STAT 100 Chance and Data Analysis Nov. 27, 2002

Review!

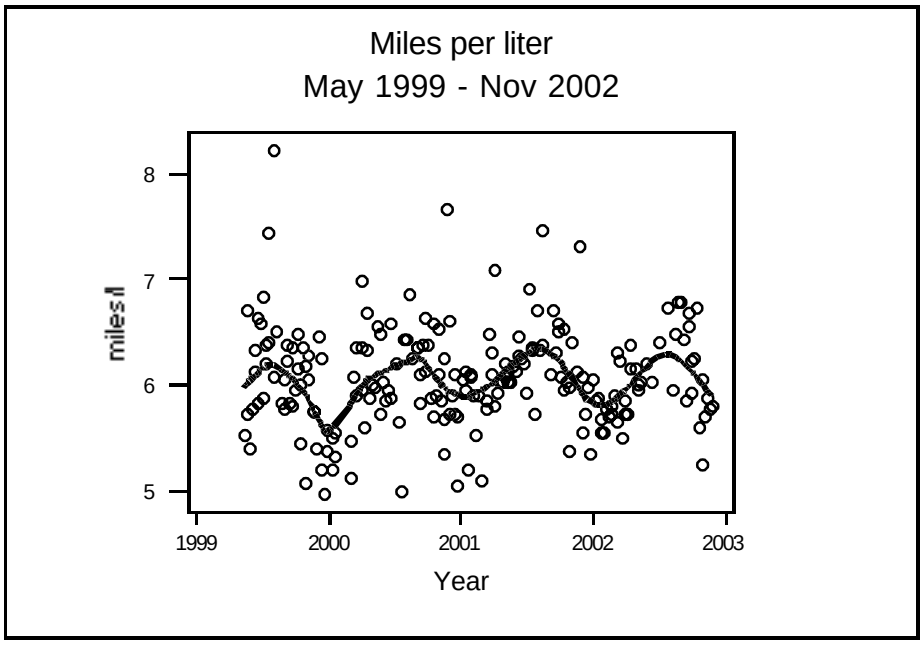
1. Time Series
   Smoothing
2. Correlation
3. Experiments and Causality
4. Regression Prediction
5. Models for Probability Distributions
6. Sampling and Estimation

----------(rest on Friday)

7. Quality Control
8. Hypothesis Testing
9. Optimization
10. **Course evaluations – Friday – Please come to class!**

1. Time Series

Recall Time Series of gasoline mileage:



There was a question of whether the lower winter mileage was caused by the wetness of the road or the outside temperature. Of course, there is a positive correlation between these two variables so it will be difficult to separate the effect of each on mileage.
For each fill up since the course started, I recorded some additional information – what proportion of the time was the road wet, and what was the temperature during the drive.
The data since September looks like this:

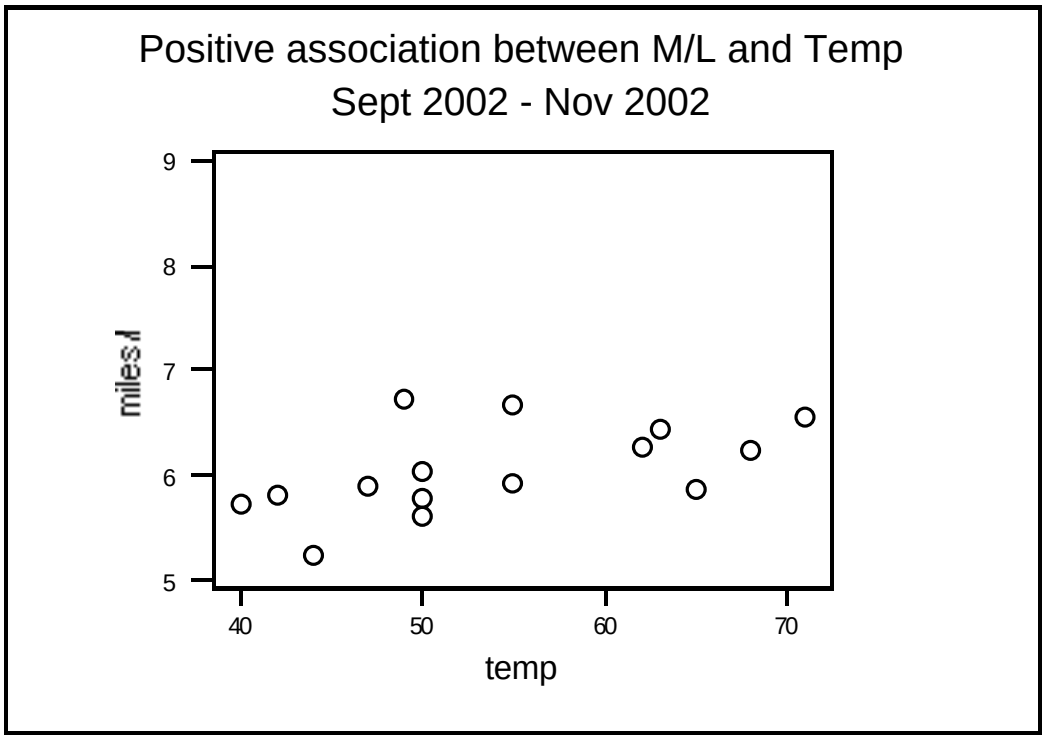
F°			wets/trp	m/l	date
63	2.0	11	0.18	6.43	2002.69
65	0.0	6	0.00	5.85	2002.70
71	0.0	8	0.00	6.55	2002.72
55	2.0	10	0.20	6.68	2002.72
68	0.0	10	0.00	6.23	2002.74
55	0.0	8	0.00	5.92	2002.76
62	2.0	10	0.20	6.26	2002.77
49	4.0	9	0.44	6.72	2002.78
50	0.0	8	0.00	5.59	2002.80
50	2.5	10	0.25	6.04	2002.82
44	2.5	10	0.25	5.23	2002.84
40	0.0	6	0.00	5.71	2002.84
47	5.0	10	0.50	5.88	2002.87
50	3.0	12	0.25	5.78	2002.89
42	1.5	4	0.37	5.80	2002.90

