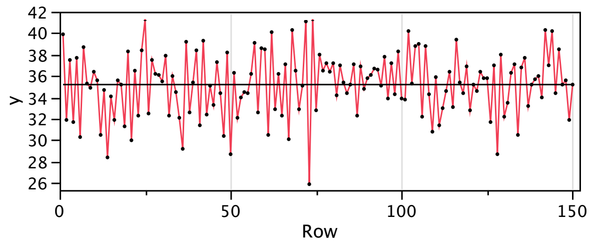
Insurance 260
Spring, 2009

Solutions, Assignment 4

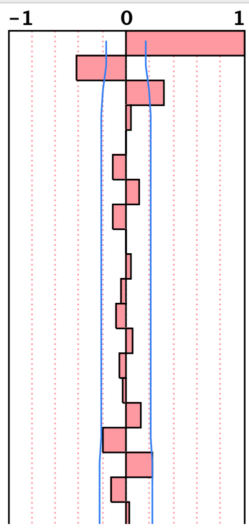
9.6 The following figures show plots of the data and the correlation functions. The sequence plot suggests a stationary process (no long-run drift). The SAC and SPAC both damp down quickly, also suggesting a stationary process.

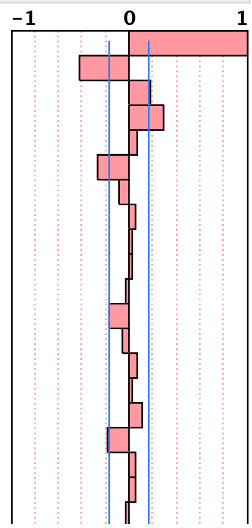


To get an initial guess of the order of the process, note that the SAC "cuts off" after two lags and the SPAC slowly damps down, suggesting a moving average process of order 2 (MA(2)). The couple of "large" correlations near lag 17 are probably the result of sampling fluctuations.

Time Series Basic Diagnostics

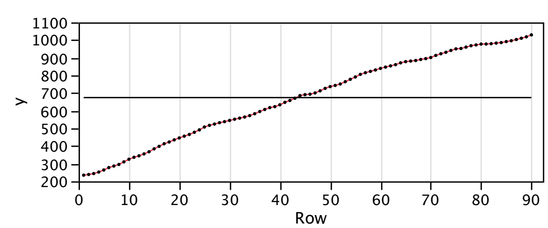
Lag	AutoCorr	Partial
0	1.0000	1.0000
1	-0.4151	-0.4151
2	0.3187	0.1769
3	0.0489	0.2884
4	0.0038	0.0747
5	-0.1142	-0.2672
6	0.1086	-0.0826
7	-0.1101	0.0499
8	0.0000	0.0301
9	0.0365	0.0346
10	-0.0423	-0.0260
11	-0.0831	-0.1644
12	0.0592	-0.0548
13	-0.0627	0.0679
14	-0.0332	0.0268
15	0.1241	0.1184
16	-0.1926	-0.1849
17	0.2212	0.0501
18	-0.1237	0.0542
19	0.0322	-0.0243





10.1 This and the next few questions refer to fitting an ARI(1,1) model to the toothpaste sales data from Table 9.7.

The sequence plot at the right shows that these sales are clearly non-stationary and require differencing. The output below summarizes the estimates from the JMP Time Series platform for the differences. The estimate of φ_1 is 0.65. Since $|\varphi_1| < 1$, the fitted model is stationary. (The estimate of δ obtained by JMP 3.0902 is slightly different from the estimate shown in the text 3.023 obtained by Minitab.



Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
AR1	1	0.6516255	0.0803564	8.11	<.0001*	3.09022114
Intercept	0	8.8704003	0.8187502	10.83	<.0001*	

Small numerical deviations such as these are typical in time series analysis because there is not a simple formula that all software uses to estimate the fitted model; rather, the estimates are obtained by iterating a series of approximations until "convergence" is obtained.)

10.2 The estimate of μ given by JMP is 8.87, labeled as the intercept in the summary of the fitted model. Notice that the estimate of μ is related to the estimate of the constant δ in the fitted model, $\mu = \delta/(1-\varphi_1)$.

