Statistical Computation
Math 475

Jimin Ding

Department of Mathematics
Washington University in St. Louis
www.math.wustl.edu/ jmding/math475/index.html

November 14, 2013

Part V

Analysis of Variance

Outline I

Examples

- Example 1: The effects of four coooking temperatures on the fluffiness of omelets prepared from a mix was studied by randomly assigning five packages of mix to each of four cooking terperatures.
  This is an experimental study since the covariate of interest is controlled by the experimenter.

- Example 2: A study of the effects of education of sales people on the sales volumes of randomly selected salesmen from a company.
  This is an observational study since the data were obtained without controlling the covariate of interest.

Terminology in ANOVA:

- factor: categorical predictor;
- factor levels: different values of predictors;
- treatment: a factor level;
- one-way ANOVA/single factor ANOVA: including only one predictor;
- qualitative/quantitative factor
- experimental (basic) unit of study: representativieness
- ANOVA model I: fixed factor levels;
  ANOVA model II: random factor levels.

Model Setup

Let Y_{ij} be the response variable in the jth trial for the ith factor level.

$$Y_{ij} = \mu_i + e_{ij}, \quad i = 1, \cdots, r, \ \ j = 1, \cdots, n_i,$$

where μ_i's are the parameters that describe the means of the ith level and e_{ij}'s are independently identically distributed with mean 0 and variance σ^2. Here r is the number of treatments (factor levels) and n_i is the number of observations in each treatment (each level).

We assume the following assumptions for this ANOVA model:

- All observations have the same variance. (Homogeneity)
- All observations are independent. (Independence)
- The error term is normally distributed. (Normality)

Estimation

LSE:

$$Q = \sum_i \sum_j (Y_{ij} - \mu_i)^2$$

$$\Rightarrow \hat{\mu}_i = \bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij},$$

$$\Rightarrow \hat{\sigma}^2 = \text{MSE} = \frac{1}{n_T - r} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i)^2,$$

where $n_T = \sum_{i=1}^{r} n_i$.
Note that the MLE for μ is same as the LSE and $\hat{Y}_{ij} = \hat{\mu}_i$.

Relationship with Regression

Regression function describe the nature of the statistical relationship between the mean response and the levels of the predictor variables. In ANOVA model,

- when the predictors are qualitative, there is no fundamental difference between ANOVA and regression with dummy variables;
- when the predictors are quantitative, ANOVA requires no specification on the nature of the statistical relation, while regression requires this specification.
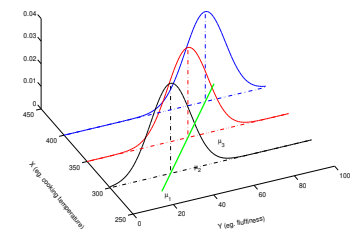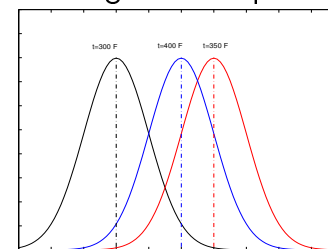
Sum Square Tables

Total deviation of each observation around the overall mean $Y_{ij} - \bar{Y}_{..}$ can be decomposed as:

$$Y_{ij} - \bar{Y}_{..} = \bar{Y}_{i.} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{i.},$$

where the first term is the deviation of the estimated factor level mean around the overall mean and the second term is deviation of each observation around its respective estimated factor level mean, which reflects the remained uncertainty after utilizing the factor level information.

Sum Square Tables

$$Y_{ij} - \bar{Y}_{..} = \bar{Y}_{i.} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{i.},$$

Now we square and sum them up for all observations to get a quantity to measure those variation:

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{r} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2,$$

SS :	$SSTO$	$=$	SSR	$+$	SSE
df :	$n_T - 1$	$=$	$r - 1$	$+$	$n_T - r$
MS :			MSR		MSE

$E(MSE) = \sigma^2$

$E(MSR) = \sigma^2 + \frac{1}{r-1}\sum n_i(\mu_i - \bar{\mu}_.)^2, \quad \bar{\mu}_. = \frac{1}{n_T}\sum n_i\mu_i.$

F-test

$H_0 : \mu_i = \mu_0, \quad i = 1, \cdots, r$ vs. H_a : at least one of $\mu_i \neq \mu_0$.

