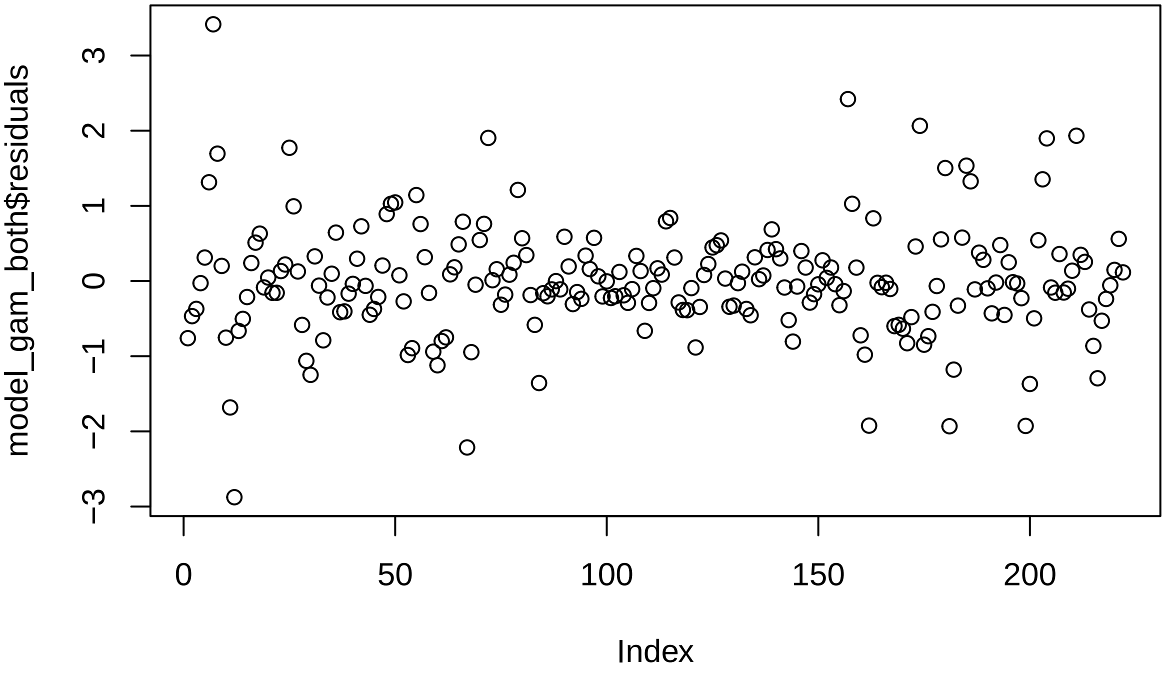
GAM on affordability

2023-07-05

I start by running the model for males and females

```
model_gam_both <- gam(
  Prevalence_both ~
    Year +
    Country+
    s(HDI, bs = 'cr') +
    s(GDP, bs = 'cr') +
    s(Education, bs = 'cr') +
   Affordability,
  data = data_gam
)
#summary(model_gam_both)
#plot(model_gam_both)
plot(model_gam_both$residuals)
```

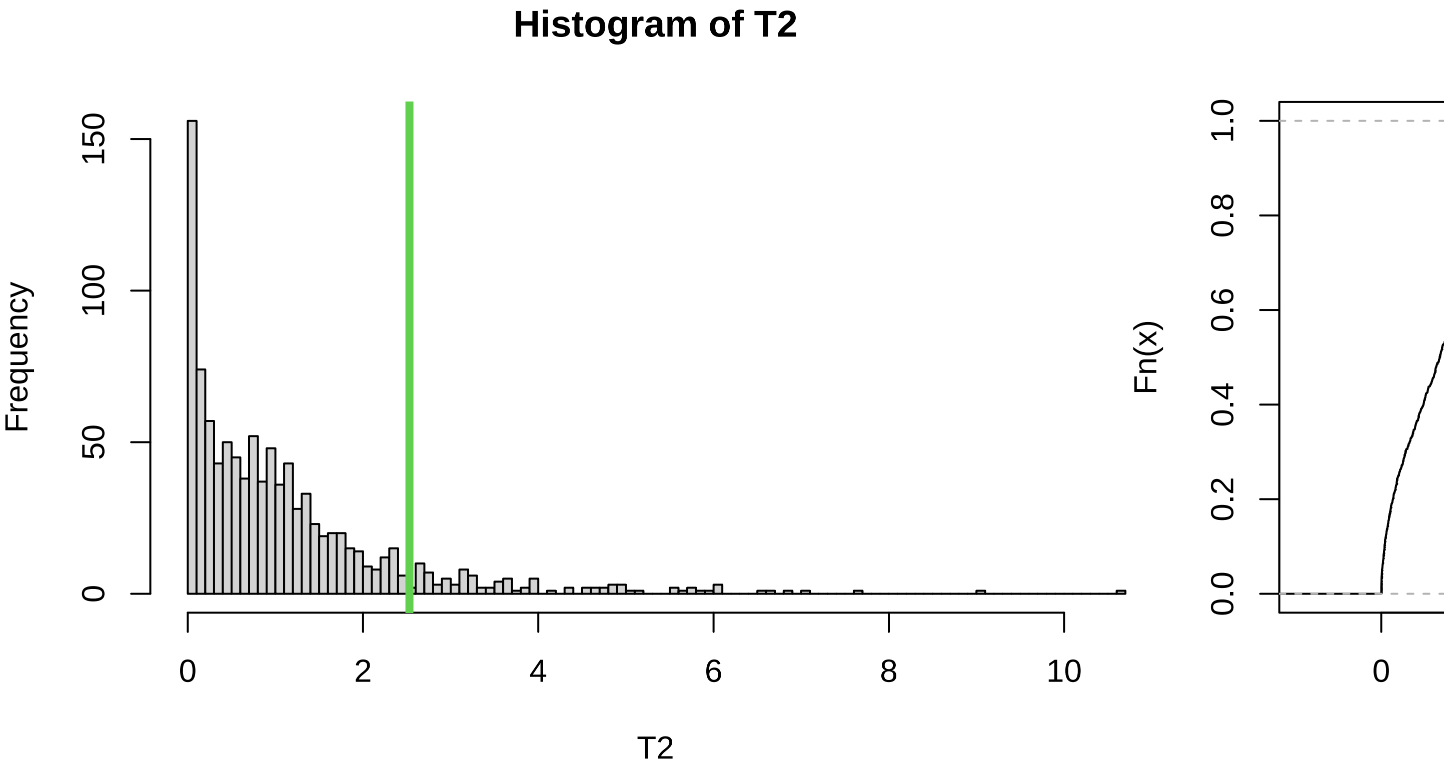


The countries clearly explain the majority of the variance in the data Let's see if first we can remove the GDP from the model H0:GDP term is 0, vs H1 is different than zero

```
summ<-summary(model_gam_both)
T0<-abs(summ$s.table[2,3])
T0
```

```
## [1] 2.529643
```

```r
model_gam_noGDP<-gam(
  Prevalence_both ~
    Year +
    Country+
    s(HDI, bs = 'cr') +
    s(Education, bs = 'cr') +
    Affordability,
  data = data_gam
)
#summary(model_gam_noGDP)

res<-model_gam_noGDP$residuals
fitted.values<-model_gam_noGDP$fitted.values

set.seed(seed)
T2<-numeric(B)
n<-nrow(data_gam)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res[permutation]
  response.perm <- fitted.values + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      s(data_gam$GDP, bs = 'cr') +
      s(data_gam$Education, bs = 'cr') +
      data_gam$Affordability
  )
  T2[perm] <- abs(summary(model.perm)$s.table[2,3])
}
diagnostic_permutation(T0,T2)
```



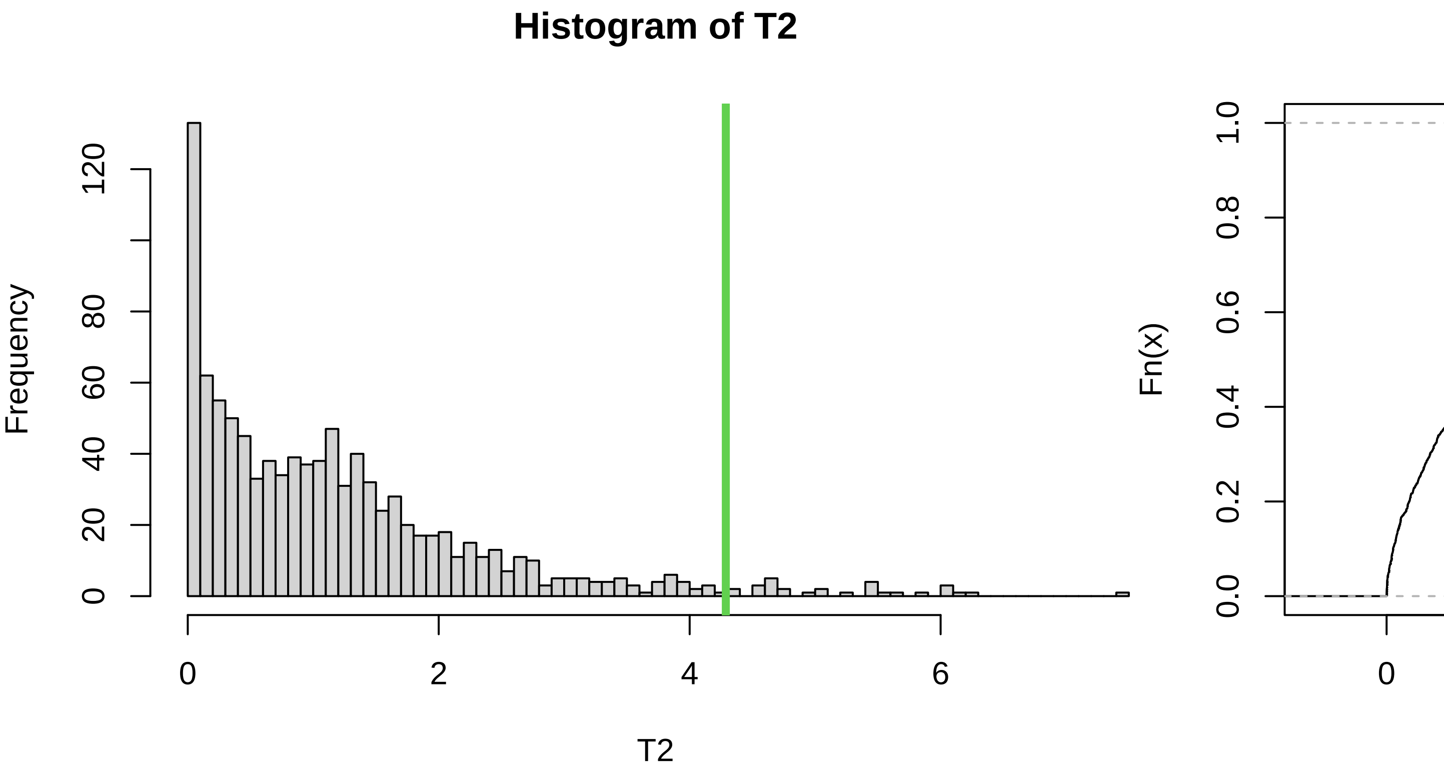
```
## p-value:  0.099
```

We cannot reject the null hypothesis that the GDP is 0

```
model_gam_noGDPnoEDU<-gam(
  Prevalence_both ~
    Year +
    Country+
    s(HDI, bs = 'cr')+
    Affordability,
  data=data_gam
)

res<-model_gam_noGDPnoEDU$residuals
fitted.values<-model_gam_noGDPnoEDU$fitted.values

T0<-summary(model_gam_noGDP)$s.table[2,3]
T0
```

```
## [1] 4.287971
```

```
set.seed(seed)
T2<-numeric(B)
n<-nrow(data_gam)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res[permutation]
  response.perm <- fitted.values + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      s(data_gam$Education, bs = 'cr') +
      data_gam$Affordability
  )
  T2[perm] <- abs(summary(model.perm)$s.table[2,3])
}
diagnostic_permutation(T0,T2)
```



