Démarche Statistique 2025-2026
Contrôle des connaissances - Partie 2 (durée : 1h)

26 janvier 2026

Nom Prénom :

Tous les documents sont autorisés. La calculatrice est autorisée. L'usage du téléphone portable est interdit, même pour sa fonction calculatrice.

On s'intéresse dans la suite à la prédiction de rendement de maïs à partir de variables agroclimatiques, tel que proposé par un consortium piloté par l'Université Paris Saclay, dans le cadre d'un *data challenge*.

Les données sont des moyennes mensuelles par département, pour 57 années et 94 départements français (au total 3394 lignes dans le tableau des données), des 55 variables suivantes :

- `yield_anomaly` : variable à prédire représentant l'anomalie de rendement de maïs (une valeur positive indique un rendement plus élevé qu'attendu, une valeur négative indique une perte de rendement par rapport à la valeur attendue), exprimée en tonne par ha.
- `ETP_1 ...ETP_9` : Evapotranspiration potentielle moyenne mensuelle (1=janvier, 9=septembre)
- `PR_1 ...PR_9` : Précipitation cumulée mensuelle (1=janvier, 9=septembre)
- `RV_1 ...RV_9` : Rayonnement moyen mensuel (1=janvier, 9=septembre)
- `SeqPR_1 ...SeqPR_9` : Nombre de jours de pluie mensuel (1=janvier, 9=septembre)
- `Tn_1 ...Tn_9` : Température minimale journalière moyenne mensuelle (1=janvier, 9=septembre)
- `Tx_1 ...Tx_9` : Température maximale journalière moyenne mensuelle (1=janvier, 9=septembre)
- `IRR` : variable catégorielle dont les valeurs (1, 2, ... , 5) sont des niveaux de surface agricole irriguée dans chaque département. La valeur 1 indique une fraction faible de surface agricole irriguée, la valeur 5 indique une fraction élevée de surface irriguée.

Dans une optique de visualisation des corrélations entre les 54 variables agroclimatiques quantitatives impliquées dans cette problématique, on utilise dans un premier temps l'Analyse en Composantes Principales :

```
dta_pca <- PCA(dta,quanti.sup=1,quali.sup=2,graph=FALSE)
round(dta_pca$eig[1:15,],digits=2)
```

```
        eigenvalue percentage of variance cumulative percentage of variance
comp 1       15.68                 29.05                               29.05
comp 2        6.15                 11.38                               40.43
comp 3        4.27                  7.90                               48.33
comp 4        3.34                  6.19                               54.52
comp 5        2.87                  5.31                               59.83
comp 6        2.42                  4.49                               64.32
comp 7        2.17                  4.02                               68.34
comp 8        1.98                  3.68                               72.01
comp 9        1.86                  3.45                               75.46
comp 10       1.70                  3.15                               78.61
comp 11       1.66                  3.07                               81.68
comp 12       1.28                  2.38                               84.06
comp 13       0.87                  1.60                               85.66
comp 14       0.80                  1.49                               87.15
comp 15       0.68                  1.27                               88.41
```

Question 1

D'après les résultats ci-dessus, à partir de quel rang une composante principale ne synthétise plus suffisamment d'information pour qu'il soit intéressant de l'étudier ? (justifiez votre réponse)

Réponse

On cherche à donner une interprétation aux premières composantes principales :

```
ord <- order(abs(dta_pca$var$cor[,1]),decreasing=TRUE)
round(dta_pca$var$cor[ord[1:5],1:3],digits=2)
```

```
      Dim.1 Dim.2 Dim.3
Tn_8   0.79  0.13  0.05
Tn_6   0.78  0.25  0.11
Tx_8   0.76 -0.04  0.01
ETP_3  0.75 -0.02 -0.21
ETP_8  0.74 -0.42  0.10
```

```
ord <- order(abs(dta_pca$var$cor[,2]),decreasing=TRUE)
round(dta_pca$var$cor[ord[1:5],1:3],digits=2)
```

```
     Dim.1 Dim.2 Dim.3
RV_7  0.43 -0.69  0.26
RV_8  0.46 -0.68  0.12
RV_1  0.46 -0.66 -0.12
RV_2  0.41 -0.66  0.08
Tn_2  0.42  0.58 -0.11
```

```
ord <- order(abs(dta_pca$var$cor[,3]),decreasing=TRUE)
round(dta_pca$var$cor[ord[1:5],1:3],digits=2)
```

```
        Dim.1 Dim.2 Dim.3
SeqPR_6 -0.34 -0.05 -0.51
PR_6    -0.27 -0.23 -0.50
RV_4     0.31 -0.21 -0.47
RV_5     0.49 -0.34 -0.46
Tn_1     0.42  0.56  0.43
```

Question 2

D'après les résultats ci-dessus, proposez une interprétation de la 1ère composante principale (que signifient une forte et une faible valeur de cette composante ?). Peut-on dire qu'une situation (un département x une année) de valeur forte de cette composante principale correspond à un été frais ?

