Applied Statistical Methods II

Chapters #22 23 in KNNL

Analysis of Covariance and unbalanced design

ANCOVA

- Assume that we have $i = 1, \ldots, r$ treatment groups.
- We have $j = 1, \ldots, n$ subjects per treatment group.
- We have a continuous variable X_{ij} that is associated with the outcome.
- The common formulation of the single-factor covariance model is:
    - $Y_{ij} = \mu_{\cdot} + \tau_i + \gamma \left(X_{ij} - \overline{X}_{\cdot\cdot} \right) + \epsilon_{ij}$
    - ϵ_{ij} iid $\sim N(0, \sigma^2)$
- We mean-center covariates out of convenience so that $E(\bar{Y}_{\cdot\cdot}) = \mu_{\cdot}$ if $\sum_i \tau_i = 0$.

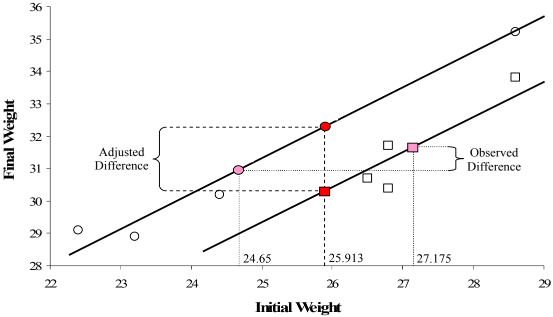
Why use ANCOVA?

- In ANCOVA model, the main interest is the treatment effect.
- Controlling the covariate X in the model
  - reduces the error variance and increases power to test treatment.
  - can access the treatment effect for similar levels of X.
- ANCOVA is particularly important in non-randomized studies where X values can be different among treatments.
- In completely randomized studies, fitting a single-factor ANOVA model without controlling for X is unbiased, but ANCOVA is still preferred if X is strongly correlated with Y.

A small sample example

- Assume that oyster's final weight (Y) is a linear function of its initial weight (X), and the location also has an effect on the final weight (two black curves).
- Treatment is location:
  - Treatment 1 - circle.
  - Treatment 2 - square.
- Although eight oysters are randomized to two locations, i.e., the expected values of the initial weight are the same in two groups.
- However, the observed $\bar{X}$ values are quite different due to the small sample size.

Example figure



Example cont.

- Pink shows the estimated treatment effect, using single factor ANOVA.
- We can see that the estimate is in the wrong direction if one misses the important factor X. Fortunately, it's not likely to reach statistical significance due to large variations.
- If one fits an ANCOVA model (i.e., adding X), then the estimated treatment effect will be in the right direction and will be more likely to reach significance, because of the reduced error variance.

Assumptions

- The ANCOVA model presented has several assumptions.
- There is a constant variance.
  - Can check with residual plots.
- The treatment effect is the same for any level of the concomitant variable X: there is no treatment by concomitant interaction.
- Normality.
  - Can assess through QQ-plots.
- Correct functional form of the concomitant variable.
  - Can check with residual plots.
  - Can have concomitant variable in a polynomial form.
  - Can have several concomitant variables.

Fitting the Model and Inference

- Model fitting and inference are performed by viewing ANCOVA as a general regression model.
- Main test of interest is $H_0 : \tau_1 = \cdots = \tau_r$.
- Assume the model $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$.
- X is $n \times p$ and does not have to be full rank.
- We will use the general linear testing to do inference.
- Consider testing $H_0 : A\beta = 0$ where A is $q \times p$.
    - A has rank q and each row of $A\beta$ is estimable.

- For the ANCOVA Model
  - $Y_{ij} = \mu_{\cdot} + \tau_i + \gamma \left(X_{ij} - \overline{X}_{\cdot\cdot} \right) + \epsilon_{ij}$
  - We want to test $H_0 : \tau_1 = \cdots = \tau_r$.
  - $F^* = \frac{SSR(\tau|X)/(r-1)}{SSE(\tau,X)/(n-r-1)} \sim F_{r-1,n-r-1}$
  - Let's look at the promotions application.

Cracker example

- A company wants to know the effect of three different promotions on cracker sales.
- Treatment is promotion:
  - Treatment 1 - sampling product in store.
  - Treatment 2 - extra regular shelf space.
  - Treatment 3 - display case.
- Each promotion is implemented in 5 randomly chosen stores.
- Also know the sales from the store in the previous period.

Finding the best promotions.

