STATISTICS 210 – Fall 2009 Homework 8

Handed out with solutions: Friday December 4, 2009

Reading: Dec. 4 Factorial experiments (overview) (Ch. 19-20, 23-24)

NOTE: Final Exam – Last year's final exam is posted on the website (data analysis and in-class portions). This year's exam is Friday December 11 at 8am in the usual classroom.

NOTE: Solutions to this homework are available on the course website.

1. **Model building and interpretation:** The output below shows results for three models (labelled MODEL I, II, and III) fit to the state data on average SAT scores. Recall that the response variable is average SAT score in a state and the most significant explanatory variables include yrs (average years of coursework), rank (average rank of students), and expend (per pupil expenditure in the state). In this question we consider the effect of region. We have created 4 indicator variables (east, south, midwest, west) so that for example east = 1 if a state is in the East region and 0 otherwise. We also create interactions of the 4 indicator variables with "expend" (the names are eexp, sexp, mwexp, wexp). Model I ignores region.

   (a) Model II only includes 3 of the region indicators. Why is the 4th indicator omitted? Explain.

   (b) How would you interpret each of the region coefficients in Model II.

   (c) Are there significant differences in SAT scores across regions after accounting for yrs, rank and expend? Carry out a statistical test that answers this question. Interpret your result.

   (d) Model III introduces the region-expend interaction terms. Interpret the meaning of the coefficients in this model.

   (e) Are there significant differences in the SAT score/expenditure slope across regions after accounting for yrs and rank? Carry out a statistical test that answers this question. Interpret your result.

MODEL I

Analysis of Variance

Source	DF	Sum Squares	Mean Square	F Value	Pr > F
Model	3	214303	71434	103.63	<.0001
Error	46	31708	689.30877		
Corrected Total	49	246011			

Parameter Estimates

Variable	DF	Param.Estim.	Std.Error	t Value	Pr > |t|
Intercept	1	-303.72430	97.84154	-3.10	0.0033
yrs	1	26.09523	5.38945	4.84	<.0001
expend	1	1.86087	0.63511	2.93	0.0053
rank	1	9.82579	0.59870	16.41	<.0001

MODEL II

Analysis of Variance

Source	DF	Sum Squares	Mean Square	F Value	Pr > F
Model	6	224442	37407	74.58	<.0001
Error	43	21569	501.60183		
Corrected Total	49	246011			

Parameter Estimates

Variable	DF	Param.Estim.	Std.Error	t Value	Pr > |t|
Intercept	1	-172.40201	88.61110	-1.95	0.0583
east	1	12.82253	13.81044	0.93	0.3583
south	1	-23.93399	10.19752	-2.35	0.0236
midwest	1	18.32333	9.54798	1.92	0.0616
yrs	1	19.21123	5.66607	3.39	0.0015
expend	1	0.74578	0.64282	1.16	0.2524
rank	1	9.89272	0.74882	13.21	<.0001

MODEL III

Analysis of Variance

Source	DF	Sum Squares	Mean Square	F Value	Pr > F
Model	9	226772	25197	52.39	<.0001
Error	40	19239	480.97198		
Corrected Total	49	246011			

Parameter Estimates

Variable	DF	Param.Estim.	Std.Error	t Value	Pr > |t|
Intercept	1	-139.26223	95.88586	-1.45	0.1542
east	1	-24.14804	46.48210	-0.52	0.6063
south	1	-116.54555	45.46911	-2.56	0.0142
midwest	1	-33.65744	55.96162	-0.60	0.5509
yrs	1	15.20484	6.09900	2.49	0.0169

```
rank        1    10.50864    0.79493   13.22   <.0001
expend      1    -0.03665    0.76775   -0.05   0.9622
eexp        1     1.78760    1.83056    0.98   0.3347
sexp        1     4.91571    2.40208    2.05   0.0473
mwexp       1     2.21139    2.40891    0.92   0.3641
```

2. **Two-factor factorial:** The data set that we consider here is from an experiment regarding childrens' memory. A random sample of 36 fourth-graders from one city were used in the experiment. Two factors are varied: type of reinforcement given during learning (none or verbal) and time of isolation (20, 40, or 60 minutes) before memory testing. Students were told to memorize a paragraph and given positive verbal reinforcement or no reinforcement while learning it according to their treatment assignment. Then students were isolated for the specified amount of time. There were 6 students randomly assigned to each of the six treatment groups. The response is a score measuring the student's memory for the learned paragraph. The data are contained in the file "paragraphdata.txt" (or paragraphdata.xls) with the first column indicating the level of reinforcement (none or verbal), the second column indicating the isolation time (20, 40, 60), and the third column giving the observed memory score. The first column is nonnumeric; the first row contains variable labels. Also a sample SAS program for a two-factor experiment is provided on the web site.

   (a) Plot the response versus isolation time. On this plot indicate the location of the mean response for each treatment combination. Connect the means for all treatments with no reinforcement (and then connect the means for all treatments with verbal reinforcement). Based on the graph what significant effects do you expect to find? (You can find the means using the MEANS REINF ISOL REINF*ISOL; command in PROC GLM where REINF and ISOL are the names of the two factors.)

