**UNIT 4 HOMEWORK!!!**

Below is a data set that has the year, the average number of melanomas per 100,000 people and the number of sunspots that year. We will first analyze just the melanoma data. ***Please include your code for all questions.***

**data** mel;

input Year Melanoma Sunspot;

datalines;

1936 1 40

1937 0.9 115

1938 0.8 100

1939 1.4 80

1940 1.2 60

1941 1 40

1942 1.5 23

1943 1.9 10

1944 1.5 10

1945 1.5 25

1946 1.5 75

1947 1.6 145

1948 1.8 130

1949 2.8 130

1950 2.5 80

1951 2.5 65

1952 2.4 20

1953 2.1 10

1954 1.9 5

1955 2.4 10

1956 2.4 60

1957 2.6 190

1958 2.6 180

1959 4.4 175

1960 4.2 120

1961 3.8 50

1962 3.4 35

1963 3.6 20

1964 4.1 10

1965 3.7 15

1966 4.2 30

1967 4.1 60

1968 4.1 105

1969 4 105

1970 5.2 105

1971 5.3 80

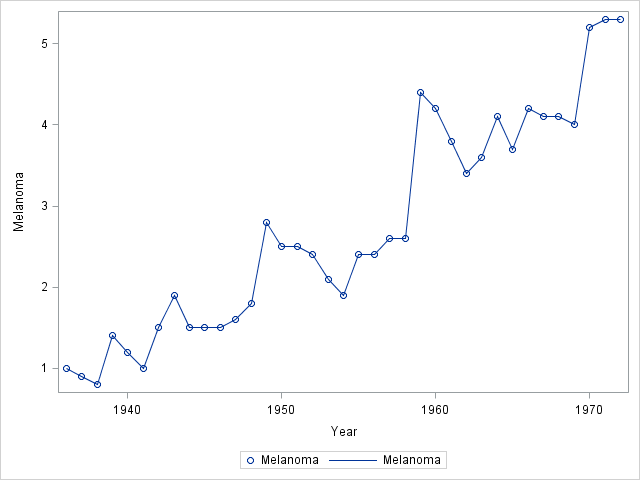
1972 5.3 65

;

1. We always want to plot the data first. Please plot the data with time on the x axis and melanoma per 100,000 people on the y axis. First of all, does it look like the mean of the rate of melanoma is increasing over time? What could be causing this … time or something possibly something else? This second question is just for thought … there are of course many possible answers.

It appears the mean is in fact increasing over time. However, I would be cautious to assume there is a relationship between the year and the amount of people with melanoma. With the global population increasing substantially every year, it is possible the number of individuals with melanoma is actually remaining constant when looking at the global percentage of people with melanoma. Nonetheless, more research is needed.

