Applied Statistical Analysis I

Multiple linear regression

Hannah Frank
frankh@tcd.ie

Department of Political Science
Trinity College Dublin

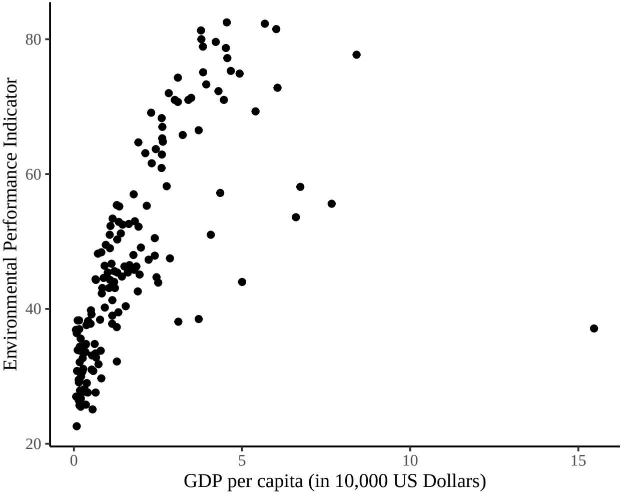
November 22, 2023

Today's Agenda

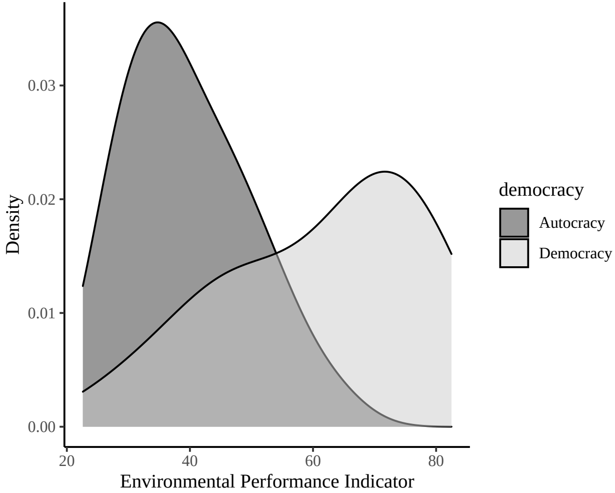
(1) Lecture recap

(2) Tutorial exercises: What is the relationship between education and Euroscepticism?

Income and environmental protection



Regime type and environmental protection



Categorical independent variables

How to include categorical independent variables with more than two levels?

Categorical independent variables

$$\mathit{Environmental\ performance}_i = \alpha + \beta_1 * \mathit{Income}_i + \beta_2 * \mathit{Region}_i + \epsilon_i$$

```
## table(qog_data$ht_region)
##
##                     Eastern Europe (1)              Latin America(2)
##                               28                                20
##     North Africa & the Middle East (3)        Sub-Saharan Africa (4)
##                               20                                49
##   Western Europe and North America (5)                 East Asia (6)
##                               27                                 6
##                    South-East Asia (7)                South Asia (8)
##                               11                                 8
##                        The Pacific (9)            The Caribbean (10)
##                               12                                13
```

Categorical independent variables

```r
# Load package
library(fastDummies)

# Create dummy variables for categorical variable
qog_data <- dummy_cols(qog_data,
                       select_columns = c("ht_region"))

# Print first 5 rows in dataset
head(qog_data[c("ht_region_1",
           "ht_region_2",
           "ht_region_3",
           "ht_region_4",
           "ht_region_5",
           "ht_region_6",
           "ht_region_7",
           "ht_region_8",
           "ht_region_9",
           "ht_region_10")], 5)
```

```
##   ht_region_1 ht_region_2 ht_region_3 ht_region_4 ht_region_5
## 1           0           0           0           0           0
## 2           1           0           0           0           0
## 3           0           0           1           0           0
## 4           0           0           0           0           1
## 5           0           0           0           1           0
##   ht_region_6 ht_region_7 ht_region_8 ht_region_9 ht_region_10
## 1           0           0           1           0            0
## 2           0           0           0           0            0
## 3           0           0           0           0            0
## 4           0           0           0           0            0
## 5           0           0           0           0            0
```

