Week 10: Linear Regression III

POP88162 Introduction to Quantitative Research Methods

Tom Paskhalis

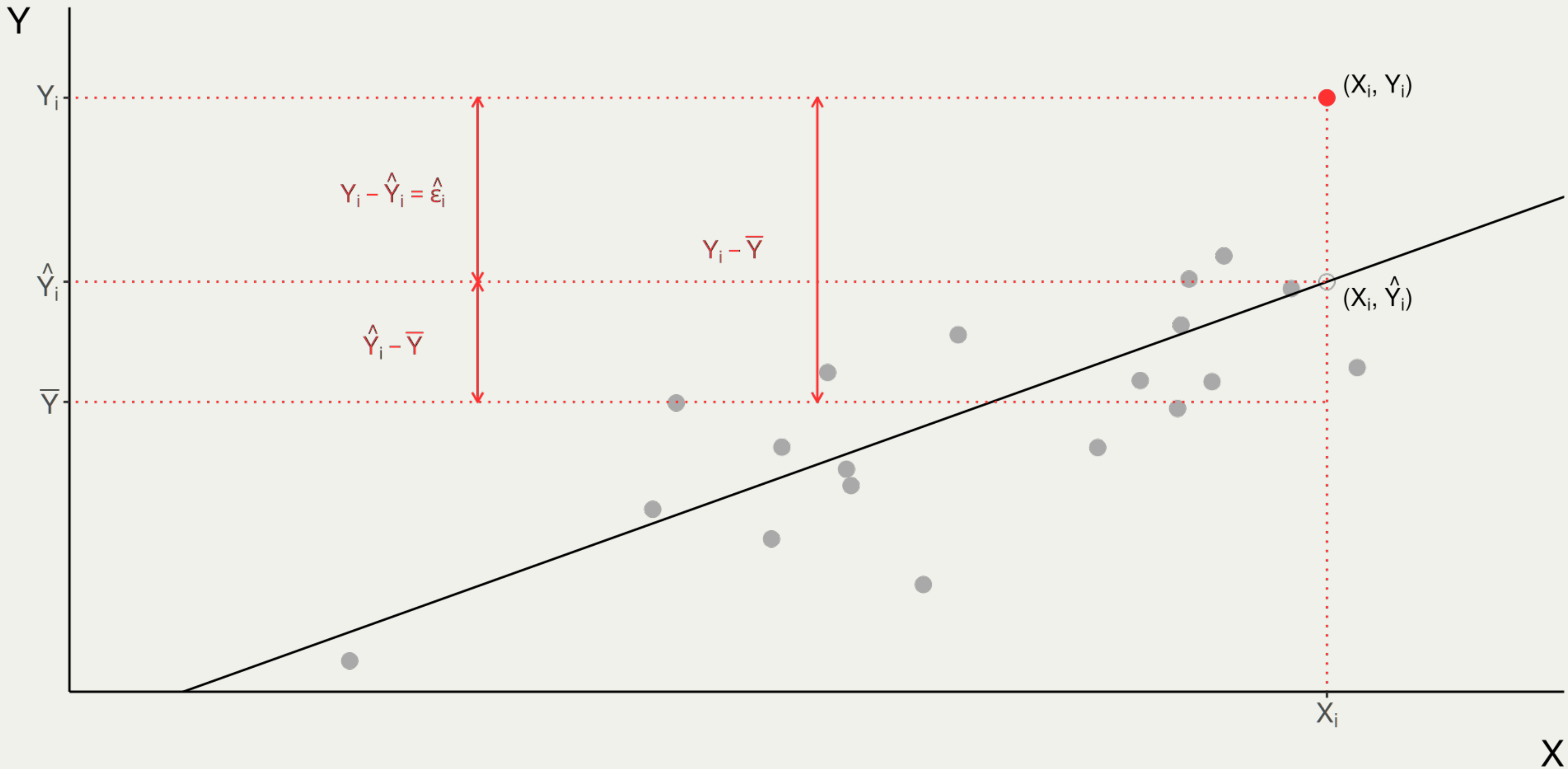
Department of Political Science, Trinity College Dublin

Topics for Today

- F-test
- Dummy Variables
- Panel Data
- Interactions

Previously…

Review: Variation in Y



Review: Total Variation in Y

- Decomposition of the total sum of squares of Y:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$TSS = SSE + ESS$$

where:

- Y_i - observed values of Y in the sample
- $\hat{Y}_i$ - fitted (predicted) values of Y_i given values of X_i in the sample
- $\bar{Y}$ - mean value of Y in the sample

Review: Model Fit

- R^2 (pronounced R-squared) - **coefficient of determination**.
- Provides a one number summary of model fit.
- Measure of the **proportional reduction in error** by the model.
- Proportional reduction relative to what?
  - Baseline prediction: $\bar{Y}$
  - Baseline prediction error: $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - Model prediction: $\hat{Y}_i$
  - Model prediction error: $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
  - $TSS - SSE$ - reduction in prediction error by the model

Review: Multiple Linear Regression Model

- We can express the population **multiple (multivariate) linear regression model** as:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

where:

- Observations $i = 1, \ldots, n$
- Y is the dependent variable
- $X_{1i}, \ldots, X_{ki}$ are k independent variables
- α is the **intercept** or **constant**
- $\beta_1, \ldots, \beta_k$ are **coefficients**
- ϵ_i is the **error term**

Hypothesis Testing for Multiple Coefficients

F-Test

- We have seen how t-test can be used to test the null hypothesis that the population coefficient of a single explanatory variable is 0.
- To test whether a set of coefficients for a set of explanatory variables are all zero, we need a different test:
- The F-test is used to test multiple-coefficient hypotheses where:
  - Null hypothesis: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ - in the population all coefficients $\beta_1, \beta_2, \ldots \beta_k$ are 0.
  - Alternative hypothesis: H_a : at least one of $\beta_1, \beta_2, \ldots \beta_k$ coefficients is not 0 in the population.

Nested Models

- Under an F-test, the null and alternative hypotheses specify two linear models for:

$$M_0 : Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{k-1} X_{k-1i} + \epsilon_i$$
$$M_a : Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{k-1} X_{k-1i} + \beta_k X_{ki} + \epsilon_i$$

- The model M_a thus contains all of the explanatory variables in M_0, as well as some additional one.
- The model M_0 is said to be nested in model M_a.

F-Statistic

- F test statistic for the null hypothesis is then:

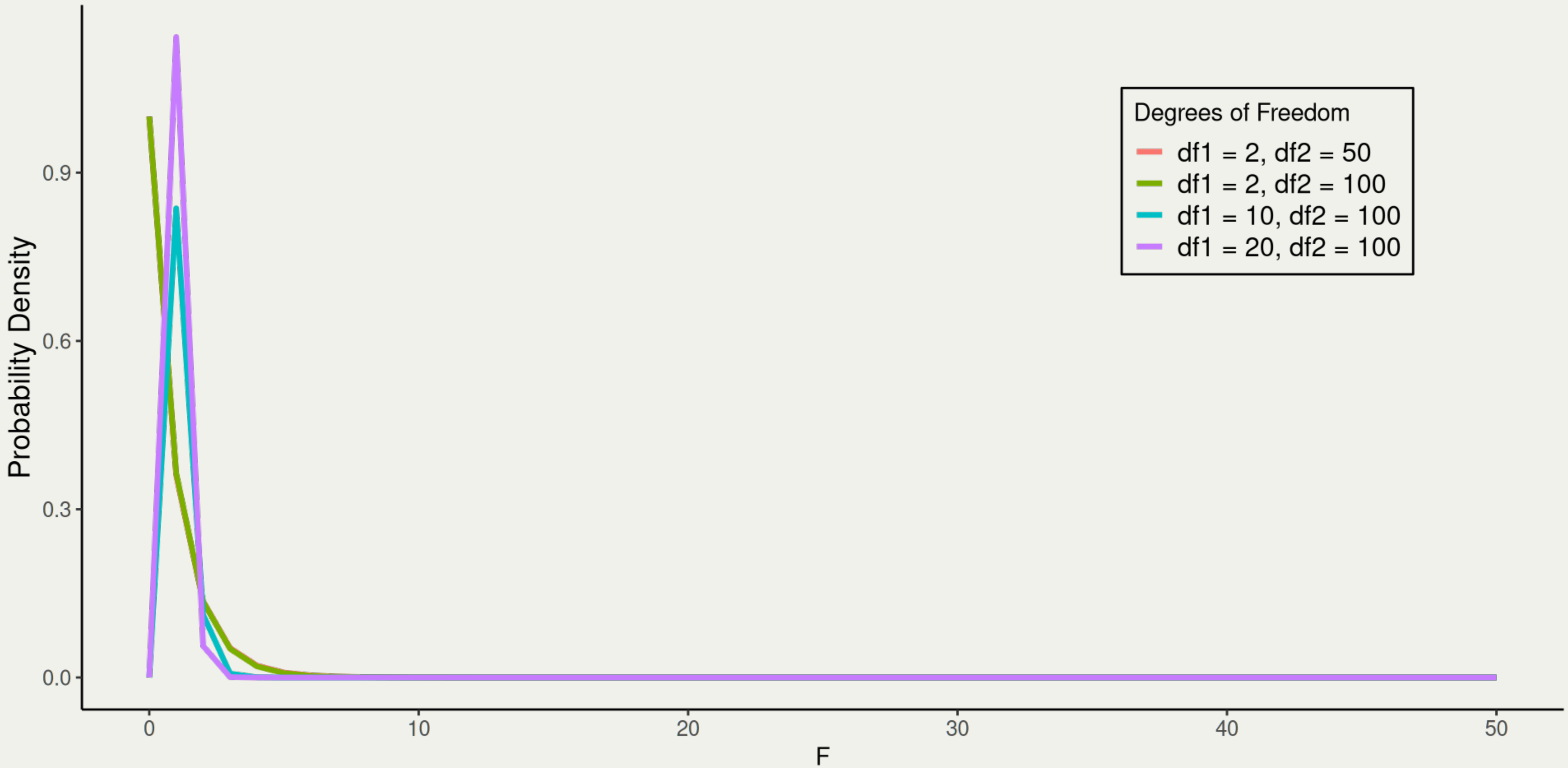
$$
\begin{aligned}
F &= \frac{(SSE_0 - SSE_a)/(k_a - k_0)}{SSE_a/(n - (k_a + 1))} \\
&= \frac{(R_a^2 - R_0^2)/(k_a - k_0)}{(1 - R_a^2)/(n - (k_a + 1))} \\
&= \frac{R^2_{change}/df_{change}}{(1 - R_a^2)/(n - (k_a + 1))}
\end{aligned}
$$

where:

- SSE_a and R_a^2 are SSE and R^2 for the full model M_a
- SSE_0 and R_0^2 are SSE and R^2 for the restricted model M_0.
- The total number of coefficients are $k_a = k$ and $k_0 = k - 1$ for M_a and M_0 respectively.

F Distribution

- The sampling distribution of the F-statistic under the null hypothesis is the F distribution with $k_a - k_0$ and $n - (k_a + 1)$ degrees of freedom.



Example: Democracy & Economy

- Let's start with a baseline model:

$$log(GDP)_i = \alpha + \beta_1 Democracy_i + \epsilon_i$$

```r
lm_fit_1 <- lm(log(gdp_per_capita) ~ democracy, data = democracy_gdp_2020)
summary(lm_fit_1)
```

```
Call:
lm(formula = log(gdp_per_capita) ~ democracy, data = democracy_gdp_2020)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9242 -0.8475 -0.1055  0.9133  3.0186 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.9801     0.1564  51.039  < 2e-16 ***
democracy     1.1000     0.1990   5.527 1.18e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.28 on 173 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1501,	Adjusted R-squared:  0.1452 
F-statistic: 30.54 on 1 and 173 DF,  p-value: 1.183e-07
```

Example: Democracy & Economy

- Adding an additional explanatory variable (log regime longevity) to our model:

$$log(GDP)_i = \alpha + \beta_1 Democracy_i + \beta_2 log(Longevity)_i + \epsilon_i$$

```r
lm_fit_2 <- lm(log(gdp_per_capita) ~ democracy + log(democracy_duration), data = democracy_gdp_2020)
summary(lm_fit_2)
```

```

Call:
lm(formula = log(gdp_per_capita) ~ democracy + log(democracy_duration),
    data = democracy_gdp_2020)

Residuals:
     Min       1Q   Median       3Q      Max
-2.85521 -0.87765 -0.07444  0.82037  3.10558

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.71571    0.34602  16.519  < 2e-16 ***
democracy                1.17576    0.17567   6.693 2.96e-10 ***
log(democracy_duration)  0.62745    0.08795   7.134 2.60e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.128 on 172 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.3441,	Adjusted R-squared:  0.3365
```

