Chapter 3: Testing for a Unit Root

January 29, 2022

Contents

In chapter 2, we introduced the two broad classes of stochastic processes that may characterize our time series: stationary and non-stationary. Stationary processes are characterized by mean reversion and constant variances and covariances. Successive values of stationary processes are predictable based on their past behavior. Importantly, if our time series are stationary, inference is based on standard limiting distributions.

Non-stationary processes have means, variances, and/or covariances that depend on time. Non-stationary processes typically contain trends, either deterministic or stochastic (or both). Deterministic trends evolve with the addition (or subtraction) of a fixed amount in each successive period. If a process contains only a deterministic trend, it is stationary after filtering out the trend. In contrast, stochastically trending processes follow a random walk through time where each step is determined by a random shock (possibly plus drift and/or trend) and the effects of shocks persist indefinitely, accumulating over time. Time series containing stochastic trends are called unit root processes and can be made stationary by first-differencing. Unit root processes necessitate special care for a number of reasons. As we outlined in Chapter 2, unit root processes are prone to the spurious regression problem, appearing related when they are not. In addition, hypothesis tests involving unit root processes tend to have nonstandard limiting distributions.

A single time series can contain any or all of these features, as well as seasonality and structural breaks. Many diagnostic tests are available to help arbitrate among these different possibilites, and, in particular, to determine whether a process contains a unit root. But arbitrating between stationary or trend stationary process and unit root processes is made difficult by (a) the low power of available tests and (b) uncertainty over the proper form of each test, e.g. which deterministic features to include in the test and how to specify lag length or lag truncation parameters for the test. Features of the sample at hand, including most significantly the size and window of the sample, can also make the task harder. But whether a time series contains a unit root is a question of particular importance for the chapters that follow. In this chapter, we introduce the tools used to determine whether a time series contains a unit root and a strategy to help you make informed and defensible decisions about the presence of a unit root. With this decision in hand, you can select a strategy for modeling the relationship between time series.

The first step is to determine the types of processes that may reasonably have generated the data in order to determine the question to pose of the data in the attendant hypothesis tests. Mechanically, this involves deciding whether the data plausibly contain a trend and whether, absent a trend, they plausibly have an expected value of zero. These decisions are informed by theory and time series plots and determine the deterministic features to include in any testing procedure. In step two, the analyst must apply a test or, as we recommend, a set of tests. We describe the most commonly used tests and highlight the strengths and weaknesses of each.We present guidelines for making these decisions. Application of the tests requires adopting systematic strategies recommended for selecting lag lengths (p) or lag truncation (l) values required for the tests. In step three, the test results are interpreted holistically. We stress that analysts should report all decisions in the testing procedure, including competing findings. This insures any uncertainty is transparent to readers.

1 The Unit Root Question

The question of whether a time series contains a unit root is easiest to understand in the context of the simple AR(1) process described in Chapter 2.

$$y_t = \phi y_{t-1} + \varepsilon_t \tag{1}$$

where ε_t is a white noise process and there is no constant ($c = 0$) because the mean of the series is zero. The parameter ϕ in Equation 1 is the autoregressive (AR) parameter that governs the relationship between contemporaneous (y_t) and lagged values (y_{t-1}) of the variable. The closer ϕ is to one, the stronger the relationship and the more persistent the time series. If $|\phi| < 1$, the process is stationary. If $|\phi| = 1$, the process contains a unit root.

Of course, the process describing the evolution of the time series may be more complex. It may contain a constant, a constant and a trend, and/or seasonality; an event (or events) may impart a temporary or permament effect on the series; and the error process need not be white noise. Regardless of these features of the data, the central question for the analysts remains the same. Is $|\phi| < 1$ or is $|\phi| = 1$? If $|\phi| = 1$, the process contains a unit root. In this section we discuss the role of the constant and trend in determining the behavior of stationary and unit root processes. We will consider seasonality and structural breaks below.

Table 1 presents three common data generating processes (DGPs) – those without a constant or trend, those with a constant, and those with a constant and linear trend – and describes the nature of the process when the time series is stationary ($|\phi| < 1$) and when it contains a unit root ($|\phi| = 1$). The first row of the table describes a process generated by Equation 1; it is a mean-zero stationary process if $|\phi| < 1$ and a unit root process if $|\phi| = 1$. The addition of a constant (row 2) implies the mean of the stationary process is non-zero ($\bar{y}_t = c/(1-\phi)$). The addition of a constant to a unit root process creates drift. The errors accumulate over time and the value c is added to y_t in each period. This imparts a deterministic linear trend in the time series. Adding a linear trend (row 3) to the DGP results in either a trend stationary process or a unit root process with trend and drift. In the latter case, y_t will contain a quadratic trend because $c + \delta\frac{t(t+1)}{2}$ is added to y_t in each period.

Table 1: Stationary and Unit Root Processes

Data Generating Process	$\|\phi\| < 1$ (Stationary)	$\|\phi\| = 1$ (Unit Root)
$y_t = \phi y_{t-1} + \varepsilon_t$	Mean-zero stationary autoregressive process	Unit root process
$y_t = c + \phi y_{t-1} + \varepsilon_t$	Stationary autoregressive process with mean $c/(1-\phi_1)$	Unit root process with drift c
$y_t = c + \delta t + \phi y_{t-1} + \varepsilon_t$	Trend stationary autoregressive process	Unit root process with drift c and (quadratic) trend t

Each row of the table specifies an AR(1) model but the deterministic components in the DGP differ. We will refer to the deterministic terms included in the model collectively, as D_t. This allows us to write the general form of Equation 1 as

$$y_t = D_t + \phi y_{t-1} + e_t. \tag{2}$$

where D_t includes any deterministic features of the DGP. D_t is either empty (row 1), contains a c (row 2), or both c and δt (row 3).

Note that we have also replaced ε_t in Equation 1 with e_t in Equation 2. This is to allow for the possibility that the error term contains additional dynamics. Perhaps e_t contains additional autoregressive (AR) or moving average (MA) terms. The process could also contain seasonal components or the errors could be heteroscedastic. All these dynamic features of the data are contained in e_t. Equation 2 can represent a variety of different processes, encompassing a host of different features. This expression will serve as the foundation for the unit root tests we present below.

There are two insights from Table 1 and Equation 2 that will be important for understanding unit root tests. First, the unit root question concerns the value of ϕ in Equation 2, regardless of the form of the deterministics. If $|\phi| = 1$, the process contains a unit root. If $|\phi| < 1$, the series is stationary.

Second, the implications of deterministic terms in a model depend on whether or not it contains a unit root. Inclusion of a constant implies a stationary process with a mean of zero or a unit root process with drift. Inclusion of a constant and trend implies a trend stationary process or a unit root with drift and trend. This will become important for the specification of hypothesis tests.

2 Step One: Identify the Null Hypothesis

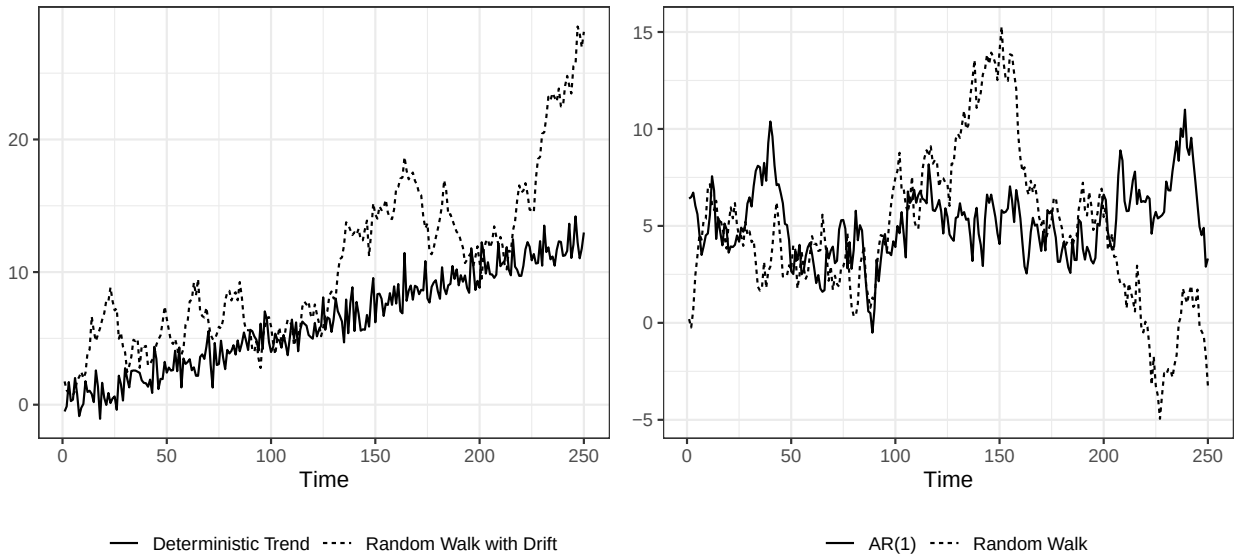
The unit root question boils down to whether $|\phi| = 1$. But to test hypotheses about ϕ, you have to make decisions about D_t. You must decide which types of processes could have plausibly generated your data. That is, which cells of Table 1 could represent your DGP? This determines which null hypotheses you will pose and which test regressions you will specify to test them. There are two things to consider. Is the time series plausibly trending and does the time series plausibly have a mean of zero? Time series plots and theory can provide answers to these questions.

2.1 Is the series plausibly trending?

The key feature of a tending process is that it systematically increases or decreases over time. You should have good intuition about whether trending behavior is plausible. For example, we expect federal spending to increase over time because of ever-growing demands for government spending. With public opinion data, on the other hand, it is difficult to imagine that attitudes toward the president, the economy, or gun control, will become systematically more (or less) positive as time marches on. Opinion series are measured as proportions or percentages. This means they are bounded on both sides. Presidential approval cannot go above 100% or below 0 %. Even if a series appears to trend in sample, an underlying trend is unlikely.

Time series plots will typically provide visual evidence of trending behavior when it exists. Figure 1 shows two pairs of simulated time series. On the left-hand side, the solid line represents a trend stationary series ($\phi = .90$, $c = 0$, $\delta = 0.05$), while the dashed line represents a unit root series with drift ($\phi = 1$, $c = 0.05$, $\delta = 0.0$). Both series systematically increase over time. Absent knowledge of the DGP, it can be difficult to tell which is which, but visual inspection eliminates some possibilities. It is obvious that neither of the series are stationary around a zero or non-zero mean because they are both increasing over time. This also eliminates the possibility that either process is a unit root without drift.

Figure 1: Simulated Time Series with and without Deterministic Trends



On the right-hand side of Figure 1, the solid line depicts a stationary process ($\phi = 0.90$, $c = 0$, $\delta = 0$) and the dashed line shows a unit root without drift ($\phi = 1$, $c = 0$, $\delta = 0$). Again, visual inspection allows us to rule out some possibilities. Neither time series shows a tendency to systematically increase or decrease over time, so there is no chance that either series is trend stationary. Both series exhibit a considerable amount of persistance – $\phi = 0.90$ for the stationary process and $\phi = 1$ for the unit root process – but neither series appears to be trending, so we can rule out the possibility that either can be characterized as a unit root with drift.

The distinct behavior of the series across the two panels and the similarities within each panel suggest how you should frame the questions you pose with your unit root tests. If a time series exhibits trending behavior, like the panel on the left, you want to know whether that time series can be classified as a trend stationary process (row 3, column 1 of Table 1) or a unit root with drift (row 2, column 2). But if a series does not trend, the natural question to ask is whether the time series is a mean stationary process (rows 1 and 2, column 1) or a unit root process without drift (row 1, column 2).