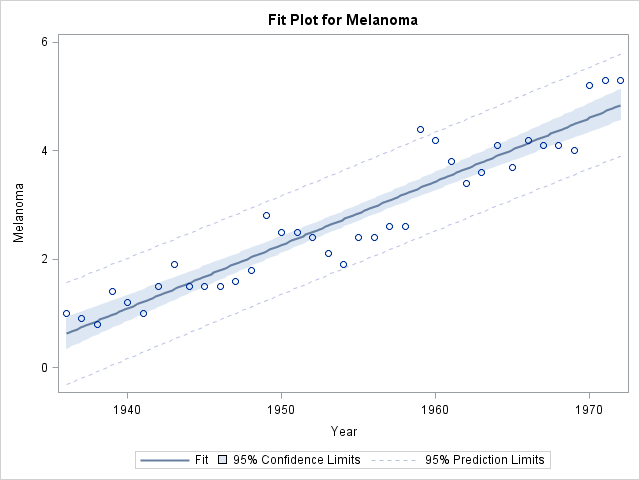
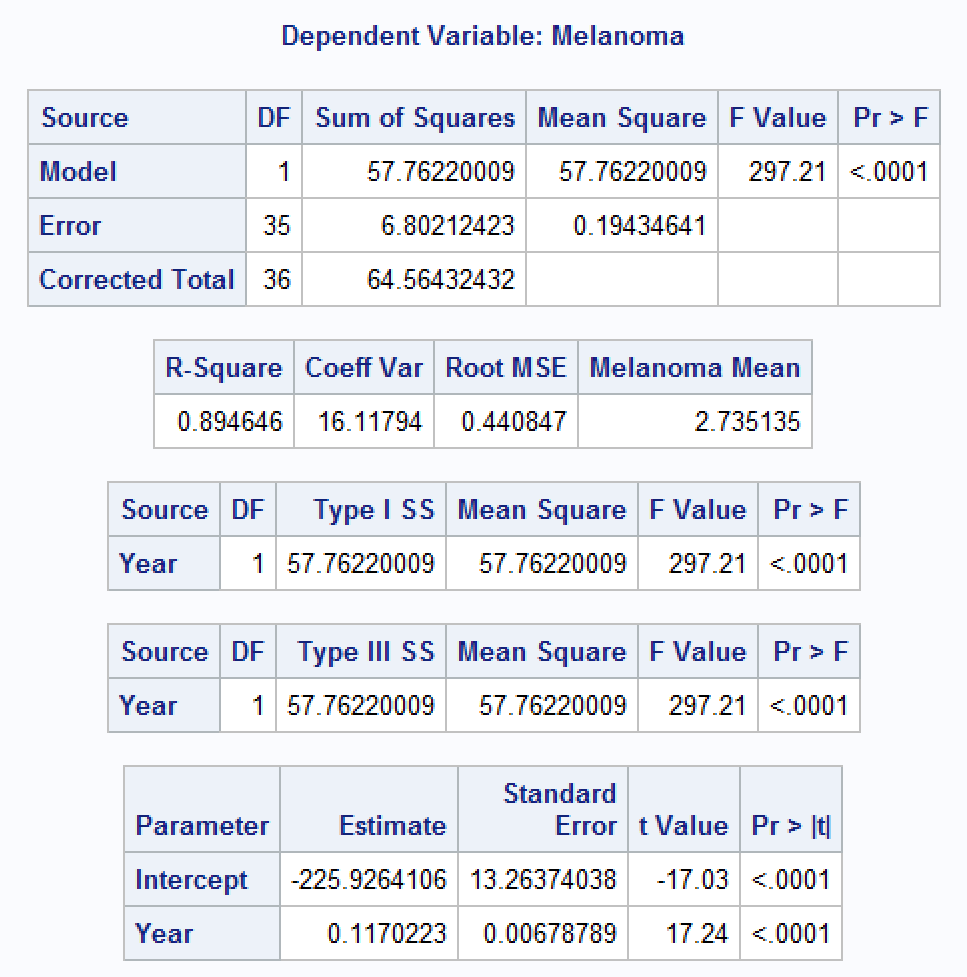




1. Use proc glm to fit a simple linear regression model: . Cut and paste the parameter estimate table and plot of the data with the regression line superimposed (default output from proc glm). Make sure that the plot has “year” on the x-axis. Note in your mind the estimates and standard errors of the intercept and slope (you don’t need to write anything down for this.) In addition, please answer this question in writing: “Does it appear that the series is going on extended “runs” in which the rate of melanomas tends to be above or below the mean for an extended amount of time (suggesting correlated residuals with time (serially correlated residuals))?

The runs appear to be within 1 – 3 years, except for the middle years. Those extend past 4 years and are not common.

