ST518, Analysis of Covariance

Jason A. Osborne

N.C. State University

November 7, 2019

1. Introduction
2. General Linear ANCOVA Model
3. Exercise
4. Partial test for treatment effect
5. Adjusted means
   - SE of adj. mean
6. LSMEANS statement for adjusted means
7. Assumptions
8. ANCOVA plot

Analysis of covariance, ANCOVA

Covariates are predictive responses. Associations between covariates z and main response variable of interest y can be used to reduce unexplained variation σ^2.

An nutrition example

Nutrition scientist conducted expt. to evaluate effects of four vitamin supplements on weight gain of lab animals. Experiment conducted in a CRD with $N = 20$ animals randomized to $a = 4$ supplement groups, each with sample size $n \equiv 5$. Response variable of interest is weight gain, but calorie intake z measured concomitantly.

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	$s_4 = 10.9$

$s_1^2 \sim 144$ 225 100 120

Q: Is there evidence of a vitamin supplement effect?

$$MS(\text{diet}) = \frac{\sum_{1}^{4}\left(\sum_{1}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2\right)}{4-1} = 5 \text{ variance among trt avgs} = 5(53.2)$$

```
proc glm ;
   class diet;
   model y=diet;
run;
```

```
The GLM Procedure

Class              Levels     Values
diet                    4     1 2 3 4

Dependent Variable: y

                                    Sum of
Source                  DF         Squares     Mean Square    F Value    Pr > F
Model                    3      797.800000      265.933333       1.82    0.1836
Error                   16     2334.400000      145.900000
Corrected Total         19     3132.200000
```

- No evidence of effect
- ~~Evidence of no effect~~

H_0: no effect H_a: μ_i's of interest

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

But calorie intake z was measured concomitantly:

Diet	y	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	73	530	3	65	470	4	42	510
1	53	470	2	63	420	3	67	480	4	34	430

Q: How and why could these new data be incorporated into analysis?

A: ANCOVA can be used to reduce unexplained variation.

Model, given z_i,

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

where x_{ij} is an indicator variable for subject i receiving vitamin supplement j:

$$x_{ij} = \begin{cases} 1 & \text{subject } i \text{ receives supplement } j \\ 0 & \text{else} \end{cases}$$

and errors $\boxed{E_i \overset{iid}{\sim} N(0, \sigma^2).}$

Exercise: specify the parametric mean weight gain for the first subject in each treatment group, conditional on their caloric intakes.

Exercise

$\beta_0 + \beta_1 \qquad + \beta_2 350$

$$E(Y_1) = ? \quad \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_z \cdot Z_1 + E(\epsilon_1)$$

$$E(Y_6) = ? \quad \beta_0 + \qquad \beta_2 \qquad + \beta_z 400$$

$$E(Y_{11}) = ?$$

$$E(Y_{16}) = ? \quad \beta_0 \qquad + \beta_z 530$$

Proceeding with MLR analysis of this general linear model:

```
                          The GLM Procedure
                      Class Level Information
                  Class          Levels    Values
                  diet                4    1 2 3 4

Dependent Variable: y

                                       Sum of
 Source                      DF       Squares    Mean Square  F Value  Pr > F
 Model                        4   1951.680373     487.920093     6.20  0.0038
 Error                       15   1180.519627      78.701308
 Corrected Total             19   3132.200000

           R-Square     Coeff Var      Root MSE        y Mean
           0.623102      15.11308      8.871376      58.70000

 Source                      DF     Type I SS    Mean Square  F Value  Pr > F
 diet                         3    797.800000     265.933333     3.38  0.0463
 z                            1   1153.880373    1153.880373    14.66  0.0016

 Source                      DF   Type III SS    Mean Square  F Value  Pr > F
 diet                         3   1537.071659     512.357220     6.51  0.0049
 z                            1   1153.880373    1153.880373    14.66  0.0016
```

MS(E) was 145.9

4 avgs differ signif'cantly after adj for effect of z

To test for a diet effect: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, use the type III F-ratio, on 3 and 15 numerator and denominator degrees of freedom. (Note that this is a comparison of nested models.)

Q: Conclusion?