Example: F Statistic

$$
\begin{aligned}
F &= \frac{(R_a^2 - R_0^2)/(k_a - k_0)}{(1 - R_a^2)/(n - (k_a + 1))} \\
&= \frac{(0.344 - 0.15)/(2 - 1)}{(1 - 0.344)/(175 - (2 + 1))} \\
&= \frac{0.194}{0.004} = 50.87
\end{aligned}
$$

And the associated p-value:

```r
1 - pf(50.87, df1 = 1, df2 = 172)
```

```
[1] 2.625899e-11
```

Example: Nested Models in R

- We can do an explicit comparison of nested models in R using `anova()` command:

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

```
Analysis of Variance Table

Model 1: log(gdp_per_capita) ~ democracy
Model 2: log(gdp_per_capita) ~ democracy + log(democracy_duration)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    173 283.36
2    172 218.66  1      64.7 50.893 2.603e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-test of One Coefficient

- If M_0 has only one fewer coefficient than M_a, the F-test is identical to the t-test for that coefficient.
- The F-statistic itself is equal to the t-statistic squared:

$$F = t^2$$

- The p-values from $t_{n-(k+1)}$ sampling distribution and the $F_{1,n-(k+1)}$ sampling distributions are the same.

F-test of All Coefficients

- If M_0 has no explanatory variables at all, we get the null hypothesis that none of the explanatory variables $X1, \ldots, X_k$ are associated with the response.
- This model has $k_0 = 0$, $R_0^2 = 0$ and $SSE_0 = TSS$
- The F-statistic becomes:

$$F = \frac{R^2/k}{(1 - R^2)/(n - (k+1))} = \frac{ESS/k}{SSE/(n - (k+1))}$$

- This test is provided in R by default, but failing to reject the null is rare unless:
  - There are very few observations.
  - The explanatory variables are little more than random noise.

F-test in practice

- You must use the same data points for both models!
- The null will be rejected if there is evidence that *any* of the parameters are non-zero.
- Therefore, it is most useful for examining groups of variables that have some logical relationship to one another.

Categorical Predictors

Binary X

- We looked at incorporating *binary* X (independent variable) into linear regression model.
- *Binary* variables are nominal (categorical) variables that take only 2 values.
- Usually *binary* variable takes:
  - 1 when some attribute is present, and
  - 0 otherwise
- But nominal-scale variables can take more than 2 values.

Beyond Binary

- To include categorical predictors with multiple levels we use **dummy variables**.
- The first level of a factor variable is **reference** (or **baseline**) category.
- The choice of the reference category is arbitrary.
- It does not affect the substantive conclusions.
- Coefficient is then the difference between the each level and the reference category.

Example: Democracy & Education

> Does democratization increase education provision?

- The provision of primary education can have different purposes:
  - Redistribution;
  - Fostering regime loyalty;
  - Building national identity;
  - Expanding pool of educated labour, etc.
- Some of these goals might be favoured by democratic regimes.
- But some might be pursued equally by non-democracies.

Source

Paglayan (2021), Lee & Lee (2016) Boix, Miller and Rosato (2013), (2020)

Example: Democracy & Education

- First, let's examine a linear relationship between primary education provision and democracy.
- Our dependent variable Y is school enrollment rate (SER).
- SER is the percentage of the school-aged population who are enrolled in primary education.
- It ranges between 0 and 100.
- We use a binary indicator for democracy ($0 = Autocracy$, $1 = Democracy$).
- We obtain the data for countries between 1820 and 2010 (Paglayan 2021; Lee & Lee 2016)).
- In our model we also choose to control for the region where the country is located:

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

Example: Geopolitical Regions

- In this model our control variable (Z) is $region_i$.
- It is a nominal-scale (categorical) variable which has 6 levels.
- Specifically, $region_i$ takes the value:
  - 1 - Advanced economies
  - 2 - Asia and the Pacific ($Asia$)
  - 3 - Eastern Europe (EE)
  - 4 - Latin America and the Caribbean (LA)
  - 5 - Middle East and North Africa ($MENA$)
  - 6 - Sub-Saharan Africa

Example: Dummy 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
⋮	⋮	⋮	⋮	⋮	⋮

- The reference category is "Advanced" economies.
- R automatically converts factor variables into dummy variables.
- The first level of the factor variable is assumed to be the reference category.

Example: Geopolitical Regions in R

