AIRPORT SYSTEMS PLANNING AND DESIGN

Assignment 1 (Forecasting Exercise) -- Answer Sheet

Prepared by Richard de Neufville

Overall Lessons:
I intended for you to take away three lessons from this exercise, and look forward to our discussion of it to see how effective it was for you. These lessons were:

1. Mechanically, to carry out a statistical analysis for forecasting, a lot of assumptions must be made, these assumptions matter in that different assumptions lead to different models, but that there is no clear way to determine which assumptions are best. Therefore, it is generally logically impossible to have much faith in any forecasting model.
2. In any event, even if one were convinced the model were right, making any forecast requires one to plug in forecast values for the drivers of the model, and those must be speculative to a degree. Thus the forecast in uncertain.
3. Therefore, because we cannot know the future within a wide range, we must develop our plans so that they can deal with a plausible variety of scenarios.

Assumption 1:
It is helpful recognize that the performance of any system is due to both internal and external factors. Applying this concept to the forecast of passenger traffic for LAX, one needs to think about the characteristics that promote traffic from the region, and other factors that promote travel to the region, such as tourism.

Many local factors can affect traffic from LAX -- population, employment, competition in terms of alternative airports or modes of communication (road, video conferences), attractiveness of air travel (prices, safety, service). There is no magic way to decide which are best conceptually. Note that variables, such as population, may stand for a large number of others which correlate well with it -- if other factors, such as the distribution of income or the participation of men and women in the work force, remain fairly constant. Population by itself is not a good explanation of traffic, which is why Boston has about 4 times as much airline traffic as Lisbon, and twice that of Sydney, all of which have about the same overall population.

Note in this connection that it is statistically wrong to use variables which are strongly correlated as explanatory variables, even if this makes sense conceptually. It is wrong because once one deals with highly correlated explanatory variables, it is difficult to have faith in their coefficients because these become interchangeable. Thus if we try to develop a statistical model of the form:

$$\text{Traffic} = a + bX + cY$$

and if: $Y = dX$

Then an infinite variety of equally good equations are possible:

$$\text{Traffic} = a + (b + cd)X = a + (c + b/d)Y$$ and any proportions in between.

This phenomenon is called multicollinearity and the easiest way to test its existence is to examine the standard errors of the coefficients. If high standard errors are observed among some coefficients and dropping one or more variables from the equations lowers the standard errors of the remaining variables, multicollinearity will usually be the source of the problem[1]. You can test this by comparing the standard errors in Forecasts 1 and 2, with and without Employment as an explanatory variable In addition to Population.

[1] Pindyck, Robert S., and Daniel L. Rubinfeld. Econometric Models and Economic Forecasts. New York: McGraw-Hill, 1991.

As to external factors, note that the traffic at many destinations is principally driven by markets elsewhere and factors that constrain this traffic. This is particularly true for tourist destinations (Guam, Hawaii, Cancun, Palma de Majorca, Canary Islands, etc.). Thus the traffic for Guam is probably mostly a function of economic conditions in Japan, the number of hotel rooms in Guam, and the attractiveness of alternative destinations. In this connection note that Los Angeles is an international and domestic tourist destination (Hollywood, Disney Land, etc.)

Assumptions 2 and 3:
As a point of departure, it should be stressed that a linear equation is no more difficult to estimate that an exponential one which can be linearized logarithmically:
$Pax = a_o X_1 (\exp a_1) X_2 (\exp a_2) \quad => \quad \ln Pax = a_0 + a_1 \ln X_1 + a_2 \ln X_2$

The choice between a linear and an exponential form properly revolves around the question of what is the underlying behavior one really wants to model. For theoretical reasons it may be generally preferable to use a non-linear form because:

- response to price is typically non-linear, and
- many factors grow exponentially with time, such as Population and Employment specifically, so that this may be convenient.

From a theoretical point of view, it would generally be desirable to segment the market into distinct groups that respond differently to incentives or to the environment. Typical segmentations of airline traffic would be:

- Business (not particularly price sensitive) versus Leisure (price sensitive)
- Local traffic (responsive to local conditions) versus Foreign traffic (responsive to conditions elsewhere)

In practice, however, these kinds of segmentations may be impractical: it may be impossible to obtain the data.

Assumption 4:
It is rarely clear where to draw the line between the periods one includes and excludes from the analysis. More data is better statistically, if the data relates to current and plausible future situations. Older data may be out-of-date.

Clearly, airline traffic in the United States changed drastically after the economic deregulation of 1978 -- a good reason to exclude prior years. However the period since was punctuated by a variety of extraordinary events:

- the 1979 second oil shock
- the air traffic control strike in the early 1980's
- the 1990/91 Gulf War.

At what point should one draw the line? There is no right answer on this one either.

Assumption 5:
It should be obvious from this exercise that it is easy to get good correlations to time series data, in many different ways.

Note, first of all, that the formula used in the Master Planning exercise does not have the highest correlation as measured by R squared. In detail this expression has two salient problems:

- none of the estimates are significant at the 90% level, as indicated by the t-statistics. This is almost certainly due to the autocorrelation between the Population and Employment variables, as evidenced by the observation that the coefficients appear significant when either is dropped out.

