Applied Statistical Analysis I

Multiple linear regression

Hannah Frank
frankh@tcd.ie

Department of Political Science
Trinity College Dublin

November 15, 2023

Today's Agenda

(1) Lecture recap

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

T-TEST FOR INDIVIDUAL COEFFICIENTS

What is the t-test for individual coefficients?

T-TEST FOR INDIVIDUAL COEFFICIENTS

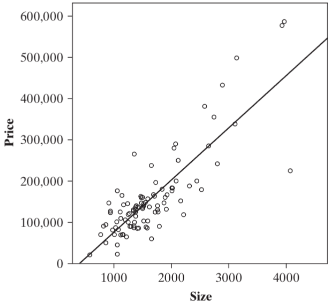
What is the t-test for individual coefficients?

- Null and alternative hypotheses:
  - there is no association between X and Y, $\beta = 0$ (H_0)
  - there is an association between X and Y, $\beta \neq 0$ (H_1)
- Test statistic: "measures the number of standard errors between the estimate and the H_0 value" (Agresti and Finlay 2009, 192).

  $t = \frac{\textit{Estimate of parameter} - \textit{Null hypothesis value of parameter}}{\textit{Standard error of estimate}}$

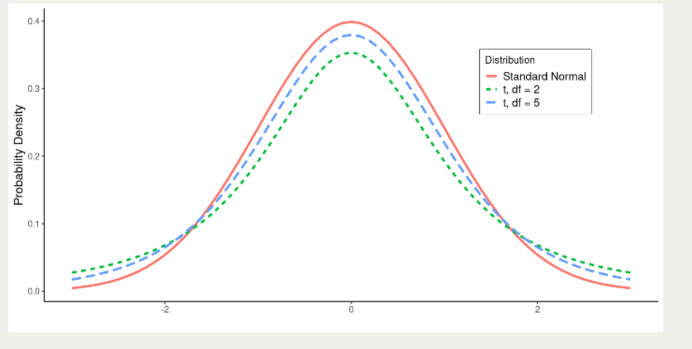
  $t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$, H_0 assumes $\beta = 0$

T-TEST FOR INDIVIDUAL COEFFICIENTS



- Is there an association between house selling price and size (Agresti and Finlay 2009, 278–279)? $Price = 50,926.2 + 126.6 * Size$
- $t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} = \frac{126.6}{8.47} = 14.95$
- How to interpret this value? How likely are we to observe data in sample (this test statistics), under the assumption that H_0 is true? $\rightarrow$ Probability distribution

T-test for individual coefficients



What is the conclusion? P-value < 0.05, We can reject H_0 with an error probability (p-value) of essentially 0%. $\rightarrow$ There is an association between house selling price and size

F-TEST

Table 3.5 Regression Output for Supervisor Performance Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	10.787	11.5890	0.93	0.3616
X_1	0.613	0.1610	3.81	0.0009
X_2	−0.073	0.1357	−0.54	0.5956
X_3	0.320	0.1685	1.90	0.0699
X_4	0.081	0.2215	0.37	0.7155
X_5	0.038	0.1470	0.26	0.7963
X_6	−0.217	0.1782	−1.22	0.2356
$n = 30$	$R^2 = 0.73$	$R_a^2 = 0.66$	$\hat{\sigma} = 7.068$	df = 23

Table 3.2 Description of Variables in Supervisor Performance Data

Variable	Description
Y	Overall rating of job being done by supervisor
X_1	Handles employee complaints
X_2	Does not allow special privileges
X_3	Opportunity to learn new things
X_4	Raises based on performance
X_5	Too critical of poor performance
X_6	Rate of advancing to better jobs

(Chatterjee and Hadi 2015, 59)

F-test

General set-up: Test whether reduced model (RM) is adequate (H_0) or full model (FM) is adequate (H_1).

