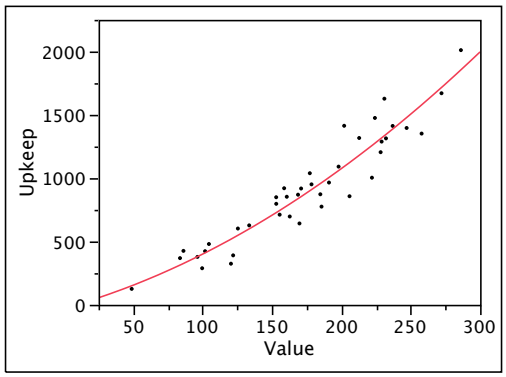
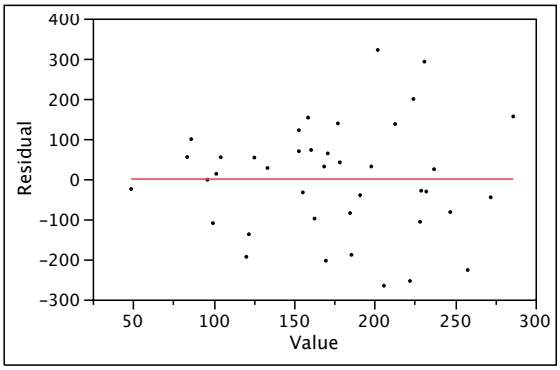
Insurance 260
Spring, 2009

Solutions, Assignment 3

5.13 The initial regression models the dependence of upkeep expenses (in dollars) relative to the value of a home (in thousands of dollars). As background, a quadratic regression has the following fit

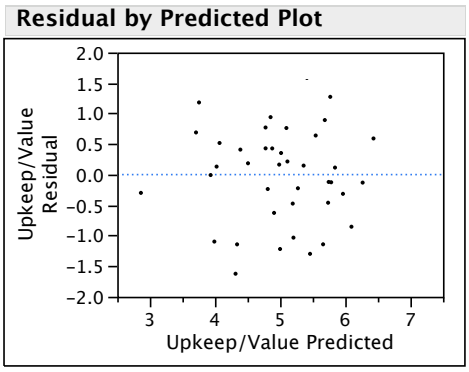


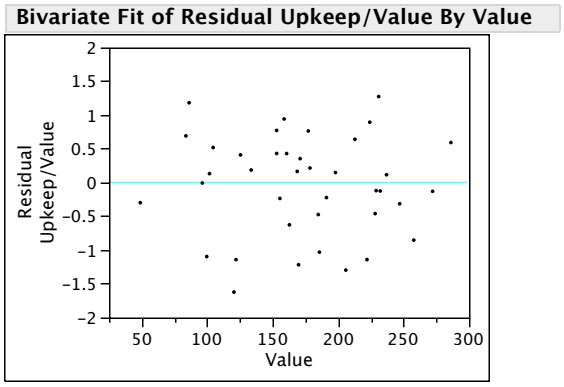


The residuals shown on the right appear to become more variable as the value of the home increases. As a remedy, the text suggests dividing the equation of the model by the value of the homes, regressing

(upkeep/value) on (1/value), a constant, and a linear term in value

The fit of this model (using multiple regression) follows, along with plots of residuals versus predicted values and residuals versus value (like the plot on the right above). The residuals seem





to have relatively constant variance in both plots (though there may be a bit of curvature, a slight "u-shaped bend" in the plot or residuals on value – but that's likely imagination).

(a) The residual variance appears more consistent after the transformation than before. In particular, the residual variance does not increase systematically as value increases.

