Principles and risks of forecasting

Robert Nau
Fuqua School of Business, Duke University
September 2014

"I have seen the future and it is very much like the present only longer." –K. Albran, *The Profit*

Introduction

Forecasting can take many forms—staring into crystal balls or bowls of tea leaves, combining the opinions of experts, brainstorming, scenario generation, what-if analysis, Monte Carlo simulation, solving equations that are dictated by physical laws or economic theories—but *statistical* forecasting, which is the main topic to be discussed here, is the art and science of forecasting from *data*, with or without knowing in advance what equation you should use.   The idea is simple: look for statistical patterns in currently available data that you believe will continue into the future.   In other words, figure out the *way* in which the future will look very much like the present, only longer.

This may sound simple, but in practice it can be quite difficult, requiring analytical skill, experience in working with data, and a good deal of background research. When you have obtained a promising data set and begun to analyze it, you may at first see complex relationships whose pattern is not obvious, or (what is often worse), you may see patterns that aren't really there.   Some important patterns may not be visible because you have not looked at the data in the right way or identified all the relevant explanatory variables or thought about all their possible connections.   These obstacles can be overcome by using the statistical tools and modeling principles discussed on this web site[1]: viewing the data from many angles before getting started, identifying candidate models that are suggested by the patterns you discover and by what you have learned from your background research, using mathematical transformations to straighten out curves or stabilize time patterns if needed, fitting models to the data and evaluating their goodness-of-fit in both absolute and relative terms, looking for evidence that the models' assumptions may be incorrect, and drawing on everything else you know about the situation in order to apply reality checks. By the end of the day you hope to come up with a model that yields useful predictions *and* whose margins of error are known *and* which tells you some things you didn't already know.

Signal vs. noise

The variable you want to forecast can be viewed as a combination of signal and noise. The signal is the predictable component, and the noise is what is left over.   The term "noise" is intended to conjure up an analogy with the sort of noise or static that you hear on a busy street or when listening to a radio station with a weak signal. In fact, audible noise and noise in your data

---

are statistically the same thing. If you digitize a noisy audio signal and analyze it on the computer, it looks just like noisy data, and if you play noisy data back through a speaker, it sounds just like audio noise.

The technical term for a data series that is pure noise is that it is a sequence of "independent and identically-distributed (i.i.d.) random variables."   The RAND( ) function in Excel generates i.i.d. random values uniformly distributed between 0 and 1 whenever it is recalculated or copied. This in turn can be used to generate other standard random variables using their inverse (INV) functions. For example, the formula =NORM.S.INV(RAND( )) generates a standard normal random variable. Crystal Ball, a leading software package for simulation, allows random functions like this to be embedded in spreadsheet models and then recalculated thousands of times to simulate uncertainty in model parameters or uncertainty about future values of random variables. However, we won't be creating our own random variables; we will be analyzing ones that arise naturally.

It is up to you to find a forecasting model (in the form of a mathematical equation) that captures the signal buried in the noise and extrapolates it in an appropriate fashion.   This is not always easy, because sometimes it is hard to separate the two.   On one hand, very complex patterns in data may look quite random until you know what you are looking for, and on the other hand, data that are completely random may appear to the naked eye to have some kind of interesting pattern. Sensitive statistical tests are needed to get a better idea of whether the pattern you see in the data is really random or whether there is some signal yet to be extracted.   If you fail to detect a signal that is really there, or falsely detect a signal that isn't really there, your forecasts will be at best suboptimal and at worst dangerously misleading.

