STAT 525 FALL 2018

Chapter 22
Analysis of Covariance

Professor Min Zhang

Background

- While the factor effects are of interest, variable X is also correlated with response Y
  - Can measure X but can't control it (block)
  - Nuisance variable X called a **covariate**
- ANCOVA adjusts Y for effect of X
- Combination of regression and ANOVA
- Without adjustment, effects of X may
  - Inflate σ^2
  - Alter treatment comparisons

Data for One-Way ANCOVA

- Y_{ij} is j^{th} observation of the response in the i^{th} level of the factor
- X_{ij} is j^{th} observation of the covariate in the i^{th} level of the factor
- $i = 1, 2, ..., r$
- $j = 1, 2, ..., n_i$

Examples

- **Pre-test/Post-test score analysis**: The change in score Y may be associated with GPA 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 GPAs and randomly assign them to a control and treatment group.
- **Weight gain experiments in animals**: If wishing to compare different feeds, the weight gain Y may be associated with the dominance of the animal. Analysis of covariance provides a way to use a herd and adjust for the dominance.

One-Way ANCOVA

- Statistical model is

$$Y_{ij} = \mu + \tau_i + \beta(X_{ij} - \overline{X}_{..}) + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots n_i \end{cases}$$

  - μ – overall mean
  - τ_i – fixed treatment effects subject to $\sum_{i=1}^{r} \tau_i = 0$
  - β – regression coefficient for the relation b/w Y and X

- Additional assumptions
  - X_{ij} is not affected by treatment
  - X and Y are linearly related
  - Constant slope (can be relaxed)

Estimation

- General Procedure:
  - Fit one-way model (Y = trt)
  - Fit one-way model (X = trt)
  - Regress residuals (resid$_Y$ = resid$_X$) for

    $$\hat{\beta} = \sum\sum (Y_{ij} - \overline{Y}_{i.})(X_{ij} - \overline{X}_{i.}) / \sum\sum (X_{ij} - \overline{X}_{i.})^2$$

  - Other model estimates are

    $$\hat{\mu} = \overline{Y}_{..}$$

    $$\hat{\tau}_i = \overline{Y}_{i.} - \overline{Y}_{..} - \hat{\beta}(\overline{X}_{i.} - \overline{X}_{..})$$

Hypotheses

- Test $H_0 : \tau_1 = \tau_2 = \ldots = \tau_r = 0$
  - Compare treatment means **after adjusting for differences among treatments due to differences in covariate levels**

$$F_0 = \frac{\text{SSM(Trt}|X)/(r-1)}{\text{SSE}/(n_T - r - 1)}$$

- Test: $\beta = 0$
  - SS regression (SSX): $\hat{\beta}^2 \sum\sum (X_{ij} - \overline{X}_{i.})^2$