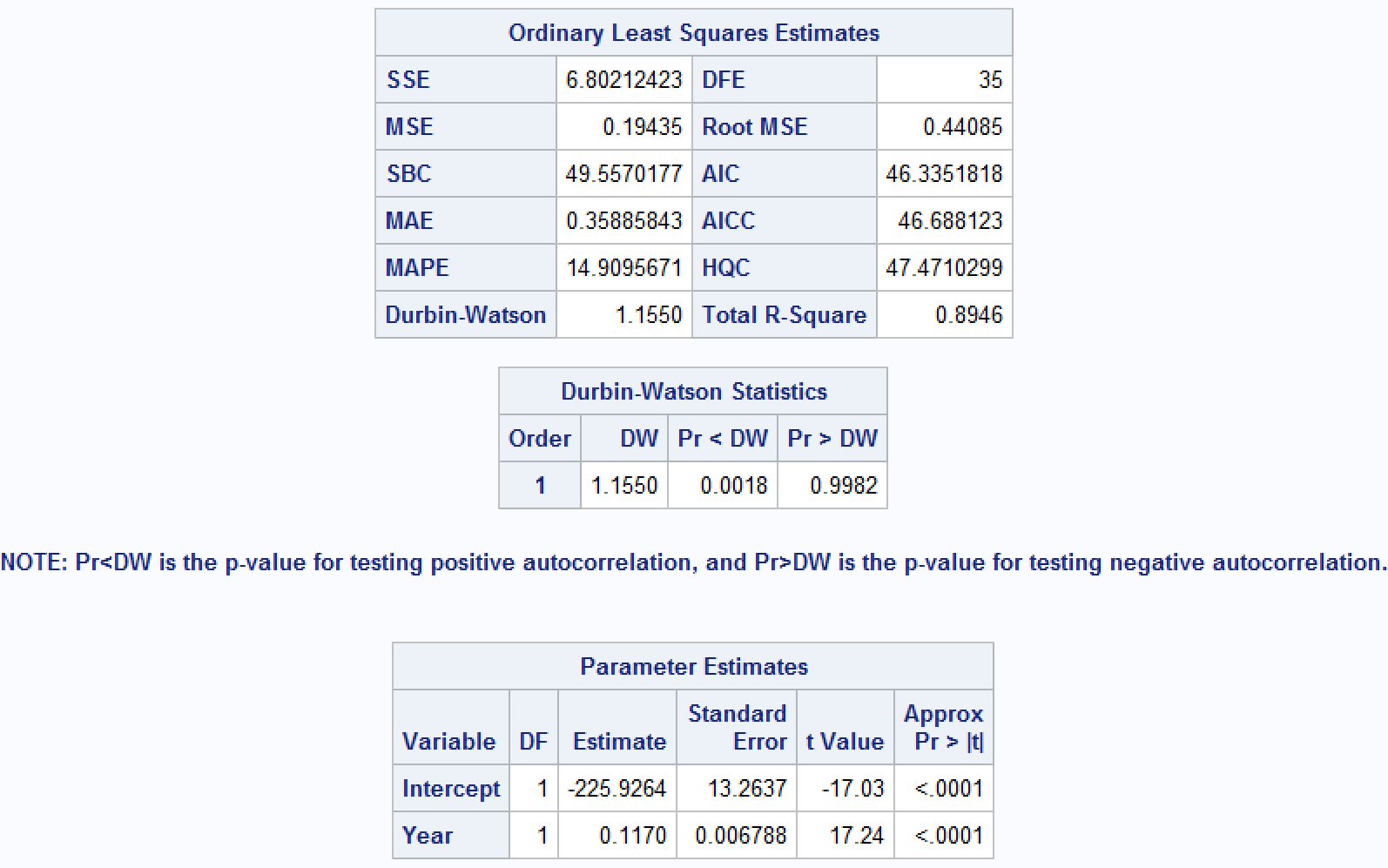
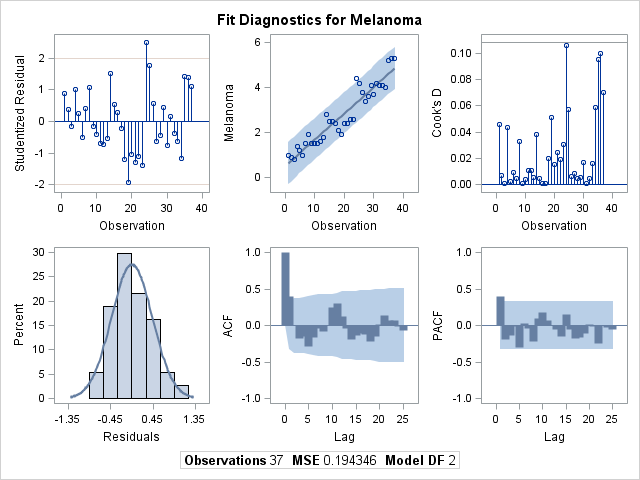






1. Now use proc autoreg to plot the same line as above (same model). Do not account for any serial correlation here. (a) Does the parameter estimate table look different than the one from the OLS model you found from proc glm above? (b) Make sure and use the “dwprob” option to get the Durbin-Watson test statistic. Does it suggest AR(1) serial correlation? Discuss in a sentence or two including the test statistic and pvalue. (c) Cut and paste the PACF and comment on any evidence it may provide into serial correlated residuals.

