SPEC REU R Resources: Visualizing & Analyzing Regressions Results with texreg

Miriam Barnum, Gaea Morales and Claudia Salas Gimenez

February 2025

In this module, we build on the foundational regression concepts covered in Regression I and shift our focus toward expanding regression analysis and presenting results in publication-quality tables. In research, regression outputs are not presented as raw statistical summaries or snapshots of R console outputs; instead, scholars use well-structured and professional tables to communicate findings clearly and effectively.

In this walkthrough, we will cover multiple linear regression models, which allow us to account for multiple explanatory variables simultaneously, introduce logistic regressions, a technique used when the dependent variable is binary (0 or 1), and learn how to create and customize publication-quality tables of regression results using the `texreg` package. To do so, we will work with the research question: Does natural resource dependency influence a country's level of democracy? By analyzing the relationship between natural resource rents and democracy indicators, we will explore how reliance on natural resources impacts governance structures. Economic wealth, measured by GDP per capita, will be included as a control variable to account for broader economic influences.

It is important to note that this modules provides a practical snapshot of different regression techniques but does not replace a formal econometrics or statistics course, where these concepts are explored in greater detail. Instead, this module aims to equip you with essential coding and interpretation skills necessary for real-world data analysis.

Initial Setup

First, set up your working directory and load the required libraries and dataset. We will use the `IDC_training_2021.RData` dataset introduced in the research paper "De Jure Powersharing 1975–2019: Updating the Inclusion, Dispersion, and Constraints Dataset" by Ziff, Barnum, Abadeer, Chu, Jao, Zaragoza, and Graham (2024).

You can access the paper and its appendix here.

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

# Load required libraries
library(dplyr)
library(ggplot2)

# Load the dataset
idc_controls <- readRDS("IDC_training_2021.rds")
```

Understanding Linear Regressions

Simple Linear Regression

As seen in Regression I, simple linear regression models the relationship between one independent variable (IV) and a dependent variable (DV). It assumes a direct, linear relationship between the two.

$$Y = \beta_o + \beta_1 X + \epsilon$$

Where:

- Y is the dependent variable.
- X is the independent variable.
- β_o is the intercept.
- β_1 is the coefficient for X, indicating its effect on Y.
- ϵ is the error term, which captures all factors affecting Y that are not included in the model (unobserved influences, measurement errors, or omitted variables that impact the dependent variable).

To address our research question, we will begin by examining whether natural resource rents (independent variable) have a significant effect on a country's level of democracy (dependent variable). Using the `IDC_training_2021.rds` data, we will use two measures to represent a country's level of democracy: electoral democracy (`v2x_polyarchy_VDEM`) and egalitarian democracy (`v2x_egaldem_VDEM`). These two indicators, derived from the Varieties of Democracy (V-Dem) dataset, capture different dimensions of democratic governance.

```
# Run simple linear regressions
## DV: electoral democracy
reg_poly <- lm(v2x_polyarchy_VDEM ~ natresource_rents_WDI, data = idc_controls)

## View regression summary
summary(reg_poly)
```

```
##
## Call:
## lm(formula = v2x_polyarchy_VDEM ~ natresource_rents_WDI, data = idc_controls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48809 -0.22919  0.00606  0.25673  0.78896
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.5645544  0.0039727  142.11   <2e-16 ***
## natresource_rents_WDI -0.0100432  0.0002778  -36.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2591 on 6346 degrees of freedom
##   (1168 observations deleted due to missingness)
## Multiple R-squared:  0.1708, Adjusted R-squared:  0.1706
## F-statistic:  1307 on 1 and 6346 DF,  p-value: < 2.2e-16
```

```
## DV: egalitarian democracy
reg_egal <- lm(v2x_egaldem_VDEM ~ natresource_rents_WDI, data = idc_controls)
```

