Le machine learning :

une alternative aux régressions en SHS ?

Rencontres « Modèles et Apprentissage
en Sciences Humaines et Sociales »

25 mai 2018, Paris

Nicolas Robette

Laboratoire de Sociologie Quantitative

(CREST-ENSAE)

Plan

a) Pourquoi chercher une alternative aux régressions ?

b) Pourquoi le « machine learning », et les forêts aléatoires en particulier ?

c) Deux exemples sur le cinéma en France

   a) La relation entre critique « intellectuelle » et succès public

   b) Les déterminants de la « survie » des réalisateurs

Les régressions

- Un outil central des analyses statistiques en sciences sociales
- Régression linéaire : « inventée » en 1897 (Yule); en sociologie à partir des 50's (Angell en 1951 aux USA ?)
- Generalized Linear Model (Nelder & Wedderburn 1972), dont la régression logistique (Berkson 1944) : depuis les 70's (80's en France)

Différents usages (1)

- Description
- Explication
- Prédiction

Différents usages (2)

- Usage « mesuré » : étudier les effets de structure
- Usage « métrologique » : hiérarchiser les facteurs explicatifs
- Usage « hyper-métrologique » : mesurer effet « pur » d'une variable, évaluation publique

[voir Bry, Robette, Roueff, 2015]

Critiques « substantielles »

- « Net effect thinking »
- Univocité des facteurs causaux
- Indépendance des individus statistiques
- « *toute* chose égale par ailleurs »
- Réversibilité des facteurs causaux

[voir Bry, Robette, Roueff, 2015]

Hypothèses du modèle linéaire (Gujarati)

1) linéarité des paramètres
2) X est non-stochastique [X values are fixed in repeated sampling]
3) erreurs de moyenne 0
4) homoscédasticité (variance égale) des erreurs
5) pas d'auto-corrélation des erreurs (covariance nulle)
6) covariance nulle entre X et erreurs
7) nb d'observations supérieur au nombre de paramètres (i.e. au nombre de variables explicatives)
8) variabilité des valeurs de X
9) spécification correcte du modèle
10) pas de multi-colinéarité parfaite
11) normalité des erreurs

Hypothèses GLM

1) linéarité (entre logits et X, mais pas entre Y et X)
2) additivité
3) erreurs de moyenne 0 ???
4) ~~homoscédasticité (variance égale) des erreurs~~
5) covariance nulle entre X et erreurs ???
6) effectifs suffisants dans les cases (adequate cell sample sizes), séparation complète
7) spécification correcte du modèle
8) pas de multicolinéarité
9) ~~normalité des erreurs~~
10) pas d'auto-corrélation des erreurs, i.e. indépendance des observations
11) distribution binomiale
12) over/underdispersion
13) adéquation de la fonction de lien
14) autre : sensibilité aux outliers, leverage et influencial observations ?

Inférence fréquentiste: les tests de significativité

- Pratique ultra-dominante…
- … mais débats récurrents [cf Poitevineau 2004] :
  - interprétations erronées et mauvais usages
  - limites et critiques

Interprétations erronées et mauvais usages

a. p-value = probabilité concernant l'hypothèse
   i. « Valid research hypothesis fantasy »
   ii. « Odds-Against-Chance Fantasy »
b. « Replicability or reliability fantasy »
c. Significativité statistique vs substantielle
d. Intensité de l'effet
e. Non-significativité et absence d'effet
f. "Statistical significance filter"
g. Comparaison de résultats
h. …

Limites et critiques

a. Nombreuses mauvaises interprétations
b. Raisonnement non naturel, contre-intuitif
c. Hypothèse nulle toujours fausse, donc inutile ; paradoxe fondamental
d. Décision binaire
e. Choix du seuil arbitraire
f. p ne mesure pas le degré de désaccord / certitude
g. Rien sur l'intensité de l'effet
h. Ne s'applique qu'aux échantillons aléatoires
i. Sensible à la taille de l'échantillon
j. Hypothèses paramétriques
k. Biais de publication
l. « Hunting with a shotgun »
m. …

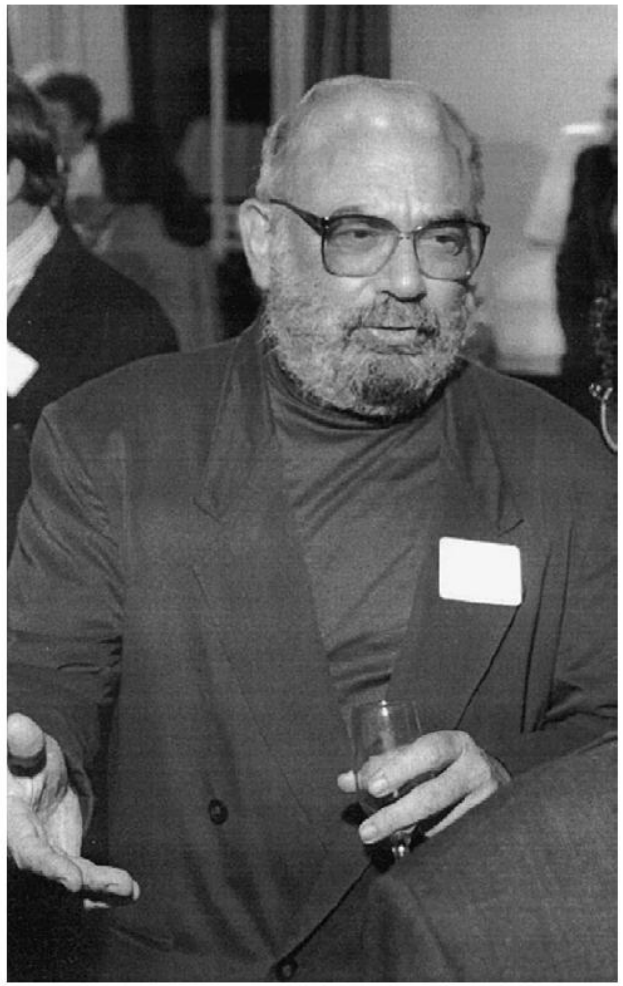
Quelles alternatives ?

- Pour l'inférence fréquentiste :
  - Régressions avec estimations *bootstrap*
  - Régressions bayésiennes
- Plus généralement :
  - Apprentissage automatique (*machine learning*)

Machine learning

- Apprentissage « supervisé » : plus proches voisins, réseaux de neurones, SVM, arbres et forêts aléatoires… (et GLM !)
    - « classification »
    - « régression »
- Apprentissage « non-supervisé » :
    - Clustering (apprentissage « non-supervisé ») : CAH, k-means…
    - Réduction de dimensions : ACP, ACM…
- …

Leo Breiman, 1928 - 2005



1954: PhD Berkeley (mathematics)

1960 -1967: UCLA (mathematics)

1969 -1982: Consultant

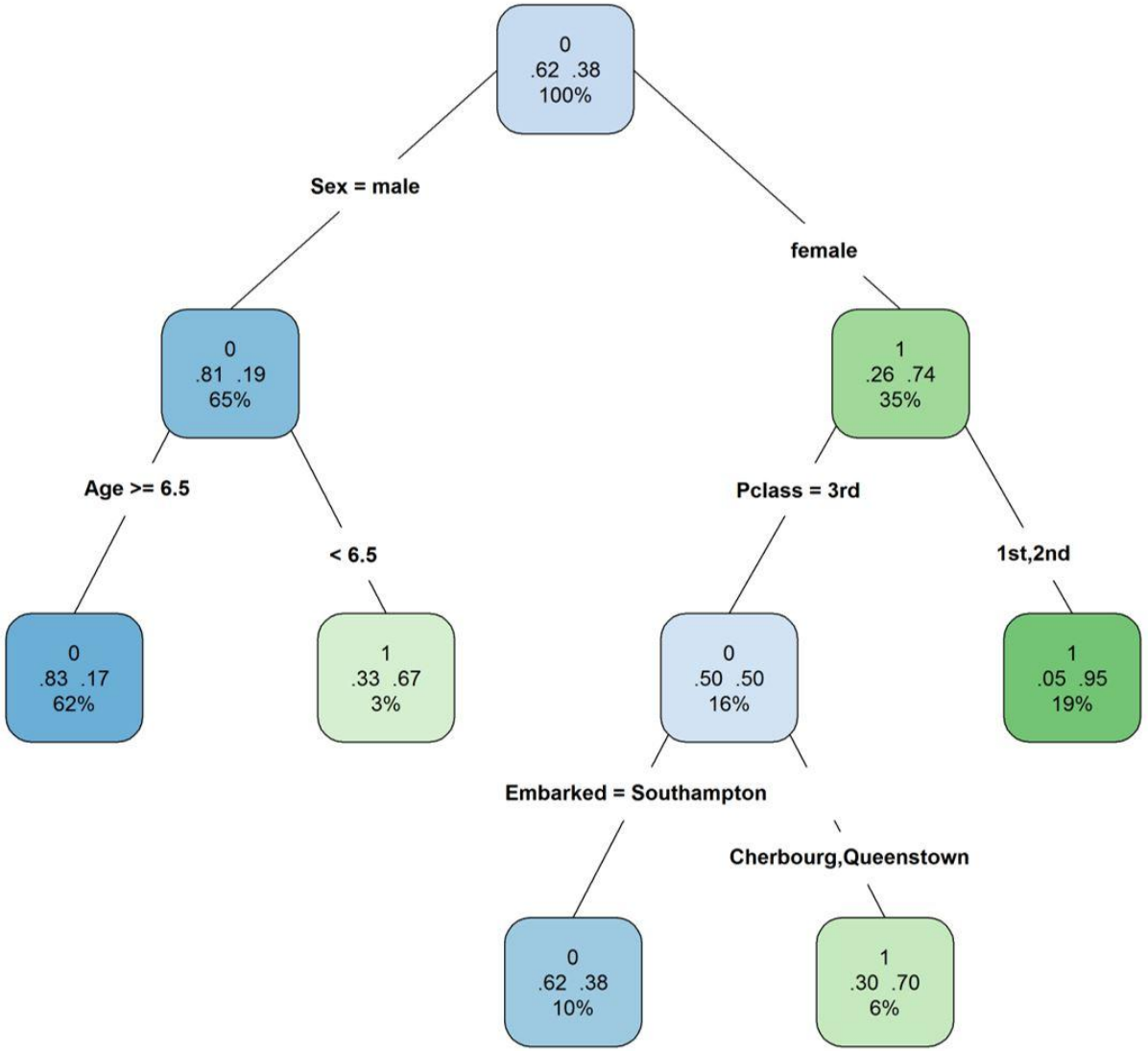
1982 - 1993 Berkeley (statistics)

1984 “Classification & Regression Trees”
(with Friedman, Olshen, Stone)

1996 “Bagging”

2001 “Random Forests”

Exemple d'arbre (*données Titanic)*



Classification & Regression Trees (CART)

- Au niveau du nœud initial (racine), on “découpe” / “segmente” / sépare (*split*) les individus en deux sous-groupes, qui forment des nœuds « filles » (*“daughter” nodes*).
- Puis chacun de ces sous-groupes est à son tour séparé, etc.
- L’objectif est de construire des sous-groupes les plus « homogènes » du point de vue de la variable à expliquer.
- Il s’agit donc d’un algorithme récursif de découpage / partition de l’espace des données en sous-régions homogènes en terme de classe.

