STAT 425 — Spring 2016 — Sections 1UG, 1GR

Homework 5

Please submit your assignment *on paper*, following the Formatting Guidelines for Homework Submission. (Even if correct, answers might not receive credit if they are too difficult to read.) Remember to include relevant computer output.

1. Consider the `ais` data set (in package `alr4`), which contains data on athletes from the Australian Institute of Sport. We will be using only the variables `Wt` (weight, kg), `Ht` (height, cm), and `Sex` (0 = male, 1 = female).

   (a) [2 pts] Consider the simple linear regression of `Wt` on `Ht` (formula `Wt ~ Ht`). Using the `boxcox` function from the package `MASS`, what (simple) transformation of `Wt` is suggested by the Box-Cox procedure? (Show the plot and also state what value of λ you would choose.)

   (b) [2 pts] Compute the RSS value (sum of the squared residuals) for the following two models: (i) the simple linear regression of `log(Wt)` on `Ht` (formula `log(Wt) ~ Ht`) and (ii) the simple linear regression of `log(Wt)` on `log(Ht)` (formula `log(Wt) ~ log(Ht)`). Which has the smaller RSS?

   (Remark: When transforming an independent variable, the transformation minimizing RSS is generally the one that should be chosen. See Weisberg 4th, Section 8.1.2.)

   (c) [2 pts] Fit the regression of `log(Wt)` on `log(Ht)`, `Sex`, *and* the interaction between `log(Ht)` and `Sex`. Give a summary of the results. Is the interaction significant?

   (d) [2 pts] Plot `log(Wt)` versus `log(Ht)` with different symbols for the males and the females. Then plot the two (*not* parallel) regression lines representing the relationship between `log(Wt)` and `log(Ht)` for the male and female athletes, according to your model from the previous part that includes an interaction. (The two lines should be on the *same* plot as the points.)

   (e) [2 pts] Fit the regression of `log(Wt)` on `log(Ht)` and `Sex`, *without* interaction. Give a summary of the results. Is `Sex` significant?

   (f) [2 pts] Plot `log(Wt)` versus `log(Ht)` as before, then plot the two (parallel) regression lines representing the relationship between `log(Wt)` and `log(Ht)` for the male and female athletes, according to your model from the previous part. Which line is higher: the male or the female?

2. The data set `turk0` (in package `alr4`) contains the results of an experiment with a completely randomized design: 35 pens of turkeys were randomly allocated into 6 groups, and each group was given a different level of supplementation with methionine.[1] You are to analyze how the average weight `Gain` (grams) of the turkeys in a pen depends on the amount `A` of methionine supplement that was assigned to the pen.

   (a) [2 pts] Is the design *balanced*? How do you know? (Hint: Examine the structure of the data. How many experimental units are in each treatment group?)

[1] See Weisberg 4th, Sec. 11.3 for a more complete description.

(b) [2 pts] Fit an appropriate ANOVA model. (Note: You will need to convert `A` into a factor variable.) Display a summary of your results.

(c) [2 pts] Produce the usual diagnostic plots for your model, and draw conclusions.

(d) [2 pts] Produce an ANOVA table.

(e) [2 pts] Test whether there are any differences among the mean weight gains of the groups (based on an F-test at $\alpha = 0.05$).

(f) [2 pts] Produce Tukey simultaneous 95% confidence intervals for all mean differences between pairs of groups.

(g) [2 pts] According to your Tukey intervals, *which pairs* of methionine levels have significantly different means (after adjusting for multiple comparisons)? (List the pairs.)

3. The data in the file `pine.dat`[2] are from an experiment to investigate how production of pine oleoresin, obtained by tapping pine trees, is affected by `shape` of the hole (1=circular, 2=diagonal slash, 3=check, 4=rectangular) and whether or not acid treatment (`trt`) was used (1=no, 2=yes). The experiment was performed with 24 pine trees as experimental units, in a completely randomized design. The response `y` is the amount (g) of resin collected from an individual tree.

