ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Ordinary Least Squares and Agnostic Regression: Why We Do the Things We Do [1]

Linear regression is the bread and butter of econometrics, and we begin the course with a quick review of it. But this review may not resemble anything you saw in previous econometrics courses, primarily because it focuses on what linear regression is, rather than on what you would like it to be. That is to say, as long as you satisfy certain trivial conditions (e.g., your matrix of regressors is full rank), you can always run a linear regression. And there is absolutely nothing wrong with doing that – regardless of problems involving "endogeneity" or "omitted variables" or "measurement error" – as long as you interpret the results appropriately. To paraphrase the National Rifle Association, "Regressions don't give biased inferences, people interpreting those regressions give biased inferences." Or, in the words of Sergeant Joe Friday, "All we want are the facts." This lecture will focus on what we refer to as "agnostic regression."

1 The Conditional Expectation Function

Consider a dependent variable, y_i, and a vector of explanatory variables, x_i. We are interested in the relationship between the dependent variable and the explanatory variables. (Josh Angrist: "What matters for empirical work, as in life, is relationships.") There are several possible reasons that we may be interested in this relationship, including:

1. Description – What is the observed relationship between y and x?

2. Prediction – Can we use x to create a good forecast of y?

[1] These notes are inspired by Josh Angrist's "Empirical Strategies" lecture notes. Any errors in transcription or interpretation are my own. The answer to this question is not going to be the Gauss-Markov Theorem.

3. Causality – What happens to y if we experimentally manipulate x?

It is generally the last item that causes rants about "exogeneity conditions" and so forth. We will ignore all that negative energy for the moment and, instead of worrying about what you can't infer, focus on what you can infer. Think positive!

Of course, few real-world relationships are deterministic. Recognizing this fact, we focus on relationships that hold "on average," or "in expectation." Given our variables y and x, we may be interested in the *conditional expectation* of y given x. That is to say, given a particular value of x, where is the distribution of y centered? This relationship is given by the *Conditional Expectation Function*, or the CEF.

$$E[y_i|x_i] = h(x_i)$$

We define the CEF residual as:

$$\varepsilon_i = y_i - h(x_i) \text{ where}$$

$$E[\varepsilon_i|x_i] = 0$$

Note that, because ε_i is the CEF residual, $E[\varepsilon_i|x_i] = 0$ holds by definition – we do not require any exogeneity assumptions regarding x_i.

Proof.

$$E[\varepsilon_i|x_i] = E[y_i - h(x_i)|x_i] = E[y_i|x_i] - E[h(x_i)|x_i]$$

$$= E[y_i|x_i] - h(x_i) = E[y_i|x_i] - E[y_i|x_i] = 0$$

To recap, the CEF residual always has zero conditional expectation. By definition. No assumptions necessary. Always.

Theorem. CEF residuals are *mean-independent* of the arguments in the CEF, x_i (see above). They are therefore orthogonal to any function of the conditioning variables.

Proof. Iterated expectations.

$$E[\varepsilon_i \cdot f(x_i)] = E[E[\varepsilon_i \cdot f(x_i)|x_i]] = E[E[\varepsilon_i|x_i]f(x_i)] = E[0] = 0$$

More importantly, the CEF is the "best" function of x that exists for predicting y (where "best" is defined in terms of expected squared loss).

Theorem. $E[y_i|x_i] = argmin_g\ E[(y_i - g(x_i))^2]$ In other words, the CEF is the function that minimizes the expected squared deviations from y_i. We say that "the CEF is the *minimum mean-square error* (MMSE) predictor for y_i given x_i."

Proof.

$$E[(y_i - g(x_i))^2] = E[((y_i - E[y_i|x_i]) + (E[y_i|x_i] - g(x_i)))^2] =$$

$$E[(y_i - E[y_i|x_i])^2 + 2(y_i - E[y_i|x_i])(E[y_i|x_i] - g(x_i)) + (E[y_i|x_i] - g(x_i))^2] =$$

$$E[E[(y_i - E[y_i|x_i])^2 + 2(y_i - E[y_i|x_i])(E[y_i|x_i] - g(x_i)) + (E[y_i|x_i] - g(x_i))^2|x_i]] =$$

$$E[E[(y_i - E[y_i|x_i])^2|x_i] + 2(E[y_i|x_i] - E[y_i|x_i])(E[y_i|x_i] - g(x_i)) + (E[y_i|x_i] - g(x_i))^2] =$$

$$E[E[(y_i - E[y_i|x_i])^2|x_i]] + E[(E[y_i|x_i] - g(x_i))^2]$$

It should be clear that choosing $g(x_i)$ such that $g(x_i) = E[y_i|x_i]$ minimizes the second term in the last line. The first term in the last line does not contain $g(x_i)$ and is therefore unaffected by our choice of $g(x_i)$. The CEF, $E[y_i|x_i]$, therefore solves $min_g\ E[(y_i - g(x_i))^2]$.