```r
## View regression summary
summary(reg_egal)
```

```
##
## Call:
## lm(formula = v2x_egaldem_VDEM ~ natresource_rents_WDI, data = idc_controls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40246 -0.19843 -0.04327  0.20213  0.57133
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.4449315  0.0035163  126.53   <2e-16 ***
## natresource_rents_WDI -0.0085963  0.0002459  -34.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2294 on 6346 degrees of freedom
##   (1168 observations deleted due to missingness)
## Multiple R-squared:  0.1615, Adjusted R-squared:  0.1614
## F-statistic:  1222 on 1 and 6346 DF,  p-value: < 2.2e-16
```

However, many real-world relationships are not perfectly linear and involve multiple factors influencing the outcome. Relying only on one predictor can lead to omitted variable bias, where the regression fails to account for other relevant factors, leading to misleading results.

Multiple Linear Regression: Addressing Omitted Variables

A multiple linear regression expands the model by including additional independent variables:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

This helps control for additional factors that could distort the relationship between the primary independent variable and the dependent variable. We call control variables to those additional explanatory variables included in a model to account for other factors influencing the dependent variable. For example, in studying the effect of natural resource rents on democracy, failing to control for GDP per capita might lead to biased results, because wealthier countries might have preexisting differences in political structures, and we need to account for that in our model.

Helpful Hint: The R syntax is follows the same structure than for simple regression analysis, adding additional independent variables using `+`:

$$lm(y\ x1 + x2 + x3, data = data)$$

```r
# Run multiple linear regressions
# DV: electoral democracy
reg_poly2 <- lm(v2x_polyarchy_VDEM ~ natresource_rents_WDI + lngdppc_WDI_PW,
                data = idc_controls)

## View regression summary
summary(reg_poly2)
```

```
##
```

```
## Call:
## lm(formula = v2x_polyarchy_VDEM ~ natresource_rents_WDI + lngdppc_WDI_PW,
##     data = idc_controls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68133 -0.15899  0.05236  0.14548  0.71979
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.2337192  0.0152726  -15.30   <2e-16 ***
## natresource_rents_WDI -0.0089605  0.0002319  -38.64   <2e-16 ***
## lngdppc_WDI_PW         0.0955603  0.0017833   53.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2149 on 6312 degrees of freedom
##   (1201 observations deleted due to missingness)
## Multiple R-squared:  0.4297, Adjusted R-squared:  0.4295
## F-statistic:  2378 on 2 and 6312 DF,  p-value: < 2.2e-16
```

```
# DV: egalitarian democracy
reg_egal2 <- lm(v2x_egaldem_VDEM ~ natresource_rents_WDI + lngdppc_WDI_PW,
                data = idc_controls)

## View regression summary
summary(reg_egal2)
```

```
##
## Call:
## lm(formula = v2x_egaldem_VDEM ~ natresource_rents_WDI + lngdppc_WDI_PW,
##     data = idc_controls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51102 -0.12004  0.02072  0.12688  0.52113
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.4107450  0.0120194  -34.17   <2e-16 ***
## natresource_rents_WDI -0.0074582  0.0001825  -40.87   <2e-16 ***
## lngdppc_WDI_PW         0.1024220  0.0014034   72.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1691 on 6312 degrees of freedom
##   (1201 observations deleted due to missingness)
## Multiple R-squared:  0.5452, Adjusted R-squared:  0.545
## F-statistic:  3783 on 2 and 6312 DF,  p-value: < 2.2e-16
```

Beyond Linear Regression: Logistic Regression

Linear regression works well when the dependent variable is continuous, but what if we need to predict a binary outcome? Suppose we want to predict whether a country is democratic (1) or non-democratic (0).

Using a standard linear model could result in predictions outside the valid range (e.g., -0.3 or 1.2), which makes no sense. This happens because linear regression assumes a continuous outcome, but when predicting binary outcomes (e.g., success/failure, democratic/non-democratic), linear models can produce unrealistic predictions beyond the [0,1] range. Logistic regression provides a solution by modeling probabilities.

Since our research question explores whether natural resource dependency influences the likelihood of a country being democratic, we will change our dependent variable to the binary variable `democracy_BX`. To run a logistic regression this in R, use the `glm()` function instead of `lm()`, and specify the type of regression using `family = binomial`.

