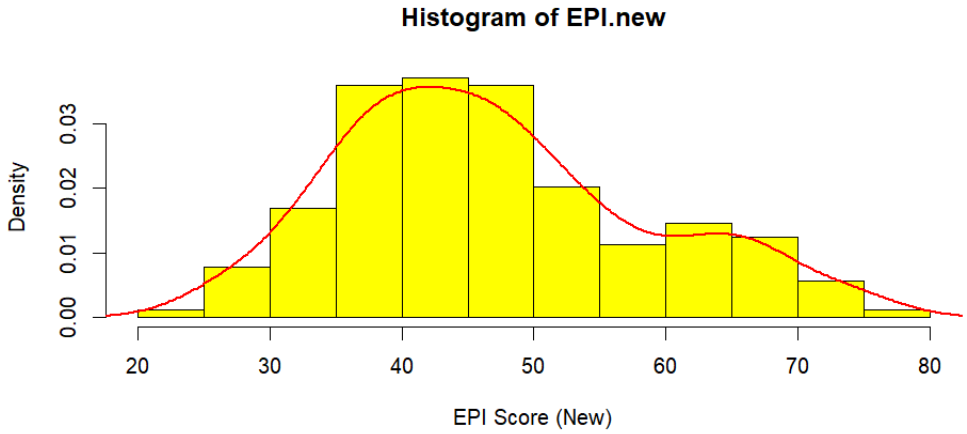
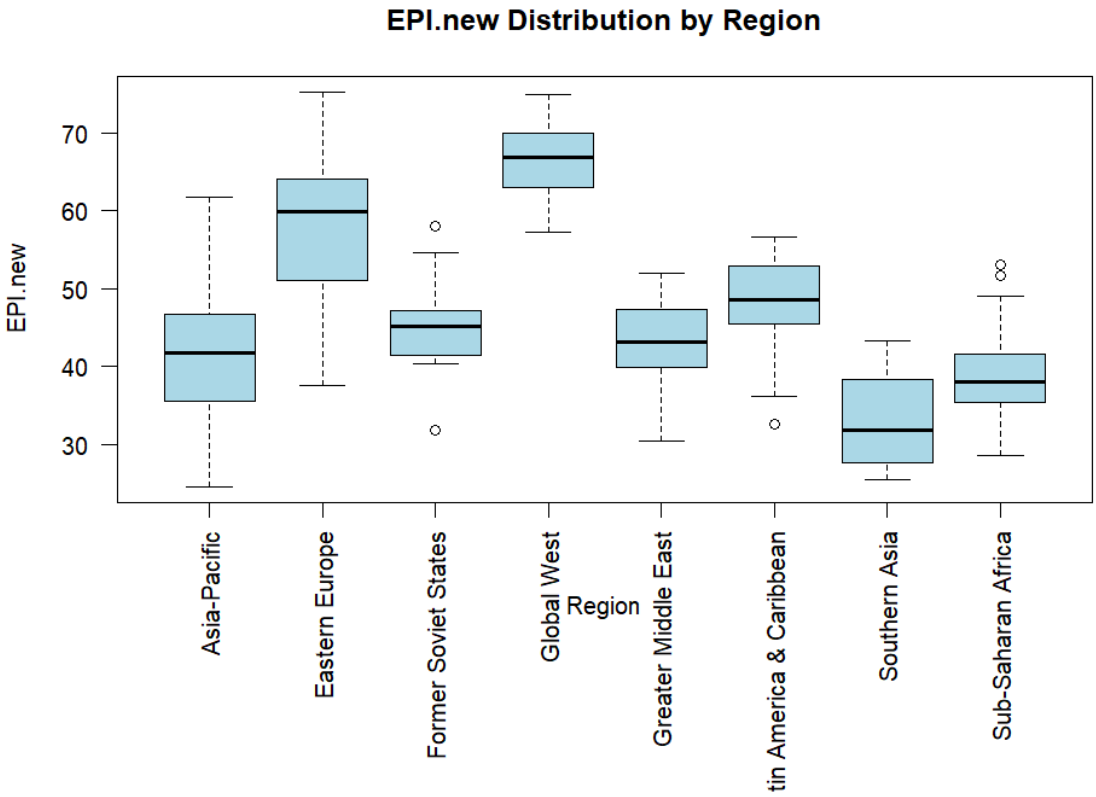
Written + Figures:

