# Application de la méthode à un sondage agricole

# Origines des données

RHoMIS
Rural Household Multi-Indicator Survey

- L'enquête RHoMIS comprend plusieurs modules centraux sur les pratiques agricoles, les moyens de subsistance et la sécurité alimentaire.
- Il existe 30 modules facultatifs supplémentaires couvrant une gamme étendue de sujets.
- La durée typique pour compléter l'enquête de base est de 45 minutes.

Modèle du sondage

# Forme des données

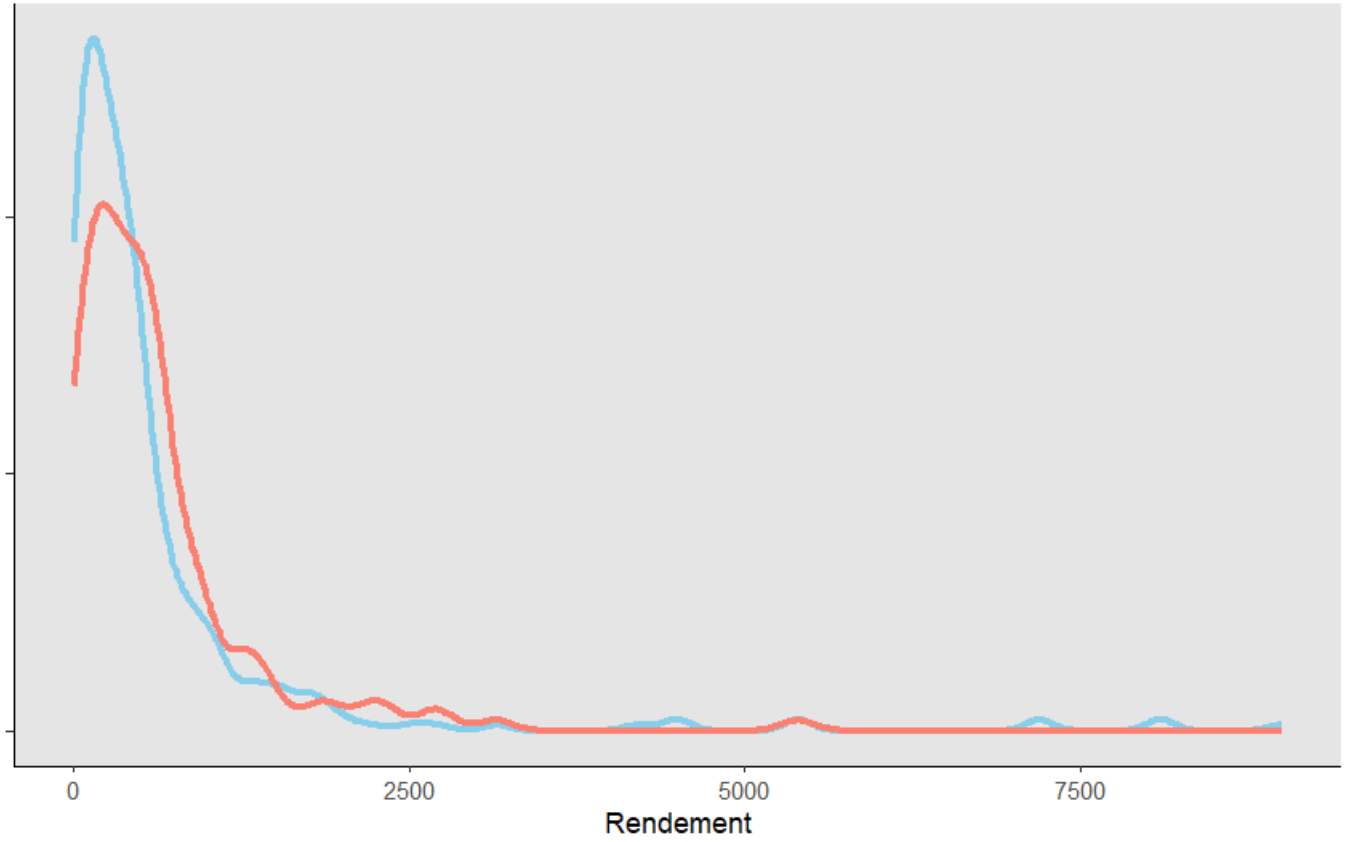
- Au départ :
  - 13310 Observations
  - 758 Variables
- Extraction d’un jeu de données selon deux critères principaux:
  - Culture principale du Maïs
  - Unité de production en Kg
- Après tri :
  - 461 Observations
  - 39 Variables

# Les 39 variables résumées :

- Caractéristiques du foyer :
  - Composition / Niveau éducation
- Exploitation :
  - Surface / Proportion dédiée à la culture principale / Utilité de la culture / Pratiques
- Indicateurs :
  - Accessibilité à la nourriture / Pauvreté / Aides

# Quel est l’effet de l’agroforesterie sur les rendements ?

# Distributions des rendements selon la pratique de l'agroforesterie




Welch Two Sample t-test

data: dfkg$crop_yield_1[dfkg$agroforestry == "y"] and dfkg$crop_yield_1[dfkg$agroforestry == "n"]

