Homework Assignment 7

Group homework (up to four members per group)
due in class on Thursday, 04/03/2014

STA 371G, Statistics and Modeling, Spring 2014

Problem 1: Wine Sales (I) (15 points)

Read the case "Northern Napa Valley Winery, Inc." in the course packet. The data file is available in the course website. The file contains the monthly wine sales for the Northern Napa Valley Winery for the period January, 1988 through August, 1996.

The goal of this case is to provide monthly forecasts for wine sales in the next twelve months, i.e. September, 1996 through August, 1997, and to combine the monthly forecasts to provide a forecast of annual sales for these twelve months.

(a) Using the appropriate time series plots, decide whether sales or the percentage change in sales, denoted PctChange, should be analyzed. Construct a regression model on PctChange that accounts for the trend and seasonality in the data.

(b) Are the residuals from this model independent? To test if the residuals are independent, you may run the following regression:

$$Residual_t = \beta_0 + \beta_1 Residual_{t-1} + \epsilon_t.$$

If you reject the null hypothesis that $\beta_1 = 0$, then consider modifying the model. (Hint: construct a trend+seasonality+AR(1) model)

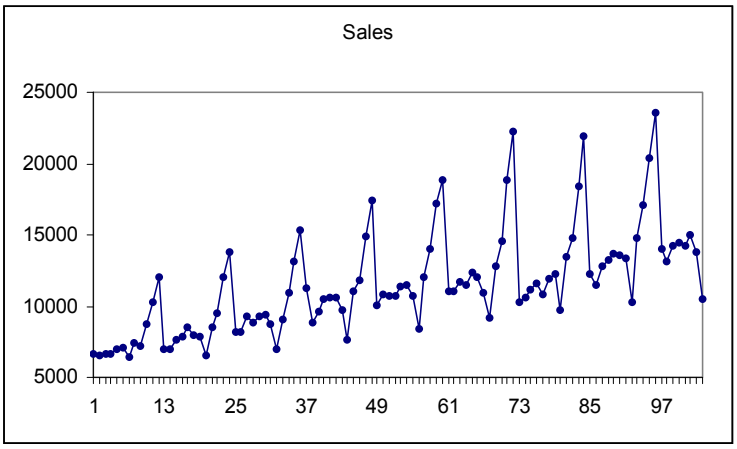
(c) Using the final model obtained above, provide numerical forecasts of wine sales in September, 1996 and October, 1996.

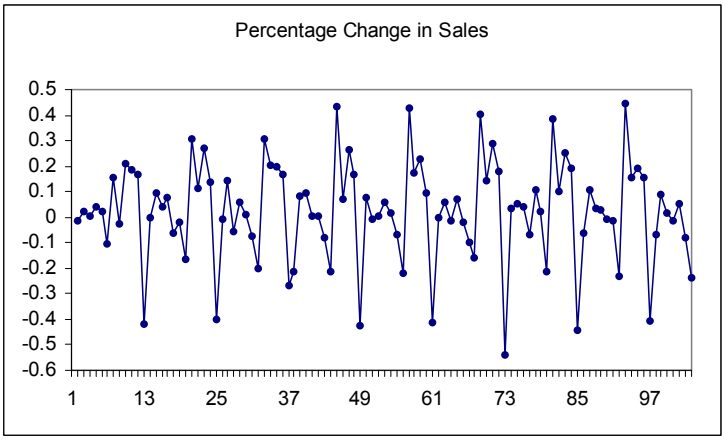
(d) Explain briefly how you would obtain forecasts of wine sales for the months November, 1996 through August, 1997. You do not have to do the actual calculations for these forecasts because they are a bit tedious to do by hand. You just need to explain briefly how you would do it.

(e) Given the forecasts for the months September, 1996 through August, 1997, explain briefly how you could get a forecast of annual sales for the year encompassing these twelve months. You do not have to do the actual calculations, just provide a brief explanation of what you would do if the twelve monthly forecasts were available.

