exploratory data analysis and models on the epi dataset

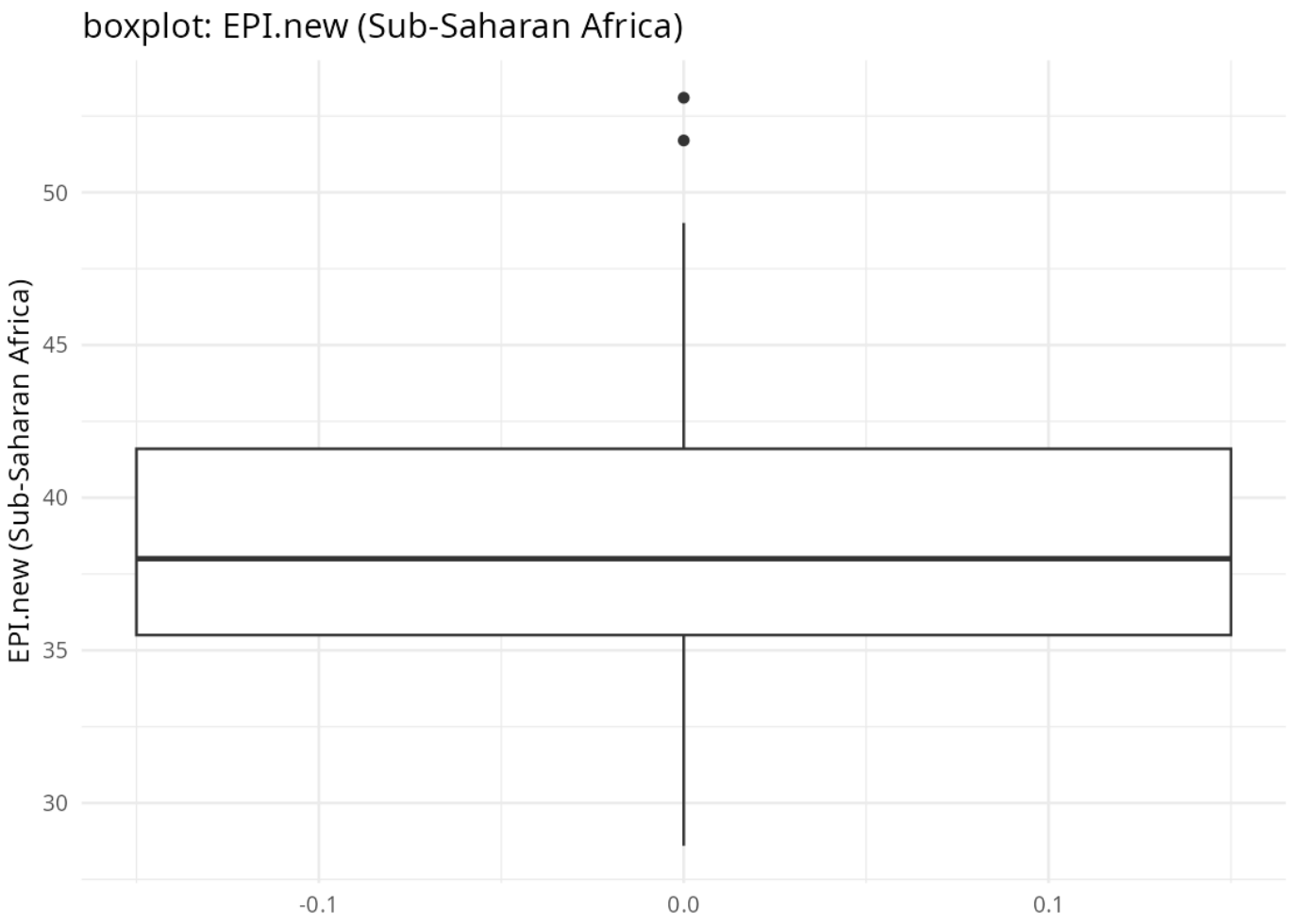
date: 2025-10-13

