SPEC REU R Resources: Visualizing Regression Results with dot-and-whisker Plots – Groupwork

Yeiyoung Choo and Claudia Salas Gimenez

March 2025

In this groupwork, we will practice generating dot-and-whisker plots to visualize regression model results using the `ggcoefstats` function from the `ggstatsplot` package. The goal of this assignment is to help you become familiar with the process of extracting and manipulating regression outputs, as well as creating customized visualizations.

For this assignment, we will analyze the relationship between oil income and political rights. Specifically, we hypothesize that higher oil income is associated with lower political rights in the subsequent year, as increased reliance on oil income might reduce incentives for democratic reforms. We will use data from the following research paper:

Aslaksen, S. (2010). Oil and democracy: More than a cross-country correlation? Journal of Peace Research, 47(4), 421–431

Additionally, we will practice extracting regression results into a tidy dataframe using the `broom` package, which prepares the data for any personalized visualization you might need beyond dot-and-whisker plots.

Initial Setup

Before starting the exercises, set your working directory, load the required packages, and import the `aslaksen2010.rds` dataset. For reference, this dataset is used in the research paper:

Aslaksen, S. (2010). Oil and democracy: More than a cross-country correlation? Journal of Peace Research, 47(4), 421–431.

```
# Set working directory
#setwd("YourFolderPath")

# Load required libraries
library(tidyverse)
library(broom)
library(ggstatsplot)
## Install ggcoefstats from CRAN if not already installed

# Load the dataset
as2010 <- readRDS("aslaksen2010.rds")
```

Exploring the Relationship Between Oil Income and Political Rights

In this groupwork, we will examine how oil income influences political rights. We hypothesize that countries with higher oil income (expressed as a share of GDP) will have lower political rights in the subsequent year.

To test this hypothesis, we will first estimate a simple regression model to establish the baseline relationship between oil income and political rights. Next, we will run a multivariate regression model by adding relevant control variables that may influence the relationship between the key variables. Finally, we will visualize the regression results using dot-and-whisker plots generated by the `ggcoefstats` function to clearly observe the effect sizes and the uncertainty surrounding our estimates.

Exercise 1: Estimating Simple Regression Model

Let's first run a simple linear regression model to analyze the impact of `oilshare` alone on political rights in the upcoming year (`pr_lead`). The `oilshare` variable is not part of the original dataset, but you can construct it by multiplying the volume of oil extraction (`oil_wb`) by the price of oil (`oilprice`), then dividing this product by the GDP, and finally multiplying the result by 100 to express `oilshare` as a percentage. Also, replace NA values with 0 for countries where `oil_wb` equals 0, since a country that does not extract any oil should have an oil income of 0. This will prevent errors in subsequent analyses.

```r
# Create oilshare variable
as2010 <- as2010 %>%
  mutate(oilshare = 100 * (oil_wb * oilprice / gdp))

# Replace NAs with 0 where oil_wb == 0
as2010$oilshare[as2010$oil_wb == 0] <- 0

# Run simple regression model
m1 <- lm(pr_lead ~ oilshare, data = as2010)

# Display results
summary(m1)
```

```
##
## Call:
## lm(formula = pr_lead ~ oilshare, data = as2010)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5240 -0.3552 -0.0240  0.3103  0.7452
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5240012  0.0057091   91.78   <2e-16 ***
## oilshare    -0.0057084  0.0003551  -16.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3556 on 4535 degrees of freedom
##   (1098 observations deleted due to missingness)
## Multiple R-squared:  0.05391,    Adjusted R-squared:  0.0537
## F-statistic: 258.4 on 1 and 4535 DF,  p-value: < 2.2e-16
```

Exercise 2: Estimating a Multivariate Regression Model

To better understand the relationship between oil income and political rights, add control variables for additional factors. In this step, we include the current level of political rights (`pr`), population (`population`) and real GDP per capita (`rgdppc`). By controlling for `pr` we can more accurately isolate the effect of oil income on future political rights, ensuring that any observed changes in `pr_lead` are not simply due to pre-existing political conditions.