10.3 The summary shows that both estimates are very significant, with p-values much less than 0.05. JMP makes you use the p-value for the estimate of μ to determine whether to retain δ in the fitted model. The relationship shown in 10.2, however, shows that testing H_0: $\delta=0$ is equivalent to testing H_0: $\mu = 0$. So long as the model is stationary ($|\varphi_1| < 1$) $\mu = 0$ if and only if $\delta = 0$.

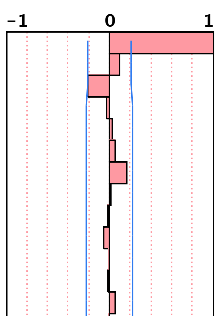
10.4 We need the autocorrelations of the residuals from this model. JMP shows these (and the partial autocorrelations) along with a sequence plot of the residuals. JMP also computes Q* or Ljung-Box statistic as part of the output and supplies the needed p-value. The formula is almost the same as that used for the simpler Box-Pierce statistic (which I expect you to know):

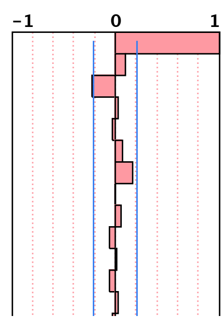
$$Q = (n\text{-}d)\, \Sigma\, r_j^2 = (90\text{-}1)(0.099^2 + 0.2067^2 + 0.0249^2 + 0.0213^2 + 0.0625^2 + 0.1664^2)$$
$$\approx 89(0.0851948) \approx 7.58$$

One then compares Q (or Q*) to the critical value of a chi-squared distribution with K (the number of squared autocorrelations) minus the number of estimated parameters (not counting the intercept δ) for the degrees of freedom. Hence use K-1 = 5. The 5% point in a chi-squared distribution with 5 degrees of freedom is 11.07. The p-value for the observed Q is 0.18.

These results are similar to the results obtained by JMP for the Ljung-Box test: it computes Q*(6) = 8.0601 with p-value 0.23.

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.0990	0.9015	0.3424	1	0.0990
2	-0.2067	4.8781	0.0872	2	-0.2186
3	-0.0249	4.9365	0.1765	3	0.0230
4	0.0213	4.9798	0.2894	4	-0.0250
5	0.0625	5.3569	0.3739	5	0.0659
6	0.1664	8.0601	0.2337	6	0.1610
7	0.0133	8.0776	0.3258	7	0.0006
8	-0.0166	8.1050	0.4233	8	0.0560
9	-0.0490	8.3477	0.4995	9	-0.0580
10	-0.0059	8.3513	0.5946	10	0.0070
11	-0.0159	8.3775	0.6791	11	-0.0606
12	0.0508	8.6491	0.7326	12	0.0335





10.5 With K=12, JMP obtains Q* = 8.6491 with p-value 0.7326. Clearly, the cumulative autocorrelation does not indicate that the model has left significant autocorrelation in the residuals.

10.6 To get the 99% intervals, set the level of confidence in the dialog that asks about whether to constrain the fit. Once you fit the ARI(1,1) model, save the results. JMP produces a data table with the forecasts and prediction intervals (rather than adding them to the original table). The following table shows the predictions. You can see that $\hat{y}_{91}$ = 1039.78 (a little larger than the value shown in the text, 1039.79) with 99% interval 1032.62 to 1046.93.

	Actual y	Row	Predicted y	Std Err Pred y	Residual y	Upper CL (0.99) y	Lower CL (0.99) y
85	991.35	85	990.81	2.78	0.54	997.97	983.66
86	996.291	86	998.04	2.78	-1.74	1005.19	990.88
87	1003.1	87	1002.60	2.78	0.50	1009.76	995.45
88	1010.32	88	1010.63	2.78	-0.31	1017.78	1003.47
89	1018.42	89	1018.11	2.78	0.31	1025.27	1010.96
90	1029.48	90	1026.79	2.78	2.69	1033.94	1019.63
91	•	91	1039.78	2.78	•	1046.93	1032.62
92	•	92	1049.58	5.36	•	1063.39	1035.76
93	•	93	1059.05	7.87	•	1079.34	1038.77
94	•	94	1068.32	10.23	•	1094.68	1041.96

(If you have 95% intervals, you can find 99% intervals or those of different coverage. Adjust the length of the interval by the ratio of the t-percentile at 95% to the t-percentile at the desired coverage.)

