MATH60604
Modélisation statistique
§ 2e Coefficient de détermination

Léo Belzile

HEC Montréal
Département de sciences de la décision

Coefficient de corrélation linéaire de Pearson

- Le coefficient de corrélation linéaire quantifie la force de la relation **linéaire** entre deux variables X et Y.
- Supposons que l'on étudie n couples d'observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, où (X_i, Y_i) sont les valeurs X et Y pour le sujet i.
- Le coefficient de corrélation de Pearson est

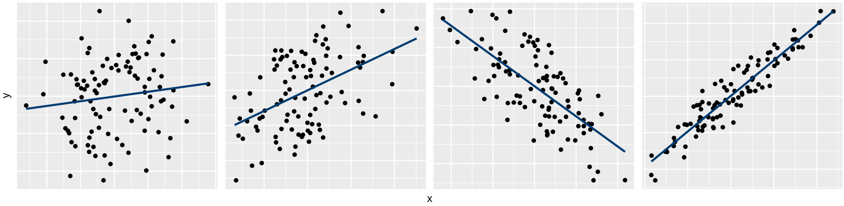
$$r = \frac{\widehat{\mathsf{Co}}(X, Y)}{\sqrt{\widehat{\mathsf{Va}}(X)\widehat{\mathsf{Va}}(Y)}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}},$$

où $\overline{X}$ et $\overline{Y}$ sont les moyennes empiriques de X et Y.

Propriétés du coefficient de corrélation linéaire de Pearson

Propriétés du coefficient de corrélation linéaire de Pearson

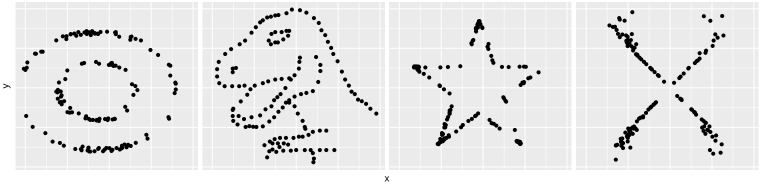
- $-1 \leq r \leq 1$
- $r = 1$ ($r = -1$) si et seulement si les n points sont alignés sur une droite de pente positive (négative). En d'autres termes, il existe deux constantes a et $b > 0$ ($b < 0$) telles que $y_i = a + bx_i$ pour tout i.



De gauche à droite, la corrélation linéaire est 0.1, 0.5, -0.75 et 0.95.

Coefficient de corrélation linéaire de Pearson

- Si $r > 0$ ($r < 0$), les deux variables sont positivement (négativement) associées, ce qui veut dire que Y augmente (diminue) en moyenne si X augmente.
- Plus $|r|$ est près de 1, moins les points sont éparpillés.
- Deux variables indépendantes sont non corrélées (l'inverse est faux).
- Une corrélation linéaire de zéro n'implique pas qu'il n'y a pas de relation entre les deux variables. Cela veut uniquement dire qu'il n'y a pas de dépendance **linéaire** entre les variables.



Les quatres jeux de données (cible, Anscombosaurus, étoile, croix) ont une corrélation linéaire identique de -0.06, mais les variables ne sont clairement pas indépendantes.

Coefficient de détermination

- Une fois le modèle linéaire ajusté, il peut être utile d'avoir un résumé qui permet de quantifier la qualité de l'ajustement.
- Le coefficient de détermination, R^2, mesure la force de la relation linéaire entre $\widehat{Y}$ et Y.
- Il représente la proportion de la variabilité de Y expliquée par les $\mathbf{X}$.
- R^2 est le carré du coefficient de corrélation entre les valeurs ajustées et la réponse, $(\widehat{Y}_1, Y_1), \ldots, (\widehat{Y}_n, Y_n)$.

Décomposition de la somme des carrés

- Supposons qu'aucune variable explicative n'est incluse dans le modèle (seulement l'ordonnée à l'origine). Dans ce cas, la valeur ajustée de Y pour toutes les observations est la moyenne empirique et la somme du carré des observations centrées est

$$\mathsf{SS}_c = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

- Si le modèle inclut les régresseurs $\mathbf{X}$, la valeur ajustée de Y_i est plutôt $\widehat{Y}_i = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_1 \mathsf{X}_{ij}$ et la somme du carré des résidus du modèle est

$$\mathsf{SS}_e = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

- La valeur de SS_e ne croît jamais quand on ajoute des variables.

Coefficient de détermination (R^2)

- R^2 mesure la proportion de la variance de Y expliquée par les régresseurs $\mathrm{X}_1, \ldots, \mathrm{X}_p$,

$$R^2 = \frac{\mathrm{SS}_c - \mathrm{SS}_e}{\mathrm{SS}_c}.$$

- Quand il y a plus d'une variable explicative, la racine carrée de R^2 est appelé coefficient de corrélation multiple.
- R^2 prend des valeurs entre 0 et 1

Interprétation du coefficient de détermination

R-carré	Coef de var	Racine MSE	intention Moyenne
0.449726	27.41959	2.264401	8.258333

- Pour le modèle qui inclut toutes les variables explicatives, $R^2 = 0.45$. Combinées, les variables expliquent 45% de la variabilité d'`intention`.
- Pour le modèle de régression linéaire simple avec `fixation` pour tout régresseur, $R^2 = 0.182$. Cela veut dire que la variable `fixation` explique 18.2% de la variabilité d'`intention`.

Avertissement au sujet de l'utilisation de R^2

- **Avertissement**: plus le nombre de régresseurs inclus est élevé, plus R^2 est grand (même si ces variables sont superflues à des fins d'inférence ou de prédictions).
- R^2 n'est donc pas un bon critère d'adéquation.
- Les logiciels rapportent parfois le coefficient de détermination ajusté, qui inclut une pénalité,

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

Le coefficient perd en interprétabilité et peut être négatif.