Régression logistique

Floryan Renuy, Desrumaux Jonathan

23/10/2020

Prédiciton du diabète

Ici, l'objectif est de décrire et prédire la présence de diabète chez un patient en fonction de certaines caractéristiques cliniques, nous verrons que l'apport de la régression logistique tient dans l'interprétation des résultats du modèle.

Objectif: Identifier les facteurs de risques associés à la présence du diabète.

Pour ce jeu de données, on constate qu'il y a 9 variables quantitatives comportant 768 individus pour chacune d'entre elles:

- **Pregnancies**, la variable désignant le nombre d'enfants de la patiente,

- **Glucose**, la variable représentant la quantité de glucose dans le sang,

- **BloodPressure**, la pression artérielle de la patiente,

- **SkinThickness**, l'épaisseur de la peau de la patiente,

- **Insulin**, représentant la quantité d'insuline pour réguler la glycémie,

- **BMI** (Body Mass Index), l'IMC de la patiente,

- **Age**, l'âge de la patiente,

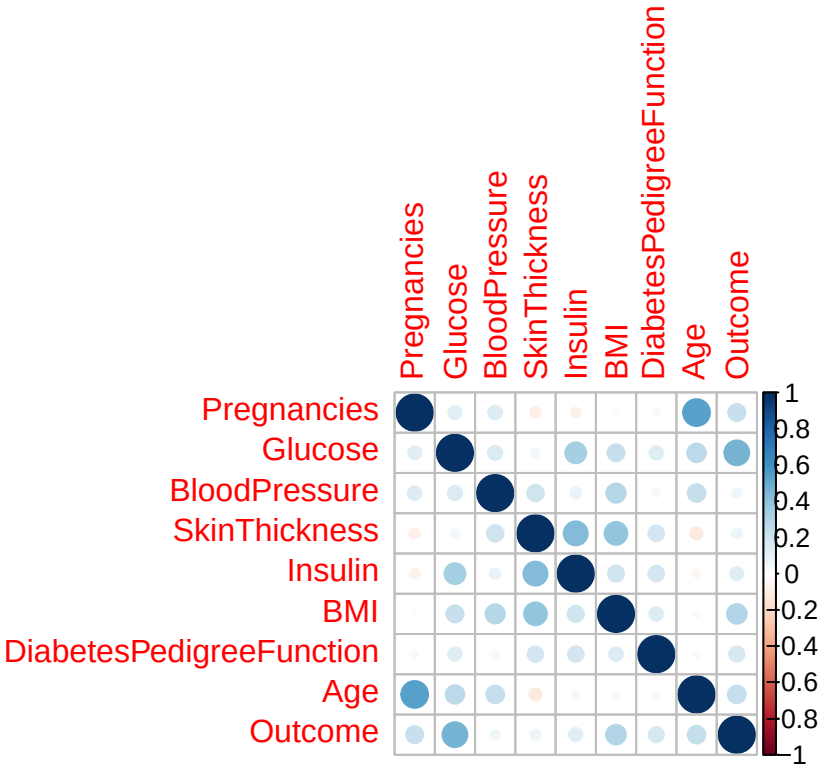
- **DiabetesPedigreeFunction**, la fonction qui évalue la probabilité d'avoir le diabète selon les antécédents familiaux,

- **Outcome**, une variable binaire (1 = la patiente a du diabète, 0 = la patiente n'a pas de diabète).

Le tableau ci-dessous dénombre les statistiques descriptives de chacune de nos 9 variables.

	Minimum	1er Q	Médiane	Moyenne	3ème Q	Maximum
Pregnancies	0	1	3	3.845	6	17
Glucose	0	99	117	120.9	140.2	199
BloodPressure	0	62	72	69.11	80	122
SkinThickness	0	0	23	20.54	32	99
Insulin	0	0	30.5	79.8	127.2	846
BMI	0	27.3	32	31.99	36.6	67.1
DPF	0.078	0.2437	0.3725	0.4719	0.6262	2.42
Age	21	24	29	33.24	41	81
Outcome	0	0	0	0.349	1	1

Voici le tableau de corrélation entre les variables:



On constate que les plus grandes corrélations s'effectue entre les couples de variables (Pregnancies, Age), (Outcome, Glucose), (Insuline, SkinThickness) et (BMI, SkinThickness).