(f) Provide a 68% prediction interval for the forecast of sales in September, 1996.

(g) Explain briefly why it is difficult to get confidence intervals for the monthly forecasts of wine sales in October, 1996 through August, 1997.


Sales
25000
20000
15000
10000
5000
1
13
25
37
49
61
73
85
97


Percentage Change in Sales
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
1
13
25
37
49
61
73
85
97

Problem 2: Wine Sales (II) (15 points)

In this problem, we will try log transformation to stabilize the variance and then build a regression model to forecast wine sales.

Run the regression

$$Log(Sales_t) = \beta_0 + \beta_1 t + \beta_2 Jan + \cdots + \beta_{12} Nov + \epsilon_t$$

where t is the trend variable and Jan, Feb, $\cdots$, Nov are dummies variables for each month of the year.

(a) What would be the predicted increase of Wine Sales (in percentage) from December 1996 to January 1997?

(b) What would be the predicted increase of Wine Sales (in percentage) from July 1997 to August 1997?

(c) Plot the residuals against time. Do you see any clear patterns in the plot? Are there any model assumptions violated?

Add the term t^2 and run the regression

$$Log(Sales_t) = \beta_0 + \beta_1 t + \beta_2 Jan + \cdots + \beta_{12} Nov + \beta_{13} t^2 + \epsilon_t$$

(d) Plot the residuals against time. Do you see any clear patterns in the plot?

(e) Using the final model obtained above, provide numerical forecasts of wine sales in September, 1996 and October, 1996.

(f) Provide a 95% prediction interval for the forecast of sales in September, 1996.

(g) (Optional) What would be the predicted increase of Wine Sales (in percentage) from March 1997 to April 1997?

(h) (Optional) According to the model, without considering the seasonal effects, do the sales increase at a faster rate as time increases?

(i) (Optional) Build an Excel spreadsheet model or write R code to forecast the sales of the next 12 months.

Problem 1: Wine Sales (I) (15 points)

The changes in Sales from month to month are increasing in dollar terms while the percentage changes in Sales (PctChange) from month to month are relatively consistent (for example, the percentage changes from December to January each year are always around 40%). Therefore, the series of percentage changes exhibits constant volatility and is "easier" to analyze.

A regression model that accounts for the trend and seasonality is

$$PctChange_t = \beta_0 + \beta_1 t + \beta_2 Jan + ...\beta_{12} Nov + \epsilon_t$$

where t is s trend variable and Jan, Feb,..., Nov are dummies for each month of the year. The results of this model are and residual plots shown below:

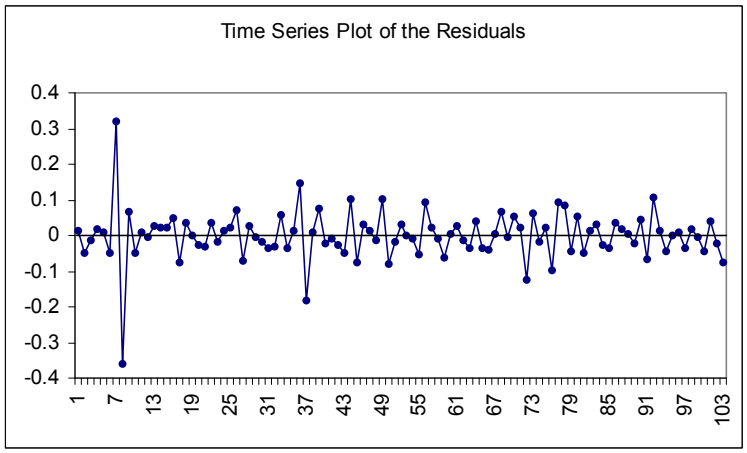
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.93321002
R Square	0.870880941
Adjusted R Square	0.853665067
Standard Error	0.074269735
Observations	103