2.2 Is it Plausible the Time Series has a Mean of Zero?

In addition to determining whether a trend is plausible, we must also decide whether the series is plausibly a mean zero stationary process. Often we can easily answer this question based on the measurement of the time series. Time series that can only take positive values cannot be mean-zero stationary processes. Only in a limited number of cases – when the analyst knows the mean of the time series is zero and the first observation is zero – can the answer to the question plausibly be yes. In fact, most unit root hypothesis tests presume a non-zero mean. However, if you are certain the data, if stationary, have a mean of zero, you can pose this question of the data: Is the process a mean-zero stationary process (row 1, column 1) or is it a unit root without drift (row 1, column 2)?

If your theoretical understanding of the process and visual inspection of the data leave you agnostic about the presence of a trend and whether the mean of a plausibly stationary process could be zero, auxiliary tests can be used to help you decide. However, the power of these tests depends on the presence

of unit roots. This presents a potential circularity problem in the testing procedure. Enders (2015, 237) emphasizes this point: "... tests for unit roots are conditional on the presence of deterministic regressors and tests for the presence of deterministic regressors are conditional on the presence of a unit root."

Once you have determined the plausible forms of the DGP, how do you specify the deterministic terms to include in the model for purposes of hypothesis testing? Unit root tests are based on the DGP *assuming the data generating process is stationary.* Our estimates of ϕ will only be consistent if D_t accurately reflects the true DGP.

Table 2 presents the appropriate specification of D_t for each question you might ask of the data. If a time series plausibly contains a trend, you specify $D_t = c + \delta t$. If the time series is not trending, you must make a decision about the mean. If the series is plausibly stationary around a mean of zero, D_t is empty, but if the series is not plausibly mean-zero stationary, or you are not sure, then $D_t = c$.

Table 2: Specifying D_t

Question	Form of Deterministics in the Test Regression
• Is the time series a trend stationary process OR is it a unit root with drift (possibly with trend)?	$D_t = c + \delta t$
• Is the time series a nonzero mean stationary process OR is it a unit root without drift?	$D_t = c$
• Is the time series a mean-zero stationary process OR is it a unit root without drift?	D_t is empty

3 Step Two: Conduct Hypothesis Tests

Once you have ascertained the types of processes that may reasonably characterize a given time series, you can conduct tests to determine whether the series contains a unit root. Entire books have been devoted to unit root and stationarity tests. The problem is that "...it is difficult for *any* [emphasis added] statistical procedure to distinguish between unit root processes and series that are highly persistent" (Enders 2015, 235).[1] Likewise, it is difficult to distinguish between a unit root with drift and a trend stationary process. These difficulties in classification arises because unit root and stationarity tests have notoriously low power for sample sizes that are common in applied work.

The *power* of a statistical test is the probability that the test will reject a null hypotheses, or yield a statistically significant result, when the null hypothesis is false (Cohen 2013, 1). Failing to reject a false null hypothesis is a type II error, or β error. The *size* of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is true. Rejecting a true null hypothesis is a type I error, or α error. In theory, the size and power of a test are established a-priori. When you specify the α-level for a test ($\alpha = .10$, $\alpha = .05$, $\alpha = .01$, etc.) you are specifying the amount of uncertainty you are

[1] Even under ideal conditions, tests will have difficulty arbitrating between a unit root and a mean stationary process with persistence. As the sample size shrinks and the DGP becomes more strongly autoregressive, the power of even the most optimal test drops substantially (Elliott, Rothenberg, and Stock 1996).

willing to accept when you pose your hypothesis and conduct the test. If $\alpha = .05$, you accept that 1 out of every 20 or 5 out of every 100 samples will yield a null result when the null hypothesis is false, but you expect to be rejecting a false null hypothesis 95% of the time. At the same time, you only expect to reject the null hypothesis 5% of the time if the null hypothesis is true. When one says that a test has poor power, they are saying that the test fails to reject false-null hypotheses too often. When one says that a test has poor size, they are saying the test rejects true-null hypotheses too often.

Unit root and statioanrity tests have *low-power against local alternatives*. Here, "low power" means they fail to reject the null hypothesis too often and "against local alternatives" means they have a hard time distinguishing series that are consistent with the null hypothesis from those that are similar to but not consistent with the null hypothesis. What this means for classification depends on the null and alternative hypotheses associated with the different tests. In Section 3.1 we consider tests with a unit root null hypothesis: the (Augmented) Dickey-Fuller test (1979), the Phillips-Perron (1988) family of tests, and the DF-GLS test (Elliott, Rothenberg, and Stock 1996). In Section 3.2 we consider the most commonly applied test of the stationary null hypothesis, the KPSS test (Kwiatkowski et al. 1992).

3.1 Unit Root Tests

Unit root tests begin with a *null hypothesis* that the series contains a *unit root*, possibly with drift and trend. The *alternative hypothesis* is that the series is stationary. Based on your knowledge of the data in step one, you either test (a) the null hypothesis of a unit root against a mean-zero stationary alternative, (b) the null hypothesis of a unit root against a nonzero-mean stationary alternative, or (c) the null hypothesis of a unit root with drift (and possibly trend) against the alternative that the process is trend stationary. The desired comparison determines the specification of D_t.

If the unit root question boils down to the value of the parameter ϕ in Equation 1, why not estimate Equation 1 and test the hypothesis that $\phi = 1$? As sensible as that approach may seem on its face, a hypothesis test of whether $\phi \leq 1$ is not a reliable indicator of whether a series was generated by a unit root process. The estimate of ϕ is biased downwards, especially in small samples. If you estimate an autoregression using unit root data, you will often estimate ϕ to be slightly less than one. Of course, even if the value of ϕ is slightly less than one, you should still be able to construct a confidence interval for ϕ or calculate a test statistic to determine whether ϕ is reliably different from 1, right? Unfortunately, its not that simple. Under the null hypothesis for a test that $\phi = 1$, the variance of the series is increasing over time and the estimates of the standard error for ϕ is unreliable. This renders any test statistic or confidence interval for ϕ invalid. These problems with the standard autoregression are what inspired the architects of our first unit root test.

3.1.1 The Dickey Fuller Test

The most widely used unit root test is the Dickey-Fuller (DF) (Dickey and Fuller 1979) test. The form of the test is based on a transformation of Equation 2 where, for now, we assume the error is white noise. We subtract y_{t-1} from both sides of the equation to produce the test regression:

$$\begin{aligned} y_t - y_{t-1} &= D_t + \phi y_{t-1} - y_{t-1} + \varepsilon_t && (3) \\ \Delta y_t &= D_t + (\phi - 1) y_{t-1} + \varepsilon_t && (4) \\ \Delta y_t &= D_t + \gamma y_{t-1} + \varepsilon_t && (5) \end{aligned}$$

We are interested in whether $\phi = 1$. But since $\gamma = \phi - 1$, testing the null hypothesis $\gamma = 0$ is the same as testing the null hypothesis $\phi = 1$. If $\phi = 1$, then $\gamma = 0$ and the series contains a unit root. If $|\phi| < 1$ then $\gamma < 0$ and y_t is stationary.[2]

Because y_{t-1} is non-stationary under the null hypothesis, we cannot use a standard t-distribution for the t-statistic on γ. We use the Dickey-Fuller distribution (White 1958). The limiting distribution is well-defined but varies with the specification of D_t and the sample size. (See Fuller 1976, Table 8.5.1). The test statistic is commonly referred to as τ if there are no deterministic terms in the test regression; τ_μ if a constant is included in the test regression; and τ_τ if a constant and a trend are included. Most statistical software packages provide critical (or p-) values based on numerical simulations of the distributions for each test.

The Dickey-Fuller test is a one-tailed hypothesis test. You compare the test statistic to the appropriate (τ, τ_μ, or τ_τ) Dickey-Fuller critical value. If the test statistic is more negative than the critical value, you reject the unit root null hypothesis. If the test statistic is closer to zero than the critical value, you fail to reject that the unit root null hypothesis.

The Augmented Dickey-Fuller Test As written above, the Dickey-Fuller test assumes the DGP takes the from of a simple AR(1) process. If the underlying DGP has a more complex ARMA(p, q) error process, the residuals in the DF regression will be serially correlated. Just as serial correlation can bias test-statistics and produce misleading inferences in standard regression models, serial correlation in the auxiliary Dickey-Fuller regressions can produce misleading results as well.

The augmented Dickey-Fuller (ADF) test solves this problem by adding enough lagged differences to the DF test equation to ensure the residuals are white noise (Said and Dickey 1984):

$$\Delta y_t = D_t + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t. \tag{6}$$

If the analyst includes too few lags (p is too small), the test tends to have poor size but relatively good power. With too many lags, however, the power of the test suffers.[3]

How do we decide how many lagged differences to include in the ADF test regression? In general, we add as many lags as we need to soak up any serial correlation in the test regression residuals. Ng and Perron (1995) recommend adopting a general-to-specific approach to selecting p. This strategy begins by selecting a maximum value for p, p_{max}, that is large enough to encompass any seasonal autoregression in the data, typically $p_{max} \geq 4$ for quarterly data and $p_{max} \geq 12$ for monthly data. Including more lags than necessary ensures that the first test regression is general enough to capture all the dynamics that exist in the data.[4] Absent seasonality, a smaller number of lags may serve this purpose. If the time series follows an AR(1) process, $p = 1$ should be sufficient, such that no lagged differences are needed in the test regression. Higher order dynamics will necessitate $p > 1$. Alternatively, Schwert (1989) suggested a data-based criterion: $p_{max} = [12 \bullet (\frac{T}{100})^{0.25}]$ where T is the number of observations.

[2]Dickey and Fuller also proposed a test of the null hypothesis $\phi = 1$ in Equation 2. The test statistic is given by $T(\hat{\phi} - 1)$ for the AR(1) case.

[3]The Dickey Fuller coefficient test of the null hypothesis $\gamma = 0$ can also be conducted in the ADF framework where it is given by $T\hat{\gamma}/(1 - \hat{\beta}_1 - \ldots - \hat{\beta}_p)$. The distribution of the test is identical to the AR(1) case.

[4]Unnecessary lags can be removed from the test regression can be iteratively removed as long as these sequential restrictions to not produce violations of the underlying regression assumption that there is no serial correlation in the regression residuals. We will discuss general-to-specific modeling in more detail in latter chapters.

The lag selection for the ADF test regression should not be taken lightly. If you include too few lags of Δy_t, the test will be biased. If you include too many lags of Δy_t, the power of the test will suffer. You should not just assume that the p_{max} you selected is sufficient given the periodicity of your data, you should test the residuals from the general model for serial correlation before you start removing the lags of Δy_t. If you reject the null hypothesis of no serial correlation, you should add lags to the test regression. We suggest using a threshold of $\alpha = .10$ rather than the conventional threshold of $\alpha = .05$ when interpreting these tests. Given what the implications of a false negative would be in this case, it is better to err on the side of caution. Another strategy uses information criteria (AIC, BIC, etc.) to select the test regression (Enders and Granger 1998). This approach is fine, but information criteria can only tell you which model fits the data best amongst those you have designated for comparison. If none of the candidate test regressions are dynamically complete, none of the regressions will produce a reliable results. You should estimate several candidate models with large orders of lags if you are using this procedure. Of course, the best approach is one that applies both of these strategies and specifies the test regression based on the preponderance of evidence. The choice of p_{max} and the logic for its selection should always be reported.

Modern computing makes it easy to implement the ADF test. Once you have selected a p_{max} for your test regression, you can estimate p ADF regressions. The most general regression includes p_{max} lags of Δy_t to account for potential serial correlation. The remaining regressions iteratively remove lags of Δy_t, with the final model including no lags of Δy_t. If we apply a serial correlation test to the residuals from each test regression, we can identify the subset of regressions that are dynamically complete; they include enough lags of Δy_t to capture all the dynamics in the data. Among these, we choose the most parsimonious specification, the one that includes the fewest lags of Δy_t. This specification search allows us to strike the best balance between size and power.