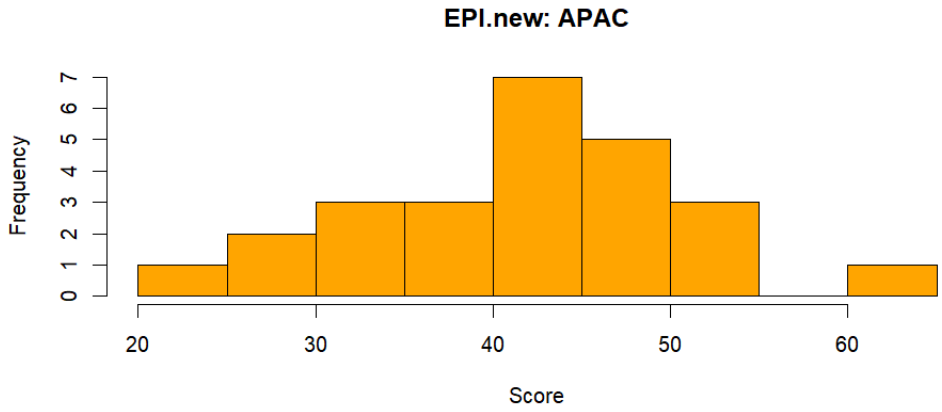
1.1. Histogram of EPI.new with density line overlayed