   (b) Obtain the analysis of variance table for the two-factor factorial model. Which effects are significant at the .05 level?

   (c) Check the assumptions of normality and constant variance using the residuals.

   (d) Interpretation:

      i. Evaluate the significance of linear (-1, 0, 1) and quadratic (-1, 2, -1) contrasts for the main effect of isolation time.

      ii. Explain why there is no need to consider a contrast for the main effect of reinforcement time.

      iii. How would you explain the interaction to a non-quantitative memory researcher.

   (e) Briefly summarize your results. How do reinforcement and isolation time effect memory? Also, describe the population for which you think these results are relevant?

3. **Three-factor factorial:**
   The table given to the right provides the results of a 3 x 3 x 2 greenhouse experiment to determine the rate of emergence of seeds from three species of legume, in three soil types, either treated or not treated with a fungicide. There are **three observations** for each combination of factors; the observed response is the number of plants out of 100 that had emerged (begun to grow) when the experiment ended.

Source	SS	d.f.	MS
species	9900.1	2	4950.1
soil	16436.1	?	?
fungicide	1932.0	?	?
spec.soil	658.4	?	?
spec.fung	194.0	?	?
soil.fung	1851.1	?	?
threeway	1069.7	4	267.41
error	3556.2	?	?
Total	35597.7	53	—

   (a) Complete the ANOVA table. Which effects are significant at the .05 level?

   (b) A table of means for each of the 18 treatment combinations is provided below. Let's use these means to understand the significant effects.

		Soil Type		
Species	Fungicide	Silt loam	Sand	Clay
Alfalfa	No	89	95	22
Alfalfa	Yes	92	90	72
Red clover	No	84	96	56
Red clover	Yes	92	97	68
Sweet clov.	No	51	66	17
Sweet clov.	Yes	59	73	40

      i. The main effects of species, soil and fungicide are significant. Use the table above to identify which species seem to emerge unusually fast or slow, which soil types are most conducive to emergence of legumes, and whether fungicides increase the rate of emergence. (Hint: You may want to compute means for each factor separately to answer this.)

ii. Which two factor interaction was most significant in the ANOVA table? Interpret this two factor interaction. (Hint: A table or picture may help).

4. **Unbalanced data:** The data in the file "uvfrogsdata.txt" (or uvfrogsdata.xls) are from an experiment to determine whether UV (ultraviolet) radiation plays a role in the disappearance of a number of species of frogs. The experiment is a 3x3 factorial design – there are three species of frog (identified by the numbers 7.5, 2.4, 1.3 in the data set – the numbers will be explained later) and three treatments (trt = 1 - no filter, trt = 2 - a filter that allows UV radiation through, trt = 3 - a filter that blocks UV radiation). For each treatment, four enclosures weres set up, each containing 150 eggs from the species. The response is the percent of frog eggs failing to hatch. (Note there is an additional column in the file for location which we don't use in this problem.) We expect the percent of eggs failing to hatch to be greatest when there is no filtering of UV (if the investigators hypothesis is correct). Note that treatment 2 is a sort of placebo ... a filter is put in place but it doesn't block out the UV radiation. The problem here is that the data are not balanced. Several of the enclosures were lost (perhaps to predation or disease). There are between 2 and 4 observations per treatment combination (23 observations in all instead of the planned 36).

   (a) Carry out the usual ANOVA analysis using PROC GLM (the generalized linear model command in Minitab).

      i. Do the type I (sequential) and type III (partial) sums of squares match in this case (as they do for balanced data)? Explain.

      ii. Which factors are found to be significant? What are the p-values?

   (b) Obtain a table of means (for each treatment combination and or each factor separately) and then:

      i. Identify one contrast of interest for the treatment main effect and test its significance.

      ii. The species are identified by numbers which indicate the presence of enzymes that repair DNA damage by UV radiation. Thus species with more enzyme should do better in general. Test a linear contrast in the species main effect.

5. **Two-factor factorial with blocking:** For each of twelve commercial fishing trips, the number of each of four types of small fish caught are recorded. The four categories correspond to two species, A and B, and two genders for each species. The twelve trips are treated as blocks since the amount caught of any one type will depend on the success of the trip as a whole. Means and standard deviations are given below along with an ANOVA table. The analysis was repeated twice, first with the actual measurements (in thousands) and then using the logarithms (base 10) of the measurements. Only the second results are provided here.

	Species A		Species B	
	Male	Female	Male	Female
mean	2.802	3.221	3.962	4.480
s.d.	0.150	0.098	0.099	0.099
n	12	12	12	12

ANOVA		
Source	d.f.	SS
species	1	17.545
sex	1	2.632
spec*sex	1	0.030
block	11	0.339
error	33	0.230

   (a) In the first ANOVA (with untransformed data) all of the "treatment" effects are highly significant. However the s.d. in the four "treatment" groups were $0.234, 0.357, 2.326, 6.689$. Do you trust the results of the first ANOVA? Explain.

   (b) Does the use of logarithms solve the problem that is mentioned in (a). Explain.

   (c) Determine which effects are significant on the logarithmic scale.

   (d) Explain the meaning of the interaction in this problem (regardless of whether it is significant or not).

   (e) Was the blocking effective? Explain.