$$F_0 = \frac{\text{SSX}/1}{SSE/(n_T - r - 1)}$$

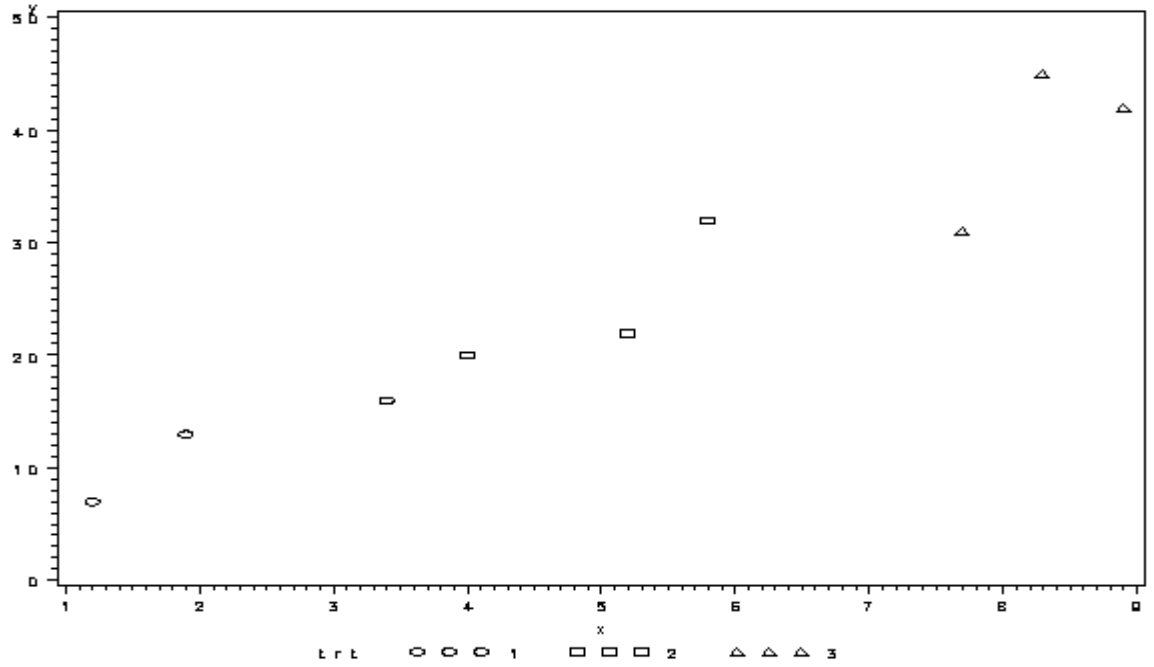
Mean Estimates

- Adjusted treatment means
  - Estimate: $\widehat{\mu}_i = \widehat{\mu} + \widehat{\tau}_i = \overline{Y}_{i.} - \widehat{\beta}(\overline{X}_{i.} - \overline{X}_{..})$
  - Expected value of Y when X is equal to the average covariate value
  - Can assume any value of X. Must make sure it is reasonable for all factor levels
  - Variance: $\widehat{\sigma}^2 \left(1/n + (\overline{X}_{i.} - \overline{X}_{..})^2 / \sum_i \sum_j (X_{ij} - \overline{X}_{i.})^2\right)$
- Pairwise differences
  - Estimate: $\widehat{\tau}_k - \widehat{\tau}_l = \overline{Y}_{k.} - \overline{Y}_{l.} - \widehat{\beta}(\overline{X}_{k.} - \overline{X}_{l.})$
  - Variance: $\widehat{\sigma}^2 \left(2/n + (\overline{X}_{k.} - \overline{X}_{l.})^2 / \sum_i \sum_j (X_{ij} - \overline{X}_{i.})^2\right)$

Example I

```
options nocenter ls=80;
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
;
proc sort data=example1; by trt;

symbol1 v=circle   i=  c=black;
symbol2 v=square   i=  c=black;
symbol3 v=triangle i=  c=black;
proc gplot data=example1;
    plot y*x=trt;
run; quit;
```



```
proc glm; class trt;
    model y=trt;
    output out=resid r=resy;
proc glm; class trt;
    model x=trt;
    output out=resid1 r=resx;

/*--- Regress Y Residuals vs X Residuals ---*/
proc glm;
    model resy=resx;
run; quit;
```

