SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

27 mars 2024

Récap

Récap

Régression linéaire

La régression linéaire modélise la façon dont la moyenne μ d'une variable de **réponse continue** Y est en fonction d'un ensemble de variables explicatives X.

$$\mu = E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Récap

Régression linéaire

La régression linéaire modélise la façon dont la moyenne μ d'une variable de **réponse continue** Y est en fonction d'un ensemble de variables explicatives X.

$$\mu = E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Régression logistique

La régression logistique modélise la manière dont les chances de « succès » p pour une variable de **réponse binaire** Y dépendent d'un ensemble de variables explicatives X.

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

Introduction du nouveau sujet

Modèles linéaires généralisés

« Generalized Linear Model (GLM) »

Pour les données qui ne suivent pas une distribution normale (comme le suppose la régression linéaire) ou une distribution binaire (comme dans la régression logistique), les modèles linéaires généralisés (GLM) offrent un cadre polyvalent grâce à l'utilisation d'autres distributions.

Modèles linéaires généralisés

Modèles linéaires généralisés

Il y a trois composants dans GLM.

- **Composant Aléatoire** : la distribution de probabilité de la variable de réponse (Y).
  - Il s'agit de la seule composante aléatoire du modèle; il n'y a pas de terme d'erreur distinct

Modèles linéaires généralisés

Il y a trois composants dans GLM.

- **Composant Aléatoire** : la distribution de probabilité de la variable de réponse (Y).
  - Il s'agit de la seule composante aléatoire du modèle; il n'y a pas de terme d'erreur distinct
- **Composant Systématique** : une combinaison linéaire de variables explicatives ou prédictives connues $(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$, où $\beta_0, \beta_1, ..., \beta_k$ sont des coefficients, et $X_1, X_2, ..., X_k$ sont des variables.

Modèles linéaires généralisés

Il y a trois composants dans GLM.

- **Composant Aléatoire** : la distribution de probabilité de la variable de réponse (Y).
  - Il s'agit de la seule composante aléatoire du modèle; il n'y a pas de terme d'erreur distinct
- **Composant Systématique** : une combinaison linéaire de variables explicatives ou prédictives connues $(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$, où $\beta_0, \beta_1, ..., \beta_k$ sont des coefficients, et $X_1, X_2, ..., X_k$ sont des variables.
- **Fonction de Lien** (η) : le lien entre les composants aléatoires et systématiques.
  - Le choix de la fonction de lien dépend de la nature de la variable dépendante, Y.
  - Par exemple, une fonction de lien `logit` est utilisée pour les distributions binomiales de Y dans la régression logistique.

Relation avec l'exemple précédent

Dans le contexte de la base de données sur le diabète chez les Pimas, un modèle de régression logistique pourrait être spécifié comme suit :

- **Composant Aléatoire** : La variable `Outcome` suit une distribution de Bernoulli, où chaque essai (patient) peut aboutir soit à un succès (avoir le diabète, codé comme 1) soit à un échec (ne pas avoir le diabète, codé comme 0).

Relation avec l'exemple précédent

Dans le contexte de la base de données sur le diabète chez les Pimas, un modèle de régression logistique pourrait être spécifié comme suit :

- **Composant Aléatoire** : La variable `Outcome` suit une distribution de Bernoulli, où chaque essai (patient) peut aboutir soit à un succès (avoir le diabète, codé comme 1) soit à un échec (ne pas avoir le diabète, codé comme 0).

- **Composant Systématique** : la combinaison linéaire des prédicteurs ou caractéristiques des patients
$\beta_0 + \beta_1 \times \text{Pregnancies} + \beta_2 \times \text{Glucose} + ... + \beta_k \times \text{Age}$

Relation avec l'exemple précédent

Dans le contexte de la base de données sur le diabète chez les Pimas, un modèle de régression logistique pourrait être spécifié comme suit :

- **Composant Aléatoire** : La variable `Outcome` suit une distribution de Bernoulli, où chaque essai (patient) peut aboutir soit à un succès (avoir le diabète, codé comme 1) soit à un échec (ne pas avoir le diabète, codé comme 0).

