30 Lecture 30: April 23

Last time

- Analysis of Variance (JF chapter 8)
  - two-way anova
  - higher-way anova

Today

- analysis of covariance (ANCOVA)
- Final exam (take-home) will be posted April 30th and due midnight May 7th.
- Course evaluation.

Analysis of Covariance

Analysis of covariance (ANCOVA) is a term used to describe linear models that contain both qualitative and quantitative explanatory variables. The method is, therefore, equivalent to dummy-variable regression, discussed in the previous lectures, although the ANCOVA model is parametrized differently from the dummy-regression model.

Covariate is a variable known to affect the response that

1. differs among EUs
2. reflects differences that exist independently of experimental treatment.

A nutrition example

A nutrition scientist conducted an experiment to evaluate the effects of four vitamin supplements on the weight gain of laboratory animals. The experiment was conducted in a completely randomized design with $N = 20$ animals randomized to $a = 4$ supplement groups, each with sample size $n \equiv 5$. The response variable of interest is weight gain, but calorie intake z was measured simultaneously.

Diet	$y(g)$	Diet	y	Diet	y	Diet	y
1	48	2	65	3	79	4	59
1	67	2	49	3	52	4	50
1	78	2	37	3	63	4	59
1	69	2	75	3	65	4	42
1	53	2	63	3	67	4	34
1	$\bar{y}_{1+} = 63$	2	$\bar{y}_{2+} = 57.8$	3	$\bar{y}_{3+} = 65.2$	4	$\bar{y}_{4+} = 48.8$
1	$s_1 = 12.3$	2	$s_2 = 14.9$	3	$s_3 = 9.7$	4	$s_4 = 10.9$

Question: Is there evidence of a vitamin supplement effect?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	3	797.8	265.9	1.823	0.184
Residuals	16	2334.4	145.9		

But calorie intake z was measured simultaneously:

Diet	$y(g)$	z	Diet	y	z	Diet	y	z	Diet	y	z
1	48	350	2	65	400	3	79	510	4	59	530
1	67	440	2	49	450	3	52	410	4	50	520
1	78	440	2	37	370	3	63	470	4	59	520
1	69	510	2	75	530	3	65	470	4	42	510
1	53	470	2	63	420	3	67	480	4	34	430

Question: How and why could these new data be incorporated into analysis?
Answer: ANCOVA can be used to reduce unexplained variation.

ANCOVA model,

$$y_{ij} = \mu + \alpha_i + \beta z_{ij} + \epsilon_{ij}$$

where μ is the reference level, α_i is the main effect of treatment, β is the partial regression coefficient, and $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The model is equivalent as the dummy-variable regression model,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_z z_i + \epsilon_i \quad \text{for } i = 1, \ldots, 20$$

Finish the table below

Source	df
Diet	
Covariate	1
Residual	
Total	

Answer:

To test for difference among treatments. The null hypothesis in terms of α_i is
$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_4 = 0$ v.s. $H_a :$ at least one $\alpha_i \neq 0$
And the null hypothesis in terms of β_i is
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ v.s. $H_a :$ at least one $\beta_i \neq 0$

Question: which two models do we compare when testing the above null hypothesis?
Answer:

Linear contrasts of means

With ANOVA (or ANCOVA) models, we do not generally test hypotheses about individual coefficients (but we can do so if we wish). For dummy-coded regressors in one-way ANOVA, a t-test or F-test of $H_0 : \alpha_1 = 0$, for example, is equivalent to testing for the difference in means between the first group and the baseline group, $H_0 : \mu_1 = \mu_m$.

Consider the one-way ANOVA model:

$$Y_{ij} = \mu_i + \epsilon_{ij}, i = 1, 2, \ldots, t, \text{ and } j = 1, 2, \ldots, n_i$$

with $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

A linear function of the group means of the form

$$\theta = c_1\mu_1 + c_2\mu_2 + \cdots + c_t\mu_t$$

is called a linear combination of the treatment means. And the c_i's are the coefficients of the linear combination. If

$$c_1 + c_2 + \cdots + c_t = \sum_{j=1}^{t} c_j = 0,$$

the linear combination is called a contrast. Contrasts with more than two non-zero coefficients are called complex contrasts.

Let two contrasts θ_1 and θ_2 be given by

$$\theta_1 = c_1\mu_1 + \cdots + c_t\mu_t = \sum_{j=1}^{t} c_j\mu_j$$

$$\theta_2 = d_1\mu_1 + \cdots + d_t\mu_t = \sum_{j=1}^{t} d_j\mu_j,$$

then the two contrasts θ_1 and θ_2 are mutually orthogonal if the products of their coefficients sum to zero:

$$c_1d_1 + \cdots + c_td_t = \sum_{j=1}^{t} c_jd_j = 0$$

θ_i and θ_j are orthogonal $\implies$ $\hat{\theta}_i$ and $\hat{\theta}_j$ are statistically independent.

