Athena User Manual

April 2020

Background and introduction

Hello! Welcome to Athena's user manual. We're so happy that you have chosen to give Athena a try.

Athena is an econometric and statistical computer tool created by John Willett and Gabe Stengel in 2020. This document details Athena's statistical functionality as of 4/29. More is being added all the time! If you have questions about Athena's functionality or think that you have found a mistake or bug anywhere, please do not hesitate to reach out to the creators. John can be reached at johnwillett7@gmail.com, and Gabe can be reached at gabrielrstengel@gmail.com.

As a general rule, if the question relates to Athena's econometric or statistical knowledge, John is the go-to. If the question concerns Athena's machine learning or natural language processing capabilities, Gabe is a better bet.

You'll notice quickly that this is not a typical user manual. Most user guides detail the ins and outs of a library's functionality. They say exactly what the library can do and what exactly to write to get it to do those things. But because Athena allows its users to write natural language as they query, it does not need a user guide of that sort. This document serves more as an educational resource for users, and it doubles as an indicator of Athena's general domains of expertise.

Contents

1 Fixed Effects

1.1 Longitudinal data

Imagine we have data $\mathbf{y}$, $\mathbf{x} \in \mathbb{R}^{NxT}$. In other words, for individuals $i = 1, 2, ...N$ we have $(y_i, x_i)_t$ for $t = 1, 2, ...T$. In economics this structure is often called panel data or longitudinal data.

For concreteness, imagine that we want to know whether the amount people work (y_i) depends on the weather (x_i). If our data was not longitudinal—that is, if we just had (y_i, x_i) from a large pool of subjects at the same time point—we would have reason to be suspicious about the results of ordinary least squares. After all, maybe people who live in hotter places work harder for reasons other than weather!

But if we have access to data that "follow" the same subjects over a period of time, then we have a powerful econometric tool at our disposal: fixed effects regression. Imagine now that we have N subjects and have observed both the number of hours people work and the temperature per week for T weeks.

1.2 Modeling the same people over time

We begin with the following T n-dimensional models:

$$y_{i,t} = \alpha_i + \beta x_{i,t} + \epsilon_{i,t}, \text{for } t = 1...T \tag{1}$$

α_i, a subject-specific intercept, is what we call "time-invariant heterogeneity" across subjects. It corresponds to those innate, unchanging characteristics of each subject that affect how much she works. Of course, α_i is only one subset of countless other variables that likely affect y_i. But even if we somehow were able to track down and include every relevant, easily-measurable quantity currently omitted from the regression, different people might still work different amounts for subject-specific reasons. Statistically, this poses a problem. What if α_i has a nonzero correlation with x_i, our covariate of interest? Perhaps smart people both work hard and enjoy living in hot places!

If you squint at it for a moment you will see that the following equation holds:

$$\bar{y_i} = \alpha_i + \beta \bar{x_i} + \bar{u_i}, \text{where } \bar{z_i} := \frac{1}{T} \sum_{t=1}^{T} z_{i,t} \tag{2}$$

Subtract and get the following model. This process is called "entity demeaning" because each mean is the mean for a specific entity (or person) rather than across entities for a specific week.

$$y_{i,t} - \bar{y_i} = \beta(x_{i,t} - \bar{x_i}) + \epsilon_{i,t} - \bar{\epsilon_i} \tag{3}$$

But look at what has happened: α_i has disappeared! This is great news for the robustness of our coefficient estimate, since we have completely eliminated omitted variable bias (the enemy of all good economics research) from

time-invariant heterogeneity. Any bias problems related to unchanging innate qualities or self-selection issues are magically—mathematically—gone. This is called an entity fixed effect, and we call $\beta = \beta_{FE}$ the fixed effects estimator.

For a more intuitive feel of why this is significant, take another look at the new model and consider the following interpretation: a β of, say, 1.5 would mean that for any given individual, her being in a week that is 1 degree hot relative to her own region's weather is associated with a 1.5 hour boost in the work-week. In other words, subjects in the study worked more when it was hotter for them.

Interestingly, if instead of de-meaning the data we regressed all of the datapoints but included entity-specific intercepts (binary variables that indicate whether or not person i is being predicted on), the math turns out to be the same. This is sometimes how computers execute this estimation technique, and it is how the open-source libraries Athena makes use of for fixed effects do it.

Time fixed effects are mathematically the same as entity fixed effects, just the other way around. Here, we introduce week dummies. Perhaps some major world event happened during week 6 of the study; by including an indicator for every week, we control for these effects.

2 Logistic Regression and Marginal Effects

2.1 Motivation and MLE

Given data $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{NxK}$ where $y_i \in \{0, 1\}$ for all i, we want to fit a model that predicts the likelihood of $y_i = 1$ based on the k covariates.

