

APPLIED STATISTICAL ANALYSIS I

Multiple linear regression

Hannah Frank
frankh@tcd.ie

Department of Political Science
Trinity College Dublin

November 8, 2023

Today's Agenda

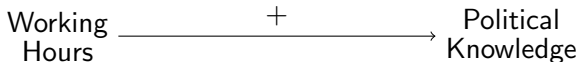
- (1) Lecture recap
- (2) Monte Carlo simulation
- (3) Tutorial exercises: What is the relationship between education and Euroscepticism?

Multiple linear regression

Why do we need multiple linear regression? And what is a multiple linear regression model?

Multiple linear regression

Why do we need multiple linear regression?



```
## Call:
## lm(formula = polknow ~ work_hours, data = samp)

## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.686 -1.760 -0.061  1.683 10.385

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.59166    1.09142  13.369  <2e-16 ***
## work_hours   0.06791    0.02640   2.572  0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 2.565 on 998 degrees of freedom
## Multiple R-squared:  0.006585, Adjusted R-squared:  0.00559
## F-statistic: 6.615 on 1 and 998 DF, p-value: 0.01025
```

How convincing is this finding?

Multiple linear regression

Why do we need multiple linear regression?

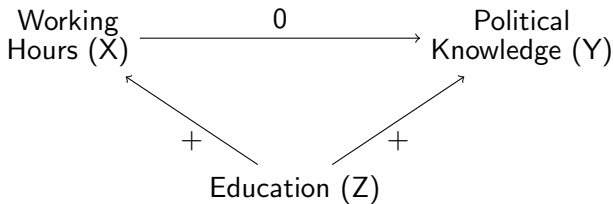


Figure: **Education as confounder**—Controlling for education is relevant, because it might drive both working hours and political knowledge. Education is causally prior to working hours.

→ **Avoid omitted variable bias.** Include relevant control variables (Z) which are correlated with both X and Y , and causally prior to X .

Multiple linear regression

Why do we need multiple linear regression?

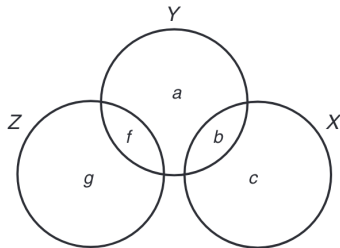


Figure 9.2. Venn diagram in which X and Z are correlated with Y, but not with each other.

“In that case – which, we have noted, is unlikely in applied research – we can safely omit consideration of Z when considering the effects of X on Y. In that figure, the relationship between X and Y – the area b – is unaffected by the presence (or absence) of Z in the model” (Kellstedt and Whitten 2018, 213).

Multiple linear regression

Why do we need multiple linear regression?

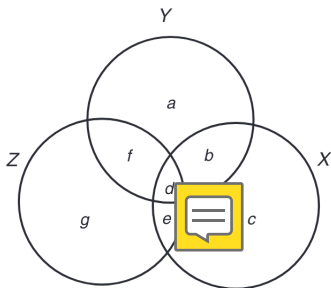


Figure 9.1. Venn diagram in which X, Y, and Z are correlated.

"If, hypothetically, we erased the circle for Z from the figure, we would (incorrectly) attribute all of the area $b + d$ to X, when in fact the d portion of the variation in Y is shared by both X and Z. This is why, when Z is related to both X and Y, if we fail to control for Z, we will end up with a biased estimate of X's effect on Y" (Kellstedt and Whitten 2018, 212).

Multiple linear regression

What is a multiple linear regression model?

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i$$

- α (intercept): expected value of Y when $X_1 = 0, \dots, X_k = 0$.
- β_1 (coefficient): expected change in Y when X_1 increases by one unit, while controlling for the remaining independent variables in the model.
- ...
- β_k (coefficient): expected change in Y when X_k increases by one unit, while controlling for the remaining independent variables in the model.

Multiple linear regression

What is a multiple linear regression model?

```
## Call:
## lm(formula = polknow ~ work_hours + edu, data = samp)

## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.7835 -1.6733  0.0035  1.5941 10.6778

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.854461   1.368601   3.547 0.000408 ***
## work_hours   0.006205   0.025623   0.242 0.808714
## edu          0.767650   0.070797  10.843 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 2.427 on 997 degrees of freedom
## Multiple R-squared:  0.1114, Adjusted R-squared:  0.1096
## F-statistic: 62.48 on 2 and 997 DF, p-value: < 2.2e-16
```

