Time Series Analytics Notes
Autoregression

The fundamental idea in forecasting is that the future depends upon the past. If we are sufficiently clever, we can figure out how. If the future does depend upon the past, then it is reasonable to expect that the immediate past – when it was still the future – depended upon its own past in much the same way as the actual future depends upon the past. So we can look at past values and ask how they depended upon *their* past in order to figure out how our future depends upon our past: To figure out how tomorrow's *unknown* price depends upon today's price, we can investigate how today's *known* price depends upon yesterday's price, and how yesterday's *known* price depends upon the price the day before, etc. This approach has the advantage of providing us with day-ahead values to analyze that are *known* – unlike the real future. It is in this sense that time series analysis is a *supervised* learning technique: The dependence of the present upon the past provides the supervision for learning how the future depends upon the present and the past.

Autoregression is a time series model that directly capitalizes upon the thinking of the preceding paragraph. In autoregression, the value of a time series depends upon its own past in a regression equation.[1] The simplest form of an autoregression is a univariate linear regression in which the value is a linear function of its immediately preceding value, with error:

$$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t,\ t = 1,2,...,n \quad \text{[Eqn 1]}$$

The errors ε_t must satisfy the requirements of a standard regression model. That is, the errors should be independent, identically distributed zero-mean normal random variables.[2] This means that the errors should be a zero-mean normal Random Sample. Formally, the errors can be defined as:

$$\varepsilon_t = Y_t - (\alpha + \beta Y_{t-1}),\ t = 1,2,...,n \quad \text{[Eqn 2]}$$

Note that the errors as defined in Eqn 2 result simply from solving Eqn 1 in terms of ε_t.

Definition. A **simple autoregression** is a time series $Y_1, Y_2, ..., Y_n, ...$ that satisfies the model in Eqn 1 and in which the errors $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n, ...$ are a zero-mean normal Random Sample.[3]

---

[1] A regression that relates a response variable to one or more of its own lagged values is also called a *dynamic* model.

[2] Dynamic models, in which past values of the response variable are also used as predictors, have an additional concern that does not arise in standard regression models, in which the predictors are completely different from the response. Namely, in a standard regression, the *x*'s are treated as though they are constants; but in a dynamic model the same values (the *y*'s) are used on both sides of the equation, although lagged on the right. On the left, these values are treated as random; on the right, they are treated as constants. This is logically inconsistent. In practical terms, it creates biased estimates. However, if the sample size is sufficiently large, and if the standard LHI(N) regression specifications hold, then the bias is small and can be ignored. See the Sidebar below for more discussion of this point.

[3] Normality of the errors is not essential for many inferences in autoregression, provided the sample size is sufficiently large, for there are versions of the Central Limit Theorem that apply to autoregression (and to regression more broadly). However, like the regular CLT, these regression CLTs do not apply to inference about individual values,

Thus, the simple autoregression is another time series model that has a RS hidden within it. In the RW, the hidden RS is the period-to-period changes in value. In autoregression, the hidden RS is the residual errors.

Using the Autoregression Model to Estimate the Mean of the Distribution of Y_t

Suppose that the autoregression model Eqn 1 has been validated:

$$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t,\ t = 1,2,...,n$$

This means that the errors ε_t are a zero-mean normal Random Sample. Suppose that Y_t has been regressed upon Y_{t-1} and the OLS coefficient estimates are $\hat{\alpha}$ and $\hat{\beta}$. Then the best estimate of the mean of the uncertainty distribution of Y_t is the plug-in estimate $E(Y_t) = \hat{\alpha} + \hat{\beta} Y_{t-1}$.

We would expect the actual mean to differ from the estimate by approximately the amount of the standard deviation of the uncertainty distribution of $\hat{\alpha} + \hat{\beta} Y_{t-1}$. This standard deviation is best left to statistical software to estimate. Sometimes this standard deviation is approximated by $RMSE / \sqrt{n}$. [4] But the latter is actually a lower bound that is exactly correct only for an estimate made at the mean of the predictor variable and is often rather far off. Most regression software has an option to print a more precise estimate of the standard deviation. For example, in SAS the options `p CLM` in the code

```
proc reg data=apts;
  model rent = area bathrooms / p CLM;
run;
```

cause printing (`p`) of 95% confidence intervals for the mean (`CLM`) *Y* (`rent`) for each combination of the *X*'s (`area bathrooms`) present in the dataset. By the Central Limit Theorem for regression, the confidence percentage is approximately correct if the sample size is sufficiently large, even if the residuals are not normally distributed. These intervals are calculated using the more precise standard deviation and also the more precise T distribution.

Using the Autoregression Model to Forecast the Next Value Y_{t+1}

Suppose that the autoregression model Eqn 1 has been validated:

$$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t,\ t = 1,2,...,n$$

This means that the errors ε_t are a zero-mean normal Random Sample. Suppose that Y_t has been regressed upon Y_{t-1} and the coefficient estimates are $\hat{\alpha}$ and $\hat{\beta}$. Then the best estimate of the mean of the uncertainty distribution of the next value Y_{t+1} is the plug-in estimate

[4] The idea of this approximation is that outside regression, the variability of individuals is σ and the variability of means is $\sigma / \sqrt{n}$; so inside regression, if the variability of individuals is RMSE, then the variability of means should be $RMSE / \sqrt{n}$, by analogy. But the analogy is a little off.