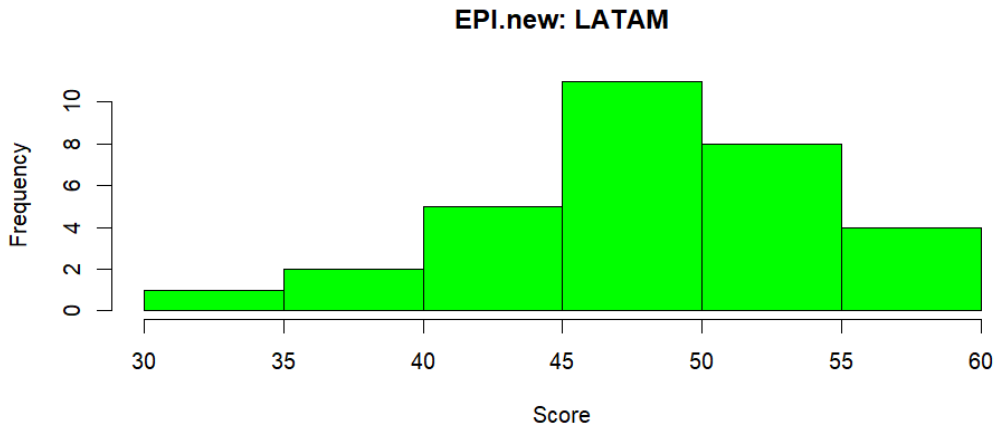


1.2. Boxplots of EPI.new for each region

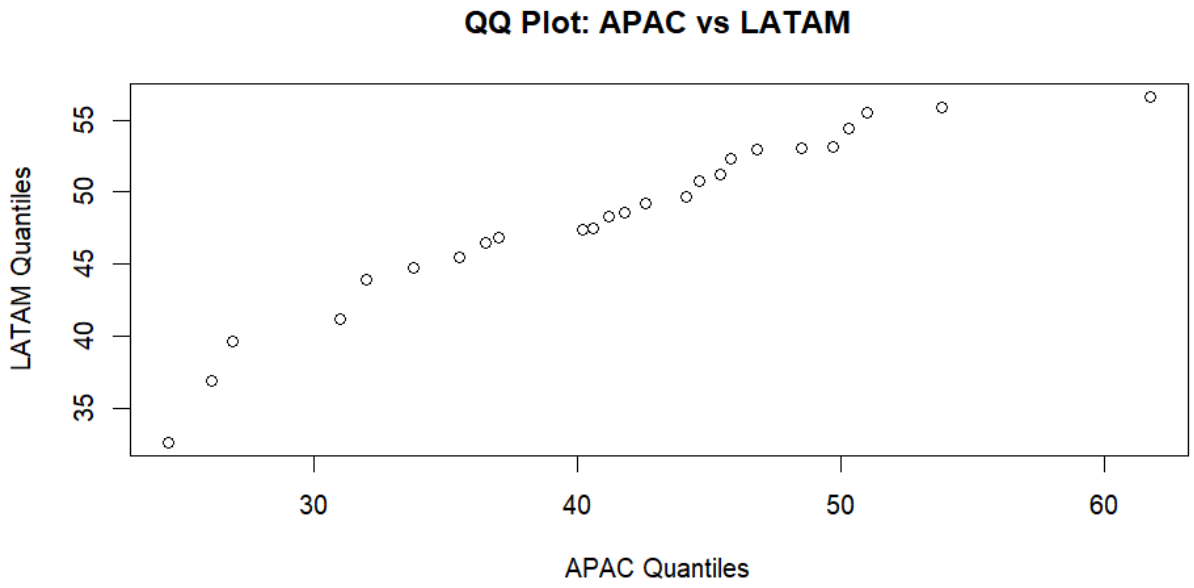


2.1. Histograms of EPI.new for each region

