SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

20 mars 2024

Plan de la séance

Plan de la séance

- Régression logistique
- Confondeur

Récap et matière à réfléxion

Récap et matière à réfléxion

Base de données “Pima Indian Diabetes”

- Variable dépendante : `Outcome`
- Variables indépendantes : `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction` et `Age`

Récap et matière à réfléxion

Base de données “Pima Indian Diabetes”

- Variable dépendante : `Outcome`
- Variables indépendantes : `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction` et `Age`

R code

```
data <- read.csv("diabetes.csv")
selected_columns <- data[, 2:6]
rows_with_zero <- apply(selected_columns, 1,
                        function(x) any(x == 0))
data_cleaned <- data[!rows_with_zero, ]
names(data_cleaned)[
  names(data_cleaned) ==
    "DiabetesPedigreeFunction"] <- "DbtPdgFunc"
```

Recap : Modèle complet

```
model_full <- lm(Outcome ~ Pregnancies + Glucose +
                   BloodPressure + SkinThickness +
                   Insulin + BMI + DbtPdgFunc + Age,
                 data = data_cleaned)
round(summary(model_full)$coefficients, 4)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.1027     0.1436 -7.6806   0.0000
## Pregnancies       0.0130     0.0084  1.5486   0.1223
## Glucose           0.0064     0.0008  7.8550   0.0000
## BloodPressure     0.0001     0.0017  0.0316   0.9748
## SkinThickness     0.0017     0.0025  0.6652   0.5063
## Insulin          -0.0001     0.0002 -0.6031   0.5468
## BMI               0.0093     0.0039  2.3907   0.0173
## DbtPdgFunc        0.1572     0.0580  2.7083   0.0071
## Age               0.0059     0.0028  2.1090   0.0356
```

Recap : Modèle ajusté

En supprimant les variables non importantes `BloodPressure`, `SkinThickness` et `Insulin` ...

```
model_reduced <- lm(Outcome ~ Pregnancies + Glucose
                      BMI + DbtPdgFunc + Age,
                    data = data_cleaned)
round(summary(model_reduced)$coefficients, 4)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.0908     0.1174 -9.2876   0.0000
## Pregnancies      0.0136     0.0083  1.6387   0.1021
## Glucose          0.0062     0.0007  8.9698   0.0000
## BMI              0.0108     0.0029  3.7636   0.0002
## DbtPdgFunc       0.1578     0.0574  2.7483   0.0063
## Age              0.0059     0.0027  2.1739   0.0303
```

Recap : Comparaison des modèles par R^2 et R^2 ajusté

```
summary(model_full)$r.squared
```

```
## [1] 0.3457734
```

```
summary(model_reduced)$r.squared
```

```
## [1] 0.3443796
```

Recap : Comparaison des modèles par R^2 et R^2 ajusté

```
summary(model_full)$r.squared
```

```
## [1] 0.3457734
```

```
summary(model_reduced)$r.squared
```

```
## [1] 0.3443796
```

```
summary(model_full)$adj.r.squared
```

```
## [1] 0.3321081
```

```
summary(model_reduced)$adj.r.squared
```