You interpret the results from the test regression that is parsimonious and dynamically complete. The interpretation is the same as before. You compare your test statistic, $\tau = \gamma/SE(\gamma)$ to the appropriate critical value: τ, τ_μ, or τ_τ. If the DF-test statistic is more negative than the critical value, you can reject the unit root null hypothesis. If the DF-test statistic is less negative than the critical value, closer to zero, then you fail to reject the unit root null hypothesis.

Dickey-Fuller Tests for Deterministic Terms What if you don't know D_t? That is, what if you are not confident in the appropriate null and alternative hypothesis based on theory and visual evidence in step one? It is important that you make an informed decision. As you add deterministic terms to the test regression, the limiting distribution for the Dickey-Fuller tests (or any of the tests we discuss below) shift further from zero, making it harder to reject the null hypothesis. This means the power of the test depends on the choice you make about D_t.[5] You do not want to include extra deterministic terms that are not part of the DGP in the test regression because the critical values will be overly conservative. On the other hand, if you omit necessary deterministics, like the trend or constant, the power of the test goes to zero as the sample size increases. An omitted constant will decrease power in proportion to the magnitude of the constant (Campbell and Perron 1991).

Dickey and Fuller (1981) provide joint hypothesis tests (F-tests) that can be used to select the form of the test regression. Dolado, Jenkinson, and Sosvilla-Rivero (1990) recommend you begin the testing procedure by estimating the least restrictive form of the test regression that may reasonably describe the DGP under the alternative hypothesis. If you want to test whether omitting the trend is consistent with

[5]For example, the asymptotic critical values for the Dickey Fuller tests are $\alpha = 0.05$ are -1.95, -2.86, and -3.41 for τ, τ_μ, and τ_τ, respectively (Fuller 1976, Table 8.5.2).

the data or whether such a restriction is untenable, you estimate the Dickey-Fuller test regression with a constant and a trend:

$$\Delta y_t = c + \delta t + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t. \tag{7}$$

After estimating this regression, test the null hypothesis $\gamma = 0$ using τ_τ. Since including the deterministic regressors reduces the power of the test, if you can reject τ_τ you can conclude the series is stationary, regardless of the form of D_t.

If you fail to reject the null hypothesis τ_τ, you need to consider the form of D_t more carefully and ask whether the trend term belongs in the model. To do so, you test the joint null hypothesis given by $\gamma = \delta = 0$ (typically referred to as ϕ_3) using an F-test. Failure to reject the null hypothesis allows omitting the trend from the model. You then respecify the test regression to include a constant but not a trend. If, however, you reject ϕ_3, then the model is a unit root with a deterministic trend, possibly also with drift. In this case y_t contains a quadratic trend.[6]

If you have determined you can exclude a trend term from the test regression, or if the least restrictive model you entertained at the outset contained only a constant, you can test whether the constant belongs in the model in a similar fashion. You begin by estimating the test regression including a constant in the model:

$$\Delta y_t = c + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta . y_{t-i+1} + \varepsilon_t. \tag{8}$$

You then test the null hypothesis $\gamma = 0$ using τ_μ. If you can reject τ_μ, you conclude the series is stationary around a non-zero mean. If you fail to reject the null, you test the joint null hypothesis given by $\gamma = c = 0$ (typically referred to as ϕ_1). If you reject this hypothesis, the process is a unit root with drift. Failing to reject this null hypothesis suggests the constant may be dropped from the test regression and that the process is a unit root without drift. But recall the earlier warning that omitting the constant from the test regression implies stationarity around a zero mean under the alternative hypothesis; using the ϕ_1 test is only reasonable if that possibility exists. Carefully consider this decision as you may fail to reject the null simply because the time series is too short to get a precise estimate of c, even if it is quite far from zero.[7]

[6] To increase your confidence in this conclusion, estimate $\Delta y_t = c + \delta t + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t$. Test the null hypothesis $\delta = 0$ using a standard t-test. (All regressors are $I(0)$.) If you reject $\delta = 0$, then there is evidence the trend belongs in the test regression. You can test the unit root null that $\gamma = 0$ using a standard t-distribution in this case, as Equation 7 contains the appropriate deterministic terms (Sims, Stock, and Watson 1990). If you fail to reject $\delta = 0$, then estimate the test regression omitting the trend term.

[7] Dickey and Fuller (1981) provide critical values for a number of other hypothesis tests based on the ADF test regressions. The ϕ_2 test is a joint test that γ, the constant, and trend term are jointly zero. Critical values are also provided for t-like test statistics for the constant and trend terms in Equations 7 and 8.

Bounded (or limited) Time Series Many time series are constructed to have either an upper limit, a lower limit, or both. For example, many public opinion series are bound between 0% and 100%. A bounded series can be confounding. In the long-run, a bounded series that appears to be trending must bounce off of the upper or lower bound and revert to a mean. A bounded series must also have finite variance for the same reason. These features mean that bounded series meet the textbook definitions for stationary series. On the other hand, Granger (2010) demonstrates the possibility of a bounded unit root series. Even though it is not possible for a bounded series to have infinite variance, the series may behave like a unit root in sample. Researchers need to be wary of citing the theoretical long-run tendencies of a series when trying to determine if the series is stationary or not. Lebo and Grant (2016) advise against using *boundedness* as an for classifying a time series as stationary. Rather, the data sample one has "in hand" should be examined and evaluated using the approaches described in this Chapter. If a series behaves like a unit root in sample, the series should be treated like a unit root.

3.1.2 The Phillips-Perron Test

Phillips and Perron (1988) propose an alternative to the ADF tests that allows more general error dynamics. The form of the test regression is identical to the non-augmented Dickey-Fuller test:

$$y_t = D_t + \gamma y_{t-1} + e_t \tag{9}$$

where e_t is stationary and may be heteroscedastic. As with the Dickey-Fuller tests, the series contains a unit root if $\gamma = 0$ and D_t is determined by the data generating process assumed under the stationary alternative hypothesis.

The difference between the Phillips-Perron (PP) and ADF tests is how each accounts for the dynamics of the error. The ADF tests account for serial correlation parametrically, by augmenting the test regressions to include p lags of Δy_{t-p}. This approach, in effect, "controls for" serial correlation. The PP tests apply a non-parametric adjustment to the t-statistic on γ to account for the dynamics in the error. This adjustment is based on the heteroskedasticity and autocorrelation consistent covariance matrix developed by Newey and West (1987), the same covariance matrix used to calculate Newey-West standard errors. Rather than "controlling for" serial correlation with lags of Δy_{t-p}, the PP tests "robustify away" serial correlation in the residual process; not by applying Newey-West standard errors to the test regression directly but using the Newey-West covariance matrix to adjust the test statistic on γ. This adjustment is what allows Phillips and Perron to estimate the regression in levels. Under the null hypotheses, the tests have the same asymptotic distributions as the DF t-statistics.[8] Because the PP tests are based on the Newey-West covariance matrix, the PP tests have the added advantage of being robust to heteroskedasticity in the error term. The ADF tests assume homoskedasticity, an assumption that may not be plausible in many applications.

With the PP test, the user does not have to specify a lag length, p, for the test regression. However, a bandwidth or lag truncation parameter, l, must be specified for the purpose of calculating the long-run variance estimate of e_t that is used to adjust the t statistic.[9]Typically one chooses a short or long lag

[8]Like the DF and ADF tests, a coefficient-based version of the PP test is also available. In the urca package in R, the coefficient test is called the Z_α test and the version based on the t-statistic is referred to as the Z_τ test. There is also a modified version of the PP test that uses the GLS detrending procedure Perron and Ng (1996, 1998).

[9]The long-run variance is estimated from the spectrum of the residuals ($\hat{e}$) at frequency zero. Kernel-based estimators or estimators based on the autoregressive spectral density may be used. Phillips and Perron (1988) used kernel-based estimators. Research has shown that either the Newey and West (1994) or Andrews (1991) data-based methods for selecting l perform well under a variety of error dynamics. The choice of kernel has little effect on the performance of the tests (Perron 1988;

truncation parameter that is based on the length of the time series, where a short lag truncation is based on the formula: $trunc(4*(n/100)^{0.25})$ and a long lag truncation is based on the formula: $trunc(12*(n/100)^{0.25})$ as given in Schwert (1989). Simulation evidence shows that this choice can have important implications. For strongly autoregressive time series, a larger lag truncation parameter is needed for the tests to have good power and size; for weak positive autocorrelated errors, a small lag truncation parameter is sufficient; but for time series with large negative moving average errors, the test is unreliable (Perron 1988, 312).[10] Phillips and Perron (1988) also report a series of adjusted t-tests for the inclusion of deterministic regressors but analysts typically rely on theory, plotting, and the Dickey-Fuller ϕ tests to determine the form of D_t.

3.1.3 The DF-GLS Test

Elliott, Rothenberg, and Stock (1996) (ERS) proposed variants of the Dickey-Fuller tests, the Dickey-Fuller Generalized Least Squares (DF-GLS) tests, that have been shown to have greater power than the DF tests in a wide variety of cases. The central difference is in the treatment of D_t. Where the DF tests estimate D_t as part of the test regression, the ERS tests use generalized least squares to estimate D_t and then use the filtered residuals as the input for a test regression that omits D_t:

$$\Delta\tilde{y}_t = \gamma\tilde{y}_{t-1} + \varepsilon_t \tag{10}$$

where $\tilde{y}_t$ is the residual series that has been detrending using GLS.[11] As with the Dickey-Fuller tests, the unit root null hypothesis is rejected if $\gamma < 0$. The form of D_t is given under the alternative hypothesis. The test regression can be augmented to accommodate serially correlated errors by adding lags of the dependent variable ($\Delta\tilde{y}_{t-p}$).

ERS recommend choosing the lag length for the test regression using the Schwartz information criterion, but a general-to-specific strategy as suggested in Section 3.1.1 is appropriate here as well.[12] The test statistic follows the Dickey-Fuller τ distribution if a constant is included in the test regression but when D_t also includes a trend, the asymptotic distribution is not Dickey-Fuller.[13] ERS (1996, 825) simulated the critical values of the test statistic in this case for a range of sample sizes and these are given in most statistical software packages.

Why go to the trouble of using this test rather than the Dickey-Fuller tests? When y_t is stationary, OLS and GLS will give asymptotically equivalent estimates of D_t. But, when y_t is strongly autoregressive or contains a unit root, the constant and trend term are poorly estimated in the Dickey-Fuller test regression; the standard error for the estimate of γ can be inflated. The GLS estimates will be more efficient in this case, which improves the power of the test. However, the DF-GLS test is particularly sensitive to the magnitude of the first observation and has been shown to be less powerful than the DF test when the first observation is large or the sample size is small (Choi 2015).

Newey and West 1994), but the best value of l for either estimator depends on this choice. Most software packages, however, do not estimate l in this fashion, but instead specify l as a function of T, in a manner such as that recommended by Schwert (1989) and given in the text.

[10] See also (Choi 2015, 37) for details.

[11] DF-GLS detrending involves first near-differencing y_t by selecting $\alpha < 1$ and subtracting αy_{t-1} from y_t. Then the near differenced y_t are regressed on a similarly transformed D_t. $\tilde{y}_t$ are the residuals from this regression. ERS found the best overall power of the test is obtained when the quasi-differencing parameter is set to $\alpha = (1 - 7/T)$ for the case of the intercept and $\alpha = (1 - 13.5/T)$ if there is an intercept and a trend. Statistical packages will adopt these default values for the test. For more details on the GLS detrending procedure see Elliott, Rothenberg, and Stock (1996).

[12] Here, again, Ng and Perron (2001) recommend using the MAIC to select l in those cases where the quasi-differenced process contains a larger moving average component.

[13] The distribution depends on the value of α used for quasi-differencing in the detrending procedure.

3.2 Stationarity Tests

The defining feature of all of the tests described so far is the unit root null hypothesis. There is another set of tests that can be used to classify series as stationary or non-stationary: stationarity tests. Like unit root tests, this class of diagnostics is defined by the null hypothesis. The null for a stationarity test is that the series is stationary, stationary around a trend or stationary around a predictable long run mean. There are several tests that have a stationary null hypothesis (See Choi (2015, 116)) but the most widely used tests, by far, are the tests developed by Kwiatkowski et al. (1992).