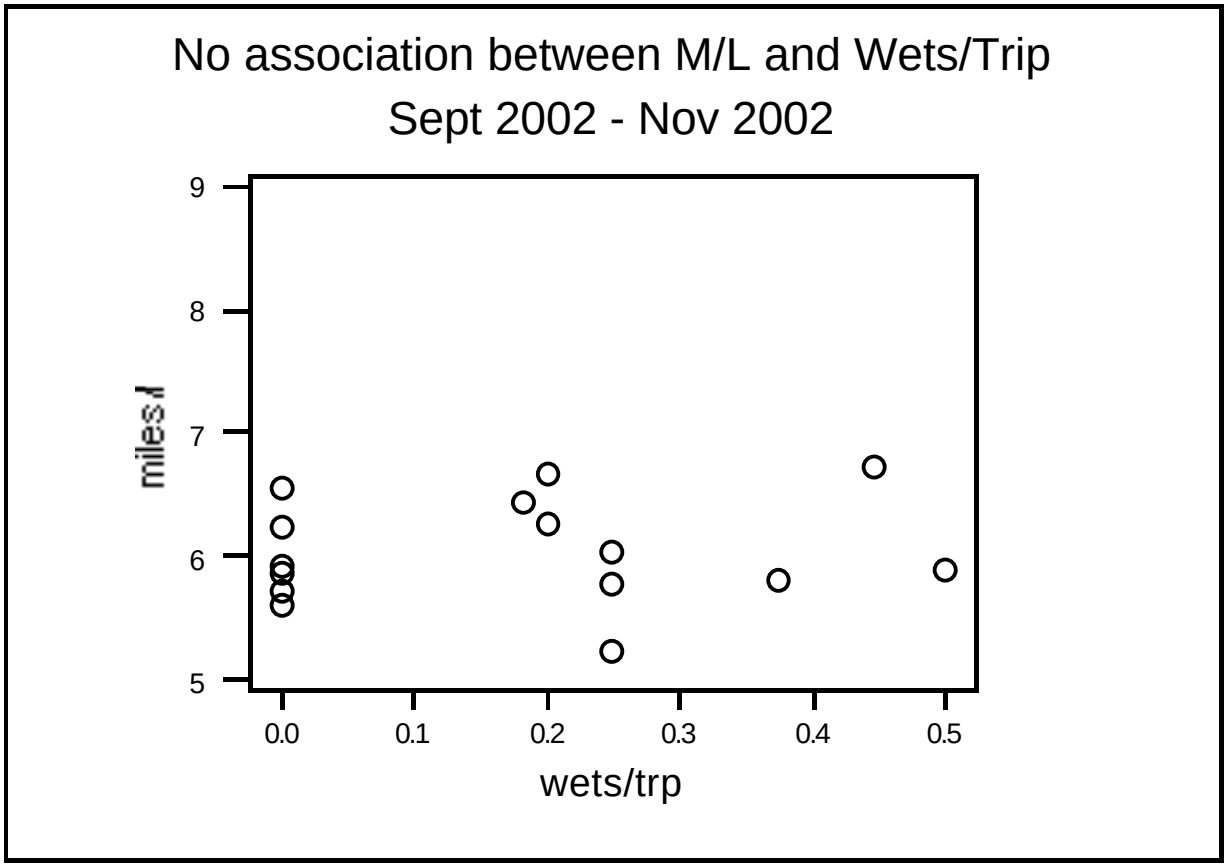
and the correlation matrix is:

2. Correlations (Pearson)

	temp	wets/trp
wets/trp	-0.488	
miles/l	0.542	0.091

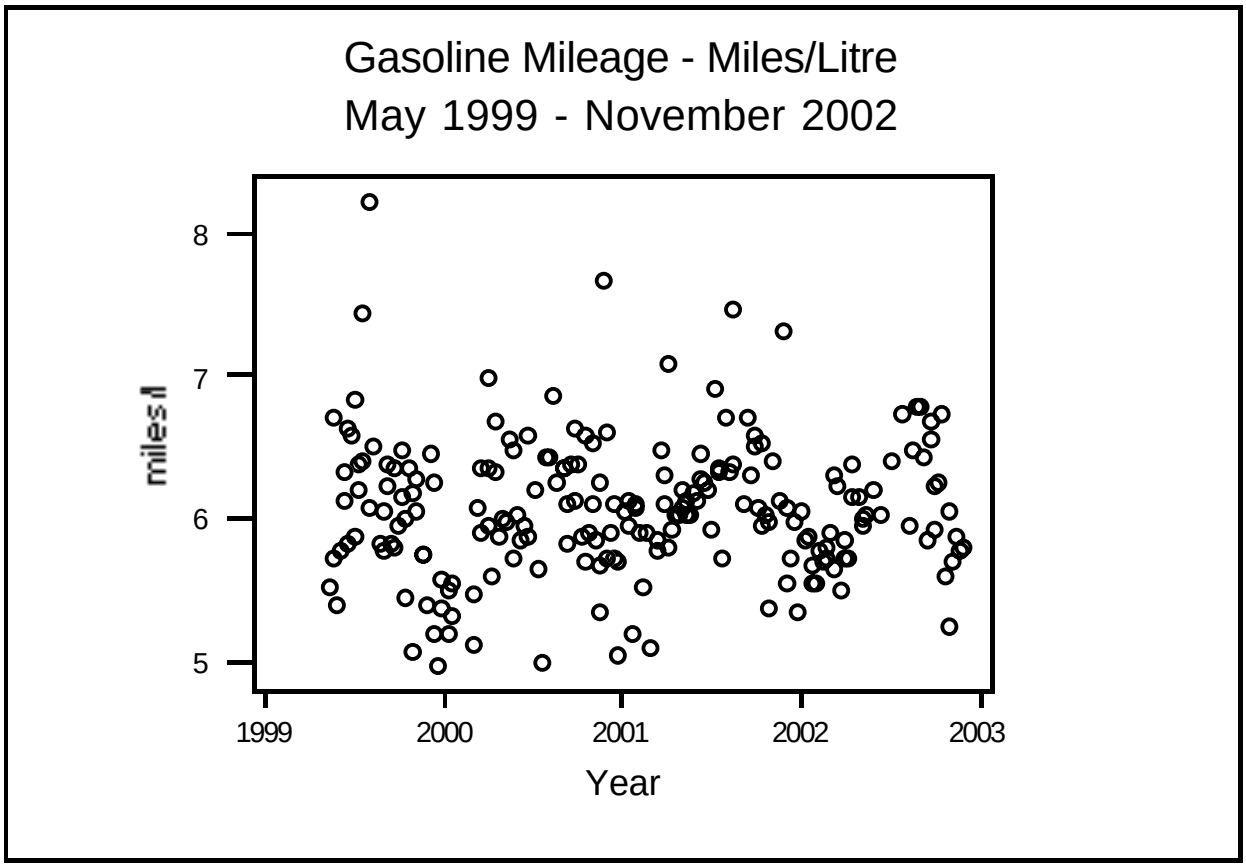
This shows that the mileage is more related to the temperature than to the wetness of the road surface. Of course, we have not ruled out other causes of seasonal mileage variation such as traffic density.


Positive association between M/L and Temp
Sept 2002 - Nov 2002
miles/l
9
8
7
6
5
40
50
60
70
temp