(b) To find the 95% prediction interval, start by finding the interval on the transformed scale. Let's call that the "expense ratio" (dollars of expense per $1,000 dollars of value). The predicted expense ratio for a $250,000 home is

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.4089246	1.32082	2.58	0.0140*
1/Value	−53.50053	83.19955	−0.64	0.5242
Value	0.0112235	0.004627	2.43	0.0203*

$$\text{Estimated Ratio} \approx 3.409 - 53.5/\text{Value} + 0.0112\,\text{Value}$$
$$= 3.409 - 53.5/250 + 0.0112 * 250 = 5.995$$

The 95% prediction interval for the expense ratio is thus about 6 ± 2(RMSE = 0.8). From here on, I will let JMP do the rest of the calculations. Allowing for extrapolation effects, the 95% prediction interval is 4.3275 to 7.6741. At $250,000, that works out to

4.3275*250 to 7.6741*250 = $1,081 to $1,919

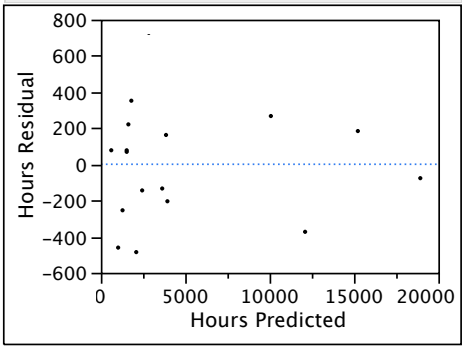
in expenses. A rather wide range, with the upper bound about twice the lower bound.

5.16 This exercise uses the hospital data considered in Assignment 2. The model now includes a dummy variable for the large hospitals noted in the previous analysis. The summary of the model shows a very large R2 and all of the individual slope estimates are statistically significant.

Summary of Fit

RSquare	0.996789
RSquare Adj	0.995718
Root Mean Square Error	363.8542
Mean of Response	4978.48
Observations (or Sum Wgts)	17

Residual by Predicted Plot



Indicator Function Parameterization

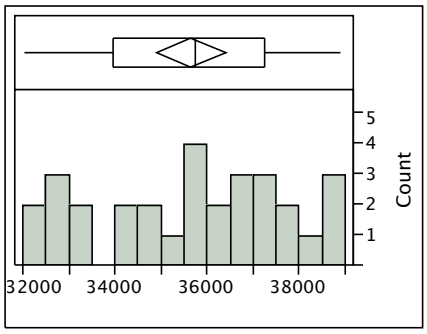
Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	2462.2164	501.9897	12.00	4.90	0.0004*
Xray	0.04816	0.01193	12.00	4.04	0.0016*
BedDays	0.7843175	0.07331	12.00	10.70	<.0001*
Length	−432.4095	93.35426	12.00	−4.63	0.0006*
Size[large]	2871.7828	573.0618	12.00	5.01	0.0003*

(a) The coefficient of the dummy variable implies that large hospitals (as defined by this dummy variable) require about 2872 more hours of labor compared to smaller hospitals at a given level of effort or demand for services (as measured by the other factors: Xrays, BedDays, and Length of stay). The effect is statistically significant, with the confidence interval for the estimated difference (about 2872 ± 2(573)) far from zero. The estimated effect, for instance, lies t=5.01 standard errors above zero.

(b) Hospital 14 used 10,343.81 hours of labor. The estimated regression assigns fitted value (plug into the regression equation) 10,077 hours to this hospital. The difference (the residual) at this hospital indicates that this hospital used 266.8 more hours than we'd expect for a large hospital. Since the RMSE of the model is 364 hours, this residual is not unusual, lying less than one SD from the fit. This hospital is not unusually inefficient for a large hospital under these conditions.

(An aside: This is not such a great way to use residuals because this hospital affects the regression fit – it is one of the few "large" hospitals that determine the slope for the dummy variable. It would have been better to fit the model without this case and then compare the prediction to the actual value. A big outlier in a regression pulls the regression toward itself, reducing the size of its residual. So-called "Studentized residuals" adjust for this effect. In this case, for example, the studentized residual for this hospital is 1.35, larger than 1, even though the y value lies within one RMSE of the fit.)

6.1 Here is the histogram and summary of the lumber data. Since the model fits a constant, we only need the information in this display to answer the text questions.