t = -0.22445, df = 411.26, p-value = 0.8225

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-194.9208 154.9705

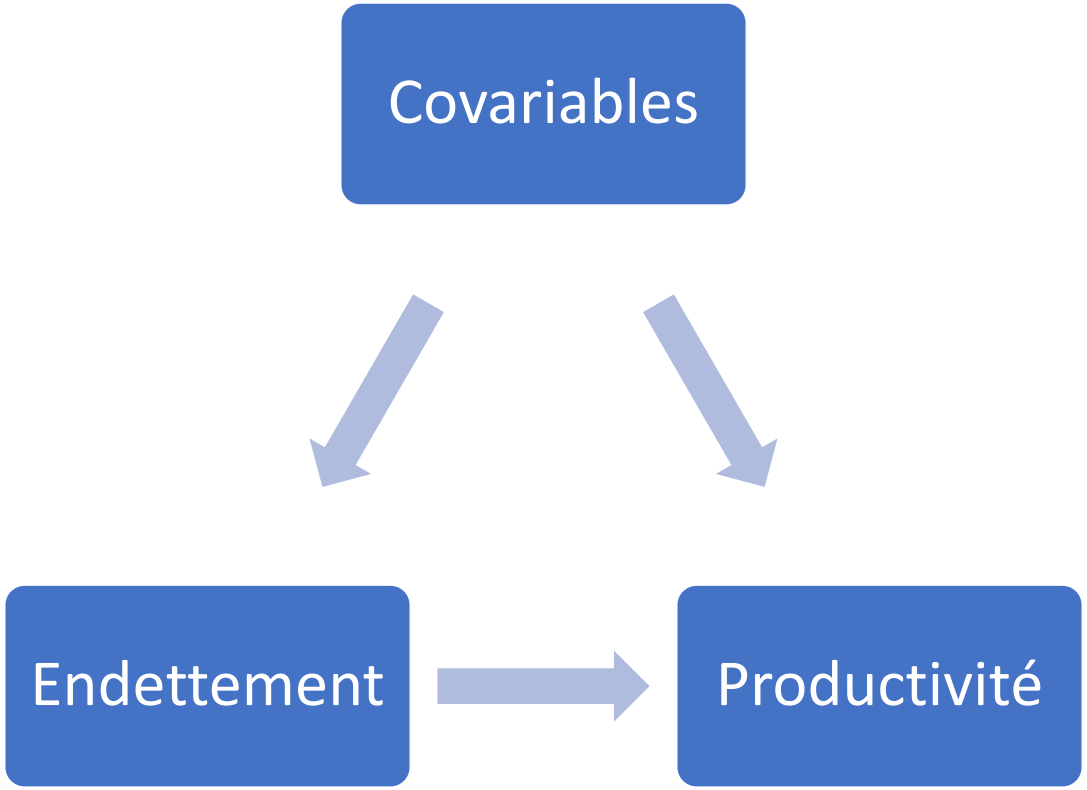
sample estimates:

mean of x mean of y

612.6214 632.5966

Avec agroforesterie : 140
Sans agroforesterie : 321
Total : 461

# Estimation du score de propension



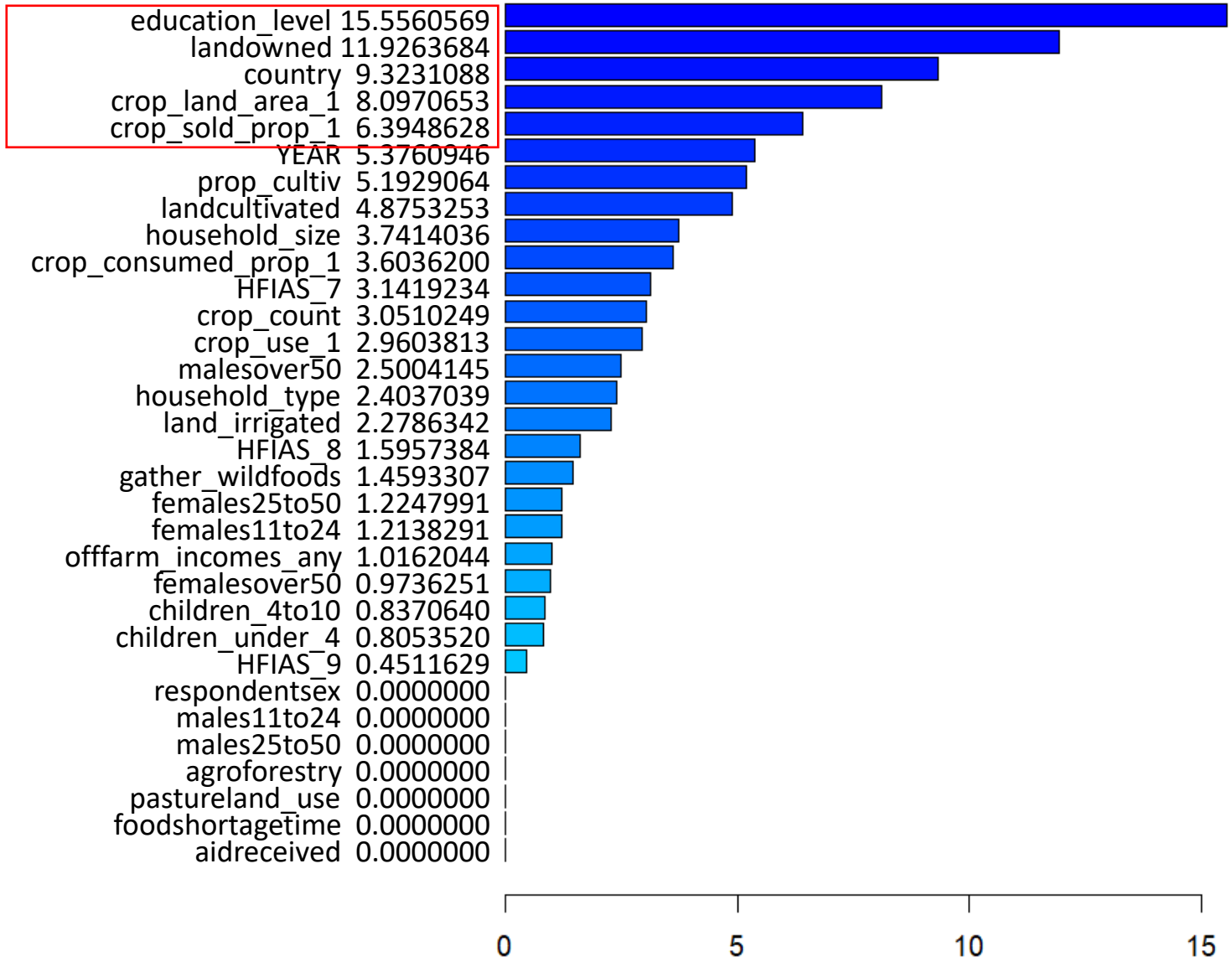
# Estimation du score de propension

gbm(traitement ~ covariables)