Some simple cases

Some of the simplest signal-noise patterns that you find in data are i.i.d. variations around a line or curve, or i.i.d. changes relative to an earlier period in time. The very simplest case is that of i.i.d. variations around a *horizontal line*, like this:
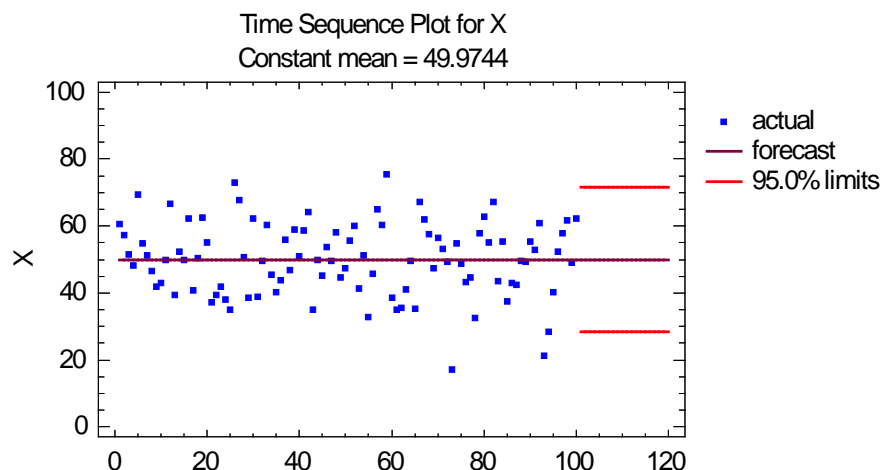
The appropriate forecasting model for this is data series is the *mean model*, in which you simply predict that series will equal its mean value in every period. This is a special case of a regression model in which there is an intercept term but no independent variables, i.e., an "intercept-only" model. A variable that measures a response to a survey question or a property of objects coming off an assembly line might look like this. Variables that have more complicated patterns might look like this after some mathematical transformation. For example, if you transform a time series of stock prices into daily or monthly *percentage changes*, it might look like the picture above
.
The mean model is not as trivial as it might first appear: you need to estimate the mean as well as the standard deviation of the variations around the mean, and the standard deviation needs to be used appropriately to calculate confidence intervals for the predictions.   The accuracy of the parameter estimates and forecasts also depends on the sample size in a systematic way. These issues, which arise in all forecasting models, are discussed in more detail later in a separate handout.

A slightly more interesting pattern is that of i.i.d. variations around a *sloping* line on a plot of your variable-of-interest versus some other variable, which indicates some degree of *correlation* between the two variables. The coefficient of correlation between two variables measures the strength of the linear relationship between them, on a scale of -1 to +1. We'll discuss that in more detail when we get to the topic of linear regression analysis.

Time Sequence Plot for Z
Linear trend = 51.5832 + 0.968143 t
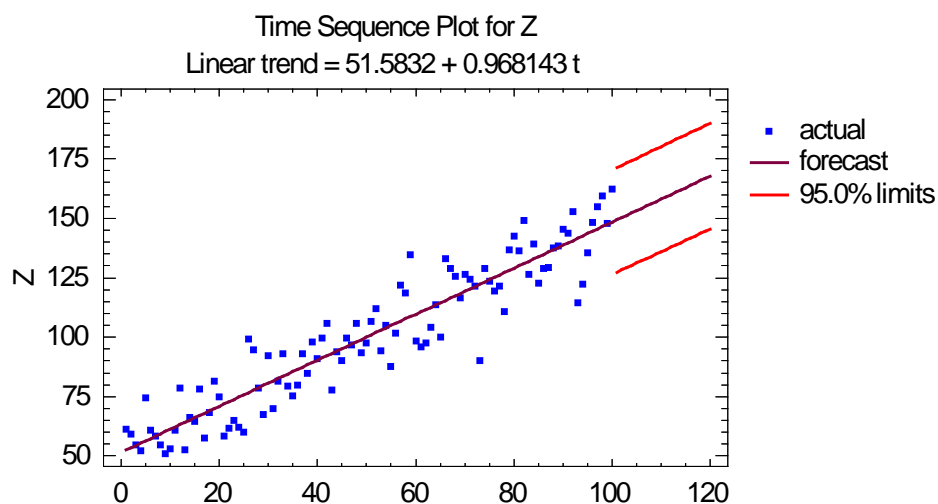
The appropriate forecasting model in this case would be a *simple regression* model. If the X axis is the time index, it is called a *trend line* model. If you transform the variable by "de-trending," i.e., subtracting out the trend line, it becomes a time series that looks like it was generated by the mean model.[2]

---

[2] In terms of its statistical properties, a series that has been de-trended by subtracting out an estimated trend line is not *exactly* like a series that is a sampled from a random process with zero trend. The de-trended series has an estimated trend of *exactly zero* within the data sample (because by definition it has been adjusted to zero out the

Another more interesting pattern is that of a time series that undergoes i.i.d. *changes* from one period to the next, which might look like this:

