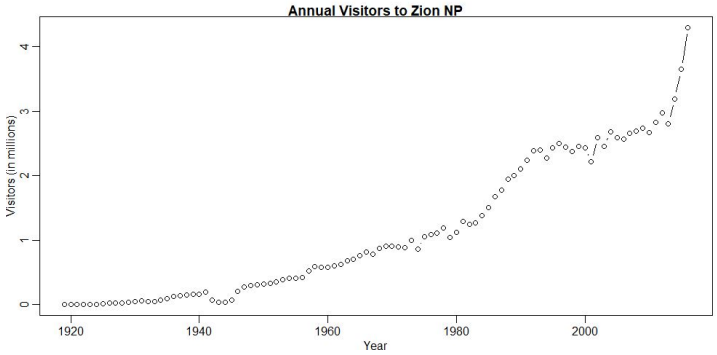
Ramifications of Autocorrelation on Prediction Performance in Linear Regression

Spencer Ebert

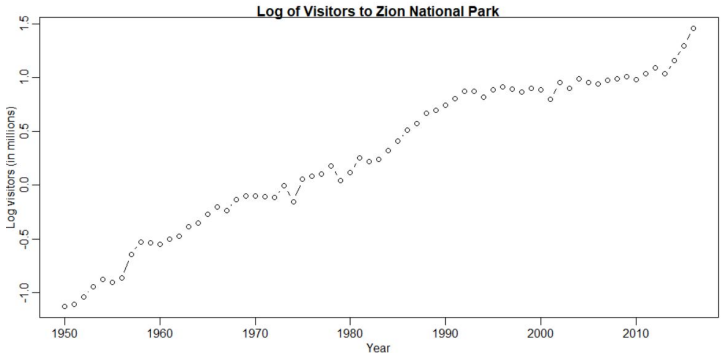
December 13, 2018

1 Motivating Example

Prediction is an important aspect of statistics and can help influence decision-making in important areas. In my Stat 330 class, we looked at annual visitors to Zion National Park and wanted to predict visitors for the upcoming years. Predicting the number of visitors for upcoming years would be important to prepare resources for the visitors like transport buses, number of employees, and trail preparation. Financial decisions would also be based off of the predicted annual visitors to the park. Here is the data for visitors to Zion National Park.



To predict future visitors we looked at the years from 1950 on and took the log of visitors because it appeared to be growing at an exponential rate. We then fit an ARIMA (1,1,1) model to the data.

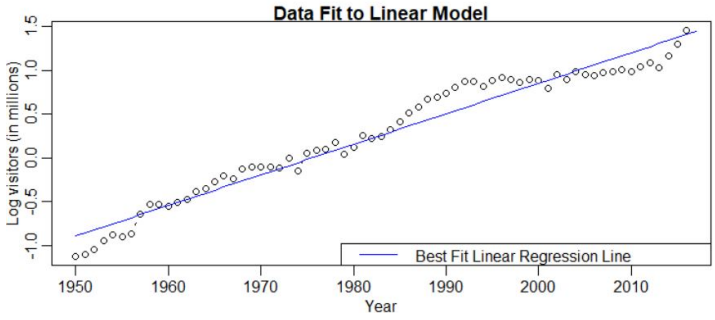


I was curious how fitting a simple linear model to the Zion data would affect prediction performance, and more generally how fitting a linear model to time series data would affect prediction interval performance. I fit the Zion data to a simple linear model where

