Introduction to Linear Models and Ordinary Least Squares

Miles D. Williams (University of Illinois at Urbana-Champaign)

July 16, 2021

keywords: OLS, linear models, math camp

1 Introduction

The most common methodological tool applied by political scientists is the additive linear regression model, and the most common method of estimating the parameters of such a model is an approach called ordinary least squares (OLS). In this guide, I'll introduce some important concepts, offer some examples, and provide some pointers, all with the goal of making linear models and OLS as intuitive and accessible as possible for newcomers to the field of political science.

2 But first some points of order...

Before we continue, we need to dispel a couple of common misconceptions. The first to keep in mind is the often ignored distinction between a ***model*** and an ***estimator***. It is not all that uncommon to see someone use the phrase "OLS model" to refer to a regression of the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

The above, however, is *not* an OLS model (there is no such thing); it is a linear model. OLS is an estimator used to estimate the values of the unknown parameters of a linear model. To miss this distinction is to miss the fact that OLS is not the only method for estimating the unknown parameters of a linear model—more on this later.

A second point concerns the use of terms like "independent" and "dependent" variables to refer to right- and left-hand side variables in a linear model, respectively. Consider a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i.$$

In the above, y_i is sometimes referred to as the "dependent" variable, and x_i and z_i are referred to as "independent" variables. The idea is that y_i *depends* (causally) on the values of x_i and z_i, and x_i and z_i are independent (random) variables that have a causal impact on the values of y_i. However, in most settings, causation may be too strong a claim—for instance, in observational settings where x_i and z_i are not assigned randomly as in a randomized controlled trial.

When we don't want to make causal, but rather *associational*, claims we might replace "dependent variable" with "outcome variable" or "response," and we might replace "independent variable" with "explanatory variable" or "predictor." It is important to be discerning about your choice of terms.

Now, on to the good stuff...

3 What is a linear model?

3.1 The basic functional form

A linear regression model is a ***statistical model*** that formulates, as a linear equation, the relationship between some response y_i and a set of predictor variables $x_{i1}, ..., x_{ik}$ for a sample of n individuals indexed $i = 1, ..., n$ as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i.$$

The values of the βs and of ϵ_i are unknown. β_0 is the intercept, and represents the mean of y_i conditional on $x_{ik} = 0$ for all k predictors. $\beta_1, ..., \beta_k$ are slope parameters that capture the rate of change in y_i given a unit change in their associated predictor variables. Finally, the values ϵ_i represent random or unexplained variation in y_i not accounted for by the set of predictors.

If you're familiar with the basics of linear algebra, the form of this equation may remind

you of another:

$$y = mx + b.$$

The above is the classic equation for a line, where x and y are variables, b is the intercept, and m is the slope—the "rise over the run." As you would have learned, b represents the value of y when $x = 0$, and m is the change in y per a one unit change in x.

A linear regression model is similar in spirit, but different in implementation, than this simple equation for a line. The idea with a linear regression model is to find the values of the slope(s) and intercept that provide the ***best fit*** for some observed data—data which is typically quite noisy, unlike the clean relationships between x and y you would have learned about in linear algebra. Whereas the solutions for m and b have deterministic analytical solutions, the values of the βs do not. They have to be calculated in reference to some rule or criterion.

Before we discuss such a rule, it will be helpful, first, to consider a linear model applied to some real-world data of interest to political scientists.

3.2 An application of a linear model

Consider an enduring question in the field of comparative political economy: the link between wealth and democracy. Say we want to know whether strength of democracy predicts greater wealth. To this end, we collect some data and put together a cross-sectional dataset of 136 countries in 2017. This dataset gives for each country its gross domestic product per capita (in 2011 dollars) and its polity 2 measure (a common, albeit contested and flawed, measure from -10 to 10 of how democratic a country is). GDP per capita will be our outcome or response variable, and polity 2 will be our explanatory variable.[1]

Figure 1 is a scatter plot of the relationship between democracy and wealth. Along the x-axis, we have countries' polity 2 scores. Along the y-axis, we have their levels of wealth, measured as the natural log of GDP per capita—it is common to take the natural log of financial data.

Visual inspection of the figure offers a sense for the relationship between democracy and wealth: more democratic countries appear to be wealthier—though there are some exceptions among more autocratic countries. Enter the ***linear regression model***...

[1] Polity 2 data come from the `PolityIV` dataset available in the `democracyData` R package. Economic data for countries come version 9.1 of the Penn World Table available via the `pwt9` R package.

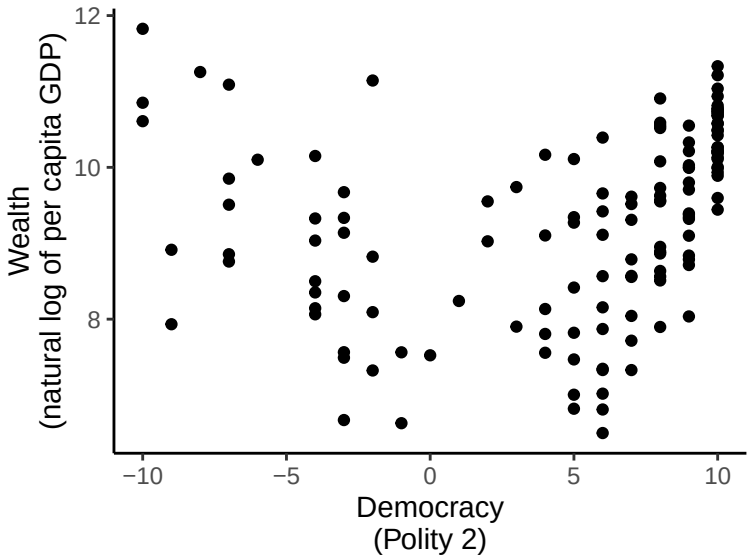

Figure 1: The relationship between democracy and wealth.

Let wealth be a linear function of democracy. Specifically:

$$\text{wealth}_i = \beta_0 + \beta_1 \text{democracy}_i + \epsilon_i,$$

where i indexes the countries in our cross-sectional dataset, β_0 is an intercept, β_1 is the slope capturing the rate of change in wealth given a change in a country's polity 2 score, and ϵ_i is variation in wealth unexplained by democracy.

This, admittedly, is a simplistic model of wealth—lots of factors influence wealth beyond polity—but it's sufficient to demonstrate the key moving pieces of linear models. The above represents a relationship or regularity in the world. A linear model takes messy data and imposes some semblance of order. It represents what's called a ***data-generating process*** that describes how democracy is linked to wealth.