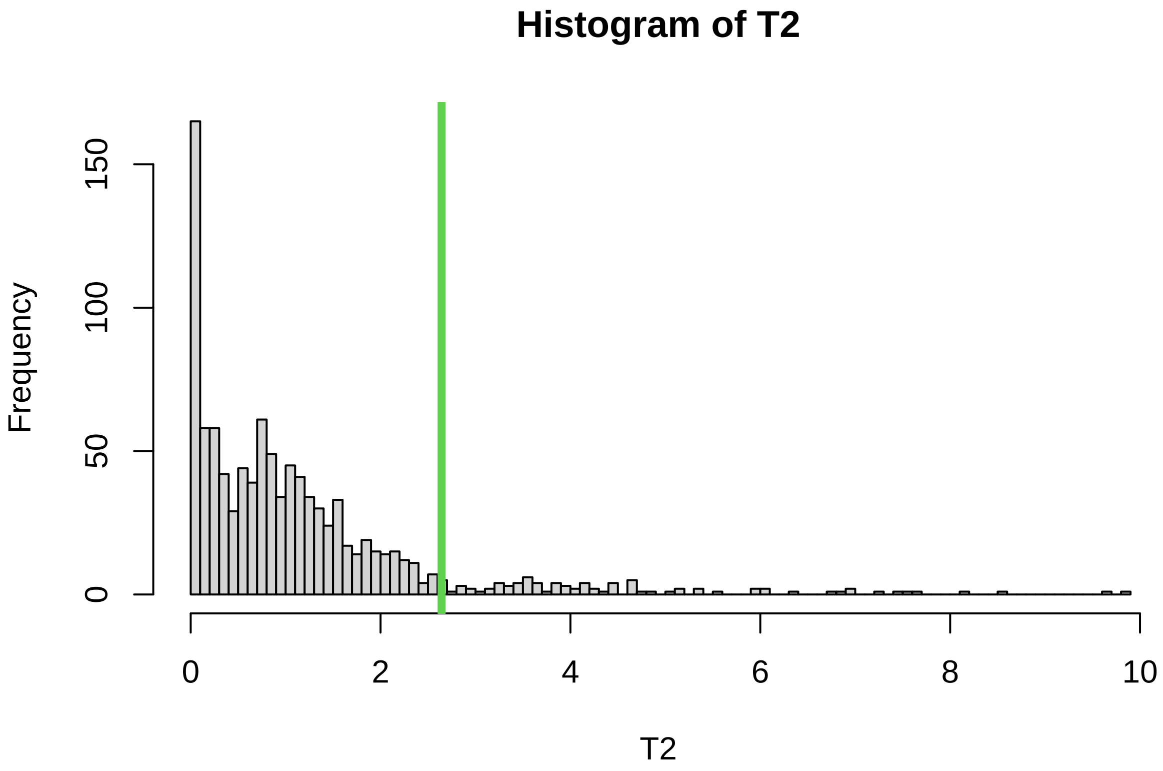
```
## p-value:  0.029
```

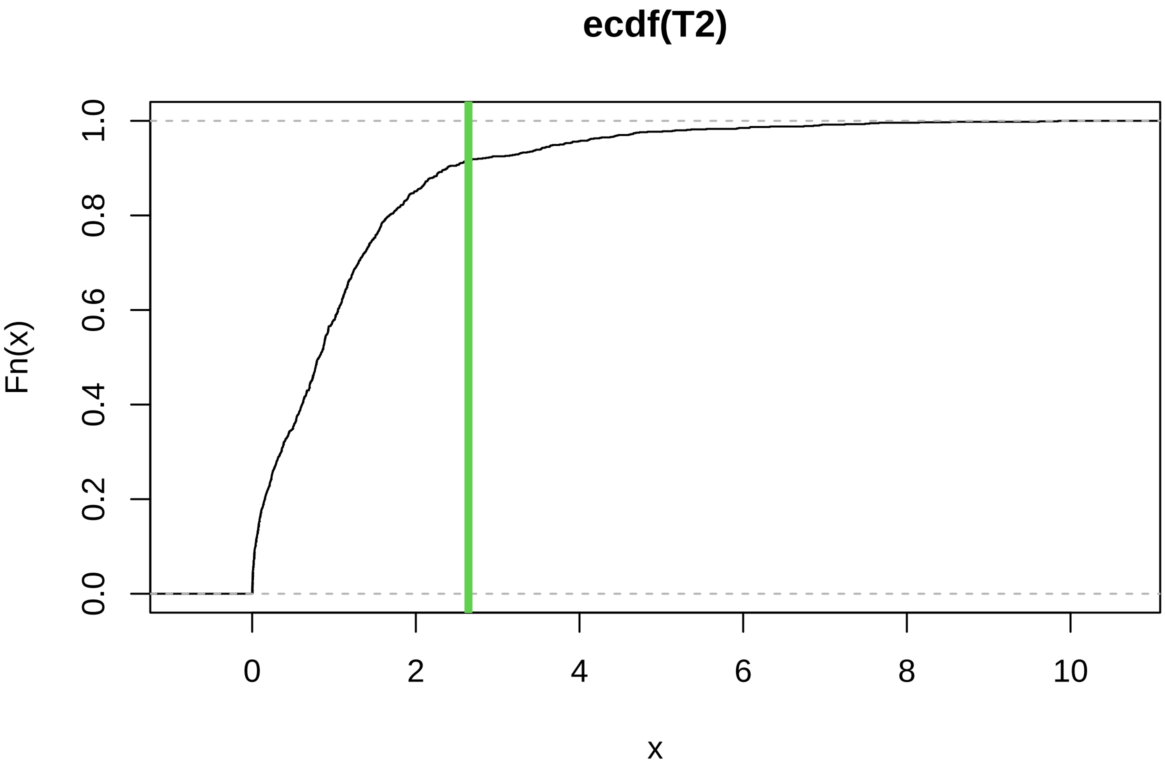
We reject the null hypothesis and keep education

```
model_gam_noGDPnoHDI<-gam(
  Prevalence_both ~
    Year +
    Country+
    s(Education, bs = 'cr')+
    Affordability,
  data=data_gam
)

summ<-summary(model_gam_noGDP)
T0<-summ$s.table[1,3]
T0
```

```
## [1] 2.642853
```

```
res<-model_gam_noGDPnoHDI$residuals
fitted.values<-model_gam_noGDPnoHDI$fitted.values

set.seed(seed)
T2<-numeric(B)
n<-nrow(data_gam)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res[permutation]
  response.perm <- fitted.values + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      s(data_gam$Education, bs = 'cr') +
      data_gam$Affordability
  )
  T2[perm] <- abs(summary(model.perm)$s.table[1,3])
}
diagnostic_permutation(T0,T2)
```

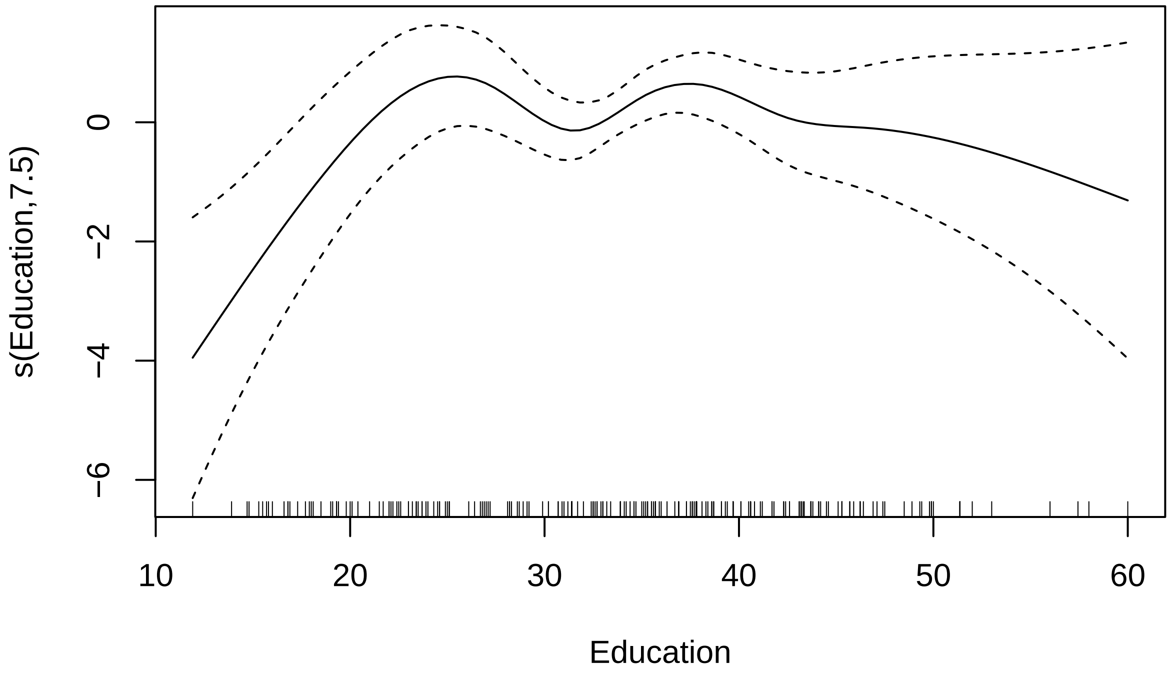




```
## p-value:  0.084
```

#We cannot reject H0 and we remove HDI, and I move to the affordability

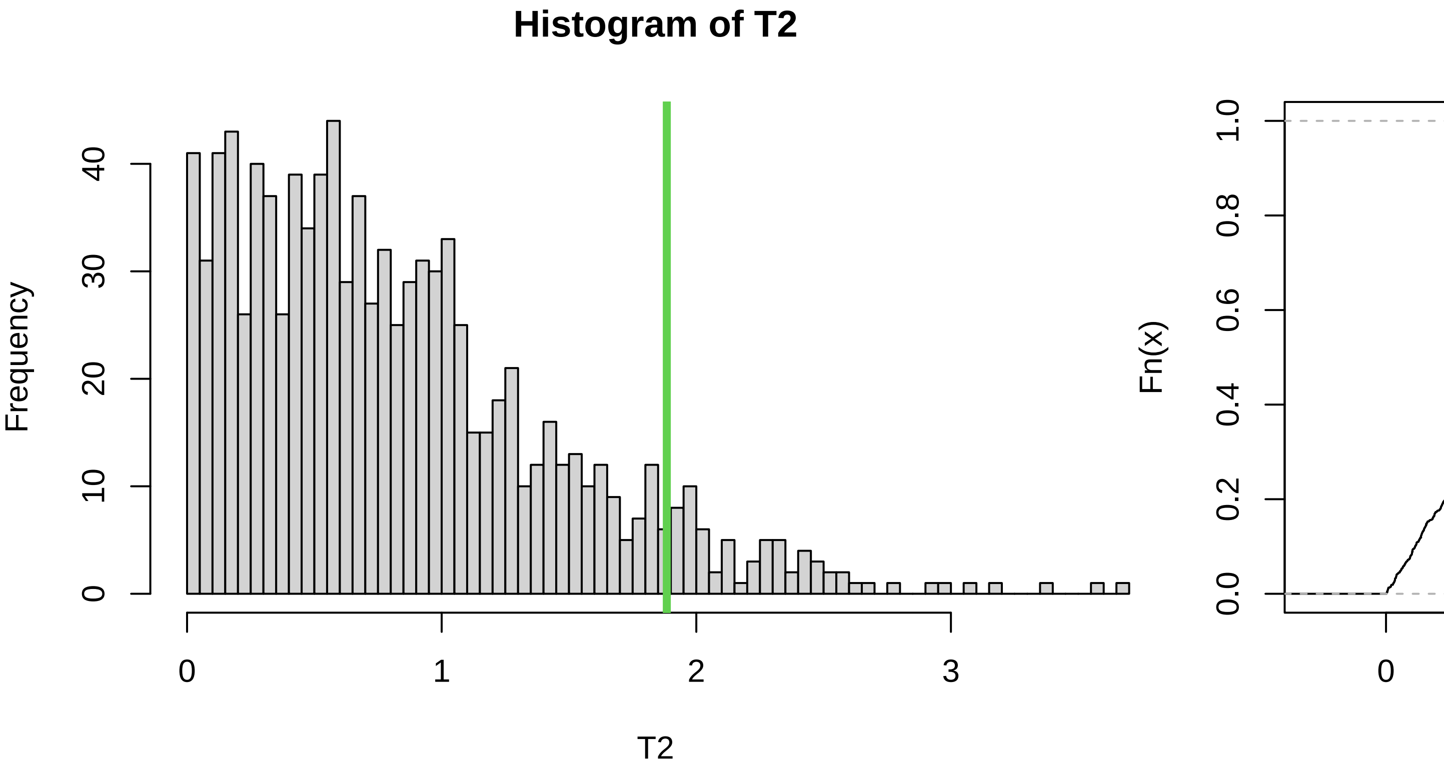
```
model.gam.noaffordability<-gam(
  Prevalence_both ~
    Year +
    Country+
    s(Education, bs = 'cr'),
  data=data_gam
)
#summary(model.gam.noaffordability)
plot(model.gam.noaffordability)
```



```r
T0<-abs(summary(model_gam_noGDP)$p.table[39,3])
T0
```

```
## [1] 1.884178
```

```r
res<-model.gam.noaffordability$residuals
fitted.values<-model.gam.noaffordability$fitted.values

set.seed(seed)
T2<-numeric(B)
n<-nrow(data_gam)

for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res[permutation]
  response.perm <- fitted.values + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      s(data_gam$Education, bs = 'cr') +
      data_gam$Affordability
  )
  T2[perm] <- abs(summary(model.perm)$p.table[39,3])
}
diagnostic_permutation(T0,T2)
```



```
## p-value:  0.068
```

