Component 1: Multiple POLS, Statistical Capacity Indicies

Sevastian Sanchez

2025-09-17

1 All Countries, Preliminary Analysis (SPI x SDGs)

SPI = Statistical Performance Index (0-100, continuous) SDG = Sustainable Development Goals (0-100, continuous) DI = EIU Democracy Index/Score (0-10, continuous) log_gdppc = Log(GDP Per Capita)

1.0.1 LOAD FINAL MERGED CSV

```
# set working directory
setwd("~/Documents/GitHub/QMSS_Thesis_Sanchez")

# load libraries
source("packages.R")

# load final merged df
merged <- read_csv("data/Main CSV Outputs/comp1_data.csv")  # updated

# refer to 'wrangled/adjust_output.Rmd' to make necessary
# adjustments
```

2 COMPONENT 1: COMPARING SPI & SCI X VARIABLES

2.1 Preliminary Analysis: Correlation & Naive OLS Models [FINALIZED]

2.1.1 Correlation Analysis: SPI, SCI, DI & SDG Composite Scores

H0: Null, there is no relationship
H1: there is a relationship between overall SPI and SDG composite scores

```
# Correlation coefficients & R-squared values for SDG and
# SPI/SCI/DI

# SDG ~ SPI
correlation_sdg_spi <- cor(merged$sdg_overall, merged$spi_comp,
    use = "complete.obs")
# R-squared value
R2_sdg_spi <- correlation_sdg_spi^2
```

```r
# SDG ~ SCI
correlation_sdg_sci <- cor(merged$sdg_overall, merged$sci_overall,
    use = "complete.obs")
# R-squared value
R2_sdg_sci <- correlation_sdg_sci^2

# SDG ~ DI
correlation_sdg_di <- cor(merged$sdg_overall, merged$di_score,
    use = "complete.obs")
# R-squared value
R2_sdg_di <- correlation_sdg_di^2

# SPI ~ DI
correlation_spi_di <- cor(merged$spi_comp, merged$di_score, use = "complete.obs")
# R-squared value
R2_spi_di <- correlation_spi_di^2

# SCI ~ DI
correlation_sci_di <- cor(merged$sci_overall, merged$di_score,
    use = "complete.obs")
# R-squared value
R2_sci_di <- correlation_sci_di^2

# pasting correlation results
paste("Correlation coefficient:", correlation_sdg_spi, "(SDG ~ SPI),",
    correlation_sdg_sci, "(SDG ~ SCI),", correlation_sdg_di,
    "(SDG ~ DI),", correlation_spi_di, "(SPI ~ DI),", correlation_sci_di,
    "(SCI ~ DI)")
```

```
## [1] "Correlation coefficient: 0.784880374521699 (SDG ~ SPI), 0.6465025625156 (SDG ~ SCI), 0.67264392
```

```r
# pasting R-sq results
paste("R-squared value:", R2_sdg_spi, "(SDG ~ SPI),", R2_sdg_sci,
    "(SDG ~ SCI),", R2_sdg_di, "(SDG ~ DI),", R2_spi_di, "(SPI ~ DI),",
    R2_sci_di, "(SCI ~ DI)")
```

```
## [1] "R-squared value: 0.616037202309323 (SDG ~ SPI), 0.417965563339237 (SDG ~ SCI), 0.45244985581842
```

SDG ~ SPI: Correlation coefficient: 0.784880, R-squared value: 0.616037 SDG ~ SCI: Correlation coefficient: 0.646503, R-squared value: 0.417966 SDG ~ DI: Correlation coefficient: 0.672644, R-squared value: 0.452450 SPI ~ DI: Correlation coefficient: 0.676171, R-squared value: 0.4572067 SCI ~ DI: Correlation coefficient: 0.477767, R-squared value: 0.2282613

The results demonstrates that SDG performance (SDG) composite scores are most strongly associated with the Statistical Performance Index (SPI), which shows a high correlation coefficient (0.78) and explains about 62% of the variance in SDG scores. Both the Statistical Capacity Index (SCI) and the Democracy Index (DI) also exhibit moderate positive correlations with SDG scores (0.65 and 0.67, respectively), accounting for 42% and 45% of the variance. This could indicate that countries with higher statistical performance and stronger democratic institutions tend to achieve better SDG outcomes, with SPI emerging as the most influential predictor among the indices examined.

Both SPI and SCI are also positively correlated with the Democracy Index, with SPI (0.68) showing a stronger association than SCI (0.48). These findings highlight the interconnectedness of statistical capacity and governance quality, suggesting that improvements in national statistical systems and democratic governance are mutually reinforcing factors in advancing sustainable development.

The correlation analysis provides a useful overview of the relationships between these indices, but further analysis is needed to understand the causal mechanisms and the impact of these indices on SDG outcomes.