2 Regression and the CEF: Why We Regress

Clearly the CEF has some desirable properties in terms of summarizing the relationship between x_i and y_i and making predictions about y_i given x_i. In particular, we have seen that it is the MMSE predictor of y_i. But what does this have to do with linear regression, and why might we want to use linear regression?

2.1 Reason the First: Regression-CEF Theorem

Theorem. If the CEF is linear, then the regression of y_i on x_i estimates the CEF. Formally, if $E[y_i|x_i] = x_i\gamma$, then $\gamma = E[x_i'x_i]^{-1}E[x_i'y_i]$ (which is what the regression coefficient converges to).

Proof.

$$E[x_i'x_i]^{-1}E[x_i'y_i] = E[x_i'x_i]^{-1}E[E[x_i'y_i|x_i]] = E[x_i'x_i]^{-1}E[x_i'E[y_i|x_i]] =$$

$$E[x_i'x_i]^{-1}E[x_i'x_i\gamma] = E[x_i'x_i]^{-1}E[x_i'x_i]\gamma = \gamma$$

Of course, there is no reason the CEF has to be linear. Two of the most common sufficient conditions for a linear CEF are: (1) joint normality of x_i and y_i or (2) a *saturated model* for discrete regressors. A saturated model is one in which you estimate a separate parameter for each point in the support of x_i (e.g., you have a separate dummy variable for each unique value of the vector x_i in your data set). This is more common in empirical work than joint normality.

In most cases, however, the CEF is not linear. But we still run regressions anyway. Why do we do this? One reason is that it is computationally tractable and that we understand its properties both when it is correctly specified and under misspecification (or, at least, we understand its properties under misspecification better than we understand the properties of other estimators). Nevertheless, there are good theoretical reasons to regress as well.

2.2 Reason the Second: BLP Theorem

Theorem. If you want to predict y_i, and you limit yourself to linear functions of x_i, then $x_i\beta = x_iE[x_i'x_i]^{-1}E[x_i'y_i]$ is the best linear predictor (BLP) of y_i in a MMSE sense. Formally, $\beta = E[x_i'x_i]^{-1}E[x_i'y_i] = argmin_b\ E[(y_i - x_ib)^2]$.

Proof.

$$\partial E[(y_i - x_i b)^2]/\partial b = 2E[x_i'(y_i - x_i b)] = 0$$

$$E[x_i'y_i] - E[x_i'x_i]b = 0$$

$$b = E[x_i'x_i]^{-1}E[x_i'y_i] = \beta$$

If you're limiting yourself to linear combination of x_i, then linear regression gives you the best predictor of y_i. Of course, this isn't a big surprise given that the OLS estimator is derived by minimizing the sample analog of $E[(y_i - x_i b)^2]$. Regardless, this property is nice if you're in the business of forecasting, but it's not as useful if your interest is in estimating the CEF as a summary of the underlying relationship between y_i and x_i. Which brings us to our third reason to regress (arguably the best reason).

2.3 Reason the Third: Regression Approximation Theorem

Theorem. The MMSE linear approximation to the CEF is $\beta = E[x_i'x_i]^{-1}E[x_i'y_i]$. Formally, $\beta = E[x_i'x_i]^{-1}E[x_i'y_i] = argmin_b\ E[(E[y_i|x_i] - x_i b)^2]$.

Proof.

$$\partial E[(E[y_i|x_i] - x_i b)^2]/\partial b = 2E[x_i'(E[y_i|x_i] - x_i b)] = 0$$

$$E[E[x_i'y_i|x_i]] - E[x_i'x_i]b = 0$$

$$b = E[x_i'x_i]^{-1}E[x_i'y_i] = \beta$$

So regression provides the best linear approximation to the CEF, even when the CEF is non-linear. Regression can therefore give you a pretty decent approximation of the CEF as long as you don't try to extrapolate beyond the support of x_i.

3 Discussion

If your object of interest is the CEF, then linear regression is a good tool for estimating it. Specifically, it is the best linear predictor in terms of minimizing the mean squared error from the CEF. More importantly, this result depends on absolutely nothing. In particular, it does not depend on:

Whether your data are i.i.d.