It is not significant at alpha=0.1, I can remove it At the net of the analysis it seems that the affordability significantly and negatively on prevalence

```
model.final.both<-gam(
  data_gam$Prevalence_both ~
    data_gam$Year +
    data_gam$Country+
    s(data_gam$Education, bs = 'cr')
)
#summary(model.final.both)
```

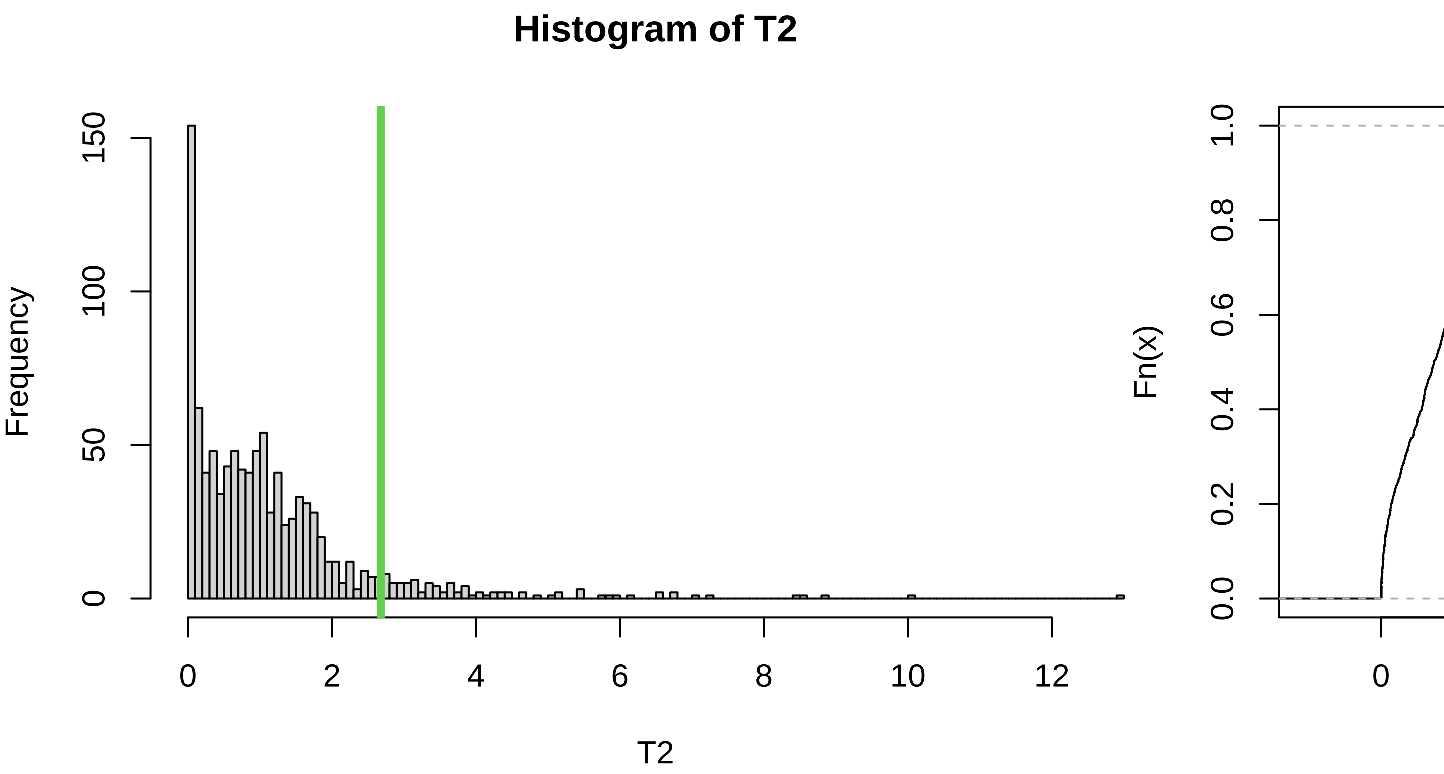
Now I try on males

```
model_gam_males <- gam(
  Prevalence_males ~
    Year +
    Country+
    s(HDI, bs = 'cr') +
    s(GDP, bs = 'cr') +
    s(Education_males, bs = 'cr') +
    Affordability,
  data = data_gam
)
#summary(model_gam_males)
```

```r
#plot(model_gam_males)

summ_males<-summary(model_gam_males)
T0.males<-abs(summ_males$s.table[2,3])
T0.males
```

```
## [1] 2.678014
```

```r
model_gam_noGDP.males<-gam(
  Prevalence_males ~
    Year +
    Country+
    s(HDI, bs = 'cr') +
    s(Education_males, bs = 'cr') +
    Affordability,
  data = data_gam
)
#summary(model_gam_noGDP.males)

res.males<-model_gam_noGDP.males$residuals
fitted.values.males<-model_gam_noGDP.males$fitted.values


T2.males<-numeric(B)
n<-nrow(data_gam)
set.seed(seed)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res.males[permutation]
  response.perm <- fitted.values.males + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      s(data_gam$GDP, bs = 'cr') +
      s(data_gam$Education_males, bs = 'cr') +
      data_gam$Affordability
  )
  T2.males[perm] <- abs(summary(model.perm)$s.table[2,3])
}
diagnostic_permutation(T0.males,T2.males)
```



```
## p-value:  0.089
```

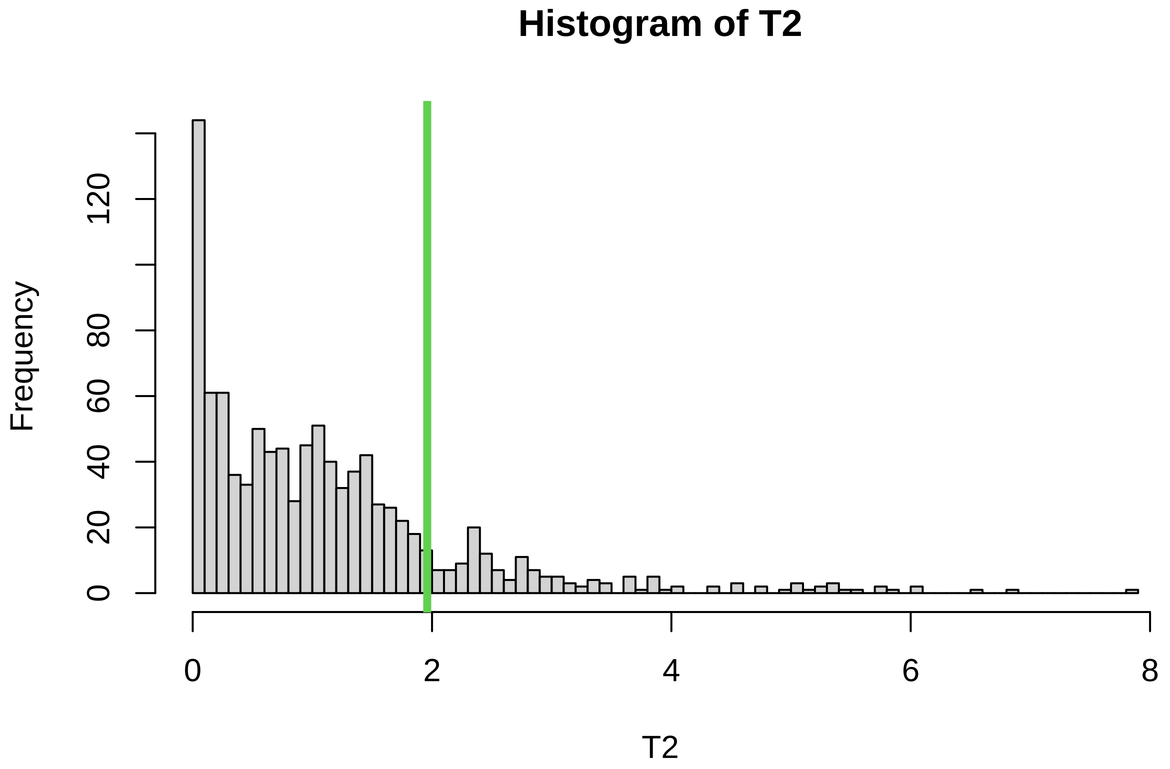
We cannot reject H0, hence we can simplify the model and remove GDP

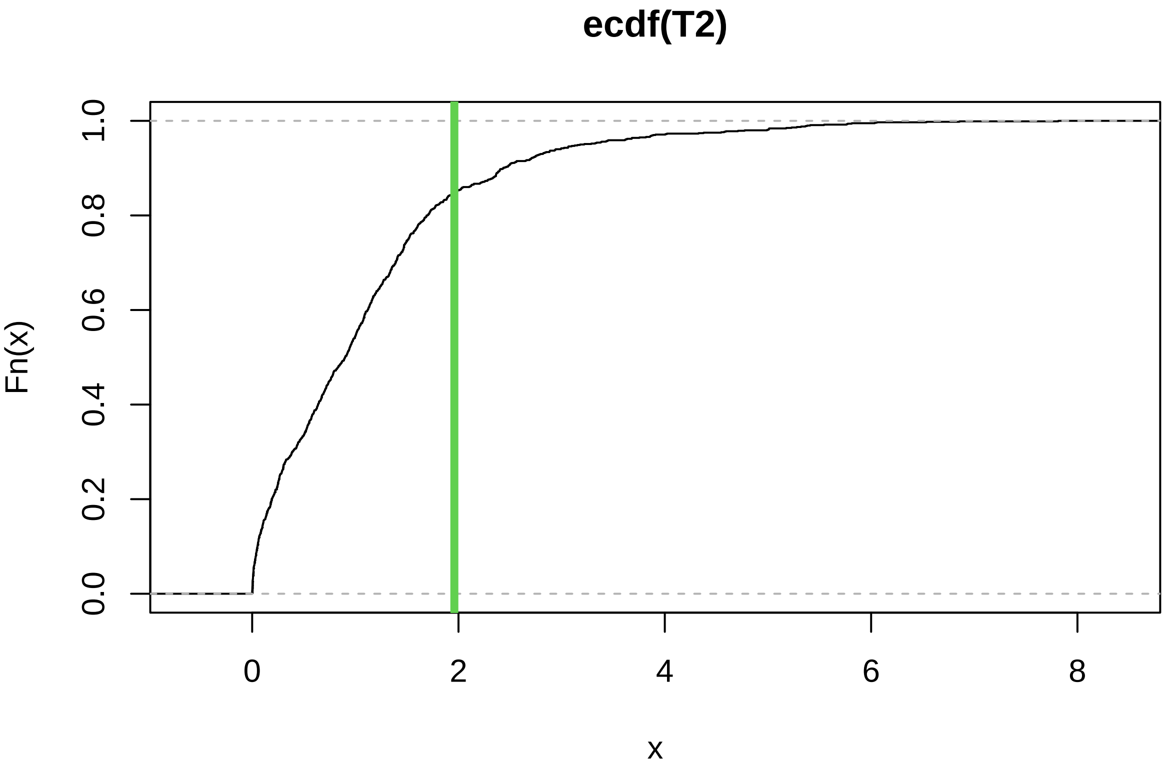
Now we try to remove the HDI variable

```
model.gam.noHDI<-gam(
  Prevalence_males ~
    Year +
    Country+
    s(Education_males, bs = 'cr')+
    Affordability,
  data = data_gam
)
#summary(model.gam.noHDI)
T0<-summary(model_gam_noGDP.males)$s.table[1,3]
T0
```

```
## [1] 1.959907
```

```
res<-model.gam.noHDI$residuals
fitted.values<-model.gam.noHDI$fitted.values

set.seed(seed)
T2<-numeric(B)
n<-nrow(data_gam)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res[permutation]
  response.perm <- fitted.values + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      s(data_gam$Education_males, bs = 'cr') +
      data_gam$Affordability
  )
  T2[perm] <- abs(summary(model.perm)$s.table[1,3])
}
diagnostic_permutation(T0,T2)
```





```
## p-value:  0.151
```

