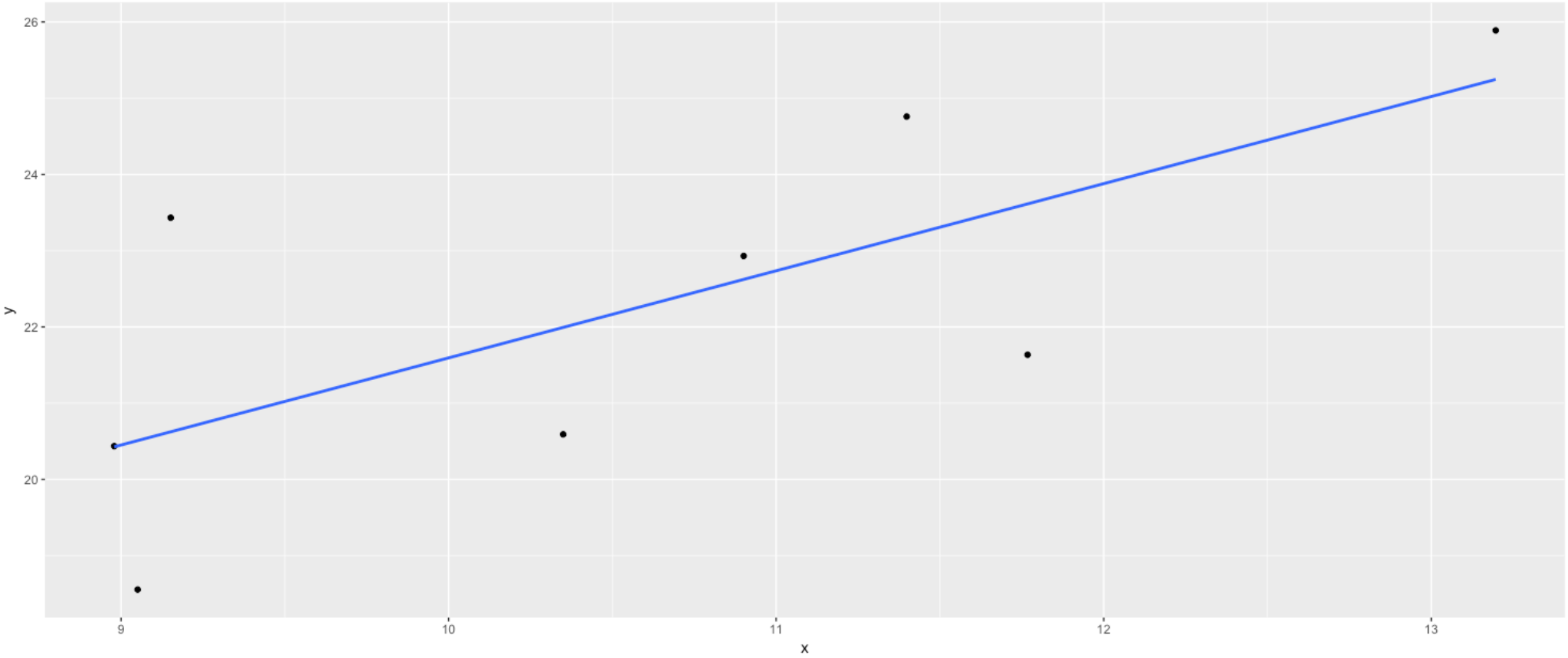
Le modèle de régression linéaire

- Réponse (VD) : 1 variable quantitative continue (notée Y)
- Prédicteurs (VI) : p variables quantitatives continues (notées $X_1, \cdots, X_p$)
- Il s'agit d'une **modélisation multivariée.**
- Le modèle s'écrit pour les i individus de l'échantillon:

$$Y_i = b_0 + b_1 X_{1,i} + \cdots b_p X_{p,i} + \varepsilon_i$$

- On veut estimer les coefficients $(b_0, b_1, \ldots, b_p)$
- On utilise la méthode des moindres carrés ordinaires (OLS : Ordinary Least Square en anglais)

Exemple graphique



Lien corrélation et régression

- Le coefficient de corrélation linéaire est lié au modèle de régression simple (1 seul prédicteur)
- Pour envisager un modèle de régression linéaire il faut :
  - Avoir une corrélation entre la variable réponse (VD) et les prédicteurs (VI)
  - De pas avoir trop de corrélations entre les prédicteurs (partage d'information)
  - Il est impératif de regarder la matrice de corrélation

Exemple

On veut prédire un score DepressT en fonction de scores AnxT, HostT, PhobT, ParT, PsyT.
Première étape : étude de la matrice de corrélation

Correlation Matrix

Pearson Correlations

		DepressT	AnxT	HostT	PhobT	ParT	PsyT
DepressT	Pearson's r	—					
	p-value	—					
AnxT	Pearson's r	0.541	—				
	p-value	< .001	—				
HostT	Pearson's r	0.594	0.495	—			
	p-value	< .001	< .001	—			
PhobT	Pearson's r	0.523	0.538	0.474	—		
	p-value	< .001	< .001	< .001	—		
ParT	Pearson's r	0.674	0.501	0.508	0.542	—	
	p-value	< .001	< .001	< .001	< .001	—	
PsyT	Pearson's r	0.690	0.422	0.500	0.531	0.748	—
	p-value	< .001	< .001	< .001	< .001	< .001	—