Pour notre variable Outcome, on constate que ce sont les variables Glucose et BMI qui sont le plus corrélées avec elle.

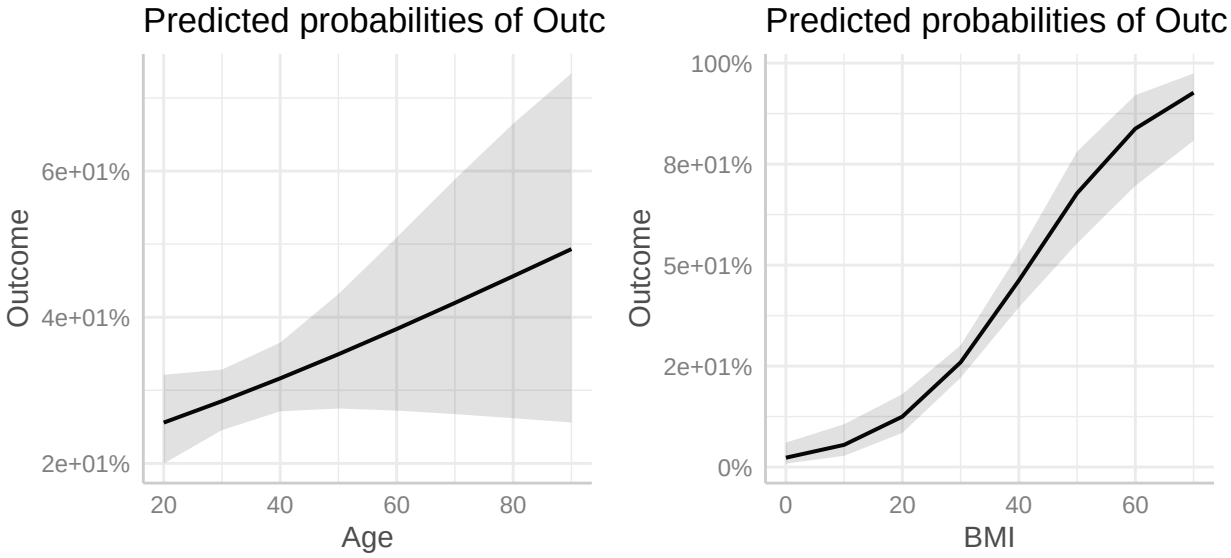
Après avoir effectuer une régression logistique de la variable Outcome, on constate que ce sont le variables Pregnancies, Glucose, BMI et DiabetePedigreeFunction qui sont le plus significatives, autrement dit, le fait dêtre enceinte, d'avoir un taux de glucose précis dans le sang, d'avoir un Indice de Masse Corporel précis et d'avoir des antécédents familiaux pour des problèmes de diabète entraîneraient un risque au patient d'être atteint de diabète. Les variables Insulin, SkinThickness et Age n'ont pas l'air d'être significatives. La variable BloodPressure semble lègérement significative.

De manière plus approndie en considérant les odds-ratio définis dans le tableau suivant :

	Odds-ratio
Pregnancies	1.131
Glucose	1.036
BloodPressure	0.987
SkinThickness	1.001
Insulin	0.999
BMI	1.094
DPF	2.573
Age	1.015

On retrouve bien que les variables Pregnancies, Glucose, BMI et DiabetePedigreeFunction ont un fort impact sur la présence de diabète chez un patient. Ce tableau nous permet de mettre des chiffres sur cet impact, par exemple, à chaque fois qu'une patiente admet une grossesse supplémentaire par rapport à une patiente qui ne va tomber enceinte qu'une seule fois, le risque de développer du diabète est multiplié par 1,131. En revanche pour la variable Bloodpressure, on constate que lorsque la pression artérielle du patient augemente, le risque de développer du diabète est multiplié par 0,987, ce qui baisse le menace, de même pour la variable Insulin, ce qui est plutôt cohérent car le rôle de l'insuline est de réguler la glycémie dans le sang, donc de la remettre dans un état stable, d'où la valeur très proche de 1 de l'odds-ratio pour cette variable.

