WORK.MILK

Northwestern University

School of Continuing Education

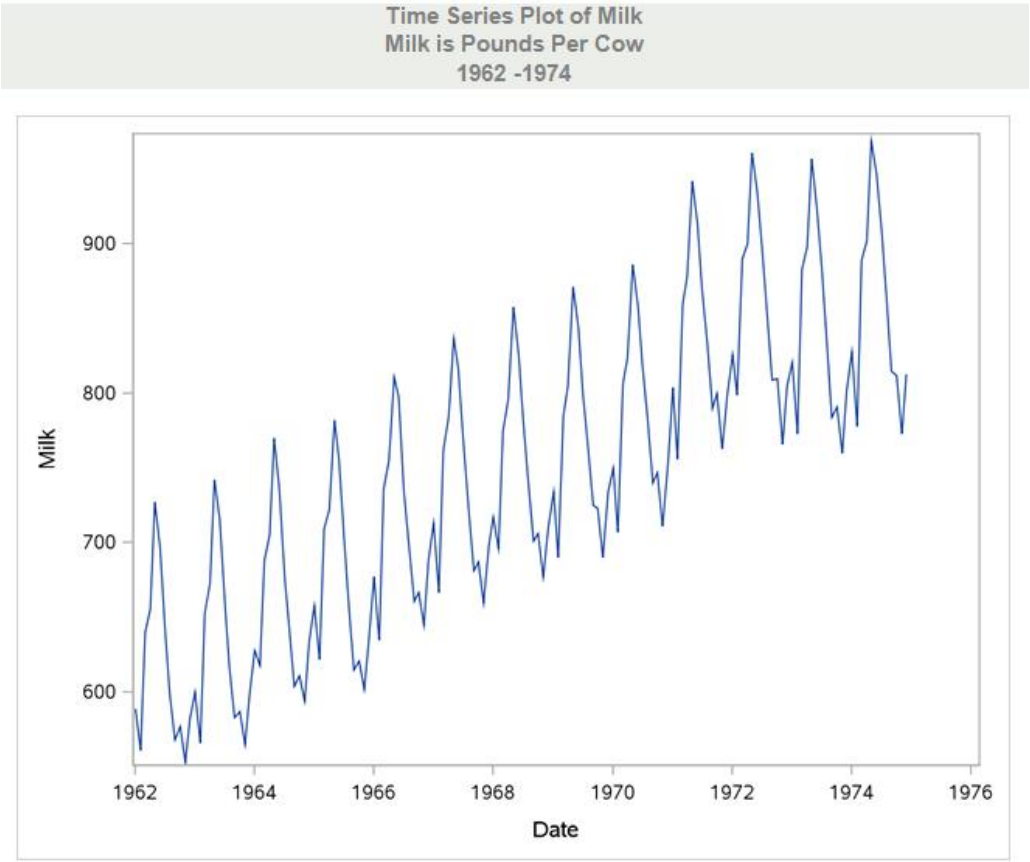
Final Exam-Predict 411-Winter 2013

(Please use extra sheets to complete your answers as needed)

1. (25 points): The attached Excel file contains data from Cryer (1986) and gives the average monthly milk production per cow (for a certain state in the U.S.) for the period January 1962 to December 1975 (n=168). Your task is to conduct a thorough Box-Jenkins analysis on the data and to address the questions which follow (Eliminate the data for 1975 and keep it as a hold out sample):

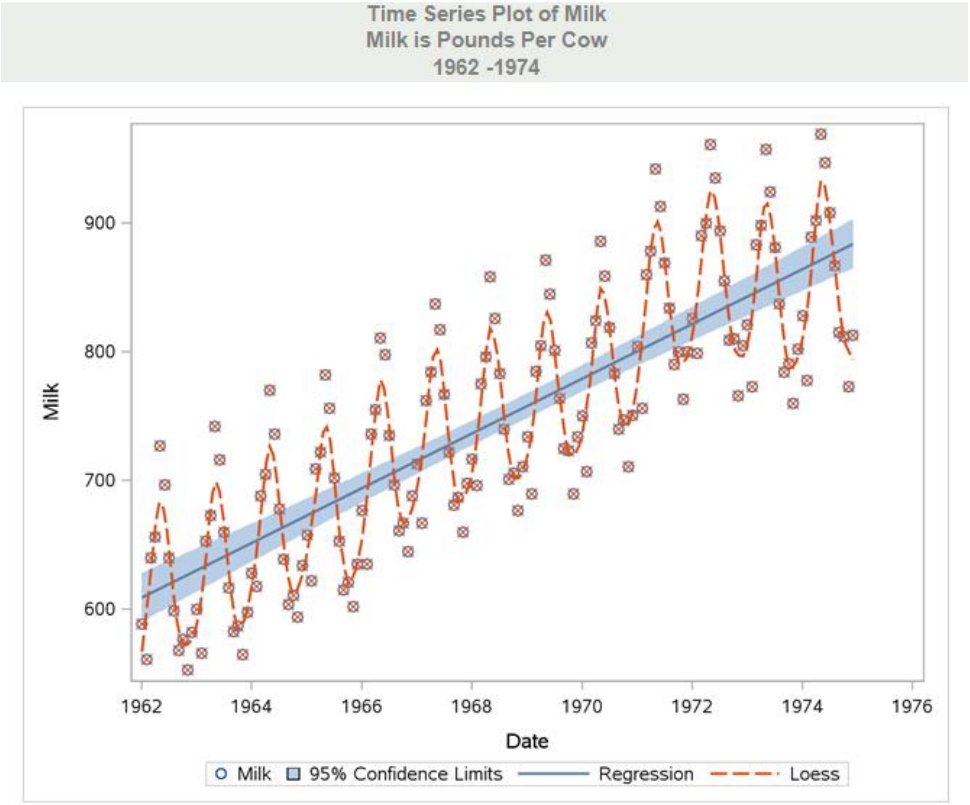
   Reference: J. D. Cryer, ``Time Series Analysis", Duxbury Press (1986)

a. Plot a time series graph of the given data. Please provide all the labels and appropriate titles. Interpret the plot. (3 points)