Avantages

• Applicables aux régressions et aux classifications, autrement dit la variable à expliquer peut être continue ou catégorielle. Possibilité également de traiter des données censurées (modèles de durée, Cox, etc.), matrices de distance...

• Les **interactions** sont au cœur du modèle. De plus, on peut analyser des interactions non-linéaires (*highly non-linear interactions*), d'ordre élevé (*high order*).

• Très faciles à interpréter, lorsque l'arbre n'est pas trop grand.

- La représentation de l'arbre permet d'analyser quelles variables sont importantes, et où elles le sont (cf interactions).
- Les nœuds finaux (*terminal nodes*) suggèrent une partition naturelle des observations en groupes homogènes.

• Pas d'hypothèses sur les distributions statistiques (normalité, etc.) = non-paramétriques.

• Robuste face aux variables redondantes (cf multicolinéarité).

• La sélection des variables est automatique.

• Traitement indifférencié selon le type des variables explicatives, ie prennent en compte variables continues et catégorielles sans aucun souci.

• Robuste face aux données aberrantes ; solutions pour les données manquantes (cf *surrogate variables*), et non suppression comme dans les régressions

• Rapidité et capacité à traiter des très grandes bases.

Limites

- Précision : d'autres algorithmes ont de meilleures performances
- Stabilité : sensibles aux variations d'échantillons
- Frontières des découpages binaires pour les variables continues

Bagging (= Boostrap AGGregatING)

1. On construit un échantillon "*bootstrap*" à partir des données = tirage au sort de n observations, avec remise.
2. On construit un arbre à partir de cet échantillon.
3. On répète 1 et 2 un grand nombre de fois, souvent plusieurs centaines : on obtient donc un ensemble d'arbres.
4. On combine / agrège ces arbres, par la valeur modale – i.e. le « vote » - (pour la classification) ou la moyenne (pour la régression).

Random Forest

1. Faire pousser un arbre à partir d'un échantillon *bootstrap* des données de départ (ie d'apprentissage).
2. À chaque nœud:
   i. Sélectionner *mtry* variables au hasard parmi les *M* variables possibles (tirage au sort indépendant à chaque nœud).
   ii. Trouver la meilleure segmentation à partir de ces *mtry* variables.
3. Faire pousser l'arbre à sa profondeur maximale (pas d'élagage).
4. Reproduire les étapes 1 à 3 un grand nombre de fois (500, par exemple)
5. Combiner les arbres par vote/moyenne pour obtenir les valeurs prédites de chaque observation.

Avantages et limites

- Hérite de beaucoup des avantages de CART
- Avec en plus:
  - Qualité de prédiction
  - Stabilité
  - « Frontières » plus douces pour les variables explicatives continues
- Mais interprétation pas toujours aisée : pas de représentation graphique, ni d'arbre « moyen »

Hypothèses des régressions « classiques »?

1. linéarité : **NON** (la forme de la relation fonctionnelle entre Y et X (continue) est déterminée inductivement, sans hypothèse a priori)
2. distribution du terme d'erreur : **NON**
   i. moyenne=0
   ii. homoscédasticité
   iii. normalité
   iv. pas d'auto-corrélation, ie indépendance des observations
   v. covariance nulle avec X
3. problèmes de multicolinéarité : **NON**
4. adequate cell sample sizes: **NON**
5. inférence fréquentiste : **NON**

Les données « Travelling »

- Ensemble des longs-métrages sortis en salles en France entre 2000 et 2016 (N = 9772)
    - Convention avec le CNC
    - Appariement avec de multiples autres sources
    - Informations sur la production, la réception, les réalisateurs…

- Des caractéristiques particulières
    - Données exhaustives
    - Observations non indépendantes
    - Variables nombreuses (et multicolinéarité)

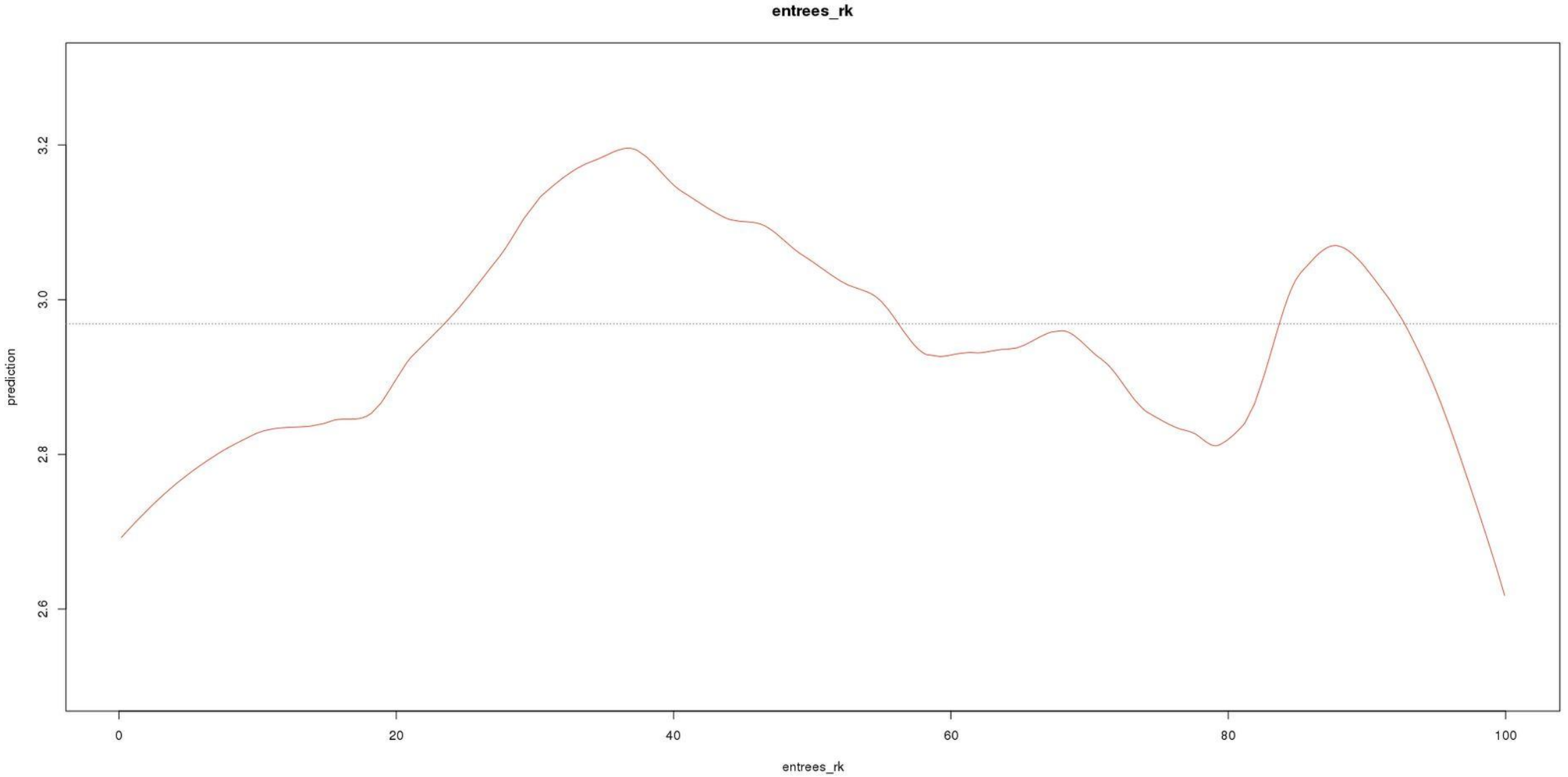
Exemple 1 : la critique est-elle snob ?

- Quel est le degré d'autonomie de la critique « intellectuelle » par rapport aux jugements du public ?

- Corrélation entre critique et nombre d'entrées = nulle

  (Pearson = -0,031 ; Spearman = -0,038)

« Effet » brut des entrées



Exemple 1 : la critique est-elle snob ?

- Quel est le degré d'autonomie de la critique « intellectuelle » par rapport aux jugements du public ?

- Corrélation entre critique et nombre d'entrées = nulle

  (Pearson = -0,031 ; Spearman = -0,038)

- Qu'en est-il lorsqu'on prend en compte la segmentation du marché ?

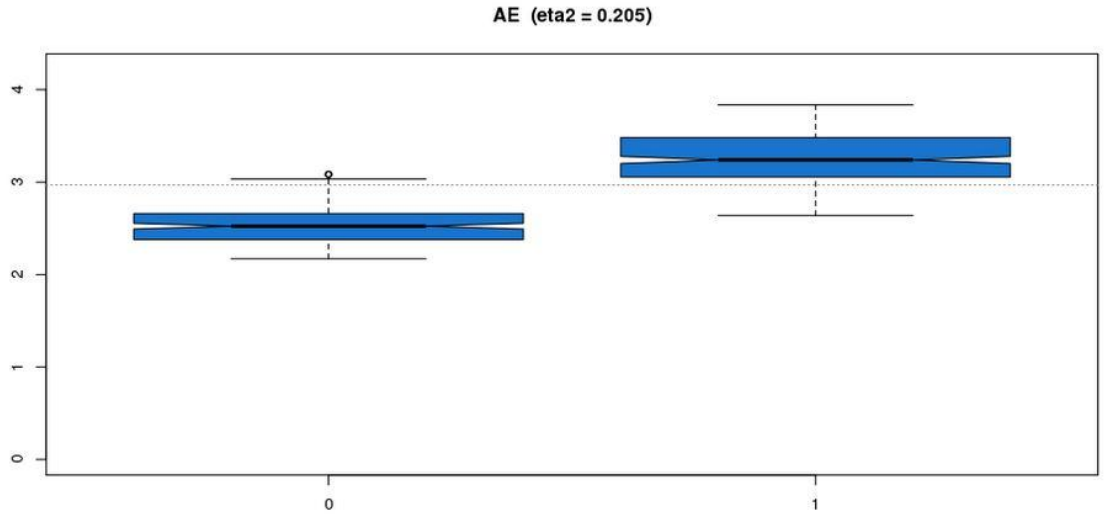
  Selon ressources économiques, film d'auteur / non, genre, origine géographique

