Assignment Analysis of Variance

A report containing solutions to the problems below is to be uploaded in the Blackboard as a single pdf-file, due no later than the deadline for this assignment. The report should be well-organized, concise, well readable, including relevant plots, tables, R-code and conclusions.

I. Computational problems

1. Generate two data sets of lengths n_1 and n_2, respectively, from two normal distributions having the same variance σ^2 but different means μ_1 and μ_2, respectively. Choose suitable values for n_1, n_2, σ^2, μ_1 and μ_2.

   (i) Use the `R` function `t.test` with the setting `var.equal=TRUE` to conduct a two-sample t-test for equality of the population means.

   (ii) Next, conduct a one-way ANOVA on the same data with a factor having two levels (corresponding to the two data sets). Obtain the ANOVA table.

   (iii) Compare the p-values and the values of the test statistics for the two different tests. Determine (and explain) the relationship between the two-sample t-test and the one-way ANOVA F-test with two factor levels.

   (iv) Suppose you know the variance σ^2. Can you use this knowledge to improve your testing procedure(s)? If so, demonstrate this by performing relevant `R`-calculations.

2. The dataset `chickwts` is a data frame included in the standard `R` installation. To view it, just type `chickwts` at the `R` prompt. This data frame contains 71 observations on newly-hatched chicks which were randomly allocated among six groups. Each group was given a different feed supplement for six weeks, after which their weight (in grams) was measured.

   (i) Use `R` to make side-by-side boxplots of the distribution of chick weights with respect to the feed supplement. Compare the plots and comment on what you observe.

   (ii) Use `R` to conduct a one-way ANOVA (with appropriate constrain) on the chick `weight` with respect to the six feed supplements, *without* using the `aov` or `lm` functions. That is, create the response vector $\mathbf{Y}$ and the incidence matrix $\mathbb{X}$ in `R`. Then use matrix operations in `R` to obtain the least-squares estimates of the parameters (also an unbiased estimator of σ^2) and obtain all the quantities needed to complete an ANOVA table.

   (iii) Based on the table, conduct a test to determine whether the factor feed supplement influences the chick weight. Use the `aov` (or `lm`) and `anova` functions to perform the same analysis and compare your results.

   (iv) Finally, check the model assumptions by using relevant diagnostic tools.

3. Hemoglobin is measured (g/100 ml.) in the blood of brown trout after 35 days of treatment with four rates of sulfamerazine (the daily rates of 0, 5, 10 and 15 g of sulfamerazine per 100 pounds of fish). Two methods (I and II) of administering the sulfamerazine were used. Ten fish were measured for each rate and each method. The data is given in file `hemoglobin.txt`.

(i) Obtain the ANOVA table and test the null hypothesis H_{AB} that there is no interaction between the two factors.

(ii) Obtain and report the parameter estimates for the model containing both factors and their interaction.

(iii) If H_{AB} is not rejected, obtain new parameter estimates under the additive model, and test for the presence of each main effect. Comment on your findings.

(iv) Check the model assumptions (for each considered model) by using relevant diagnostic tools.

4. A quality engineer at a textile company is interested in comparing the breaking strength of a synthetic fiber made on 3 different machines, thickness of the fiber is also measured. The data is contained in file `fiber.txt`.

   (i) Investigate whether the factor machine influences the fiber strength by performing a statistical test, without taking the thickness into account.

   (ii) Investigate whether the factor `machine` influences the fiber `strength` by performing a statistical test, now including `thickness` as an (additional) explanatory variable into the analysis (hint: use the additive model). Comment on your findings.

   (iii) Investigate whether the dependence of the fiber strength on thickness is similar for all three machines. First do this graphically, next confirm this by performing an appropriate statistical test. What is the estimated fiber strength (possibly different for different machines) for a fiber with the average thickness?

   (iv) Which of the two models, without or with variable `thickness`, do you prefer? For the chosen model, check the model assumptions by using relevant diagnostic tools.

II. Theoretical problems

1. Consider the one-way ANOVA model under the constraint $\mu = 0$. Confirm that $\hat{\alpha}_i$ are unbiased estimators of α_i, $i = 1, \dots I$. Show that $S_\omega - S_\Omega = \sum_{i=1}^{I} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$. Determine $\mathbb{E}(S_\omega + S_\Omega)$ and $\mathrm{Var}(S_\Omega - S_\omega)$.

2. Consider the one-way ANOVA under the constraint $\sum_{i=1}^{I} \alpha_i = 0$, with a balanced design. Show that $\hat{\mu}$ and $\hat{\alpha}_i$ are unbiased estimators of μ and α_i, respectively. Determine $\mathrm{Var}(\hat{\mu})$, $\mathrm{Var}(\hat{\alpha}_i)$ and $\mathrm{Cov}(\hat{\mu}, \hat{\alpha}_i)$.

3. Show that $\hat{\gamma}_{ij}$ is an unbiased estimator of the parameter γ_{ij} in the two-way ANOVA model.

4. Suppose that in the two-way ANOVA model the parameter σ^2 is known. Propose tests (you are allowed to use σ^2) for the presence of main effects and interaction.