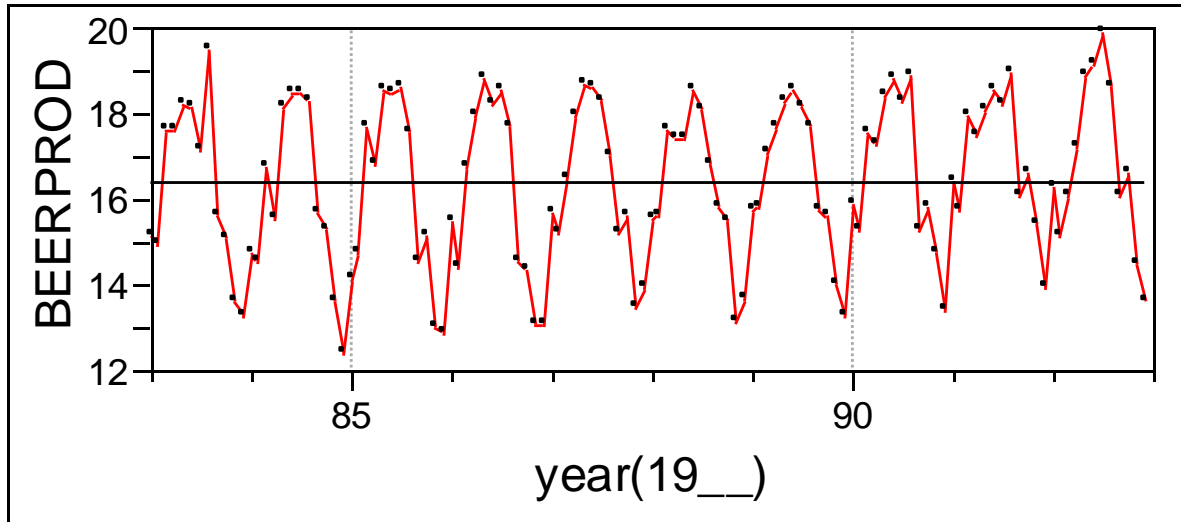


Lecture 19 Time Series, Part II

Trends and Seasonality

- Many Time-Series contain trends and/or seasonal effects.
- These need to be incorporated into the analysis.

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



Mean	16.41
Std	1.83
N	120

Decomposition of Time Series:

Additive Decomposition: $Y_t = T_t + S_t + E_t$

1. Long term trend T_t : Long-term, relatively smooth pattern or direction exhibited by a series
2. Seasonal variation S_t : Refers to *cycles* that occur over short repetitive calendar periods. The term “seasonal variation” may refer to the four traditional seasons or to systematic patterns that occur during a month, a week, a day, or hours.
3. For example, demand for restaurant meals features
“seasonal variation” = variation through the day.
4. Random variation E_t .

Example:

Beer Production series clearly contains seasonal variation.

It appears to also exhibit a modest upward trend over time. (We'll investigate that soon.) And,

It contains random variation.

Note: A multiplicative decomposition model is sometimes a better fit to reality:

Multiplicative Decomposition:

$$Y_t = T_t S_t E_t .$$

A Multiplicative Model is additive on the log scale:

$$\log Y_t = \log T_t + \log S_t + \log E_t$$

Polynomial Models for the Trend Component T_t :

A natural approach to modeling the trend component is to use polynomials for the time trend, *e.g.*,

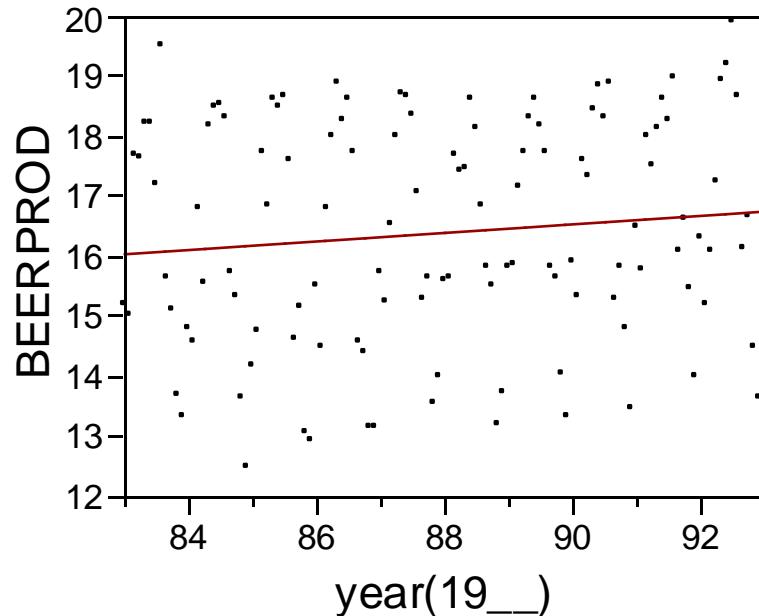
Linear trend: $T_t = \beta_0 + \beta_1 t$