The logistic regression approach uses maximum-likelihood estimation (MLE). In the general case, we assume we have n independent observations drawn from a density $f(data_i, \boldsymbol{\theta})$ for some unknown parameter vector $\boldsymbol{\theta}$, and we set up the likelihood equation in this way:

$$L(\theta) := \prod_{i=1}^{n} f(data_i; \boldsymbol{\theta}) \tag{4}$$

Taking the log we get:

$$ln(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} lnf(data_i; \boldsymbol{\theta}) \tag{5}$$

Enforcing a handful of (weak) mathematical criteria, one can show that the MLE estimator of the parameter vector is not only consistent but, in the limit, Gaussian:

$$\hat{\boldsymbol{\theta}}^{MLE} \sim N(\boldsymbol{\theta}, \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}^{MLE}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}) \tag{6}$$

2.2 Logistic regression

We wish to estimate $\boldsymbol{\theta}$ with some k-dimensional column vector of coefficients $\hat{\boldsymbol{\beta}}$. Define a function ψ such that $\psi(\hat{\boldsymbol{\beta}}^T \boldsymbol{X_i}) = \mathbb{P}(Y_i = 1 | \boldsymbol{X_i})$, where $X_1 = 1$ to

allow an intercept. Trivially, assuming these probabilities exist, the likelihood of any result given its outcome is $\mathbf{1}_{Y_i=1}\mathbb{P}(Y_i = 1|\boldsymbol{X_i}) + \mathbf{1}_{Y_i=0}\mathbb{P}(Y_i = 0|\boldsymbol{X_i})$. Substituting in our probability estimates, we can get the likelihood function:

$$ln(L(\boldsymbol{\beta})) = \sum_{i=1}^{n}(\boldsymbol{Y_i}ln(\psi(\hat{\boldsymbol{\beta}}^T\boldsymbol{X_i})) + (1 - \boldsymbol{Y_i})ln(1 - \psi(\hat{\boldsymbol{\beta}}^T\boldsymbol{X_i})) \tag{7}$$

With modern numerical optimization software, this is easy enough to maximize given a function ψ. (In its MLE implementations, Athena routinely makes use of gradient descent algorithms available through an array of online machine learning and optimization libraries.) So all that is left to do here is choose a function. Given the binary response variable setting we want a function whose range is in $[0, 1]$. In economics, the most common choices are $\boldsymbol{\Phi}$—the standard normal CDF (Probit regression)—and the logistic function. For this project we chose the latter because it is more common in cross-disciplinary settings and because it has more support from the online libraries of which we made use.

$$F(\alpha) = \frac{e^{\alpha}}{1 + e^{\alpha}} \tag{8}$$

The logistic function, as a reminder, looks as in Equation 4.8 above.

2.3 Marginal effects

In traditional machine learning settings, the prediction estimates themselves are commonly the quantities of interest. But in the social sciences, we are not so concerned (speaking broadly) in generating sensible predictions for $\hat{\boldsymbol{y}}$. Instead, we tend to be more interested in our estimated weight vector $\hat{\boldsymbol{\beta}}$! Our chief aim is to study the size and significance of the relationships between the features we select and the dependent variable. For example, economists are not likely to wonder how well unemployment, GDP, and consumption spending predict mental health outcomes as a group. But they would be interested in how much explanatory power unemployment has in determining mental health outcomes when we control for GDP and consumption spending.

With a linear regression model, we are immediately in good shape, because we can simply read off the values of the coefficients and rest assured that those numbers are "apples-to-apples" with the dependent variable. But when we use a nonlinear function transformation like in logistic regression, the coefficients no longer correspond to absolute changes in our dependent variable.

The solution to this problem most commonly used in social sciences is to consider the marginal effects of coefficient changes. We take the derivative of the logistic function to get the following estimates, where $\boldsymbol{X_i}$ denotes an arbitrary row of data.

$$\frac{\partial\mathbb{P}(Y_i = 1|\boldsymbol{X_i})}{\partial X_{j,i}} = \frac{e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{X_i}}}{(1 + e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{X_i}})^2}\beta_j \tag{9}$$

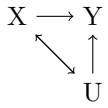
Of course, the values of these partial derivatives depend on the values of all the covariates. The most common choices for comparisons are to evaluate them at the means of the variables or to calculate the average partials.

3 Instrumental Variables

3.1 General Overview

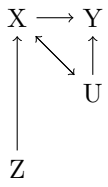
Suppose that we are interested in the relationship between Y and X but that there are a group of variables $\boldsymbol{U}$ that may be correlated with both Y and X in a confounding way. If we have access to any of the variables in $\boldsymbol{U}$, we should put them into our model as controls. But some variables are difficult to control for, especially ones that have to do with people's personalities, talents, and work ethics.

We often represent these with what are called "dependency graphs" like the one below. Imagine we are running a study at Princeton: we are interested in whether spending time in Dillon (X) actually makes people more physically fit (Y). (Assume we have some objective measure of fitness, perhaps to do with one's heart health.) We have access to the requisite data, but we face a significant bias problem. People who go to Dillon a lot may be more likely to do other things that also contribute to health. So there is this other variable out there—a general attitude toward fitness—that is correlated with our variables of interest. Looking at the following graph, it seems we are stuck. How can we possibly extricate the magnitude of that X-Y arrow from the grasp of those confounding $\boldsymbol{U}$ variables?



But wait: We know which residential colleges our students live in! Surely living in Wilson or Whitman, which are very close to Dillon, is well-correlated with the frequency with which one works out at Dillon. Whitmanites can roll out of bed and limp over to Dillon every morning. If you are in Rocky or Forbes, that is not a very fun walk.

The best part about this realization is that residential college assignment is random. So it is not correlated with any variables that might confound our results. Calling residential college assignment Z for the moment, we have a nice new graph:



Critically, Z is correlated with Y only through X. In our case, we assume that one's residential college affects her physical fitness only through the frequency with which she attends Dillon. (We may begin to question this assumption if we consider Butler College's proximity to Studio '34 pizza! We will revisit this concern.)

Here is the intuition behind instrumental variable regression: If higher values of Y are associated with higher values of Z, then we have reason to believe that X is related to Y in a causal way, not just a correlative way through the omitted variables in $\boldsymbol{U}$. If this is not obvious, imagine we found that people in Whitman were far fitter than their peers in Forbes. This seems now to be reasonable evidence that going to Dillon is good for your physical health.