3.3 We have a model; how do we make predictions?

Models are wonderful things, but they're nothing but window dressing without some way to identify their unknown parameters. To do this, we need an ***estimator***, which you can think of as a rule or procedure for selecting appropriate values of unknown quantities. As the title of this guide suggests, the rule we most often use to estimate the parameters of linear models is ***ordinary least squares*** (OLS). However, this is not our only option.

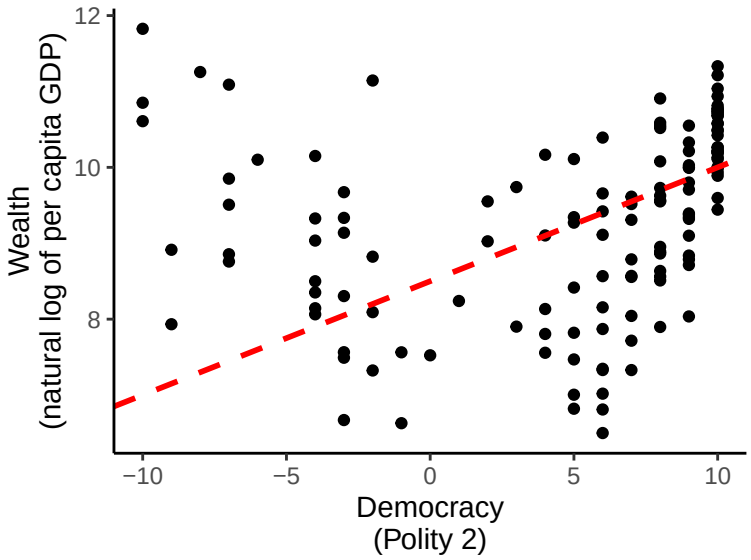

Figure 2: MBG (my best guess) estimate of the line of best fit.

One (perhaps unreliable) estimator is MBG, otherwise known as *my best guess*. MBG estimates the slope and intercept terms by, you guessed it, looking at the data and guessing what the best fitting slope and intercept are. Figure 2 shows my best guess for the model of wealth with a red dashed line. The fit doesn't look all that bad, if I do say so. For my best guess, I selected $\beta_0 = 8.5$ and $\beta_1 = 0.15$.

Of course, MBG is probably not the most scientifically rigorous choice. My selection was based purely on my subjective judgment. I picked some values that *looked* right; however, you might have chosen differently.

While MBG may not be entirely reliable, OLS is—at least in the sense that if I use OLS, and if you use OLS, we both with arrive at the same estimates of our unknown parameters.

The "least squares" part of OLS comes from the fact that the estimator finds the values of β_0 and β_1 that minimize the *sum of the squared residuals* (SSR). The residuals are just the leftover variance in a response unexplained by model predictors. These are the estimated values of the error term. While ϵ_i is the unobserved error of a linear model, $\hat{\epsilon}_i$ are the calculated residuals that equal the observed difference between a predicted value of the response, $\hat{y}_i$, and the observed value, y_i. For instance, for the MBG estimator, the residual for country i would be calculated as

$$\hat{\epsilon}_i = \text{wealth}_i - \widehat{\text{wealth}}_i,$$

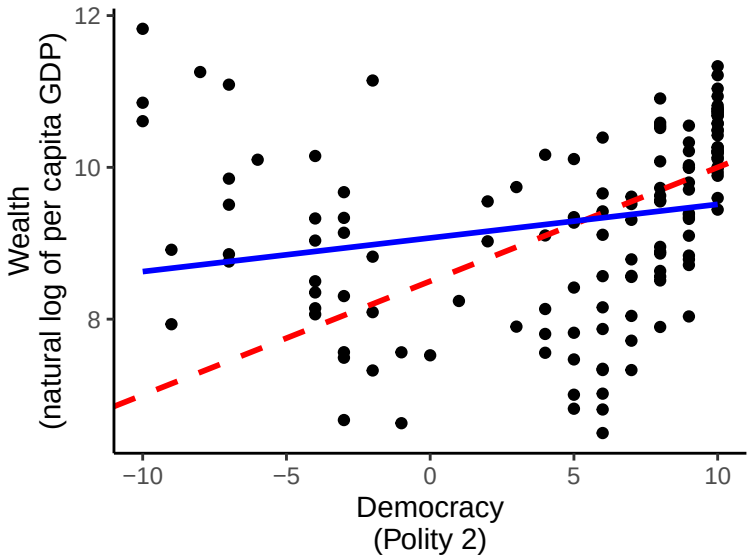

Figure 3: MBG vs. OLS.

where predicted values of wealth for country i are

$$\widehat{\text{wealth}}_i = 8.5 + 0.15 \times \text{democracy}_i.$$

These are the values that lie along the red dotted line in Figure 2.

The criteria OLS uses to select β values is the SSR:

$$\text{SSR} = \sum_i \hat{\epsilon}_i^2.$$

The values OLS chooses are those that minimize this ***objective function*** (that's what we call a function that we want to either minimize or maximize with respect to some parameters). In the case of democracy and wealth, OLS finds the values of β_0 and β_1 that minimize the sum of the squared differences between predicted values of wealth and observed values of wealth.

The OLS fit for democracy and wealth is shown in Figure 3 in blue. Note how different it is from the MBG estimate—the intercept is higher ($\beta_0 = 9.07$), and the slope less severe ($\beta_1 = 0.04$). Unsurprisingly, the SSR for this solution is much lower than what I achieved with MBG: 190.54 as opposed to 243.3. OLS didn't agree with my best guess, because my best guess didn't sufficiently minimize the SSR for wealth.

Optimizing on SSR is of course not the only valid approach to calculating β_0 and β_1. An

alternative to OLS is least absolute deviations (LAD). LAD is similar to OLS, but rather than find the slope and intercept that minimize the sum of the squared residuals, LAD finds the parameter values that minimize the sum of the absolute values of the residuals (SAVR):

$$\text{SAVR} = \sum_i |\hat{\epsilon}_i|.$$

This difference may seem trivial, but it's not. Taking the absolute values of the residuals, rather than squaring them, means that LAD is not as sensitive to observations that fall far away from the line of best fit. You can see the difference this makes by looking at Figure 4. The green dashed line is the LAD fit. Note that it falls between the MBG and OLS fits. This is because it is slightly less sensitive to the wealthy autocratic outliers in the data than is OLS.

While LAD is a robust method, in political science it is rare to see LAD applied. This is not necessarily because the approach is uniformly inferior to OLS—to the contrary, a major strength of LAD its lower sensitivity to outlier observations in the data. However, OLS has some strengths that LAD does not. For instance, OLS has a closed-form analytical solution (which we will discuss in greater detail below). LAD estimates, conversely, have to be computed using a numerical optimizing algorithm. In other words, we have to have a computer try out a bunch of different combinations of intercept and slope values to find the best. For a relatively small dataset of 136 country observations with only two parameters to find, this is no big deal. But, for much larger datasets with many more parameters to estimate, computing power can become an issue. OLS, because it has a closed-form analytical solution (that's just a fancy way of saying we can use an equation to find the parameter values), is much faster to implement with statistical software.

