

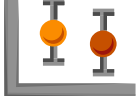
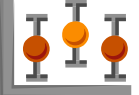
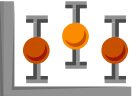
TD7

Régression linéaire multiple

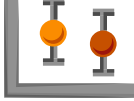
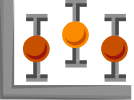
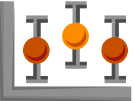
Lotje van der Linden

l.vanderlinden@cogsci.nl

Emploi du temps

Séance	Date	Thème	
1	04/02/2016	Introduction Les statistiques descriptives	
2	11/02/2016	Le test t : Comparer 2 moyennes	
3	25/02/2016	L'ANOVA simple : Comparer >2 moyennes d'une seule VI	
4	03/03/2016	L'ANOVA factorielle : Comparer >2 moyennes de plusieurs VI	
5	10/03/2016	Continuation ANOVA Preparer le fichier csv	
6	17/03/2016	La régression simple et corrélation 1 prédicteur	
7	24/03/2016	La régression multiple Plusieurs prédicteurs	
8	31/03/2016	(Les tests non-paramétriques) Révisions	
9	14/04/2016	Examen sur table	

Emploi du temps

Séance	Date	Thème	
1	04/02/2016	Introduction Les statistiques descriptives	
2	11/02/2016	Le test t : Comparer 2 moyennes	
3	25/02/2016	L'ANOVA simple : Comparer >2 moyennes d'une seule VI	
4	03/03/2016	L'ANOVA factorielle : Comparer >2 moyennes de plusieurs VI	
5	10/03/2016	Continuation ANOVA Preparer le fichier csv	
6	17/03/2016	La régression simple et corrélation 1 prédicteur	
7	24/03/2016	La régression multiple Plusieurs prédicteurs	
8	31/03/2016	(Les tests non-paramétriques) Révisions	
9	14/04/2016	Examen sur table	

Planning

- Aujourd'hui

Partie 1 : • Théorie : • La régression linéaire multiple
Pause
Partie 2 : • Informatique : • Exercices
TER

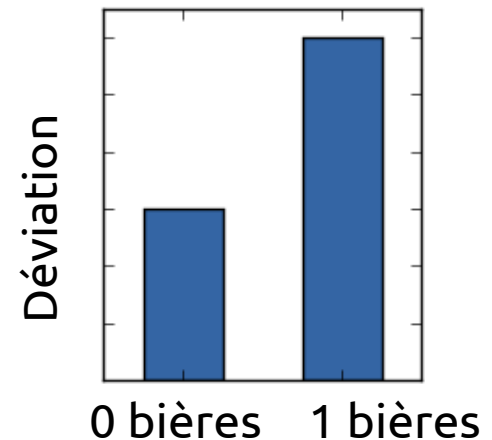
La régression linéaire simple

1 variable explicative

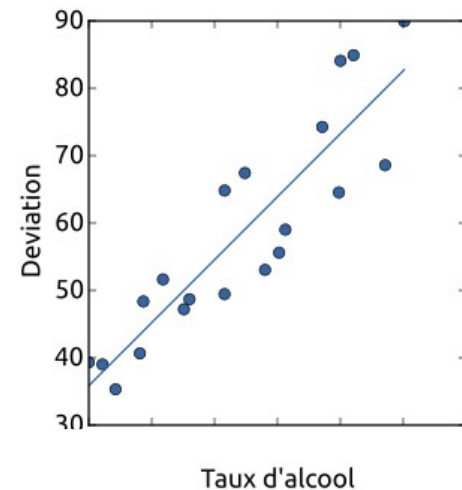
La question de recherche

- Deux types de questions de recherche :
 - **Différences :**
 - Comparer deux ou plusieurs moyennes entre elles
 - La VI a plusieurs modalités
 - **Relations :**
 - Étudier les rapports / les relations / les liens entre plusieurs variables
 - Les variables sont continues

Le groupe qui a bu 2 bières montre plus de déviation que le groupe qui n'a pas bu de bière.



Le taux de l'alcool, est-il corrélé avec les compétences de conduit? Si ou, dans quel sens ?

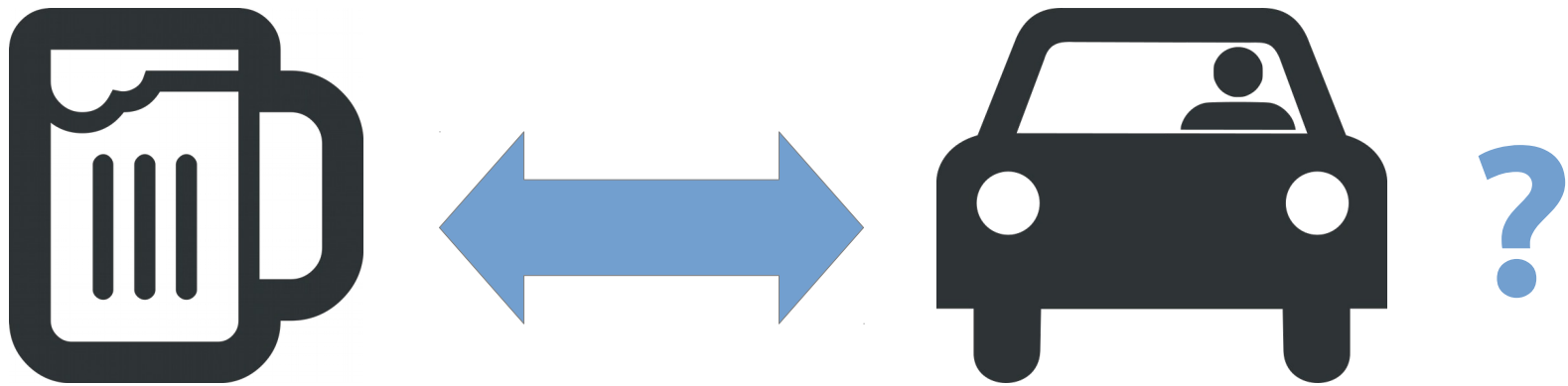


La relation entre deux variables

- Corrélation
 - Mesure la force de l'association entre deux variables
- La régression linéaire simple
 - Prédit une variable expliquée sur la base d'une variable explicative
 - Retourne une équation de régression qui décrit la relation entre la variable expliquée et la variable explicative

La relation entre deux variables

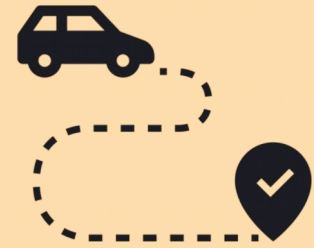
- La problématique :
 - Existe-t-il une relation entre le taux d'alcool et la conduit de voiture ?



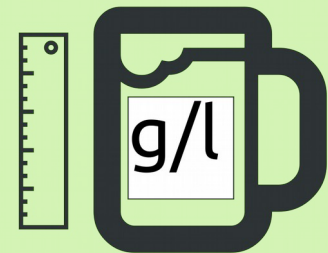
La relation entre deux variables

- Les variables :
 - On recueille deux variables continues au même moment
 - Une variable **expliquée**
 - Une variable **explicative**
 - Parfois arbitraire
- Question de recherche :
 - Les deux variables, sont-elles corrélées ?

Variable 1
Déviation de la conduite
sur un parcours (en cm.)