```
# Run logistic regression with binary democracy
log_reg_binary_dem <- glm(democracy_BX ~ natresource_rents_WDI + lngdppc_WDI_PW,
                      family = binomial,
                      data = idc_controls)

## View regression summary
summary(log_reg_binary_dem)
```

```
##
## Call:
## glm(formula = democracy_BX ~ natresource_rents_WDI + lngdppc_WDI_PW,
##     family = binomial, data = idc_controls)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.821692   0.188019  -25.64   <2e-16 ***
## natresource_rents_WDI -0.091799   0.003575  -25.68   <2e-16 ***
## lngdppc_WDI_PW         0.673890   0.023001   29.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8601.3  on 6205  degrees of freedom
## Residual deviance: 6439.5  on 6203  degrees of freedom
##   (1310 observations deleted due to missingness)
## AIC: 6445.5
##
## Number of Fisher Scoring iterations: 5
```

Creating Publication-Quality Regression Tables

While the `summary()` function in R provides detailed regression model outputs, these results are not formatted for publication. To present findings effectively in academic papers, policy reports, or professional settings, we need to create well-structured tables that clearly communicate regression results. The `texreg` package allows us to generate publication-ready tables in HTML, Word, and LaTeX formats.

Helful Hint: If you haven't installed `texreg` yet, make sure to run `install.packages("texreg")` before running loading the package.

```
# Load texreg package
library(texreg)
```

Generating Basic HTML Tables

Let's start by creating basic HTML tables. The `htmlreg()` function produces an HTML table that can be easily copied into Word documents. The basic syntax of the function is:

$$htmlreg(model\_list, file = "output.html")$$

where:

- `model_list` refers to the list of regression models that will be stored in the table.
- `file = "output.html"` indicates the name of the output file and the output type.

Let's try to print the simple regression models and multiple regression model we ran at the beginning of this walkthrough in a well polished table:

```
# Generate a basic HTML regression table
htmlreg(list(reg_poly, reg_poly2, reg_egal, reg_egal2),
        file = "democracyregs_21.html")
```

Why Show Both Simple and Multiple Regression Results?

In academic research, it is often important to present the results of both simple and multiple regression analyses. Simple regression models provide a baseline understanding of the relationship between a single independent variable and the dependent variable. They help identify whether a particular variable has any noticeable effect on the outcome before accounting for other factors.

Then, multiple regression models allow researchers to account for additional variables, providing a clearer picture of how each independent variable influences the dependent variable when other factors are held constant, helping to isolate relationships more accurately.

By comparing the results from simple and multiple regression models, researchers can better assess the robustness of their findings and determine whether additional explanatory variables significantly alter the observed relationships.

Customize HTML Table

While this table is great, we can make the table clearer, more readable and informative by adding additional elements such as a title, descriptive headers to separate models into meaningful groups, clear coefficient names for better interpretability, significance level indicators (stars) to denote statistical significance, or captions for additional context.

Note: `stars` are used to denote the significance levels of the coefficients (whether a coefficient is statistically significant and at what level). If regression coefficients are marked with stars at significance levels, let's say `p = 0.05, 0.01, 0.001`), it indicates stronger statistical evidence against the null hypothesis at the 95%, 99% and 99.99% confidence levels, suggesting robust results. The default levels are `p = 0.1, 0.05, 0.01`, but significance levels can be adjusted.

```
# Generate a customized HTML regression table
htmlreg(list(reg_poly, reg_poly2, reg_egal, reg_egal2),
          file = "democracyregs_21.html",
          caption = "Democracy and Resource Rents",
          caption.above = TRUE,
          custom.header = list("Electoral Democracy" = 1:2,
                               "Egalitarian Democracy" = 3:4),
          ## Labels models to indicate the dependent variable used
          custom.coef.names = c("Intercept", "Resource rents", "ln(GDPpc)"),
          ## Rename coefficients
          stars = c(0.001, 0.01, 0.05),
```

```
          ## Define significance levels
          custom.note = "%stars. Electoral and egalitarian democracy are indices from
          V-Dem; binary democracy is from Boix, Miller, and Rosato.")
          ## Add note explaining the sources of the democracy indices used
```

Generating Regression Tables for LaTeX

So far, we've been using the HTML format because it is easily compatible with Word documents. However, another widely used format in academic research is LaTeX. LaTeX is a high-quality typesetting system designed for producing professional documents, particularly those containing mathematical equations, complex tables, and citations. It is commonly used for academic papers and journal publications.

To generate a LaTeX-formatted regression table in R, simply switch to the `texreg` function and save the output as a `.tex` file, which can be directly integrated into a LaTeX document. The code syntax is the same as for the `htmlreg()` function.