Réponse

La variable à prédire, `yield_anomaly`, a été intégrée dans l'Analyse en Composantes Principales en tant que variable supplémentaire :

```
round(dta_pca$quanti.sup$cor,digits=2)
```

```
              Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
yield_anomaly -0.05  0.16 -0.25 -0.08  0.15
```

Question 3

D'après les résultats ci-dessus, quelle composante principale est la plus liée à la variable à prédire ? Diriez-vous qu'une anomalie de rendement élevée est associée à un mois de juin chaud ?

Réponse

La variable mesurant le niveau de surfaces agricoles irriguées, `IRR`, a aussi été intégrée dans l'Analyse en Composantes Principales en tant que variable supplémentaire :

```
round(dta_pca$quali.sup$eta2,digits=2)
```

```
    Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
IRR  0.23  0.01     0     0  0.01
```

```
round(dta_pca$quali.sup$coord,digits=2)
```

```
      Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
IRR_1 -2.38  0.25 -0.12 -0.05  0.22
IRR_2 -1.07  0.09 -0.11 -0.08  0.03
IRR_3  1.67 -0.26  0.19  0.09 -0.15
IRR_4  2.56 -0.09  0.03  0.05 -0.12
IRR_5  1.41 -0.15  0.09  0.04 -0.11
```

Question 4

D'après les résultats ci-dessus, diriez-vous qu'une situation (département x année) associée à une forte valeur de la 1ère composante principale correspond aussi à un niveau élevé de surface agricole irriguée ? (justidiez votre réponse)

Réponse

Dans une première approche pour prédire `yield_anomaly`, on cherche à construire un modèle à partir de tout ou partie des 54 variables explicatives quantitatives. On présente ci-après les 10 premières lignes de la table des coefficients estimés du modèle intégrant les 54 variables explicatives :

```
fullmod <- lm(yield_anomaly~.,data=dta[,-2])
summary(fullmod)$coef[1:10,]
```

```
               Estimate Std. Error    t value    Pr(>|t|)
(Intercept) -0.55071507   0.434870 -1.2663891 2.05462e-01
ETP_1        0.00498927   0.189669  0.0263051 9.79016e-01
ETP_2        0.23357983   0.186782  1.2505500 2.11186e-01
ETP_3        1.04177247   0.179582  5.8010919 7.20301e-09
ETP_4       -0.74563848   0.172719 -4.3170682 1.62709e-05
ETP_5        0.48790679   0.181291  2.6912945 7.15301e-03
ETP_6        0.47368456   0.171489  2.7621883 5.77296e-03
ETP_7       -0.73789709   0.144153 -5.1188341 3.24809e-07
ETP_8       -0.97122845   0.163980 -5.9228385 3.48546e-09
ETP_9        0.60858321   0.174969  3.4782297 5.11187e-04
```

Question 5

Comment expliqueriez-vous à un.e non-statisticien.ne le résultat du test associé à l'effet de ETP_1 dans le tableau ci-dessus ?