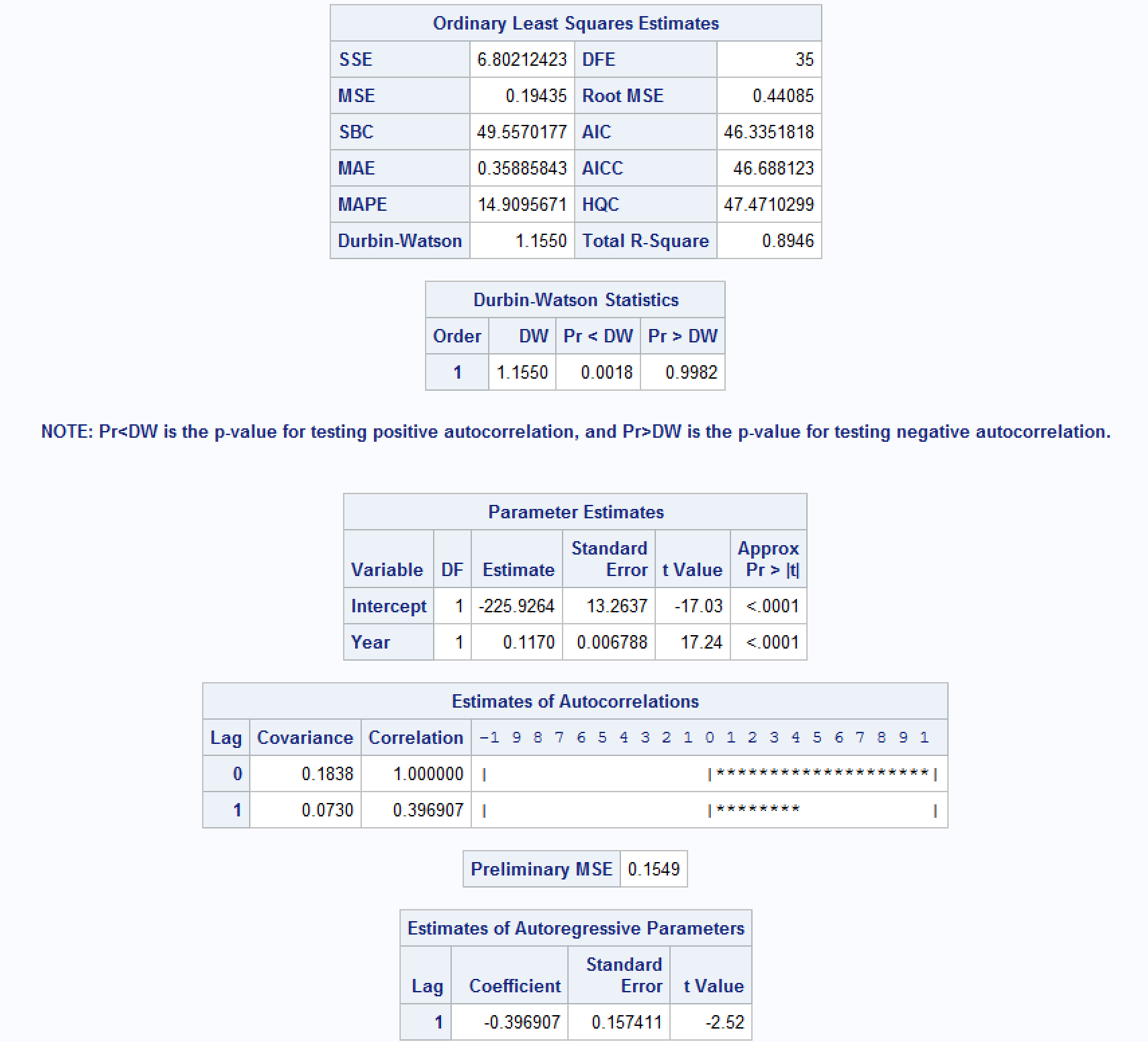
The parameter estimates produced by both procedures are identical to each other. The reason for this is because the models are unchanged. The DW statistic of 1.1550 and p-value 0.0018 suggest there is a positive autocorrelation between the data. The PACF indicate an AR(1) serial correlation is present in the model.