$E(Y_{t+1}) = \hat{\alpha} + \hat{\beta}Y_t$. The best forecast of the next value is the estimated mean of the distribution of the next value Y_{t+1}. So the forecast of the next value Y_{t+1} is the plug-in estimate $\hat{Y}_{t+1} = \hat{\alpha} + \hat{\beta}Y_t$.

We would expect the actual next value to differ from the estimate by approximately the amount of the standard deviation of the uncertainty distribution of Y_{t+1}. This standard deviation can be estimated by the root mean-square-error (RMSE) of the regression, which is approximately the standard deviation of the residuals. So

$$\text{Forecast} = \hat{Y}_{t+1} = \hat{\alpha} + \hat{\beta}Y_t \pm \text{RMSE}$$

The error margin ± RMSE provides approximately a 68% prediction (confidence) interval for the next individual value *if the N specification of the regression model holds*. RMSE is a bit of an underestimate for the margin of error. Most regression software has an option to print a slightly better estimate of the margin of error. For example, in SAS the options `p CLI` in the code

```
proc reg data=apts;
  model rent = area bathrooms / p CLI;
run;
```

cause printing (`p`) of 95% confidence/prediction intervals for each individual (`CLI`) observation in the dataset. These intervals are calculated using the more precise standard deviation and also the more precise T distribution and assume approximate normality of the residuals.

Two Special cases of Autoregression

Suppose that Y_t is an autoregressive time series. Thus,

$$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t,\ t = 1,2,...,n$$

and the errors ε_t are a zero-mean normal RS.

Case 1. Suppose that $\beta = 1$. Then $Y_t = \alpha + 1Y_{t-1} + \varepsilon_t$. Hence, $Y_t - Y_{t-1} = \alpha + \varepsilon_t$. That is, the changes in the time series are a constant + a zero-mean normal RS. But a constant + a zero-mean RS is still a RS! Adding a constant α does not make a RS unlevel, nor does it change the standard deviation, nor affect the independence. So the changes are a RS. Thus, if $\beta = 1$, the autoregression is a RW.

Case 2. Suppose that $\beta = 0$. Then $Y_t = \alpha + 0Y_{t-1} + \varepsilon_t$. Hence, $Y_t = \alpha + \varepsilon_t$. That is, the time series is a constant + a zero-mean normal RS. But a constant + a zero-mean RS is still a RS! Adding a constant α does not make a RS unlevel, nor does it change the standard deviation, nor affect the independence. Thus, if $\beta = 0$, the autoregression is a RS.

These two simple observations give us further insight into the roles of the RS and RW: They are at opposite ends of a spectrum of possible ways that a time series can depend upon the past!

- In the RW, there is complete dependence upon the past, with new values being only independent random variation around the past value.
- In the RS, there is no dependence upon the past; new values are drawn independently of and without regard for the values of the past.

Next, consider an autoregression in which $\beta \neq 1$ or 0. In this case, there is partial dependence upon the past. Furthermore, $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$ implies that $Y_t - Y_{t-1} = \alpha + (\beta - 1)Y_{t-1} + \varepsilon_t$. Since we are assuming that $\beta \neq 1$, then the left-hand-side changes in the time series depend partially upon the past. In this case, the past has useful information about how the time series will change in the future. In the case of the RW, the past has no useful information about the value of future changes. It is important to be able to tell which time series have useful information about the future direction of a time series and which do not. Testing whether $\beta = 1$ is therefore an important task. Tests of the hypothesis that $\beta = 1$ are called **unit root tests**. There are a number of such tests. The best-known is the Dickey-Fuller test.

RW?? RS (ß = 0 implies no dependence on past?)

In the current context, however, a simple regression test of H_0: $\beta = 1$ may be good enough. This is a T-test. To perform the test, run the regression of Y_t on Y_{t-1} and verify that the residuals satisfy L,H,I,(N). Get the estimate $\hat{\beta}$ of the slope, and the estimated standard error $\hat{\sigma}_\beta$ of the slope. Then calculate the T-statistic = $\dfrac{\hat{\beta} - 1}{\hat{\sigma}_\beta}$. Reject H_0: $\beta = 1$ if the T-statistic is too far from 0 [5] (rule of thumb: < -2 or > +2, which correspond roughly to a p-value of 0.05.) This tests whether the *direction* (change) of the time series can be forecast from the past. (Yes, if $\beta = 1$ can be rejected.)

?WHAT? ??WHY??

To test whether the *level* of the time series can be forecast from the past, test H_0: $\beta = 0$. This is a T-test. To perform the test, run the regression of Y_t on Y_{t-1} and verify that the residuals satisfy L,H,I,(N). Get the estimate $\hat{\beta}$ of the slope, and the estimated standard error $\hat{\sigma}_\beta$ of the slope. Then calculate the T-statistic = $\dfrac{\hat{\beta} - 0}{\hat{\sigma}_\beta} = \dfrac{\hat{\beta}}{\hat{\sigma}_\beta}$. Reject H_0: $\beta = 0$ if the T-statistic is too far from 0 (rule of thumb: < -2 or > +2, which correspond roughly to a p-value of 0.05.) Conclude that the level of the time series can be forecast if $\beta = 0$ can be rejected. Every software program that does regression prints out the results of this test.

Example. At ground level, ozone is a major air pollutant. Environmental authorities would like to be able to predict ozone levels in order to issue pollution advisories and implement ameliorative measures. To what extent do daily ozone levels depend upon their immediate past? The following is some core SAS output of a regression of daily ozone on its lag, i.e. a first-order autoregression. The data are from Houston for a 4-month period in summer, 1982 – a time of very high pollution. Ozone is measured in parts per hundred million.