Types of effects

Consider the following two-way ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$
$$i = 1, 2 = a \text{ and } j = 1, 2 = b \text{ and } k = 1, 2, \ldots, 7 = n.$$

$\epsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Parameter constraints: $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i (\alpha\beta)_{ij} = 0$ for each j and $\sum_j (\alpha\beta)_{ij} = 0$ for each i.

- Factor A: AGE has $a = 2$ levels - A_1 : younger and A_2 : older
- Factor B: GENDER has $b = 2$ levels - B_1 : female and B_2 : male

Three kinds of effects in this 2×2 design:

1. Simple effects are simple contrasts.
   - $\mu(A_1B) = \mu_{12} - \mu_{11}$ - simple effect of gender for young folks.
   - $\mu(AB_1) = \mu_{21} - \mu_{11}$ - simple effect of age for women.
2. Interaction effects are differences of simple effects: $\mu(AB) = \mu(AB_2) - \mu(AB_1) = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})$
   - difference between simple age effects for men and women
   - difference between simple gender effects for old and young folks
   - interaction effect of AGE and GENDER.
3. Main effects are averages or sums of simple effects

$$\mu(A) = \frac{1}{2}(\mu(AB_1) + \mu(AB_2))$$

$$\mu(B) = \frac{1}{2}(\mu(A_1B) + \mu(A_2B))$$

Sampling distribution of linear contrast estimates

For a linear contrast

$$\theta = c_1\mu_1 + \cdots + c_t\mu_t$$

The *best* estimator for a contrast of interest can be obtained by substituting treatment group sample means $\bar{y}_{i+}$ for treatment population means μ_i in the contrast θ:

$$\hat{\theta} = c_1\bar{Y}_{1+} + c_2\bar{Y}_{2+} + \cdots + c_t\bar{Y}_{t+}$$

Example

Recall the binding fraction data that investigate binding fraction for several antibiotics using $n = 20$ bovine serum samples:

Antibiotic	Binding Percentage	Sample mean
Penicillin G	29.6 24.3 28.5 32.0	28.6
Tetracyclin	27.3 32.6 30.8 34.8	31.4
Streptomycin	5.8 6.2 11.0 8.3	7.8
Erythromycin	21.6 17.4 18.3 19	19.1
Chloramphenicol	29.2 32.8 25.0 24.2	27.8

Consider the pairwise contrast comparing penicillin (population) mean to Tetracyclin mean:

$$\theta = \mu_1 - \mu_2 = (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 + (0)\mu_5$$

Obtain a point estimator of θ.
Answers:

Question: How good is this estimate? In other words, how much uncertainty associated with the estimate?

We want to characterize the sampling distribution of $\hat{\theta}$. According to our model setup, Y_{ij} follow normal distributions. $\hat{\theta}$ is a linear function of Y_{ij}, so that $\hat{\theta}$ follows a normal distribution. We want to derive the mean and variance (the two sufficient statistics) to characterize the normal distribution that $\hat{\theta}$ follows:

$$\hat{\theta} \sim \mathcal{N}(\theta, Var(\hat{\theta}))$$

Derive expressions for the mean and the variance:

Therefore, the standard error:

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})} = \sqrt{\sigma^2 \sum_{j=1}^{t} \frac{c_j^2}{n_j}}$$

which is estimated by

$$\widehat{SE}(\hat{\theta}) = \sqrt{MS[E] \sum_{j=1}^{t} \frac{c_j^2}{n_j}}$$

To test $H_0 : \theta = \theta_0$ (often 0) versus $H_1 : \theta \neq \theta_0$, use t-test:

$$t = \frac{\hat{\theta} - \theta_0}{\widehat{SE}(\hat{\theta})} \overset{H_0}{\sim} t_{N-t}$$

At level α, the critical value for this test is $t(N-t, \alpha/2)$ and $100(1-\alpha)\%$ confidence interval for a contrast $\theta = \sum c_j \mu_j$ is given by

$$\sum c_j \bar{Y}_{j+} \pm t(N-t, \alpha/2) \sqrt{MS[E] \sum \frac{c_j^2}{n_j}}$$

Multiple Comparisons

Let's first review type I and type II errors.

	H_0 is True	H_0 is False
Don't reject H_0	Probability $1-\alpha$	Probability β
Reject H_0	Probability α	Probability $1-\beta$

- Type I error: rejection of a true null hypothesis (false positive).
- Type II error: failure to reject a false null hypothesis (false negative).
- Type I error rate or significance level (α): the probability of rejecting the null hypothesis given the null hypothesis is true.
- Type II error rate (β): the probability of failure to reject the null hypothesis given the null hypothesis is false. $1-\beta$ gives the power of a test.

