1 Multiple Regression

1.1 Motivation: Omitted Variables

Suppose you are interested in understanding how wages are related to years of education, so you look at the model

$$wage = \beta_1 + \beta_2 educ + \eta.$$

For now, think of η as being the typical disturbance term. The interpretation is that we want to explain *wage* with *educ* and "other stuff" captured in η.

Now ask yourself: of the "other stuff" in η that explains wage, is any of that also correlated with education? I am strongly inclined to say "yes." Take IQ for example. I would expect a higher IQ to explain a higher wage; but I also suspect that there is a correlation between IQ and years of education (e.g. college students have a higher IQ than the general public). So when we consider someone with more education, we are also likely considering someone with a higher IQ. This is problematic because β_2 in the regression above is implicitly telling us the effect of education *and* of IQ on wage, and therefore β_2 does not isolate the effect of education on wage.

In other words, *we are failing to hold IQ constant when considering different levels of education*, and consequently we are getting both the effect of higher education *and* the effect of higher IQ in our estimate of β_2. This relationship is illustrated in Figure 1.

That we fail to include a variable that is correlated with both the independent and dependent variable means our estimate for β_2 will be **biased**, that is, $E[b_2] \neq \beta_2$. We refer to this as **omitted variable bias**. Technically this is consequence of violating classical OLS assumption 2 (see below), i.e. zero conditional mean, because $E[v|educ] \neq 0$.

So how do we progress? Simple: just stick IQ into the regression as well. Our improved model is thus of the form

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \epsilon.$$

Now when we take the partial derivative with respect to education, we are explicitly holding IQ constant by definition of a partial derivative. Therefore

$$\frac{\partial wage}{\partial education} = \beta_2$$

gives the relationship between education and wage where IQ is being *controlled for*.

Of course, there are probably other omitted variables as well. In a laboratory exper-

iment, ideally all of these factors can be controlled for if the experiment is properly designed. But we are limited to the data we observe, which may or may not contain all relevant variables. (Probably won't.) Thus, even if we control for a bunch of variables, we still can never be certain that we have fully determined the direct relationship between any x and y.

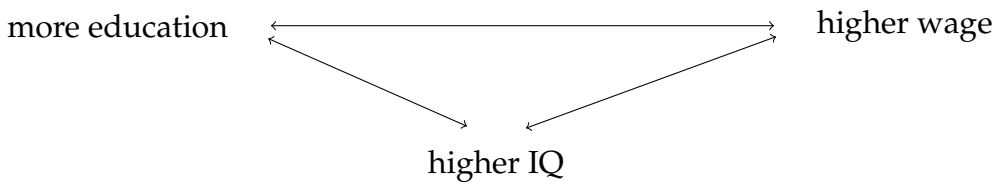

FIGURE 1: More education is correlated with higher wage, but it's also correlated with higher IQ. If we do not hold IQ constant, then we are not accurately characterizing the relationship between education and wage.

On the other hand, if an omitted regressor is correlated with y but not with x, then omitting it is fine. The omitted regressor is still part of ϵ because it is something that explains y but isn't in the regression. But because it isn't correlated with x, the zero conditional mean assumption $E[\epsilon|x] = 0$ still holds, and therefore estimates are still unbiased and consistent.

For instance, consider again $wage = \beta_1 + \beta_2 educ + \eta$. Tall people on average earn a higher wage than short people, so height is relevant in explaining *wage*: it's part of v, one of the "other things" that explain wage. But tall people are not on average more educated than short people, so height is not correlated with *educ*. In this case there is no omitted variable bias from omitting height: changes in education do not mean we are implicitly considering changes in height.

To so summarize:

- If a variable is relevant (it explains y) and is correlated with any included regressors, then it is a *confounding variable*: omitting it from the regression violates OLS assumption 2 and estimates suffer from the omitted variable bias and are inconsistent.

- If a variable is relevant (it explains y) and is not correlated with any included regressors, then omitting it from the regression is fine: estimates are unbiased and consistent.