The estimated model is $\text{Ozone}_t = 6.87446 + 0.40419\ \text{Ozone}_{t-1}$. The t-Value (4.82) rejects the hypothesis that $\beta = 0$. The hypothesis that $\beta = 1$ is also rejected, with a t-Value of (0.40419 – 1)/0.08381 = -7.1091. Thus, this model is neither a RS nor a RW. It is an autoregression, in which tomorrow's ozone depends partly upon today's ozone. On average, the value of ozone can be forecast to within about ±5 (4.9991) parts per 100,000,000. So if today's

This is the RMSE (b/c Forecast). 68% Confidence in prediction

[5] If the T-statistic is too far from zero, then $\hat{\beta}$ is too far from 1.

ozone level is 10, the forecast of tomorrow's ozone would be 6.87446 + 0.40419*10 = 10.9164 ±4.9991. The current ozone level explains (informally speaking) about Rsquare = 16.24% of tomorrow's value. So there is ample room to improve this model by incorporating other predictive variables.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	581.27172	581.27172	23.26	<.0001
Error	120	2998.93320	24.99111		
Corrected Total	121	3580.20492			

Root MSE	4.99911	R-Square	0.1624
Dependent Mean	11.54918	Adj R-Sq	0.1554
Coeff Var	43.28542		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	6.87446	1.06976	6.43	<.0001
lagOzone		1	0.40419	0.08381	4.82	<.0001

An Alternative View of Autoregression

Consider the model

$$Y_t = \alpha + u_t,\ t = 1,2,...,n \quad \text{[Eqn 3]}$$

in which the u_t are error terms but are *not* a zero-mean normal RS. Instead, the errors exhibit autoregressive behavior of their own:

$$u_t = \rho u_{t-1} + \varepsilon_t,\ t = 1,2,...,n \quad \text{[Eqn 4]}$$

in which the ε_t error terms are a zero-mean normal RS. This model still contains a hidden RS (ε_t), but it is hiding one more level further down in the model specification than we have previously considered.

- Eqn 1 is a model for a time series in which the observed Y_t time series exhibits autoregressive dependence upon its own past.
- Eqn 3 and Eqn 4 provide a model for a time series in which the unobserved error terms u_t exhibit autoregressive dependence upon their own past.

What is the relationship between the Eqn 1 autoregression and the Eqn 3/Eqn 4 autoregression? I will next demonstrate that they are broadly equivalent. This is an important observation because it provides an alternative parameterization of autoregressive models: Instead of estimating coefficients of lags of *Y*, we can estimate coefficients of lags of the errors. This is the approach that SAS takes to estimating autoregressive models in its PROC AUTOREG. It also gives further insight into the nature of autoregressive models: The errors in autoregression contain hidden explanatory power that can be extracted by regressing the errors on their lags. In other words,

autocorrelation can be viewed as resulting from the omission of explanatory variables from the regression.

First, suppose that time series model Eqn 3/Eqn 4 holds. Plug Eqn 4 into Eqn 3 to get $Y_t = \alpha + u_t$ = $\alpha + \rho u_{t-1} + \varepsilon_t$ [6]. Then lag 1 time period in Eqn 3 to get $Y_{t-1} = \alpha + u_{t-1}$. Multiply the latter by ρ to get $\rho Y_{t-1} = \rho\alpha + \rho u_{t-1}$. Subtract the latter from $Y_t = \alpha + \rho u_{t-1} + \varepsilon_t$ to get $Y_t - \rho Y_{t-1} = \alpha - \rho\alpha + \varepsilon_t$. Then $Y_t = (\alpha - \rho\alpha) + \rho Y_{t-1} + \varepsilon_t$. This has the same form as Eqn 1, in that the current value depends linearly upon the past, and the errors ε_t are a zero-mean normal RS. So the autoregressive model Eqn 3/Eqn 4 is also autoregressive in the sense that autoregression has been defined (Eqn 1 model).

