Formation Statistiques - GLM : exercices

Gilles San Martin - gilles.sanmartin@gmail.com

Septembre 2013

Contents

Note préliminaire
Les premiers exercices qui suivent utilisent des faux jeu de données simulés qui du coup se comportent de manière idéale. Dans la réalité, l'analyse de tels jeux de données devrait souvent se faire différemment. Par exemple les données de comptage suivent rarement une distribution normale (sauf pour des moyennes élevées). Lorsqu'on fait des suivi de sites au cours du temps on garde en général les mêmes sites d'année en année mais il faut alors tenir compte de l'effet site dans l'analyse, etc...

Exercice 1 : simulation de jeux de données simples

Simulez 2 jeux de données :

- l'un représente la relation linéaire négative entre deux variables quantitatives continues
- l'autre la relation entre une variable continue et une variable qualitative à 10 niveaux possibles

Analysez ces deux jeux de données et vérifiez que le modèle estime correctement les paramètres que vous avez choisis. Changez le nombre de répétitions, la variance résiduelle,... et visualisez l'effet sur les paramètres, leurs erreurs standard et le R^2.
Faites une représentation graphique des données, des valeurs prédites et de leurs erreurs standard.

Exercice 2 : régression linéaire simple

Pour suivre l'évolution des populations d'une espèce de papillon, on sélectionne chaque année 10 sites aléatoirement répartis en Belgique. Sur chaque site on réalise plusieurs transects au cours de la saison et on en retire une abondance moyenne.

- Analysez ces données et faites une représentation graphique des données, du modèle et des erreurs standard des prédictions.
- Quel est le % de variance expliquée par le modèle ?
- Quel était le nombre de papillons moyen par transect en 2000 (estimation du modèle) ?
- Qu'elle est la tendance estimée dans cette population en nombre de papillons par an ? Calculez un intervalle de confiance à 90% sur cette valeur avec les fonctions de R.
- Qu'elle est la tendance estimée en % de diminution sur 10 ans ?
- A l'aide de la méthode bootstrap (non paramétrique) estimez un intervalle de confiance à 95% sur le % de diminution en 10 ans (NB : il n'existe pas de méthode paramétrique directe pour ce faire)
- A l'aide de la méthode bootstrap (non paramétrique) estimez un intervalle de confiance sur les valeurs prédites par le modèle. Faites en une représentation graphique (NB : ici on pourrait utiliser une méthode paramétrique à la place).

```
d <- data.frame (
    nb = c(70.6, 82.8, 67.5, 103.9, 84.9, 67.7, 87.3, 91.1, 88.6, 75.4,
           101.7, 84.8, 69.7, 45.8, 95.9, 78.3, 78.8, 93.2, 91.3, 87.9,
           91.8, 89.7, 79.1, 48.2, 87.3, 77.2, 75.7, 55.9, 70.8, 84.3, 97.4,
           75.5, 82.8, 76.2, 56.3, 70.8, 71.1, 76.1, 93.5, 88.4, 73.5, 72.2,
           86.5, 84.3, 65.7, 65.4, 81.5, 87.5, 74.3, 89.2, 81, 65.8, 80.1,
           58.1, 96.5, 104.7, 69.5, 59.3, 83.5, 73, 110, 73.4, 84.3, 74.4,
           62.9, 76.8, 46.9, 96, 76.3, 106.6, 80.1, 62.4, 82.2, 59, 54.2,
           77.4, 66.4, 73, 74.1, 64.2, 63.5, 70, 89.7, 49.1, 80.9, 77, 87.9,
           67.4, 77.6, 76, 62.9, 89.1, 88.4, 81.5, 94.8, 79.4, 51.9, 62.4,
           52.6, 63.9, 60.7, 70.6, 56.3, 72.4, 60.2, 96.5, 80.8, 83.7, 75.8,
           95.2),
    year = c(2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000,
             2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2002,
             2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2003, 2003,
             2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2004, 2004, 2004,
             2005, 2005, 2005, 2005, 2005, 2005, 2006, 2006, 2006, 2006, 2006,
             2004, 2004, 2004, 2004, 2004, 2004, 2004, 2005, 2005, 2005, 2005,
             2006, 2006, 2006, 2006, 2006, 2007, 2007, 2007, 2007, 2007, 2007,
             2007, 2007, 2007, 2007, 2008, 2008, 2008, 2008, 2008, 2008, 2008,
             2008, 2008, 2008, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009,
             2009, 2009, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010,
             2010))
```

Exercice 2bis : micro analyse de puissance par simulation

Exercice subsidiaire pour ceux que la question intéresse. Pour des cas aussi simples, il existe des méthodes analytiques mais l'avantage des simulation est leur flexibilité et la possibilité de les appliquer dans pratiquement tous les cas.
On a commencé un programme de monitoring depuis 3 ans. Chaque année, on évalue l'abondance d'une espèce sur 5 sites échantillonnés au hasard.
On aimerait savoir combien de sites on devrait échantillonner chaque année pour pouvoir détecter une diminution de 3 individus par an (30 individus en 10 ans) avec une puissance de 0.8 (on détecte une pente significative dans 80 % des cas).

L'idée générale de l'analyse de puissance par simulation est la suivante : on génère des jeux de données selon un modèle qui a des caractéristiques posées par l'expérimentateur (dans notre cas la pente = -3) et d'autres estimées si possible sur base d'un jeu de données existant (dans notre cas : l'intercept et la variance résiduelle). On génère ces jeux de données un grand nombre de fois et on regarde combien de fois on détecte un effet significatif (c'est la puissance). Ensuite on recommence en faisant varier la taille d'échantillon et on examine son effet sur la puissance.

Voici comment procéder :

- Calculer la régression linéaire sur les données ci-dessous et récupérer l'intercept (alpha) et l'erreur standard des résidus (sigma). La pente beta est fixée à -3. Vous devez aussi générer les x, 10 années, avec pour commencer une seule observation par année.
- Générez un faux jeu de données avec ces paramètres, faites en l'analyse (lm) et récupérez la p valeur de la pente (dans le summary du modèle) puis stockez la.
- Recommencez l'opération un grand nombre de fois (200 fois pex)
- Recommencez à nouveau les deux points précédents mais en faisant varier le nombre de répétitions par année (de 1 à 15 devrait suffire). Idéalement les p-valeurs générées par ces simulations se stockent dans une matrice dont chaque colonne correspond à un nombre de répétitions différent.
- Finalement, comptez la proportion de p-valeurs en dessous du seuil alpha 0.05 pour chaque taille d'échantillon et regardez à partir de combien de données cette valeur devient >0.8.

```
d <- data.frame(
    y = c(81.2, 105.5, 74.9, 147.9, 109.9, 74.4, 113.6, 121.1, 116.3,
          89.8, 143.4, 109.7, 79.4, 31.6, 131.7),
    year = rep(0:2, each = n))
```

Exercice 3 : ANOVA1 et comparaisons multiples

On a mesuré la diversité entomologique dans des vergers cultivés selon 5 modalités : vergers privés sans aucune intervention, vergers en permaculture, vergers bio, vergers cultivés en lutte intégrée et vergers en culture intensive. On considère que l'indice de diversité mesuré a une distribution à peu près normale.