Mise en œuvre du modèle

- On va obtenir des estimateurs des coefficients de la régression ainsi que d'autres indicateurs importants pour juger de la qualité du modèle :
  - Table d'Anova du modèle.
  - % de variance expliqué.
  - Test de significativité des coefficients.
- L'analyse du modèle devra toujours se faire selon ces trois étapes

Table d'ANOVA du modèle

- On teste ici l'hypothèse ci-dessous (c'est à dire qu'aucun des prédicteurs n'apporte d'information sur la VD) :

$$H_0 : b_1 = b_2 = \ldots = b_p$$

- Il s'agit d'un test de Fisher comme pour une Anova :

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	6829	5	1365.78	40.63	< .001
	Residual	4337	129	33.62		
	Total	11166	134			

Donc l'un des b au moins est différent de 0 :

$$F(5, 129) = 40.63, p < .001$$

% de variance expliquée, RMSE

- Parmi les indicateurs de qualité du modèle on a :
  - R^2 (généralisation du coefficient de corrélation à plusieurs prédicteurs) il est compris entre 0 et 1, plus il est proche de 1 plus le % de variance expliqué est important
  - R_a^2 appelé R2 ajusté comme le R^2 compris entre 0 et 1, est moins sensible au nombre de prédicteurs.
  - RMSE : Root Mean Square Error
- Ils sont donnés par les formules suivantes :

$$R^2 = \frac{SCE_{Mod}}{SCE_{Total}} \qquad R_a^2 = 1 - \frac{s_{Res}^2}{s_{Total}^2} \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} = s_{Res}$$

Différence entre R^2 et R_a^2 :

- Le R^2 augmente si le nombre de prédicteur augmente (vaut 1 si on a autant de prédicteurs que d'individus dans l'échantillon).
- Le R^2 ne peut être utilisé que pour comparer deux modèles ayant le même nombre de prédicteurs (sinon on utilise le R_a^2).

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.782	0.612	0.597	5.798

Tests de significativité des coefficients

Ce sont des tests de Student, ils testent l'hypothèse :

$$H_0 : b_j = 0$$

(c'est-à-dire le prédicteur j n'apporte pas d'information sur la VD) :

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	8.819	3.869		2.280	0.024
	AnxT	0.157	0.067	0.166	2.354	0.020
	HostT	0.195	0.061	0.222	3.205	0.002
	PhobT	0.036	0.068	0.039	0.536	0.593
	ParT	0.182	0.077	0.208	2.347	0.020
	PsyT	0.298	0.077	0.333	3.875	< .001

Interprétation des coefficients

- On définit deux types de coefficients :
  - Les b_j qui sont les coefficients estimés
  - Les β_j qui sont les coefficients standardisés : on a centré et réduit la VD et l'ensemble des prédicteurs et on a estimé le modèle avec ces nouvelles variables.
- La significativité des coefficients est la même (du fait de la linéarité de la relation). De fait on aura toujours $\beta_0 = 0$.
- Par exemple on peut conclure du tableau précédent : β_1 est l'effet marginal, compte tenu des variables HostT, PhobT, ParT et PsyT, de la variable AnxT sur la DepressT : autrement dit pour des scores HostT, PhobT, ParT et PsyT constants, une augmentation de 1 point du score AnxT induira une augmentation de β_1 fois son écart type (du fait de la standardisation) du score de DepressT.

Exemple d'interprétation

- Interprétation du coefficient de AnxT : tous les autres prédicteurs étant constants (« **toutes choses par ailleurs égales** ») une augmentation de 1 point de score de AnxT induit une augmentation du score DepressT de 1.6 points:

$$\begin{aligned} \beta_1 \times s_{AnxT} &= .166 \times 9.652 \\ &= 1.60 \end{aligned}$$

Corrélation partielle

- On considère deux modèles :

$$\begin{aligned} Mod_1 : Y &\sim X_1 + X_2 + \ldots + X_p \\ Mod_0 : Y &\sim X_2 + \ldots + X_p \end{aligned}$$

- Corrélation partielle de Y et X_1 : on veut connaitre la corrélation de X_1 avec Y en contrôlant les effets des autres prédicteurs. On définit par son carré :

$$R^2_{(X_1,Y).X_2,\ldots,X_p} = \frac{SCE_{Mod_1} - SCE_{Mod_0}}{SCE_{Res_0}}$$

Corrélation semi-partielle

- On définit son carré par :

$$R^{2,p}_{(X_1,Y).X_2,\ldots,X_p} = \frac{SCE_{Mod_1} - SCE_{Mod_0}}{SCE_{Total}}$$

- Il s'agit donc d'une généralisation du coefficient R^2 au seul prédicteur X_1 en contrôlant l'effet des autres.