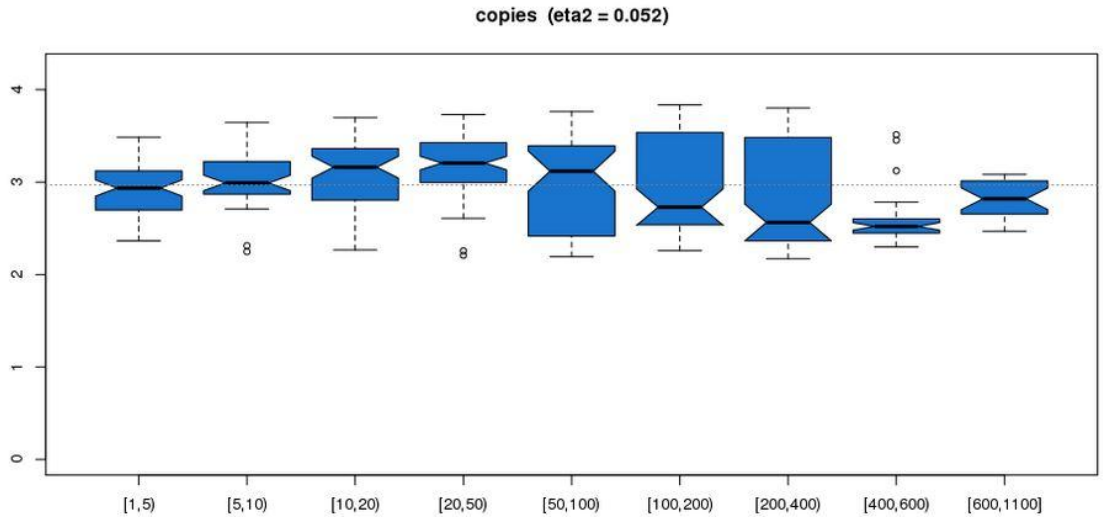
Exemple 1 : la critique est-elle snob ?

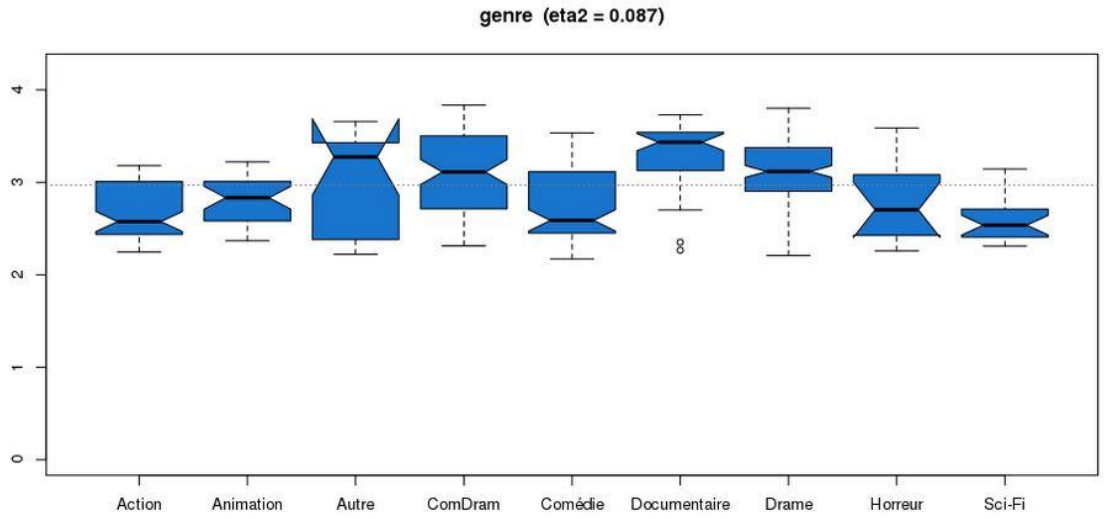
- Les données
  - N = 9772
  - Données exhaustives
  - Observations non-indépendantes
  - Multicolinéarité

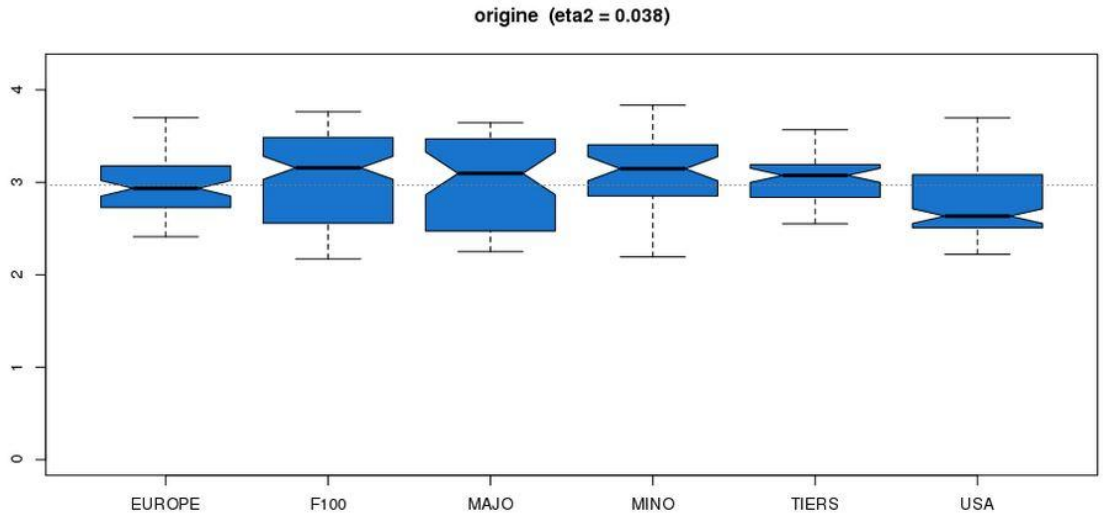
- Spécification du problème
  - Variable dépendante = critique « intellectuelle » (continue)
  - Variable indépendante principale = nombre d'entrées en salles (continue)
  - Variables indépendantes « de structure » = nombre de copies en première semaine (9 modalités), label « art et essai » (2), genre (9), origine géographique (6)