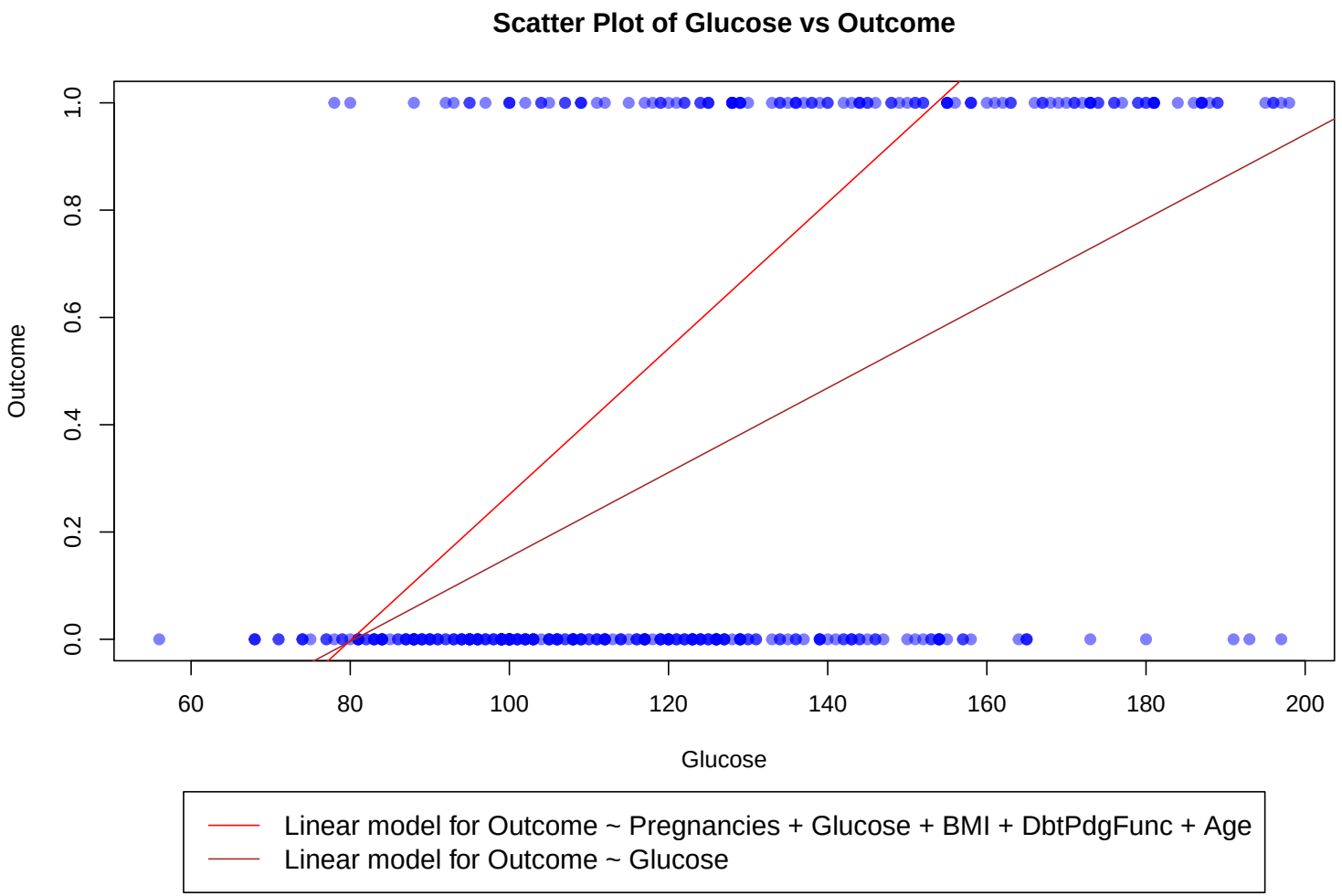
```
## [1] 0.3358872
```

Recap : Intervalles de confiance de β's

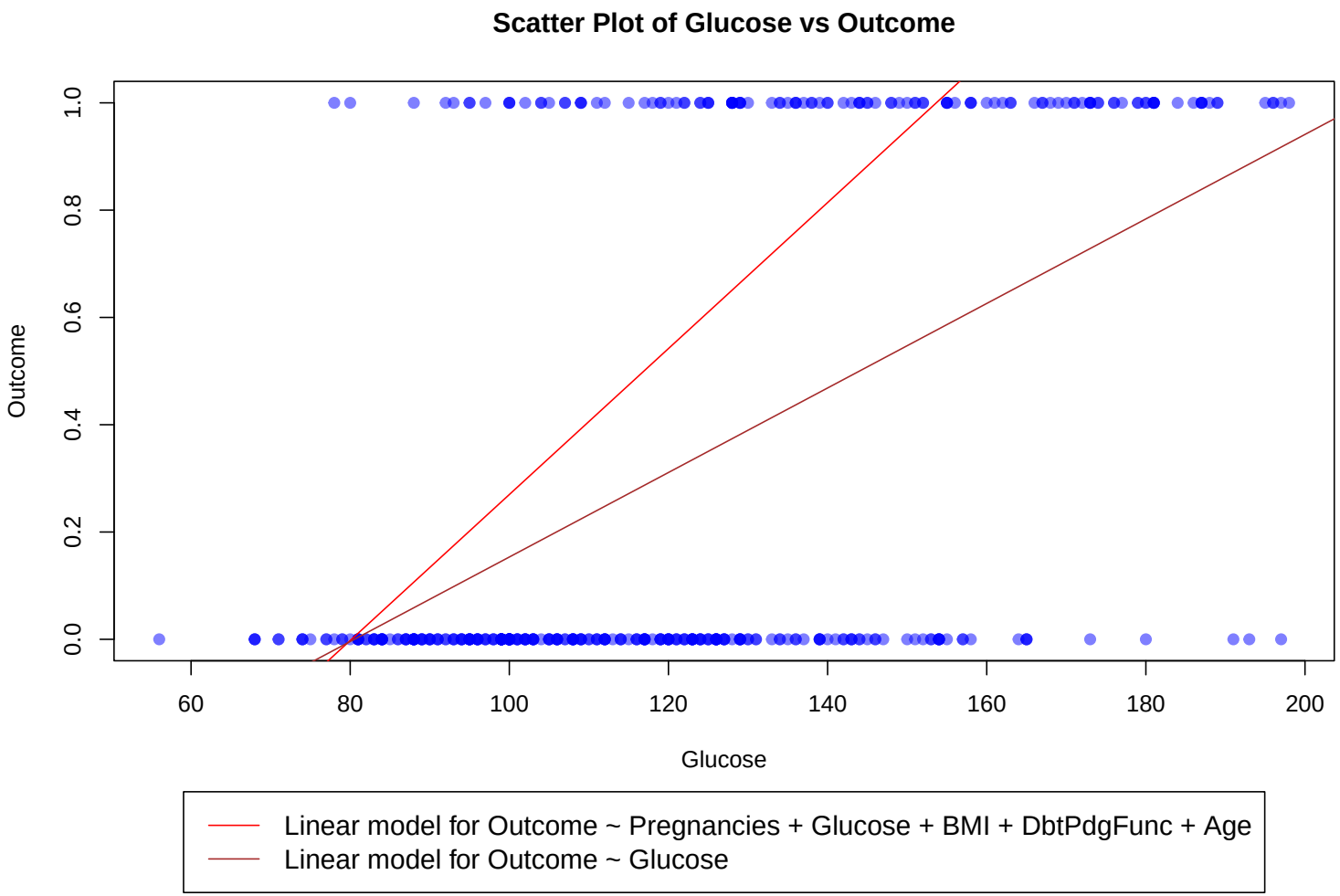
```
round(confint(model_reduced, level = 0.95), 4)
```

```
##                 2.5 %  97.5 %
## (Intercept) -1.3217 -0.8599
## Pregnancies -0.0027  0.0299
## Glucose      0.0048  0.0075
## BMI          0.0051  0.0164
## DbtPdgFunc   0.0449  0.2708
## Age          0.0006  0.0113
```

Recap : Visualisation des résultats



Recap : Visualisation des résultats



La réponse (`y` ou `Outcome`) n'est pas une variable continues, mais binaire, soit 0 ou 1 et non une valeur intermédiare.

Attention!