- Okay, so there is a difference.
- What is the best promotion or promotions?
- This is a multiple comparisons problem.
- Can use Tukey, Bonferroni and Scheffé, if we want.
- We must use LSMEANS in SAS rather than MEANS.
  - MEANS computes the collapsed means, $\overline{Y}_{i\cdot\cdot}$.
  - LSMEANS gives the least squares estimators at the mean of other factors/covariates.
  - Same for one-way ANOVA models.
  - Same for two-way ANOVA with balanced data.
  - Different for unbalanced two-way ANOVA.
  - Different for ANCOVA.

Recall Tukey MCP for One-Way ANOVA

- $Y_{ij} = \mu_i + \epsilon_{ij}$
- $i = 1, \ldots, r$
- $s^2 \sim \chi^2_\nu$ is an estimate of σ^2 that is independent of Y_i.
- Let $w = \max(\hat{\mu}_i) - \min(\hat{\mu}_i)$.
- Studentized range is $q_{r,v} = \frac{w}{s}$.
- We can numerically find the distribution of q.
- Let $\hat{D}_{ik} = \hat{\mu}_i - \hat{\mu}_k$.
- If we have equal sample sizes, the family of confidence intervals $\hat{D}_{ik} \pm Ts(\hat{D}_{ik})$, $T = \frac{1}{\sqrt{2}} q_{r,\nu}(1-\alpha)$, has at least $1-\alpha$ coverage.
- Coverage is exactly $1-\alpha$ if data are balanced.
  - Conservative if there is unbalanced data.
  - Usually much less conservative than Sheffé or Bonferroni.

Tukey for ANCOVA

- Assume that we have n subjects per treatment.
- $\hat{\tau}_i - \hat{\tau}_j \pm s q_{r,v}(1-\alpha)\sqrt{\frac{1}{n} + \left(\overline{X}_{i\cdot} - \overline{X}_{j\cdot}\right)^2 / 2S_{XX}}$
- s^2 is the MSE.
- $S_{XX} = n^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n} (X_{ij} - \overline{X}_{i\cdot})^2$
- The family of confidence intervals will be conservative for estimating all pairwise comparisons, but less than other approaches.

Two-Factor ANCOVA

- Method is easily generalizable to two-factors.
- Consider example to determine effects of flower variety and moister level on yield of saleable flowers.
  - Two levels of variety: LP and WB
  - Two levels of moister: low and high.
- Will use plot size as a concomitant variable (X).
- Each setting is given to six flowers.
- 2×2 factorial study.
- $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \gamma(X_{ijk} - \overline{X}_{...}) + \epsilon_{ijk}$

Looking forward:

- Unbalanced Two-Way ANOVA
  - Chapter 23
  - Cannot separate SSTR into SSA, SSB, and SSAB

Let's Remember Balanced Two-Way ANOVA

- We have two factors: A and B.
- A has 'a' levels and B has 'b' levels.
- *n* subjects receive each combination of A and B.
  - Have $a \times b$ unique treatment groups.
  - $n_T = a \times b \times n$.
- Observe Y_{ijk}, $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n$
- $Y_{ijk} - \overline{Y}_{\ldots} = \left(\overline{Y}_{ij\cdot} - \overline{Y}_{\ldots}\right) + \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)$
- Square and sum: SSTO = SSTR + SSE
  - $SSTO = \sum_{ijk} \left(Y_{ijk} - \overline{Y}_{\ldots}\right)^2$
  - $SSTR = n\sum_{ij} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{\ldots}\right)^2$
  - $SSE = \sum_{ijk} \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2$
  - Cross terms are zero since $\sum_k Y_{ijk} = \sum_k \overline{Y}_{ij\cdot}$.

Decomposing SSTR

- With balanced data, can further decompose SSTR.
- $\overline{Y}_{ij\cdot} - \overline{Y}_{\cdots} =$
  $\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right) + \left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right) + \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)$
- Square and sum: SSTR = SSA + SSB + SSAB
  - $SSA = nb\sum_i \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2$
  - $SSB = na\sum_j \left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2$
  - $SSAB = n\sum_{ij} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)^2$
- The cross terms die out. Example on next slide.

