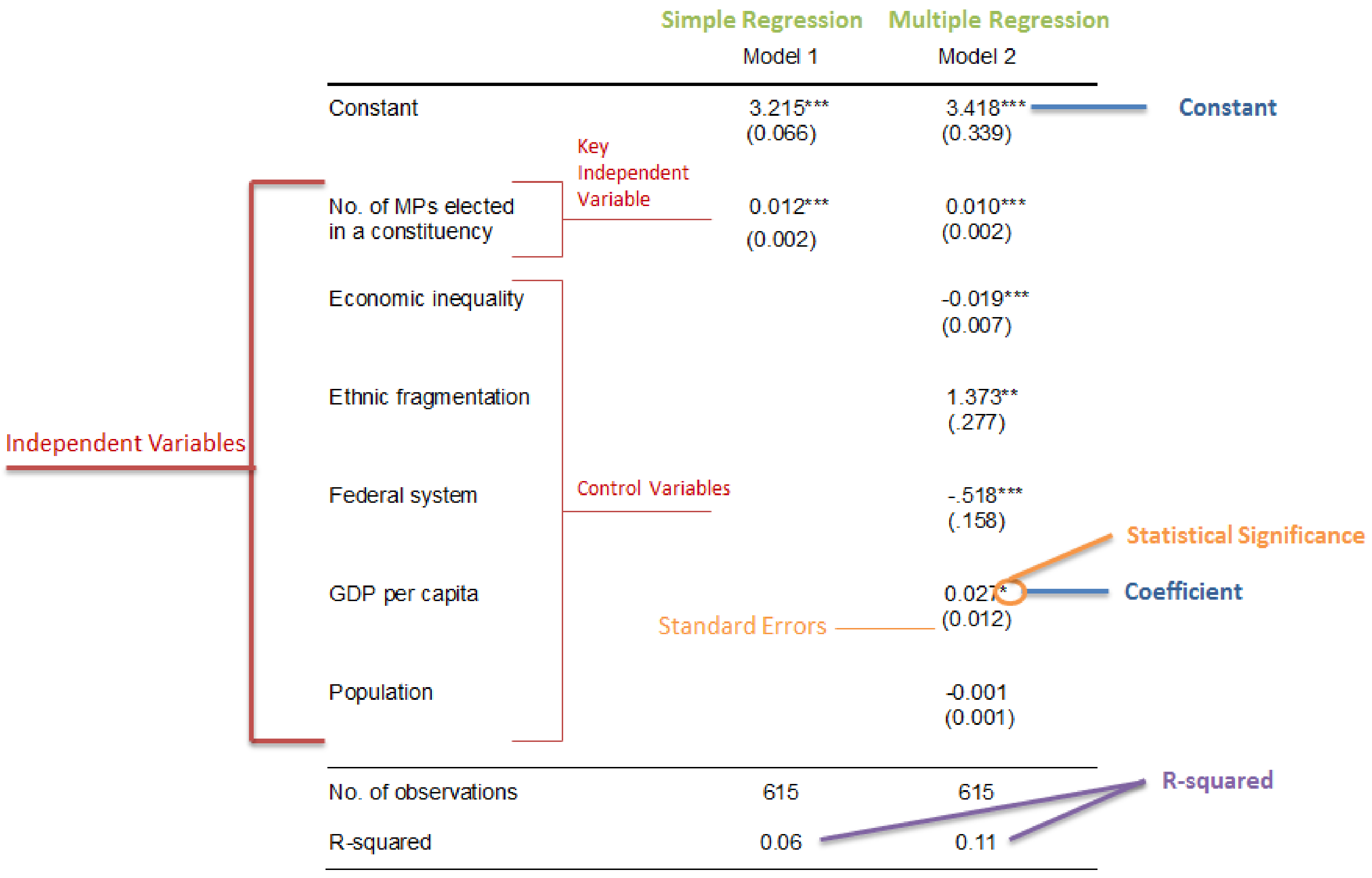
ANNOTATED REGRESSION OUTPUT:
How to understand and interpret regression tables?