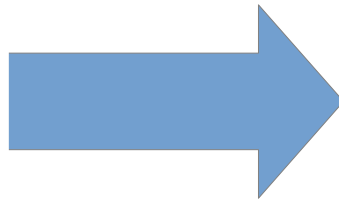
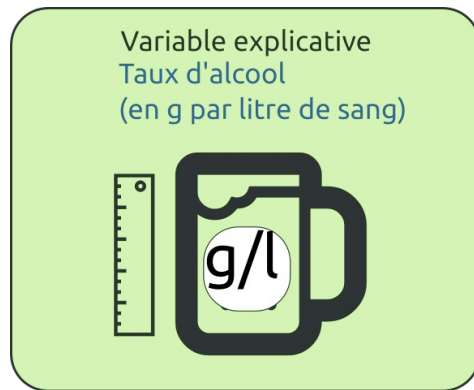


Variable 2
Taux d'alcool
(en g par litre de sang)



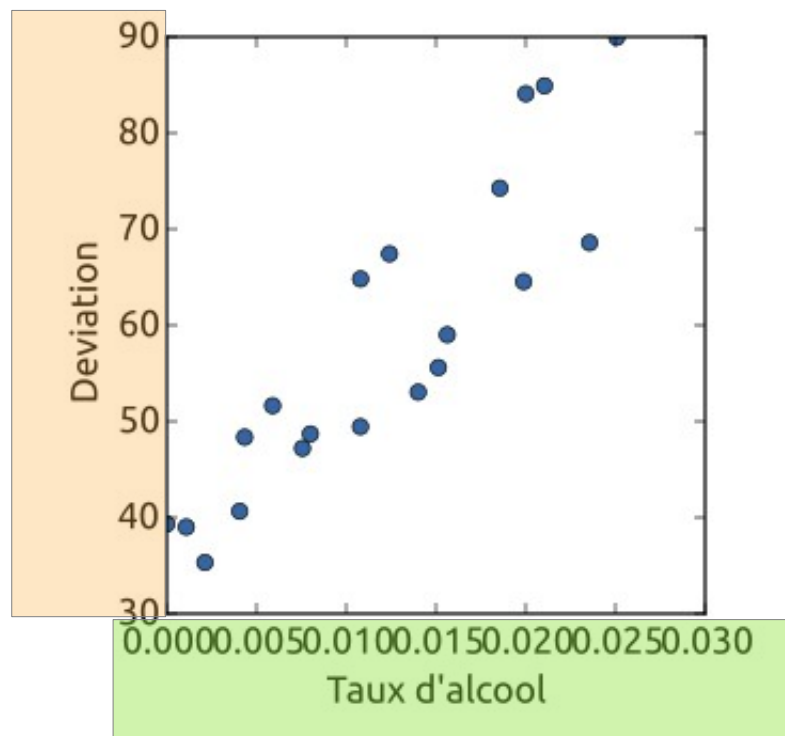
La relation entre deux variables

- La problématique :
 - Pourrait-on prédire la déviation sur un test de conduit à partir du taux d'alcool ?
 - (Si variable explicative est invoquée → pas de causalité)



La relation entre deux variables

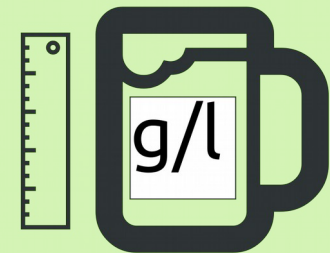
- La représentation des données :
 - Nuage de points → 1 point par sujet
 - **Abcisse** : la variable explicative
 - **Ordonnée** : la variable expliquée



Variable 1
Déviation de la conduite
sur un parcours (en cm.)

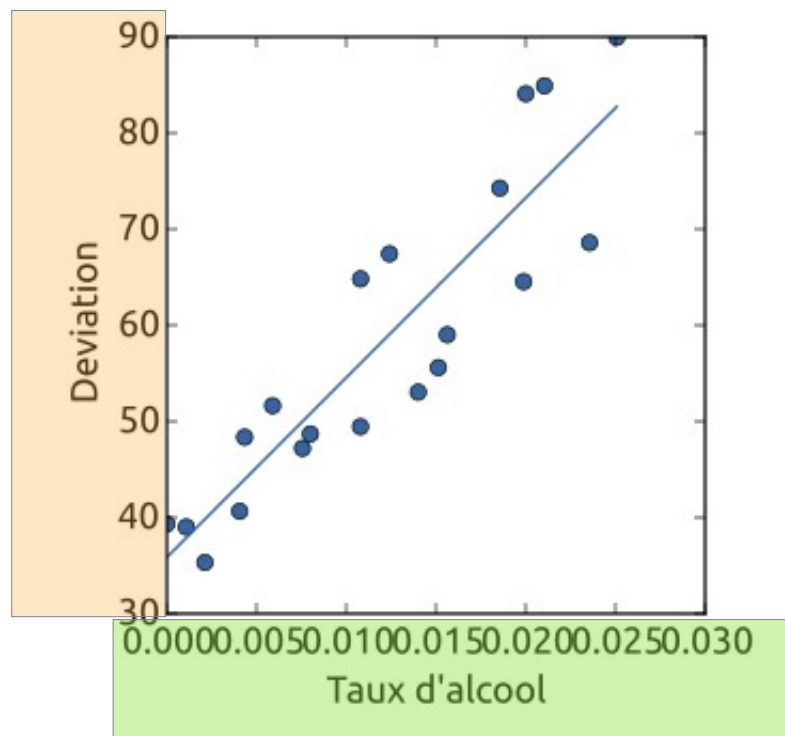


Variable 2
Taux d'alcool
(en g par litre de sang)



La relation entre deux variables

- La régression linéaire simple :
 - Peut-on prédire les valeurs sur **l'ordonnée**
 - à partir des valeurs sur **l'abscisse** ?



Variable 1
Déviation de la conduite
sur un parcours (en cm.)



Variable 2
Taux d'alcool
(en g par litre de sang)