Moments	
Mean	35651.867
Std Dev	2037.3599
Std Err Mean	371.96933
upper 95% Mean	36412.629
lower 95% Mean	34891.104
N	30
Sum Wgt	30
Sum	1069556
Variance	4150835.5
Skewness	-0.234385
Kurtosis	-1.011446
CV	5.7145954
N Missing	0

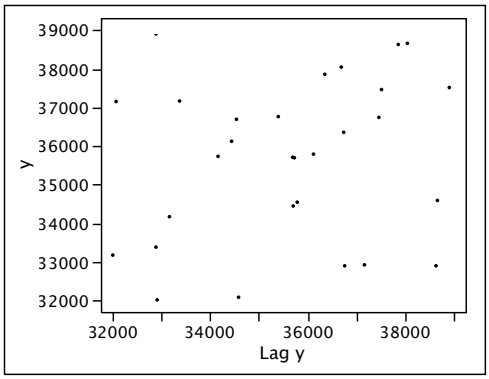
(a) There does not appear to be a trend in the plot, so we are not far off in treating the data as a sample from a single population. (We need to see the lag-plot in part "c" to check for dependence.)

(b) The forecast is the mean. The rough 95% prediction interval is then the mean ± two SDs of the data around the mean, as though predicting a random draw from the distribution shown in the histogram, or 35562 ± 2 (2037). (Don't use the confidence interval for μ, however. You can see from the diamond in the plot that this range is far too narrow when it comes to describing the variation in the data itself.)

A more precise answer (it was okay for grading to stop at the previous interval) takes account of estimating μ. In particular, we should use a t-statistic with 29 deg. freedom and scale up the SD by a factor of sqrt(1+1/30). These do not make a huge difference (a few percent), even with a relatively small sample:

Rough interval:	35562 ± 2*2037	= 31,488 to 39,636
"Precise" interval	35562 ± 2.045*2037*sqrt(1+1/30)	= 31,327 to 39,797

(c) We need the scatterplot of y_t on y_{t-1} to see the autocorrelation. The plot (next page) shows that there's nothing going on, no autocorrelation to be found.

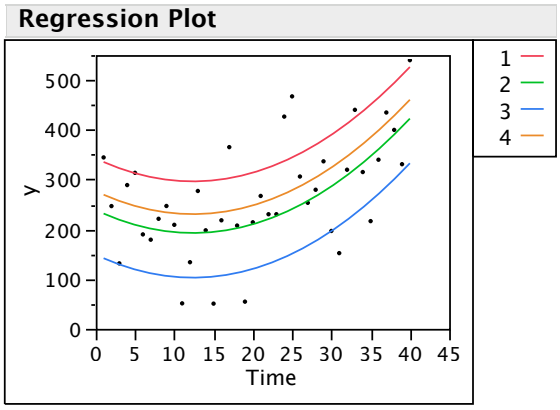


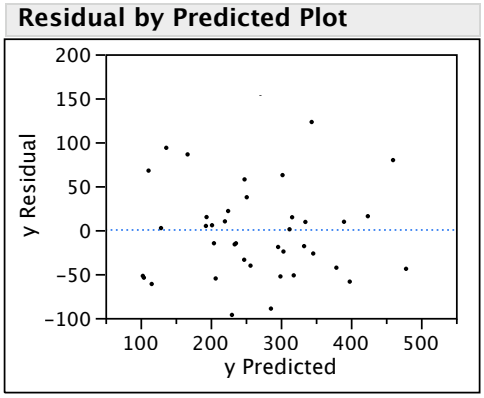
6.4 Energy costs of a school, in $100s.

(a) Quarterly energy demand is clearly seasonal. The sequence plot also shows a trend that appears to bend gradually. We cannot tell whether there's also autocorrelation until we fit the model and inspect the residuals.

(b) The variation shown in Figure 6.34 of the text appears steady over this time period. Logs do not appear needed. Again, we'll know more when we fit the regression and inspect residuals.

(c) The following summarize the JMP analysis of this model. Pretty wild plot of the trends.