	Model 1	Model 2
Constant	3.215*** (0.066)	3.418*** (0.339)
No. of MPs elected in a constituency	0.012*** (0.002)	0.010*** (0.002)
Economic inequality		-0.019*** (0.007)
Ethnic fragmentation		1.373** (.277)
Federal system		-.518*** (.158)
GDP per capita		0.027* (0.012)
Population		-0.001 (0.001)
No. of observations	615	615
R-squared	0.06	0.11

Simple Regression
Multiple Regression
Constant
Key Independent Variable
Independent Variables
Control Variables
Statistical Significance
Coefficient
Standard Errors
R-squared

Dependent variable = effective number of parties elected in a parliament
Observations = 615 elections in 82 democracies, 1945-2005

VARIABLES AND THEIR TYPES

The first column lists **independent variables (X)**, which are also frequently referred to as ‘explanatory variables’, ‘covariates’ and ‘predictor variables’. In our models, the number of MPs elected in a constituency, economic inequality, ethnic fragmentation, federal system, GDP per capita and population are all independent variables. The researchers included them in the regression models because there are theoretical reasons to believe that they can have an impact on the **dependent variable (Y),** which is the effective number of parties in parliament. Model 1 has only one independent variable (No. of MPs elected in a constituency), whereas Model 2 has six independent variables (No. of MPs elected in a constituency, economic inequality, ethnic fragmentation, federal system, GDP per capita and population). This means that Model 1 is a **simple regression model**, whereas Model 2 is a **multiple regression model**.

AIM OF REGRESSION ANALYSIS

The main goal of regression analysis is to estimate the effects of one or more independent variables on the dependent variable. Here, our key research interest is how the number of MPs elected in a constituency affects the number of parties in parliament. Number of MPs is therefore our **key independent variable**, i.e. the independent variable that is of particular interest. It is common for researchers, however, to include additional independent variables (**control variables**), which might also influence the dependent variable but are not of main interest to the researcher. In Model 2, economic inequality, ethnic fragmentation, federal system, GDP per capita and population are all control variables.

OMITTED VARIABLE BIAS

The main reason for including 'controls' in the model is to reduce the so-called **omitted variable bias**, which can occur when a model is incorrectly specified, i.e., when researchers leave out one or more explanatory variables that have an effect on both the dependent and independent variables. In our model, for example, we include a control variable 'Ethnic fragmentation' which – if unaccounted for – may confound the 'true' relationship between the number of MPs in a constituency (our key independent variable) and the effective number of parties (our dependent variable). Not only could high ethnic fragmentation lead to a higher number of parties in parliament (for example, there could be one party in parliament for each ethnic group in society), but highly fragmented societies might also adopt a specific type of electoral system (such as a proportional electoral system with high district magnitude). If these assumptions are correct and we omit ethnic fragmentation from the model, the results in our regression results table might be flawed, as we failed to account for an explanatory variable that has a very relevant impact on our dependent and another independent variable. It is therefore crucial that researchers include ethnic fragmentation in the model and show that the relationship between the key X and Y is not 'confounded' or 'spurious' but remains even after 'controlling' (accounting) for ethnic fragmentation. Not surprisingly, control variables are often called **confounding variables**. In our case, the estimates of the key X remain similar even after we control for ethnic fragmentation and three other potential confounders. We therefore can be reasonably confident that ethnic fragmentation (and GDP per capita, population, federal system and economic inequality) is not 'driving' (causing) the estimates in Model 1.