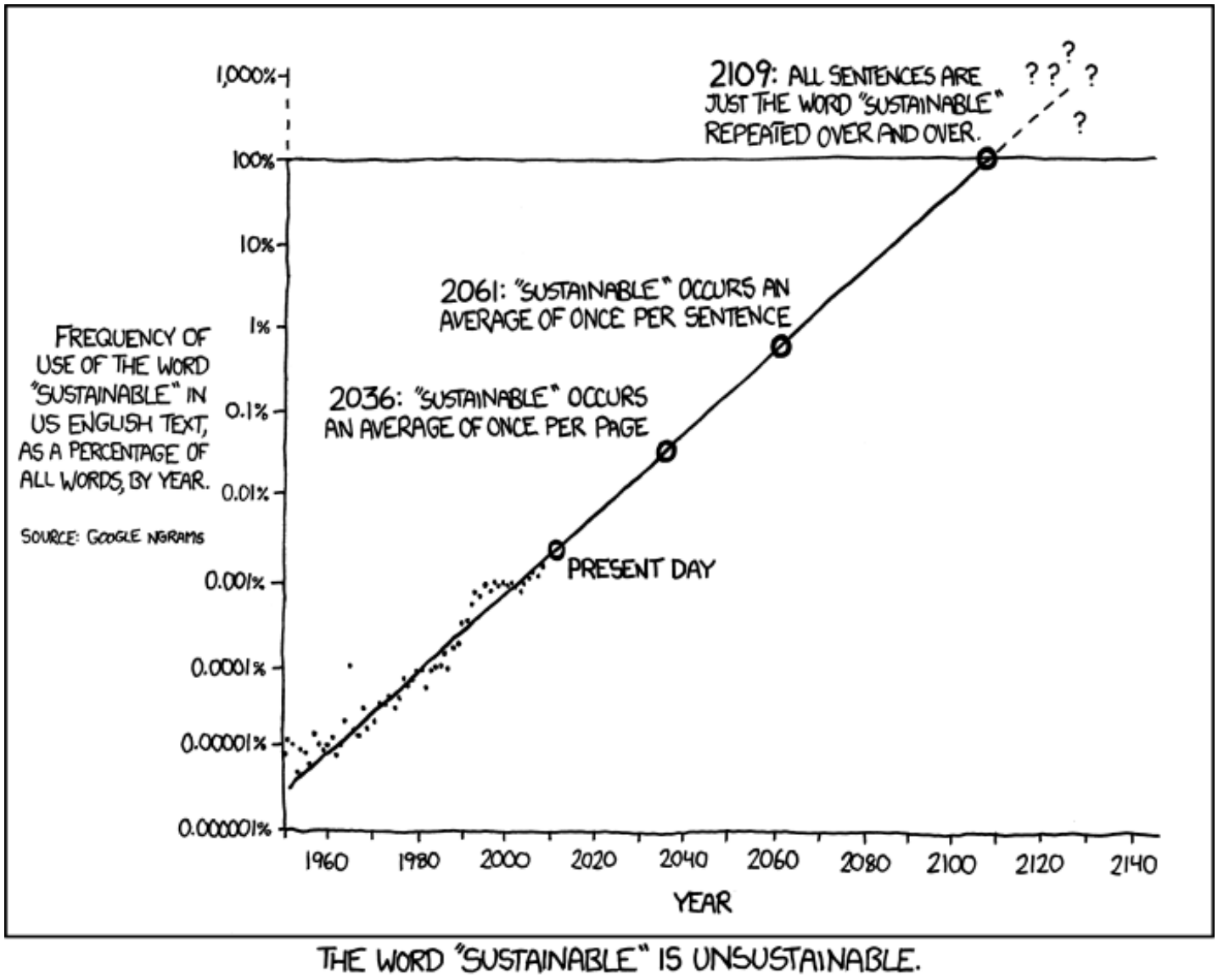
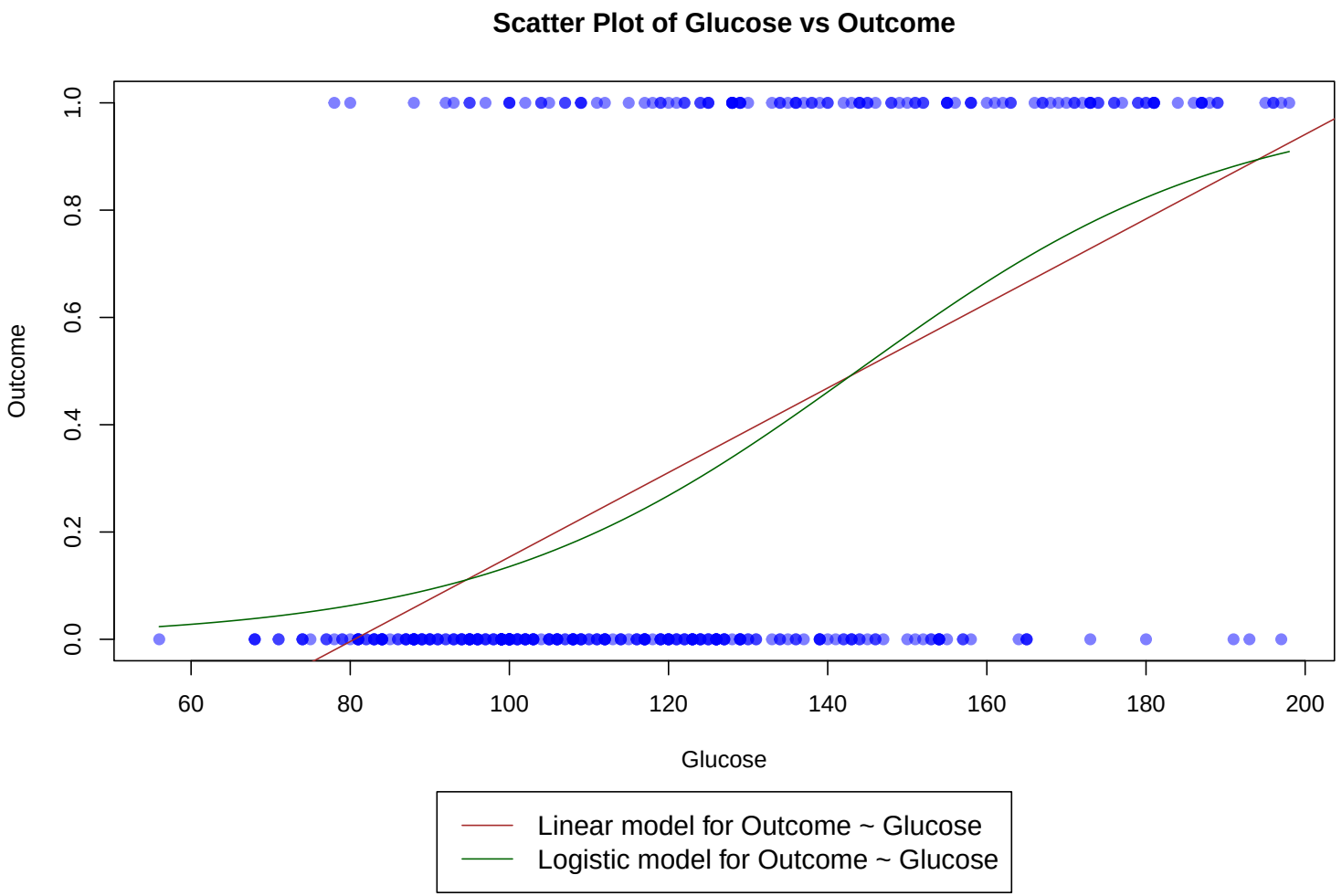


Figure 1: Extrapolation - "Sustainable is unsustainable."

Recap : Visualisation des résultats



Une alternative est la régression logistique, fournissant un résultat sous forme de **probabilité** que y soit 0 ou 1.