2.1.2 NAIVE OLS Models (Component 1): Comparing SPI & SCI Variables on SDG Performance

Finding estimated impact of variables on SDG status prior to adding controls or robust SEs

```r
# 1. OLS for SPI and SDG - Overall
ols_spi_naive <- lm(sdg_overall ~ spi_comp, data = merged)
summary(ols_spi_naive)
```

```
##
## Call:
## lm(formula = sdg_overall ~ spi_comp, data = merged)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.3175  -4.4186   0.5969   4.4301  20.1684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.94626    0.72064   48.49   <2e-16 ***
## spi_comp     0.47806    0.01048   45.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.338 on 1298 degrees of freedom
##   (2040 observations deleted due to missingness)
## Multiple R-squared:  0.616,  Adjusted R-squared:  0.6157
## F-statistic:  2083 on 1 and 1298 DF,  p-value: < 2.2e-16
```

```r
# 2. OLS for SCI and SDG - Overall
ols_sci_naive <- lm(sdg_overall ~ sci_overall, data = merged)
summary(ols_sci_naive)
```

```
##
## Call:
## lm(formula = sdg_overall ~ sci_overall, data = merged)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0240  -4.9307   0.2307   4.8361  18.8180
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.88189    0.71150   47.62   <2e-16 ***
## sci_overall  0.39209    0.01021   38.40   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.166 on 2053 degrees of freedom
##   (1285 observations deleted due to missingness)
## Multiple R-squared:  0.418,  Adjusted R-squared:  0.4177
## F-statistic:  1474 on 1 and 2053 DF,  p-value: < 2.2e-16
```

```r
# 3. Multiple Regression with both SPI and SCI
ols_multiple_naive <- lm(sdg_overall ~ spi_comp + sci_overall, data = merged)
summary(ols_multiple_naive)
```

```
##
## Call:
## lm(formula = sdg_overall ~ spi_comp + sci_overall, data = merged)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5483  -5.4484   0.4037   4.7941  17.9050
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.86438    1.27744  28.075  < 2e-16 ***
## spi_comp     0.28779    0.03369   8.542  < 2e-16 ***
## sci_overall  0.15311    0.03232   4.738  2.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.671 on 593 degrees of freedom
##   (2744 observations deleted due to missingness)
## Multiple R-squared:  0.4651, Adjusted R-squared:  0.4633
## F-statistic: 257.8 on 2 and 593 DF,  p-value: < 2.2e-16
```

```r
# stargazer summary table of models: ols_spi_naive, ols_sci_naive, ols_multiple_naive.
stargazer(
  ols_spi_naive,
  ols_sci_naive,
  ols_multiple_naive,
  title = "Naive OLS Models: SPI & SCI x SDG",
  dep.var.caption = "Dependent Variable:",
  dep.var.labels = "SDG Composite Score",
  covariate.labels = c("SPI Composite Score", "SCI Composite Score", "Intercept"),
  column.labels = c("SPI Only", "SCI Only", "SPI + SCI"),  # <-- Model labels at the top
  omit.stat = c("f", "ser"),
  digits = 4,
  type = "text"
  #out = "component_1/figures/naive_mods_sdgs_spi_sci_tab.html"  # Save as HTML file
)
```

```
##
## Naive OLS Models: SPI & SCI x SDG
## ====================================================
##                           Dependent Variable:
##                     --------------------------------
```

```
##                            SDG Composite Score
##                         SPI Only   SCI Only  SPI + SCI
##                           (1)        (2)        (3)
## ----------------------------------------------------
## SPI Composite Score 0.4781***             0.2878***
##                      (0.0105)              (0.0337)
##
## SCI Composite Score            0.3921***  0.1531***
##                                 (0.0102)   (0.0323)
##
## Intercept           34.9463*** 33.8819*** 35.8644***
##                      (0.7206)   (0.7115)   (1.2774)
##
## ----------------------------------------------------
## Observations          1,300      2,055       596
## R2                   0.6160     0.4180     0.4651
## Adjusted R2          0.6157     0.4177     0.4633
## ====================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

ols_spi_naive: 0.47806 (p-value $<$ 0.01)
ols_sci_naive: 0.39081 (p-value $<$ 0.01)
ols_multiple_naive: spi: 0.28779 (p-value $<$ 0.01); sci: 0.15311 (p-value $<$ 0.01)

The impact of SCI on SDG and SPI on SDG are statistically significant, in all models. SPI appears to have a greater impact on SDGs compared to that of SCI, regardless of the model. All of this is without controls or accounting multiple time periods of the same subject (i.e., countries).

2.1.3 NAIVE OLS Models (Component 2): DI variable on SPI & SDG Performance

Finding estimated impact of DI on SPI and SDG status prior to adding controls or robust SEs

```r
# 1. OLS for SPI and DI - Overall
ols_spi_di_naive <- lm(spi_comp ~ di_score, data = merged)
summary(ols_spi_di_naive)
```

```
##
## Call:
## lm(formula = spi_comp ~ di_score, data = merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.200  -7.910   0.980   8.098  37.873
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.9418     0.9185   43.48   <2e-16 ***
## di_score      5.0122     0.1548   32.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.11 on 1245 degrees of freedom
##   (2093 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.4572, Adjusted R-squared:  0.4568
## F-statistic:  1049 on 1 and 1245 DF,  p-value: < 2.2e-16
```

```r
# 2. OLS for SDG and DI - Overall
ols_sdg_di_naive <- lm(sdg_overall ~ di_score, data = merged)
summary(ols_sdg_di_naive)
```

```
##
## Call:
## lm(formula = sdg_overall ~ di_score, data = merged)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0231  -5.5688   0.2781   5.8145  25.0253
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.68921    0.41777  114.15   <2e-16 ***
## di_score     3.22301    0.07032   45.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.816 on 2542 degrees of freedom
##   (796 observations deleted due to missingness)
## Multiple R-squared:  0.4524, Adjusted R-squared:  0.4522
## F-statistic:  2100 on 1 and 2542 DF,  p-value: < 2.2e-16
```

```r
# stargazer summary table of models: ols_spi_di_naive, ols_sdg_di_naive.
stargazer(
  ols_spi_di_naive,
  ols_sdg_di_naive,
  title = "Naive OLS Models: DI & SPI/SDG",
  dep.var.caption = "Dependent Variable:",
  dep.var.labels = c("SDG Composite Score", "SPI Composite Score"),
  #covariate.labels = c("di_score", "Intercept"),
  #column.labels = c("SPI Only", "SCI Only", "SPI + SCI"),  # <-- Model labels at the top
  omit.stat = c("f", "ser"),
  digits = 4,
  type = "text"
  #out = "component_1/figures/naive_mods_sdgs&spi_di_tab.html"  # Save as HTML file
)
```