Categorical independent variables

```r
# Run regression model
m2 <- lm(epi_epi ~ income +
            ht_region_1 + ht_region_2 + ht_region_3 +
            # no region 4 (Sub-Saharan Africa) = reference category.
            ht_region_5 + ht_region_6 + ht_region_7 + ht_region_8 +  ht_region_9 +
            ht_region_10, data = qog_data)

# Print results
summary(m2)
```

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.3992     1.1296  28.683  < 2e-16 ***
income         1.7410     0.4061   4.287 3.23e-05 ***
ht_region_1   18.4245     1.8769   9.817  < 2e-16 ***
ht_region_2   11.6208     2.0362   5.707 6.01e-08 ***
ht_region_3    9.4434     2.4665   3.829 0.000189 ***
ht_region_5   35.2532     2.4854  14.184  < 2e-16 ***
ht_region_6   16.2287     3.6737   4.418 1.91e-05 ***
ht_region_7    4.1247     2.7820   1.483 0.140281
ht_region_8   -2.1694     3.2676  -0.664 0.507774
ht_region_9        NA         NA      NA       NA
ht_region_10  11.0665     3.5607   3.108 0.002257 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.528 on 149 degrees of freedom
  (35 observations deleted due to missingness)
Multiple R-squared:  0.7897,Adjusted R-squared:  0.777
F-statistic: 62.16 on 9 and 149 DF,  p-value: < 2.2e-16
```

Categorical independent variables

```r
# Use relevel to code dummy variables on the fly
# specify region 4 (Sub−Saharan Africa) = reference category
m3 <− lm(epi_epi ~ income + relevel(as.factor(ht_region), ref = "4"),
             data = qog_data)

# Print results
summary(m3)
```

```
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                32.3992     1.1296  28.683  < 2e-16 ***
income                                      1.7410     0.4061   4.287 3.23e-05 ***
relevel(as.factor(ht_region), ref = "4")1  18.4245     1.8769   9.817  < 2e-16 ***
relevel(as.factor(ht_region), ref = "4")2  11.6208     2.0362   5.707 6.01e-08 ***
relevel(as.factor(ht_region), ref = "4")3   9.4434     2.4665   3.829 0.000189 ***
relevel(as.factor(ht_region), ref = "4")5  35.2532     2.4854  14.184  < 2e-16 ***
relevel(as.factor(ht_region), ref = "4")6  16.2287     3.6737   4.418 1.91e-05 ***
relevel(as.factor(ht_region), ref = "4")7   4.1247     2.7820   1.483 0.140281
relevel(as.factor(ht_region), ref = "4")8  -2.1694     3.2676  -0.664 0.507774
relevel(as.factor(ht_region), ref = "4")10 11.0665     3.5607   3.108 0.002257 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.528 on 149 degrees of freedom
  (35 observations deleted due to missingness)
Multiple R-squared:  0.7897,Adjusted R-squared:  0.777
F-statistic: 62.16 on 9 and 149 DF,  p-value: < 2.2e-16
```

Under control of income, Eastern Europe has an Environmental Performance Index score of 18.4245 scale points higher than Sub-Saharan Africa.

Interactions

What are interactions?

Interactions

The association between X on Y might vary depending on the value of a third variable M (=Moderator):

$$\hat{Y}_i = \alpha + \beta_1 X_i + \beta_2 M_i + \beta_3 (X_i M_i) + \epsilon_i$$

The interpretation of the regression coefficients changes:

- α is the expected value of Y when $X = 0$ and $M = 0$
- β_1 is the change in Y when X increases by one unit, when $M = 0$
- β_2 is the change in Y when M increases by one unit, when $X = 0$
- β_3 is the *interaction term* of X and M

Rearrange terms:

$$\hat{Y}_i = \alpha + \beta_2 M_i + (\beta_1 + \beta_3 M_i) X_i + \epsilon_i$$

β_3 is the *added* increase in β_1, if M increases by one unit.

Categorical by continuous interaction

$$\textit{Environmental Performance}_i = \alpha + \beta_1 \textit{Income}_i + \beta_2 \textit{Regime Type}_i + \beta_3 \textit{Income}_i * \textit{Regime Type}_i + \epsilon_i$$