Introduction au nouveau sujet

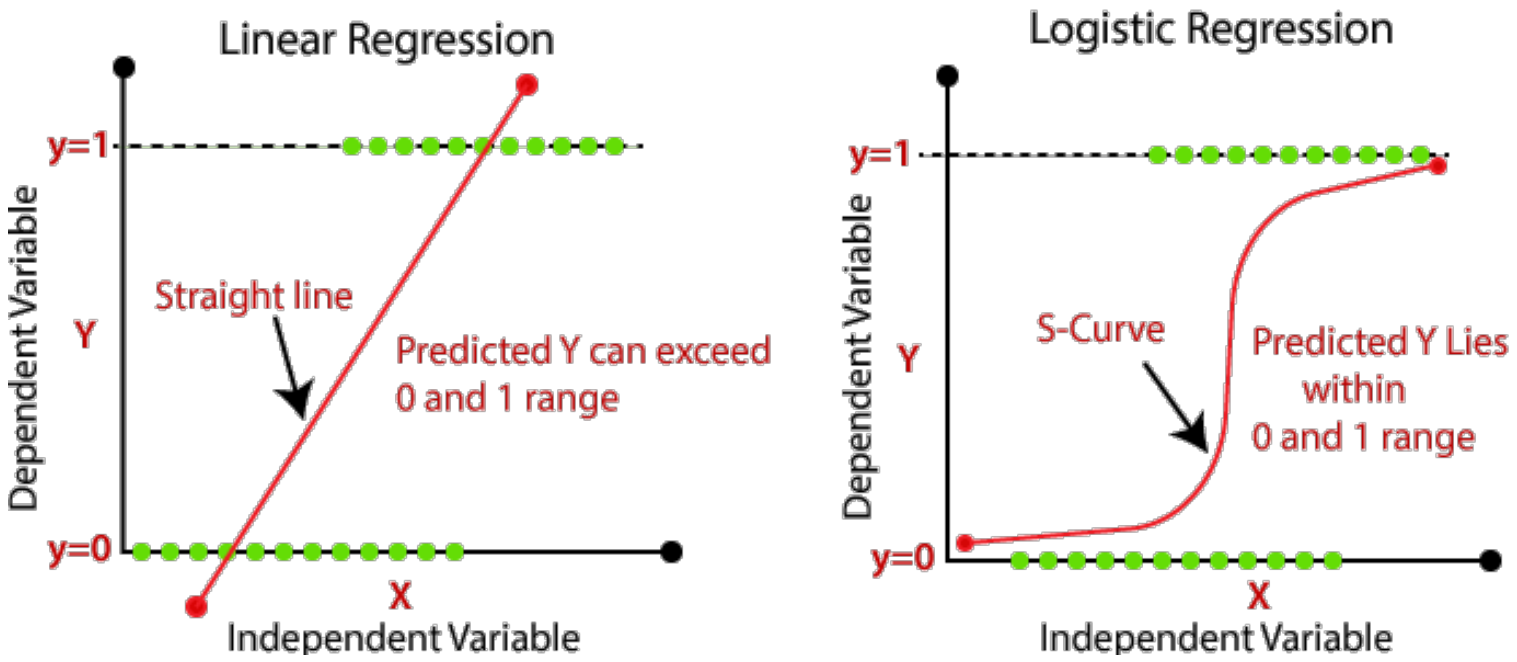


Figure 2: Régression linéaire vs. logistique

Régression logistique

Régression logistique : Modèle

La régression logistique modélise la probabilité d'un résultat binaire basée sur une ou plusieurs variables prédictives. Cela est particulièrement utile lorsque la variable dépendante ne peut prendre que deux résultats possibles (succès ou échec).

Régression logistique : Modèle

La régression logistique modélise la probabilité d'un résultat binaire basée sur une ou plusieurs variables prédictives. Cela est particulièrement utile lorsque la variable dépendante ne peut prendre que deux résultats possibles (succès ou échec).

Le modèle de régression logistique est basé sur **la fonction logit, le logarithme naturel du rapport de cotes**.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

- p est la probabilité d'une des issues (réponses),
- $X_1, X_2, ..., X_k$ sont les variables prédictives.
- $\beta_1, \beta_2, ..., \beta_k$ représentent le changement dans le log des cotes de l'issue pour un changement unitaire dans les variables prédictives.

Régression logistique : Inférence

Inférence sur les Coefficients : Les tests d'hypothèse sur $\beta_1, \beta_2, ..., \beta_k$ sont réalisées pour déterminer si les prédicteurs sont significativement associés à l'issue.

Régression logistique : Inférence

Inférence sur les Coefficients : Les tests d'hypothèse sur $\beta_1, \beta_2, ..., \beta_k$ sont réalisées pour déterminer si les prédicteurs sont significativement associés à l'issue.

Méthode d'Estimation : Les coefficients sont estimés en utilisant l'Estimation du Maximum de Vraisemblance (MLE) afin de trouver les coefficients qui maximisent la vraisemblance d'observer les données de l'échantillon.

Régression logistique : Inférence

Inférence sur les Coefficients : Les tests d'hypothèse sur $\beta_1, \beta_2, ..., \beta_k$ sont réalisées pour déterminer si les prédicteurs sont significativement associés à l'issue.

Méthode d'Estimation : Les coefficients sont estimés en utilisant l'Estimation du Maximum de Vraisemblance (MLE) afin de trouver les coefficients qui maximisent la vraisemblance d'observer les données de l'échantillon.

Interprétation en Rapport de Cotes : Un rapport de cotes supérieur à 1 indique une augmentation des cotes de l'issue avec une augmentation unitaire du prédicteur, et vice versa.

Régression logistique : Choix des variables

Guides généraux

- Commencez avec un **cadre théorique** ou des recherches antérieures pour identifier les prédicteurs potentiels.

Régression logistique : Choix des variables

Guides généraux

- Commencez avec un **cadre théorique** ou des recherches antérieures pour identifier les prédicteurs potentiels.
- Prenez en compte la **signification statistique** des variables dans les analyses préliminaires.

Régression logistique : Choix des variables

Guides généraux

- Commencez avec un **cadre théorique** ou des recherches antérieures pour identifier les prédicteurs potentiels.
- Prenez en compte la **signification statistique** des variables dans les analyses préliminaires.
- Vérifiez la **multicollinéarité** parmi les prédicteurs, car une forte collinéarité peut déformer l'estimation et l'interprétation des coefficients.

Régression logistique : Choix des variables

Guides généraux

- Commencez avec un **cadre théorique** ou des recherches antérieures pour identifier les prédicteurs potentiels.
- Prenez en compte la **signification statistique** des variables dans les analyses préliminaires.
- Vérifiez la **multicollinéarité** parmi les prédicteurs, car une forte collinéarité peut déformer l'estimation et l'interprétation des coefficients.
- Évitez d'inclure trop de variables, surtout dans de petits ensembles de données, pour prévenir le **surajustement**.

Régression logistique : Choix des variables

Guides généraux

- Commencez avec un **cadre théorique** ou des recherches antérieures pour identifier les prédicteurs potentiels.
- Prenez en compte la **signification statistique** des variables dans les analyses préliminaires.
- Vérifiez la **multicollinéarité** parmi les prédicteurs, car une forte collinéarité peut déformer l'estimation et l'interprétation des coefficients.
- Évitez d'inclure trop de variables, surtout dans de petits ensembles de données, pour prévenir le **surajustement**.

Méthodes de sélection séquentielles

- Ajouter ou retirer des prédicteurs basés sur des critères tels que l'AIC ou le BIC.

Critère d'Information d'Akaike (AIC)

L'AIC est une mesure de la qualité **relative** d'un modèle statistique pour un ensemble de données, et basé sur le concept d'entropie d'information.

Critère d'Information d'Akaike (AIC)

L'AIC est une mesure de la qualité **relative** d'un modèle statistique pour un ensemble de données, et basé sur le concept d'entropie d'information.

- $AIC = 2k - 2\ln(L)$
  - k est le nombre de paramètres dans le modèle et
  - L est la vraisemblance du modèle.

L'AIC pénalise les modèles pour leur complexité (nombre de paramètres), aidant ainsi à éviter le surajustement.

Critère d'Information d'Akaike (AIC)

L'AIC est une mesure de la qualité **relative** d'un modèle statistique pour un ensemble de données, et basé sur le concept d'entropie d'information.

- $AIC = 2k - 2\ln(L)$
  - k est le nombre de paramètres dans le modèle et
  - L est la vraisemblance du modèle.