- $\sum_i \overline{Y}_{i..} = a\overline{Y}_{...}$
- $\sum_j \overline{Y}_{.j.} = b\overline{Y}_{...}$

$$\begin{aligned}
\sum_{ijk}(\overline{Y}_{i..} - \overline{Y}_{...})(\overline{Y}_{.j.} - \overline{Y}_{...}) &= n\sum_i\sum_j(\overline{Y}_{i..} - \overline{Y}_{...})(\overline{Y}_{.j.} - \overline{Y}_{...}) \\
&= n\sum_i(\overline{Y}_{i..} - \overline{Y}_{...})\sum_j(\overline{Y}_{.j.} - \overline{Y}_{...}) \\
&= n\sum_i(\overline{Y}_{i..} - \overline{Y}_{...})(\sum_j \overline{Y}_{.j.} - b\overline{Y}_{...}) \\
&= 0
\end{aligned}$$

For Unbalanced Data

- Now consider the unbalanced data setting.
- We have two factors: A and B.
- A has *a* levels and B has *b* levels.
- n_{ij} subjects receive A=i and B=j
  - $n_T = \sum_{i,j} n_{ij}$.
- Observe Y_{ijk}, $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n_{ij}$
- $Y_{ijk} - \overline{Y}_{\ldots} = \left(\overline{Y}_{ij\cdot} - \overline{Y}_{\ldots}\right) + \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)$
- Square and sum: SSTO = SSTR + SSE
  - $SSTO = \sum_{ijk} \left(Y_{ijk} - \overline{Y}_{\ldots}\right)^2$
  - $SSTR = \sum_{ij} n_{ij} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{\ldots}\right)^2$
  - $SSE = \sum_{ijk} \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2$
  - Cross terms are zero since $\sum_k Y_{ijk} = \sum_k \overline{Y}_{ij\cdot}$.

ANOVA Decomposition Not As Nice

$$\begin{aligned}\overline{Y}_{...} &= \sum_{ijk} \frac{Y_{ijk}}{\sum_{ij} n_{ij}} \\ &\neq \sum_{i} \frac{\overline{Y}_{i..}}{a} \neq \sum_{j} \frac{\overline{Y}_{.j.}}{b}\end{aligned}$$

- For unbalanced data, overall mean is not the average of group means.
- Cross terms will not cancel so that $SSTR \neq SSA + SSB + SSAB$.
- Type I and Type III ANOVA tables will be different.

Testing Group Effects

- By viewing the two-way ANOVA model as a regression, can still do the general F-tests.
- Consider a full regression model: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$.
- $\beta = [\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b]'$ (this is an additive model)
- To test H_0: no factor B effect: can be written as a general linear test and use F test
  - $F^* = \frac{SSE(A) - SSE(A,B)}{b-1} \Big/ \frac{SSE(A,B)}{n_T - a - b + 1}$
  - Under H_0, $F^* \sim F_{b-1, n_T - a - b + 1}$

Growth example

- Children with slow growth are administered growth hormone.
- We want to know how sex and severity of depression of bone development affect growth.
- Outcome is difference in growth rate before and during treatment in cm/month.
- Sex is male or female.
- Depression is severe, moderate, or mild.
- Observational study:
  - Unbalanced design.
  - Tried to have 3 per groups but had drop-outs.

Difference in the LSMEANS and MEANS statements

To estimate the mean of the *i*th level of factor A:

- Means statement: $\hat{\mu}_{i\cdot}^{M} = \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} / \left(\sum_{j=1}^{b} n_{ij} \right)$.
- LSMEANS statement: $\hat{\mu}_{i\cdot}^{L} = \sum_{j=1}^{b} \hat{\mu}_{ij} / b$, where $\hat{\mu}_{ij}$ is the BLUE of μ_{ij}.
- Difference in group means: $\hat{\mu}_{i\cdot}^{M} - \hat{\mu}_{i'\cdot}^{M}$ and $\hat{\mu}_{i\cdot}^{L} - \hat{\mu}_{i'\cdot}^{L}$.
- Same for balanced data, different for unbalanced.
- SAS examples.

Zero Cells

- Zero cells are when one of the treatment combinations has no observations.
- Sometimes this is due to the science:
  - Study to look at % of tumor shrinkage after chemo.
  - You have 4 different types of cancers: lung, head, liver, ovarian.
  - Want to look at effects in males and females.
  - Fully biological males cannot have ovarian cancer.
- Sometimes they occur due to drop out.
  - Only have one female with severe depression in previous study.
  - What if she was not in the study?

Dealing with Zero Cells

- If you are fitting a model with interaction, you loose a degree of freedom. The denominator degrees of freedom is $n_T - (ab - 1)$.
- Cannot draw inference on cells with missing data. But can still draw inference on linear combinations of other cell means.
- If you know a priori that there is no interaction, fit the additive model.
  - Can draw inference on every group.
  - Pool information from other groups to draw inference on group with no data.
- SAS Examples.