MATH60604
Modélisation statistique
§ 2h - Colinéarité

HEC Montréal
Département de sciences de la décision

Multicolinéarité

- On dit que deux covariables X_1 et X_2 sont colinéaires si
  - X_1 et X_2 sont toutes deux corrélées avec Y
  - X_1 et X_2 sont fortement corrélées entre elles — elles contiennent essentiellement la même information.
- Il peut y avoir de la multicolinéarité entre plus de deux variables...de la même façon qu'il pourrait y avoir plus d'un facteur confondant.
- Dans ce cas, la multicolinéarité (ou colinéarité) sert à décrire le cas de figure où une (ou plusieurs) variable explicative est fortement corrélée avec une combinaison linéaire des autres covariables.
- Une conséquence nuisible de la multicolinéarité est la perte de précision dans l'estimation des paramètres, et donc l'augmentation des erreurs-type des paramètres.

Un exemple débile pour illustrer la colinéarité

- Considérez le log du nombre quotidien de locations de Bixi en fonction de la température en degrés Celcius et Farenheit (et la température en °F arrondie au degré près). Soit le modèle linéaire

$$\texttt{lognutilisateur} = \beta_0 + \beta_{\texttt{c}}\texttt{celcius} + \beta_{\texttt{f}}\texttt{farenheit} + \varepsilon.$$

- L'interprétation de $\beta_{\texttt{c}}$ est « le facteur d'augmentation du nombre de locations quotidiennes quand la température croît de 1°C, tout en gardant la température F constante »...
- Les deux unités de températures sont liées par la relation linéaire

$$1.8\texttt{celcius} + 32 = \texttt{farenheit}.$$

- Supposons que le vrai effet (fictif) de la température sur le log du nombre de locations de vélo est

$$\texttt{lognutilisateur} = \alpha_0 + \alpha_1 \texttt{celcius} + \varepsilon.$$

- Les coefficients du modèle qui n'inclut que la température Farenheit sont donc

$$\texttt{lognutilisateur} = \gamma_0 + \gamma_1 \texttt{farenheit} + \varepsilon.$$

  où $\alpha_0 = \gamma_0 + 32\gamma_1$ et $1.8\gamma_1 = \alpha_1$.
- Les paramètres du modèle postulé avec les deux variables,

$$\texttt{lognutilisateur} = \beta_0 + \beta_{\texttt{c}} \texttt{celcius} + \beta_{\texttt{f}} \texttt{farenheit} + \varepsilon,$$

  ne sont pas **identifiables**: n'importe laquelle combinaison linéaire des deux solutions donne le même modèle ajusté.

Bixi et multicolinéarité

On utilise des données tirées du site de Bixi avec la température à 16h (`rfarenheit` dénote la température Farenheit arrondie) pour expliquer le nombre de locations quotidiennes entre 2014 et 2019.

Paramètre	Estimation	Erreur type	Valeur du test t	Pr > |t|
Constante	8.844327052	0.02819099	313.73	<.0001
celcius	0.048566261	0.00135205	35.92	<.0001

Paramètre	Estimation	Erreur type	Valeur du test t	Pr > |t|
Constante	7.980926861	0.05132678	155.49	<.0001
farenheit	0.026981256	0.00075114	35.92	<.0001

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|
Constante	8.844327052	B	0.02819099	313.73	<.0001
celcius	0.048566261	B	0.00135205	35.92	<.0001
farenheit	0.000000000	B	.	.	.

Paramètre	Estimation	Erreur type	Valeur du test t	Pr > |t|
Constante	9.555086770	1.14747585	8.33	<.0001
celcius	0.088592866	0.06461502	1.37	0.1706
rfarenheit	-0.022227045	0.03587330	-0.62	0.5356

SAS imprime un avertissement en cas de colinéarité.

> *Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.*

Effet de la colinéarité

Règle générale, la colinéarité a les impacts suivants:

- Les estimés des coefficients changent drastiquement quand de nouvelles observations sont ajoutées au modèle, ou quand on ajoute/enlève des variables explicatives.
- Les erreurs-type des coefficients de la régression linéaire sont très élevées, parce que les β ne peuvent pas être estimés précisément.
- Conséquemment, les intervalles de confiance pour ces coefficients sont très larges.
- Les paramètres individuels ne sont pas statistiquement significatifs, mais le test F pour l'effet global du modèle indiquera que certaines variables sont utiles.

Comment détecter la multicolinéarité et les facteurs confondants?