```
##
## Naive OLS Models: DI & SPI/SDG
## ====================================================
##                         Dependent Variable:
##              ---------------------------------------
##              SDG Composite Score SPI Composite Score
##                      (1)                 (2)
## ----------------------------------------------------
## di_score          5.0122***           3.2230***
##                   (0.1548)            (0.0703)
##
```

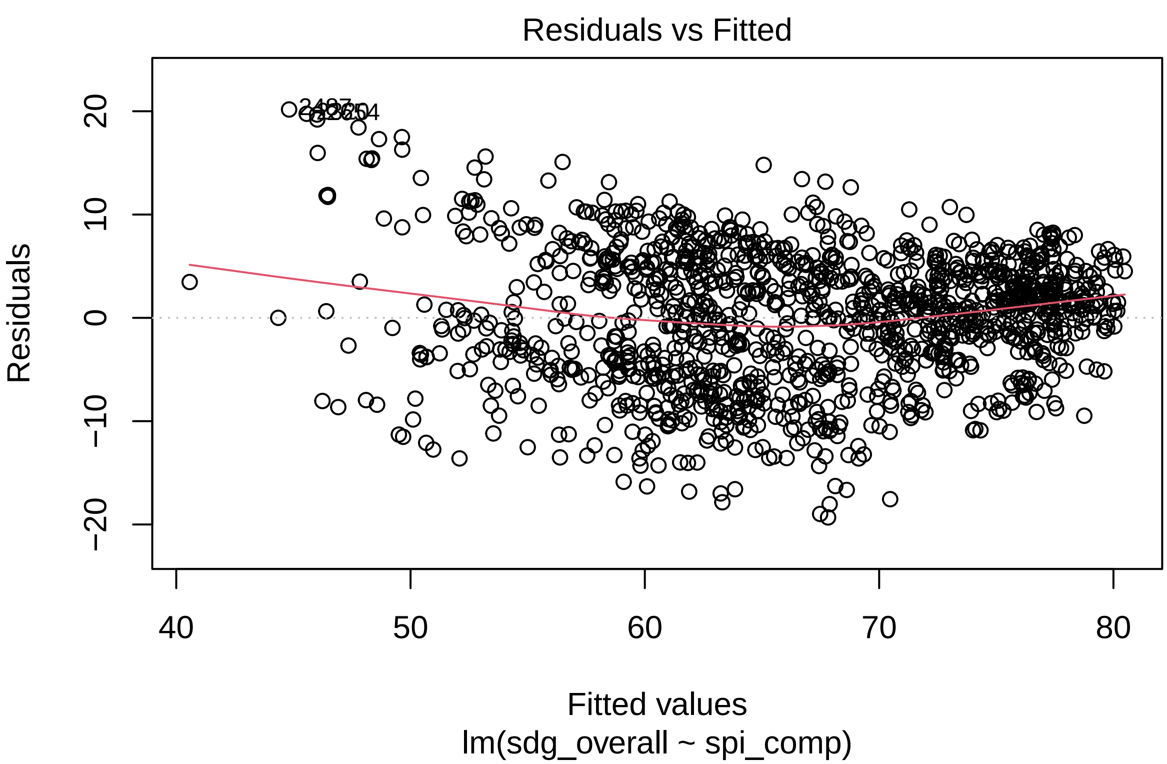
```
## Constant           39.9418***          47.6892***
##                     (0.9185)            (0.4178)
##
## -----------------------------------------------------
## Observations          1,247               2,544
## R2                   0.4572              0.4524
## Adjusted R2          0.4568              0.4522
## =====================================================
## Note:                       *p<0.1; **p<0.05; ***p<0.01
```

results ols_spi_di_naive: 5.0122 (p-value < 0.001) ols_sdg_di_naive: 3.22301 (p-value < 0.001)

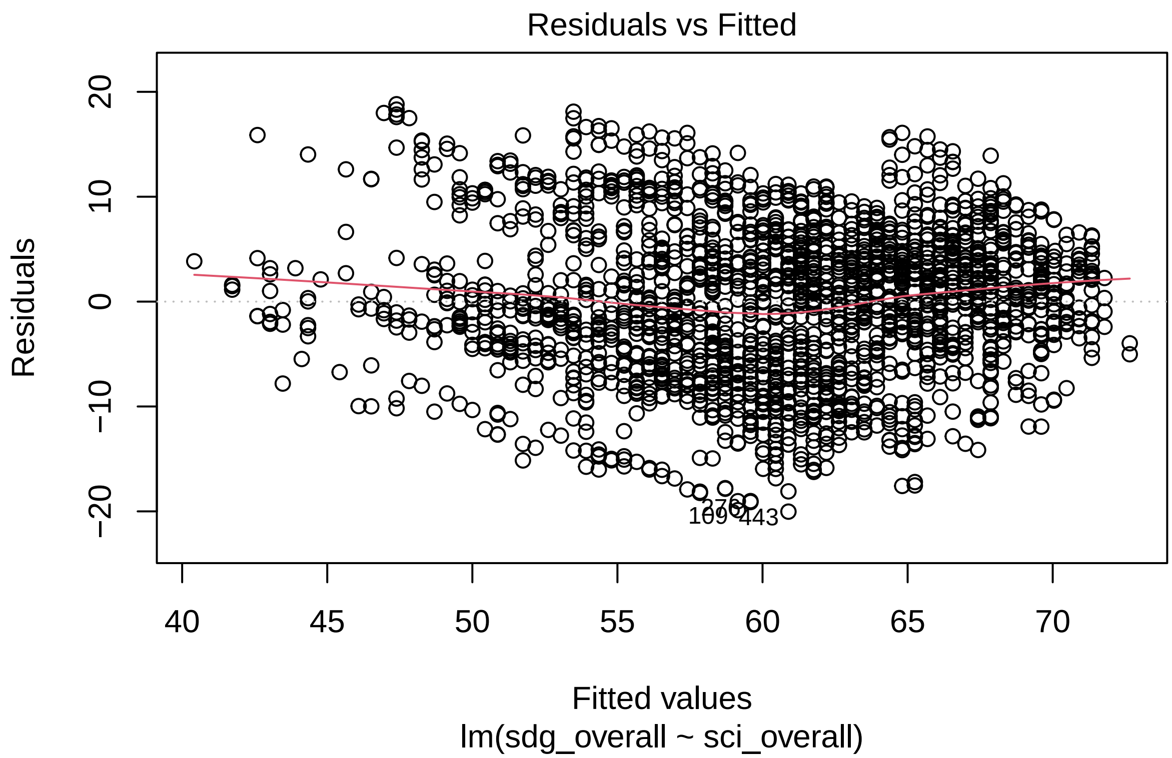
The naive OLS models indicate that the Democracy Index (DI) has a strong positive relationship with both the Statistical Performance Index (SPI) and the Sustainable Development Goals (SDG) composite scores. The coefficients suggest that a one-unit increase in DI is associated with a 5.0122 increase in SPI and a 3.22301 increase in SDG scores, both statistically significant at $p < 0.001$. This suggests that countries with higher democracy scores tend to have better statistical performance and SDG outcomes.

2.1.4 Checking for Heteroskedasticity: residual plots [no need to report]

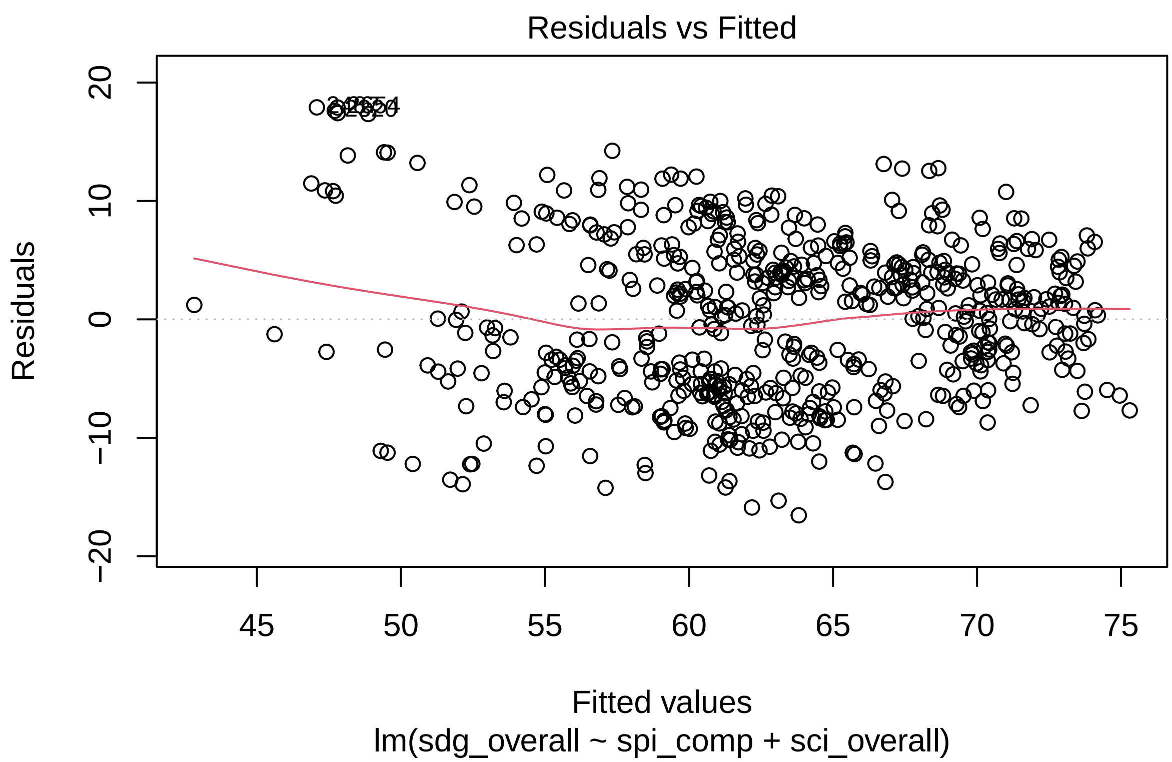
```
# residual plots
plot(ols_spi_naive, which = 1)  # SDG ~ SPI model
```



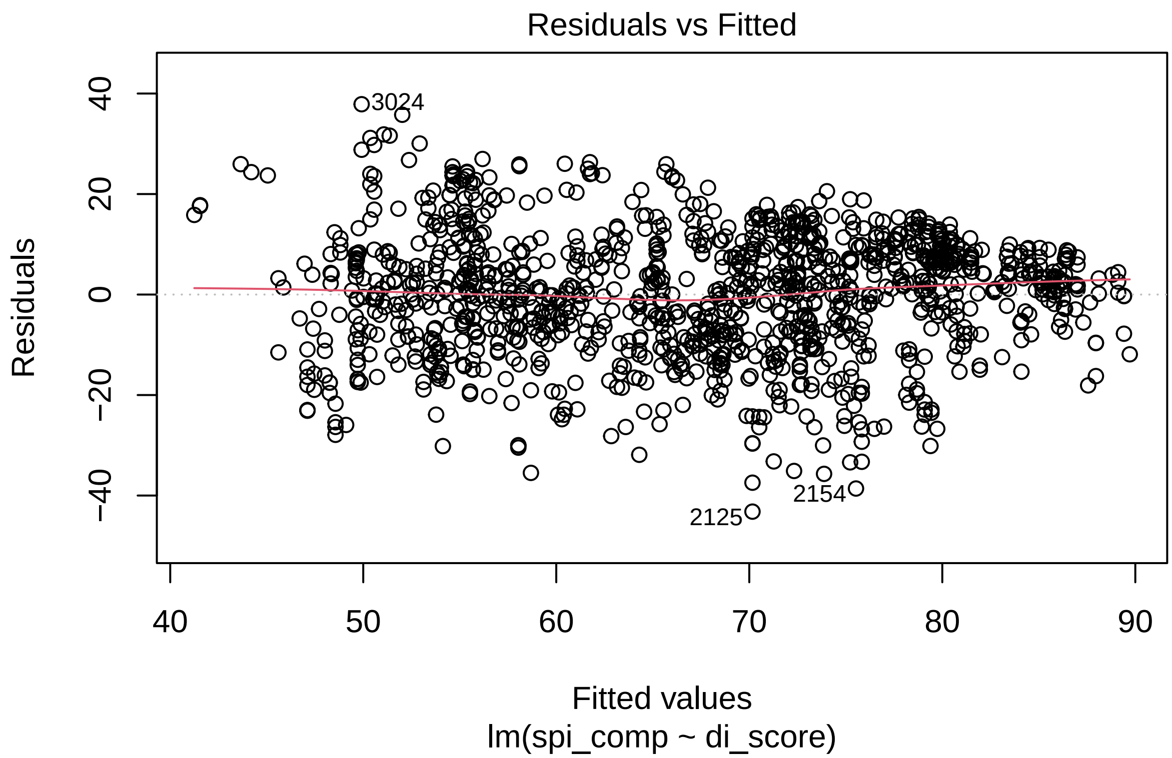