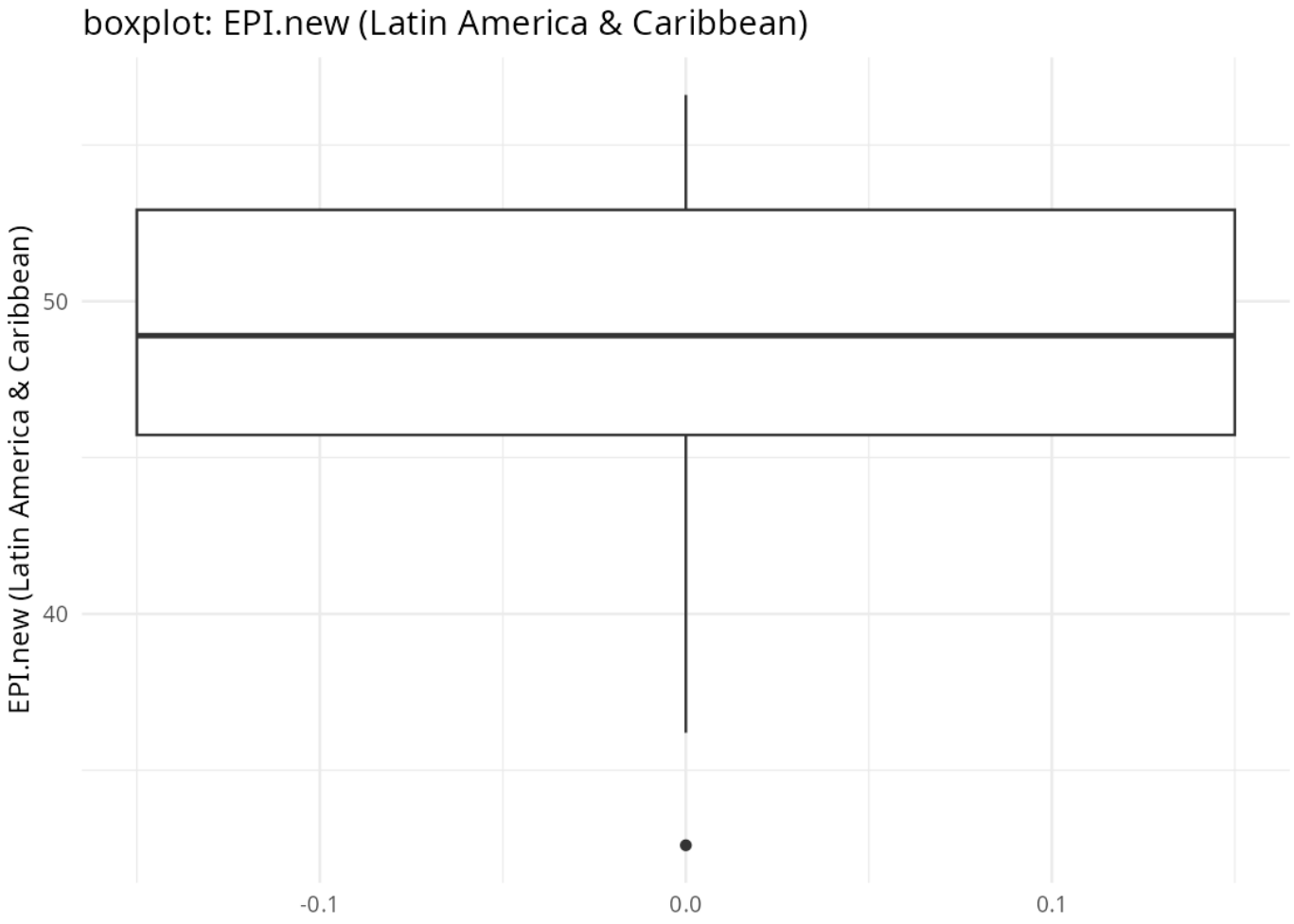
dataset and choices

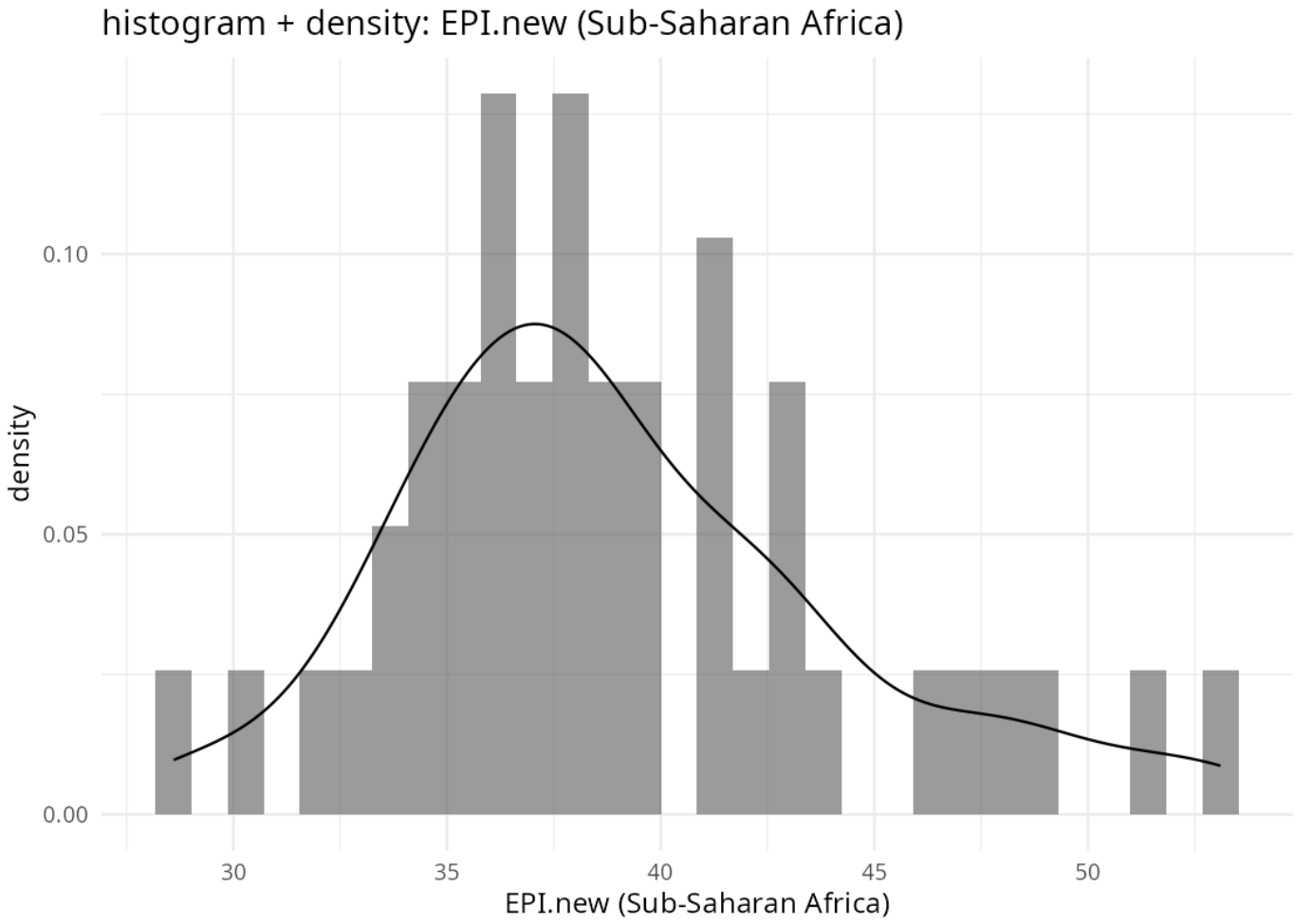
- **file**: `epi_results_2024_pop_gdp_v2.csv`
- **region column**: `region`
- **response var**: `EPI.new`
- **regions**: `Sub-Saharan Africa` vs `Latin America & Caribbean`