Additionally, compared to OLS, LAD estimates can be rather unstable, and there are even some cases where it will fail to find a unique solution. OLS on the other hand is much more stable and almost always has a unique solution. For these reasons (and more), OLS is the method of choice in most quantitative studies in political science.

3.4 Summary

In sum, a linear regression model is a statistical model of the relationship between some explanatory variable and an outcome expressed as a linear equation. In its most basic form, this model has an intercept, a slope, and an error term. Unlike the linear models you would have learned about in linear algebra, the parameters of a linear regression model do

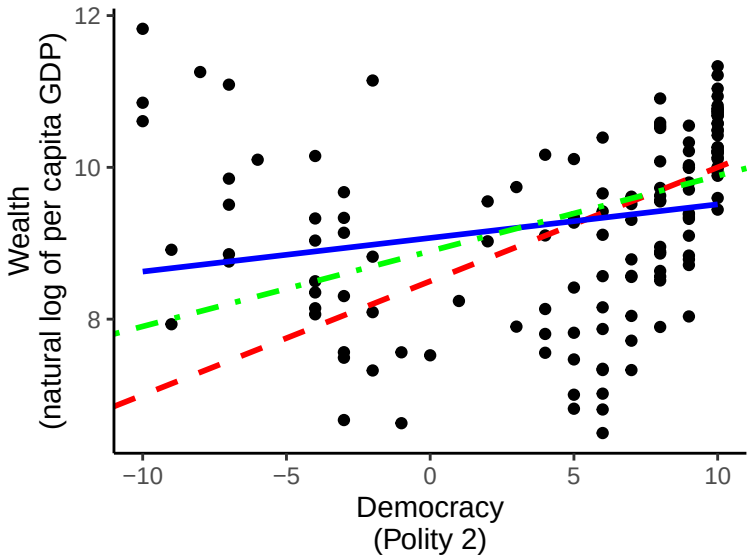

Figure 4: MBG vs. OLS vs. LAD.

not necessarily have deterministic closed-form solutions. Rather, there are numerous ways to estimate the parameters of a linear model. None of these is "right"—some are more useful than others, and each comes with its own advantages and disadvantages. OLS, however, is the most common method for estimating the parameters of a linear model. In the next section, we'll discuss OLS in greater detail.

4 Ordinary least squares

4.1 The analytical solution

As I mentioned in the previous section, OLS is the most popular method for estimating the parameters of a linear model. In this section, we'll take a deeper look at the OLS solution for model parameters.

Keeping with the example of democracy and wealth, recall that our simple linear model of wealth is given as

$$\text{weath}_i = \beta_0 + \beta_1 \text{democracy}_i + \epsilon_i,$$

where β_0 represents the wealth of a country if the polity 2 measure of democracy equals zero, and β_1 represents the change in wealth per a one unit change in the polity 2 score. The ϵ_i term represents random noise, or unexplained variation in wealth.

OLS finds the unknown parameters of this linear model such that the sum of the squared residuals (SSR) is minimized. The values of the residuals, $\hat{\epsilon}_i$, represent the leftover variation in wealth not explained as a linear function of democracy given fitted values of β_0 and β_1:

$$\widehat{\text{wealth}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{democracy}_i,$$
$$\hat{\epsilon}_i = \text{wealth}_i - \widehat{\text{wealth}}_i.$$

The clever thing about OLS is that it has a closed-form solution for the unknown parameters. We'll skip over the bit about how this solution is derived, but, suffice it to say, the analytical solution for $\hat{\beta}_1$ is simply:

$$\begin{aligned}\hat{\beta}_1 &= \frac{cov(\text{democracy}_i,\ \text{wealth}_i)}{var(\text{democracy}_i)} \\ &= \frac{\sum_i (\text{democracy}_i - mean[\text{democracy}_i])(\text{wealth}_i - mean[\text{wealth}_i])}{\sum_i (\text{democracy}_i - mean[\text{democracy}_i])^2}\end{aligned}$$

where $mean[\cdot]$ just indicates that we're taking the mean of the variable in brackets:

$$mean[\text{wealth}_i] = \frac{\sum_i \text{wealth}_i}{n}; \quad mean[\text{democracy}_i] = \frac{\sum_i \text{democracy}_i}{n}.$$

Further, the solution for the intercept is

$$\hat{\beta}_0 = mean[\text{wealth}_i] - \hat{\beta}_1 mean[\text{democracy}_i].$$

Now, these solutions are simple enough, but the math becomes a little more involved if we have multiple variables on the right-hand side of the linear model. For this reason, it is sometimes more efficient to express linear models in matrix notation:

$$\begin{bmatrix} \text{wealth}_1 \\ \vdots \\ \text{wealth}_n \end{bmatrix} = \begin{bmatrix} 1 & \text{democracy}_1 \\ \vdots & \vdots \\ 1 & \text{democracy}_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

or just

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is an $n \times 1$ matrix containing the values of wealth for the each of the n countries in the cross-sectional data, and $\mathbf{X}$ is a $n \times 2$ matrix containing a constant and the values of

democracy for the n countries in the data.

The OLS solution for the 2×1 matrix of parameters $\boldsymbol{\beta}$ is then expressed as

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The parameter values returned by this approach will be identical to those obtained using the scalar solutions shown previously. If we had multiple predictors, we would just add an additional column to $\mathbf{X}$ containing the values of each, and then the returned matrix of $\hat{\beta}$s would be calculated accordingly—that is, we would have a $\hat{\beta}$ for each column of $\mathbf{X}$.

4.2 What are OLS assumptions?

The goal of estimating a linear model is to generate "good" or consistent estimates of the unknown parameters. The extent to which OLS estimates are good hinge on the following:

1. that the data-generating process underlying the observed data is additive and linear;
2. that the explanatory variables are exogenous.

The first assumption should go without saying. If a linear, additive equation is not the best way to model an outcome as a function of predictors, then OLS estimates will not be consistent because the model itself has been mispecified.

The second assumption is similar to the first. The exogeneity of the explanatory variables has to do with whether there is some unobserved or unaccounted for confounding variable that simultaneously influences the values of the outcome and explanatory variables of interest. Take our model of wealth as a function of democracy. If there is some additional variable that both influences democracy and wealth, failure to include this variable in our model may lead OLS to return a biased estimate of the relationship between democracy and wealth. We call variables *endogenous* (the opposite of *exogenous*) if they are both caused by some unobserved variable. (As it turns out, endogeneity is a major problem in research on the link between democracy and wealth.)