3.2.1 The KPSS Test

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests differ from the unit root tests described above; a series is assumed to be mean (or trend) stationary under the null hypothesis (Kwiatkowski et al. 1992). The tests are based on residuals from a test regression where y_t is regressed on D_t:

$$y_t = D_t + e_t \tag{11}$$

and e_t is a composite error term that may contain both a unit root ($r_t = r_{t-1} + \varepsilon_t$) and a stationary (and possibly heteroscedastic) component (u_t). If y_t is stationary, it must be the case that the composite error term is stationary. This requires $\sigma_\varepsilon^2 = 0$ so that r is a constant. KPSS proposed a Lagrange Multiplier test of the null that $H_0 : \sigma_\varepsilon^2 = 0$ applied to the residuals ($\hat{e}_j$) from an OLS regression 11:

$$KPSS = (T^{-2}\Sigma_{t=1}^{T}\hat{S}_t^2)/\hat{\lambda} \tag{12}$$

where $\hat{S}_t = \Sigma_{j=1}^{t}\hat{e}_j$ is the partial sum of the residuals squared and $\hat{\lambda}^2$ is a heteroskedasticity and autocorrelation consistent (HAC) estimate of the long-run variance of e_t calculated using $\hat{e}_t$. As the partial sum diverges from the variance of the residuals, the value of the KPSS test statistic grows and we are increasingly likely to reject the stationary null. KPSS calculated quantiles of the limiting distribution of the test statistics, η_μ for $D_t = c$ and η_τ for $D_t = c + \delta t$. Software packages typically report the critical values associated with the test.

The test requires a heteroskedasticity and autocorrelation consistent (HAC) estimate of the long-run variance-covariance (VCOV) matrix for the residuals. These HAC residuals account for excess residual autocorrelation and heteroskedasticity that affect the performance of the DF tests. The standard approach applies the estimator proposed by Newey and West (1987) but Hobijn, Franses, and Ooms (2004) note an alternative described by Park and Phillips (1988). The HAC estimates of the long run variance appears in the denominator of Equation 12. Like the PP tests, HAC estimation requires you to select bandwidth (or lag truncation) values, l, for the VCOV estimates. These are typically selected using the same sample size based formulas provided by Schwert (1989): short ($l = trunc(4 * (n/100)^{0.25})$) truncation or long ($l = trunc(12 * (n/100)^{0.25})$) truncation.

The performance of a KPSS test depends on the selection of the appropriate lag truncation parameter given the features of the data. KPSS show that, for the size of the test to be acceptable, a longer lag truncation is needed with more autocorrelation in y_t. But longer lag truncation will also reduce the power of the test. As Kwiatkowski et al. (1992, 173) state, for T between 50 and 100 "...there is a clear, if unattractive, trade-off between correct size and power: choosing l large enough to avoid size distortions in the presence of realistic amounts of autocorrelation will make the tests have very little power. Only for $T \geq 200$ do we find appreciable power without the risk of very substantial size distortions." The analyst

should thus remember that for smaller samples there will be a tendency to over reject the stationary null hypothesis.[14]

3.3 Which Tests Should We Use?

Which test is best for your time series analysis? Unfortunately, there is no definitive answer to this question but there is definitive guidance on which practices should be avoided. One approach applies all of the tests indiscriminately. A careless analyst would apply all the the different tests with multiple forms of D_t and a variety of lag lengths, picking and choosing the results to report based on their prior beliefs or preferences. This strategy is likely to lead to confusion and incorrect inferences. An equally problematic approach is one where an analyst chooses to conduct a single test and report a single result with no indication for which form of the test was used or how it was selected. "I conducted a Dickey Fuller test ..." or "After testing, we concluded the series does (does not) contain a unit root ..." are hallmarks of this approach. The lack of transparency is not the only problem. Given the importance of the decisions you make about D_t, l, and the number of lags to include in your tests, a failure to apply a range of different tests could precipitate major errors in classification and, as a consequence of miss-classification, analysis.

The performance of the tests will vary from application-to-application and you are likely to find conflicting results. Each of the described tests has reasonable power if you have correctly chosen D_t, the sample size is large, and the test can accommodate the dynamics in the error process. However, the power of the tests varies as a function of several factors that are themselves difficult to discern. An ADF test may be best if the dynamics of the error follow a simple autoregressive process or if the first observation is unusual. But other tests may have more power. The PP and KPSS tests have more power than the ADF tests if the dynamics of the error are more complex and particularly if they are heteroscedastic. The DF-GLS tests, which estimate the intercept and trend more efficiently, are likely to produce more precise estimates of ϕ. The smaller standard errors mean the DF-GLS tests have better power than the ADF, PP, or KPSS tests in some circumstances but the DF-GLS tests suffer if the first observation is unusual or the sample size is small.

Rather than trying to choose the best test a-priori, it is better to apply a battery of tests and make your decision by weighing the preponderance of evidence. If the results across all the different tests are consistent for a series, you can have greater confidence that your classification decision is correct. If, as happens in most cases, you find mixed evidence when you apply a battery of tests, you should not arbitrarily pick and choose the results that suit you; instead you should use the information you have about the relative performances of the different tests to try and understand why these differences exist and which results might be more reliable given what you know about your data.

3.4 A Note on Presenting Results

After the tests have been conducted, the results should be communicated in a transparent way. You should not only identify which tests you applied and provide the conclusions you have drawn from the tests, you should also provide your audience with the information they need to draw their own conclusions. You

[14] Hobijn, Franses, and Ooms (2004, 484) explain the implications of this choice somewhat differently: "Choosing too large a bandwidth implies that the long run variance is overestimated: the test statistic becomes too small and the test will have little or no power in finite samples if one employs common nominal significance levels. On the other hand, if one chooses too small a bandwidth and the process is highly autoregressive, then the long run variance is underestimated, the test statistic becomes too large and the test is oversized."

may not be inclined to include this information in the body of a manuscript, but the information could be provided in footnotes, end notes, or appendices.

You should provide the results for your unit root tests. This includes the values of the test statistics and either the p-values for the statistics or the critical values that the statistics can be compared to. If you use tests that assume dynamic completeness of auxiliary regressions – like the ADF and DF-GLS tests – you should also report tests for residual serial correlation to demonstrate that the reported test statistics and p-values are reliable.

You should also provide information about the specification of the tests you report so that readers have the information they need to reproduce your results if they choose to. This includes a discussion of how you selected p_{max} for the ADF and DF-GLS tests and the results from each lag length (p) or lag truncation (l) selection strategy. While the general-to-specific model building strategy we describe in Chapter 5 can be applied when specifying these auxiliary regressions, others have suggested that selecting lag lengths based on information criteria. You should also consider reporting the results from each specification of D_t that is plausible based on what you observe in the data and you should also be transparent about how you decided which forms of D_t to test and which forms of D_t to ignore. Ideally your conclusions will be robust to the specification of the auxiliary regression models. Presenting full information ensures that others can assess the validity of inferences drawn from your tests. Table 8 in section 5.1.3 presents a workable template for reporting.

3.5 Structural Breaks

As discussed in Chapter 2, a structural break occurs when there is a change in the (structural) parameters that describe a data generating process. Most models and diagnostics assume that the underlying parameters are constant over the full sample period. If the structural parameters change, these assumptions are violated.

Structural breaks include a change in the level and/or trend in a time series. These shifts in mean and changes in trajectory are usually the result of a dramatic event or a major development. The introduction of a vaccine or a policy change can permanently alter a data generating process and this change will be reflected in the series. A structural break need not be permanent. In other circumstances a dramatic event may result in a temporary change in the series but the parameters that describe the data generating process have not changed. The 9/11 terrorist attacks in the United States, for example, produced a sudden shock that is visible in many economic and political time series. These outlier events are important for applied analysis becase they can comprise a significant portion of the variation in a series and a failure to account for them can confound diagnostics, hypothesis testing, and statistical inference.

When a time series is "disturbed" in some fashion by an event, the DGP includes one or more intervention variables in D_t. Assume a structural change in the process occurred at time $t = T_B$ and that any effects begin in the subsequent period. The effect may be temporary or permanent and may impact the expected value of the series or change the slope of a trend. Four types of intervention variables capture the effects of interventions in most cases.

- A *step* intervention (Ds_t) can be used to model a policy change or the introduction of a new technology. The step variable is coded zero before the event, one in the period where the event occurred, and one thereafter : $..., 0, 0, 0, 1, 1, 1, ...$ The step captures permanent upward or downward shifts in a process. For a stationary series, this would imply a new equilibrium level.

- A *pulse* intervention (Dp_t) can be used to capture an outlier event like 9/11. A pulse is coded one when the event occurs and zero otherwise. That is, the pulse variable is zero before the event, one in the period of the event, and zero in all subsequent time periods: $..., 0, 0, 0, 1, 0, 0, 0, ...$ This indicator variable captures sudden changes in the process that are not permanent.

- A *transitory* intervention variable (Dtr_t) can also be used to capture outlier events but the dynamics implied by the transitory variable are slightly different from a pulse. A transitory variable is coded zero before an event, one in the period when the event occurs, negative one in the next period, and zero thereafter: $..., 0, 0, 0, 1, -1, 0, 0, 0, ...$ The transitory variable captures course correction. An event causes a sudden increase or decrease in a variable but, in the next period, forces pull the series back toward the pre-intervention level. For a stationary series a transitory intervention implies immediate - or almost immediate - recovery; a pulse intervention implies a sudden change but a stationary series will return to its pre-intervention equilibrium gradually, over time.

- A *trend* intervention variable (Dt_t) implies a change in the trend of a trending series or the introduction of a trend to a non-trending series. A series that is not trending before the intervention will be trending after. This kind of variable could also be used to capture the change in a trending variable; the series is trending at one rate before the intervention and at a different rate after the intervention. The trend intervention is coded zero before the event that causes the change in trend, one in the period where the event occurs, and then increases afterwards: $..., 0, 0, 0, 1, 2, 3, 4, ...$ Trend interventions can be used to capture changes in trajectory.

A process may be affected by multiple interventions and an event may have combinations of effects on the process, necessitating inclusion of multiple intervention variables.

The implication of each intervention variable for the trajectory of a time series depends on whether the process is stationary or contains a unit root. Table 3 extends the information in Table 1 to include processes that contain structural breaks. It describes the nature of a stationary AR(1) process and a unit root process containing each type of intervention variable. Nothing prohibits these interventions from being included in a model without a constant, but we include one here, as this is the typical case.

The DGP presented in row 1 contains the step intervention Ds_t. The intervention variable essentially separates the series into pre- and post-event periods. If the process is stationary, the mean changes from the pre- ($\bar{y}_t = c/\phi$) to the post-intervention period ($\bar{y}_t = (c + \eta)/\phi$). However, if the series contains a unit root, the first period is characterized by a drift of c while the second period drift is $c + \eta$. In both periods, the series will exhibit a tendency to increase or decrease over time, but the rate of growth will differ.

The DGP in row 2 contains the pulse intervention Dp_t. If the process is stationary, this intervention will capture a transitory, or single-period, spike in y_t equal to η (the effects on y_t will decay over time). If the process contains a unit root, the process shifts by η, and the effect of the event is permanently incorporated into the series.

Consider the DGP in row 3, which includes the transitory intervention variable Dtr_t. If the series is stationary, η is added to y_t when the event occurs and subtracted in the subsequent period, producing a transitory "blip" – a change followed by oppositely signed change – in the data. However, if the process contains a unit root, y_t will shift up by η in the period after the event and shift down by η in the next period, thus the effect is temporary.

Finally, in row 4, the DGP includes c, t and Dt_t. The inclusion of Dt_t in a trend stationary process causes a "broken" trend in y_t, where the slope of the trend changes after the intervention. In a unit root process the effect is to change the magnitude of the (quadratic) trend.