Summary of Fit

RSquare	0.744324
RSquare Adj	0.706725
Root Mean Square Error	60.47257
Mean of Response	265.5458
Observations (or Sum Wgts)	40

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	276.63631	35.0485	34.00	7.89	<.0001*
Time	-7.458255	3.396031	34.00	-2.20	0.0350*
Time*Time	0.301231	0.080304	34.00	3.75	0.0007*
Quarter[1]	65.770648	27.15916	34.00	2.42	0.0209*
Quarter[2]	-37.87011	27.0958	34.00	-1.40	0.1713
Quarter[3]	-127.6113	27.05743	34.00	-4.72	<.0001*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	361966.91	72393.4	19.7962
Error	34	124335.67	3656.9	Prob > F
C. Total	39	486302.58		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Quarter	3	3	195194.21	17.7922	<.0001*
Time	1	1	17637.94	4.8232	0.0350*
Time*Time	1	1	51456.18	14.0709	0.0007*

i) The first dummy variable repeats 1,0,0,0, 1,0,0,0,... and so forth. The second runs 0,1,0,0, 0,1,0,0. The third begins 0,0,1,0, 0,0,1,0...

ii) The overall model is statistically significant. The overall F = 19.8 reported in the Anova summary has p-value much less than 0.05. Since we entered the quarters as a bundle (Q1,Q2,Q3), we should test them that way as well using the partial F test. JMP provides the partial F for Quarter,. The partial F for Quarter (see the effect test output) gives F = 17.79 with p-value < 0.0001. For time, we should also bundle Time and Time^2 as one and find the partial F. This is useful to avoid the nasty effects of collinearity. It is more common, however, to look to see whether the separate t-statistics for Time and time^2 are statistically significant (they are, though the linear component is close to 0.05 due to collinarity).
[To obtain the partial-F for time, remove both from the model and fit a regression on just Quarter. The R^2 from that fit is 0.3787. Hence, the F-statistic is
F = (0.7443 - 0.3787)/(1-0.7443) * (40-6)/2 = 24.3
That's big – much bigger than the cutoff at about 3.3 (for F with 2 and 34 d.f.).]

iii) For periods 41 and 42, the predictions are (in hundreds of dollars)
$\hat{y}_{41}$ =276.6 - 7.458*41+0.301*41*41+65.77≈ 542.573
$\hat{y}_{42}$ =276.6 - 7.458*42+0.301*42*42-37.87 ≈ 456.458

iv) Using JMP, the model gives the following predictions (2nd column) and 95% prediction intervals (3rd and 4th columns) for the next four quarters (you can see the effects of

41	542.987813	400.943821	685.031806
42	456.890977	312.328277	601.453676
43	385.29614	237.907614	532.684667
44	531.656304	381.152128	682.160479

rounding in iii). These intervals are considerably wider than those produced by $\hat{y}$±2 RMSE (because of the effects of extrapolation). For example, for $\hat{y}_{44}$, the "naive" interval is
531.656 ± 2 * 60.473 = 410.71 to 652.602
(That's shorter by about 60*$100.)

v) The Durbin-Watson statistic indicates autocorrelation of about 0.55, which is statistically significant.

Durbin-Watson

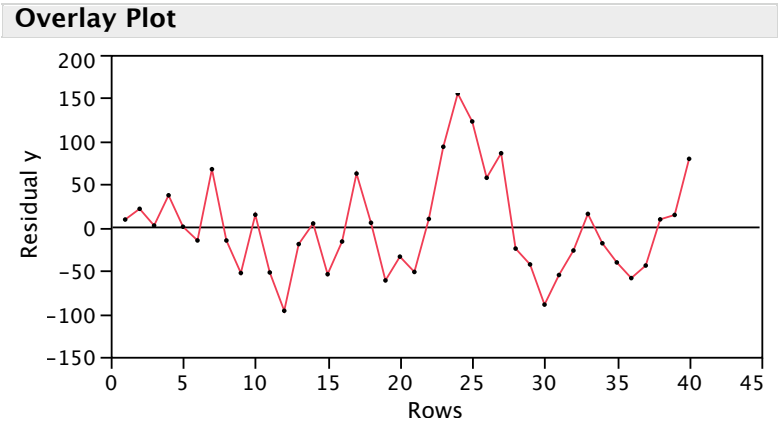
Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
0.8396892	40	0.5544	<.0001*