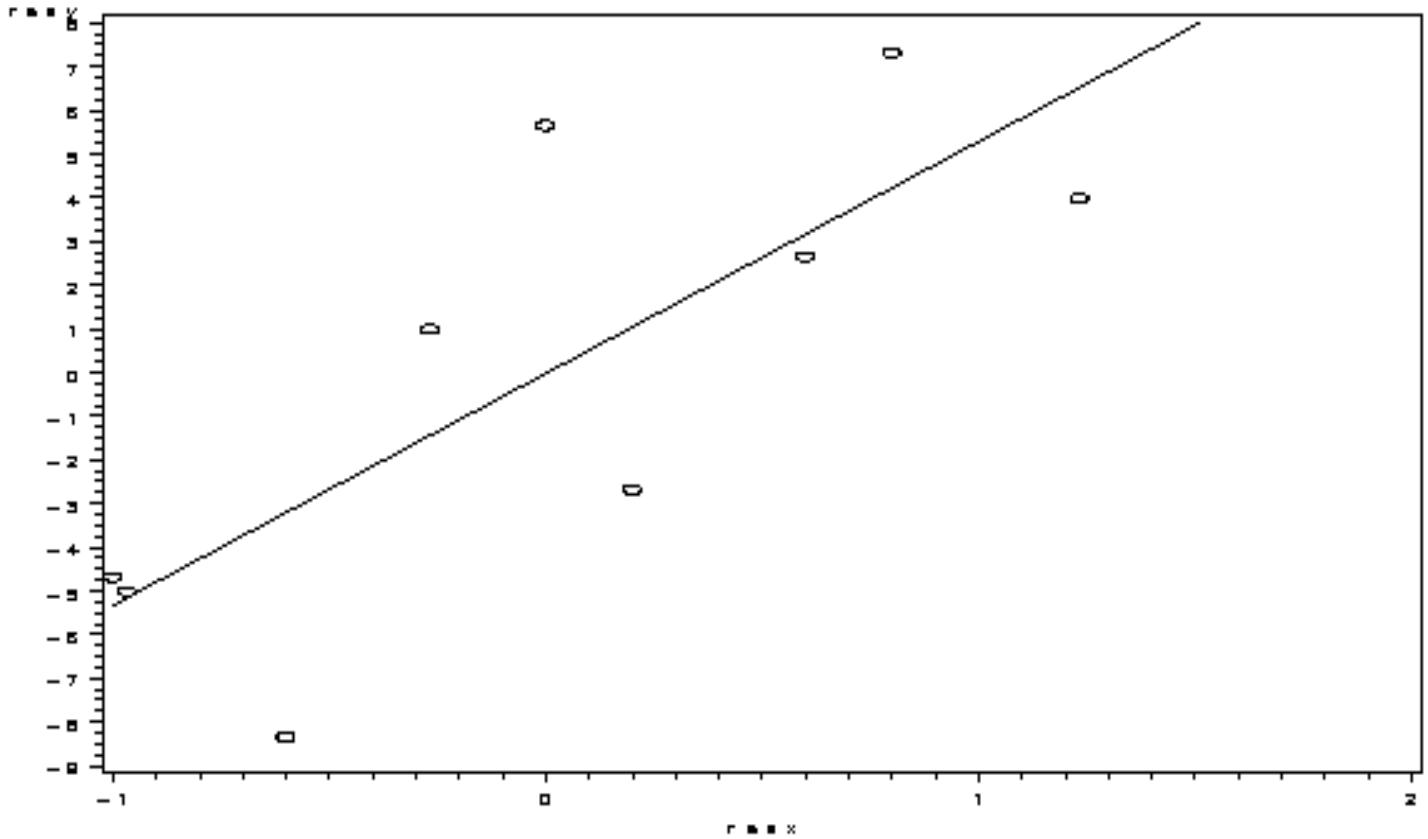
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	138.2699594	138.2699594	10.18	0.0153
Error	7	95.0633739	13.5804820		
Corrected Total	8	233.3333333			

R-Square	Coeff Var	Root MSE	resy Mean
0.592586	1.03728E17	3.685171	3.5527E-15

Source	DF	Type I SS	Mean Square	F Value	Pr > F
resx	1	138.2699594	138.2699594	10.18	0.0153

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.000000000	1.22839018	0.00	1.0000
resx	5.297699594	1.66027872	3.19	0.0153

```
/* Partial Regression Plot */
symbol1 v=circle i=rl;
proc gplot data=resid1;
    plot resy*resx;
run; quit;
```



```
proc glm data=example1;
    class trt;
    model y=trt x / solution;
run; quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1260.936626	420.312209	22.11	0.0026
Error	5	95.063374	19.012675		
Corrected Total	8	1356.000000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
trt	2	1122.666667	561.333333	29.52	0.0017
x	1	138.269959	138.269959	7.27	0.0430

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	2	3.2122606	1.6061303	0.08	0.9203
x	1	138.2699594	138.2699594	7.27	0.0430

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-4.637573297 B	16.49828508	-0.28	0.7899
trt	1	5.159224177 B	12.56372645	0.41	0.6983
trt	2	2.815741994 B	7.39601943	0.38	0.7191
trt	3	0.000000000 B	.	.	.
x		5.297699594	1.96446828	2.70	0.0430

```
/*-- WARNING : DO NOT USE MEANS STATEMENT --*/
proc glm data=example1;
    class trt; model y=trt x;
    means trt /lines tukey;
run; quit;
```

```
Tukey's Studentized Range (HSD) Test for y

Alpha                                      0.05
Error Degrees of Freedom                      5
Error Mean Square                      19.01267
Critical Value of Studentized Range     4.60173
Minimum Significant Difference           11.585

Means with the same letter are not significantly different.

