Multiple Regression

The textbook by FPP offers a nice, intuitive discussion of regression and other statistical tools. Unfortunately, their discussion deviates from the usual notation-heavy presentation in political science. The standard presentation in political science borrows heavily from econometrics. Since POLS 309 focuses on the regression model, we're going to dive in to the notation a bit in these notes. Think of it as a preview for POLS 309.

Simple Regression Model

FPP represent the regression model as $y = mx + b$, where we understand y as the "predicted value" or "average value" of y rather than the outcome variable itself.

The standard econometric notation represents the same idea slightly differently, as $y = \beta_0 + \beta_1 x + u$. There are a few differences:

1. In econometric notation, y represents the observed outcome variable, not the predicted or average outcome as FPP's notation.
2. The intercept is denoted as β_0 rather than b and is typically written first. Similarly, the slope is denoted as β_1 rather than m.
3. The term u is new. It represents the error–the vertical distance between the regression line and the points. This allows y to represent the observed outcome rather than the predicted or average outcome. u is refered to as the "error," "error term," or "disturbance."

The observed variables y and x have different names depending on the author and discipline.

y	x
Dependent Variable	Independent Variable
Outcome Variable	Explanatory Variable
Response	Predictor
Regressand	Regressor
	Feature
	Covariate

Economists draw a sharp distinction between the *actual* regression model $y = \beta_0 + \beta_1 x + u$ and the estimated regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. This distinction is important for statistical theory, because it allows methodologists to evaluate approached to estimating β_0 and β_1. The idea is that there is some true model–the actual relationship between x and y–but the researcher can only find *estimates* of β_0 and β_1. The estimated is distinguished from the true value by adding a "hat," so that the estimate of the true value β_0 is denoted as $\hat{\beta}_1$.

Similarly, u is the distance between the observed values of y and the *true* regression line $\beta_0 + \beta_1 x$. $\hat{u}$ represents the distance between observed values of y and the *estimated* regression line $\hat{\beta}_0 + \hat{\beta}_1 x$.

In order to justify the least squares estimator (i.e., using the line that minimizes the r.m.s. error), econometricians make several assumptions about the model $y = \beta_0 + \beta_1 x + u$. However, in my view, these assumptions are not particularly important because least squares tends to work well even when the assumptions are not met. Also, we don't yet have the probability theory we need to describe these assumptions.

Multiple Regression Model

We can add variables by simply expanding the econometric notation to two variables, so that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Note that we have two explanatory variables now, so we attach subscripts 1 and 2 to distinguish them.

Similarly, we could expand the model to three explanatory variables, so that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$. In general, we can have k explanatory variables in the model, so that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$.

The interpretation of the slope coefficients β_1, β_2, ..., and β_k remains the same as our interpretation of m. As x_i increases by one unit, the average value of y increases by β_i units, where y equals 1, 2, ..., or k.

Review Exercises

1. Explain what the `lm()` function does and how it works. In particular, what is the first argument to `lm()`? The second? What does `lm()` output (or return)?
2. What does the function `coef()` do? In particular, what is the first (and only) argument to `coef()`? What does it output?
3. What does the function `residuals()` do? In particular, what is the first (and only) argument to `residuals()`? What does it output? How can you use the output to calculate the r.m.s. error of the regression?
4. What function do you use to add a linear regression line to a scatterplot? What arguments do we typically supply and why?
5. Explain how to adjust the alpha transparency of the points in a scatterplot. Explain how to jitter the points.
6. What two model summaries can we use to guess the predictive ability of the model? How do we calculate each in R?
7. What does the function `predict()` do? In particular, what is the first argument to `predict()`? The second? What does it output? How do you store the output of `predict()` as a variable in the prediction set?