Whether you treat your regressors as random variables or fixed quantities.

Whether your regressors are correlated with the CEF residuals (by definition, they are not, since the residuals are mean-independent of any function of the conditioning variables).

Whether the CEF is linear or not.

Whether your dependent variable is continuous, discrete, non-negative, or anything else.

Regression is therefore remarkably robust as an estimation tool, provided that you interpret it for what it actually is — an approximation of the conditional expectation function — rather than what you might like it to be (an estimate of a causal relationship). So if you're only interested in description or prediction, we can probably end the class right here.

4 Machine Learning

When talking about *prediction*, we would be remiss not to think about machine learning (ML). Mullainathan and Spiess (2017) give a nice overview of potential applications of ML in applied econometrics. Because ML focuses on prediction, not causality, it is generally a complement to, rather than a substitute for, good research designs.

ML uses sophisticated algorithms to predict an outcome, y. As with regression, we predict y using a function of x. The key tension in any ML algorithm is the balance between extracting as much information from the data as possible while avoiding overfitting. The

intuition behind overfitting is easy to see with linear regression. Since regression minimizes the sum of squared residuals, it can always do weakly better with additional regressors. Thus regression will always fit a non-zero coefficient to any additional regressors that you add, until it runs out of degrees of freedom (at which point it will perfectly predict the outcome).

ML deals with overfitting by penalizing models with too much complexity (regularization). In the linear regression example, that means we would penalize regression models that fit too many coefficients. Later in the course we will discuss an algorithm that does exactly this, the Least Absolute Shrinkage Selection Operator (LASSO). For the moment, just think of ML algorithms as algorithms that try to predict y using arbitrarily complicated functions of x, while still avoiding the overfitting that is inherent with these complex functions.

So how does ML, with its focus on prediction, relate to *causal* inference? At the moment, only tangentially. Mullainathan and Spiess give three examples of potential ML applications in applied econometrics. First, we could use ML to preprocess data or generate new data sets. For example, language is something that has historically been difficult to code numerically, but in some cases ML algorithms can extract meaning or intent from natural language (e.g., classifying text reviews as positive or negative). Likewise, some data processing steps that formerly required human interaction become automated (e.g., optical character recognition).

Second, ML can be useful for some estimation steps that pertain to prediction. While overfitting is not a first-order concern in most applied econometrics studies, it is an issue with certain estimators (e.g., many weak instruments, an issue we will talk about later). The most common example is the question of how many control variables to include in a regression. On the one hand, including too many controls consumes degrees of freedom and results in overfitting (of controls). On the other hand, excluding key controls can cause bias. We will discuss a ML technique later in the course for selecting regression controls.

Finally, prediction itself can be policy relevant in some cases. For example, we may want to predict which individuals are most likely to take up treatment (targeting). Or we may be interested in predicting future behavior or success. A judge, for example, may want to

predict an individual's recidivism, or a university admissions officer may want to predict an applicant's likelihood to graduate college. In these cases, one of which we discuss next, ML techniques have the potential to improve upon the predictive capabilities of linear regression.[2]

5 Application: Predicting College Success

Geiser and Santelices (2007) use high school GPA, standardized test scores (SAT), and other covariates to predict college performance (college GPA) using linear regression for UC freshman entering between Fall 1996 and Fall 1999. The results from this exercise are listed in Table 4 of their article, reproduced below. They find that, in this sample, high school GPA is a more effective predictor of college GPA than any other measure. In particular, it is much more effective than SAT I (the standard SAT). This can be seen in at least two ways. First, in comparing Model 1 – which uses high school GPA as a predictor – and Model 2 – which uses SAT I as a predictor – we see that Model 1 has a much higher R^2; in other words, high school GPA is explaining much more of the variation in college GPA than SAT I score is (Note: This may be the *only* time in this course that you will hear reference to R^2. In general it is *not* an interesting statistic in answering policy-relevant questions.) We also see, in Model 7, that the standardized coefficient on high school GPA is substantially larger than the standardized coefficient(s) on SAT I (the standardized coefficient is a normal regression coefficient that has been rescaled to indicate how many standard deviations y changes with a one standard deviation change in x). Moving up one standard deviation in the high school GPA distribution is therefore much more beneficial for college GPA (in a predictive sense) than moving up one standard deviation in the SAT I score distribution.