We call the residential college placement variable a statistical "instrument" because it is used as an intermediary between the variables of interest. This approach dates back at least to Wright (1928). For an instrument to be valid, it must pass two tests.

1. It must be "exogenous." That is, it must be uncorrelated with $\boldsymbol{U}$ and affect Y only through X. (In our case, we must assume that Forbesians' nearness to Wawa does not affect their physical fitness.)

2. It must be "relevant." That is, it must have a nonzero correlation with X. (If we found out that students in Whitman and Forbes go to Dillon the same amount, then we have not helped ourselves at all!)

3.2 Regression in the univariate case

How would we actually run a regression with instruments in it? Our life is very easy if we have just one regressor x_i and one instrument Z. Consider the following model.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{10}$$

Notationally, β_1 is not the coefficient in a regression of Y on X but rather the theoretical unbiased coefficient that we mean to estimate using our instrument Z. In this simple case, we can—if we've had our coffee—simply do the following.

$$\begin{aligned} cov(y_i, Z_i) &= cov(\beta_0 + \beta_1 x_i + \epsilon_i, Z_i) \\ &= \beta_1 cov(x_i, Z_i) + cov(u_i, Z_i) \end{aligned} \tag{11}$$

But from our first criterion:

$$cov(u_i, Z_i) = 0. \tag{12}$$

So:

$$\beta_1 = \frac{cov(y_i, Z_i)}{cov(x_i, Z_i)} \tag{13}$$

We can calculate this quantity from the data without too much trouble. Let us reflect now on why, from a mathematical standpoint, we require instruments that are correlated with our independent variable. Notice that we are dividing by the covariance of X and Z, so if this equals zero we will not be able to proceed. And even when covariance is nonzero, we can run into really noisy coefficient estimates if X and Z are not very well-correlated. (We typically desire not just relevant instruments but "strong" instruments. We will discuss this later.)

3.3 Two-stage least squares

But what do we do if we have more than one independent variable that is potentially correlated with our model's error term? It turns out we can instrument for multiple variables at the same time. The only restriction is that we need at least as many instruments as we have endogenous regressors.

(As a reminder, this word "endogenous" just means that a variable is correlated with the error term or, in other words, suffers from bias from omitted variables. Its opposite is the word "exogenous," which is what we want our instruments to be.)

The solution to the multivariate instrument case lies running two layers of linear regression in an approach called "two-stage least squares," or 2SLS. In the first stage, we regress our endogenous covariates on the instruments. In the second, we regress the fitted values for the endogenous covariates on the dependent variable. The coefficients from the second model are the ones we keep.

We will now take a quick look at the math, sticking first with the univariate case for easy intuition's sake. Our two models are the following.

$$\begin{aligned} x_i &= \lambda_0 + \lambda_0 z_i + u_i, \text{with } \boldsymbol{E}[u_i|z_i] = 0 \\ y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \text{with } \boldsymbol{E}[\epsilon_i|x_i] = 0 \end{aligned} \tag{14}$$

As a reminder, β_1 is not the coefficient estimate we would get in a regression of Y on the endogenous X but rather the hypothetical causal coefficient in the X-Y relationship that we mean to estimate through Z. Otherwise, we could not assume $\boldsymbol{E}[\epsilon_i|x_i] = 0$. Substituting in we get the following, where $v_i := \beta_1 u_i + \epsilon_i$ is the error term of the combined model.

$$\begin{aligned} y_i &= \beta_0 + \beta_1(\lambda_0 + \lambda_0 z_i + u_i) + \epsilon_i \\ &= \beta_0 + \beta_1(\lambda_0 + \lambda_0 z_i) + +\beta_1 u_i + \epsilon_i \\ &= \beta_0 + \beta_1(\lambda_0 + \lambda_0 z_i) + v_i \end{aligned} \tag{15}$$

This is a useful result because X is sort of a dual figure in terms of its correlation with Y. Some of its correlation is due to omitted variables for which

we want to control, and some is due to the causal correlation we are interested in. Thinking back to our Dillon example, we just want to capture that variation in X that is due to residential college placement and not due to general health attitudes. The terms $\lambda_0 + \lambda_0 z_i$ fit this bill, since they are the fitted values of X that are uncorrelated with ϵ_i. Mathematically, this means $\boldsymbol{E}[\epsilon_i|\lambda_0 + \lambda_0 z_i] = 0$. So we know that $\boldsymbol{E}[v_i|\lambda_0 + \lambda_0 z_i] = 0$ as well! Therefore, we can proceed with this new regression model and know that, if our assumption that $\boldsymbol{E}[\epsilon_i|z_i] = 0$ is correct, the β_1 we find in the second stage will be an unbiased estimate of X's causal impact on Y.

This success suggests the following multivariate 2SLS procedure. Say we have k endogenous covariates $x_1, ...x_k$ and p instrumental variables $z_1, ...z_p$ with $p \geq k$. We regress each endogenous covariate on all of the instruments to get fitted values. Then we regress Y on the k fitted values. (Technical note: Any exogenous control variables are included in both stages.)

The result is an incredibly powerful econometric tool. Famous examples of instrument use include Angrist and Krueger (1992), Angrist and Levy (1997), and Levitt (1996).

3.4 Testing for exogeneity (the "J-test")

When used correctly, instruments are uniquely useful econometric assets. But they come with a very difficult-to-defend assumption: that they are uncorrelated with the error term (with omitted variables) in the second stage regression. Even in our Dillon gym, res-college example—even with an instrument that was randomly assigned—we were concerned about bias in the results from the ways in which residential colleges might affect fitness other than through proximity to Dillon.