Time Sequence Plot for Y
Random walk with drift = 0.434184
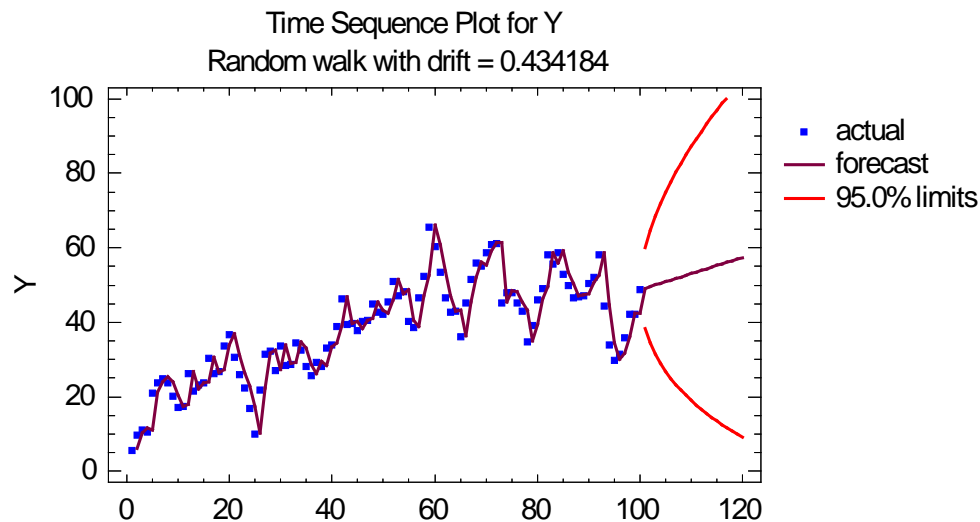
The appropriate forecasting model for this series is the *random walk* model, so-called because the variable takes random steps up and down as it goes forward. This model is of profound importance in financial data analysis. If you transform the variable by computing its period-to-period *changes* (the "first difference" of the time series), it becomes a time series that is described by the mean model. What is particularly interesting about the random walk model is the precise way in which the confidence limits for the forecasts get wider at longer forecast horizons, which is central to the theory of option pricing in finance.

Notice that there appear to be some interesting patterns in this graph, e.g., a series of regular peaks and valleys from period 60 onward. This is just a "statistical illusion": the time series was artificially generated using i.i.d. random steps. This is typical of random walk patterns—they don't look as random as they are! You need to analyze the statistical properties of the steps (in particular, the correlation between the size of the step in one period and the sizes of steps in preceding periods) to determine if they are truly random or if they have non-random properties such as "momentum" or "mean-reversion" or "seasonality."

---

estimated trend), whereas a series that is sampled from a random process with no trend will generally have a *non-zero* but not statistically significant estimated trend due to sampling error.

Another commonly seen pattern in time series data is that of i.i.d. variations around a *seasonal* pattern, which might look like this:
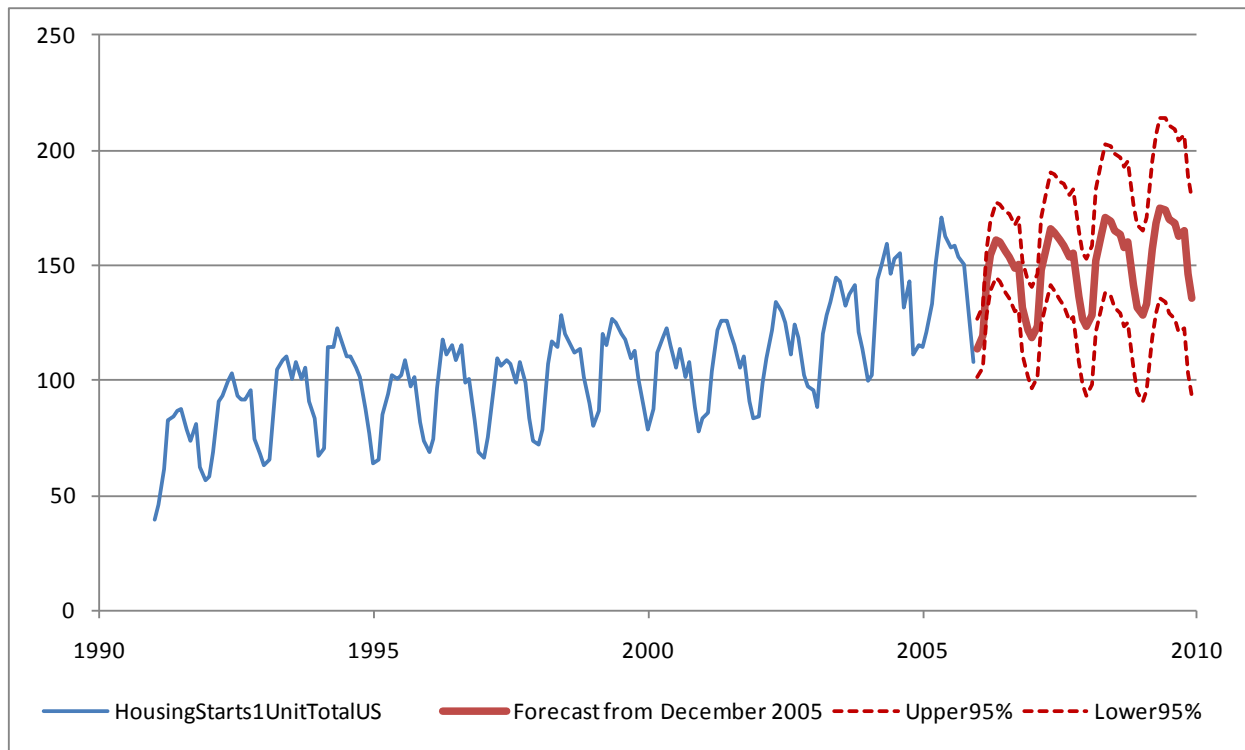