```
plot(ols_sci_naive, which = 1)  # SDG ~ SCI model
```



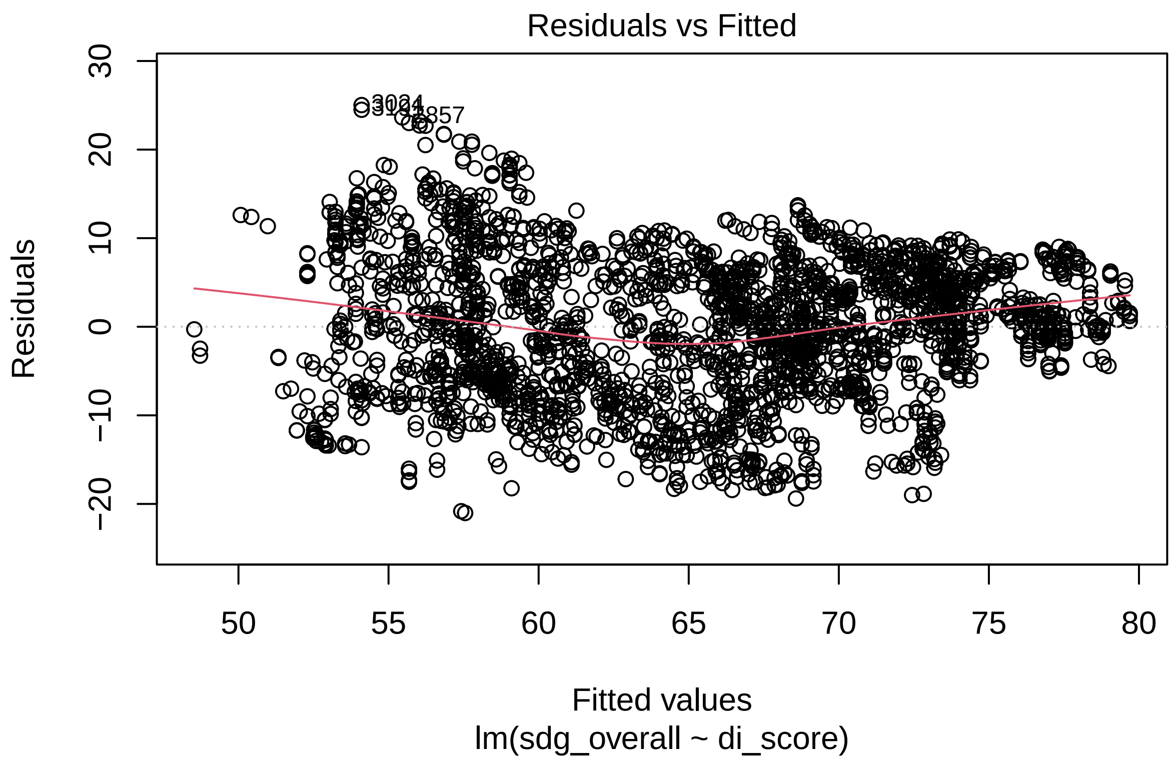