           Mean      N    trt
A        39.333      3    3
B        24.667      3    2
C        12.000      3    1
```

- `MEANS` statement reports group means as $\widehat{\mu}_i = \bar{Y}_i$, ignoring the effect of covariate X;
- `MEANS` statement compares $\mu_i = \mu + \tau_i + \beta(X_{i\cdot} - \bar{X}_{\cdot\cdot})$;
- `LSMEANS` statement reports estimates of $\mu_i = \mu + \tau_i$, and compares them.

```
/*-- LSMEANS PROVIDES THE ADJUSTED MEANS --*/
proc glm data=example1;
    class trt; model y=trt x;
    lsmeans trt / tdiff adjust=tukey;
run; quit;
```

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|

Dependent Variable: y

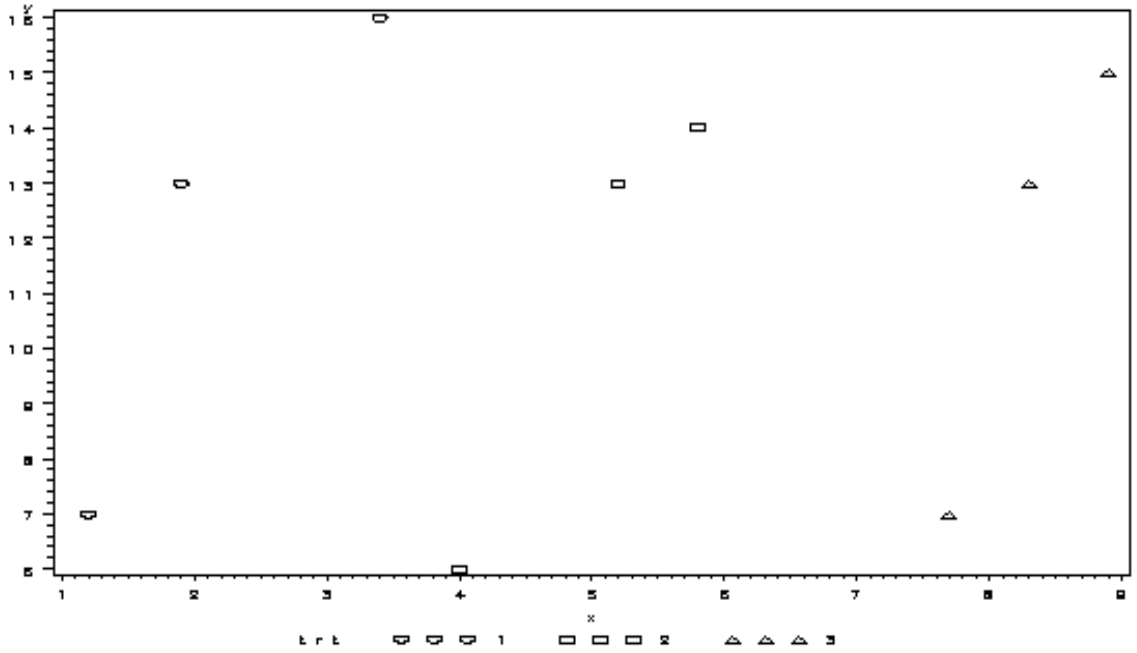
i/j	1	2	3
1		0.354685 0.7373	0.410644 0.6983
2	-0.35468 0.7373		0.38071 0.7191
3	-0.41064 0.6983	-0.38071 0.7191	

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

Example II

```
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 sort data=example2; by trt;

symbol1 v=circle   i=  c=black;
symbol2 v=square   i=  c=black;
symbol3 v=triangle i=  c=black;
proc gplot data=example2;
    plot y*x=trt;
run; quit;
```