« Effets » bruts des variables de structure

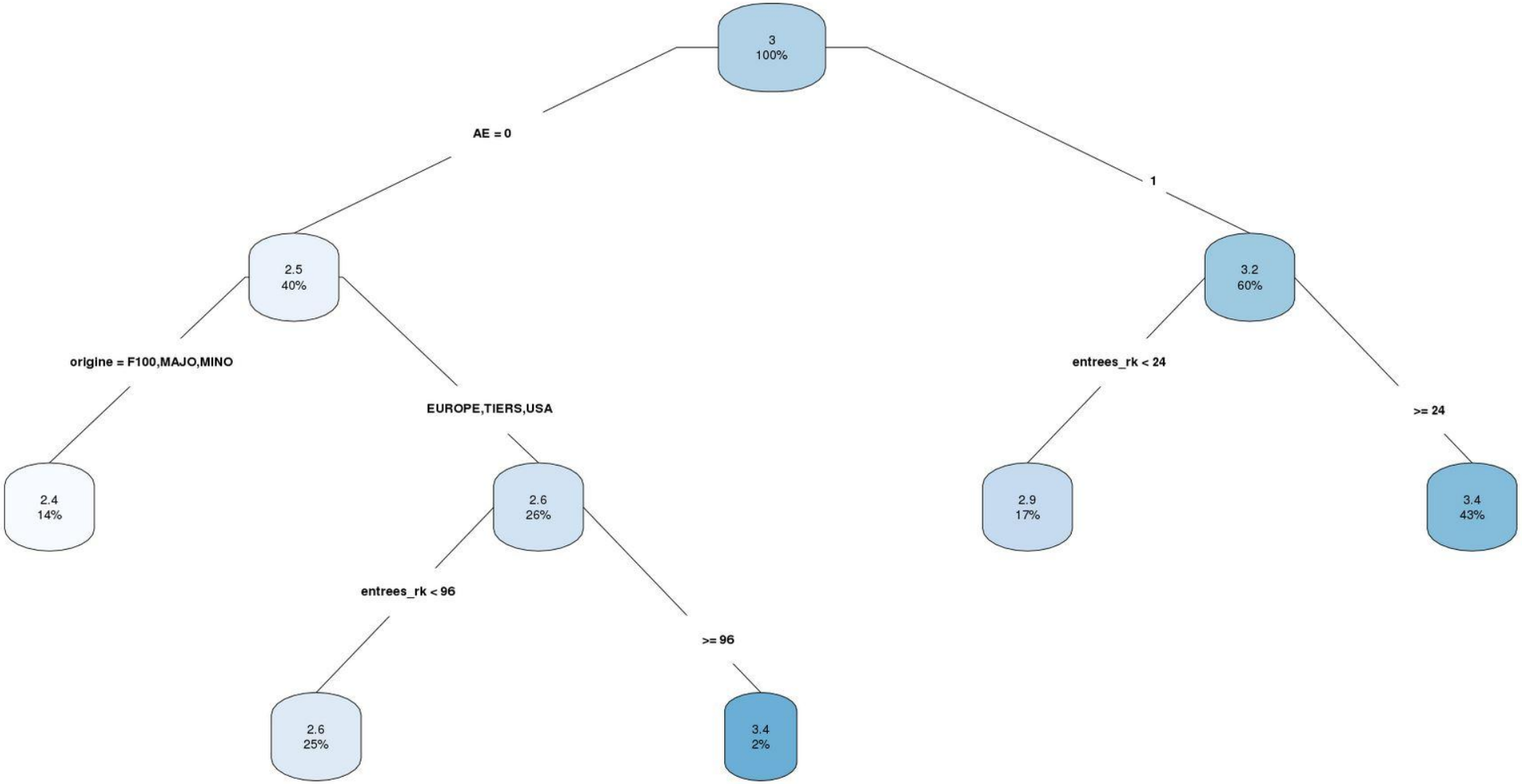




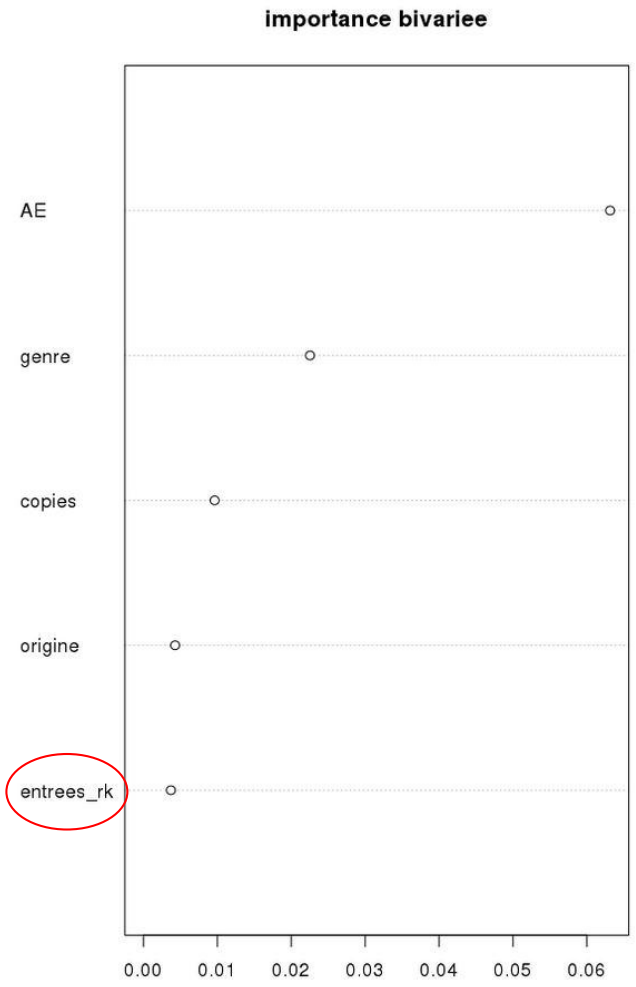




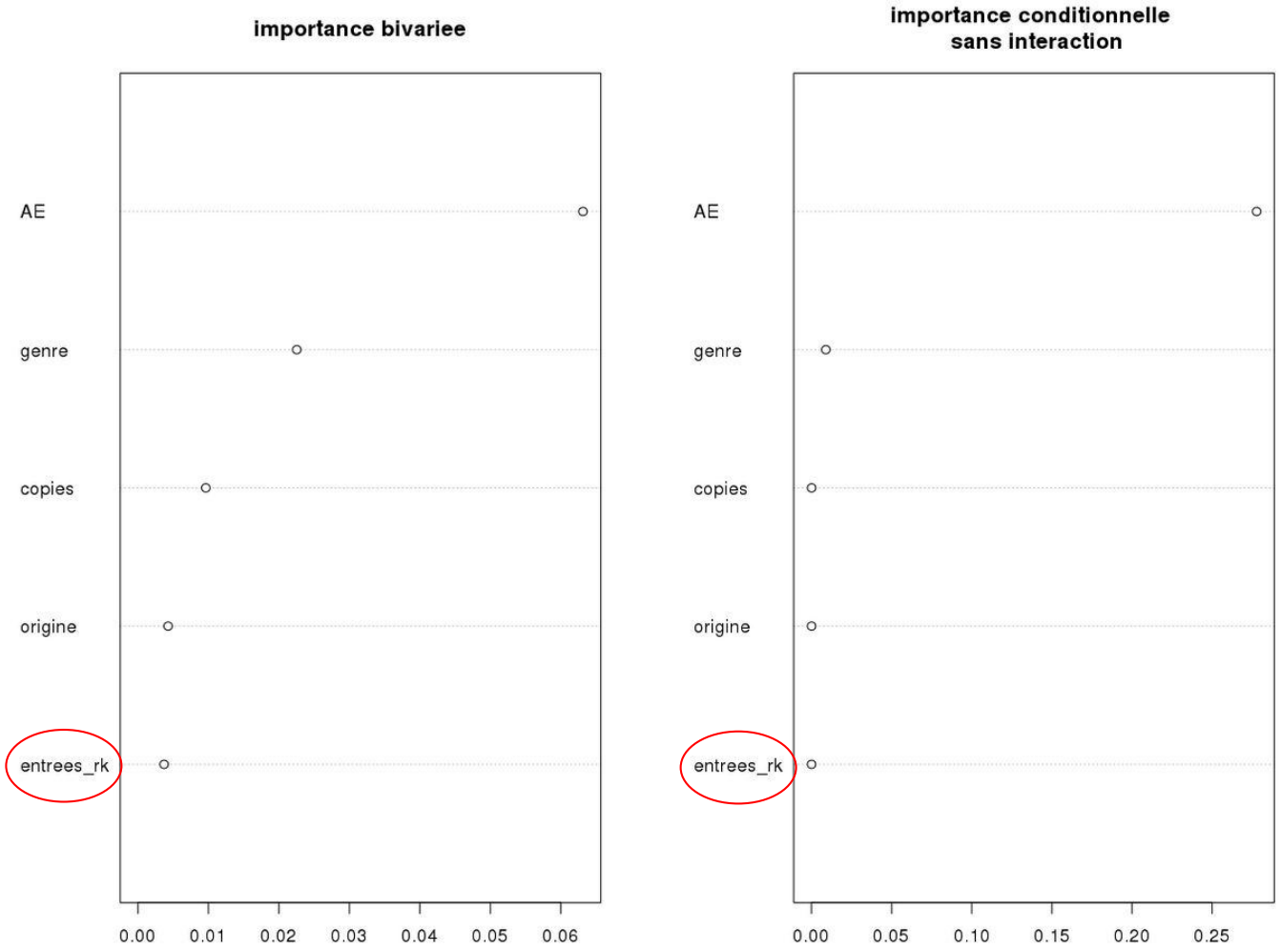
Arbre de régression



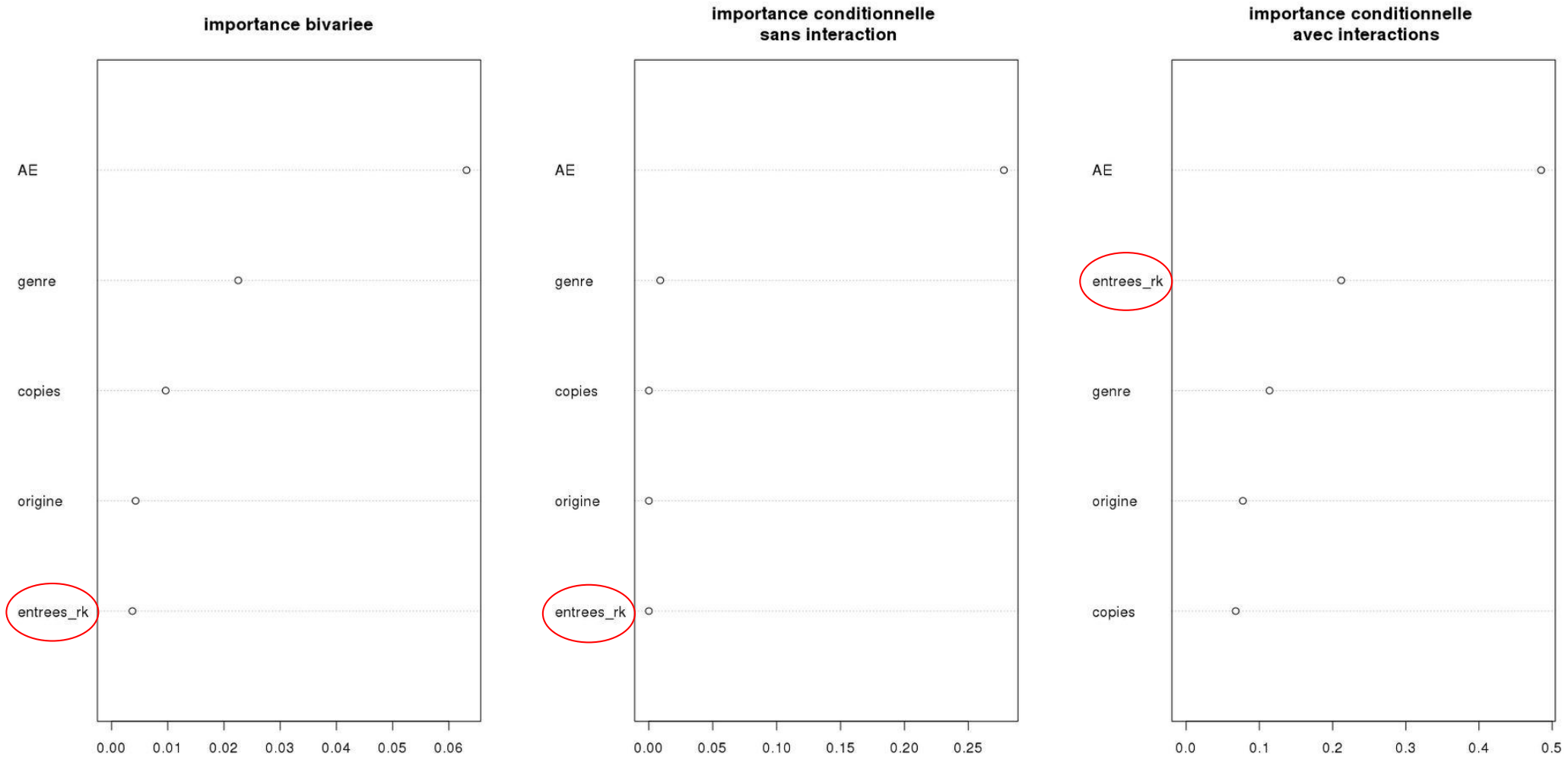
Importance des variables