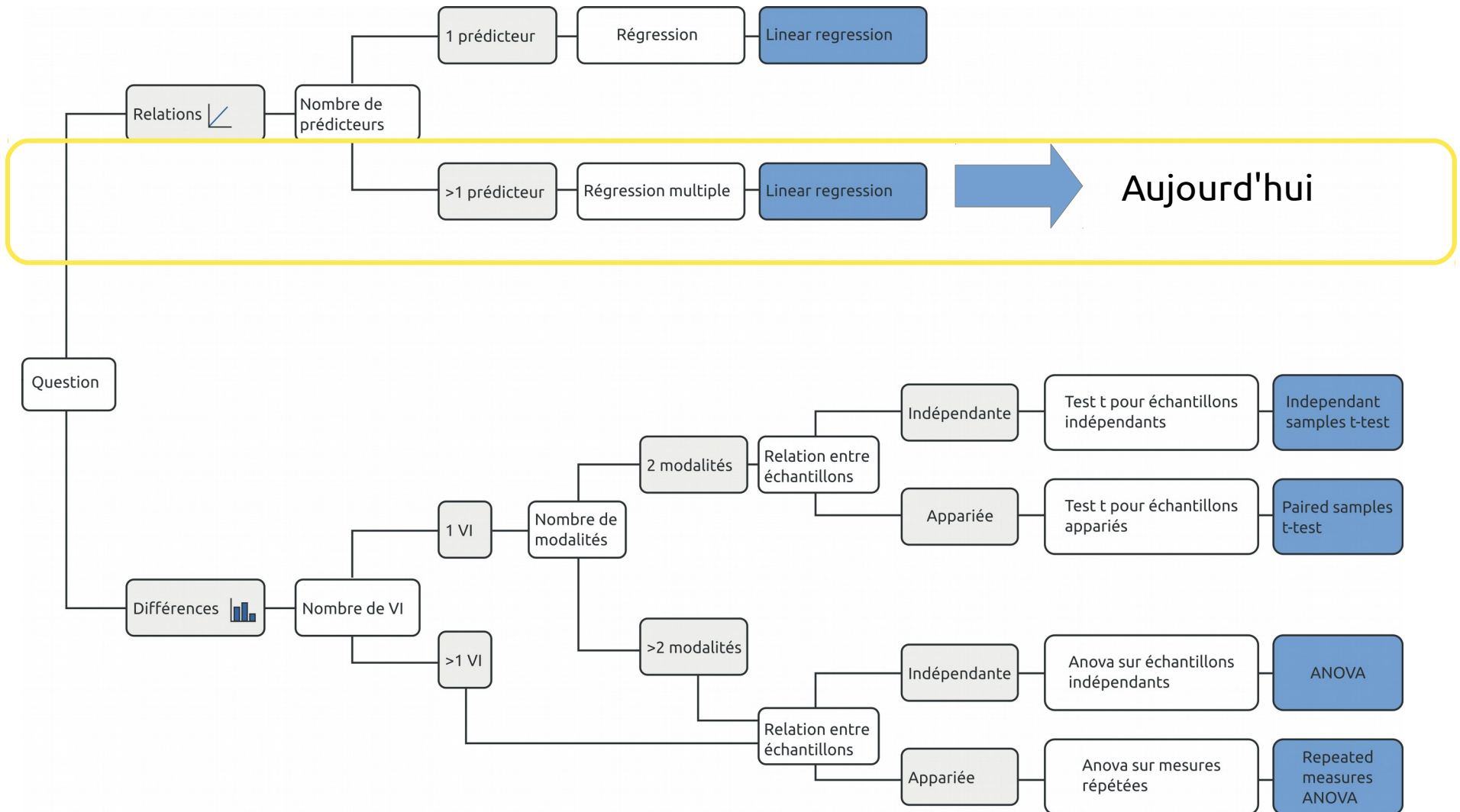
La régression linéaire multiple

Plusieurs variable explicative

La terminologie

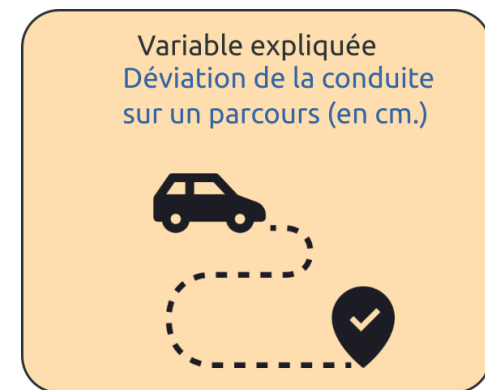
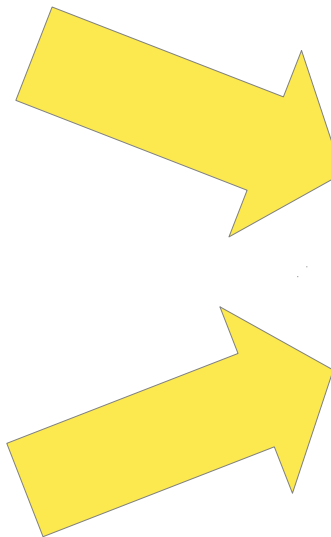
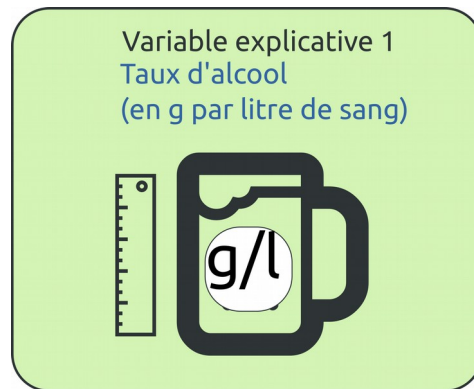
- La régression linéaire simple versus la régression linéaire multiple
 - « **simple** » → indique qu'on a une seule variable explicative
 - « **multiple** » → indique qu'on a au moins deux variables explicatives

Arbre de décision



La régression linéaire multiple

- Pourrait-on prédire la déviation sur un test de conduit sur la base du taux d'alcool et la vitesse ?



La régression linéaire multiple

- Objectifs:
 - Prédire la variable expliquée sur la base de plusieurs variables explicatives
 - Prédire la déviation du parcours sur la base de taux d'alcool et la vitesse
 - Evaluer l'importance relative de chaque variable explicative
 - Evaluer l'importance relative de la variable Taux d'alcool et la variable Vitesse

La régression linéaire multiple

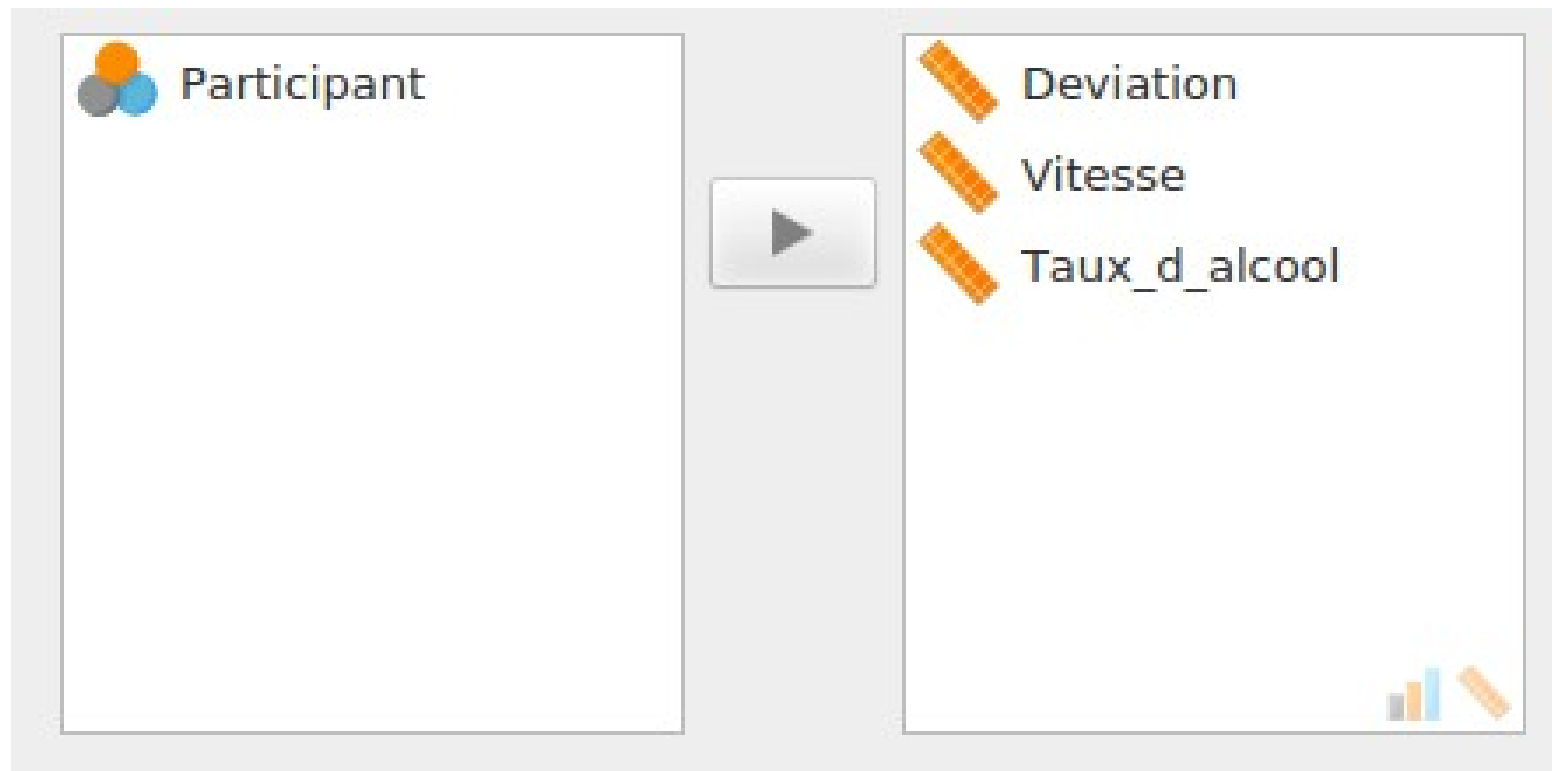
- Les statistiques inférentielles :
 - H_0 : Il n'existe pas de relation linéaire entre la combinaison des variables explicatives et la variable expliquée
 - H_1 : La combinaison des variables explicatives est associée significativement à la variable expliquée