Intuitively, if after introducing the factor level the remaining deviation of observations around the estimated mean (SSR) is relatively small, comparing with the deviation explained by introducing the treatments (SSE), then it implies that the factor level does provide more information and so the means of different levels are different.

$\Leftrightarrow$ If SSR is comparatively larger than SSE, it is less possible that H_0 is true and we should reject H_0.

F-test

$H_0 : \mu_i = \mu_0, \quad i = 1, \cdots, r$ vs. H_a : at least one of $\mu_i \neq \mu_0$.

Under normality assumption, we know $\frac{SSR}{\sigma^2} \sim \chi^2_{r-1}$ and $\frac{SSE}{\sigma^2} \sim \chi^2_{n_T-r}$. This lead to

$$F - \text{value} = \frac{MSR}{MSE} = \frac{\frac{SSR}{\sigma^2}/(r-1)}{\frac{SSE}{\sigma^2}/(n_T-r)} \sim F(r-1, n_T - r)$$

So we reject the null hypothesis for extreme large F-value (those with small p-values) and conclude that at least one of treatments is different from the others.

Remark: This F-test is same as F-test for goodness-of-fit we learned in regression.

$$MSR = \frac{SSR}{r-1} = \frac{SSTO - SSE}{(n_T-1) - (n_T-r)} = \frac{SSE(R) - SSE(F)}{df_R - df_F}.$$

Confidence Interval and Hypothesis Testing for Level Mean

Because

$$E(\bar{Y}_{i.}) = \mu_i, \quad Var(\bar{Y}_{i.}) = \sigma^2/n_i,$$

the $1 - \alpha$ confidence interval for μ_i is:

$$\mu_i \pm t(1 - \alpha/2, n_T - r)\sqrt{MSE/n_i}.$$

To test:

$$H_0 : \mu_i = c \quad \text{v.s.} \quad H_a : \mu_i \neq c,$$

$$t - \text{value} = \frac{\bar{Y}_{i.} - c}{\sqrt{MSE/n_i}} \sim t_{n_T - r}.$$

Find p-value and draw a conclusion.

Contrasts

A special type of linear combinations of level means:

$$L = \sum_{i=1}^{r} c_i \mu_i, \quad \text{where} \sum_{i=1}^{r} c_i = 0.$$

For examples:
- The difference between the first two level means $L = \mu_1 - \mu_2$;
- Comparison between the three standard treatments and the new treatment $L = \frac{\mu_1 + \mu_2 + \mu_3}{3} - \mu_4$.

To test:

$$H_0 : L = 0 \quad \text{v.s.} \quad H_a : L \neq 0.$$

We use

$$t - \text{value} = \frac{\hat{L}}{se\{\hat{L}\}} = \frac{\sum_{i=1}^{r} c_i \bar{Y}_{i.}}{\sqrt{MSE(\sum_{i=1}^{r} c_i^2/n_i)}}$$

Multiple Comparisons - Tukey's Method

- Tukey (Duncan, Tukey-Kramer): all pairwise comparisons of factor level means.

$$\begin{aligned} H_0 : & \quad \mu_i - \mu_j = 0, \text{for all } i \neq j \\ H_a : & \quad \text{at least one pair is non-zero.} \end{aligned}$$

$$\frac{\hat{L}}{se\{\hat{L}\}} = \frac{\bar{Y}_{i.} - \bar{Y}_{j.}}{\sqrt{MSE(1/n_i + 1/n_j)}} \sim \frac{1}{\sqrt{2}} q(1 - \alpha; r, n_T - r),$$

where $q(r, v)$ stands for Studentized Range distribution.