- Réorganisez les niveaux du facteur verger de façon à les avoir dans le même ordre que dans l'énoncé
- Faites une représentation graphique des données brutes (boxplot)
- Construisez un modèle adapté à ces données, interprétez les résultats et faites une représentation graphique des données, des valeurs prédites et de leurs erreurs standard
- Faites les comparaisons multiples de tous les vergers par rapport aux vergers "privés" (similaire à un test de Dunnett).
- Faites toutes les comparaisons entre les paires de verger (similaire à un test de Tukey). Certaines comparaisons avec les vergers privés sont devenues non significatives alors qu'elles l'étaient avec le test de Dunnett. Pourquoi ? Déduisez de ces comparaisons la représentation compacte basée sur des lettres (cld). Vous pouvez les ajouter au graphe précédent en explorant le contenu de l'objet cld (fonction str) et en utilisant la fonction mtext qui permet d'ajouter du texte dans les marges.
- Faites les comparaisons multiples suivantes (en contrôlant pour le risque d'erreur global) au moyen d'une matrice de contrastes:
  - verger privé vs verger perma (-10.557)
  - verger bio vs verger intégré (3.841)
  - verger intégré vs verger intensif (9.062)
  - diversité moyenne (privé et perma) vs verger bio (14.612)
  - diversité moyenne (privé et perma) vs verger intégré (18.453)
  - diversité moyenne (privé, perma et bio) vs diversité moyenne (intégré, intensif) (18.113)

  Calculez en suite les moyennes observées pour chaque type de verger et recalculez à la main sur base de ces moyennes les comparaisons demandées ci-dessus. Vous devriez retrouver exactement les mêmes valeurs qu'avec la matrice de contrastes (par exemple en prenant la moyenne des vergers privés moins la moyenne des vergers "perma", vous devriez obtenir une différence de 10.557).

```
d <- data.frame(
    diversité = c(34.7576055036875, 19.664575452664, 26.5066470600983,
                  -9.07418621577609, 7.5747481328663, 33.0363353529454, 16.5515602391921,
                  1.72020886529488, 16.6132880054403, 14.6597730257354, 2.13684380147872,
                  14.0096813308418, 13.3424766321721, 20.3867529036576, 35.7569643538246,
                  4.33828703365535, 38.9339429651601, 24.5328275002211, 36.9769699183864,
                  19.495011070936, 5.80452932091219, 15.0558846198543, 15.3567127658806,
                  7.75753664603512, 13.8376785183751, 9.31077053768117, 23.4485343254044,
                  13.8580656622167, 41.9091701978149, 24.3843083929896, 11.6137852772016,
                  2.14644152678566, 13.743359774925, 14.9924516364314, 20.2254391855278
    ),
    verger = c("perma", "intégré", "bio", "intensif", "intensif", "privé",
               "bio", "intensif", "bio", "bio", "intensif", "bio", "intégré",
               "bio", "privé", "intensif", "perma", "privé", "perma", "intensif",
               "intensif", "intégré", "intégré", "intensif", "intégré",
               "intégré", "bio", "bio", "perma", "privé", "intensif", "intensif",
               "intégré", "intégré", "privé"))
```

Exercice 4 : 2 variables explicatives qualitatives (ANOVA2 sans interaction)

On veut étudier comment réagi un ravageur à 3 techniques culturales. Cinq champs ont été sélectionnés, divisés en 3 parties et chaque type de technique culturale (A, B, C) a été attribuée à une des parcelles au hasard. On a mesuré ensuite le niveau de dégâts en surface attaquée en m^2 sur une zone de $100m^2$ située au centre de chaque parcelle. Cependant, l'essai a raté sur certaines parcelles et on a donc quelques valeurs manquantes (NA).

Sans tenir compte de l'effet site

- Faites une représentation graphique des données brutes en fonction des pratiques culturales sans tenir compte des sites (pex un boxplot)
- Comparez l'effet des techniques culturales sur le niveau de dégats (sans tenir compte des sites) au moyen d'un modèle linéaire (ANOVA).
- Faites une représentation graphique des données brutes, des valeurs prédites et de leur erreur standard.
- Interprétez les résultats grâce au graphique et aux sorties du modèle.
- On voudrait effectuer des comparaisons multiples pour évaluer quelle technique culturale a un effet différent des autres. Est-ce que çà a un sens ici ?
- Calculez les dégâts moyens pour chaque type de technique culturale et comparez les aux valeurs prédites par le modèle.

En tenant compte de l'effet site

- Essayez de faire une représentation graphique des données brutes montrant les valeurs pour chaque site et chaque technique culturale. Vous pouvez utiliser par exemple la fonction boxplot (syntaxe utilisant les formules). Vous pouvez aussi faire un graphique plus complexe en collant les facteurs "site" et "cult".
- Comparez l'effet des techniques culturales sur le niveau de dégâts *en tenant compte des sites* au moyen d'un modèle linéaire (ANOVA2).
- Faites une représentation graphique des données brutes, des valeurs prédites et de leur erreur standard, pour chaque type de pratique culturale et indépendamment du site (pour un site moyen).
- Interprétez les résultats grâce au graphique et aux sorties du modèle.
- Effectuez des comparaisons multiples pour évaluer quelle technique culturale a un effet différent des autres, indépendamment du site?
- Calculez les dégâts moyens pour chaque type de technique culturale et comparez les aux valeurs prédites par le modèle. Elles sont différentes, pourquoi ?

```r
d <- data.frame(
    degats = c(27, 37.2, 40.1, 58.2, NA, 71.5, 4.4, 7.7, 16.8, 50.9, 52.4, NA, 13.4,
               19, 26.7),
    site = c("1", "1", "1", "2", "2", "2", "3", "3", "3", "4", "4", "4", "5", "5", "5"),
    cult = c("A", "B", "C", "A", "B", "C", "A", "B", "C", "A", "B", "C",  "A", "B", "C"))
```

Exercice 5 : régression multiple

Suite à l'expérience précédente (ex 4), on aimerait bien comprendre quels sont les facteurs qui expliquent les différences entre les sites. On suspecte que le paysage autour du champ peut influencer le niveau de dégât. On a donc sélectionné aléatoirement 30 champs, présentant des pratiques culturales similaires, on a mesuré au centre de chacun la surface attaquée sur un carré de 10 x 10 m. On a ensuite mesuré la surface des différent types d'occupations du sol que l'on pense pouvoir influencer le niveau de dégâts dans un rayon de 2km autour du centroïde du champ : pommes de terre (pdt), céréales, betterave, colza, prairies, maïs, forêts, autres sites semi-naturels.