1.2 Example: Wages

Import `wages.csv` into R. It contains, you guessed it, information about (monthly) wages, education, IQ, and some other stuff. If we regress wages on education, the result is

$$\widehat{wage} = 139.12 + 61.59 \times educ.$$

This implies that someone with one more year of education would be expected to have a higher monthly wage by \$61.59. But as discussed earlier, this is implicitly including the effect of a higher IQ, since the model above fails to control for IQ. We control for IQ by regressing wage on both education and IQ. By doing so, we expect the effect of education to be lower because now the effect isn't being exaggerated by a higher IQ. Indeed,

$$\widehat{wage} = -131.67 + 44.27educ + 4.95IQ.$$

So as predicted, the estimated effect of education on wage drops from 61.59 to 44.27. Before controlling for IQ, our estimate of β_2 had an *upward bias*.

The relevant R commands and output are shown in Figure 2 on the next page.

2 Classical OLS Assumptions

For OLS to "work" by default, we need the following conditions to hold given dependent variable y and the set of regressors $x_2, x_3, \ldots, x_k$. Note that we have $k-1$ regressors because we started at x_2. Therefore we will be estimating k things because we are also estimating the intercept coefficient. Hence we will have $n-k$ degrees of freedom when we do inference.

1. **CNLRM1: Correct Linear Model.** The true model is linear and correctly specified as

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon. \tag{1}$$

   Intuition: if we estimate a population model that's actually of a different form, then our estimates are probably garbage. (Keep in mind that any x_j might be nonlinear.)