- the sign of the coefficient on yield is “wrong” in that from a theoretical point of view the number of passengers should decrease when yields rise, rather than the other way around. What is the explanation for this? A likely answer is that when the demand for seats is high, the airlines raise prices. (Any of you trying to get seats for a popular concert will notice that you’ll have to pay more than usual). In this case, as in many others, there is two-way causality: price rises affect demand, but the level of demand affects prices.

Second, note that adding more “explanatory” variables always improves a correlation. This is because the process of least-squares is an optimization process, and a new variable can never hurt (if it were going to, the optimization process would give it a coefficient of zero). Thus dropping a variable between equations 1 and 2 obviously decreases R squared. Note however, although you were not required to do the exercise, that the very small decrease in R squared due to dropping Employment as an explanatory variable does not indicate the inherent importance of Employment, because of the correlation between Population and Employment. If Population had been dropped out of Forecast 1, the resulting decrease in R squared would also have been small. (Try it.)

The best forecast of the four specified is the one that is exponential with time. Noting that the coefficient “r” in the expression indicates the continuous compounding rate per unit time, so that the best fit for traffic at LAX over the period is 3.16% per year. (This arises, of course, because so many of the explanatory variables, such as population, tend to grow at a compound rate.) Any of you could have calculated this in a day’s work -- yet the LAX forecasting exercise that you have seen probably cost about $500,000 or so. Think about whether a more complicated effort is really worth paying for.

The correlation of LAX traffic with the “irrelevant” variables is also in general good, because so many of these variables also grow exponentially with time. See the attached table.

Assignment 6:
The forecast in any event depends on forecast values of the “explanatory variables”. The exercise comes down to substituting several forecasts for a single one! This is not a clear advantage.

The forecasts of the explanatory variables are not unique, as the Master Plan for LAX demonstrates. (Many other forecasting exercises refer, explicitly or implicitly, to a single set of forecasts of the explanatory variables, but this does not demonstrate that there is only one forecast for these data -- only that they make use of only one.)

Using the different forecasts leads to a wide range of possible forecasts. As you have shown yourselves in the exercise. Furthermore, these projections are based on extrapolating past trends -- and we know that changes in trends occur routinely, and that these changes further alter forecasts. Because of this, an enormous effort on precision for the long-term model itself may not be justified as a practical matter.

Conclusion:
Given that the projections can be within a wide range, and that there is really no solid basis for picking any one projection as the best, it would seem appropriate to plan around the plausible range of forecasts. One might choose to focus on a mid-point forecast as a reference point, but the planning process should specifically explore the implications of higher or lower levels of activity.

Irrelevant variables for 2004

Railroad Length in 19th century Germany	0.996
Population of China	0.99
Population of France	0.99
Square root of time	0.97
Abortions in Quebec	0.97
US Retail Sales	0.97
Car Accidents in South Korea	0.97
Crime in South Dakota	0.94
Internet Traffic	0.92
South Korea cellphones	0.89
Sulphuric Acid Production in China	0.87
Red Sox Attendance	0.64
Japanese Travelers overseas	0.62
Green Bay Packer wins	0.36

Irrelevant Variables for 2002

Irrelevant Variable Chosen	**Resulting R^2**
Population of Buffalo County, NE	0.996
Population of Iran	0.994
World Population	0.991
Articles Concerning Biological Information Published	0.990
Unenrolled Secondary Students in Honduras	0.989
Population of India	0.988
TV Advertising $	0.985
Agricultural Vehicles in Switzerland	0.975
Population of Brazil	0.975
Life Expectancy at Birth in Chile	0.945
Mass of Permanently Discharged Commercial Spent Nuclear Fuel from Light Water Reactors	0.933
New Zealand Credit Card Billings per Month in NZ$	0.922
Worldwide Bicycle Production	0.860
Women in Congress	0.724
Elementary and Secondary Teachers in the US	0.588
Kg Fish per Capita	0.575

Correlations between the Number of Passengers at Los Angeles International Airport from 1975 and 1995, and "Irrelevant" variables selected by students in class exercise (1998)

Variable Selected (US Data unless noted)	R Squared
Disposable Income in Hawaii	0.999
Computing Power (Moore's Law)	0.995
Population of Mexico	0.990
Soybean Production	0.989
Automobiles	0.982
Rural Population	0.974
Data on a Praying Mantis experiment	0.973
Persons employed in department stores	0.964
Analyst's Height in Centimeters	0.963
Ratio of Senior Citizens in Japan	0.957
Poultry consumption	0.947
Private Investment in Non-residential structures	0.916
Internet traffic (1991-1994)	0.909
Anthropogenic Atmospheric Emissions	0.909
Consumer Price Index (1955-1975)	0.908
Half of Analyst's father's age	0.908
Crime rate	0.885
January Department Store Price Index	0.884
Median Sales Price of a Home	0.884
Prison Population in Oregon	0.868
Sales of Cigarettes	0.768
Passengers on Japanese Bullet Train	0.752
Rice Production	0.480
LAX 3 variable model	0.923
(Note: about half the 1 variable models with irrelevant data did better! Many more of these would have done better if they had 3 variables!!)	