THE UNIVERSITY OF TEXAS AT AUSTIN
McCombs School of Business

STA 372.5 **Spring 2018**

HOMEWORK #4 – Due Wednesday, February 21

1. Problem #3 on the 2017 midterm exam.

2. The file STA372_Homework4_Question2.dat on the *Data sets* page of the Canvas class website contains 40 quarters of sales data for The Gap (in thousands of dollars). The columns in the dataset are (in order) *Time*, *Quarter* and *Sales*.

   The Gap sales data analyzed in this problem is the same data that was analyzed in homework #3. You should use the R script you created for homework #3 and add on the necessary code at the bottom of that script to do this problem.

   Hint: For help in writing the additional portion of the R script required for this problem, see the R script used in class to run a lagged regression for the Wal-Mart data (i.e., see pages 16 and 17 of the lecture note: "Lagged *y*-values")

   (a) Regress $\log(A_t)$ against t and t^2, where $\log(A_t)$ is seasonally adjusted $\log(Sales_t)$. Use *ggplot2* to plot the residuals from this regression and the R command *acf* to compute their autocorrelation function. Are the residuals from this regression independent? This regression as well as the plot of the residuals and the calculation of their autocorrelation function were already done in homework #3 so you can use the same R script for part (a) here as used in homework #3.

   (b) Estimate the coefficients in the model

   $$\log(A_t) = \alpha + \beta_1 \log(A_{t-1}) + \beta_2 t + \beta_3 t^2 + \varepsilon_t \qquad \varepsilon_t \text{ iid } N(0, \sigma_\varepsilon^2)$$

   using the *lm* command in R. As a check, the estimated regression model you should get is:

   $$\log(A_t) = 4.683 + 0.608 \log(A_{t-1}) + 0.024t - 0.000133t^2 .$$

   (c) Plot the residuals from the regression in part (b) and their autocorrelation function. Are the residuals independent? What does this imply about whether there is information left in the residuals?

   (d) How does the standard deviation of the error terms compare for the models in parts (a) and (b)? What is the implication of this?

(e) Compute the in-sample forecasts for $\log(A_t)$ using the model in part (b). Use *ggplot2* to plot $\log(A_t)$ vs. *Time* and (*In-sample forecasts of log*(A_t)) vs. *Time* on the same plot.

(f) Compute the in-sample forecasts for $Sales_t$. Use *ggplot2* to plot $Sales_t$ vs. *Time* and (*In-sample forecasts of* $Sales_t$) vs. *Time* on the same plot. Do the in-sample forecasts of *Sales* do a good job tracking the actual *Sales* values?

(g) Compute the forecast and 95% confidence interval for *Sales* in quarter 41.

(h) Compute the forecast and 90% confidence interval for *Sales* in quarter 42. Note that part (h) is asking for a 90% confidence interval, not a 95% confidence interval.