2. **CNLRM2: Zero Conditional Mean.** The disturbance term has zero mean condi-

```
wages <- read.csv("wages.csv")
library(stargazer)

ols1 = lm(wage ~ educ, data = wages)
stargazer(ols1, type = "text")

ols2 = lm(wage ~ educ + iq, data = wages)
stargazer(ols2, type = "text")
```

```
> ols1 = lm(wage ~ educ, data = wages)
> stargazer(ols1, type = "text")

================================================
                        Dependent variable:
                    ----------------------------
                               wage
------------------------------------------------
educ                          61.586***
                               (5.931)

Constant                      139.117*
                              (81.160)
------------------------------------------------
Observations                    852
R2                             0.113
Adjusted R2                    0.112
Residual Std. Error      380.658 (df = 850)
F Statistic          107.817*** (df = 1; 850)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01


> ols2 = lm(wage ~ educ + iq, data = wages)
> stargazer(ols2, type = "text")

================================================
                        Dependent variable:
                    ----------------------------
                               wage
------------------------------------------------
educ                          44.268***
                               (6.852)

iq                            4.954***
                               (1.019)

Constant                      -131.671
                              (97.555)
------------------------------------------------
Observations                    852
R2                             0.137
Adjusted R2                    0.135
Residual Std. Error      375.686 (df = 849)
F Statistic          67.168*** (df = 2; 849)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

FIGURE 2: When IQ is included in the regression (and therefore controlled for), we find that education explains less about wage.

tional upon the regressors, that is,

$$E[\epsilon|x_2, \ldots, x_k] = 0. \tag{2}$$

Intuition: think of the disturbance term as being the mistake of the model. If we expect the mistake to be non-zero on average, then our model is probably garbage. This condition is equivalent to saying that u is uncorrelated with all of the regressors.

More technically, it allows us to go from

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon$$

$$E[y|x_2, \ldots, x_k] = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k, \tag{3}$$

the latter being the interpretation of the regression line itself (i.e. the conditional expectation of y given our regressors).

3. **CNLRM3: Homoskedasticity.** The conditional variance of the disturbance term is constant and finite, that is,

$$\text{Var}(\epsilon|x_2, \ldots, x_k) = \sigma_\epsilon^2 < \infty. \tag{4}$$

There isn't much economic intuition here; it's mostly a technical assumption, albeit an unrealistic one, that offers a convenient starting point for rigorous analysis. In practice it is violated frequently, which is not difficult to deal with (as explained later). This condition is illustrated in Figure 3.

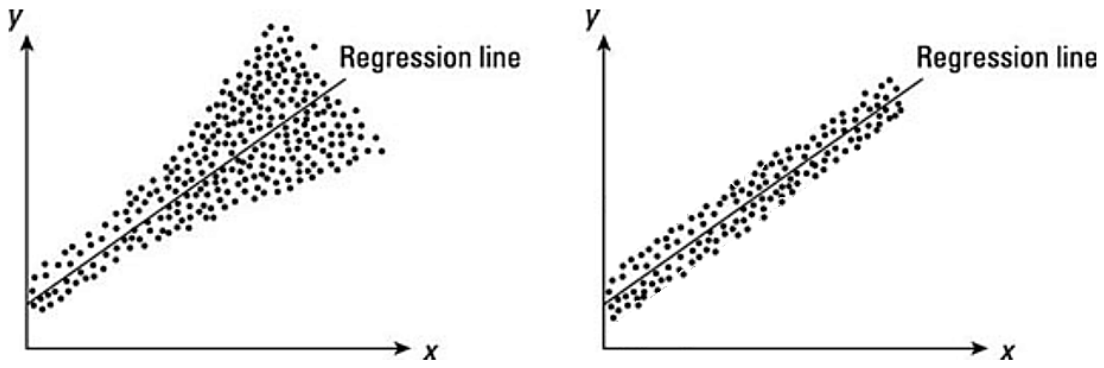

FIGURE 3: The figure on the left is an example of heteroskedasticity; the right an example of homoskedasticity. The left is heteroskedastic because the variation around the regression line gets bigger as x increases. Good luck envisioning this in higher dimensions.

4. **CNLRM4: Independent Errors.** Errors for different observations are statistically independent, that is,

$$\epsilon_i \perp \epsilon_j \quad \text{whenever } i \neq j.$$

Intuition: if model disturbances are correlated, then there is some underlying pattern that we are overlooking, so our results are probably garbage.

As an example of a violation, suppose we look at ECN 102 final exam scores in all of 2017; that means we're looking at ECN 102 final exam scores for three different professors. Problem is, different professors write exams of differing difficulty. Hence we would expect a lenient professor's students to do better than the regression predicts (so we'd have correlation among observations with positive ϵ), and we expect a challenging professor's students to do worse than what the regression predicts (so we'd have correlation among observations with negative ϵ). This is called **clustering** because each professor's final exam forms a cluster of students. (Note that students in different clusters, however, are independent from each other.)

5. **CNLRM5: Normality of Errors.** Errors are normally distributed with variance σ^2, i.e.,

$$\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right). \tag{5}$$

This is another technical assumption for "nice" results, explained below. In practice it can be weakened, but it is necessary for inference on small sample sizes.

6. **CNLRM6: No Perfect Multicollinearity.** There exists no exact linear relationship between explanatory samples. Furthermore, the number of observations must be greater than the number of explanatory variables (plus constant term), i.e. $n \geq k$.

Intuition: if there is such a perfect relationship between two or more regressors, then we can't "untangle" the effect of each regressor. In other words, it's like including the same regressor twice, and that redundancy breaks the OLS solution technique.

3 Implications of OLS Assumptions