4.3 What are NOT OLS assumptions?

For OLS estimates to be consistent, linearity and additivity, as well as exogeneity of the explanatory variables, are all that is required. It is not uncommon, however, to see

some add that the error term in the model is normally, and independently and identically distributed (iid). This, however, is not actually a requisite for consistent OLS estimates. The normality of the error term is more often a mathematical property of OLS estimation; not an assumption. Further, iid of the error term is only relevant for making statistical inferences about estimated model parameters; not estimation of the parameters themselves. However, it is true that in the special case where normality and iid of the error term hold, the OLS estimator is identical to the maximum likelihood estimator—yet another (model-based) approach to estimating the parameters of a linear regression model.

Linear independence of the right-hand side variables also is sometimes listed as an assumption of OLS. This, too, is not quite true. If linear independence is violated (that is, if your data suffer from multicollinearity), this is problematic in so far as it precludes estimation of model parameters. If two identical variables are included on the right-hand side of a linear model, OLS will not be able to identify parameters for both. This does not violate any assumptions of OLS, but it can make estimating a linear model problematic. Including a variable, or set of variables, that add little to no new information that can be used to predict an outcome variable is simply unwise—but it does not violate any assumptions.

5 OLS and statistical inference

As scientists, we are often interested in testing hypotheses when we do our research. To do this, we almost never are interested only in fitting a linear regression model—we also want to make statistical inferences about the model parameters we've estimated.

In our linear model predicting wealth as a function of democracy, we observed in Figure 3 that the slope of the OLS line of best fit was positive—specifically, it was $\hat{\beta}_1 = 0.04$. However, an important question is whether this estimate is statistically distinguishable from zero. We can answer this question with the help of ***standard errors***. These capture how precisely the point estimates for model parameters have been estimated and are used in the calculation of test statistics and, ultimately, p-values, which quantify how surprised we should be to see $\hat{\beta}_1$ take the value that it does if its true value were zero.

5.1 Calculating standard errors

Standard errors are calculated from a mathematical object called the *variance-covariance matrix*. This matrix contains estimates of the variances of the βs calculated via OLS along

its diagonal, and the covariances of each of the βs in its off-diagonal elements.

Calculating the variance-covariance matrix for classical statistical inference for OLS is quite simple to do, though it does require a number of assumptions to hold—assumptions that, in your own research, you almost never will be in a position to make. These assumptions are that the data are (1) independently and (2) identically distributed (iid). If this is the case, then the solution for the variance-covariance matrix of the model parameters is simply

$$\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) \\ cov(\hat{\beta}_0, \hat{\beta}_1) & var(\hat{\beta}_1) \end{bmatrix},$$

where $\hat{\sigma}^2$ is given as

$$\hat{\sigma}^2 = \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{n-k} = \frac{\sum_i \epsilon_i^2}{n-k} = \frac{\text{SSR}}{n-k}.$$

The value k in the above represents the number of parameters to be estimated in the linear regression model.

To get the standard errors for the βs, all we need to do is extract the diagonal elements of the above matrix, and take their square root:

$$\begin{bmatrix} se(\hat{\beta}_0) & se(\hat{\beta}_1) \end{bmatrix} = \sqrt{\text{diag}\,[\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]}.$$

In most applied settings, iid assumptions are rather heroic, and so the above variance-covariance matrix will not be efficient—it won't give you the appropriate estimates of the parameters' standard errors. In some cases, observations in the data may be dependent—such as when observations are clustered within groups. In other cases, the errors may not be identically distributed—a problem otherwise known has heteroskedasticity. Sometimes, both problems may exist simultaneously.

Focusing on the issue of identically distributed errors, if we'd rather not impose such a restriction, a heteroskedasticity robust variance-covariance matrix can be calculated as follows:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} var(\beta_0) & cov(\beta_0, \beta_1) \\ cov(\beta_0, \beta_1) & var(\beta_1) \end{bmatrix}.$$

This is what is known as the HC0 variance-covariance estimator (HC = heteroskedasticity consistent), or the White estimator, so named for a paper by Halbert White published in *Econometrica* in 1980.

In practice, however, few actually use the raw HC0 errors. It's often necessary to adjust

for finite sample bias associated with the HC0 variance-covariance matrix. So, researchers often apply a degrees of freedom adjustment to the above to calculate what are called HC1 standard errors. Their solution is just the HC0 variance-covariance matrix multiplied by the number of observations over the number of observations minus the number of model parameters:

$$\begin{aligned}\text{HC1} &= \frac{n}{n-k}\text{HC0} \\ &= \frac{n}{n-k}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

There are also HC2, HC3, and more types of robust variance-covariance estimators—but we'll worry about those another day. HC1 are are appropriate most of the time, in most observational settings.

While HC estimators are robust to heteroskedasticity, it may also be appropriate to adjust for dependence in the the distribution of the errors. Clustering of observations is common in various types of data. For instance, if we had outcomes on student performance on a standardized test, we might worry that test scores within a given classroom are interdependent. We would need to cluster our standard errors by classroom to account for this. As another example, in the case of democracy and wealth, suppose we had a panel dataset for our 136 countries that ran from 2007 to 2017. Because GDP per capita for one country in one year is going to depend on its GDP per capita in years prior, we would need to cluster our standard errors by country to account for within-country dependence in wealth.

The solution for clustered errors is similar to that for robust errors, but with the caveat that the observation specific errors are replaced with group-specific errors. Here's the solution for CR0 errors (the clustered version of HC0 errors):

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'\text{diag}[\boldsymbol{\epsilon}_g\boldsymbol{\epsilon}_g']\mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1},$$

where g indexes the groups (clusters) in the data.

Just like with HC0 versus HC1, errors, best-practice is to apply a degrees of freedom adjustment to account for the number of observations, model predictors, and groups in the data. This gives us what are called CR1 errors, the clustered counterpart of HC1 errors:

$$\text{CR1} = \frac{G}{G-1}\frac{n-1}{n-k}\text{CR0}.$$

For our cross-section of countries, we only need HC1 errors, which we can calculate very

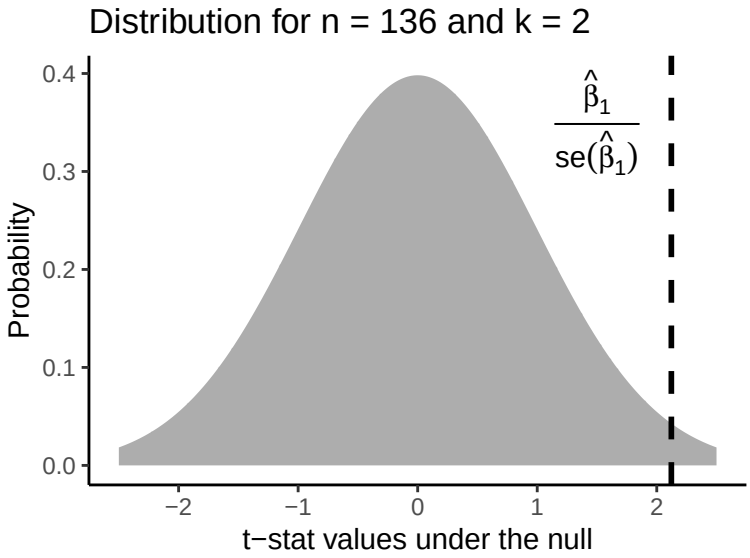

Figure 5: t-distribution under the null for the slope coefficient on democracy.

easily with statistical software. For the intercept and slope of our linear model of wealth, these are, respectively, $se(\hat{\beta}_0) = 0.16$ and $se(\hat{\beta}_1) = 0.02$.