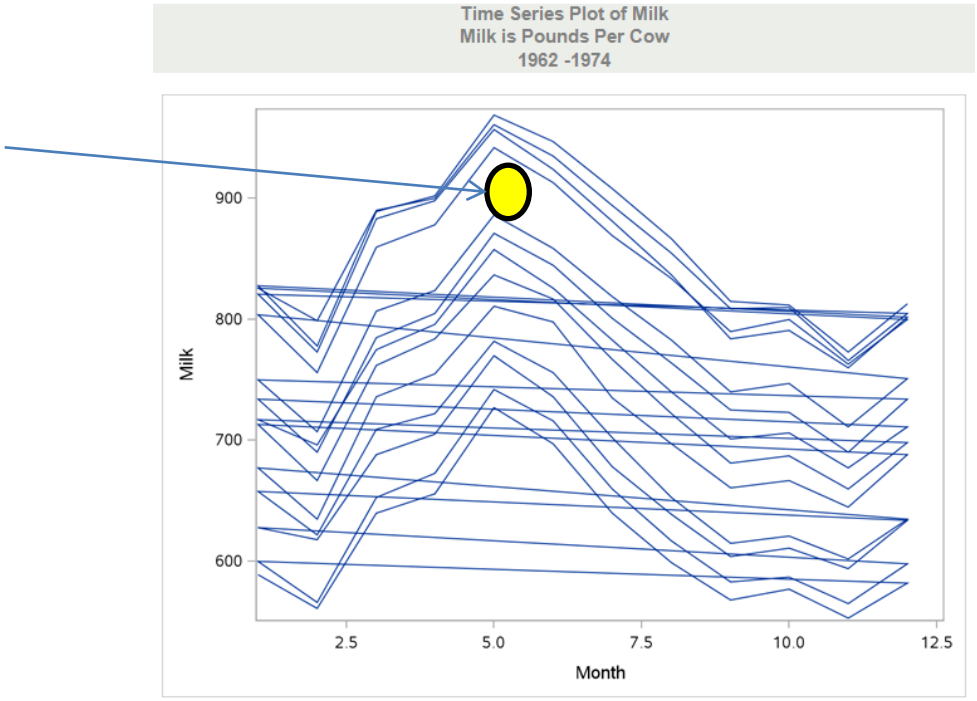



On the Y-Axis, one can see the milk yield, which is pounds per cow. On the X-Axis, one can see date, which is expressed over a monthly time period over 13 years. Notice how this plot leaves out data for 1975 because it will be analyzed at a further time. From this plot, one cannot ascertain if this is a national, state, or regional yield for farmers. Initially, as time increases so too does the milk yield. In addition, it would appear that January and May have yields that spike.

b. Does the data appear to be stationary? Explain why or why not? Is there a need for a variance stabilizing transformation? Explain why or why not? Is there evidence of seasonality in the data set? Explain why or why not? (2 points)



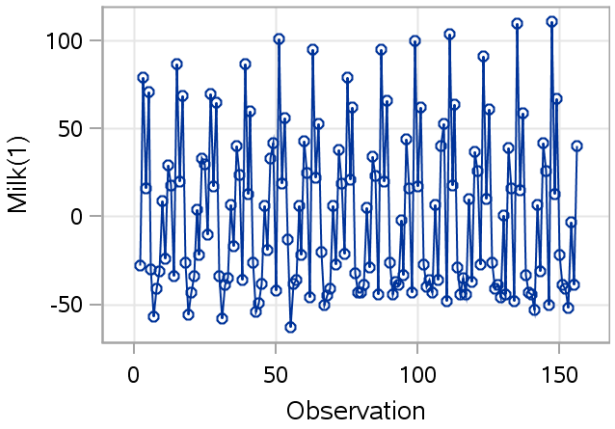

According to Wang, the definition of stationary is, "where a time series has a constant mean and not trend over time." The regression line is helpful for visually seeing that overtime there is an upward trend, which shows that the data is not stationary, because the average will not be zero. The seasonality can clearly be seen through the Loess smoother. Every year, the production starts rather low than around the middle of the year it reaches its peak, then it decreases. One can see this year over year with a gradual increase as time increases.



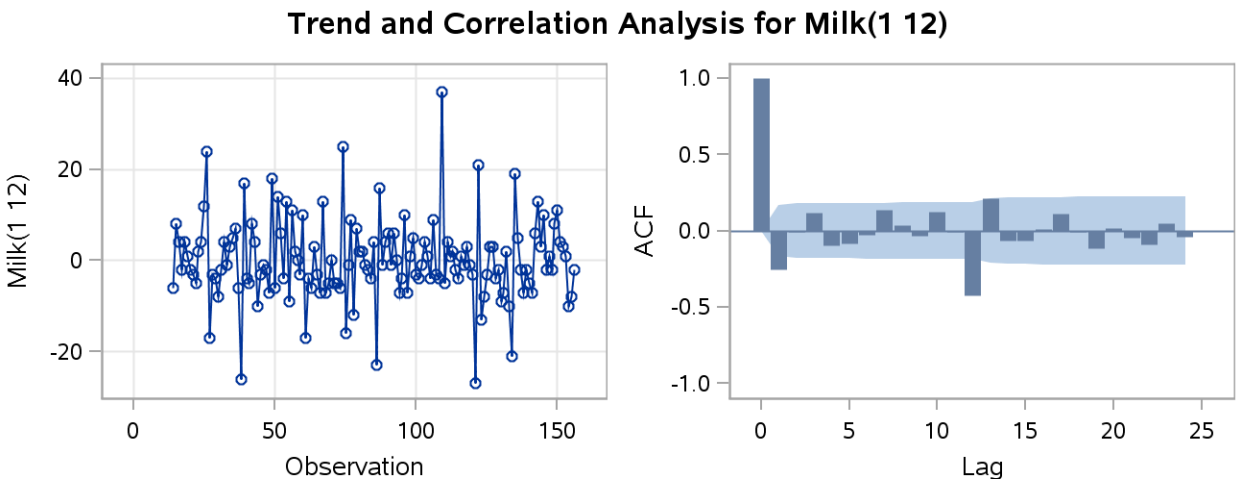

The variance on the other hand looks to be stable. The time series plot above shows the years expressed as months stacked on top of one another. One can see that the variance year over year remains roughly the same. What appears to change year over year is the increase in yield, and not the variance. One can see that at year 1970-1971 there was an increase in yield. If the variance did increase over time, the last year would have a similar initial milk yield in January to that of the first years yield, and then as the months progressed it would spike much higher and then regress back to the first years final year yield. As a result, I do not believe there is a need for a variance stabilizing technique.