The reduced model is nested within the full model $\rightarrow$ compare "the goodness of fit that is obtained when using the full model, to the goodness of fit that results using the reduced model".

$$F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

(Chatterjee and Hadi 2015, 71–72)

F-test

$$F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

* Sum of squared errors (SSE), denotes *lack of fit* $\rightarrow$ $SSE(RM) - SSE(FM)$ "represents the increase in the residual sum of squares due to fitting the reduced model".
* We use the ratio, weighted by "respective degrees of freedom to compensate for the different number of parameters involved in the two models".
* p=number of IVs full model, n=number of observations, k=number of paraneters reduced model

(Chatterjee and Hadi 2015, 71–72)

F-test

Two versions of the F-test

1. "All the regression coefficients are zero".
2. "Some of the regression coefficients are zero".

(Chatterjee and Hadi 2015, 71)

F-test for all coefficients

What is the F-test for all coefficients?

F-test for all coefficients

"All the regression coefficients are zero."

* Reduced model (RM): $Y = \beta_0 + \epsilon$

  all slopes are equal to zero, $\beta_k = 0$ (H_0) $\rightarrow$ the null model performs better

* Full model (FM): $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \epsilon$

  at least one slope is different from zero, $\beta_p \neq 0$ (H_1) $\rightarrow$ the full model performs better

$$F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)} = \frac{[SST - SSE]/p}{SSE/(n-p-1)} = \frac{SSR/p}{SSE/(n-p-1)}$$

"Because the least squares estimate of β_0 in the reduced model is $\bar{y}$, the residual sum of squares from the reduced model is SSE(RM)=SST." "reduced model has one regression parameter and the full model has p+1 regression parameter". "Because SST=SSR+SSE, we can replace SST-SSE by SSR"

(Chatterjee and Hadi 2015, 73)

F-test for all coefficients

Table 3.5 Regression Output for Supervisor Performance Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	10.787	11.5890	0.93	0.3616
X_1	0.613	0.1610	3.81	0.0009
X_2	−0.073	0.1357	−0.54	0.5956
X_3	0.320	0.1685	1.90	0.0699
X_4	0.081	0.2215	0.37	0.7155
X_5	0.038	0.1470	0.26	0.7963
X_6	−0.217	0.1782	−1.22	0.2356
$n = 30$	$R^2 = 0.73$	$R_a^2 = 0.66$	$\hat{\sigma} = 7.068$	df = 23

Table 3.7 Supervisor Performance Data: Analysis of Variance (ANOVA) Table

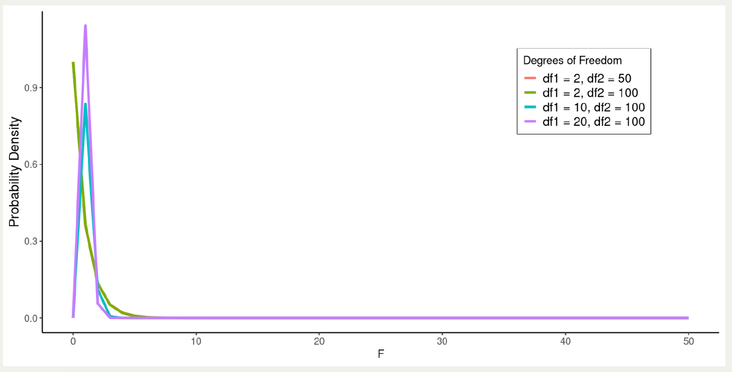
Source	Sum of Squares	df	Mean Square	F-Test
Regression	3147.97	6	524.661	10.5
Residuals	1149.00	23	49.9565	

$$F = \frac{SSR/p}{SSE/(n-p-1)} = \frac{3147.97/6}{1149.00/23} = 10.50$$

How to interpret this value? How likely are we to observe data in sample (this test statistics), under the assumption that H_0 is true?
$\rightarrow$ Probability distribution

(Chatterjee and Hadi 2015, 75)

F-test for all coefficients



What is the conclusion? P-value < 0.05, We can reject H_0 with an error probability (p-value) of essentially 0%. $\rightarrow$ The full model performs better, "not all $\beta's$ can be taken as zero"

(Chatterjee and Hadi 2015, 75).

Partial F-test

What is the F-test for some coefficients?