5.2 From standard errors to p-values

With our parameters and their standard errors estimated, the next step is to determine whether $\hat{\beta}_1$ is statistically distinguishable from zero. One way that we do this is to compute a test statistic (which has a known distribution) and then calculate the probability of observing a statistic of the value we calculated by random chance—more precisely, under the *null hypothesis* that the true value of $\beta_1 = 0$. This probability is called a p-value.

The test statistic that we compute for OLS estimates is called a t-statistic. For our estimate of the slope in our model of wealth, this is:

$$t_1 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}.$$

Once we have t, we then need to compare it with the t-distribution under the null hypothesis. This is the asymptotically derived probability distribution of the t-statistic when we assume that the true value of our estimated parameter is zero. Figure 5 shows what this distribution looks like, and notes with a vertical line where our calculated t-statistic for the slope on democracy falls. This value is 2.12.

We can see from the figure that the probability of this test statistic is pretty low. How low? To answer this question, we first need to make a decision about whether we want to perform a one-tailed or two-tailed test. For the latter, we're interested in calculating the probability of observing some $|t| \geq |t_1|$. For the former, we're interested in calculating the probability of observing some $t \geq t_1$. A one-tailed test takes into account the direction of an estimate, while a two-tailed test just takes into account its magnitude, regardless of direction. In most cases, we're interested in two-tailed tests. However, there are some cases where one-tailed tests are appropriate.

For a two-tailed test, we compute the probability of observing a t-statistic as extreme as the one we observed for democracy's relationship with wealth as:

$$p_1 = 1 - 2 \times F(|t_1|),$$

where $F(\cdot)$ is the cumulative distribution function for the probability density function for the t-statistic. We multiply this value by 2 since it is two-sided. For $\hat{\beta}_1$, this probability is $p_1 = 0.036$.

This is what's known as a p-value. Smaller p-values suggest that a statistic is a lower probability event under the null hypothesis. Usually, we set some criterion value of p such that if p is less than or equal to it we ***reject the null***—that is to say, consider an estimate statistically distinguishable from zero. We call the level of a test the α-level, and by convention, this is set at $\alpha = 0.05$. As a rule of thumb, t-values at least as large as 1.96 will meet this criterion. In the case of the relationship between democracy and wealth, we would reject the null since p_1 is less than 0.05.

5.3 Summarizing regression results

A summary of our OLS estimates and our inferences about them are shown in Table 1. In most research papers, you'll see some version of this table (what we call a regression table) summarizing results from a single or several regression models.

Included here are the estimates for the intercept, or constant term, and the estimated coefficient for the relationship between democracy and wealth from our linear regression model. The robust standard errors for each parameter are shown in parentheses.

It is also common practice to signify the "significance" of an estimate with a *. Since our coefficient for democracy is statistically distinguishable from zero at the $p \leq 0.05$ level, we

Table 1: OLS Estimates with Robust S.E.s

	Model of Wealth
Constant	$9.07\ (0.16)^{***}$
Democracy (Polity 2)	$0.04\ (0.02)^{*}$
Num. obs.	136
RMSE	1.19

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

note this by including a * next to the coefficient.

6 Extensions of the linear model

6.1 Nonlinear functional forms

A linear additive model is the most common functional form in regression analysis. However, it can be easily extended to accommodate more complicated relationships in a dataset. Take the relationship between democracy and wealth. As the scatter plots shown in the previous section suggest, we can easily model wealth as a linear function of countries' polity 2 scores. There were some wealthy autocratic outliers that were not well predicted by our model, but these cases may be exceptions; not the rule.

Or are they? What if the relationship between democracy and wealth is quadratic, rather than linear?

Fitting a quadratic model is quite easy. All we need to do is add a squared democracy score to our original model,

$$\text{wealth}_i = \beta_0 + \beta_1 \text{democracy}_i + \beta_2 \text{democracy}_i^2 + \epsilon_i$$

and then we can use OLS to find β_0, β_1, and β_2. With the fitted parameters, we can then plot the OLS predicted values of wealth as a function of democracy, as in Figure 5.

As it turns out, the quadratic model is a fairly good fit for the data. Technically, the updated model is still a linear regression in the sense that it is additive with respect to the β parameters. The form of the relationship between the explanatory variable and the outcome, however, is nonlinear. Many other complicated forms can be expressed as linear models as well.

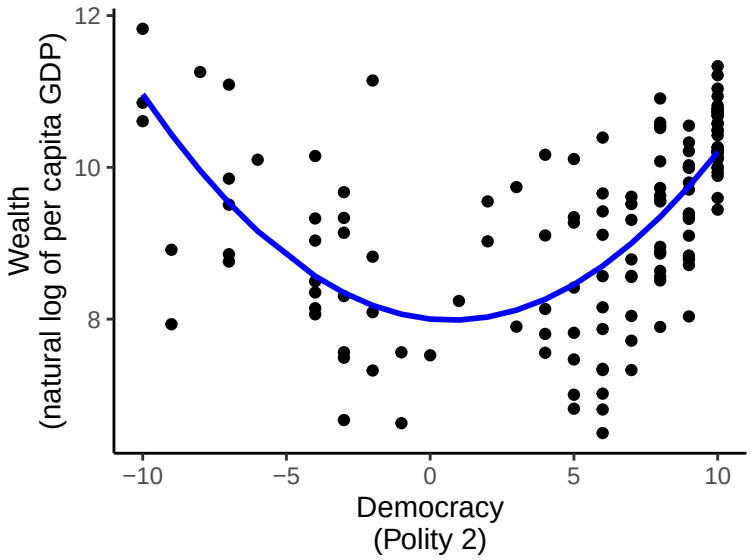

Figure 6: A quadratic relationship between democracy and wealth?