```r
plot(ols_multiple_naive, which = 1)   # SDG ~ SPI + SCI model
```




```r
plot(ols_spi_di_naive, which = 1)   # SPI ~ DI model
```




```r
plot(ols_sdg_di_naive, which = 1)  # SDG ~ DI model
```




```r
# U-shaped residuals detected, suggests non-linearity of
# x-variable terms. Additional tests reconfirm
# non-linearity (See Breusch-Pagan Test below).
```

2.2 TEST 1: Pooled OLS & Clustered Robust Standard Errors – COMPARING MEASURES [DONE]

Methodology: Pooled OLS Models & Clustered Robust (Huber-White) Standard Errors All variables of statistical capacity (SPI & SCI) will be compared on a base pooled OLS regression model structure. Pooled OLS recognizes the panel-like structure allowing to index by specific country and year (country-year). Regular OLS, assumes independence of observations which is not suitable given the repeated waves of country-year over the course of multiple consecutive years. Furthermore, it is customary to apply clustered-group robust standard errors to account for heteroskedasticity and within-unit correlation of countries over many time points.

H0: Null, there is no relationship between SPI and SDG composite scores H1: There is a statistically significant relationship between SPI and SDG composite scores

```r
# 1. OLS for SPI and SDG - Overall
ols_spi <- plm(formula = sdg_overall ~ spi_comp + di_score +
    log_gdppc + factor(year), model = "pooling", index = c("country_code",
    "year"), data = merged)
summary(ols_spi, vcov = vcovHC(ols_spi, cluster = "group", type = "HC1"))
```

```
## Pooling Model
##
## Note: Coefficient variance-covariance matrix supplied: vcovHC(ols_spi, cluster = "group", type = "HC
##
## Call:
## plm(formula = sdg_overall ~ spi_comp + di_score + log_gdppc +
##     factor(year), data = merged, model = "pooling", index = c("country_code",
##     "year"))
##
## Unbalanced Panel: n = 156, T = 2-8, N = 1236
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -12.152624  -3.128759  -0.026545   3.098090  13.880684
##
## Coefficients:
##                  Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)      18.96922    2.51816  7.5330 9.596e-14 ***
## spi_comp          0.28735    0.03847  7.4695 1.524e-13 ***
## di_score          0.20585    0.28491  0.7225  0.470115
## log_gdppc         3.31206    0.47490  6.9743 5.018e-12 ***
## factor(year)2017 -0.41907    0.14216 -2.9478  0.003261 **
## factor(year)2018 -0.95926    0.24016 -3.9942 6.877e-05 ***
## factor(year)2019 -0.63082    0.26247 -2.4034  0.016393 *
## factor(year)2020 -0.65548    0.36042 -1.8186  0.069210 .
## factor(year)2021 -2.19599    0.51780 -4.2410 2.393e-05 ***
## factor(year)2022 -2.10963    0.49295 -4.2796 2.018e-05 ***
## factor(year)2023 -2.49491    0.52514 -4.7509 2.264e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Total Sum of Squares:    125210
## Residual Sum of Squares: 27858
## R-Squared:      0.77752
## Adj. R-Squared: 0.7757
## F-statistic: 78.0072 on 10 and 155 DF, p-value: < 2.22e-16
```

```r
# 2. OLS for SCI and SDG - Overall
ols_sci <- plm(formula = sdg_overall ~ sci_overall + di_score +
    log_gdppc + factor(year), model = "pooling", index = c("country_code",
    "year"), data = merged)
summary(ols_sci, vcov = vcovHC(ols_sci, cluster = "group", type = "HC1"))
```