Partial F-test

"Some of the regression coefficients are zero".

* Reduced model (RM): $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$

  subset of slopes is equal to zero, $\beta_k = 0$ (H_0) $\rightarrow$ the reduced model performs better

* Full model (FM): $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \epsilon$

  at least one slope in the subset is different from zero, $\beta_p \neq 0$ (H_1) $\rightarrow$ the full model performs better

$$F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

(Chatterjee and Hadi 2015, 77)

Partial F-test

Table 3.5 Regression Output for Supervisor Performance Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	10.787	11.5890	0.93	0.3616
X_1	0.613	0.1610	3.81	0.0009
X_2	−0.073	0.1357	−0.54	0.5956
X_3	0.320	0.1685	1.90	0.0699
X_4	0.081	0.2215	0.37	0.7155
X_5	0.038	0.1470	0.26	0.7963
X_6	−0.217	0.1782	−1.22	0.2356
$n = 30$	$R^2 = 0.73$	$R_a^2 = 0.66$	$\hat{\sigma} = 7.068$	df = 23

Table 3.7 Supervisor Performance Data: Analysis of Variance (ANOVA) Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	3147.97	6	524.661	10.5
Residuals	1149.00	23	49.9565	

(Chatterjee and Hadi 2015, 75)

Partial F-test

Table 3.8 Regression Output from the Regression of Y on X_1 and X_3

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	3042.32	2	1521.1600	32.7
Residuals	1254.65	27	46.4685	

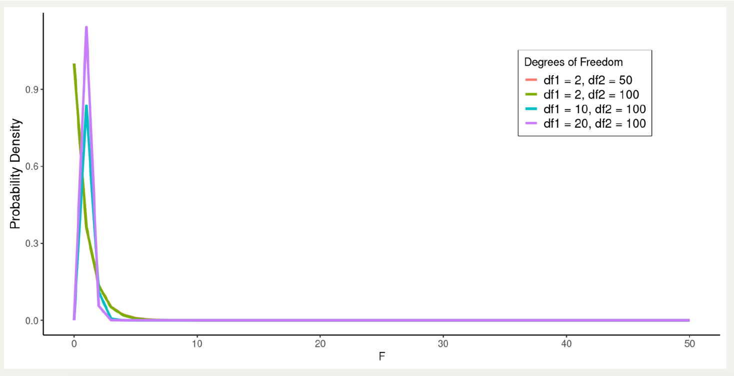
Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	9.8709	7.0610	1.40	0.1735
X_1	0.6435	0.1185	5.43	< 0.0001
X_3	0.2112	0.1344	1.57	0.1278
$n = 30$	$R^2 = 0.708$	$R_a^2 = 0.686$	$\hat{\sigma} = 6.817$	df = 27

$$F = \frac{[1254.65 - 1149]/4}{1149/23} = 0.0528$$

How to interpret this value? How likely are we to observe data in sample (this test statistics), under the assumption that H_0 is true? $\rightarrow$ Probability distribution

(Chatterjee and Hadi 2015, 76)

Partial F-test



What is the conclusion? P-value > 0.05, We cannot reject H_0. $\rightarrow$ The reduced model performs better. "The variables X_1 and X_3 together explain the variation in Y as adequately as the full set of six variables" (Chatterjee and Hadi 2015, 77).

Binary independent variables

How to include binary independent variables in multiple linear regression?