6.2 Linear models with more than one explanatory variable

Almost never in your career as a political scientist will your regression analysis center on a linear model with only one right-hand side variable. Even if you're primarily interested in the relationship between a single variable and outcome, you will often need to "control for" other variables in your analysis. When we include multiple variables in our equation, we call the linear model a ***multiple regression model***.

There are two types of control variables that you'll likely include:

1. *confounding variables* that have relationships with your explanatory variable of interest and your response variable.
2. variables that are *prognostic* of the response, but independent of your explanatory variable.

The first type of control variable is included to reduce bias in your estimates. The second type of control variable is included to improve the efficiency (precision) of your estimates.

Take our country cross-sectional dataset. In addition to collecting data on wealth and polity, say we also collected data on a measure called ***human capital***. The human capital index (HCI) is derived from a per person measure years of schooling and economic returns to education. Might human capital also influence per capita GDP?

We can easily expand our model of wealth to include human capital:

$$\text{wealth}_i = \beta_0 + \beta_1 \text{democracy}_i + \beta_2 \text{human capital}_i + \epsilon_i.$$

The OLS estimate for democracy returned by this approach will now be adjusted to account for human capital. In technical terms, the estimate on democracy will represent the residual linear relationship between democracy and wealth, after partialing (subtracting) out the variation in wealth and in democracy explained as a linear function of human capital.

The above may seem rather dense, but the implications of controlling for covariates is not all that complicated in principal. It may be helpful to consider an alternative, but functionally equivalent, approach to controlling for a covariate.

The estimate of β_1 in the above can be recovered using the matrix solution shown in a previous section, or it can be recovered by doing the following:

1. First, fit a regression model where wealth is just a function of human capital:

$$\text{wealth}_i = \eta_0 + \eta_1 \text{human capital}_i + v_i.$$

2. Second, fit a regression model where democracy is a function of human capital:

$$\text{democracy}_i = \gamma_0 + \gamma_1 \text{human capital}_i + \varepsilon_i.$$

3. Third, take the calculated residuals from the above equations, and regress the residual variation in wealth ($\hat{v}_i = \text{wealth}_i - \widehat{\text{wealth}}_i$) on the residual variation in democracy ($\hat{\varepsilon}_i = \text{democracy}_i - \widehat{\text{democracy}}_i$):

$$\hat{v}_i = \beta_0 + \beta_1 \hat{\varepsilon}_i + \epsilon_i.$$

$\hat{\beta}_1$ for the the above regression is our estimate of the marginal linear relationship between democracy and wealth.

For another way of thinking about what "controlling for" means, consider the scalar closed-form solution for β_1 discussed previously. For the simple model of wealth considered earlier, $\hat{\beta}_1$'s OLS solution was just:

$$\hat{\beta}_1 = \frac{\sum_i (\text{democracy}_i - mean[\text{democracy}_i])(\text{wealth}_i - mean[\text{wealth}_i])}{\sum_i (\text{democracy}_i - mean[\text{democracy}_i])^2},$$

Table 2: OLS Estimates with Robust S.E.s

	Model of Wealth
Constant	$5.50\ (0.21)^{***}$
Democracy (Polity 2)	$-0.02\ (0.02)$
Human Capital (HCI)	$1.47\ (0.09)^{***}$
Num. obs.	136
RMSE	0.70

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

which is just the covariance in democracy and wealth over the variance in democracy:

$$\hat{\beta}_1 = \frac{cov(\text{democracy}_i,\ \text{wealth}_i)}{var(\text{democracy}_i)}.$$

When we control for other variables in our analysis, we just replace the sample means for our outcome and explanatory variable of interest in the above formulations with conditional means:

$$\widehat{\text{wealth}}_i = mean[\text{wealth}_i|\ \text{human capital}_i],$$
$$\widehat{\text{democracy}}_i = mean[\text{democracy}_i|\ \text{human capital}_i].$$

The solution for $\hat{\beta}_1$ is thus now

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_i(\text{democracy}_i - \widehat{\text{democracy}}_i)(\text{wealth}_i - \widehat{\text{wealth}}_i)}{\sum_i(\text{democracy}_i - \widehat{\text{democracy}}_i)^2} \\ &= \frac{cov(\text{democracy}_i - \widehat{\text{democracy}}_i,\ \text{wealth}_i - \widehat{\text{wealth}}_i)}{var(\text{democracy}_i - \widehat{\text{democracy}}_i)}.\end{aligned}$$

OLS estimates for the multiple regression model of wealth are shown in Table 2. Interestingly, when we control for human capital, the estimate for democracy changes sign and falls short of statistical significance.

7 Goodness of fit and model selection

The final concept we'll cover in this guide is "goodness of fit" (GOF). When we estimate a linear model via OLS, sometimes we not only want to make inferences about our estimated intercept and slope(s), we also want to assess the overall performance of our model.

This can be helpful, for instance, when adjudicating between alternative linear model specifications.

Take our model of wealth. Given the data and variables we've considered thus far, we might imagine four alternative specifications:

$$(1) \quad \text{wealth}_i = \beta_0 + \beta_1 \text{democracy}_i + \epsilon_i,$$
$$(2) \quad \text{wealth}_i = \beta_0 + \beta_1 \text{democracy}_i + \beta_2 \text{human capital}_i + \epsilon_i,$$
$$(3) \quad \text{wealth}_i = \beta_0 + \beta_1 \text{democracy}_i + \beta_2 \text{democracy}_i^2 + \epsilon_i,$$
$$(4) \quad \text{wealth}_i = \beta_0 + \beta_1 \text{democracy}_i + \beta_2 \text{democracy}_i^2 + \beta_3 \text{human capital}_i + \epsilon_i.$$

Knowing which of these is "right" is impossible to know, but we can use GOF to determine which is best.

For linear models, two measures of GOF are often used. The first is an R^2 statistic, which captures the proportion of variation in a response explained by model predictors. The equation for R^2 is

$$R^2 = 1 - \frac{SSR}{SST},$$

where SST is

$$SST = \sum_i (y_i - \bar{y})^2.$$

That's equivalently just the sum of the squared residuals from a model that only includes an intercept.

Usually, we aren't interested in the raw R^2 but rather an adjusted-R^2 that accounts for degrees of freedom. The adjusted-R^2 is calculated as

$$\text{adjusted-}R^2 = 1 - \frac{(n-1)SSR}{(n-k)SST},$$

where n is the number of observations and k is the number of model parameters.