Mais on constate que la variable Age pourrait avoir également un impact sur la présence de diabète chez le patient, nous allons essayer de confirmer cette hypothèse en retranchant cette variable en plage de périodes (20-30 ans, 30-40 ans, ...). Nous allons effectuer la même procédure pour la variable BMI, représentant l'indice de masse corporelle de la patiente, nous allons la découper selon des indices précis (0-20, 20-40, ...).

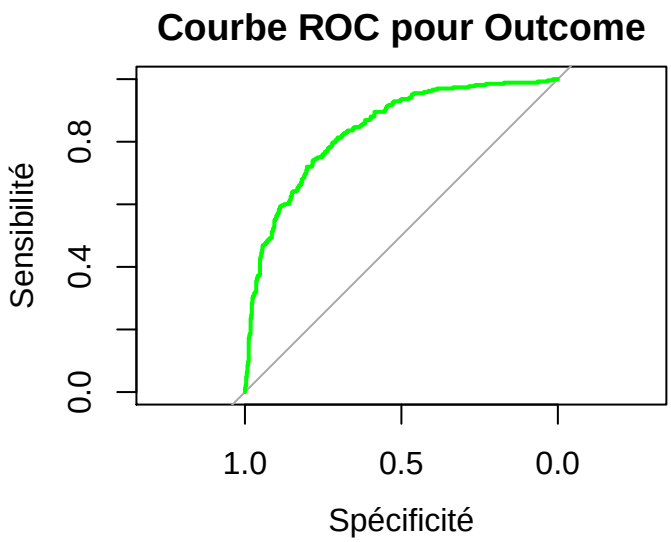


Dans les graphiques ci-dessus, on trace le risque d'avoir du diabète en fonction dans un premier temps de la variable Age puis de la variable BMI (courbe noire), l'esapce grisâtre représente l'intervalle de confiance.

On constate grâce aux graphiques que le risque de diabète augmente lorsque l'âge de la patiente augmente et de même pour l'Indice de Masse Corporelle. Donc plus la patiente devient âgée, plus le rsique de diabète sera élévé et plus la patiente sera corpulente plus le risque de diabète sera élevé. Il y a même de plus gros risques encore pour une patiente qui serait de plus en plus corpulente, on voit que l'intervalle de confiance est plus restreint que pour l'âge et que la courbe augmente beaucoup plus vite.

En conclusion les patientes sont beaucoup plus susceptibles d'être atteinte de diabète par de multiples grossesses, par leur taux de glucose dans le sang, par leur fort IMC, au cours de leur vie (plus elles vieillissent, plus le risque est élevé) et si elles ont des antécédents familiaux de cette maladie.

La courbe ROC ci-dessous nous montre que notre modèle est tout à fait correct, nous pouvions néanmoins le rendre meilleur en supprimant des variables non significatives comme Insulin ou encore SkinThickness.



Prédiciton du mode de contraception

Dans cet exercice on veut identifier les différents facteurs influençant la prise d'une contraception chez la femme. Cette étude a été réalisée sur 1473 femmes d'Indonésie. Dans un second temps, on voudrait étudier les différences entre une contraception court-terme et une contraception long-terme.

Objectif: Identifier les facteurs influençant la prise d'une contraception.