(d) The SAS output (shown in the text) estimates the autoregression coefficient φ and adjusts the rest of the estimates for the presence of this term.

i) The estimate is 0.594 and is significant since its p-value (0.0003) is less than 0.05.

ii) Yes, though the term for Q2 indicates that Q2 and Q4 are no longer significantly different. The significance of the time trend is also reduced with the presence of the lagged explanatory variable.

iii) The predictions are given in the output from SAS and you can simply read them off. For example, $\hat{y}_{41} = 605$ with prediction interval 507 to 704 (quite a bit higher than the ordinary regression). To see why the predictions are higher, just plot the residuals in time order. The last one is about 100 above the fit...



It is useful (*though not part of the assigned exercise*) to compare the reported SAS output to the results obtained by our simple "add the residuals to the fit" procedure. The following summarizes the fit with the residuals added to the model. The coefficient of the lagged residuals is basically the estimate of φ reported in the text (0.59408). The regression estimate is 0.5945 with similar t-statistic as well. The RMSE is similar to the reported fit of the model by SAS.

Summary of Fit

RSquare	0.826371
RSquare Adj	0.793815
Root Mean Square Error	51.02982
Mean of Response	263.5241
Observations (or Sum Wgts)	39

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	278.55175	31.8896	8.73	<.0001*
Time	-8.178034	3.142054	-2.60	0.0139*
Time*Time	0.3238279	0.072881	4.44	<.0001*
Quarter[1]	70.831157	23.58715	3.00	0.0052*
Quarter[2]	-37.41152	22.86777	-1.64	0.1116
Quarter[3]	-127.3594	22.833	-5.58	<.0001*
Lag Residuals	0.5945033	0.150023	3.96	0.0004*

As for predictions, the first prediction is $\hat{y}_{41} = 605.7$. To get the rest (via JMP), the extend the column of residuals with estimates using φ^j times the last residual:

$$79.5 * (0.594, 0.592^2, 0.594^3) = (47.2, 28.1, 16.7)$$

After you fill these in (shaded yellow below in the excerpt of the data table), JMP computes the predictions as $\hat{y}_{42} = 497$, $\hat{y}_{43} = 415$, and $\hat{y}_{44} = 556$. The reported SAS predictions are 506, 427, and 570. Before you think these are large differences, you've got to take into account the accuracy of these estimates. To build the prediction intervals, get JMP to compute the SE of individual predictions. Then add a column to take account of extrapolating the residuals. The extrapolation effect is the cumulative sum $1 + \varphi^2 + \varphi^4$ etc that we have seen previously. These terms are (in the column labeled "Extrapolate Residual")

time = 41 factor = 1
time = 42 factor = 1 + φ^2 = 1.353
time = 42 factor = 1 + φ^2 + φ^4 = 1.478
time = 42 factor = 1 + φ^2 + φ^4 + φ^6 = 1.522

The t-statistic is $t_{.025,32}$ = 2.037 (about 2). The prediction intervals formed as