Réponse

Grâce à la fonction stepwise du package RcmdrMisc de R, on recherche le meilleur modèle de régression linéaire de yield_anomaly sur un sous-ensemble des variables explicatives :

```
select <- RcmdrMisc::stepwise(fullmod,
                              direction="forward/backward",
                              criterion="BIC",trace=FALSE)
```

```
Direction:  forward/backward
Criterion:  BIC
```

```
summary(select)$coef
```

```
               Estimate Std. Error  t value    Pr(>|t|)
(Intercept) -0.53575394 0.30800946 -1.73941 8.20546e-02
PR_7         0.15540741 0.02132336  7.28813 3.89746e-13
PR_8         0.09022385 0.01851921  4.87191 1.15628e-06
ETP_3        1.20091100 0.14812572  8.10738 7.16931e-16
Tx_6        -0.14920435 0.01638239 -9.10760 1.40961e-19
Tx_5         0.08712099 0.01053425  8.27026 1.90232e-16
ETP_8       -0.91168119 0.14588394 -6.24936 4.63830e-10
RV_1         0.01668730 0.00170945  9.76181 3.22847e-22
Tx_4         0.18803668 0.01878726 10.00873 2.94100e-23
Tn_1         0.04793497 0.01009758  4.74718 2.14876e-06
PR_5        -0.11477077 0.01761277 -6.51634 8.27928e-11
SeqPR_5      0.84871330 0.15468823  5.48661 4.40019e-08
PR_9         0.06280281 0.01188418  5.28457 1.34017e-07
ETP_4       -0.39456772 0.06637217 -5.94478 3.05127e-09
ETP_9        0.86411128 0.13946684  6.19582 6.50018e-10
ETP_6        0.29647152 0.05811329  5.10161 3.55408e-07
PR_6         0.07309904 0.01751799  4.17280 3.08491e-05
RV_3        -0.01080616 0.00175516 -6.15678 8.29936e-10
RV_8         0.01172887 0.00245509  4.77736 1.85198e-06
RV_9        -0.00995535 0.00234016 -4.25414 2.15571e-05
PR_3        -0.07175832 0.01563166 -4.59058 4.58191e-06
Tn_4        -0.12359457 0.02167997 -5.70086 1.29481e-08
Tx_3        -0.04429322 0.01185866 -3.73509 1.90755e-04
Tn_2         0.02499665 0.00734931  3.40122 6.78676e-04
ETP_7       -0.69805078 0.12483751 -5.59168 2.42829e-08
RV_7         0.00807720 0.00234322  3.44705 5.73690e-04
SeqPR_7     -0.63847277 0.18014105 -3.54429 3.99023e-04
Tn_7         0.05302557 0.01456662  3.64021 2.76513e-04
Tx_8        -0.04023110 0.01350829 -2.97825 2.91966e-03
```

Question 6

D'après les résultats ci-dessus, quel est le mois au cours duquel le niveau d'évapotranspiration semble le plus influencer l'anomalie de rendement ? Une forte anomalie de rendement est-elle plutôt associée à une faible ou une forte évapotranspiration lors de ce mois ?

Réponse

Question 7

Comment expliqueriez-vous à un.e non-statisticien.ne qu'il faut préférer le modèle `select` *au modèle* `fullmod` *?*

Réponse

On compare maintenant le modèle sélectionné au modèle complet, par un test d'analyse de la variance :

```
anova(select,fullmod)
```

```
Analysis of Variance Table

Model 1: yield_anomaly ~ PR_7 + PR_8 + ETP_3 + Tx_6 + Tx_5 + ETP_8 + RV_1 +
    Tx_4 + Tn_1 + PR_5 + SeqPR_5 + PR_9 + ETP_4 + ETP_9 + ETP_6 +
    PR_6 + RV_3 + RV_8 + RV_9 + PR_3 + Tn_4 + Tx_3 + Tn_2 + ETP_7 +
    RV_7 + SeqPR_7 + Tn_7 + Tx_8
Model 2: yield_anomaly ~ ETP_1 + ETP_2 + ETP_3 + ETP_4 + ETP_5 + ETP_6 +
    ETP_7 + ETP_8 + ETP_9 + PR_1 + PR_2 + PR_3 + PR_4 + PR_5 +
    PR_6 + PR_7 + PR_8 + PR_9 + RV_1 + RV_2 + RV_3 + RV_4 + RV_5 +
    RV_6 + RV_7 + RV_8 + RV_9 + SeqPR_1 + SeqPR_2 + SeqPR_3 +
    SeqPR_4 + SeqPR_5 + SeqPR_6 + SeqPR_7 + SeqPR_8 + SeqPR_9 +
    Tn_1 + Tn_2 + Tn_3 + Tn_4 + Tn_5 + Tn_6 + Tn_7 + Tn_8 + Tn_9 +
    Tx_1 + Tx_2 + Tx_3 + Tx_4 + Tx_5 + Tx_6 + Tx_7 + Tx_8 + Tx_9
  Res.Df  RSS Df Sum of Sq     F   Pr(>F)
1   3365 2389
2   3339 2347 26     42.33 2.316 0.000173 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 8

Comment expliqueriez-vous l'hypothèse nulle du test ci-dessus à un.e non-staticien.ne ?

Réponse

Question 9

Pourquoi la colonne `Df` du tableau ci-dessus contient la valeur 26 (à quoi correspond cette valeur) ?

Réponse

Question 10

Selon vous, le résultat ci-dessus doit-il nous conduire à enlever des variables sélectionnées ou au contraire à en ajouter ? (Justifiez cotre réponse)

Réponse