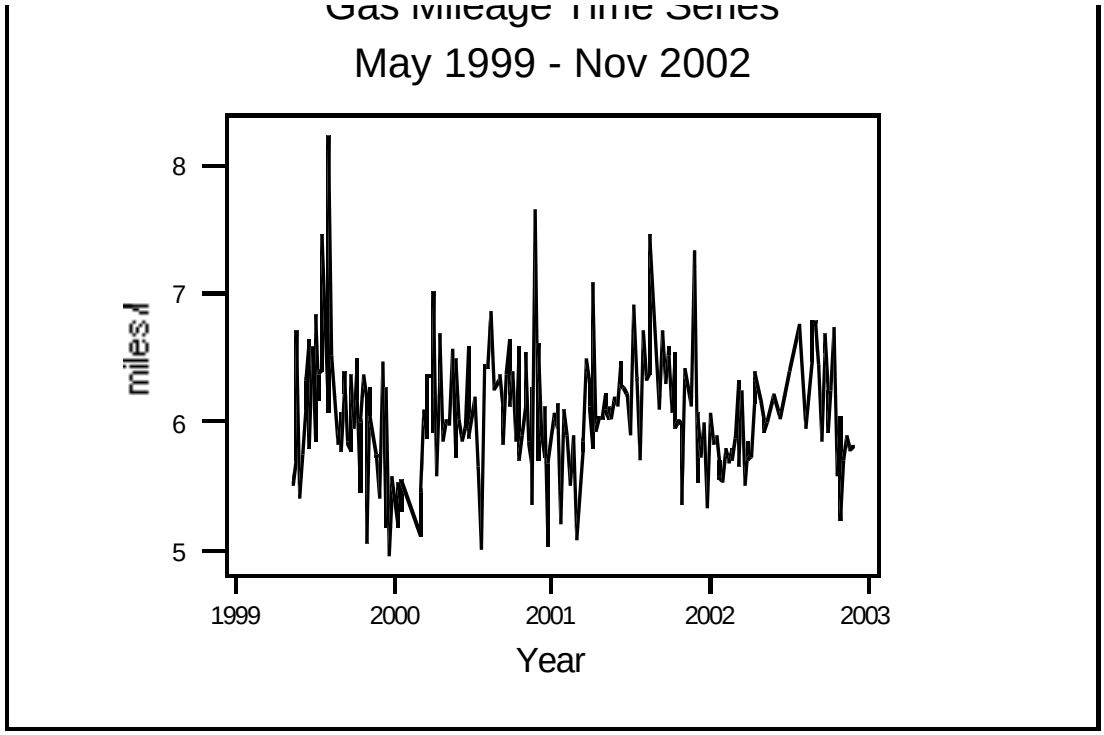

No association between M/L and Wets/Trip
Sept 2002 - Nov 2002
miles/l
9
8
7
6
5
0.0
0.1
0.2
0.3
0.4
0.5
wets/trp

1. (Continued) Time Series Smoothing

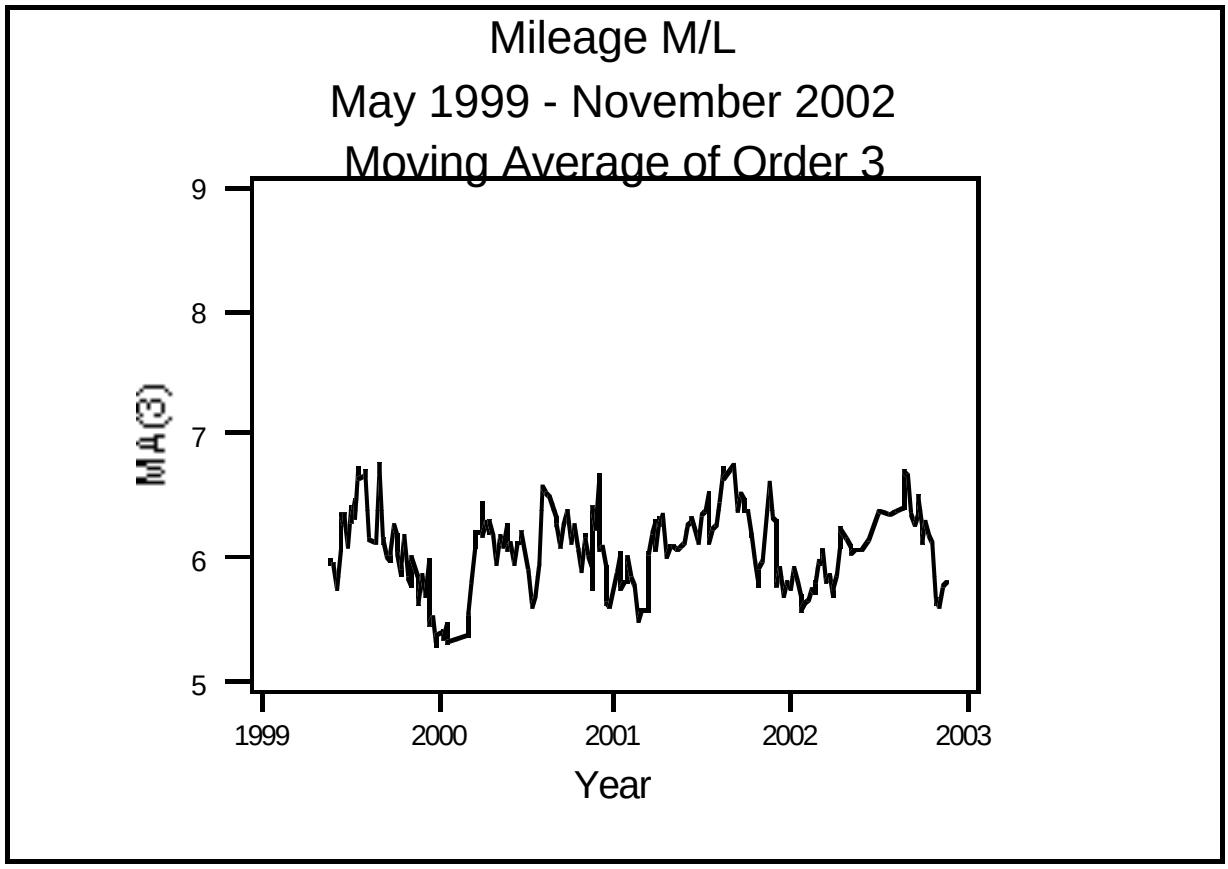
Let us now re-visit the gas mileage data to consider the effect of smoothing the data.