L'AIC pénalise les modèles pour leur complexité (nombre de paramètres), aidant ainsi à éviter le surajustement.

- Une valeur AIC plus basse indique un meilleur modèle.
- Lors de la comparaison de modèles, la valeur absolue de l'AIC n'est pas aussi importante que la différence entre les valeurs AIC de différents modèles.
- Des modèles avec un AIC différant de plus de 2 sont généralement considérés comme ayant des preuves substantielles contre le modèle avec l'AIC le plus élevé.

Critère d'Information Bayésien (BIC)

Le BIC, dérivé de la probabilité bayésienne, introduit une pénalité plus forte pour le nombre de paramètres dans le modèle.

Critère d'Information Bayésien (BIC)

Le BIC, dérivé de la probabilité bayésienne, introduit une pénalité plus forte pour le nombre de paramètres dans le modèle.

- $BIC = \ln(n)k - 2\ln(L)$
  - n est le nombre d'observations,
  - k est le nombre de paramètres, et
  - L est la vraisemblance du modèle.

Le BIC a tendance à pénaliser plus lourdement la complexité que l'AIC, surtout à mesure que la taille de l'échantillon augmente.

Critère d'Information Bayésien (BIC)

Le BIC, dérivé de la probabilité bayésienne, introduit une pénalité plus forte pour le nombre de paramètres dans le modèle.

- $BIC = \ln(n)k - 2\ln(L)$
  - n est le nombre d'observations,
  - k est le nombre de paramètres, et
  - L est la vraisemblance du modèle.

Le BIC a tendance à pénaliser plus lourdement la complexité que l'AIC, surtout à mesure que la taille de l'échantillon augmente.

- Une valeur BIC plus basse indique un meilleur modèle.
- La règle de décision pour comparer les modèles avec le BIC est similaire à l'AIC.
- Une différence de 6 ou plus est considérée comme une preuve forte contre le modèle avec le BIC le plus élevé.

R code

```r
AIC(logistic_model)
```

```
## [1] 390.666
```

```r
BIC(logistic_model)
```

```
## [1] 398.6085
```

```r
# calcul du BIC par
# la fonction AIC avec l'argument k = log(n)
AIC(logistic_model, k = log(nrow(data_cleaned)))
```

```
## [1] 398.6085
```

AIC vs. BIC

L'AIC se concentre davantage sur l'adéquation (goodness of fit) du modèle. Il est issu de la théorie de l'information et vise à choisir un modèle qui explique le mieux les données, même s'il comprend plus de paramètres.

Le BIC est dérivé de la probabilité bayésienne et est plus concerné par l'identification du vrai modèle parmi l'ensemble des candidats. Il part du principe qu'il existe un vrai modèle et tente de s'en rapprocher.

AIC vs. BIC

L'AIC se concentre davantage sur l'adéquation (goodness of fit) du modèle. Il est issu de la théorie de l'information et vise à choisir un modèle qui explique le mieux les données, même s'il comprend plus de paramètres.

Le BIC est dérivé de la probabilité bayésienne et est plus concerné par l'identification du vrai modèle parmi l'ensemble des candidats. Il part du principe qu'il existe un vrai modèle et tente de s'en rapprocher.

Différences clés dans l'utilisation

- Complexité : L'AIC peut sélectionner des modèles plus complexes, tandis que le BIC a tendance à favoriser des modèles plus simples.
- But : L'AIC est mieux adapté aux modèles axés sur la prédiction, tandis que le BIC est plus approprié pour les modèles visant à expliquer la structure sous-jacente.

Démarches de sélection des variables de modèle

Exemple : Sélection progressive

Examiner la corrélation de toutes les variables indépendantes avec la variable dépendante `Outcome`

```
## [1] "Pregnancies : 0.2566"
## [1] "Glucose : 0.5157"
## [1] "BloodPressure : 0.1927"
## [1] "SkinThickness : 0.2559"
## [1] "Insulin : 0.3014"
## [1] "BMI : 0.2701"
## [1] "DbtPdgFunc : 0.2093"
## [1] "Age : 0.3508"
## [1] "Outcome : 1"
```

- Ajouter d'abord `Glucose`, puis `Age`, `Insulin`, `BMI`, ... etc.

Exemple : Sélection progressive

```
model1 <- glm(Outcome ~ Glucose,
              data=data_cleaned, family=binomial)
AIC(model1)
```

```
## [1] 390.666
```

Exemple : Sélection progressive

```
model1 <- glm(Outcome ~ Glucose,
              data=data_cleaned, family=binomial)
AIC(model1)
```

```
## [1] 390.666
```

```
model2 <- glm(Outcome ~ Glucose + Age,
              data=data_cleaned, family=binomial)
AIC(model2)
```

```
## [1] 376.6897
```

- Garder `model2` et continuer à ajouter des variables

Exemple : Sélection progressive

```
model3 <- glm(Outcome ~ Glucose + Age + Insulin,
              data=data_cleaned, family=binomial)
AIC(model3)
```

```
## [1] 378.6714
```

- Retourner au model2 et ne pas inclure Insulin

Exemple : Sélection progressive

```
model3 <- glm(Outcome ~ Glucose + Age + Insulin,
              data=data_cleaned, family=binomial)
AIC(model3)
```

```
## [1] 378.6714
```

- Retourner au model2 et ne pas inclure Insulin
- Essayer la prochaine variable BMI

```
model4 <- glm(Outcome ~ Glucose + Age + BMI,
              data=data_cleaned, family=binomial)
AIC(model4)
```

```
## [1] 362.3656
```

- Garder model4 et continuer à ajouter des variables

Exemple : Sélection progressive

Le reste du processus sera consacré aux travaux pratiques.

Modèle à jour

```
round(summary(model4)$coefficients, 4)
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.6773     1.0419 -9.2884    0e+00
## Glucose       0.0363     0.0049  7.3913    0e+00
## Age           0.0541     0.0132  4.0854    0e+00
## BMI           0.0779     0.0201  3.8697    1e-04
```

Exemple : Sélection régressive

Commencer par une modèle complet

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -10.0407     1.2177 -8.2458   0.0000
## Pregnancies     0.0822     0.0554  1.4823   0.1383
## Glucose         0.0383     0.0058  6.6351   0.0000
## BloodPressure  -0.0014     0.0118 -0.1200   0.9045
## SkinThickness   0.0112     0.0171  0.6568   0.5113
## Insulin        -0.0008     0.0013 -0.6317   0.5276
## BMI             0.0705     0.0273  2.5798   0.0099
## DbtPdgFunc      1.1409     0.4274  2.6692   0.0076
## Age             0.0340     0.0184  1.8470   0.0647
```

- Éliminer `BloodPressure`, `SkinThickness`, `Insulin`, `Pregnancies`, ... une variable à la fois

Exemple : Sélection régressive