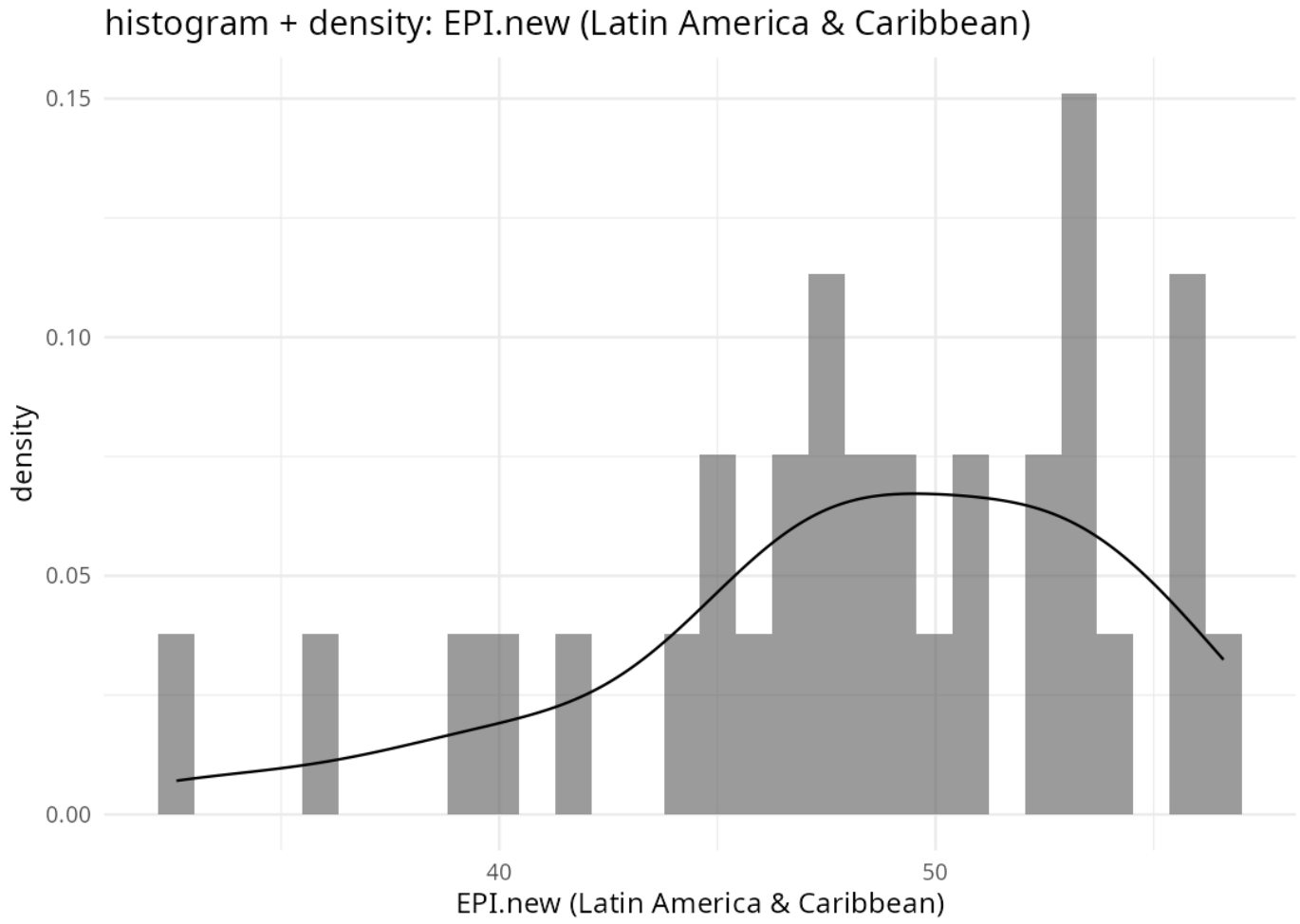
1) variable distributions

1.1 boxplots and histograms (with density!)