Luckily, we have a way to test this assumption. But we need more instruments than we have endogenous covariates. (Now p is strictly greater than k.)

Intuitively, if we had one X and two Zs, we could run 2SLS with both of them. The math was clear: If both are valid instruments, we should get the same estimate for β_1. So if we get radically different estimates, then we know that one of them is biased.

You may already see the snag. If we get radically different estimates, we have no way to tell which is right.

In the case of multiple covariates and instruments, we make use of a χ^2 (chi-squared) distribution. The details are more technical than they are enlightening—failing the litmus test for inclusion in this section—but here is the gist. If a set of instruments are exogenous, then they are uncorrelated with the second-stage error term ϵ_i. So if we had access to ϵ_i, we could regress it on all of the Zs and run an F-test (which uses a χ^2 distribution as its null) to test that they are jointly zero. We do not, of course, have access to ϵ_i, but we can use the fitted values from the second stage and run the test. (We need more instruments than covariates because the null distribution ends up having $p - k$ degrees of

freedom.) This approach is variously called a J-test, Sargan's J-test, and an overidentification test. See Sargan (1958), Sargan (1975), and Fumio (2000).

We hit the same snag here, though. Because our fitted values for the error term come from a regression with all of the instruments, we can never know which instrument causes a failed result, and even if we get a passing result, we cannot know whether all the instruments are valid or just biased in roughly the same way.

3.5 Testing for weak instruments

An instrument Z needs a nonzero covariance with X for the mathematics even to work. But even a small nonzero covariance will cause very high-variance coefficient predictions, since it is in the denominator of the OLS estimator. (See Jaegar and Baker (1995) and Stock and Wright (2002).)

The rule of thumb is that if you regress all the instruments on an endogenous covariate, you should get a joint F-statistic greater than ten.

3.6 Finding instruments

The problem of locating instruments is of tremendous importance to economists. A strong instrument story can help build a career. How, from a computational perspective, can we automate the search for good instrumental variable candidates in a dataset?

Remember that a good instrument is one that is both exogenous and relevant. The relevance piece is easy. Any variable that is well-correlated with X will do.

At first, the exogeneity piece seems a bit hopeless. We have this exogeneity test at our disposal, but given a failure result it cannot tell us which instrument failed! And even worse, given a successful result, we cannot be sure whether all the instruments are valid or just equivalently biased.

Think back to our Dillon example. We suspect that residential college placement is uncorrelated with the amount of junk food students eat and therefore correlated with fitness only through Dillon, but given the proximities of Forbes to Wawa, Butler to Studio '34, and Rocky to Nassau, we are not so sure!

But let us imagine now that Dillon is rolling out a golden voucher program to encourage people to come (and to make some money on the side). Holders of a voucher can enter through a side door, change in a private locker room, and warm up in a private cardio room. To test the program, Dillon is randomly selecting a few hundred lucky Princeton students to get vouchers for free before anyone else has the option of buying one.

This instrument is, in my opinion, more clearly exogenous than our first one. It is both randomly selected and plausibly uncorrelated with any fitness behaviors other than Dillon attendance. So consider the following procedure: Run 2SLS with the residential college instrument, run 2SLS with the voucher instrument, and then compare the estimates. If they are roughly the same, then we know that our residential college instrument was exogenous all along!

In this example, of course, this is not very useful. (If we already have an exogenous instrument whose 2SLS estimate we trust, where is the need for validating another instrument?) But recall that in the multivariate case, we need multiple instruments. Imagine we had 12 endogenous covariates and one instrument known to be exogenous. At the moment, we cannot run an instrumental variable regression.

But consider the following selection approach. We have dependent variable Y, endogenous regressors $X_1...X_k$, instrument candidates $Z_1...Z_n$ for $n >> k$, and instrument Z_0 which is known for intuitive reasons to be uncorrelated with the variation in Y that is unexplained by the X variables. The goal is to find a lean set of strong, valid instruments from the dataset in a computationally efficient way.

The first aim is to find those instruments from the n candidates that satisfy our exclusion restriction. We run a J-test with the endogenous covariates and the $n+1$ instruments (including Z_0). Then we take out Z_1 and run it again; then we add Z_1 back and take out Z_2; then we add Z_2 back and take out Z_3; etc. Once we get to the end of this list, we take them out two at a time, and then three at a time, and off we go.

This approach is a good one because we know that the very first "pass" we get is the optimal bundle, meaning the most inclusive group of exogenous instruments possible.

Proof: We are trying to find the optimal subset O of the set $\{Z_1, Z_2, Z_3, ..., Z_n\}$. Imagine we are in the j^{th} round of pruning the group, and we get our first pass with some set S of $n-j+1$ instruments from the total $\binom{n}{j+1}$ round j tests. Suppose that $S \neq O$. There are three possible explanations for this.

1. $|O| > |S|$. But this cannot be true, because we would have tested O in round i for $i < j$. Since O by definition receives a pass when it is reached, $|O| \leq |S|$.

2. $|O| < |S|$. This cannot be true. O, by definition, is the most inclusive group of exogenous instruments in the dataset. So there is no valid subset R such that $|R| > |O|$. Therefore, $|S| \leq |O|$

3. $|O| = |S|$, but $O \neq S$. This implies that we can get to O from S by exchanging $b : b > 0$ instruments between S and S^c. Consider one such exchange: Z_u from S for a Z_v that is in both S^c and O. If Z_u is in S, it is exogenous; if Z_v is in O, it is also exogenous. Therefore, there is a larger set of exogenous instruments that includes all the other instruments in S, Z_u, and Z_v. This implies $|O| > |S|$, a contradiction.