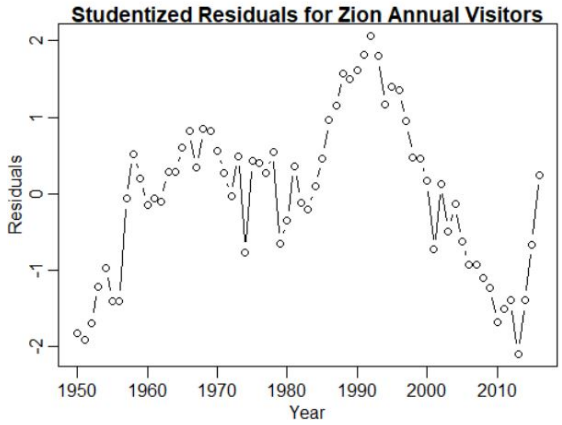
$$log(Visitors) = \beta_0 + \beta_1 Year + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \tag{1}$$

One of the assumptions for linear models is no autocorrelation among the response variable. In the case for visitors to Zion, visitors for each year are strongly correlated with the previous year. If I fit a linear model to the data then I would be overlooking the correlation from year to year. This results in only the linear trend and variance of the data to influence the prediction intervals while completely ignoring the autocorrelation.

To diagnose the problem, the first thing to look at is the graph of the data and a plot of the residuals and see if it seems to be autocorrelated. For the Zion visitor data, it seems as though each year is influenced by the last, because of the trend seen in my plot.



For the residuals, it seems that they follow a particular pattern rather than being independent.



The second thing to diagnose an autocorrelation problem is to perform the Durbin-Watson test. When performing the Durbin-Watson test of autocorrelation on Zion visitors data I got

```
lag Autocorrelation D-W Statistic p-value
  1         0.85481     0.2426811       0
Alternative hypothesis: rho != 0
```

The p-value is less than 0.001 so there is significant autocorrelation from year to year. This further emphasizes the findings that Zion annual visitors are highly correlated.

If I move forward ignoring the autocorrelation and fit a linear model to the data, I hypothesize that prediction interval coverage would be off.

2 Simulation Study

In my simulation study, I wanted to see how often a prediction interval for a linear model contained the true value of the next data point given the data was autocorrelated. I also wanted to examine how the strength of the correlation, and size of the data affects the prediction interval coverage.

To examine the prediction interval coverage I followed these 6 steps.

1. Specify how many simulations to run (N), the size of the data (n), and the correlation between each iteration (ρ)

2. Simulate data from an AR(1) model with $n + 1$ data points
   Note: We want to have one extra data point to see how often it falls in the prediction interval
   Equation for the AR(1) Model where y is the simulated data and t is the index

$$y_{t+1} = \rho y_t + \epsilon_{t+1}, \quad -1 < \rho < 1 \tag{2}$$

3. Take the first n data points (excluding y_{n+1}) and fit a linear model where
   $y = \beta_0 + \beta_1 Index + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$
   $Index$ is the number of the data point (1,2,3,...,n)

4. Calculate the 95% prediction interval for $Index = n + 1$ by using the linear model

5. Return whether y_{n+1} was in the calculated prediction interval

6. Repeat steps 2-5 N times and return the proportion of times that the prediction interval contained the true data point

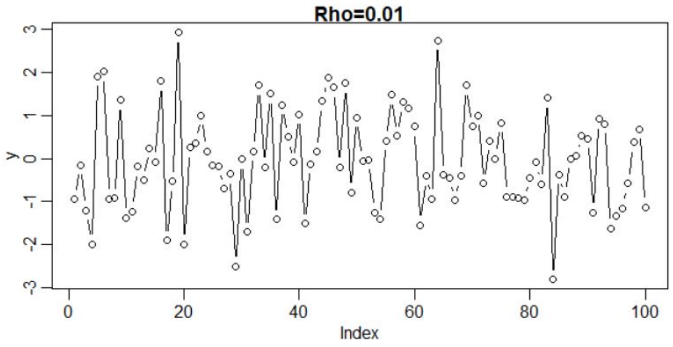
To compare how the correlation affects the prediction interval coverage, I chose 5 values for ρ: 0.01, 0.8, 0.9, 0.99, and 0.9999. I chose $\rho = 0.01$ to be the control for my experiment, because it is basically no correlation and the function simulating AR(1) data doesn't take in 0 values. Having $\rho = 0.01$ should result in about a 95% success rate in prediction performance, because the autocorrelation isn't violated for the linear model. I then chose my other values to see how the prediction changes as ρ gets close to 1. For the sample sizes (n) I chose 50, 100, 200, and 500 so that I could see if success rate increased as the sample size got larger.