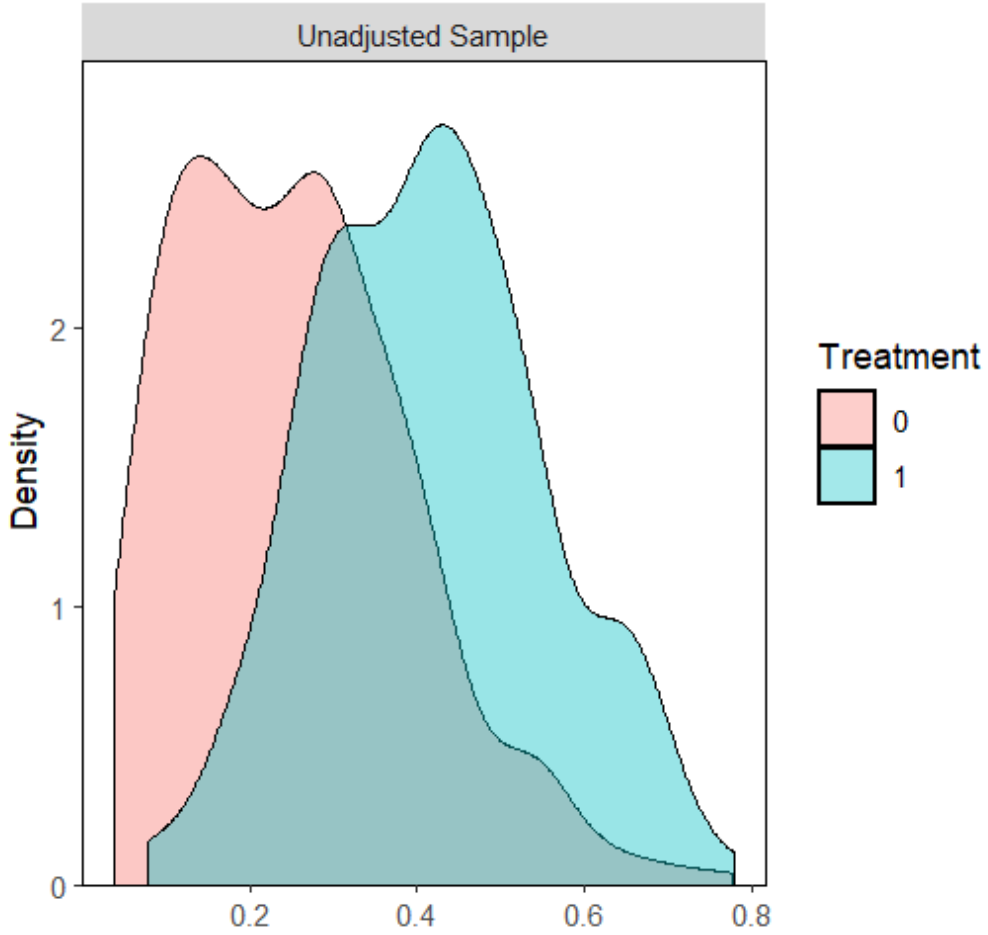
treat ~  YEAR + country + respondentsex + household_type + education_level +children_under_4+children_4to10 + males11to24 + females11to24 + males25to50 + females25to50 + malesover50 + femalesover50 + landcultivated + landowned + crop_count + crop_land_area_1 + crop_use_1 + crop_consumed_prop_1 + crop_sold_prop_1 + land_irrigated + agroforestry + pastureland_use + gather_wildfoods + foodshortagetime +HFIAS_9 +HFIAS_8 + HFIAS_7 + aidreceived + offfarm_incomes_any + household_size + prop_cultiv

# Contributions au modèle

>50%



# Scores estimés



# Appariement

```
Appariement = Match(Y = dfkg$crop_yield_1, Tr = dfkg$treat,
                    X = dfkg$pscoregbm,
                    estimand = "ATT",
                    M = 2, # Nombre de voisins
                    caliper = 0.20,
                    replace=TRUE,
                    ties=FALSE)
```