Quadratic trend: $T_t = \beta_0 + \beta_1 t + \beta_2 t^2$

Cubic trend: $T_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$

etc

Beer Example: Linear trend



Linear Fit

$$\text{BEERPROD} = 10.37 + 0.06867 \text{ year}(19__)$$

Summary of Fit

RSquare	0.012
RMSE	1.839
Observations	120

Analysis of Variance

Source	DF	Sum Sq	Mean Sq	F Ratio
Model	1	4.71	4.71	1.39
Error	118	399.27	3.38	Prob > F
C. Total	119	403.99		0.2402

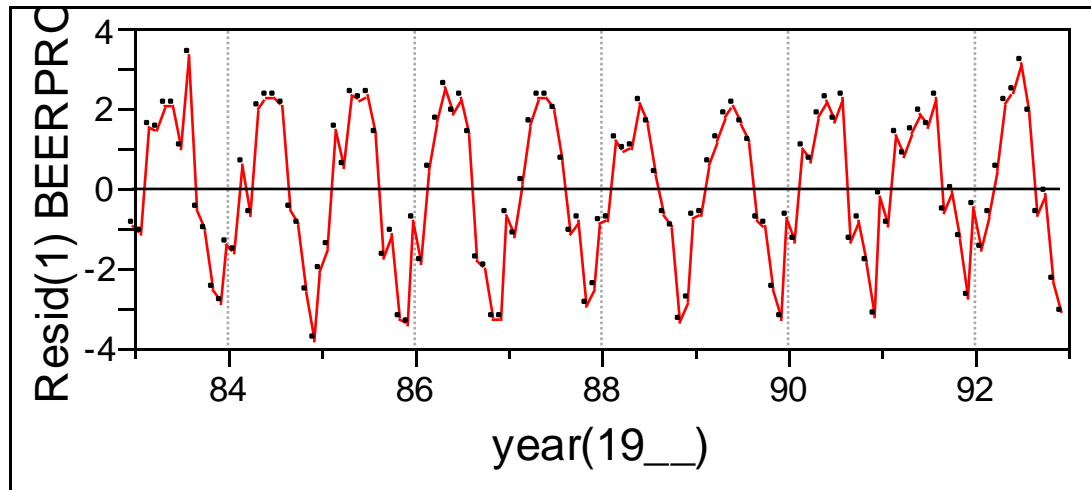
The t-test that the linear trend is not 0 is **not significant**. However, the test is **not valid** because the errors are not independent.

[They contain seasonal effects and perhaps also time-series auto-regressive effects]

Seasonal Pattern of Residuals

The seasonal pattern of the data was easy to see in the first time-series plot.

It shows up equally well in the time-series plot of the residuals from the linear trend fit:



Mean	3.72e-15
Std Er	1.824
N	120

[This pattern does not show up quite so clearly in the ordinary residual plot from the linear regression because the residual points are not connected. Try it to see! Or try to find the pattern in the regression plot on p.5!]

Modeling Seasonal Variation through Categorical Variables

In order to capture the seasonal variation, as well as the (possible) linear trend:

- Create a **categorical** [*or ‘nominal’*] variable that labels the “season” (or its relevant segments) and include this season variable in the Model.
- For the beer production data, the seasons are indexed according to the different months.
- This gives a mixed type of model containing a regression variable (linear trend) and a categorical variable (Month).
- Here is the analysis ---

Tabular Output from Seasonal + Linear Model

Summary of Fit

RSquare	0.916
RMSE	0.562

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	370.08	30.84	97.31
Error	107	33.91	0.317	Prob > F
C. Total	119	403.99		<.0001