Therefore, $S = O$.

Once one has the most inclusive group of plausibly exogenous instruments, LASSO can be used to find an optimally lean set.

4 Time Series

4.1 Motivation and stationarity

We observe variable X_t for $t = 1, 2, 3, ..., T$ Our goal is to predict X_{T+1} based on its previous values.

First, some notation. We call X_{t-1} a "lagged value" of X_t and X_{t-i} the "i^{th} lag" of X_t. We call a random process "stationary" if its distribution is constant over time. Mathematically, stationarity means that:

$$\begin{aligned} \boldsymbol{E}[X_t] =& \boldsymbol{E}[X_r] \text{ for all } t, r. \\ cov(X_t, X_{t-i}) =& cov(X_r, X_{r-i}) \text{ for all } t, r. \end{aligned} \tag{16}$$

As we will discuss, stationarity does not mean that the distribution of X_t cannot depend on its lagged values. We allow and in fact want this dependency. This is called serial correlation: When X_{r-1} is above $\boldsymbol{E}[X_t]$, X_r is also likely to be above $\boldsymbol{E}[X_t]$ (or perhaps below). So stationarity is really more of a broad distributional concept. The most precise definition I have seen for the word came to me from a lecture write-up from Princeton's Dr. Plagborg-Moller. Borrowing it here: $Y_t = \{Y_1, Y_2, ...\}$ is stationary if, for any p, the distribution of $\{Y_s, Y_{s+1}, ...Y_{s+p}\}$ does not depend on s.

4.2 Univariate autoregression

A p^{th}-order autoregression is denoted AR(p) and has the following model form with stationary stochastic process X_t.

$$X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_p X_{t-p} + \epsilon_t \tag{17}$$

Our assumption is that X_t depends on the p lagged values before it. We estimate the weights with ordinary least squares.

4.3 Testing stationarity

If the underlying distribution of X_t changes over time, we cannot impose on it a constant weight vector $\boldsymbol{\beta}$, since if we do we will have $\boldsymbol{E}[\epsilon_t|X_t] \neq \boldsymbol{E}[\epsilon_s|X_s]$ for $s \neq t$. This means that our usual OLS criteria are not met, but it may not be obvious when you have a first look at the data. The issue of non-stationarity is particular biting when we have multiple time series variables. "Spurious regression" occurs when OLS incorrectly indicates that independent non-stationary variables have a linear relationship. There is a great deal of literature surrounding this problem (see Granger (1974) and Kao (1999)), but we will not discuss it at length here because our product does not currently support any methods related to it. It is mentioned only to underscore the importance of having sophisticated methods for checking a process's stationarity.

In AR(1), the process is stationary if and only if $|\beta| < 1$. Intuitively, if $|\beta| > 1$, the process has a memory. A positive shock at $t = s$ will ripple further

and further for all $t > s$, amplifying the shock and changing the distribution. If $|\beta| < 1$, shocks are absorbed and X trends back to its expectation every time it deviates. Notice that even a standard stochastic random walk is not well-behaved enough to be considered a stationary process.

The story is more complicated in the AR(p) case. Consider the solutions to the following.

$$\beta_1 w + \beta_2 w^2 + \beta_3 w^3 + ... + \beta_p w^p = 1 \tag{18}$$

One can show that if all the solutions w, which may be complex, are "outside the unit circle" on the complex plane, the stochastic process is stationary. (See Shumway and Stoffer (2010) for explanation; original derivation unknown.) We are therefore interested in tests to tell us whether all the roots are inside the unit circle. The most common such "unit root test" comes originally from Dickey and Fuller (1979); its modern form is termed the "augmented Dickey-Fuller test." Denoting $X_s - X_{s-1} = \Delta X_s$, we consider the following p-lag model.

$$\Delta X_t = \lambda_0 + \lambda_1 X_{t-1} + \beta_1 \Delta X_{t-1} + \beta_p \Delta X_{t-p} + \epsilon_t \tag{19}$$

Next we find a p-value for the following test-statistic relative to the null distribution engineered by Dickey and Fuller.

$$\hat{t} = \frac{\hat{\lambda_1}}{SE(\hat{\lambda_1})} \tag{20}$$

The null hypothesis is that $\lambda_1 = 0$, or that the process is non-stationary because the lagged value X_{t-1} has no predictive value for ΔX_t beyond the p lagged differences. A rejection of this test is evidence that there is no unit root present, which means that the process is stationary and that we can comfortably carry on with autoregression.

Notice that if a process is non-stationary, there are still approaches that will allow us to do OLS:

$$\Delta ln(X_t) = ln(X_t) - ln(X_{t-1}) \tag{21}$$

A common one is considering log differences (as above) and seeing whether the random process $\Delta ln(X_t)$ is stationary.

4.4 Vector autoregression

Often we are interested in using the past values of other variables to predict X_t. Consider another time series variable Y_t, with analogous support on $t = 1, 2, 3, ...$ We use the following models.

$$\begin{aligned} X_t &= \beta_{1,0} + \beta_{1,1} X_{t-1} + ... + \beta_{1,p} X_{t-p} + \lambda_{1,1} Y_{t-1} + ... + \lambda_{1,p} Y_{t-p} + \epsilon_{1,t} \\ Y_t &= \beta_{2,0} + \beta_{2,1} X_{t-1} + ... + \beta_{2,p} X_{t-p} + \lambda_{2,1} Y_{t-1} + ... + \lambda_{2,p} Y_{t-p} + \epsilon_{2,t} \end{aligned} \tag{22}$$

This approach is called vector autoregression (VAR) because we are regressing a vector of time series variables on their past values. See Hamilton (1994).

