

Lecture 18 Time Series (Part I) STAT102

Time Series Data:

Data gathered on a single individual (person, firm and so on) over a sequence of time periods, which may be hours, days, weeks, months, quarters, years, etc.

A time series consists of observations $Y_t, t = 1, \dots, T$.

Usual Primary Goal of Time Series Analysis:

Produce forecasts of future values of the time series.

Usual Secondary Goal of Analysis:

Understand the (random) structure of the process that created the series.

Discover the effect on the series of any independent covariates.

Such structure is often also needed to produce best forecasts.

Examples:

1. Many companies attempt to predict the future demand for their product and their share of the market.
2. Governments want to forecast the future values of interest rates, unemployment rates and percentage increases in the cost of living.
3. Investors and investment firms want to predict the future values of financial assets (stocks, bonds, etc.)
4. Economists want to forecast mortgage interest rates, demand for housing, the cost of building materials, the usage of gasoline and electricity, etc.
5. Meteorologists want to predict up-coming temperature and rainfall patterns.
6. Prediction of future population size is an important (but complicated) time series issue

Orientation

- Lectures 18 - 20 contain an introduction to time series modeling
- Theory will be presented along-side some basic examples
- There are several statistical methods for analyzing Time-series data
 - We use methods built from **regression** modeling –

Other methods are studied in more advanced courses on Time Series

- Regression methods often produce similar results to these other methods, but need less background to understand and employ correctly
- But, the regression methods we study may be less precise and less flexible for complex modeling situations
- *Dielman*, Chapter 7, discusses a third suite of ad-hoc methods that are more fragmented in approach. We will mostly not discuss these methods.
- Hence you will need to mostly rely on our notes for this topic, though we will use some data sets and problems from the text.

Plotting Time Series

- Time Series are usually plotted as a function of “time” (t).
- Dot plots with connecting line segments are customary.
- These are produced in JMP in the “Time Series” platform.

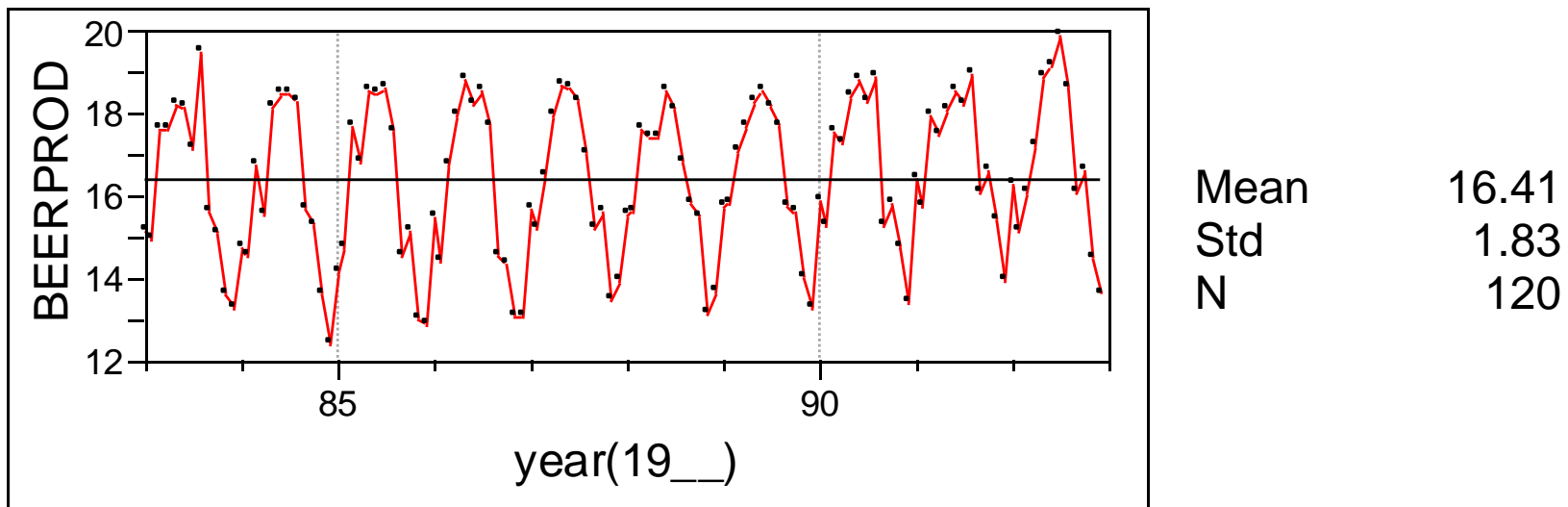
To get there click **Analyze** → **Modeling** → **Time Series**

