MATH 204 EXAMINATION 2006

SOLUTIONS

1. For this question I will use the following notation:

`store`	:	for the store factor predictor
`shelf`	:	for the shelf space as a factor predictor
`shelfx`	:	for the shelf space treated as a continuous covariate

Using this notation, the SPSS output corresponds to the following models (in the same order as the output starting at the bottom of page 6). The quantity k is the total number of parameters fitted in the model apart from the intercept, and EDF is the error degrees of freedom. Here $n = 36$, and note that

$$\text{EDF} = n - k - 1 \qquad \therefore \qquad k = n - \text{EDF} - 1$$

	Model	k	EDF	SSE
M0	Null	0	35	7661.639
M1	`store`	5	30	1469.500
M2	`store` + `shelf`	10	25	1307.694
M3	`shelfx`	1	34	7516.379
M4	`store` + `shelfx`	6	29	1324.240
M5	`store` + `shelfx` + `store .shelfx`	11	24	1223.314

For this question, we have no ANOVA tables, so we will use the method of ANOVA-F testing for nested models, and the test statistic

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

where SSE_R is the error sum of squares for the **Reduced Model**, specified using $g+1$ parameters including the intercept, and SSE_C is the error sum of squares for the **Complete Model**, specified using $k+1$ parameters including the intercept.

If the reduced model is an adequate simplification of the complete model, then

$$F \sim \text{Fisher-F}(k-g, n-k-1)$$

Note here that

$$k - g = (n - g - 1) - (n - k - 1) = \text{EDF}_R - \text{EDF}_C$$

so the $k - g$ quantities can be deduced directly from the EDFs.

(a) We wish to assess the effect of `shelf` (ie shelf-space as a factor predictor), allowing for the effect of `store` . This suggests that we should compare the nested models M1 and M2, with M2 as the Complete model. Therefore we compare

$$F = \frac{(1469.500 - 1307.694)/(10-5)}{1307.694/25} = 0.619$$

with the Fisher-F$(k-g, n-k-1) \equiv$ Fisher-F$(5, 25)$ distribution. From the tables on page 14, Fisher-F$_{0.05}(5, 25) = 2.60 > 0.619$. Therefore we have no reason to reject M1 as an adequate simplification of M2, and conclude that there is **no evidence** that shelf space (when fitted as a factor predictor) has an effect on sales.

(b) The list of post-hoc test p-values (marked as "Sig." in the final column in the Multiple Comparisons table on page 7) reveals that there are **no** significant differences between any of the levels - all p-values are greater than 0.05.

(c) We wish to assess the effect of `shelfx` (ie shelf-space as a covariate), allowing for the effect of `store` . This suggests that we should compare the nested models M1 and M4, with M4 as the Complete model. Therefore we compare

$$F = \frac{(1469.500 - 1324.240)/(6-5)}{1324.240/29} = 3.181$$

with the Fisher-F$(1, 29)$. From the tables on page 14, Fisher-F$_{0.05}(1, 29) = 4.18 > 3.181$. Therefore we have no reason to reject M1 as an adequate simplification of M4, which implies that there is **no evidence** that shelf-space (when fitted as a continuous covariate) has an effect on sales. For completeness we also compare the nested models M1 and M5 using

$$F = \frac{(1469.500 - 1223.314)/(11-5)}{1223.314/24} = 0.805$$

but as Fisher-F$_{0.05}(6, 24) = 2.51 > 0.805$, again, M1 is an adequate simplification.

Thus it appears that shelf-space is not a useful variable for explaining sales, whether it is fitted as a factor predictor or a covariate.

(d) There is no discrepancy; all parts give the same conclusion. For model M4, the table on page, we see a p-value of 0.085 attached to the coefficient of `shelfx` , which is almost significant at $\alpha = 0.05$, but not quite. The results indicate that a linear relationship between sales and shelf-space is implied, but is not statistically significant. Note that we cannot easily compare M2 with M4, as the models are not nested.

Note: I do not know what the official solution to this question is; I cannot find any evidence in the data when analyzed as above. Clearly the model where a shelf-space is fitted as a covariate produces an almost significant result - fewer parameters are being fitted - but the result is still not significant at $\alpha = 0.05$

(e) To answer the question of whether the effect of shelf-space as a linear covariate changes from store to store, we compare models M4 and M5. Therefore we compare

$$F = \frac{(1324.240 - 1223.314)/(11-6)}{1223.314/24} = 0.396$$

with the Fisher-F$(5, 24)$. From the tables on page 14, Fisher-F$_{0.05}(5, 24) = 2.62 > 0.396$. Therefore we have no reason to reject M4 as an adequate simplification of M5, which implies that there is **no evidence** that shelf-space (when fitted as a continuous covariate) has a different effect on sales in different stores

(f) In my opinion, the most appropriate model is model M1, which fits a **constant but different mean level for each store**. The table on page 6, because of the SPSS contrast parameterization that is adopted, we combine the estimates for the intercept and for store 1 to get the prediction, that is, predicted sales are $19.167 + 11.500 = 30.667$.

