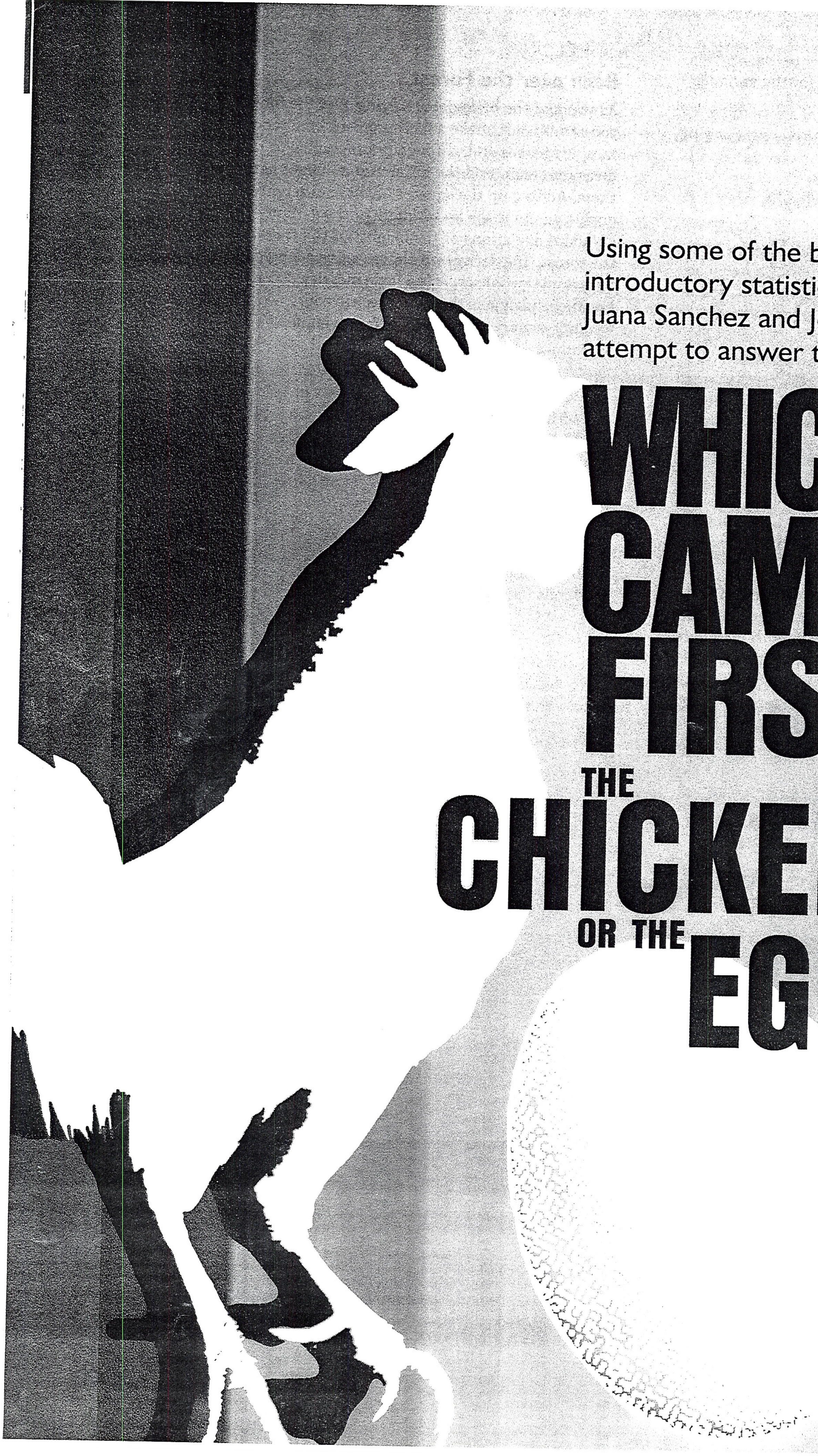

Using some of the basic concepts of introductory statistics, statisticians Juana Sanchez and Jean Wang attempt to answer the question,

WHICH CAME FIRST THE CHICKEN OR THE EGG?