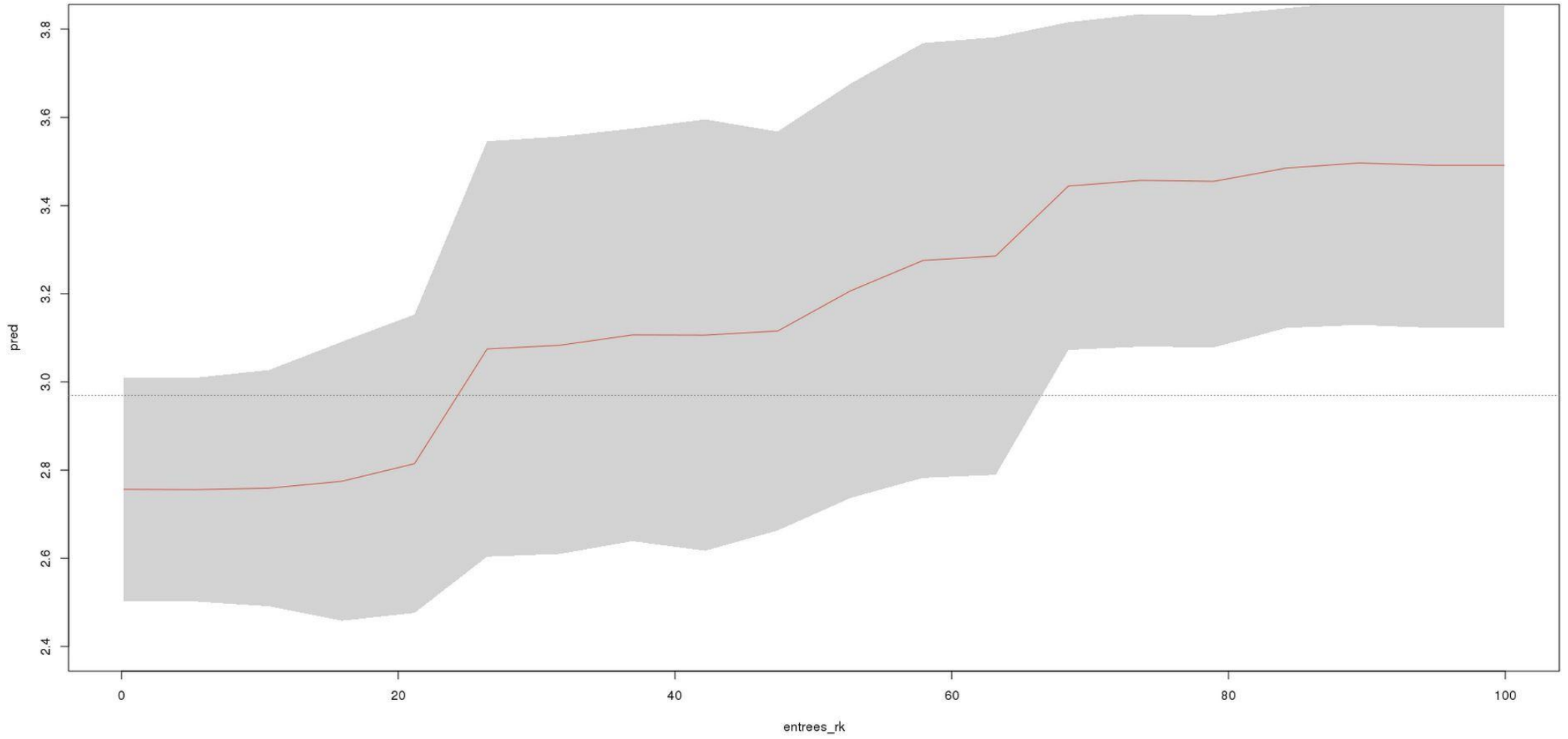
Importance des variables



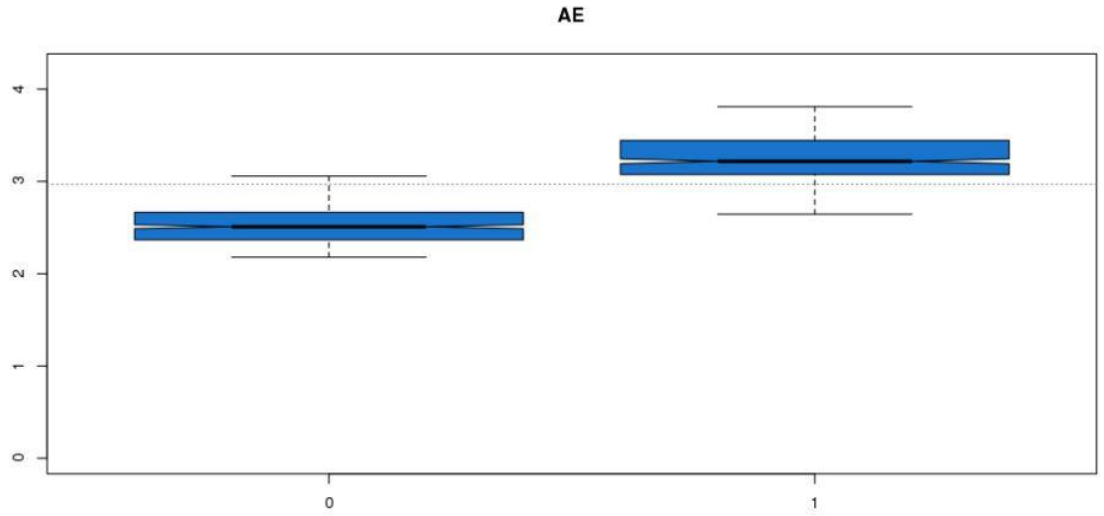
Importance des variables

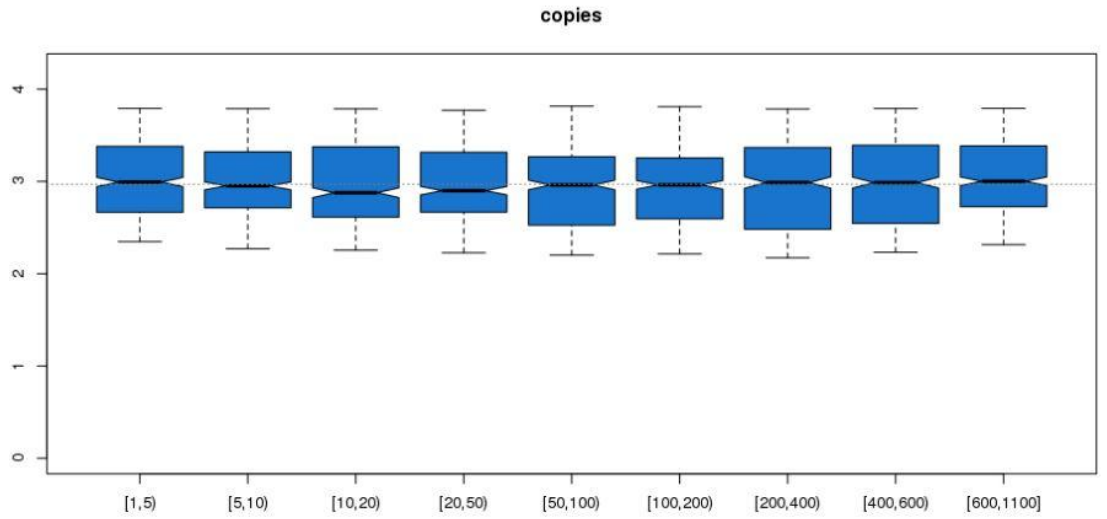


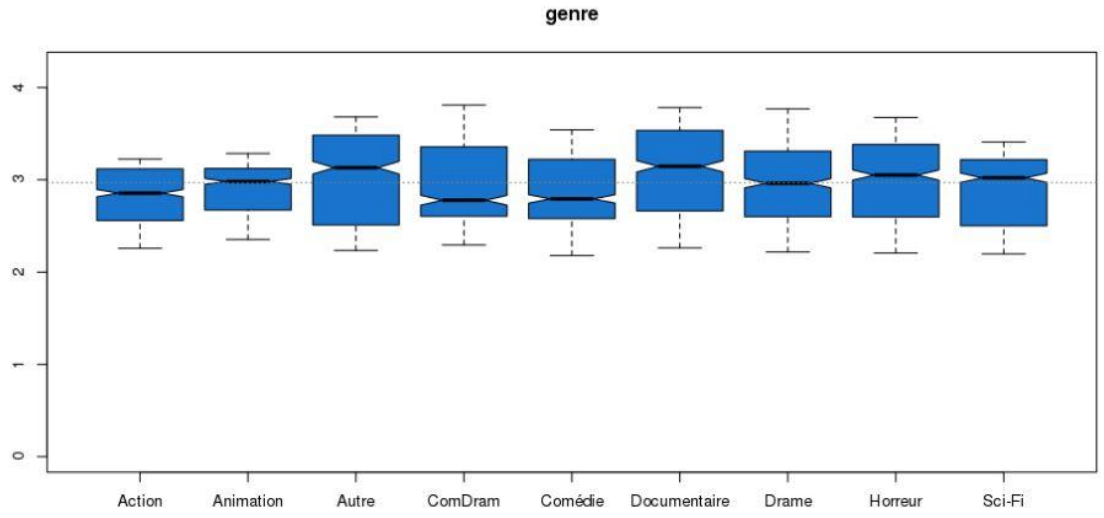
« Effet » conditionnel du nombre d'entrées