```
## Pooling Model
##
## Note: Coefficient variance-covariance matrix supplied: vcovHC(ols_sci, cluster = "group", type = "HC
##
## Call:
## plm(formula = sdg_overall ~ sci_overall + di_score + log_gdppc +
##     factor(year), data = merged, model = "pooling", index = c("country_code",
##     "year"))
##
## Unbalanced Panel: n = 117, T = 7-13, N = 1515
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -13.59286  -2.94502  -0.19244   3.10513  13.47045
##
## Coefficients:
##                   Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)       1.881498   2.950772  0.6376 0.5238125
## sci_overall       0.237633   0.030195  7.8699 6.751e-15 ***
## di_score         -0.082517   0.267958 -0.3079 0.7581657
## log_gdppc         5.508454   0.474053 11.6199 < 2.2e-16 ***
## factor(year)2008 -1.250095   0.235714 -5.3034 1.306e-07 ***
## factor(year)2010 -1.018682   0.297008 -3.4298 0.0006204 ***
## factor(year)2011 -1.505211   0.360971 -4.1699 3.222e-05 ***
## factor(year)2012 -1.035280   0.390527 -2.6510 0.0081104 **
## factor(year)2013 -0.955591   0.417556 -2.2885 0.0222452 *
## factor(year)2014 -0.551208   0.430294 -1.2810 0.2003903
## factor(year)2015  0.264456   0.411581  0.6425 0.5206229
## factor(year)2016  0.742872   0.418737  1.7741 0.0762535 .
## factor(year)2017  0.955586   0.452615  2.1113 0.0349154 *
## factor(year)2018  1.485885   0.452008  3.2873 0.0010350 **
## factor(year)2019  1.888027   0.462382  4.0833 4.675e-05 ***
## factor(year)2020  2.982542   0.483598  6.1674 8.911e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    129110
## Residual Sum of Squares: 33042
## R-Squared:      0.74409
## Adj. R-Squared: 0.74152
## F-statistic: 48.9042 on 15 and 116 DF, p-value: < 2.22e-16
```

```r
# 3. Multiple Regression with both SPI and SCI
ols_multiple <- plm(formula = sdg_overall ~ spi_comp + sci_overall +
    di_score + log_gdppc + factor(year), model = "pooling", index = c("country_code",
    "year"), data = merged)
summary(ols_multiple, vcov = vcovHC(ols_multiple, cluster = "group",
    type = "HC1"))
```

```
## Pooling Model
##
## Note: Coefficient variance-covariance matrix supplied: vcovHC(ols_multiple, cluster = "group", type =
##
## Call:
## plm(formula = sdg_overall ~ spi_comp + sci_overall + di_score +
##     log_gdppc + factor(year), data = merged, model = "pooling",
##     index = c("country_code", "year"))
##
## Unbalanced Panel: n = 114, T = 4-5, N = 567
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -10.75847  -2.78830  -0.18876   2.55368  12.56261
##
## Coefficients:
##                    Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)        1.968133   3.053081  0.6446  0.519427
## spi_comp           0.124233   0.056824  2.1863  0.029211 *
## sci_overall        0.138766   0.048370  2.8688  0.004276 **
## di_score          -0.392001   0.275704 -1.4218  0.155638
## log_gdppc          5.792850   0.463222 12.5055 < 2.2e-16 ***
## factor(year)2017 -0.072062   0.188465 -0.3824  0.702339
## factor(year)2018  0.049536   0.357330  0.1386  0.889794
## factor(year)2019  0.399866   0.397804  1.0052  0.315245
## factor(year)2020  1.020984   0.620722  1.6448  0.100567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    44419
## Residual Sum of Squares: 10649
## R-Squared:      0.76027
## Adj. R-Squared: 0.75683
## F-statistic: 76.3702 on 8 and 113 DF, p-value: < 2.22e-16
```

2.2.1 Results in Stargazer Table & CSVs

Results: The results of the OLS models with robust standard errors indicate that both SPI and SCI have a statistically significant positive relationship with SDG composite scores. The p-values for both SPI and SCI are less than 0.001, indicating strong evidence against the null hypothesis of no relationship for either model. Holding all else constant (log GDP per capita, democracy score and year), SPI and SCI exhibit positive moderate and statistically significant relationships with SDG status.

ols_spi: 0.28735 (p-value < 0.001)*** ols_sci: 0.237633 (p-value < 0.001)*** ols_multiple: spi: 0.124233 (p-value < 0.05)* sci: 0.138766 (p-value < 0.01)**

When compared in separate models, SPI has a greater impact on SDG status (0.28735) than SCI (0.237633). This suggests that a one-unit increase in SPI is associated with a larger improvement in SDG outcomes compared to a one-unit increase in SCI, holding all controls constant.

Interestingly, the opposite holds true in a multiple regression model containing both SPI and SCI. SPI's impact on SDG status (0.124233) (net of SPI) is less than that of SCI's (0.138766) (net of SCI), holding all controls constant. When together, the coefficients represent the unique impact of each predictor variable (measures of statistical capacity) on SDG status, net of all other variables.

Model 1 (ols_spi) does not control for SCI and model 2 (ols_sci) does not control for spi – this is okay. SPI is the predecessor of the SCI, sharing/data overlap, and so it is expected to have significant statistical correlation (multicollinearity). This is possibly what explains the reduction of both regression coefficients in model 3: 0.28735 to 0.124233 for SPI (**% decrease); and from 0.237633 to 0.138766 for SCI** (% decrease). This indicates that both variables capture much of the same underlying relationship with SDG performance.

However, the fact that both SPI and SCI remain significant when included together (model 3), although SCI less so than SPI, with a high adjusted R-sq (0.741525), which suggests they capture different dimensions of statistical capacity that independently contribute to SDG status.

2.2.2 Checking for Multicolinearity: VIF of SPI & SCI [DONE]

```r
# Checking correlation between SPI and SCI
cor(merged$spi_comp, merged$sci_overall, use = "complete.obs")
```

```
## [1] 0.8276634
```

```r
# Checking VIF (Variance Inflation Factor) in Model 3
vif(ols_multiple)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## spi_comp     4.316894  1        2.077714
## sci_overall  3.754613  1        1.937682
## di_score     1.598314  1        1.264244
## log_gdppc    1.485085  1        1.218641
## factor(year) 1.302733  4        1.033611
```

```r
# Datatable
vif_vals <- vif(ols_multiple)  # returns a named vector
tidy_vif <- enframe(vif_vals, name = "term", value = "vif")

print(tidy_vif)
```

```
## # A tibble: 5 x 2
##   term         vif[,"GVIF"] [,"Df"] [,"GVIF^(1/(2*Df))"]
##   <chr>               <dbl>   <dbl>                <dbl>
## 1 spi_comp             4.32       1                 2.08
## 2 sci_overall          3.75       1                 1.94
## 3 di_score             1.60       1                 1.26
## 4 log_gdppc            1.49       1                 1.22
## 5 factor(year)         1.30       4                 1.03
```