Exemple

- On regarde pour le prédicteur AnxT :

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	6643	4	1660.67	47.73	< .001
	Residual	4523	130	34.79		
	Total	11166	134			
1	Regression	6829	5	1365.78	40.63	< .001
	Residual	4337	129	33.62		
	Total	11166	134			

Note. Null model includes HostT, PhobT, ParT, PsyT

Part And Partial Correlations

Model		Partial	Part
1	AnxT	0.203	0.129
	HostT	0.272	0.176
	PhobT	0.047	0.029
	ParT	0.202	0.129
	PsyT	0.323	0.213

Sélection de variables (I)

- La sélection de variables est l'une des problématique liée aux modèles linéaires. Il existe plusieurs types de procédures :
  - Procédures automatiques (déconseillées) : stepwise, forward, backward.
  - Procédures hiérarchisées : on introduit des groupes de variables et on teste la significativité de l'apport du nouveau groupe de variable.

Remarque : Une variable incluse dans un modèle présente un intérêt pour le chercheur donc la retirer de celui-ci doit être « réfléchi ».

Sélection automatique

- Il faut avoir un critère de sélection : ce sera la variation du F qui sera celui utilisé (il en existe d'autres qui conduisent à des modèles plus ou moins parcimonieux)
- Exemple Sélection forward :

Model Summary

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2
1	0.810	0.656	0.653	5.376	0.656	253.365	1	133
2	0.820	0.673	0.668	5.259	0.017	6.965	1	132

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	13.193	3.072		4.295	< .001
	GSIT	0.774	0.049	0.810	15.917	< .001
2	(Intercept)	14.836	3.069		4.834	< .001
	GSIT	0.886	0.064	0.927	13.889	< .001
	SomT	−0.160	0.061	−0.176	−2.639	0.009

Régression hiérarchisée (HR)

$$Mod_0 : DepressT \sim AnxT + HostT + PhobT$$
$$Mod_1 : DepressT \sim AnxT + HostT + PhobT + ParT + PsyT$$

Model Summary

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p
0	0.682	0.465	0.453	6.754	0.465	37.93	3	131	< .001
1	0.782	0.612	0.597	5.798	0.147	24.37	2	129	< .001

Note. Null model includes AnxT, HostT, PhobT

$$F_{change} = \frac{n-k-j}{j} \times \frac{SCE_{Mod_1} - SCE_{Mod0}}{SCE_{Res_1}}$$

Suite HR

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	5190	3	1730.10	37.93	< .001
	Residual	5975	131	45.61		
	Total	11166	134			
1	Regression	6829	5	1365.78	40.63	< .001
	Residual	4337	129	33.62		
	Total	11166	134			

Note. Null model includes AnxT, HostT, PhobT

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
0	(Intercept)	17.113	4.258		4.019	< .001
	AnxT	0.227	0.076	0.240	2.995	0.003
	HostT	0.328	0.067	0.373	4.863	< .001
	PhobT	0.203	0.074	0.217	2.740	0.007
1	(Intercept)	8.819	3.869		2.280	0.024
	AnxT	0.157	0.067	0.166	2.354	0.020
	HostT	0.195	0.061	0.222	3.205	0.002
	PhobT	0.036	0.068	0.039	0.536	0.593
	ParT	0.182	0.077	0.208	2.347	0.020
	PsyT	0.298	0.077	0.333	3.875	< .001

Problème de multi-colinéarité

- Il peut arriver qu'un prédicteur « résume » l'information portée par les autres il va dans ce cas annuler les effets de ceux-ci : il s'agit de problème de multi-colinarité.
- On ajoute le prédicteur GSIT dans le modèle à 5 prédicteurs :

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	12.240	3.580		3.419	< .001
	AnxT	−0.099	0.078	−0.105	−1.274	0.205
	HostT	0.073	0.060	0.083	1.209	0.229
	PhobT	−0.045	0.064	−0.048	−0.710	0.479
	ParT	0.020	0.077	0.023	0.265	0.791
	PsyT	0.131	0.077	0.147	1.713	0.089
	GSIT	0.704	0.134	0.736	5.267	< .001

VIF (variance inflation factor)

- Il permet de détecter les problèmes de multi-colinéarité, on doit avoir VIF<5 :

Collinearity Statistics	
Tolerance	VIF
0.370	2.701
0.533	1.877
0.539	1.856
0.324	3.091
0.339	2.953
0.128	7.840

Validation du modèle

- De nombreuses conditions doivent être vérifiées une fois le modèle établi :
  - Beaucoup de ces conditions sont liées aux résidus (homogénéité de leur variance, normalité, indépendance, …)
- On ne s'intéressera qu':
  - à la présence de points influents et ou aberrants : ils peuvent en l'occurrence avoir un impact sur l'estimation des coefficients.
  - à la matrice de variance covariance des résidus
  - à la normalité des résidus

Points aberrants et leviers

Casewise Diagnostics

Case Number	Std. Residual	DepressT	Predicted Value	Residual
9	2.609	70.00	55.26	14.74
12	2.685	80.00	64.75	15.25
38	−2.015	54.00	65.37	−11.37
47	−2.274	44.00	56.29	−12.29
48	−2.600	49.00	63.93	−14.93
64	−2.171	49.00	61.30	−12.30
68	−2.874	42.00	58.32	−16.32
108	2.224	70.00	57.68	12.32
128	−2.215	49.00	61.66	−12.66
135	2.789	76.00	60.72	15.28

Modèles d'ANOVA

Le modèle d'ANOVA

- Réponse (VD) : 1 variable quantitative continue (notée Y)
- Prédicteurs (VI) : p variables qualitatives (notées $F_1, \cdots, F_p$)
- Il s'agit d'une **modélisation multivariée.**
- Le modèle est plus compliqué à écrire mais on peut l'écrire comme un modèle de régression.