Binary independent variables

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

```
## Call:
## lm(epi_epi ~ 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

$$\hat{Y}_i = \alpha + \beta_1 * Regime\ Type_i + \beta_2 * Income_i$$

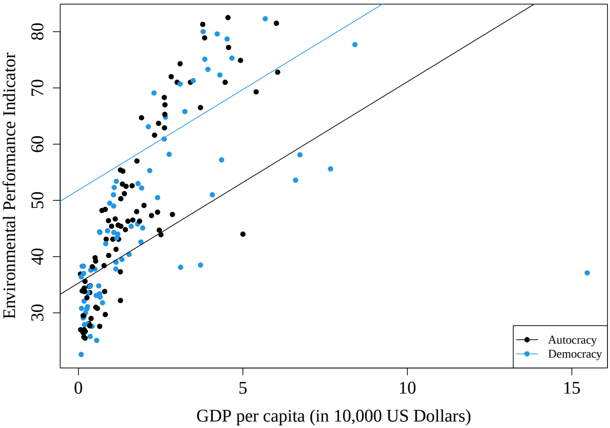
Model for Autocracies:

$$\hat{Y}_i = 35.303 + (16.527 * Regime\ Type_i) + (3.579 * Income_i)$$
$$\hat{Y}_i = 35.303 + (16.527 * 0) + (3.579 * Income_i)$$
$$\hat{Y}_i = 35.303 + (3.579 * Income_i)$$

Model for Democracies:

$$\hat{Y}_i = 35.303 + (16.527 * Regime\ Type_i) + (3.579 * Income_i)$$
$$\hat{Y}_i = 35.303 + (16.527 * 1) + (3.579 * Income_i)$$
$$\hat{Y}_i = 51.83 + (3.579 * Income_i)$$

Binary independent variables



Categorical independent variables

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

Categorical independent variables

Country	X_{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	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
⋮	⋮	⋮	⋮	⋮	⋮

$$School\ enrollment\ rate = \alpha + \beta_1 Democracy_i + \beta_2 Region_{EE} + \beta_3 Region_{LA} + \beta_4 Region_{MENA} + \beta_5 Region_{Sub-Saharan} + \epsilon_i$$

$\rightarrow$ Include binary/dummy variables for all levels minus one (=reference category).

- α (intercept): expected value of Y when $X = 0$
- 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).

```
# Code dummy variables on the fly
# specify region Sub-Saharan Africa = reference category
lm <- lm(primary_ser ~ democracy + relevel(as.factor(region), ref="Sub-Saharan
      Africa"), data = paglayan2021)

# Print model output
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?

$$\text{Education (X)} \xrightarrow{\quad - \quad} \text{Euroscepticism (Y)}$$

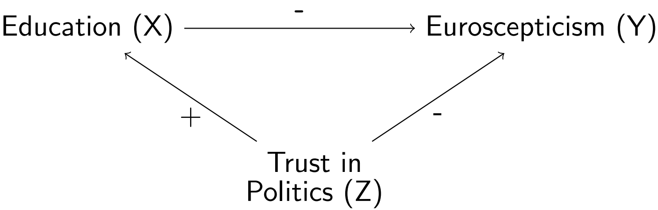
Hypothesis$_1$: The higher the years of education, the lower the level of Euroscepticism.

What is the relationship between education and Euroscepticism?

$$
\begin{tikzcd}
\text{Education (X)} \arrow[rr, "-"] & & \text{Euroscepticism (Y)} \\
& \text{Income (Z)} \arrow[ul, "+"] \arrow[ur, "-"'] &
\end{tikzcd}
$$

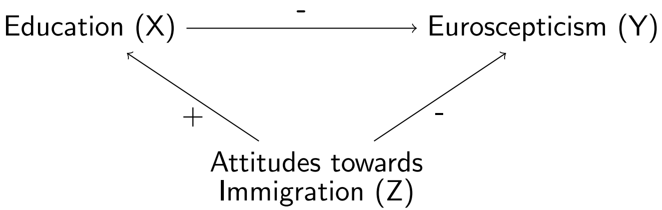
Hypothesis$_2$: The higher the income, the lower the level of Euroscepticism. $\rightarrow$ Economic dimension

What is the relationship between education and Euroscepticism?



Hypothesis$_3$: The higher the trust in politics, the lower the level of Euroscepticism. → Political dimension

What is the relationship between education and Euroscepticism?



Hypothesis$_3$: The more positive attitudes towards immigration, the lower the level of Euroscepticism. → Cultural dimension

References I

Agresti, Alan, and Barbara Finlay. 2009. *Statistical methods for the social sciences.* Essex: Pearson Prentice Hall.

Chatterjee, Samprit, and Ali S. Hadi. 2015. *Regression analysis by example.* Somerset: Wiley.