c. Explain the approach one would take to convert this time series to a stationary time series? Implement the approach and provide evidence that the time series is now stationary. (2 points)

Differencing is a technique that transforms the time series to a different time series in order to eliminate the non-stationarity. For example, X(t) equals the variable X over specific intervals. When differencing, x(t) is transformed to d(t) and d(t) equals the difference between consecutive values of x(t). This is considered the first difference. The second difference is d(1) (t) – d(1) (t-1), and can progress on to additional differences (statistics.com). From the data set, the goal is to get the mean around zero and to eliminate the seasonality.




After first order differencing, the mean is right around zero, but the seasonality is still present. At this point, one could move on to picking an ARIMA model, but the seasonality that is still present will greatly affect the autocorrelation check for residuals. Essentially, the seasonality pattern has useful information in it that will make the model less accurate if it is not dealt with. The model would not pass the autocorrelation check of residuals, thus I will deal with the seasonality now. From the monthly time series plot above, one can see that the seasonality is based on an annual occurrence starting in January and ending in December. Thus, the seasonality is based on a twelve month pattern. Differencing at the $12^{th}$ lag should get rid of the seasonality.




Differencing at both the 1 and 12 lag seems to have assuaged the issues of the upward trend as well as the seasonality. The goal of differencing is to choose the lowest order coupled with a stationary mean where the time series fluctuates around zero and the ACF plot decline to zero quickly. Notice how in the ACF plot it declines quickly to zero, and there are a few plots that are outside of the confidence ban and the scatter plot still shows signs of mild seasonality. Time series modeling is an art and science, and at this point the data can be accurately used to build a Box-Jenkins ARIMA model.

d. Conduct appropriate analysis to identify an optimal ARIMA model for the data set? Provide details of every step you take towards reaching the optimal model. (10 points)

Stated in my answer for question c the data is stationary and the seasonality has been removed, which is a rather large part of picking an optimal model.

Modeling time series data utilizing Box-Jenkins techniques follows four main steps:

1. Model Identification – Based on ACF and PACF plots as well as diagnostics
2. Model Estimation – Conditional Least Squares (based on minimized sum of squared residuals for stochastic models)
3. Diagnostic Checking – Assessment of variance, AIC, as well as statistically valid coefficients.
4. Forecasting

Model Identification is centered on the acronym ARIMA which means Autoregressive Integrated Moving Average. Within this acronym, there are three models that can be used to fit time series data. Before delving into the models, the one major assumption of ARIMA is that the data has been transformed into a stationary time series, which means the data has a mean of zero and no trend overtime. As seen above, this has been accomplished.

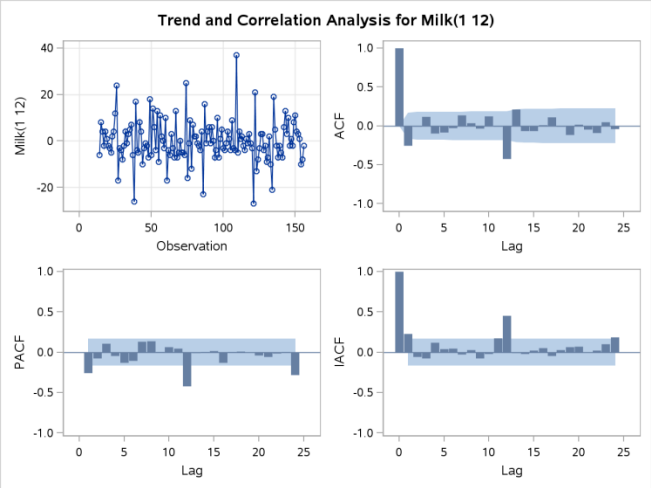
Now that the time series is stationary, I will assess the white noise test to validate that the data is not white noise, which tests whether there is actually any useful data to model.

| Autocorrelation Check for White Noise | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
| 6 | 14.38 | 6 | 0.0257 | -0.257 | -0.007 | 0.118 | -0.098 | -0.087 | -0.026 |
| 12 | 48.63 | 12 | <.0001 | 0.137 | 0.034 | -0.031 | 0.122 | -0.006 | -0.426 |
| 18 | 59.07 | 18 | <.0001 | 0.211 | -0.067 | -0.064 | 0.009 | 0.108 | -0.009 |
| 24 | 63.81 | 24 | <.0001 | -0.116 | 0.016 | -0.046 | -0.089 | 0.050 | -0.041 |

All the lags pass the P-value statistic at the 95% confidence interval, but one should note that lags one through six do have some white noise but they still pass. The autocorrelation check for white noise statistically demonstrates that we can reject the null hypothesis that the autocorrelations in all the lags are jointly zero. While the stationarity is not ideal, for the purposes of building this model it will suffice. The next step is to fit the model.