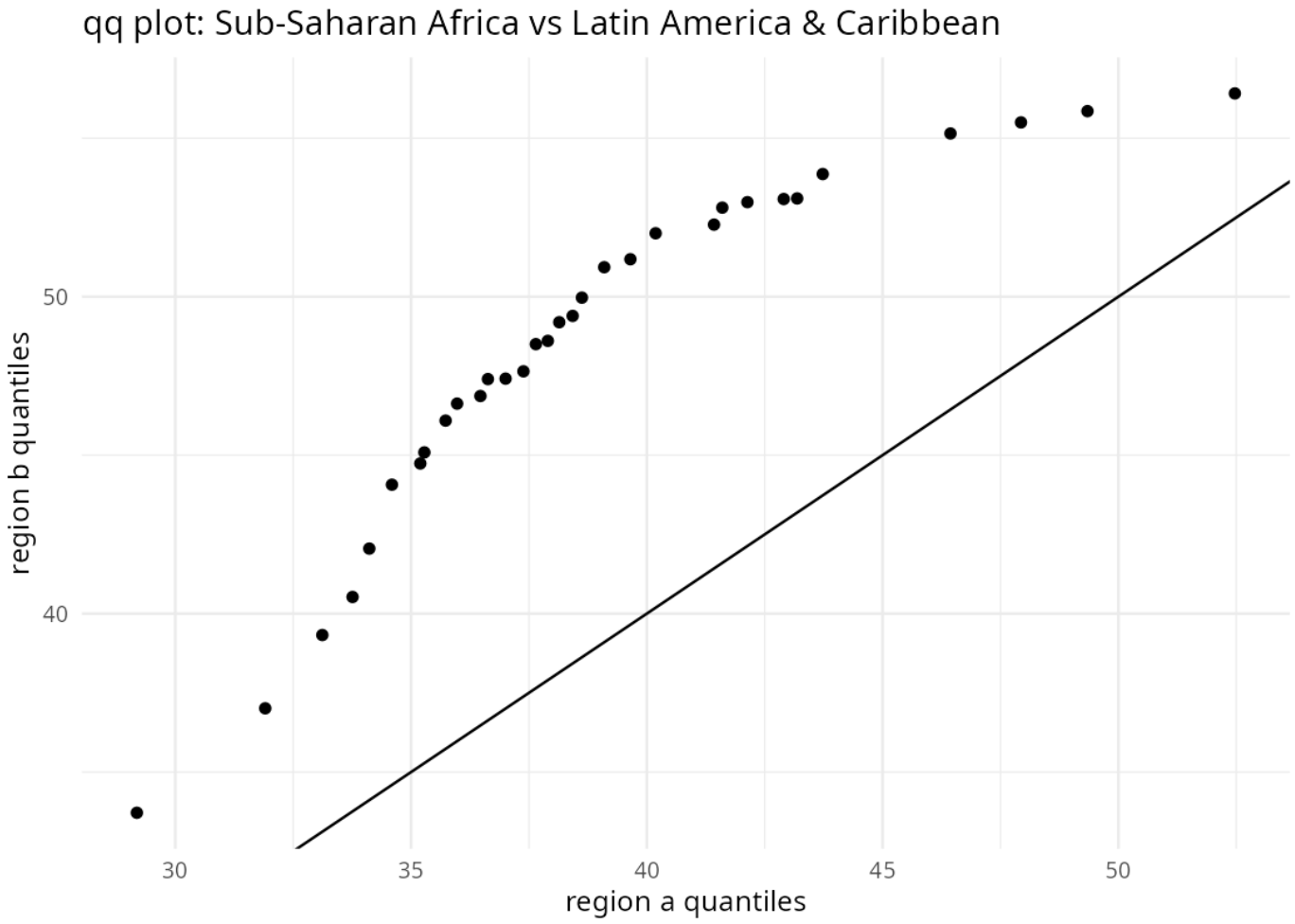



boxplot: EPI.new (Latin America & Caribbean)
EPI.new (Latin America & Caribbean)
50
40
-0.1
0.0
0.1


histogram + density: EPI.new (Sub-Saharan Africa)
density
0.10
0.05
0.00
30
35
40
45
50
EPI.new (Sub-Saharan Africa)



1.2 qq plot (two-sample)