```r
table(paglayan2021$region)
```

```
          Advanced Economies            Asia and the Pacific 
                         936                             624 
              Eastern Europe Latin America and the Caribbean 
                         312                             975 
  Middle East and North Africa              Sub-Saharan Africa 
                         507                             897 
```

```r
paglayan2021$region <- as.factor(paglayan2021$region)
```

```r
table(as.numeric(paglayan2021$region))
```

```
  1   2   3   4   5   6 
936 624 312 975 507 897 
```

```r
levels(paglayan2021$region)
```

```
[1] "Advanced Economies"               "Asia and the Pacific"           
[3] "Eastern Europe"                   "Latin America and the Caribbean"
[5] "Middle East and North Africa"     "Sub-Saharan Africa"             
```

Example: Geopolitical Regions in R

```r
paglayan2021_2010 <- subset(paglayan2021, year == 2010)
paglayan2021_2010 <- paglayan2021_2010[order(paglayan2021_2010$country),]
```

```r
head(paglayan2021_2010[,c("country", "region")])
```

```
         country                          region
3549 Afghanistan            Asia and the Pacific
1716     Albania                  Eastern Europe
3081     Algeria    Middle East and North Africa
1014   Argentina Latin America and the Caribbean
4173   Australia              Advanced Economies
1521     Austria              Advanced Economies
```

```r
head(model.matrix( ~ region, data = paglayan2021_2010))
```

```
     (Intercept) regionAsia and the Pacific regionEastern Europe
3549           1                          1                    0
1716           1                          0                    1
3081           1                          0                    0
1014           1                          0                    0
4173           1                          0                    0
1521           1                          0                    0
     regionLatin America and the Caribbean regionMiddle East and North Africa
3549                                     0                                  0
1716                                     0                                  0
3081                                     0                                  1
1014                                     1                                  0
4173                                     0                                  0
1521                                     0                                  0
     regionSub-Saharan Africa
3549                        0
1716                        0
3081                        0
```

1014	0
4173	0

Example: Linear Regression with Categorical X

```r
lm_fit_3 <- lm(primary_ser ~ democracy + region, data = paglayan2021_2010)
summary(lm_fit_3)
```

```

Call:
lm(formula = primary_ser ~ democracy + region, data = paglayan2021_2010)

Residuals:
     Min       1Q   Median       3Q      Max 
-28.6254  -0.9387   0.8562   1.7045  10.9446 

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            96.0829     2.2650  42.421  < 2e-16 ***
democracy                               3.0921     1.7851   1.732  0.08636 .  
regionAsia and the Pacific             -0.8279     2.4412  -0.339  0.73524    
regionEastern Europe                   -0.8918     2.9448  -0.303  0.76264    
regionLatin America and the Caribbean  -0.8423     1.9634  -0.429  0.66884    
regionMiddle East and North Africa      2.4460     2.8361   0.862  0.39053    
regionSub-Saharan Africa               -7.0275     2.1680  -3.241  0.00162 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Panel Data

Panel Data on SER

- For the purposes of illustrating dummy variables I subset the data to observation for only one year: 2010.
- The original data includes an observation for each country for every 5-year period (quinquennial):

Country	Y_{SER}	X_{year}	X_{region}	$X_{democracy}$
Afghanistan	0	1820	2	0
Afghanistan	0	1825	2	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Albania	0.15	1820	3	NA
Albania	0.19	1825	3	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- What can we learn from this kind of data?

Pooled Regression

- As a first idea, we can just treat each country-year as a distinct data point.
- This is called a **pooled** model: we treat all the country-years into a common pool of data points.
- This specification would be, essentially, equivalent to the one we used before:

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

```r
lm_fit_4 <- lm(primary_ser ~ democracy + region, data = paglayan2021)
summary(lm_fit_4)
```

```

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
```

Beyond Pooled Regression

- Why might we want to not do pooled regression?
- We are assuming that the relationship between democracy and education is the same:
  - within countries across years
  - across countries within a year
- If we run take our pooled regression, and add dummy variables for each year…
- We can "hold year constant"
- It means that we are going to just estimate the relationship between democracy and education within each year, and then average that relationship over the years.