- Si les variables sont exactement colinéaires, SAS et R en enlèvera une (SAS imprime la même remarque que lorsque vous déclarez des variables catégorielles à l'aide de `class`).
  - Les variables qui ne sont pas **parfaitement** colinéaires (par exemple arrondies) ne seront pas détectées par le logiciel et poseront problème.
- On peut regarder la **corrélation linéaire** entre variables explicatives et les changements dans les estimés des paramètres pour les régressions avec et sans certaines variables.
- Quand il y a plus de deux variables multicolinéaires, la détection est moins facile.
- Une variable explicative peut être corrélée avec une combinaison linéaire d'autres variables sans forcément avoir une corrélation très forte avec les variables individuelles.

Facteur d'inflation de la variance

- Un autre outil pour détecter la multicolinéarité est le facteur d'inflation de la variance (VIF).
- Pour une variable explicative donnée X_j, son VIF est

$$\mathrm{VIF}(j) = \frac{1}{1 - R^2(j)}$$

  où $R^2(j)$ est le R^2 du modèle obtenu en régressant X_j sur les autres variables explicatives.
- On parle parfois de facteur de tolérance, $\mathrm{TOL} = 1 - R^2(j)$, soit la réciproque du `VIF`.

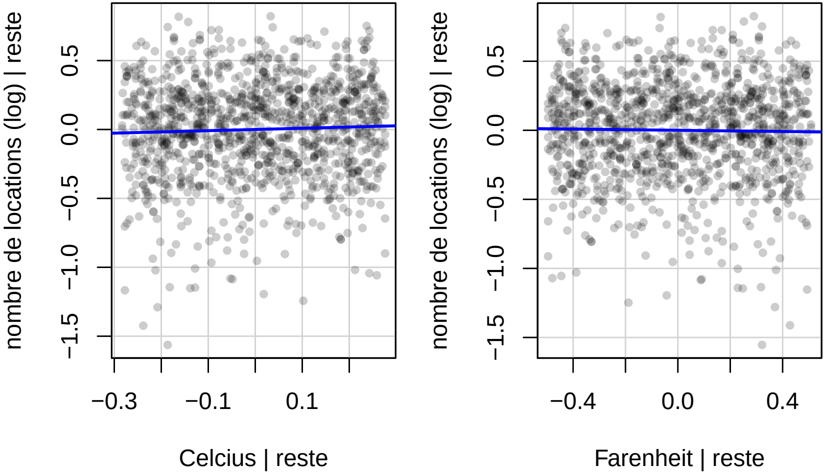
Quand est-ce que la colinéarité est problématique?

- $R^2(j)$ représente la proportion de la variance de X_j qui est expliquée par les autres prédicteurs.
- Il n'y a pas de concensus mais, règle générale,
  - $\mathrm{VIF}(j) > 4$ ou $\mathrm{TOL} < 0.25$ si $R^2(j) > 0.75$
  - $\mathrm{VIF}(j) > 5$ ou $\mathrm{TOL} < 0.2$ si $R^2(j) > 0.8$
  - $\mathrm{VIF}(j) > 10$ ou $\mathrm{TOL} < 0.1$ si $R^2(j) > 0.9$.

Multicolinéarité et Bixi

- La valeur de la statistique F pour le test de significativité globale (omise de la sortie) du modèle linéaire simple avec température Celcius est 1292 avec une valeur-p de moins de 0.0001; cela suggère que la température est un excellent prédicteur (5% d'augmentation du nombre d'utilisateurs pour chaque augmentation de 1°C).
- En revanche, dès qu'on inclut Celcius et Farenheit (arrondi au degré près), les coefficients individuels ne sont plus statistiquement significatifs à niveau 5%.
- Qui plus est, le signe du coefficient de `rfarenheit` est différent de celui du modèle avec `farenheit`!
- Remarquez que les erreurs-type de Celcius sont 48 fois plus grandes dans le modèle avec les deux variables.
- Les facteur d'inflations de la variance de `celcius` et `rfarenheit` sont énormes (2282) et permet de diagnostiquer le problème.

Diagrammes de régression partielle pour Bixi



Exemple fictif d'un modèle avec un problème de multicolinéarité

- Voici un exemple fictif avec 100 observations d'une variable réponse Y avec cinq variables explicatives X_1 à X_5
- Les données sont dans la base de données `simcolineaire.sas7bdat`.
- En réalité, les valeurs de Y ont été générées aléatoirement selon le modèle

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 + \varepsilon$$

- Le paramètre associé à chaque variable explicative est 1.