Estimate... -33.047

SE..... 88.568

T-stat..... -0.37312

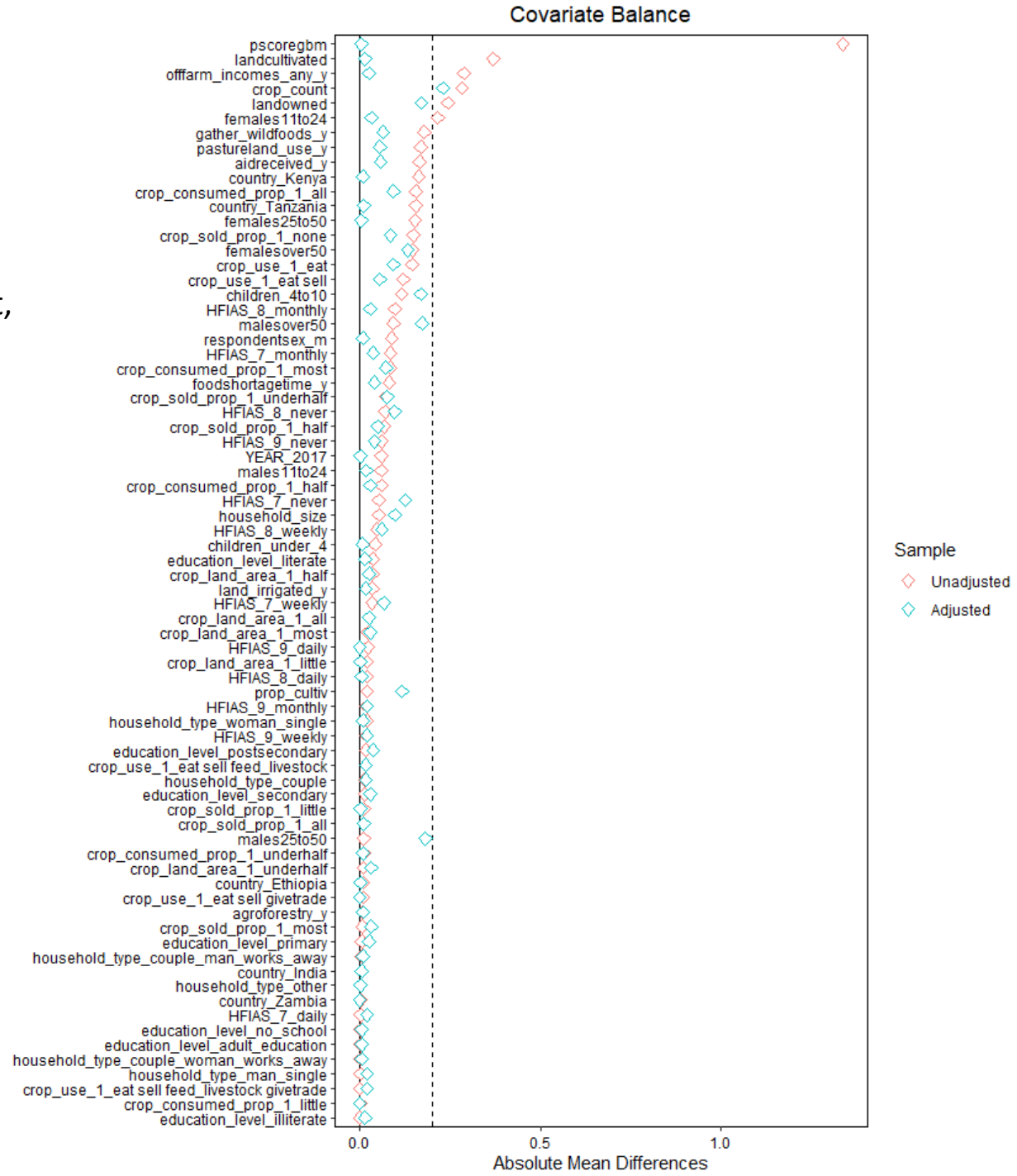
p.val..... 0.70906

Original number of observations..... 461

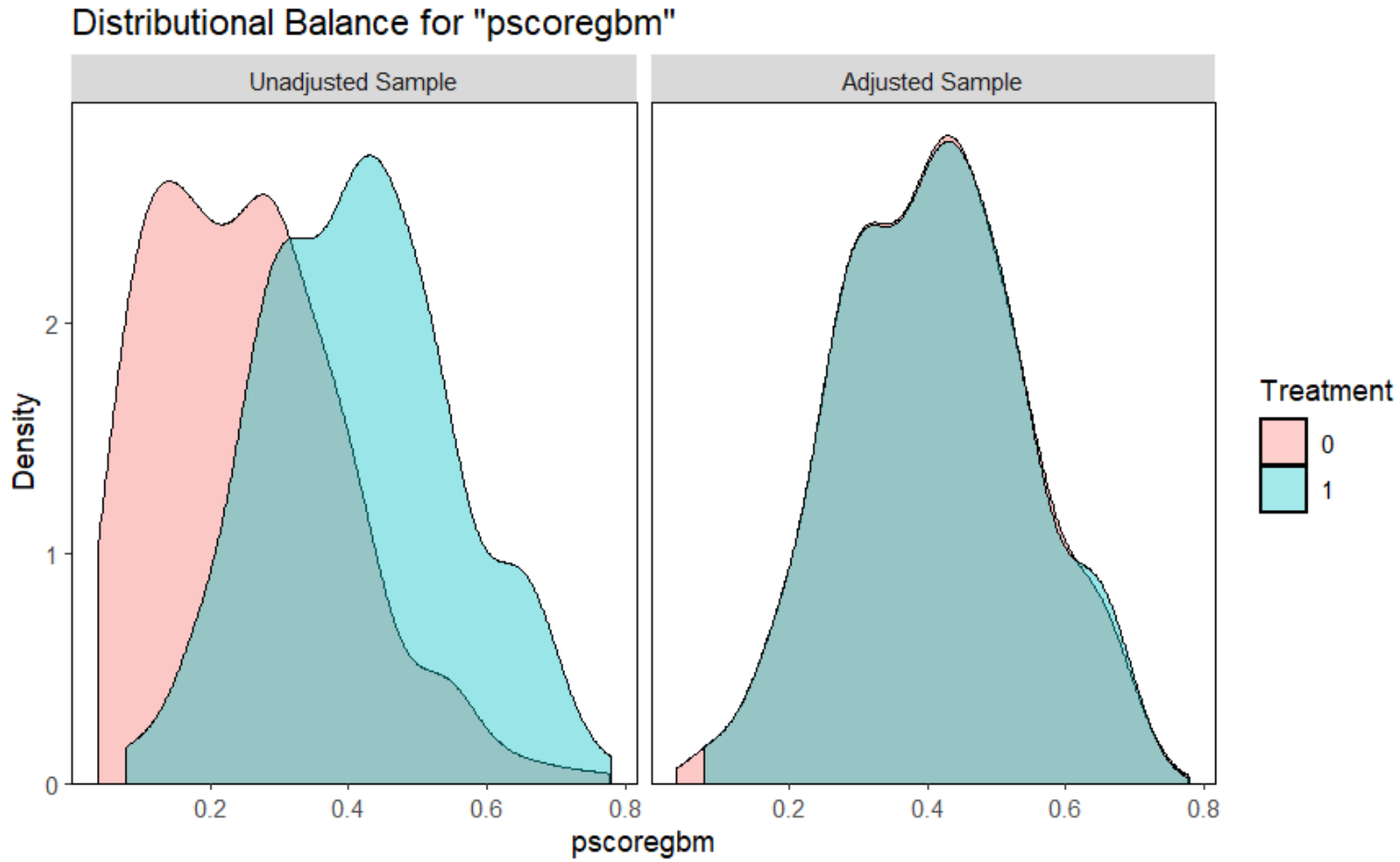