Chickens play a huge role in our lives. As well as sometimes acting as pets or ending up sitting on plates, chickens show up in pop culture. They had feature roles in the hit movie "Chicken Run"; they frequently guest star in popular games such as "The Legend of Zelda" and "Final Fantasy"; and they take center stage in the classic dilemma of causality: "Which came first, the chicken or the egg?"

Let's use statistics to solve this dilemma. We will not delve into the philosophical issues. For that, you can watch last year's CBS News Video, "Was the Chicken or Egg 1st?" Rather, let's use time series analysis of historical data from the United States chicken industry to shed some light on the direction of causality.

A time series is a sequence of observations that are ordered in time. Some common examples include daily temperatures, weekly stock prices, and monthly employment figures. In a time series, the value of a variable for today often depends on its value in the past. In effect, the variable has some memory of its past and, perhaps, some memory of the past of other variables. Thus, the nature of time series data is different from the data usually studied in statistics courses—the observations are not independent. But, like many other datasets, time series data can be explained with models. In order for a data series to be correctly modeled using time series analysis, it must be stationary. A stationary time series is one who's mean, variance, and covariances do not change over time.

In our study, the question of causality can be rephrased as "Does the chicken depend on the egg, or does the egg depend on the chicken?" We can use a vector autoregressive model of chicken and egg data to answer that question. Although this and other time series data analysis methods are advanced concepts, the basic ideas come from the fundamentals of estimation, hypothesis testing, p-values, and regression analysis.

The United States Chicken Industry

The main products of the poultry industry in the United States are broilers (chicken for meat) and table eggs (eggs for cooking). Being the world's largest producer of poultry meat, 14% of the total United States' annual poultry production is exported. The United States is also the second-largest egg producer in the world.

Although it has become a highly specialized agricultural business nowadays, the commercial poultry industry was made up of millions of small backyard farms before the 1950s, when meat was a byproduct of egg production. Today, poultry products account for about 10% of all farm revenue, and the industry has been transformed almost completely from a fragmented, home-owned industry to a highly organized, vertically integrated industry linking all production decisions from farm to market.

Chicken and Egg Time Series Data

One of the two variables we can study is monthly chickens hatched with the intended purpose of becoming broilers. We will call this variable *hatched*. The other variable we can study is *eggs*, but not table eggs; rather, we will study broiler eggs.

Often the best way to begin a statistical study is to plot the data, and time series analysis is a good example of this. So, let's look at the plots in Figure 1. The data span the years between 1975 through 2002. Because the values of the time series, *hatched* and *eggs*, display an upward trend with inconsistent variability, the time series are not stationary. This is common for real-world time series data. There is also obvious seasonality in the data—a repeating periodic effect at approximately the same time each year.

For many nonstationary time series, the trend can be removed by differencing the data (i.e., subtracting consecutive values of the variable.) Then, the model is built using the changes in the variable from time period to time period, instead of the original values of the variable.

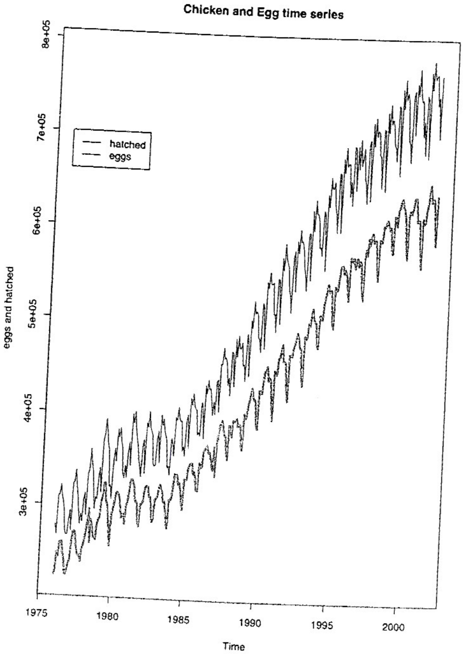

FIGURE 1. Number of broiler chickens hatched (top) in thousands and number of broiler eggs in incubators (bottom), in thousands, on the first day of the month

Differencing

Differencing is an easy and effective method to help stabilize a nonstationarity time series. Simple differences are differences taken one period apart. Seasonal differences are differences taken 12 periods apart. In some time series, we need to do both simple and seasonal differencing.