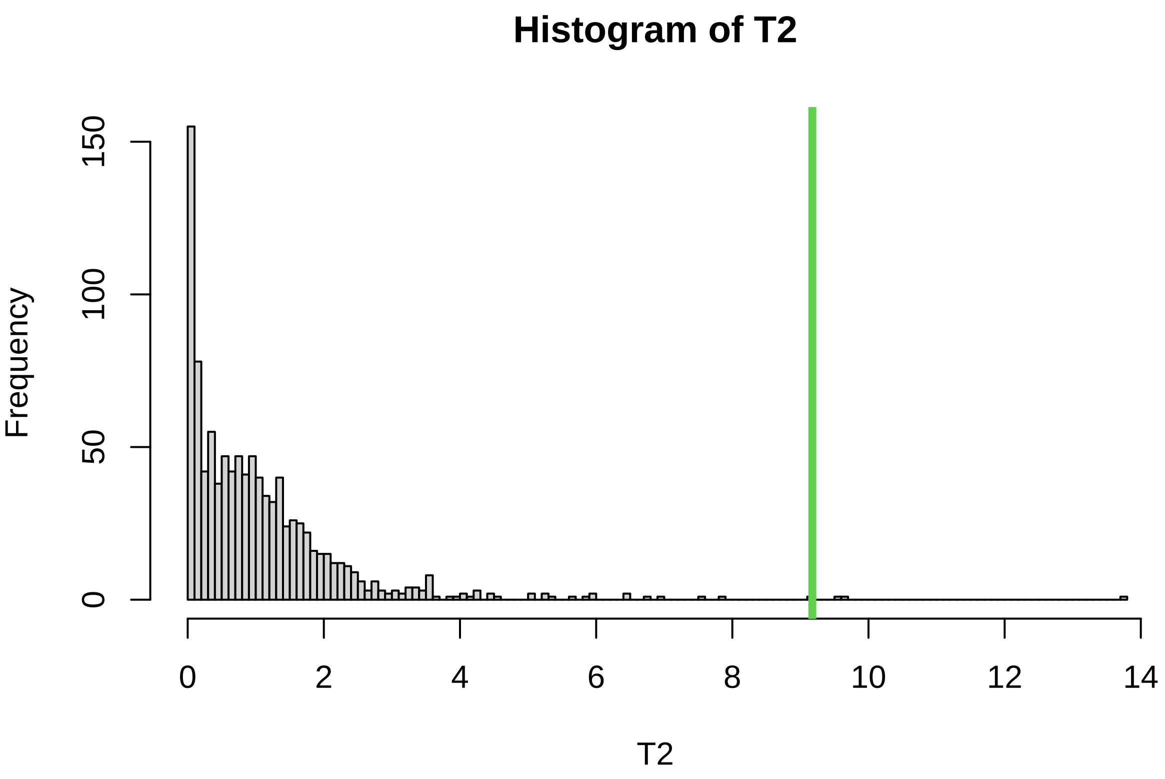
We cannot reject H0, we can remove HDI.

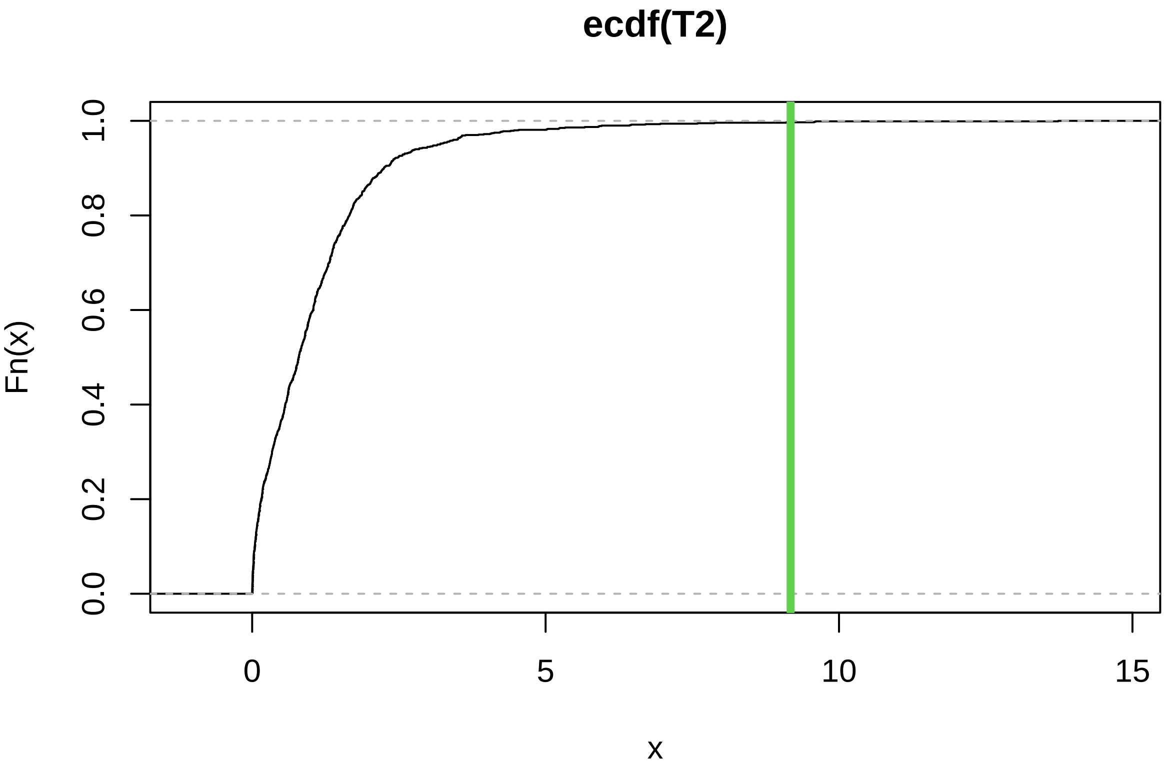
I now try to remove education

```
model.gam.noedu.males<-gam(
  Prevalence_males ~
    Year +
    Country+
    Affordability,
  data = data_gam
)
#summary(model.gam.noedu.males)

T0<-abs(summary(model.gam.noHDI)$s.table[1,3])
T0
```

```
## [1] 9.175003
```

```
res<-model.gam.noedu.males$residuals
fitted.values<-model.gam.noedu.males$fitted.values

set.seed(seed)
T2<-numeric(B)
n<-nrow(data_gam)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res[permutation]
  response.perm <- fitted.values + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$Education_males, bs = 'cr') +
      data_gam$Affordability
  )
  T2[perm] <- abs(summary(model.perm)$s.table[1,3])
}
diagnostic_permutation(T0,T2)
```





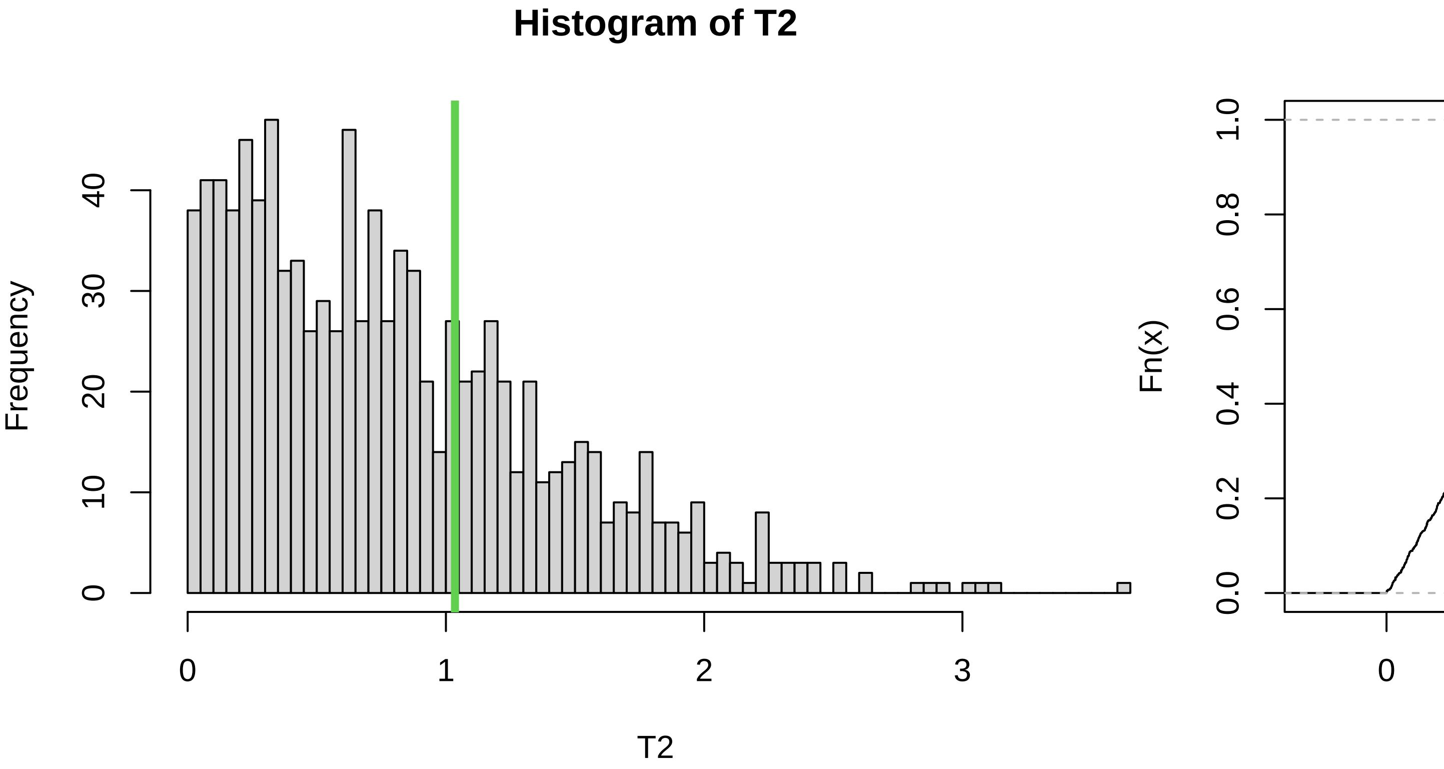
```
## p-value:  0.003
```

We cannot reduce the model and we should keep the education. Now we try to remove affordability

```
model.gam.noaffordability<-gam(
  Prevalence_males ~
    Year +
    Country+
    s(Education_males, bs = 'cr'),
  data = data_gam
)
#summary(model.gam.noaffordability)
T0<-abs(summary(model.gam.noHDI)$p.table[39,3])
T0
```

```
## [1] 1.035014
```

```
res.males<-model.gam.noaffordability$residuals
fitted.values.males<-model.gam.noaffordability$fitted.values


T2<-numeric(B)
n<-nrow(data_gam)

for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res.males[permutation]
  response.perm <- fitted.values.males + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$Education_males, bs = 'cr') +
      data_gam$Affordability
  )
  T2[perm] <- abs(summary(model.perm)$p.table[39,3])
}
diagnostic_permutation(T0,T2)
```



```
## p-value:  0.306
```

So we can remove affordability So we have that we cannot say affordability impacts on males