4.5 Granger causality

Given the results of a multivariable time series model, how do we measure the correlation between several processes?

For this project we chose the "Granger causality" approach because it has considerable support from online libraries and because it is commonly applied in mathematical and engineering fields outside of economics. See Granger (1969) and Eichler (2012).

Consider the following multivariable time series model, with $k+1$ processes $\{X_t, Y_{1,t}, Y_{2,t}, ..., Y_{k,t}\}$. Each process has a unique number of lags.

$$\begin{aligned} X_t = \beta_0 &+ \beta_1 X_{t-1} + ... + \beta_p X_{t-p_1} \\ &+ \lambda_{1,1} Y_{1,t-1} + ... + \lambda_{1,p} Y_{1,t-p_1} \\ &+ \lambda_{2,1} Y_{2,t-1} + ... + \lambda_{2,p} Y_{2,t-p_2} \\ &+ \lambda_{3,1} Y_{3,t-1} + ... + \lambda_{3,p} Y_{3,t-p_3} \\ &... \\ &+ \lambda_{k,1} Y_{k,t-1} + ... + \lambda_{k,p} Y_{k,t-p_k} + \\ &+ \epsilon_t \end{aligned} \tag{23}$$

We are interested in whether variable $Y_{i,t}$ "Granger causes" X_t. We run an F-test the following joint null hypothesis:

$$\lambda_{i,1} = \lambda_{i,2} = ... = \lambda_{i,p_i} = 0 \tag{24}$$

In other words, the null states that the lags of $Y_{i,t}$ do not help with predictions of X_t. Notice that despite the name, Granger causality is just a correlational test and has debatably no power in determining causality. See Granger (1980), Granger (2004), and Mariusz (2015).

4.6 Lag length

Return to our simplest times series model: AR(p).

$$X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_p X_{t-p} + \epsilon_t \tag{25}$$

It is common—although, perhaps, philosophically dubious—to think of every well-behaved stochastic process as having a "true" lag length p. If a process's true memory is six time periods, then the coefficient on a seventh or eighth lagged value will equal 0.

In this view, p itself is a parameter of interest. How do we find a sensible estimate $\hat{p}$ using the data? There are three common approaches in economics and engineering information theory. Our product supports all of them.

4.5.6.i: Normal test statistic approach

Begin with a large model, perhaps AR(25). Test whether $\beta_{25} = 0$. If you cannot reject that it does not, move down to AR(24). Repeat until you find a significant coefficient.

This is an intuitive but flawed approach. If your p-value is 0.05, then 5% of the time you check a zero coefficient you end up accepting it, ending up with overly large models.

A fix to this problem: if you stop at AR(23), continue down and make sure that β_{22}, β_{21}, etc. also are significantly different from zero. If they are not, you may have good reason to suspect that that β_{23} was a false positive.

4.5.6.ii: Akaike Information Criterion
The Akaike Information Criterion (AIC) is a general measure of out-of-sample prediction error that is often used to determine time series lag lengths. (See Akaike (1974).) Given an AR(p) model with T time points, AIC is defined in the following way.

$$\begin{aligned} AIC(p) &= ln(\frac{SSR(p)}{T}) + (p+1)\frac{2}{T} \\ &= -2ln(\Lambda) + 2p \end{aligned} \tag{26}$$

$SSR(p)$ is the sum of squared residuals of AR(p) and $\Lambda = L(data|\hat{\boldsymbol{\beta}})$ is the log-likelihood function of the fit.

Given this definition, we choose $p = \hat{p_{AIC}}$, where:

$$\hat{p_{AIC}} = \underset{p\in[0,p_{max}]}{\text{argmin}} AIC(p) \tag{27}$$

The intuition here is that we are assigning a penalty term that penalizes including many lags. Finding the minimum $\hat{p_{AIC}}$ means finding a model that has good fit relative to the number of regressors (the number of lags).

4.5.6.iii: Bayesian Information Criterion The Bayesian Information Criterion (BIC) is analogous to AIC but with the following form.

$$\begin{aligned} BIC(p) &= ln(\frac{SSR(p)}{T}) + (p+1)\frac{ln(T)}{T} \\ &= -2ln(\Lambda) + 2ln(T)p \end{aligned} \tag{28}$$

It is sometimes called the Schwarz criterion; it came from Schwarz (1978). Again we are using a penalty term and choosing the minimum:

$$\hat{p_{BIC}} = \underset{p\in[0,p_{max}]}{\text{argmin}} BIC(p) \tag{29}$$

4.5.6.iv: Discussion of approaches
Even with the proposed updated choice heuristic, the test-statistic approach is not very systematic and so is not recommended by Athena. Still, though, it seems to be the first choice of many economists, so it is included in the product.

Both the AIC and BIC approaches are commonly used by sophisticated statisticians in different prediction scenarios. The BIC estimate has the advantage of being "consistent," meaning that under certain assumptions its expected value is the true lag length, which is not the case for AIC.

But AIC is preferred by Athena for a few reasons. First, the "true" lag length is largely a hypothetical object. Vrieze (2012) ran a simulated data experiment comparing AIC and BIC when the true length was included in the candidate models. Although BIC's expected value was more consistent in hovering around the true length, it was also more likely to pick a bad model than AIC, which was slightly biased upward but was consistently around the right answer. AIC tends to deliver lower mean squared error, making it a more precise tool than BIC. In addition, Stone (1977) showed that AIC is asymptotically equivalent to a cross-validation model choice procedure.