- Fichier de données

	Variable expliquée	Var. explicative 2	Variable explicative 1
	 Deviation	 Vitesse	 Taux_d_alcool
1	61	99	0.0099
2	58	93	0.01302
3	67	91	0.03094
4	41	110	0.0242
5	44	91	0.00819
6	52	106	0.00636

JASP

- Corrélations



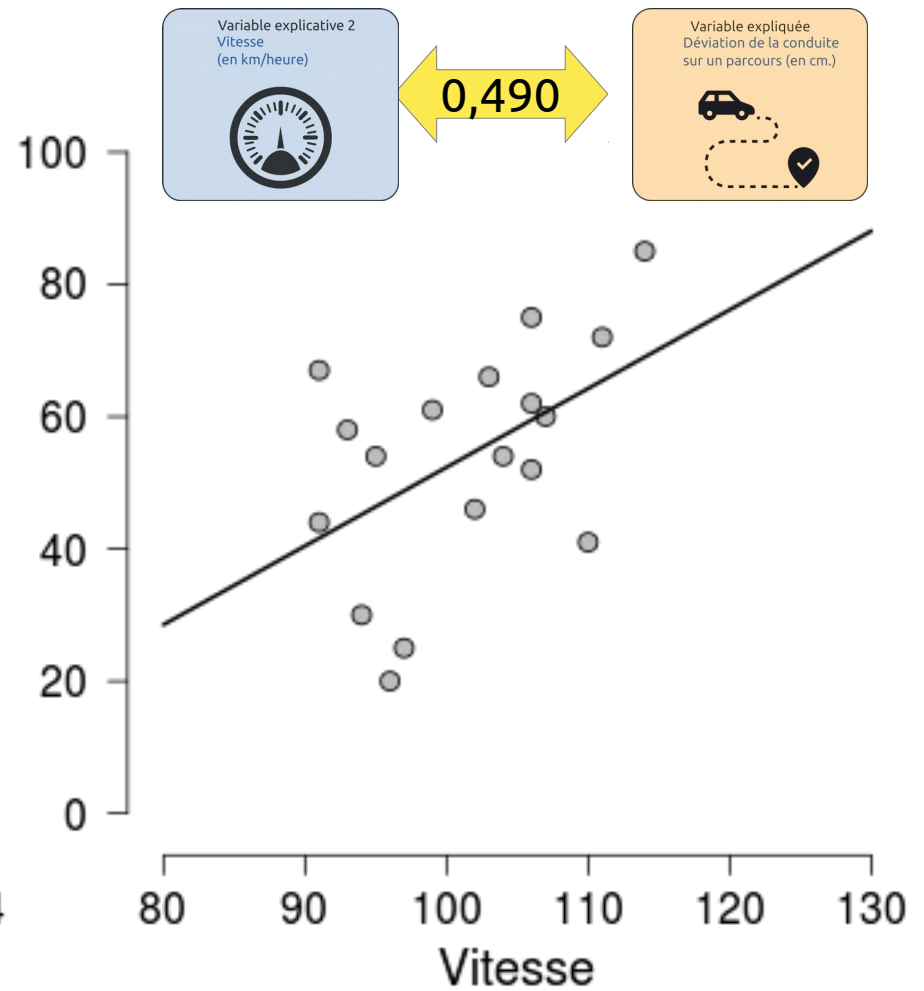
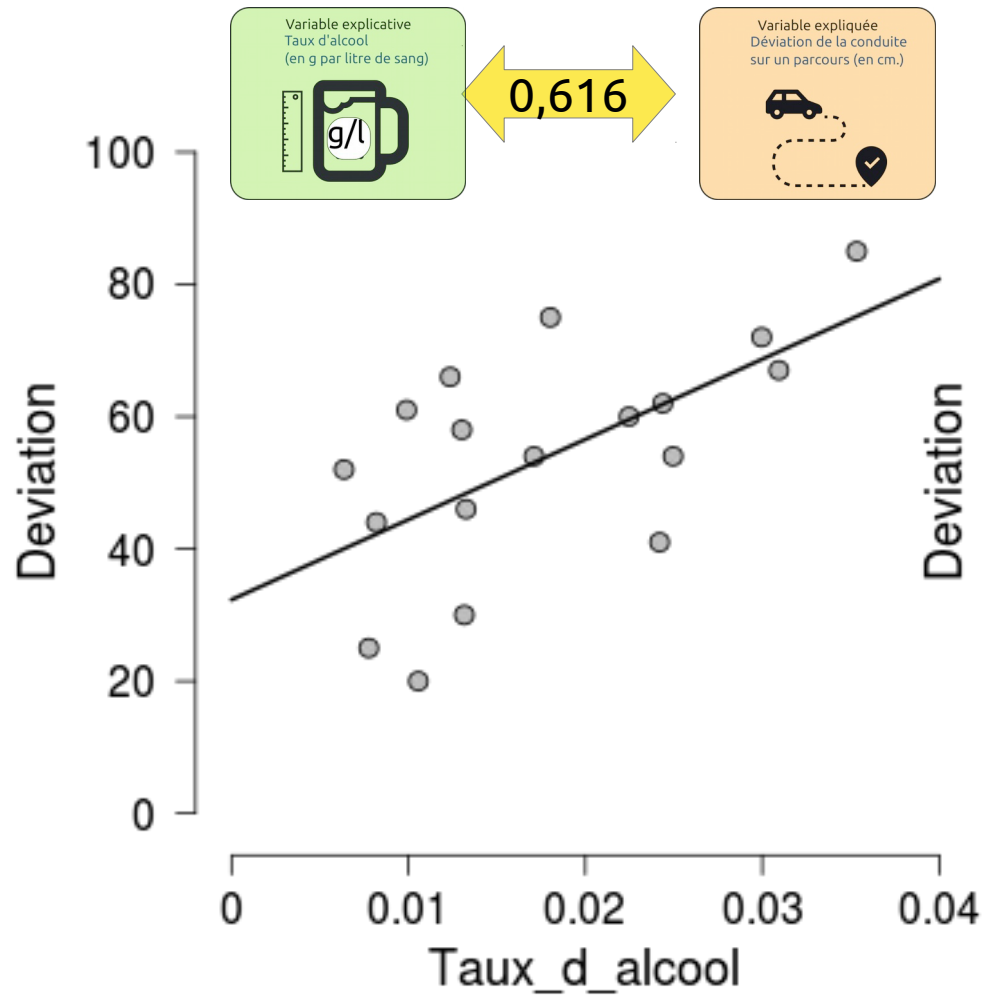
- Corrélations

