ENGINEERING TRIPOS PART IIA 2004

Solutions to Module 3E3
Modelling Risk
Principal Assessor: Dr H Jiang
Second Assessor: Dr S Scholtes

Exam Cribs for 3E3 Modelling Risk (May 2004)

1(a) The Chi-square goodness-of-fit test can be used to determine whether the sample data generated by the random number generator is of the expected probability distribution.

1(b)(i) The key uncertainty is the demand. You may have records for past sales. Sales forecasting could be done through a regression analysis over the past data. The sales could be of a particular probability distribution such as Poisson distributions.

1(b)(ii) To solve this problem using simulation, first select a value for s, then set up an experiment in which the value D is generated from the probability distribution obtained in (i). Finally, the cost associated with these particular values of s and D is calculated. This experiment is replicated independently n times, producing n observations of the cost. A confidence interval for the mean cost may be computed from these data. The whole process is then repeated for other values of s. These estimates of the cost can be plotted and the plot can be used to locate the optimal value of s.

1(c) The trend is the long-term sweep or general direction of movement in a time series. Seasonality indicates that a time series exhibits a regular or repeating pattern.

1(d)(i) X_t: The observation

E_t: The base level

T_t: The per-period trend

S_t: Seasonality factor

$F_t(k)$: The forecast of period $t+k$ carried out at the end of period t

- E_t is a convex combination of two terms, where X_t/S_{t-c}, the deseasonalised observation, is an estimate of the base obtained from the current period, and $E_{t-1}+T_{t-1}$ is our base level estimate before observing X_t.
- T_t is a convex combination of two terms. The first term, $E_t - E_{t-1}$, is an estimate of trend from the current period given by the increase in the smoothed base from period $t-1$ to period t. The second term, T_{t-1}, is our previous estimate of the trend.
- S_t is a convex combination of two terms. X_t/E_t is an estimate of period t's seasonality and S_{t-c} is the most recent estimate of period t's seasonality (the last and the same season).
- $F_t(k)$ is set to equal to $(E_t + kT_t)S_{t+k-c}$, where E_t is the base level, kT_t is the accumulated trend, and S_{t+k-c} is the most recent estimate of period $(t+k)$'s seasonality.

1(d)(ii) The given data sre shown in the table below.

year	index	X_t	E_t	T_t	S_t	F_t
2003	1		300	50		
	2	380				
	3					
	4					*
2004	1					
	2					*
	3					
	4					

$S_{-3} = 0.9, S_{-2} = 0.95, S_{-1} = 0.95, S_0 = 1.2,$
$\alpha = 0.2, \beta = 0.4, \gamma = 0.5$
$c = 4$

When $t = 2$,

$$\begin{aligned} E_t &= \alpha X_t/S_{t-c} + (1-\alpha)(E_{t-1}+T_{t-1}) \\ &= 0.2 \times 380/0.95 + 0.8(300+50) \\ &= 360 \end{aligned}$$

$$
\begin{aligned}
T_t &= \beta(E_t - E_{t-1}) + (1-\beta)T_{t-1} \\
&= 0.4(360 - 300) + 0.6 \times 50 \\
&= 54
\end{aligned}
$$

$$
\begin{aligned}
S_t &= \gamma X_t / E_t + (1-\gamma)S_{t-c} \\
&= 0.5 \times 380/360 + 0.5 \times 0.95 \\
&= 1.0028 \\
&\approx 1
\end{aligned}
$$

When $k = 2$, the forecast for the fourth quarter of 2003 is

$$
\begin{aligned}
F_t(k) &= (E_t + kT_t)S_{t+k-c} \\
&= (360 + 2 \times 54) \times 1.2 \\
&= 561.6(\text{billion dollars})
\end{aligned}
$$

1(d)(iii)

When $k = 4$, the forecast for the second quarter of 2004 is

$$
\begin{aligned}
F_t(k) &= (E_t + kT_t)S_{t+k-c} \\
&= (360 + 4 \times 54) \times 1.0 \\
&= 576.0(\text{billion dollars})
\end{aligned}
$$

2(a) The linear regression model is $y = a + bx + \varepsilon$. The objective of the least square criterion is to minimise the square deviation between the actual value of the dependent variable and the predicted value based on the regression model.

$$\max \quad \sum (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = a + bx_i$, and a and b are decision variables. Substituting, the objective becomes

$$\max \quad \sum (y_i - a - bx_i)^2$$

Let $z = \sum (y_i - a - bx_i)^2$. The optimal solution of the above minimisation problem must satisfy the following optimality condition:

$$\begin{cases} \frac{\partial z}{\partial a} = 0 \\ \frac{\partial z}{\partial b} = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} \sum 2(y_i - a - bx_i) \times (-1) = 0 \\ \sum 2(y_i - a - bx_i) \times (-x_i) = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} \sum y_i - na - b \sum x_i = 0 \\ \sum x_i y_i - a \sum x_i - b \sum x_i^2 = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} \sum x_i \sum y_i - na \sum x_i - b(\sum x_i)^2 = 0 \\ n \sum x_i y_i - na \sum x_i - nb \sum x_i^2 = 0 \end{cases}$$

$$\Longrightarrow \sum x_i \sum y_i - n \sum x_i y_i = ((\sum x_i)^2 - n \sum x_i^2) b$$