Original number of treated obs..... 175

Matched number of observations..... 171

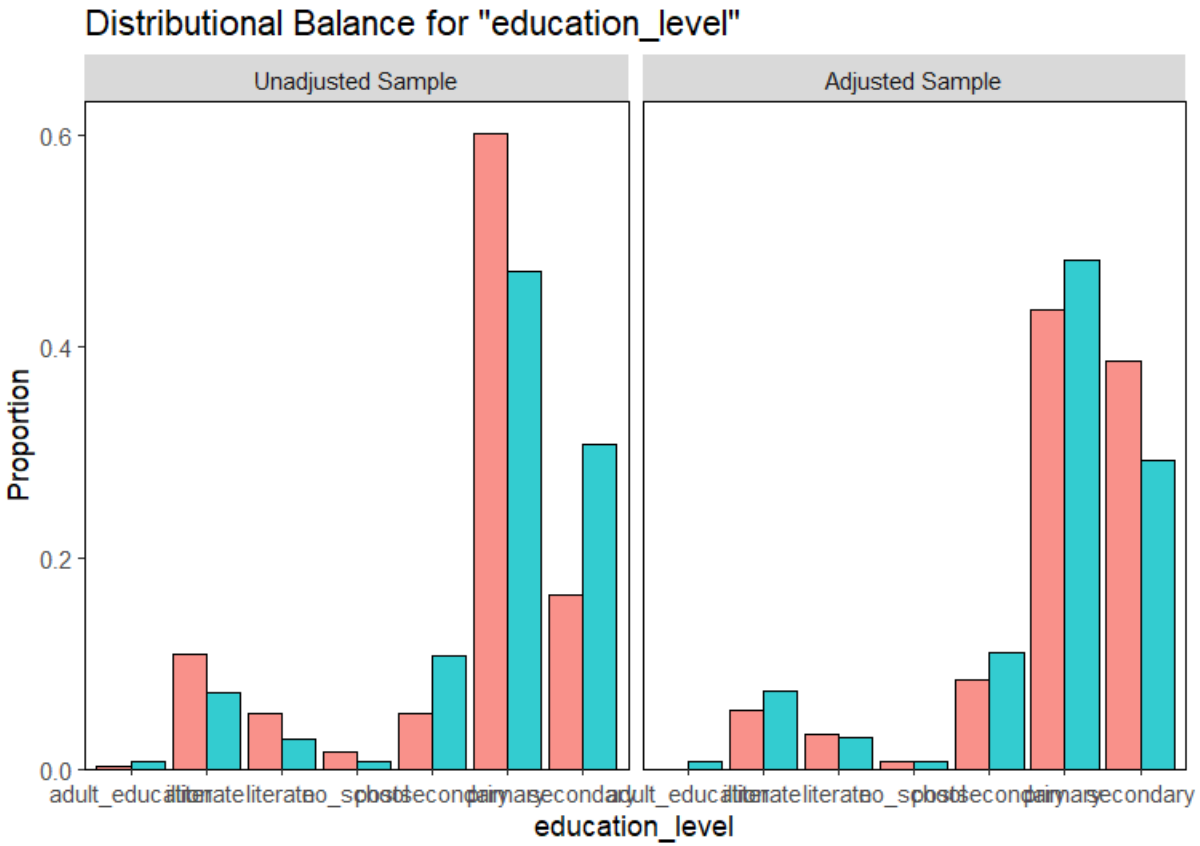
Matched number of observations (unweighted). 342