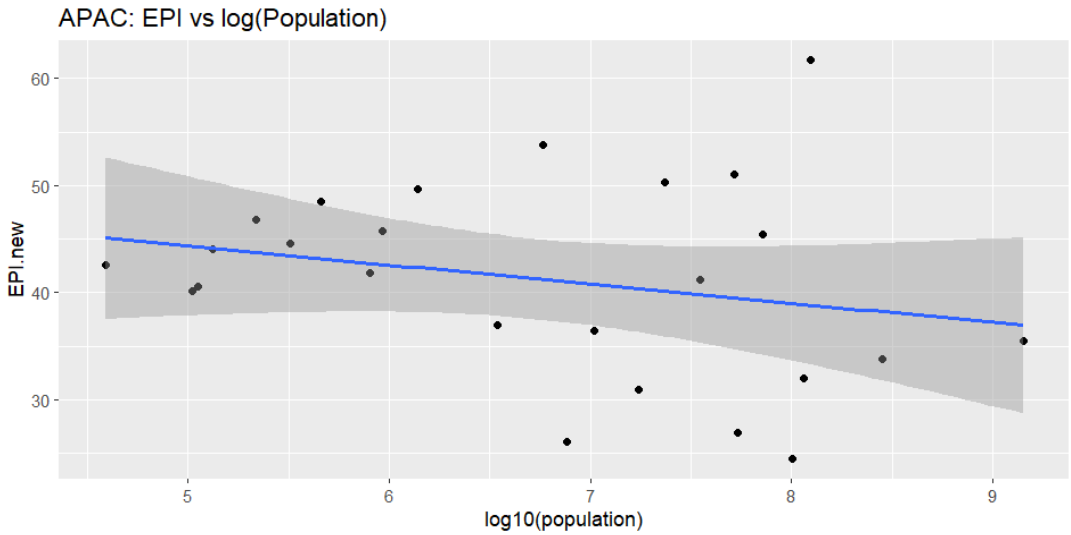


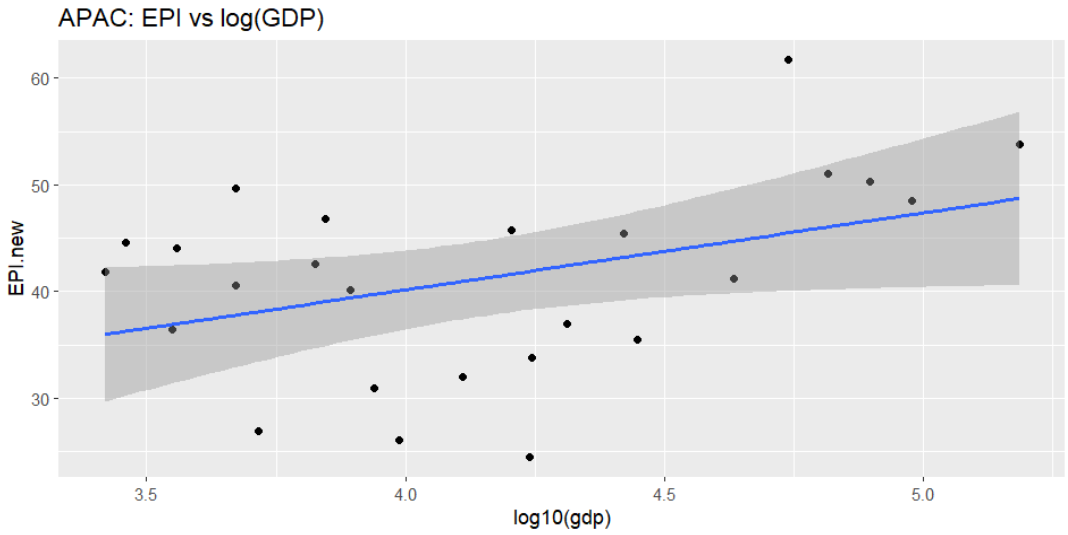


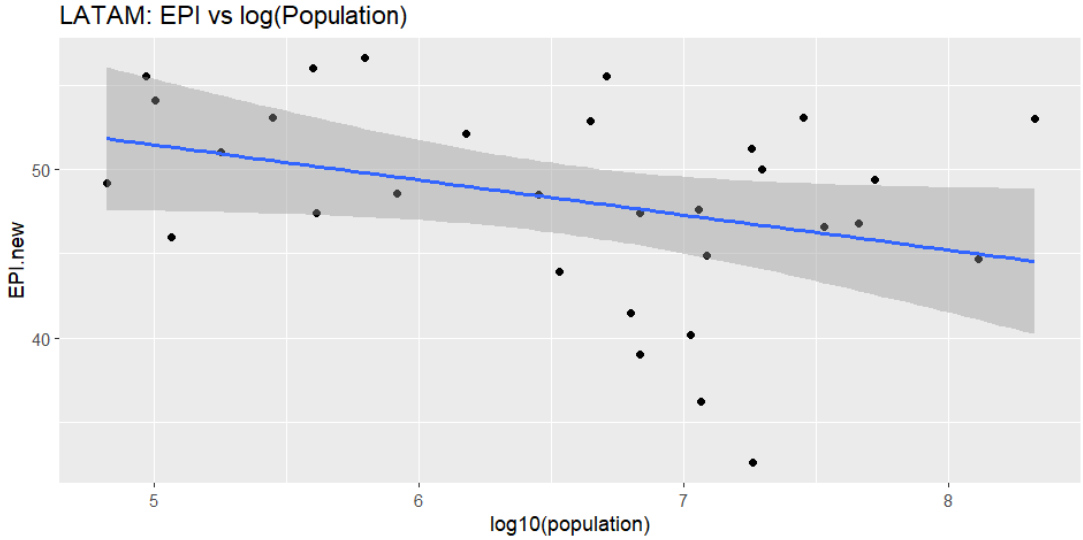
2.2. QQ Plot for EPI.new between APAC and LATAM