$$\hat{y}_{41} \pm 2.037 * \text{sqrt(factor)} * \text{SE(indiv)}$$

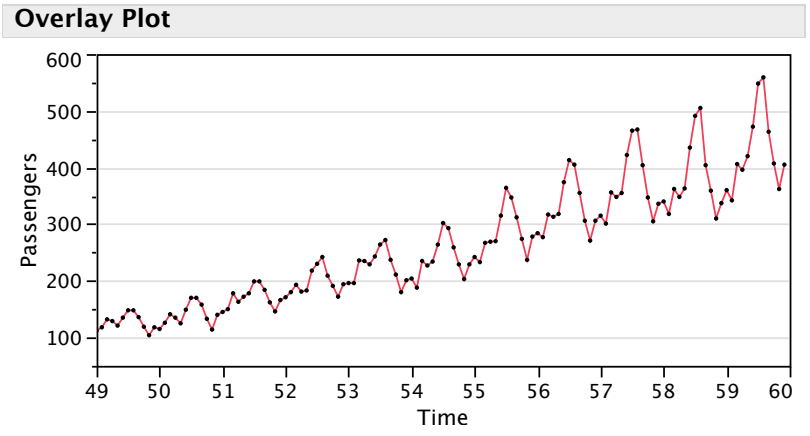
are shown in the excerpt of the spreadsheet below. (The intervals in the text from SAS seem a bit too short, as if not adjusted fully for the effects of extrapolation. In particular, the length of the first interval from SAS is about 200, whereas the first interval from the regression has length closer to 250.)

Time	Residual Energy Cost	Lag Residuals	Pred Formula Energy Cost	StdErr Indiv Energy Cost	Extrapolate Residual	Lower PI	Upper PI
35	-40.574	-18.369	250.730	54.703	•	•	•
36	-58.755	-40.574	379.702	55.094	•	•	•
37	-44.177	-58.755	455.186	56.376	•	•	•
38	9.330	-44.177	371.719	56.441	•	•	•
39	14.475	9.330	330.338	56.948	•	•	•
40	79.504	14.475	478.160	57.716	•	•	•
41	47.200	79.504	605.704	61.834	1.000	479.753	731.655
42	28.100	47.200	496.956	61.044	1.353	352.322	641.589
43	16.700	28.100	415.000	61.911	1.478	261.687	568.314
44	•	16.700	555.577	63.137	1.522	396.916	714.238

5 International Air Traffic

There are a number of ways to model these data. I'll follow the approach we have used in several examples. (A nice alternative considers month-to-month percentage changes.)

(a) The time plot shows an upward trend, a strong seasonal pattern, and increasing variation around the level. Looks like a log will be needed.



(b) The following output summarizes the fit of log passengers on time, with seasonal dummies and a quadratic trend. The model also has an adjustment for autocorrelation, using the lag of the residuals. (You may or may not have used the quadratic component of the trend, but your model definitely needs the time trend,

Summary of Fit

RSquare	0.993709
RSquare Adj	0.992949
Root Mean Square Error	0.034621
Mean of Response	5.492399
Observations (or Sum Wgts)	131

seasonal terms, and adjustment for autocorrelation.)

(i) The overall model is statistically significant, as shown by the F statistic in the Anova summary.

(ii) The two trend components and the lag residuals are all statistically significant, with very large t-stats and small p-values. The seasonal terms collectively are significant as shown by the effect test (partial F test).

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	21.961240	1.56866	1308.697
Error	116	0.139043	0.00120	Prob > F
C. Total	130	22.100282		<.0001*

