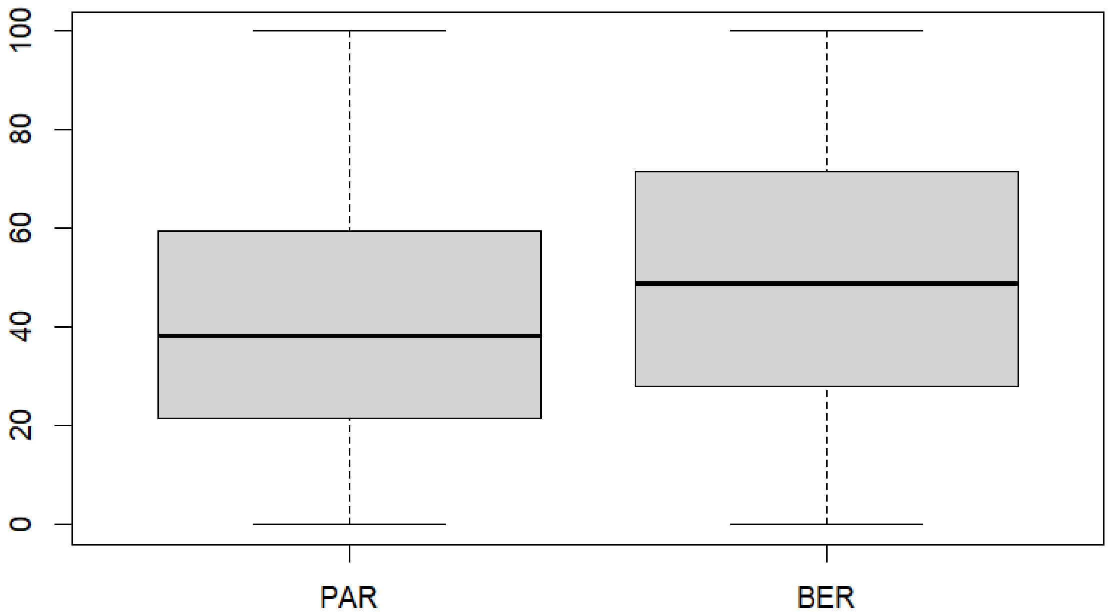
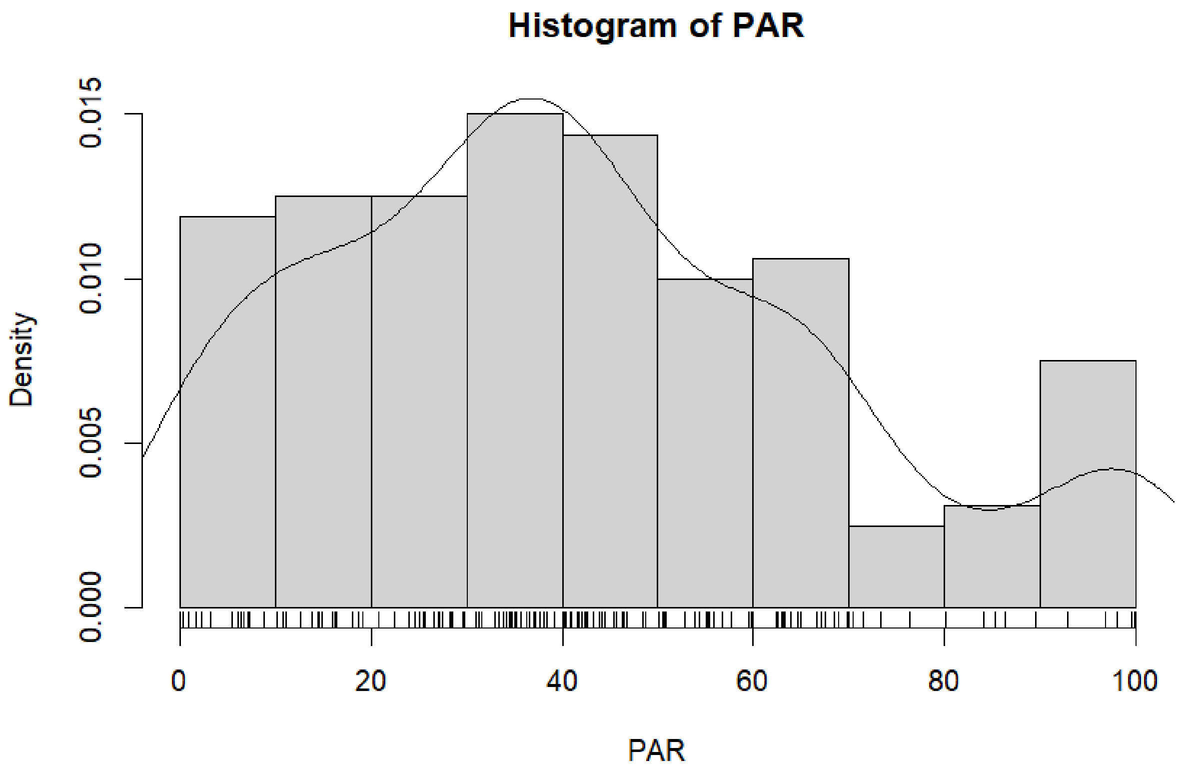
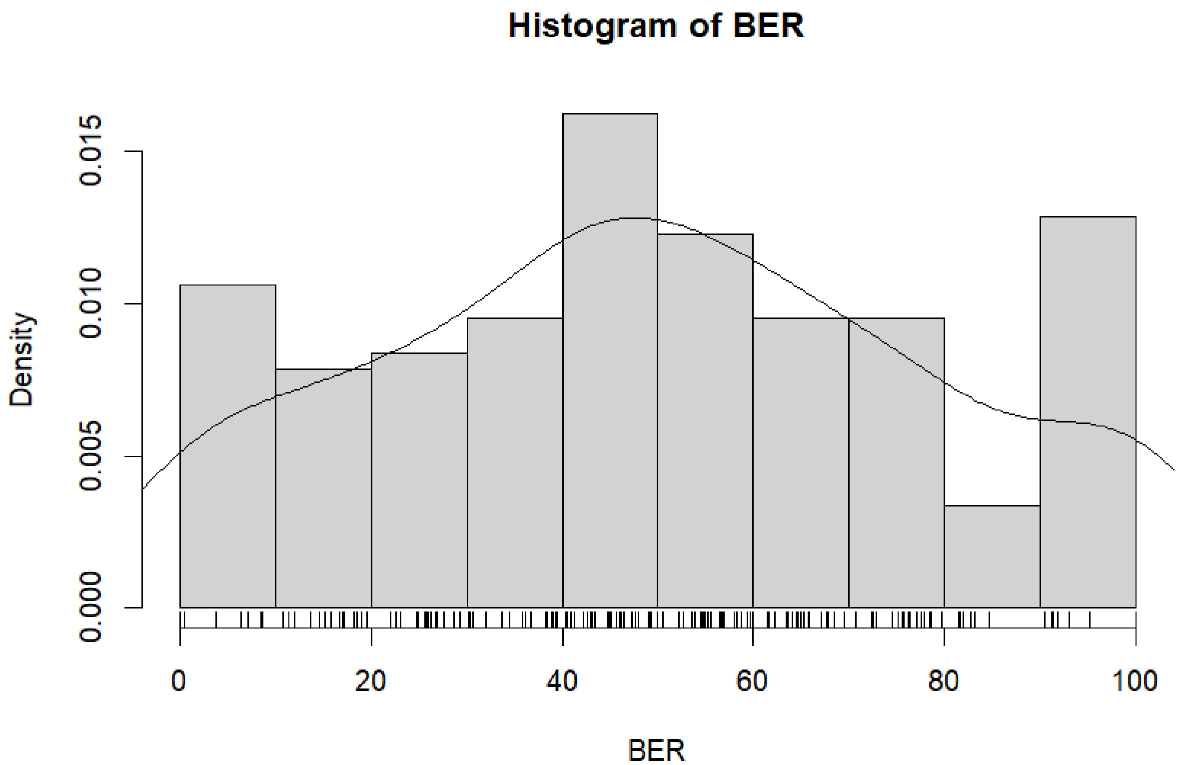
Ryan Lee
Data Analytics
October 14, 2025

Assignment 2

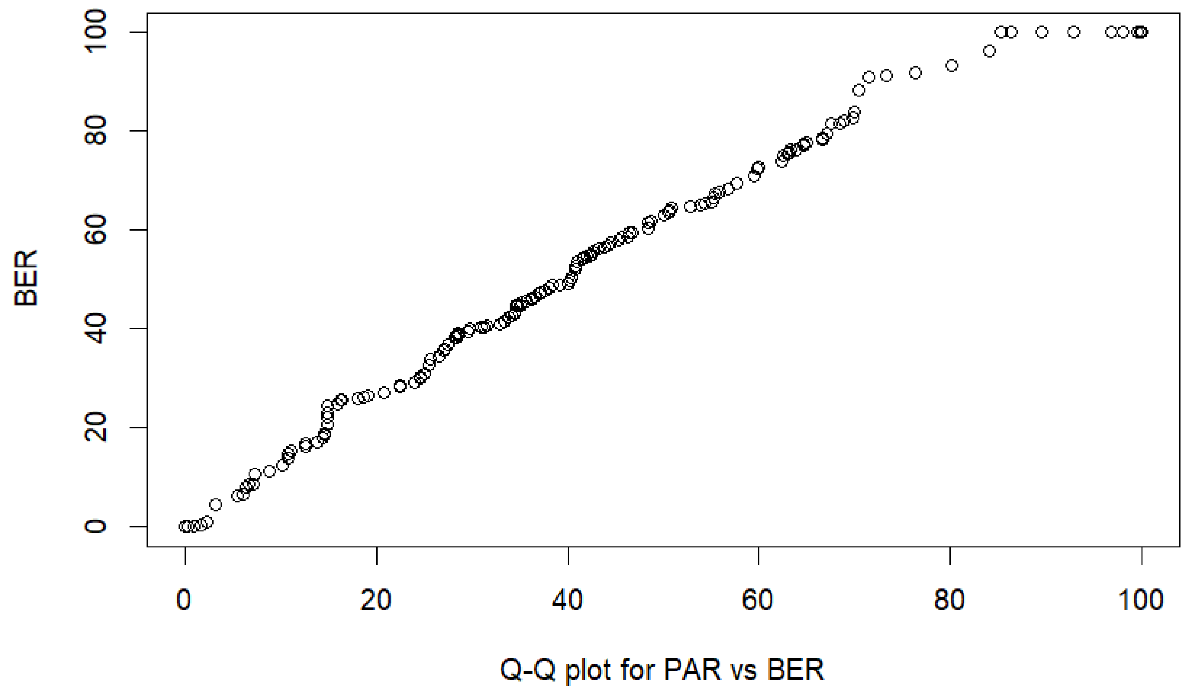
Variable Distribution



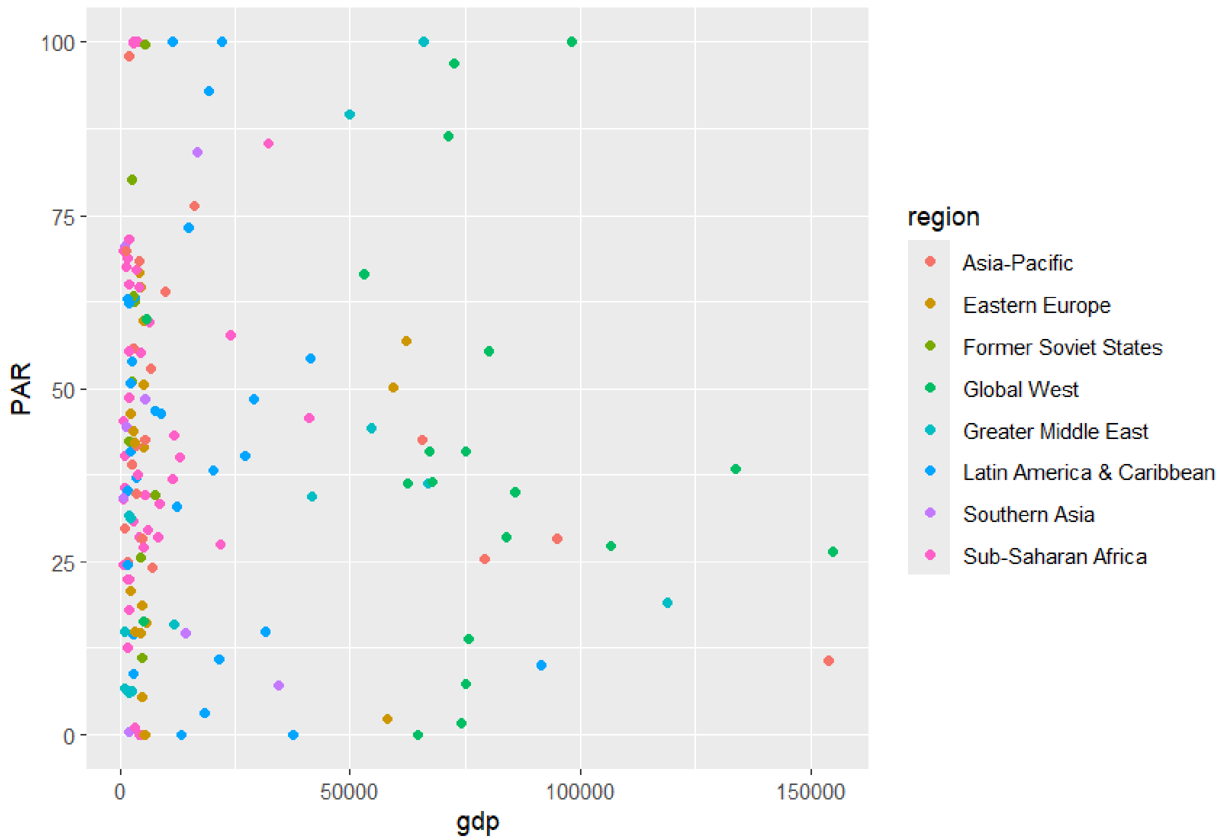

Histogram of PAR
Density