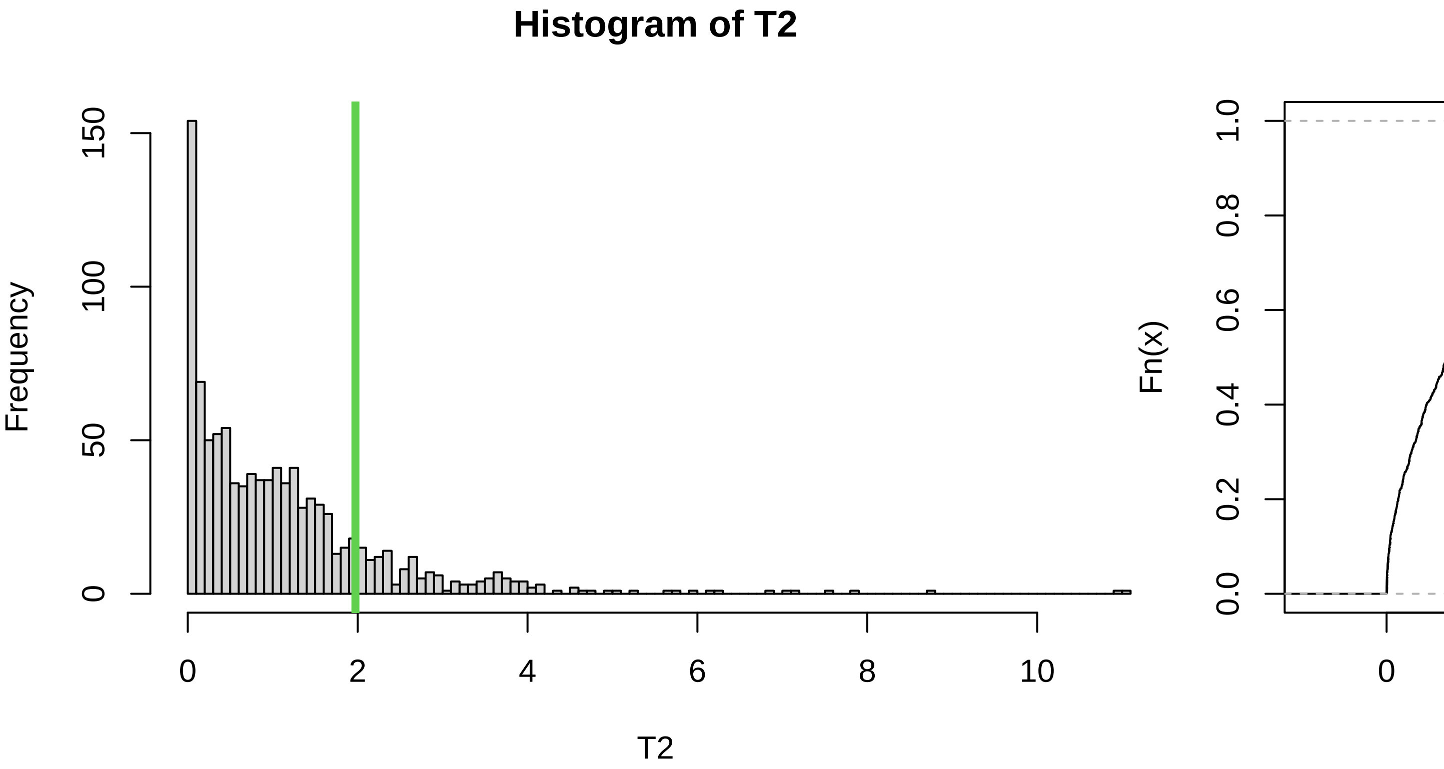
We now try to use the same approach on females

We start by checking if we can remove GDP

```
model_gam_females1 <- gam(
  Prevalence_females ~
    Year +
    Country+
    s(HDI, bs = 'cr') +
    s(GDP, bs = 'cr') +
    s(Education_females, bs = 'cr') +
    Affordability,
    data = data_gam
)
#summary(model_gam_females1)
#plot(model_gam_females1)


# Compute the distributions
summ_females<-summary(model_gam_females1)
T0.females<-summ_females$s.table[2,3]
T0.females
```

```
## [1] 1.973924
```

```
model_gam_noGDP.females<-gam(
  Prevalence_females ~
    Year +
    Country+
    s(HDI, bs = 'cr') +
    s(Education_females, bs = 'cr') +
    Affordability,
    data = data_gam
)
#summary(model_gam_noGDP.females)

res.females<-model_gam_noGDP.females$residuals
fitted.values.females<-model_gam_noGDP.females$fitted.values


T2.females<-numeric(B)
n<-nrow(data_gam)

for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res.females[permutation]
  response.perm <- fitted.values.females + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      s(data_gam$GDP, bs = 'cr') +
      s(data_gam$Education_females, bs = 'cr') +
      data_gam$Affordability
  )
  T2.females[perm] <- abs(summary(model.perm)$s.table[2,3])
}
diagnostic_permutation(T0.females,T2.females)
```



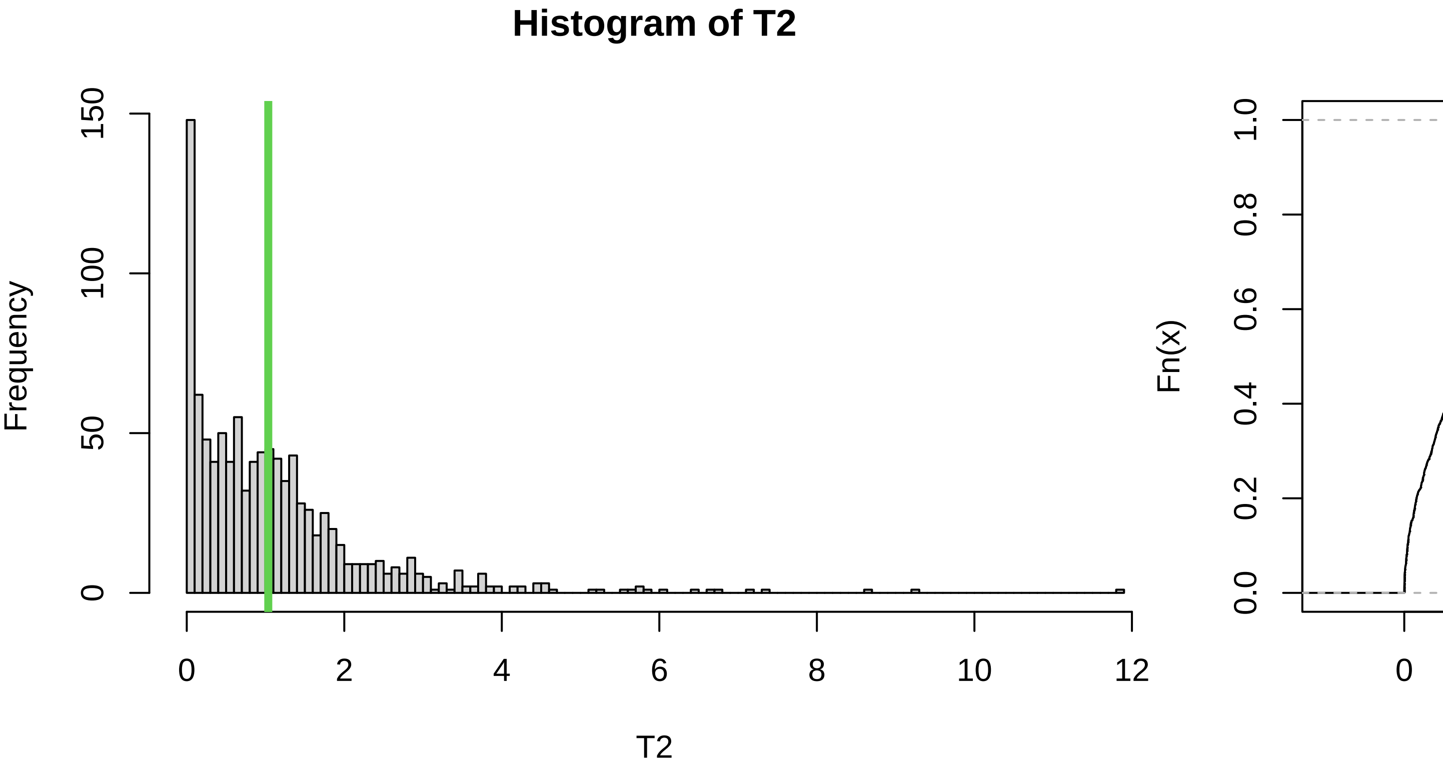
```
## p-value:  0.166
```

We cannot reject H0, hence we can simplify the model, so we remove the GDP. Now we try to remove education

```
model_gam_females.noedu <- gam(
  Prevalence_females ~
    Year +
    Country+
    s(HDI, bs = 'cr') +
    Affordability,
  data = data_gam
)
#summary(model_gam_females.noedu)
summ_females<-summary(model_gam_noGDP.females)

T0.females<-abs(summ_females$s.table[2,3])
T0.females
```

```
## [1] 1.780988
```

```
res.females<-model_gam_females.noedu$residuals
fitted.values.females<-model_gam_females.noedu$fitted.values

T2.females<-numeric(B)
n<-nrow(data_gam)

for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res.females[permutation]
  response.perm <- fitted.values.females + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      s(data_gam$Education_females, bs = 'cr') +
      data_gam$Affordability
  )
  T2.females[perm] <- abs(summary(model.perm)$s.table[2,3])
}
diagnostic_permutation(T0,T2.females)
```



```
## p-value:  0.419
```

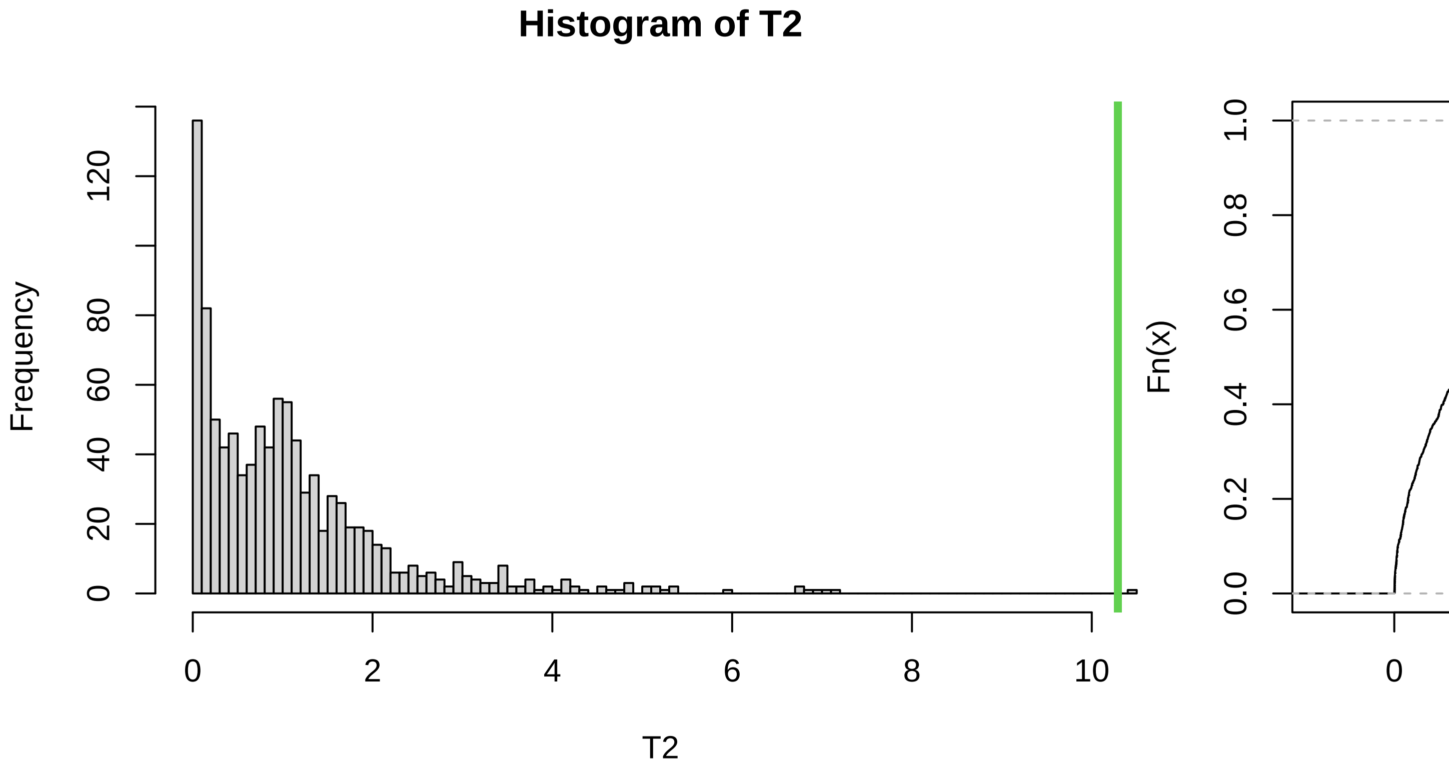
So we cannot reject H0, we remove education. Now we try to remove HDI

```
model_gam_females.nohdi <- gam(
  Prevalence_females ~
    Year +
    Country+
    Affordability,
  data = data_gam
)
#summary(model_gam_females.nohdi)
```

```
T0<-summary(model_gam_females.noedu)$s.table[1,3]
T0
```

```
## [1] 10.29072
```

```
res.females<-model_gam_females.nohdi$residuals
fitted.values.females<-model_gam_females.nohdi$fitted.values

T2.females<-numeric(B)
n<-nrow(data_gam)
set.seed(seed)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res.females[permutation]
  response.perm <- fitted.values.females + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      data_gam$Affordability
  )
  T2.females[perm] <- abs(summary(model.perm)$s.table[1,3])
}
diagnostic_permutation(T0,T2.females)
```



```
## p-value:  0.001
```