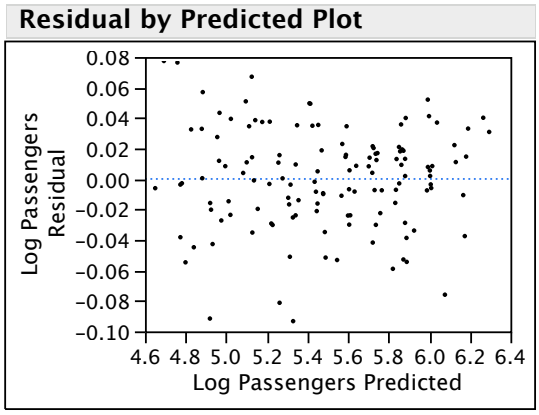
Effect Tests

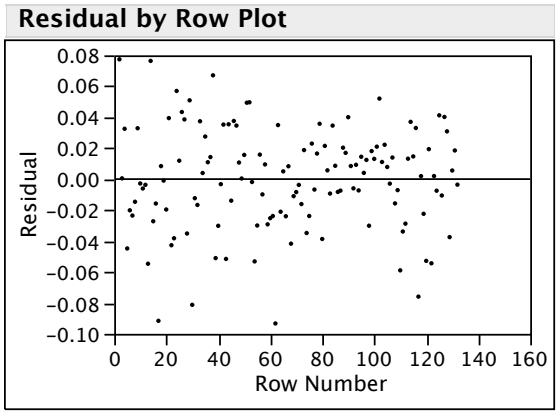
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Time	1	1	0.2095490	174.8220	<.0001*
Time*Time	1	1	0.1174699	98.0025	<.0001*
Month	11	11	2.0438706	155.0141	<.0001*
Lag Log Residuals	1	1	0.1342325	111.9871	<.0001*

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	−11.35005	1.01113	116.00	−11.23	<.0001*
Time	0.4917635	0.037193	116.00	13.22	<.0001*
Time*Time	−0.003377	0.000341	116.00	−9.90	<.0001*
Month[Jan]	0.014015	0.015147	116.00	0.93	0.3567
Month[Feb]	0.0015006	0.014784	116.00	0.10	0.9193
Month[Mar]	0.1377443	0.01478	116.00	9.32	<.0001*
Month[Apr]	0.0954608	0.014777	116.00	6.46	<.0001*
Month[May]	0.0912879	0.014773	116.00	6.18	<.0001*
Month[Jun]	0.2137546	0.014771	116.00	14.47	<.0001*
Month[Jul]	0.3141016	0.014768	116.00	21.27	<.0001*
Month[Aug]	0.3070308	0.014766	116.00	20.79	<.0001*
Month[Sep]	0.1660306	0.014765	116.00	11.25	<.0001*
Month[Oct]	0.024932	0.014764	116.00	1.69	0.0940
Month[Nov]	−0.115806	0.014763	116.00	−7.84	<.0001*
Lag Log Residuals	0.6962059	0.065789	116.00	10.58	<.0001*

(c) Here are the overall residual plots from this model.

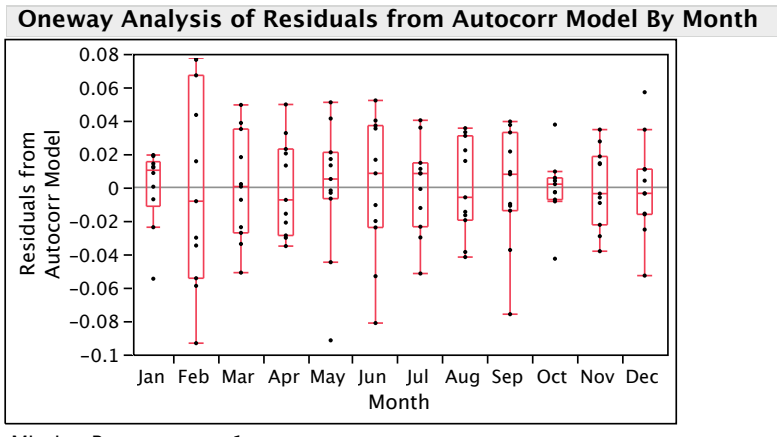




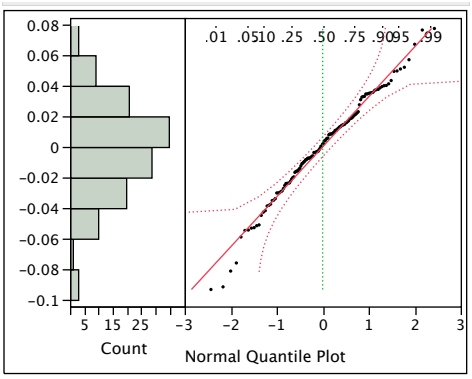
(i)

The DW test has p-value near 0.7; there's no further dependence of this type after adding the lag residual term. We can check further by plotting the residuals on their lag. Nothing particularly interesting shows up. This model captures the dependence of adjacent residuals. (There could be other residual patterns, such as from year to year.)