- Explorez le jeu de données. Vous pouvez utiliser la fonction pairs2 fournie dans le script "mytoolbox.R" (utilisez la fonction source pour charger le contenu du script en mémoire). Est-ce que toutes les données sont OK ? Si certaines valeurs vous semblent impossibles remplacez les par des valeurs manquantes.
- Examinez les corrélations entre les variables explicatives. Calculez également les VIFs d'un modèle contenant toutes les variables explicatives. Que constatez-vous ? Quelle solutions pourriez-vous adopter ?
- Une fois les dispositions prises pour éviter des problèmes potentiels, construisez un modèle prédictif du niveau de dégâts en fonction des variables paysagères et interprétez les sorties du modèle.
- Quelle est la variable explicative qui a l'effet le plus important en termes biologique/agronomique sur le niveau de dégât ? Standardisez les variables explicatives (enlever la moyenne, diviser par l'écart type, ou utiliser la fonction scale) et estimez à nouveau le modèle. Interprétez les résultats.
- Pour chaque variable explicative faites un graphique représentant les données et les valeurs prédites en contrôlant pour les autres variables explicatives.
- Comparez le R^2 et le R^2 ajusté. Sont-ils très différents ? Pourquoi ?
- Quel est le % de variance additionnel expliqué par chaque variable explicative ? (pour chaque variable estimer un modèle sans cette variable et faire la différence avec le R^2 du modèle complet)

```
d <- structure(list(pdt = c(31.86, 44.65, 68.74, 108.98, 24.2, 107.81,
113.36, 79.3, 75.49, 7.41, 24.72, 21.19, 82.44, 46.09, 92.38,
59.72, 86.11, 119.03, 45.6, 93.29, 112.16, 25.46, 78.2, 15.07,
32.07, 46.33, 1.61, 45.89, 104.36, 40.84), colza = c(8.32, 98.13,
113.11, 32.33, 20.32, 4.07, 21.45, 77, 2.75, 1, 47.12, 97.67,
45.15, 45.7, 31.79, 52.72, 54.91, 64.88, 79.88, 13.52, 26.2,
94.54, 11.74, 85.18, 26.14, 32.15, 60.57, 22.63, 52.73, 80.38
), prairie = c(34.51, 94.6, 49.08, 105.96, 112.86, 5.47, 63.37,
107.09, 66.17, 54.79, 114.82, 54.4, 81.31, 68.72, 12.35, 107.98,
29.53, 5.05, 39.35, 114.54, 106.74, 83.14, 76.86, 119.31, 78.68,
85.02, 65.29, 71.3, 34.7, 17.65), forêt = c(5.69, 31.11, 30.46,
31.17, 43.05, 32.02, 0.47, 11.63, 33.3, 25.71, 34.68, 27.25,
14.14, 46.17, 14.62, 41.86, 14.31, 13.34, 9.34, 11.61, 15.83,
15.13, 7.95, 2, 10.94, 40.53, 26.28, 45.73, 41.57, 2.29), céréale = c(25.6,
46.49, 60.39, 124.94, 27.5, 99.6, 118.24, 86.68, 81.25, 4.36,
39.83, 25.09, 76.23, 23.95, 103.63, 59.27, 85.95, 128.47, 53.82,
99.23, 121.35, 33.28, 78.95, 4.83, 38.26, 45.77, 0.05, 31.18,
99.58, 45.02), betterave = c(17.06, 60.43, 59.17, 99.78, 4.23,
105.08, 110.21, 73.01, 74.43, 11.69, 16.94, 8.25, 74.65, 46.21,
90.86, 52.69, 98, 122.43, 50.67, 90.36, 114.4, 45.53, 88.32,
12.04, 21.81, 43.66, 0.38, 47.2, 105.82, 44.46), maïs = c(27.78,
91.83, 67.78, 106.81, 114.41, 26.05, 68.9, 91.91, 57.93, 49.45,
129.51, 58.72, 86.12, 70.04, 5.68, 129.42, 35.5, 18.55, 47.77,
108.87, 93.93, 80.52, 64.55, 110.57, 71.18, 64.78, 75.34, 73.14,
21.04, 32.7), naturel = c(12.42, 35.28, 46.73, 4.02, 4000, 39.61,
8.15, 3.43, 24.84, 12.36, 27.52, 12.27, 2.49, 47.14, 29.01, 40.21,
6.65, 0.33, 3.22, 47.85, 17.84, 7.77, 1.34, 8.89, 0.53, 18.81,
34.91, 30.38, 41.34, 11.75), dégats = c(35.49, 26.46, 60.72,
32.86, 9.34, 72.89, 70.27, 33.8, 64.4, 25.6, 19, 42.85, 37.66,
13.5, 91.67, 1.49, 71.37, 84.46, 60.63, 41.68, 66.79, 15.86,
61.09, 25.94, 31.8, 1.35, 26.68, 19.29, 67.34, 59.77)), .Names = c("pdt",
```

```
    "colza", "prairie", "forêt", "céréale", "betterave", "maïs",
    "naturel", "dégats"), row.names = c(NA, -30L), class = "data.frame")
```

Exercice 6 : interaction simple entre 2 variables qualitatives

On a mesuré le rendement de froment (variable "y", en tonnes par hectares) sur une trentaine de champs semés avec 3 variétés différentes (A, B, C) et cultivés avec ou sans labour.

- Faites une représentation rapide des données (pex boxplot)
- Estimez un modèle prédictif du rendement
- Calculez les valeurs prédites, leurs erreurs standard et faites une représentation graphique du modèle et des données
- Interprétez les coefficients du modèle.
- Est-ce que le rendement est différent entre les champs labourés et non labourés ? Est-ce que le rendement est différent entre les variétés ? Quels tests globaux (comparaisons de modèles emboîtés) pourriez-vous faire ici et quelle serait leur signification ?
- Construisez la matrice de contrastes non indépendants permettant de faire les comparaisons multiples suivantes :
  - comparer chaque variété entre labour et non labour
  - comparer les variétés entre elles dans les champs labourés
  - comparer les variétés entre elles dans les champs non labourés

```
d <- structure(list(var = structure(c(1L, 1L, 1L, 1L, 1L, 1L, 1L,
1L, 1L, 1L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 3L, 3L, 3L,
3L, 3L, 3L, 3L, 3L, 3L), .Label = c("A", "B", "C"), class = "factor"),
culture = structure(c(2L, 1L, 2L, 2L, 1L, 1L, 2L, 2L, 1L,
1L, 1L, 1L, 1L, 1L, 1L, 1L, 2L, 1L, 1L, 2L, 2L, 2L, 2L, 2L,
2L, 1L, 1L, 1L, 1L, 1L), .Label = c("Labour", "Non Labour"
), class = "factor"), y = structure(c(11.37, 9.18, 11.16,
13.6, 9.33, 8.18, 12.49, 12.74, 9.58, 8.69, 10.51, 9.39,
8.38, 6.79, 10.12, 8.96, 5.98, 9.94, 9.82, 6.59, 11.92, 11.78,
11.07, 9.01, 11.62, 10.94, 10.84, 9.53, 10.52, 11.42), .Dim = c(30L,
1L), .Dimnames = list(c("1", "2", "3", "4", "5", "6", "7",
"8", "9", "10", "11", "12", "13", "14", "15", "16", "17",
"18", "19", "20", "21", "22", "23", "24", "25", "26", "27",
"28", "29", "30"), NULL))), .Names = c("var", "culture",
"y"), row.names = c(NA, -30L), class = "data.frame")
```

Exercice 7 : interaction entre une variable quantitative et une variable qualitative

On a suivi chaque année depuis 1990 l'abondance d'une espèce de papillon sur un site fauché chaque année. En 2000, on décide de changer de mode de gestion en passant à du pâturage extensif que l'on pense être plus adapté et on continue à suivre la population jusqu'en 2010. Chaque année on compte les papillons le long de 5 transects indépendants et positionnés aléatoirement sur le site. Pour simplifier le modèle on utilisera la moyenne par transect (on pourrait utiliser les données brutes mais alors il faudrait inclure une variable explicative "date"). La question est de savoir si la population a réagi positivement au changement de mode de gestion sur ce site.