INTERPRETATION OF THE COEFFICIENTS

The **coefficients** are the estimates of the **magnitude** (size) of the effect of the Xs on the Y. They quantify how much Y changes when the independent variable increases by one unit while holding all other Xs constant. For example, we estimate in Model 2 that one unit increase in the number of MPs in a constituency leads to a 0.010 unit <u>increase</u> in the effective number of parties in parliament (Y) holding economic inequality, ethnic fragmentation, federal system, GDP per capita and population constant. Notice that the sign in front of the coefficient indicates the direction of the effect. While the number of MPs in a constituency has a positive effect on the number of parties, the federal system variable has a negative coefficient sign (**-**0.518) and therefore a negative effect. The coefficient of the federal system variable can be interpreted as follows: holding all other independent variables constant, having a federal state structure as opposed to not having a federal state structure leads to a 0.518 decrease in the effective number of parties. Please note that we phrase the interpretation of the federal system variable differently from the interpretation of the Number of MPs variable, as the latter is a continuous variable whereas the former is a dichotomous (aka dummy, taking a value of either 0 or 1) variable.

Please note as well that, whenever you interpret an independent variable out of a multiple regression model, you need to add ‘holding all other variables constant’ to your interpretation, to make clear that more than one variable has been included in the analysis.

CONSTANT

The **constant (intercept)** denotes the expected value of the dependent variable if we hold all independent variables at 0. Note, however, that this interpretation is not always meaningful. For example, a country’s population size is unlikely to be 0. It therefore might be better to refer to the constant as a number which tells you where the regression line (or plane) meets the Y-axis.

REGRESSION EQUATION

As we already mentioned above, the regression analysis helps us to understand how one or more variables (X) affect another (Y). More specifically, it helps the researcher to estimate how the Y changes when one X is varied (changes its value) while the other Xs are held constant (fixed). To this end, researchers estimate the following equations:

$Y = a + b * X_,$ which is the simple regression equation, used for simple regression models

and

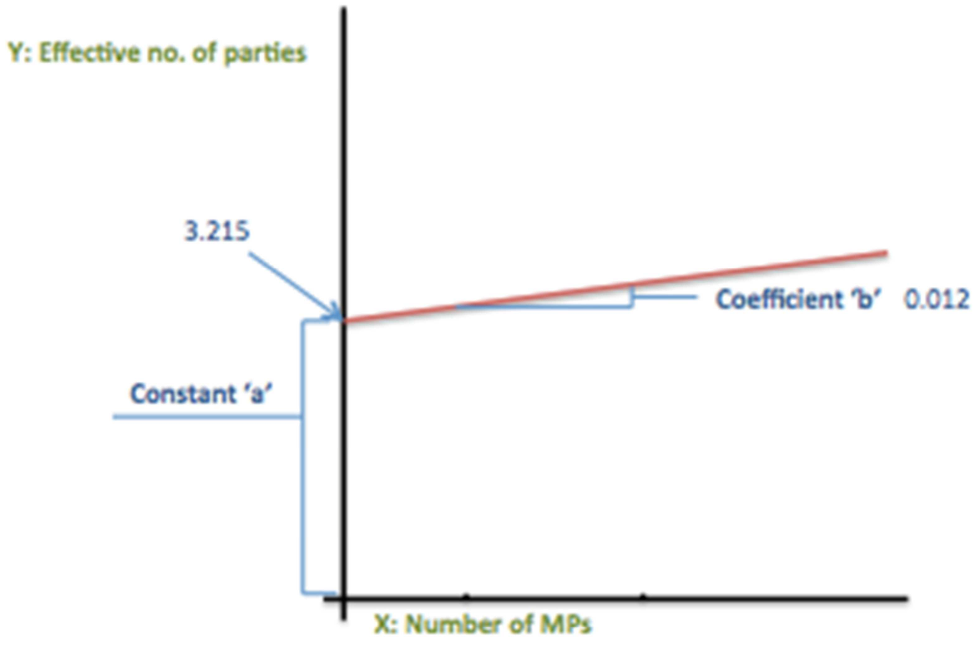
$Y = a + b_1 * X_1 + b_2 * X_2 + b_3 * X_3$ ..., which is the multiple regression equation, used for multiple regression models.