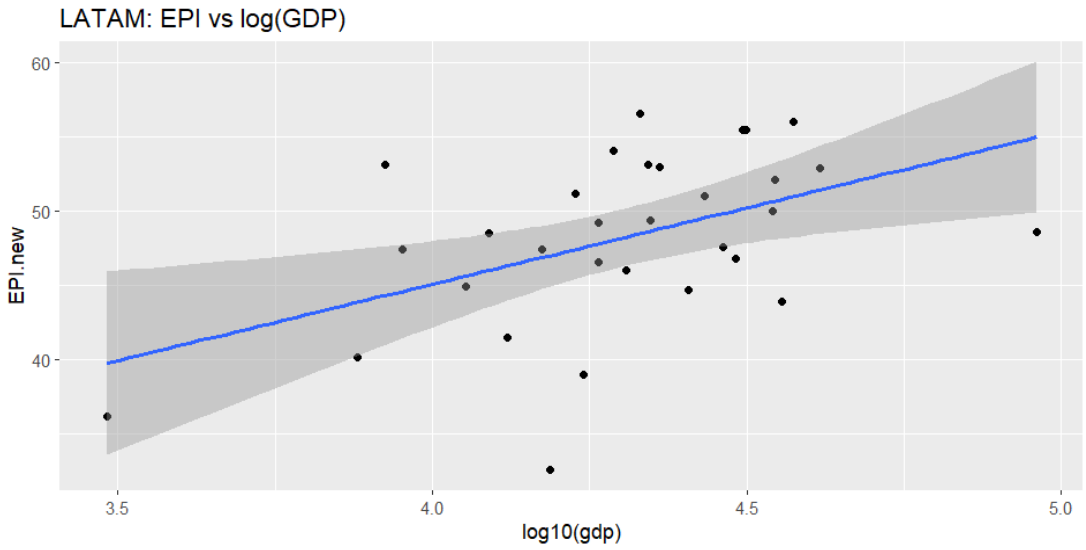


3.1. Population and GDP against EPI.new for APAC and LATAM









3.2. Linear Models

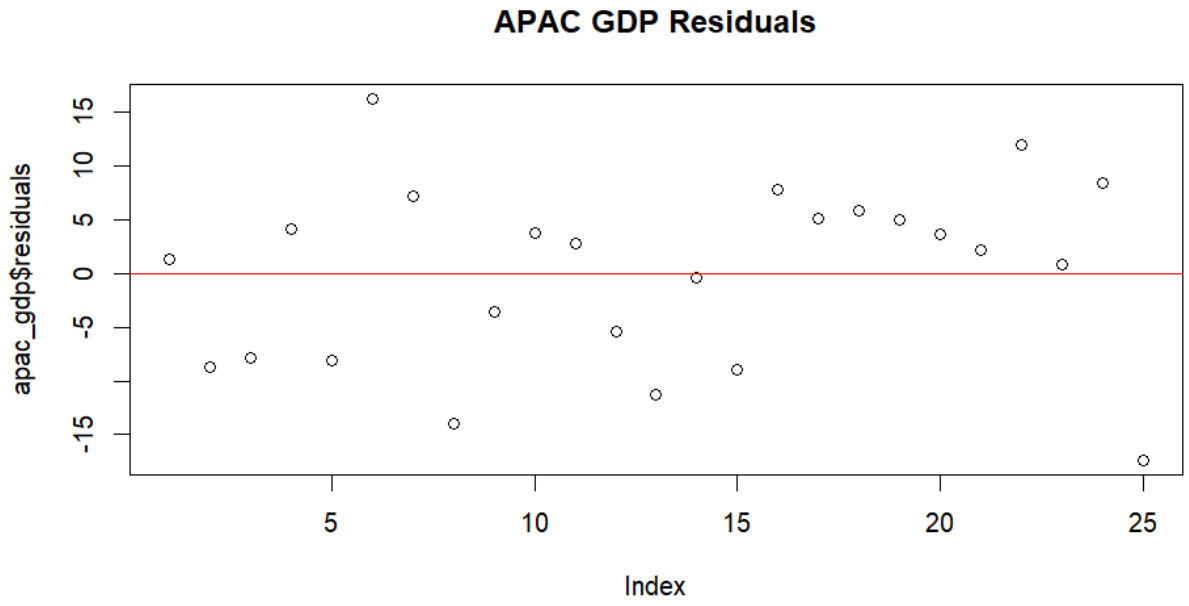
APAC GDP Model Summary

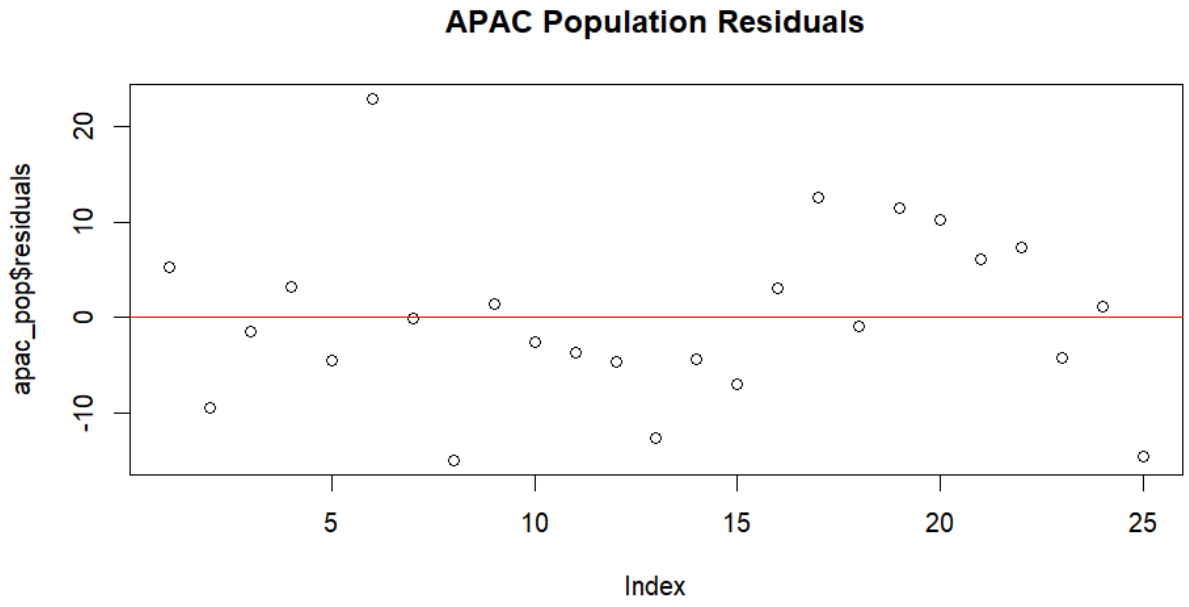
```
Call:
lm(formula = EPI.new ~ log10(gdp), data = apac)

Residuals:
    Min      1Q  Median      3Q     Max
-17.395  -7.896   2.187   5.069  16.195

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.313     14.273   0.793   0.4361
log10(gdp)     7.214      3.414   2.113   0.0457 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.542 on 23 degrees of freedom
Multiple R-squared:  0.1626,   Adjusted R-squared:  0.1262
F-statistic: 4.465 on 1 and 23 DF,  p-value: 0.04566
```

APAC Population Model Summary

```
Call:
lm(formula = EPI.new ~ log10(population), data = apac)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9132  -4.4214  -0.9645   5.3048  22.8464

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         53.299     10.084   5.285  2.3e-05 ***
```

```
log10(population)    -1.785      1.470  -1.214    0.237
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.049 on 23 degrees of freedom
Multiple R-squared:  0.06021,  Adjusted R-squared:  0.01935
F-statistic: 1.474 on 1 and 23 DF,  p-value: 0.2371
```