10.11 This exercise uses the sales of shampoo data from Table 9.10. (This is a short series, with n = 30 cases.) If you fit Model 1, an AR(2) model with a constant, using JMP (without constraints), you get these results:

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
AR1	1	0.32359	0.14264	2.27	0.0315*	38.1491554
AR2	2	0.57763	0.14914	3.87	0.0006*	
Intercept	0	386.18170	29.89687	12.92	<.0001*	

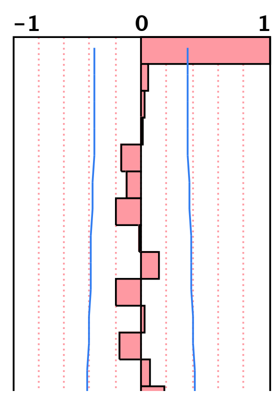
JMP does not find a problem with the fitting process. JMP estimates a rather different mean (386 compared to 332.6) but obtains similar estimates for φ_1 and φ_2. Several problems that are reasonable to notice include these: (the question asks for 2)

(1) The shown output (p 486) indicates substantial collinearity between the two estimates (see correlations of the parameter estimates; the correlation of the two estimates is -0.934). That's a problem, though not surprising since these are lags of the same time series.

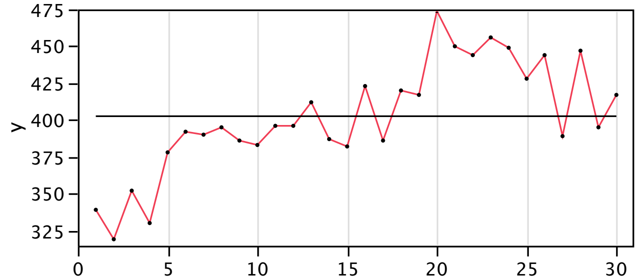
(2) There's also a fairly large residual autocorrelation at lag 6 (-0.19). It's larger than the others, but this output shows that its not significant using Q*. Note the wide interval on the residual autocorrelations caused by the short length of the time series.

(3) I'd add the fact that the sequence plot of the data indicate that there's likely a trend. The model that's been fit assumes that the process is stationary. That assumption seems at odds with the data.

Lag	AutoCorr	Ljung-Box Q	p-Value
0	1.0000	.	.
1	0.0553	0.1012	0.7504
2	0.0239	0.1208	0.9414
3	0.0107	0.1249	0.9887
4	-0.1520	0.9785	0.9130
5	-0.1147	1.4835	0.9150
6	-0.1906	2.9369	0.8167
7	-0.0112	2.9421	0.8903
8	0.1346	3.7324	0.8804
9	-0.1921	5.4202	0.7962
10	0.0214	5.4422	0.8598
11	-0.1720	6.9375	0.8041
12	0.0675	7.1808	0.8454



10.12 This question concerns the fit of an ARI(1,1) model (without an intercept) to the sales of shampoo data. The following output shows the JMP analysis of this model. First note that both the Bayes and Akaike criteria prefer the nonstationary ARI(1,1) model (So does a common-sense look at the sequence plot of the data ... it's got a strong upward trend.)



Model Comparison

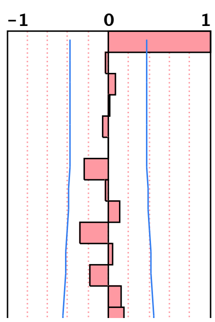
Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	AIC Rank	SBC Rank
AR(2) No Constrain	27	533.63209	278.06274	282.26633	0.605	272.06274	2	2
ARI(1, 1) No Intercept No Constrain	28	511.67676	264.67795	266.04525	0.610	262.67795	1	1

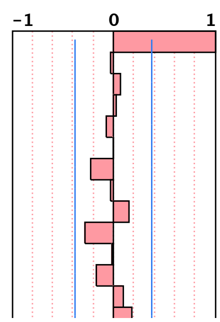
Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1	1	-0.6292289	0.1400727	-4.49	0.0001*

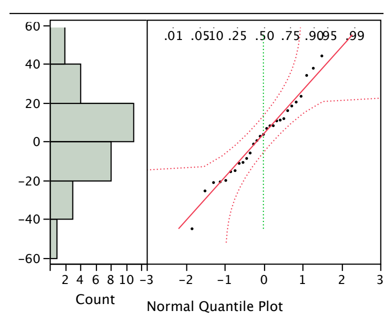
The estimated parameter is highly statistically significant. Continuing to the residuals, neither residual correlation function indicates significant remaining autocorrelation. (Again, with such a short series, there's a lot of sampling fluctuation in these estimates.