```r
AIC(modelf) # modèle complète
```

```
## [1] 362.0212
```

Exemple : Sélection régressive

```
AIC(modelf) # modèle complète
```

```
## [1] 362.0212
```

```
model5 <- glm(Outcome ~ Pregnancies + Glucose +
                SkinThickness + Insulin + BMI +
                DbtPdgFunc + Age,
              data=data_cleaned, family=binomial)
AIC(model5)
```

```
## [1] 360.0356
```

- Ne pas garder `model5` (ne pas retirer `BloodPressure`), mais essayer de retirer les autres variables

Exemple : Sélection régressive

```
model6 <- glm(Outcome ~ Pregnancies + Glucose +
                BloodPressure + Insulin + BMI +
                DbtPdgFunc + Age,
              data=data_cleaned, family=binomial)
AIC(model6)
```

```
## [1] 360.452
```

Exemple : Sélection régressive

```
model6 <- glm(Outcome ~ Pregnancies + Glucose +
                BloodPressure + Insulin + BMI +
                DbtPdgFunc + Age,
              data=data_cleaned, family=binomial)
AIC(model6)
```

```
## [1] 360.452
```

```
model7 <- glm(Outcome ~ Pregnancies + Glucose +
                BloodPressure + SkinThickness +
                BMI + DbtPdgFunc + Age,
              data=data_cleaned, family=binomial)
AIC(model7)
```

```
## [1] 360.4183
```

- Ne retirer ni SkinThickness ni Insulin

Exemple : Sélection régressive

Mais, si on essayait de retirer ces trois variables (`BloodPressure`, `SkinThickness` et `Insulin`) en même temps . . .

```
model8 <- glm(Outcome ~ Pregnancies + Glucose +
                BMI + DbtPdgFunc + Age,
              data=data_cleaned, family=binomial)
AIC(model8)
```

```
## [1] 356.8851
```

- Selon l'AIC, le modèle s'améliore.

Le reste du processus sera consacré aux travaux pratiques.

Présentation des modèles

Les **intervalles de confiance** des variables statistiquement significatives ne couvrent pas 0.

```
## Waiting for profiling to be done...

##                    2.5 %  97.5 %
## (Intercept)   -12.5490 -7.7614
## Pregnancies    -0.0260  0.1920
## Glucose         0.0273  0.0500
## BloodPressure  -0.0245  0.0221
## SkinThickness  -0.0224  0.0448
## Insulin        -0.0034  0.0018
## BMI             0.0177  0.1253
## DbtPdgFunc      0.3209  1.9972
## Age            -0.0015  0.0709
```

Récap et matière à réfléxion

Multicollinéarité

Corrélations modérées entre variables indépendantes

```
cor(data_cleaned$Pregnancies, data_cleaned$Age)
```

```
## [1] 0.6796085
```

```
cor(data_cleaned$Glucose, data_cleaned$Insulin)
```

```
## [1] 0.581223
```

```
cor(data_cleaned$SkinThickness, data_cleaned$BMI)
```

```
## [1] 0.6643549
```

Comparaison des modèles

```
model2 <- glm(Outcome ~ Glucose + Age,
              data=data_cleaned, family=binomial)
AIC(model2)
```

```
## [1] 376.6897
```

```
model3 <- glm(Outcome ~ Glucose + Age + Insulin,
              data=data_cleaned, family=binomial)
AIC(model3)
```

```
## [1] 378.6714
```

Pourquoi l'ajout de la variable 'Insulin' augmente-t-il l'AIC ?

Comparaison des modèles

```
model2 <- glm(Outcome ~ Glucose + Age,
              data=data_cleaned, family=binomial)
AIC(model2)
```

```
## [1] 376.6897
```

```
model3 <- glm(Outcome ~ Glucose + Age + Insulin,
              data=data_cleaned, family=binomial)
AIC(model3)
```