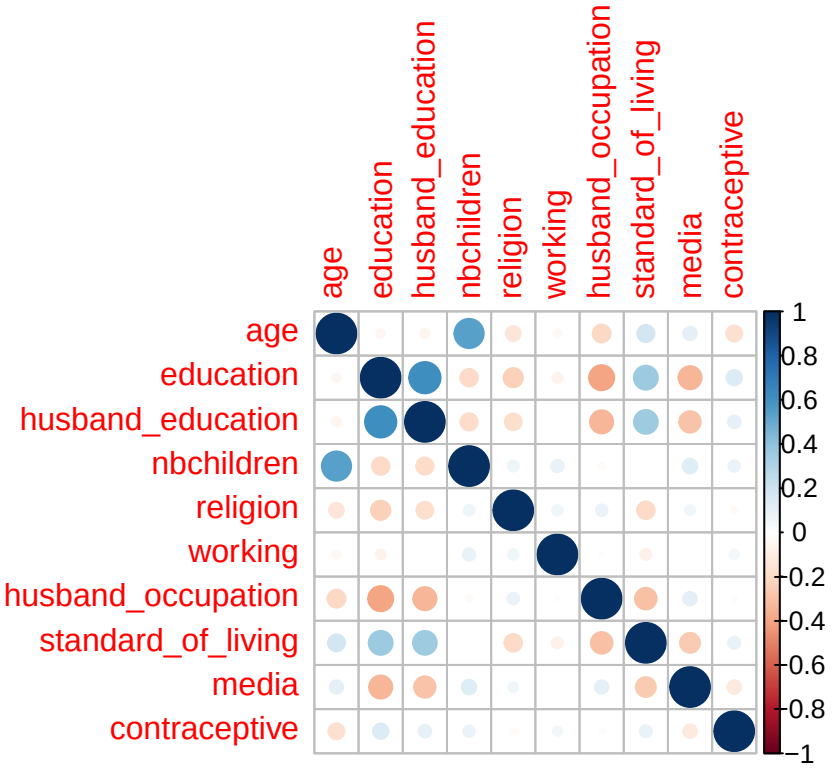
Pour ce jeu de données, on constate qu'il y a 10 variables quantitatives et 1473 observations comme dit auparavant:

- **age**, L'âge en années des femmes étudiées,
- **education**, le niveau d'éducation (codé de 1 : faible à 4 : élevé),
- **husband_education**, le niveau d'éducation du mari (codé de 1 : faible à 4 : élevé),
- **nbchlidren**, le nombre d'enfants,
- **religion**, la religion de la femme (1 : musulmane, 0 : aucune),
- **working**, si une femme travaille (1 : oui, 0 : non),
- **husband_occupation**, le niveau d'occupation du mari (codé de 1 : faible à 4 : élevé),
- **standart_of_living**, le niveau de vie du ménage (codé de 1 : faible à 4 : élevé),
- **media**, l'exposition aux média (1 : oui, 0 : non),
- **contraceptive**, le type de contraception (1 : aucune, 2 : court-terme, 3 : long-terme).

Le tableau ci-dessous dénombre les statistiques descriptives de chacune de nos 10 variables.

	Minimum	1er Q	Médiane	Moyenne	3ème Q	Maximum
Age	0	1	3	3.845	6	17
Education	0	99	117	120.9	140.2	199
husband_education	0	62	72	69.11	80	122
nbchildren	0	0	23	20.54	32	99
Religion	0	0	30.5	79.8	127.2	846
working	0	27.3	32	31.99	36.6	67.1
husband_occupation	0.078	0.2437	0.3725	0.4719	0.6262	2.42
standart_of_living	21	24	29	33.24	41	81
Media	0	0	0	0.349	1	1
Contraceptive	0	0	0	0.349	1	1

On peut voir ici la moyenne, le minimum, le maximum de chacune des caractéristiques prisent en compte chez les femmes étudiées.
Par exemple, on peut voir qu'au sein des 1473 femmes, la plus jeune a 16 ans, la plus âgée a 49 ans et la moyenne est de 32 ans et 6 mois. On peut faire la même analyse pour la caratéristique, on peut voir que le minimum est 0 (la femme n'a pas de religion) et le maximum est 1 (la femme a une religion) et la moyenne est de 0.85 ce qui veut dire que que notre étude comporte plus de femmes étant musulmanes. On peut aussi voir grâce à la variable contraceptive, qu'il y a quasiment la moitié des femmes qui utilisent une contraception à court-terme.