Table 3: Stationary and Unit Root Processes with Structural Breaks

Data Generating Process	$\|\phi\| < 1$ (Stationary)	$\phi = 1$ (Unit Root)
$y_t = c + \phi y_{t-1} + \eta Ds_t + \varepsilon_t$	Mean stationary autoregressive process with a change in the mean	Unit root process with a change in the magnitude of the drift
$y_t = c + \phi y_{t-1} + \eta Dp_t + \varepsilon_t$	Mean stationary autoregressive process with a temporary intervention	Unit root process with drift c and one time shift
$y_t = c + \phi y_{t-1} + \eta Dtr_t + \varepsilon_t$	Mean stationary autoregressive process with a temporary intervention in which the series adds (or subtracts) η to the series in one period and then subtracts (adds) η in the next period	Unit root process with drift c and temporary effect on a unit root process
$y_t = c + \phi y_{t-1} + \delta t + \eta Dt_t + \varepsilon_t$	Trend stationary autoregressive process with a change in the slope of the trend	Unit root process with drift c and change in the magnitude of the (quadratic) trend

$D_{tr} = 1$ at $t = T_B + 1$ and $D_{tr} = -1$ at $t = T_B + 2$; $D_p = 1$ if $t = T_B + 1$; $D_s = 1$ if $t > T_B$ and zero otherwise; $Dt_t = \{1, 2, 3, \ldots\}$ beginning at time $T_B + 1$; and $D_T = t - T_B$ for $t > T_B$ and zero otherwise; T_B denotes the known time period in which the event (or break) occurred.

Like the DGPs in Table 1, where the effects of c and δt on the evolution of a time series differ depending on whether $|\phi| < 1$ or $|\phi| = 1$, the effect of the interventions depend on whether the series is a stationary or unit root process. The effects of any intervention variable cumulate in a unit root process, e.g., interventions that have a temporary effect on a stationary process have a permanent effect on a unit root process.

Structural breaks complicate unit root tests. For example, with a structural break, you can fail to reject the unit root null hypothesis, even when both the pre- and post-break portions of the series are stationary. To see why, imagine a stationary time series with two distinct sub-periods separated by an event that caused a change in the mean, such that the process contains Ds_t. Ignoring Ds_t and fitting a trend line through the entire sample will produce a positive (negative) slope if the shock increased (decreased) the mean. The end result is that the estimated value of ϕ in a simple autoregression is necessarily biased toward unity. As ϕ grows closer to unity, the process becomes more like a random walk and unit root tests become increasingly likely to fail to reject the null hypothesis. How, then, do we proceed to test for

a unit root in the presence of a structural break? The answer depends on whether the date of the break is known.[15]

3.5.1 Unit Root Testing with Structural Breaks

Structural break when the date is known If we know when the structural break occurred – perhaps we suspect 9/11 fundamentally changed views of privacy – the analyst might choose to apply any of the above unit root tests separately to the pre- and post-break data. However, because this reduces the size of the sample available for the test(s), the power of the test(s) can be quite poor, unless the analyst is fortunate enough to have a large sample in both periods. A second solution is to use the full sample period and to augment D_t to account for the break (Perron 1989, 1990; Perron and Vogelsang 1992). The basic assumption is that outlying events can be modeled as one-time changes in the deterministic part of the process after which variants of the Dickey-Fuller test can be used to test the hypothesis that $\phi = 1$.[16]

As above, we assume a structural change in the process occurred at time $t = T_B$ and that any effects begin in that period. We can test several hypotheses about the nature of the time series pre- and post-break. Perron (1989) provides a test to arbitrate among three types of pre- and post-break behavior. In each case, the process is assumed to be a unit root with drift under the null hypothesis and a trend stationary process under the alternative. Variants of the test allow for one-time shifts in the levels of the process, shifts in the drift or slope of the trend, or both. Perron (1990) and Perron and Vogelsang (1992) provide additional tests when the time series does not contain a deterministic trend under either the null or alternative hypothesis. They test the null hypothesis that the time series is a unit root process with a one-time shift while the alternative is that the time series is mean stationary with a change in mean occurring with the break.

The various tests proposed by Perron can all be understood within a single framework. We first consider the general case and then three additional cases likely to be of interest.

Case 1. Of the tests proposed by Perron, the most general form describes a series with trending behavior where the intervention causes a change in the nature of the trend. Thus, two possibilities are likely. Either y_t is a unit root with drift (given by c) that has undergone a one-time level shift (given by μ_1) as well as a change in the magnitude of the drift (given by μ_2) at the break, T_B:

$$H_0{:}y_t = c + y_{t-1} + \mu_1 Dp_t + \mu_2 Ds_t + \varepsilon_t. \tag{13}$$

Or the process is trend stationary with a shift in the intercept (given by μ_2) and the slope of the trend given by (μ_3):

$$H_A{:}y_t = c + \delta t + \mu_2 Ds_t + \mu_3 Dt_t + \varepsilon_t \tag{14}$$

where, to simplify our presentation, we assume white noise errors. However, the dynamics of the error may be stationary and invertible ARMA processes, in which case the effects of the break will be distributed over time. The first equation gives the null hypothesis for the test and the second gives the alternative hypothesis for the test.

[15] Many extensions of these tests have been proposed to account for multiple break points or when the break has a gradual effect on the time series. See (Carrion-i Silvestre, Kim, and Perron 2009).

[16] This method is referred to as the innovational outlier method. Its primary feature is that the intervention is assumed to follow the same dynamic structure as the error process and therefore for the effects of the break to be gradual. Alternative methods exist for "additive outliers" whose effects are assumed to be instantaneous, i.e., there is a single effect at the breakpoint. The additive outlier method models the break and the residuals are used as input for standard unit root tests.

Consider the effect of an event that occurred at $T_B = 100$. If y_t is generated by the null then in periods $t = 1$ through $t = 99$ the DGP is given by $\Delta y_t = c + \varepsilon_t$, in period $t = 100$ the process shifts by $\mu_1 + \mu_2$, and for the remainder of the time series the magnitude of the drift is given by $c + \mu_2$. In contrast, if y_t is generated by a trend stationary process, the DGP through $t = T_B$ is given by $y_t = c + \delta t + \varepsilon_t$ while post-break it is given by $y_t = c + \mu_2 + (\delta + \mu_3)t + \varepsilon_t$, such that both the intercept and the slope of the trend line change after the break.

Case 2. In the second case, we assume that the break either caused a change in the magnitude of the drift term in the random walk plus drift (the null hypothesis) or a change in the slope in a trend stationary process (the alternative hypothesis). Thus we omit Dp_t from the null hypothesis given by Equation 13 and we omit Ds_t from the alternative hypothesis given by Equation 14:

$$H_0: \quad y_t = c + y_{t-1} + \mu_2 Ds_t + \varepsilon_t \tag{15}$$

$$H_A: \quad y_t = c + \delta t + \mu_3 Dt_t + \varepsilon_t \tag{16}$$

Consider again the effect of an event that occurred at $t = 100$. In case 2, y_t is generated as a unit root with drift given by c through time $t = 99$ after which the magnitude of the drift changes permanently to $c + \mu_2$. If y_t is generated under the alternative hypothesis, the process is trend stationary both pre and post break but the slope of the trend changes from δ to $\delta + \mu_3$ at $t = 100$.

Case 3. Next, we consider the case in which we assume that the break either caused a jump in the level of a unit root process (with drift) or a jump in the intercept of a stationary process. The former is the null hypothesis; the latter is the alternative hypothesis. Thus we omit Ds_t in Equation 13 and we omit Dt_t from Equation 14:

$$H_0: \quad y_t = c + y_{t-1} + \mu_1 Dp_t + \varepsilon_t \tag{17}$$

$$H_A: \quad y_t = c + \delta t + \mu_2 Ds_t + \varepsilon_t \tag{18}$$

Continuing with our example, now under H_0 the process is a unit root with drift given by c for the full period of analysis, but at $T = 100$ there is a shift in y_t given by μ_1. Under H_A the intercept shifts from c to $c + \mu_2$ at $T = T_B$ and the slope of the trend is unchanged.

Case 4. In the final case, we assume that y_t does not contain a deterministic trend and so omit c from the formulation of the null hypothesis in Equation 13 and we omit t and Dt_t from the alternative hypothesis in Equation 14:

$$H_0: \quad y_t = y_{t-1} + \mu_1 Dp_t + \varepsilon_t \tag{19}$$

$$H_A: \quad y_t = c + \mu_2 Ds_t + \varepsilon_t \tag{20}$$

Under case 4, the process is either a unit root (without drift) with a one time jump in the level at $t = T_B$ given by μ_1 or a mean stationary process with a shift in the intercept given by μ_2, also at $t = T_B$.

In none of these cases are the null and alternative hypotheses nested. This means that having determined the appropriate null and alternative hypotheses, we conduct the selected test by combining

both to form a test regression that maintains all terms appearing in either the null or alternative hypothesis.

$$\textbf{Case 1:} \quad y_t = c + \delta t + \mu_1 Dp_t + \mu_2 Ds_t + \mu_3 Dt_t + \phi y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \varepsilon_t \tag{21}$$

$$\textbf{Case 2:} \quad y_t = c + \delta t + \mu_2 Ds_t + \mu_3 Dt_t + \phi y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \varepsilon_t \tag{22}$$

$$\textbf{Case 3.} \quad y_t = c + \delta t + \mu_1 Dp_t + \mu_2 Ds_t + \phi y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \varepsilon_t \tag{23}$$

$$\textbf{Case 4.} \quad y_t = c + \mu_1 Dp_t + \mu_2 Ds_t + \phi y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \varepsilon_t \tag{24}$$

In each case, p lagged differences are included in the model to ensure the residuals are white noise. We test the null hypothesis using the t statistic for $\phi = 1$. The distribution of the t-statistic depends on the proportion of observations before and after the break, $\lambda = T_{B-1}/T$, and the form of the test regression. Percentiles of the limiting distribution for various values of λ are provided by Perron (1989) for the first three cases and by Perron (1990) for case 4. **?** update these critical values and provide distributions for cases when the location(s) of the break point(s) are not known.

Structural break date unknown. If you do not know the date of the structural break, the data are used to identify the break point and then inferences are drawn in a similar fashion. The standard test was proposed by Zivot and Andrews (1992), who recommended a sequential approach to testing for a unit root in which a series of ADF test regressions are estimated, each including a simple pulse dummy variable for a candidate break point. The analyst draws inference from the ADF test regression where the t-statistic from the ADF test is at its minimum (i.e., when it is most negative). By selecting this break date, the test maximizes the chance that the null hypothesis will be rejected and the analyst concludes the series is stationary.[17] Critical values for the Zivot-Andrews test statistic are more negative than standard critical values, however, which may make it difficult to reject the unit root null hypothesis. Zivot and Andrews (1992) provide critical values for the test.

3.5.2 Stationarity Tests with Structural Breaks

With a structural break, the KPSS test is too likely to reject the stationary null. Lee and Strazicich (2001) proposed augmenting D_t in the test regression to account for a structural break occurring at a known time-point in similar fashion to that suggested in Section 3.5.1. To test the null hypothesis that a process, y_t, is mean stationary with a one-time shift in the level against the alternative of a unit root process, we augment the test regression to include D_L where $D_L = 1$ for $t \geq T_B$ and zero otherwise. T_B is the known break date:

$$y_t = c + \pi_1 D_L + e_t. \tag{25}$$

In order to allow for a trend stationary series with a one-time shift in the level and a change in the slope of the trend, we augment the test regression to include D_L and D_T, where $Dt_t = t - T_B$ for $t \geq T_B$ and zero otherwise:

$$y_t = c + \delta t + \pi_1 Ds_t + \pi_2 Dt_t + e_t. \tag{26}$$

[17] This test also differs from the case in which the break date is known in that the effect of the break is assumed to be instantaneous rather than having a dynamic effect on the process.