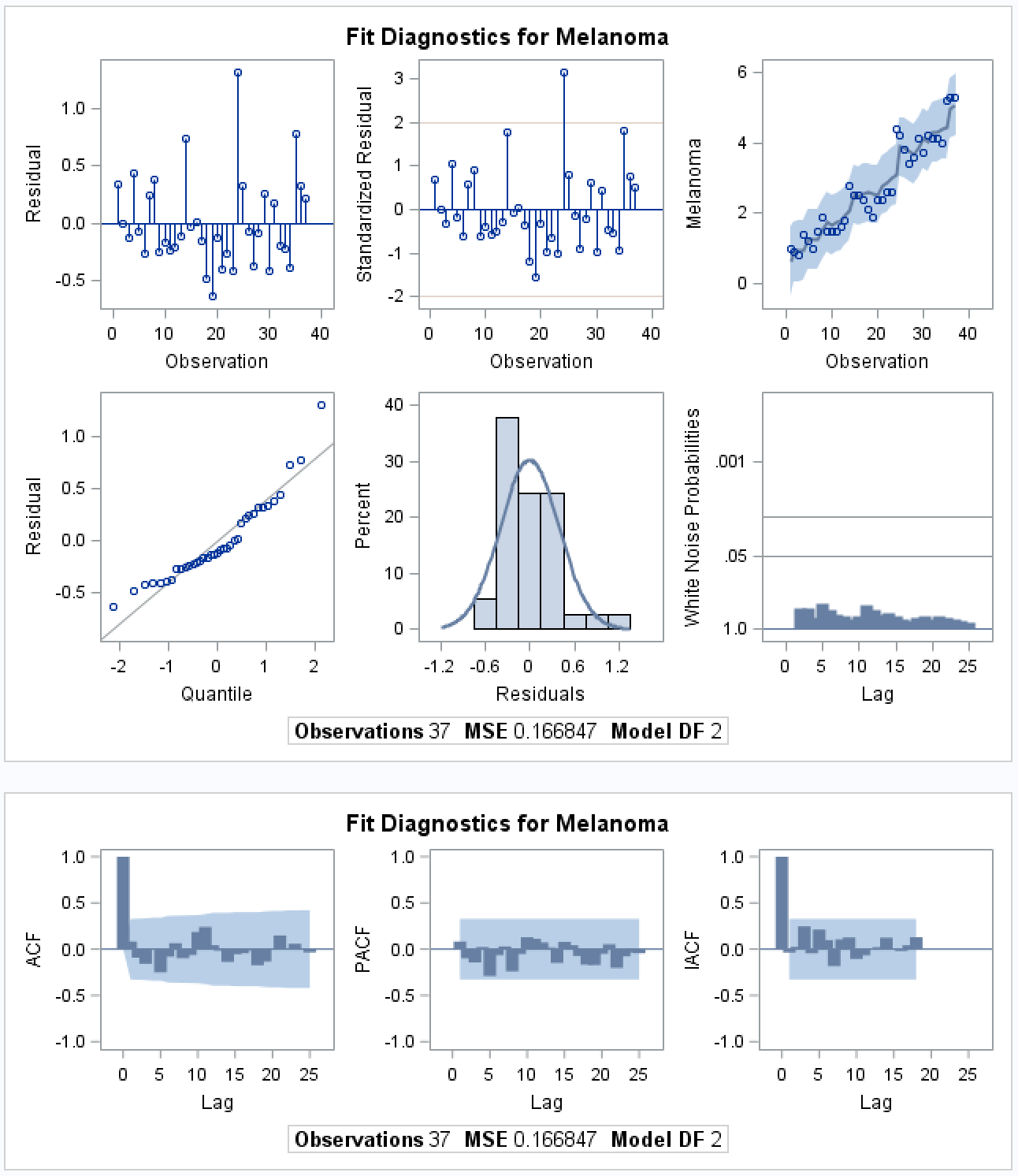
 

1. Now use proc autoreg to fit a model that accounts for an AR(1) correlation structure in the residuals. (a) Cut and paste the Yule-Walker parameter estimate table and compare and contrast in a few sentences the slope and intercept estimates and their standard errors with the OLS estimates from above. (b) Report the estimate of the first serial correlation coefficient. (c) Make sure and use the “dwprob” option again to get the Durbin-Watson test statistic. Does it suggest any remaining serial correlation in the series? Discuss in a sentence or two including the test statistic and pvalue. (d) Cut and paste the PACF and comment on any evidence it may provide into any remaining serial correlation. (e) Compare the MSE from the OLS and Yule Walker models. (f) Compare the AIC from the OLS and Yule Walker models. Which one does the AIC favor?