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.0237	0.0180	0.8932	1	-0.0237
2	0.0749	0.2047	0.9027	2	0.0744
3	0.0175	0.2153	0.9751	3	0.0210
4	-0.0595	0.3426	0.9869	4	-0.0646
5	0.0027	0.3428	0.9968	5	-0.0030
6	-0.2311	2.4305	0.8762	6	-0.2243
7	-0.0237	2.4534	0.9306	7	-0.0336
8	0.1136	3.0063	0.9340	8	0.1506
9	-0.2785	6.4927	0.6898	9	-0.2818
10	0.0411	6.5727	0.7651	10	-0.0071
11	-0.1821	8.2283	0.6927	11	-0.1647
12	0.1210	9.0019	0.7028	12	0.0920



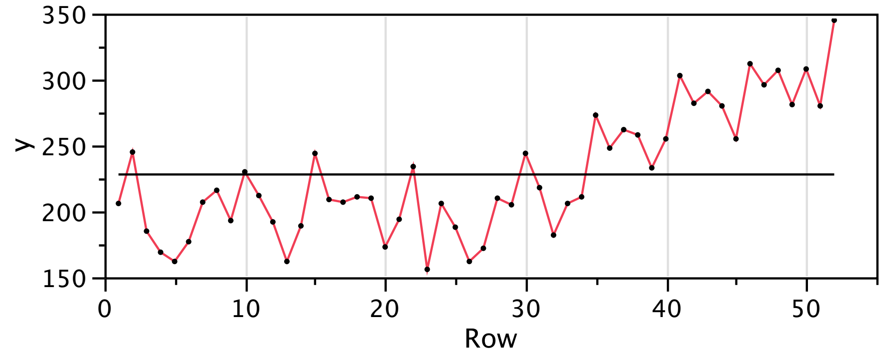


The normal quantile plot of the 29 residuals also seems OK: the residuals could be a sample from a normal population. Hence, there's no problem indicated in these results. About all one might do is consider fitting several other models and using the selection criteria to compare them. With such a short series (and no indicated problems) that's not likely to be useful

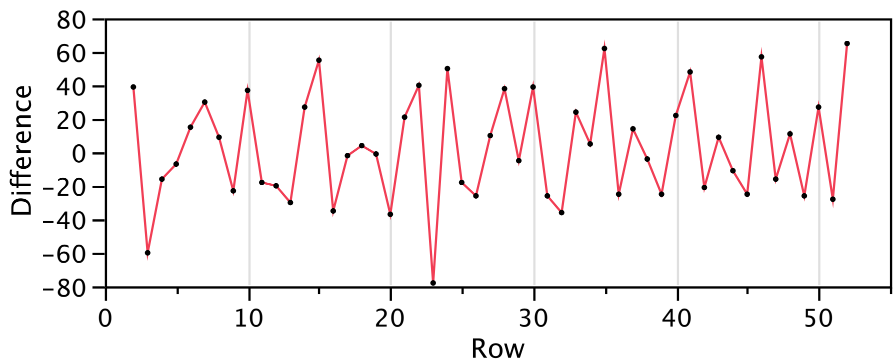


10.13 This exercise concerns data from Table 8-1 on sales of thermostats. This exercise points out that ARIMA models offer an alternative to exponential smoothing. In particular, exponential smoothing such as the trend-corrected method discussed in this exercise are special cases of ARIMA models.

(a) Here are the modeling results. Start with the sequence plot. The data seem stationary initially, then began to trend upward (not unlike recent global temperatures). This trend suggests that we need to difference the data.

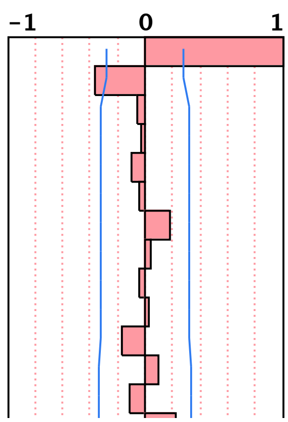


The plot of the differences shows that the differenced data are stationary.