> Let $Y_1, Y_2, \cdots, Y_r \sim_{iid} N(\mu, \sigma^2)$, $w = max(Y_i) - min(Y_i)$ and s^2 be an estimate of σ^2 with v degree of freedom and independent of Y_i's. Then w/s follows Studentized Range distribution with degree of freedom r and v. The p-value can be checked from a simulated table.

Multiple Comparisons - Tukey's Method

Remark: When the design is balanced ($n_i \equiv n$), Tukey procedure is an exact test for all possible pairwise comparisons. Otherwise, it is a conservative test. In general, this procedure is good of data snooping.

Multiple Comparisons - Scheffé Method

- Scheffé: all possible contrasts among the factor level means.

$$H_o: \qquad L = \sum_{i=1}^{r} c_i \mu_i = 0, \text{for all } c_i's.$$
$$H_a: \quad \text{at least one set of} c_i's \text{ leads non-zero} L.$$

$$\frac{\hat{L}}{se\{\hat{L}\}} = \frac{\sum_i c_i \hat{\mu}_i}{\sqrt{MSE(\sum_i c_i^2/n_i)}}$$
$$\sim \sqrt{(r-1)F(1-\alpha; r-1, n_T - r)}.$$

Remark: It is good for data snooping and it is conservative.

Multiple Comparisons - Bonferroni

- Bonferroni: the particular set of pairwise comparisons, contrasts, or linear combinations specified by the user.

$$\frac{\hat{L}}{se\{\hat{L}\}} \sim t(1 - \alpha/2g; n_T - r),$$

where g is the number of interested tests.

Multiple Comparisons - Bonferroni

Remark:
- It is easy to employ and so widely used for general multiple comparisons.
- Bonferroni test is less conservative than "Scheffé" test when the number of contrasts of interest is about the same as the number of factor levels.
- When all pairwise comparisons are of interest, it usually leads to wider confidence intervals than Tukey procedure.
- It is exact when the family of tests are independent.

Rule of Thumb:
Choose the narrowest confidence intervals.

Modified Levene Test:

$$H_0: \quad Var(Y_{1j}) = Var(Y_{2j}) = \cdots = Var(Y_{rj})$$
$$H_a: \quad \text{at least one of them is different.}$$

Test statistics:

$$d_{ij} = |Y_{ij} - \tilde{Y}_i|,$$
$$F - \text{value} = \frac{MSR_d}{MSE_d} \sim F(r-1, n_T - r)$$

where $\tilde{Y}_i$ is the observed median of the ith level and

$$MSR = \frac{\sum_i n_i(\bar{d}_{i.} - \bar{d}_{..})^2}{r-1},$$
$$MSR = \frac{\sum_i \sum_j (d_{ij} - \bar{d}_{i.})^2}{n_T - r}.$$

Modified Levene Test:

Remark:

- Levene test is robust to the normality assumption and does not require equal sample size.
- Heterogeneity can be cured by using weighted LS.
- In regression, homogeneity assumption is usually tested by the modified Levene test on the grouped residuals.

More model diagnostic issues can be found in the reading material. (Applied linear statistical models, P776-P777; The analysis of variance by Scheffé, 1959.)

Other Possible Formulation:

$$Y_{ij} = \bar{\mu}. + \tau_i + e_{ij}, \quad i = 1, \cdots, r, \quad j = 1, \cdots, n_i,$$

where $\bar{\mu}.$ is the overall mean (cross all treatments) and $\tau_i = \mu_i - \bar{\mu}.$ describes the individual treatment deviation from the overall mean, which is subject to $\sum_i \tau_i = 0$ (or $\sum_i w_i \tau_i = 0$).

Note: Here we have one more parameter but also one more constrain.

Two-Factors Model Setup

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \quad i = 1, \cdots, a, j = 1, \cdots, b, k = 1, \cdots, n_{ij},$$

where i is for factor A and j is for factor B.