The test statistic is given as in Equation 12 for the case without a break. However, the distribution of the test statistic varies with the proportion of observations before the break, T_{B-1}/T. Table 1 in Lee and Strazicich (2001) provides critical values for both η_μ and η_τ for a variety of values of T_{B-1}/T.[18]

3.5.3 Seasonality in Unit Root Tests

Seasonality and Unit Root Tests Seasonality can affect the performance of unit root rests. For obvious reasons, a failure to account for higher order dependence at seasonal lags can undermine the inferences from DF and DF-GLS tests because both assume the residuals are white noise. Including high-order lags to control for seasonality can reduce the power of the tests in small to moderate samples. You not only are estimating additional nuisance parameters, you are also losing a potentially large number of observations by including multiple seasonal lags (Ghysels 1990). If you are using monthly data, for example, you would lose 24 observations if you include two seasonal lags.

Ghysels and Perron (1993) show that the PP tests perform better in these scenarios. The non-parametric adjustments used to obviate the problems with higher-order correlations do not require lags to be included in the test regressions. Dickey-Full tests can still be used to test for unit roots if an appropriate number of autoregressive terms are included to augment the model and whiten the residuals but this approach can produce non-trivial size distortions (Ghysels, Lee, and Noh 1994). If the results from the PP tests suggest a different classification than the ADF and DF-GLS tests, you should err on the side of the PP test because it does not include a large number of lags; though you should take care to set your lag truncation paramater l at a lag length long enough to account for all of the unmodeled dynamics in the data.

You might be tempted to apply unit root tests after filtering out the seasonal components of the data, using a seasonal ARMA filter or some other seasonal adjustment procedure. This approach should be avoided. Ghysels and Perron (1993) show that seasonal adjustments reduce the power of the DF and PP tests. Seasonal filters induce persistence that biases estimates of γ and affects the distribution of the test statistics (Ghysels 1990). You are better served applying tests to the unfiltered data, accounting for seasonal dynamics using the standard methods used to account for serial correlation.

Seasonal Unit Roots and Seasonal Unit Root Tests A time series y_t is said to have a *seasonal unit root* if differencing the series $(1-L^S)$ times produces a stationary series (Choi 2015, 164). The S exponent over the lag operate L is used to denote the seasonal periodicity of the series: $S = 4$ for quarterly data, $S = 12$ for monthly data, and so on. The quantity $(1 - L^S)$ suggest the order (1) and periodicity of the *seasonal differencing* required to render the series stationary. Just as regressions involving two unit root series are subject to the spurious regression problem, so are regressions involving two seasonal unit roots. And just as unit roots must be differenced to create stationary series (Δy_t), seasonal unit roots must be seasonally differenced to create stationary series ($(1 - L^S)y_t$).

A number of tests have been developed to identify seasonal unit roots. Dickey, Hasza, and Fuller (1984) propose tests for seasonal unit roots that parallel the standard DF tests. A seasonally differenced y_t is regressed on the seasonally lagged y_{t-s} and the test statistic on y_{t-s} is used to test the null hypothesis of a seasonal unit root. The critical values of the Dickey-Hasza-Fuller (DHF) tests are, like the critical

[18] In the case where the break point is unknown, Lee and Strazicich (2001) propose a minimum stationarity test in which a sequence of (infimum) KPSS tests are estimated, each allowing for a potential break point. The test regression that minimizes the test statistic is selected as the basis for inference. Following this procedure ensures that the chosen regression is least likely to reject the stationary null hypothesis. See Lee and Strazicich (2001) for further details.

values for the DF tests, non-standard with the shape of the distribuions and the extremity of the associated critical values changing based on D_t and S. The test regressions for the DHF tests can also be augmented like the ADF tests, adding lags of $(1 - L^S)y_t$ to account for serial correlation in the residuals. Alternative seasonal unit root and stationarity tests have been proposed by Hylleberg et al. (1990), Canova and Hansen (1995), and others; Hylleberg (1992) and Choi (2015) provide reviews of this literature and discuss the most commonly used tests.

4 Step Three: Interpret the Results Holistically

Armed with test results, we are ready to draw inferences. Unfortunately, no unit root or stationarity test is optimal without full knowledge of the dynamics of a time series. This is why we conduct multiple tests but the different tests can support contradictory conclusions and there are no hard and fast rules you can use to arbitrate among competing results. Another complication is that the features of your sample – including the length of the sample, the sample window, and the value of the initial observation – affect the ability of the tests to discern between stationary and unit root processes. This casts doubt on any conclusions based on these hypothesis tests. To navigate these uncertainties, you should not only consider the results of the different tests but also consider which tests are most reliable given the features of your data.

4.1 How Should You Draw Inferences from Your Test Results?

You will often find that inferences from the full set of hypothesis tests are at odds. While inconsistent test results engender uncertainty about whether a time series is mean (or trend) stationary or contains a unit root, your knowledge of the relative power and size of the tests can help put your competing results in perspective.

None of the tests are bulletproof. The unit root and stationarity tests we describe have low power in small to moderate samples, particularly when the data are strongly autoregressive. As a practical matter, this means that we are too likely to fail to reject the unit root null hypothesis with the DF, PP, and DF-GLS tests and too likely to fail to reject the stationarity null with the KPSS tests. If all tests indicate a time series is stationary, it is likely a stationary process. If all of the tests indicate a series contains a unit root or the KPSS test produces the only inconsistent result, you should proceed more cautiously. If your sample size is small, you should be skeptical of the validity of these tests. If the sample size is large, and you are confident you specified D_t correctly and selected an appropriate lag length or lag truncation parameter (l), you can be more confident in the results.[19]

When you are not confident about your classification of your series, it can be difficult to proceed with your analysis. If your variables can all be classified as stationary, the models we describe in Chapters 5 and 8 are appropriate. If your variables can all be classified as non-stationary, the models we describe in Chapter 6 and 9 are appropriate. If you do not know whether your variables can be classified as stationary or non-stationary or if you find yourself classifying some variables as stationary and other variables as non-stationary, you will have trouble relying on any of these methods. Chapter 7 introduces a strategy for inference that incorporates uncertainty about classification into the hypothesis testing procedure. This approach can be used whether your data are stationary, non-stationary, and in situations where you are unsure. Alternately, you could procede with separate analyses that cover the different possibilities for

[19]The level of significance used to evaluate the tests is an important consideration as well.

the stationarity of your data and then compare the results. Being transparent about your conflicting test results and the inferences that rely on them are keys to presenting your findings honestly.

4.2 How do Features of the Sample Influence Test Results?

As we alluded to above, sample size is an important consideration in the evaluation of hypothesis tests. Other features of your data can also be important. The sample window – the period of time covered by the sample data – including the values of the first and last observation, can have an outsized influence on test results. We consider each in turn.

4.2.1 Sample size

The length of a series provides information about the reliability of unit root and stationarity tests. When time series are short ($T < 50$), or moderately short and highly persistent ($T < 100$), there may not be enough information in the data for any of the tests we described. While limited information hampers the power of all statistical tests, this is a particularly pernicious problem for unit root tests that have low power. In the extreme, where the process is white noise ($\phi = 0$), each of the tests above is likely to produce correct inferences, even with fairly short samples. If $\phi = .85$, the optimal unit root test will correctly reject the null hypothesis less than 20% of the time for $T = 50$ and an α level of 0.05 (Podivinsky and King 2000). For samples of size 500, the power of unit root tests is generally quite good, unless ϕ is very close to one.

The implication for the analyst with small samples is clear. If you can reject the unit root null hypothesis, you can be quite confident the data are stationary, conditional on having selected the correct form of D_t and an appropriate lag length or lag truncation value. However, if you fail to reject the unit root null hypothesis, you should be cautious – interpret the tests with care and a healthy dose of skepticism, particularly if your results are not consistent across tests.

4.2.2 Sample window

The sample window refers to the period of time covered by the data. Your sample window may be atypical for any number of reasons. A structural break may significantly affect the variation in the sample. Even accounting for the break in hypothesis tests, the proportion of variance in-sample due to the break may be so large that hypothesis tests are unreliable. Or it may be that, as luck would have it, the dominant features of the data in-sample are inconsistent with our knowledge of the process more generally. This case presents a different type of challenge. If you wish to explain the behavior of a process in an unusual time period, you proceed with hypothesis tests per usual, but it would be a mistake to draw inferences about the process outside the sample window.

It may also be the case that the first or last observations contain the minimum or maximum value in the sample. The importance of the magnitude of the first observation in the data is often ignored, but the first observation has an outsized effect on unit root tests. If the time series has just begun – the first observation in the sample is the first observation in fact – the initial observation of an autoregressive process will impart a trend that will dominate its behavior in sample so that the mean, variance, and covariances will depend on time, even if the process is stationary, unless the sample size is quite large (how large depends on the persistence in the series). The same is true if the sample is small. The DF-GLS test is particularly sensitive to the magnitude of the first observation and has been shown to be less powerful

than the DF test when the first observation is large or the sample size is small (Choi 2015). This fact, can account for discrepancy in the results of the tests.

In short, it is important to identify the dominant features of the data in-sample and take these into account when drawing inferences from hypothesis tests that discriminate between stationary and unit root processes. The tests we have presented can account for many of these features (namely trends and structural breaks) but not all (namely small samples and sample windows with atypical behavior). The careful analyst will need to bear these in mind when interpreting hypothesis tests.

5 Examples

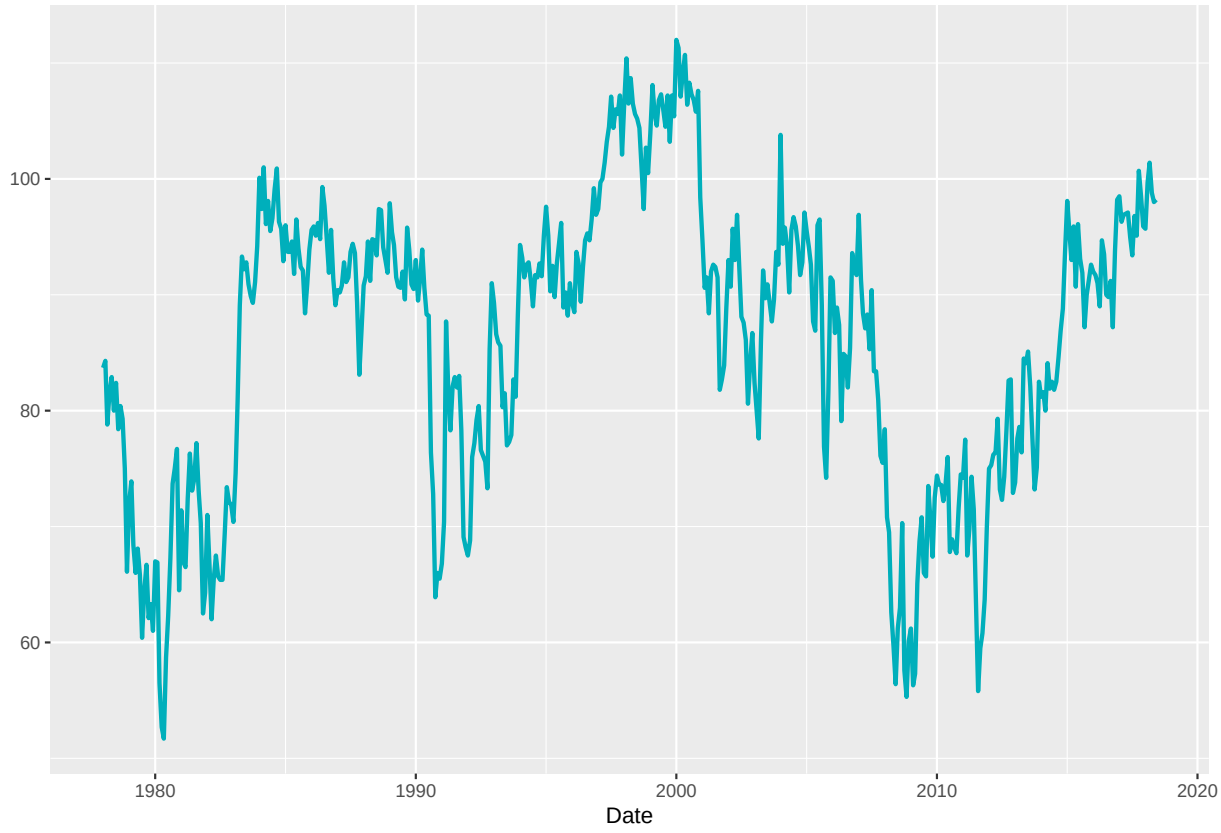
The testing procedures we outline above are complicated and especially difficult to think about in the abstract. You should carefully choose your test based on a careful examination of your time series and you should be transparent about your decisions. We illustrate a testing strategy with two examples. First, we determine which deterministic features to include in the test regression: whether to include a constant, trend, and/or dummy variables to capture the effects of interventions. Next, we conduct a full set of hypothesis tests, adopting recommended, systematic strategies for selecting lag lengths or lag truncation values for the tests. Finally, we interpret the results holistically, paying attention again to the plausible characterizations of the time series and its behavior in sample, as well as considering the length of the sample, the nature of the initial observation, and the strengths and weaknesses of the tests.