Anova 1 facteur

- On va faire un modèle Y~F (3 modalités : 1,2,3)
- On peut faire une Anova (L3) :

ANOVA – Y

Cases	Sum of Squares	df	Mean Square	F	p
F	61.535	2	30.768	363.1	< .001
Residual	6.525	77	0.085		

Note. Type III Sum of Squares

Ce qu'il ne faut pas faire

- Les modalités de F sont codées 1,2,3 on ne peut pas faire directement un modèle de régression :

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	30.46	1	30.462	63.19	< .001
	Residual	37.60	78	0.482		
	Total	68.06	79			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	4.350	0.211		20.603	< .001
	F	−0.781	0.098	−0.669	−7.950	< .001

Régression sur Indicatrices

- Il va falloir créer deux nouvelles variables (on en crée J-1 où J est le nombre de modalités de F). Ces nouvelles variables appelées indicatrices ne comporteront que des 0 et des 1, ici pour la première lorsque F=2 on aura des 1 et des 0 sinon, pour la deuxième lorsque F=3 on aura des 1 et des 0 sinon.

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	61.535	2	30.768	363.1	< .001
	Residual	6.525	77	0.085		
	Total	68.061	79			

On retrouve le même SCE que dans l'ANOVA

Interprétation des coefficients

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	4.053	0.058		69.61	< .001
	F=2	−2.068	0.079	−1.085	−26.23	< .001
	F=3	−1.561	0.082	−0.784	−18.96	< .001

Descriptives – Y

F	Mean	SD	N
1	4.053	0.450	25
2	1.985	0.114	30
3	2.492	0.232	25

$$Y = \mu + \alpha_j$$

On estime donc 3 coefficients pour un facteur à 3 modalités

Contrastes Orthogonaux

- On pose $C_1 = 2(F = 3) - (F = 1) - (F = 2)$ et $C_2 = (F = 2) - (F = 1)$ on obtient le test de contraste en modélisant

$$x \sim C_1 + C_2$$

Contrasts

Difference Contrast – F

Comparison	Estimate	Std. Error	t	p
2 – 1	−2.068	0.079	−26.231	< .001
3 – 1, 2	−0.527	0.070	−7.497	< .001

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	2.843	0.033		87.031	< .001
	C2	1.034	0.039	0.927	26.231	< .001
	C1	−0.176	0.023	−0.265	−7.497	< .001

ANOVA 2 facteurs

- On considère deux facteurs F (2 modalités) et G (3 modalités).

$$Y = \mu + \alpha_j + \beta_k + \gamma_{jk}$$

ANOVA – Y

Cases	Sum of Squares	df	Mean Square	F	p
F	59.050	2	29.525	351.407	< .001
G	0.070	1	0.070	0.838	0.363
F * G	0.239	2	0.119	1.422	0.248
Residual	6.217	74	0.084		

Note. Type III Sum of Squares

Ligne 1 teste $H_0 : \alpha_j = 0$ pour tout j

Ligne 2 teste $H_0 : \beta_k = 0$ pour tout k

Ligne 3 teste $H_0 : \gamma_{jk} = 0$ pour tout j, k

Ecriture sous la forme d'un modèle de régression

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	61.843	5	12.369	147.2	< .001
	Residual	6.217	74	0.084		
	Total	68.061	79			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	3.940	0.084		47.085	< .001
	F=2	−1.963	0.124	−1.030	−15.815	< .001
	F=3	−1.428	0.114	−0.718	−12.522	< .001
	G=2	0.217	0.116	0.117	1.870	0.065
	F=2&G=2	−0.205	0.161	−0.096	−1.272	0.207
	F=3&G=2	−0.263	0.165	−0.098	−1.599	0.114

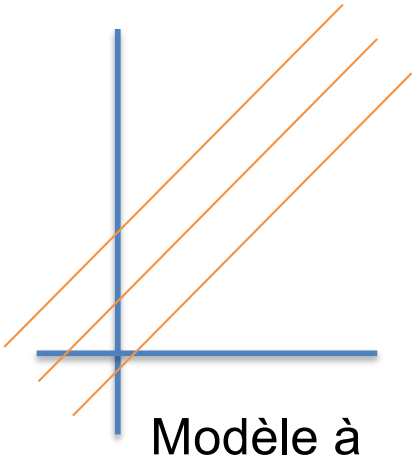
Descriptives - Y

F	G	Mean	SD	N
1	1	3.940	0.551	12
	2	4.157	0.320	13
2	1	1.977	0.142	10
	2	1.989	0.101	20
3	1	2.512	0.241	14
	2	2.466	0.228	11

