Additional note 7

Lecture 18

1. **Revision questions of L17**

   - If a categorical factor has numerical labels and so can be treated as continuous variable, it can be fitted to a one-way ANOVA and a simple linear regression. How to compare the two models in terms of prediction and model fit?

     Answer: Adv: A simple linear regression gives the trend of changes in Y across x and so it can be used to predict Y at the levels of x other than the levels for the given categories in the experimental design.
     Disadv: The simple linear regression model has 1 df for Treatment sum of squares (TSS) which is less than the $t-1$ df for one-way ANOVA. The RSS for simple linear regression is often higher.

   - How to enable the same RSS for the regression and one-way ANOVA model?

     Answer: Consider polynomial regression model (orthogonal or not) of order $t-1$.

   - How to calculate the 1df component of TSS in the one-way ANOVA model that corresponds to the TSS of the simple linear regression model?

     Answer: $\frac{\left(\sum\limits_{i=1}^{t} c_i \bar{Y}_i\right)^2}{\sum\limits_{i=1}^{t} \frac{c_i^2}{n_i}} = \beta^2 S_{xx} = \text{LinRegSS}$ where $c_i = n_i(x_i - barx)$.

2. This lecture introduces two-way ANOVA and shows how it is linked to one-way ANOVA. The a levels of factor A and b levels of factor B form ab levels for the combination of factor A and factor B. If we treat these ab levels as coming from one factor called AB, the TSS from this one-way ANOVA so estimated is more than as the TSS of two-way ANOVA with only main (first order) effects but is the same as the TSS of two-way ANOVA with interaction effect as well. We show that the TSS from one-way ANOVA for the factor AB with ab levels can be partition into SS due to A alone, SS due to B alone and SS due to the interaction of factor A and factor B.

3. Note the way how the mean adjusted SS due to Y called TotSS (total SS) with df $n-1$ be splitted into SS due to A, SS due to B, SS due to AB (interaction) (the sum of these 3 SS is TSS captured by the model) and RSS (not captured by the model), and so do df, and how the different SS can be calculated (the last two pages). These calculations help the understanding of the model structure. Note that we assume complete and balanced design in which all combinations have observations of equal numbers.

Lecture 19

1. **Revision questions of L18**

   - When the ab combinations of treatments from two factors (A & B) is treated as one factor of ab treatments and fitted to one-way ANOVA, is this RSS_1 greater, less or equal to the RSS_2 of two-way ANOVA with A & B main effects?

     Answer: RSS_1 for one-way ANOVA is less, ie $\text{RSS}_1 < \text{RSS}_2$ or $\text{TSS}_1 > \text{TSS}_2$.

   - What does the extra TSS_1 relative to TSS_2 account for? What is the corresponding df?

     Answer: It accounts for the interaction effect between factor A and factor B. The df for this interaction effect is the difference in df between TSS_1 and TSS_2 which is $(ab-1)-(a-1)-(b-1) = ab - a - b + 1 = (a-1)(b-1)$.

- What is the name of this design?

  Answer: This design is called factorial design.

2. The TSS of one-way ANOVA with ab treatments can be splitted into SS due to factor A alone, SS due to factor B alone and the SS due to the interaction of factor A and B.

3. Interaction plots visualize the change of Y across levels of a factor at each level of the second factor. If there similar trends, it reveals absence of interaction effect between factor A and B. Otherwise, say the trends even cross-over which indicate inconsistency, the two factors are likely to interact.

4. For complete and balance (also called orthogonal) design, the SS_A and SS_B are independent of and will be changed by the order factor A and B entered into the model. The mean adjusted projection matrices $\mathbf{P}_{W_A}$ and $\mathbf{P}_{W_B}$ are orthogonal such that $\mathbf{P}_{W_A}\mathbf{P}_{W_B} = \mathbf{0}$.

5. However when some treatment combinations are not observed (incomplete) and have different replicates, the factor first entered into the two-way ANOVA model is marginal so that its SS is the same as one-way ANOVA with that factor. The factor next entered has SS adjusted for the first factor.

Lecture 20

1. **Revision questions of L19**

   - What is interaction plot? What does it tell?

     Answer: The change of Y across levels of one factor for each level of the second factor. If the lines show similar trends (like parallel to each other), there is not interaction effect. If the line show different trends (like cross-overs), there is interaction effect.

   - What is orthogonal design? What is the impact of this design on the SS of factor A and B?

     Answer: It refers to complete and balance design, ie all treatment combinations are observed at the same number of replicates. SS of factor A and B are the same, independent of the order they entered into the model. Product of mean adjusted projection matrices is zero.

   - What if the design is not orthogonal? What is the impact on the SS of factor A and B?

     Answer: Non-orthogonal design refers to design not complete or balance. SS of each factor depends on the order factors are entered into the model. The first factor entered is marginal, equal to the SS of one-way ANOVA of that factor and the second factor entered adjusted for the first factor.

2. There are three types of normality test apart from the usual boxplot and QQ plot.

   The first type is the Pearson Chi-square goodness of fit test comparing the observed and expected counts in certain equal-width intervals of the data range.

   The second type compares the empirical distribution function with the tested distribution function. Kolmogorov-Smirnov (KS) test considers the maximum of such difference whereas Cramer-von Mises test considers integral of such difference with uniform weight and Anderson-Darling test considers weight function $1/(F(x) \times (1 - F(x))$ which gives heavier weights at the two ends of the data range.

   Lastly Shapiro-Wilk test takes a different approach using order statistic. The idea is similar to the simulation based Shapiro-Francia test which considers R^2 of the points in the QQ plot. Some results suggest that Shapiro-Wilk test often performs the best and next come Anderson-Darling test.