After running the simulation 10,000 times for each set of ρ and n, I got the prediction interval coverage in this table.

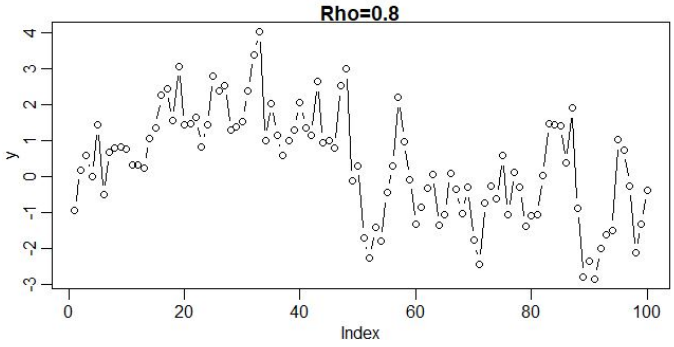
95% Prediction Interval Coverage					
n \ ρ	0.01	0.80	0.90	0.99	0.9999
50	0.9536	0.9157	0.8858	0.8403	0.8418
100	0.9538	0.9276	0.9092	0.8462	0.8344
200	0.9483	0.9424	0.9290	0.8463	0.8255
500	0.9488	0.9446	0.9405	0.8721	0.8328

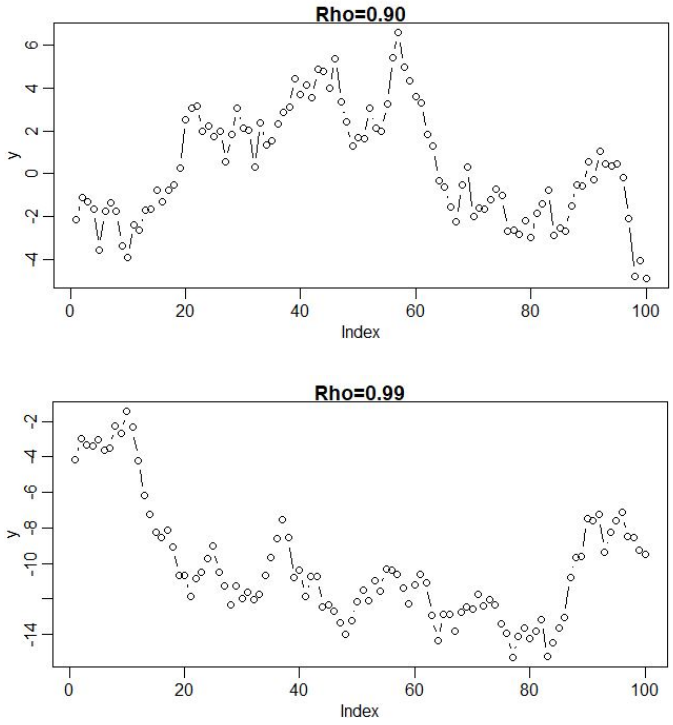
It's easy to see that when $\rho = 0.01$ that the linear model 95% prediction interval does a good job predicting the next value with a proportion around 0.95 where it should be. This comes as no surprise because the data isn't violating the no autocorrelation assumption for linear models.

A graph of the simulated data for $\rho = 0.01$ helps illustrate the effect of low autocorrelation on prediction performance. The data looks pretty random, indicating that there is no autocorrelation. This results in more accurate prediction performance as seen in the simulation.