```
## [1] 378.6714
```

Pourquoi l'ajout de la variable 'Insulin' augmente-t-il l'AIC ?

- Il y a déjà la variable 'Glucose'.

Sélection des prédicteurs

Il faut tenir compte de la **connaissance du domaine**.

L'insuline est l'hormone responsable de la régulation du taux de glycémie (glucose dans le sang).

`Glucose` : la concentration de glucose plasmatique mesurée 2 heures après un test de tolérance au glucose oral.

`Insulin` : l'insuline sérique 2 heures après le début du test, en micro-unités par millilitre (mu U/ml).

Sélection des prédicteurs

Il faut tenir compte de la **connaissance du domaine**.

L'insuline est l'hormone responsable de la régulation du taux de glycémie (glucose dans le sang).

`Glucose` : la concentration de glucose plasmatique mesurée 2 heures après un test de tolérance au glucose oral.

`Insulin` : l'insuline sérique 2 heures après le début du test, en micro-unités par millilitre (mu U/ml).

`Glucose` et `Insulin` dans le processus métabolique

- Chez un individu en bonne santé, une augmentation du niveau de glucose déclenche la libération d'insuline.
- Mais, en cas de résistance à l'insuline (un précurseur du diabète), cette relation est perturbée.

Sélection des prédicteurs

Glycémie (`Glucose`) et Diabète (`Outcome`)

Il existe généralement une forte corrélation positive.

- Des niveaux élevés de glucose sont souvent indicatifs du diabète, car l'incapacité du corps à utiliser efficacement l'insuline entraîne une élévation du taux de glycémie.

Insuline (`Insulin`) et Diabète (`Outcome`)

Cette relation peut être plus complexe.

- Aux premiers stades du diabète de type 2, les niveaux d'insuline peuvent être élevés car le corps essaie de compenser l'augmentation de glycémie.
- Avec le temps (dans les stades avancés du diabète), le pancréas peut produire moins d'insuline, conduisant à des niveaux d'insuline plus faibles.

Introduction au nouveau sujet

Diabète : Le corps ne peut pas traiter correctement les niveaux de glucose dans le sang. Cela est dû soit à une production insuffisante d'insuline par le pancréas (diabète de type 1) soit à une utilisation inefficace de l'insuline produite (diabète de type 2 et cas de "Pima Indian Diabetes").

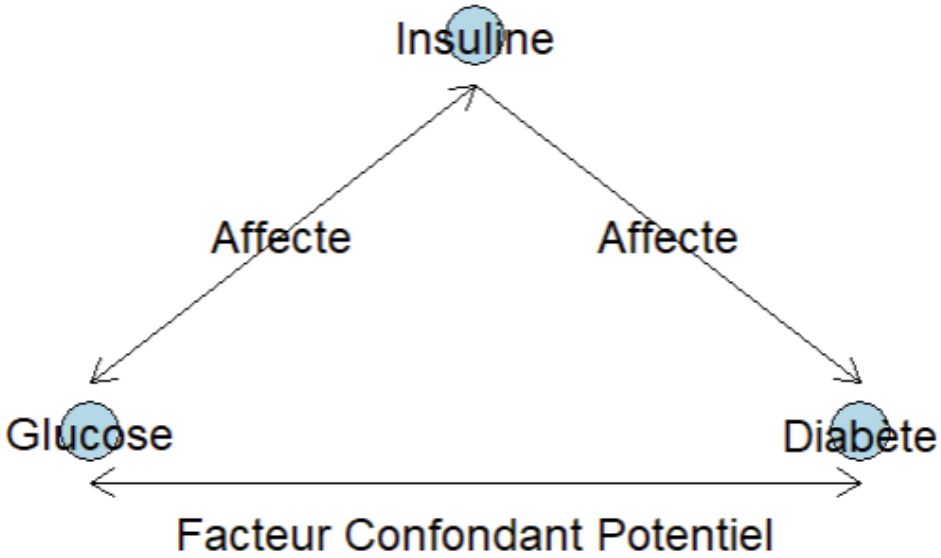


Figure 3: Confusion dans la relation

Confondeur

Confondeur

Un confondeur (ou facteur de confusion) est une variable qui influence à la fois la réponse et le prédicteur.

Ce facteur peut conduire à une interprétation trompeuse de la relation entre les variables étudiées car il affecte le résultat, mais il ne s'agit pas de la principale variable d'intérêt.

Confondeur

Un confondeur (ou facteur de confusion) est une variable qui influence à la fois la réponse et le prédicteur.

Ce facteur peut conduire à une interprétation trompeuse de la relation entre les variables étudiées car il affecte le résultat, mais il ne s'agit pas de la principale variable d'intérêt.

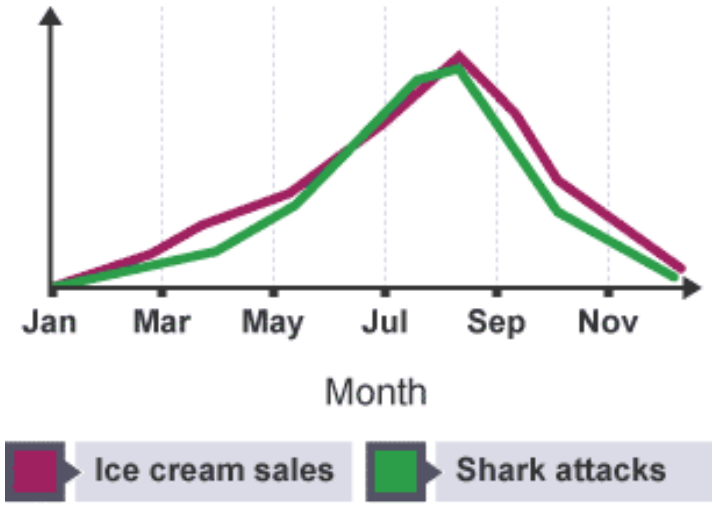


Figure 4: La vente de la glace a-t-elle un impact sur une attaque de requin ?

Confondeur

Sans tenir compte de l'augmentation de la température (confondeur), il peut y avoir une association fallacieuse.

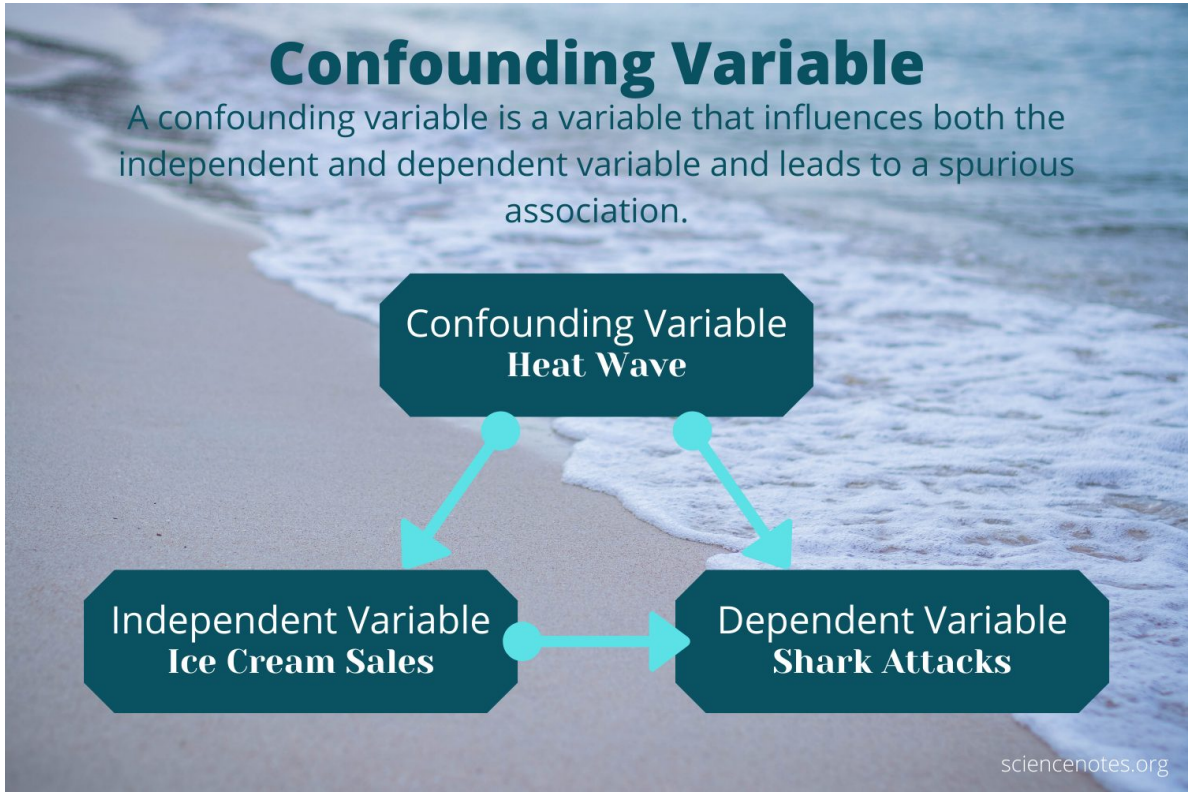


Figure 5: Confondeur dans la relation

D'autre exemples dans la base de données

Indice de Masse Corporelle (IMC) peut être influencé par des facteurs de mode de vie.

- Un IMC élevé pourrait suggérer un mode de vie incluant moins d'activité physique ou un régime alimentaire pouvant contribuer à la prise de poids, deux facteurs de risque du diabète.

D'autre exemples dans la base de données

Indice de Masse Corporelle (IMC) peut être influencé par des facteurs de mode de vie.

- Un IMC élevé pourrait suggérer un mode de vie incluant moins d'activité physique ou un régime alimentaire pouvant contribuer à la prise de poids, deux facteurs de risque du diabète.

Âge est un facteur connu du risque de diabète.