5.1 The Index of Consumer Sentiment

For our first example, we consider the Index of Consumer Sentiment (ICS). Each month since January of 1978, the University of Michigan's Survey Research Center has surveyed 500 respondents in order to gauge the economic evaluations and behavior of consumers in the United States. The ICS is meant to tap "optimism and certainty" or "pessimism and uncertainty" felt by consumers. It is calculated from five questions which ask respondents for their assessments of their personal finances (retrospective and prospective), national economic performance (one and five years out), and (current) buying conditions (Kellstedt, Linn, and Hannah 2015; Curtin 2008).[20] For each of the five indicators, a series is created by summing the percentage of positive responses, subtracting the percentage of negative responses, and adding 100. The ICS averages these individual values, weighting each indicator equally. For our analysis we consider monthly values of the ICS from January 1978 through June of 2018 for $T = 486$. We present a graph of the ICS for this time period in Figure 2. A visual inspection tells us a lot.

[20]The specific survey questions are 1. Do you think now is a good or bad time for people to buy major household items? [good time to buy; uncertain, depends; bad time to buy]. 2. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago? [better;same;worse]. 3. Now turning to business conditions in the country as a whole. Do you think that during the next twelve months, we'll have good times financially or bad times or what? [good times; uncertain;bad times]. 4. Looking ahead, which would you say is more likely? That in the country as a whole we'll have continuous good times during the next five years or so, or that we'll have periods of widespread unemployment or depression, or what? [good times;uncertain;bad times]. 5. Now looking ahead do you think that a year from now, you (and your family living there) will be better off financially, or worse off, or just about the same as now? [better;same;worse].

Figure 2: Michigan Index of Consumer Sentiment, January 1978-July 2018 ($T = 486$)



5.1.1 Step 1: Determine the Types of Processes that May Reasonably Have Generated the Data

Does the series contain a deterministic trend or a stochastic trend? If we begin by thinking about how the data were generated, there is no reason to expect the ICS contains a deterministic trend. Evaluations of the economy are likely to be relatively more optimistic and pessimistic as economic performance warrants. The behavior of the ICS is consistent with our understanding of the US economy. The time series ranges from 51.7 to 112.0, with the minimum value occurring in early 1980 amidst high inflation and the maximum value occurring in January of 2000 after a long period of economic growth.[21] Examining Figure 2, we observe evidence that the ICS does not contain a deterministic trend. We see a local minimum that corresponds with the great recession of 2008, which is followed by a gradual improvement in evaluations. The series has no tendency to increase (or decrease) by a fixed amount over time. Consumer sentiment is, however, clearly persistent; lower values tend to be followed by lower values and higher values tend to be followed by higher values. It is not obvious whether the series is stationary or whether its wandering behavior indicates a stochastic trend.

Is it plausible the time series has a mean of zero? As noted, the ICS ranges from 51.7 to 112.0 over this period. It is not plausible to have a mean of zero.

[21] Theoretically the series has a minimum value of 0 and a maximum value of 200, although it never approaches these values.

What do we know about the occurrence of events or shocks during the sample period? It might be reasonable to expect large negative shocks to the economy, such as the 2008 recession, significantly alter the evolution of the ICS or that the ICS is affected by the party in the White House (or changes in the party in the White House). Indeed, the ICS drops precipitously with the onset of the 2008 recession. This does not seem to lead to a new mean in the series so, for the sake of this example, we do not explore the possibility that this is a structural break in the series.

These observations lead us to conduct hypothesis tests under the assumption that D_t includes (only) a constant. Thus, under the unit root null hypothesis the ICS is a pure random walk while under the alternative hypothesis the ICS is a mean stationary process. For the KPSS test, the series is a mean stationary process under the null and a pure random walk under the alternative.

5.1.2 Step 2: Conduct Hypothesis Tests

Which hypothesis tests should we use? We estimate each of the unit root and stationarity tests described below. For each we present detailed results designed to be used to select the appropriate form of the test regression and we illustrate how this decision does or does not effect our inference.

How Should We Select the Lag Length or Lag Truncation Parameter for the Test?

The Dickey Fuller Test. We begin with the (Augmented) Dickey Fuller test. We begin by selecting $p_{max} = 12$. This choice allows for the possibility of seasonality in the data. We then estimate the ADF test for specifications including lags 12, 11, 10,..., to zero. Table 4 presents the test statistic, τ_μ, the associated p-value of the test statistic, the AIC, the corrected AIC, and the BIC, the t-statistic on the highest lagged difference included in the test regression, the p-value for the Breusch-Godfrey test of the null hypothesis of no serial correlation in the residuals of the test regression, here for 12 lags, as well as the number of observations used for the test. We report the information criteria to illustrate how inferences from the test depend on the strategy chosen for lag length selection.

It is easy to see that the inference we draw from the test depends on the lag length, p, selected for the test regression. For tests with lag lengths three or greater, we would fail to reject the null hypothesis at the 0.05 level. For tests with smaller numbers of lags, we reject the null hypothesis that the ICS is a unit root process. So which is the appropriate number of lags? We use the general-to-specific model selection strategy to select p.

The general-to-specific strategy requires identifying the test regression in which the t-statistic for the highest order lagged difference included in the regression exceeds $\pm$ 1.60. In this case that is for $p = 12$. Then we test whether the residuals are serially correlated. The Breusch-Godfrey test fails to reject the null hypothesis of no serial correlation. Thus, using the general-to-specific model selection strategy, the model with $p = 12$ should be used. Based on the evidence in Table 4, we can see that removing the 12th lag does not induce serially correlated errors, so we might select the model with $p = 5$, the next highest order lagged difference term to have a t-statistic exceeding $\pm$1.60. However, it is also the case that removing this lag does not induce significant autocorrelation. In fact, it is not until the test regression with $p = 1$, that significant serial correlation remains in the test regression residuals. We might thus draw our inference from the test regression with $p = 2$ where $\tau_\mu = -2.979$. The Dickey Fuller p-value for the test is 0.038. Thus we would conclude that we can reject the null hypothesis the series is a pure random walk in favor of the alternative that the series is mean stationary.

Using an information criterion to select p, however, leads us to select either $p = 5$ (AIC and AIC_c) or $p = 0$ (BIC). In either case, we fail to reject the null that the residuals are white noise at the 0.05 level using the Breusch-Godfrey test, although selecting a model with $p = 0$ we can reject the null at 0.10. For $p = 5$, $\tau_\mu = -2.371$, which has a p-value of 0.15 and for $p = 0$ $\tau_\mu = -3.362$ with a p-value of .01.

Table 4: (Augmented) Dickey-Fuller Test Results: Michigan Index of Consumer Sentiment, January 1978 - June 2018, $D_t = (1, 0)$

Test Lags	τ_μ	τ_μ (p-value)	AIC	AIC-c	BIC	Diff t-value	Breusch-Godfrey (p-value)	Obs
12	−2.619	0.090	2624.939	2625.990	2687.326	2.107	0.710	473
11	−2.384	0.147	2627.493	2628.410	2685.721	0.582	0.543	473
10	−2.336	0.161	2625.841	2626.634	2679.910	−0.740	0.605	473
9	−2.440	0.131	2624.403	2625.082	2674.312	1.003	0.762	473
8	−2.342	0.159	2623.432	2624.005	2669.182	1.045	0.668	473
7	−2.240	0.192	2622.546	2623.022	2664.137	−1.170	0.589	473
6	−2.388	0.146	2621.939	2622.328	2659.371	0.311	0.449	473
5	−2.371	0.151	2620.038	2620.348	2653.310	−2.230	0.460	473
4	−2.654	0.083	2623.060	2623.301	2652.173	−0.651	0.168	473
3	−2.764	0.064	2621.488	2621.668	2646.443	−1.351	0.147	473
2	−2.979	0.038	2621.328	2621.457	2642.124	−1.947	0.090	473
1	−3.300	0.015	2623.135	2623.220	2639.771	−0.154	0.041	473
0	−3.362	0.013	2621.159	2621.210	2633.636		0.068	473

Note: The Breusch-Godfrey test for serial correlation was calculated on 12 lags. The column labeled "Diff t-value" contains the t statistic for the coefficient on the highest order lagged difference in the test regression.

DF-GLS test. We turn next to the DF-GLS test. Continuing to use $p_{max} = 12$, we proceed in like fashion to the ADF test. Table 5 presents the results for the DF-GLS test for 12 to zero lags. Here we report the value of the test statistic and its accompanying p-value along with (for now) the BIC. The BIC selects $p = 7$. The value of the DF-GLS statistic for this test regression is -2.11 (p=0.24).

Phillips-Perron Table 6 presents the test results for the PP test with both a short and long lag truncation parameter. Recall that this choice determines the number of lags that are included in the calculation of the long-run variance of the series used to make a non-parametric adjustment to the t-test from a Dickey-Fuller test regression with no lagged dependent variables. The values in this example are $l = 4$ and $l = 17$. The p-value associated with the test statistics is less than 0.05 in both cases. Thus, both test results lead us to the same inference: the monthly ICS time series is mean stationary.[22]

[22]No information on the kernel or lag length calculation are given for the ur.pp (or ur.kpss) test(s) in R. However, the values chosen for the test are consistent with those recommended in conjunction with the Bartlett kernel.

Table 5: DFGLS Test Results: Michigan Index of Consumer Sentiment, January 1978 - June 2018, $D_t = (1, 0)$

Test Lags	Test Statistic	p-value	BIC
12	−2.530	0.109	2681.831
11	−2.322	0.165	2685.641
10	−2.223	0.198	2689.800
9	−2.327	0.164	2690.275
8	−2.206	0.205	2690.268
7	−2.110	0.241	2690.088
6	−2.262	0.185	2691.548
5	−2.246	0.190	2690.424
4	−2.533	0.108	2694.384
3	−2.643	0.085	2693.169
2	−2.869	0.050	2693.829
1	−3.196	0.021	2698.139
0	−3.289	0.016	2696.605

Table 6: Phillips-Perron Test Results: Michigan Index of Consumer Sentiment, January 1978 - June 2018, $D_t = (1, 0)$

Test	Test Lags	Test Statistic	p-value
Z_τ	short	−2.984	0.037
	long	−3.094	0.028

KPSS test. We present results for the KPSS test in Table 7 for both the short and short lag truncation values, here (5 and 17). Recall that the size of the KPSS test tends to be poor such that we may reject the null of mean stationarity even though using each of the tests above we could reject the unit root null hypothesis. The results of the KPSS test using the long truncation parameter suggest the ICS is mean stationary ($0.24 < 0.463$). However, we would reject the null of mean stationarity using a shorter lag truncation parameter ($0.66 > 0.463$).