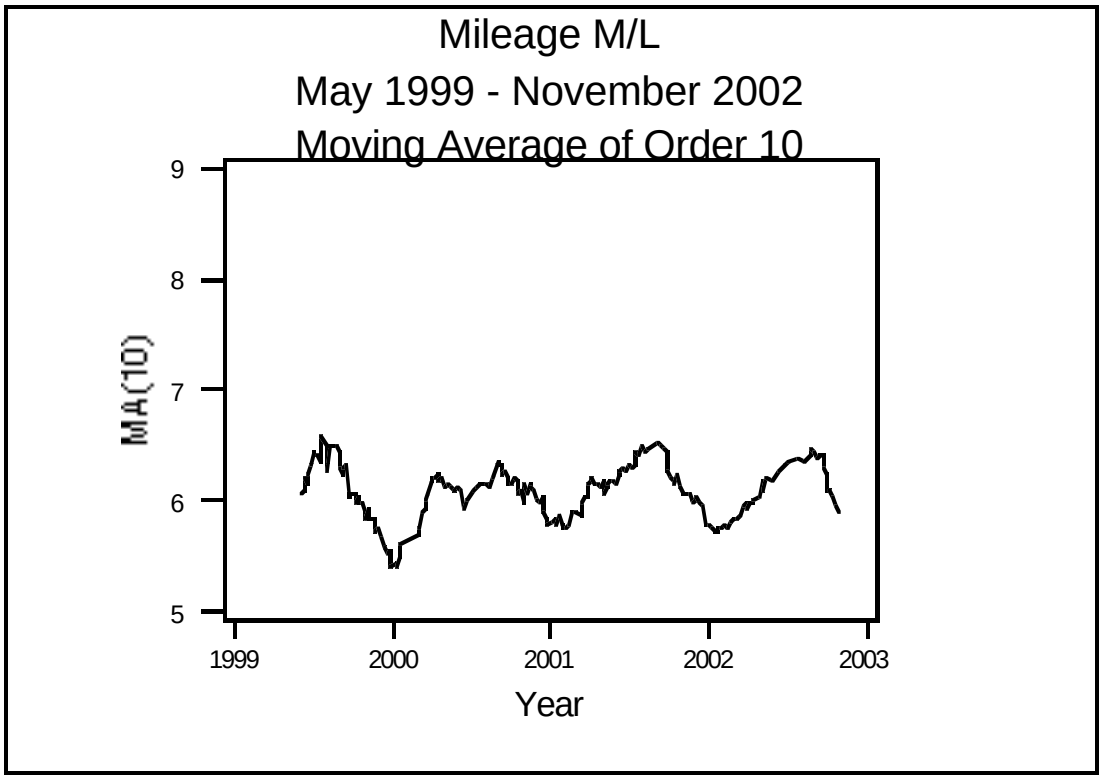
Connecting the data points,



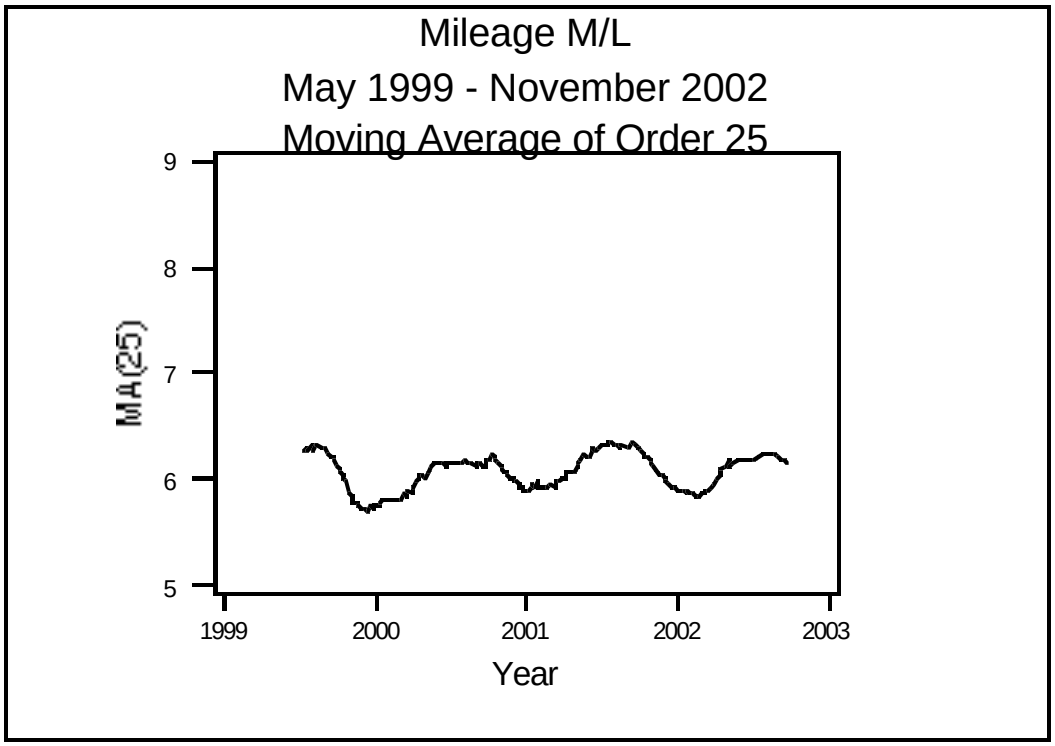
Now consider the moving average of "order 3" which means we first plot the average of the first 3 values, then we plot the average of the values in positions 2 to 4, and then the average of the ones in position 3 to 5, and so on. The time that we plot the average of the first three values is the time of the second value, and the time at which we plot the average of the values in positions 2 to 4, is the time of reading 3, and so on. The result is