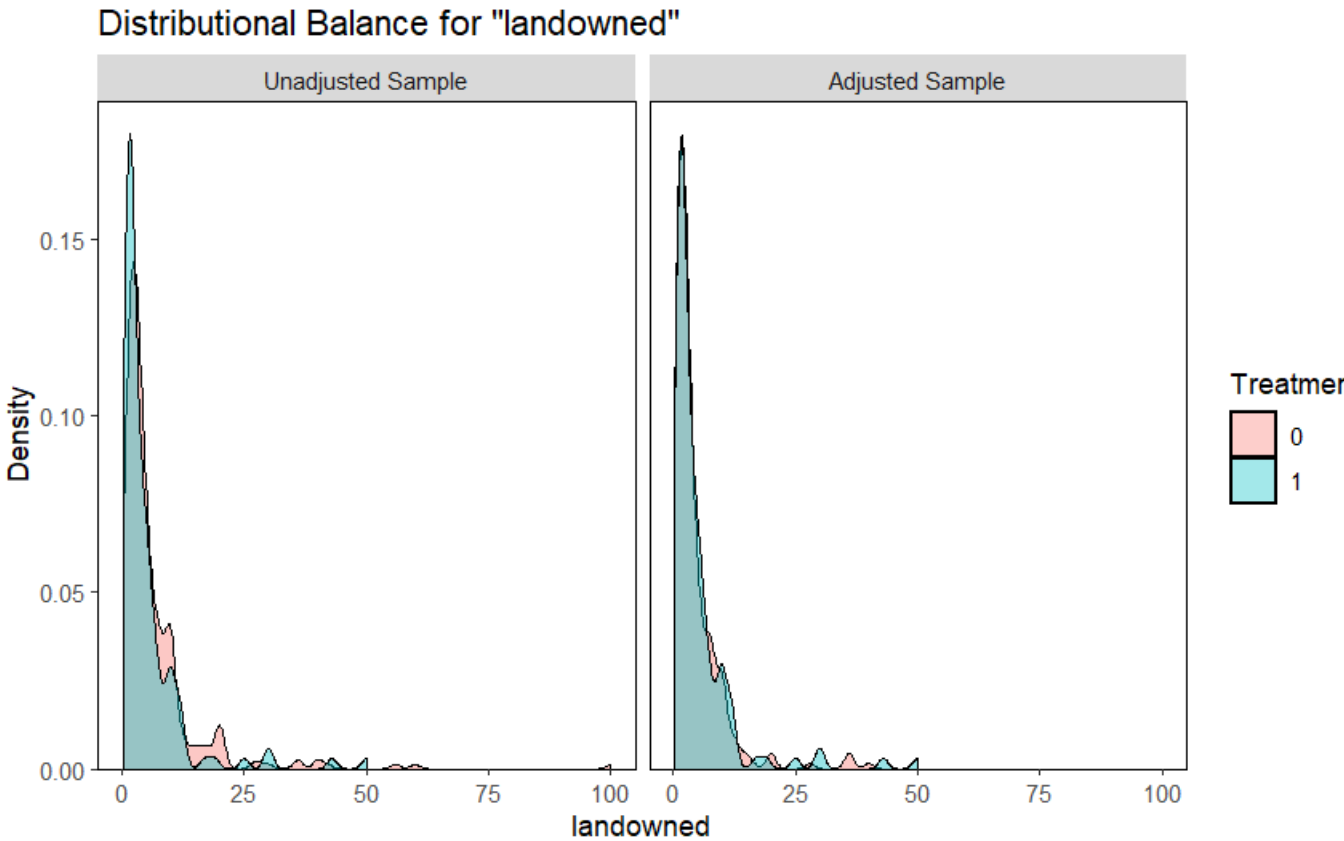


# Distributions des variables les plus représentatives du modèle av/ap ajustement

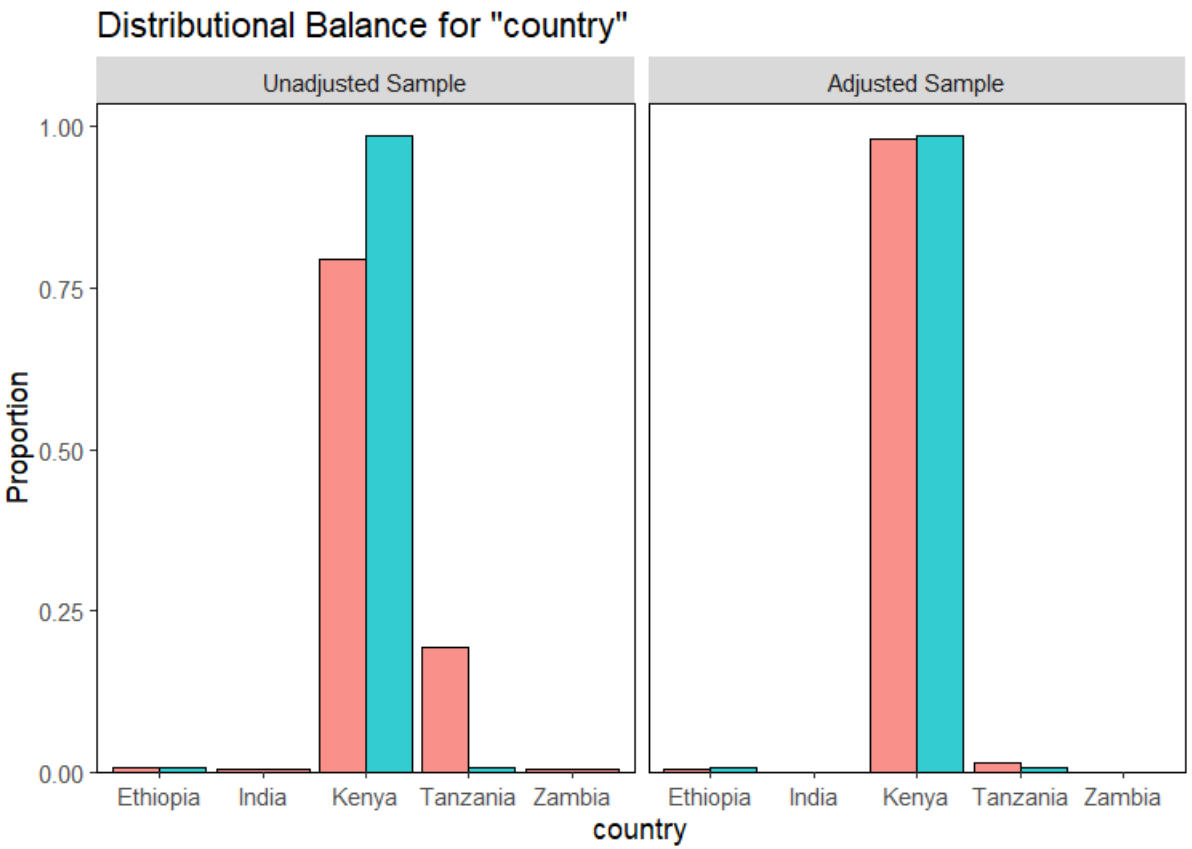


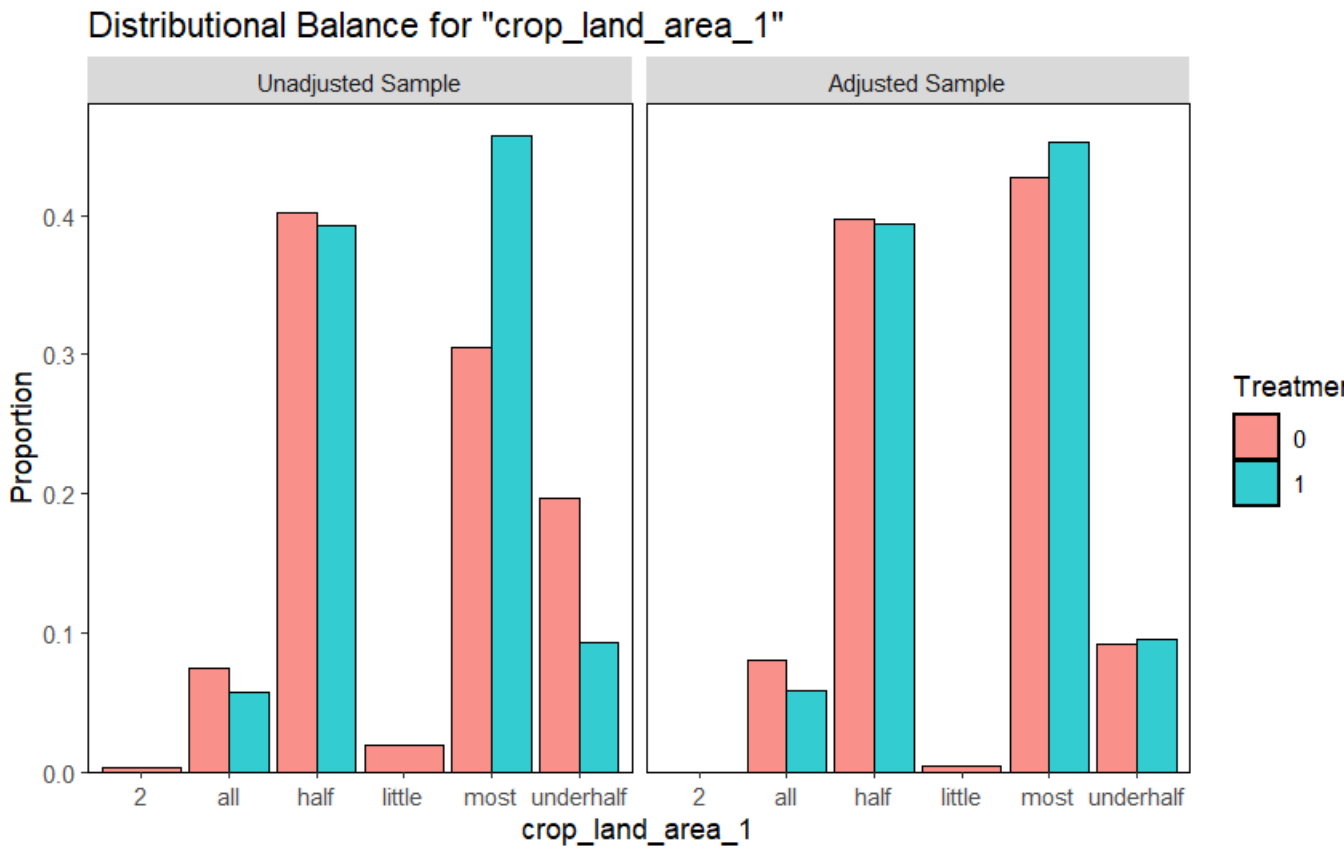
# Distributions des variables les plus représentatives du modèle av/ap ajustement