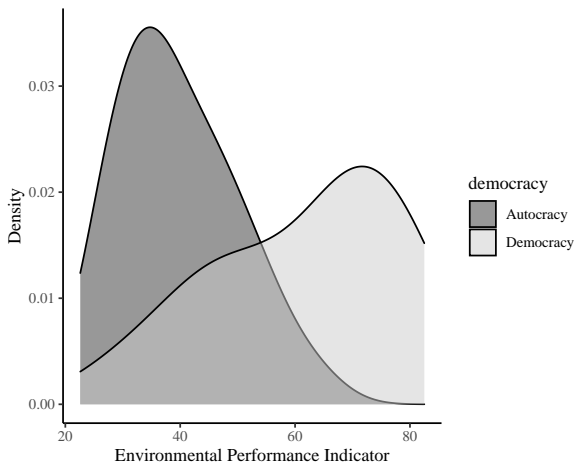
The effect of working hours *disappears*.

→ Controlling for working hours, with every additional year of education, the political knowledge increases by 0.76765 scale points.

Categorical independent variables

What is the reference category?

What is the reference category?



$$Environmental\ Performance_i = \alpha + \beta_1 * Regime\ Type_i$$

What is the reference category?

Dummy variables should take value 0 and 1 for easy interpretation →
Re-code existing variables.

```
1 # Import data from Quality of Government dataset
2 qog_data <- read.csv("qog-bas-cs-jan21.csv")
3
4 # Generate dummy variable for regime type as factor variable — democracy
5 # vdem_polyarchy ranges between 0 and 1; cutoff at 0.7
6 # Countries with score equal or above 0.7 are democracies, those below autocracies
7 qog_data$democracy <- factor(ifelse(qog_data$vdem_polyarchy >= 0.7, 1, 0))
8
9 # Define levels of democracy in factor variable
10 levels(qog_data$democracy) <- c("Autocracy", "Democracy")
11
12 # Summarize generated dummy variable
13 summary(qog_data$democracy)
```

##	Autocracy	Democracy	NA's
##	119	54	21

What is the reference category?

```
1 # Generate dummy variable for regime type as factor variable — autocracy
2 qog_data$autocracy <- factor(ifelse(qog_data$vdem_polyarchy < 0.7, 1, 0))
3
4 # Define levels of autocracy in factor variable
5 levels(qog_data$autocracy) <- c("Democracy", "Autocracy")
6
7 # Print first 10 rows in dataset
8 head(qog_data[c("democracy", "autocracy")], 10)
```

	democracy	autocracy
1	0 Autocracy	1 Autocracy
2	0 Autocracy	1 Autocracy
3	0 Autocracy	1 Autocracy
4	<NA>	<NA>
5	0 Autocracy	1 Autocracy
6	<NA>	<NA>
7	0 Autocracy	1 Autocracy
8	1 Democracy	0 Democracy
9	1 Democracy	0 Democracy
10	1 Democracy	0 Democracy



What happens if we run:

$$\text{Environmental Performance}_i = \alpha + \beta_1 \text{Democracy}_i + \beta_2 \text{Autocracy}_i + \epsilon_i$$

What is the reference category?

$$\text{Environmental Performance}_i = \alpha + \beta_1 \text{Democracy}_i + \beta_2 \text{Autocracy}_i + \epsilon_i$$

```
1 # Fit regression model
2 ml_trap <- lm(eps_emi ~ democracy + autocracy, data = qog_data)
3
4 # Print results
5 summary(ml_trap)
```

```
lm(formula = eps_emi ~ democracy + autocracy, data = qog_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-34.107	-8.860	-0.610	9.293	26.190

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	39.610	1.138	34.80	<2e-16 ***
democracy1	22.098	2.002	11.04	<2e-16 ***
autocracy1	NA	NA	NA	NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Violates assumption of no perfect multicollinearity (essentially a data problem) →
One category needs to be excluded = reference category. Interpretation of the model is relative to the reference category.

Binary independent variables

How to include binary independent variables in multiple linear regression?

Binary independent variables

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

```
## Call:
## lm(eps_emi ~ democracy + income, data = qog_data)

## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.563  -6.502   0.498   6.773  20.198