After simple and seasonal differencing, we can see in Figure 2 that *hatched* and *eggs* fluctuate around a constant mean of zero and the variance looks relatively stable with a few extreme values.

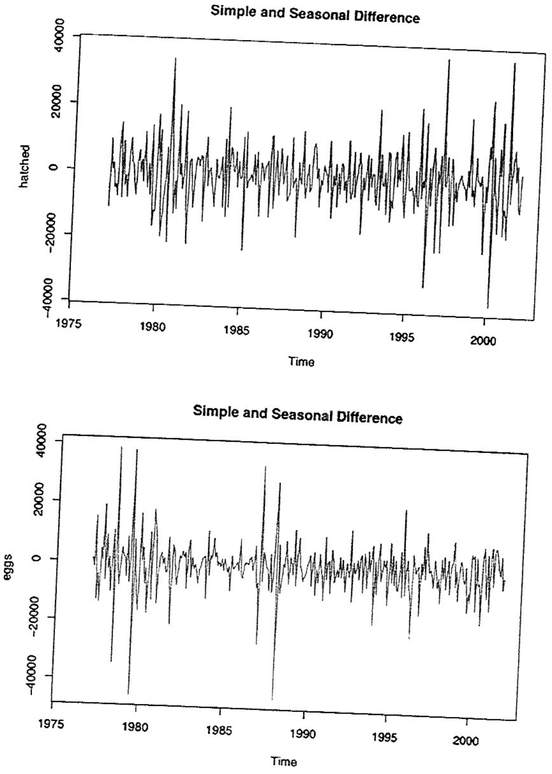

FIGURE 2. Stationary time series after simple and seasonal differencing for *hatched* and *eggs*

But in time series analysis, we do not rely on only our eyes. We also look at two plots that play a prominent role in understanding a time series: the autocorrelation function and the partial autocorrelation function.

Autocorrelation is the association between values of the same variable over time. The autocorrelation coefficient (ρ_k) measures the autocorrelation between two values in a time series k time periods apart. The autocorrelation function (ACF) plots the autocorrelation coefficents values for k from 0 and up.

Partial autocorrelation is the association between times series values separated by k time periods with the effects of the intermediate observations eliminated. A plot of these values is the partial autocorrelation function (PACF).

Autocorrelation Functions

If we want to know whether the "memory" of a time series goes as far back as k months, that is whether the value of *hatched* this month depends on what happened k months ago, then we can test:

$$H_0 : \rho_k = 0$$

$$H_a : \rho_k \neq 0$$

where ρ_k is the autocorrelation between the value of the series at time t and its value k periods before. In the graph in Figure 3, at each lag k (the vertical axis numbers), we test this null hypothesis. The two bands in the graphs represent two standard errors in the sampling distribution of the sample autocorrelation coefficient r_k. If a spike is past two standard errors, this means the p-value for the test at that lag k is smaller than 0.05, and, therefore, we can reject the null hypothesis. If so, the spike is significant, and we say there is memory or correlation between values of the variable at time t and at time $t-k$.

