Regression Part II

- One-factor ANOVA
- Another dummy variable coding scheme
- Contrasts
- Multiple comparisons
- Interactions

One-factor Analysis of variance

- Categorical Explanatory variable
- Quantitative Response variable
- *p* categories (groups)
- H_0: All population means equal
- Normal conditional distributions
- Equal variances

Dummy Variables

- You have seen
  - Indicator dummy variables with intercept
  - Effect coding (with intercept)
- **Cell means coding** is also useful at times.

A common error

- Categorical explanatory variable with *p* categories
- *p* dummy variables (rather than *p-1*)
- And an intercept

- There are *p* population means represented by *p+1* regression coefficients - not unique

But suppose you leave off the intercept

- Now there are *p* regression coefficients and *p* population means
- The correspondence is unique, and the model can be handy -- less algebra
- Called **cell means coding**

Cell means coding: *p* indicators and no intercept

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Drug	x_1	x_2	x_3	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	0	$\mu_1 = \beta_1$
B	0	1	0	$\mu_2 = \beta_2$
Placebo	0	0	1	$\mu_3 = \beta_3$

Add a covariate: x_4

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Drug	x_1	x_2	x_3	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
A	1	0	0	$\beta_1 + \beta_4 x_4$
B	0	1	0	$\beta_2 + \beta_4 x_4$
Placebo	0	0	1	$\beta_3 + \beta_4 x_4$

Contrasts

$$c = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$$

$$\widehat{c} = a_1\overline{Y}_1 + a_2\overline{Y}_2 + \cdots + a_p\overline{Y}_p$$

$$\text{where } a_1 + a_2 + \cdots + a_p = 0$$

Overall F-test is a test of p-1 contrasts

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

a_1	a_2	a_3	a_4
1	-1	0	0
0	1	-1	0
0	0	1	-1

$$c = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$$

In a one-factor design

- Mostly, what you want are tests of contrasts,
- Or collections of contrasts.
- You could do it with any dummy variable coding scheme.
- Cell means coding is often most convenient.
- With $\boldsymbol{\beta}=\boldsymbol{\mu}$, test H_0: $\mathbf{L}\boldsymbol{\beta}=\mathbf{h}$

- Can get a confidence interval for any single contrast using the *t* distribution.

Multiple Comparisons

- Most hypothesis tests are designed to be carried out in isolation
- But if you do a lot of tests and all the null hypotheses are true, the chance of rejecting at least one of them can be a lot more than α. This is **inflation of the Type I error probability**.
- Otherwise known as the curse of a thousand t-tests.
- Multiple comparisons (sometimes called follow-up tests, post hoc tests, probing) try to offer a solution.

Multiple Comparisons

- Protect a *family* of tests against Type I error at some *joint* significance level α
- If all the null hypotheses are true, the probability of rejecting at least one is no more than α

Multiple comparison tests of contrasts in a one-factor design

- Usual null hypothesis is $\mu_1 = \ldots = \mu_p$.
- Usually do them after rejecting the initial null hypothesis with an ordinary F test.
- The big three are
  - Bonferroni
  - Tukey
  - Scheffé

Bonferroni

- Based on Bonferroni's inequality

$$Pr\left\{\cup_{j=1}^{k} A_j\right\} \leq \sum_{j=1}^{k} Pr\{A_j\}$$

- Applies to *any* collection of k tests
- Assume all k null hypotheses are true
- Event A_j is that null hypothesis j is rejected.
- Do the tests as usual
- Reject each H_0 if $p < 0.05/k$
- Or, adjust the p-values. Multiply them by k, and reject if $pk < 0.05$

Bonferroni

- Advantage: Flexible – Applies to *any* collection of hypothesis tests.
- Advantage: Easy to do.

- Disadvantage: Must know what all the tests are before seeing the data.
- Disadvantage: A little conservative; the true joint significance level is *less* than α.