The estimated standard deviation, s, yields the standard error of the prediction **within group** as $s/\sqrt{6}$; in the general linear model,

$$s^2 = \frac{SSE}{EDF} = \frac{1469.500}{30} = 48.983$$

so the standard error of the prediction is $\sqrt{48.983/6} = 2.857$; note that this can be deduced by using the arbitrariness of the store labelling - in this balanced case, the standard error of the prediction for any one group mean must be the same as for all other group means.

(g) It it not possible to fit the interaction between `store` and `shelf` as we do not have sufficient data - the number of replicates in each of the factor-level combinations is one. Therefore the interaction model if fitted will yield a sum of squared errors and error degrees of freedom equal to zero.

2. The design of the experiment means that a **two independent samples** analysis is required, as a random sample of advertisements is selected from each publication. Given the tables provided, we will carry out a two sample **Mann-Whitney-Wilcoxon** test; specifically, we will use the small sample version based on the Wilcoxon statistic, W, the sum of the ranks for observation from publication 2. Here the sample sizes are even, so we can label the samples arbitrarily; we take the *Newsweek* data as observations from population 2.

   Thus $n_1 = n_2 = 6$, so $n = 12$. The ranks are computed as below

Group	1	1	1	1	1	1	2	2	2	2	2	2
y	8.20	9.23	9.92	11.16	11.55	15.75	3.12	4.88	5.12	7.67	9.66	10.21
Rank	5	6	8	10	11	12	1	2	3	4	7	9

   Hence $R_1 = 52$, $R_2 = 26$. Thus, according to our convention, $W = 26$.

   Using the tables provided, we find the critical values

$$T_L = 26 \qquad T_U = 52.$$

   and thus $W = T_L$, and W lies within the lower segment of the rejection region. For the hypothesis

$$H_0 \ : \ \eta_1 = \eta_2$$

   we

$$\begin{aligned} \textbf{Reject } H_0 \text{ against } H_a \ &: \ \eta_1 > \eta_2 \quad \text{as} \quad W \leq T_L \\ \textbf{Do not reject } H_0 \text{ against } H_a \ &: \ \eta_1 < \eta_2 \quad \text{as} \quad W < T_U \\ \textbf{Reject } H_0 \text{ against } H_a \ &: \ \eta_1 \neq \eta_2 \quad \text{as} \quad W \leq T_L \end{aligned}$$

   at the $\alpha = 0.025, 0.025$ and 0.05 significance levels respectively.

   Thus we reject the hypothesis of equal population medians, or more generally the hypothesis of equal distributions, in favour of the hypothesis that advertisements in *Scientific American* are more difficult to read than those in *Newsweek*.

3. This question addresses the analysis of categorical data in a contingency table, and the use of a Chi-squared test for independence.

   (a) The expected number of cases under the assumption of independence is

   $$\widehat{n}_{ij} = \frac{n_{i.} n_{.j}}{n}$$

   where $i = 4, j = 3$, and $n_{i.}$ and $n_{.j}$ are the row and column sums for row i and column j respectively. Here, from the SPSS output on page 9, $n_{4.} = 185, n_{.3} = 106$, and $n = 400$, so

   $$\widehat{n}_{43} = \frac{185 \times 106}{400} = 49.025.$$

   Note that the table on the output is rounded to one decimal place; note also that the expected counts totals in rows and columns must equal the actual row and column totals. Hence to this level of approximation, we could find $\widehat{n}_{43}$ using the third row as

   $$\widehat{n}_{43} = 185 - 104.5 - 31.5 = 49.0$$

   (b) The degrees of freedom missing from the table is

   $$(r-1)(c-1) = 3 \times 2 = 6.$$

   Using the table on page 13, the $\alpha = 0.0005$ tail quantile for the Chisquared(6) distribution is 18.55. Therefore we **reject** the null hypothesis of independence, and conclude that there is a difference between the incidence of melanoma at the different sites.

4. (a) Using the first table on page 11, we can conclude by inspecting the t-statistics and p-values that both `Dist` and `Days` are significant in the model. There seems to be no significant multicollinearity between these variables. Thus the simplest model we will consider is

$$\texttt{Dist} + \texttt{Days}$$

To test whether `Month` should be added, we use the ANOVA-F testing procedure for nested models; as

$$k - g = (n - g - 1) - (n - k - 1)$$

we have

$$F = \frac{(106.080 - 65.417)/(83 - 72)}{65.417/72} = 4.069$$

to be compared with the Fisher-F$(k - g, n - k - 1) \equiv$ Fisher-F$(11, 72)$ distribution. From the tables on page 14, Fisher-F$_{0.05}(11, 72) \doteq$ Fisher-F$_{0.05}(10, 80) = 1.95 < 4.069$. Therefore we reject the reduced model as an adequate simplification, and conclude that `Month` should be included, that is, the simplest adequate model checked so far is

$$\texttt{Dist} + \texttt{Days} + \texttt{Month}.$$

We call this model $\mathrm{M_a}$.