- Treatment/Cell mean: $\mu_{ij} = \bar{\mu}.. + \alpha_i + \beta_j + (\alpha\beta)_{ij}$
  - Overall mean : $\bar{\mu}.. = \sum_{i=1}^{a} \sum_{j=1}^{b} \mu_{ij}/(ab)$;
  - Main effect of factor A: $\alpha_i = \bar{\mu}_i. - \bar{\mu}.., \quad i = 1, \cdots, a$;
  - Main effect of factor B: $\beta_j = \bar{\mu}._j - \bar{\mu}.., \quad j = 1, \cdots, b$;
    where $\bar{\mu}_i. = \sum_{j=1}^{b} \mu_{ij}/b$ and $\bar{\mu}._j = \sum_{i=1}^{a} \mu_{ij}/a$ are A factor and B factor level means.
    Note : $\sum_{i=1}^{a} \alpha_i = 0; \quad \sum_{j=1}^{b} \beta_j = 0.$
  - Interaction:
    $(\alpha\beta)_{ij} = \mu_{ij} - (\bar{\mu}.. + \alpha_i + \beta_j), \quad i = 1, \cdots, a, j = 1, \cdots, b.$
    Note : $\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0; \quad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0.$
- Additive model: $\mu_{ij} = \bar{\mu}.. + \alpha_i + \beta_j$
- Full factorial model: $\mu_{ij} = \bar{\mu}.. + \alpha_i + \beta_j + (\alpha\beta)_{ij}$

Interactions

- Exam whether the difference between two mean responses for any two levels of factor B is same as for all levels of factor A. (Check whether the treatment mean curves for the different factor levels are parallel.)
- If all $n_{ij} \equiv 1$, we should not use the interaction term.
- Sometimes when two factors interact, the interaction effects are so small that they are considered to be unimportant interactions. Importance of the interaction term is difficult to judge sometimes. This decision is not a statistical decision and should be made by the subject area specialist. The unimportant interactions are usually excluded from the model to gain power in testing main effects.

Interactions

Remark: Even though the computation of unbalanced ANOVA is much harder than balanced ANOVA, the statistical software has no difficulty handling the computation for us. But non-orthogonality introduced by unbalanced design may cause difficulty in understanding the ANOVA output.

Estimation

We minimize
$$Q = \sum_i^a \sum_j^b \sum_k^{n_{ij}} [Y_{ijk} - \mu_{..} - \alpha_i - \beta_j - (\alpha\beta)_{ij}]^2,$$

such that $\sum_i^a \alpha_i = 0$, $\sum_j^b \beta_j = 0$, $\sum_i^a (\alpha\beta)_{ij} = 0$ for $j = 1, \cdots, b$ and $\sum_j^b (\alpha\beta)_{ij} = 0$ for $i = 1, \cdots, a$.

Then the LSE is:
$$\begin{aligned}
\hat{\mu}_{..} &= \bar{Y}_{...}, \\
\hat{\alpha}_i &= \bar{Y}_{i..} - \bar{Y}_{...}, \\
\hat{\beta}_j &= \bar{Y}_{.j.} - \bar{Y}_{...}, \\
(\hat{\alpha\beta})_{ij} &= \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}.
\end{aligned}$$

Prediction: $\hat{Y}_{ijk} = \bar{Y}_{ij.}$.

Testing

Sum Square Decomposition for a balance ANOVA:
$$\begin{aligned}
SSTO &= \sum_{ijk} (Y_{ijk} - \bar{Y}_{...})^2 = SSE + SSR \\
SSE &= \sum_{ijk} (Y_{ijk} - \hat{Y}_{ijk})^2 = \sum_{ijk} (Y_{ijk} - \bar{Y}_{...})^2, \\
SSR &= \sum_{ijk} (\hat{Y}_{ijk} - \bar{Y}_{...})^2 \\
&= nb \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 + na \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2 \\
&\quad + n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \\
&= SSA + SSB + SSAB
\end{aligned}$$