0.000
0.005
0.010
0.015
0
20
40
60
80
100
PAR



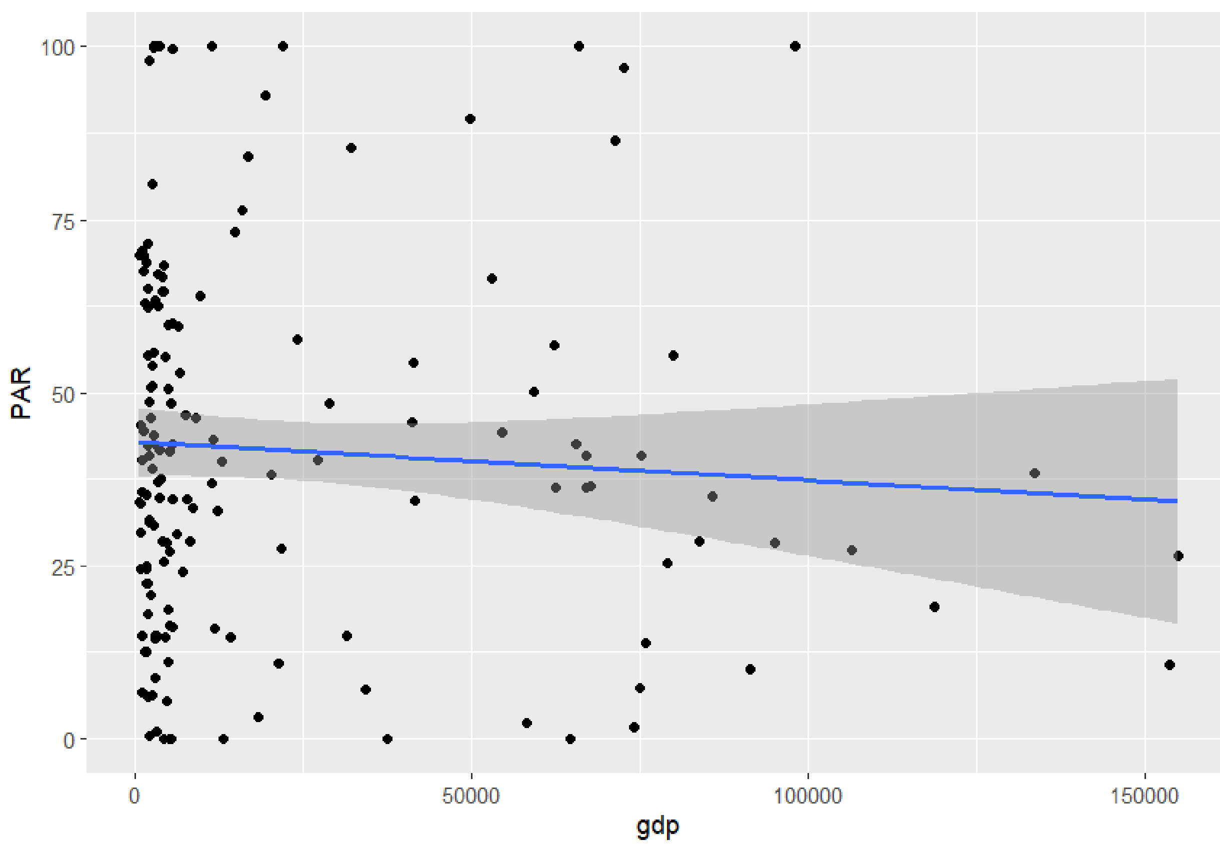
Histogram of BER
Density
0.000
0.005
0.010
0.015
0
20
40
60
80
100
BER



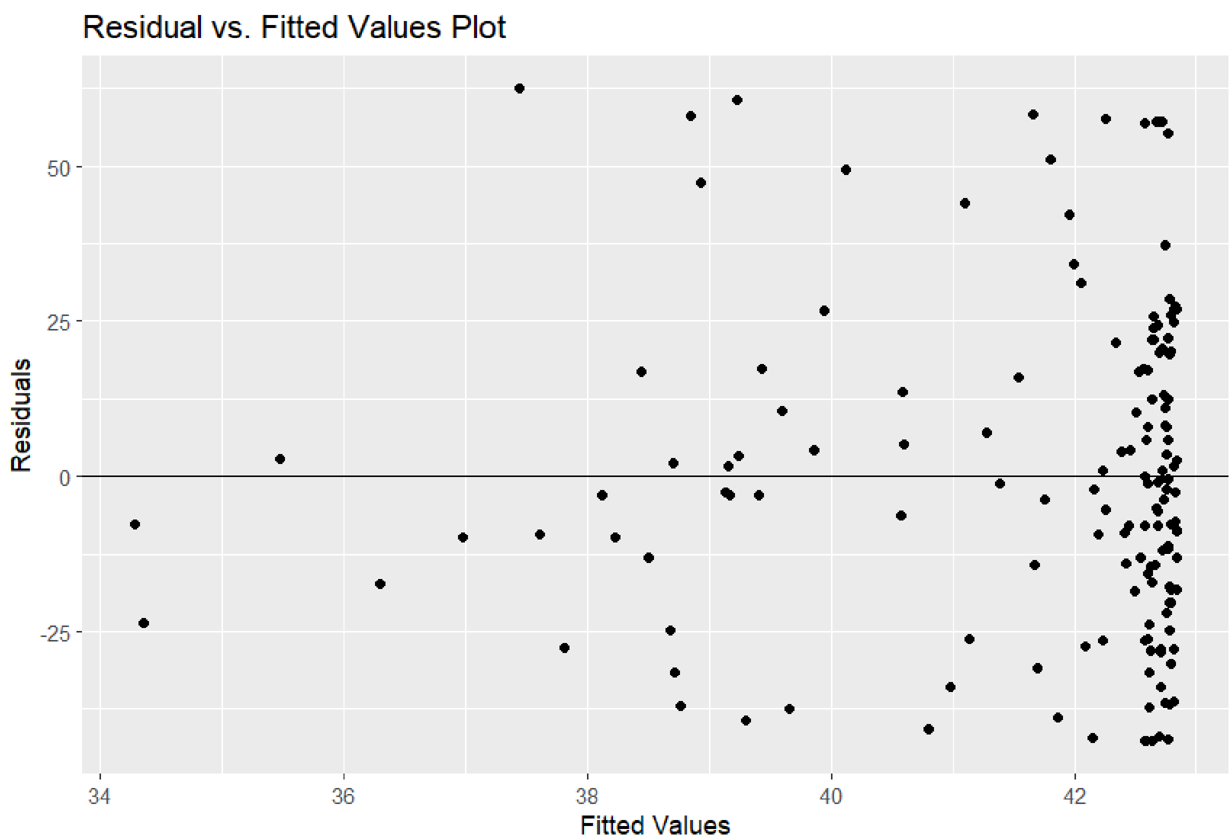
Linear Models



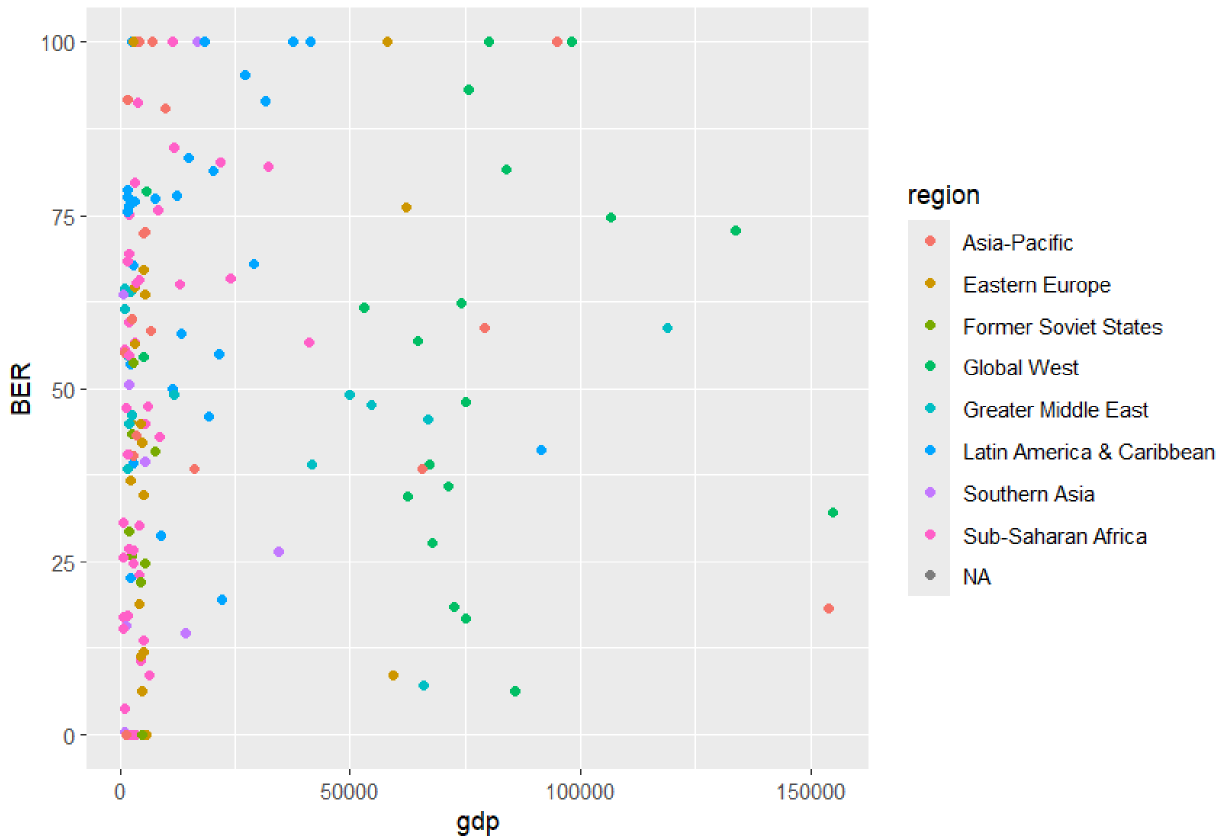

```
> summary(lin.mod0)

Call:
lm(formula = PAR ~ gdp, data = epi.data)

Residuals:
    Min      1Q  Median      3Q     Max
-42.643 -19.388  -2.628  17.284  62.563

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.288e+01  2.566e+00  16.711   <2e-16 ***
gdp         -5.548e-05  6.535e-05  -0.849    0.397
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.87 on 157 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.00457,  Adjusted R-squared:  -0.00177
F-statistic: 0.7208 on 1 and 157 DF,  p-value: 0.3972
```



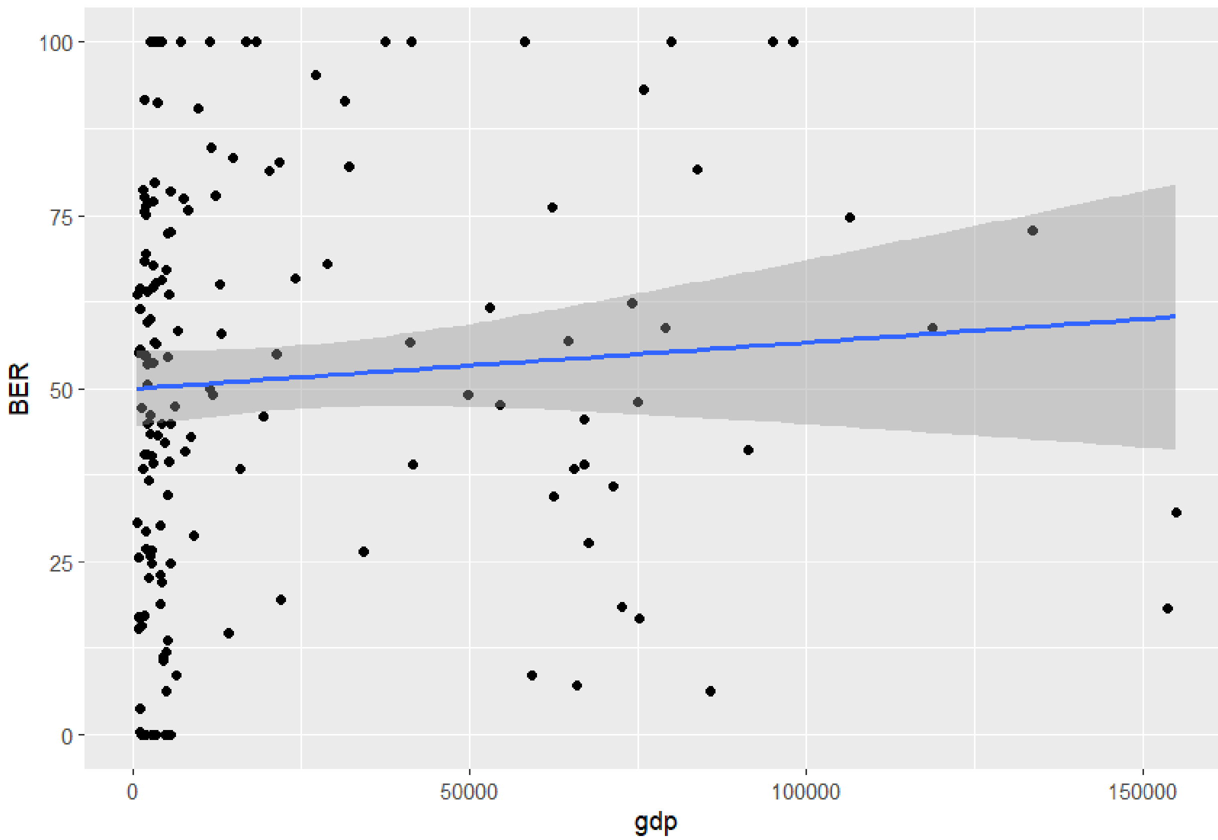
PAR
100
75
50
25
0
0
50000
100000
150000
gdp



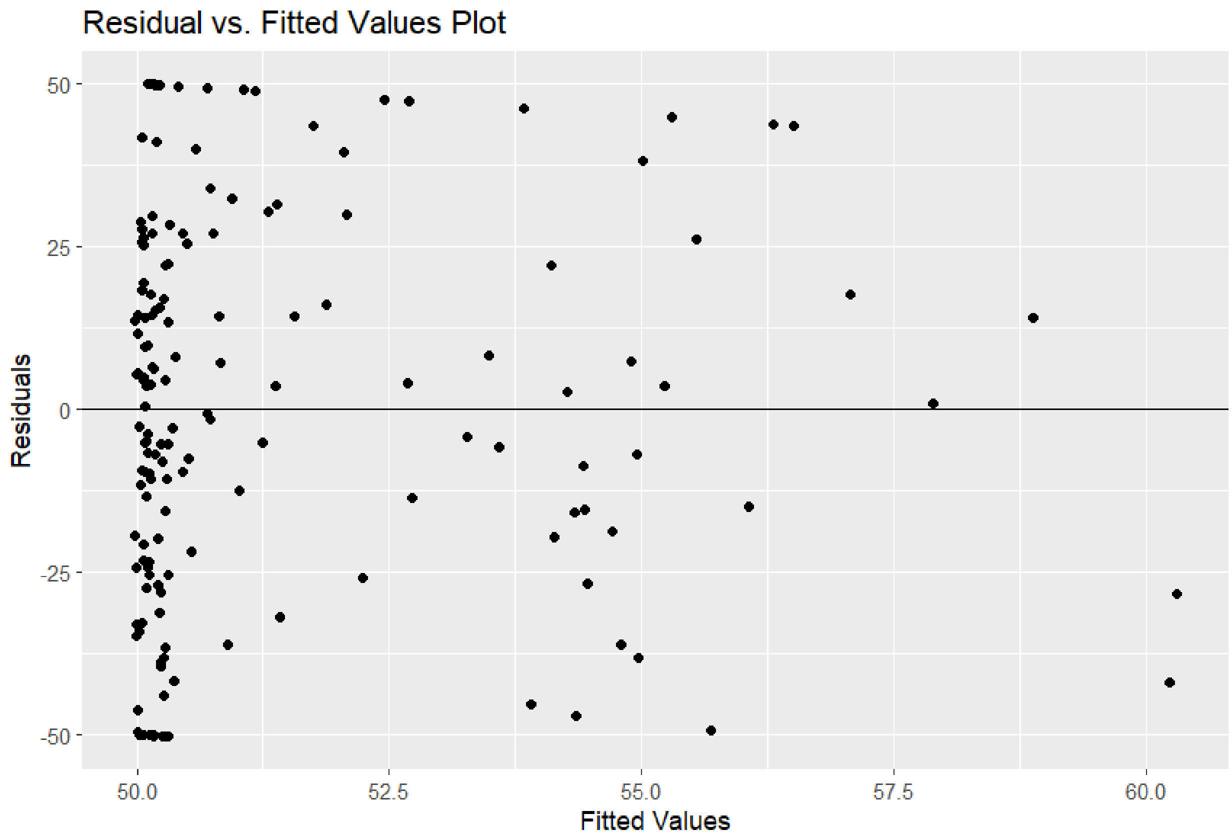
Residual vs. Fitted Values Plot
Residuals
50
25
0
-25
34
36
38
40
42
Fitted Values




```
> summary(lin.mod1)

Call:
lm(formula = BER ~ gdp, data = epi.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.312 -23.344   0.421  22.155  49.892 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.994e+01  2.786e+00  17.927   <2e-16 ***
gdp         6.697e-05  7.094e-05   0.944    0.347    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.17 on 157 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.005645, Adjusted R-squared:  -0.0006882 
F-statistic: 0.8913 on 1 and 157 DF,  p-value: 0.3466
```



BER
100
75
50
25
0
0
50000
100000
150000
gdp



Observing both models here, it's very clear that these models aren't very good because visually seems like these models are most likely not linear. However, between the two models, the BER linear model is much better because it produces a lower p-value and high t-statistic.