Two Examples

1. Beer Production
Plot contains seasonal effects,
and perhaps a linear trend
2. Housing Starts
A “pure” time series

Example 1: Beer Production

Data: Monthly U.S. beer production, in **millions** of barrels, January 1983 through Dec 1992: (see *Dielman*, “Beer11”)

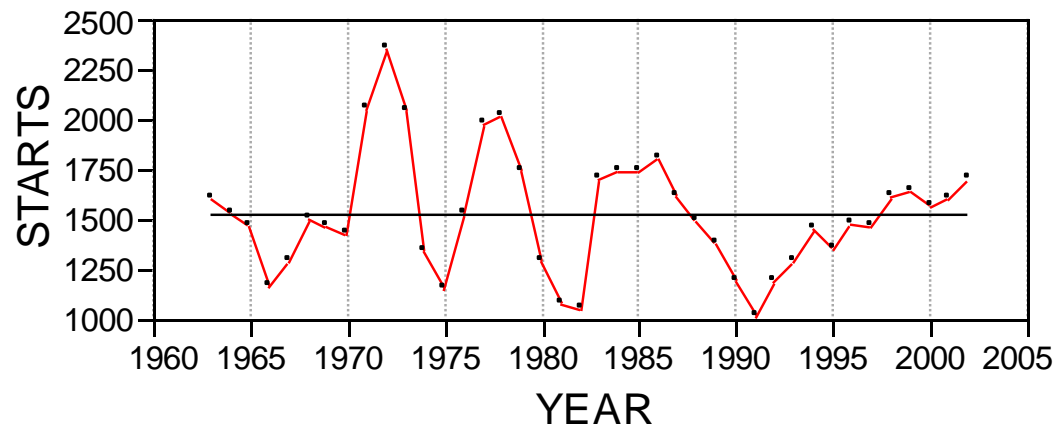


Time Series Plot

- Note the cyclical pattern (*aka* a “seasonal” pattern)
- There also may be an increasing trend over time (*Can you see it?*)

Example 2: Housing Starts

- Data: *Dielman*'s data file "HSTARTS3" contains data on the **annual** number of housing starts (in 1000s) from 1963 to 2002.
- *Dielman* cites US Dep't of Commerce. Similar data from US Census Bureau gives figures that are about 20% smaller. We'll use *Dielman*'s data.
- Here is a time-series plot of the data:



- This appears to be a “**pure**” time series (no seasonal or [linear] time trend)

“Pure” Time Series

- Denote the time-series as $Y_t, t = 1, \dots, T$.
- Write it as

$$Y_t = \mu + E_t$$

where μ denotes the mean level.

- Then E_t is a “centered” [= mean 0] time series.
- Suppose we want to use Y_1, \dots, Y_T to forecast Y_{T+1} .
- Assume no linear (or other) drift and no seasonal effect.
- **IF** the random variation is also independent over time, then there is no information about E_{T+1} in E_1, \dots, E_T ,
- **Hence**, it is **then** right to use the forecast $\hat{E}_{T+1} = 0 = \mathbf{E}(E_{T+1} | E_T)$, and hence to estimate Y_{T+1} as $\hat{Y}_{T+1} = \hat{\mu} \triangleq \bar{Y}$.

Autocorrelation

- If the random variation is **dependent** over time, then we can make use of Y_1, \dots, Y_T to forecast Y_{T+1} .
- The easiest useful situation of this sort is when each Y_{t+1} depends linearly on Y_t , with additional independent random variation. [*ie*, Each $E(E_{t+1} | E_t)$ depends linearly on E_t]
- The correlation between Y_{t+1} and Y_t is called first order **autocorrelation**.
- Autocorrelation is often denoted by the letter ρ . This symbol is also used for correlation in other settings as well.
- The corresponding statistical model is called an AR(1) model.
- We can use ordinary regression on Y_T to forecast Y_{T+1} .
- Here's how it's done in the housing starts example:

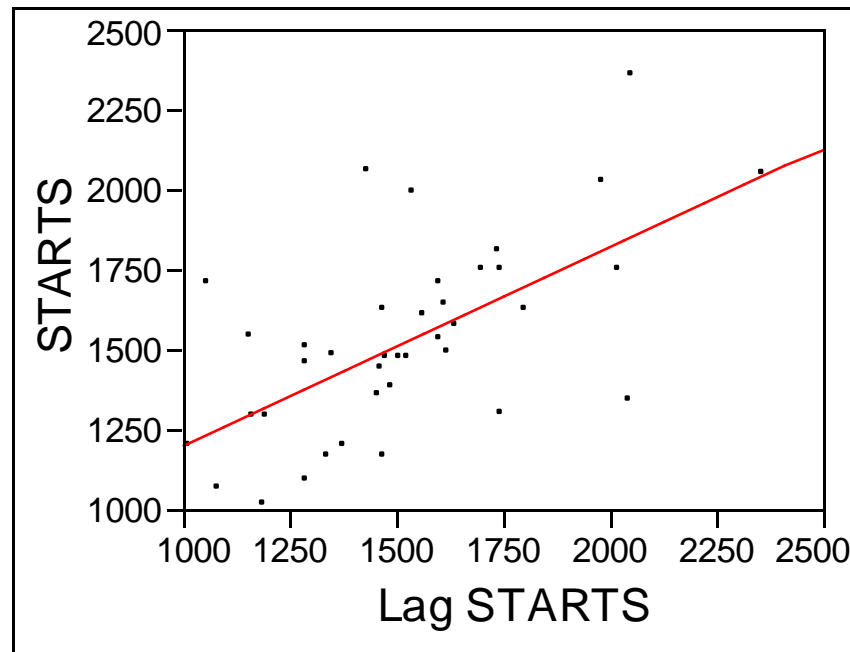
Using Regression in JMP to Analyze an AR(1) Time Series

- Think of the original time series variable as $Y = \text{STARTS}$.
- Think of the model as $Y_{t+1} = \beta_0 + \beta_1 Y_t + e_t : e_t \text{ indep., etc.}$
- Create a new column of variables with the formula
Lag(“STARTS”, 1).
- The values of this variable will be Y_{t-1} .
 - Thus the first entry will be empty (since $t - 1 = 0$).
 - The second entry will be Y_1 . *Etc.*
- Use this as X and run an ordinary regression of Y on X.
- The result will give an estimation equation of the form
 $Y_{t+1} = b_0 + b_1 Y_t$ that can be used to forecast Y_{T+1}
- It will also have an estimate $R = \widehat{\text{Corr}}(Y_{t-1}, Y_t)$.
- This is an estimate of the model parameter ρ .

Time Series Analysis of HSTARTS

Here is the output of the analysis described on the previous page:

The created variable Y_{t-1} has been called “Lag STARTS”



Linear Fit: $STARTS = 581.35 + 0.621 \text{ Lag STARTS}$

So, $b_0 = 581.35$ and $b_1 = 0.621$ in the pred'n. eqt. $Y_{T+1} = b_0 + b_1 Y_T$. cont

Summary of Fit

RSquare	0.3825
Mean of Response	1528.7
Observations	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1320324	1320324	22.92
Error	37	2131559	57610	Prob > F
C. Total	38	3451884		<.0001

Notes:

1. The RSquare value yields $\hat{\rho} = \sqrt{\text{RSquare}} = \sqrt{.3825} = 0.618$.
2. The sample size here is only $n = 39 = T - 1$ since one observation has been lost in creating LagSTARTS
3. The F-Ratio is testing $H_0 : \rho = 0$ versus the 2-sided alternative $H_1 : \rho \neq 0$.
4. We reject $H_0 : \rho = 0$ and conclude that the Y_t are **dependent**.

Forecasting Based on the Lagged Regression Analysis

- The estimation equation from the regression is

$$\text{EstSTARTS} = 581.35 + 0.621 \text{ Lag STARTS}$$

- In symbols this is

$$\hat{Y}_t = 581.35 + 0.621 \times Y_{t-1}.$$

- This can be used to forecast Y_{T+1} .

- *For example*, in 2002: STARTS = 1704.9 = Y_T

- So the forecast for $T+1 = 2003$ is

$$\begin{aligned}\hat{Y}_{T+1} &= 581.35 + 0.621 \times Y_T \\ &= 581.35 + 0.621 \times \mathbf{1704.9} = \mathbf{1639.7}\end{aligned}$$

Forecasting Further into the Future

- The preceding supplies a forecast of Y_{T+1} .
- It is also possible to get forecasts of Y_{T+2} , Y_{T+3} , etc.
- Some care is needed to get it right. *See the optional notes.*
- The long term average of Y is estimated as

$$\hat{\mu} = \frac{b_0}{1 - b_1}.$$

- Then Y_{T+r} , $r = 1, \dots$ are forecasted as

$$\hat{Y}_{T+r} = \hat{\mu} + b_1^r (Y_T - \hat{\mu}).$$

- Thus $\hat{Y}_{T+2} = 1533.0 + b_1^2 (Y_T - 1533.0) = \mathbf{1599.2}$.

The influence of Y_T fades quickly, since $|b_1| < 1$ and the projection r steps into the future will be $Y_{T+r} = \hat{\mu} + b_1^r (Y_T - \hat{\mu}) \rightarrow \hat{\mu}$ as $r \rightarrow \infty$.