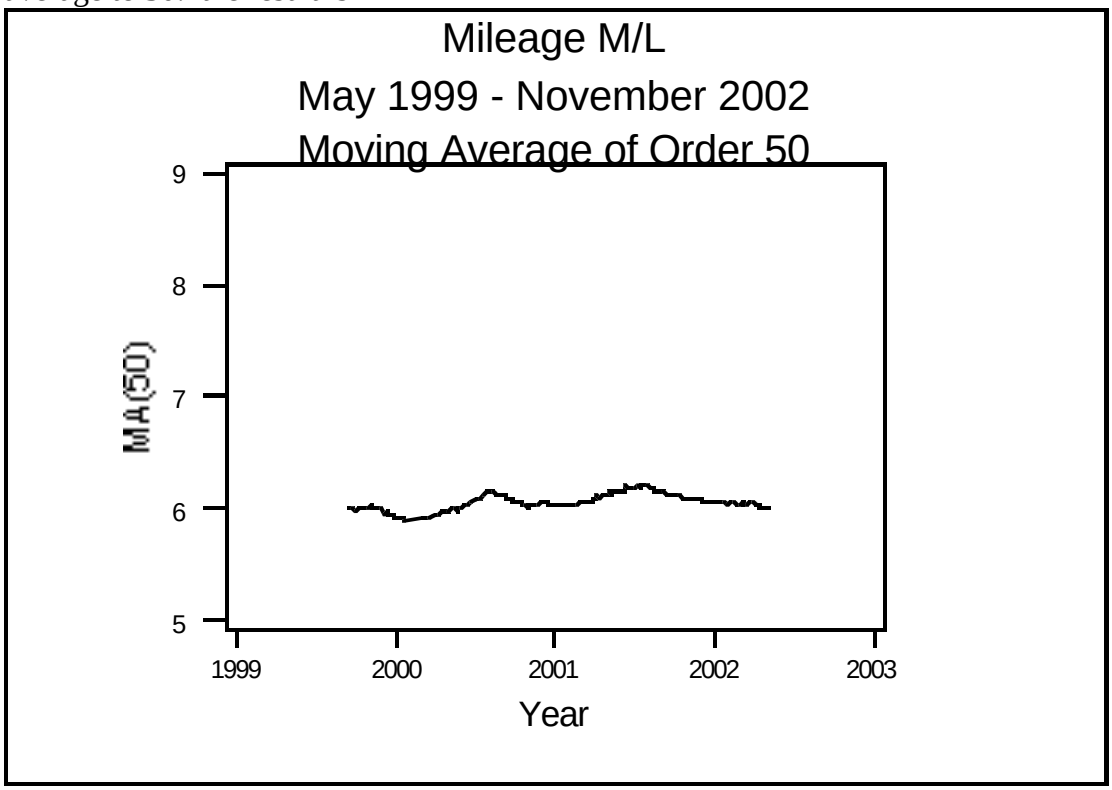
Now what happens if we average more than 3, say 10 values. This produces the moving average of order 10. It looks like this:



and of order 25,



The seasonal pattern is now very clear. Lets keep increasing the order of the moving average to 50: the result is



We have lost the seasonality by smoothing too much!

There is a right amount of smoothing and it is mostly determined subjectively.

