<https://studium.umontreal.ca/pluginfile.php/3279545/mod_resource/content/2/co/Chap1_4_Application.html>

<https://colab.research.google.com/drive/1O2FRugTyXU3rqKOqCiUQq5bllnOGJXmy>

Bonjour,

Nous allons maintenant mettre en pratique les concepts vus tout au long de ce cours, à commencer par la régression simple avec une variable indépendante continue.

Des capsules d’exemples d’application sont accompagnées de fichiers de syntaxe et de données vous permettant de reproduire les analyses présentées dans la vidéo. Je vous encourage à essayer d’effectuer les étapes de l’analyse simultanément au fur et à mesure que je vous les présenterais afin d’acquérir une meilleure maitrise de la syntaxe R

Donc l'ensemble de ces syntaxes vont prendre la même forme, c'est-à-dire, la syntaxe va commencer avec un ensemble de blocs jaunes. En fait ces ensembles de données vont être insérées dans la syntaxe. Donc dans cet exemple-ci, on voit qu'on a une série, une colonne de chiffres qui représentent des cas différents. Ensuite, on a un certain nombre de noms d'états. Donc on va d'abord charger des données dans SAS, avant de faire l'exemple de régression simple à proprement parler. On va aussi travailler à partir de la bibliothèque ''Work'', c'est la deuxième chose importante à se rappeler. Donc tout ce qu'on va faire avec les exemples ne sera pas sauvegardé nulle part dans l'ordinateur. On va donc charger des données de manière temporaire, exécuter des exemples, je vais vous décrire une interprétation d'un tableau et puis, par la suite, quand on va fermer SAS, rien de cela ne sera sauvegardé. Donc notre premier exemple, on va utiliser une base de données qui constitue, en fait, des États américains. On va s'intéresser particulièrement au taux de chômage et puis au niveau de l'éducation, aux dépenses effectuées en éducation au sein de chacun des états. On va essayer de voir s'il y a une relation entre ces deux variables. Donc on va d'abord y aller étape par étape, on va charger nos données. Pour ça, il faut sélectionner l'ensemble du code pour charger les données. C'est vraiment sélectionner et dérouler vers le bas jusqu'à la fin de cet ensemble de données. C'est important de sélectionner le dernier mot-clé ''run'', qui va faire exécuter l'ensemble de cette procédure. D'abord, on va charger des données, on sélectionne ça et on appuie, toujours dans SAS pour faire exécuter des procédures, on va appuyer sur notre petit bonhomme qui court. C'est normal qu'il ne s'est rien produit. Souvent, on s'attend à ce qu'il y ait quelque chose qui apparaisse lorsqu'on fait exécuter une procédure. Mais dans SAS, comme je le rappelle, il y a beaucoup de choses qui se passent à l'arrière. Donc pour valider que tout s'est bien déroulé, je vais aller appuyer sur l'onglet ''Journal'' et puis on va aller observer qu'est-ce qui s'est passé. Donc en changeant d'onglet, dans le bas, on est maintenant dans le journal. Et puis on voit que la procédure ''data regsimple'' a bien été exécutée. Il y a un ensemble de variables qui ont été créées. En fait, il y a 51 observations qui ont été chargées et 12 variables. Donc ça confirme qu'on a bien chargé nos données. Pour contre-valider la chose, on peut aller sur le côté gauche, dans l'explorateur de SAS, on va aller double-cliquer sur la bibliothèque, aller voir la bibliothèque ''Work'' et constater que ''Regsimple'', c'est bien l'ensemble de données qu'on vient de charger, se trouve dans la bibliothèque ''Work''. Parce que c'est notre première fois, je donne plus de détails. Ce sont des choses que vous devriez naturellement aller vérifier à chaque fois que vous ouvrez SAS et charger des données, que les données se retrouvent bien dans la bibliothèque où vous les avez placées pour pouvoir travailler. Je retourne maintenant dans ma syntaxe pour pouvoir effectuer la régression en tant que tel. Donc la première chose à faire, lorsqu'on charge une nouvelle base de données, pour bien se familiariser, on va