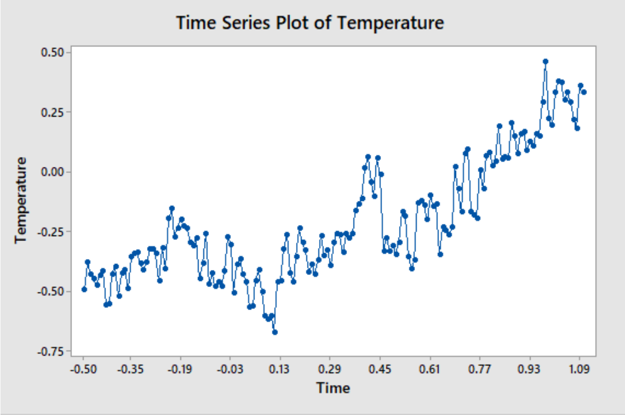
Graduate Project: Chapter 15

Nick Handelman
4/30/18

1 Problem 7

1.1 Visual Analysis



Graph 1.1 - Time Series Plot of Temperature

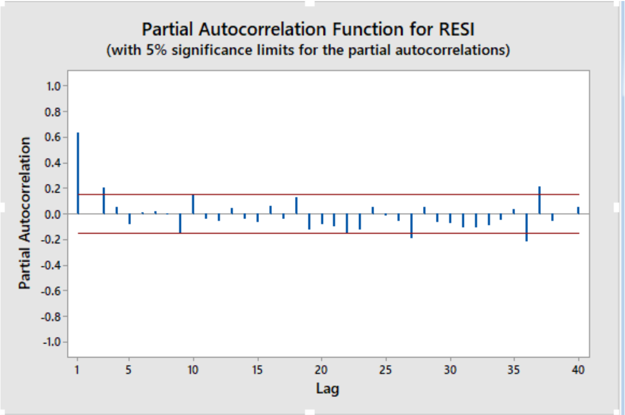
Graph 1.1 visually indicates that serial correlation may be present since temperature values at each time are typically close to those of their neighbors.

1.2 Unadjusted Regression

$Temperature = -0.4079 + 0.1522Time + 0.4794Time^2$ with $Time = \frac{Year-1900}{100}$
The p-values for the coefficients of $Time$ and $Time^2$ were both less than 0.001 This gives strong evidence that neither of the coefficients is 0.
I stored the residuals from this fit in a column to calculate the first serial correlation coefficient r_1.

1.3 Calculating the First Serial Correlation Coefficient r_1

Minitab has a built in partial autocorrelation function (PACF) that calculates $r_1 = 0.630478$ using the residuals from the regression fit.



Graph 1.2 - Problem 7 PACF

1.4 Is Serial Correlation Present?

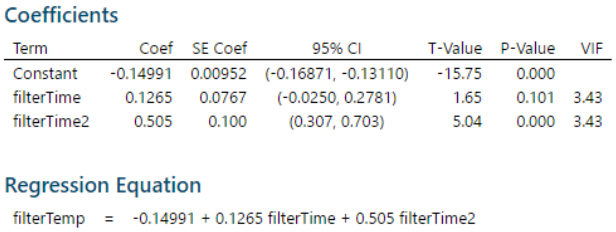
There are 161 residuals, so the ”Easy, Large Sample Test For Serial Correlation” described in section 15.4.1 is applicable.
The test statistic is $Z = r_1\sqrt{n} = 0.630478 * \sqrt{161} = 7.9998$ and the $Pr(|Z| > 7.9998)$ is virtually zero. Thus, there is very strong evidence that serial correlation is present.

1.5 Is the First Order Autoregression (AR(1)) Model Adequate?

Refer back to Graph 1.2. The red lines give the standard deviation significance limits. The only partial autocorrelation that is significantly beyond the limits is for lag 1 (i.e. r_1). For the other lags, the evidence that the partial autocorrelation isn't 0 is not strong. Thus, AR(1) is adequate.

1.6 Filtering and Adjusted Regression

Since the AR(1) model is adequate, filtering can be applied. I applied the equations $V_t = Y_t - r_1 Y_{t-1}$ and $U_{1t} = t - r_1(t - 0.01)$ and $U_{2t} = t^2 - r_1(t - 0.01)^2$ to adjust for the serial correlation. A new regression is run on the filtered data and the results are in Figure 1.3.



Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-0.14991	0.00952	(-0.16871, -0.13110)	-15.75	0.000	
filterTime	0.1265	0.0767	(-0.0250, 0.2781)	1.65	0.101	3.43
filterTime2	0.505	0.100	(0.307, 0.703)	5.04	0.000	3.43

Regression Equation

filterTemp = -0.14991 + 0.1265 filterTime + 0.505 filterTime2

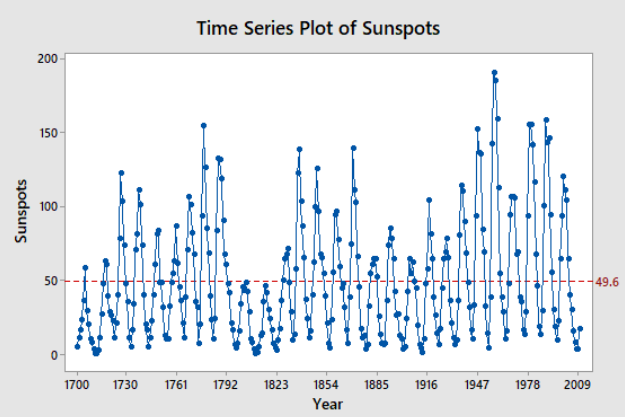
Figure 1.3 - Adjusted Regression

1.7 Conclusions

This is an observational study, so a causal relationship can't be claimed.

2 Problem 9

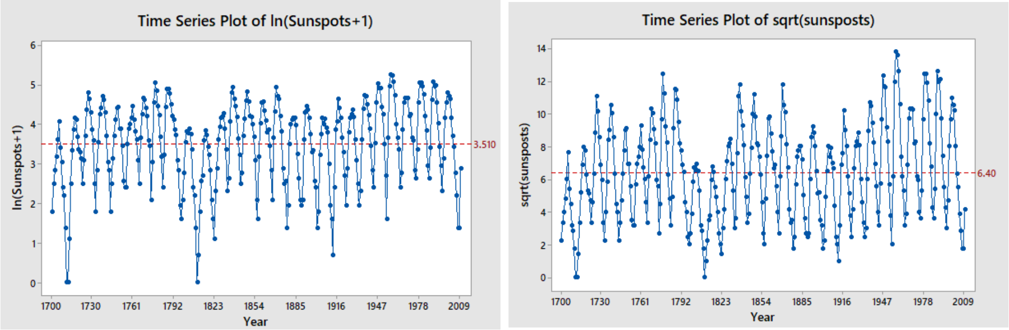
2.1 Construct a Time Series Plot



Graph 2.1 - Time Series Plot of Sunspots

The greatest number of sunspots is close to 200, while the smallest number of sunspots is 0. This discrepancy suggests a log transform (after adding 1 to each observation). The data are counts (the number of sunspots) which suggests a square root transform. Also, it appears that both the mean and variance are trending upward over time. This suggests that the time series is not stationary and a transformation would be useful.

2.2 Transformation



Graphs 2.2 - Time Series Plots of Transformed Sunspot Counts

Graphs 2.2 shows the time series plots of transformed sunspot counts. In both plots, the long-term trends in both mean and variance appear to be less pronounced than in the untransformed data. However, the square root plot appears more stationary since it appears to better fit the subjective criteria of ”turning the plot upside down and seeing if it looks the same”.