Fixed Effects

```r
lm_fit_5 <- lm(formula = primary_ser ~ democracy + region + as.factor(year), data = paglayan2021)
```

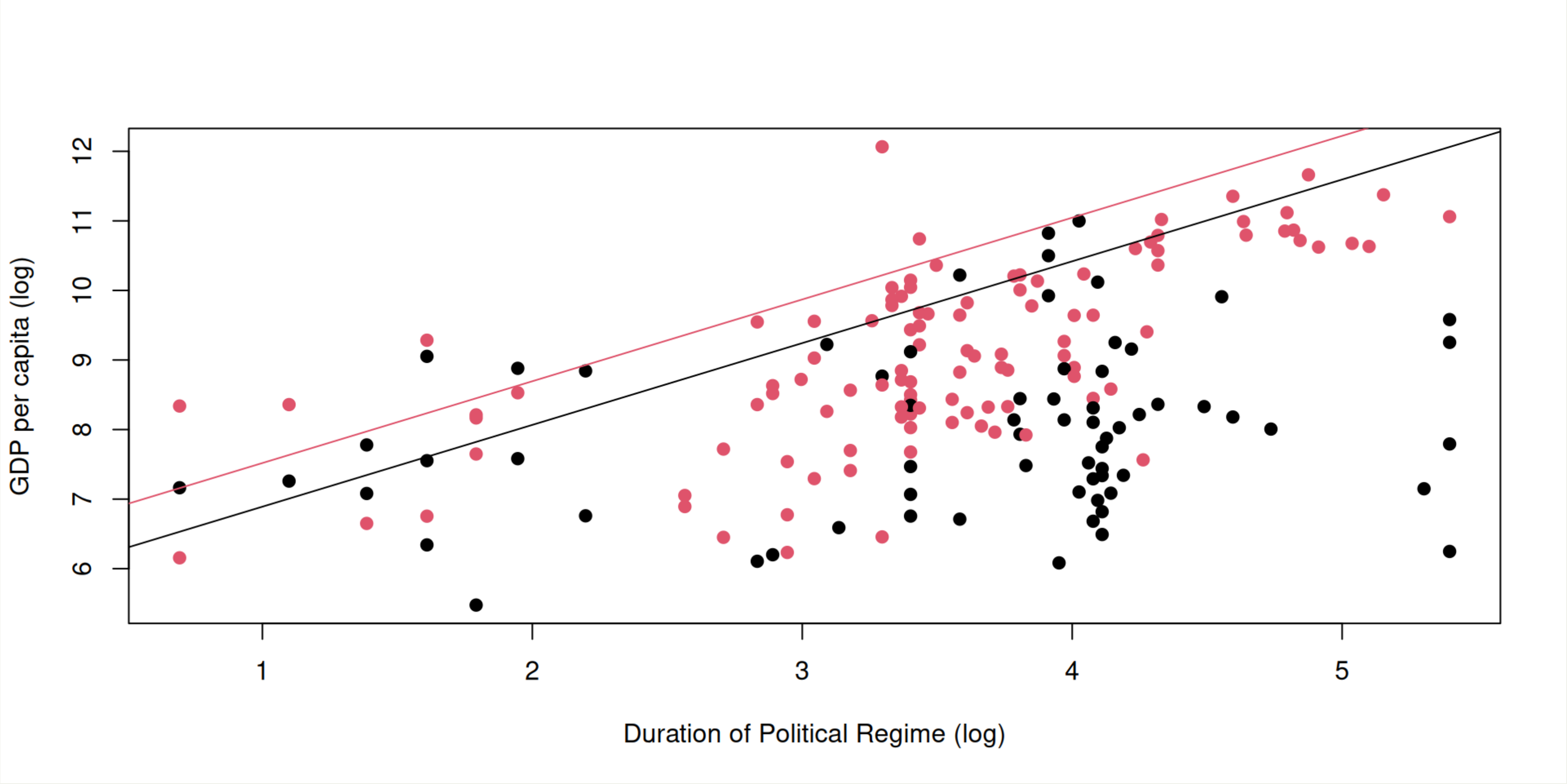
	Pooled	**FEs**
(Intercept)	51.123	30.215
	(1.351)	(4.211)
democracy	41.291	9.874
	(1.351)	(1.143)
regionAsia and the Pacific	−12.164	−34.053
	(2.053)	(1.510)
regionEastern Europe	9.928	−9.723
	(2.503)	(1.802)
regionLatin America and the Caribbean	−16.153	−28.349
	(1.567)	(1.128)
regionMiddle East and North Africa	1.326	−31.701
	(2.393)	(1.807)
regionSub-Saharan Africa	−3.063	−43.322
	(2.143)	(1.717)
Year	No	Yes
Num.Obs.	2496	2496
R2	0.371	0.696
R2 Adj.	0.370	0.690