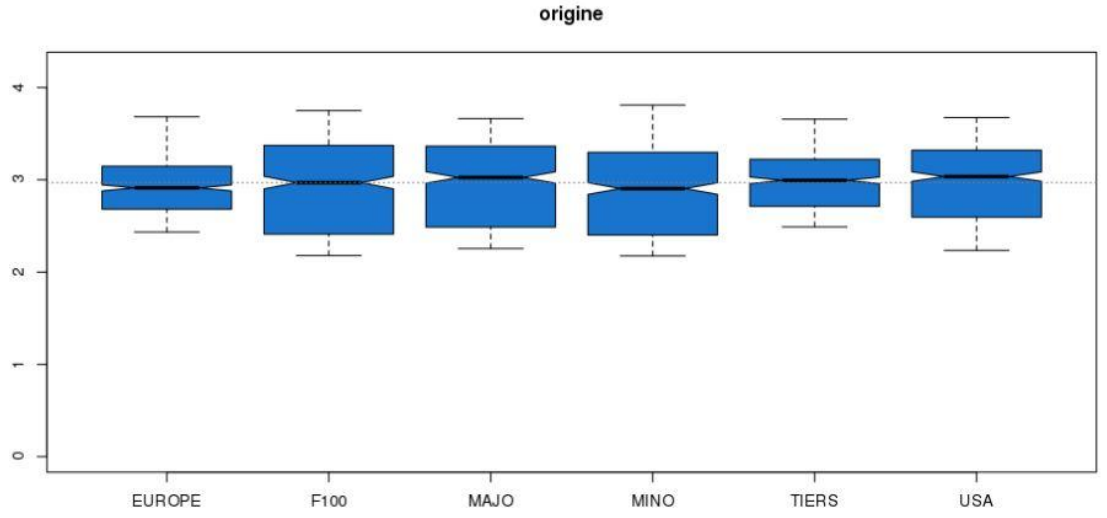


« Effets » conditionnels des variables de structure





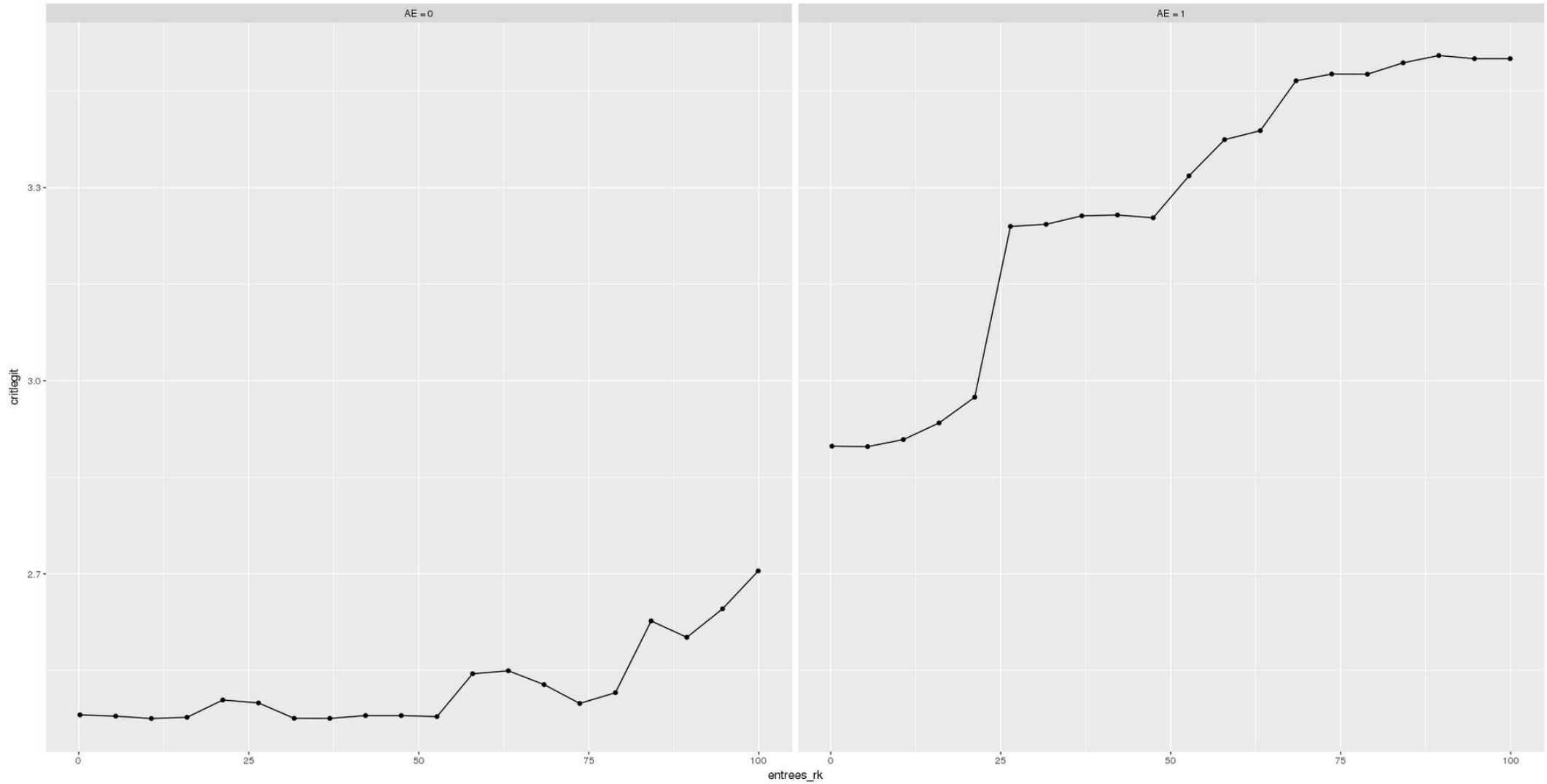




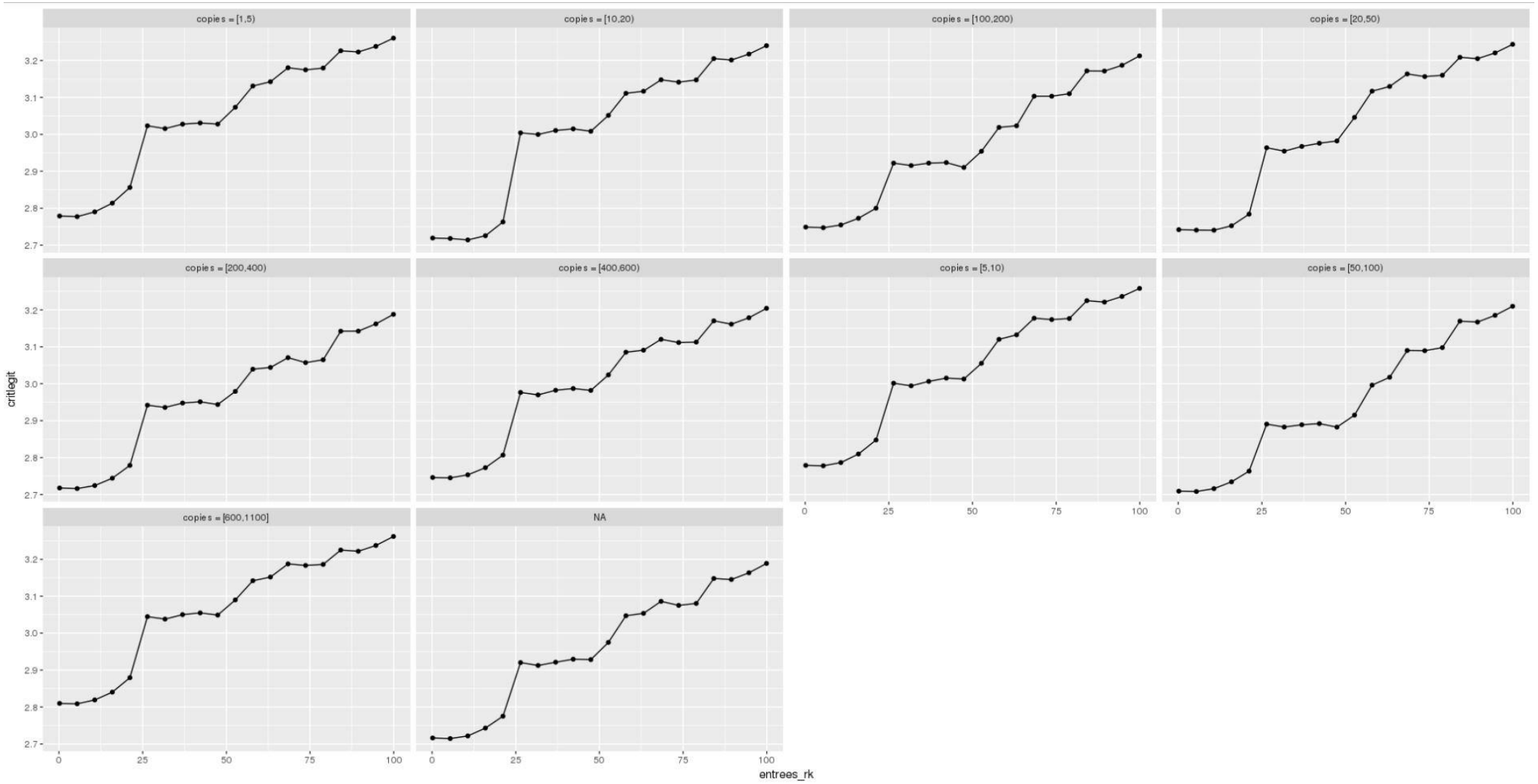
Interactions

V1	V2	Imp.V1	Imp.V2	Additive	Paired	Difference
AE	entrees_rk	-0.491	-0.209	-0.701	-0.570	0.131
entrees_rk	genre	-0.209	-0.107	-0.316	-0.229	0.087
AE	origine	-0.491	-0.075	-0.567	-0.507	0.059
entrees_rk	copies	-0.209	-0.062	-0.271	-0.226	0.045
AE	genre	-0.491	-0.107	-0.599	-0.636	-0.037
entrees_rk	origine	-0.209	-0.075	-0.284	-0.253	0.031
copies	genre	-0.062	-0.107	-0.169	-0.146	0.023
genre	origine	-0.107	-0.075	-0.182	-0.161	0.021
copies	origine	-0.062	-0.075	-0.137	-0.117	0.020
AE	copies	-0.491	-0.062	-0.554	-0.548	0.006

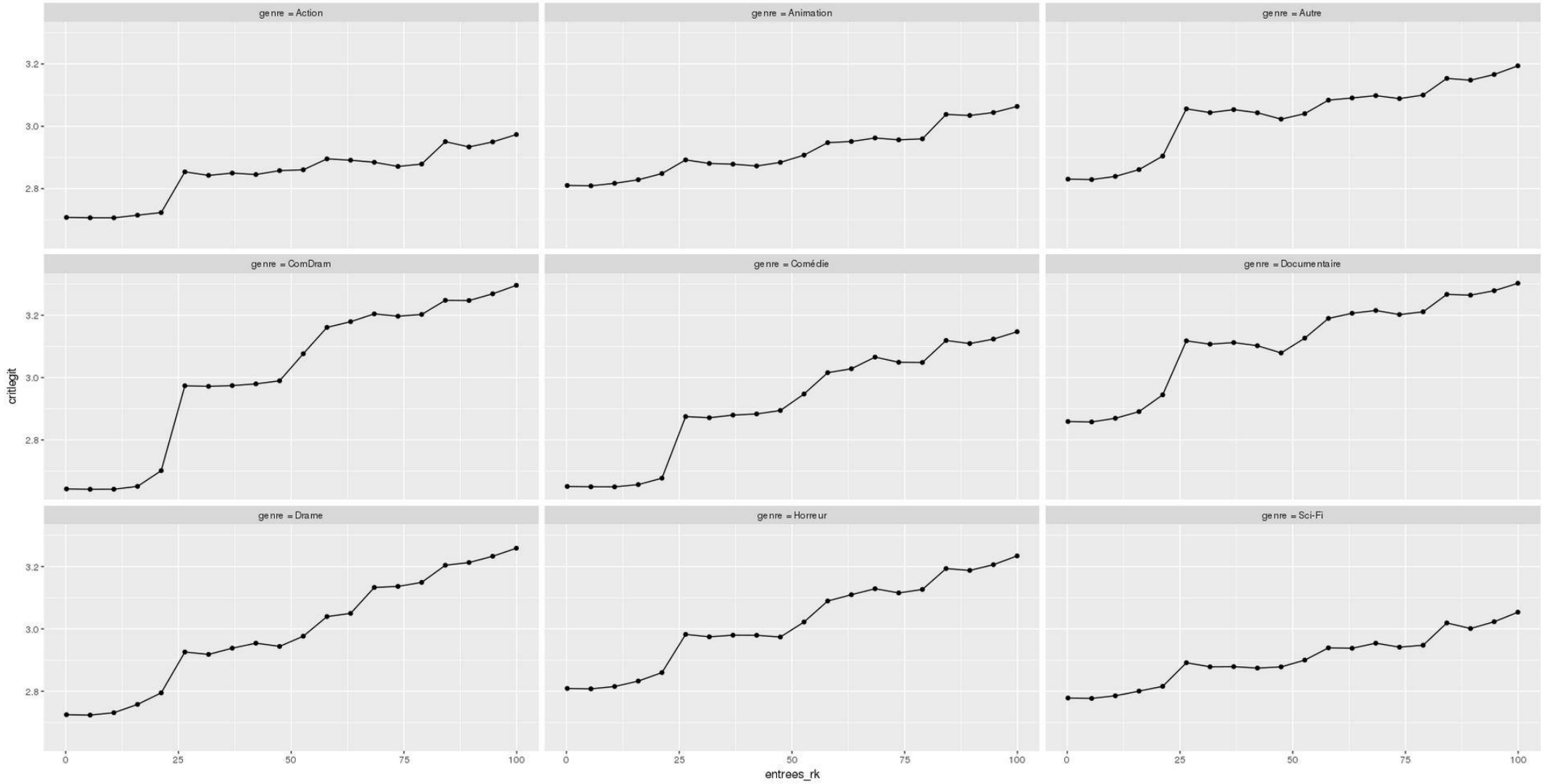
Interaction ‘‘AE * entrées’’