Hence

$$\begin{aligned} b &= \frac{\sum x_i \sum y_i - n \sum x_i y_i}{(\sum x_i)^2 - n \sum x_i^2} \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(\sum x_i)^2 - n\bar{x}^2} \\ &= \frac{S_{xy}}{S_{xx}} \end{aligned}$$

Since $\sum y_i - na - b \sum x_i = 0$,

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} = \bar{y} - b\bar{x}.$$

For this question,

$$\begin{aligned} b &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(\sum x_i)^2 - n\bar{x}^2} \\ &= \frac{122224 - 1340121/11}{52729 - 579121/11} \\ &= 394.8/81.6 \\ &\approx 4.83 \end{aligned}$$

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ &= 1761/11 - 4.83 \times 761/11 \\ &= 160.09 - 4.83 \times 69.18 \\ &\approx -174.05 \end{aligned}$$

Therefore the regression equation is

$$\hat{y} = -174.05 + 4.83x$$

When $x = 67$,

$$\hat{y} = -174.05 + 4.83x = -174.05 + 4.83 \times 67 \approx 149.54$$

When $x = 73$,

$$\hat{y} = -174.05 + 4.83x = -174.05 + 4.83 \times 73 \approx 178.54$$

2(b) In the regression model $y = a + bx + \varepsilon$, a represents the intercept, and b the slope. a, b and ε are all random variables. a and b in the regression equation $\hat{y} = a + bx$ are determined by minimising the least square from the sampling data (x_i, y_i) and (x_i, y_i) change with sample. Samples are drawn randomly from the whole population. Therefore, a and b are random variables. ε is the error term, which is assumed to be of $N(0, \sigma^2)$.

2(c) R-square statistic is defined by

$$R^2 = \frac{RSS}{TSS} = \frac{(S_{xy})^2}{S_{xx}S_{yy}} = \frac{(\sum(x_i y_i) - n\bar{x}\bar{y})^2}{(\sum x_i^2 - n\bar{x})(\sum y_i^2 - n\bar{y})}$$

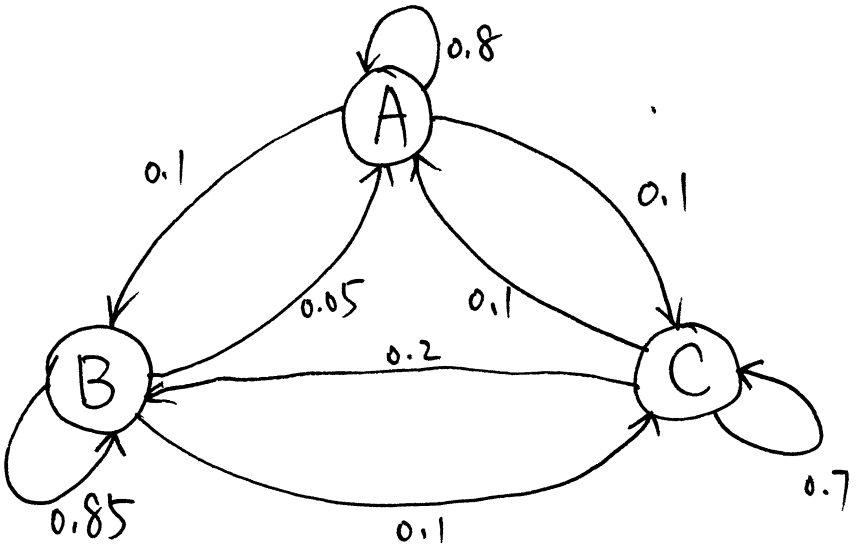
R-square statistics is also called the coefficient of determination. It is the proportion of the total variation in the dependent variable that is explained by its relationship with the independent variable in the regression line. $R^2 \in [0, 1]$. The bigger the R^2 is, the better the linear regression line fits the data.

3(a) The probability distribution of the customer's next car purchase is $(0.8, 0.1, 0.1)$ from companys A, B and C respectively.

If at least one of the next two cars she buys will be a car from company A, then she either buys her next car from company A or buys her second next car from company A but buys her next car from companys B or C. By probability rules, the probability that at least one of the next two cars she buys will be a car from company A is

$$0.8 + 0.10 \times 0.05 + 0.10 \times 0.10 = 0.815$$
$$(A \to A \quad A \to B \to A \quad A \to C \to A)$$

3(b) The transition network for this example is shown in the graph below.



Since $A \to B \to C \to A$, all states are in the same class. This Markov chain is irreducible because all states communicate. This Markov chain is regular because all element in the transition matrix are positive.

