Department of Statistics

Paper 475.330 | **1997 Exam Solutions** | **Date: Oct 21**

1. (a) Errors are assumed to be (i) independent, (ii) come from a Normal distribution, and (iii) have constant variance. We would check these assumptions using (i) a plot of residuals versus lagged residuals, (ii) a Normal probability plot, and (iii) a plot of residuals versus fitted values.

   (b) Adjusted R^2, the residual mean square error ($\hat{\sigma}^2$), and Mallow's C_p statistic.

   (c) The odds ratio is given by
   $$\frac{1601 \times 412368}{510 \times 162527} = 7.96$$

   (d) If we have 2 binary factors, A and B, then the odds ratio of A given B is the same as the odds ratio of B given A. This property makes the odds ratio useful for case-control (retrospective) studies because we can calculate the odds ratio for having the disease given some risk factor even though we cannot calculate the probability or relative risk of having the disease.

   (e) Assuming we have grouped data we can do a goodness of fit test for the null hypothesis that the model is adequate:
   $$\text{P-value} = \text{pr}(\chi^2 \geq 13.3), \quad \chi^2 \sim \text{Chi-square(df} = 14)$$
   which (using the table from page 10) gives a P-value of > 0.20. Thus we conclude the model is adequate.

   (f) From the fitted model we have
   $$\begin{aligned} \text{logit}(\hat{\pi}) &= -2 + 1 \times 2 = 0 \\ \hat{\pi} &= \exp(0)/(1 + \exp(0)) = 0.50 \end{aligned}$$

   (g) When fitting log-linear models we assume the response has a Poisson distribution and use the log link function.

2. Although, we did not talk specifically about "causal diagrams", I have given answers to parts (a) and (d).

   (a) This part is essentially asking you how do you believe the variables are related to each other (which variables cause changes in other variables).

   We may argue that due to advances in building cars and roads we would expect changes in both RATE and SPEED as YEAR increases. Further, we may argue that the oil embargo caused speed limits to be lowered thus reducing SPEED. Also it is reasonable to assume that SPEED has an effect on RATE. Thus we have YEAR affecting both SPEED and RATE and EMB affecting SPEED, and SPEED affecting RATE.

   (b) not applicable.

   (c) not applicable.

(d) It would make sense to carry out a Durbin-Watson test since the data was collected over time (serial correlation is possible in this case).

3. (a) The P-values for **pen** differ because the terms are added in different orders for the two tables. In the first table **pen** is added last so the P-value corresponds to the hypothesis H_0: the different pens do not affect the response given that we have taken into account the effects of the other 3 factors (**x**, **food**, and **sex**). The second table **pen** is added after **x** but before **food**, and **sex** and so the P-value corresponds to H_0: the different pens do not affect the response given that we have taken into **x** (weights at the beginning of the experiment) but ignoring the effects of **food**, and **sex**.

   (b) The purpose of this experiment was to compare the different feeding treatments. The other 3 variables were included to reduce the variability and thus make these comparisons more precise. We are interested in comparing feeding treatments taking into account the effects of **x**, **pen**, and **sex**. The line for **food** in table 2 tests H_0 the different feeding treatments do not affect the response given we take into account the other 3 regressors.

   (c) Note: if I was to ask this type of question I would give you the output from "dummy.coef".

```
> dummy.coef(lm(y~x+pen+sex+food))
$"(Intercept)":
 (Intercept)
     5.53694


$x:
          x
 0.09127417


$pen:
 1         2          3         4         5
 0 0.5542283 -0.1767501 0.4626962 0.4833144


$sex:
 M      F
 0 0.4293


$food:
 A          B      C
 0 -0.4431468 -0.673
```

The predicted growth rate for a male pig in pen 1 with an intial weight of 45 is:

$$\begin{aligned}\hat{y} &= 5.537 + 0.09127(45) + 0 + 0 - 0.4431 \\ &= 9.20\end{aligned}$$

You can do the same calculation from the output to summary.

(d) The most important assumption about the *form* of the model is that there are no interactions between the regressors. The best way to check this is to fit the model that contains all the 2 factor interactions and see if any are significant. (You should also check for constant variance, Normal errors, outliers, etc. but I don't think that is what this question is asking).

4. (a) The data would be put into 3 columns. The first column would contain the purity index values, the second would be a factor indicating whether the standard or modified process was used, and the third would be the binary response where 1 =fault and 0 = no fault.

```
purity   process   y
7.2      stand     0
7.2      mod       0
6.3      stand     1
6.3      mod       0
```

(b) All these models are modelling the probability of a fault.

Model 1 only uses the process as an explanatory variable. The estimated coefficient would indicate the probability of a fault is less for the modified process but the standard error is nearly as large as the estimated coefficient. The P-value indicates the difference is not statistically significant.

Model 2 only uses purity as an explanatory variable (same line model). The estimated coefficient indicates that as purity increases the probability of a fault decreases. The standard error is small enough that this effect is statistically significant.

Model 3 uses both the process and purity as explanatory variables but no interaction (parallel lines model). The coefficients for both process and purity have the same signs as for Model 1 or 2 and so the same interpretations are valid. Again purity is significant but process is not.

Model 4 uses both the process and purity as explanatory variables and includes their interaction(separate lines model). The coefficient for processMod indicates that the intercept for the modified process is 4.3379 less than that for the standard process but the standard error is large enough that this is not significant. The coefficient for the interaction indicates the change in the coefficient for purity between the standard and modified processes. For the standard process the coefficient for purity is -0.8684 whereas for the modified process it is $-0.8684 + 0.4744) = -0.394$. However, this difference is not significant.

(c) We have no evidence that the the probability of a fault is different for the two processes. None of the coefficients involving process for our fitted models were significantly different from 0 so we cannot eliminate the possibility the 2 processes have the same probability of a fault. However, the fitted models do indicate that our best guess is that the probability of a fault is smaller for the modified process – we may be able to obtain evidence of this if we were to collect more data.

(d) For model 3;

$$\begin{aligned}\text{logit}(\hat{\pi}) &= 4.4608 - 0.6042(8) = -0.3728 \\ \hat{\pi} &= \exp(-0.3728)/(1 + \exp(-0.3728)) = 0.408\end{aligned}$$

5. (a) I would select

```
Y~C+S+I+S:I
```

Using the Chi-square table we can see that adding S:I to the model C + S + I results in a significant reduction in residual deviance. Adding either C:I or C:S to the model C + S + I + S:I does not result in a significant reduction in deviance.

(b) This model indicates that level of satisfaction is related to level of influence but not to level of contact with other residents. Level of influence is not related to level of contact with other residents.

(c) For this data we need the model with all three interactions in it:

```
Y~C+S+I+C:I+C:S+S:I
```

If we start at the bottom of the table we conclude it is not feasible to drop any of the interactions (there is a significant increase in residual deviance if we do).

(d) The model in (c) indicates that all three factors are related to each other for tower block dwellers. Satisfaction is related to both the level of influence on management and the level of contact with other residents. And level of influence on management is related to level of contact with other residents.