Note: In this step, make sure to transform population and real GDP per capita into logarithmic scales to simplify the scale differences. Name these new variables `lpop` (logged population) and `lrgdppc` (logged real GDP per capita).

```
# Transform population and real GDP per capita into their logarithmic scales
as2010 <- as2010 %>%
  mutate(lpop = log(population),
         lrgdppc = log(rgdppc))

# Run multiple regression model
m2 <- lm(pr_lead ~ pr + oilshare + lpop + lrgdppc , data = as2010)

# Check results
summary(m2)
```

```
##
## Call:
## lm(formula = pr_lead ~ pr + oilshare + lpop + lrgdppc, data = as2010)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98118 -0.02211 -0.00128  0.01230  0.80286
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0847748  0.0206169  -4.112 4.00e-05 ***
## pr           0.9318744  0.0056244 165.684  < 2e-16 ***
## oilshare    -0.0006740  0.0001210  -5.568 2.73e-08 ***
## lpop         0.0012616  0.0009902   1.274    0.203
## lrgdppc      0.0129660  0.0017616   7.360 2.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09944 on 4238 degrees of freedom
##   (1392 observations deleted due to missingness)
## Multiple R-squared:  0.9262, Adjusted R-squared:  0.9261
## F-statistic: 1.329e+04 on 4 and 4238 DF,  p-value: < 2.2e-16
```

Exercise 2.1: Expanding Multiple Regression Model

Finally, extend the multivariate regression model by adding education (`educ`) and a variable that accounts for the degree to which a country is integrated into the global economy (`open`) as additional predictors. Run the updated regression model and save it as `m3`.

```
# Run multiple regression model
m3 <- lm(pr_lead ~ pr + oilshare + lpop + lrgdppc + educ + open, data= as2010)

# Check results
summary(m3)
```

```
##
## Call:
## lm(formula = pr_lead ~ pr + oilshare + lpop + lrgdppc + educ +
##     open, data = as2010)
```

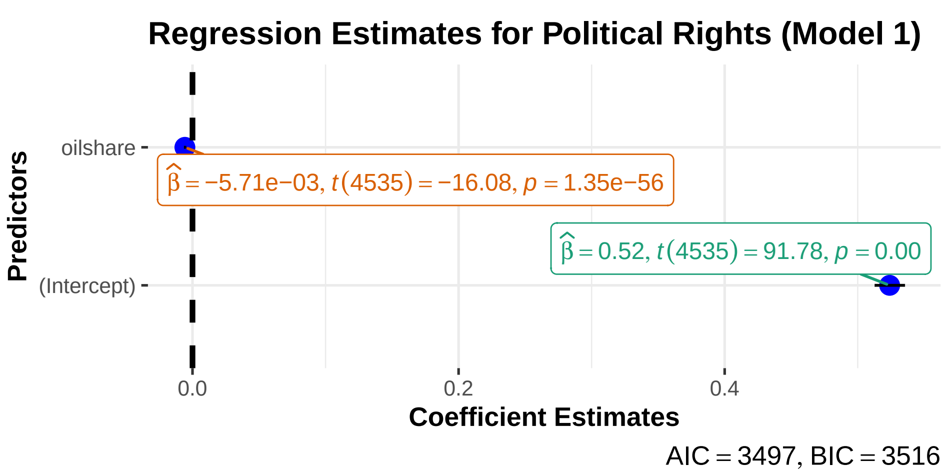
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97980 -0.02469 -0.00333  0.01592  0.78150
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.290e-03  3.348e-02   0.098 0.921727
## pr           9.076e-01  7.223e-03 125.648  < 2e-16 ***
## oilshare    -5.974e-04  1.684e-04  -3.547 0.000395 ***
## lpop        -1.754e-03  1.429e-03  -1.228 0.219599
## lrgdppc      8.368e-03  3.143e-03   2.662 0.007804 **
## educ         4.818e-03  1.130e-03   4.263 2.08e-05 ***
## open        -1.119e-04  4.727e-05  -2.366 0.018030 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1024 on 3143 degrees of freedom
##   (2485 observations deleted due to missingness)
## Multiple R-squared:  0.9188, Adjusted R-squared:  0.9186
## F-statistic:  5928 on 6 and 3143 DF,  p-value: < 2.2e-16
```

Exercise 3: Generating Dot-and-Whisker Plot with `ggstatsplot`

Exercise 3.1: Dot-and-Whisker Plot for a Single Model

Create a dot-and-whisker plot for the single regression models using the `ggcoefstats` function. As seen in the Data Visualization III Walkthrough, this function converts your regression results into a tidy dataframe (with coefficient estimates, standard errors, and confidence intervals) and plots each estimate as a dot with whiskers.

