

Analysis of Covariance

Applied Regression and Other Multivariable Methods
Sections 15-1 - 15-7

20

Examples

- **Pretest/Posttest score analysis:** The gain in score Y may be associated with the pretest score X . Analysis of covariance provides a way to “handicap” each student. That way, one does not need to find a group of students with similar pretest scores and randomly assign them to a control and treatment group. Similar to analyzing difference in scores.
- **Weight gain experiments in animals:** If wishing to compare different feeds, the weight gain Y may be associated with the original weight of the animal. Analysis of covariance provides a way to use a herd and adjust for the varying original weights.
- **Comparing competing drug products:** The effect of the drug Y after two hours (measured on a scale from 1 to 10) may be associated with the initial mental and physical shape of the subject. Variables describing the initial mental and physical shape may be used as covariates.

20-2

Background

- Consider comparing two treatments
- Will do regression using dummy variable Z
- Another factor X is correlated with Y
- Nuisance factor X called a covariate

- ANCOVA adjusts Y for effect of covariate X before comparing treatments
- Compares each treatment when X is set at \bar{X}
- Without adjustment, effects of X may
inflate σ^2
alter treatment comparisons

20-1

Regression Approach

- Have covariate X and dummy variable Z
 - Fit the model
- $$Y = \beta_0 + \beta_1 X + \beta_2 Z + E$$
- Use partial F test of $H_0 : \beta_2 = 0$ to determine if the two treatments are different
 - Adjusted means are

$$\begin{aligned}\bar{Y}_0(\text{adj}) &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \\ \bar{Y}_1(\text{adj}) &= (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 \bar{X}\end{aligned}$$

- Can also be expressed

$$\begin{aligned}\bar{Y}_0(\text{adj}) &= \bar{Y}_0 - \hat{\beta}_1(\bar{X}_0 - \bar{X}) \\ \bar{Y}_1(\text{adj}) &= \bar{Y}_1 - \hat{\beta}_1(\bar{X}_1 - \bar{X})\end{aligned}$$

- Approach can be extended to more than two treatments
- This approach assumes constant slope between X and Y

20-3

ANOVA Approach

- Consider single covariate in one way ANOVA (k trt)
- Statistical model is

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + E_{ij} \quad \begin{cases} i = 1, 2, \dots, k \\ j = 1, 2, \dots, n_i \end{cases}$$

- Additional assumptions

X_{ij} not affected by treatment

X and Y are linearly related

Constant slope

- General Procedure:

Fit one-way model ($Y = \text{trt}$)

Fit one-way model ($X = \text{trt}$)

Regress residuals on each other

This is same approach/description we discussed partial correlation (i.e., comparing Y vs X_1 adjusting for X_2)

20-4

Analysis of Covariance

- Test $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$
 - Compare treatment means after adjusting for differences in covariate levels
 - Trt and covariate not orthogonal (order of fit matters)

$$F_0 = \frac{SS(\text{trt}|X)/k - 1}{SSE/(N - k - 1)}$$

- Adjusted treatment means

– Estimate $\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i = \bar{Y}_i - \hat{\beta}(\bar{X}_i - \bar{X})$

– Variance: $\hat{\sigma}^2 (1/n_i + (\bar{X}_i - \bar{X}_{..})^2 / \sum \sum (X_{ij} - \bar{X}_i)^2)$

- Test: $\beta = 0$

– Sum of Squares regression ($SS(X)$): $\beta^2 \sum \sum (X_{ij} - \bar{X}_i)^2$

$$F_0 = \frac{SS(X)/1}{SSE/(N - k - 1)}$$

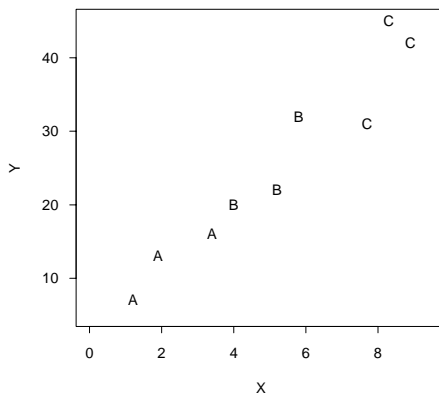
20-5

Analysis of Covariance

Two Examples

- 1 No treatment differences

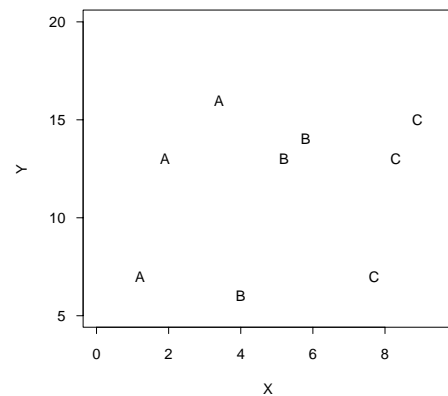
- Positive linear relationship
- Covariate larger in each group
- Thus, appears to be treatment difference



20-6

- 2 Treatment differences exist

- Positive linear relationship
- Covariate larger in each group
- Thus, no apparent treatment difference



Using SAS

```
options nocenter ls=80; goptions colors=(none);

data example1;
  input trt x y @@;
  z1 = 0; z2 = 0;
  if trt = 1 then z1 = 1;
  if trt = 2 then z2 = 1;
cards;
  1 1.2 7 1 1.9 13 1 3.4 16
  2 4.0 20 2 5.2 22 2 5.8 32
  3 7.7 31 3 8.3 45 3 8.9 42
;

proc sort; by trt;
symbol1 v=circle; symbol2 v=square; symbol3 v=triangle;
proc gplot; plot y*x=trt;

proc glm; class trt;
model y=trt; means trt / tdiff lsd;

proc glm; class trt;
model y=trt x; lsmeans trt / tdiff adjust=t;

proc reg;
model y = x z1 z2;
/* z1 coef is diff between adj means of trt 1 vs 3
   z2 coef is diff between adj means of trt 2 vs 3
   Coef z1 minus coef z2 is diff between trt 1 vs 3
   The test command can be used to test hypotheses */
trt1vs3: test z1=0;
trt2vs3: test z2=0;
trt1vs2: test z1-z2=0;
run;
quit;
```