Tukey (HSD)

- Based on the distribution of the largest mean minus the smallest.
- Applies only to pairwise comparisons of means.
- If sample sizes are equal, it's most powerful, period.
- If sample sizes are not equal, it's a bit conservative.

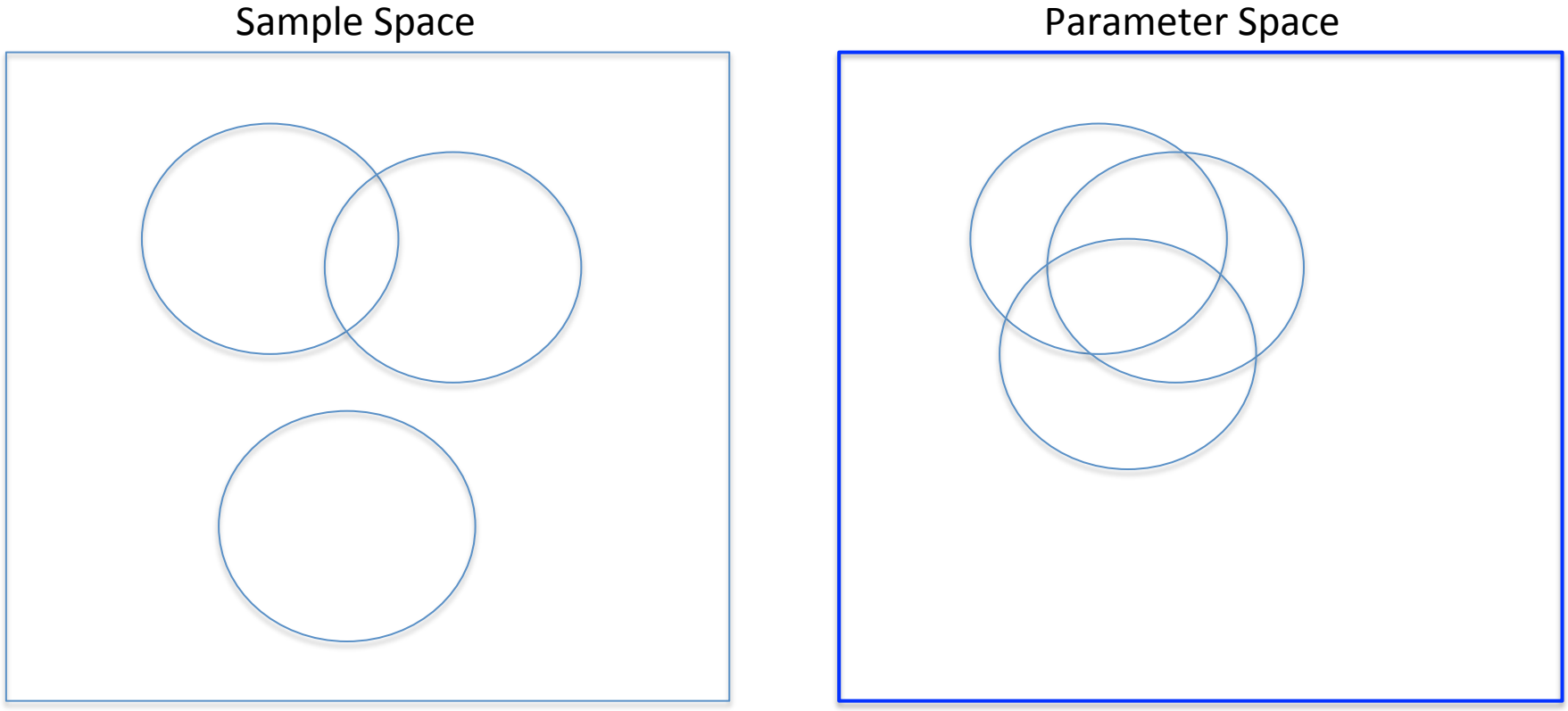
Scheffé

- Find the usual critical value for the initial test. Multiply by p-1. This is the Scheffé critical value.
- Family includes *all* contrasts: Infinitely many!
- You don't need to specify them in advance.
- Based on the union-intersection principle.