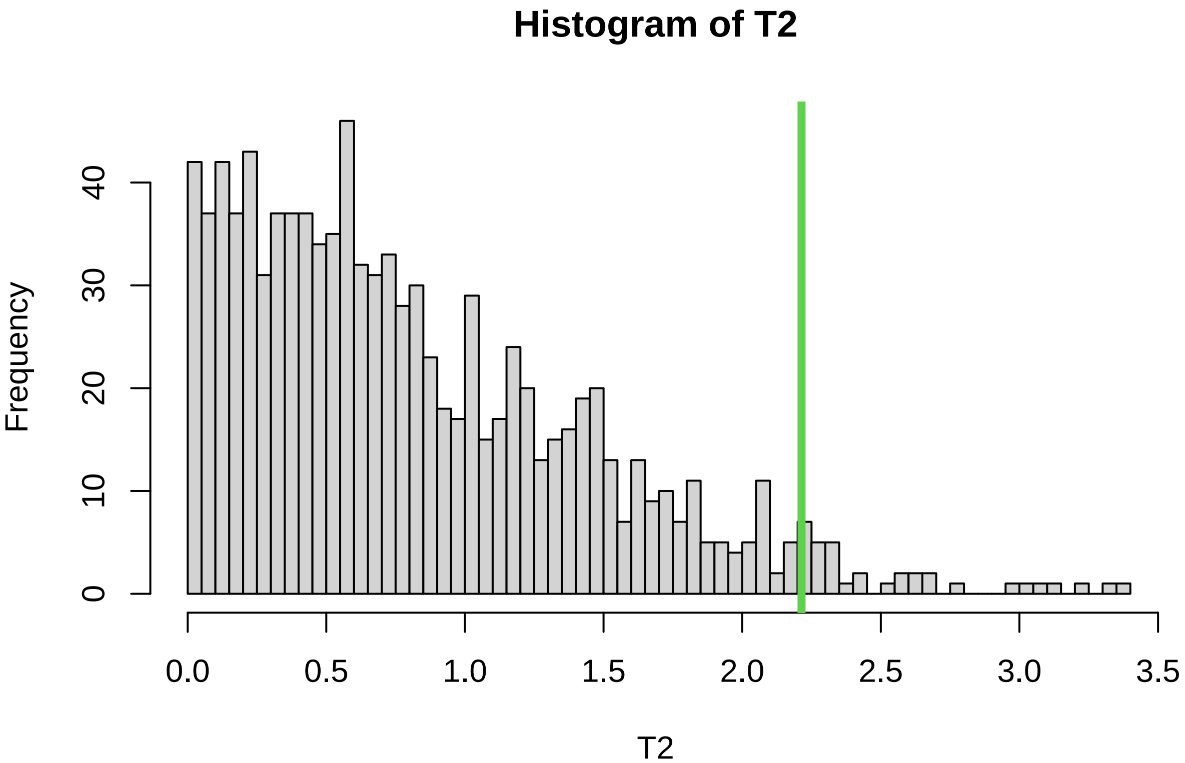
We reject H0, cannot remove HDI

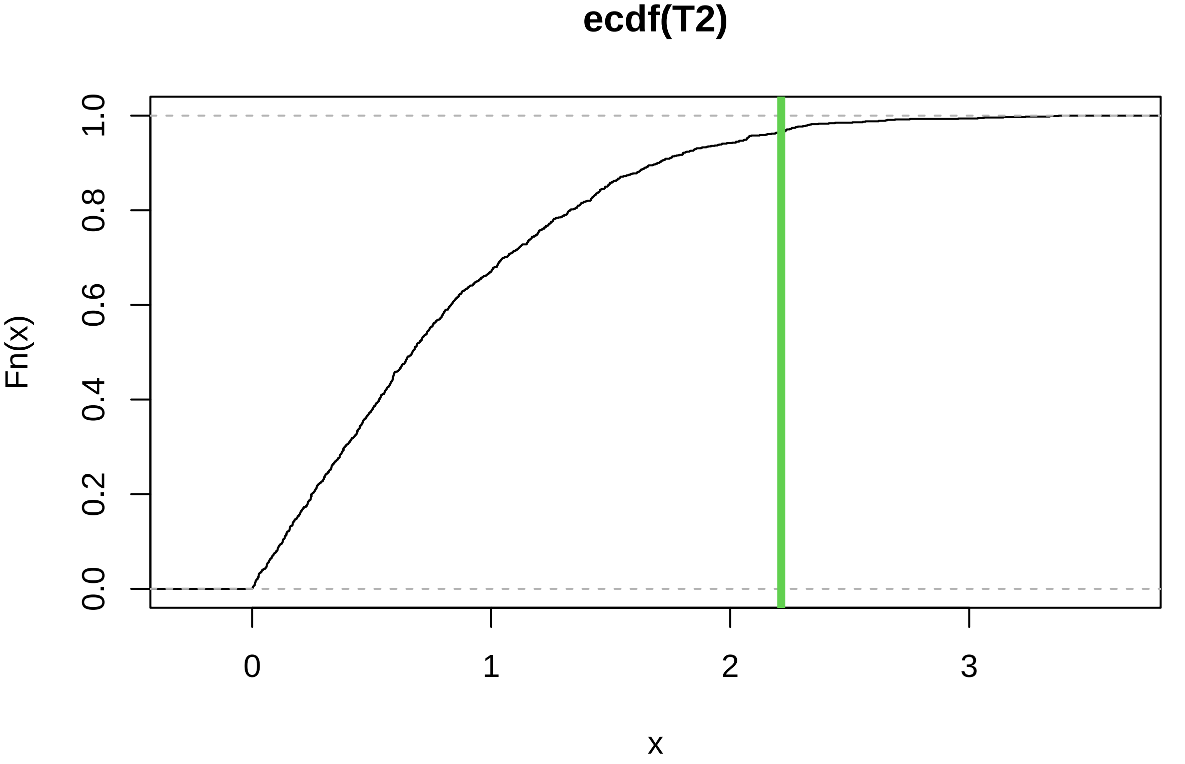
Now we try to remove affordability

```
model_gam_females.noaffordability <- gam(
  Prevalence_females ~
    Year +
    Country+
    s(HDI, bs = 'cr'),
  data = data_gam
)
#summary(model_gam_females.noaffordability)
#plot(model_gam_females.noaffordability)

summ_females<-summary(model_gam_females.noedu)
T0.females<-abs(summ_females$p.table[39,3])
T0.females
```

```
## [1] 2.214045
```

```
res.females<-model_gam_females.noaffordability$residuals
fitted.values.females<-model_gam_females.noaffordability$fitted.values

T2.females<-numeric(B)
n<-nrow(data_gam)
set.seed(seed)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res.females[permutation]
  response.perm <- fitted.values.females + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      data_gam$Affordability
  )
  T2.females[perm] <- abs(summary(model.perm)$p.table[39,3])
}

diagnostic_permutation(T0.females,T2.females)
```





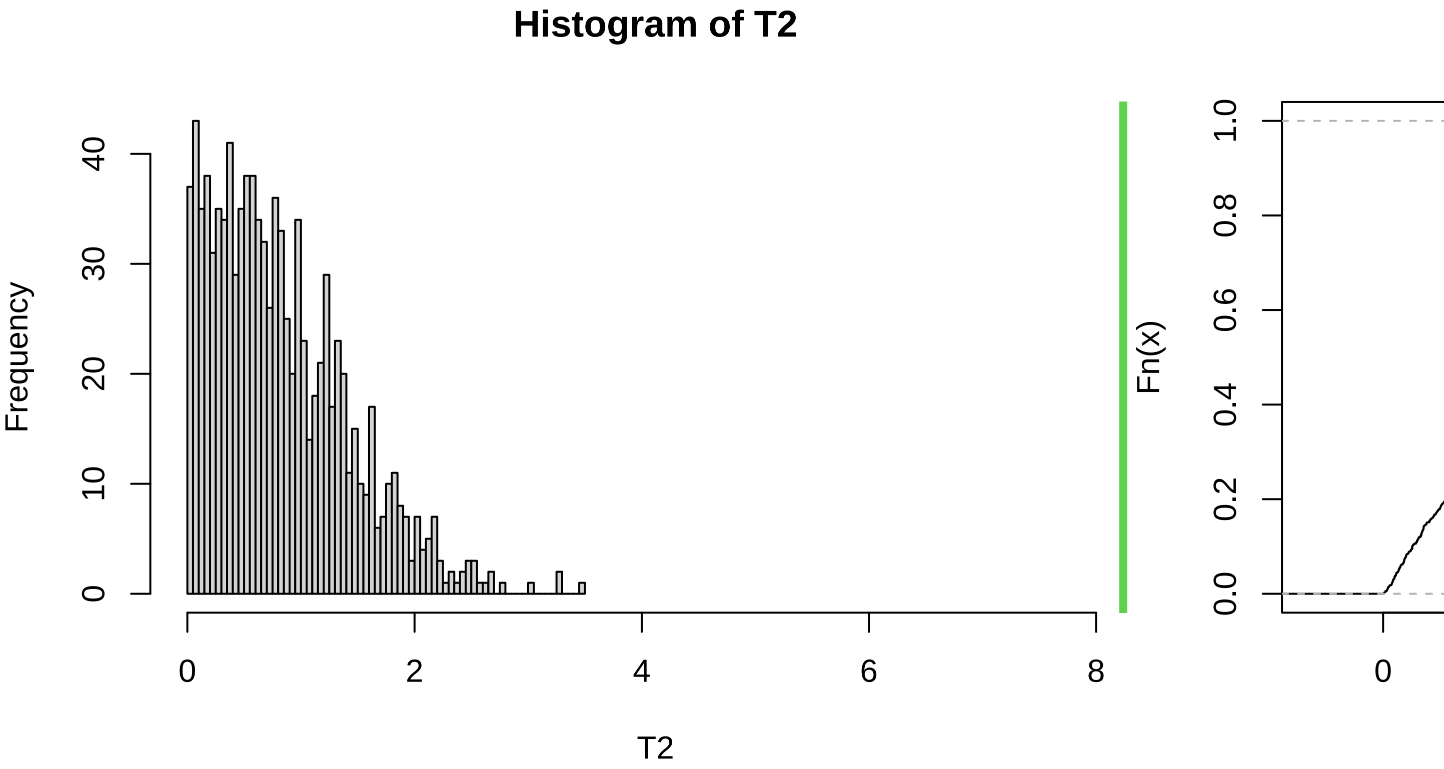
```
## p-value:  0.034
```

I can reject H0 and I keep affordability Finally, I try to remove the year

```r
model_gam_females.noyear <- gam(
  Prevalence_females ~
    Country+
    s(HDI, bs = 'cr')+
    Affordability,
  data = data_gam
)
#summary(model_gam_females.noyear)
#plot(model_gam_females.noaffordability)

summ_females<-summary(model_gam_females.noaffordability)
T0.females<-abs(summ_females$p.table[2,3])
T0.females
```

```
## [1] 8.23889
```