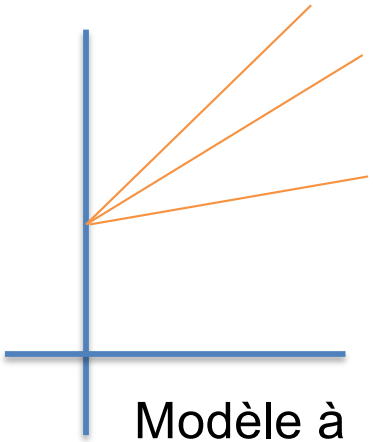
ANCOVA

Le modèle d'ANCOVA

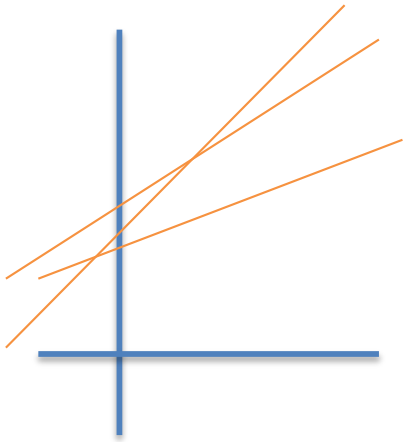
- Réponse (VD) : 1 variable quantitative
- Prédicteurs : variables quantitatives et qualitatives
- On va s'intéresser aux modèles avec un prédicteur X continu et un facteur A à j modalités. Un modèle d'ANCOVA sera alors de l'un des types suivants :



Modèle à pentes constantes



Modèle à intercepts constants



Modèle général

Exemple 1 : sans interaction

- On veut faire un modèle PVLoss~DepressT+Gender.
- Dans ce cas on parle de modèle sans interaction (première équation page précédente). On dit aussi à pentes constantes.

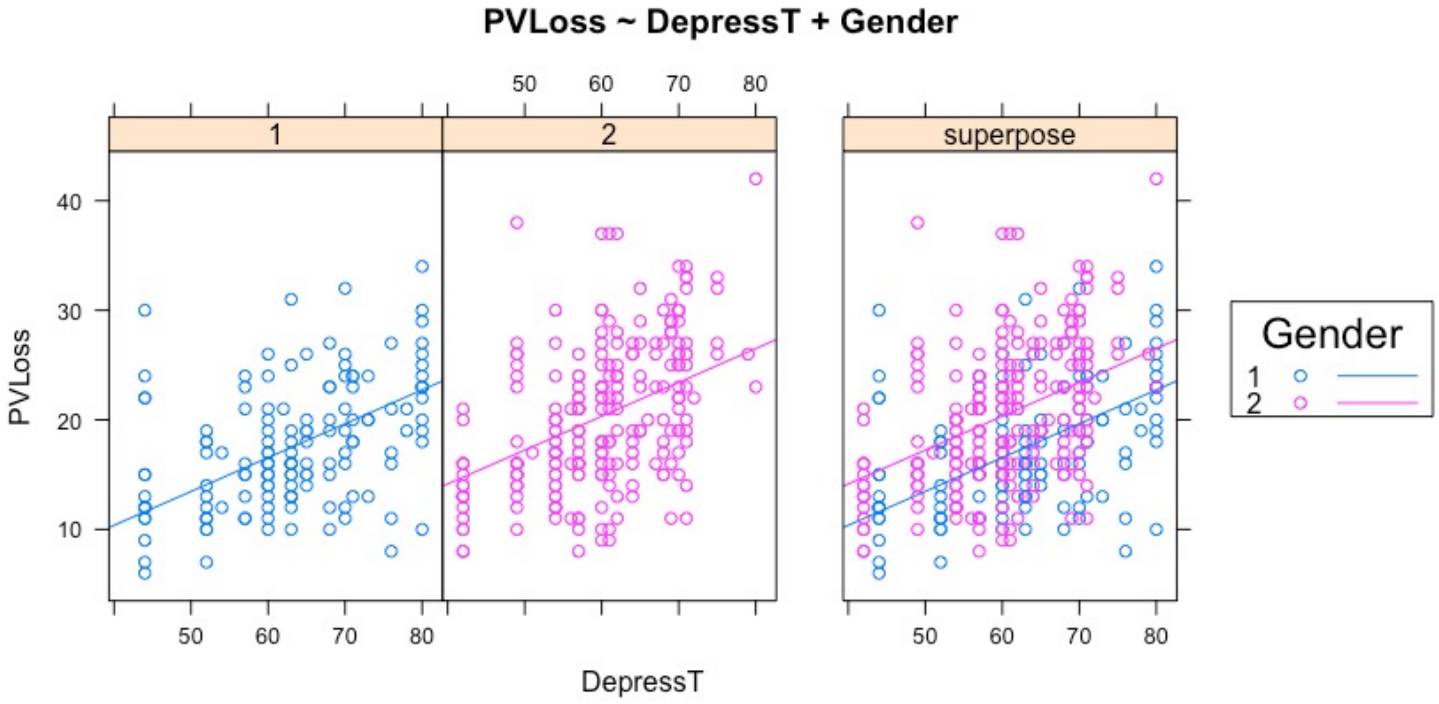


Table du modèle

- On a comme dans le cas des modèles de régression et d'ANOVA une table de Fisher du modèle :

