ECON 3040 - Interaction terms

Ryan T. Godwin

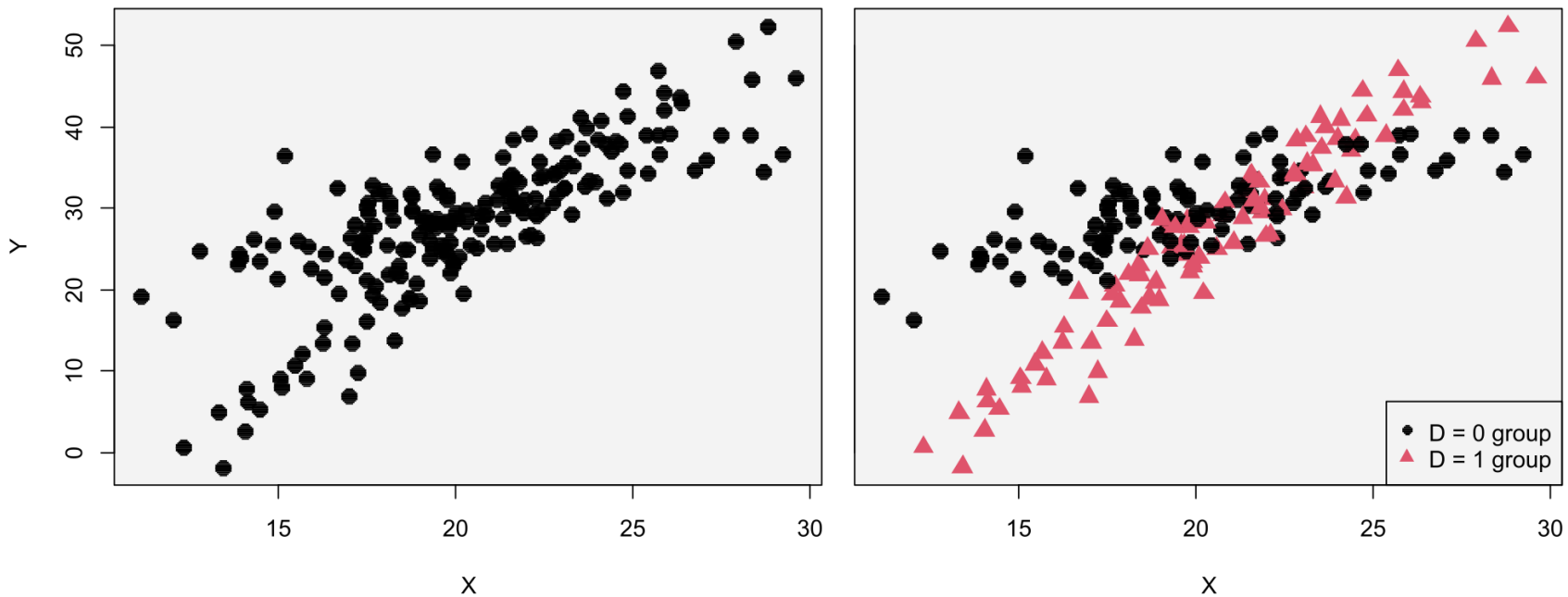
University of Manitoba

Types of interaction terms

- Interaction terms allow for a type of non-linear effect between variables.
- They are useful when the effect of X on Y may depend on a different X or D variable.
- The interaction term ($D \times X$) allows for a different linear effect between the two groups (the groups defined by D).
- Both of the variables in the interaction term can be dummy variables ($D_1 \times D_2$), or both of the variables in the interaction can be continuous ($X_1 \times X_2$), but the latter situation is somewhat rare and we do not discuss it here.

Simple example

Figure: Same data is plotted in both panels. In the right panel, we use a dummy variable D to colour code the data points, revealing that there are separate regression lines for each group.