Table 7: KPSS Test Results: Michigan Index of Consumer Sentiment, January 1978 - June 2018, $D_t = (1, 0)$

Lags	Test Statistic	5% critical-value
short	0.665	0.463
long	0.250	0.463

5.1.3 Step 3: Interpret the Results Holistically

Below we summarize the results from the selected test regressions, including information necessary for readers to evaluate the testing procedure. In particular, we report the length (and dates) of the sample period, the value for p_{max}, as well as the justification for the chosen value, the level of significance used for inference, the inference from each test, and the software package in which the tests were conducted.[23]

How Should We Draw Inferences from the Set of Hypothesis Test Results? We can now draw inference and assess our level of confidence that the ICS is a mean stationary as opposed to unit root process. We first note that inference varies both within unit root and the KPSS tests and across them. Within the unit root tests, the Dickey-Fuller and Phillips-Perron tests are at odds. However, the Phillips-Perron tests *generally* have more power and we might therefore choose to infer the process is mean stationary. The KPSS test itself leads to different inferences depending on selection of l. Two factors lead us to weigh inference from the long lag truncation more heavily. First, as noted above, KPSS showed that the more autocorrelated the original series, the longer the lag truncation parameter value needs to be for the test to have an acceptable size, i.e., to avoid rejecting the stationary null too frequently. Second, given that we have a relatively long time series, we should not have to worry about any potential power distortions from selecting a long lag truncation parameter. Inference in this case also favors the conclusion that the ICS is mean stationary.

How do Features of the Sample Influence Test Results? In conducting the analysis of the ICS we are fortunate in two respects. First, the length of the sample is quite long. However, given that the ICS is quite persistent, even 486 time points may not be enough for the tests to discriminate effectively between a mean stationary process and a unit root. Second, (and related to the first in this case) the sample window covers a long enough time period that events, such as the recession of 2008, do not overwhelm the variation of the series in sample. This limits the influence of specific events on test results. But given that we have not accounted for the recession (yet), we should not put too much weight on our findings to date.

Our takeaway from this analysis is that it is more likely the ICS is mean stationary rather than a unit root process, but we would not be comfortable placing a wager on this decision. In fact, the final

[23] Note that if any auxiliary tests are performed as part of the testing procedure these should be included in the table.

Table 8: Unit Root and Stationarity Tests: ICS, January 1978 through July, 2018 ($T = 486$)

Test	Model Selection Strategy	l	Test Statistic	p-value/ Critical Value	Bresch Godfrey (p-value)	Decision*	Inference
Dickey-Fuller (τ_μ)							
	GS	12	-2.619	0.090	0.710	Fail to Reject	Unit Root
	AIC_c	5	-2.371	0.151	0.460	Fail to Reject	Unit Root
DF-GLS							
	GS						
	BIC						
Phillips-Perron (Z_τ)							
	short	4	-2.984	0.037	NA	Reject	Mean Stationary
	long	17	-3.094	0.028	NA	Reject	Mean Stationary
KPSS (η_μ)							
	short	4	0.665	0.463	NA	Reject	Unit Root
	long	17	0.250	0.463	NA	Fail to Reject	Mean Stationary

Note: $D_t = (1, 0)$ for all tests such that the null hypothesis for the DF, DF-GLS, and PP tests is that the ICS is a unit root (without drift); the alternative hypothesis is that the ICS is mean stationary. The null and alternative hypotheses are reversed for the KPSS test. $p_{max} = 12$ to accommodate potential seasonal autocorrelation.
$* \ \alpha = 0.05$ is used to make the decision to reject or fail to reject the null hypothesis.
All tests were conducted in R using the urca package. Short lag truncation for both the PP and KPSS tests is based on the formula: $trunc(4 * (T/100)^{0.25}) = 4$; long lag truncation is based on the formula: $trunc(12 * (T/100)^{0.25}) = 17$ as given in Schwert (1989) and used in urca.

conclusion we draw is that we cannot be certain the correct classification of the ICS. In Chapter 5 we will use time series regression models to analyze the relationship between the ICS and a number of economic and political variables assuming the series is stationary. In Chapter 7, we will illustrate how to assess the same relationship accounting for our uncertainty in the classification of the ICS.

6 Conclusion

Although statistical software packages make stationarity and unit root testing seem easy, it is a complicated process to know the exact form the tests should take and then to make the right inferences. The decisions about the unit root question are consequential as the following chapters will show. Once you've decided your data are stationary, you can worry less about spurious findings and use the simpler models we present in Chapter 5. With unit root data, however, the possibility of cointegration requires careful exploration that we cover in Chapter 6. But one lesson of this chapter, especially with our ICS example, is that even with a lot of data, it can be very hard to definitively know the data generating process of your time series. This can lead you to try multiple assumptions while testing your hypotheses or to try models, such as the bounds methods in Chapter 7, that rely on fewer assumptions.

References

Andrews, D. 1991. "Heteroskedasticity and Autocorrelation Consistent Covariant Matrix Estimation." *Econometrica* 59(3): 817–858.

Campbell, John Y, and Pierre Perron. 1991. "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots." *NBER macroeconomics annual* 6: 141–201.

Canova, Fabio, and Bruce E Hansen. 1995. "Are seasonal patterns constant over time? A test for seasonal stability." *Journal of Business & Economic Statistics* 13(3): 237–252.

Carrion-i Silvestre, Josep Lluís, Dukpa Kim, and Pierre Perron. 2009. "GLS-based unit root tests with multiple structural breaks under both the null and the alternative hypotheses." *Econometric theory* 25(6): 1754–1792.

Choi, In. 2015. *Almost All about Unit Roots*. Cambridge University Press.

Cohen, Jacob. 2013. *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Academic Press.

Curtin, Richard. 2008. "Consumer Sentiment Index." *Encyclopedia of Survey Research Methods* 2: 135–8.

Dickey, David A., and Wayne A. Fuller. 1979. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association* 74: 427–431.

Dickey, David A, and Wayne A Fuller. 1981. "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root." *Econometrica: Journal of the Econometric Society* pp. 1057–1072.

Dickey, David A, David P Hasza, and Wayne A Fuller. 1984. "Testing for unit roots in seasonal time series." *Journal of the American Statistical Association* 79(386): 355–367.

Dolado, Juan J, Tim Jenkinson, and Simon Sosvilla-Rivero. 1990. "Cointegration and unit roots." *Journal of economic surveys* 4(3): 249–273.

Elliott, G., T. Rothenberg, and J. H. Stock. 1996. "Efficient Tests for an Autoregressive Unit Root." *Econometrics* 64(July): 813–836.

Enders, Walter. 2015. *Applied Econometric Time Series*. 4th ed. New York: Wiley and Sons.

Enders, Walter, and Clive William John Granger. 1998. "Unit-root tests and asymmetric adjustment with an example using the term structure of interest rates." *Journal of Business & Economic Statistics* 16(3): 304–311.

Fuller, Wayne A. 1976. *Introduction to Statistical Time Series*.

Ghysels, Eric. 1990. "Unit-root tests and the statistical pitfalls of seasonal adjustment: the case of US postwar real gross national product." *Journal of Business & Economic Statistics* 8(2): 145–152.

Ghysels, Eric, and Pierre Perron. 1993. "The effect of seasonal adjustment filters on tests for a unit root." *Journal of Econometrics* 55(1-2): 57–98.

Ghysels, Eric, Hahn S Lee, and Jaesum Noh. 1994. "Testing for unit roots in seasonal time series: some theoretical extensions and a Monte Carlo investigation." *Journal of econometrics* 62(2): 415–442.

Granger, C.W.J. 2010. "Some Thoughts on the Development of Cointegration." *Journal of Econometrics* 158: 3–6.

Hobijn, Bart, Philip Hans Franses, and Marius Ooms. 2004. "Generalizations of the KPSS-test for stationarity." *Statistica Neerlandica* 58(4): 483–502.

Hylleberg, Svend. 1992. *Modelling Seasonality*. Oxford, UK: Oxford University Press.

Hylleberg, Svend, Robert F Engle, Clive WJ Granger, and Byung Sam Yoo. 1990. "Seasonal integration and cointegration." *Journal of econometrics* 44(1-2): 215–238.

Kellstedt, Paul M, Suzanna Linn, and A Lee Hannah. 2015. "The Usefulness of Consumer Sentiment: Assessing Construct and Measurement." *Public Opinion Quarterly* 79(1): 181–203.

Kwiatkowski, D., P.C.B. Phillips, Peter Schmidt, and Yongcheol Shin. 1992. "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root." *Journal of Econometrics* 54(October): 159–78.

Lebo, Matthew J, and Taylor Grant. 2016. "Equation Balance and Dynamic Political Modeling." *Political Analysis* 24(1): 69–82.

Lee, Junsoo, and Mark Strazicich. 2001. "Testing the Null of Stationarity in the Presence of a Structural Break." *Applied Economics Letters* 8(6): 377–382.

Newey, Whitney K, and Kenneth D West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covaraince Matrix." *Econometrica* 55(3): 703–708.

Newey, Whitney K, and Kenneth D West. 1994. "Automatic Lag Selection in Covariance Matrix Estimation." *The Review of Economic Studies* 61(4): 631–653.

Ng, Serena, and Pierre Perron. 1995. "Unit Root Tests in ARMA Models with Data-dependent Methods for the Selection of the Truncation Lag." *Journal of the American Statistical Association* 90(429): 268–281.

Ng, Serena, and Pierre Perron. 2001. "Lag length selection and the construction of unit root tests with good size and power." *Econometrica* 69(6): 1519–1554.

Park, Joon Y, and Peter CB Phillips. 1988. "Statistical inference in regressions with integrated processes: Part 1." *Econometric Theory* 4(3): 468–497.

Perron, Pierre. 1988. "Trends and Random Walks in Macroeconomic Time Series: Further Evidence from a New Approach." *Journal of Economic Dynamics and Control* 12(2-3): 297–332.

Perron, Pierre. 1989. "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis." *Econometrica: Journal of the Econometric Society* pp. 1361–1401.

Perron, Pierre. 1990. "Testing for a unit root in a time series with a changing mean." *Journal of Business & Economic Statistics* 8(2): 153–162.

Perron, Pierre, and Serena Ng. 1996. "Useful Modifications to Some Unit Root Tests with Dependent Errors and their Local Asymptotic Properties." *The Review of Economic Studies* 63(3): 435–463.

Perron, Pierre, and Serena Ng. 1998. "An Autoregressive Spectral Density Estimator at Frequency Zero for Nonstationarity Tests." *Econometric theory* 14(5): 560–603.

Perron, Pierre, and Timothy J Vogelsang. 1992. "Testing for a Unit Root in a Time Series with a Changing Mean: Corrections and Extensions." *Journal of Business & Economic Statistics* 10(4): 467–470.

Phillips, Peter CB, and Pierre Perron. 1988. "Testing for a Unit Root in Time Series Regression." *Biometrika* 75(2): 335–346.

Podivinsky, Jan M, and Maxwell L King. 2000. "The Exact Power Envelope of Tests for a Unit Root.".

Said, E, and David A Dickey. 1984. "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order." *Biometrika* 71(3): 599–607.

Schwert, G William. 1989. "Tests for Unit Roots: A Monte Carlo Investigation." *Journal of Business & Economic Statistics* 7(2): 147–159.

Sims, Christopher A, James H Stock, and Mark W Watson. 1990. "Inference in linear time series models with some unit roots." *Econometrica: Journal of the Econometric Society* pp. 113–144.

White, John S. 1958. "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case." *The Annals of Mathematical Statistics* pp. 1188–1197.

Zivot, Eric, and Donald W K Andrews. 1992. "Further Evidence on the Great Crash, the Oil-price Shock, and the Unit-root Hypothesis." *Journal of Business & Economic Statistics* 10: 251–270.