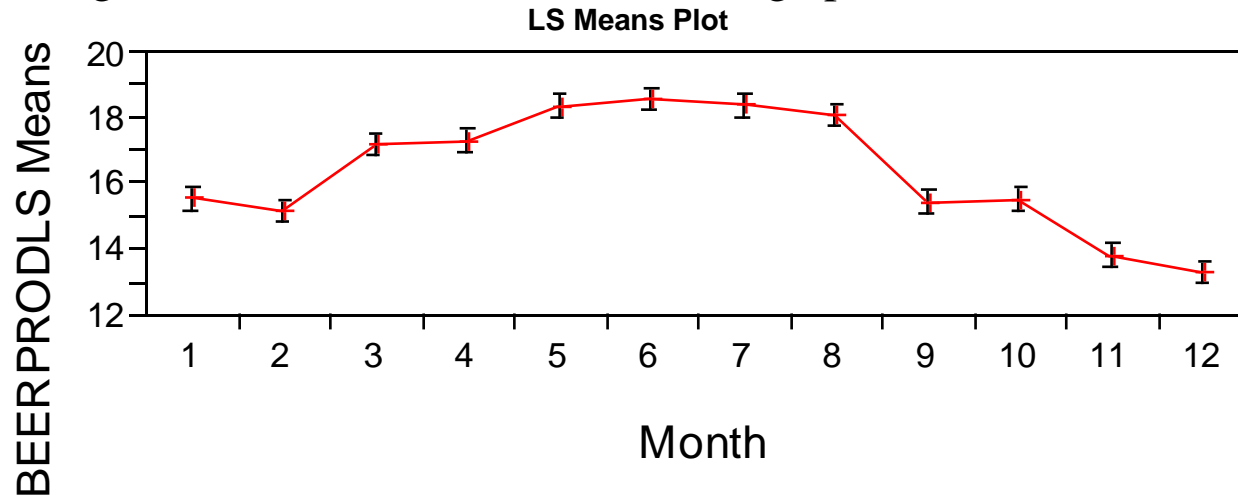
Effect Tests

Source	DF	Sum of Squares	F Ratio	Prob > F
year(19__)	1	8.59	27.10	<.0001
Month	11	365.36	104.80	<.0001

- The **Effect test** shows that *[now]* the **means for the different months** are not all the same (**P-value < .0001**). *[after controlling for slope]*
- It also shows that the **slope coefficient** is significant after controlling for the seasonal variation (months). (Again, **P-value < .0001**.)

Graphical Output from Seasonal + Linear Model

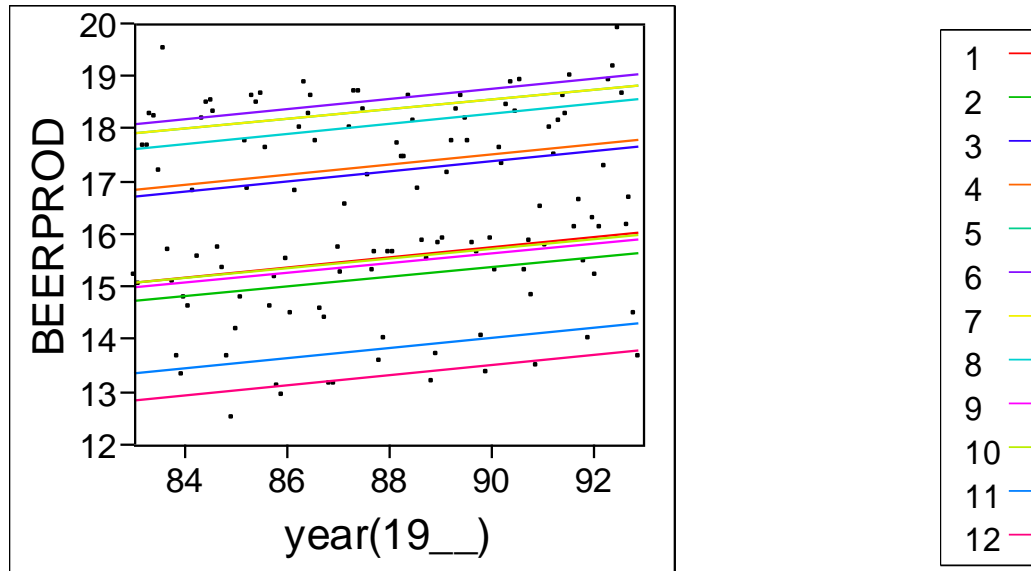
- The following plot shows the seasonal effects. [To get the plot click the red triangle next to month, above the leverage plot, and click LS Means plot]:



- Beer production is the highest over the summer months and the lowest over the winter months. December is the lowest month.