On constate que les plus grandes corrélations s'effectuent entre les couples de variables (nbchlidren et age), (husband_education et education). Ce qui veut dire par exemple qu'une femme ayant plus d'enfant est en général plus âgée, ce qui semble à priori logique, mais également si un homme est bien éduqué alors la femme l'est aussi. Et inversement, si la femme est mal ou peu éduquée, l'homme l'est aussi.

Pour notre variable contraceptive qui est notre caractéristique à étudier, on peut voir qu'elle est principalement en lien avec l'âge et l'éducation.

Après avoir effectué une régression logistique sur la variable “contraceptive”, on constate que lorsque la femme prend une contraception à court-terme c’ela signifie qu’elle à un bon niveau d’éducation et de vie au sein de son ménage (avec quelques enfants). De même lorsque la femme prend plutôt une contraception à long-terme c’est qu’elle à un niveau de vie plutôt moyen avec une éducation à nouveau moyenne (bien inférieure à celle de la contraception à court-terme) et quelques enfants à charge. On peut donc voir que le niveau d’éducation du mari et même son occupation n’ont peu ou pas d’impacts sur la prise d’une contraception, tout comme l’âge de l’individu. En ce qui concerne la religion et l’exposition aux médias, on peut voir que plus une femme est exposée aux médias, plus elle sera tentée de ne pas prendre de contraception, de même si elle est de religion musulmane.

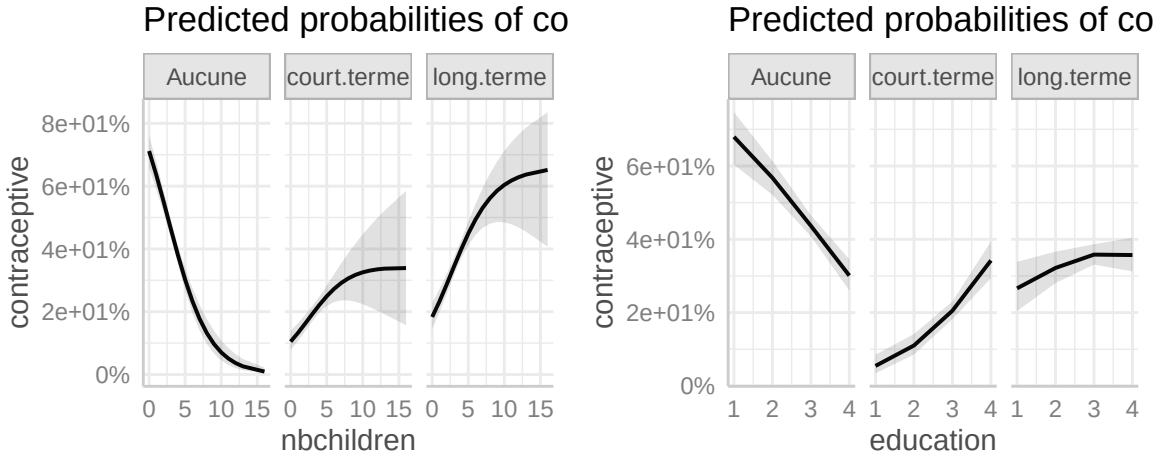
On affiche les odds-ratio dans les tableaux suivants avec à gauche le mode de contracpetion à court-terme et à droite le mode de contraception à long-terme:

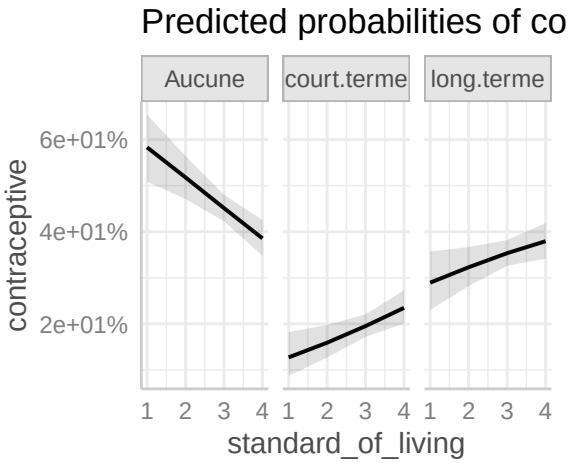
Variables	Odds-ratio
Age	0.955
Education	2.421
Husband_education	0.919
Nbchlidren	1.412
Religion	0.620
Working	1.031
Husband_occupation	0.924
Standart_of_living	1.408
media	0.643