Y is the dependent variable, *X* denotes independent variables, *a* is the constant and *b* refers to the coefficients of the independent variables.

Using the results in Model 2, the above regression equation can therefore be re-written as:

*Effective number of parties = 3.418 + 0.010 * Number of MPs + (- 0.019) * Economic Inequality + 1.373*Ethnic Fragmentation + …*

The following graph illustrates the estimated relationship between our key X and the Y as estimated in Model 1:



As we increase the number of MPs in a constituency, the effective number of parties increases too. With 0 MPs in the district, the estimate of the effective number of parties equals 3.215. Note, however, that this would only be a meaningful interpretation if we had districts with 0 MPs, which, in real life, is highly unlikely. We therefore will leave the constant without interpretation. The graph also shows that a one unit increase in the number of MPs in a constituency (e.g. moving from 1 MP to 2 MPs) leads to a 0.012 unit increase in the effective number of parties. Since our model is linear, the 0.012 increase in our Y also occurs when we increase the number of MPs from 2 to 3, 3 to 4, 4 to 5 and so on.

STATISTICAL SIGNIFICANCE

While obtaining the estimates of the effect of X on Y is one of the main goals of regression analysis, it is equally important to estimate how sure the researcher can be that their results are genuine and not due to a 'luck.' For this purpose, researchers develop statistical tests and calculate how likely it is that the estimated coefficients are genuinely different from zero. The stars indicate the **level of statistical significance** for every variable in the model. If an X is statistically significant, it means that it has a relevant effect on the Y (it matters for the phenomenon we are trying to explain). If an X is not statistically significant, it means that it does not have a relevant effect on the Y (it does not matter for the phenomenon we are trying to explain). If there are no stars next to the coefficient (see the population variable), the independent variable is not statistically significant (despite the fact that the coefficient itself is different from zero). The more stars there are next to the coefficient, the higher is the level of statistical significance, i.e. the more confident we can be that the independent variable has a genuine effect on the dependent variable and, hence, that our estimate is not due to random chance. However, keep in mind that even highly significant result can be spurious due to omitted variable bias (see above). There are different ways of denoting statistical significance. While some researchers use letters and others symbols, the most common is a ‘star system’, where three stars indicate high significance and one star indicates borderline significance. Three stars indicate that the variable has a statistically significant effect at the 99% level, i.e. we can be confident at the 99% significance level that the independent variable has a relevant effect on the dependent variable. Two stars indicate the 95% level of statistical significance and one star indicates the 90% level of statistical significance.

STANDARD ERROR

Every coefficient has its own **standard error**, which is displayed in parentheses underneath the coefficient. The calculation of the standard error may be complicated and standard errors can be difficult to understand. It is best to think of the standard error as a measure of error in the calculation of every coefficient. A more advanced definition is that the standard error is an estimate of the standard deviation of the coefficients. The smaller the standard error in comparison to the coefficient size, the more confident we can be that there is a relevant relationship between the independent and dependent variable. In this context, note that standard errors which are large in comparison to the coefficient size will always result in low statistical significance or even no statistical significance at all. This is not very surprising given that researchers base their calculations of statistical significance on the ratio of the coefficient to its standard error.

R-SQUARED

Finally, researchers estimate the explanatory power of their models, i.e. how closely their models fit the data. **R-squared** therefore tells you how much of the variation in the dependent variable is explained by all the independent variables included in the model (as opposed to statistical significance which refers to the relationship between an X and the Y). R-squared values range between 0 and 1, and thus can be easily translated into percentages of explained variation of the dependent variable. Here, Model 1 explains 6% of the variation of the dependent variable, whereas Model 2 explains 11% of the variation of the dependent variable. The fact that we can explain more variation in our Y in Model 2 is not surprising given that Model 2 includes several additional statistically significant Xs, which clearly improved the overall ‘fit of the model’ to the data.