This sort of pattern is seen in retail sales, air travel, tourism, housing construction, power and water consumption, and many other measures of economic activity and environmental conditions. The noise is often not immediately apparent on such a graph, because the random variations are usually small in comparison to the seasonal variations, but the random variations are very important in seasonally-adjusted terms. This particular time series consists of **monthly US total housing starts for single-unit structures from January 1991 to December 2005, in thousands**. If you look closely, you will see that the seasonal variations are not identical and there are also changes in short-term trends due to business-cycle effects, but over the long run the trend and seasonal pattern appear to be reasonably consistent.

For the purposes of this chart, four years of monthly forecasts and 95% confidence limits were generated for the years 2006-2009 based on data up to December 2005 using one of the industry-standard seasonal forecasting models that we will meet later in the course: a so-called "ARIMA" model. The forecasts and confidence limits look quite reasonable based on the historical pattern in the data—that is, they seem to pass the "eyeball test". Alas, history did not turn out this way!

Another important type of pattern is that of correlation*s* among *many* different variables. You are probably familiar with X-Y scatterplots in which one variable is plotted against another. When working with many variables, we often want to look at an array of scatterplots between all pairs of variables, which is a so-called *scatterplot matrix*. A scatterplot matrix is the visual companion of a *correlation matrix* that shows all pairwise coefficients of correlation between variables. Most statistical analysis software (including RegressIt, our Excel add-in), will draw

scatterplot matrices for you.)   For example, consider a data set that consists of facts about houses sold in a community in a given year, and suppose the objective is to study how the selling price is correlated with features of the house such as its age, its square footage, its tax value, and the number of features it includes out of some set of desirable features.   The scatterplot matrix for this data set looks like this:
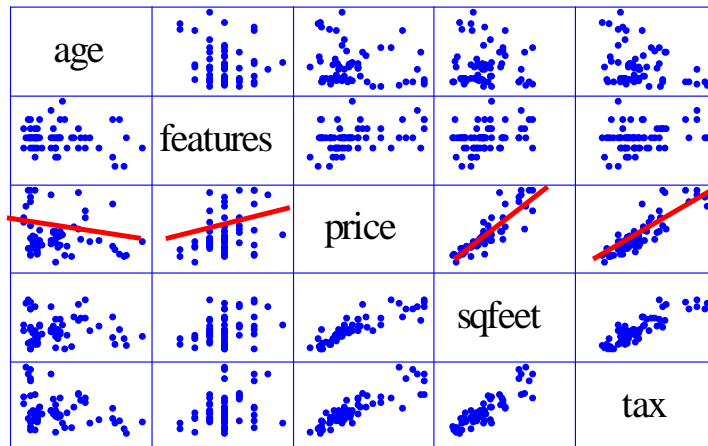
By examining the individual scatterplots you can see at a glance whether there are strong linear or nonlinear relationships among any pairs of variables and you can also quickly identify qualitative properties of their individual distributions, such as whether they are randomly scattered or have only a few possible values or have any extreme values ("outliers"). Here the selling price is seen to be strongly linearly related to square footage and tax valuation—which is not surprising—and less strongly related to age and features of the house. These relationships are highlighted by the red lines drawn on the plots in the row for which price is the Y-axis variable.[3] You can also see that the "features" variable has only 8 possible values, which are evidently integers from 0 to 7 since this is a counting variable Complex patterns like this are usually fitted by *multiple regression* models. (The red lines here are not regression lines, just hand-drawn lines that have been added for visual emphasis. The software we will use in this course can add regression lines to scatterplot matrices if desired.)