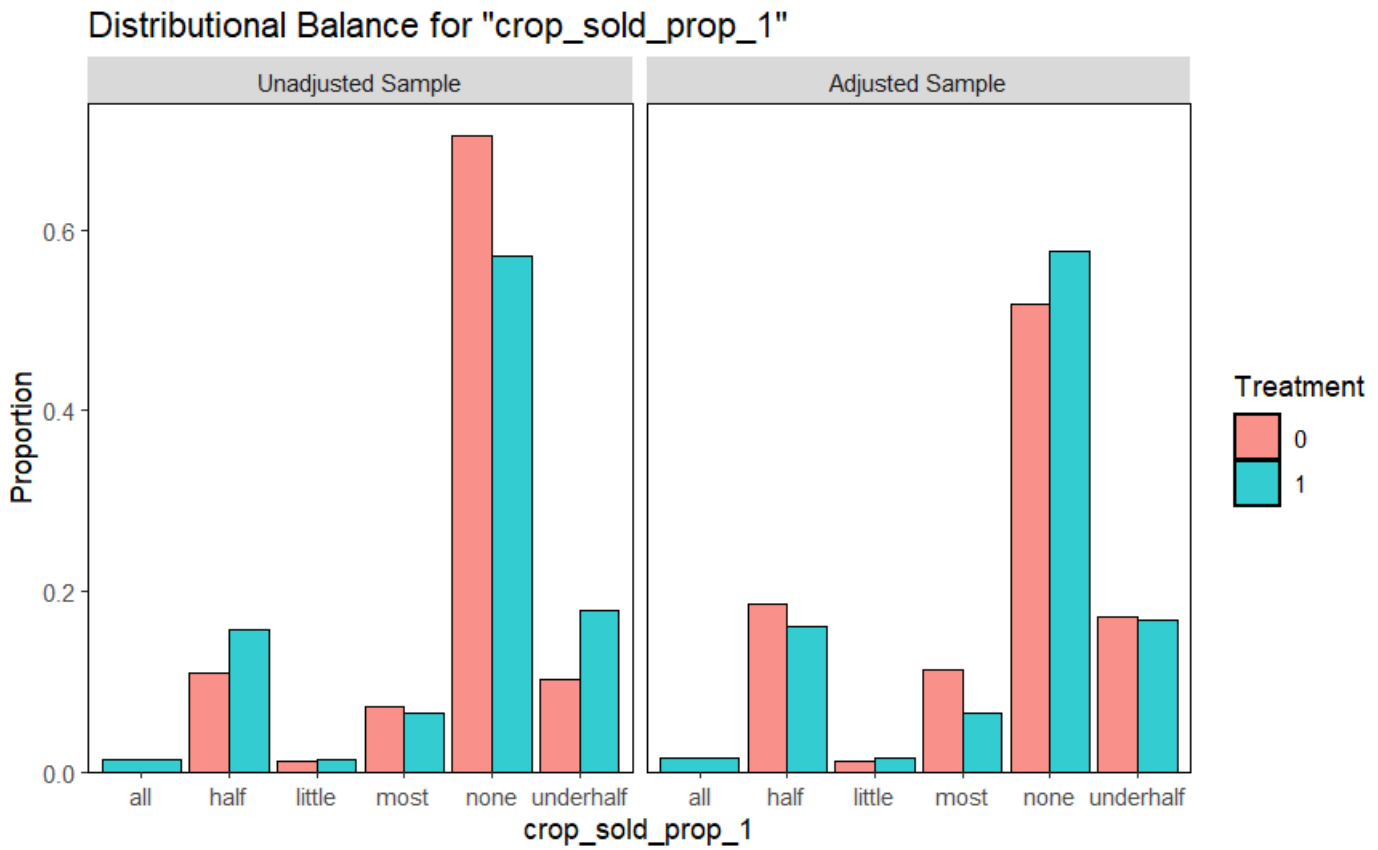



# Distributions des variables les plus représentatives du modèle av/ap ajustement

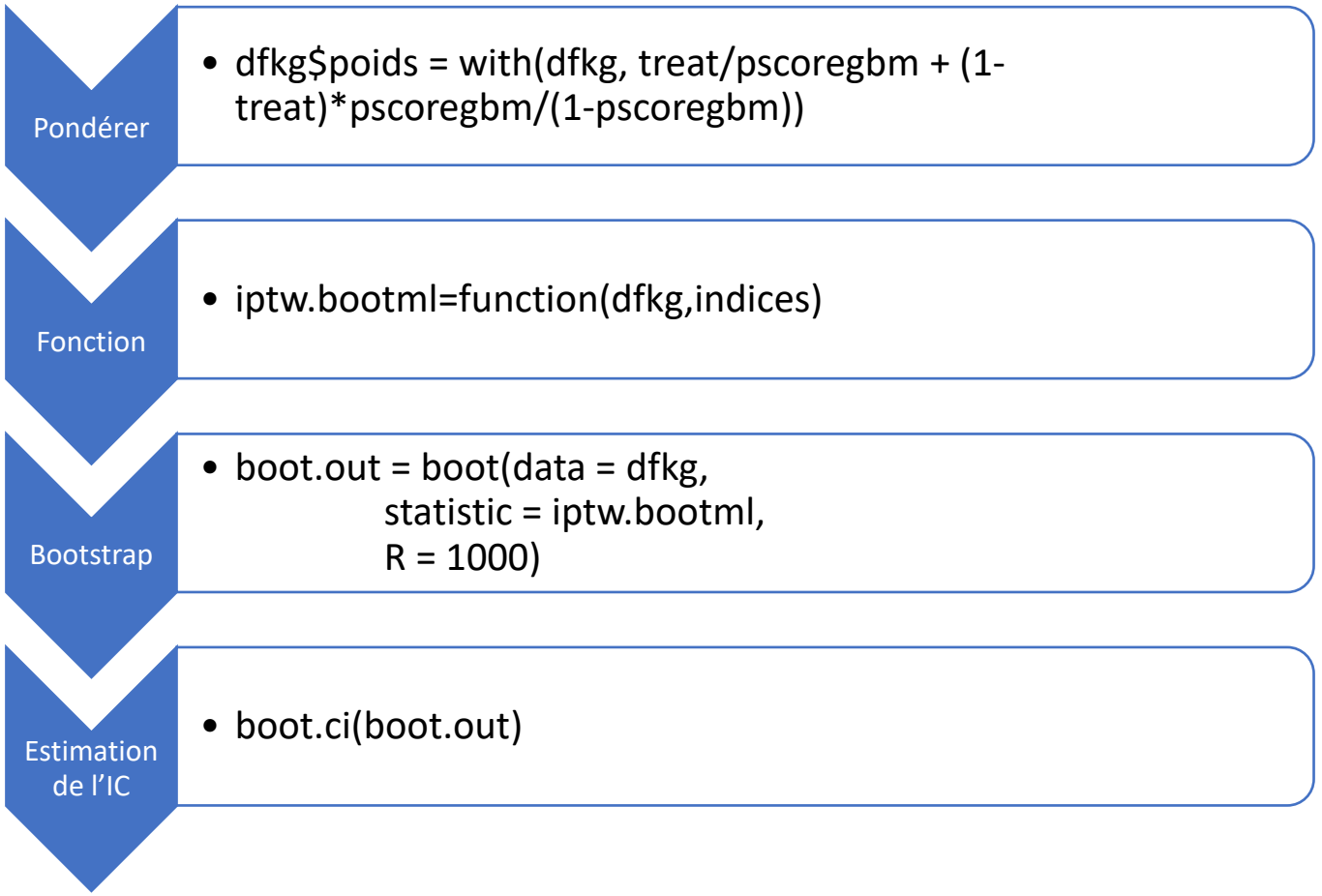


# Distributions des variables les plus représentatives du modèle av/ap ajustement



# Pondération inverse



# Estimation de l'IC avec Bootstrap

**Pondérer**

- dfkg$poids = with(dfkg, treat/pscoregbm + (1-treat)*pscoregbm/(1-pscoregbm))

**Fonction**

- iptw.bootml=function(dfkg,indices)

**Bootstrap**

- boot.out = boot(data = dfkg,
  statistic = iptw.bootml,
  R = 2000)

**Estimation de l'IC**

- boot.ci(boot.out)

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.out)

Intervals :
Level Normal Basic Studentized
95% (-188.94, 149.34 ) (-189.61, 148.69 ) (-178.83, 174.02 )

Level Percentile BCa
95% (-188.64, 149.66 ) (-188.52, 149.98 )
Calculations and Intervals on Original Scale

# Double robustesse

- Modèle du score de propension
  - propension = gbm(agroforestry ~ Covariables)
- Modèle du rendement
  - model_rendement <- gbm(crop_yield_1 ~ Variables)
- Dupliquas du jeu de données
  - TreatOui = dfkg ; TreatNon = dfkg
  - TreatOui$agroforestry=1 ; TreatNon$agroforestry=0
- Prédiction des valeurs du modèle et erreurs
  - TreatOui$yieldpredict = predict(model_rendement, newdata = TreatOui)
  - TreatOui$residuals = TreatOui$crop_yield_1 - TreatOui$yieldpredict
  - TreatNon$yieldpredict = predict(model_rendement, newdata = TreatNon)
  - TreatNon$residuals = TreatNon$crop_yield_1 - TreatNon$yieldpredict

# Double robustesse

- Estimateur E Y a=1
  - TreatOui$estimation = with(TreatOui,(yieldpredict + (treat / pscoregbm) * residuals))
- Estimateur E Y a=0
  - TreatNon$estimation = with(TreatNon,(yieldpredict + ((1 - treat) / (1 - pscoregbm)) * residuals))

# Double robustesse

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.out)

Intervals :
Level Normal Basic
95% (-143.5, -105.4 ) (-143.2, -104.4 )

Level Percentile BCa
95% (-122.1, -83.2 ) (-135.4, -105.1 )
Calculations and Intervals on Original Scale