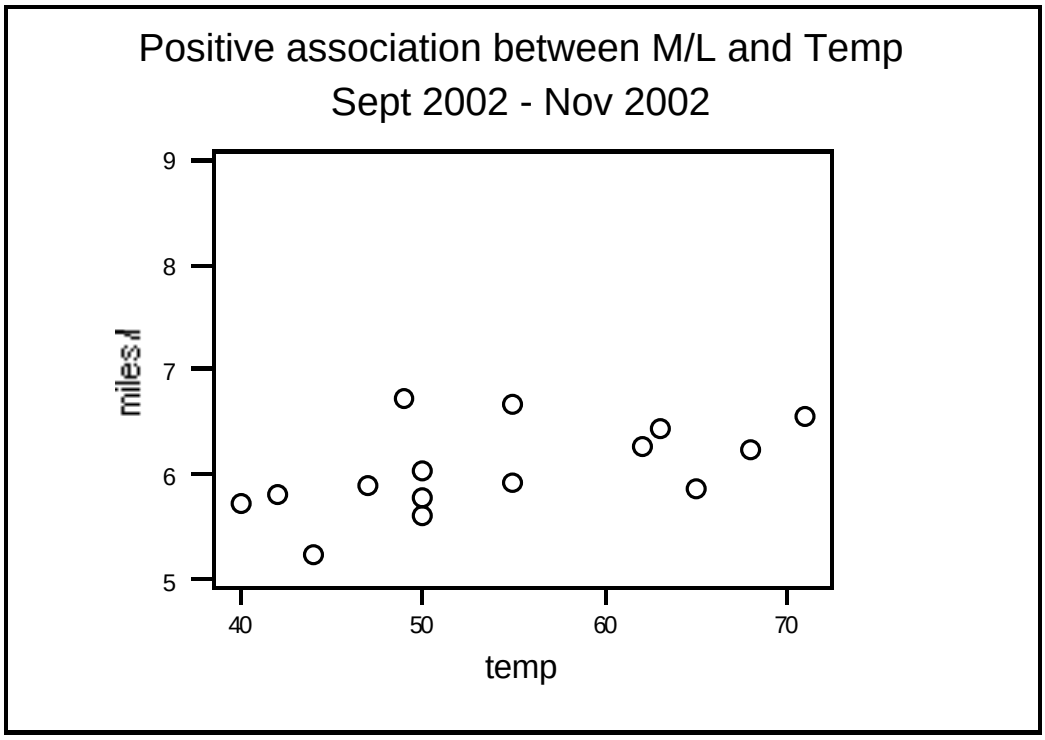
Notice another problem with a large order moving average. We lose the beginning and end of the series. The larger the order of the moving average, the more3 points we lose.

The moving average is an easy smoothing method to understand, but it is quite flawed. There are better ways (like lowess) but they are more complicated and beyond the scope of this course. The important thing is to understand the role of smoothing and to understand that there can be too much and too little for displaying (or finding) and underlying pattern.

3. Experiments and Causality:

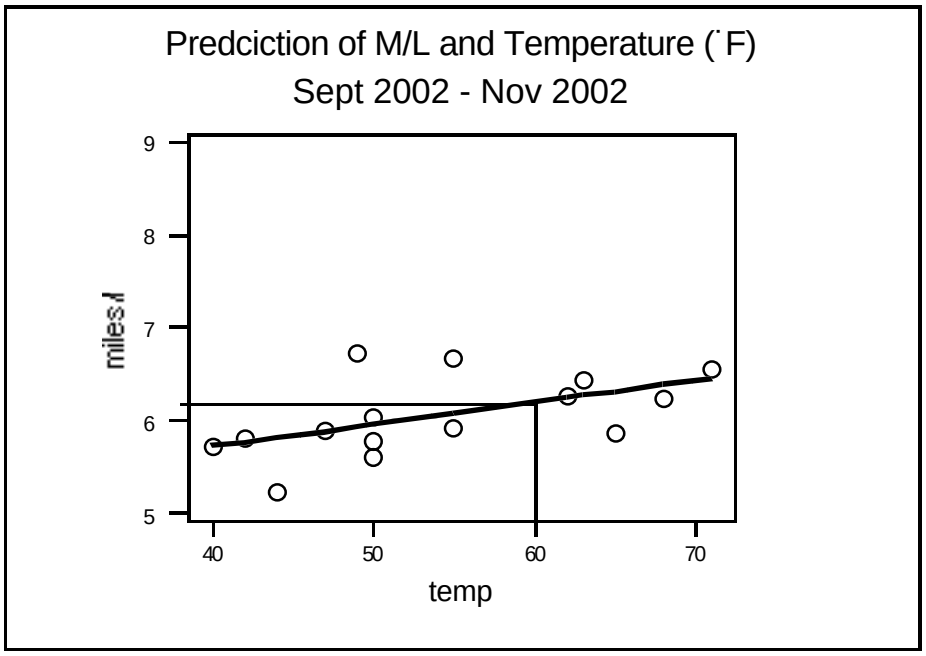
How would my data on mileage be collected as a true experiment to determine if it is traffic or temperature that reduces mileage?