2.3 PACF of the Square Root Transformed Data

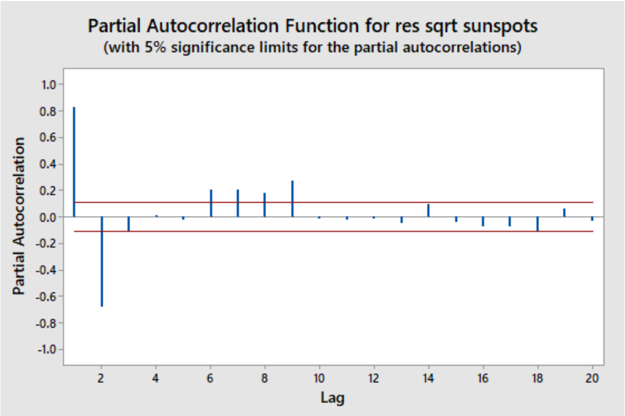


Figure 2.3 - PACF of the Square Root Transformed Data

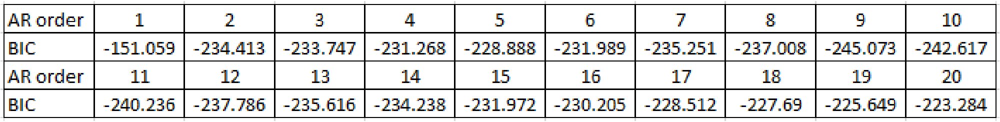
In figure 2.3, the red lines give the standard deviation significance limits. The partial autocorrelations that are significantly beyond the limits are for lags 1 and 2. For the other lags, the evidence that the partial autocorrelation isn't 0 is not strong. There also doesn't appear to be a pattern suggesting a moving average component. Thus, AR(2) is an adequate model.

2.4 BICs for the Square Root Transformed Data

I calculated the BIC for each AR(p) model through AR(20) according to the equation:

$$BIC = (p+1)log(311) + 311 * \sum_{j=1}^{p} log(1-\hat{\phi}_j^2) \tag{1}$$

p is the order of the AR model and $\hat{\phi}_j^2$ is the jth partial autocorrelation. There is 1 parameter (the mean) and 311 data points. The results are given in Table 2.4.

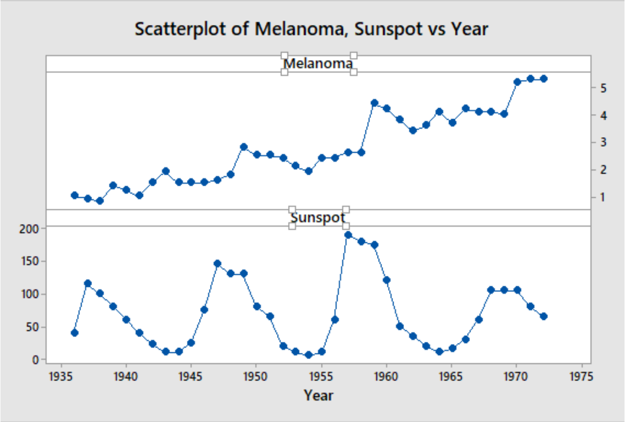


AR order	1	2	3	4	5	6	7	8	9	10
BIC	-151.059	-234.413	-233.747	-231.268	-228.888	-231.989	-235.251	-237.008	-245.073	-242.617
AR order	11	12	13	14	15	16	17	18	19	20
BIC	-240.236	-237.786	-235.616	-234.238	-231.972	-230.205	-228.512	-227.69	-225.649	-223.284

Table 2.4 - BICS of AR(p) Models Through p = 20

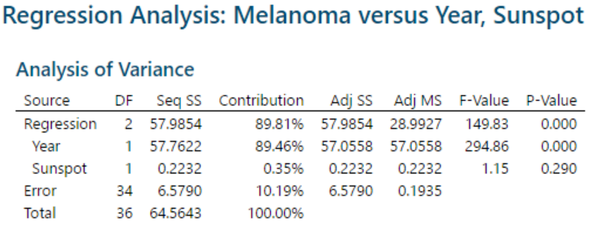
The results suggest that AR(9) gives the best fit for the data, since that model has the smallest BIC at -245.073. The AR(2) model is close with a BIC of -234.413.