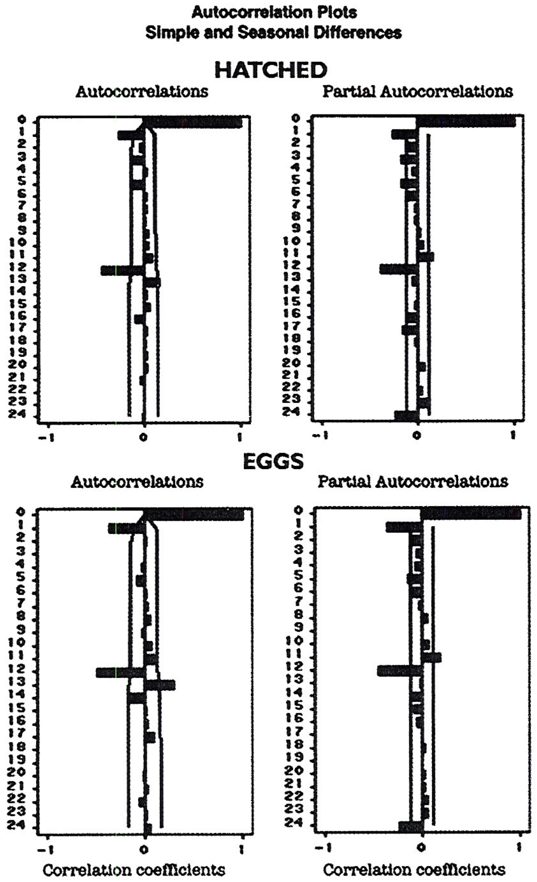

FIGURE 3. Sample autocorrelation and partial autocorrelation functions (sample ACF and sample PACF) for the variables *hatched* and *eggs*

Looking at the sample autocorrelation and partial autocorrelation functions in Figure 3, we can see that the ACF has a significant spike at lag 1 and the PACF exponentially dies down. Therefore, based on these traits, the model for the variable *hatched* is a moving average process of order 1 for nonseasonal lags. However, as there is also a spike in the ACF at 12 months and the PACF shows the seasonality dying down at 24 months, we also have a moving average model at the seasonal lag. Our final model for the variable *hatched* is MA (1, 12), which is shorthand for saying it is a moving average model using lagged variables at one month and 12 months.

Looking at the other time series, *eggs,* the original values display an upward, positive trend, so it also is not stationary (see Figure 1). We can transform *eggs* by taking simple and seasonal differences so the values of the time series fluctuate around a constant mean of zero for stationarity (see Figure 2). The sample autocorrelation and partial autocorrelation functions for the differenced *eggs* data look similar to the ACF and PACF for differenced *hatched* data (see Figure 3). That indicates the model to use for *eggs* is the same as for *hatched:* MA (1, 12).

We are lucky; the change in *eggs* and the change in *hatched* follow the same model. This will make it easier for us to find which comes first, the chicken or the egg. But we are not quite ready for that yet.

Vector Autoregression

The question of which variable leads—precedes in time—in the movement of two stationary time series has been studied often in economics. For example, in the context of predicting stock market prices, one question can be whether the price of a stock that trades in both the United States and, let's say, Germany, is such that the United States price leads the German price or the German price leads the United States price during overlapping trading periods (i.e., during the hours both markets are simultaneously open).

Questions such as this can be answered with a technique used in econometric analysis called vector autoregression (VAR). It is a method that can help us determine the time precedence between variables. If we find that one variable consistently precedes another in time, that would be evidence supporting a possible causal relationship.

In our case, we want to see whether the number of broilers *hatched* causes the number of broiler *eggs* or the number of broiler *eggs* causes the number of broilers *hatched.* For this, we need two regression equations: one to regress *hatched* on *eggs* and the other to regress *eggs* on *hatched.* As this is a time series model, we want to estimate these two regression equations taking into account that there could be causality in either direction. So, we estimate the two equations together with a common variance-covariance matrix for both. This is different from separately estimating them.

ECONOMETRICS

Econometrics literally means "economic measurement." It is the branch of economics that uses statistical methods to study empirically relationships in economic data. Regression analysis and time series analysis provide the foundation for econometric investigations.