```
proc glm data=example2;
    class trt;
    model y=trt x / solution;
run; quit;
```

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

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-25.56540370	B	6.66848712	-3.83	0.0122
trt	1	27.84618854	B	5.07816707	5.48	0.0028
trt	2	14.13644565	B	2.98941739	4.73	0.0052
trt	3	0.00000000	B	.	.	.
x		4.48579161		0.79402382	5.65	0.0024

```
proc glm data=example2;
    class trt;
    model y=trt x;
    lsmeans trt / tdiff;
run; quit;
```

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 H0: LSMean(i)=LSMean(j) / Pr > |t|

Dependent Variable: y

i/j	1	2	3
1		5.133597	5.483512
		0.0037	0.0028
2	-5.1336		4.72883
	0.0037		0.0052
3	-5.48351	-4.72883	
	0.0028	0.0052	

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

```
proc glm data=example2;
    class trt;
    model y=trt x;
    lsmeans trt / tdiff adjust=tukey;
run; quit;
```

```
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

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

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

             Dependent Variable: y
i/j               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
```

Summary on the Two Examples

- Both emphasize how covariate can change the treatment comparisons. Usually it just reduces the MSE.
- Example I: No treatment differences
  - Positive linear relationship
  - Covariate larger in each group
  - Thus, appears to be treatment difference
- Example II: Treatment differences exist
  - Positive linear relationship
  - Covariate larger in each group
  - Thus, no apparent treatment difference

Nonconstant Slope

- Can allow for different slope by including interaction

$$y_{ij} = \mu + \tau_i + (\beta + (\beta\tau)_i)(x_{ij} - \overline{x}_{..}) + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots n_i \end{cases}$$

- In SAS, simply add interaction term
- Provides test for nonconstant slope
- Can also build model for other relationships between X and Y (e.g., quadratic)

Example I

```
proc glm;
    class trt;
    model y=trt x / solution;
    lsmeans trt / tdiff;

proc glm;
    class trt;
    model y=trt x trt*x / solution;
    lsmeans trt / tdiff;
run; quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1260.936626	420.312209	22.11	0.0026
Error	5	95.063374	19.012675		
Corrected Total	8	1356.000000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	2	3.2122606	1.6061303	0.08	0.9203
x	1	138.2699594	138.2699594	7.27	0.0430

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

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-4.637573297	B	16.49828508	-0.28	0.7899
trt	1	5.159224177	B	12.56372645	0.41	0.6983
trt	2	2.815741994	B	7.39601943	0.38	0.7191
x		5.297699594		1.96446828	2.70	0.0430

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-36.75000000	B	49.83227932	-0.74	0.5143
trt	1	40.60356201	B	50.39772400	0.81	0.4794
trt	2	31.65476190	B	53.63535098	0.59	0.5966
x		9.16666667	B	5.99345810	1.53	0.2236
x*trt	1	-5.40677221	B	6.79395005	-0.80	0.4843
x*trt	2	-3.21428571	B	7.16355259	-0.45	0.6841

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

---

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

- Again, `LSMEANS` reports estimates of $\mu_i = \mu + \tau_i$ for the model on Slide 22-19.

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

Dependent Variable: y

i/j	1	2	3
1		0.354685 0.7373	0.410644 0.6983
2	-0.35468 0.7373		0.38071 0.7191
3	-0.41064 0.6983	-0.38071 0.7191	

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

Dependent Variable: y

i/j	1	2	3
1		-0.22548 0.8361	0.591 0.5961
2	0.225476 0.8361		0.781205 0.4917
3	-0.591 0.5961	-0.78121 0.4917	

Multi-Factor ANCOVA

- Can incorporate covariate into any model
- For two-factor model

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \beta(x_{ijk} - \overline{x}_{...}) + \epsilon_{ijk}$$

- Constant slope **for each** ij **combination**
- Can include interaction terms to vary slope
- Plot y vs x for each combination

Chapter Review

- One-way analysis of covariance
  - Data
  - Model
  - Inference
- Multi-Factor analysis of covariance
- Diagnostics and remedies