Another advantage of AIC, Burnham and Anderson (2002) note, is that its values are easily translatable to the estimated probabilities that the corresponding lags are the correct length. If we consider b models, denote their AIC values $AIC_1, AIC_2, ..., AIC_b$, and denote the minimum value among them AIC^*, then:

$$g_i := e^{\frac{1}{2}(AIC^* - AIC_i)} \tag{30}$$

g_i is the probability that i is the true lag length proportional to the probability that the the length corresponding to AIC^* is the true lag length. For example, if AIC^* was achieved by the q-lag model and $g_4 = 0.25$, then q is four times as likely as 4 to be the correct lag length.

For more information about this debate and trade-off, see Burnham and Anderson (2004) and Gelman (2017).

5 Selected other stochastic prediction techniques

5.1 Principle component analysis

Principle component analysis (PCA) is a common exploratory step in prediction procedures and machine learning problems across many sciences. It is particularly useful when there is a large amount of data and dimensionality reduction is desired, perhaps for visualization or clustering purposes. Though its applications are new, the approach itself is not. Credit for it tends to be given to Pearson (1901) and Hotelling (1933).

PCA takes data of many dimensions and orthogonally projects them to a new coordinate system, typically of many fewer dimensions. The procedure minimizes information loss by choosing coordinate axes that capture the most variance in the data.

Say we have a typical data matrix $\boldsymbol{D} \in \mathbb{R}^{n \times k}$, meaning we have n observations of k variables. We find $\boldsymbol{D}^T\boldsymbol{D}$ and then de-mean each column. It turns out that $\boldsymbol{D}^T\boldsymbol{D}$ is a covariance matrix of the our variables. Usually the next step—and at the moment Athena mandates this step—is to normalize the columns as

well by dividing by their standard deviations. (Theoretically if you had particular reason to believe that the high-variance features of the data were the most relevant, you might want to skip this step.) Call this normalized matrix $\boldsymbol{Z}$ and, since it is symmetric, orthogonally diagonalize it using its eigendecomposition:

$$\boldsymbol{Z} = \boldsymbol{S}\boldsymbol{X}\boldsymbol{S}^{-1} \tag{31}$$

Recall that the diagonal entries of $\boldsymbol{X}$ are the eigenvalues $(\lambda_1, \lambda_2, ...)$ of $\boldsymbol{Z}$. (These are the singular values of $\boldsymbol{D}$.) The corresponding columns of $\boldsymbol{S}$, which are the rows of $\boldsymbol{S}^{-1}$ since it is orthogonal, form the orthonormal eigenbasis for $\boldsymbol{Z}$ with each eigenvector corresponding (in index) to the diagonal element of $\boldsymbol{X}$.

The trick is that the largest eigenvalues correspond to the vector coordinates of the axes that capture the most variation in the data. Finding the largest eigenvalues in $\boldsymbol{X}$, build $\boldsymbol{S}^*$ by rearranging the eigenvector columns in order of the relative magnitude of the eigenvalues to which they correspond (meaning the eigenvector with the largest eigenvalue is now the first column).

The next step is to calculate the new data matrix by changing its coordinate space.

$$\boldsymbol{Z}^* = \boldsymbol{Z}\boldsymbol{S}^* \tag{32}$$

At this point a user needs to choose how many of these "principle components" she would like to keep. This decision is most often informed both by the user's goals (if it is visualization, she may just keep two components) and by the variance of the coordinate variation itself. If the data's information is distributed relatively evenly across the orthonormal basis of $\mathbb{R}^k$, then we may want to keep as many dimensions as possible. Note that we can judge the information loss from dropping a coordinate dimension directly. Denoting the variance explained by component j as VAR_j, we can leverage the following relationship.

$$\text{VAR}_j = \frac{\lambda_j}{\sum_{i=1}^{k} \lambda_i} \tag{33}$$

5.2 Poisson regression (GLM)

An approach falls into the generalized linear model (GLM) category if it loosens the common OLS restriction that error terms be Gaussian. Poisson regression is one such model. In this regression approach, the dependent variable is assumed to be distributed Poisson, and an MLE technique is used.

Recall that a counting process is Poisson if an "arrival" is no more likely at any one point than any other. One can derive the Poisson distribution by taking the infinitesimal limit of a Bernoulli process. Speaking informally, a Poisson process "flips a coin" every instant for whether that instant will carry an arrival, with the probability of an arrival equaling at every instant the "right type" of 0 such that the expected value of the process over some amount of time is some well-behaved finite number.

Therefore, the response variable Y must be a counting process (all y from $Y \in \mathbb{Z}^{\text{nonneg}}$), and it is assumed to have the following property for all $k > 0$ and for

some λ.

$$\mathbb{P}(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{34}$$

We assume that the parameter λ is determined by the covariates $X_1, ...X_k$. The relationship may be direct or log-linear. The following two models are both common. Athena uses the first.

$$\begin{aligned} \lambda &= \boldsymbol{\beta}^T \boldsymbol{X} \\ \lambda &= e^{\boldsymbol{\beta}^T \boldsymbol{X}} \end{aligned} \tag{35}$$

Next, we maximize the usual log-likelihood. If we observe $data = \boldsymbol{d} \in \mathbb{R}^n$:

$$\mathbb{P}(d_i|\boldsymbol{\beta}) = \frac{(\boldsymbol{\beta}^T \boldsymbol{X})^{d_i} e^{(-\boldsymbol{\beta}^T \boldsymbol{X})}}{d_i!} \tag{36}$$