The second metric we use is an F-statistic. Like the t-statistic for linear regression coefficients, the F-statistic has a known theoretical distribution, against which we can assess the probability of observing an F-value under the null hypothesis that the linear model is no better than a model that' reduced to a single intercept. Say, for example, we wanted to calculate F for equation 1 listed above. We first would calculate SSR for this model. Then,

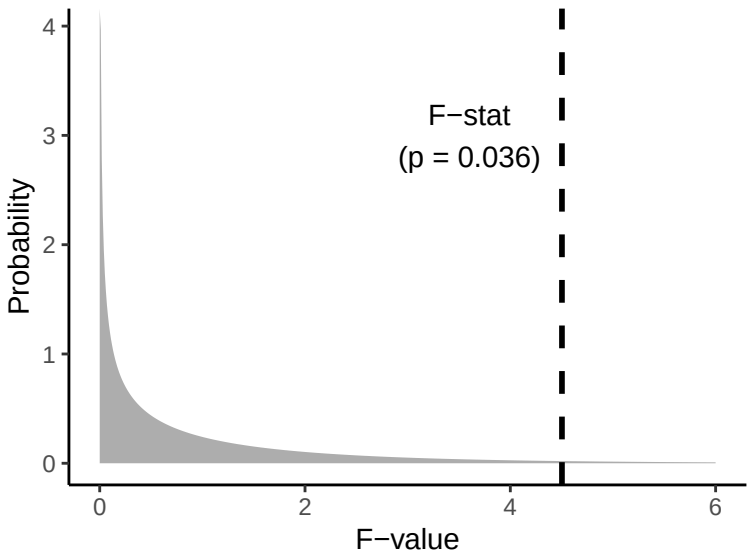

Figure 7: F-statistic for model of wealth relative to distribution under the null.

we would calculate SST. We then compute F as

$$F = \frac{SST - SSR}{k-1} / \frac{SSR}{n-k}$$

Performing an F-test lets us determine whether we should reject the hypothesis that

$$\text{wealth}_i = \beta_0 + \beta_1 \text{democracy}_i + \epsilon_i$$

can be reduced to

$$\text{wealth}_i = \beta_0 + \epsilon_i.$$

Figure 7 shows the F-value for the simple model of wealth as a function of democracy. As suggested, the observed value of 4.5 is statistically significant at the $p < 0.05$ level ($p = 0.036$). This means we can have some confidence that wealth as a linear function of democracy is a better fit for the data than wealth as a function of a constant.

We of course have a number of regression models to choose from, the results of which are shown in Table 3. Can we adjudicate among these to determine which is the best?

We have a few options. First, we can compare the adjusted-R^2 estimates for each of the models. Presumably, the model that explains the greatest variation in wealth is a good candidate for best fit. By this criterion, model 4, which models wealth as a linear function of human capital and a quadratic function of polity, is best.

Table 3: OLS Estimates for Different Models of Wealth

	Model 1	Model 2	Model 3	Model 4
Contant	9.07^{***}	5.50^{***}	8.00^{***}	5.61^{***}
	(0.16)	(0.21)	(0.18)	(0.19)
Democracy (Polity 2)	0.04^{*}	-0.02	-0.04	-0.05^{**}
	(0.02)	(0.02)	(0.02)	(0.02)
Democracy^2			0.03^{***}	0.01^{***}
			(0.00)	(0.00)
Human Capital (HCI)		1.47^{***}		1.20^{***}
		(0.09)		(0.08)
Adj. R^2	0.04	0.67	0.41	0.74
Num. obs.	136	136	136	136
F statistic	4.50	178.37	48.15	154.24

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Difference in Sum of Sq.

	Model 1	Model 2	Model 3	Model 4
Model 1	0	124.83*	74.79*	140.05*
Model 2	-124.83*	0	-50.04	15.22*
Model 3	-74.79*	50.04	0	65.26*
Model 4	-140.05*	-15.22*	-65.26*	0

Note:

* denotes if Pr(F-stat) is less than 0.05.

We can also perform a series of F-tests where, instead of comparing each model to one reduced to only an intercept, we compare each to the other models in the data. Modifying the F-statistic to accommodate this is quite easy with statistical software. Pairwise comparisons for each of the models in the data are shown in Table 3.

The results are consistent with the adjusted-R^2 values and suggest that Model 4 is superior to the alternative specifications. Specifically, for the F-test, we can reject the hypothesis that wealth modeled as a quadratic function of democracy and a linear function of human capital can be reduced to any of the other, simpler, specifications.