- **Composant Systématique** : la combinaison linéaire des prédicteurs ou caractéristiques des patients
$\beta_0 + \beta_1 \times \text{Pregnancies} + \beta_2 \times \text{Glucose} + ... + \beta_k \times \text{Age}$

- **Fonction de Lien** : le lien logit $\eta = \ln\left(\frac{p}{1-p}\right)$

Relation avec l'exemple précédent : R code


```
glm_pima <- glm(Outcome ~ Pregnancies + Glucose +
                  BloodPressure + SkinThickness +
                  Insulin + BMI + DbtPdgFunc + Age,
                    data=data_cleaned,
                 family=binomial(link="logit"))
round(summary(glm_pima)$coefficients, 4)
```

```
##                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
```

Relation avec l'exemple précédent : R code

Le code family=binomial(link="logit") spécifie la famille et la fonction de lien à utiliser dans le modèle.

- La famille `binomial` indique à la fonction `glm()` que la variable dépendante est binaire (dans ce cas, ayant le diabète ou non, souvent codée comme 1 ou 0).
- La partie `link="logit"` spécifie que la fonction de lien `"logit"` doit être utilisée, ce qui est standard pour la régression logistique qui modélise le log des cotes de la probabilité du résultat comme une combinaison linéaire des variables prédictives.

Régression de Poisson

La régression de Poisson est particulièrement adaptée pour modéliser des données de comptage où la variable de réponse Y représente le nombre de fois qu'un événement se produit dans un intervalle ou un espace fixe.

Régression de Poisson

La régression de Poisson est particulièrement adaptée pour modéliser des données de comptage où la variable de réponse Y représente le nombre de fois qu'un événement se produit dans un intervalle ou un espace fixe.

- **Composant Aléatoire** : Y suit une distribution de Poisson avec la moyenne λ des données de comptage non négatives.

Régression de Poisson

La régression de Poisson est particulièrement adaptée pour modéliser des données de comptage où la variable de réponse Y représente le nombre de fois qu'un événement se produit dans un intervalle ou un espace fixe.

- **Composant Aléatoire** : Y suit une distribution de Poisson avec la moyenne λ des données de comptage non négatives.
- **Composant Systématique** : $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

Régression de Poisson

La régression de Poisson est particulièrement adaptée pour modéliser des données de comptage où la variable de réponse Y représente le nombre de fois qu'un événement se produit dans un intervalle ou un espace fixe.

- **Composant Aléatoire** : Y suit une distribution de Poisson avec la moyenne λ des données de comptage non négatives.
- **Composant Systématique** : $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
- **Fonction de Lien** : $\eta = \ln(\lambda)$

Régression de Poisson

La régression de Poisson est particulièrement adaptée pour modéliser des données de comptage où la variable de réponse Y représente le nombre de fois qu'un événement se produit dans un intervalle ou un espace fixe.

- **Composant Aléatoire** : Y suit une distribution de Poisson avec la moyenne λ des données de comptage non négatives.
- **Composant Systématique** : $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
- **Fonction de Lien** : $\eta = \ln(\lambda)$

Exemple : modélisation du nombre d'arrivées de clients dans un magasin en une heure en fonction de facteurs tels que le jour de la semaine et les activités promotionnelles

Régression de Poisson : Exemple

La base de données `InsectSprays` est composé de deux colonnes :

- `count` : Le nombre d'insectes observés dans chacune des unités expérimentales agricoles.
- `spray` : Un facteur indiquant l'insecticide utilisé (A, B, C, D, E ou F).

```
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

Régression de Poisson : R code

L'argument family=poisson(link="log") spécifie qu'on utilise la régression de Poisson avec une fonction de lien `log`.

```
mod_count <- glm(count ~ spray, data=InsectSprays,
                 family=poisson(link="log"))
```

Régression de Poisson : R code

L'argument family=poisson(link="log") spécifie qu'on utilise la régression de Poisson avec une fonction de lien `log`.

```
mod_count <- glm(count ~ spray, data=InsectSprays,
                 family=poisson(link="log"))
```

Les coefficients indiquent l'efficacité des niveaux de spray par rapport au facteur A.

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

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.6741     0.0758 35.2744   0.0000
## sprayB        0.0559     0.1057  0.5284   0.5972
## sprayC       -1.9402     0.2139 -9.0711   0.0000
## sprayD       -1.0815     0.1507 -7.1789   0.0000
## sprayE       -1.4214     0.1719 -8.2677   0.0000
## sprayF        0.1393     0.1037  1.3433   0.1792
```

Travail en cours du projet

Travail en cours du projet

Éléments de la présentation aujoud'hui

- Problématique
- Objectifs du projet
- Méthodologie
- Retombées prévues

Ordre des présentations

- Le 27 mars : C - D - E
- Le 3 avril : D - E - C
- Le 10 avril : E - C - D