3(c) The probability distribution of the states tends to a limit as the number of transitions tends to infinity. This is because this Markov chain is boh irreducible and a periodic (implied by the regularity property of this Markov chain). The steady state probability distribution π satisfies the equation below in addition to $\sum \pi_i = 1$.

$$\pi = \pi P$$

where P is the probability transition matrix. It follows that

$$\begin{cases} 0.8\pi_1 + 0.05\pi_2 + 0.1\pi_3 = \pi_1 \\ 0.1\pi_1 + 0.85\pi_2 + 0.2\pi_3 = \pi_2 \\ 0.1\pi_1 + 0.1\pi_2 + 0.7\pi_3 = \pi_3 \\ \pi_1 + \pi_2 + \pi_3 = 1 \end{cases}$$

The solution is $\pi_1 = 0.25$, $\pi_2 = 0.5$, $\pi_3 = 0.25$. Hence the limiting distribution is $(0.25, 0.5, 0.25)$.

3(d) It can be seen that the Markov chain with the new transition matrix is also irreducible and aperiodic. Therefore, the probability distribution of the states tends to a limit. Similarly the steady state probability distribution x satisfies $\sum x_i = 1$ and the following equation

$$x = xQ$$

where Q is the new probability transition matrix. It follows that

$$\begin{cases} 0.85x_1 + 0.1x_2 + 0.15x_3 = x_1 \\ 0.1x_1 + 0.8x_2 + 0.1x_3 = x_2 \\ 0.05x_1 + 0.1x_2 + 0.75x_3 = x_3 \\ x_1 + x_2 + x_3 = 1 \end{cases}$$

The solution is $x_1 = 4/9$, $x_2 = 1/3$ and $x_3 = 2/9$. Therefore, the limiting probability distribution of the states is $(4/9, 1/3, 2/9)$.

The annual profit without warranty is

$$(16000 - 10000) \times 0.25 \times N = 1500N$$

where N is the market size.

The annual profit with warranty is

$$(16000 - 10000 - 1000) \times \frac{4}{9} \times N = 5000 \times \frac{4}{9} \times N$$

which is greater than $1500N$. Therefore, Company A should institute the five-year warranty since it is more profitable according to the above analysis.

4(a) Traffic intensity is defined by $\rho = \lambda/(s\mu)$, where λ and μ are the arrival rate and the service rate, and s is the number of servers in the queueing system. ρ is the expected fraction of time the individual servers are busy. It is the same as the utilisation of servers.

A queueing system is stationary if the probability distribution of the states of the system remains the same over time.

When traffic intensity is one, the system won't reach a stationary state. $\rho = 1$ indicates that on average every server is busy 100%. However, customers arrive randomly and sometimes servers do not have customers to serve. That means that the queue will eventually blow up.

4(b) Two systems have the different average waiting times. The channel system with the constant service time has a shorter average waiting time than the one with an exponential service time because the latter has a larger variability in the system. The Pollaczek-Khintchine formula reads

$$W_q = \frac{\lambda(S^2 + E^2)}{2(1 - \lambda E)},$$

where E and S are the mean and the standard deviation of the service time distribution, and λ is the arrival rate. The same conclusion could be reached from the above formula. The values of E and λ are the same for both occasions, but the values of S are different ($S = 0$ when the service time is constant).

4(c)(i) The queueing system has five states with probabilities

$$P_0 = 1/16, P_1 = 4/16, P_2 = 6/16, P_3 = 4/16, P_4 = 1/16.$$

$$L = \sum_i iP_i = 0 \times 1/16 + 1 \times 4/16 + 2 \times 6/16 + 3 \times 4/16 + 4 \times 1/16 = 2$$

$L = 2$ indicates that on average, the number of patients in the centre is about 2 (including those who are served by the doctors and those who are waiting for service).

4(c)(ii)

$$L_q = \sum_{i>s}(i-s)P_i = (3-2)\times 4/16 + (4-2)\times 1/16 = 6/16 = 3/8$$

4(c)(iii) The expected number of patients being servered is

$$L - L_q = 2 - 3/8 = 13/8.$$

4(c)(iv) Little's formulae read

$$L_q = \lambda W_q, \qquad L = \lambda W.$$

If $\lambda = 4$, then

$$W_q = L_q/\lambda = \frac{3}{8}/4 = 3/32(\text{hours})$$
$$W = L/\lambda = 2/4 = 1/2(\text{hours})$$

4(c)(v)

$$W = W_q + \frac{1}{\mu}$$

which shows that

$$\frac{1}{\mu} = W - W_q = 1/2 - 3/32 = 13/32.$$

Then the expected service time per patient is 13/32 hours.

4(c)(vi) The traffic intensity is

$$\rho = \frac{\lambda}{s\mu} = \frac{4}{2\times 32/13} = 13/16.$$

Hence

$$1 - \rho = 1 - 13/16 = 3/16 \approx 0.1875$$

This implies that about 18.75% of time, Andy and Bob do things other than serve patients.