When fitting a model to time series data, a specific nomenclature is followed to keep track of the specific model. The ARIMA model has three factors involved, (p,d,q), where p= the order of autoregressive components that are statistically significant, d= the number of differencing that has been done on the data, and q= the order of the moving average (JMU.edu). In the model selection, the nomenclature will follow a (p,d,q) list of the ideal model selected. Analyzing the inferential statistics AIC is an informal method to assess the model fit. This statistic can be used to compare different sets of variables. Higher values mean a worse fit to the data. As in Ordinary Least Squares, AIC is used to penalize models that have more variables. The ARMA model is comprised of terms from both AR and MA. Now the questions of how many terms get included as well as which model is used are answered. Depending on the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) values, a specific model will be used. Both of these functions are similar to what the r-value represents in OLS. Visually the ACF and PACF are great ways to discern if an AR, MA ARMA model is appropriate. The

similarity between AR and MA is the use of lagged terms, but the difference is the AR model uses the lags from the actual timer series where the MA model uses lags from the noise or residuals (colorado.edu). Similar to OLS, the goal of fitting the correct model is a minimized sum of square errors, and statistical validation. Depending on the graphical output and diagnostics, either an AR, MA, or ARMA model will be used. I will fit all the models and work through the output to discern the best model. The initial correlation analysis, to the left, points to an MA model. When the ACF has one large spike, and the PACF diminishes over time, this points to utilizing the MA model. Bear in mind that through this modeling phase a small sum of squares errors is desired.




The first model will follow the nomenclature of a (1,1,0) X (1,1,0), which means that this is a 1 lag autoregressive model with one differencing period with another autoregressive lag term with another differencing component at the twelfth month. The variance is 63.255 and the AIC is 1001.83. In addition to these values, the model is statistically significant at the 95% confidence level for both AR coefficients. The autocorrelation check of residuals show that important information has not been left, because the residuals are white noise. This model has rather impressive numbers, but it needs to be compared to the other models.

An ARIMA model is comprised of (1,1,1) X (1,1,1). Utilizing conditional least squares the autoregressive coefficients are not statistically significant, as well as the first moving average coefficient. The variance is 57.516 and the AIC is 990.17, and the residuals pass for white noise. While these values are better than the AR model, the total ARIMA model is not statistically significant.

The MA model (0,1,1) X (0,1,1) is the most parsimonious as well as statistically valid. The model is statistically significant, variance is .56.823, and the AIC is 986.5. In addition, the autocorrelation check of residuals shows that the remaining values are white noise.

All three models were fit and based on the conditional least squares model estimation the diagnostics showed that the Moving Average model (0,1,1) X (0,1,1) was the best fit.

e. Use the optimal model to forecast the milk production for 1975. How does the forecasted values compare to the observed values? (Seen as the highlighted section) Do the same for another model (a non-optimal model) and compare its results to the one obtained from the optimal model. How do the results compare? (6 points)

| MA Model (Optimal) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Actual | Predicted | Difference | SqDiff | Total | Div by obs | |
| 1975 | 834 | 838.5504 | 4.5504 | 20.70614 | | | |
| 1975 | 782 | 793.9491 | 11.9491 | 142.781 | | | |
| 1975 | 892 | 898.7077 | 6.7077 | 44.99324 | | | |
| 1975 | 903 | 912.87 | 9.87 | 97.4169 | | | |
| 1975 | 966 | 976.0417 | 10.0417 | 100.8357 | | | |
| 1975 | 937 | 949.5253 | 12.5253 | 156.8831 | | | |
| 1975 | 896 | 907.9241 | 11.9241 | 142.1842 | | | |
| 1975 | 858 | 867.2198 | 9.2198 | 85.00471 | | | |
| 1975 | 817 | 818.0684 | 1.0684 | 1.141479 | | | |
| 1975 | 827 | 820.0869 | -6.9131 | 47.79095 | | | |
| 1975 | 797 | 783.3822 | -13.6178 | 185.4445 | | | |
| 1975 | 843 | 823.2742 | -19.7258 | 389.1072 | | | |
| | | | | | 1414.289 | 117.8574 | MSE |

| AR Model (As explained above) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Actual | Predicted | Difference | SqDiff | Total | Div by obs | |
| 1975 | 834 | 834.4973 | 0.4973 | 0.247307 | | | |
| 1975 | 782 | 785.287 | 3.287 | 10.80437 | | | |
| 1975 | 892 | 895.7577 | 3.7577 | 14.12031 | | | |
| 1975 | 903 | 909.5847 | 6.5847 | 43.35827 | | | |
| 1975 | 966 | 972.848 | 6.848 | 46.8951 | | | |
| 1975 | 937 | 945.74 | 8.74 | 76.3876 | | | |
| 1975 | 896 | 904.8296 | 8.8296 | 77.96184 | | | |
| 1975 | 858 | 862.3758 | 4.3758 | 19.14763 | | | |
| 1975 | 817 | 809.8356 | -7.1644 | 51.32863 | | | |
| 1975 | 827 | 811.3195 | -15.6805 | 245.8781 | | | |
| 1975 | 797 | 775.89 | -21.11 | 445.6321 | | | |
| 1975 | 843 | 816.72 | -26.28 | 690.6384 | | | |
| | | | | | 1722.4 | 143.5333 | MSE |

It can be seen that the MA model is a better fit based on the Mean Square Errors. Furthermore, if one were to fit the ARMA model it would have a lower MSE, but remember that the coefficients were not statistically significant.

| ARMA Model | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Actual | Predicted | Difference | SqDiff | Total | Div by obs | |
| 1975 | 834 | 838.0734 | 4.0734 | 16.59259 | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1975 | 782 | 793.5164 | 11.5164 | 132.6275 | | | |
| 1975 | 892 | 898.3668 | 6.3668 | 40.53614 | | | |
| 1975 | 903 | 912.4583 | 9.4583 | 89.45944 | | | |
| 1975 | 966 | 975.3894 | 9.3894 | 88.16083 | | | |
| 1975 | 937 | 948.5275 | 11.5275 | 132.8833 | | | |
| 1975 | 896 | 906.9291 | 10.9291 | 119.4452 | | | |
| 1975 | 858 | 866.1413 | 8.1413 | 66.28077 | | | |
| 1975 | 817 | 816.7949 | -0.2051 | 0.042066 | | | |
| 1975 | 827 | 819.0478 | -7.9522 | 63.23748 | | | |
| 1975 | 797 | 782.3508 | -14.6492 | 214.5991 | | | |
| 1975 | 843 | 822.3126 | -20.6874 | 427.9685 | | | |
| | | | | | 1391.833 | 115.9861 | MSE |

f. Write the full form of the equation resulting from the optimal model. (2 points)