While ρ goes up, the prediction performance gets worse as I suspected. With $\rho = 0.9999$, the 95% prediction interval calculated by the linear model only succeeds in covering the real value about 83% of the time. It's surprising to me though that the prediction interval coverage isn't that far off until you start getting to a correlation of 0.90. It's easier to see a possible reason why when looking at the simulated correlated data for 0.8, 0.9, and 0.99.





At $\rho = 0.80$ the data looks correlated, but you can still see some randomness to the data, which makes it so the autocorrelation doesn't affect the linear model prediction performance too badly for the next observation. On the other hand, when you have $\rho = 0.90$ the data is a lot more correlated and doesn't have as much randomness, thus resulting in worse prediction performance. There is even more correlation when $\rho = 0.99$ resulting in the prediction performance going further down.

For the size of the simulated data points n, I noticed that the prediction performance got slightly better as n went up for most values of ρ. Although, when $\rho = 0.9999$ the number of simulated data points didn't seem to have an effect on the prediction performance. I expected that n wouldn't affect the prediction performance significantly, but these results indicate a possible positive effect of n on prediction. The effect of n seems to be greatest when ρ is around 0.9, but it doesn't have a significant effect at the extremes $\rho = 0.01, 0.9999$. More investigation into power and the relationship between linear models and time series would be needed to understand the cause of this possible relationship, but for the purposes of this analysis we see that as autocorrelation goes up the prediction performance gets significantly worse.

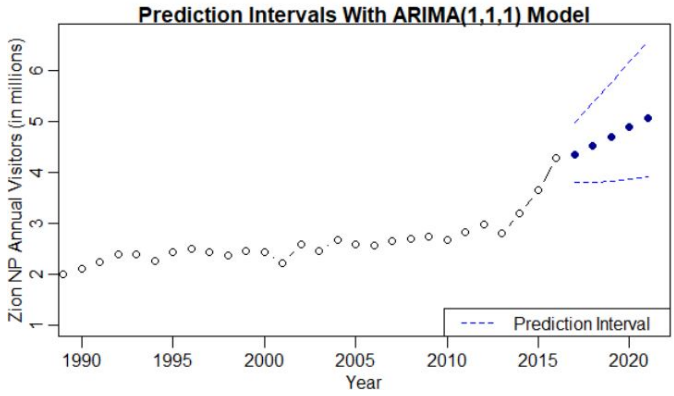
3 Advice for Statistical Practice

When fitting data to a linear model, it is important to check if there is autocorrelation in the data. In a time series, this is usually the case. You can check autocorrelation by looking at a graph of the data, residuals, and performing the Durbin-Watson of autocorrelation. When looking for future predictions, fitting a linear model to data may be easier, but it overlooks the relationships in the response variable, especially for time series data.

One thing I found interesting from my simulation is that when $0 < \rho < 0.8$ the 95% prediction interval for the next observation is fairly accurate. When ρ isn't too large it may be valid to fit a linear model to autocorrelated data if you are looking for prediction intervals on the next value. (This is only for the specific case of looking at Prediction Intervals and may not be applicable to other applications of linear models.)

For an unsuspecting statistician, fitting a linear model to time series may not affect prediction interval coverage in many cases, but in cases where the autocorrelation is high the prediction interval coverage will be misleading and not have the expected coverage. To be on the safe side, it's important to fit a model that takes into account autocorrelation, rather than blindly fitting a linear model to the data. This is especially true when current important decisions are dependent on future predictions.

The Zion National Park annual visitors data clearly showed that there is autocorrelation, and the autocorrelation looked pretty high, therefore, fitting a linear model to the data would distort the prediction interval. Because of this, I fit the Zion annual visitors to an ARIMA(1,1,1) model and then calculated prediction intervals with the appropriate fit.



If a simple linear model was fit, then there would be a greater likelihood that the park would over prepare or under prepare. A Zion National Park leader could have more confidence in the annual visitor predictions given from an ARIMA(1,1,1) model rather than from a linear model, and prepare for upcoming years accordingly.