ANCOVA - PVLoss

Cases	Sum of Squares	df	Mean Square	F	p
Gender	1211	1	1210.75	39.48	< .001
DepressT	3128	1	3127.61	101.99	< .001
Residual	11408	372	30.67		

Note. Type III Sum of Squares

Estimations des coefficients

- On a de plus le tableau des coefficients :

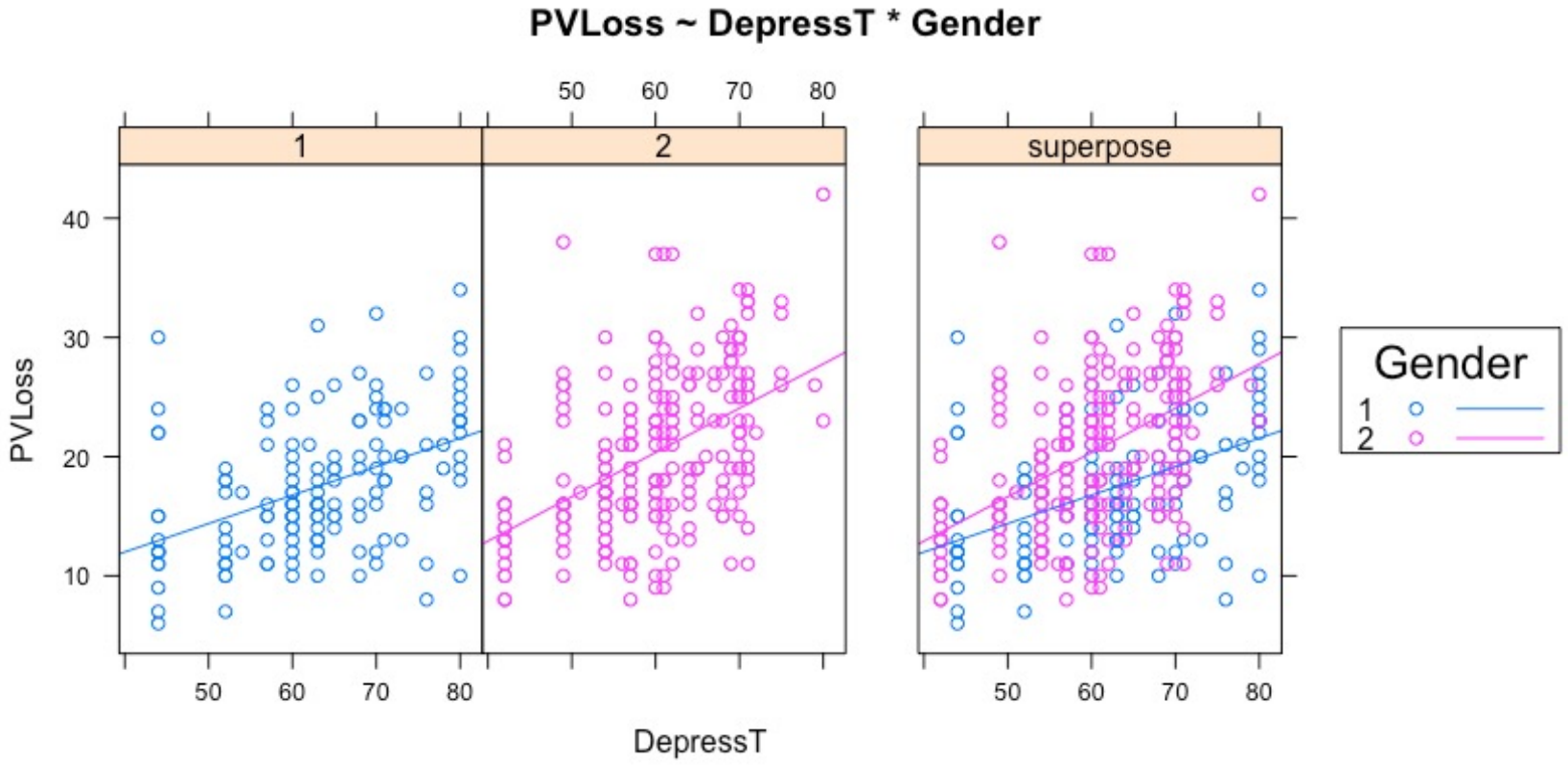
	b	SE(b)	t	p
(Intercept)	-1.9865	1.9862	-1.000	0.318
DepressT	0.3090	0.0306	10.099	<.001
Gender2	3.7728	0.6004	6.283	<.001

$$PVLoss = -1.9865 + .3090 DepressT + 3.7728 \mathbb{I}_{Gender=2}$$

- On a donc deux équations de régression :
- Pour Gender=1 : $PVLoss = -1.9865 + .3090 DepressT$
- Pour Gender=2 : $PVLoss = 1.7863 + .3090 DepressT$

Exemple 2 : avec interaction

- Dans ce modèle on suppose que l'effet de la variable DepressT sur PVLoss est différente selon Gender :