Pearson Correlations

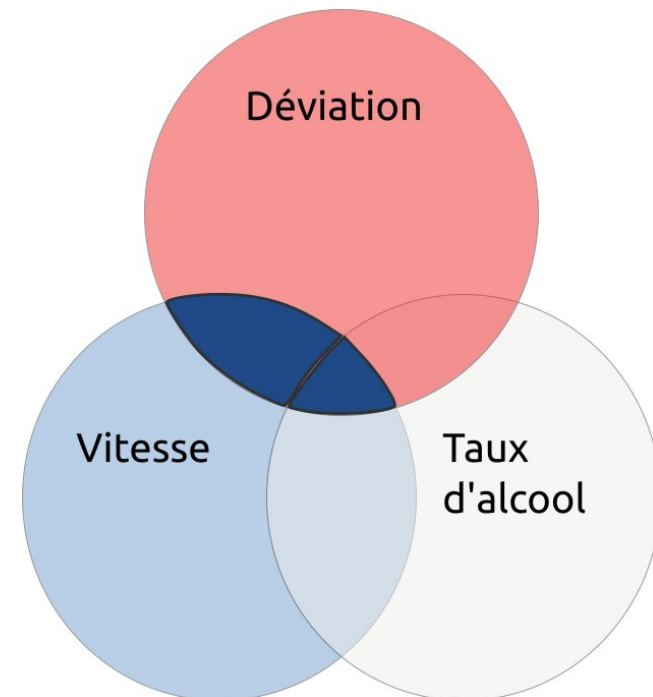
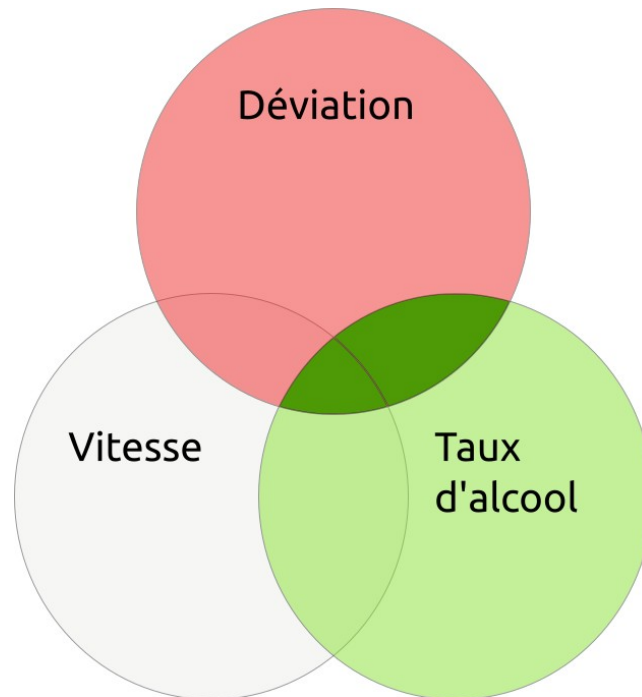
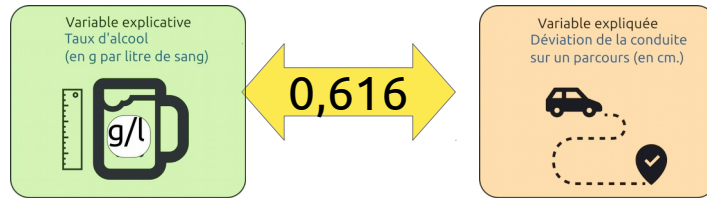
	Deviation	Vitesse	Taux_d_alcool
Deviation	—	0.490*	0.616**
Vitesse		—	0.514*
Taux_d_alcool			—

* $p < .05$, ** $p < .01$, *** $p < .001$

- Corrélations

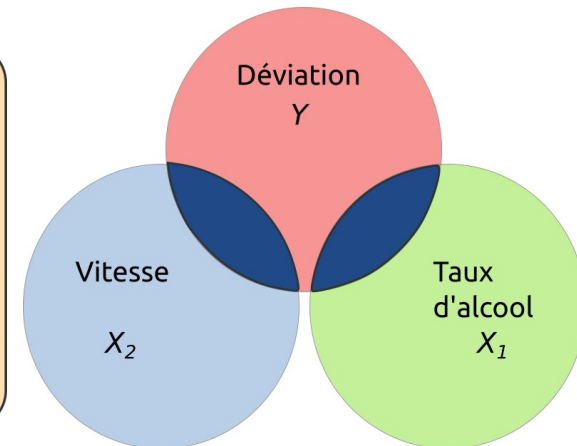
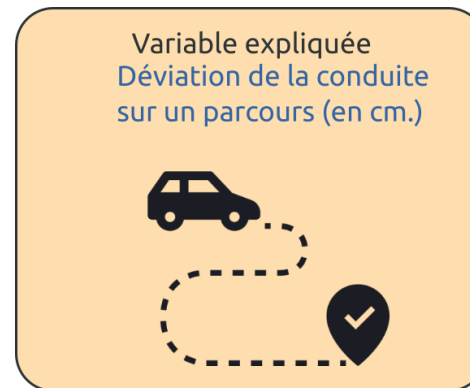
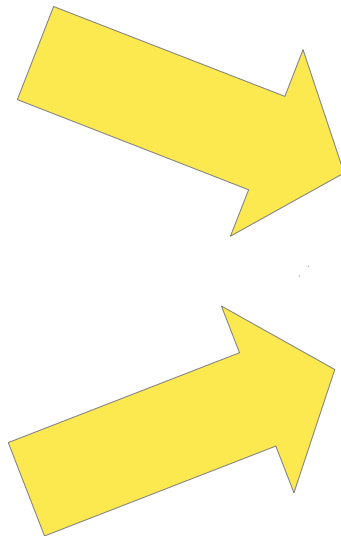
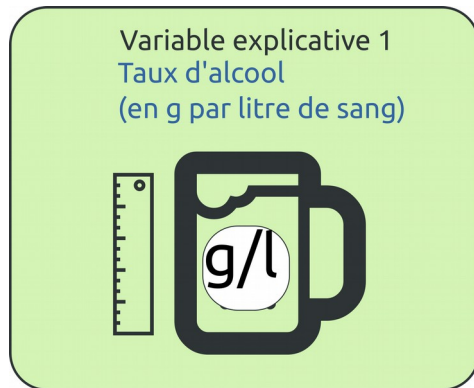


- Corrélations



La régression linéaire multiple

- Modèle avec plusieurs variables explicatives

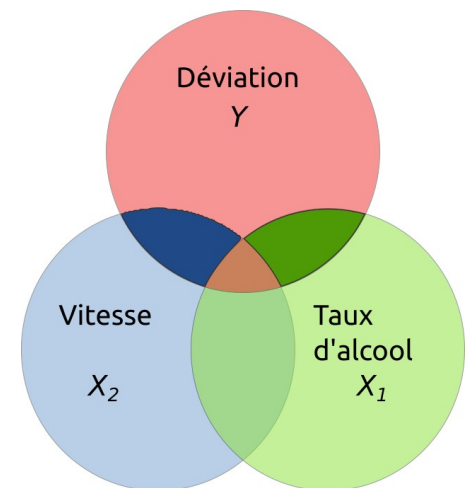
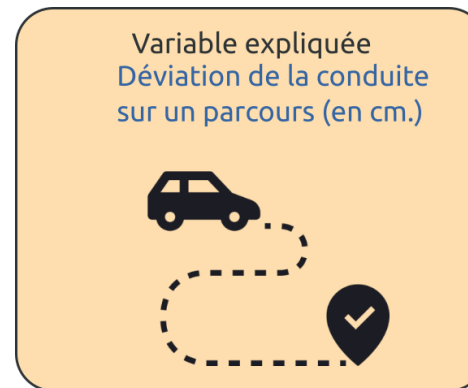
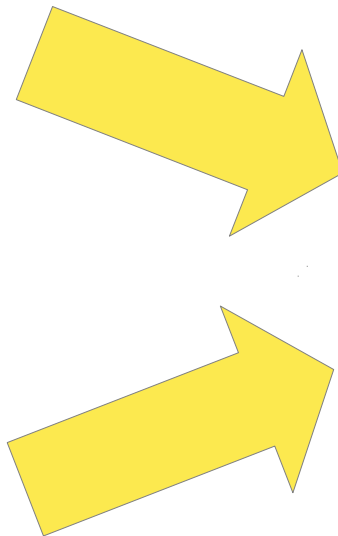
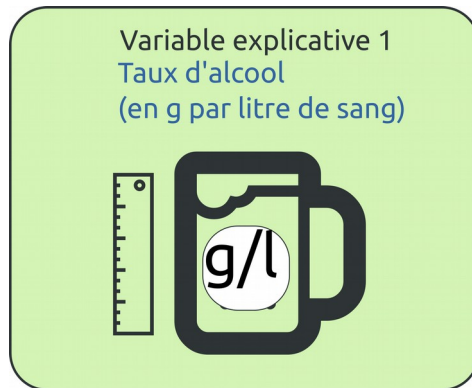


