STAT 400 Data Analysis September 26, 2005

Today: Intro to Bimbo Data Set and Problem
Seasonal Adjustment.
Comments on Mercedes Assignment

Programs to look at annual trends.

```
plot(WkD)
a=loess(WkD~Row)
b=predict(a)
plot(a,type="l",col="red")
lines(b,type="l",col="blue",lwd=3)

plot(WkS)
a=loess(WkS~Row)
b=predict(a)
plot(a,type="l",col="red")
lines(b,type="l",col="blue",lwd=3)
```

Details for Future:

Bimbo Bakery Assignment:

To hand in Oct 3. (Wed)

The data set consists of 53 weeks of daily data (Mon-Sat each week): The number of loaves of bread delivered to a particular retail outlet are recorded, as well as the number of loaves that were unsold at the end of the day. (In other words we know the actual sales). Loaves cost $0.50 to make and sell for $1.00. The cost to the bakery of a loaf being unsold on the day it was delivered is $0.25 (as it would be if half the old loaves were sold at half price, and the rest thrown out.) The cost of a loaf being demanded by a customer when it there are none left is $2.00 per customer (since there is a chance the retail outlet may lose the customer.)

The objective is to advise the bakery on whether its delivery amounts are optimal to maximize gross profits to the bakery. For example, if 100 loaves are delivered, and 90 are sold, gross profit is 90*(1.00-0.50) – 10*(0.25) = $42.50. If 100 loaves are delivered but 110 were demanded, the gross profit would be 100*(1.00-0.50)-10*(2.00) = $30.00.

The catch is that we do not know the demand in those cases where the number of unsold loaves is zero. How can we use the data to judge whether the delivery amounts should be increased or decreased?

This is a situation where we need to do some modeling to help fill in the missing data. One way to do this is to guess the demand distribution, simulate the effect it would have over the year of the data using the delivery amounts, and compare the simulated sales distribution with the actual sales distribution. By trial and error, we can fit a demand distribution to the data in this way. Once we have a good fit, we can again run through the year of data increasing or decreasing the delivery amounts by a certain percentage. By computing the gross profit for several percentage changes in the delivery amount, we can estimate the optimal percentage change (that maximizes gross profit), and recommend this to the bakery.

A preliminary problem is to deal with the seasonal component of the data. Actually, with one year of data, we have no way of knowing if it is really "seasonal" but if there is a trend in the data, we need to allow for it – the reason is that we want to consider the 53 Mondays (or any particular day) as if they were independent. If there were a time-trend over the series of Mondays, this independence would not hold. For simplicity of description, let's call this trend seasonal. A Monday in January would differ from a Monday in July by not only random error, but also by this seasonal trend. The way to do this is to fit the trend, and subtract it from the data, but add back in the yearly average so the numbers look like the seasonally adjusted numbers.

The data to smooth could be sales or deliveries. Since sales are truncated by deliveries, I think it is best to use deliveries as the seasonal indicator. Then adjust the sales by the same proportion that the deliveries are adjusted. (Have you a better idea?). Once you have the sales and deliveries adjusted for seasonality, you can use the modeling ideas mentioned above.

Again, the steps are:

Guess the demand distribution for your day.

Simulate 53 values from this distribution.

Use these 53 values along with the 53 (actual adjusted) deliveries to compute what the simulated sales (SS) would be on each of these 53 days. Then compare this SS distribution with the distribution of the actual observed adjusted sales. A close match means you have a good guess of the demand distribution. A poor match means you should try some different parameters in your guessed model or perhaps even a different model. This comparison can be done with a Q-Q plot or overlaying two ecdf plots.

I'll give you some programs to help with this via e-mail and these notes.

Of course, this is just the first step. But the next step is a bit easier. It is to use your demand distribution to assess the annual profit for the company (for this product at this retail outlet). Then by re-running this analysis with the deliveries modified by a ratio r (eg. .95 or 1.05) you can see how to choose r to maximize the profit. A graph is a good way to do this.

What you can start doing right away is to look at the sales distribution and think about guessing the shape and parameters of the underlying demand distribution.

The seasonal adjustment: First look at the D trend (weekly)

```
a=loess(WkD~Row)
 b=predict(a)
 plot(WkD,type="l",col="red")
 lines(b,type="l",col="blue",lwd=3)
```

Then adjust rel to mean. Then compare the original and adj WkD.

```
mean(WkD)
WkD.adj=678+(WkD-b)
plot(WkD.adj,type="l",col="green",lwd=2)
lines(WkD,type="l",col="red",lwd=2)
```

Now show that loess on the WkD.adj is flat.

```
a=loess(WkD.adj~Row)
b=predict(a)
plot(a,type="l",col="red",lwd=2)
lines(b,type="l",col="blue",lwd=2)
```

Comments on the Mercedes Data Assignment:

Everyone figured out how to use the software to get a reasonable graph. The descriptions of what the graph showed varied a bit but were generally well done. The definition of the exercise was deliberately vague to encourage thinking about what would be appropriate. I am giving almost full marks to everyone! Nevertheless, there are several points to be made that should be kept in mind for future analyses …

- Graphs need titles, and axes should have understandable labels.
- Many graphs for analysis, few graphs for presentation. Must choose a value of alpha. Less is More! Put trial-and-error in an Appendix. Same for residual plots.
- Residual plot is a good idea to check choice of alpha.
- Choice of alpha depends on the information sought. Seasonal will smooth over annual anomalies.
- Descriptions should use whole sentences!
- If litres/100km is "consumption", km/litre is something else, perhaps "fuel efficiency." Use the words that suit the data graphed.
- Good to attach software commands used – helps to inform marker.
- Use of context – starting with an order 2 moving average reduces the variability without losing information. (Big fill-up errors are unlikely to be adjacent in the original data.)

- The interpretation should also use the context of the data. While the loess smooth shows a clear seasonal trend, it is not clear what aspect of the season is the cause – temperature, rainfall, traffic, or tire pressure? Might be worth stating is report ….
- Fitting a sin curve subsequent to the loess analysis might make sense if the phase of the pattern were important to estimate… i.e. at what time of year is the fuel efficiency the lowest? This might give a clue to the cause
- When more info is available about the context of a data set, this info should be sought. For example, the air conditioning was not working throughout the period. Also, almost all driving was the same daily route Tsawwassen <-> SFU. Tire pressure maintenance was much less than monthly. Fill-ups were at a wide variety of service stations, so fill-up effectiveness would be variable. Engine temperature based on the coolant was about 80°C, so ambient temp varying from 5-25°C might not make too much difference to the engine efficiency.
- Non-symmetry of residuals is a feature worth mentioning. It relates to the measurement process which may be useful.