3 Problem 14



Graph 3.1 - Time Series Plots of Sunspots vs. Year and Melanoma vs. Year

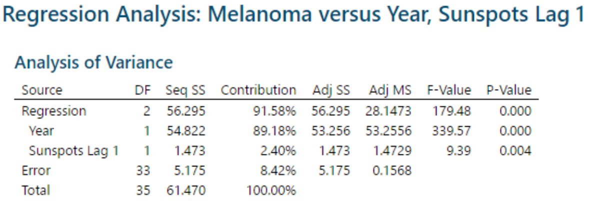
After a cursory glance, I see a pattern where the incidence of melanoma spikes about 1-3 years after a spike in the number of sunspots. Also, I see a pattern where a decrease in the number of sunspots is followed about 1-3 years later by a decrease or weak increase in the incidence of melanoma. The figures in 3.2 are Analysis of Variance tables of the regression of the incidence of melanoma on year and number of sunspots. In all cases, year has the most explanatory power. In the top left table, the sunspot p-value of 0.290 indicates there isn't enough evidence to claim that the incidence of melanoma has a significant relationship to the number of sunspots in that year (lag 0). In the top right and bottom left tables, the opposite is true for the number of sunspots in the previous year (lag 1) and 2 years previous (lag 2).



Regression Analysis: Melanoma versus Year, Sunspot

Analysis of Variance

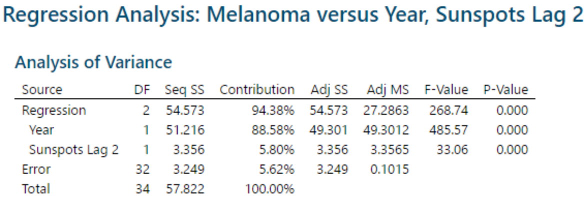
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	57.9854	89.81%	57.9854	28.9927	149.83	0.000
Year	1	57.7622	89.46%	57.0558	57.0558	294.86	0.000
Sunspot	1	0.2232	0.35%	0.2232	0.2232	1.15	0.290
Error	34	6.5790	10.19%	6.5790	0.1935		
Total	36	64.5643	100.00%				



Regression Analysis: Melanoma versus Year, Sunspots Lag 1

Analysis of Variance

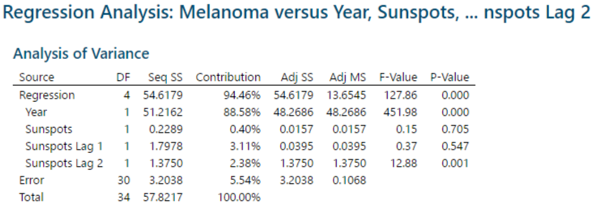
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	56.295	91.58%	56.295	28.1473	179.48	0.000
Year	1	54.822	89.18%	53.256	53.2556	339.57	0.000
Sunspots Lag 1	1	1.473	2.40%	1.473	1.4729	9.39	0.004
Error	33	5.175	8.42%	5.175	0.1568		
Total	35	61.470	100.00%				



Regression Analysis: Melanoma versus Year, Sunspots Lag 2

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	54.573	94.38%	54.573	27.2863	268.74	0.000
Year	1	51.216	88.58%	49.301	49.3012	485.57	0.000
Sunspots Lag 2	1	3.356	5.80%	3.356	3.3565	33.06	0.000
Error	32	3.249	5.62%	3.249	0.1015		
Total	34	57.822	100.00%				



Regression Analysis: Melanoma versus Year, Sunspots, ... nspots Lag 2

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	4	54.6179	94.46%	54.6179	13.6545	127.86	0.000
Year	1	51.2162	88.58%	48.2686	48.2686	451.98	0.000
Sunspots	1	0.2289	0.40%	0.0157	0.0157	0.15	0.705
Sunspots Lag 1	1	1.7978	3.11%	0.0395	0.0395	0.37	0.547
Sunspots Lag 2	1	1.3750	2.38%	1.3750	1.3750	12.88	0.001
Error	30	3.2038	5.54%	3.2038	0.1068		
Total	34	57.8217	100.00%				

Figures 3.2 - Melanoma vs. Year and Sunspots Lagged 0, 1 and 2 Years

In the bottom right table, the explanatory variables are year, number of sunspots lag 0, number of sunspots lag 1 and number of sunspots lag 2. The p-values for sunspots and sunspots lag 1 indicate the removal of these variables (one at a time) from the model does not have a significant effect on the new model's explanatory power. The removal of sunspots lag 2 does have a significant effect however since its p-value is 0.001. Therefore, this table suggests that the incidence of melanoma, after accounting for the year, is most strongly related to the number of sunspots 2 years previous.