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      35.3027      1.1269  31.327 < 2e-16 ***
## democracyDemocracy 16.5270      1.8409   8.978 9.08e-16 ***
## income              3.5793      0.4266   8.390 2.92e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 9.982 on 154 degrees of freedom
## (37 observations deleted due to missingness)
## Multiple R-squared:  0.6175, Adjusted R-squared:  0.6126
## F-statistic: 124.3 on 2 and 154 DF,  p-value: < 2.2e-16
```

In comparison to autocracies (= reference category), democracies have a 16.5270 scale point higher score on the Environmental Performance Index, under control of income.

Binary independent variables

$$\hat{Y}_i = \alpha + \beta_1 * \text{Regime Type}_i + \beta_2 * \text{Income}_i$$

Model for Autocracies:

$$\hat{Y}_i = 35.303 + (16.527 * \text{Regime Type}_i) + (3.579 * \text{Income}_i)$$

$$\hat{Y}_i = 35.303 + (16.527 * 0) + (3.579 * \text{Income}_i)$$

$$\hat{Y}_i = 35.303 + (3.579 * \text{Income}_i)$$

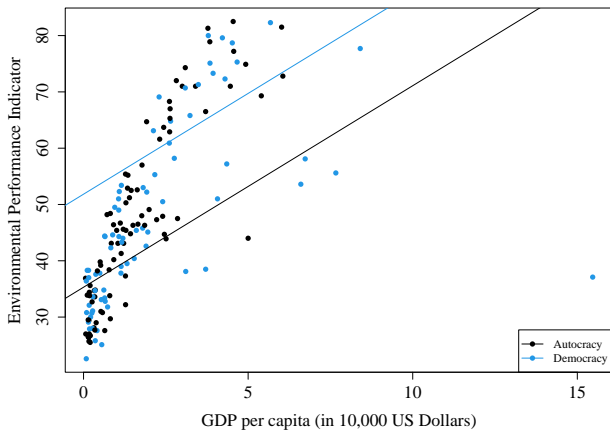
Model for Democracies:

$$\hat{Y}_i = 35.303 + (16.527 * \text{Regime Type}_i) + (3.579 * \text{Income}_i)$$

$$\hat{Y}_i = 35.303 + (16.527 * 1) + (3.579 * \text{Income}_i)$$

$$\hat{Y}_i = 51.83 + (3.579 * \text{Income}_i)$$

Binary independent variables



Categorical independent variables

How to select the reference category?

How to select the reference category?

```
1 # Run regression model with democracy variable
2 m1_dem <- lm(epi_epi ~ income + democracy, data = qog_data)
3
4 # Run regression model with autocracy variable
5 m1_aut <- lm(epi_epi ~ income + autocracy, data = qog_data)
6
7 # Get regression table with stargazer
8 stargazer(m1_dem, m1_aut)
```

	<i>Dependent variable:</i>	
	epi_epi	
	(1)	(2)
income	3.579*** (0.427)	3.579*** (0.427)
democracy1	16.527*** (1.841)	
autocracy1		-16.527*** (1.841)
Constant	35.303*** (1.127)	51.830*** (1.892)
Observations	157	157
R ²	0.618	0.618
Adjusted R ²	0.613	0.613
F Statistic (df = 2; 154)	124.331***	124.331***
Note:	* p<0.1; ** p<0.05; *** p<0.01	

How to select the reference category?

Model 1 for Autocracies:

$$\hat{Y}_i = 35.303 + (16.527 * \textit{Regime Type}_i) + (3.579 * \textit{Income}_i)$$

$$\hat{Y}_i = 35.303 + (16.527 * 0) + (3.579 * \textit{Income}_i)$$

$$\hat{Y}_i = 35.303 + (3.579 * \textit{Income}_i)$$

Model 2 for Autocracies:

$$\hat{Y}_i = 51.830 + (-16.527 * \textit{Regime Type}_i) + (3.579 * \textit{Income}_i)$$

$$\hat{Y}_i = 51.830 + (-16.527 * 1) + (3.579 * \textit{Income}_i)$$

$$\hat{Y}_i = 35.303 + (3.579 * \textit{Income}_i)$$

→ Mathematically identical models.



How do we select the reference category?



How to select the reference category?