Testing

$$\begin{array}{ccccccccc}
SSTO &=& SSA &+& SSB &+& SSAB &+& SSE \\
df: && a-1 && b-1 && (a-1)(b-1) && ab(n-1)
\end{array}$$

F test:
- Interactions:
  $H_0 : (\alpha\beta)_{ij} = 0$, for all i, j;   v.s. H_a : Not all $(\alpha\beta)_{ij}$ are zeros.
  $$F - \text{value} = \frac{MSAB}{MSE} \sim F_{(a-1)(b-1),ab(n-1)}$$

- Main effect for factor A:
  $H_0 : \alpha_i = 0$, for all i;   v.s. H_a : Not all α_i are zeros.
  $$F - \text{value} = \frac{MSA}{MSE} \sim F_{(a-1),ab(n-1)}$$

Three-Way Layout

- Generalization from two-way layout.
- Three-way interactions should be included carefully since the interpretation of those high-way interactions is difficult sometimes and models with too many interactions are usually unstable.

Nested Design

In the factorial studies, every level of one factor appears with each level of every other factor. The factors are hence said to be crossed. However, sometimes factors are nested. For example,

- Two instructors from each of three schools, which located in St.Louis, Chicago and Kansas City respectively, were selected to test some training method. Both effect of school (factor A) and effect of instructor (factor of B) are of interest. Note that an instructor can hardly teach in schools at St.louis and at Chicago simultaneously.

Nested Design

- An analyst was interested in the effects of community (factor A) and neighborhood (factor B) on the spread of information about new products. Information was obtained from 5 families in each neighborhoods within selected communities. Note that the neighborhood 1 in one community is different from the neighborhood 1 in another community.

Nested Design

Remarks:

- Nested factors are frequently encountered in observational studies (example 2), where the researcher cannot manipulate the factors under study, or in experiments where only some factors can be manipulated (example 1).
- The distinction between crossed and nested design is subtle sometimes. For example, in the example 2, if we label the neighborhood by the income level, then the income level $10K - 20K$ in one community is same as the the income level $10K - 20K$ in another community. Hence the income level becomes a crossed factor and this leads to a crossed factorial design.

Model Setup

Nested two-way ANOVA:

$$Y_{ijk} = \bar{\mu}.. + \alpha_i + \beta_j(\alpha_i) + e_{ijk},$$

$$i = 1, \cdots, a, j = 1, \cdots, b_i, k = 1, \cdots, n_{ij}$$

ANOVA Table

Sum Square Decomposition:

$$
\begin{aligned}
SSTO &= \sum_{ijk}(Y_{ijk} - \bar{Y}...)^2 \\
&= \sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot} + \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} + \bar{Y}_{i\cdot\cdot} - \bar{Y}...)^2 \\
&= \sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot})^2 + \sum_{ij} n_{ij}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot})^2 \\
&\quad + \sum_i (\sum_{j=1} n_{ij})(\bar{Y}_{i\cdot\cdot} - \bar{Y}...)^2 \\
&= SSE + SSB(A) + SSA \\
df : N_T - 1 &= (N_T - \sum_{i=1}^{a} b_i) + (\sum_{i=1}^{a} b_i - a) + (a-1)
\end{aligned}
$$

Nested Design

F test:

- In the case of levels of factor B is fixed and of interest:
  Test A: $\frac{MSA}{MSE}$
  Test B(A): $\frac{MSB(A)}{MSE}$
- In the case of levels of factor B is random and nuisance:
  Test A: $\frac{MSA}{MSB(A)}$
  Test B(A): $\frac{MSB(A)}{MSE}$

Repeated Measures

The same subject (person, store, experimental animals, plants, etc.) is used in each of the treatment under study. The subject serves as a block, and experimental units within a block may be viewed as the different occasions when a treatment is applied to the subject. For example,

- Each of 200 persons who have headaches is given 2 different drugs and a placebo, for two weeks each, with the order of the drugs randomized for each person.
- In a weight study, 100 overweight persons are given the same diet and their weight measured at the end of each week for 12 weeks to access the weight loss over time.

