Analysis of Covariance

Design of Experiments - Montgomery
Section 14-3

Background

- Consider factor x which is correlated with y
- Can measure x but can't control it (block)
- Nuisance factor x called a covariate

- ANCOVA adjusts y for effect of covariate x
- Combination of regression and analysis of variance
- Without adjustment, effects of x may

  inflate σ^2

  alter treatment comparisons

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.

Model Description

- Consider single covariate in CRD
- Statistical model is

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

- Additional assumptions

  x_{ij} not affected by treatment

  x and y are linearly related

  Constant slope

- General Procedure:

  Fit one-way model (y = trt)

  Fit one-way model (x = trt)

  Regress residuals (residuals1 = residuals2)

  Model estimates are

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

  $\hat{\beta} = E_{xy}/E_{xx} = \sum\sum(y_{ij} - \overline{y}_{i.})(x_{ij} - \overline{x}_{i.})/\sum\sum(x_{ij} - \overline{x}_{i.})^2$

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

Analysis of Covariance

- Test $H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$
  - Compare treatment means after adjusting for differences among treatments due to differences in covariate levels
  - Trt and covariate not orthogonal (order of fit matters)

$$F_0 = \frac{\text{SS(trt}|x)/a-1}{\text{SS}_\text{E}/(N-a-1)}$$

- Adjusted treatment means
  - Estimate $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \overline{y}_{i.} - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{..})$
  - Variance: $\hat{\sigma}^2 \left(1/n + (\overline{x}_{i.} - \overline{x}_{..})^2 / \sum\sum (x_{ij} - \overline{x}_{i.})^2\right)$

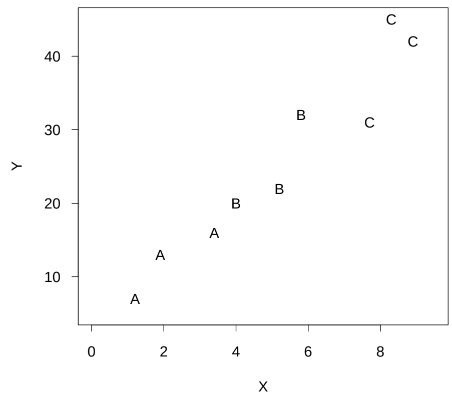
- Test: $\beta = 0$
  - Sum of Squares regression (SS$_x$): $\hat{\beta}^2 \sum\sum (x_{ij} - \overline{x}_{i.})^2$

$$F_0 = \frac{\text{SS}_x/1}{SS_\text{E}/(N-a-1)}$$

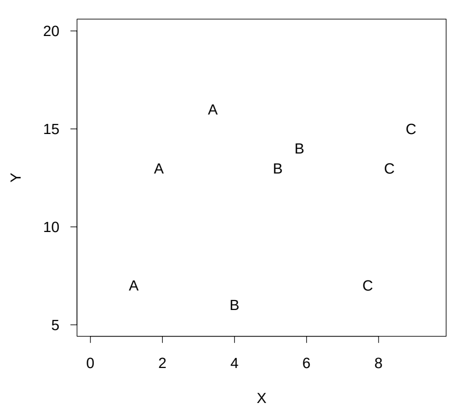
Analysis of Covariance

Two Examples

1 No treatment differences
  - Positive linear relationship
  - Covariate larger in each group
  - Thus, appears to be treatment difference



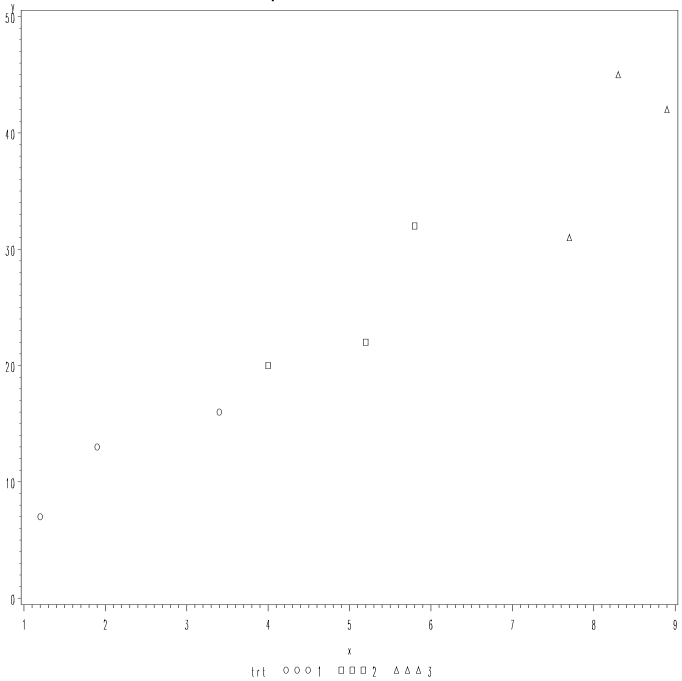
2 Treatment differences exist
  - Positive linear relationship
  - Covariate larger in each group
  - Thus, no apparent treatment difference