```
# write.csv(tidy_vif, file =
# 'component_1/figures/vif_results.csv')
```

Colinearity: The correlation between SCI and SPI is about 0.8277. When placed within the same model, SCI inflated the standard error of SPI from 0.013079 to 0.027054. SCI had a similar reaction from the SPI with its standard error increasing from 0.0095864 to 0.024096.

VIF: Such multicollinearity is moderately reflected in the VIF test which accounts for all x variables in the model instead of just the two measures of statistical capacity (SCI & SPI).

VIF Results: term vif Df GVIF^(1/(2*Df)) spi_comp 4.316894 1 2.077714 sci_overall 3.754613 1 1.937682 di_score 1.598314 1 1.264244 log_gdppc 1.485085 1 1.218641 factor(year) 1.302733 4 1.033611

Overall there reveals no severe multicollinearity (all GVIF $<$ 5). There is moderate correlation between statistical capacity measures (spi_comp and sci_overall) with SPI moderately inflated by a factor of 4.32 and SCI inflated by a factor of 3.75. Nevertheless, it is acceptable to include both in the same model as doing so will not severely impact estimates with both factors less than 5.0. Even so, there are significant limitations in either model that warrant strong consideration, including sample size, and longitudinal suitability. All other variables show minimal multicollinearity concerns.

2.2.3 Checking for Heteroskedasticity: Breusch-Pagan Test [DONE]

This validates the need for integrating robust standard errors in our models

```
# Breusch-Pagan Test for Heteroskedasticity
bp_spi <- bptest(ols_spi)  # in objects for presentability
bp_sci <- bptest(ols_sci)
bp_multiple <- bptest(ols_multiple)

# combine for data frame
bp_tests <- list(ols_spi = bp_spi, ols_sci = bp_sci, ols_multiple = bp_multiple)

# Tidy all tests and add a 'model' column
bptests_results <- bp_tests %>%
    map_df(~tidy(.x), .id = "model")

print(bptests_results)
```

```
## # A tibble: 3 x 5
##   model        statistic  p.value parameter method
##   <chr>            <dbl>    <dbl>     <dbl> <chr>
## 1 ols_spi          207.   7.05e-39        10 studentized Breusch-Pagan test
## 2 ols_sci           86.6 4.22e-12        15 studentized Breusch-Pagan test
## 3 ols_multiple      31.1 1.33e- 4         8 studentized Breusch-Pagan test
```

```
# write.csv(bptests_results, file =
# 'component_1/figures/bp_heterosked_results.csv')
```

Model: BP statistic p-value ols_spi 207. = 7.05e-39
ols_sci 86.6 = 4.22e-12 ols_multiple 31.1 = 1.33e- 4

The Breusch-Pagan Test was applied to test to see whether residuals are constant across observations, which signals unaccounted non-linear relationships, especially with macro factors such as GDP Per Capita and

Population in the models. This is important because Ordinary Least Squares models assume constant error variance. In such a complex world of diverse cultural and ever-changing political structures across almost 200 countries, cross-national data, especially in development, is rarely ever linear. Accordingly, this test evaluates the extent of such non-linearity among specified predictors.

As such, results indicate strong evidence of heteroskedasticity in all three models. The small p-values in all models indicates that the variance of residuals are not constant across observations in all three models. This reinforces the motivation behind applying robust standard errors, which have been integrated to all OLS models. Without Robust SEs, there is a risk of inflated t-statistics, leading to false significance and misinterpretation of results.

Despite the improvement from 206.5 (SPI) and 86.6 (SCI) to 31.1 (Both), there still remains statically significant heteroskedasticity in the combined model. Both statistical capacity measures create a better-specified model (ols_multiple), though not enough to eliminate heteroskedasticity entirely.

2.2.4 Missing Data Structure & Interpretations [DONE]

Systematic, non-random missing data pattern: SPI has near complete country data coverage (165 out of 168 countries with an SDG score), but with a stubborn temporal limitation (2016-2023). On the other hand, SCI has longer temporal coverage (2004-2020) but lacks reporting on high-income countries focusing primarily on the developing world (123 out of 168 countries with an SDG score).

2.2.5 AIC/BIC Checking Fit [DONE]

H0: Null, SCI model or Combined model > SPI model H1: SPI model > SCI model & combined model

```r
# switching to lm for AIC & BIC tests [POLS]
ols_spi_lm <- lm(sdg_overall ~ spi_comp + di_score + log_gdppc + factor(year), data = merged)
ols_sci_lm <- lm(sdg_overall ~ sci_overall + di_score + log_gdppc + factor(year), data = merged)
ols_multiple_lm <- lm(sdg_overall ~ spi_comp + sci_overall + di_score + log_gdppc + factor(year), data =

# Switching to lm for AIC & BIC tests [FE]
fe_spi_lm <- lm(sdg_overall ~ spi_comp + di_score + log_gdppc + factor(year) + factor(country_code), da
fe_sci_lm <- lm(sdg_overall ~ sci_overall + di_score + log_gdppc + factor(year) + factor(country_code),
fe_multiple_lm <- lm(sdg_overall ~ spi_comp + sci_overall + di_score + log_gdppc + factor(year) + facto

# AIC & BIC tests - POLS models
AIC(ols_spi_lm, ols_sci_lm, ols_multiple_lm)
```

```
##                 df      AIC
## ols_spi_lm      12 7382.060
## ols_sci_lm      17 9003.178
## ols_multiple_lm 10 3292.001
```

```r
BIC(ols_spi_lm, ols_sci_lm, ols_multiple_lm)
```

```
##                 df      BIC
## ols_spi_lm      12 7443.496
## ols_sci_lm      17 9093.672
## ols_multiple_lm 10 3335.405
```

```r
# AIC & BIC tests - FE models
AIC(fe_spi_lm, fe_sci_lm, fe_multiple_lm)
```

```
##                 df      AIC
## fe_spi_lm      167 2588.303
## fe_sci_lm      133 4337.547
## fe_multiple_lm 123 1073.202
```