(0,1,1)x(0,1,1).

w_t=(y_t-y_(t-1)-(y_(t-12)-y_(t-13)

w_t=(1-.79B)(1-.75B_12)e_t .21 * .25

w_t=(1-.75B_12-.79B+ .59B_13)e_t

w_t=e_t-.75e_(t-12)-.79e_(t-1) + .59e_(t-13)

This is the final notation for the MA model:
y_t=y_(t-1)+y_(t-12)-y_(t-13) + e_t-.75e_(t-12)-.79e(t-1)+.59e(t-13)

(0,1,1)x(0,1,1).

w_t=(y_t-y_(t-1)-(y_(t-12)-y_(t-13)

w_t=(1-.25B)(1-.61B_12)e_t

w_t=(1-.61B_12-.25B+ .15B_13)e_t

w_t=e_t-.61e_(t-12)-.25e_(t-1) + .15e_(t-13)

This is the final notation for the MA model:
y_t=y_(t-1)+y_(t-12)-y_(t-13) + e_t-.61e_(t-12)-.25e(t-1)+.15e(t-13)

2. (10 points): Consider two structural models given by the following system of equations (Note: These are two independent models):

Model 1

$$Y_1 = \alpha_1 + \alpha_2 Y_2 + \alpha_3 X_1 + \alpha_4 X_2 + u_1$$
$$Y_2 = \beta_1 + \beta_2 Y_3 + \beta_3 X_2 + u_2$$
$$Y_3 = \gamma_1 + \gamma_2 Y_2 + u_3$$

Model 2

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_2 + u_1$$
$$Y_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u_2$$

For each system:

a. Determine which equations are under-identified, just-identified, and over-identified. Give justification for your responses.

$Y_1 = \alpha_1 + \alpha_2 Y_2 + \alpha_3 X_1 + \alpha_4 X_2 + u_1$ =Under-identified, based on the fact that X1 and X2 are exogenous and there are no excluded exogenous variables.

$Y_2 = \beta_1 + \beta_2 Y_3 + \beta_3 X_2 + u_2$ = Just-identified, based on the fact that the endo and excluded exo genous variables are equal.

$Y_3 = \gamma_1 + \gamma_2 Y_2 + u_3$ = Over-identified, there is only one endogenous variable in the equation and two excluded exogenous variables.

$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_2 + u_1$ = Over Identified, there are 2 excluded exogenous variables and one included endogenous variable.

$Y_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u_2$ = Just Identified, there are 0 excluded exogenous variables and 0 included endogenous variables.

b. Explain how you would estimate the identified equations (both the just- and over-identified equations)

One can substitute the known values for the unknown coefficients per equation and that will estimate the equations correctly.

3. (10 points): Using a sample of 545 full-time workers in the United States, a researcher is interested in whether women are systematically underpaid compared to men. First, she estimates the average hourly wages in the sample for men and women, which are $5.91 and $5.09, respectively.

   a. Do these numbers answer the question of interest? Why not?

   The question of interest is, "Are women systematically underpaid compared to men?"
   There are really two questions being asked here:
   1: Are women underpaid compared to men?
   2: Are women <u>systematically</u> underpaid compared to men?
   A synonym for Systematically is Scientifically.
   In order to answer this question, one must first ascertain whether women are underpaid compared to men, and then establish a systematic approach to figuring this out.
   Using a sample of 545 full-time workers to represent the 113.8 million full time workers (Statstica.com) in the US is to statistically small to be a significant sample. Furthermore, one does not know the estimation method used, as well as whether one can infer a "systematic" pattern of underpayment from the sample size. Do certain industries underpay more than others, how did one delineate sexism versus age, years of experience, demographic, and education? These are all questions that massively skew the results.

   b. How could one (at least partially) correct for this?

One could partially correct for the errors in this experiment by delineating the region, county, specific industry, education level, years of experience, and demographic information. Furthermore, it would be a lot easier to answer the question of, "Are women underpaid compared to men?" through this study. The researcher simply does not have the structure to answer the systematic part of the initial question.

The researcher also runs a simple regression of an individual's wage on a male dummy, equal to 1 for males and 0 for females. This gives the results reported in Table 1 below:

Table 1: Hourly wages explained from gender: OLS results

| Variable | Estimate | Standard Error | t-ratio |
|---|---|---|---|
| Constant | 5.09 | 0.58 | 8.78 |
| Male | 0.82 | 0.15 | 5.47 |

N=545 s=2.17 $R^2$=26%

c. How can you interpret the coefficient estimate of 0.82? How do you interpret the estimated intercept of 5.09?

For the male coefficient a 1 unit increase will increase Y by .82 plus the intercept, and a 1 unit decrease will decrease Y by .82 plus the intercept. Bear in mind that the model would actually look like 5.09 + (.82 x X).

The intercept is the average value of Y, the response variable, when all the coefficients equal 0. In other words the intercept is the simple average without adding in the male variable. In some equations, the intercept is arbitrary but for this model the intercept is very important because this model uses a dummy variable. If there is not a male in the equation, the hourly rate is just 5.09(Hour).

d. How do you interpret the $R^2$ value?

The coefficient of determination measure how well the regression line estimates the actual data points. Thus, it would be stated that 26% of the variation in the dependent variable can be explained by the independent variable.

e. Explain the relationship between the coefficient estimates in the table and the average wage rates of males and females.

Given that this OLS is utilizing a dummy variable, both 1 and 0 have a value. In the model, 0 represents the lack of being male, thus one is a female. Notice how the intercept is simply the hourly wage for female workers. Now, if one adds the constant with the coefficient value for male it equals the average hourly wage for men. One could also make male the intercept, but then female would have a negative value.

f. A student is unhappy with this model because "a female dummy is omitted from the model." Comment upon this criticism.