Reject Ho at any level .0049 or above

FYI: model was fit with the following code:

```
proc glm;
   class diet;
   model y=diet z;
   means diet;            unadjusted
   lsmeans diet/stderr;   adjusted means ~ estimated mean gain if we could balance the design wrt caloric intake
run;
```

NOTE: the drop in $\sqrt{MSE}$ (was $\hat{\sigma} \approx 12g$ is $\hat{\sigma} \approx 9g$)

Adjusted and unadjusted means

Recall the sample mean weight gains for the four diets (generated by the `means diet;` statement in `proc glm`):

```
                     The GLM Procedure

Level of        ------------y------------    ------------z-------------
diet       N          Mean        Std Dev          Mean          Std Dev

1          5    63.0000000     12.2678441    442.000000       58.9067059
2          5    57.8000000     14.8727940    434.000000       61.0737259
3          5    65.2000000      9.6540147    468.000000       36.3318042
4          5    48.8000000     10.8949530    502.000000       40.8656335
```

avg = 461.5

These means y are computed without taking z into account, so they are called unadjusted means.

= $\bar{z}$

Unadjusted means do not make any adjustment for the facts that

1. caloric intake may vary by diet (presumably by chance, not because of diet)
2. weight gain depends on caloric intake

Adjusted means

Adjusted means are estimated mean weight gains at a common reference value (sample mean, $\bar{z}$) of the covariate, z.

Here, $\bar{z} = (442 + 434 + 468 + 502)/4 = 461.5$. The adjusted means are then just

$$
\begin{aligned}
\bar{y}_{1,a} &= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_z(461.5) \\
\bar{y}_{2,a} &= \hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_z(461.5) \\
\bar{y}_{3,a} &= \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_z(461.5) \\
\bar{y}_{4,a} &= \hat{\beta}_0 + \hat{\beta}_z(461.5)
\end{aligned}
$$

"means adjusted to the average calorie intake, $\overline{z} = 461.5$"

SOLUTION option in MODEL statement of PROC GLM produces (nonuniquely estimable) parameter estimates that correspond to parameterization with diet 4 effect set to 0:

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		$\hat\beta_0$ -35.66310108	B	22.41252629	-1.59	0.1324
diet	1	$\hat\beta_1 =$ 24.29519136	B	6.19932022	3.92	0.0014
diet	2	20.44121688	B	6.35678835	3.22	0.0058
diet	3	$\hat\beta_3 =$ 22.12060844	B	5.80625371	3.81	0.0017
diet	4	0.00000000	B	.	.	.
z		$\hat\beta_z =$ 0.16825319		0.04394140	3.83	0.0016

```
NOTE: The X'X matrix has been found to be singular, and a generalized
      inverse was used to solve the normal equations.  Terms whose
 estimates are followed by the letter 'B' are not uniquely estimable.
```

Substitution of $\hat\beta$ into the expressions for adjusted means yields

$\hat\beta_0 \quad \hat\beta_1 \quad \hat\beta_z \quad \bar z$

$$
\begin{aligned}
n_1 = 5 \qquad \bar{y}_{1,a} &= -35.7 + 24.3 + 0.17(461.5) = 66.3 \qquad (\hat{SE} = 4.06)\\
\bar{y}_{2,a} &= -35.7 + 20.4 + 0.17(461.5) = 62.4 \qquad \hat{SE} = 4.15\\
\bar{y}_{3,a} &= -35.7 + 22.1 + 0.17(461.5) = 64.1 \qquad \hat{SE}\ 3.98\ *\\
n_4 = 5 \qquad \bar{y}_{4,a} &= -35.7 + +0.17(461.5) = 42.0
\end{aligned}
$$

* smallest

SLR

$$\boxed{\mathrm{Var}(\hat\beta_1) = \frac{\sigma^2}{\sum(x_i-\bar x)^2}} \qquad \hat\beta_1 = \boxed{\frac{\sum(x_i-\bar x)y_i}{\sum(x_i-\bar x)^2}}$$

$$\underset{(20\times1)}{Y} = \boxed{X}\underset{(5\times1)}{\beta} + \underset{(20\times1)}{E} \sim MVN(0, \sigma^2 I)$$

Standard errors of $\bar{y}_{j,a}$

Consider $\bar{y}_{2,a}$. What vector c is needed so that $c'\hat{\beta} = \bar{y}_{2,a}$?

$$\bar{y}_{2,a} = (1, 0, 1, 0, 461.5)\begin{pmatrix}\hat\beta_0\\ \hat\beta_1\\ \hat\beta_2\\ \hat\beta_3\\ \hat\beta_Z\end{pmatrix} = 62.9$$

What is the standard error of $c'\hat{\beta}$?