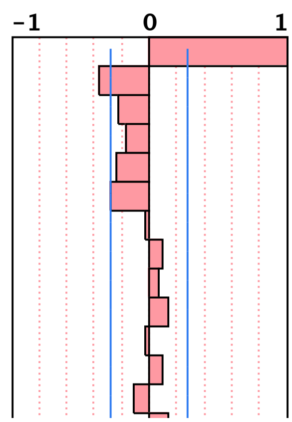


The SAC and SPAC suggest a moving average of order 1 (SAC cuts off after 1, whereas SPAC drift downward). So, try a few like this and see what happens.

Lag	AutoCorr
0	1.0000
1	-0.3651
2	-0.0592
3	-0.0336
4	-0.0933
5	-0.0413
6	0.1860
7	0.0482
8	-0.0386
9	0.0345
10	-0.1698
11	0.1027
12	-0.1155



Lag	Partial
0	1.0000
1	-0.3651
2	-0.2221
3	-0.1713
4	-0.2372
5	-0.2732
6	-0.0219
7	0.0935
8	0.0722
9	0.1430
10	-0.0331
11	0.1047
12	-0.1103



AIC prefers the ARIMA(2,1,2) model (all of the φs and θs are statistically significant).

BIC prefers the more parsimonious IMA(1,1) model. (We could remove constants from both.) Let's see how the residuals of this simpler model appear: Does it omit much autocorrelation?

Model Comparison

Model	DF	AIC	SBC	RSquare	-2LogLH	AIC Rank	SBC Rank
ARIMA(1, 1, 1) No Constrain	48	488.78737	494.58284	0.650	482.78737	4	2
ARIMA(2, 1, 2) No Constrain	46	485.04739	494.70652	0.695	475.04739	1	3
ARIMA(3, 1, 3) No Constrain	44	488.27267	501.79545	0.700	474.27267	3	4
IMA(1, 1) No Constrain	49	487.32839	491.19205	0.646	483.32839	2	1

The sequence plot of the residuals does not have a pattern, and the residual correlations are not statistically significant. We can use the IMA(1,1) model.

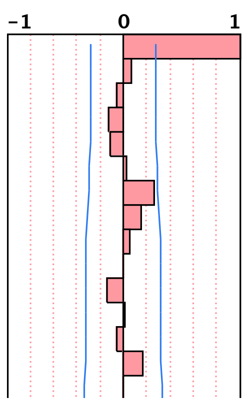
Parameter Estimates

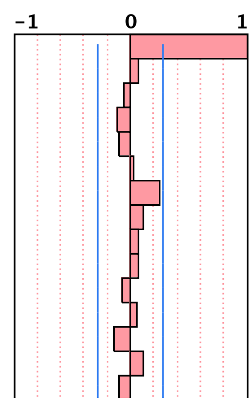
Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
AR1	1	0.998011	0.131844	7.57	<.0001*	0.91829602
AR2	2	-0.437270	0.141035	-3.10	0.0033*	
MA1	1	1.796515	0.267886	6.71	<.0001*	
MA2	2	-1.000000	0.296162	-3.38	0.0015*	
Intercept	0	2.090556	1.531908	1.36	0.1790	

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
MA1	1	0.7059458	0.100257	7.04	<.0001*	2.24654944
Intercept	0	2.2465494	1.188734	1.89	0.0647	

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.0713	0.2749	0.6001	1	0.0713
2	-0.0575	0.4573	0.7956	2	-0.0629
3	-0.1248	1.3341	0.7210	3	-0.1170
4	-0.1147	2.0902	0.7192	4	-0.1032
5	0.0306	2.1451	0.8287	5	0.0317
6	0.2712	6.5641	0.3630	6	0.2497
7	0.1580	8.0971	0.3241	7	0.1185
8	0.0570	8.3011	0.4046	8	0.0706
9	0.0010	8.3012	0.5041	9	0.0732
10	-0.1409	9.6094	0.4754	10	-0.0687
11	0.0145	9.6236	0.5645	11	0.0507
12	-0.0504	9.7994	0.6336	12	-0.1343
13	0.1747	11.9715	0.5300	13	0.1122
14	-0.0035	11.9724	0.6085	14	-0.0998





(b)

Why use ARIMA models rather than exponential smoothing: flexibility. We are not so confined to a few special cases, but can select from a broader collection of models. There's a benefit in simple models: fewer choices, but automatic selection criteria take the pain out of that.