Interaction ‘‘copies * entrées’’



Interaction "genre * entrées"

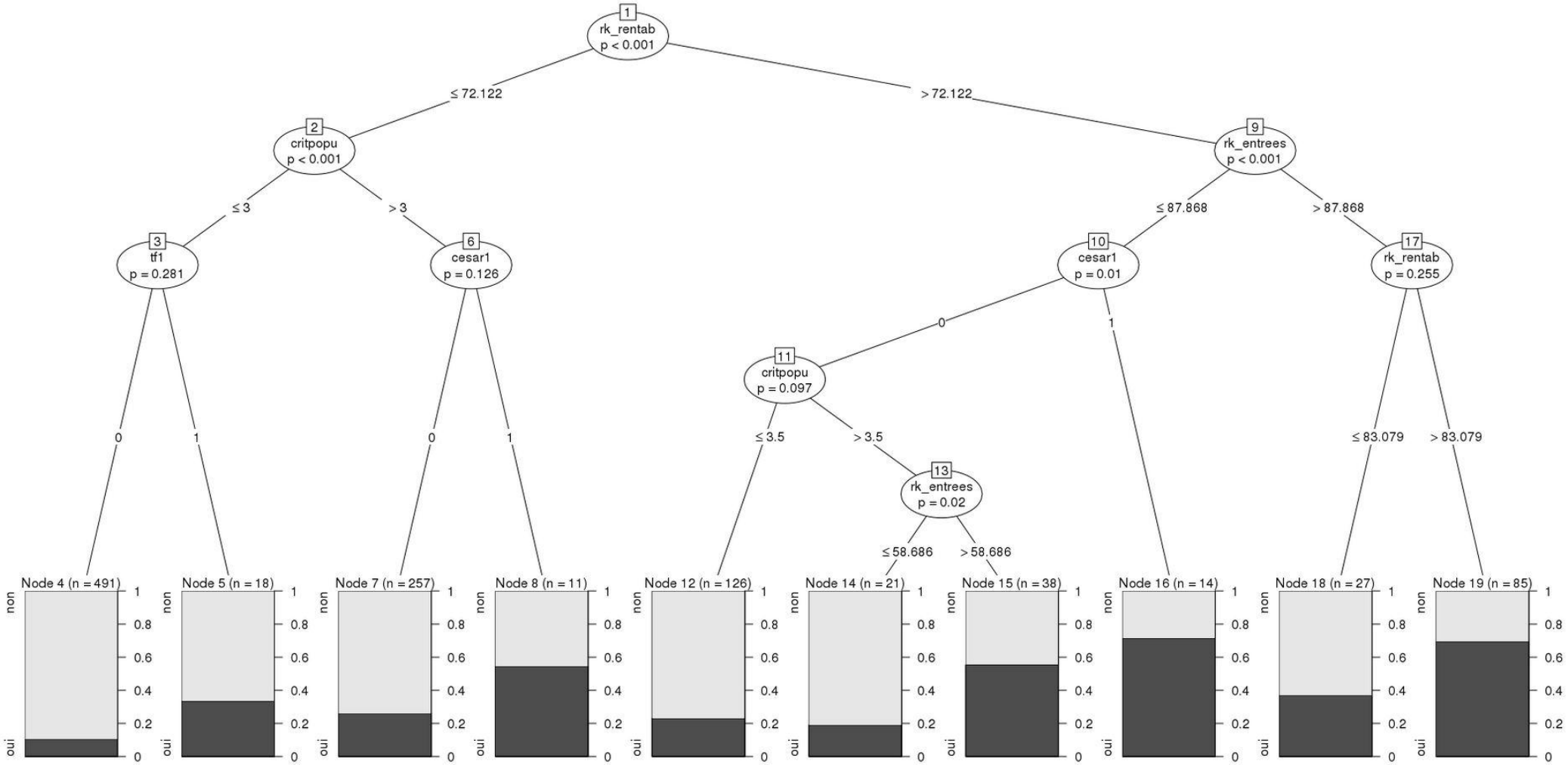


Exemple 2 : la survie des réalisateurs

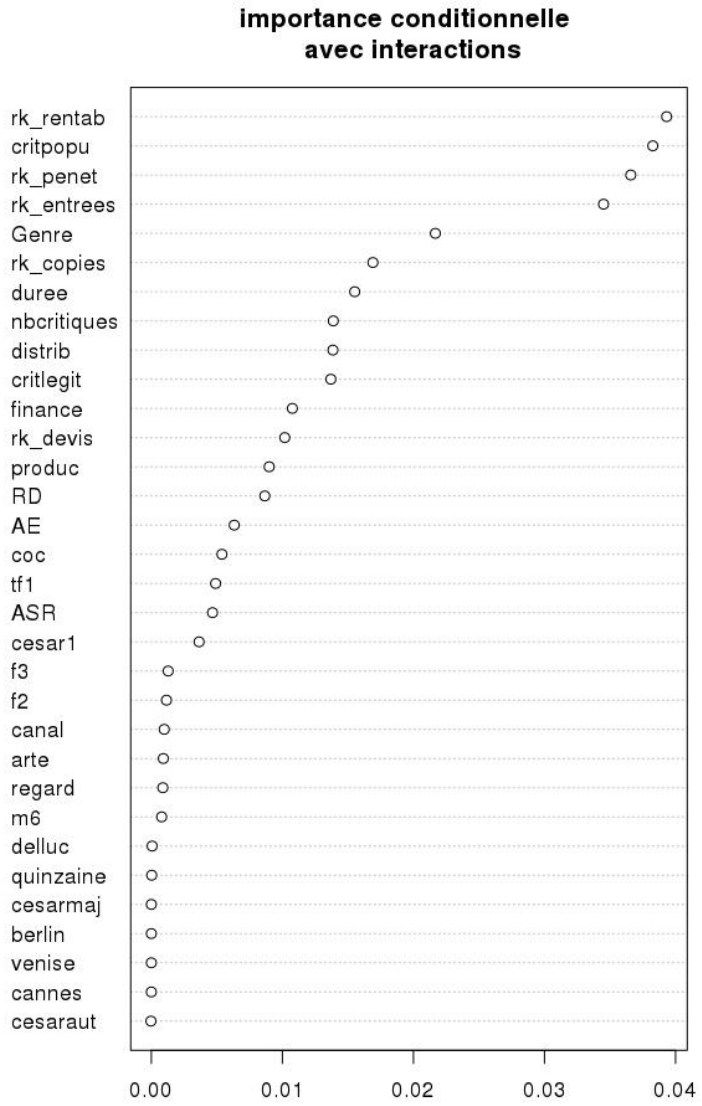
- Les données
  - Les 1ers films d'initiative française sortis en France entre 2000 et 2011 (N = 1088)
  - Données exhaustives
  - Observations non-indépendantes
  - **Nombreuses variables (M = 32)** et multicolinéarité

- Spécification du problème
  - Variable dépendante = réalisation d'un 2nd film dans les 5 ans (0/1)
  - 32 variables indépendantes (continues ou catégorielles) = caractéristiques du 1er film (production, réception)