Standard errors in parentheses

Interactions

Example: Democracy & Economy

- The previous model assumed that the relationship between regime longevity and log GDP per capita is the same in democracies and autocracies.



Interactions

- There is an **interaction** between two explanatory variables, if the relationship between (either) one of them and the response variable depends on the value of the other.

- We can build this intuition into the linear regression model by including the **product** of two explanatory variables in our model.

- We can test whether there is in fact evidence of an interaction by doing a hypothesis test on the coefficient on the product of the explanatory variables.

Interaction Term

- The simple model we have been studying assumes 'constant associations' (i.e. the relationship between X and Y does not depend on other X's).

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- We can relax the assumption of constant association by adding the product of explanatory variables to a model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

Example: Interacting Democracy & Longevity

- In the model we studied before:

$$log(GDP)_i = \alpha + \beta_1 Democracy_i + \beta_2 log(Longevity)_i + \epsilon_i$$

- Increasing regime longevity by 1 is associated with an increase in log GDP per capita by β_2 regardless of the regime type.

- Consider the model with interaction:

$$log(GDP)_i = \alpha + \beta_1 Democracy_i + \beta_2 log(Longevity)_i + \beta_3 Democracy_i \times log(Longevity)$$

- What happens now if we increase $log(Longevity)_i$ by 1?

Example: Interacting Democracy & Longevity

$$log(GDP)_i = \alpha + \beta_1 Democracy_i + \beta_2 log(Longevity)_i + \beta_3 Democracy_i \times log(Longevity)_i -$$

- $log(Longevity)_i$ increases by 1:
  - If $Democracy_i = 0$, in non-democracies, expected $log(GDP)_i$ increases by β_2
  - If $Democracy_i = 1$, in democracies, expected $log(GDP)_i$ increases by $\beta_2 + \beta_3$
- The increase in expected $log(GDP)_i$ when $log(Longevity)_i$ increases now depends on the value of another variable.