ANOVA

	df	*SS*	*MS*
Regression	12	3.34837965	0.279031637
Residual	90	0.49643942	0.005515994
Total	102	3.84481907	

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.157470265	0.02944145	5.348590662
Trend	3.68785E-06	0.000246581	0.014955915
M1	-0.570578369	0.037135686	-15.36469168
M2	-0.186275436	0.03610208	-5.159687039
M3	-0.087203218	0.036096185	-2.415856902
M4	-0.140005831	0.036091974	-3.879140326
M5	-0.129681557	0.036089447	-3.593337342
M6	-0.143401209	0.036088604	-3.973586979
M7	-0.215269701	0.036089447	-5.964893344
M8	-0.323050544	0.036091974	-8.950758554
M9	0.177770498	0.037142235	4.786208989
M10	-0.011330671	0.037138142	-0.305095255
M11	0.076837582	0.037135686	2.069103611



To test if the residuals are independent we can run the following regression:

$$Res_t = \beta_0 + \beta_1 Res_{t-1} + \epsilon_t$$

The results are: so that we concluded that there are still some time dependence left in the

Regression Statistics	
Multiple R	0.527050734
R Square	0.277782476
Adjusted R Square	0.270560301
Standard Error	0.059866007
Observations	102

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.000248968	0.005927945	0.041999054
Res_Lag1	-0.529935429	0.08544852	-6.201809353

residuals as they are indeed correlated in time (see the estimate and t-stat for b_1).

We can try and solve for this problem by adding a auto-regressive term $PctChange_{t-1}$ to the model:

$$PctChange_t = \beta_0 + \beta_1 t + \beta_2 Jan + ...\beta_{12} Nov + \beta_{13} PctChange_{t-1} + \epsilon_t$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95227415
R Square	0.906826056
Adjusted R Square	0.893061724
Standard Error	0.063790003
Observations	102

ANOVA

	df	*SS*	*MS*
Regression	13	3.485117544	0.268085965
Residual	88	0.358086476	0.004069164
Total	101	3.84320402	

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.280095053	0.033009542	8.485275346
PctChangeLag1	-0.530913837	0.091125533	-5.826180847
Trend	3.84334E-05	0.000214927	0.17882105
M1	-0.611405293	0.032656111	-18.72253832
M2	-0.531815718	0.067074575	-7.92872292
M3	-0.226795197	0.039184906	-5.78782038
M4	-0.227031787	0.034411673	-6.597522537
M5	-0.244773938	0.036757496	-6.659157072
M6	-0.253045077	0.036262046	-6.978234945
M7	-0.33223031	0.036929226	-8.99640596
M8	-0.478199918	0.040864934	-11.70196234
M9	-0.029501219	0.047787731	-0.617338763
M10	0.042321586	0.033199757	1.274755879
M11	0.030060624	0.03289106	0.91394514

Now, we can once again check the independence of the residuals by running the regression:

$$Res_t = \beta_0 + \beta_1 Res_{t-1} + \epsilon_t$$

The results are:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.173991334
R Square	0.030272984
Adjusted R Square	0.020477762
Standard Error	0.059092814
Observations	101

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.000544777	0.005880589	0.092639864
Res_Lag1	-0.175505942	0.099832364	-1.758006466

Now we can't reject the hypothesis that the residuals are independent through time!

(c) September 1996 is month $t = 105$. The model is (b) allows us to predict the percentage growth from August 1996 to September 1996 via:

$$\widehat{PctChange}_{105} = 0.28 + (0.000038 \times 105) + (-0.029 \times 1) + (-0.5309 \times -0.240) = 0.381$$

This implies the following prediction for September:

$$Sales_{105} = Sales_{104} \times (1 + \widehat{PctChange}_{105}) = 10522 \times (1 + 0.381) = 14530$$

To predict the percentage growth for October we use:

$$\widehat{PctChange}_{106} = 0.28 + (0.000038 \times 106) + (0.042 \times 1) + (-0.5309 \times 0.381) = 0.121$$

where the red term represent the predicted growth for september. Therefore, the prediction for Sales in October is:

$$Sales_{106} = Sales_{105} \times (1 + \widehat{PctChange}_{106}) = 14530 \times (1 + 0.121) = 16289.$$

(e) The forecasts for September, 1996 through August, 1997 are added together to obtain a forecast of annual sales for the year encompassing these twelve months.