$$R^2 = r1^2 + r2^2?$$

Si taux d'alcool et vitesse ne sont pas
corrélés entre eux : oui

La régression linéaire multiple

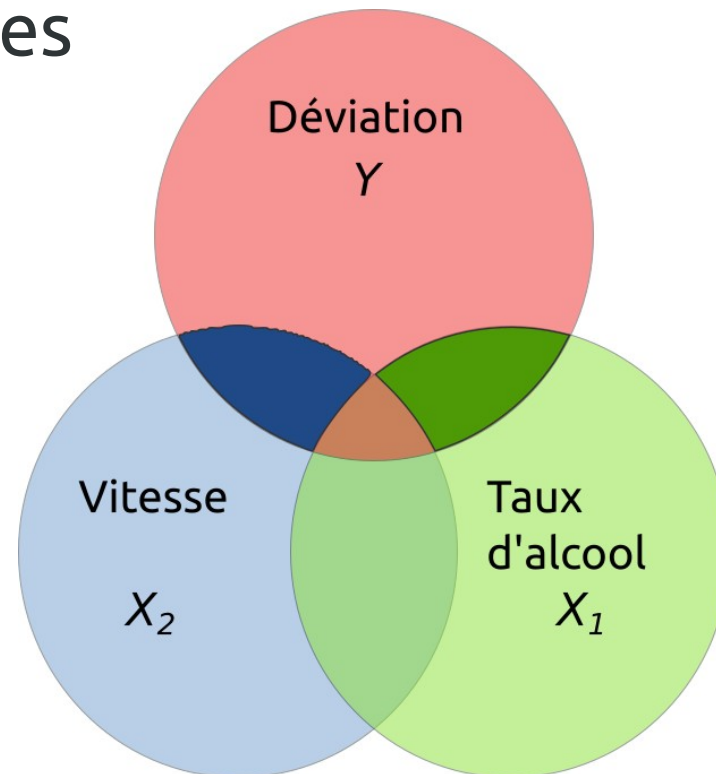
- Modèle avec plusieurs variables explicatives



$$R^2 = r_1^2 + r_2^2?$$

Si taux d'alcool et vitesse sont
corrélés entre eux : non !

- La régression linéaire multiple évalue :
 - La contribution unique de chaque variable explicatives
 - lorsque toutes les autres variables sont maintenues constantes



JASP

The image shows the JASP software interface for configuring a statistical model. On the left, a list of variables is available for selection: 'Participant' (represented by three colored circles), 'Deviation' (represented by a grid), 'Vitesse' (represented by a grid), and 'Taux_d_alcool' (represented by a grid). On the right, the 'Dependent' variable field is empty, with a play button to its left and a small orange icon to its right. Below this, the 'Method' is set to 'Enter' in a dropdown menu. The 'Covariates' section is also empty, with a play button to its left and a small orange icon to its right.

Participant

Deviation

Vitesse

Taux_d_alcool

Dependent

Method Enter

Covariates

JASP

The image shows the JASP software interface with a model setup for a participant. On the left, a box labeled "Participant" contains three colored circles (orange, grey, blue). On the right, the "Dependent" variable is set to "Deviation". Below this, the "Method" is set to "Enter". Under the "Covariates" section, two variables are listed: "Taux_d_alcool" and "Vitesse". Each variable name is preceded by a small orange icon representing a variable. The interface is light grey with white text and icons.

Participant

Dependent

Deviation

Method Enter

Covariates

Taux_d_alcool

Vitesse

- Régression linéaire multiple

Model Summary ▼

Model	R	R ²	Adjusted R ²	RMSE
1	0.648	0.420	0.343	14.060

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	2149	2	1074.4	5.435	0.017
	Residual	2965	15	197.7		
	Total	5114	17			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t-value	p
1	intercept	-21.452	52.870		-0.406	0.691
	Taux_d_alcool	974.207	451.160	0.495	2.159	0.047
	Vitesse	0.572	0.557	0.236	1.028	0.320

- Régression linéaire multiple

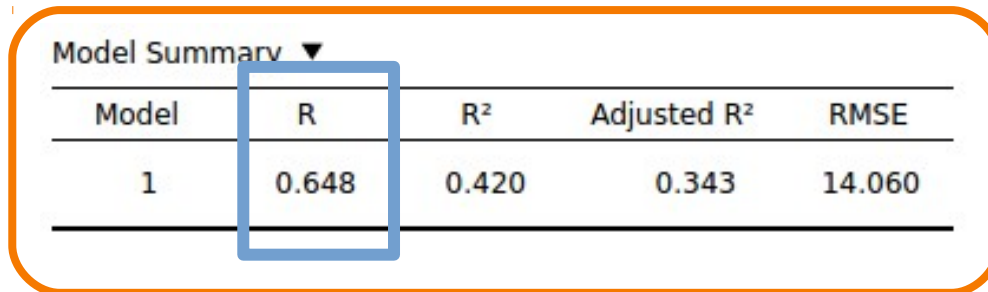
Model Summary ▼				
Model	R	R ²	Adjusted R ²	RMSE
1	0.648	0.420	0.343	14.060

Model summary :

L'influence collective des deux variables explicatives sur la variable expliquée →

L'impact collectif de Vitesse et Taux d'alcool sur la Déviation du parcours

- Régression linéaire multiple



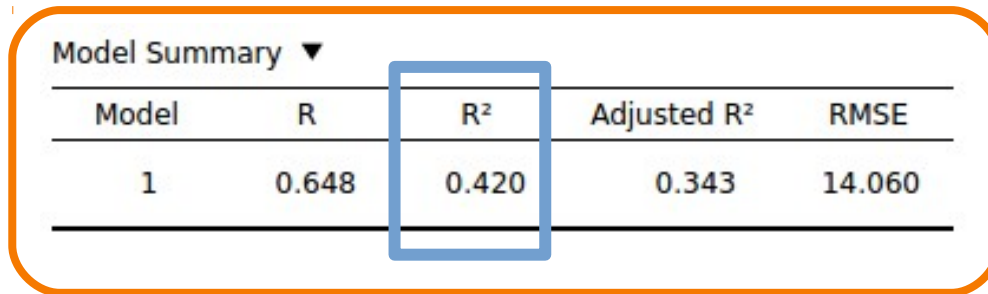
The image shows a screenshot of the 'Model Summary' table in JASP. The table has five columns: Model, R, R², Adjusted R², and RMSE. There is one row for Model 1. The value of R (0.648) is highlighted with a blue box. The entire table is enclosed in an orange rounded rectangle.