Simple example

To illustrate the usefulness of interaction terms, we use a *fake* data set. The variables are:

- Y - the dependent variable
- X - an explanatory variable
- D - a dummy variable

The data is plotted above. Let's begin by estimating a simple model:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \epsilon \tag{1}$$

Simple example

In R we can use:

```
summary(lm(Y ~ X + D), data=mydata)
```

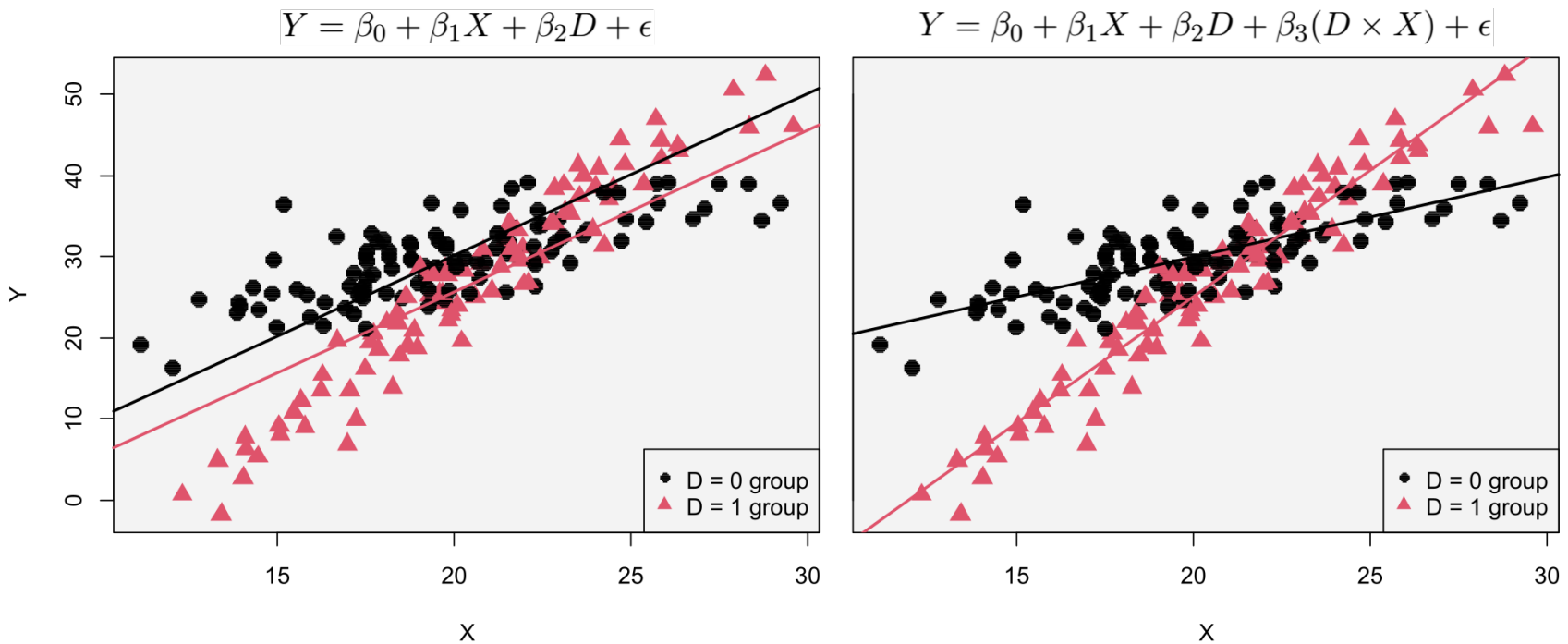
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.67535    2.00733  -4.820 2.86e-06 ***
X            1.99131    0.09807  20.304  < 2e-16 ***
D           -4.59618    0.72893  -6.305 1.85e-09 ***
```

Results:

- b_0 = -9.68. This is the intercept for the $D = 0$ group.
- b_1 = 1.99. An increase in X of 1 leads to an average increase in Y of 1.99. This is the marginal effect of X on Y.
- b_2 = -4.60. The $D = 1$ group Y values are 4.60 less than the $D = 0$ group, on average. The intercept shifts down by this amount for the $D = 1$ group, so that their intercept is $b_0 + b_2$ = -9.68 - 4.60 = -14.28.

Simple example

Figure: Left panel model (equation 1) uses a dummy variable, which allows for a different intercept for the two groups. Right panel model (equation 2) uses a dummy variable and an *interaction term*, which allows for a different intercept and *different slope*.



Simple example

The estimated model is shown in Figure 2 (left panel). The $D = 1$ group's regression line is 4.60 lower. We have two different regression lines for the two different groups, but they both have the same slope. We want them to have different slopes!

Dummy-continuous interaction

Ideally, we would like a separate regression line for the two groups, since the effect of X on Y may differ for the two. We need something new: an *interaction term*. This will allow for two separate marginal effects (slopes) for the two groups.

Dummy-continuous interaction term:

When X is a continuous variable and D is a dummy variable, $D \times X$ is a new variable called an *interaction term*. It allows for the effect of X on Y to differ between the two groups defined by the dummy.

Dummy-continuous interaction

Putting the *interaction term* into the model gives us:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (D \times X) + \epsilon \tag{2}$$

where $D \times X$ is the interaction term, and is a new variable that is created by multiplying the other two variables together. To see how model 2 allows for two separate lines, consider what the population model is for $D = 0$, and separately for $D = 1$.

Population model for $D = 0$

Let's substitute in the value $D = 0$ into equation 2 and get the population model for the first group:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2(0) + \beta_3(0 \times X) + \epsilon \\ &= \beta_0 + \beta_1 X + \epsilon \end{aligned} \tag{3}$$

From equation 3, we can see that the intercept is β_0 and the slope is β_1.

Population model for $D = 1$