- Un âge plus avancé est associé à un risque accru de diabète de type 2, à cause des changements dans le métabolisme et éventuellement de mode de vie au fil du temps.

D'autre exemples dans la base de données

Indice de Masse Corporelle (IMC) peut être influencé par des facteurs de mode de vie.

- Un IMC élevé pourrait suggérer un mode de vie incluant moins d'activité physique ou un régime alimentaire pouvant contribuer à la prise de poids, deux facteurs de risque du diabète.

Âge est un facteur connu du risque de diabète.

- Un âge plus avancé est associé à un risque accru de diabète de type 2, à cause des changements dans le métabolisme et éventuellement de mode de vie au fil du temps.

Fonction du Pedigree du Diabète (ou prédisposition génétique au diabète) prend en compte l'histoire du diabète chez les proches et la relation génétique de ces proches avec le sujet.

Résumé

Il est crucial de toujours inclure les confondeurs connus dans le modèle (`BMI`, `Age` et `DiabetesPedigreeFunction`).

Mais, il faut être prudent ...

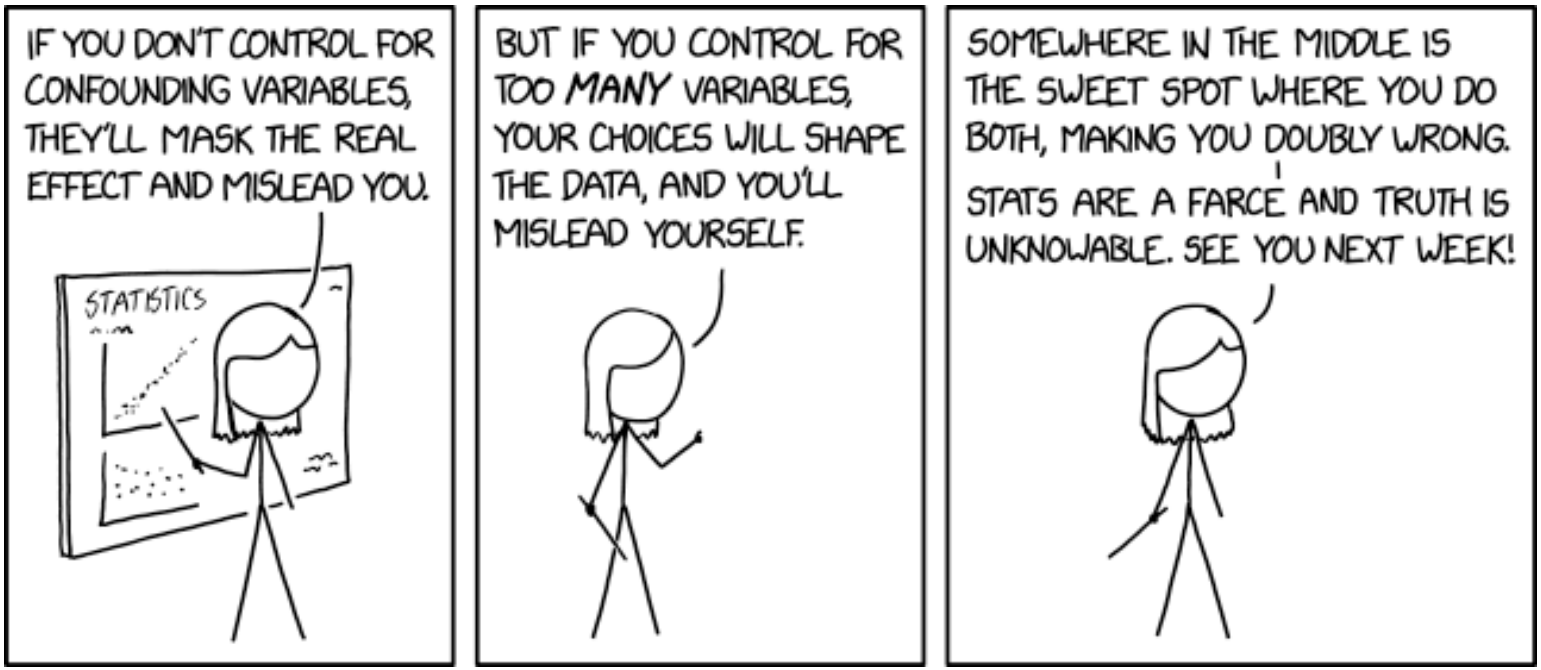


Figure 6: Rappel concernant les confondeurs

Erreurs d'identification des confondeurs

Ajuster trop de variables qui ne sont pas des confondeurs peut entraîner un **surajustement**, obscurcissant les vraies associations ou créant de fausses associations.

Erreurs d'identification des confondeurs

Ajuster trop de variables qui ne sont pas des confondeurs peut entraîner un **surajustement**, obscurcissant les vraies associations ou créant de fausses associations.

Attention à la **collinéarité**, où deux variables prédictives ou plus sont fortement corrélées. Ajuster l'une peut inadvertamment ajuster les autres, conduisant à des résultats trompeurs.

Erreurs d'identification des confondeurs

Ajuster trop de variables qui ne sont pas des confondeurs peut entraîner un **surajustement**, obscurcissant les vraies associations ou créant de fausses associations.

Attention à la **collinéarité**, où deux variables prédictives ou plus sont fortement corrélées. Ajuster l'une peut inadvertamment ajuster les autres, conduisant à des résultats trompeurs.

Il est important de **différencier un confondeur d'un modificateur d'effet** (interaction).

- Un modificateur d'effet change la direction ou la force de l'association entre l'exposition et le résultat selon ses niveaux.
- Un confondeur est une influence externe à contrôler.

Travaux pratiques

Travaux pratiques

Continuez à travailler avec la base de données “Pima Indian Diabetes”.

Avec `Outcome` comme variable dépendante, utilisez le reste des paramètres comme variables indépendantes pour créer un modèle de régression logistique.

Parmi les variables indépendantes, identifiez des confondeurs possibles et ajustez-les.

Employez les méthodes de sélection progressive et régressive.

Comparez différents modèles avec l'AIC et le BIC.

Quelles variables apportent une contribution importante au modèle (sont des bons prédicteurs pour la réponse `Outcome`) ?

Projet 2

Projet 2

Vous devez faire une présentation de votre projet, partager votre écran en expliquant et exécutant vos codes R devant vos collègues et interpréter les résultats.

La proposition du projet doit contenir les éléments suivants :

- Problématique
- Objectifs du projet
- Méthodologie
- Retombées prévues

La date limite pour la proposition est le 27 mars 2024.

Projet 2

De plus, vous devez compter les démarches suivantes :

- Sources des données
- Visualisation des données
  - Distribution des données de chaque variable
  - Relation parmi l'ensemble des variables

La date limite pour partager vos analyse exploratoire des données est le 3 avril 2024.

Projet 2

Dernièrement, vous allez tenir compte la rétroaction des vos collègues et la professeure afin de créer un bon modèle de régression de votre choix. Vous devez expliquer :

- Modèle de régression (linéaire, logistique ou d'autres)
- Interprétation des résultats
- Comparaison avec d'autres études
- Limitations de votre projet
- Impliquations pratiques

La date de la présentation finale est le 10 avril 2024.