```
# Generate a LaTeX regression table
texreg(list(reg_poly, reg_poly2, reg_egal, reg_egal2),
          file = "democracyregs_21.tex",
          caption = "Democracy, Wealth, and Resource Rents",
          caption.above = TRUE,
          custom.header = list("Electoral Democracy" = 1:3,
                               "Egalitarian Democracy" = 4:6),
          custom.coef.names = c("Intercept", "Resource rents", "ln(GDPpc)"),
          stars = c(0.001, 0.01, 0.05),
          custom.note = "%stars. Electoral and egalitarian democracy are indicies from
          vDem; binary democracy is from Boix, Miller, and Rosato.")
```

Omitting Certain Coefficients

When conducting regression analysis, scholars often need to account for unobserved factors that vary across countries, industries, or time periods, such as language, religion, or geography. These factors, if left unaccounted for, could bias our results and lead to misleading conclusions. One common solution is to use fixed effects, a technique that allows us to control for these unobserved characteristics by introducing dummy variables for each group.

Essentially, fixed effects control for unobserved factors that remain constant within a given group but vary across groups. These factors may influence the dependent variable but are not directly included in the regression model. By using fixed effects, we can isolate the impact of the independent variables on the dependent variable while controlling for these constant group-specific differences. For example:

- **Time fixed effects** → Control for global trends shocks or trends that affect all observations in the dataset in the time period being controlled for (usually year). Events like macroeconomic shocks and global trends (e.g., economic crises, technology booms, COVID-19 pandemic) would be accounted for with time-fixed effects.
- **Country fixed effects** → Control for differences across countries. For instance, each country has unique historical, institutional, and cultural factors that influence democracy but do not change much over time, and by using country fixed effects, we ensure that differences in democracy levels due to these historical and structural factors do not bias our estimates.

To incorporate these dummy variables into a regression, convert the relevant categorical variable (e.g., country) into a factor using `as.factor()`, and include it in the regression formula in `lm()`.

```
# Include country fixed effects in regression
reg_poly_fe <- lm(v2x_polyarchy_VDEM ~ natresource_rents_WDI + lngdppc_WDI_PW +
                    as.factor(country), data = idc_controls)
```

```
## Regression model with country fixed effects

summary(reg_poly_fe)
```

Because including dummy variables for each country could result in numerous coefficients, adding all of them to the table can make such unreadable. To create a cleaner table, we can omit these coefficients and instead add a row in the table that indicates we included fixed effects in the model.

The `omit.coef` argument in `htmlreg()` allows us to exclude coefficients that match a certain pattern (e.g., all variables containing the word `"country"`). Additionally, the `custom.gof.rows` argument lets us manually a row to indicate that fixed effects were included.

```
# Generate a table omitting country fixed effects coefficients
htmlreg(list(reg_poly, reg_poly2, reg_poly_fe),
        file = "democracyregsfe_21.html",
        caption = "Democracy, Wealth, and Resource Rents",
        caption.above = TRUE,
        custom.header = list("Electoral Democracy" = 1:4),
        custom.coef.names = c("Intercept", "Resource rents", "ln(GDPpc)"),
        stars = c(.05, .01, .001),
        custom.note = "%stars. Electoral and egalitarian democracy are indices from
        vDem; binary democracy is from Boix, Miller, and Rosato.",
        omit.coef = "country",
        ## Omits coefficients related to country fixed effects
        custom.gof.rows = list("Country FE" = c("No", "No", "Yes")))
        ## Indicates what models include fixed effects
```

Conclusion

In this module, we explored multiple linear regression and logistic regression to analyze how natural resource dependency affects democracy. We also learned how to present results using publication-quality regression tables to enhance clarity and professionalism in research.

Moving forward, we will keep practicing multivariate regression analysis and logistic regression, learn how extract regression results and interpret them, and explore more functions for creating high-quality regression tables to refine your data presentation skills.