General principle

- The union of the critical regions is the critical region of the overall test.
- The intersection of the null hypothesis regions is the null hypothesis region of the overall test.
- So if all the null hypotheses in the family are true, the parameter is in the null hypothesis region of the overall test.
- And the probability of rejecting at least one of the family null hypotheses is α, the significance level of the overall test.

Critical region is union of critical regions
Null hypothesis is intersection of null hypotheses



Actually all you need is containment

- The union of critical regions of tests in the family must be *contained* in the critical region of the overall (initial) test, so if any test in the family rejects H_0, the overall test does too.

- In this case the probability that at least one test in the family will wrongly reject H_0 is $\leq \alpha$.

Scheffé are union-intersection tests

- Follow-up tests *cannot* reject H_0 if the initial F-test does not. Not quite true of Bonferroni and Tukey.
- If the initial test (of p-1 contrasts) rejects H_0, there is a contrast for which the Scheffé test will reject H_0 (not necessarily a pairwise comparison).
- Adjusted p-value is the tail area beyond F/(p-1) using the null distribution of the *initial* test.

Which method should you use?

- If the sample sizes are nearly equal and you are only interested in pairwise comparisons, use Tukey because it's most powerful

- If the sample sizes are not close to equal and you are only interested in pairwise comparisons, there is (amazingly) no harm in applying all three methods and picking the one that gives you the greatest number of significant results. (It's okay because this choice *could be* determined in advance based on number of treatments, α and the sample sizes.)

- If you are interested in contrasts that go beyond pairwise comparisons and you can specify *all* of them before seeing the data, Bonferroni is almost always more powerful than Scheffé. (Tukey is out.)

- If you want lots of special contrasts but you don't know in advance exactly what they all are, Scheffé is the only honest way to go, unless you have a separate replication data set.

Interactions

- Interaction between independent variables means “It depends.”
- Relationship between one explanatory variable and the response variable *depends* on the value of the other explanatory variable.
- Can have
  - Quantitative by quantitative
  - Quantitative by categorical
  - Categorical by categorical

Quantitative by Quantitative

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \\ E(Y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \end{aligned}$$

For fixed x_2

$$E(Y|\mathbf{x}) = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

Both slope and intercept depend on value of x_2

And for fixed x_1, slope and intercept relating x_2 to E(Y) depend on the value of x_1

Quantitative by Categorical

- One regression line for each category.
- Interaction means slopes are not equal
- Form a product of quantitative variable by each dummy variable for the categorical variable
- For example, three treatments and one covariate: x_1 is the covariate and x_2, x_3 are dummy variables

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &\quad + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon \end{aligned}$$

General principle

- Interaction between A and B means
  - Relationship of A to Y depends on value of B
  - Relationship of B to Y depends on value of A
- The two statements are formally equivalent

Make a table

$$E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
2	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0+\beta_2)+(\beta_1+\beta_4)x_1$
2	0	1	$(\beta_0+\beta_3)+(\beta_1+\beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

What null hypothesis would you test for

- Equal slopes
- Comparing slopes for group one vs three
- Comparing slopes for group one vs two
- Equal regressions
- Interaction between group and x_1

What to do if H_0: $\beta_4=\beta_5=0$ is rejected

- How do you test Group "controlling" for x_1?
- A reasonable choice is to set x_1 to its sample mean, and compare treatments at that point.

- How about setting x_1 to sample mean of the group (3 different values)?
- With random assignment to Group, all three means just estimate $E(X_1)$, and the mean of all the x_1 values is a better estimate.

Categorical by Categorical

- Soon
- But first, an example of multiple comparisons.

Copyright Information