2) linear models

full: EPI.new ~ gdp

full: EPI.new ~ gdp + population

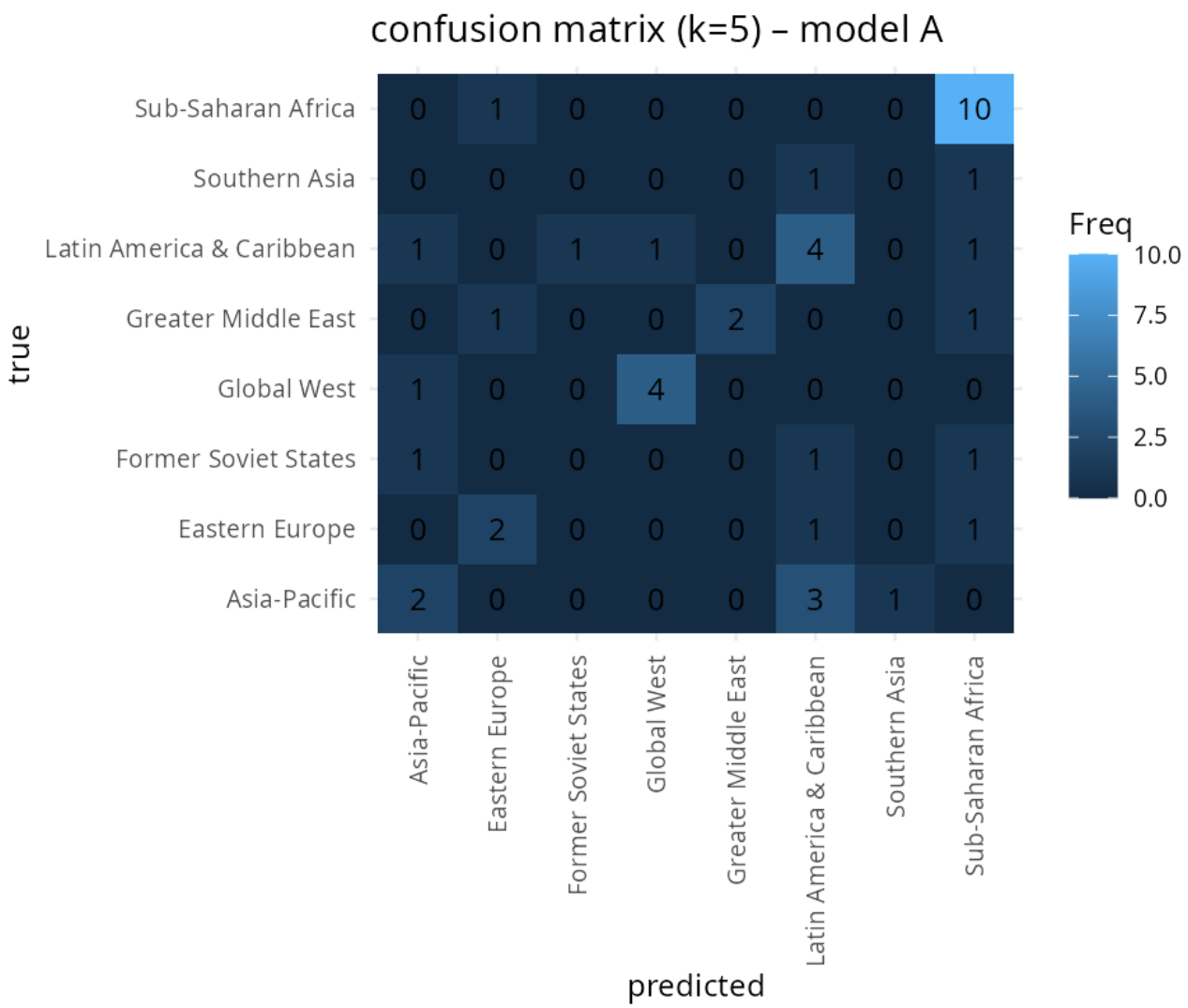
2.2 same models on one region (comparison)

on region `Sub-Saharan Africa`, the better model is **region Sub-Saharan Africa: EPI.new ~ gdp + population** (r^2=0.361, aic=265.4, bic=272.7).

3) classification (knn, label = region)