Given that this model is using a dummy variable, 0 represents female, thus the intercept is the female coefficient.

In addition, there are no other variables in the model, but if one were to add additional variables it would be assumed that the coefficients would be interpreted as the value (x) given that all the other variables remain constant.

g. Using the results in Table 1, test the hypothesis that men and women have, on average, the same wage rate, against the one-sided alternative that women earn less. State the assumptions required for this test to be valid.

H0:=M=F

H1:F<M

This is a left tailed test.

The t-test equals 5.09/.58 = 8.78 the p-value is =.036098

The assumption of a p-value is that the truth of the null hypothesis is that the finding was the result of chance alone. In addition there are the standard assumptions for OLS to be valid.

Linearity Assumption: The response Y to the predictors X is assumed to be linear in its regression parameters. (Meaning it conforms to a straight line), 94 Chaterjee.

The errors are assumed to be independently and identically distributed normal random variables each with mean zero and a common variance. This implies four other sub-assumptions.

a. The error has a normal distribution, this can be seen visually through graphs.
b. The errors have mean zero.
c. The errors have the same variance o2, known as constant variance assumption. When this is not the case, it will be discussed in Chap 7.
d. The errors are independent of each other, thus they have no covariance. Chapter 8 covers when this assumption is not met.

Assumptions of the Predictors:

e. The predictor variables are non-random, meaning they are assumed fixed or selected in advance.
f. The values are measured without error.
g. The predictor variables are assumed to be linearly independent of each other.

All observations are equally reliable and have an approximately equal role in determining the regression results, 96 Chaterjee.

If the assumptions hold, the null hypothesis can be rejected at the 95% confidence level.

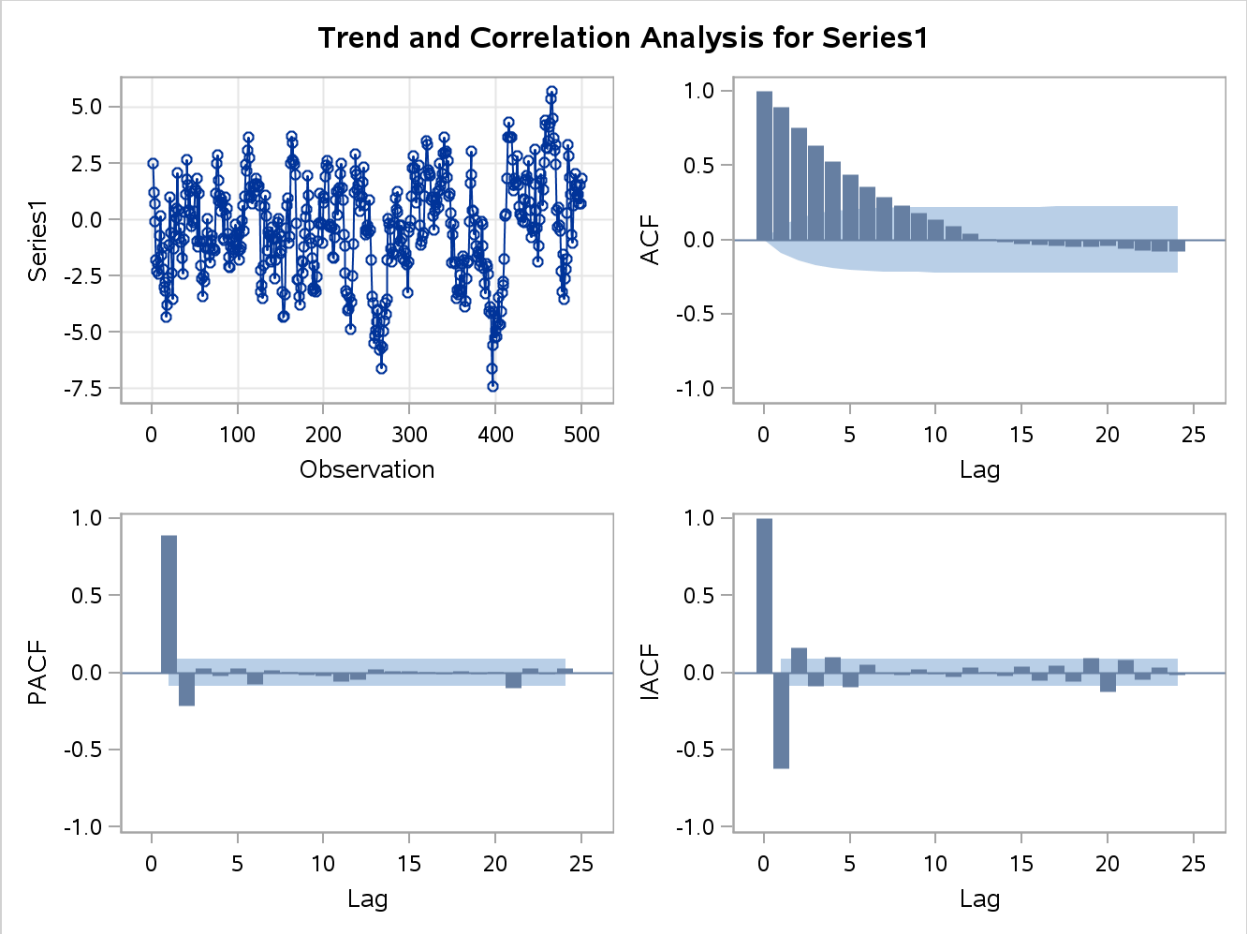
4. (10 points): Given the model $y_t = \alpha_0 + \varphi_1 y_{t-1} + \varepsilon_t$, answer the following questions.

   a. Assume that $\varepsilon_t \sim N(0, \sigma^2)$. What type of time series model is this?

   Given the error term, the time series model is autoregressive1.

   b. Graph the value of $y$ against $t$ for 10 periods when $\varepsilon_1 = 0.2$, $\varphi_1 = 0.8$, and $a_0 = 0$.