Using SAS

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

proc glm; class trt;
 model y=trt;
 output out=resid r=resy;
proc glm; class trt;
 model x=trt;
 output out=resid1 r=resx;
proc glm; model resy=resx;
symbol1 v=circle i=rl;
proc gplot; plot resy*resx;
run;

proc glm data=example1;
 class trt; model y=trt x / solution;
 means trt /lines lsd;
 lsmeans trt / tdiff adjust=t;
run;
```

Scatterplot of X vs Y



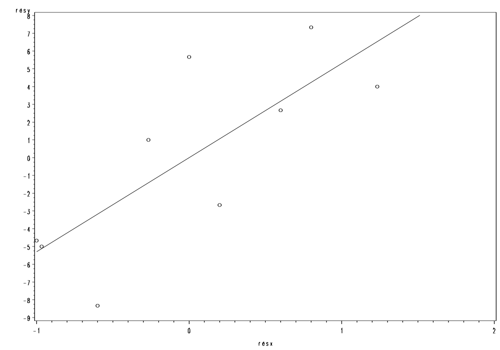
The GLM Procedure
Dependent Variable: resy

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



Dependent Variable: y

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

t Tests (LSD) for y

Alpha	0.05
Error Degrees of Freedom	5
Error Mean Square	19.01267
Critical Value of t	2.57058
Least Significant Difference	9.1518

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

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|

Dependent Variable: y

i/j	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.

```
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 data=example2;
 class trt; model y=trt x / solution;
 lsmeans trt / tdiff;
 lsmeans trt / tdiff adjust=bon;
run;
------------------------------------------------------------------------------
                                     Sum of
Source                    DF        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

                                       Standard
Parameter            Estimate             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
```



```
The GLM Procedure
Least Squares Means
                         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
                          0.0037        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.


Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

      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
                          0.0110        0.0083
   2       -5.1336                     4.72883
            0.0110                      0.0156
   3      -5.48351      -4.72883
            0.0083        0.0156
```


Regression Approach to ANCOVA

- Consider the following model ($a = 3$)

$$y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \epsilon_j$$

$j = 1, 2, \ldots N$

$X_{1j} = 1$ if Trt 1 and $X_{1j} = -1$ if Trt 3

$X_{2j} = 1$ if Trt 2 and $X_{2j} = -1$ if Trt 3

$X_{3j} = (x_j - \overline{x}_{..})$ x is the covariate

- Trt 1: $y_j = \beta_0 + \beta_1 + \beta_3(x_j - \overline{x}_{..}) + \epsilon_j$
- Trt 2: $y_j = \beta_0 + \beta_2 + \beta_3(x_j - \overline{x}_{..}) + \epsilon_j$
- Trt 3: $y_j = \beta_0 - \beta_1 - \beta_2 + \beta_3(x_j - \overline{x}_{..}) + \epsilon_j$
- Results in estimates

$\hat{\mu} = \hat{\beta}_0 \quad \hat{\tau}_1 = \hat{\beta}_1 \quad \hat{\tau}_2 = \hat{\beta}_2 \quad \hat{\beta} = \hat{\beta}_3$

Nonconstant Slope in ANCOVA

- Statistical model for constant slope is

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

- Can allow for different slope by including interaction

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

- In SAS, simply add interaction term into model
- Provides test for nonconstant slope

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
;

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

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;
```



```
                                     Sum of
Source                      DF      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

                                       Standard
Parameter           Estimate              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
-------------------------------------------------------------------------
                                     Sum of
Source                      DF      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

                                       Standard
Parameter           Estimate              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
```



```
                          LSMEAN
trt        y 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      0.410644
                          0.7373        0.6983
   2      -0.35468                     0.38071
            0.7373                      0.7191
   3      -0.41064      -0.38071
            0.6983        0.7191

--------------------------------------------------------------------

                          LSMEAN
trt        y LSMEAN      Number
1        23.2379068           1
2        25.5925926           2
3        10.5092593           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                    -0.22548         0.591
                          0.8361        0.5961
   2      0.225476                    0.781205
            0.8361                      0.4917
   3        -0.591      -0.78121
            0.5961        0.4917
```


Analysis of Covariance

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

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

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