Warning : BCa Intervals used Extreme Quantiles
Some BCa intervals may be unstable

# Comparaison des résultats

- T-test :
  - (-194.92 ; 154.97)
- Appariement :
  - (−203.74 ; 137.65)
- Pondération :
  - (-188.64,  149.66 )
- Double robustesse :
  - (-122.1,  -83.2 )

# Validation croisée k-fold

```r
# Définition des hyperparamètres à tester
grid <- expand.grid(n.trees = c(100,1000, 2000),
                    shrinkage = c(0.01, 0.05),
                    interaction.depth = c(3,5,7,9,11),
                    n.minobsinnode = c(5,10,20,40))
dfkg$treat = as.factor(dfkg$treat)

# Définition du contrôle de la validation croisée
ctrl <- trainControl(method = "cv",  # Méthode de validation croisée
                     number = 10,      # Nombre de folds de la validation croisée
                     verboseIter = TRUE)

# Validation croisée avec caret
metrics = c("Accuracy", "ROC")
cv_results <- train(treat ~  YEAR + country + respondentsex + household_type + edu
                      females11to24 + males25to50 + females25to50 + malesover50 +
                      crop_land_area_1+crop_use_1+
                      crop_consumed_prop_1+crop_sold_prop_1+land_irrigated+agrofor
                      pastureland_use+gather_wildfoods+
                      foodshortagetime+HFIAS_9+HFIAS_8+HFIAS_7+
                      aidreceived+offfarm_incomes_any+household_size+prop_cultiv,
                    data = dfkg,
                    method = "gbm",
                    trControl = ctrl,
                    tuneGrid = grid)

# Affichage des résultats
print(cv_results$results)
cv_results$results[which.max(cv_results$results$Accuracy),]
```

# Rappel

- Estimer le score de propension
  - Régression logistique
  - Machine Learning
- Appariement
  - Vérifier les caractéristiques du nouvel échantillon
- IPTW
  - Pondérer
  - Modèle pondéré du rendement
  - Obtention de l'IC par Bootstrap
- Double robustesse
  - Définir le modèle de rendement
  - Estimer selon les formules
  - Obtenir l'IC par Bootstrap