Variables	Odds-ratio
Age	0.899
Education	1.477
Husband_education	1.058
Nbchlidren	1.419
Religion	0.722
Working	1.180
Husband_occupation	1.199
Standart_of_living	1.256
media	0.622

Ces odds-ratio nous permettent de voir à nouveau les liens entre notre variable étudiée “contraceptive” et les variables qui l’expliquent au mieux. Ici on retrouve que les variables “Education”, “nbchlidren” et “Standart_of_living” ont un impact sur la prise de contraception que ce soit à court-terme comme à long-terme. Par exemple, ces tableaux nous montre que si une femme passe de 2 enfants à 3 enfants en ayant une contraception à court-terme alors la prise de contraception est multipliée par 1.412, c’est donc trés probable qu’une femme passe d’une contraception à court-terme à une contraception à long-terme.

Regardons l’impact de ces trois variables à travers quelques graphiques.

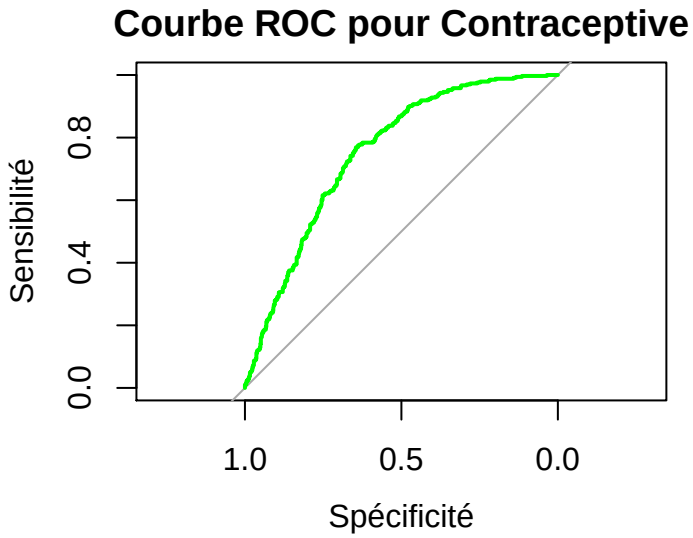




Dans ces graphiques on peut voir l'impact de nos 3 variables jugées plus intérréssantes (nombre d'enfants, education de la femme, niveau de vie). Par exemple, on s'interresse à la variable "nbchlidren". Parmis nos 1473 femmes étudiées, on regarde celle qui ont 0 enfants, on peut donc voir qu'il y a 70% d'entre elles qui n'ont aucune contraception, 10% qui ont une contraception à court-terme et 20% qui ont une contraception à long-terme.

Interressons nous maintenant au cas où il n'a pas de contraception, on peut donc voir que 80% des femmes ont 0 enfants ,seulement 7% des femmes ont 10 enfants et aucune contraception ce qui équivaut à environ 40 femmes. On peut donc voir que la courbe dans la partie "Aucune" est décroissante, donc plus une femme a d'enfants, plus elle prendra une contraception. On peut voir que les parties "court-terme" et "long-terme" ont des courbes croissantes, donc plus une femme aura d'enfant, plus elle sera tentée de prendre une contraceptio et particuliérement une contraception à "long-terme"

Nous allons regarder l'efficacité de notre modèle grâce la courbe ROC :



On peut donc voir que la courbe est plutôt moyenne, elle pourrait être améliorée en supprimant les variables "age", "working" et "husband_occupation", qui sont très peu significatives. Ce qui représente 3 variables sur 9. Ainsi, un tiers des variables n'est pas utile dans notre analyse.