```r
# Run regression model with interaction term
int_m2 <- lm(epi_epi ~ income + democracy + income*democracy, data = qog_data)

# Print results
summary(int_m2)
```

```
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                37.1474     1.0684  34.768  < 2e-16 ***
## income                      2.1902     0.4532   4.833 3.24e-06 ***
## democracyDemocracy          3.4490     2.7819   1.240    0.217
## income:democracyDemocracy   5.1029     0.8686   5.875 2.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.046 on 153 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.6879, Adjusted R-squared:  0.6818
## F-statistic: 112.4 on 3 and 153 DF,  p-value: < 2.2e-16
```

Categorical by continuous interaction

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               37.1474     1.0684  34.768  < 2e-16 ***
## income                     2.1902     0.4532   4.833 3.24e-06 ***
## democracyDemocracy         3.4490     2.7819   1.240    0.217    
## income:democracyDemocracy  5.1029     0.8686   5.875 2.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.046 on 153 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.6879, Adjusted R-squared:  0.6818
## F-statistic: 112.4 on 3 and 153 DF,  p-value: < 2.2e-16
```

- The average Environmental Protection Index (EPI) for poor (Income=0) autocracies is 37.1474 scale points *(α)*.
- For autocracies, with every additional 10,000 USD of income, the EPI increases by 2.1902 scale points *(β_1)*. → Income effect for autocracies
- For poor democracies, the EPI is 3.4490 scale points higher, in comparison to poor autocracies *(β_2)*.
- For democracies, with every additional 10,000 USD of income, the EPI increases by 7.2931 scale points *($\beta_1 + \beta_3$=2.1902+5.1029=7.2931)*. → Income effect for democracies

Categorical by continuous interaction

Model for Autocracies (democracy = 0)

$\hat{Y}_i = 37.1474 + (2.1902 * Income_i) + (3.4490 * Regime\ Type_i) + (5.1029 * Income_i * Regime\ Type_i)$

$\hat{Y}_i = 37.1474 + (2.1902 * Income_i) + (3.4490 * 0) + (5.1029 * Income_i * 0)$

$\hat{Y}_i = 37.1474 + (2.1902 * Income_i)$

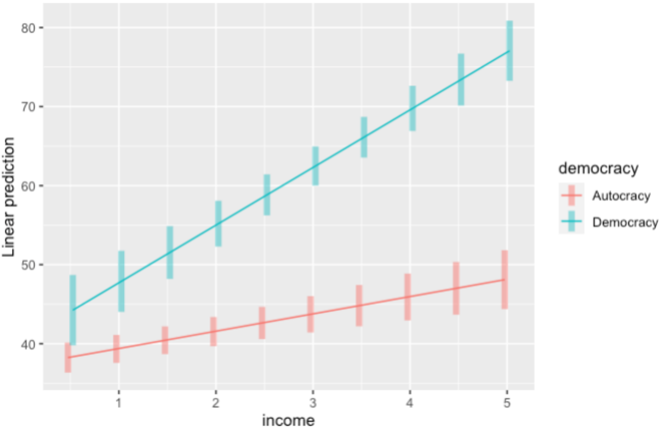
Model for Democracies (democracy = 1)

$\hat{Y}_i = 37.1474 + (2.1902 * Income_i) + (3.4490 * Regime\ Type_i) + (5.1029 * Income_i * Regime\ Type_i)$

$\hat{Y}_i = 37.1474 + (2.1902 * Income_i) + (3.4490 * 1) + (5.1029 * Income_i * 1)$

$\hat{Y}_i = 40.5964 + (7.2931 * Income_i)$

Categorical by continuous interaction



Non-linear effects

Model a curvilinear (=curved lines) relationship between an independent variable and the dependent variable.