UNIVERSITÉ DE NANTES

Table d'ANOVA

- L'ajout de l'interaction est significative donc on a bien un effet différent entre les filles et les garçons de DepressT sur PVLoss

ANCOVA - PVLoss

Cases	Sum of Squares	df	Mean Square	F	p
DepressT	3038.41	1	3038.41	100.073	< .001
Gender	40.47	1	40.47	1.333	0.249
DepressT ✻ Gender	143.66	1	143.66	4.732	0.030
Residual	11264.22	371	30.36		

Note. Type III Sum of Squares

Table des Coefficients

	b	SE(b)	t	p
(Intercept)	2.44515	2.83842	0.861	0.3895
DepressT	0.23875	0.04438	5.379	<.001
Gender=2	-4.38350	3.79697	-1.154	0.2490
DepressT:Gender=2	0.13268	0.06100	2.175	0.0302

- Pour Gender=1 : $PVLoss = 2.445 + .2388DepressT$
- Pour Gender=2 : $PVLoss = 2.445 + .3714DepressT$
- On en déduit que pour les filles la dépression a un impact plus fort sur le sentiment de vulnérabilité que les garçons.

Retour sur la méthode

- On teste toujours en premier le modèle le plus général (c'est-à-dire avec interaction) et en fonction des résultats on teste un modèle à pentes (***slope***) ou à ordonnées à l'origine (***intercept***) constants.
- On peut aussi dès le départ choisir un modèle et le tester selon l'hypothèse de recherche que l'on veut tester

Régression logistique binaire

Le modèle binaire

- La variable réponse Y est de type binaire (modalités 0 et 1). La modalité 1 représentera le groupe expérimental.
- On dispose de p prédicateurs de types quantitatifs ou qualitatifs.
- On veut modéliser la probabilité que Y=1 sachant une valeur des prédicteurs $X = (X_1, \cdots, X_p)$.
- On cherche aussi une relation linéaire mais comme il s'agit d'une probabilité (nombre entre 0 et 1) il est nécessaire d'utiliser une fonction dite fonction de lien dans la modélisation. Soit $p(x) = P(Y = 1|X = x)$
- La fonction usuelle de lien pour le modèle logistique binaire est le logit .

Estimation des coefficients du modèle

- En utilisant un lien logit on cherche à estimer les coefficients

$$logit\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

On ne peut pas dans ce cas utiliser les moindres carrés. Ce sont des méthodes numériques qui sont utilisées.

Exemple smartphone

- Etude marketing de 2014
- Variable réponse (achat) : êtes-vous prêt à acheter le smartphone ?
- Prédicteurs : âge, sexe, smart (possède déjà un smartphone).

Test global du modèle

- Test de déviance : il confronte deux modèles le modèle H0 qui n'a que la constante contre le modèle H1 qui est le modèle ayant toutes les prédicteurs. Ici on teste le modèle H1 avec uniquement l'âge et on obtient

$$\chi^2(1) = 215.008, p < .001$$

- Ce test remplace la table d'ANOVA du modèle linéaire.
- On peut aussi utiliser d'autres critères : AIC ou BIC.

Model summary

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2
H_0	1630	1632.064	1637.154	1199					
H_1	1415	1419.056	1429.236	1198	215.008	< .001	0.132	0.164	0.243