Model Summary ▼				
Model	R	R ²	Adjusted R ²	RMSE
1	0.648	0.420	0.343	14.060

R Coefficient de corrélation multiple

Coefficient de corrélation entre l'ensemble des variables explicatives et la variable expliquée

- Régression linéaire multiple



The image shows a screenshot of the 'Model Summary' table in JASP. The table has five columns: Model, R, R², Adjusted R², and RMSE. The first row of data shows Model 1 with R = 0.648, R² = 0.420, Adjusted R² = 0.343, and RMSE = 14.060. The R² value is highlighted with a blue box, and the entire table is enclosed in an orange rounded rectangle.

Model Summary ▼				
Model	R	R ²	Adjusted R ²	RMSE
1	0.648	0.420	0.343	14.060

R^2

La variance expliquée

La proportion de la variance en déviation qui pourrait être expliquée par le modèle

Ici : 42 % de la déviation est expliqué par le modèle contenant Taux d'alcool et Vitesse comme valeurs explicatives

- Régression linéaire multiple

Table d'ANOVA :

Le modèle, prédit-il la déviation de façon significative ?

ANOVA						
Model		Sum of Squares	df	Mean Square	F	p
1	Regression	2149	2	1074.4	5.435	0.017
	Residual	2965	15	197.7		
	Total	5114	17			

Oui ! $F(2,15) = 5,435$; $p = 0,017$)

- Régression linéaire multiple

L'équation :

Les deux variables explicatives, prédisent-elles la déviation du parcours de façon significative ?

Taux d'alcool : Oui !

Après d'avoir contrôlé pour la vitesse/ en assumant que la vitesse est constante, le taux d'alcool prédit la déviation

Coefficients						
Model		Unstandardized	Standard Error	Standardized	t-value	p
1	intercept	-21.452	52.870		-0.406	0.691
	Taux_d_alcool	974.207	451.160	0.495	2.159	0.047
	Vitesse	0.572	0.557	0.236	1.028	0.320

- Régression linéaire multiple

L'équation :

Les deux variables explicatives, prédisent-elles la déviation du parcours de façon significative ?

Taux d'alcool : Oui !

Si le taux d'alcool augmente, la déviation du parcours augmente aussi ($t = 2,159$; $p = 0,047$).

Coefficients						
Model		Unstandardized	Standard Error	Standardized	t-value	p
1	intercept	-21.452	52.870		-0.406	0.691
	Taux_d_alcool	974.207	451.160	0.495	2.159	0.047
	Vitesse	0.572	0.557	0.236	1.028	0.320

- Régression linéaire multiple

L'équation :

Les deux variables explicatives, prédisent-elles la déviation du parcours de façon significative ?

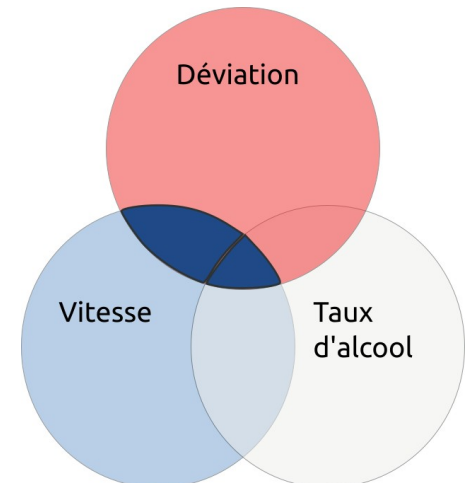
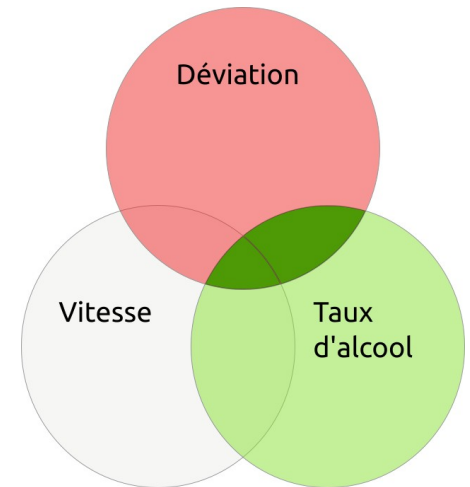
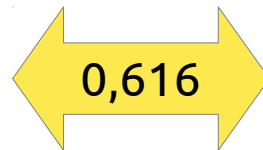
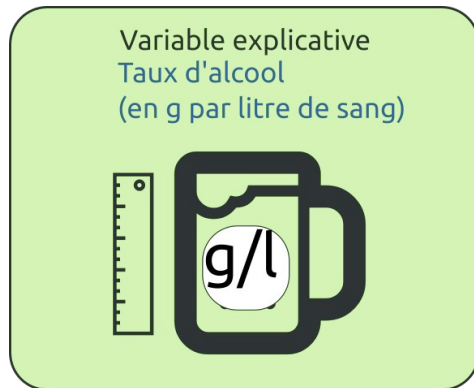
Vitesse : Non !

Après d'avoir contrôlé pour le taux d'alcool / en assumant que le taux d'alcool est constant, la vitesse ne prédit pas la déviation

Coefficients						
Model		Unstandardized	Standard Error	Standardized	t-value	p
1	intercept	-21.452	52.870		-0.406	0.691
	Taux d alcool	974.207	451.160	0.495	2.159	0.047
	Vitesse	0.572	0.557	0.236	1.028	0.320