I did 500 to get a good glimpse of the trend. This plot suffers from increased variance.




   c. Draw an appropriate ACF and PACF plot for the model given in this question.

5. (15 points): Consider the following OLS regression between the 1975 Wages for 428 married women versus their actual experience in the labor market and their years of education (1976 Panel Study of Income Dynamics, Mroz(1987).

$$\log(wage) = -0.400 + 0.0160 \times Exper + 0.1095 \times Educ$$

The data set was analyzed using SAS. Partial output in tabular form is presented below

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr>F |
|---|---|---|---|---|---|
| Model | **A 2** | **D 33.13246** | 16.56623 | **G 37.017** | **H <.0001** |
| Error | **B 425** | **E 190.196** | 0.44752 | | |
| Corrected Total | **C 427** | **F 223.32846** | | | |

Root MSE: I **.668969**

R-Square: **J .**148357

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t value | Pr>\|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -0.40017 | 0.19037 | | |
| Exper | 1 | 0.01567 | 0.00402 | **3.898** | **M One Tailed P-value =.079936** |
| Educ | 1 | 0.10949 | 0.01417 | **L 7.726888** | **N One Tailed P-value = .040967** |

a. Calculate the values associated with the letters **A through N**.
   A 2 independent variables thus DF = 2
   B n-k-1= 428-2-1=425
   C 425+2=427
   D Mean Square * DF = 16.56623*2=33.13246
   E Mean Square * DF = .44752*425=190.196
   F SS(W) + SS(B) = 223.32846
   G MSR/MSE = 16.56623/.44752 = 37.017
   H <.0001 This involves F-value Numerator and Denominator
   I Square root of the Mean Square Error = square root of .44752
   J R-squared = SSR / SST = 33.13246/223.32846=.148357

K P.E. / S.E. =

***I am calculating the P-value as a one tailed p-value*****

b. Interpret the coefficients associated with Exper and Educ.

g. $\log(wage) = -0.400 + 0.0160 \times Exper + 0.1095 \times Educ$

For a one unit change in Exper, log(wage) changes by (.0160 X 100% = 1.6%) holding all other explanatory variables constant.

For a one unit change in Educ, log(wage) changes by (.1095 X 100% = 10.95%) holding all other explanatory variables constant.

6. (10 points) A criminologist is interested in studying the following question: "Is the death penalty applied in a racially discriminatory fashion?" To answer this question, data were collected for 100 death penalty cases in the State of Georgia. Logistic regression was used with the binary dependent variable *death penalty* against a number of independent variables. The analysis is set up to obtain the predicted probability of getting the death penalty (*death penalty* = 1). The independent variables were defined as follows:

   *blkdef* = 1 if black defendant; 0 otherwise.

   *whtvict* = 1 if white victim; 0 otherwise.

   *aggcirc* = number of aggravating circumstances.

   *fevict* = 1 if female victim; 0 otherwise.

   *stranger* =1 if stranger victim; 0 otherwise.

   *multvic* = 1 if 2 or more victims; 0 otherwise.

   *multstab* = 1 if multiple stabs; 0 otherwise.

   *yngvict* = 1 if victim 12 or younger; 0 otherwise.

   A partial output table is given below:

| Parameter | DF | Estimate | Standard Error | Wald Chi Square | P-Value |
|---|---|---|---|---|---|
| Intercept | 1 | -3.5675 | 1.1243 | 10.0682 | 0.0015 |
| blkdef | 1 | -0.5308 | 0.5439 | 0.9526 | 0.3291 |
| whtvict | 1 | 1.5563 | 0.6161 | 6.382 | 0.0115 |
| aggcirc | 1 | 0.373 | 0.1963 | 3.6096 | 0.0574 |
| fevict | 1 | 0.3707 | 0.5405 | 0.4703 | 0.4928 |
| stranger | 1 | 1.7911 | 0.5386 | 11.0577 | 0.0009 |
| multvic | 1 | 0.1999 | 0.745 | 0.072 | 0.7885 |
| multstab | 1 | 1.4429 | 0.7938 | 3.3047 | 0.0691 |
| yngvict | 1 | 0.1232 | 0.9526 | 0.0167 | 0.8971 |

a. Holding all other variables constant and using a type I error rate of 5%, are black defendants more likely to get the death penalty than white defendants? Why or why not? Interpret the coefficient for *blkdef.*

Interpreting coefficients for Logistic Regression (LR) is not as straight forward as in Ordinary Least Squares Regression. In the LR model above a logit coefficient of -.5308 can be interpreted as -.5308 log odds increase for every 1-unit increase in the explanatory variable, assuming all the other coefficients are held constant (Allison). It is really hard to conceptualize a -.5308 log-odds increase, which can be explained by the fact that LR captures a nonlinear relationship. From the Maximum Likelihood Estimates output, what can be gleaned is the statistical significance as well as the sign of the estimate. A positive or negative sign indicates the direction of the relationship (uoregon.edu). In addition to assessing the sign of a coefficient, analyzing the statistical significance is important. P-values assess the probability that your sample results are chance or extreme given that the null hypothesis is true. As the p-value increases, the probability increases that the sample estimate is based on pure chance. Lower p-values are an indicator of a statistically solid coefficient. Given that P-value is beyond the 95 confidence interval this variable is not statistically significant.

The log-odds ratio is a far better output for understanding the coefficients. The odds ratio is simply computed by taking the natural log of e^estimate. In addition to the odds, this calculation is also adjusted since it controls for other variables. Once the odds ratio has been calculated, the output is easier to understand. For example, the odds ratio for blkdef can be calculated as 2.7182 (e) ^ -.5308 = .5881. This can be interpreted as an individual that has prior convictions has .5881 times the odds of recidivism than non-prior conviction. I personally prefer probability to odds. This can be translated from the following calculation, probability = .5881/(1 + .5881) = .37. Even though the variable is not statistically significant, black defendants only have a .37 probability of receiving the death penalty versus nonblack defendants.

b. Calculate the odds ratio of getting the death penalty for a defendant whose crime was against a white defendant. Is this odds ratio statistically significant using a type 1 error rate of 5%. Interpret the odds ratio.