Second, suppose that time series model Eqn 1 holds. Then $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$, in which the ε_t are i.i.d. By "growing" this model forward in time from initiation, we have
$Y_1 = \alpha + \beta Y_0 + \varepsilon_1$, then substituting this into
$Y_2 = \alpha + \beta Y_1 + \varepsilon_2 = \alpha + \beta(\alpha + \beta Y_0 + \varepsilon_1) + \varepsilon_2 = \alpha + \beta\alpha + \beta^2 Y_0 + \beta\varepsilon_1 + \varepsilon_2$, and so on:
$Y_3 = \alpha + \beta Y_2 + \varepsilon_3 = \alpha + \beta(\alpha + \beta\alpha + \beta^2 Y_0 + \beta\varepsilon_1 + \varepsilon_2) + \varepsilon_3 =$
$\alpha + \beta\alpha + \beta^2\alpha + \beta^3 Y_0 + \beta^2\varepsilon_1 + \beta\varepsilon_2 + \varepsilon_3$,
…,
$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t = \alpha(1 + \beta + \beta^2 + \cdots + \beta^{t-1}) + \beta^t Y_0 + \beta^{t-1}\varepsilon_1 + \beta^{t-2}\varepsilon_2 + \cdots + \beta\varepsilon_{t-1} + \varepsilon_t$.
Now, if we assume that *t* is large (a long time has passed since the time series started) and that $0 < \beta < 1$, then the magnitude of $\beta^t Y_0$ will be small and negligible and $\alpha(1 + \beta + \beta^2 + \cdots + \beta^{t-1})$ will be approximately $\frac{\alpha}{1-\beta}$ by the formula for the sum of a geometric series. So
$Y_t \cong \frac{\alpha}{1-\beta} + \beta^{t-1}\varepsilon_1 + \beta^{t-2}\varepsilon_2 + \cdots + \beta\varepsilon_{t-1} + \varepsilon_t$. Now, define an "error" term
$u_t = \beta^{t-1}\varepsilon_1 + \beta^{t-2}\varepsilon_2 + \cdots + \beta\varepsilon_{t-1} + \varepsilon_t$. Then lag the definition of u_t (replace *t* by *t* - 1):
$u_{t-1} = \beta^{t-2}\varepsilon_1 + \beta^{t-3}\varepsilon_2 + \cdots + \beta\varepsilon_{t-2} + \varepsilon_{t-1}$ and substitute into u_t to get
$u_t = \beta(\beta^{t-2}\varepsilon_1 + \beta^{t-3}\varepsilon_2 + \cdots + \beta\varepsilon_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \beta u_{t-1} + \varepsilon_t$, which matches Eqn 4, with $\beta = \rho$.
Substituting, we have $Y_t \cong \frac{\alpha}{1-\beta} + \beta u_{t-1} + \varepsilon_t = \frac{\alpha}{1-\beta} + u_t$. Since $u_t = \beta u_{t-1} + \varepsilon_t$, then the latter form of the model ($Y_t \cong \frac{\alpha}{1-\beta} + u_t$) matches the form of Eqn 3. This model is autoregressive in the errors. The point is that a model that is autoregressive in *Y* can be expressed as a model that is autoregressive in its errors.

[6] This form of the autoregressive model provides the insight, noted in the preceding paragraph, that the known (past) values of the errors have useful information about the next *Y* and can, in fact, be used as predictors of the next *Y*. It also provides justification for requiring that the errors in regression be independent. If they are not, then you are leaving explanatory power unused on the table, for the (lags of the) residuals could be incorporated as additional predictors and boost the explanatory power.

Thus, models Eqn 1 and Eqn 3/Eqn 4 really describe the same autoregression. It is broadly correct to say that autoregression in Y_t with RS errors is the same as a RS model in Y_t with autoregressive errors. This remains broadly correct when additional time series (like X_t) are added to the models as predictors.

The two views of autoregression are equivalent. In analyzing autoregressive models, we may use whichever view is easier or more convenient or insightful.

Sidebar. {*This sidebar shows that the value of β in autoregression is approximately the autocorrelation of Y_t.*}

A small point in the immediately preceding argument was the hypothesis that $0 < \beta < 1$ so that $\beta^t Y_0$ will be small. It turns out that if model Eqn 1 holds, then β is approximately the autocorrelation of the time series Y_t. To see this, recall that in any simple regression, the slope = $\rho_{y,x} \frac{\sigma_y}{\sigma_x}$. In Eqn 1, the slope is β. If Eqn 1 is estimated by regression, then $\rho = Corr(Y_t, Y_{t-1})$, which is the autocorrelation, and $\sigma_y = stdev(Y_t)$ and $\sigma_x = stdev(Y_{t-1})$. Now these two standard deviations are almost the same because the set of numbers used to calculate $\sigma_y = stdev(Y_t)$ differs from the set of numbers used to calculate $\sigma_x = stdev(Y_{t-1})$ by only one value. So $\frac{\sigma_y}{\sigma_x} = \frac{stdev(Y_t)}{stdev(Y_{t-1})} \approx 1$. Thus the slope $\approx \rho 1 = \rho$, which is the autocorrelation. So unless the autocorrelation is perfect, then $-1 < \rho < 1$, so that $\beta^t Y_0 \approx \rho^t Y_0$ is small.
End of sidebar.

Detection of Autoregression in the Errors

Suppose that you want to run a regression to predict Y_t in a time series. How can you determine whether the errors exhibit autoregression? This is an important question because if the errors do exhibit autoregression, then the errors cannot be independent. In that case, the "I" specification for OLS regression would be violated, and a different model should be proposed. Here are three procedures to detect autocorrelation in the errors:

- To test for autoregression in a time series, first run the regression *as though it is valid*. Get the residuals u_t from that regression. Then regress the residuals upon their lag: That is, run the regression model $u_t = \beta u_{t-1} + \varepsilon_t$.[7] Then use the standard T-test for H_0: $\beta = 0$. If you can reject the null hypothesis, then autoregression is likely present in the residuals.

[7] This is a *no-intercept* regression model. In SAS, use the option NOINT in the model statement of PROC REG or PROC AUTOREG. It probably would not hurt to use the intercept model, but the intercept will likely be around zero, since the mean residual is zero.

- Alternatively, you can run the regression and get the residuals u_t from that regression. Then calculate the autocorrelation of those residuals. Then test H_0: $\rho = 0$ by rejecting the null hypothesis if the autocorrelation of the residuals lies outside the range ($-2/\sqrt{n}$, $+2/\sqrt{n}$) for an approximate 0.05 significance level.[8]
- The most famous test for autoregression in the residuals is the Durbin-Watson test. To perform this procedure, first run the regression as though it is valid and get the residuals u_t from the regression. Then calculate the Durbin-Watson statistic

$$d = \frac{\sum_{t=2}^{n}(u_t - u_{t-1})^2}{\sum_{t=1}^{n}(u_t)^2}$$

Reject H_0: $\rho = 0$ if d is either too big or too small.