Exemple fictif de multicolinéarité

Voici d'abord la matrice des corrélations entre toutes ces variables.

code SAS pour la corrélation

```
proc corr data=infe.simcolineaire noprob;
var y x1-x5;
run;
```

Coefficients de corrélation de Pearson, N = 100

	Y	X1	X2	X3	X4	X5
Y **Y**	1.00000	0.45184	0.45549	0.64572	0.41047	0.34706
X1 **X1**	0.45184	1.00000	0.05607	0.68896	0.14553	0.01874
X2 **X2**	0.45549	0.05607	1.00000	0.64534	0.07247	-0.02981
X3 **X3**	0.64572	0.68896	0.64534	1.00000	0.15883	0.00667
X4 **X4**	0.41047	0.14553	0.07247	0.15883	1.00000	0.11266
X5 **X5**	0.34706	0.01874	-0.02981	0.00667	0.11266	1.00000

Modèle linéaire simple pour l'exemple fictif

- La corrélation entre Y et chaque variable explicative est significative et positive.
- Par conséquent, si on ajustait séparément les cinq modèles avec une seule variable explicative à la fois, le paramètre de la variable serait significatif et positif à chaque fois. Ceci est cohérent avec le vrai modèle qui a généré les données.
- Ceci démontre aussi qu'il y a assez d'observations pour bien estimer les paramètres et avoir les bonnes conclusions quant à leurs effets, du moins lorsque qu'on les considère un à la fois.

Exemple fictif de multicolinéarité

- En revanche, X_1, X_2 et X_3 sont très corrélées entre elles, ce qui peut causer un problème de multicolinéarité.
- Ajustons d'abord le modèle contenant toutes les variables explicatives avec `proc reg` tout en demandant les diagnostics de multicolinéarité.

Code SAS pour calculer les facteurs d'inflation de la variance

```
proc reg data=infe.simcolineaire;
model y=x1-x5 / vif;
run;

proc glm data=infe.simcolineaire;
model y=x1-x5 / ss3 solution tolerance;
run;
```

Procédures `reg` versus `glm` en SAS

La procédure `reg` permet également d'ajuster des modèles linéaires dans SAS.

- Par défaut, les graphiques sont imprimés (option `plots=diagnostics` dans `glm`).
- La procédure `reg` a des fonctionnalités pour la sélection de modèle (pas utile en inférence).
- Le tableau des coefficients est également imprimé (option `solution` dans `glm`).
- En revanche, la procédure `reg` ne permet pas d'inclure des variables catégorielles: ces dernières doivent **obligatoirement** être encodées à l'aide de variables indicatrices binaires (`0-1`) (**erreur fréquente**).
- Dans SAS, on peut utiliser l'option `vif` dans la procédure `reg` ou `tol` (réciproque du VIF) avec les procédures `reg` et `glm`.

Estimés des paramètres et des VIF

Résultats estimés des paramètres

Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Inflation de variance
Intercept	Intercept	1	-0.76110	2.43241	-0.31	0.7551	0
X1	X1	1	0.43149	0.45829	0.94	0.3488	3.75609
X2	X2	1	0.68894	0.45638	1.51	0.1345	3.38306
X3	X3	1	1.94048	0.77306	2.51	0.0138	6.42789
X4	X4	1	1.06329	0.24587	4.32	<.0001	1.04162
X5	X5	1	1.14430	0.23231	4.93	<.0001	1.01507

Variable dépendante : Y

Tolérances

Variable	Tolérance de Type I	Tolérance de Type II
Intercept	100	6.3051461638
X1	1	0.2662342154
X2	0.9968564885	0.2955903718
X3	0.1560669286	0.1555721533
X4	0.9722559013	0.9600474856
X5	0.9851577945	0.9851577945

Adéquation pour l'exemple fictif de multicolinéarité

- Dans l'ensemble, le modèle semble adéquat; le R^2 est de 62%.
- En revance, les coefficients X_1 et X_2 ne sont pas significatifs une fois les autres variables prises en compte.
- Les facteurs d'inflation de la variance VIF de X_3 est grand (6.43) et ceux de X_1 et X_2 oscillent entre 3 et 4.
- Ceci indique également un problème potentiel de multicolinéarité. La précision dans l'estimation de ces paramètres n'est pas aussi bonne que s'il n'y avait pas de multicolinéarité.
- Notez que le VIF est une mesure individuelle. Elle ne nous dit pas quelles variables sont corrélées entre elles.