Statistics 206

Homework 8

Due : Nov. 30, 2015, In Class

1. **Regression formulations of one-way ANOVA model**.

   (a) Consider the *cell means formulation* discussed in class:

   $$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \cdots, I, j = 1, \cdots, n_i.$$

   Express this model as a linear regression model:

   $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

   Specify $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. What is $\mathbf{X}^T\mathbf{X}$ and what is $\mathbf{X}^T\mathbf{Y}$? What is the LS estimator of $\boldsymbol{\beta}$? Derive the fitted values and residuals.

   (b) Consider the alternative formulation used by `R`:

   $$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, I, j = 1, \cdots, n_i, \quad \alpha_1 = 0.$$

   Express this model as a linear regression model by specifying $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. Compare it with the linear regression model with $I - 1$ indicator variables for factor levels (with level 1 as the reference class). What do you find?

2. **Decomposition of total sum of squares in one-way ANOVA**.

   (a) Show that $SSTO = SSTR + SSE$, where
   - $SSTO := \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\cdot\cdot})^2$.
   - $SSTR := \sum_{i=1}^{I} n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$.
   - $SSE := \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2$.

   (b) Show that
   - $\sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\cdot\cdot}) = 0$. (So $d.f.(SSTO) = n_T - 1$)
   - $\sum_{i=1}^{I} n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) = 0$. (So $d.f.(SSTR) = I - 1$)
   - For each $i = 1, \cdots, I$, $\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot}) = 0$. (So $d.f.(SSE) = n_T - I$)

3. **Expectation and Sampling distribution of SS in one-way ANOVA**.

   (a) Show that $E[SSTR] = (I - 1)\sigma^2 + \sum_{i=1}^{I} n_i (\mu_i - \mu_\cdot)^2$, where

   $$\mu_\cdot = \frac{1}{n_T} \sum_{i=1}^{I} n_i \mu_i, \quad n_T = \sum_{i=1}^{I} n_i.$$

   (Hints: (i) $\overline{Y}_{i\cdot} \sim N(\mu_i, \sigma^2/n_i)$ and are independent with each other; (ii) $\bar{Y}_{\cdot\cdot} = \frac{1}{n_T} \sum_{i=1}^{I} n_i \bar{Y}_{i\cdot} \sim N(\mu_\cdot, \frac{\sigma^2}{n_T})$; (iii) For a random variable Z: $E(Z^2) = Var(Z) + (E(Z))^2$.

(b) Show that $E[SSTO] = (n_T - 1)\sigma^2 + \sum_{i=1}^{I} n_i(\mu_i - \mu_\cdot)^2$. (Hint: Recall $SSE \sim \sigma^2\chi^2_{n_T - I})$.

(c) **(Optional Problem).** Under $H_0 : \mu_1 = \cdots = \mu_I$, show that SSTO $\sim \sigma^2\chi^2_{(n_T-1)}$. Due to independence of SSE and $SSTR$ (see Problem 4), therefore under H_0, SSTR $\sim \sigma^2\chi^2_{(I-1)}$.

(Hints: (i) Under H_0, $\mu_i \equiv \mu_\cdot$ for $i = 1, \cdots, I$; (ii) So $Y_{ij} - \overline{Y}_{\cdot\cdot} = (Y_{ij} - \mu_i) - (\overline{Y}_{\cdot\cdot} - \mu_\cdot)$. Therefore, without loss of generality, we can assume $\mu_i \equiv 0$ for $i = 1, \cdots, I$, i.e., $Y_{ij} \sim_{i.i.d.} N(0, \sigma^2)$; (iii) Note $\overline{Y}_{\cdot\cdot} = \frac{1}{n_T}\sum_{ij} Y_{ij}$.)

4. **Independence of SSE and SSTR in one-way ANOVA.** Show the following:

(a) For each $i = 1, \cdots, I$, $Y_{ij} - \overline{Y}_{i\cdot}$ $(j = 1, \cdots, n_i)$ and $\overline{Y}_{i\cdot}$ are independent. *(Hint: Calculate their covariance.)*

(b) For $i \neq i'$, $Y_{i'j} - \overline{Y}_{i'\cdot}$ $(j = 1, \cdots, n_{i'})$ and $\overline{Y}_{i\cdot}$ are independent .

(c) SSE and $\overline{Y}_{i\cdot}$ $(i = 1, \cdots, I)$ are independent. *(Hint: Use the fact that if two sets of random variables are independent, then any function of the first set is independent with any function of the second set.)*

(d) SSE and $SSTR$ are independent.

Notes: Indeed we have already had the results in problems 2, 3, 4 from linear regression theories (Are you able to recognize them?). Here, we show that they can be derived without using matrix algebra. This is due to the special structure of the ANOVA model: Recall that the design matrices in problem 1 are of very special forms.

5. **One-way ANOVA case study in R.**

Notes: You do not need to hand in this problem on Monday.

A company uses six filling machines of the same make and model to place detergent into cartons that show a label weight of 32 ounces. The production manger has complained that the six machines do not place the same amount of fill into the cartons. A consultant requested that 20 filled cartons be selected randomly from each of the six machines and the content of each carton weighted. The observations were recorded in terms of the deviations of weights from 32 ounces. The data is under Resources/Homework/filling.txt: The first column is the observation, the second column is the index for the filling machine and the third column is the index for the carton. Consider fitting the one-way ANOVA model to this data.

(a) What is the response variable? What is the factor? How many levels are there for this factor? Name the design of this study.

(b) Draw side-by-side box plots of the response for the factor levels. Do the factor level means appear to differ? Does the variability of the observations within each factor level appear to be approximately the same for all factor levels?

(c) Fit the one-way ANOVA model. Draw residual versus fitted value plot and residual Q-Q plot. Comment on model assumptions. Are remedial measures needed? *(Hint: When using the `lm` function, remember to declare the factor as `factor`.)*

(d) Obtain the estimated factor levels means and obtain the ANOVA table.

(e) Test whether or not the mean fill differs among the six machines at level 0.05. State the null and alternative hypotheses, the decision rule and the conclusion.

(f) Construct a 99% confidence interval for μ_2. If we are interested in all factor level means, what multiple comparison procedure shall we use? What would be the corresponding 99% confidence interval for μ_2?

(g) How many pairwise comparisons of factor levels means are there? If we are interested in all these pairwise comparisons, what multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? Construct the corresponding 95% confidence intervals for all pairwise comparisons. At familywise significance level 0.05, which pairs of factor level means should be declared as being different?

(h) What if we are only interested in 6 **pre-specified** pairwise comparisons, which multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? What if we are only interested in the 6 pairwise comparisons that show the most differences in the data (i.e corresponding to the six largest $|\widehat{D}|$), which procedure shall we use and what are the corresponding C.Is?

(i) If we are interested in all possible contrasts, which multiple comparison procedure shall we use? Construct the corresponding 95% confidence interval for the contrast:

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}.$$

Test whether or not $L = 0$ with family-wise-error-rate controlled at 0.05. What if we are interested in 20 pre-specified contrasts (including L), which multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? Construct the corresponding 95% confidence interval for L (assuming L is in this prespecified set of contrasts). What do you find?