The examples shown above illustrate some of the most basic patterns we will look for in our data—means and standard deviations that may or may not be stable, trends that may or may not be linear, walking that may or may not be random, seasonality, cyclical patterns, correlations among many variables—and the corresponding forecasting models that might be used to explain and predict them. These basic models will serve as building-blocks for more sophisticated models that we will study later on.

---

[3] In a scatterplot matrix, a given variable is the *Y-axis* variable in all of the plots in its corresponding *row*, and it is the X-axis variable in all of the plots in its corresponding column.

Risks of forecasting

Forecasting is a risky business: "if you live by the crystal ball you end up eating broken glass." There are three distinct sources of forecasting risk and corresponding ways to measure and try to reduce them. I'll state them here in general terms, and we will discuss them in more detail later in the context of the mean model and other general models such as regression and ARIMA.

(i)   **Intrinsic risk** is random variation that is beyond explanation with the data and tools you have available. It's the "noise" in the system. The intrinsic risk is usually measured by the "standard error of the model," which is the estimated standard deviation of the noise in the variable you are trying to predict.[4] Although there is always some intrinsic risk (the future is always to some extent unpredictable), your estimate of its magnitude can sometimes be reduced by refining a model so that it finds additional patterns in what was previously thought to be noise, e.g., by using more or better explanatory variables. This doesn't necessarily mean the original model was "wrong," but merely that it was too simplistic or under-informed.

(ii)   **Parameter risk** is the risk due to errors in estimating the parameters of the forecasting model you are using, under the assumption that you are fitting the correct model to the data in the first place. This is usually a much smaller source of forecast error than intrinsic risk, if the model is really correct. Parameter risk is measured in terms of the "standard errors of the coefficient estimates" in a forecasting model—for example, the standard error of the estimated slope of a trend line. Parameter risk can be reduced in principle by obtaining a larger sample of data. However, when you are predicting *time series*, more sample data is not always better. Using a larger sample might mean including older data that is not as representative of current conditions. No pattern really stays the same forever, which is known as the "blur of history" problem.

The standard error of the *forecast* that a model ultimately produces is computed from a formula that involves the standard error of the model and the standard errors of the coefficients, thus taking into account *both* the intrinsic risk and the parameter risk. *The standard error of the forecast is always greater than the standard error of the model*, i.e., the standard error of the model is a lower bound on forecast accuracy. How much greater it is depends on the standard errors of the various coefficients and also on how far the values of the independent variables are from their own mean values. The errors in coefficient estimates become relatively more important when making forecasts under extreme conditions or for very distant points in time.

To the extent that you can't reduce or eliminate intrinsic risk and parameter risk, *you can and should try to quantify* them in realistic terms, so as to be honest in your reporting and so that appropriate risk-return tradeoffs can be made when decisions are based on the forecast.

---

[4] In general the term "standard error" refers to the estimated *standard deviation of the error* that is being made in estimating a coefficient or predicting a value of some variable.

(iii) **Model risk** is the risk of choosing the wrong model, i.e., making the wrong assumptions about whether or how the future will resemble the past. *This is usually the most serious form of forecast error, and there is no "standard error" for measuring it, because every model assumes itself to be correct.* Model risk can be reduced by following good statistical practices, which I will emphasize throughout this course. In fact, you might say that this course is mostly about how to choose the right forecasting model rather than the wrong one. If you follow good practices for exploring the data, understanding the assumptions that are behind the models, and testing the assumptions as a routine part of your analysis, you are much less likely to make serious errors in choosing a model. *The risk of choosing the wrong model is very high if you try to rely on simplistic rules ("biggest R-squared") or on automatic forecasting software, without understanding your own data, systematically exploring it, using your own judgment and experience, and carefully testing the model assumptions.* There is no magic formula—that's why you should study these notes.