VECTOR AUTOREGRESSION

Vector autoregression is a technique in the econometrician's tool kit for analyzing the dynamic properties of an economic system. One application is to estimate the direction of a possible causal relationship between two variables, X and Y. Least squares regression is used to examine the autocorrelation ("self-correlation") due to the time-dependence of each of the variables. First, the analysis is performed with X as the dependent variable and its historical values and the Y values as the independent variables. Then, the process is repeated with Y as the dependent variable with its historical values and the X values as independent variables. Comparing the results of the two regressions can show the direction of possible causality.

A simple vector autoregression for our problem would be a model such as this:

$$X_{1t} = \phi_{11}X_{1,\,t-1} + \phi_{12}X_{2,\,t-1} + \varepsilon_{1t}$$
$$X_{2t} = \phi_{21}X_{1,\,t-1} + \phi_{22}X_{2,\,t-1} + \varepsilon_{2t}$$

where X_{1t} is *hatched* and X_{2t} is *eggs*, both variables are stationary with a mean of zero, and ϕ_{ij} are constants we estimate by regression.

Looking at the above model, we notice that if ϕ_{12} is zero, but ϕ_{21} is not zero, there is no feedback from X_2 to X_1. Thus, X_{1t} (*hatched*) does not depend on the lagged value of *eggs*, but X_{2t} (*eggs*) does depend on the lagged value of *hatched*. This would indicate any causality goes in only one direction.

Vector Autoregressive Model for *Hatched and Eggs*

Based on the structure we found in the sample ACF and PACF, the bivariate VAR model is:

$$hatched_t = -0.2391\ hatched_{t-1} - 0.0043\ eggs_{t-1} - 0.3768\ hatched_{t-12} - 0.0956\ eggs_{t-12}$$

p-value = 0.000 p-value = 0.9464 p-value = 0.0001 p-value = 0.1331

$$eggs_t = 0.1237\ hatched_{t-1} - 0.3614\ eggs_{t-1} + 0.0543\ hatched_{t-12} - 0.5001\ eggs_{t-12}$$

p-value = 0.0169 p-value= 0.0001 p-value = 0.2904 p-value = 0.0001

where $hatched_t$ is the value at time t of the seasonal difference of the first difference for the variable *hatched* and $eggs_t$ is the value at time t of the seasonal difference of the first difference for the variable *eggs*. The p-values of the coefficients correspond to the test of the null hypothesis that a coefficient is equal to zero. A p-value ≥ 0.05 means the coefficient is not significantly different from zero.

Comparing the p-values highlighted in orange, we can see that *hatched* last month (one-month lag) affects *eggs* this month, but no lag of *eggs* affects *hatched* in the present month. This means that while the number of broilers previously *hatched* affects the number of broiler *eggs* in incubators now, the number of broiler *eggs* incubating previously does not affect the number of broilers *hatched* now.

In "economic-speak," that means the number of chickens *hatched* is a leading indicator for the number of *eggs*, but the number of *eggs* is not a leading indicator for the number of chickens *hatched*. So, in our dilemma of causality, the chicken comes first!

Our conclusion makes sense in the economic context. A downturn in broilers is probably an indication of a sluggish chicken meat market, perhaps due to factors such as a recession in the economy, maybe some pandemic of avian flu, or some other economic factor. If this is the case, it does not make economic sense to keep the number of eggs in incubators at the previous level. Why incubate eggs that will give chickens that will not be sold? Consequently, we would expect the number of eggs in incubators to go down. So, as the demand for chicken meat goes up or down, the number of eggs in incubators should follow.

Chicken or the Egg?

As we have seen, time series analysis is fun and makes use of the basic concepts we learn studying introductory statistics: estimation, test of hypotheses, p-values, and regression. We just adapt the basic principles to the circumstances present in a time series modeling problem.

For chicken farmers, it is useful to know that the number of chickens hatched is a leading indicator for the number of eggs in incubators. For all of us concerned with the dilemma of causality, it is nice to see how quantitative methods can help us answer a classic dilemma: Which comes first, the chicken or the egg? Conditional on the vector autoregressive model that we used, in the economic decision chain of the United States poultry industry, the data indicate that chickens come first. ◕