2.7182 (e) ^ 1.5563= 4.74 is the odds ratio for a defendant whose crime was against a white victim compared to a non-white defendant. The probability is 4.74/(1+4.74) 83%, which in my opinion is quite high. This odds ratio is statistically significant using a type 1 error rate of 5%.

c. What is the predicted probability of getting the death penalty for a black defendant who kills a white (female) victim who is a stranger with two aggravating circumstances, multiple victims, multiple stabs, and a victim younger than 12 years of age? What would the prediction be if all that changed was that the defendant was not black?

P= (-3.5675) + 1.5563 + .3707 + (-.5308) + 1.7911 + (.373*2=.746) + (.1999) + (1.4429) + (.1232) = 2.1318
2.7182^2.1318=8.43  8.43/(1+8.43)= 89%

Predicted probability for not black:
P= (-3.5675) + 1.5563 + .3707 + (-.5308) + 1.7911 + (.373*2=.746) + (.1999) + (1.4429) + (.1232) =2.6626
2.7182^2.6626 =14.33 14.33/(1+14.33)= 93%

d. The regression coefficients for *multvic* and *yngvict* are not statistically significant. Make an argument for why we would include these independent variables in the logistic regression model even though their regression coefficients are nonsignificant.

There may be a non-statistical reason to keep the variables in the model. Perhaps the manager overseeing the model construction for political reasons wanted the variables in the equation. These variables also emotionally exacerbate a crime which could be appealing to the prosecutor for a jury presentation.

7. (20 points) Short answer questions.

a. Explain what is meant by omitted variable bias.

Omitted variable bias (OVB) is when control variables are omitted from the model. Specifically, the bias occurs in the OLS estimator. In order for the bias to occur, two things must occur for the missing factor: 1, it must be a determinant of the response variable, 2 – correlated with the repressor (s) independent variable (s). Both of these must occur for OVB to occur.

In particular, how does omitted variable bias impact the regression coefficients?
The regression coefficients residuals correlate, which is a violation of the OLS assumptions. Specifically, the coefficients residuals are related to the omitted variable, which makes the results less reliable, bias, and inconsistent.

Discuss one technique to mitigate omitted variable bias.

Instrumental variable estimator is an approach that is used in presence of endogenous regressors. In a nut shell, this allows for the creation of an "instrumental variable" that is a consistent estimator for the slope, is correlated with "x" and uncorrelated with the error term. There are a few ways to get this "magical" variable through utilizing scalar regressors, the Wald estimator, and two-stage least squares depending on the model.

b. What are the main differences between GLS and FGLS estimation techniques?
Generalized Least Squares (GLS) is often used when the presence of heteroscedasticity and autocorrelation is present in a dataset. The assumptions for GLS are more relaxed than OLS, and include an assumption that the errors are nonspherical. There is a rather large shortcoming with using GLS in that one must know the variance-covariance matrix of the disturbances, which are never actually known (Eastern Michigan). Feasible Generalized Least Squares (FGLS) takes the sample data and estimates a variance-covariance matrix. Given that the data is heteroscedastic and specifically this set is working within time series, it is assumed the errors are correlated and thus non-spherical. FGLS essentially transforms the residuals, similar to Weighted Least Squares, such that they are spherical and satisfy the assumptions of linear regression. FGLS transforms the model with the purpose of obtaining more efficient estimators and standard errors than OLS.

c. Discuss two uses of the Hausman's Specification Test.
The Hausman's Specification Test can be used to statistically assess using a Full Model or Reduced Model, and assess estimator's verses other estimators. In addition, it also helps to evaluate if a model relates to the data.

d. Discuss the main differences between fixed and random effects models.
Fixed effects and random effects are specific model techniques that are used to analyze panel data. Fixed effect (FE) models allow for different intercepts per subject in the data set, but assumes that the slopes are constant, ie parallel to one another. FE modeling has quite a few benefits, but one of the major weaknesses is the assumption that the subjects represent the entire population. From my experience, one rarely is able to capture an entire population, thus the conclusions drawn from FE can only be inferred on the data studied. RE modeling assumes random distribution based on the differences from the subjects. This assumption broadens the scope drawn from the model to include larger populations than just the studied data. The caveat for this model is the assumption that the unobserved heterogeneity is independently distributed from the subjects, which is very hard to satisfy in reality. In essence, the RE model is very similar to the pooled model except that an analysis is conducted between the error terms. Based on the error term, one will discern whether or not to utilize the FE model or the pooled model.

e. For a sample of 600 married females, we are interested in explaining participation in market employment from exogenous characteristics in $\mathbf{x}_i$ (age, family composition, education). Let $y_i = 1$ if person $i$ has a paid job and 0 otherwise. We estimate a probit regression model using this data set. Suppose you have a person with $\mathbf{x}_i^T \beta = 2$ where $\beta$ is the estimated vector of regression coefficients. What is your prediction for her labor market status, $y_i$? Give the exact predicted probability.

97.5% --> Given that the distribution curve includes everything to the left.

f. What is meant by overdispersion as it relates to Poisson regression models? How does this impact a model's regression coefficients? A model's standard errors? Discuss two ways of mitigating this.

Poisson regression is based on a Poisson distribution, which is a discrete or binary likelihood distribution. I think of this similar to logistic regression in that the distribution is non-linear. The distribution gives a probability for a specific count of events that happen in sequence with a prior-known mean rate with independence from past events. In Poisson regression the mean is equal to the variance. Overdispersion occurs when the variance is greater than the mean. Thus, Poisson regression should not be used.
The effects of overdispersion on the coefficients are accurate, but not as precise as the p-values make them to appear. Thus, the standard errors are conservative when overdispersion is present.

Negative Binomial allows for more variability in the data, and is based on a different probability.
Zeroes can cause overdispersion, and one approach is to use Zero-inflated Poisson (ZIP). This approach allows for the inflation of zeroes and predicts a pragmatic percent for the zeroes.