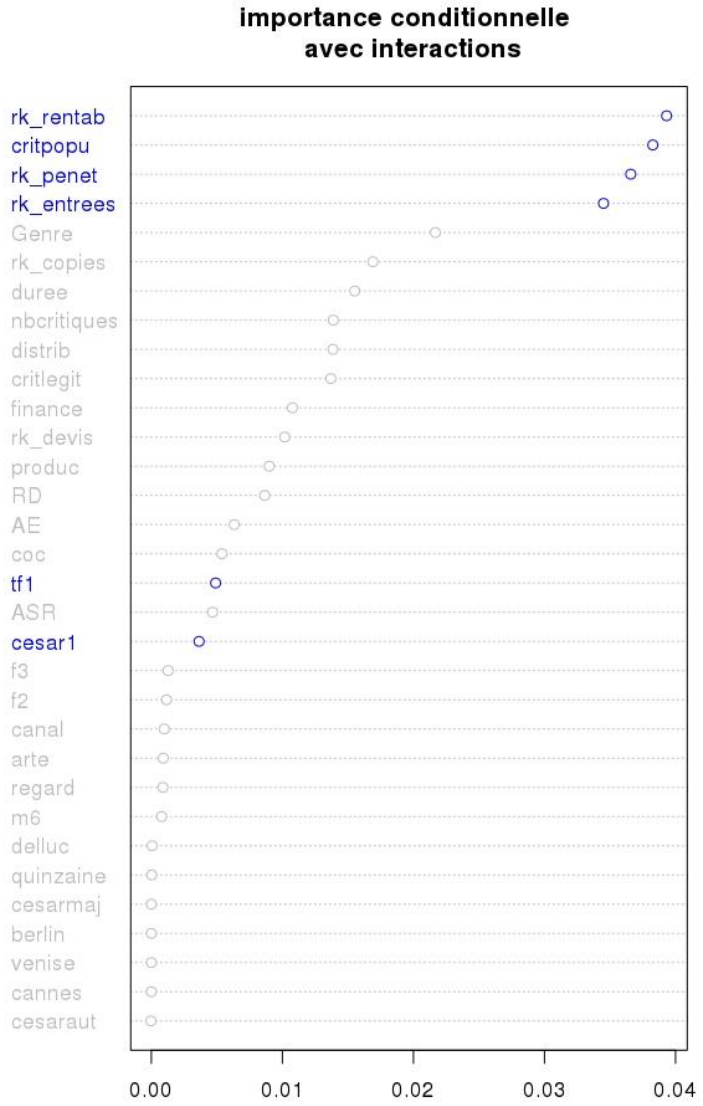
Arbre de classification



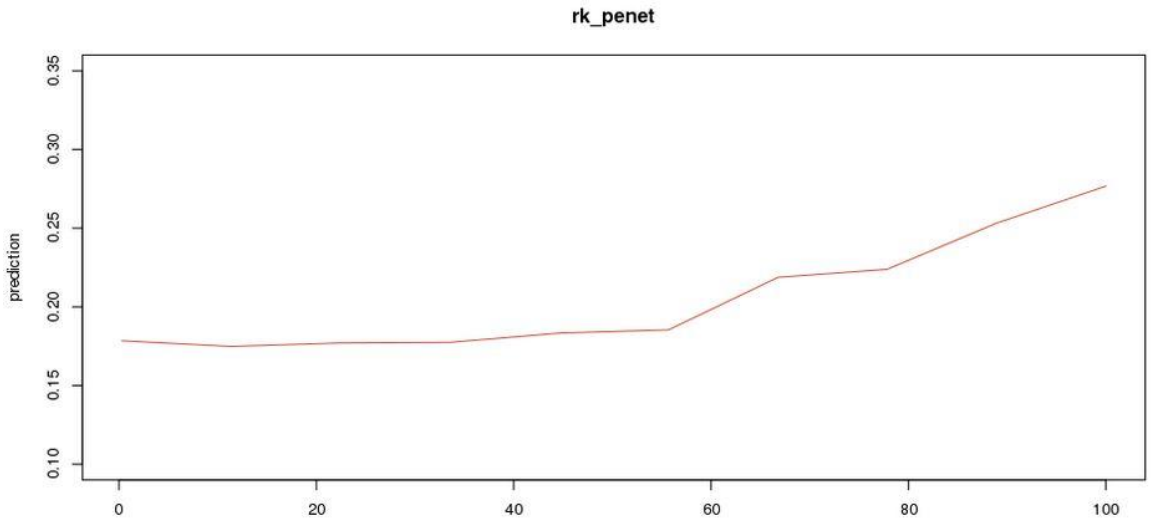
Importance des variables

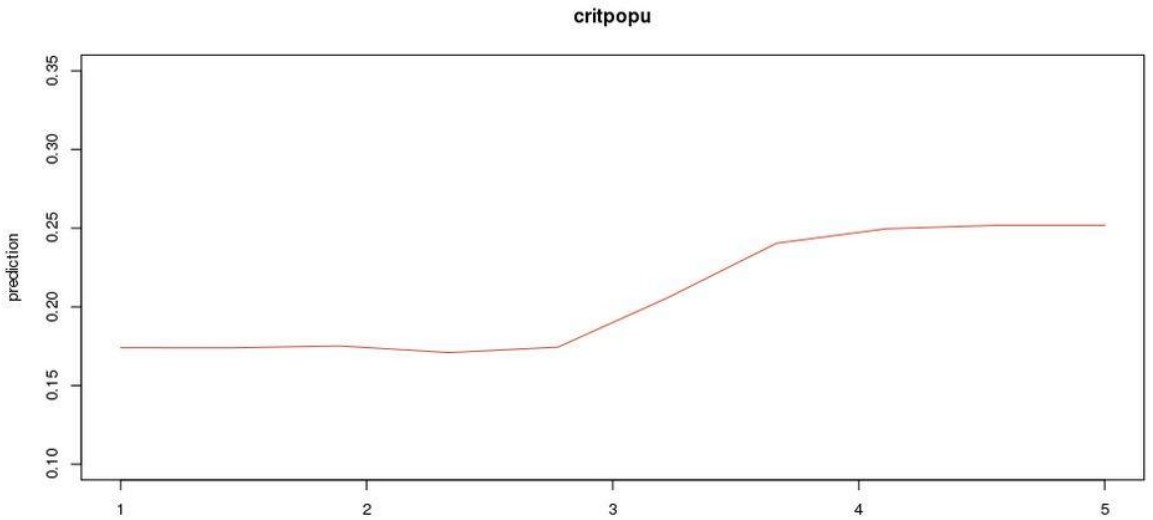


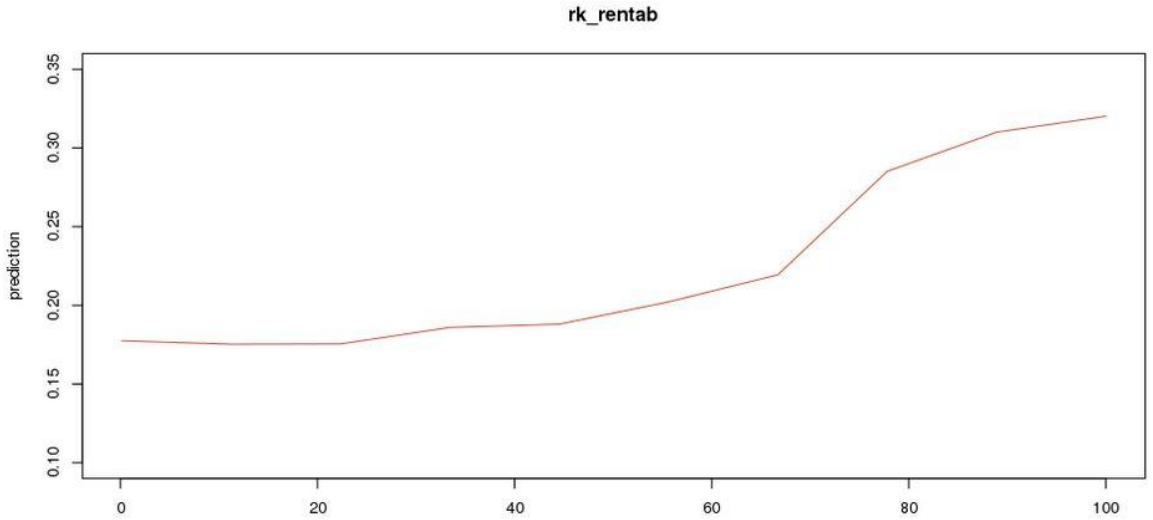
Importance des variables (après sélection)

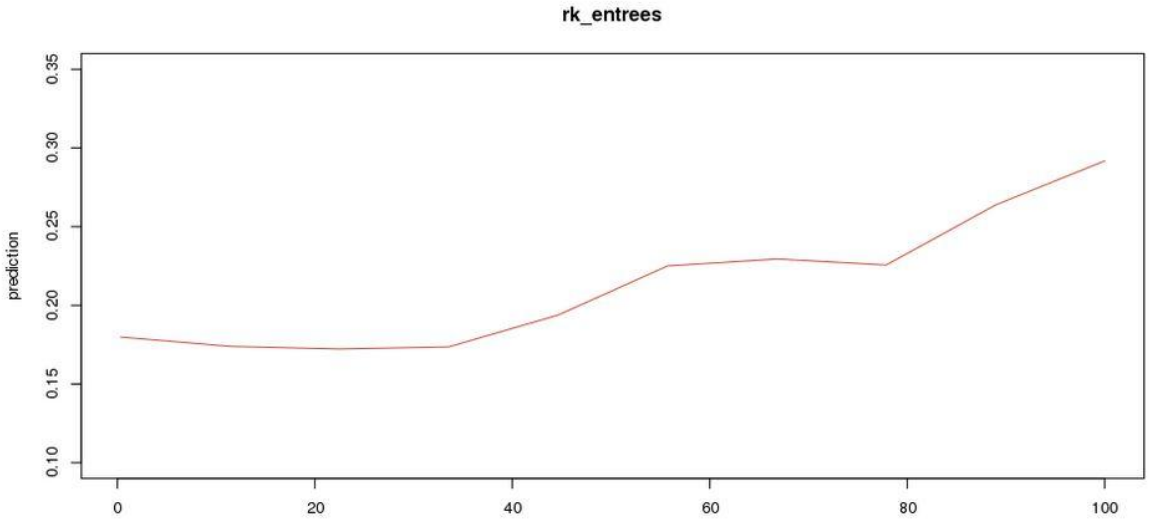


Principaux effets conditionnels









Au final

- Un **cadre unique** quel que soit le type de variable à prédire (continue, catégorielle, ou même censurée).
- Les effets des variables « candidates à l'explication » (i.e. prédictrices) ne sont pas indépendants par postulat, ils ne s'additionnent pas mais se conjuguent : les **interactions** sont au cœur du modèle.
- Pas d'**hypothèses paramétriques** (linéarité, normalité, hétéroscédasticité, etc.).
- On peut avoir un grand nombre de variables prédictrices ; pas de problème de **multicolinéarité**.
- Les interprétations ne sont pas fondées sur **des tests de significativité**, dont les nombreuses limites sont connues (sensibilité au nombre d'observations ; hypothèse d'un échantillonnage aléatoire ; etc.)
- Peu de **paramètres**, qui influent relativement peu sur la performance du modèle mais permettent de faire jouer le curseur entre effets bruts / conditionnels et entre présence / non d'interactions
- Des **outils d'interprétations** qui répondent aux principaux usages des régressions en SHS : arbres, « partial dependence » (effets conditionnels), importance (et sélection) des variables, détection d'interactions…

Références

• Leo Breiman, Jerome Friedman, Richard Olshen, Charles Stone (1984). *Classification and Regression Trees*. Wadsworth.

• Leo Breiman (1996). "Bagging Predictors". *Machine Learning*, 24, 123-140.

• Leo Breiman (2001). "Random Forests". *Machine Learning*, 45, 5-32.

• Site web de Breiman:

https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

• Présentation de Adele Cutler :

http://www.math.usu.edu/adele/RandomForests/UofU2013.pdf

• Carolin Strobl, James Malley, Gerhard Tutz (2009). "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests". *Psychological Methods*, 14(4), 323-348. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927982/

• Trevor Hastie, Rob Tibshirani, Jerome Friedman (2009). *Statistical Learning*. Springer.

Packages R

- **rpart** : arbres CART
- **rpart.plot** : représentations plus jolies des arbres CART
- **randomForest** : RF « à la Breiman »
- **randomForestSRC** : RF « à la Breiman » élargies aux modèles de durée + qq outils utiles (interactions, partial dependence plots...)
- **party** : RF avec « inférence conditionnelle »
- **adabag** : bagging, boosting
- **gbm** : boosting
- **mlr** : package « tout-inclus », possibilité de parallélisation, mais syntaxe un peu plus difficile