You can see that most of these assumptions are close analogues to the simple regression, the exception being CNLRM6. You will not be surprised then to learn that the implications are largely the same as well.

- Assumptions CNLRM1-2 imply that OLS estimates are unbiased, so that $E[b_j] = \beta_j$.

- Assumptions CNLRM1-4 imply that OLS estimates are consistent, so that $b_j \xrightarrow{p} \beta_j$ as $n \to \infty$. Furthermore, assumptions CNLRM1-4 imply that OLS is the **b**est **l**inear **u**nbiased **e**stimator, or **BLUE**. When we say "best," we mean they have the smallest standard errors and hence precision of inference is the most accurate.

- Adding CNLRM5 implies that OLS is the **b**est **u**nbiased **e**stimator, or **BUE**, even when compared to nonlinear methods. Furthermore, it implies that

$$t \equiv \frac{b_j - \beta_j}{\text{se}(b_j)} \sim T(n-k)$$

  is exactly true for any β_j, even for small samples; without CNLRM5 it is only approximately true if the sample size is large enough. (Therefore CNLRM5 is required for inference on small samples.) We are estimating k things, which is why we have $n-k$ degrees of freedom.

- Assumption CNLRM6 is always required; in the presence of perfect multicollinearity, the regression cannot be executed. Accordingly, this is usually just implicitly assumed because otherwise it's game over and we should just give up and go home. (Actually, there's usually a very easy fix for it, shown in a bit.)

4 Including Irrelevant Regressors

Suppose we accidentally include some regressor x_ℓ that does not explain y at all, thereby making it irrelevant. Well, because it's irrelevant, its coefficient will be $\beta_\ell = 0$, and therefore the population regressions

$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon,$$
$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \beta_\ell x_\ell + \epsilon,$$

are actually identical. So nothing is violated by including the irrelevant regressor: the results are still unbiased and consistent, provided OLS assumptions 1 and 2 hold for everything else. The problem is that OLS still has to try to estimate β_ℓ if x_ℓ is included, which is just adding noise to the estimation process. This means the regression will be less precise (i.e. higher standard errors). But you're usually better off with less precise estimates than biased ones, so most researchers err on the side of including regressors that might be irrelevant.

5 Multiple Regression Inference

Under CNLRM1-4, the t-statistic regarding regressor x_j is given by

$$t = \frac{b_j - \beta_j}{\text{se}(b_j)}, \tag{6}$$

and it is drawn from an approximate $T(n-k)$ distribution.Inference proceeds in the usual way. There is no rule of thumb for how large n needs to be for the approximation to be adequate. If CNLRM5 holds, then t is drawn from exact $T(n-k)$ distribution.

If either CNLRM3 or CNLRM4 fail, then the typical standard errors are not valid. We can oftentimes use one of the following alternatives, however.

- use **heteroskedasticity-robust standard errors** if only CNLRM3 fails
- use **cluster-robust standard errors** if CNLRM4 alone fails because of suspected clustering in variable x
- use **heteroskedasticity and autocorrelation-consistent (HAC) standard errors** if using time series data.

We will have tests for these, but hold onto that thought for another day.

6 Coefficients of Determination

6.1 Overall Significance

When we did a bivariate regression with just one regressor, we asked: does x actually explain anything about y? In other words, we tested $H_0 : R^2 = 0$ against $H_1 : R^2 > 0$ using an F-test.

But now we have multiple regressors, all of which contribute to explanatory power of the model. Which is to say, when we test $H_0 : R^2 = 0$ against $H_1 : R^2 > 0$, we're actually testing the *combined explanatory power of all regressors in the model*. This is called a **test of overall significance**: we want to determine whether our entire suite of regressors jointly explain something, *anything*, about y.

Let's be a bit more concrete. Suppose the regression is $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$. We want to test whether the combined explanatory power of x_2 and x_3 and x_4 is zero or

non-zero. This test of overall significance has hypotheses

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \qquad (H_0 : R^2 = 0)$$
$$H_1 : \text{at least one of } \beta_2, \beta_3, \beta_4 \neq 0. \qquad (H_1 : R^2 > 0)$$

Ergo if we reject the null hypothesis, then we conclude that the regression is significant overall: the combination of regressors has at least some explanatory power.

The good news is, we use the same F-statistic as before, given by

$$F \equiv \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(k-1, n-k), \tag{7}$$

where k is the number of coefficients being estimated. In this example, $k = 4$ because the model estimates β_1 through β_4. Do note that this F-statistic is only valid when disturbances are homoskedastic (i.e. when CNLRM3 holds); otherwise we'll have to use a *heteroskedasticity-robust* version, which is more difficult than anything we'd calculate in this class (it requires matrix algebra).

6.2 Adjusted R-Squared

A problem with the default R-squared formula is that it *always* increases when you add an additional explanatory variable—even if that explanatory variable is totally irrelevant—due to statistical noise. This is not a desirable property: you don't want to use a measurement that gives an impression of having additional explanatory power when you add a variable that doesn't actually explain anything.