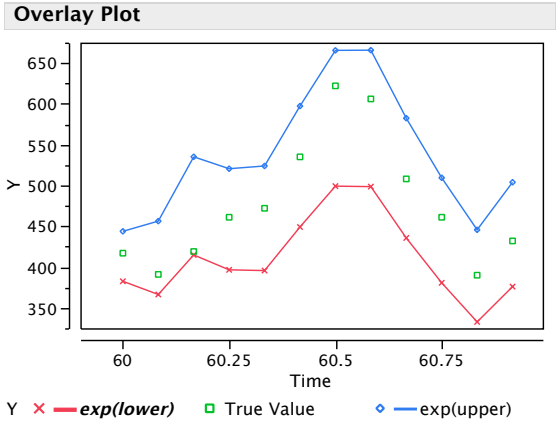
(ii) The assumption of equal variance seems questionable. The residuals appear to have less variation from around row 80 to row 110 (around 1955 through 1957), suggesting a period of less variation in demand. The variation by month may also change, as shown in these comparison boxplots. Generally, October through January have less residual variation than others, but it's hard to tell with only 11 years of data.



(iii) For normality, we use the normal quantile plot, which looks fine for these residuals.



(d) For predictions, extend the residuals into the forecast period so that JMP can do the calculations. The estimated autocorrelation (φ) is 0.696, so we have to multiply the last residual (0.112) by this value raised to powers. These values are shown in the accompanying data table shown below. This plot shows the prediction intervals and actual data. All of the intervals cover the values in 1960, though you might argue that they should after seeing how wide the intervals are.



(e) This is a tough one to solve completely, and JMP does not give enough information for all of the calculations. For the prediction, you have to exponentiate and then add. That gives an estimate of about 5,636 thousand. The actual total is 5,714. That's not too hard. The hard part is getting a prediction interval. It's hard because

(a) Our variance estimates are on the log scale, not the count scale.

(b) The prediction errors are not independent since we have used the same model for all

These two make it difficult to get an interval for the total by "analytic" means. These difficulties have led to simulation based estimates obtained by repeating the modeling over-and-over to see how the results change from sample to sample. A possible choice would be to use the sum of

	Time	Residual Log Passengers	Lag Log Residuals	Pred Formula Log	Extrapolate Factor	StdErr Indiv Log	Lower PI	Upper PI	exp(lower)	True Value	exp(upper)
133	60	0.007795	0.011192	6.021	1.000	0.037	5.947	6.095	382.565	417	443.769
134	60.08333	0.005420	0.007795	6.013	1.484	0.037	5.904	6.123	366.388	391	456.290
135	60.16667	0.003776	0.005420	6.155	1.719	0.037	6.028	6.282	414.826	419	535.085
136	60.25	0.002628	0.003776	6.119	1.833	0.037	5.983	6.255	396.573	461	520.484
137	60.33333	0.001829	0.002628	6.121	1.888	0.038	5.981	6.261	395.670	472	523.842
138	60.41667	0.001273	0.001829	6.250	1.914	0.038	6.107	6.392	449.090	535	597.271
139	60.5	0.000886	0.001273	6.357	1.927	0.038	6.213	6.501	499.134	622	665.507
140	60.58333	0.000617	0.000886	6.356	1.934	0.038	6.212	6.501	498.530	606	665.738
141	60.66667	0.000429	0.000617	6.222	1.937	0.038	6.077	6.367	435.626	508	582.378
142	60.75	0.000299	0.000429	6.088	1.938	0.038	5.942	6.233	380.666	461	509.358
143	60.83333	0.000208	0.000299	5.953	1.939	0.038	5.807	6.099	332.773	390	445.630
144	60.91667	•	0.000208	6.076	1.939	0.038	5.930	6.222	375.981	432	503.879

the lower bounds and the sum of the upper bounds. That's going to be very wide, not in line with how close the total of the predictions comes to the actual total.