(b) When `Temp` replaces `Month` , the same method of comparison gives

$$F = \frac{(106.080 - 78.313)/(83 - 82)}{78.313/82} = 29.074$$

to be compared with the Fisher-F$(k - g, n - k - 1) \equiv$ Fisher-F$(1, 82)$ distribution. From the tables on page 14, Fisher-F$_{0.05}(1, 82) \doteq$ Fisher-F$_{0.05}(1, 80) = 3.96 < 29.074$. Therefore we reject the reduced model as an adequate simplification, and conclude that `Temp` should be included, that is, the model

$$\texttt{Dist} + \texttt{Days} + \texttt{Temp}$$

also fits better than the model `Dist + Days` . This is also confirmed by the p-value for `Temp` in the regression model ($p < 0.001$). We denote this model $\mathrm{M_b}$.

(c) Looking at the original data table on page 9, it is clear that the average monthly temperature provides **identical** information to month number. That is, `Temp` is an exact (linear) function of month. Looking at the sums of squared errors, we see that for both the models

$$\texttt{Dist} + \texttt{Days} + \texttt{Month}$$

and

$$\texttt{Dist} + \texttt{Days} + \texttt{Temp} + \texttt{Month}$$

have identical SSE values (65.417) for each model. Thus `Temp` explains all the effect of `Month` ; we have that the covariates are perfectly multicollinear - this also shows up in the scatterplot on page 10.

(d) It is hard to assess this graph, but it seems that there might be (i) some outliers (extremely large in magnitude residuals) and (ii) a slight positive incline in the plot of predicted values versus residuals. Thus the assumption of Normality might be inappropriate, and the assumption of constant variance might also be inappropriate. It also seems that the residuals are patterned, in that they seem more likely to be negative when `KmPerl` is small, and positive when `KmPerl` is large, which indicates a deficiency in the model, although not necessarily in the assumptions about the random error.

(e) The proportion of variation explained by model $\mathrm{M_b}$ is

$$R^2 = \frac{SSE_{\text{Null}} - SSE_{\mathrm{M_b}}}{SSE_{\text{Null}}} = \frac{184.798 - 65.417}{184.798} = 0.646$$

(f) The best fitting model is either $\mathrm{M_a}$ or $\mathrm{M_b}$. We assess this by inspecting the adjusted R^2 quantities

$$\text{Adj } R^2_{\mathrm{M}} = 1 - \frac{SSE_{\mathrm{M}}/(n-k-1)}{SSE_{\text{Null}}/(n-1)}$$

For Model $\mathrm{M_a}$:

$$\text{Adj } R^2_{\mathrm{M_a}} = 1 - \frac{SSE_{\mathrm{M_a}}/(n-k-1)}{SSE_{\text{Null}}/(n-1)} = 1 - \frac{65.417/72}{184.796/85} = 0.582$$

For Model $\mathrm{M_b}$:

$$\text{Adj } R^2_{\mathrm{M_b}} = 1 - \frac{SSE_{\mathrm{M_b}}/(n-k-1)}{SSE_{\text{Null}}/(n-1)} = 1 - \frac{79.313/82}{184.796/85} = 0.555$$

Thus the prediction is found by using the formula with estimates from model $\mathrm{M_a}$

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \texttt{Dist} + \widehat{\beta}_2 \texttt{Days} + \widehat{\beta}_3 \texttt{Month}$$

where `Month` is month 4, and $\widehat{\beta}_3$ is the coefficient estimated for that level of the factor predictor. From the table on page 11, we have

$$5.676 + (0.004 \times 300) + (-0.111 \times 0) + 1.131 = 8.007$$

In reality, it is **not possible** to report a standard error for this prediction. For **independent** random variables $X_1, \ldots, X_k$ and constants $a_1, \ldots, a_k$, the formula

$$\text{Var}\left[\sum_{i=1}^{k} a_k X_k\right] = \sum_{i=1}^{k} a_k^2 \text{Var}[X_k]$$

gives the variance of the sum in terms of the sum of the variables. However, this result does not work for **dependent** variables, so the calculation

$$\text{s.e.}(\widehat{y}) = \sqrt{0.544^2 + 300^2 \times 0.001^2 + 0.622^2} = 0.879$$

that takes the sums of the squared standard errors appropriately scaled **is not valid** here, as the coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ are probabilistically dependent.

(g) There are two redundant parameters in the estimation of `Month` factor level coefficients.

- The first is redundant as it is essentially reported as the "Intercept" in the first row of the tables. SPSS sets the baseline group as the highest labelled factor level, and uses a contrast parameterization, that is, estimates differences from baseline. The intercept is reporting the mean level of the baseline group.
- The second parameter is redundant as `Month` is perfectly multicollinear with `Temp` . There are only twelve different parameters to represent the effect of the twelve months; the model

  $$\texttt{Dist} + \texttt{Days} + \texttt{Temp} + \texttt{Month}$$

  tries to fit fourteen (one intercept, one for `Temp` and twelve for month). Thus the software is forced to set two parameters to zero.