

Week 10: Linear Regression III

POP88162 Introduction to Quantitative Research Methods

In this tutorial we will look at the data analysed by [Paglayan \(2021\)](#).

Download and read in the dataset `Paglayan2021_Simplified.csv` dataset. Remember, you will need to change the file path to the one that is correct for your computer.

```
paglayan2021 <- read.csv("../data/Paglayan2021_Simplified.csv")
```

For details on the data presented see [replication materials](#) for the article. Here we are looking at the simplified version of the `Education_LeeLee_Democracy_MAIN.dta` dataset.

It might be helpful to also check the original dataset. Note that it is saved as Stata (`.dta`) rather than comma-separated (`.csv`) file. In order to read in the original file make sure that you have package `haven` installed first.

```
paglayan2021_original <- haven::read_dta("../data/Paglayan2021.dta")
```

Start with usual checks of that dataset using `dim()`, `str()`, `head()` and `tail()` commands.

F Test

Let's start by fitting a simple bivariate linear regression modelling an association between political regime and student enrollment rate (SER)

$$SER_i = \alpha + \beta_1 Democracy_i + \epsilon_i$$

Here we saved the fitted model object under the name `lm_fit_1`. Now we can use `summary()` function to print out detailed model output.

```
lm_fit_1 <- lm(primary_ser ~ democracy, data = paglayan2021)
summary(lm_fit_1)
```

```
##
## Call:
## lm(formula = primary_ser ~ democracy, data = paglayan2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -82.137 -25.485   4.083  13.083  55.523
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44.4769     0.7515   59.19  <2e-16 ***
## democracy    43.5206     1.2706   34.25  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.27 on 2494 degrees of freedom
```

```
## (1755 observations deleted due to missingness)
## Multiple R-squared: 0.3199, Adjusted R-squared: 0.3196
## F-statistic: 1173 on 1 and 2494 DF, p-value: < 2.2e-16
```

What is the F-statistic and the associated p-value for this model? What model does it compare this model to?

Let's fit a multiple linear regression model with the same dependent variable, but now controlling for the region:

$$SER_i = \alpha + \beta_1 Democracy_i + \beta_2 Region_i + \epsilon_i$$

```
lm_fit_2 <- lm(primary_ser ~ democracy + region, data = paglayan2021)
summary(lm_fit_2)
```

```
##
## Call:
## lm(formula = primary_ser ~ democracy + region, data = paglayan2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -86.554 -22.469   2.598  19.613  65.030
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      51.123      1.351   37.847 < 2e-16 ***
## democracy        41.291      1.351   30.557 < 2e-16 ***
## regionAsia and the Pacific -12.164      2.053   -5.925 3.55e-09 ***
## regionEastern Europe      9.928      2.503    3.966 7.51e-05 ***
## regionLatin America and the Caribbean -16.153      1.567  -10.311 < 2e-16 ***
## regionMiddle East and North Africa      1.326      2.393    0.554  0.580
## regionSub-Saharan Africa     -3.063      2.143   -1.429  0.153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.14 on 2489 degrees of freedom
## (1755 observations deleted due to missingness)
## Multiple R-squared: 0.3713, Adjusted R-squared: 0.3697
## F-statistic: 245 on 6 and 2489 DF, p-value: < 2.2e-16
```

Now let's explicitly compare the two models:

```
anova(lm_fit_1, lm_fit_2)
```

```
## Analysis of Variance Table
##
## Model 1: primary_ser ~ democracy
## Model 2: primary_ser ~ democracy + region
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     2494 2285695
## 2     2489 2113094   5    172601 40.661 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the F-statistic and the associated p-value for this model? What is its interpretation? Which model is nested within which?

Dummy Variables

```
summary(lm_fit_2)
```

```
##
## Call:
## lm(formula = primary_ser ~ democracy + region, data = paglayan2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -86.554 -22.469   2.598  19.613  65.030
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   51.123     1.351   37.847 < 2e-16 ***
## democracy                     41.291     1.351   30.557 < 2e-16 ***
## regionAsia and the Pacific    -12.164     2.053   -5.925 3.55e-09 ***
## regionEastern Europe          9.928     2.503    3.966 7.51e-05 ***
## regionLatin America and the Caribbean -16.153     1.567 -10.311 < 2e-16 ***
## regionMiddle East and North Africa  1.326     2.393    0.554  0.580
## regionSub-Saharan Africa      -3.063     2.143   -1.429  0.153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.14 on 2489 degrees of freedom
## (1755 observations deleted due to missingness)
## Multiple R-squared:  0.3713, Adjusted R-squared:  0.3697
## F-statistic: 245 on 6 and 2489 DF, p-value: < 2.2e-16
```

What are the interpretations of the coefficients for regions? What is the reference category?

Using factor variables in R, change the reference category for region to 'Sub-Saharan Africa'. Re-fit the model.

What changed? How do the two models compare?

Interaction

Now let's look at the interaction between regime type and region and how those are associated with the school enrollment rate.

```
lm_fit_3 <- lm(primary_ser ~ democracy * region, data = paglayan2021)
summary(lm_fit_3)
```

```
##
## Call:
## lm(formula = primary_ser ~ democracy * region, data = paglayan2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -85.17 -22.02   2.84  15.30  68.20
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   53.574     1.714   31.260 < 2e-16 ***
## democracy                     37.457     2.144   17.473 < 2e-16 ***
## regionAsia and the Pacific    -14.952     2.521   -5.931 3.44e-09 ***
```

```

## regionEastern Europe          9.866      2.968   3.324 0.000901 ***
## regionLatin America and the Caribbean -21.774     2.113 -10.304 < 2e-16 ***
## regionMiddle East and North Africa   -1.090     2.703  -0.403 0.686869
## regionSub-Saharan Africa            -1.945     2.607  -0.746 0.455693
## democracy:regionAsia and the Pacific    5.518     4.657   1.185 0.236190
## democracy:regionEastern Europe        -8.532     5.916  -1.442 0.149418
## democracy:regionLatin America and the Caribbean 15.447     3.193   4.839 1.39e-06 ***
## democracy:regionMiddle East and North Africa    3.401     7.618   0.446 0.655336
## democracy:regionSub-Saharan Africa       -14.277     4.919  -2.903 0.003734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.88 on 2484 degrees of freedom
## (1755 observations deleted due to missingness)
## Multiple R-squared:  0.3835, Adjusted R-squared:  0.3808
## F-statistic: 140.5 on 11 and 2484 DF, p-value: < 2.2e-16

```

What is your substantive conclusion given this output?