   (a) [2 pts] Examine the data. How many treatment groups are there? How many experimental units are in each treatment group?

   (b) [2 pts] Fit the linear model appropriate for analysis of this experiment, with `y` as the response. Display a summary of the results. (Note: You will need to convert `shape` and `trt` into factor variables.)

   (c) [2 pts] Produce and interpret the usual diagnostic plots for your model. Do you notice any problems?

   (d) [2 pts] Perform a Box-Cox analysis on your model. (Show the graph produced by the `boxcox` function.) What simple transformation seems most appropriate?

   (e) [2 pts] Fit the linear model appropriate for analysis of this experiment, with `sqrt(y)` as the response. Display a summary of the results.

   (f) [2 pts] Create an interaction plot (with the square root of `y` as the response). Use `shape` as the x-axis factor.

   (g) [2 pts] Produce an ANOVA table for your model of part (e).

   (h) [2 pts] Using your ANOVA table, test for interaction effects. (What is your conclusion?)

   (i) [2 pts] *If it is appropriate*, test for the main effects. *If that is NOT appropriate*, briefly explain why not.

4. [GRADUATE SECTION ONLY] Consider the means model for a one-way ANOVA: $Y_{ij} = \mu_i + e_{ij}$ where the errors are independent and follow the distribution $e_{ij} \sim N\left(0, \sigma^2\right)$. Index i indicates the group, ranging from 1 to T, and index j the observation within the group, ranging from 1 to n_i. The total number of observations is $N = \sum_{i=1}^{T} n_i$.

[2]From Oehlert (2000) *A First Course in Design and Analysis of Experiments*, New York: W. H. Freeman.

Statistics and parameters associated with each group are given in the following table:

	Group			
	1	2	$\cdots$	T
Size	n_1	n_2	$\cdots$	n_T
Sample Mean	$\bar{Y}_1$	$\bar{Y}_2$	$\cdots$	$\bar{Y}_T$
Sample Variance	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\cdots$	$\hat{\sigma}_T^2$
Population Mean	μ_1	μ_2	$\cdots$	μ_T
Population Variance	σ^2	σ^2	$\cdots$	σ^2

(Assume that $n_i > 1$ for all i.)

(a) [2 pts] Fully specify the distribution of $\bar{Y}_i - \bar{Y}_k$, where $i \neq k$.

(b) [2 pts] Let Z_{ik} be the standardized version of $\bar{Y}_i - \bar{Y}_k$. What is the expression for Z_{ik}? (Hint: Subtract the mean and divide by the standard deviation.)

(c) [2 pts] Using the fact that

$$U_i \;=\; (n_i - 1)\,\frac{\hat{\sigma}_i^2}{\sigma^2} \;\sim\; \chi^2_{n_i-1}$$

and $U_1, \ldots, U_T$ are independent (because the groups are independent), show that

$$(N - T)\,\frac{\hat{\sigma}^2}{\sigma^2} \;\sim\; \chi^2_{N-T}$$

where $\hat{\sigma}^2 = \frac{1}{N-T}\sum_{i=1}^{T}(n_i - 1)\,\hat{\sigma}_i^2$.

(d) [2 pts] Since Z_{ik} and $\hat{\sigma}^2$ are independent, confirm that

$$T_{ik} \;=\; \frac{\bar{Y}_i - \bar{Y}_k - (\mu_i - \mu_k)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}} \;\sim\; t_{N-T}$$

(e) [2 pts] The previous part implies that $-t_{\alpha/2,N-T} < T_{ik} < t_{\alpha/2,N-T}$ with probability $1 - \alpha$. Use this fact to find a and b such that, with probability $1 - \alpha$,

$$a \;<\; \mu_i - \mu_k \;<\; b.$$

Some reminders:

- Unless otherwise stated, all data sets are either automatically available or can be found in either the `alr4` package or the `faraway` package in R.
- Unless otherwise stated, use a 5% level ($\alpha = 0.05$) in all tests.