The intuition for the Durbin-Watson test: The numerator of d is the total of squared differences between neighboring residuals. The denominator is a standardizing quantity. Small values of d mean that the numerator is small, which means that most residuals are close to their neighbors (which indicates positive autocorrelation: high values have neighboring values that are also high; low values have neighboring values that are also low). Large values of d mean that most residuals are far from their neighbors (which indicates negative autocorrelation: high values have neighboring values that are low; low values have neighboring values that are high). Moderate values of d mean that most residuals exhibit no dependable relationship with their neighbors (zero autocorrelation).

The range of possible values for d is 0 to 4. [See the following sidebar if you are interested in why the range for d is 0 to 4.] d close to 0 signals positive autocorrelation; d close to 4 signals negative autocorrelation; d close to 2 signals zero autocorrelation.

What are the critical values? d has a complicated distribution that depends on the values of the x predictors, if any. Durbin and Watson were able to provide bounds on the distribution that do not depend on the x values in the original regression. However, these bounds left a range of values for d in which the test is inconclusive, neither accepting nor rejecting the null hypothesis of zero autocorrelation. A further note is that the predictor variables should not include any lags of Y.[9] Tables or computer programs will calculate the DW statistic and its significance. E.g., SAS computes DW and its significance for specified lags in PROC AUTOREG.

Sidebar. Why is (0,4) the range for d? Why does d close to 0 signal positive autocorrelation; d close to 4 signal negative autocorrelation; and d close to 2 signal zero autocorrelation? To answer these questions, let us expand and approximate d:

---

[8] For small and medium-sized samples, a more accurate test can be based on Fisher's hyperbolic arc tangent transformation: If H_0: $\rho = 0$ is true, then artanh(r) = $\frac{1}{2}\log_e \frac{1+r}{1-r}$ has approximately a normal distribution with mean 0 and variance 1/(n-3), where r is the sample autocorrelation (or any other correlation). So reject H_0 if |artanh(r)| > 1.96/sqrt(n-3) for approximate 0.05 significance test.

[9] If lags of Y are included among the predictors, then a modification called the Durbin-h test is available.

$$d = \frac{\sum_{t=2}^{n}(u_t - u_{t-1})^2}{\sum_{t=1}^{n}(u_t)^2} = \frac{\sum_{t=2}^{n}u_t^{\,2} - 2\sum_{t=2}^{n}u_t u_{t-1} + \sum_{t=2}^{n}u_{t-1}^{\,2}}{\sum_{t=1}^{n}(u_t)^2} =$$

$$\frac{\sum_{t=2}^{n}u_t^{\,2}}{\sum_{t=1}^{n}(u_t)^2} + \frac{\sum_{t=2}^{n}u_{t-1}^{\,2}}{\sum_{t=1}^{n}(u_t)^2} - 2\frac{\sum_{t=2}^{n}u_t u_{t-1}}{\sum_{t=1}^{n}(u_t)^2} \approx 1 + 1 - 2\frac{\sum_{t=2}^{n}u_t u_{t-1}}{\sum_{t=1}^{n}(u_t)^2}$$ (since the numerator and denominator of the first two terms differ only by one or two of the numbers being summed)

$$\approx 2 - 2\frac{\sum_{t=2}^{n}(u_t - \bar{u}_t)(u_{t-1} - \bar{u}_{t-1})}{\sqrt{\sum_{t=1}^{n}(u_t)^2}\sqrt{\sum_{t=1}^{n}(u_t)^2}}$$ (since the mean residual $\bar{u}_t$ or $\bar{u}_{t-1}$ is zero or approximately zero)

$$\approx 2 - 2\frac{\sum_{t=2}^{n}(u_t - \bar{u}_t)(u_{t-1} - \bar{u}_{t-1})}{\sqrt{\sum_{t=2}^{n}(u_t - \bar{u}_t)^2}\sqrt{\sum_{t=2}^{n}(u_{t-1} - \bar{u}_{t-1})^2}}$$ (again since the mean residual is approximately zero and the terms substituted in the square root differ in only one number)

$\approx 2 - 2 \cdot autocorr(u_t, u_{t-1})$ (by definition of correlation). Since correlation is bounded between -1 and +1, the largest value that *d* could have is approximately 2 – 2(-1) = 4 in the case of extreme negative autocorrelation; the smallest value that *d* could have is approximately 2 – 2(+1) = 0 in the case of extreme positive autocorrelation; in the case of no autocorrelation, we expect *d* to be approximately 2 – 2(0) = 2.

End of sidebar.

The Specifications of the Autoregression Model

Figure 1 is a schematic that illustrates the simple autoregression model for three of the possible previous values of *Y*. If the previous value of *Y* is 100, for example, then the distribution of the next value of *Y* is the normal distribution shown at Y_{t-1} = 100. It is important to note that the horizontal axis is *not* time. That is, the leftmost distribution is not the distribution of *Y* at time 1, nor is the middle distribution the distribution of *Y* at time 2, nor yet is the rightmost distribution the distribution of *Y* at time 3. What Figure 1 shows is the possible distributions of *Y* when the *previous* value of *Y* is the horizontal axis.