Include X and the square of X:

$$\hat{Y}_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

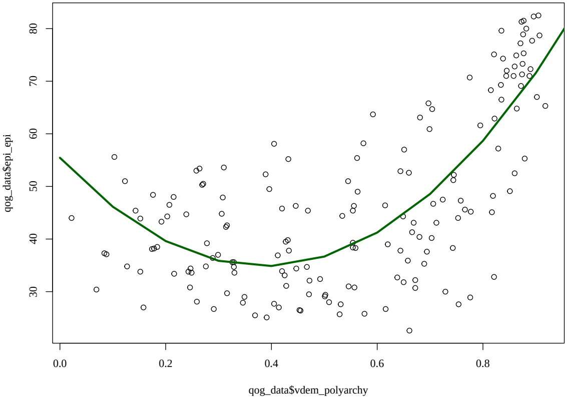
Non-linear effects

"U-shaped" relationship between democracy and environment protection?

```r
# Generate quadratic term
qog_data$sqr_vdem_polyarchy <- qog_data$vdem_polyarchy^2

# Run ols regression with quadratic term
q_m1 <- lm(epi_epi ~  income + vdem_polyarchy
             + sqr_vdem_polyarchy,
             data = qog_data)

# Print results
summary(q_m1)
```

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         39.4244     4.2944   9.180 2.82e-16 ***
income               3.0094     0.4576   6.576 7.19e-10 ***
vdem_polyarchy     -44.3531    17.7037  -2.505   0.0133 *
sqr_vdem_polyarchy  74.1559    17.0553   4.348 2.50e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.133 on 153 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.6819,Adjusted R-squared:  0.6757
F-statistic: 109.3 on 3 and 153 DF,  p-value: < 2.2e-16
```

Non-linear effects



What is the relationship between education and Euroscepticism?

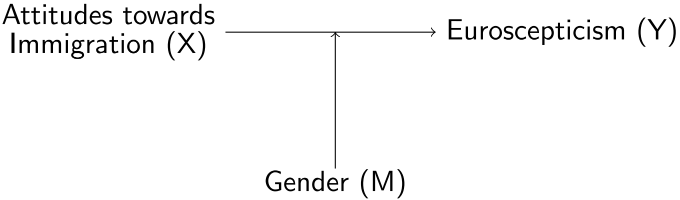
H_1: The higher the years of education, the lower the level of Euroscepticism.

H_2: The higher the income, the lower the level of Euroscepticism.

H_3: The higher the trust in politics, the lower the level of Euroscepticism.

H_4: The more positive attitudes towards immigration, the lower the level of Euroscepticism.

Does gender influence the effect of attitudes towards immigration on Euroscepticism?

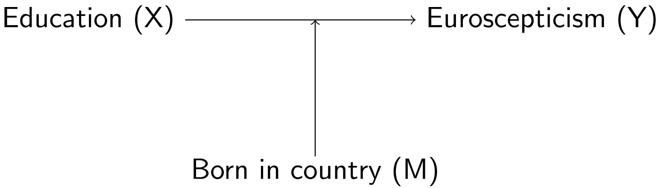


Attitudes towards Immigration (X)

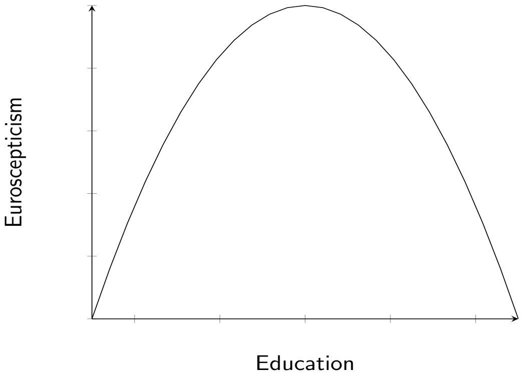
Euroscepticism (Y)

Gender (M)

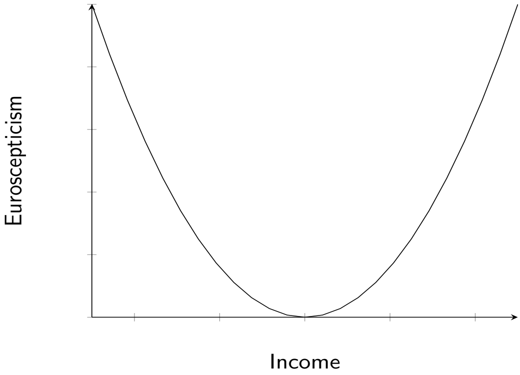
Does whether the person was born in the country influence the effect of education on Euroscepticism?



Is the effect of education on Euroscepticism inverted U-shaped?



Is the effect of income on Euroscepticism U-shaped?


Euroscepticism
Income