20-7

The GLM Procedure Least Squares Means

trt	y LSMEAN	LSMEAN Number
1	27.8342355	1
2	25.4907533	2
3	22.6750113	3

Least Squares Means for Effect trt
t for H0: LSMean(i)=LSMean(j) / Pr > |t|

i/j	Dependent Variable: y		
	1	2	3
1		0.354685	0.410644
		0.7373	0.6983
2	-0.35468		0.38071
	0.7373		0.7191
3	-0.41064	-0.38071	
	0.6983	0.7191	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

*** The upper number in each cell of the table is the t statistic

*** The lower number is the P-value for a two-sided test

20-8

The REG Procedure

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1260.93663	420.31221	22.11	0.0026
Error	5	95.06337	19.01267		
Corrected Total	8	1356.00000			

Root MSE	4.36035	R-Square	0.9299
Dependent Mean	25.33333	Adj R-Sq	0.8878
Coeff Var	17.21192		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-4.63757	16.49829	-0.28	0.7899
x	1	5.29770	1.96447	2.70	0.0430
z1	1	5.15922	12.56373	0.41	0.6983
z2	1	2.81574	7.39602	0.38	0.7191

Test trt1vs3 Results for Dependent Variable y

Mean Square				
Source	DF	Square	F Value	Pr > F
Numerator	1	3.20609	0.17	0.6983
Denominator	5	19.01267		

Test trt2vs3 Results for Dependent Variable y

Mean Square				
Source	DF	Square	F Value	Pr > F
Numerator	1	2.75571	0.14	0.7191
Denominator	5	19.01267		

Test trt1vs2 Results for Dependent Variable y

Mean Square				
Source	DF	Square	F Value	Pr > F
Numerator	1	2.39182	0.13	0.7373
Denominator	5	19.01267		

20-9

Using SAS

```
options nocenter ls=80;

data example2;
  input trt x y @@;
cards;
  1 1.2 7 1 1.9 13 1 3.4 16
  2 4.0 6 2 5.2 13 2 5.8 14
  3 7.7 7 3 8.3 13 3 8.9 15
;

proc glm;
class trt;
model y=trt;

proc glm;
class trt;
model y=trt x;
lsmeans trt / tdiff adjust=tukey;
run;
quit;
```

20-10

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	100.6915501	33.5638500	10.81	0.0126
Error	5	15.5306721	3.1061344		
Corrected Total	8	116.2222222			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
trt	2	1.55555556	0.77777778	0.25	0.7877
x	1	99.13599459	99.13599459	31.92	0.0024

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	2	94.55407736	47.27703868	15.22	0.0075
x	1	99.13599459	99.13599459	31.92	0.0024

Adjustment for Multiple Comparisons: Tukey-Kramer

trt	y LSMEAN	LSMEAN Number
1	25.4075327	1
2	11.6977898	2
3	-2.4386558	3

Least Squares Means for Effect trt
t for HO: LSmean(i)=LSmean(j) / Pr > |t|

i/j	Dependent Variable: y		
	1	2	3
1		5.133597	5.483512
		0.0084	0.0064
2	-5.1336		4.72883
	0.0084		0.0119
3	-5.48351	-4.72883	
	0.0064	0.0119	

20-11

Nonconstant Slope in ANCOVA

- Statistical model for constant slope is

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + E_{ij} \quad \begin{cases} i = 1, 2, \dots, k \\ j = 1, 2, \dots, n_i \end{cases}$$

- Can allow for different slope by including interaction

$$Y_{ij} = \mu + \alpha_i + (\beta + (\beta\alpha)_i)(X_{ij} - \bar{X}) + E_{ij} \quad \begin{cases} i = 1, 2, \dots, k \\ j = 1, 2, \dots, n_i \end{cases}$$

- In SAS, simply add interaction term into model
- Can be done in REG or GLM
- Provides test for nonconstant slope

20-12

Using SAS

```
options nocenter ls=75;

data example1;
input trt x y @@;
cards;
1 1.2 7 1 1.9 13 1 3.4 16
2 4.0 20 2 5.2 22 2 5.8 32
3 7.7 31 3 8.3 45 3 8.9 42
;

/* With GLM, do not need to define interaction as a
variable prior to doing analysis */
proc glm;
class trt;
model y=trt x trt*x;
lsmeans trt / tdiff;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1278.409474	255.681895	9.89	0.0441
Error	3	77.590526	25.863509		
Corrected Total	8	1356.000000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	2	20.5146998	10.2573499	0.40	0.7034
x	1	149.7599282	149.7599282	5.79	0.0953
x*trt	2	17.4728475	8.7364237	0.34	0.7374

trt	LSMEAN		
	y LSMEAN	Number	
1	23.2379068	1	
2	25.5925926	2	
3	10.5092593	3	

Least Squares Means for Effect trt
t for HO: LSmean(i)=LSmean(j) / Pr > |t|

i/j	Dependent Variable: y		
	1	2	3
1		-0.22548	0.591
		0.8361	0.5961
2	0.225476		0.781205
	0.8361		0.4917
3	-0.591	-0.78121	
	0.5961	0.4917	

20-13

20-14