This is where the **adjusted R-squared**, denoted $\overline{R}^2$, comes into play. It adds a "penalty" every time an additional regressor is added. If the increase in R^2 is larger than the penalty, then $\overline{R}^2$ will increase and we suspect that the explanatory power of the model has improved. If the penalty is larger than the increase in R^2, then $\overline{R}^2$ will decrease and we suspect that the explanatory power of the model has not improved and therefore the new regressor should probably be omitted. (We will test this more formally later.)

The formula for adjusted R-squared is given by

$$\overline{R}^2 = 1 - \frac{\text{RSS}}{\text{TSS}}\frac{(n-1)}{(n-k)} = 1 - \frac{s_e^2}{s_y^2}, \tag{8}$$

although a more intuitive formulation is

$$\overline{R}^2 = R^2 - \frac{k-1}{n-k}(1-R^2), \tag{9}$$

where the second term is the penalty term. If a new regressor is added, then R^2 will go up, but so will the penalty term $(k-1)/(n-k)$. Whichever change is larger determines whether $\overline{R}^2$ goes up or down overall.

7 Dummy Variables

7.1 Definition of Dummy Variable

We might be interested in seeing how different categories affect the dependent variable. For instance, we might want to see if someone working in an urban environment earns more than someone working elsewhere. To analyze, we construct a **dummy variable** that is equal to either zero or one. An urban worker would have value $urban = 1$, and a non-urban worker would have value $urban = 0$. Accordingly, we would run the regression

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 urban + \epsilon.$$

The coefficient β_4, then, would tell you the expected difference in monthly wage for an urban worker compared to a non-urban worker. Another way of thinking about it is, β_4 captures the expected change in wage if a worker moves from a non-urban environment to an urban environment, that is, if *urban* changes from 0 to 1.

7.2 Dummy Variable Trap

Notice in the preceding example that there are two categories, but only one dummy variable. In general, if you have m categories, then you must include exactly $m-1$ dummy variables; the category you omit is called the **reference category**. Including dummy variables for all categories results in the **dummy variable trap**, which is a source of perfect multicollinearity that breaks OLS estimation. So always use one fewer dummy than there are categories (or drop the intercept; this is less common).

Here's a silly example to illustrate why things go wrong. People become really loyal to stupid things that don't matter, for example, which brand of cola they drink.[1] They

[1] Blind taste test? People can't tell the difference.

either drink Coke and only Coke; or Pepsi and only Pepsi; or, for the purposes of this example, RC Cola and only RC Cola.[2]

We want to see how many cavities people get from drinking a beverage that is used to remove rust from nails. We record their preference in the following manner:

$$\textit{choice} = 1 \text{ if Coke}, \quad \textit{choice} = 2 \text{ if Pepsi}, \quad \textit{choice} = 3 \text{ if RC Cola}.$$

Now define dummies for all categories. Let $d_1 = 1$ for choosing Coke; $d_2 = 1$ for choosing Pepsi; and $d_3 = 1$ for choosing RC Cola. Then the possible values for each dummy are

$$\begin{aligned} \textit{choice} = 1 \quad &\implies \quad d_1 = 1,\ d_2 = 0,\ d_3 = 0, \\ \textit{choice} = 2 \quad &\implies \quad d_1 = 0,\ d_2 = 1,\ d_3 = 0, \\ \textit{choice} = 3 \quad &\implies \quad d_1 = 0,\ d_2 = 0,\ d_3 = 1. \end{aligned}$$

Notice that in all three cases, $d_1 + d_2 + d_3 = 1$. And therefore, say, $d_1 = 1 - d_2 - d_3$. This is perfect multicollinearity because one of our regressors (d_1) can be perfectly explained by a linear relationship of two other regressors (d_2 and d_3). So if we try to regress *cavities* on d_1, d_2, and d_3, then OLS explodes and we're all doomed.

Except you can just remove any one of the three dummies from the regression, then all is well and well is all for all. The coefficients of the model are then seen as being *relative* to the reference category. To that end, consider the model where we omit the Coke dummy variable d_1, given by

$$\textit{cavities} = \beta_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon.$$

Let us interpret each coefficient.

- β_1: how many cavities are associated with being a Coke drinker (reference category);
- β_2: how many more (or less, if negative) cavities are associated with being a Pepsi drinker instead of a Coke drinker;
- β_3: how many more (or less, if negative) cavities are associated with being an RC Cola drinker instead of a Coke drinker.

In the case of the urban workers, β_4 captures how much higher of a wage a person receives if they work in an urban environment relative to working in a non-urban environment (the reference category).

[2]No one actually drinks RC Cola, do they?

7.3 Example: Wages

Again using `wages.csv`, let us consider the regression proposed earlier,

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 urban + \epsilon.$$

OLS estimation yields

$$\widehat{wage} = -213.28 + 41.58 educ + 4.92 IQ + 169.01 urban.$$

As shown in the R output below in Figure 4, the p-value for β_4 indicates statistically significance, so we conclude that an urban worker is expected to earn a monthly wage \$169.01 higher than that of a non-urban worker.