Then we get to the following, where the log factorial term $-log(d_i!)$ is dropped because it does not change in $\boldsymbol{\beta}$.

$$\mathbb{P}(d_i|\boldsymbol{\beta}) = L(\boldsymbol{d}, \boldsymbol{\beta}) = \sum_{i=1}^{n} (d_i log(\boldsymbol{\beta}^T \boldsymbol{X}) - \boldsymbol{\beta}^T \boldsymbol{X}) \tag{37}$$

5.3 Hidden regimes (Markov switching regression)

Recall that Markov chains are used to describe stochastic processes with short memories. Say $X_t \in D \; \forall \; t$, where D is a finite set of integers. (Each element of D is typically dubbed a "state" or "regime.") We call X_t a Markov variable if for all t, all k, and all possible lagged values:

$$\mathbb{P}(X_t = k|X_{t-1}, X_{t-2}, X_{t-3}, ...) = \mathbb{P}(X_t = k|X_{t-1}) \tag{38}$$

That is, the likelihood of X_t being in any state s depends on only which state X_{t-1} was in. For a time-homogeneous Markov chain, where the probabilities do not change, it is common to represent these transition probabilities in an object called a stochastic transition matrix. If there are two states (like $D = \{1, 2\}$), it may look like the following.

$$\boldsymbol{P} = \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix} \tag{39}$$

By convention, $\boldsymbol{P}_{i,j} = \mathbb{P}(j|i)$. So here, the probability of going to state 2 from state 1 is 0.7, and the probability of remaining in state 1 is 0.3. Of course, the probability that, say, $X_3 = 1$ depends on whether X_1 equaled 1 or 2. But for some Markov processes (the ones we are typically interested in), over time this probability of being in a certain state ceases to depend on the initial state. Consider the limit:

$$\pi(s) := \lim_{t \to \infty} \mathbb{P}(X_t = s|X_1 = i) \text{ , for } s, i \in D \tag{40}$$

We call a chain "ergodic" if this limit exists for each state s and does not depend on i. It turns out that a Markov chain is ergodic if and only if it is irreducible (meaning all states are reachable from all other states, perhaps in multiple moves) and aperiodic (meaning it does not alternate between subsets of D with probability 1).

The hidden regime regression model, proposed first by Hamilton (1989), assumes that a time series variable Y is governed by a Markov process, typically an ergodic one. (We tend to switch to the word "regimes" to describe states of this model.) So Y switches from state to state at every observation with certain transition probabilities, and at every state it has a certain distribution. The aim of the model is to estimate these quantities.

If we assume 2 regimes and 1 lag, the model looks like this.

$$\begin{aligned} &\text{In state 1: } y_t = \mu_1 + \beta_1 y_{t-1} + \epsilon_t \\ &\text{In state 2: } y_t = \mu_2 + \beta_1 y_{t-1} + \epsilon_t \end{aligned} \tag{41}$$

Here, μ_1 and μ_2 are the state-specific intercepts of the model, and β_1 is the usual AR(1) lag coefficient. If the timing of the regime switches is known, this reduces to two simple OLS procedures. But in the case of interest, we do not know when there is a switch in regime. So we introduce a random variable s_t that corresponds to the state of Y at time t and consolidate to a more general model.

$$y_t = \mu_{s_t} + \boldsymbol{\beta}^T \boldsymbol{x} + \boldsymbol{w}^T \boldsymbol{z} + \epsilon_t \tag{42}$$

Note that $\boldsymbol{x}$ is a vector of variables with state-invariant weights $\boldsymbol{\beta}$, and $\boldsymbol{z}$ is a vector of variables with state-dependent weights $\boldsymbol{w}$. To build the likelihood function, we need to estimate the probability that $s_t = k$ given y_t and the model parameters $\boldsymbol{\alpha}$.

$$\mathbb{P}(s_t = k | y_t; \boldsymbol{\alpha}) = \frac{f(y_t | s_t = k, y_{t-1}; \boldsymbol{\alpha}) \mathbb{P}(s_t = k | y_{t-1}; \boldsymbol{\alpha})}{f(y_t | y_{t-1}; \boldsymbol{\alpha})} \tag{43}$$

We get this result by applying Bayes' theorem and conditioning on the one lag y_{t-1}. We desire the log-likelihood function:

$$L(\boldsymbol{\alpha}) = \sum_{t=1}^{T} log f(y_t | y_{t-1}; \boldsymbol{\alpha}) \tag{44}$$

To find the terms of the sum, we need to combine the equation above with another explicit formulation of the conditional state probability.

$$\mathbb{P}(s_t = i | y_{t-1}; \boldsymbol{\alpha}) = \sum_{j=1}^{k} \mathbb{P}(s_t = i | s_{t-1} = j, y_{t-1}; \boldsymbol{\alpha}) \mathbb{P}(s_t = j | y_{t-1}; \boldsymbol{\alpha}) \tag{45}$$

This model and its variants are most sanely employed not just when there are underlying shifts in circumstance present throughout the data but when

those shifts are predictable and informative in terms of their order. Initial testing of Athena with economics peers revealed an over-eagerness to estimate with hidden regimes. Owing to its complexity and nonlinearity, the results of this model are not as easy to interpret economically as OLS, so it is better saved for situations in which prediction accuracy is more important than establishing a causal link between economic variables. When it is possible, users are encouraged to include states explicitly as dummy variables. Users should also remember that time-homogeneous Markov states and transition probabilities are essentially meaningless for non-stationary variables.