The model produced a Y-intercept that was lower than the OLS model and also carries a higher error coefficient. The first serial correlation coefficient is -0.396907. As previous mentioned, the Durbin-Watson statistic is 1.1550, which indicates there is a positive autocorrelation and that indication is supported by the p-value 0.0018. The MSE and AIC is lower for the Yule-Walker Estimates, which would favor the Yule-Walker model.

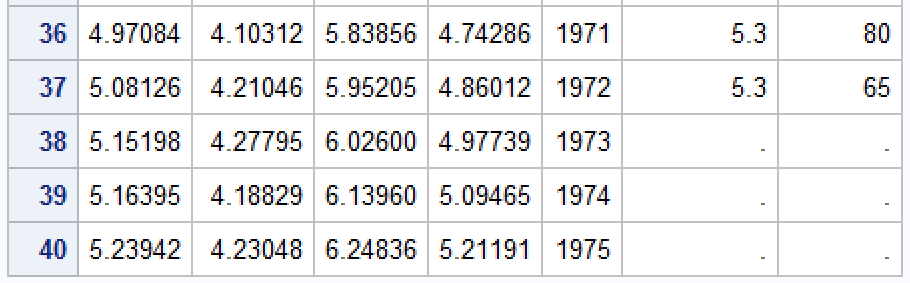


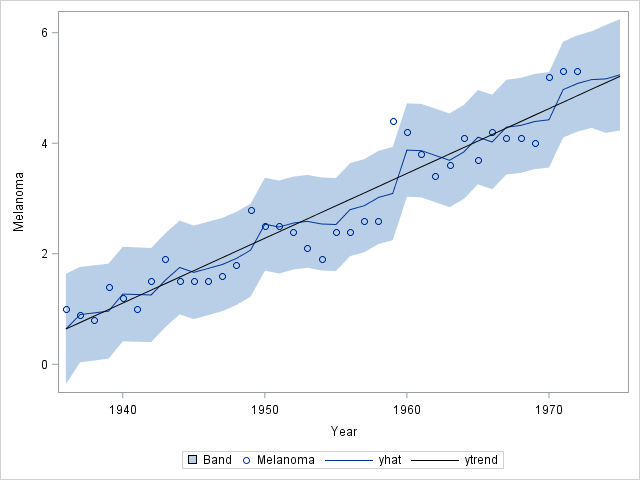




1. Use your model to predict the number of melanomas from 1973 - 1975. Include the prediction with 95% confidence limits as well as a plot of the series and the predictions.









**BONUS:**

(Up to 5 pts) Looking back at the series in question 1, scientists back in the 70’s surmised that melanomas may be related to sunspots. Is there any evidence that melanoma incidence is related to sunspot activity in the same year, or to sunspot activity in the previous one or even two years? Useful Hint: Check Example 1 of this function: <http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000212547.htm>