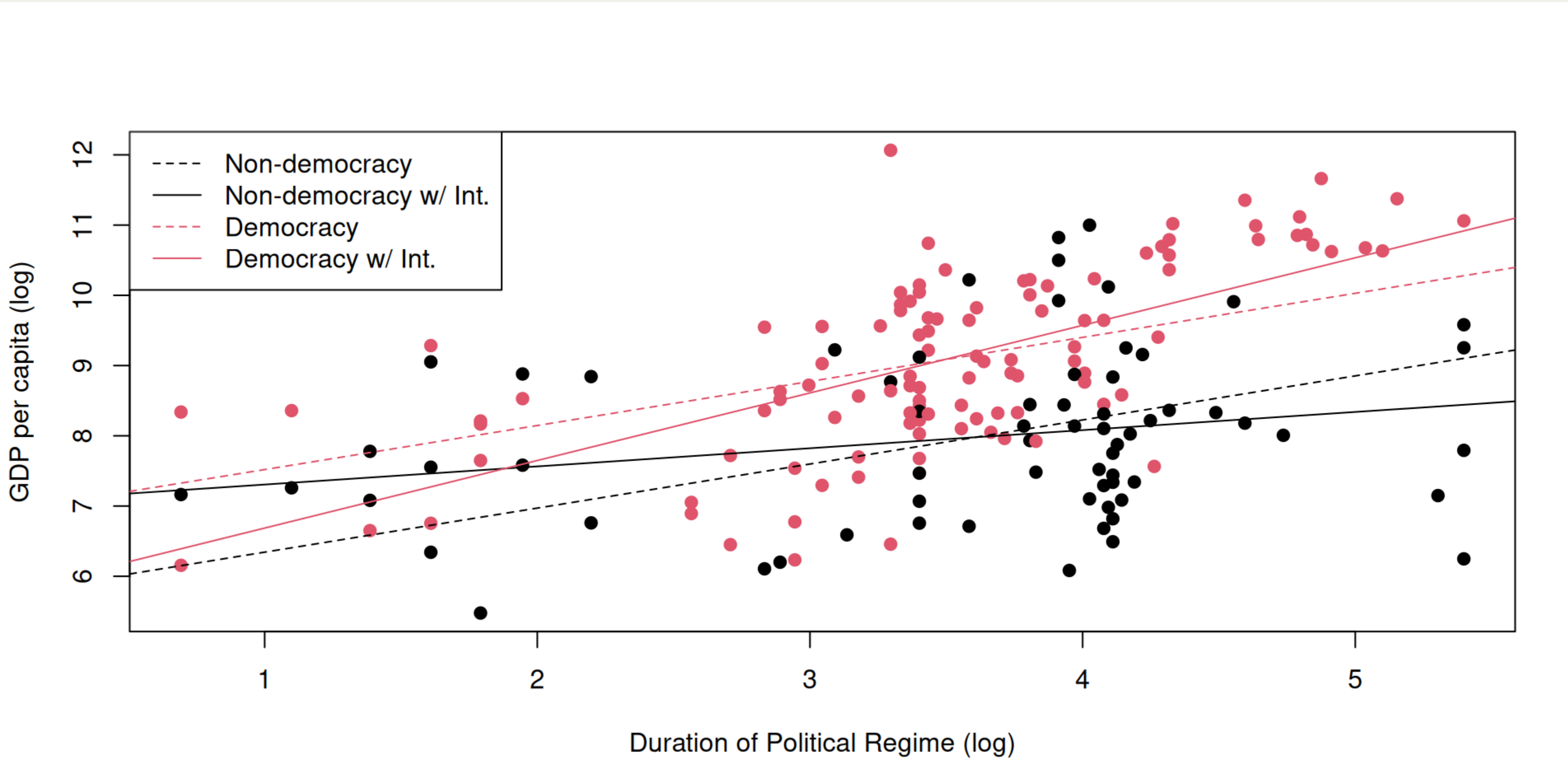
Example: Interactions in R

```r
lm_fit_6 <- lm(log(gdp_per_capita) ~ democracy * log(democracy_duration), data = democracy_gdp_2020)
summary(lm_fit_6)
```

```

Call:
lm(formula = log(gdp_per_capita) ~ democracy * log(democracy_duration),
    data = democracy_gdp_2020)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4386 -0.7792  0.0015  0.7448  3.1699 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         7.0479     0.4590  15.355  < 2e-16 ***
democracy                          -1.3234     0.6206  -2.133   0.0344 *  
log(democracy_duration)             0.2583     0.1218   2.120   0.0355 *  
democracy:log(democracy_duration)   0.7037     0.1682   4.183 4.59e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.077 on 171 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.405,	Adjusted R-squared:  0.3946 
F-statistic:  38.8 on 3 and 171 DF,  p-value: < 2.2e-16
```

Example: Plotting Interactions

Plot | Code



Testing for Interactions

- Similarly to other coefficients, we can test that $\beta_3 = 0$.
- In other words, we are testing whether the association between log regime longevity and log GDP per capita in different in democracies and non-democracies.
- The t-statistic for the interaction term is 4.183, which corresponds to a p-value of 0.0000459.

Regression Assumptions

Regression Assumptions

- The linear regression model makes several assumptions about the data:

  1. **Linearity**: The relationship between the response and explanatory variables is linear.
  2. **Independence**: The residuals are independent of each other.
  3. **Homoscedasticity**: The variance of the residuals is constant.
  4. **Normality**: The residuals are normally distributed.
  5. **No multicollinearity**: The explanatory variables are not too highly correlated with each other.

- We can check these assumptions using diagnostic plots.
- The most useful of such plots is the *residuals vs fitted values* graph.

Next

- Workshop:
  - RQ Presentations III
- Next week:
  - Causation