How do you know when your model is good, or at least not obviously bad? One very basic test of your model is whether its errors really look like pure noise, i.e., independent and identically distributed random variables. If the errors are *not* pure noise, then by definition there is some pattern in them, and you could make them smaller by adjusting the model to explain that pattern. We will study a number of statistical tests that are used to determine whether the errors of a forecasting model are truly random. However, random-looking errors that are obtained in fitting the past data do not necessarily translate into realistic forecasts and confidence limits for what will happen in the future if the model's assumptions about the future are wrong.

As an example of model risk, consider what actually happened to US housing starts after 2005. Here is a chart showing the actual history together with the long-term point forecasts from the earlier chart. The forecasts tracked the first 5 months of 2006 pretty well, and then… boom!
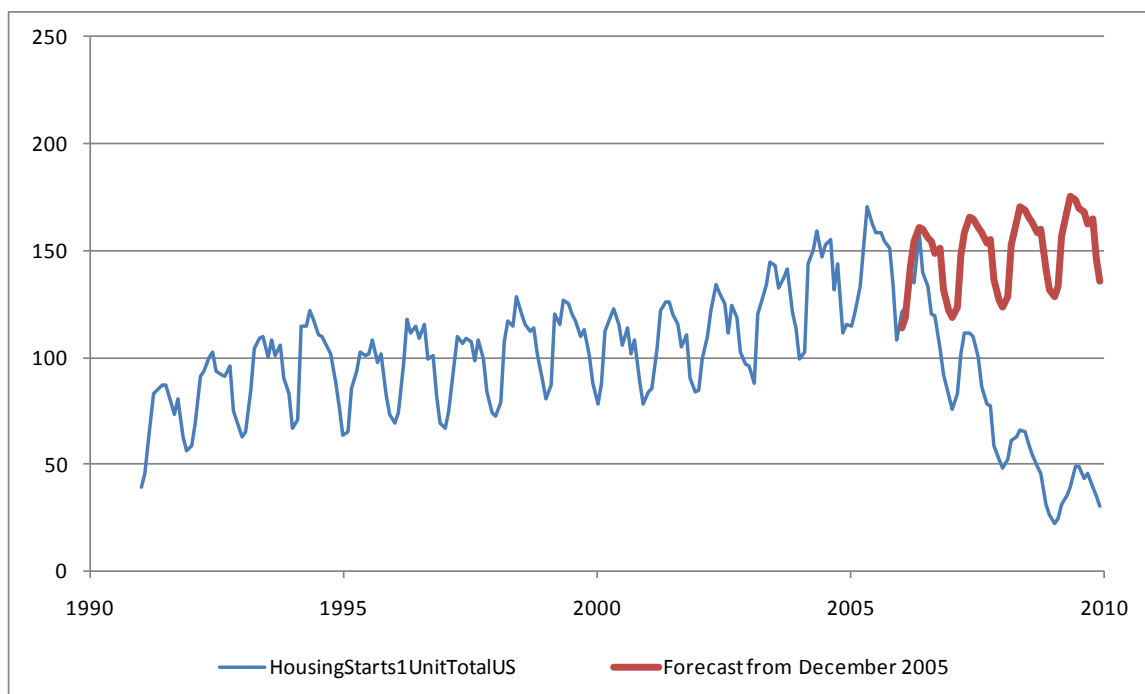
Actually, if you look at a longer historical sample—say, going back to 1959—you can see that the fairly consistent upward trend that was observed over the 15-year period from 1991 to 2005 was not likely to continue indefinitely, so a statistical extrapolation of that pattern should not necessarily have been expected to provide a realistic long-term forecast of what would happen in the coming years. And if you look closely at the years 2003-2004-2005, you can see the sort of sudden upward acceleration that was followed by cyclical downturns at several times in the more distant past, i.e., the telltale growth sign of a "bubble". A more careful analysis of the "fundamentals" of the housing market would also have raised red flags. The earlier downturns weren't quite as bad as what started in mid- 2006, though!

US Structures With 1 Unit: New Privately Owned Housing Units Started; T