8 Appendix: code used for the analysis

```
# Set options:
knitr::opts_chunk$set(echo = FALSE,
                      message = FALSE,
                      warning = FALSE,
                      fig.height = 3, fig.width = 4)

# Load packages:
library(tidyverse) # for grammar
library(estimatr)  # for linear models with robust SEs
library(texreg)    # for regression tables
library(kableExtra)# for tidy tables

# --------------------------
# LOAD AND MERGE THE DATA
# --------------------------

# Penn data
pwt <- pwt9::pwt9.1 %>%
  filter(year %in% 2017) %>%
  select(country, year, pop, rgdpna, hc) %>%
  mutate(code = countrycode::countrycode(country, "country.name", "iso3c"))

# Polity data
pty <- democracyData::politylV %>%
  filter(year %in% 2017) %>%
  select(politylV_country, year, polity2) %>%
  rename(country = politylV_country) %>%
  mutate(code = countrycode::countrycode(country, "country.name", "iso3c"))

# Merge
ds <- inner_join(
  pwt, pty,
  by = c("year", "code")
```

```
) %>%
  select(-country.y) %>%
  rename(country = country.x) %>%
  select(code, everything()) %>%
  na.omit

# Make wealth variable:
ds <- ds %>%
  mutate(
    wealth = log(rgdpna / pop)
  )

# ----------
# ANALYSIS
# ----------

# Make figure 1:
p <- ggplot(ds) +
  aes(polity2, wealth) +
  geom_point() +
  labs(
    x = "Democracy\n(Polity 2)",
    y = "Wealth\n(natural log of per capita GDP)"
  ) +
  theme_classic()
p # plot

# Make figure 2
p <- p +
  geom_abline(
    slope = 0.15,
    intercept = 8.5,
    lty = 2,
    col = "red",
    size = 1
  )
```

```r
p # plot

# Make figure 3:
fit <- lm(wealth ~ polity2, ds) # fit model
p <- p +
  geom_line(
    aes(polity2, fitted(fit)),
    size = 1,
    col = "blue"
  )
p # plot

# Get SSR for my best guess and ols:
mbg <- round(sum((ds$wealth - 8.5 - 0.15 * ds$polity2)^2), 2)
ols <- round(sum(resid(fit)^2), 2)

# Make simple objective function for LAD:
lad <- function(x, y, b) {
  yhat <- b[1] + b[2] * x
  AD <- sum(abs(yhat - y))
  return(AD)
}

# Use numerical optimizer to find LAD parameter estimates:
opt <- optim(
  fn = lad,
  par = c(0, 0),
  y = ds$wealth,
  x = ds$polity2
)

# Make figure 4:
p +
  geom_abline(
    intercept = opt$par[1],
    slope = opt$par[2],
```

```
    col = "green",
    size = 1,
    lty = 4
  )

# Compute HC1 variance-covariance matrix:
 # The components:
n <- nrow(ds) # sample size
k <- 2        # no. of parameters
X <- cbind(constant = 1, x = ds$polity2)
              # the data matrix
e <- resid(lm(wealth ~ polity2, ds))
              # the residuals
I <- diag(nrow = n, ncol = n)
              # identity matrix

 # Calculate White errors:
HC0 <- solve(t(X)%*%X)%*%t(X)%*%(e%*%t(e) * I)%*%X%*%solve(t(X)%*%X)

 # Apply degrees of freedom adjustment for HC1 errors:
HC1 <- (n / (n - k)) * HC0

 # Take the squareroot of the diagnonal to get standard errors:
ses <- sqrt(diag(HC1))

# Plot of the t-distribution:
x <- seq(-2.5, 2.5, by = 0.01)
t <- dt(x, df = n - k)
ggplot() +
  aes(x, ymin = 0, ymax = t) +
  geom_ribbon(
    alpha = 0.4
  ) +
  geom_vline(
    xintercept = cov(ds$polity2, ds$wealth) / var(ds$polity2) / ses[2],
    lty = 2,
```

```
    size = 1
  ) +
  labs(
    x = "t-stat values under the null",
    y = "Probability",
    title = "Distribution for n = 136 and k = 2"
  ) +
  annotate(
    "text",
    x = 1.5,
    y = 0.35,
    label = expression(frac(hat(beta)[1], 'se'*(hat(beta)[1])))
  ) +
  theme_classic()

# The estimate of beta_1
b1 <- cov(ds$polity2, ds$wealth) / var(ds$polity2)

# The t-stat
t1 <- b1 / ses[2]

# The p-value
p <- 2 * pt(t1, df = n - k, lower.tail = F)

# Fit model with lm_robust:
fit <- lm_robust(
  wealth ~ polity2,
  data = ds,
  se_type = "stata" # These are HC1 errors
)

# Make regression table:
texreg(
  fit,
  custom.model.name = "Model of Wealth",
  custom.coef.map =
```

```
    list("(Intercept)" = "Constant",
         "polity2" = "Democracy (Polity 2)"),
  include.ci = F,
  caption = "OLS Estimates with Robust S.E.s",
  caption.above = T,
  single.row = T,
  include.rsquared = F,
  include.adjrs = F
)

# Get predictions from a quadratic model:
fit <- fitted(
  lm(wealth ~ polity2 + I(polity2^2), ds)
)

# Scatter plot with quadratic fit:
ggplot(ds %>% mutate(fit = fit)) +
  aes(polity2, wealth) +
  geom_point() +
  geom_line(
    aes(polity2, fit),
    size = 1,
    col = "blue"
  ) +
  labs(
    x = "Democracy\n(Polity 2)",
    y = "Wealth\n(natural log of per capita GDP)"
  ) +
  theme_classic()

# Fit the four regression models:
base <- lm_robust(
  wealth ~ polity2,
  ds,
  se_type = "stata"
)
```

```r
basecon <- lm_robust(
  wealth ~ polity2 + hc,
  ds,
  se_type = "stata"
)
quad <- lm_robust(
  wealth ~ polity2 + I(polity2^2),
  ds,
  se_type = "stata"
)
quadcon <- lm_robust(
  wealth ~ hc + polity2 + I(polity2^2),
  ds,
  se_type = "stata"
)

# Make regression table of model with control:
texreg(
  basecon,
  custom.model.name = "Model of Wealth",
  custom.coef.map =
    list("(Intercept)" = "Constant",
         "polity2" = "Democracy (Polity 2)",
         "hc" = "Human Capital (HCI)"),
  include.ci = F,
  caption = "OLS Estimates with Robust S.E.s",
  caption.above = T,
  single.row = T,
  include.rsquared = F,
  include.adjrs = F
)

# Plot the F-distribution:
dfr <- nrow(ds) - 1
dff <- nrow(ds) - 2
n   <- nrow(ds)
```

```r
x   <- seq(0, 6, 0.01)
basef <- base$fstatistic
f   <- df(x, df1 = basef[2], df2 = basef[3])
p   <- pf(basef[1], basef[2], basef[3], lower.tail = F)
ggplot() +
  aes(
    x, ymin = 0, ymax = f
  ) +
  geom_ribbon(alpha = 0.4) +
  geom_vline(
    xintercept = basef[1],
    lty = 2,
    size = 1
  ) +
  annotate(
    "text",
    x = basef[1]*.8,
    y = 3,
    label = "F-stat\n(p = 0.036)"
  ) +
  labs(
    x = "F-value",
    y = "Probability"
  ) +
  theme_classic()

# Regression table with all four models:
texreg(
  list(base, basecon, quad, quadcon),
  include.ci = F,
  include.rsquared = F,
  include.rmse = F,
  include.fstat = T,
  custom.coef.map = list(
    "(Intercept)" = "Contant",
    "polity2" = "Democracy (Polity 2)",
```

```
    "I(polity2^2)" = "Democracy$^2$",
    "hc" = "Human Capital (HCI)"
  ),
  caption = "OLS Estimates for Different Models of Wealth",
  caption.above = T
)

# Re-estimate models with lm()
# [I can't use anova() for F-tests on lm_robust objects]
base <- lm(wealth ~ polity2, ds)
basecon <- lm(wealth ~ polity2 + hc, ds)
quad <- lm(wealth ~ polity2 + I(polity2^2), ds)
quadcon <- lm(wealth ~ polity2 + I(polity2^2) + hc, ds)

# Run a loop to make all pairwise comparisons between models:
modlist <- list(base, basecon, quad, quadcon)
modgrid <- expand.grid(modlist1 = modlist, modlist2 = modlist)
ftest <- matrix(0, 4, 4)
colnames(ftest) <- rownames(ftest) <- c("Model 1", "Model 2", "Model 3", "Model 4")
for(i in 1:nrow(modgrid)) {
  test <- anova(modgrid$modlist1[[i]], modgrid$modlist2[[i]])
  ftest[i] <- paste0(round(test$`Sum of Sq`[2],2),
                     if(test$`Pr(>F)`[2] <= 0.05 &
                        !is.na(test$`Pr(>F)`[2])) "*" else "")
}

# Make the table of pairwise comparisons:
kable(ftest,
      "latex",
      caption = "Difference in Sum of Sq.",
      booktabs = T) %>%
  footnote("* denotes if Pr(F-stat) is less than 0.05.") %>%
  cat()
```