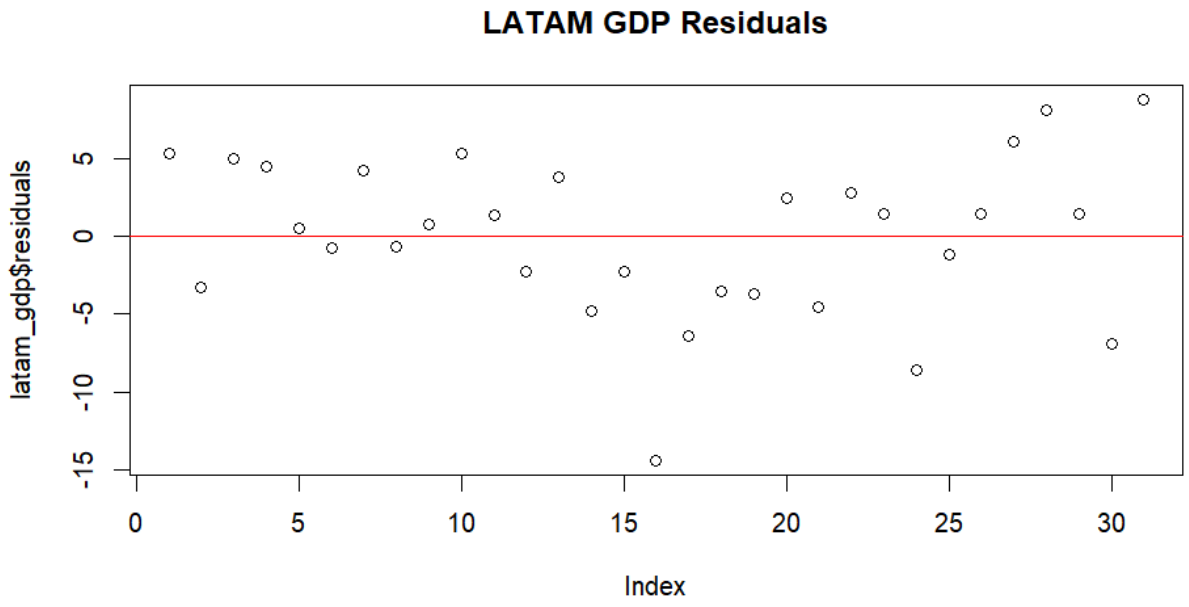


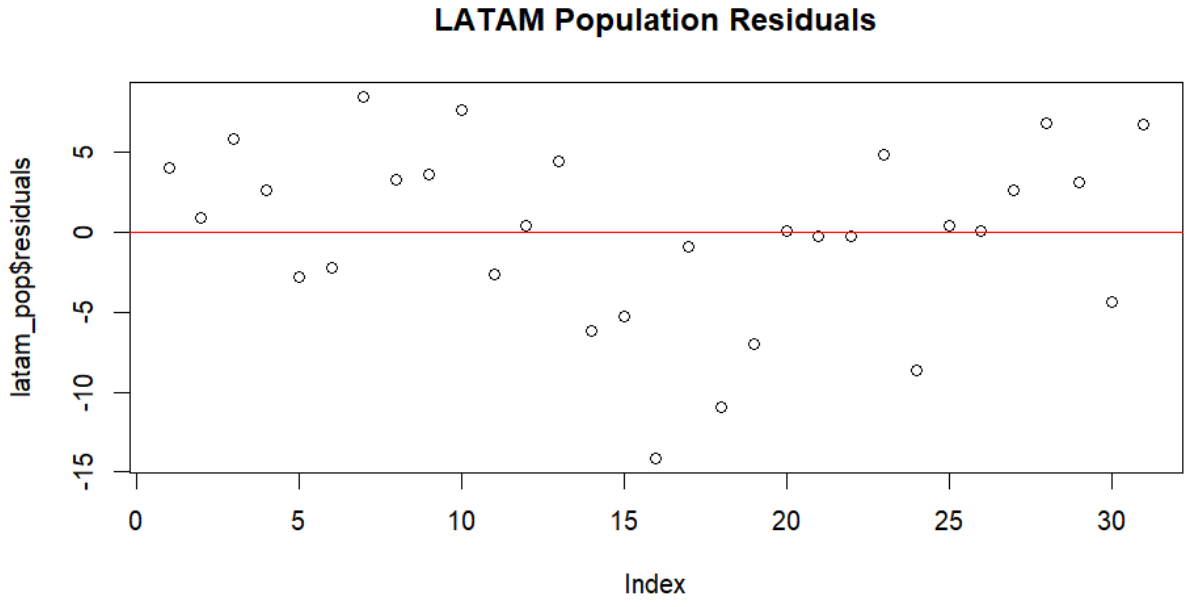
LATAM GDP Model Summary

```
Call:
lm(formula = EPI.new ~ log10(gdp), data = latam)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4179  -3.4131   0.7471   3.9835   8.7854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.937     15.085   0.261  0.79593
log10(gdp)    10.289      3.499   2.941  0.00637 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.249 on 29 degrees of freedom
Multiple R-squared:  0.2297,   Adjusted R-squared:  0.2031
F-statistic: 8.648 on 1 and 29 DF,  p-value: 0.006371
```