$$\widehat{\mathrm{Var}}(c'\hat\beta) = c'\widehat{\mathrm{Var}}(\hat\beta)c = c'\underline{\hat\sigma^2(X'X)^{-1}}c$$

$$= c' MS(E)(X'X)^{-1}c$$

To get SAS to produce the adjusted means and estimated standard errors, use an `LSMEANS` statement for the factor `diet` and a `STDERR` option:

```
                    The GLM Procedure
                  Least Squares Means

                                        Standard
diet          y LSMEAN                     Error     Pr > |t|

1           66.2809372                 4.0588750       <.0001
2           62.4269627                 4.1473443       <.0001
3           64.1063543                 3.9776677       <.0001
4           41.9857458                 4.3482563       <.0001
```

Concerns:
Aside from the usual residual-based checks for model adequacy, does treatment affect the covariate? To check this, one could carry out a one-way ANOVA treating z as a response variable and check for a diet effect on the mean of z:

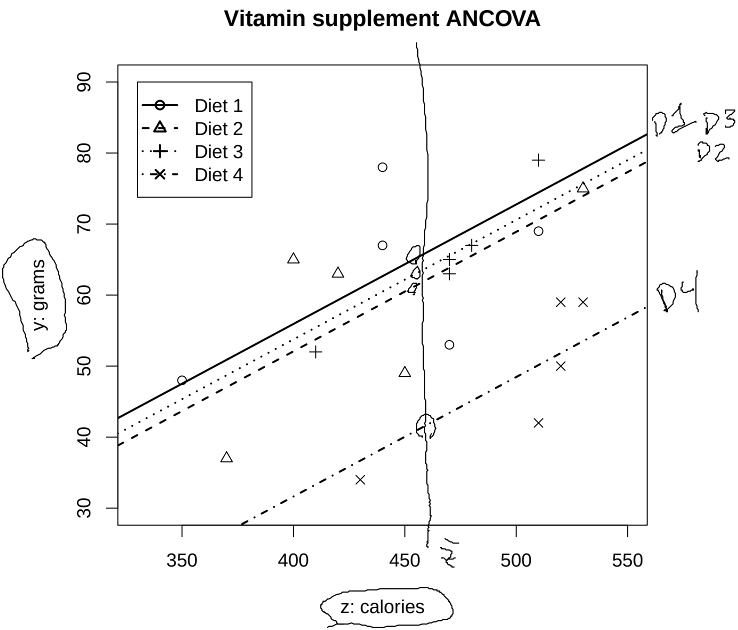
```
                               The GLM Procedure
Dependent Variable: z
                                           Sum of
 Source                      DF          Squares     Mean Square   F Value   Pr > F
 Model (diet)                 3      14095.00000      4698.33333      1.84   0.179
 Error                       16      40760.00000      2547.50000
 Corrected Total             19      54855.00000
```

A: No evidence that treatment affects covariate.

Q: Among the diets, which we've concluded are different, what are the differences? (Look at the means, have a guess.)

Q: If you are a lab animal and you want to gain weight, which diet(s) would you choose?

Q: Why are the standard errors for the adjusted means different?
Q: Which adjusted means require the most adjustment?



Vitamin supplement ANCOVA
Diet 1
Diet 2
Diet 3
Diet 4
y: grams
z: calories
30
40
50
60
70
80
90
350
400
450
500
550
D1 D3
D2
D4

```
vitsupp.dat <- read.table("vitsupp.txt",header=TRUE)
vitsupp.dat$z <- 10*vitsupp.dat$z
pdf(file="vitsupp1.pdf")
par(cex=1.2)
attach(vitsupp.dat)
plot(z,y,pch=Diet,main="Vitamin supplement ANCOVA",xlab="z: calories",
     ylab="y: grams", xlim=c(330,550),ylim=c(30,90))
legend(330,90,legend=c("Diet 1","Diet 2","Diet 3","Diet 4"),pch=1:4,
     lty=1:4,lwd=2)

vitsupp.fit <- lm(y~as.factor(Diet)+z)
betahat <- coef(vitsupp.fit)
abline(betahat[1],betahat[5],lwd=2)
abline(sum(betahat[1:2]),betahat[5],lwd=2,lty=2)
abline(sum(betahat[c(1,3)]),betahat[5],lwd=2,lty=3)
abline(sum(betahat[c(1,4)]),betahat[5],lwd=2,lty=4)

dev.off()
```