Need to have treatments (high and low traffic; high and low temperature) assigned at random to trips. Tough to do! Can only do an observational study selecting trips occurring under the four combinations – but there is always the possibility that some other cause associated with these conditions is the real cause. The correlation in the scatterplot below does not prove a causal link. Correlation does not prove causation.



4. Regression and prediction

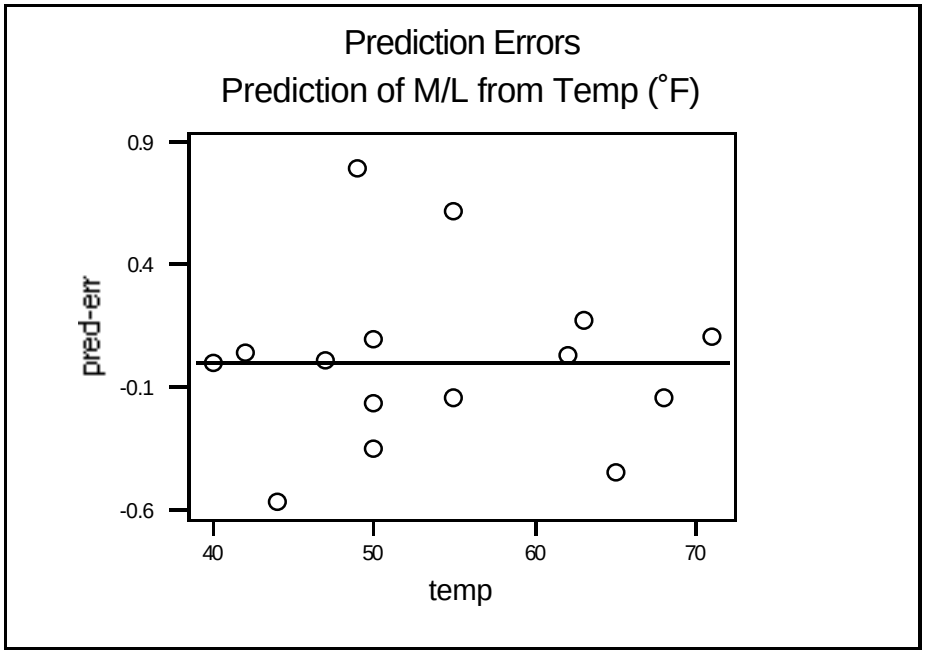
The above relationship does allow a prediction of miles/l at any particular temperature, even if it is not a causal relationship.



AT 60˚F (about 16˚C) the miles per litre would be about 6.2.

How accurate is this prediction?

Look at the prediction errors for this prediction line:



Does the regression line minimize the sum of squared errors? The sum of squared errors in this case is 14.3. Any other line would give residuals that have a larger sum of squared errors. This prediction line is called the "**least squares regression line**".

5. Models for Probability Distributions

Normal – models averages and sums – 68% 95% 99.7% within 1 2 3 SDs of mean
e.g IQs are Normal mean 100 SD 15. What proportion of people have IQ less than 115?
ANS: 84% (68% in 85 to 115, and 16% = (1/2)(100-68) below 85.

Binomial – models number of successes in n 0-1 trials with prob 1 = p. Closely related to proportion of successes which has mean p and SD = $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$

Geometric – models number of 0-1 trials until first head,
where probability of head=p. In survival applications "head" is the end of the duration, like a traffic accident. In survival applications, it models durations when the hazard (probability of ending in the next time period) is constant

Uniform – models situations where all possibilities are equally likely – like digits in a serial number, or points in a square.

Weibull - In survival applications, it models durations when the hazard (probability of ending in the next time period) is increasing (as it would when things wear out – like cars.).

6. Sampling and Estimation

Population – usually not observed – e.g. regular marijuana users among SFU students
is a 0-1 population in this case
Sample – selected at random from the population of interest, sample size n. e.g. n=100
Parameter - aspect of population that is of special interest – e.g. proportion of users, p, a number between 0 and 1.
Numerical summary of sample – sample mean (or sample proportion) and sample SD.
Estimation: of a population parameter, e.g. p by a sample summary value (called a "statistic" by theoreticians).
in this case the natural estimate is the sample proportion (15%?)

Summarize the estimate AND its SD; .15 and $\frac{\sqrt{.15(1-.15)}}{\sqrt{100}}$ = .036.

So estimate is .15 ± .036.

Sampling with and without replacement:

If N=15000, and n=100, and if sampling is without replacement, the SD is actually $\frac{\sqrt{.15(1-.15)}}{\sqrt{100}}$ times $\left(\sqrt{1-\frac{n-1}{N-1}} = \sqrt{1-\frac{99}{14999}} = 0.997\right)$ = .036 again

Sampling without replacement and with replacement result in the same kind of samples when n is much less than N.

One interesting result of this is that a sample of 100 provides as good an estimate from a population of 15,000 as it does from a population of 15,000,000.

Next time **"Course Evaluations"**

This is a new course – evaluations especially important this time.