Graphical Output, Cont.

- We can also view the regression lines for each month:



- Note that these lines are parallel. (WHY? –What happens if you cross month and year?)
- Also, now that we have included the month variable, the estimated linear trend is significant. (*ie*, beer production increased an average of $0.093 * 1000000 = 84,000$ barrels each year, for each fixed month).

Further Fit-Model Analyses

To see if a quadratic time trend is needed, we fit a multiple regression model of beer production on time that also includes the possible factor **(Year – its mean)²** and Month.

Summary of Fit					
	RSquare			0.920	
	Root Mean Square Error			0.551	
Effect Tests					
Source	DF	Sum of Squares	F Ratio	Prob > F	
year(19__)	1	8.59	28.28	<.0001	
(Year-its mean)^2	1	1.71	5.64	0.0193	
Month	11	365.75	109.47	<.0001	

The quadratic trend is statistically significant, but not by a lot. (The estimated coefficient is 0.016 from the parameter estimates table. This has a positive sign as one might expect. (WHY?))

R^2 increases by only a very small amount from the linear to the quadratic model.

Correspondingly the RMSE decreases by only a very small amount – from RMSE=0.562 to RMSE=0.551 – about a 2% decrease.

Thus the quadratic model is only **slightly** more accurate; hence we will omit the quadratic term for simplicity, without sacrificing much accuracy.

Forecasting with the Fit Model output

The time period covered in the sample is January 1983 through Dec 1992. We can forecast beer production for the future using the coefficients from our Fit Model analysis.

Selected coefficients from the Linear + seasonal model are

Parameter Estimates	
Term	Estimate
Intercept	8.220
year(19__)	0.09315
Month[3]	0.796
Month[6]	2.1644

Hence, for a future time period t in month k ,

$$\hat{Y}_t = \hat{E}(Y_t | Year(19 __) = t, Month = k) = \hat{\beta}_{\text{intercept}} + \hat{\beta}_{\text{Year}} \times t + \hat{\beta}_{\text{Month}[k]}$$

Predicted beer production for June 1993:

$$8.220 + .09315 \times (93 + 5/12) + 2.1644 = 19.09$$

Predicted beer production for March 1993:

$$8.220 + .09315 \times (93 + 2/12) + .7956 = 17.70$$

etc

Beer Production (cont.)

- For the Beer data the proposed model so far is

$$Y_t = T_t + S_t + E_t \text{ where}$$

T_t is a linear time trend (or a quadratic trend).

S_t is a seasonal effect depending on the month for t .

E_t is the random variation.

- We already fit T_t and S_t using the Fit Model option.
- The residuals are E_t . So far we've assumed these are independent. This **might be** true
- **But it also might be that they are themselves autocorrelated!!**
- If so, time-series ideas should be included in the model

Including an autoregressive effect into the E_t :

- Create a column with the lag-one values of Y_t . These are thus denoted as Y_{t-1} .
- Fit a Model with dependent variable Y_t and independent variables T_t , S_t , and Y_{t-1} .
- Symbolically, this is the model

$$Y_t = \beta_0 + \beta_1 T_t + \beta_{S_t} + \beta_3 Y_{t-1} + e_t$$

where the e_t denote independent, homoscedastic, normal errors.

Here is the output of this analysis:

Summary of Fit

RSquare	0.9194
Root Mean Square Error	0.5557
Mean of Response	16.4229
Observations	119

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Model	13	370.05	28.466	92.170	
Error	105	32.43	0.309		
C. Total	118	402.48			<.0001

Effect Tests

Source	DF	Sum of Squares	F Ratio	Prob > F
Month	11	182.06	53.591	<.0001
year(19__)	1	4.22	13.670	0.0003
lag1(BeerProd)	1	1.48	4.781	0.0310

- The Summary of Fit table shows $R^2=0.919$. This is only slightly larger than the value of $R^2=0.916$ in the season+linear trend model.
- Hence the addition to that model of the AR(1) dependence is not very important, even if it is statistically significant.
- Actually, the AR(1) part of the model is statistically significant (after controlling for the season+linear trend);
- but only with Pvalue = 0.0310.