If the distributions of *Y* in Figure 1 were arranged in time order, they might look like Figure 2 (just one of the possible ways they could be arranged). In Figure 2, the distribution of *Y* at time *t* = 1 is the leftmost distribution, the distribution of *Y* at time *t* = 2 is the middle distribution, and the distribution of *Y* at time *t* = 3 is the rightmost distribution. The distributions in the time order shown here are *not* linearly related.

- At time 1, the hypothetical previous value of *Y* was 300; so the distribution of *Y* at time 1 looks like the *right*most distribution of *Y* in Figure 1. In Figure 1 the value of the previous *Y* for the rightmost distribution is 300.
- Similarly, at time 2, the hypothetical previous value of *Y* was 100; so the distribution of *Y* at time 2 looks like the *left*most distribution of *Y* in Figure 1. In Figure 1 the value of the previous *Y* for the leftmost distribution was 100.

- Likewise, at time 3, the hypothetical previous value of *Y* was 200; so the distribution of *Y* at time 3 looks like the middle distribution of *Y* in Figure 1. In Figure 1 the value of the previous *Y* for the middle distribution was 200.

Figure 1. The simple autoregression model

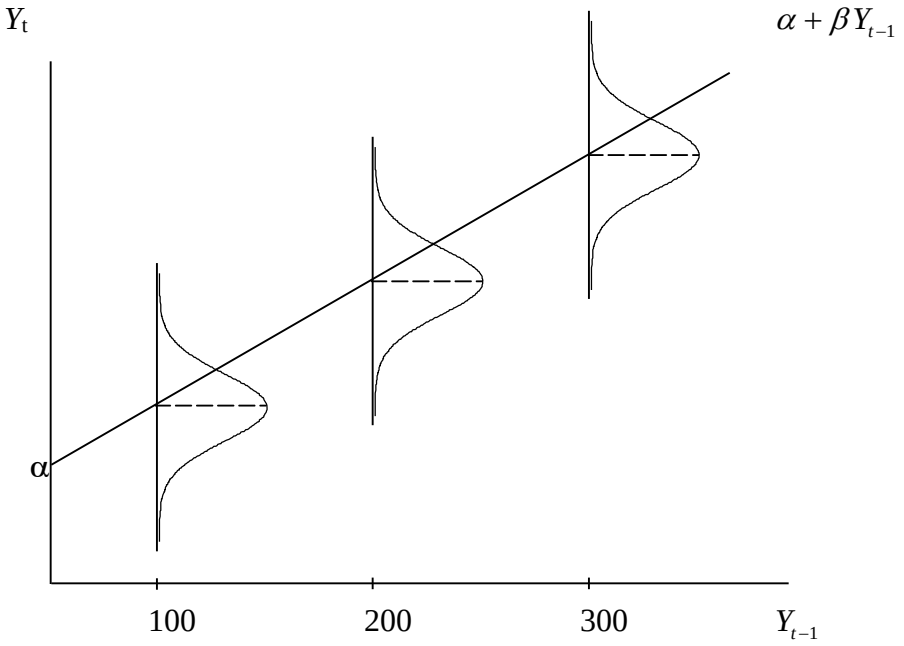
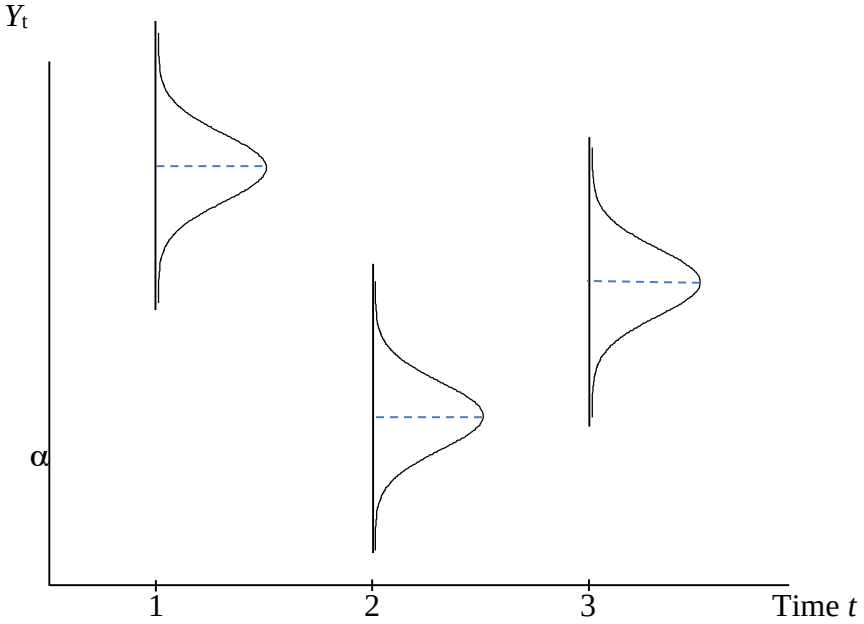


Figure 2. The simple autoregression model in time order



It is worth investing further effort to understand better the specifications of the autoregression model. As with the RS and RW, it is helpful to expand the i.i.d. definition into the three L,H,I specifications and add a fourth – N, for normality. It turns out that there are two equivalent ways

to express the specifications – one in terms of Y_t and the other in terms of the errors. Both sets of specifications yield further understanding of autoregression. So at the risk of overtaxing your patience, I will discuss both sets. The following are the four specifications in terms of Y_t for the simple autoregression model:

- **L** – **L**inearity (or **L**evel). This says that the means of the distributions of Y_t lie on a straight line $\alpha + \beta Y_{t-1}$. That is, the expected location of the next *Y* is a linear function of the last *Y*. This is the *linear* aspect of **L**.[10] In Figure 1, the L specification is illustrated by the straight line crossing the vertical axis at α and slanting upward to the right. That straight line connects the means of all of the possible distributions of Y_t whatever the previous value (Y_{t-1} - the horizontal axis) may have been. So to find out what the mean of Y_t is, locate the previous value of *Y* on the horizontal axis and read up to the value on the line – or just plug the previous value of *Y* into $\alpha + \beta Y_{t-1}$. Please note carefully that the **L** specification does **not** say that the means of *Y* are linear in time, but rather are linear in the preceding value of *Y*.
- **H** – **H**omoscedasticity. This says that the standard deviations of the distributions of Y_t are the same for every possible value of the last *Y*. In Figure 1, the H specification is represented by all of the normal distributions having the same spread. Thus, each normal distribution in Figure 1 is a carbon copy of every other normal distribution. The distributions merely change their locations by sliding along the slanted mean line $\alpha + \beta Y_{t-1}$.
- **I** – **I**ndependence. This says that the draws from the distributions of Y_t are made independently at every time *t*. This specification is a bit subtle. It may not mean quite what you think it means. It does *not* mean that the value of Y_t is independent of the value of Y_{t-1}. It means that once the preceding value Y_{t-1} is known, then that determines the mean of the next distribution, namely $\alpha + \beta Y_{t-1}$, and from that point onward, Y_{t-1} has *no further* effect on the value of Y_t that is to be drawn. Another way to say this is that the value of Y_{t-1} has no effect on whether the value of Y_t that is to be drawn will be above or below the mean $\alpha + \beta Y_{t-1}$. That is, the value of Y_{t-1} has no effect on the *deviation* of Y_t from its mean $\alpha + \beta Y_{t-1}$: the deviation Y_t - ($\alpha + \beta Y_{t-1}$) is independent of Y_{t-1} – that is, the residual is independent of Y_{t-1}.
- **N** – Normality. This is a new specification. It joins our old friends L,H,I as a fourth specification. It says that the probabilities for the distributions of Y_t are normal. **N** is illustrated in Figures 1 and 2 by the normal curves shown for each draw of Y_t. **N** is specified mainly for convenience in deriving the mathematical properties of autoregression. Advanced mathematical analysis[11] shows that if the sample size is sufficiently large – and "large" can be astonishingly small (roughly $n > 30$) – many of the regression properties hold approximately whether or not the distributions of Y_t are

[10] There is also a **L**evel aspect. See the restatement of LHIN in the next set of bullets.
[11] The Central Limit Theorem applied in regression.

normal. The most important exception is for assessing the confidence probability of a margin of error for a forecast. Yet even there, if N fails, we can still assess the confidence empirically by counting cases in the error range and dividing by *n*.

The above discussion of LHIN is about the distribution of Y_t: the means of the distributions of Y_t are linear, the standard deviations of the distributions of Y_t are homoscedastic, the draws from the distributions of Y_t are independent, the distributions of Y_t are normal. Yet the definition of autoregression on page 1 of these notes identifies the specifications of autoregression in terms of the errors. The defining specifications are not in terms of Y_t; the defining specifications state that the errors are a zero-mean normal RS.

So what is the connection between LHIN in terms of Y_t in the preceding bulleted discussion and the definition in terms of errors? Answer: They are equivalent! If Y_t satisfies LHIN as discussed above, then the errors are a zero-mean normal RS. Conversely, if the errors are a normal RS, then Y_t satisfies LHIN as above. So each of the LHIN conditions in the bullets above can be restated equivalently in terms of errors. It is worthwhile to lay out the connection explicitly.

But first, recall the definition of the errors:

$$\varepsilon_t = Y_t - (\alpha + \beta Y_{t-1}),\ t = 1,2,...,n \quad \text{[Eqn 2]}$$

Note that the errors as defined in Eqn 2 result simply from solving Eqn 1 in terms of ε_t.

So here are LHIN restated in terms of the errors:

- **L** – **L**evel (or **L**inearity). This says that the means of the distributions of the errors ε_t are all zero. This is clear from Eqn 2. For if L holds, then the mean of Y_t is $\alpha + \beta Y_{t-1}$. So the mean of $\varepsilon_t = Y_t - (\alpha + \beta Y_{t-1})$ must be 0.
- **H** – **H**omoscedasticity. This says that the standard deviations of the distributions of the errors ε_t are all the same. This is clear from Eqn 2. For if H holds, then the standard deviations of Y_t are all the same. Subtracting a constant from Y_t does not change the standard deviation.[12] So the standard deviation of $\varepsilon_t = Y_t - (\alpha + \beta Y_{t-1})$ is the same as the standard deviation of Y_t. Thus the errors ε_t are homoscedastic.
- **I** – **I**ndependence. If I holds, then the draws from the distributions of Y_t are made independently at every time *t* in the sense discussed in the above bullet on I. Subtracting a constant $(\alpha + \beta Y_{t-1})$[12] from each Y_t does not affect the independence of draws. Furthermore, it was noted in the above bullet on I that the residual errors $\varepsilon_t = Y_t - (\alpha + \beta Y_{t-1})$ are independent of previous values Y_{t-1} - and hence of previous errors $\varepsilon_{t-1} = Y_{t-1} - (\alpha + \beta Y_{t-2})$, etc.
- **N** – Normality. Subtracting a constant from each value in a normal distribution does not affect the normality of the distribution. It just changes the mean. So if N holds for Y_t, N must hold for $\varepsilon_t = Y_t - (\alpha + \beta Y_{t-1})$.