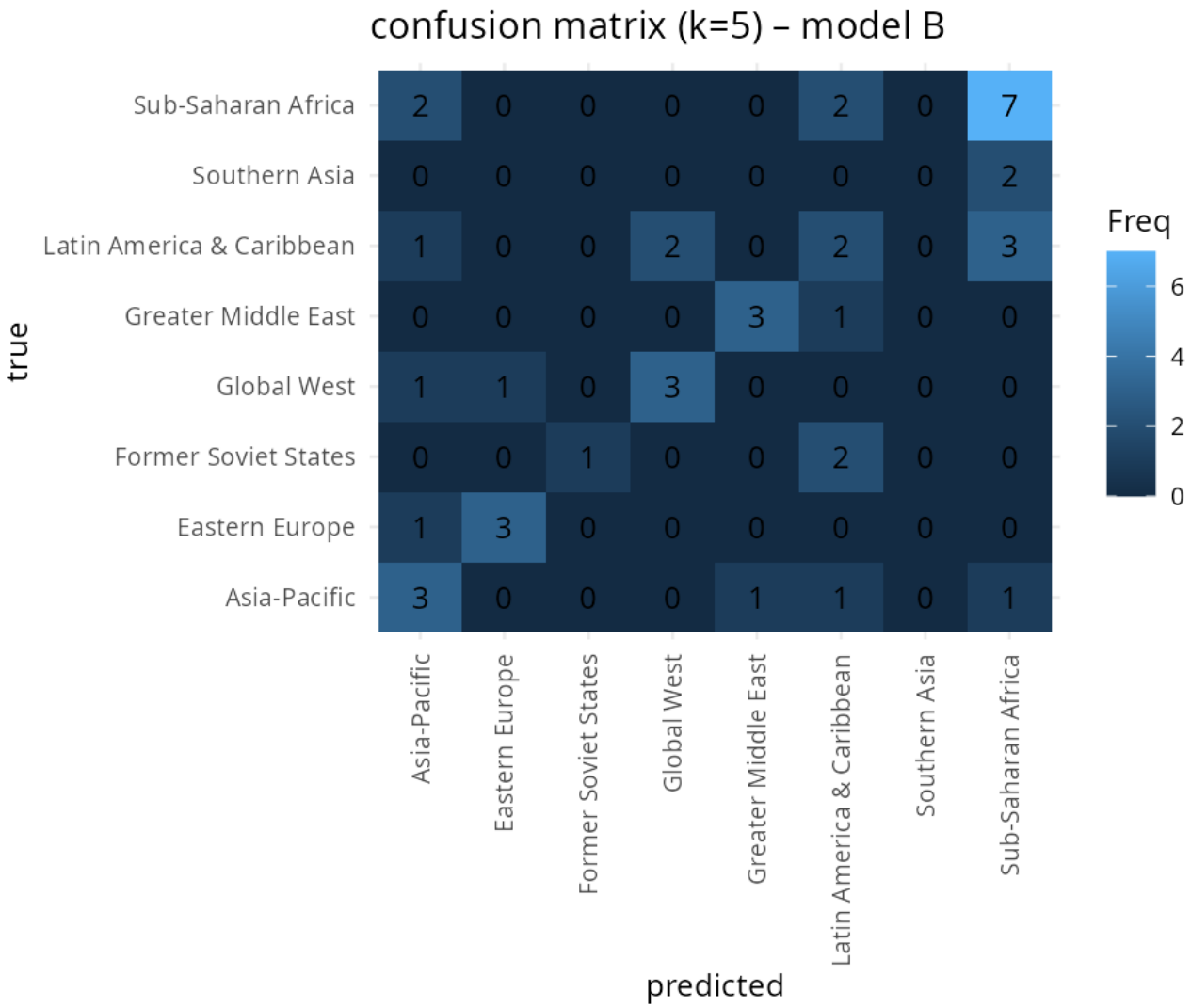
model A

- **k**: 5 | **accuracy**: 0.5581 | **test n**: 43 variables: `c("AGR.new", "AIR.new", "APO.new")`



model B

- **k**: 5 | **accuracy**: 0.5116 | **test n**: 43 variables: `c("BCA.new", "BDH.new", "CBP.new")`


confusion matrix (k=5) – model B
true
Sub-Saharan Africa
Southern Asia
Latin America & Caribbean
Greater Middle East
Global West
Former Soviet States
Eastern Europe
Asia-Pacific
2 0 0 0 0 2 0 7
0 0 0 0 0 0 0 2
1 0 0 2 0 2 0 3
0 0 0 0 3 1 0 0
1 1 0 3 0 0 0 0
0 0 1 0 0 2 0 0
1 3 0 0 0 0 0 0
3 0 0 0 1 1 0 1
Asia-Pacific
Eastern Europe
Former Soviet States
Global West
Greater Middle East
Latin America & Caribbean
Southern Asia
Sub-Saharan Africa
predicted
Freq
6
4
2
0