Repeated Measures

- In 8 randomly selected supermarket stores, 3 price levels have applied on grapefruit for three one-week periods, with the order of the 3 price levels randomly assigned for each store. The sales data are collected to study the relationship between grapefruit sales and the price at which grapefruits are offered.

Key: subject
- a randomly selected unit
- independent with each other
- serves as their own control so that the treatments can be compared within the unit

Repeated Measures

- Advantage:
  All sources of variability between subjects are excluded from the experimental errors, because any two treatments can be compared directly for each subject. Only variation within subjects enters the experimental error. Hence the experiment saves the experimental unit (samples) and achieves larger power at lower cost.

- Disadvantage:
  There might be interference among treatments (carryover effect). Sometimes the order of the treatments applied will affect the response variable. For example, in judging 5 different soup recipes, a bland recipe may get a higher rating when preceded by a highly spiced recipe than when preceded by a blander recipe.

Split-Plot Designs

Split-plot designs were originally developed for agricultural experiments. Consider an experiment to study the effects of two varieties of yam and three watering levels on yield, using different field as blocks. Each field is divided into two plots to randomly assign different yams and then each plot is split into three subplots to assign different amount water. (This is different from complete randomized block design in which 6 combinations of yam and watering level would be randomly assigned into each field.)

The primary interesting factor usually is randomized within each block first to reduce the variability within a block.

A split-plot design for two-factors studies can be viewed as a special repeated measures design. But only the interactions related with primary interesting factor are included in the model.

Full ANCOVA Model

ANCOVA models include both categorical and continuous covariates. For example, when there is one categorical predictor (with a levels) and one continuous predictor X, the model can be written as, for $i = 1, \cdots, a, j = 1, \cdots, n$,

$$Y_{ij} = \bar{\mu}.. + \alpha_i + \beta(X_{ij} - \bar{X}..) + \gamma_i(X_{ij} - \bar{X}_{i\cdot}) + e_{ij},$$

where

- μ describes overall mean,
- α_i describes the effect of the categorical predictor at level i and is subject to $\sum_i \alpha_i = 0$,
- β is the overall regression coefficient between continuous predictor X and continuous response Y,
- and γ_i is the level specific regression coefficient in level i, which is subject to $\sum_i \gamma_i = 0$.
- The error term $e_{ij} \sim_{i.i.d} N(0, \sigma^2)$.

ANCOVA

When the effect of continuous variable X on Y is same in all levels of categorical predictor, the model can be reduced to an additive model with $\gamma_i = 0$ for all i. That is:

$$Y_{ij} = \bar{\mu}.. + \alpha_i + \beta(X_{ij} - \bar{X}..) + e_{ij},$$

In the $x - y$ plots, the regression lines of different categorical levels are parallel.

MANOVA

- Response variable is a vector instead of a scaler.
- In both MANOVA and repeated measurements designs, response variables come from the same subject and is corrected within subjects. However, in repeated measurements design, the response variable is the same variable measured at different conditions; in MANOVA, the response variables might be different measurements (such as weight, height, blood samples, etc.).
- Instead of a univariate F-value, we will have a multivariate F-value (Wilk's Lambda or Hotelling's trace or Pillai's criterion) based on a comparison of the error covariance matrix and the effect covariance matrix.

MANOVA

- The advantage of MANOVA over ANOVA is that the former takes into account the dependence within subjects and protest Type I errors. The disadvantage is that MANOVA involves more complicated design and might be hard to interpret which covariate affects each response variable.
- One degree of freedom is lost for each additional dependent variable.
- Remark: Including highly corrected response variables in MANOVA might not be better than using one response variable in ANOVA.

Reading Assignment

Textbook: Applied Statistics and the SAS Programming Language,
Chap 7, Chap 8

References

- Neter, Kutner, Nachtsheim and Wasserman, 2003, *Applied Linear Statistical Models, 5th Ed.*
- Montgomery, 2004, *Design and Analysis of Experiments, 6th Ed.*