Now, let's consider all simple (pairwise) contrasts for the binding fraction data with $t = 5$ antibiotic treatments of the form $\theta = \mu_i - \mu_j$.

- We have $\begin{pmatrix} 5 \\ 2 \end{pmatrix} = 10$ tests for significance each at level $\alpha = 0.05$
- what is the probability of committing at least one type I error?

We need to consider the familywise error rate (fwe) when testing k contrasts:

$$fwe = \Pr(\text{at least one type I error})$$

Methods for simultaneous inference for multiple contrasts include

- Bonferroni
- Scheffé
- Tukey

When the number of comparisons is in the hundreds or thousands (e.g. genome-wide association studies), and FWE control is hopeless, more manageable type I error rate is the False Discovery Rate (FDR):

$$FDR = E(\frac{\text{Falsely rejected null hypotheses}}{\text{Number of rejected null hypotheses}})$$

Bonferroni correction

Suppose interest lies in exactly k contrasts. The Bonferroni adjustment to α controls fwe is

$$\alpha_{bonferroni} = \frac{\alpha}{k}$$

and simultaneous 95% confidence intervals for the k contrasts are given by

$$a_1\bar{Y}_{1+} + \cdots + a_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{a_j^2}{n_j}}$$

$$b_1\bar{Y}_{1+} + \cdots + b_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{b_j^2}{n_j}}$$

$$\cdots$$

$$k_1\bar{Y}_{1+} + \cdots + k_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{k_j^2}{n_j}}$$

where ν denotes df for error.

Example: for the binding fraction example, consider only pairwise comparisons with Penicillin:

$$\theta_1 = \mu_1 - \mu_2, \theta_2 = \mu_1 - \mu_3, \theta_3 = \mu_1 - \mu_4, \theta_4 = \mu_1 - \mu_5$$

We have $k = 4, \alpha_{bonferroni} = 0.05/k = 0.0125$ and $t(\frac{\alpha_{bonferroni}}{2}, 15) = 2.84$. Substitution leads to

$$t(\frac{\alpha_{bonferroni}}{2}, 15)\sqrt{MS[E]\left(\frac{1^2}{4} + \frac{(-1)^2}{4} + \frac{0^2}{4} + \cdots + \frac{0^2}{4}\right)}$$

$$=2.84\sqrt{(9.05)\frac{2}{4}} = 6.0$$

so that **simultaneous** 95% confidence intervals for θ_1, θ_2, θ_3 and θ_4 take the form

$$\bar{y}_{1+} - \bar{y}_{i+} \pm 6.0$$

Scheffé

Another method to construct **simultaneous** 95% confidence intervals for **ALL** contrasts, use

$$\sum_{j=1}^{t} c_j \bar{y}_{j+} \pm \sqrt{(t-1)(F^*)MS[E]\sum_{j=1}^{t}\frac{c_j^2}{n_j}}$$

where $F^* = F(\alpha, t-1, N-t)$. For a pairwise comparisons of means, μ_j and μ_k, this yields

$$\bar{y}_{j+} - \bar{y}_{k+} \pm \sqrt{(t-1)(F^*)MS[E](1/n_j + 1/n_k)}$$

Using $\alpha = 0.05$, need to specify

- t (from the design)
- F^* (same critical value as for $H_0 : \alpha_i \equiv 0$).
- $MS[E]$ (from the data)
- $\bar{y}_{j+}$, $\bar{y}_{k+}$
- n_j, n_k (from the data)

For binding fraction data,

$$\sqrt{(t-1)(F^*)MS[E](\frac{1}{n_j}+\frac{1}{n_k})} = \sqrt{(5-1)(3.06)9.05(\frac{1}{4}+\frac{1}{4})} = 7.44$$

If any two sample means differ by more than 7.44, they differ significantly.

Tukey

Tukey's method is better than Scheffé's method when making **all pairwise** comparisons in balanced designs ($n = n_1 = n_2 = \cdots = n_t$). It is conservative, controlling the experimentwise error rate, and has a lower type II error rate in these cases than Scheffé. (It is more powerful.)

For simple contrasts of the form

$$\theta = \mu_j - \mu_k$$

to test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

reject H_0 at level α if

$$|\hat{\theta}| > q(t, N-t, \alpha)\sqrt{\frac{MS[E]}{n}}$$

where $q(t, N-t, \alpha)$ denotes α level studentized range for t means and $N-t$ degrees of freedom, the quantity $q(t, N-t, \alpha)\sqrt{\frac{MS[E]}{n}}$ is referred to as Tukey's honestly significant difference (HSD). The studentized ranges can be calculated using R function `qtukey(1 − α, t, N − t)`.