Optional

Explanation of Formula for r -step Prediction

Rewrite the original form of the model – $Y_{t+1} = \beta_0 + \beta_1 Y_t + E_{t+1}$ – as

$$Y_{t+1} = \mu + \beta_1 (Y_t - \mu) + E_{t+1}.$$

This entails $\mu - \beta_1 \mu = \beta_0$. Solving gives $\mu = \beta_0 / (1 - \beta_1)$. Hence estimate μ as

$$\hat{\mu} = \frac{b_0}{1 - b_1}.$$

Now, $E(E_{t+1}) = 0$. Hence $E(Y_{t+1} - \mu | Y_t) = \beta_1 (Y_t - \mu)$, and similarly for Y_{t+2} . So,

$$\begin{aligned} E(Y_{t+2} - \mu | Y_t) &= E(E(Y_{t+2} - \mu | Y_{t+1}) | Y_t) \\ &= E(\beta_1 (Y_{t+1} - \mu) | Y_t) = \beta_1 \times \beta_1 (Y_t - \mu) = \beta_1^2 (Y_t - \mu). \end{aligned}$$

This yields the desired prediction equation,

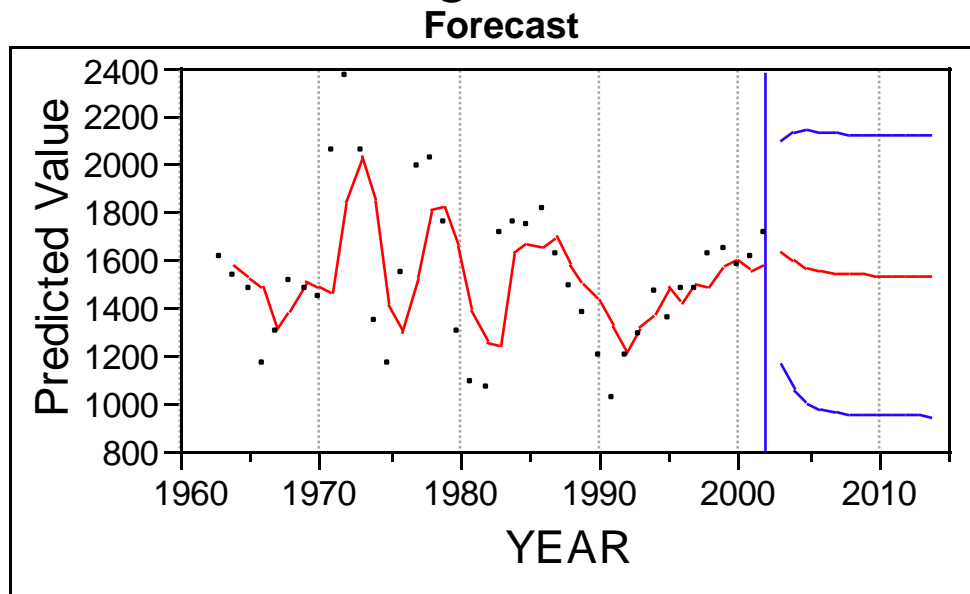
$$\hat{Y}_{T+2} = \hat{\mu} + b_1^2 (Y_T - \hat{\mu}).$$

For larger prediction gaps, r , reason similarly to get the desired equation.

Another Way to Get a Very Similar Answer

Use the Time-series AR(1) option in JMP.

- On the red-arrow drop down menu choose ARIMA
- In the box that appears enter “Autoregressive Order” = **1** and click the “Estimate” button. You’ll get



Note that this plot shows the forecasts, and even gives confidence bands for them.

Optional

Explanation of the similarity of the two methods -
and of their slight difference

We have referred to our time-series model as a “pure” model.

A more technical model is “stationary”.

Stationary time-series models do not change as t increases. So $\text{Var}(Y_{t+1}) = \text{Var}(Y_t)$.

Let $\rho = \text{Corr}(Y_{t+1}, Y_t)$. Standard theory for the linear model $Y_{t+1} = \beta_0 + \beta_1 Y_t + E_{t+1}$ yields

$$\beta_1 = \rho \frac{\sqrt{\text{Var}(Y_{t+1})}}{\sqrt{\text{Var}(Y_t)}} = \rho.$$

Hence we should expect to find that the estimates of β_1 and ρ satisfy

$$b_1 \approx R.$$

The Time Series analysis in JMP uses a method that takes into account this stationarity, and is not exactly Least Squares. The estimate of autocorrelation from JMP, $\hat{\rho}$, (in the “Partial” AR display) should also be close to b_1 and R .

In our data: $b_1 = 0.621$, $\hat{\rho} = 0.615$, $R = 0.619$.