Forecasting with the AR(1) Model

- The equation to forecast one-step ahead is

$$\hat{Y}_t = \beta_0 + \beta_1 T_t + \beta_{S_t} + \beta_3 Y_{t-1}$$

- In order to do so we need the relevant coefficients in the model. They are in the “Expanded Estimates table.
- Here is some of that table: with the coefficients labeled:

Expanded Estimates Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	$\beta_0 = 6.485$	1.755	3.70	0.0003
Month[1]	$\beta_{S_1} = -0.207$	0.341	-0.61	0.5457
Month[2]	$\beta_{S_2} = -1.036$	0.187	-5.55	<.0001
Month[3]	$\beta_{S_3} = 1.050$	0.204	5.14	<.0001
year(19__)	$\beta_1 = 0.074$	0.185	4.07	0.0003
lag1(BeerProd)	$\beta_3 = 0.209$	0.095	2.19	0.0310

- The equation to forecast one-step ahead is

$$\hat{Y}_t = \beta_0 + \beta_1 T_t + \beta_{S_t} + \beta_3 Y_{t-1}.$$

- Jan 1993 has values $T = 93.00$, Month = 1 , and $Y_{t-1} = 13.64$ (directly from the data table).

- Hence

$$\hat{Y}_{T=93} = 6.485 + 0.074 \times 93 + (-0.207) + 0.209 \times 13.64 = 16.01$$

- If we plug this (estimated) value into the equation for Feb '93 we get a forecast for Feb 93. ETC.
- This process yields the following forecasts for Jan '93, Feb '93 and March '93. [The process can be continued further into the future, if desired.]
- Here is a table showing the forecasts for Jan '93, Feb '93 and March '93:

Table of Forecasts for Jan – March ‘93

Previous Y or \hat{Y}	T	β_s	Formula	Forecast
13.64	93.00	-0.207	$6.485 + 0.074 \times 93$ $+ (-0.207) + 0.209 \times 13.64$	16.01
16.01	93.083	-1.036	$6.485 + 0.074 \times 93.083$ $+ (-1.036) + 0.209 \times 16.01$	15.68
15.68	93.167	1.050	$6.485 + 0.074 \times 93.167$ $+ (1.050) + 0.209 \times 15.68$	17.71

Optional Material

Evaluation of Forecasts: Out-of-Sample Validation

- Out of sample validation is a valuable way to evaluate how well competing forecasting models can be expected to work.
- This idea can be used on all sorts of prediction and forecasting models – not just on time-series.
- Here's how the method works on our time series.

Out-of-Sample Validation; General Description

1. Reserve part of the data to be the *out-of-sample* validation test set.
2. For a time series this must be the last part of the data series.
3. The remainder of the data is the *in-sample* part used to fit the statistical models and parameter values.
4. Fit the model(s) on the in-sample part.
5. For each model make predictions (=forecasts) about the values of the out of sample part.
6. Measure the squared error of these predictions
7. Compare how well the different models do in terms of their total squared prediction error.

For small models like ours this can be done manually. For larger models you need programmable software (R or S, SAS, ...), and programming ability.

Out-of-Sample Validation; Example

- We reserve the data from 1992 as our Validation Test Set.
- We fit three multiple regression models to the data from January, 1983 through December, 1991.
 1. Linear regression, no month variables
 2. Linear regression & month variables
 3. Quadratic regression & month variables
 - Later we'll look at some other models
- Then we compare the forecasted to the actual values for 1992:

Out of Sample Comparison of Three Models

Model	Sum of Squared Forecast Errors
Linear trend, no month variables	42.93
Linear trend, month variables	5.64
Quadratic trend, month variables	7.32

- **Notice** that the “Linear trend & month” model performs best.
- This happens even though the quadratic term is significant in the quadratic trend & month model! Why?
(Its P-value = .0054 in the Effect Test table for the test model)
- There may be some statistical bad luck here
- But this is not atypical of what happens when adding extra variables to incorrectly specified models, even when those variables are somewhat significant. (It doesn't usually happen with overwhelmingly significant extra variables.)

Further Explanation of Why Adding a Significant Variable to the Model Made the Forecasts Behave Worse

- This can't happen (except by very bad luck) if the Quadratic regression & month model is correct
- SO, If it wasn't just very bad statistical luck THEN
- It was because the Quadratic regression on the year & ANOVA on the month model wasn't correct as an extrapolation into the future.
- “Correct” here means that real life and the statistical model aren't the same.

(Of course then the Linear regression on the year & ANOVA on the month model isn't correct either, but somehow it isn't as badly incorrect.)