Coefficients du modèle : Odds-Ratios

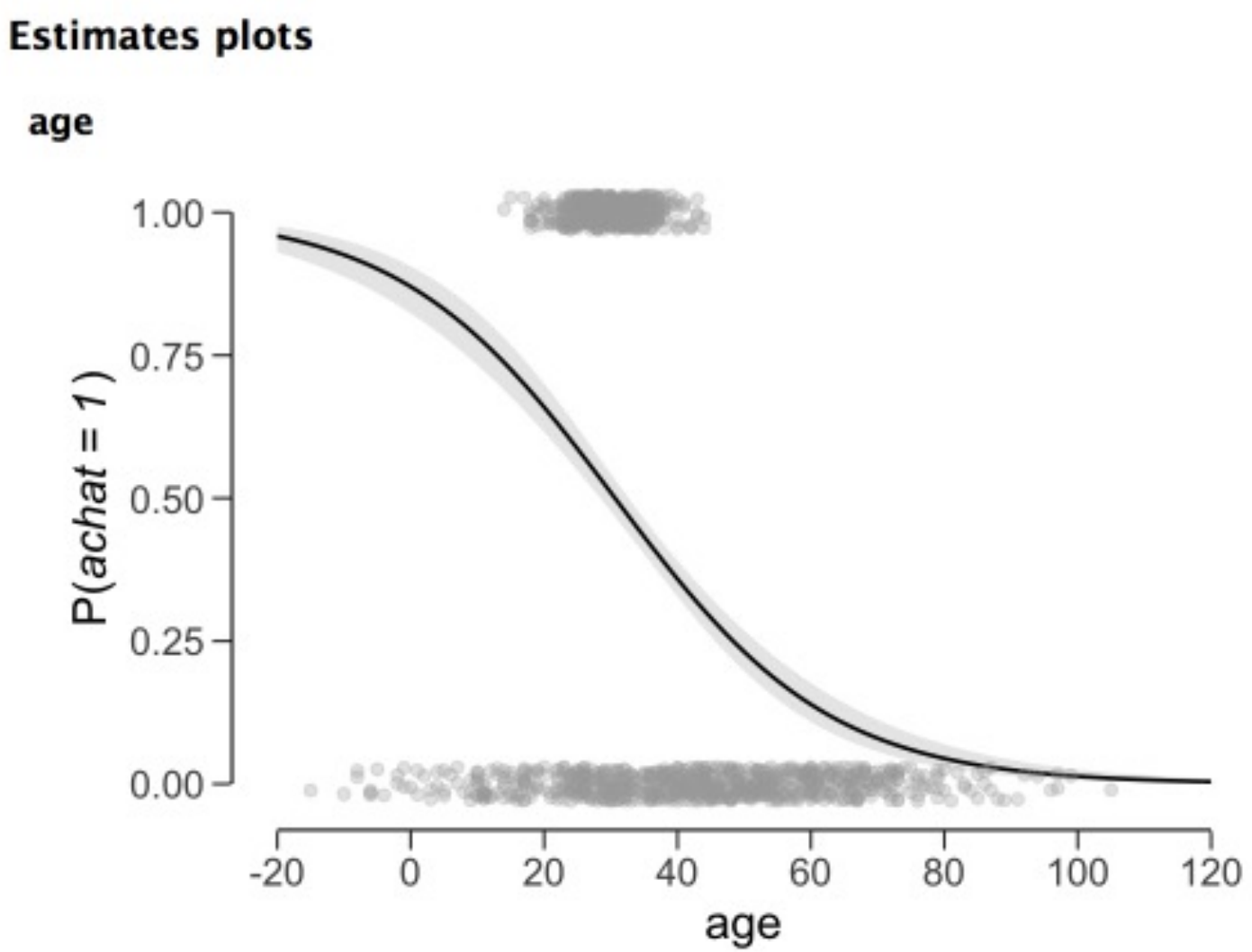
- Attention l'interprétation des coefficients est différente du modèle de régression linéaire. Ici le coefficient de l'âge est négatif donc plus on est jeune plus on a une probabilité importante de déclarer vouloir acheter un nouveau smartphone.
- L'odds-ratio va permettre de préciser cette relation une personne 1 an plus âgée aura 6% de moins de probabilité de vouloir acheter un nouveau smartphone (OR=.94)

Coefficients

						95% Confidence interval	
	Estimate	Standard Error	Odds Ratio	z	p	Lower bound	Upper bound
(Intercept)	1.902	0.185	6.696	10.27	< .001	1.539	2.264
age	−0.062	0.005	0.940	−12.27	< .001	−0.072	−0.052

Note. achat level '1' coded as class 1.

Modèle final



Matrice de confusion

- Une des qualités intrinsèque du modèle sera sa capacité à bien classer les individus (matrice de confusion) :

Confusion matrix

	Predicted	
Observed	0	1
0	514	186
1	228	272

On peut en déduire que 65.5% des individus sont bien classés (modèle peu performant).

- On définit le taux de vrais positifs (sensibilité) 514/(514+186)=.734
- On définit le taux de vrais négatifs (spécificité) : 272/(228+272)=.544

Augmentation du modèle

- On ajoute les variables sexe et smart. La première question est de savoir si ce nouveau modèle apporte plus d'information que le précédent

Model summary

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2
H_0	1415	1419.056	1429.236	1198					
H_1	1093	1101.149	1121.510	1196	321.907	< .001	0.227	0.340	0.447

Note. Null model contains nuisance parameters: age

- On peut estimer les paramètres de ce nouveau modèle :

Coefficients ▼

	Estimate	Standard Error	Odds Ratio	z	p
(Intercept)	0.456	0.223	1.577	2.047	0.041
age	−0.065	0.006	0.937	−11.633	< .001
smartp (1)	2.553	0.160	12.846	15.930	< .001
sexe (1)	0.135	0.151	1.145	0.893	0.372

Note. achat level '1' coded as class 1.

Interprétation du coefficient variable qualitative

- La variable smart est significative avec un OR=12.8, ce qui signifie qu'un individu possédant déjà un smartphone aura 12.8 fois plus de chance de répondre qu'il veut en acheter un nouveau qu'un individu n'en possédant pas.
- On peut ajouter un terme d'interaction sexe*smart.
- De nouveau la première étape consiste à regarder si l'ajout de cette variable améliore le modèle :

Model summary

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2
H_0	1093.1	1101.149	1121.510	1196					
H_1	978.7	988.684	1014.134	1195	114.465	< .001	0.105	0.152	0.516

Note. Null model contains nuisance parameters: age, smartp (1), sexe (1)

Suite du modèle

Coefficients

	Estimate	Standard Error	Odds Ratio	z	p
(Intercept)	1.316	0.242	3.727	5.445	< .001
age	−0.067	0.006	0.935	−10.995	< .001
smartp (1)	1.224	0.193	3.401	6.352	< .001
sexe (1)	−2.340	0.345	0.096	−6.782	< .001
smartp (1) * sexe (1)	3.761	0.404	43.008	9.306	< .001

Note. achat level '1' coded as class 1.