(f) The 68% Prediction Interval for September 1996 is

$$\widehat{PctChange}_{105} \pm 1 \times s = 0.381 \pm 0.064 = [0.317; 0.445].$$

Thus the 68% Prediction Interval for the wine sales in September 1996 is

$$[(1 + 0.317) * 10522; (1 + 0.445) * 10522] = [13858, 15204]$$

(g) There are two sources of error in this prediction. The first source of error is due to ϵ_{106} while the second source of error is due to using Predicted $PctChange_{105}$ instead of the actual value of $PctChange_{105}$. The uncertainty in the prediction due to the uncertainty in ϵ_{106} can be accounted for in the usual way. However, it is more difficult to account for the uncertainty in the prediction induced by using Predicted $PctChange_{105}$ instead of the actual value of $PctChange_{105}$. A similar argument explains why it is difficult to get predictive intervals for $PctChange_{107}$, ..., $PctChange_{116}$.

Problem 2: Wine Sales (I) (15 points)

$Log(Sales_t) = 9.3915 + 0.0072 * t - 0.5525 * M1 - 0.5920 * M2 - 0.5320 * M3 - 0.5224 * M4 - 0.5030 * M5 - 0.4970 * M6 - 0.5644 * M7 - 0.7605 * M8 - 0.4705 * M9 - 0.3419 * M10 - 0.1389 * M11 + \epsilon_t$

From time $t-1$ to t, the sales increases by

$$\frac{Sales_t}{Sales_{t-1}} - 1 = e^{\log(Sales_t) - \log(Sales_{t-1})} - 1$$

(a) From December 1996 to January 1997, the Wine Sales would increase by about

$$e^{0.0072 - 0.5525} - 1 = -42\%$$

(b) From July 1997 to August 1997, the Wine Sales would increase by about

$$e^{0.0072 - 0.7605 - (-0.5644)} - 1 = -17\%$$

(c) It appears that there is a nonlinear relationship between Sales and Time left unexplained. The assumptions that the errors are independent, and identically normal distributed are clearly violated.

Add the term t^2 and run the regression

$$Log(Sales_t) = \beta_0 + \beta_1 t + \beta_2 Jan + \cdots + \beta_{12} Nov + \beta_{13} t^2 + \epsilon_t$$

(d) There do not appear to be any clear patterns left.

(e) 14755 for September, 1996 and 16829 for October, 1996.

(f) The 95% prediction interval for Log(Sales) in September, 1996

$$[9.5993 - 2 * 0.05837, 9.5993 + 2 * 0.05837] = [9.4825, 9.7160]$$

The 95% prediction interval for Sales in September, 1996 is

$$[13128.75, 16581.45]$$

(g) (Optional) What would be the predicted increase of Wine Sales (in percentage) from March 1997 to April 1997?

$$\frac{Sales_{112}}{Sales_{111}} - 1 = e^{\log(Sales_{112}) - \log(Sales_{111})} - 1 = e^{\beta_1 + \beta_{13}(112^2 - 111^2) + \beta_{Apirl} - \beta_{March}} - 1$$

$$e^{0.0113 - 0.00004 * (112^2 - 111^2) - 0.51453 - (-0.524)} - 1 = 1.2\%$$

(h) (Optional) Without considering the seasonal effect, every month the sales increases by about $100(\beta_1 + 2\beta_{13} t)$ percent. As β_{13} is negative and 0 is not within its 95% confidence interval, the model predicts that the sales increase at a lower rate as time increases.

(i) (Optional) See the Excel file in the course website