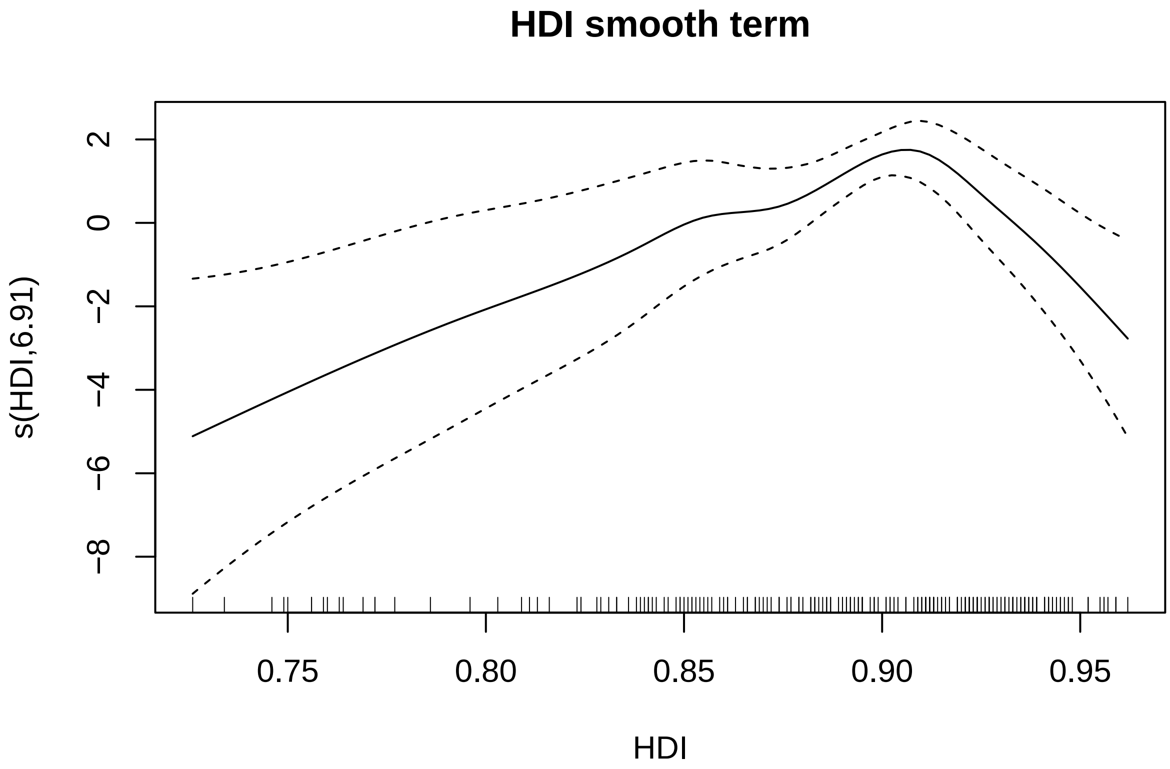
```r
res.females<-model_gam_females.noyear$residuals
fitted.values.females<-model_gam_females.noyear$fitted.values

T2.females<-numeric(B)
n<-nrow(data_gam)
set.seed(seed)
for(perm in 1:B){
  permutation <- sample(n)
  res.perm <- res.females[permutation]
  response.perm <- fitted.values.females + res.perm
  model.perm<-gam(
    response.perm ~
      data_gam$Year +
      data_gam$Country+
      s(data_gam$HDI, bs = 'cr') +
      data_gam$Affordability
  )
  T2.females[perm] <- abs(summary(model.perm)$p.table[2,3])
}
diagnostic_permutation(T0.females,T2.females)
```



```
## p-value:  0
```

We should keep the year

```
###Final model females
final_model_females<-gam(
  Prevalence_females ~
    Year +
    Country+
    s(HDI, bs = 'cr') +
    Affordability,
   data=data_gam
)
#summary(final_model_females)

plot(final_model_females,  main="HDI smooth term")
```



Bootstrap on the prediction for the final model for female prevalence

We now compute bootstrap prediction intervals on predictions for 2025, to assess which countries will reach DSG3 goal of relative 30% decrease from 2010 to 2025

```
prevf<-data_gam[data_gam$Country=="France",3][1]
prevf
```

```
## [1] 30.9
```

```
target<- prevf-0.3*prevf
target
```

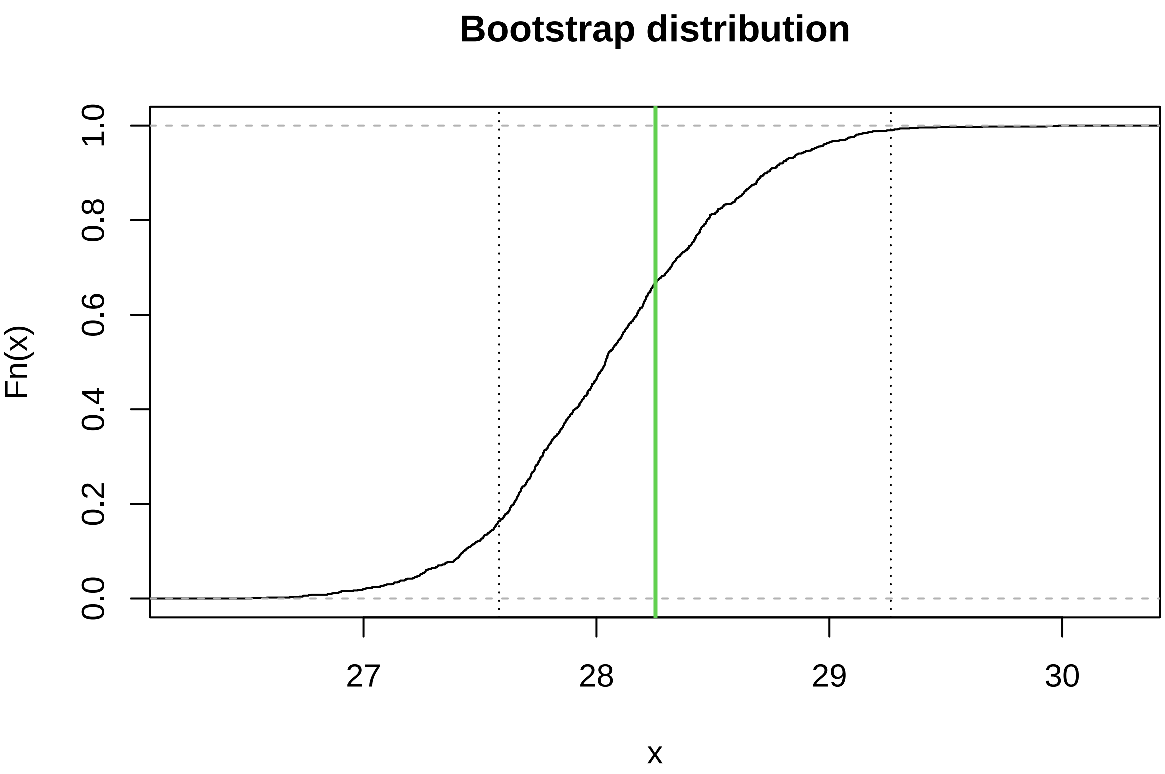
```
## [1] 21.63
```

```
new_obs<-data.frame(Country="France",Year=2025,HDI=0.902,Affordability=3.2)
predict(final_model_females,new_obs,se=TRUE)
```

```
## $fit
##        1
## 28.25355
##
## $se.fit
##         1
## 0.5584714
```

```r
xnew = new_obs
T.obs = predict(final_model_females,newdata=xnew)

fitted.obs<-predict(final_model_females, data_gam)
res.obs<- data_gam$Prevalence_females-fitted.obs #change y

T.boot <- numeric(B)
set.seed(seed)
pb = progress::progress_bar$new(total = B,
                                format = " Processing [:bar] :percent eta: :eta")
for(i in 1:B){
  response.b <- fitted.obs + sample(res.obs, replace = T)
  model.boot <- gam(response.b ~
                      Year +
                      Country+
                      s(HDI, bs = 'cr') +
                      Affordability,
                    data=data_gam)
  T.boot[i] <- predict(model.boot,newdata=xnew)
  pb$tick()
}
myalpha=0.1
diagnostic_bootstrap(T.boot, T.obs, alpha = myalpha)
```

```
## [1] "Variance:  0.262736379972276"
## [1] "Standard deviation:  0.512578169621255"
## [1] "Bias:  -0.19609271172083"
## [1] "MSE:  0.301188731562305"
##    lower   center    upper
## 27.58187 28.25355 29.26387
```



```
target
```

```
## [1] 21.63
```

France will not reach its SDG3 target for females, according to this prediction

We now try for Turkey

```
prevf<-data_gam[data_gam$Country=="Türkiye",3][1]
prevf
```

```
## [1] 16.7
```

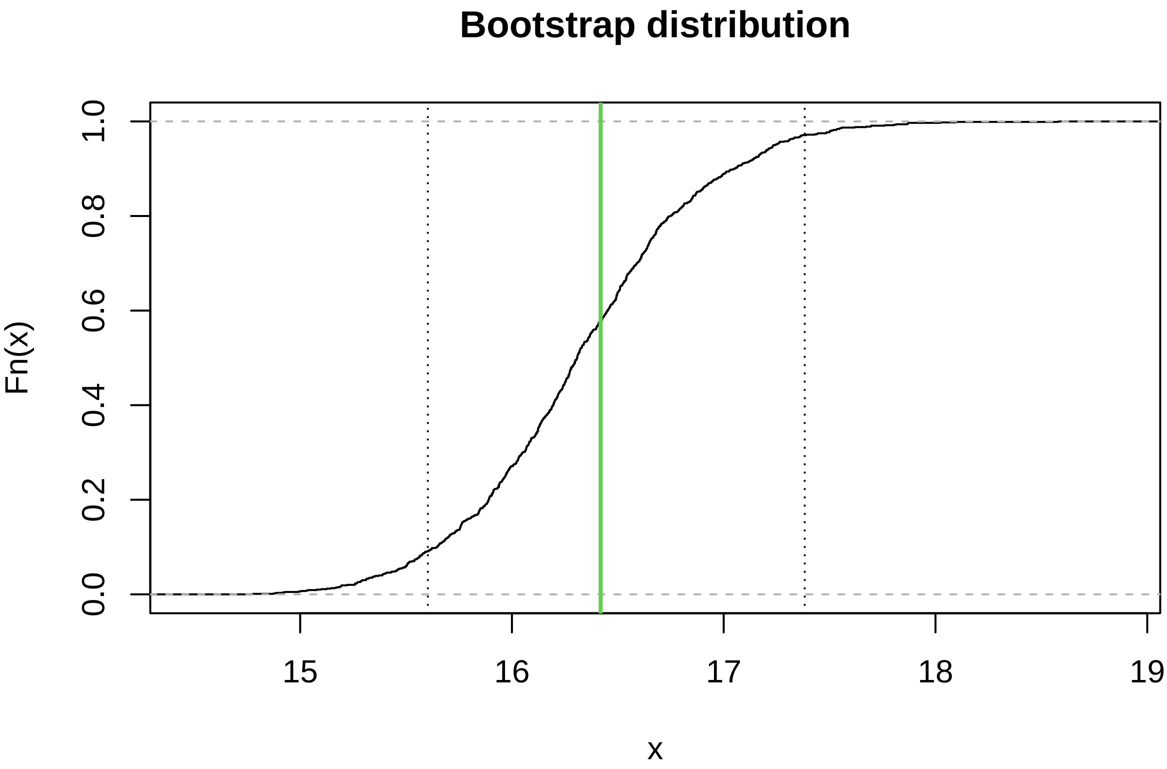
```
target<- prevf-0.3*prevf
target
```

```
## [1] 11.69
```

```
new_obs<-data.frame(Country="Türkiye",Year=2025,HDI=0.85,Affordability=3.8)
xnew = new_obs
T.obs = predict(final_model_females,newdata=xnew)

fitted.obs<-predict(final_model_females, data_gam)
res.obs<- data_gam$Prevalence_females-fitted.obs #change y

T.boot <- numeric(B)
set.seed(seed)
```

```
pb = progress::progress_bar$new(total = B,
                                format = " Processing [:bar] :percent eta: :eta")
for(i in 1:B){
  response.b <- fitted.obs + sample(res.obs, replace = T)

  model.boot <- gam(response.b ~
                          Year +
                          Country+
                          s(HDI, bs = 'cr') +
                          Affordability,
                        data=data_gam)
  T.boot[i] <- predict(model.boot,newdata=xnew)
  pb$tick()
}
myalpha=0.1
diagnostic_bootstrap(T.boot, T.obs, alpha = myalpha)
```

```
## [1] "Variance:  0.297282572103616"
## [1] "Standard deviation:  0.545236253475148"
## [1] "Bias:  -0.0910742001050018"
## [1] "MSE:  0.305577082028382"
##    lower   center    upper
## 15.60299 16.41899 17.38266
```



```
target
```

```
## [1] 11.69
```