Does this relationship answer any interesting, policy-relevant questions? Arguably, yes. If you are a UC admission officer, and you are tasked with reducing acceptance rates due to state budget cuts, then you can use the regression results to predict which students are least

[2]Whether ML meaningfully improves upon linear regression for predictive purposes depends heavily on the specific application. In the Mullainathan and Spiess application, predicting housing prices, the best ML technique that they test provides only a modest improvement over OLS.

I. Validity of Admissions Factors in Predicting Cumulative Fourth-Year GPA

We begin with findings on the relative contribution of admissions factors in predicting cumulative four-year college GPA. Table 4 shows the percentage of explained variance in cumulative fourth-year GPA that is accounted for by HSGPA, SAT I verbal and math scores, and SAT II Writing, Mathematics and Third Test scores. The estimated effects of these admissions factors on cumulative fourth-year GPA were analyzed both singly and in combination. Parents' education, family income and school API rank were also included in all of the regression models in order to control for the "proxy" effects, noted above, of socioeconomic status on standardized test scores and other admissions variables.

Table 4

Relative Contribution of Admissions Factors in Predicting Cumulative Fourth-Year GPA

	Standardized Regression Coefficients										
	High School GPA	SAT I Verbal	SAT I Math	SAT II Writing	SAT II Math	SAT II 3rd Test	Parents' Education	Family Income	School API Rank	Number	% Explained Variance
Model 1	**0.41**	x	x	x	x	x	**0.12**	**0.03**	**0.08**	59,637	20.4%
Model 2	x	**0.28**	**0.10**	x	x	x	**0.03**	**0.02**	0.01	59,420	13.4%
Model 3	x	x	x	**0.30**	**0.04**	**0.12**	**0.05**	**0.02**	**-0.01**	58,879	16.9%
Model 4	**0.36**	**0.23**	0.00	x	x	x	**0.05**	**0.02**	**0.05**	59,321	24.7%
Model 5	**0.33**	x	x	**0.24**	**-0.05**	**0.10**	**0.06**	**0.02**	**0.04**	58,791	26.3%
Model 6	x	**0.06**	-0.01	**0.26**	**0.04**	**0.12**	**0.04**	**0.02**	**-0.01**	58,627	17.0%
Model 7	**0.34**	**0.08**	-0.02	**0.19**	**-0.04**	**0.09**	**0.05**	**0.02**	**0.04**	58,539	26.5%

Boldface indicates coefficients are statistically significant at 99% confidence level.
Source: UC Corporate Student System data on first-time freshmen entering between Fall 1996 and Fall 1999.

likely to succeed. We know from the previous theorems that the CEF provides the MMSE prediction of y (college GPA) and that regression provides the MMSE linear approximation to the CEF. So in a predictive sense you are likely to do well (at least relative to alternative choices), and in this case what you care about is prediction.

These results also have policy relevance in that the University of California would like to maintain a diverse student body but is not allowed to give any weight to ethnicity as an admission criterion. UC administrators are aware, however, that weighting SATs more heavily (as is traditionally done) tends to favor Caucasians (and possibly Asians?), while weighting high school GPA more heavily tends to favor Blacks and Latinxs (in a relative sense). But will putting more weight on high school GPA and less weight on SAT scores result in a lower quality student body? The results from Table 4 indicate that it will not; in fact, if anything, it may result in a higher quality student body. This result is reassuring as UC moves away from using SAT scores for admissions.

Are the estimated relationships causal? Highly unlikely. Even after controlling for parental education and income, there are probably unobserved individual, family, neighborhood, and peer characteristics that affect college success and are correlated with high school GPA and SAT scores.[3] The regression results, however, are still useful for prediction and have interesting applications in policy-relevant questions.

One can still take issue with the results along multiple dimensions. For example, should some adjustment be done to GPA to reflect the student's choice of major?[4] Might there be other variables collected from the applicants that could improve the predictive power of the model? Nevertheless, the fact remains that the results are useful and interesting despite the fact that the coefficients do not have causal interpretations. This example makes appropriate

[3]In fact, it doesn't even make sense to talk about an experimental manipulation of high school GPA or SAT scores. The effects on college success will almost surely depend on whether the treatment entails raising these attributes through cram sessions or through mentoring programs or through intensive intervention earlier in life. The treatment is better defined as the actual intervention than as raising GPA by one point or increasing SAT scores by 100 points.

[4]I would strongly recommend against getting into this debate – it will be a great way to alienate a lot of colleagues very quickly.

use of a descriptive relationship estimated via linear regression, which is probably more than can be said for the vast majority of empirical applications in economics.