```
# Create dot-and-whisker plot using ggcoefstats
ggcoefstats(m1,
            title = "Regression Estimates for Political Rights (Model 1)",
            xlab = "Coefficient Estimates",
            ylab = "Predictors")
```

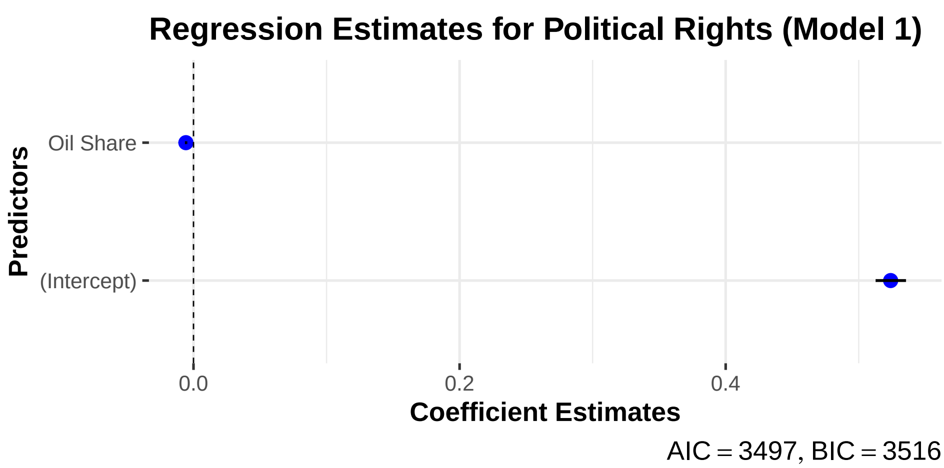


Exercise 3.2: Upgrading the Dot-and-Whisker Plot

Challenge yourself to enhance the clarity of the plot. For instance, rename the variables to be more descriptive and remove the detailed text labels.

For a complete list of all features you can modify in `ggcoefstats`, please refer to the ggcoefstats GitHub page.

```
# Upgrade the dot-and-whisker plot
ggcoefstats(x = m1,
            stats.labels = FALSE,
            ## Remove detailed text labels
            point.args = list(size = 2.0, color = "blue", na.rm = TRUE),
            vline.args = list(linewidth = 0.25, linetype = "dashed"),
            xlab = "Coefficient Estimates",
            ylab = "Predictors",
            title = "Regression Estimates for Political Rights (Model 1)") +
  scale_y_discrete(labels = c("pr" = "Political Rights (Current)",
                              "oilshare" = "Oil Share"))
```

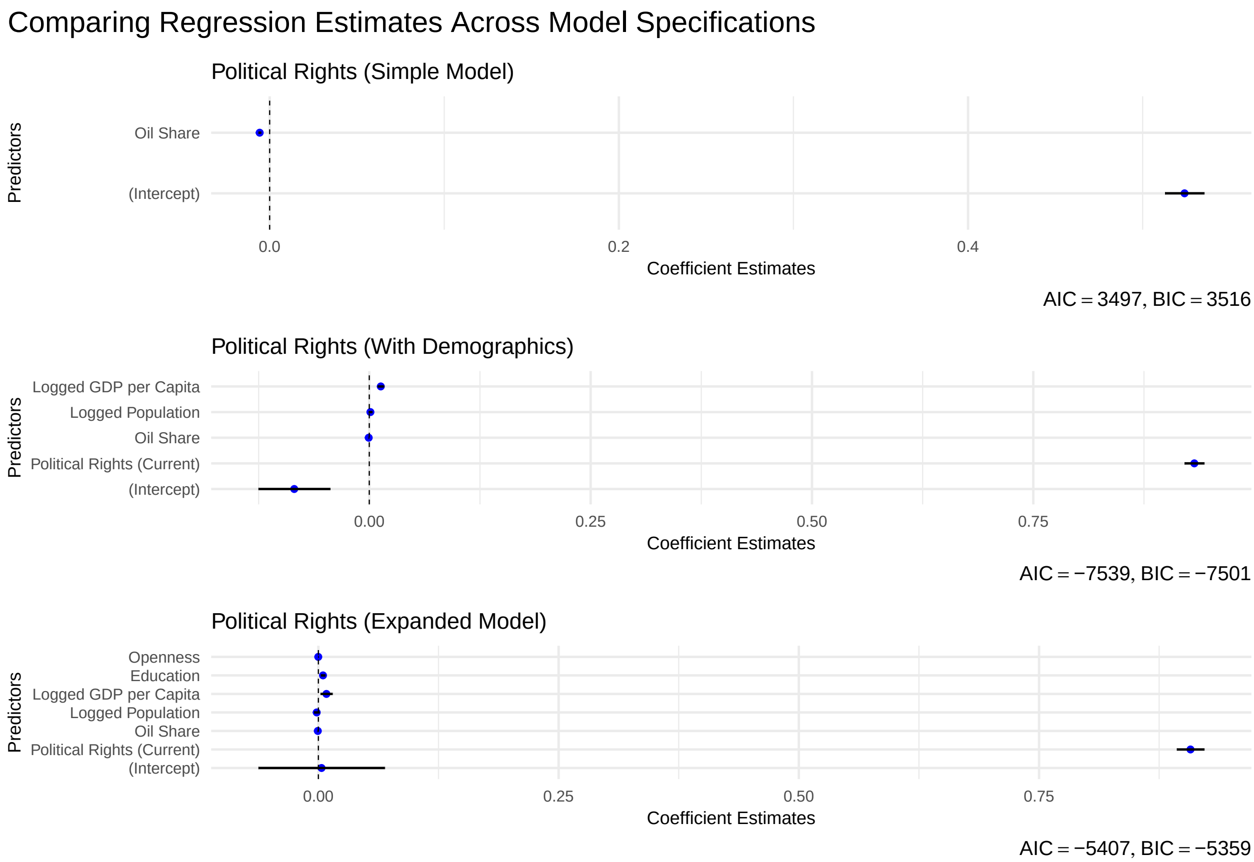


Exercise 3.3: Combined Dot-and-Whisker Plot for Multiple Models

Compare the regression estimates across models by combining dot-and-whisker plots for the single regression model and the two multivariate regression models into a single figure.