kNN Classification

```
# Display results
> cm1
Confusion Matrix and Statistics

                            Reference
Prediction                   Asia-Pacific Global West Latin America & Caribbean
Southern Asia
  Asia-Pacific                          2           0                         1
0
  Global West                           0           0                         0
0
  Latin America & Caribbean             0           0                         3
0
  Southern Asia                         0           0                         0
0
  Sub-Saharan Africa                    0           0                         0
0
                            Reference
Prediction                   Sub-Saharan Africa
  Asia-Pacific                                0
  Global West                                 0
  Latin America & Caribbean                   0
  Southern Asia                               0
  Sub-Saharan Africa                          0

Overall Statistics

               Accuracy : 0.8333
                 95% CI : (0.3588, 0.9958)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.3512

                  Kappa : 0.6667

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Asia-Pacific Class: Global West Class: Latin
America & Caribbean
Sensitivity                       1.0000                 NA
0.7500
Specificity                       0.7500                  1
1.0000
Pos Pred Value                    0.6667                 NA
1.0000
Neg Pred Value                    1.0000                 NA
0.6667
```

```
Prevalence                        0.3333                  0
0.6667
Detection Rate                    0.3333                  0
0.5000
Detection Prevalence              0.5000                  0
0.5000
Balanced Accuracy                 0.8750                 NA
0.8750
                     Class: Southern Asia Class: Sub-Saharan Africa
Sensitivity                            NA                        NA
Specificity                             1                         1
Pos Pred Value                         NA                        NA
Neg Pred Value                         NA                        NA
Prevalence                              0                         0
Detection Rate                          0                         0
Detection Prevalence                    0                         0
Balanced Accuracy                      NA                        NA
> cm2
Confusion Matrix and Statistics

                           Reference
Prediction                   Asia-Pacific Global West Latin America & Caribbean
Southern Asia
  Asia-Pacific                          0           0                         1
0
  Global West                           0           0                         0
0
  Latin America & Caribbean             2           0                         3
0
  Southern Asia                         0           0                         0
0
  Sub-Saharan Africa                    0           0                         0
0
                           Reference
Prediction                   Sub-Saharan Africa
  Asia-Pacific                                0
  Global West                                 0
  Latin America & Caribbean                   0
  Southern Asia                               0
  Sub-Saharan Africa                          0

Overall Statistics

               Accuracy : 0.5
                 95% CI : (0.1181, 0.8819)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.8999
```

```
                  Kappa : -0.2857

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Asia-Pacific Class: Global West Class: Latin
America & Caribbean
Sensitivity                       0.0000                 NA
0.7500
Specificity                       0.7500                  1
0.0000
Pos Pred Value                    0.0000                 NA
0.6000
Neg Pred Value                    0.6000                 NA
0.0000
Prevalence                        0.3333                  0
0.6667
Detection Rate                    0.0000                  0
0.5000
Detection Prevalence              0.1667                  0
0.8333
Balanced Accuracy                 0.3750                 NA
0.3750
                     Class: Southern Asia Class: Sub-Saharan Africa
Sensitivity                            NA                        NA
Specificity                             1                         1
Pos Pred Value                         NA                        NA
Neg Pred Value                         NA                        NA
Prevalence                              0                         0
Detection Rate                          0                         0
Detection Prevalence                    0                         0
Balanced Accuracy                      NA                        NA
```

```
>
```

```
> cat("Model 1 Accuracy:", round(acc1, 7), "\n")
Model 1 Accuracy: 0.8333333
> cat("Model 2 Accuracy:", round(acc2, 7), "\n")
Model 2 Accuracy: 0.5
```

Since Model 1's accuracy is greater than Model 2, Model 1 seems to be the better model. Model 1's features were ECO, BDH, and MKP new columns. Both kNN models experienced a train test split of 70% and 30% with 10-fold cross validation across varying k-values. After plotting their contingency matrices and picking out their accuracies, Model 1 displayed better results.