- Comme toujours faites un graphique de vos données brutes pour les visualiser...
- Créez une variable (facteur) permettant de distinguer les années avant ou après le changement de mode de gestion (on l'appellera "change"). On va construire un modèle du nombre de papillons en fonction de l'année, de la variable "change" et de leur interaction
- Ne serait-il pas judicieux de centrer la variable "year" et si oui, sur quelle valeur ?
- Estimez le modèle, faites les inférences et interprétez les paramètres estimés. Quelle était la tendance en nombre d'individus par an avant et après le changement de gestion (après = -4.1067 individus par an).
- Faites une représentation graphique des données et du modèle
- Quelle était la tendance en % d'individus perdus sur 10 ans avant et après le changement (avant : -13.34%, après : - 39.45%)
- Est-ce que dans cette étude, enlever le facteur "change" tout en gardant l'interaction change x years aurait du sens ? Estimez ce modèle et faites en une représentation graphique (avec les erreurs standard des prédictions)
- sur base de ce modèle évaluez les 3 hypothèses/questions suivantes en corrigeant les p-valeurs pour prendre en compte le risque global :1 - Est-ce que la pente était différente de 0 avant 2000 ?, 2 - Est-ce qu'elle l'était après 2000 ?, 3 - est-ce que la pente était différente avant et après 2000 ?
- Est-ce qu'on peut extrapoler ces résultats à d'autres sites ? Est-ce qu'on peut conclure définitivement que c'est le mode de gestion qui a aggravé la situation ?

```
d <- data.frame(
    nb = c(105.6, 117.8, 102.5, 138.9, 119.9, 101.2, 120.8,
           124.6, 122.1, 108.9, 134.7, 117.8, 102.7, 78.8, 128.9, 109.8,
           110.3, 124.7, 122.8, 119.4, 122.8, 120.7, 110.1, 79.2, 118.3,
           106.7, 105.2, 85.4, 100.3, 113.8, 126.4, 104.5, 111.8, 105.2,
           85.3, 98.3, 98.6, 103.6, 121, 115.9, 100.5, 99.2, 113.5, 111.3,
           92.7, 90.9, 107, 113, 99.8, 114.7, 106, 90.8, 105.1, 83.1, 121.5,
           126.2, 91, 80.8, 105, 94.5, 129, 92.4, 103.3, 93.4, 81.9, 92.3,
           62.4, 111.5, 91.8, 122.1, 93.1, 75.4, 95.2, 72, 67.2, 86.9, 75.9,
           82.5, 83.6, 73.7, 70.5, 77, 96.7, 56.1, 87.9, 80.5, 91.4, 70.9,
           81.1, 79.5, 63.9, 90.1, 89.4, 82.5, 95.8, 76.9, 49.4, 59.9, 50.1,
           61.4, 55.7, 65.6, 51.3, 67.4, 55.2),
    year = c(1990, 1990, 1990, 1990, 1990, 1991, 1991, 1991, 1991, 1991,
             1992, 1992, 1992, 1992, 1992, 1993, 1993, 1993, 1993, 1993, 1994,
             1994, 1994, 1994, 1994, 1995, 1995, 1995, 1995, 1995, 1996, 1996,
             1996, 1996, 1996, 1997, 1997, 1997, 1997, 1997, 1998, 1998, 1998,
             1998, 1998, 1999, 1999, 1999, 1999, 1999, 2000, 2000, 2000, 2000,
             2000, 2001, 2001, 2001, 2001, 2001, 2002, 2002, 2002, 2002, 2002,
             2003, 2003, 2003, 2003, 2003, 2004, 2004, 2004, 2004, 2004, 2005,
             2005, 2005, 2005, 2005, 2006, 2006, 2006, 2006, 2006, 2007, 2007,
             2007, 2007, 2007, 2008, 2008, 2008, 2008, 2008, 2009, 2009, 2009,
             2009, 2009, 2010, 2010, 2010, 2010, 2010)
    )
```

Exercice 8 : non linéarité et transformation de variables

Voici un jeu de données avec une variable y que l'on veut prédire en fonction de 5 variables explicatives x1 à x5. Aucune de ces relations ne sont linéaires. Le but de l'exercice est de trouver les bonnes transformations pour linéariser ces relations.