Substituting in the value $D = 1$ into equation 2, we get the population model for the other group:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2(1) + \beta_3(1 \times X) + \epsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X + \epsilon \end{aligned} \tag{4}$$

For the $D = 1$ group, the intercept is $\beta_0 + \beta_2$ and the slope is $\beta_1 + \beta_3$. The marginal effect of X on Y differs by β_3 between the two groups.

R code for an interaction term

We can include the interaction term by adding the term `I(D * X)` to the `lm()` function:

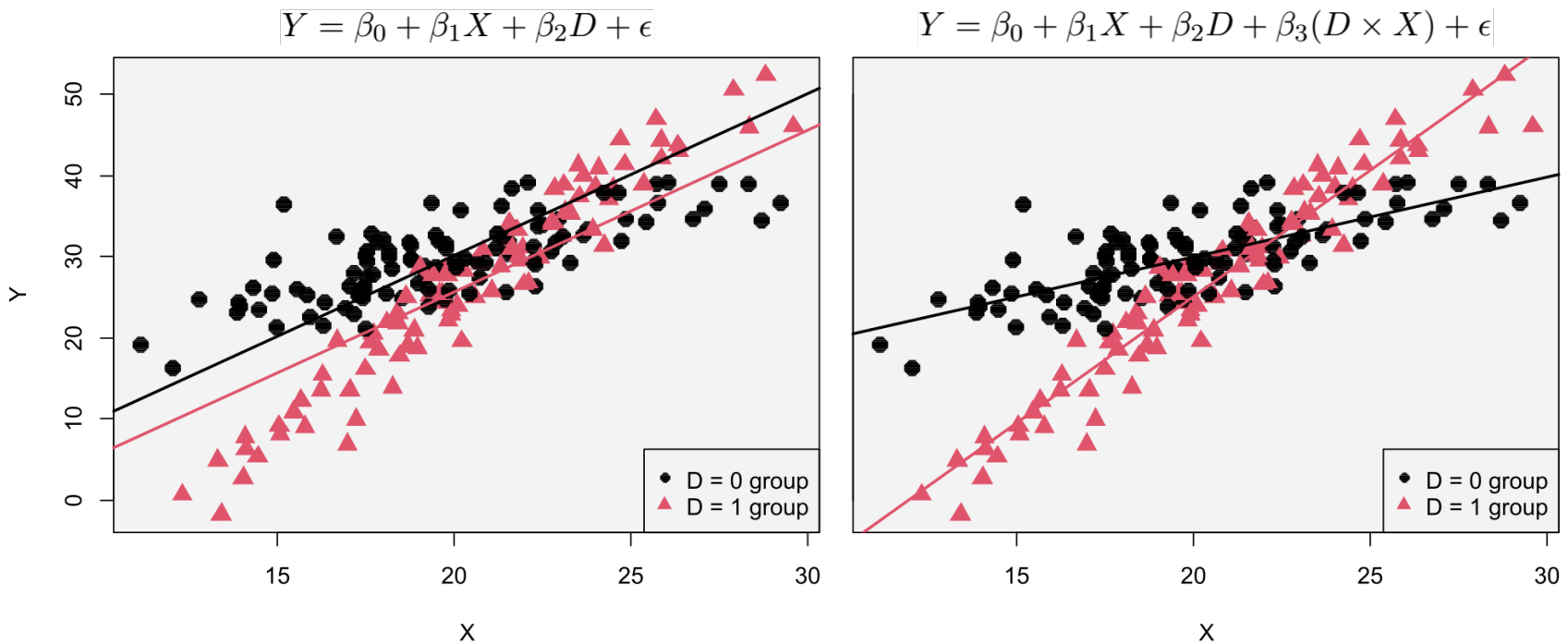
```
summary(lm(Y ~ X + D + I(D*X)), data=mydata)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.25251    1.73101   5.923  1.4e-08 ***
X             0.98663    0.08581  11.497  < 2e-16 ***
D           -47.61500    2.56503 -18.563  < 2e-16 ***
I(D * X)      2.13132    0.12499  17.052  < 2e-16 ***
```

The estimated value of $b_3 = 2.13$ means that the effect of X on Y (the slope) is 2.13 higher for the $D = 1$ group. That is, the effect of X on Y is 0.99 for $D = 0$, and $(0.99 + 2.13 = 3.12)$ for $D = 1$. The two different regression lines, with the two different slopes, are shown in the right panel of Figure 2.

Simple example

Figure: Left panel model (equation 1) uses a dummy variable, which allows for a different intercept for the two groups. Right panel model (equation 2) uses a dummy variable and an *interaction term*, which allows for a different intercept and *different slope*.



Example: land ruggedness and GDP

This example comes from "Ruggedness: The Blessing of Bad Geography in Africa", by Nunn and Puga (2012). The data is available from the authors here. The main variables in the study, for each of 170 countries, are:

- $\log(GDPpercap)$ - log real GDP per capita from 2000. This is the dependent variable, or y variable.
- $rugged$ - a Terrain Ruggedness Index that measures the amount of variation in the elevation of a country. It is a continuous variable. The higher the ruggedness, the more difficult the terrain is to traverse. This is the x variable.
- $Africa$ - a dummy variable equal to 1 if the country is in Africa. This is the D variable.

Example: land ruggedness and GDP

- Rugged terrain hinders trade and productive activities, so the higher the ruggedness of a country, the lower the GDP (a negative relationship between x and y).
- However, the authors argue that the relationship is opposite (positive) for African countries.
- The rationale is that rugged terrain offered protection from the slave trades.
- The slave trades hindered future economic development.
- For African countries, the higher the ruggedness, the higher the GDP.

Example: land ruggedness and GDP

The population model is:

$$\log(GDPpercap) = \beta_0 + \beta_1 rugged + \beta_2 Africa + \beta_3 (Africa \times rugged) + \epsilon$$

Download the data[1] and use `lm()` with an interaction term `I(cont_africa * rugged)`:

```
rug <- read.csv("https://rtgodwin.com/data/rugged.csv")
mod <- lm(log(rgdppc_2000) ~ rugged + cont_africa + I(
    cont_africa * rugged), data=rug)
summary(mod)
```

[1] As per Nunn and Puga (2012), the missing values for GDP were removed.

```
Coefficients:
                      Estimate Std.Error t value Pr(>|t|)
(Intercept)            9.22323    0.1396  66.044  < 2e-16 ***
rugged                -0.20286    0.0773  -2.621  0.00958 **
cont_africa           -1.94805    0.2272  -8.572 6.79e-15 ***
I(cont_africa*rugged)  0.39339    0.1316   2.989  0.00323 **
---
Signif.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9438 on 166 degrees of freedom
Multiple R-squared:  0.3569,  Adjusted R-squared:  0.3453
F-statistic: 30.71 on 3 and 166 DF,  p-value: 7.595e-16
```

All variables are significant. The estimate -0.20286 means that for every increase in a country's ruggedness of 1, GDP is 20.286% lower on average. But, African countries are *significantly* different. The variable `cont_africa * rugged` allows for the effect of ruggedness to be different between the two groups, and it is significant with a p-value of 0.00323. For African countries, an increase of ruggedness of 1 leads to an *increase* in GDP of $-20.286\% + 39.339\% = 19.053\%$.

Figure: Data is from Nunn and Puga (2012). Log real GDP per capita (from 2000) for 170 countries, and a measure of the ruggedness of the terrain in each country. A model with a dummy variable for African countries, and an interaction term with the dummy and ruggedness, is estimated. The interaction term allows for a different effect of difficult terrain on GDP, depending on whether the country is African or not.