```
# Combine the dot-and-whisker plots for models m1 and m2
combine_plots(
  plotlist = list(
    # Plot for Model 1: Simple Model
    ggcoefstats(x = m1,
                stats.labels = FALSE,
                point.args = list(size = 1.0, color = "blue", na.rm = TRUE),
                vline.args = list(linewidth = 0.25, linetype = "dashed"),
                xlab = "Coefficient Estimates",
                ylab = "Predictors",
                title = "Political Rights (Simple Model)") +
      scale_y_discrete(labels = c("pr" = "Political Rights (Current)",
                                  "oilshare" = "Oil Share")) +
      theme_minimal() +
      theme(
        plot.title = element_text(size = 10),
```

```r
        axis.title = element_text(size = 8),
        axis.text = element_text(size = 7)
      ),
    # Plot for Model 2: With Demographic Controls
    ggcoefstats(x = m2,
                stats.labels = FALSE,
                point.args = list(size = 1.0, color = "blue", na.rm = TRUE),
                vline.args = list(linewidth = 0.25, linetype = "dashed"),
                xlab = "Coefficient Estimates",
                ylab = "Predictors",
                title = "Political Rights (With Demographics)") +
      scale_y_discrete(labels = c("pr" = "Political Rights (Current)",
                                  "oilshare" = "Oil Share",
                                  "lpop" = "Logged Population",
                                  "lrgdppc" = "Logged GDP per Capita")) +
      theme_minimal() +
      theme(
        plot.title = element_text(size = 10),
        axis.title = element_text(size = 8),
        axis.text = element_text(size = 7)
      ),
    # Plot for Model 3: Expanded Model
    ggcoefstats(x = m3,
                stats.labels = FALSE,
                point.args = list(size = 1.0, color = "blue", na.rm = TRUE),
                vline.args = list(linewidth = 0.25, linetype = "dashed"),
                xlab = "Coefficient Estimates",
                ylab = "Predictors",
                title = "Political Rights (Expanded Model)") +
      scale_y_discrete(labels = c("pr" = "Political Rights (Current)",
                                  "oilshare" = "Oil Share",
                                  "lpop" = "Logged Population",
                                  "lrgdppc" = "Logged GDP per Capita",
                                  "educ" = "Education",
                                  "open" = "Openness")) +
      theme_minimal() +
      theme(
        plot.title = element_text(size = 10),
        axis.title = element_text(size = 8),
        axis.text = element_text(size = 7)
      )),
  plotgrid.args = list(nrow = 3),
  annotation.args =
    list(title = "Comparing Regression Estimates Across Model Specifications"))
```



Bonus Exercise

Extend Model 3 by adding additional control variables from the data, and run a new regression model (Model 4) that includes all predictors from Model 3 plus these extra controls. Then, update your dot-and-whisker plot to compare all regression models. Provide a brief discussion on how the inclusion of the new control variables affects the coefficient estimates.

Exercise 4: Extracting and Customizing Tidy Regression Results

While `ggcoefstats` provides clear visualizations, you might want to further customize your output or create alternative visualizations. To do this, extract the model results into a tidy dataframe using the `broom` package. Use the `tidy()` function to obtain a dataframe with regression coefficients, standard errors, and p-values.

```
# Extract tidy regression results for each model
n1df <- tidy(m1)
n2df <- tidy(m2)
n3df <- tidy(m3)
```

Conclusion

In this groupwork, we examined how oil income influences political rights by running multiple regression analyses and visualizing the results with dot-and-whisker plots generated by the `ggcoefstats` function from the `ggstatsplot` package. We began with a simple model and progressively added additional controls to observe how the estimates evolved. We then combined the plots from different models into a single figure to facilitate direct comparisons.

Moving forward, we will continue to build on these concepts in the homework assignment, providing further opportunities to practice and refine your data visualization skills for displaying regression results.