- Explorez graphiquement les relations 2 à 2 entre y et les 5 variables explicatives. Vous pouvez faire un graphique pour chaque combinaison (dans une boucle par exemple) mais vous pouvez aussi utiliser la fonction pairs2 (après avoir sourcé le script mytoolbox.R) qui vous permettra de vérifier en plus qu'il n'y a pas trop de corrélation entre les variables explicatives
- Faites un modèle simple de y en fonction des 5 variables explicatives sans interactions. Notez le R^2 et l'erreur standard résiduelle. Faites un graphique des résidus en fonction des valeurs prédites et des graphiques permettant d'examiner la distribution des résidus. Vous pouvez utiliser plot(votremodèle) ou la fonction diagplot(votremodèle) après avoir sourcé le script(mytoolbox.R). Interprétez-les.
- Faites un graphique des résidus en fonction de chaque variable explicative. Vous pouvez utiliser la fonction diagplot2 (dans mytoolbox.R). Essayez de trouver une transformation/méthode permettant de linéariser la variable dont la relation est la plus clairement non linéaire (x5 pour commencer). Refaites des plots résidus vs variables explicatives et recommencez jusqu'à avoir linéarisé les 5 relations. Aidez-vous de la règle de Mosteller & Tukey vue dans la partie théorique.
- Lorsque vous obtenez votre modèle final, vérifiez qu'il est bon et comparez les R^2 et l'erreur standard résiduelle avec les valeurs obtenues dans le modèle de départ. Faites une représentation graphique des données et du modèle pour chaque variable explicatives. Faites ces graphiques sur l'échelle d'origine (pex y ~ x1 et pas y~x1^3).

```
d <- structure(list(y = c(-149.6, -95.1, -307.6, -319.7, -590.5, -247.2,
-713.4, -295.8, -502.2, -401.9, -425.3, -262.1, -400.7, -364.7,
-299.1, -272.3, 99, -187, -293.6, -355.8, -63.6, -377.9, 203.2,
-261.4, -456.4, -234.6, -170.7, -11.3, -212.9, 36.2, -658.1,
-340.7, -191.7, -241.3, -481.6, -758.9, -104.9, -350.5, -101.2,
-169.6, -3.5, -306.7, -360.4, -593.8, -233.6, -364.9, -387.1,
41.5, -884.3, -505.6, -283.4, -358.2, -371.4, -519.5, -263.7,
-257.3, -496.4, -497.1, 20, -158.9, -256.2, -50, -589.1, -540,
-245.7, -261.3, -288.3, 86.2, -161.6, -57.8, -257.5, -338.3,
-353.2, -261.6, -361.5, -441.2, -494.7, -486.4, -567.5, -441.3,
-223.3, -455.1, -185.6, -475.4, -725.7, -486.5, -351.1, -183.5,
-171.9, -220.5, -159.6, -193.8, -342.3, -329.7, -157.4, -615.9,
-391.5, -369.2, -430.4, -455.8), x1 = c(5.8, 7.6, 1.7, 8.5, 9.4,
5, 2.5, 2.6, 7.9, 1.9, 1.5, 1.9, 5.6, 4, 7.7, 1.7, 8.8, 6.6,
8.7, 3.9, 9.4, 2.6, 5, 4.1, 9.5, 4, 7.3, 6.1, 3.9, 9, 6.2, 6.3,
4.7, 5, 1.5, 0.1, 4.3, 4.1, 9.2, 4, 9.4, 3.3, 4.5, 0.1, 3.7,
7.6, 1.4, 6.6, 0.1, 9, 8.9, 6.4, 4.2, 2.7, 6.6, 5.3, 3.2, 5.7,
5.6, 3.3, 3.1, 4.7, 2.9, 0.4, 1.6, 4.9, 1.2, 9.1, 8.4, 3.7, 4.4,
4.3, 7.8, 2.8, 4.7, 8.3, 9.4, 8.4, 5.8, 4.3, 5.6, 1.5, 9.9, 1.1,
0.9, 4, 4.8, 6.3, 4, 8.7, 7.3, 9.8, 1.5, 6.1, 6, 3.1, 2.9, 3.1,
0.8, 0.9), x2 = c(6.4, 9.5, 8.3, 5.2, 1.5, 6.5, 6.2, 8.6, 6,
1.5, 8.3, 7.9, 7.2, 2.9, 4.6, 0.6, 6.5, 2.1, 1.5, 6.2, 8.6, 6.7,
9.1, 1, 7.7, 7.9, 1.6, 4.1, 0.6, 0, 4.9, 6, 9.1, 0.2, 2.7, 2,
1, 4.2, 0.1, 3.9, 0, 4.8, 1, 8.2, 4.7, 7.1, 5.8, 9.4, 2.2, 1.7,
8.8, 4.5, 7.3, 5.6, 2.4, 5.2, 4.1, 3.3, 9.7, 4.2, 1.9, 5.6, 6.5,
7.7, 2.7, 8.8, 2.6, 0.6, 2.4, 7.8, 8.3, 8.7, 0.4, 6.3, 4.6, 0.3,
2.6, 3.7, 6.4, 0.9, 2.3, 2.3, 7.5, 0.5, 0.6, 6.5, 7.6, 0.4, 8.1,
0.2, 8, 0.6, 4.6, 3.2, 7.6, 3.3, 4.1, 4.9, 7.2, 8.2), x3 = c(2.3,
9.6, 6.2, 2, 6.8, 0.8, 7.9, 6.8, 9.9, 5.1, 0.8, 6.2, 8.7, 8.9,
3.5, 4.4, 1.3, 2.9, 8.4, 5.9, 3.6, 4.5, 0.2, 2.8, 2.8, 7.5, 1.4,
2.3, 1.7, 0.5, 7, 1.2, 7.4, 8.9, 8.6, 4.3, 1.1, 6.4, 2, 3.7,
1, 9.9, 4.8, 8.8, 1.3, 3.9, 4.8, 8.1, 6.7, 6.8, 9.2, 8.4, 1.5,
9.8, 6.6, 4.8, 4.6, 2.2, 7.9, 1.7, 3.5, 0.9, 2.1, 8.8, 2.1, 6.9,
2.5, 0.1, 8.2, 0.8, 2.4, 4.7, 6.6, 0.1, 7.3, 4.7, 3.7, 4.9, 4.3,
6.9, 2, 9.9, 4.1, 5.5, 2, 3.8, 4, 7.7, 8.4, 7, 6.8, 5.1, 7.2,
3.9, 5.7, 8.4, 6.5, 3.9, 4.5, 9.4), x4 = c(5.7, 8.9, 5.9, 3.4,
1.6, 7.7, 4.9, 3.8, 3.6, 2.6, 6.1, 6.9, 9.1, 9.3, 3.1, 7.4, 9.6,
9.8, 5.9, 1, 3.8, 7.2, 3.8, 6.7, 2.1, 6.3, 3, 9.9, 1.7, 8, 1.1,
0.2, 3.7, 7.7, 2.2, 5.5, 5.7, 5.1, 9.1, 7.1, 6.1, 7.4, 5.4, 8.4,
7.2, 7.9, 7.4, 6.1, 6.7, 4.1, 0.9, 6.7, 2.9, 0, 5.9, 0.4, 7.5,
0.3, 9, 6.3, 5, 7.2, 1.9, 3.5, 8.7, 5.4, 7, 4.2, 9.5, 2.4, 6,
4.5, 8.2, 1.2, 1.5, 6.5, 4.1, 2.7, 2.7, 7.7, 9.6, 6.4, 8.6, 8.6,
1.9, 3.3, 4.6, 9.2, 7.7, 8, 6.3, 4.8, 7.9, 1.8, 6.5, 6.9, 3.2,
3.1, 2.7, 8.7), x5 = c(6.4, 9.2, 3.3, 8.9, 9.8, 9.1, 0.1, 2.7,
1.2, 3.7, 10, 3.9, 1, 2, 2.1, 4.9, 5.2, 8.5, 7.3, 6.8, 6.8, 8.4,
6.3, 7.4, 10, 3.8, 7, 4.3, 4.3, 3, 0.1, 1.2, 2.3, 4.7, 4.6, 8.3,
5.5, 2.9, 1.9, 4.7, 5, 4.6, 8, 3.1, 8.2, 9.5, 2.5, 4.1, 9.1,
9.4, 8.2, 7.9, 9.1, 2.7, 7, 4.5, 0.7, 0.6, 1.1, 4.1, 3.5, 3.9,
10, 4.9, 2.8, 1.8, 4.4, 5.2, 3.3, 4.8, 8.7, 1.1, 1.2, 1.3, 4,
0.6, 0.3, 9.3, 0.3, 1.3, 1.2, 3.3, 1.4, 1.7, 10, 9.1, 8.6, 6.6,
5.5, 7.3, 6.2, 3.6, 3.6, 2.4, 3.7, 9.6, 3.7, 7.3, 4.6, 8.3)), .Names = c("y",
"x1", "x2", "x3", "x4", "x5"), row.names = c(NA, -100L), class = "data.frame")
```

Exercice 9a : GLM logistique à une variable explicative

On a relevé la présence ou l'abence d'une espèce de plante dans 100 sites pour lequels on dispose également d'une mesure de l'humidité du sol sur une échelle continue de 1 à 12 (très sec à aquatique).
Faites un modèle prédictif de la présence de cette espèce et une réprésentation graphique des données et du modèle.
Au moyen des valeurs prédites, déterminez dans quelle gamme de valeurs d'humidité, on a 80% de chance de trouver l'espèce (réponse : 7.3 - 9.5).

```
d <- data.frame(
    hum = c(3.92, 5.09, 7.3, 10.99, 3.22, 10.88, 11.39,
      8.27, 7.92, 1.68, 3.27, 2.94, 8.56, 5.23, 9.47, 6.47, 8.89, 11.91,
      5.18, 9.55, 11.28, 3.33, 8.17, 2.38, 3.94, 5.25, 1.15, 5.21,
      10.57, 4.74, 6.3, 7.6, 6.43, 3.05, 10.1, 8.35, 9.74, 2.19, 8.96,
      5.52, 10.03, 8.12, 9.61, 7.08, 6.83, 9.68, 1.26, 6.25, 9.06,
      8.62, 6.25, 10.47, 5.82, 3.69, 1.78, 2.09, 4.48, 6.7, 8.28, 5.48,
      11.04, 4.23, 6.05, 4.66, 8.16, 3.84, 6.26, 9.43, 1.93, 10.63,
      4.73, 10.23, 4.81, 4.67, 6.24, 10.81, 10.51, 5.29, 9.55, 11.57,
      5.78, 8.84, 5.4, 4.58, 9.33, 3.23, 8.82, 2.34, 3.7, 2.58, 3.64,
      1.65, 8.07, 10.64, 9.57, 9.77, 6.01, 5.51, 9.92, 7.65),
    pres = c(0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1,
     0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1,
     1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
     0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,
     0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1)
    )
```

Exercice 9b : GLM logistique à deux variables explicatives

On a relevé la présence ou l'abence d'une espèce de plante dans 300 sites pour lequels on dispose également d'une mesure de l'humidité du sol sur une échelle continue de 1 à 12 (très sec à aquatique) et d'acidité du sol sur une échelle continue de 1 à 9 (très acide à très basique). NB : Dans ce cas l'analyse par GLM reste faiable mais pour ce genre de modèles les Generalized Additive Models (GAM) peuvent vite se montrer assez utiles.