```r
BIC(fe_spi_lm, fe_sci_lm, fe_multiple_lm)
```

```
##                 df      BIC
## fe_spi_lm      167 3443.282
## fe_sci_lm      133 5045.529
## fe_multiple_lm 123 1607.066
```

```r
# Results into lists
aic_vals <- c(
  AIC(ols_spi_lm),
  AIC(ols_sci_lm),
  AIC(ols_multiple_lm),
  AIC(fe_spi_lm),
  AIC(fe_sci_lm),
  AIC(fe_multiple_lm)
)

bic_vals <- c(
  BIC(ols_spi_lm),
  BIC(ols_sci_lm),
  BIC(ols_multiple_lm),
  BIC(fe_spi_lm),
  BIC(fe_sci_lm),
  BIC(fe_multiple_lm)
)

# Model names
model_names <- c("ols_spi", "ols_sci", "ols_multiple", #POLS
                 "fe_spi", "fe_sci", "fe_multiple") #FE

# Combine into dataframe
aic_bic_comp1_results <- data.frame(
  model = model_names,
  AIC = aic_vals,
  BIC = bic_vals
)
print(aic_bic_comp1_results)
```

```
##          model      AIC      BIC
## 1      ols_spi 7382.060 7443.496
## 2      ols_sci 9003.178 9093.672
## 3 ols_multiple 3292.001 3335.405
## 4       fe_spi 2588.303 3443.282
## 5       fe_sci 4337.547 5045.529
## 6  fe_multiple 1073.202 1607.066
```

```
# saving to results_csv
#write.csv(aic_bic_comp1_results, file = "component_1/figures/aic_bic_results.csv")
```

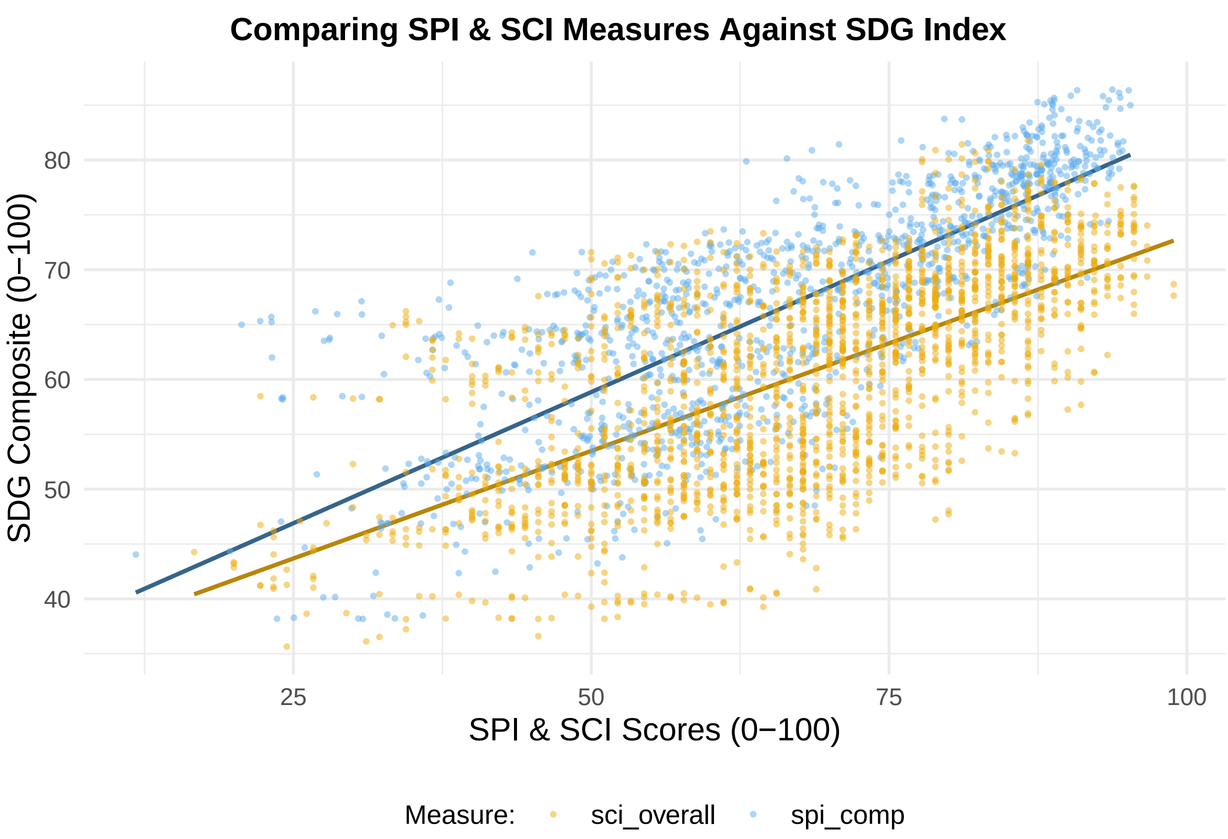
AIC/BIC Results model AIC BIC m1: ols_spi 7382.060 7443.496 m2: ols_sci 9003.178 9093.672 m3: ols_multiple 3292.001 3335.405 m4: fe_spi 2588.303 3443.282 m5: fe_sci 4337.547 5045.529 m6: fe_multiple 1073.202 1607.066 note: all models have different number of observations, ols_multiple containing the least.

Adjusted R-squares m1: ols_spi: 0.7757 [best fit] m2: ols_sci: 0.74152 m3: ols_multiple: 0.75683 m4: fe_spi
m5: fe_sci
m6: fe_multiple

2.2.6 Selecting Best Model [DONE]

Best fit: ols_spi (Adj Rsq: 0.7725) (AIC/BIC: 7382.060, 7443.496) (n=1082) To determine which predictor (i.e., SPI or SCI) is stronger –also considering if they're stronger together (ols_multiple)– this investigation considers (a) descriptive statistics (i.e., number of observations); (b) adjusted R2 estimates; and (c) AIC/BIC model scores. These measures are compared across models that sustain the same controls, however, statistical checks are weighted against frequency of observations, given anticipated differences in the number of years and mis-alignment of particular start to end year periods.

2.2.7 Visual Analysis of Fit: SCI & SPI x SDG



Note The SPI regression line is expected to appear higher in terms of SDG Score compared the SCI model because SPI countries include higher-income countries. As previously mentioned, the SCI soley focuses on lower to upper-middle income countries (146 countries over 17 years).