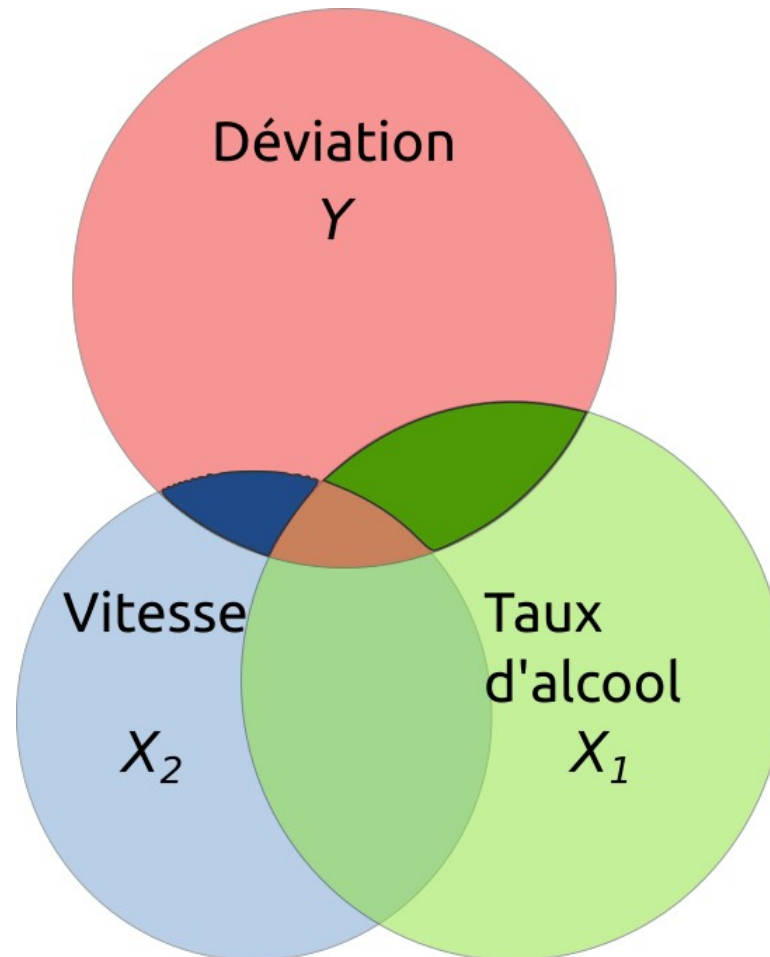
La régression linéaire multiple

- Interprétation

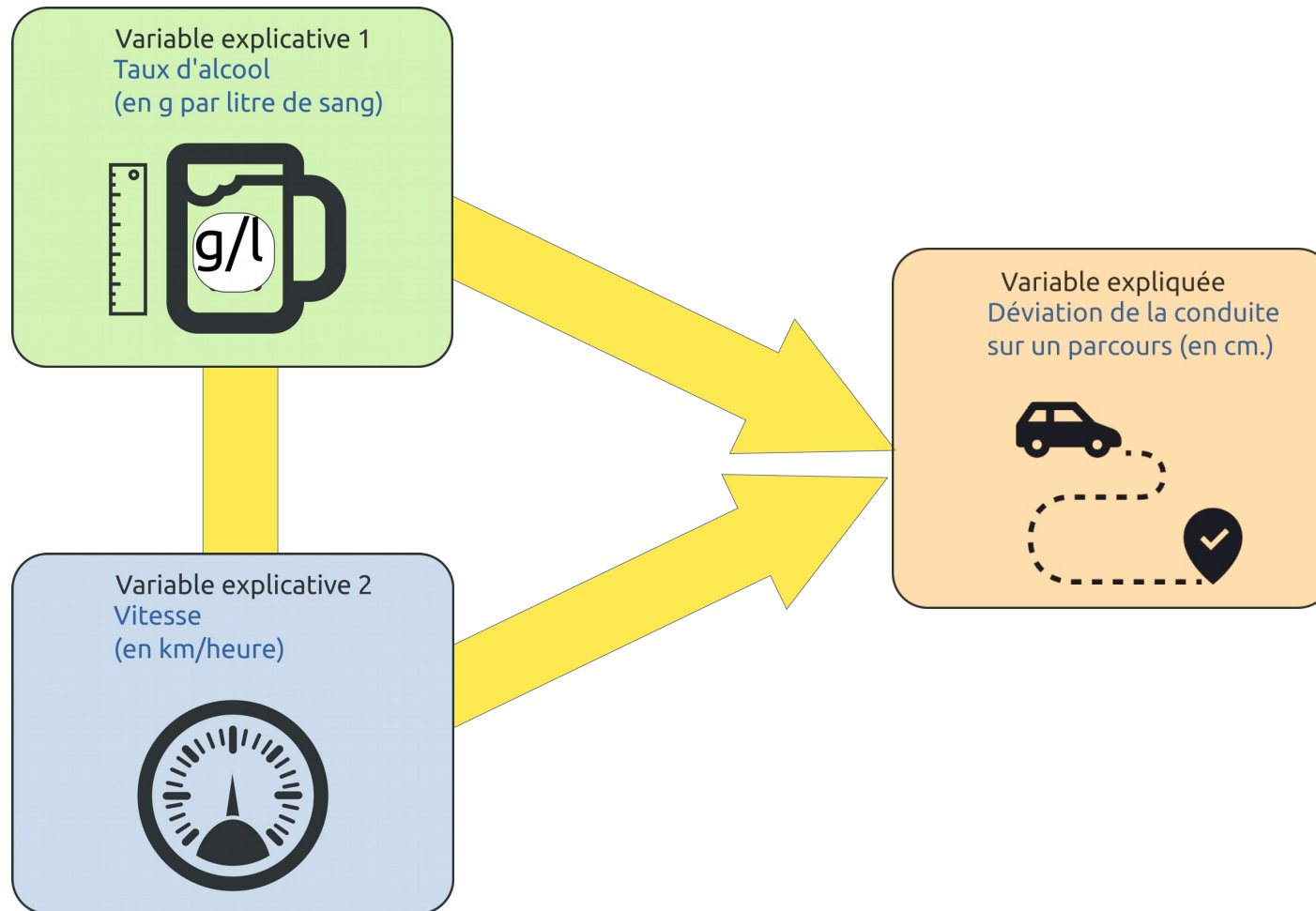


La régression linéaire multiple

- Interprétations



Conclusion



Nous avons effectué une régression linéaire multiple pour prédire la déviation du parcours sur la base du taux de l'alcool et la vitesse de conduit. Cette analyse a révélé une équation de régression significative $F(2,15) = 5,435 ; p = 0,017$) avec une R^2 de 0,420.

Le taux d'alcool est un prédicteur significatif de la déviation ($t = 2,159 ; p < 0,047$). Plus le taux d'alcool augmente, plus la déviation augmente. La vitesse n'est pas un prédicteur significatif ($p = 0,320$).

La régression linéaire multiple

- Si on voudrait tester une interaction :
 - Par ex : Est-ce que l'effet de l'alcool sur la conduite devient plus grand si la vitesse augmente ?
 - Pour spécifier l'interaction dans une régression multiple, on est obligé de créer 'à la main' une nouvelle variable, par exemple Alcool*Vitesse, en faisant le produit des deux variables Alcool et Vitesse (cf. Exercice 3)

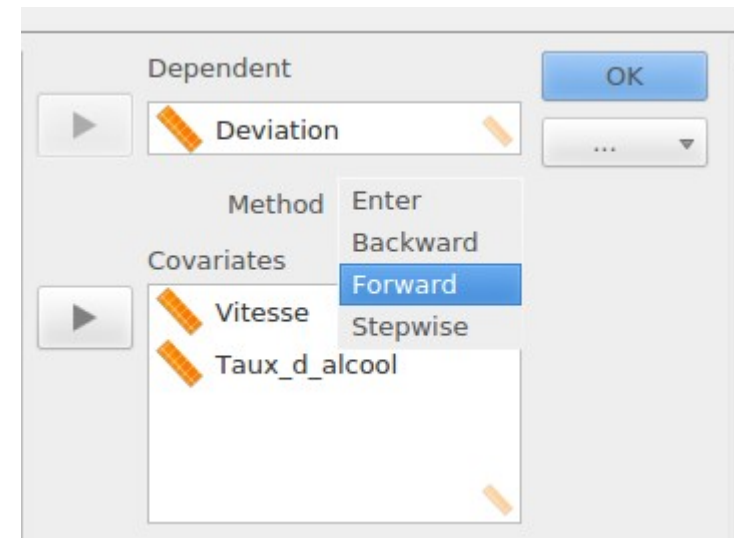
Participant	Deviation	Vitesse	Taux_d_alcool	Alcool_x_Vitesse
1	61	99	0.0099	0.9801
2	58	93	0.01302	1.21086
3	67	91	0.03094	2.81554
4	41	110	0.0242	2.662
5	44	91	0.00819	0.74529
6	52	106	0.00636	0.67416
7	46	102	0.01326	1.35252

La régression linéaire multiple

- Inclure des variables nominales.
 - Par ex : Sexe, a deux modalités : Homme et Femme
 - On est obligé de créer des « dummy variables » (aussi : variable « muette » ou « indicatrice »)
 - Si la variable a 2 modalités :
 - On code les modalités comme variable binaire : 0 et 1 (cf. Exercice 3)
 - Si la variable a >2 modalités :
 - Plus compliqué
 - On n'a pas le droit de les coder comme 0, 1 et 2
 - Il faut créer plusieurs colonnes « dummy »

La régression linéaire multiple

- Les méthodes de sélection
 - Défaut : La régression avec entrée forcée (*enter*)
 - Toutes les variables à évaluer sont entrées au même moment
 - Le test F évalue l'ensemble du modèle
 - La contribution unique de chaque variable explicatives
 - La régression avec entrée progressive
 - L'inclusion dans le modèle de chaque variable explicative dépend de sa contribution mathématique à l'amélioration du modèle
 - JASP : 3 méthodes progressives :
 - La méthode descendante / régression par élimination (*backward*)
 - La méthode ascendante (*forward*)
 - La méthode pas-à-pas (*stepwise*)



Exercices