```
1 # Run regression model with democracy variable
2 m1 <- lm(eps_eps ~ income + democracy, data = qog_data)
3
4 # Get regression table with stargazer
5 stargazer(m1)
```


Dependent variable:	
	eps_eps
democracy1	16.527*** (1.841)
income	3.579*** (0.427)
Constant	35.303*** (1.127)
Observations	157
R ²	0.618
Adjusted R ²	0.613
F Statistic (df = 2; 154)	124.331***
Note: * p<0.1; ** p<0.05; *** p<0.01	

In comparison to autocracies (= reference category), democracies have a 16.5270 scale point higher score on the Environmental Performance Index, under control of income.


Categorical independent variables

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


Categorical independent variables




Country	Region
Afghanistan	Asia
Albania	EE
Algeria	MENA
Argentina	LA
Australia	Advanced
⋮	⋮



Country	X_{region}
Afghanistan	2
Albania	3
Algeria	5
Argentina	4
Australia	1
⋮	⋮



Country	X_{Asia}	X_{EE}	X_{LA}	X_{MENA}	$X_{Sub-Saharan}$
Afghanistan	1	0			0
Albania	0	1	0	0	0
Algeria	0	0	0	1	0
Argentina	0	0	1	0	0
Australia	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮



$$\text{School enrollment rate} = \alpha + \beta_1 \text{Democracy}_i + \beta_2 \text{Region}_{EE} + \beta_3 \text{Region}_{LA} + \beta_4 \text{Region}_{MENA} + \beta_5 \text{Region}_{Sub-Saharan} + \epsilon_i$$

→ Include binary/dummy variables for all levels minus (=reference category).

- α (intercept): expected value of Y when $X_k = 0$
- β (coefficient): expected change in Y for $X = 1$, in comparison to reference category

Categorical independent variables

→ Convert into factor variable, then R automatically generates dummy variables, with first level as reference category (or change with relevel-function).

```
1 # Code dummy variables on the fly
2 # specify region Sub-Saharan Africa = reference category
3 lm <- lm(primary_ser ~ democracy + relevel(as.factor(region), ref="Sub-Saharan
  Africa"), data = paglayan2021)
4
5 # Print model output
6 summary(lm)
```

Call:

```
lm(formula = primary_ser ~ democracy + relevel(as.factor(region),
  ref = "Sub-Saharan Africa"), data = paglayan2021)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	48.060	1.796	26.754	< 2e-16 ***
democracy	41.291	1.351	30.557	< 2e-16 ***
ref = "Sub-Saharan Africa")Advanced Economies	3.063	2.143	1.429	0.153007
ref = "Sub-Saharan Africa")Asia and the Pacific	-9.101	2.437	-3.734	0.000192 ***
ref = "Sub-Saharan Africa")Eastern Europe	12.991	2.825	4.599	4.46e-06 ***
ref = "Sub-Saharan Africa")Latin America and the Caribbean	-13.090	2.073	-6.315	3.20e-10 ***
ref = "Sub-Saharan Africa")Middle East and North Africa	4.389	2.695	1.629	0.103515

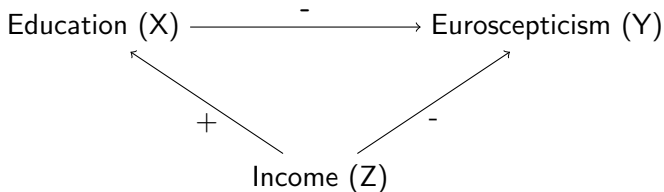
Under control of regime type, Eastern Europe has a student enrollment rate of 12.991 percentage points higher than Sub-Saharan Africa.

What is the relationship between education and Euroscepticism?

Education (X) —————⁻————→ Euroscepticism (Y)

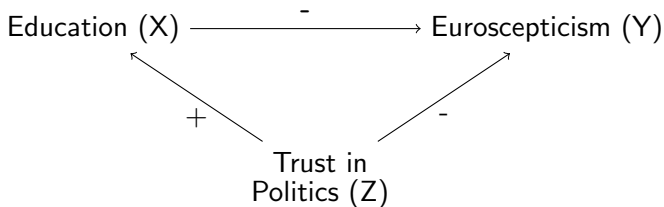
Hypothesis₁: The higher the years of education, the lower the level of Euroscepticism.

What is the relationship between education and Euroscepticism?



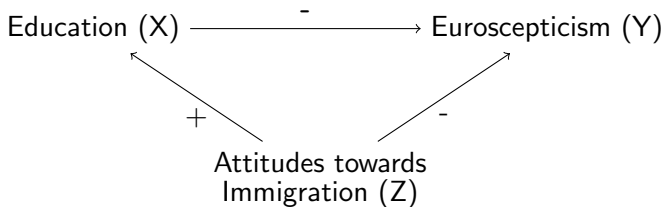
Hypothesis₂: The higher the income, the lower the level of Euroscepticism. → Economic dimension

What is the relationship between education and Euroscepticism?



Hypothesis₃: The higher the trust in politics, the lower the level of Euroscepticism. → Political dimension

What is the relationship between education and Euroscepticism?



Hypothesis₃: The more positive attitudes towards immigration, the lower the level of Euroscepticism. → Cultural dimension

References I



Kellstedt, Paul M., and Guy D. Whitten. 2018. *The fundamentals of political science research*. Cambridge: Cambridge University Press.