- Faites un modèle prédictif de la présence de cette espèce et une réprésentation graphique classique des données et du modèle sous forme de courbe (fonction lines).
  Pour la représentation graphique, vous pouvez par exemple représenter les valeurs prédites en fonction de l'humidité pour des valeurs de ph de 3, 5, 7 et en fonction du ph pour des valeurs d'humidité de 5, 7, 9 et 11.
- la représentation graphique précédente n'est pas optimale dans ce cas. L'idéal est une représentation en pseudo-3D. Pour ce faire estimez les valeurs prédites de probabilité présence en fonction de toutes les combinaisons possibles de humidité et ph. Pour construire une matrice avec toutes ces combinaisons possibles, vous pouvez vous aider de la fonction expand.grid. Vous pouvez enuite faire eune représentation avec hum en x, ph en y et des couleurs pour représenter la probabilité prédite. Vous pouvez utiliser la fonction image ou filled.contour. Vous pouvez aussi explorer les fonctions persp ainsi que persp3D du package rgl. Voir les exemples dans les transsparents de la formation R (partie sur les graphiques). Pour utiliser ces fonctions vous devrez transformer votre vecteur de valeurs prédites en une matrice dont les colonnes correspondent aux valeurs de hum, le lignes aux valeurs de ph et la matrice elle même contient les valeurs prédites. (voir exemples de image avec le jeu de données volcano)
- Au moyen des valeurs prédites, déterminez dans quelle gamme de valeurs d'humidité, et de pH on a 80% de chance de trouver l'espèce (réponse : humidité : 7.1 - 10.2, ph : 1.8 - 4.6).

```
d <- structure(list(hum = c(4.55, 1.49, 9.05, 6.42, 3.59, 2.39, 4.08,
11.14, 7.44, 4.86, 10.83, 4.03, 3.43, 4.31, 10.76, 2.88, 2.39,
2.6, 11.22, 1.58, 1.73, 9.69, 7.12, 3.83, 1.66, 9.38, 11.84,
3.56, 7.78, 10.11, 8.39, 4.98, 11.62, 10.44, 9.11, 8.96, 6.45,
1.1, 9.7, 1.85, 7.16, 8.67, 9.72, 9.32, 8.24, 6.65, 7.12, 10.9,
6.05, 11.98, 9.35, 5.03, 10.36, 8.84, 6.72, 1.47, 11.56, 5.98,
6.73, 3.06, 1.21, 3.12, 4.91, 4.17, 11.5, 7.54, 10.62, 8.96,
6.04, 8, 4.33, 11.54, 11.3, 2.72, 5.09, 10.08, 10.84, 1.16, 9.17,
1.71, 6.9, 6.49, 6.22, 3.95, 3.22, 9.51, 11.54, 7.35, 2.83, 1.9,
8.86, 2.6, 4.59, 11.7, 2.5, 4.45, 2.15, 2.27, 11.83, 7.79, 7.96,
3.84, 8.75, 3.2, 10.14, 3.64, 11.15, 11.94, 5.77, 6.08, 9.63,
10.34, 3.92, 3.03, 11.76, 4.47, 11.24, 6.06, 7.22, 9.31, 9.48,
5.25, 3.45, 2.07, 11.21, 4.71, 1.97, 2.25, 2.23, 3.64, 11.17,
10.4, 11.96, 7.31, 11.1, 11.69, 10.51, 5.11, 5.32, 4.33, 10.63,
11.12, 9.41, 11.5, 6.78, 9.94, 6.19, 11.01, 10.72, 9.64, 2.41,
4.19, 10.42, 4.7, 6.72, 2.99, 9.69, 7.2, 11.01, 11.16, 4.2, 5.82,
9.11, 11.38, 5.1, 1.22, 9.81, 1.09, 6.57, 2.14, 5.1, 9.78, 3.07,
8.78, 5.07, 4.87, 6.45, 4.48, 11.17, 5.48, 3.48, 2.73, 8.59,
5.49, 3.12, 9.95, 6.01, 9.87, 7.37, 11.68, 1.55, 3.73, 9.62,
7.28, 5.83, 7.04, 1.65, 5.72, 2.57, 2.03, 2.46, 5.02, 6, 6.91,
3.04, 5.94, 3.82, 7.8, 10.69, 4.5, 10.12, 11.58, 6.97, 10.03,
11.38, 10.17, 9.75, 3.51, 3.01, 3.24, 4.87, 10.74, 2.23, 11.25,
8.68, 9.47, 8.77, 9.58, 7.07, 5.65, 5.42, 9.52, 10.15, 2.38,
2.12, 11.6, 4.15, 7.82, 7.73, 7.18, 4.24, 9.13, 9.55, 7.71, 7.02,
10.47, 6.9, 4.6, 10.95, 1.04, 8.01, 10.6, 3.34, 2.3, 1.37, 9.61,
10.51, 8.69, 1.8, 10.72, 5.61, 11.18, 9.61, 1.81, 2.76, 11.94,
10.88, 1.25, 7.43, 10.26, 3.63, 3.8, 11.78, 7.03, 3.29, 9.21,
6.38, 3.88, 9.85, 8.52, 4.54, 1.97, 1.27, 1.8, 1.71, 2.72, 3.14,
6.87, 4.46, 5.17, 9.58, 8.48, 2.29, 10.2, 8.1, 11.19, 2.9, 6.76,
4, 10.18), ph = c(8.24, 1.99, 8.67, 4.76, 6.33, 8.5, 2.18, 2.86,
8.47, 8.73, 5.39, 2.91, 3.88, 6.53, 2.08, 8.24, 2.11, 3.83, 2.31,
3.08, 2.5, 7.44, 8.81, 4.63, 8.66, 6.61, 2.35, 5.06, 8.01, 1.72,
3.21, 8.95, 7, 3.02, 3.19, 4.4, 1.38, 5.68, 8.15, 2.51, 1.5,
1.06, 6.11, 5.27, 3.04, 6, 8.92, 6.04, 4.95, 2.6, 2.75, 1.7,
2.89, 2.58, 5.28, 4.49, 3.55, 1.33, 1.33, 1.54, 1.43, 2.48, 5.76,
4.54, 5.49, 1.23, 3.21, 7.77, 3.21, 2.1, 6.3, 2.26, 3.55, 2.22,
4.08, 5.71, 3.73, 4.67, 4.23, 2.36, 5.46, 1.47, 2.61, 7.56, 7.71,
8.69, 3.61, 5.75, 5.46, 8.92, 5.15, 4.67, 2.79, 1.61, 3.64, 8.39,
7.15, 1.8, 6.9, 1.06, 8.14, 1.82, 5.73, 3.73, 4.01, 8.12, 8.47,
4.3, 5.22, 1.69, 8.27, 7.21, 5.58, 8.11, 4.69, 4.54, 2.58, 4.02,
5.85, 5.35, 1.1, 8.96, 8.05, 6.99, 7.53, 7.27, 6.88, 4.44, 5.79,
1.29, 6.68, 4.06, 7.63, 5.94, 2.9, 2.8, 8.77, 2.74, 8.35, 3.46,
5.36, 3.87, 6.17, 8.22, 8.43, 7.44, 1.86, 5.46, 2.93, 8.06, 7.46,
5.7, 4.7, 7.78, 8.17, 2.03, 1.03, 2.03, 2.46, 8.88, 7.09, 7.94,
8.86, 1.17, 7.96, 4.6, 6.18, 6.51, 7.93, 4.28, 6.91, 6.39, 3.88,
2.96, 1.33, 8.23, 8.77, 5.92, 3.96, 5.94, 5.66, 1.48, 8.67, 6.73,
7.69, 4.35, 6.7, 3.62, 3.86, 2.54, 3.63, 3.33, 3.8, 4.9, 8.73,
5.19, 7.41, 5.65, 6.04, 7.75, 6.31, 5.16, 3.65, 5.93, 5.65, 5.76,
1.47, 6.92, 4.94, 8.99, 5.33, 1.33, 2.62, 3.88, 1.93, 5.47, 2.88,
3.5, 5.01, 3.99, 3.93, 1.02, 6.07, 4.29, 3.84, 5.89, 4.95, 6.7,
8.96, 4.97, 6.99, 7.73, 5.44, 5.74, 2.22, 1.15, 4.09, 7.72, 2.58,
6.75, 7.97, 2.79, 4.25, 3.7, 3.25, 8.94, 6.15, 1.23, 3.27, 7.52,
6.43, 2.48, 1.23, 7.85, 5.65, 2.56, 4.74, 3.86, 1.52, 8.46, 4.96,
2.71, 2, 6.9, 5.37, 4.84, 6.09, 3.16, 6.66, 5.34, 3.42, 3.73,
8.22, 7.93, 7.64, 5.86, 4.69, 5.43, 6.23, 7.92, 8.5, 1.05, 6.15,
4.1, 2.04, 6.17, 4.38, 2.18, 4.17, 3.12, 3.79, 6.55, 5.06, 1.7,
4.08, 8.88, 8.03, 5.98, 8.94, 7.91), pres = c(0L, 0L, 0L, 1L,
```

```
0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L,
0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 1L, 1L, 1L,
0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 1L, 0L,
1L, 1L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L,
0L, 1L, 0L, 0L, 1L, 0L, 0L, 0L, 1L, 0L, 1L, 0L, 1L, 0L, 0L, 0L,
0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L,
0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L,
0L, 1L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 1L,
0L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L,
0L, 0L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 0L,
0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 0L,
0L, 0L, 0L, 0L, 0L, 1L, 0L, 1L, 1L, 0L, 0L, 0L, 1L, 0L, 0L, 0L,
0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L,
0L, 1L, 0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 0L, 1L, 1L, 0L, 0L, 0L,
0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 1L, 0L, 1L,
1L, 0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 0L, 1L, 0L, 1L, 0L, 0L,
0L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 1L,
0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L,
0L, 1L, 1L, 0L, 0L, 0L, 0L, 0L)), .Names = c("hum", "ph", "pres"
), row.names = c(NA, -300L), class = "data.frame")
```