[12] $(\alpha + \beta Y_{t-1})$ is effectively a constant, since Y_{t-1} will be known at time *t* when the next draw is made to get Y_t.

Thus, the first statement of LHIN (for Y_t) implies that the errors in an autoregression are a zero-mean normal random sample. It is also easy to show that LHIN of the errors implies LHIN of the *Y*'s.

The reason that equivalency of LHIN of the *Y*'s and LHIN of the errors is important is that it is more convenient to validate autoregression by testing LHIN of the errors than by testing LHIN of Y_t. The reason is that every time we run autoregression, the software provides us with estimates of the errors in the form of the observed residuals. We can therefore use the output residuals as proxies for the errors in order to validate autoregression by testing whether the residuals satisfy LHIN.

Sidebar. {*This is a note on some technical issues involved in regressing a Y variable upon a function of itself (its lag). You should be aware that there are issues, but you should not obsess over them.*}

In the autoregression model $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$, the lagged response is used as a predictor. Predictors that are lagged responses are not completely benign. Predictors that are lagged responses have special problems. These problems arise from regression's treatment of predictor variables as constants.

The specifications of autoregression and regression in general are just LHIN. But regression carries an additional, often tacit, assumption that the predictor variables are constants – or at least that the predictors can be treated *as though* they are constants. This is true if the values of the predictor variables are in fact constants, and may be plausible under the regression model, which views the regression equation as the *conditional* mean of *Y* given the *X*'s. This conditional view makes the predictor variables effectively constant, even if they were in fact collected as random variables. This approach can usually be justified if the response variable and the predictor variables are, in fact, different variables – like *Y* and *X*.

However, with lagged responses as predictors, the same values appear on both sides of the regression equation – treated as random variables on the left-hand side, but as constants on the right. This schizophrenic treatment is logically inconsistent. The fiction that the predictors can be treated as constant or as conditionally constant breaks down. But regression is so useful that people do it anyway! Could that hurt you? What if you use lagged responses as predictors, you verify LHIN, and you estimate the regression equation with standard OLS (ordinary least squares) regression methodology? Then estimates of the coefficients α and β will be biased and their standard errors will also be biased. However, these biases will shrink toward zero as the sample size grows. If the sample size is sufficiently large, coefficient tests and confidence intervals will be approximately correct if done in the usual manner with the reported output. The take-away: OLS regression still works, but large sample sizes may be necessary to compensate for coefficient bias and bias in standard errors. Modifications to OLS methodology can correct for these problems, but those correction methods are beyond the scope of this course.

A word of caution: The immediately preceding remarks apply to the autoregression model $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$, in which the errors satisfy LHIN. If the errors have a more exotic structure – in particular, if the errors are not independent, then we need to worry! If the autoregression model has correlated errors, then OLS estimates of α and β are biased and remain so in large data sets. The bias never goes away, no matter how large your dataset, and the

amount of bias can be substantial. More advanced time series methods exist for dealing with this troublesome situation. We will not go into those in this course.
End of Sidebar.

Summary

Autoregression is a regression model in which one or more lags of the response variable are used as predictor variables. There may or may not be other predictor variables. The simplest form of an autoregression is

$$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t,\ t = 1,2,...,n$$

in which the error terms ε_t satisfy the standard L,H,I,N specifications of a regression model – i.e., the error terms are a mean zero normal RS.

Use the autoregressive model to **estimate the mean of the distribution of** Y_t by plugging the previous value Y_{t-1} into the estimated regression equation:

$$\text{Estimate of mean of } Y_t = \hat{Y}_t = \hat{\alpha} + \hat{\beta} Y_{t-1} \pm \text{st err of regression estimate}$$

The expected margin of error is best left to statistical software to compute (e.g., SAS)

Use the autoregressive model to **forecast the next value** Y_{t+1} by plugging the current value Y_t into the estimated regression equation:

$$\text{Forecast of value of } Y_{t+1} = \hat{Y}_{t+1} = \hat{\alpha} + \hat{\beta} Y_t \pm \text{RMSE}$$

The expected margin of error is approximately the RMSE of the regression and can be more precisely approximated by statistical software (e.g., SAS).

Both the RS and RW are special cases of autoregression. The RS is autoregression in which $\beta = 0$ (no dependence upon the past). The RW is autoregression in which $\beta = 1$ (complete dependence upon the past).

There are two views of autoregression models:

- Y depends upon lag(Y) and the errors are i.i.d., or
- Y does not depend upon lag(Y) but the error depends upon lag(error).

These two views are mathematically equivalent. So you can use whichever form is convenient.

You can test for autoregression in the errors of a regression in at least three ways:

- Run the regression, calculate the residuals, regress the residuals on their lag and test the coefficient of the "*X*" variable [i.e., lag(residuals)] = 0.
- Run the regression, calculate the residuals, compute the autocorrelation of the residuals, and test the autocorrelation of residuals = 0.
- Run the regression, calculate the Durbin-Watson statistic from the residuals, and reject zero autocorrelation of residuals if DW is too far from 2.