d'abord produire les moyennes. On va y aller en séquences, on va produire les moyennes de nos deux variables qui nous intéressent. On va ensuite faire un graphique de dispersion pour observer la relation qu'on s'attend à voir dans la régression. Donc on va observer ce graphique. Et puis, ensuite, on va faire l'analyse de régression en tant que tel. Donc j'y vais avec la première procédure. Je vous invite à consulter la documentation de SAS ou celle, encore une fois, dans le matériel pédagogique du cours, pour pouvoir avoir les détails sur chacun des mots-clés des syntaxes. Dans les capsules ici, je vais simplement sélectionner ou vous mentionner le code et exécuter. Donc le premier bloc de syntaxe, en fait, simplement, nous sert à produire les moyennes de nos deux variables. Donc, encore une fois, je sélectionne. Il y a toujours le petit bonhomme qui court pour aller exécuter la procédure ou un raccourci, c'est le bouton F3 sur votre clavier. Je vais souvent faire ça, ça va beaucoup plus rapidement. Donc comme vous voyez à l'écran, tout ce qui est apparu pour l'instant en fait, c'est un tableau de moyenne. Donc on a nos deux variables, ''scs134'' ça représente le montant mensuel dépensé en éducation au sein de chacun des états et ''jbs170'', c'est notre variable ''taux de chômage''. Donc on voit le n, il y a 51 États américains. Et puis le minimum, le maximum de chacune des variables, moyennes et écarts-types. Donc première étape : produire nos tableaux de moyennes. Deuxièmement, on va vérifier la distribution de nos variables, voir si on respecte le postulat de normalité. Donc à savoir si nos deux variables sont distribuées normalement ou si certaines transformations pourraient être utiles à faire. Donc je sélectionne le deuxième bloc de codes, le bloc pour créer des histogrammes. Et j'exécute la procédure. Donc le premier tableau, ce sera l'histogramme du taux de chômage des États américains. Alors on a la distribution qui n'est pas tout à fait normale. On va quand même continuer avec notre exemple. Mais, à part un état américain, qui semble avoir un taux de chômage beaucoup plus élevé, donc à peu près 15, que l'ensemble des autres, la partie gauche serait relativement normale. Donc on voit à l'oeil que la majorité des états se situent entre 2.5 et 7.5% de taux de chômage. On déroule vers le haut, parce que SAS, quand on exécute une procédure, il nous amène directement à la fin. Donc on va souvent avoir à dérouler vers le haut pour retrouver les autres sorties qui ont été produites. Donc en déroulant vers le haut, on a, finalement, le premier graphique d'histogramme qu'on a demandé à être produit, qui est celui du nombre de dépenses mensuelles en éducation. Alors on voit que cette distribution est un peu plus normale que le tableau précédent. Donc rien à dire de plus pour l'instant dans ces deux histogrammes. Alors on va poursuivre avec le graphique de dispersion qui nous montre la relation entre les deux. Je sélectionne donc notre troisième bloc de syntaxe, qui va nous servir à produire le graphique de dispersion, et j'exécute la procédure. Alors on se trouve ici avec un très grand graphique. On voit mal, je vais le monter un peu. En fait, je vais le cacher parce qu'on s'intéresse plus au point en haut. Sur l'axe des X, on a la dépense mensuelle en éducation, donc la variable éducation. Sur l'axe des Y, on a notre variable dépendante qui est le taux de chômage. Donc on voit qu'il semble y avoir une tendance à augmenter légèrement. Plus les dépenses en éducation augmentent, plus le taux de chômage augmente également. Alors il y a quelques points, par-contre, qui semblent sortir du nuage de tendance générale. C'est-à-dire, il y a un point tout en haut, qu'on voyait tantôt dans notre histogramme aussi, un état américain avec un taux de chômage très élevé, autour de 15. Donc on voit ce point-là se retrouve ici, qui est très loin du nuage de points. Donc la régression, encore une fois, ça sert à qualifier, par