Turkey will not reach its SDG3 target for females, according to this prediction

We now try for Norway

```
prevf<-data_gam[data_gam$Country=="Norway",3][1]
prevf
```

```
## [1] 25.8
```

```
target<- prevf-0.3*prevf
target
```

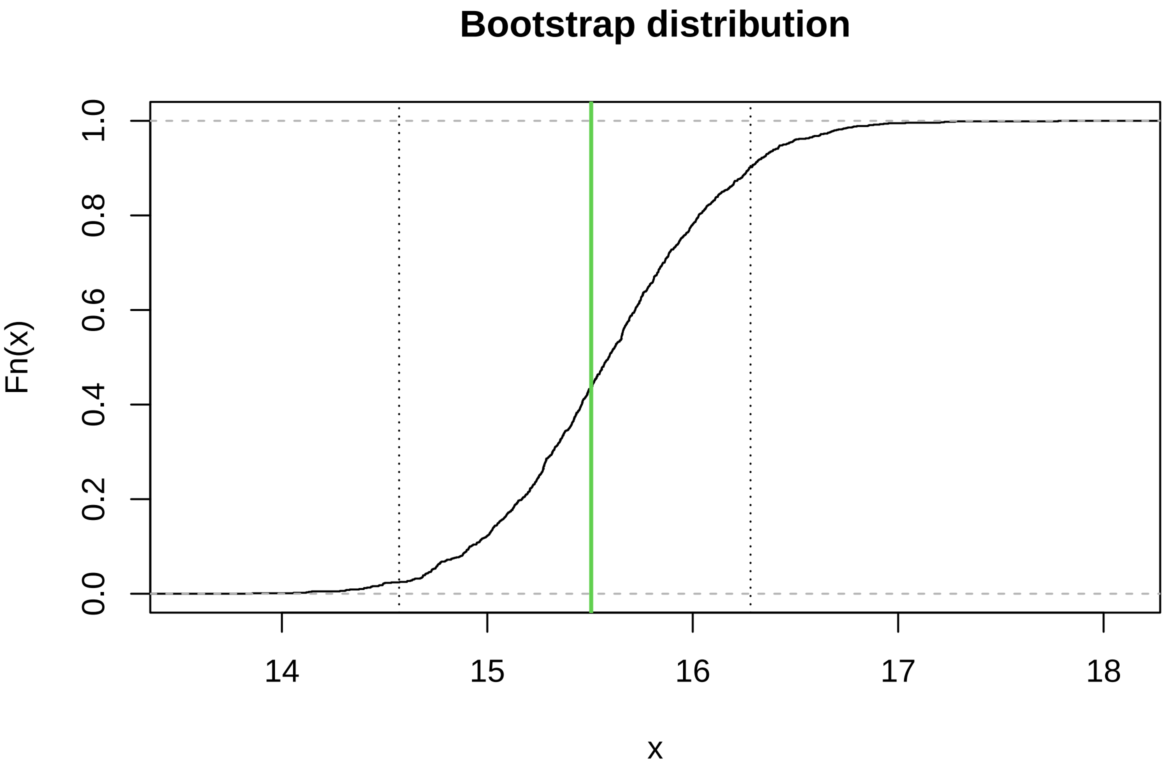
```
## [1] 18.06
```

```
new_obs<-data.frame(Country="Norway",Year=2025,HDI=0.966,Affordability=2.22)
predict(final_model_females,new_obs,se=TRUE)
```

```
## $fit
##        1
## 15.50573
##
## $se.fit
##         1
## 0.6730505
```

```
xnew = new_obs
T.obs = predict(final_model_females,newdata=xnew)

fitted.obs<-predict(final_model_females, data_gam)
res.obs<- data_gam$Prevalence_females-fitted.obs

T.boot <- numeric(B)
set.seed(seed)
pb = progress::progress_bar$new(total = B,
                                format = " Processing [:bar] :percent eta: :eta")
for(i in 1:B){
  response.b <- fitted.obs + sample(res.obs, replace = T)
  model.boot <- gam(response.b ~
                      Year +
                      Country+
                      s(HDI, bs = 'cr') +
                      Affordability,
                    data=data_gam)
  T.boot[i] <- predict(model.boot,newdata=xnew)
  pb$tick()
}
myalpha=0.1
diagnostic_bootstrap(T.boot, T.obs, alpha = myalpha)
```

```
## [1] "Variance:  0.279283484705486"
## [1] "Standard deviation:  0.528472785207986"
## [1] "Bias:  0.0907654661997466"
## [1] "MSE:  0.287521854559944"
##    lower   center    upper
## 14.57067 15.50573 16.28126
```



```
target
```

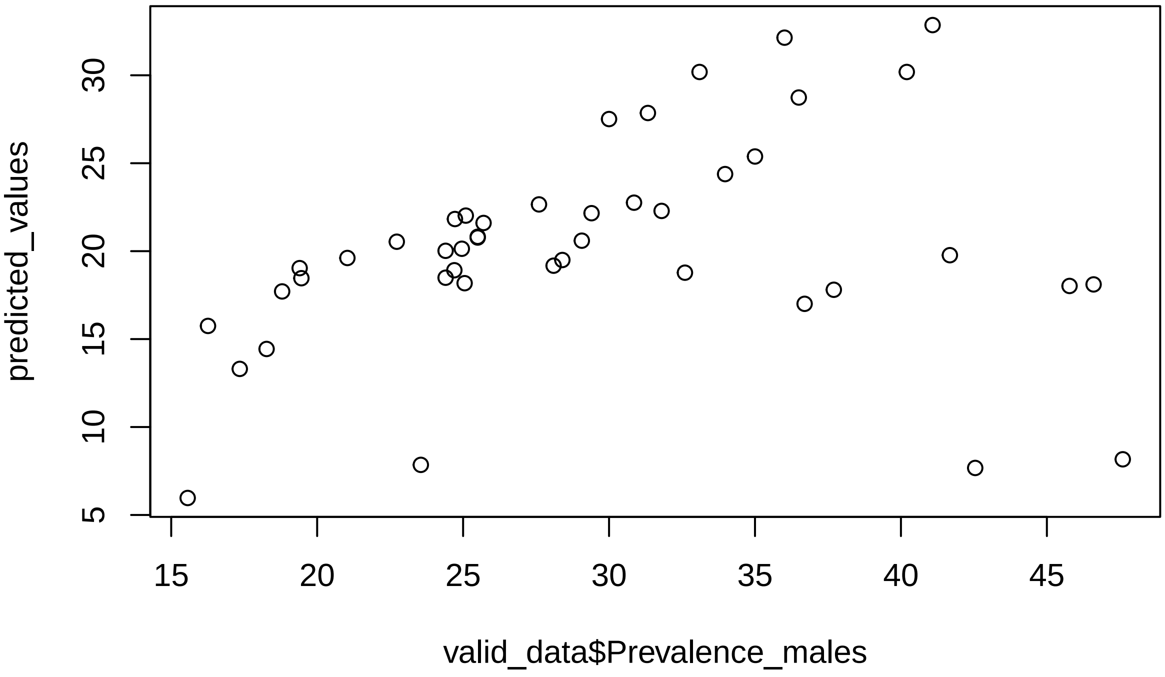
```
## [1] 18.06
```

Norway will probaby reach the target

EXTRA:

First using MPE metric defined as follows

```
# Specify the number of folds for cross-validation
num_folds <- 5

# Create an empty vector to store the evaluation metric for each fold
evaluation_metric <- numeric(num_folds)

# Perform cross-validation
set.seed(2022)  # For reproducibility

# Create indices for cross-validation folds
folds <- sample(rep(1:num_folds, length.out = nrow(data_gam)))

mpe <- function(actual, predicted) {
  n <- length(actual)
  mpe_val <- (1/n) * sum(abs((actual - predicted)) / actual) * 100
  return(mpe_val)
}

# Iterate over  each fold
for (i in 1:num_folds) {
  # Split the data into training and validation sets based on the fold
  train_data <- data_gam[folds != i, ]  # Training set
  valid_data <- data_gam[folds == i, ]  # Validation set

  # Fit the GAM model with GCV using the training set
  gam_model <- gam(Prevalence_females ~
                     Year +
                     Country+
                     s(HDI, bs = 'cr') +
                    # s(Education_females, bs = 'cr') +
                     Affordability,
                   data = train_data)

  # Make predictions on the validation set
  predicted_values <- predict(gam_model, newdata = valid_data)

  # Calculate the evaluation metric (e.g., MSE) for the fold
  evaluation_metric[i] <- mpe(valid_data$Prevalence_females, predicted_values)
}
plot(valid_data$Prevalence_males,predicted_values)
```



```
# Calculate the average evaluation metric across all folds
average_metric <- (mean(evaluation_metric))

# Print the average evaluation metric
cat("Average Evaluation Metric:", average_metric, "\n")
```

```
## Average Evaluation Metric: 4.993702
```

```
# 0.0001605886
```

Then using MSE

```
data_gam$HDI_MHI_clustering<-as.factor(data_gam$HDI_MHI_clustering)
#######
model_gam_both <- gam(
  Prevalence_both ~
    Year +
    s(HDI, bs = 'cr') +
    s(GDP, bs = 'cr') +
    s(Education, bs = 'cr') +
    HDI_MHI_clustering+
    Affordability,
  data = data_gam
)

summary(model_gam_both)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Prevalence_both ~ Year + s(HDI, bs = "cr") + s(GDP, bs = "cr") +
##     s(Education, bs = "cr") + HDI_MHI_clustering + Affordability
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          65.60589  198.01242   0.331  0.74075
## Year                 -0.02168    0.09814  -0.221  0.82538
## HDI_MHI_clustering2   6.92649    2.08259   3.326  0.00105 **
## HDI_MHI_clustering3   6.18255    2.61381   2.365  0.01899 *
## HDI_MHI_clustering4   8.75109    2.92276   2.994  0.00311 **
## Affordability        -1.77336    0.60248  -2.943  0.00364 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df      F  p-value
## s(HDI)       7.031  8.054  8.401  < 2e-16 ***
## s(GDP)       8.024  8.653 11.383  < 2e-16 ***
## s(Education) 4.312  5.341  5.153 0.000162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.757   Deviance explained = 78.4%
## GCV = 13.936  Scale est. = 12.344    n = 222
```

```
set.seed(2022)  # For reproducibility

# Create indices for cross-validation folds

data_gam$Prevalence_females<-data_gam$Prevalence_females
num_folds=5
folds <- sample(rep(1:num_folds, length.out = nrow(data_gam)))
evaluation_metric<-numeric(num_folds)
# Iterate over each fold
for (i in 1:num_folds) {
  # Split the data into training and validation sets based on the fold
  train_data <- data_gam[folds != i, ]  # Training set
  valid_data <- data_gam[folds == i, ]  # Validation set

  # Fit the GAM model with GCV using the training set
  gam_model <- gam(Prevalence_females ~
                     Year+
                     Country+
                     s(HDI, bs = 'cr') +
                     Affordability,
                   data = train_data
                   )
  # Make predictions on the validation set
  predicted_values <- predict(gam_model, newdata = valid_data)
  #predicted_values <- plogis(predicted_values_l)

  #Calculate the evaluation metric (e.g., MSE) for the fold
  evaluation_metric[i] <- mean((valid_data$Prevalence_females - (predicted_values))^2)
}

# Calculate the average evaluation metric across all folds (e.g. RMSE)
average_metric <- (sqrt(mean(evaluation_metric)))

# Print the average evaluation metric
cat("Average Evaluation Metric:", average_metric, "\n") #1.250058
```

```
## Average Evaluation Metric: 1.267663
```