Exercice 10 : Vaches laitières - modèle mixte gaussien en carré latin

Note préliminaire
Les premiers exercices de cette série correspondent à des designs expérimentaux classiques avec à la fois des facteurs fixes et des facteurs aléatoires. Pour pouvoir comparer les différentes approches on a tenté de les analyser à la fois avec un modèle fixe, une anova mixte classique (fonction aov) et des modèles mixtes. Dans ces cas parfaitement ballancés, avec des erreurs à distribution normale, et des variables explicatives parfaitement indépendantes, les anova mixtes et les modèles mixtes devraient donner des résultats proches.

Ennoncé de l'exercice 10
On veut comparer l'effet de deux aliments B et C par rapport à un aliment témoin A sur la production laitière. On a fourni ces aliments à 6 vaches pendant 3 périodes. Chaque vache a reçu chaque aliment une seule fois pendant une des 3 périodes dans un orde différent. On s'est arrangé pourqu'à chaque période chaque aliment soit attribué à 2 vaches sur les six (design en carré latin). Chaque période représente 4 semaines d'observations avec une période d'adaptation préalable. C'est un exemple Tiré de Dagnelie 2003 Principes d'expérimentation Exemple 8.6. L'ouvrage est disponible en ligne : http://www.dagnelie.be/

```
d <- data.frame(
    aliment = c("B", "C", "A", "B", "A", "C", "A", "A", "B", "C", "C",
                "B", "C", "B", "C", "A", "B", "A"),
    periode = paste("per", rep(1:3, each = 6), sep="_"),
    vache = paste("vache", rep(1:6, times = 3), sep="_"),
    lait = c(21.7,20.4,25.9,23.5,19.1,19.0,19.1,19.9,24.0,20.5,17.0,19.1,16.4,
             19.5,22.2,21.9,20.0,17.6),
    aliment_t0 = c( rep(0,6), "B", "C", "A", "B", "A", "C", "A", "A",
                    "B", "C", "C", "B")
)
```