```
wages <- read.csv("wages.csv")
library(stargazer)

ols3 = lm(wage ~ educ + iq + urban, data = wages)
stargazer(ols3, type = "text")
```

```
> ols3 = lm(wage ~ educ + iq + urban, data = wages)
> stargazer(ols3, type = "text")

================================================
                        Dependent variable:
                    ---------------------------
                               wage
------------------------------------------------
educ                          41.581***
                               (6.729)

iq                            4.920***
                               (0.998)

urban                        169.014***
                              (28.133)

Constant                     -213.282**
                              (96.559)

------------------------------------------------
Observations                    852
R2                             0.172
Adjusted R2                    0.169
Residual Std. Error      368.154 (df = 848)
F Statistic           58.661*** (df = 3; 848)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

FIGURE 4: The p-value for β_4 (urban) indicates statistically significance at 1%, so we conclude that an urban worker is expected to earn a monthly wage \$169.01 higher than that of a non-urban worker.

8 Interactions

8.1 Marginal Effects

When we have multiple regressors, we might be interested in how they, um, interact with each other when it comes to explaining the dependent variable. A regression with interactions will look something like

$$y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 xz + \epsilon \quad \Longrightarrow \quad \widehat{y} = b_1 + b_2 x + b_3 z + b_4 xz,$$

where xz is the interaction term. The idea is that x might affect y differently depending on what value z is, and vice versa. That is, the **marginal effect** of x on $\widehat{y}$ is given by

$$\frac{d\widehat{y}}{dx} = b_2 + b_4 z.$$

When we consider marginally larger x, we expect y to be marginally different by $b_2 + b_4 z$.

8.2 Example: Foreign Aid and Dictatorships

You might be interested in how foreign aid affects education funding in undeveloped countries, so you run the regression

$$educ = \beta_1 + \beta_2 aid + \epsilon.$$

The coefficient β_2 tells you the association between an additional dollar of foreign aid received and education funding for the average undeveloped country; the marginal effect is constant: one more dollar of foreign aid is associated with $\$\beta_2$ more education funding.

We suspect, however, that the effect of foreign aid is different depending whether the undeveloped country is democratic or ruled by a dictator. Introduce the dummy variable $dictator = 0$ for democracy and $dictator = 1$ for dictatorship and run the regression

$$educ = \beta_1 + \beta_2 aid + \beta_3 (aid \times dictator) + \epsilon.$$

In this formulation, the effect of foreign aid depends on the value of *dictator* (i.e. the interaction of regressors). The marginal effect of foreign aid on education funding is

$$\frac{\partial educ}{\partial aid} = \beta_2 + \beta_3 \times dictator.$$

If the country is a democracy, then the marginal effect of foreign aid on education funding is just β_2. If the country is a dictatorship, then the marginal effect is $\beta_2 + \beta_3$. A natural hypothesis is that $\beta_3 < 0$, or in words: dictatorships that receive foreign aid don't seem to allocate as much of that foreign aid into education when compared to a democracy. (A more nuanced approach would try to measure the degree of dictatorship instead of a binary designation, but you get the picture.)