LATAM Population Model Summary

```
Call:
lm(formula = EPI.new ~ log10(population), data = latam)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1581  -2.7015   0.4061   3.7906   8.4601

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         61.857      6.952   8.898  8.7e-10 ***
log10(population)   -2.080      1.049  -1.984   0.0568 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.612 on 29 degrees of freedom
Multiple R-squared:  0.1195,   Adjusted R-squared:  0.08912
F-statistic: 3.935 on 1 and 29 DF,  p-value: 0.05683
```





3.3. Compare the models for both regions and very briefly describe which one is a better fit and why you think that is the case.

```
APAC Model Fit Comparison
GDP model R-squared: 0.1625709
Population model R-squared: 0.06021212

LATAM Model Fit Comparison
```

```
GDP model R-squared: 0.2297099
Population model R-squared: 0.1194797
```

In both the APAC and LATAM regions, the GDP model is the better fit. In APAC, the GDP model explains about 16.3% of the variance in EPI scores compared to 6% for population. In LATAM, the GDP model explains 23% of the variance versus 11.9% for population. Higher GDP levels often correlate with increased national capacity to invest in environmental infrastructure while population size does not necessarily correlate with environmental quality.

4.1. Train kNN Model

Optimal k = 3

Confusion Matrix:

	Actual		
Predicted	Asia-Pacific	Eastern Europe	Latin America & Caribbean
Asia-Pacific	6	0	1
Eastern Europe	0	7	2
Latin America & Caribbean	0	1	6

Accuracy: 0.826087 = 82.6%

4.2. Train another model with 3rd variable

Confusion Matrix:

	Actual		
Predicted	Asia-Pacific	Eastern Europe	Latin America & Caribbean
Asia-Pacific	4	0	2
Eastern Europe	1	6	0
Latin America & Caribbean	1	2	7

Accuracy: 0.7391304 = 73.9%

4.3. In 1-2 sentences explain which model performs better and why you think that is the case.

Model 1 (log population, log GDP, EPI score) achieved 82.6% accuracy, and Model 2 achieved 73.9% accuracy. I think that Model 1 performs better because the overall Environmental Performance Index captures a broader set of geographic indicators that provide stronger separation between these specific regions than the Ecosystem Vitality (ECO.new) variable alone.

Resources

- Module 4 Slides - https://tw.rpi.edu/sites/default/files/2026-01/Data_Analytics2026Spring_group1_module4_intro_analytics_methods.pdf
- Module 5 Slides - https://tw.rpi.edu/sites/default/files/2026-02/Data_Analytics2026Spring_Group2_module5_Regression_Classification_Methods_1.pdf