une droite, ce nuage de points. Donc on s'attend à voir une relation probablement significative qui va être relativement positive. À chaque fois qu'on avance, qu'on dépense plus en éducation, on a un taux de chômage un peu plus élevé. Donc on va aller exécuter notre régression en tant que tel et on va voir qu'est-ce que ça nous donne. Donc le quatrième bloc de code, l'analyse de régression. Donc l'analyse de régression va toujours faire ça de la même manière. Dans le code, on va écrire sur la ligne ''model'' notre variable dépendante à gauche d'un ''='' (ici jbs170, c'est le taux de chômage) qu'on veut prédire par les dépenses en éducation. Alors en sélectionnant le code, et en exécutant, encore une fois avec F3 sur votre clavier ou le petit bonhomme qui court dans la barre de raccourcis. Donc dans la sortie, on est tout de suite à la fin. Je vais remonter jusqu'au début, ici la procédure, pour passer chacun des tableaux avec vous. Donc quand on exécute une régression dans SAS, on a d'abord quelques détails descriptifs. Donc premièrement, il nous dit qu'on a utilisé la procédure ''reg'', spécifique à la variable dépendante qu'on a demandée, c'est ''jbs170''. C'est bien ce qu'on a demandé. Donc c'est important de valider, surtout dans les premiers temps, pour vous assurer que vous n'avez pas inverser, par exemple, les variables. Donc on a 51 observations lues, le n de notre échantillon, chose qui est importante à valider, on a bien 51 états américains. Ensuite, on a un tableau sur l'analyse de variance. Je vais passer un peu plus rapidement, mais ça nous décrit en fait la somme des carrés inter-groupes et intra-groupes. Donc le ''erreur'' c'est ce qui se trouve à être la somme des carrés inter-groupes. La ligne ''model'', c'est la somme des carrés intra-groupes. Donc le test F est significatif, ce qui nous indique qu'une de nos variables va l'être aussi dans notre modèle. C'est un modèle de résidus simples, donc on sait tout de suite que notre seule variable sera significative. Ensuite, on descend dans le deuxième tableau. On a une statistique qui nous intéresse particulièrement, c'est le R carré, qu'on appelle ''R2'' aussi. Donc le R carré est 0,0968, ce qui implique environ 10% de la variance du taux de chômage serait expliqué par les dépenses en éducation. Je vais descendre un peu, dérouler notre menu pour aller voir les ''Valeurs estimées des paramètres'', qui est le tableau en tant que tel qui contient les résultats de notre régression. Alors sur la ligne ''Intercept'', on voit aussi la constante dans d'autres types de logiciels, donc on trouve ici la valeur estimée de l'endroit, en fait, que la droite croise. Pour un état qui dépenserait aucun dollar en éducation, on s'attendrait à un taux de chômage de 2,86%. Donc ça c'est notre valeur qu'on va utiliser si on veut calculer des valeurs prédites. La deuxième ligne ''scs134'' se trouve à être les résultats associés à notre variable. Donc on voit une valeur estimée qui est positive. Tel qu'on s'attendait dans le graphique de dispersion, ça augmente un peu, donc .20, .21 disons. Et puis cette variable, le ''p'' c'est le p=0.0263, on voit que c'est plus petit que 0.05. Donc la relation va être significative. Donc on conclurait de cet exemple que plus les dépenses en éducation augmentent, plus le taux de chômage aussi augmente dans chacun des états américains. Alors ce sont des données fictives, ce qui explique le sens de cette relation qui est un peu surprenant. Mais tout de même! Alors ça conclut notre premier exemple sur la régression. Dans la prochaine capsule, on va continuer avec cet exemple, mais on va explorer les résidus. Parce qu'on a vu, si on se rappelle, qu'il y avait un état américain avec un taux de chômage particulièrement élevé. Alors on va examiner ça dans notre prochaine capsule, voir quelles sont les conséquences sur l'estimation qu'on vient de faire de cet état américain.

Lien Youtube : <https://www.youtube.com/watch?v=tXeGuimsw20>