Exercice 11 : Modèle mixte binomial et sélection de modèle

On veut étudier les facteurs environnementaux qui peuvent expliquer (ou du moins prédire) la présence de la sole dans un estuaire portugais. (données du chapitre 21, de Zuur et al. 2007).
La présence de l'espèce et les variables environnementales ont été notés sur 65 points d'échantillonage répartis dans 4 zones (Area).
Construisez un modèle prédictif de la présence de l'espèce. Est-ce qu'un modèle mixte serait adapté dans ce cas et si oui lequel ? Faut-il utiliser toutes ces variables explicatives ? N'y a-t-il pas des variables redondantes/colinéaires ? Les relations sont-elles bien linéaires ? N'y a-t-il pas de points extrèmes ?
Une fois que vous avez un modèle complet satisfaisant, utilisez les méthodes de sélection de modèles pour déterminer quelles sont les variables les plus importantes pour prédire la présence de cette espèce.
Faites une représentation graphique du modèle sur base des "model averaged coefficients".

```
d <- structure(list(Sample = 1:65, season = c(1L, 1L, 1L, 1L, 1L,
1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L,
1L, 1L, 1L, 1L, 1L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L,
2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L,
2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L), month = c(5L,
5L, 5L, 5L, 5L, 5L, 5L, 5L, 5L, 5L, 5L, 5L, 6L, 6L, 6L, 6L,
6L, 6L, 6L, 6L, 6L, 6L, 6L, 6L, 6L, 7L, 7L, 7L, 7L, 7L, 7L, 7L,
7L, 7L, 7L, 7L, 7L, 7L, 8L, 8L, 8L, 8L, 8L, 8L, 8L, 8L, 8L, 8L,
8L, 8L, 8L, 9L, 9L, 9L, 9L, 9L, 9L, 9L, 9L, 9L, 9L, 9L, 9L, 9L
), Area = c(2L, 2L, 2L, 4L, 4L, 4L, 3L, 3L, 3L, 1L, 1L, 1L, 1L,
2L, 2L, 2L, 4L, 4L, 4L, 3L, 3L, 3L, 1L, 1L, 1L, 1L, 2L, 2L, 2L,
4L, 4L, 4L, 3L, 3L, 3L, 1L, 1L, 1L, 1L, 4L, 4L, 4L, 3L, 3L, 3L,
2L, 2L, 2L, 1L, 1L, 1L, 1L, 3L, 3L, 3L, 2L, 2L, 2L, 4L, 4L, 4L,
1L, 1L, 1L, 1L), depth = c(3, 2.6, 2.6, 2.1, 3.2, 3.5, 1.6, 1.7,
1.8, 4.5, 6, 4, 4, 2.7, 2.2, 2.5, 2.8, 4, 3.2, 2.3, 1.7, 2, 6.5,
3.5, 9, 2.5, 2.3, 2, 2.4, 2.4, 2, 1.6, 1.3, 1.25, 1.7, 4.5, 4.25,
4, 5, 2.25, 3.7, 2.2, 2, 1.7, 1.7, 2.7, 2.5, 2, 7, 5, 6, 6.5,
1.5, 1.5, 1.5, 3.2, 2.5, 3.5, 2.8, 2.1, 4, 5, 4.8, 1.4, 2.5),
temperature = c(20L, 18L, 19L, 20L, 20L, 20L, 19L, 17L, 19L,
21L, 22L, 22L, 22L, 21L, 21L, 21L, 21L, 21L, 22L, 20L, 20L,
21L, 22L, 24L, 23L, 23L, 23L, 23L, 23L, 24L, 24L, 25L, 27L,
26L, 26L, 25L, 25L, 25L, 25L, 24L, 24L, 24L, 24L, 24L, 24L,
22L, 22L, 23L, 25L, 25L, 25L, 25L, 23L, 23L, 23L, 19L, 19L,
18L, 17L, 19L, 19L, 21L, 21L, 22L, 21L), salinity = c(30L,
29L, 30L, 29L, 30L, 32L, 29L, 28L, 29L, 12L, 17L, 3L, 2L,
28L, 29L, 28L, 30L, 30L, 30L, 30L, 30L, 31L, 4L, 7L, 12L,
16L, 29L, 29L, 30L, 25L, 27L, 29L, 27L, 25L, 25L, 10L, 12L,
14L, 4L, 30L, 32L, 32L, 30L, 32L, 34L, 30L, 30L, 30L, 14L,
15L, 7L, 5L, 27L, 30L, 31L, 30L, 34L, 33L, 33L, 31L, 31L,
10L, 16L, 19L, 5L), transparency = c(15L, 15L, 15L, 15L,
15L, 7L, 15L, 10L, 10L, 35L, 30L, 35L, 35L, 10L, 10L, 5L,
10L, 5L, 40L, 20L, 15L, 10L, 40L, 20L, 35L, 40L, 25L, 10L,
15L, 25L, 20L, 30L, 20L, 20L, 20L, 60L, 80L, 50L, 60L, 25L,
25L, 40L, 20L, 15L, 15L, 30L, 30L, 30L, 70L, 80L, 50L, 40L,
20L, 15L, 15L, 20L, 20L, 5L, 15L, 7L, 10L, 30L, 50L, 30L,
25L), gravel = c(3.74, 1.94, 2.88, 11.06, 9.87, 32.45, 6.77,
22.61, 4.45, 3.54, 2.14, 3.1, 2.96, 3.64, 1.03, 1.33, 10.84,
8.43, 31.49, 5.08, 24.13, 3.37, 1.42, 1.77, 3.72, 0.62, 0.46,
1.63, 0.04, 7.88, 6.6, 31.29, 8.59, 28.59, 7.15, 1.96, 1.2,
6.85, 1.05, 2.62, 2.82, 1.96, 11.94, 7.66, 27.23, 9.2, 30.08,
4.9, 6.85, 1.41, 2.73, 3.2, 2.01, 0.21, 5.49, 7.33, 10.68,
26.22, 6.45, 30, 8.13, 10.47, 3.62, 1.51, 1.27), large_sand = c(13.15,
4.99, 8.98, 11.96, 28.6, 7.39, 14.55, 34.15, 15.43, 15.16,
36.89, 37.9, 40.24, 13.08, 3.14, 7.08, 11.22, 26.45, 6.46,
12.11, 37.24, 17.03, 12.53, 37.71, 39.08, 35.59, 16.22, 3.98,
```

```
3.01, 14.37, 23.7, 5.81, 9.33, 33.23, 13.12, 9.55, 40.71,
43.24, 32.19, 13.1, 2.33, 0.34, 16.88, 25.9, 10.34, 13.43,
35.5, 10.79, 5.99, 36.57, 41.21, 31.68, 8.92, 0.01, 4.03,
13.31, 24.88, 10.57, 9.2, 35.74, 7.88, 10.34, 35.92, 44.54,
27.4), med_fine_sand = c(11.93, 5.43, 16.85, 21.95, 19.49,
9.43, 9.88, 6.5, 7.88, 44.88, 44.33, 35.27, 36.81, 10.34,
4.42, 14.11, 20.7, 19.06, 9.38, 11.78, 6.44, 4.47, 41.57,
49.1, 38.03, 34.54, 7.08, 1.57, 13.51, 21.93, 21.41, 12.04,
14.51, 9.31, 0.91, 44.92, 51.56, 37.79, 31.19, 9.48, 0.55,
8.58, 22.42, 25.87, 11.2, 13.07, 8.53, 3.7, 40.65, 50.33,
36.82, 30.84, 7.87, 2.49, 5.5, 26.61, 26.46, 12.83, 16.02,
8.3, 1.29, 43.04, 45.72, 34.83, 29.44), mud = c(71.18, 87.63,
71.29, 55.03, 42.04, 50.72, 68.8, 36.75, 72.24, 36.42, 16.65,
23.73, 20, 72.94, 91.4, 77.48, 57.24, 46.06, 52.67, 71.03,
32.19, 75.13, 44.48, 11.43, 19.17, 29.25, 76.24, 92.82, 83.44,
55.82, 48.29, 50.86, 67.57, 28.87, 78.82, 43.57, 6.53, 12.12,
35.57, 74.8, 94.31, 89.11, 48.76, 40.57, 51.23, 64.31, 25.89,
80.6, 46.52, 11.69, 19.24, 34.28, 81.19, 97.29, 84.98, 52.76,
37.98, 50.38, 68.33, 25.96, 82.7, 36.14, 14.75, 19.12, 41.89
), Solea_solea = c(0L, 0L, 1L, 0L, 0L, 0L, 1L, 1L, 0L, 1L,
0L, 1L, 1L, 0L, 1L, 1L, 0L, 0L, 0L, 0L, 1L, 0L, 1L, 1L, 1L,
1L, 1L, 0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 1L, 0L, 1L, 1L, 1L,
0L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 1L, 1L, 1L, 1L, 0L, 0L, 0L,
0L, 0L, 1L, 0L, 0L, 0L, 1L, 0L, 0L, 0L)), .Names = c("Sample",
"season", "month", "Area", "depth", "temperature", "salinity",
"transparency", "gravel", "large_sand", "med_fine_sand", "mud",
"Solea_solea"), class = "data.frame", row.names = c(NA, -65L))
```