20. THE ANALYSIS OF COVARIANCE

20.1 INTRODUCTION

The analysis of covariance is a technique whose purpose is to reduce the amount of error in the experimental results (i.e. an error control technique). This method is applicable when there is one or several variables, x_1 , x_2 ,... which take values for each expermental unit. These values cannot be controlled in the experimental context and we have the values they assume for each experimental unit. Such variables are referred to as *covariates* or concommitant variables. The analysis of covariance technique removes the effects of these variables thus changing our estimate of experimental error. If the effects of the covariates are large then removing their effects tends to reduce our estimate of error. For example, consider an experiment to assess the efficacy of 4 different types of clinical treatment of anorexia as measured by weight gain over a specified period. It seems reasonable that the final weight of the subject will depend on initial weight as well. If we cannot control for the initial weights of the subjects in the experiment then this value is a covariate.

We consider here the simplest problem where we have one qualitative factor design and one quantitative covariate for which we observe at most an affine (linear) dependence. The analysis of this situation should make clear how to proceed in more complicated contexts.

Denote the factor A and the concommitant variable by X, suppose we obtain data

$$\begin{array}{lll} A_1: & x_{11} \ldots x_{1n} \ldots\ldots & A_a: \quad x_{a1} \ldots x_{an} \\ & Y_{11} \ldots y_{1n} \ldots\ldots & \quad\quad y_{a1} \ldots y_{an} \end{array}$$

The linear model is given by

$$\boxed{E[y|x] = \beta_1 w_1 + \ \ldots\ + \beta_a w_a + \beta x}$$

where $w_i = 1$ if the response is from treatment i and is zero otherwise.
This gives design matrix:

$$W = \begin{vmatrix} 1 & 0 & \ldots & 0 & x_1 \\ 1 & . & & . & . \\ 1 & . & & . & . \\ 1 & 0 & & . & x_{1n} \\ 0 & 1 & & . & \\ 0 & 1 & & . & \\ . & 1 & & 0 & \\ . & 0 & & 1 & \\ . & . & & 1 & \\ 0 & 0 & \ldots & 1 & x_{aa} \end{vmatrix}$$

If we wish to test the hypothesis that there is no cause-effect relationship between therapy and weight gain

i.e. $H_0 : \beta_1 = \beta_2 = ... = \beta_a$

we construct the following ANOVA table:

Source	DF
Mean	1
Affine (Linear)	1
A	a-1
Error	(n-1)(a)-1
Total	na

We test H_0 by comparing [MS(A)/MS(E)] with [F(a-1), (n-1)(a-1)] distribution. Generally we are not interested in testing whether there is an effect due to the covariate.

In this simple context we can obtain a formulae for the least squares estimates and the relevant SS.

These are:

$b_i = \frac{T_i(y)}{n} - \frac{bT_i(X)}{n}$ where T_i(y) is the total of all y data values from A_i.

$$b = \frac{S_{xy} - \frac{\sum T_i(x)T_i(y)}{n}}{S_{xx} - \frac{\sum T_i(x)}{n}}$$

$$SS_{xx} = \sum (x_i - \bar{x})^2 \qquad S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

This gives us the following computational formula.

Source	SS
Mean	$\frac{G^2}{na}$
Affine (linear)	$\frac{(n-1)s_{xy^2}}{s_{x^2}}$
A	$\frac{\sum T_{i^2}(y)}{n} + \frac{(x'y - \sum T_i(x)T_i(y))^2}{n}$ / $x'x - \frac{\sum T_{i^2}(x)}{n}$
Error	SS (Affine) - SS (Mean) Subtraction
	$y'y$

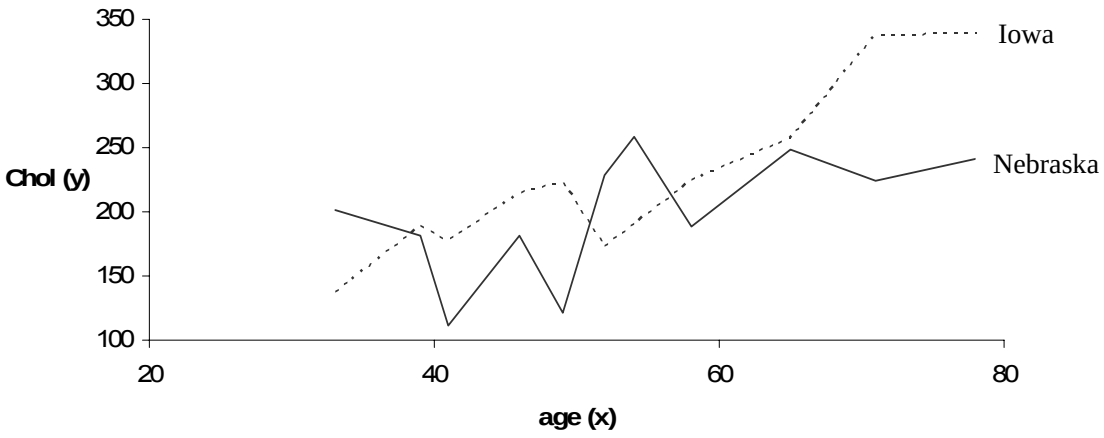
20.2 COVARIANCE

Swanson, P. et al, *J. of Gerontolog*y, 10, 41-47, 1955, examined how cholesterol concentration varied in women. Two states were considered, Iowa and Nebraska. Age was also recorded since it is known to influence cholesterol. The data are given below for a sample of 22 women.

AGE AND CONCETRATION OF CHOLESTEROL (MG/100 ML) IN THE BLOOD SERUM OF IOWA AND NEBRASKA WOMEN

Iowa. n = 11		Nebraska. n = 11	
Age X	Cholesterol Y	Age X	Cholesterol Y
46	181	18	137
52	228	44	173
39	182	33	177
65	249	78	241
54	259	51	225
33	201	43	223
49	121	44	190
76	339	58	257
71	224	63	337
41	112	19	189
58	189	42	214

A graph of the data shows the linearity of the model.



Computer Implementation

```
data list/ state 1 age 3-4 chol 6-8
1 46 181
1 52 228
. . .
. . .
. . .
2 19 189
2 42 214
end data
```

If you have already entered the above data into the data editor, open a new syntax file and type in the below commands.

```
MANOVA
  CHOL BY STATE (1, 2) WITH AGE
 /PRINT HOMEGENEITY (BARTLETT COCHRAN)
 /NOPRINT PARAM(ESTIM)
  /PLOT CELLPLOTS
  /RESIDUALS CASEWISE PLOTS
  /OMEANS TABLES (STATE)
  /PMEANS TABLES ( STATE )
 /METHOD=UNIQUE
 /ERROR WITHIN+RESIDUAL.
```

Once you have completed typing the commands, highlight the entire text and press the small black error located on the tool bar.

The means for the variables are given below.

Adjusted and Estimated Means

Variable .. CHOL

CELL	Obs. Mean	Adj. Mean	Est. Mean	Raw Resid.	Std. Resid.
1	207.727	197.099	207.727	.000	.000
2	214.818	225.446	214.818	.000	.000

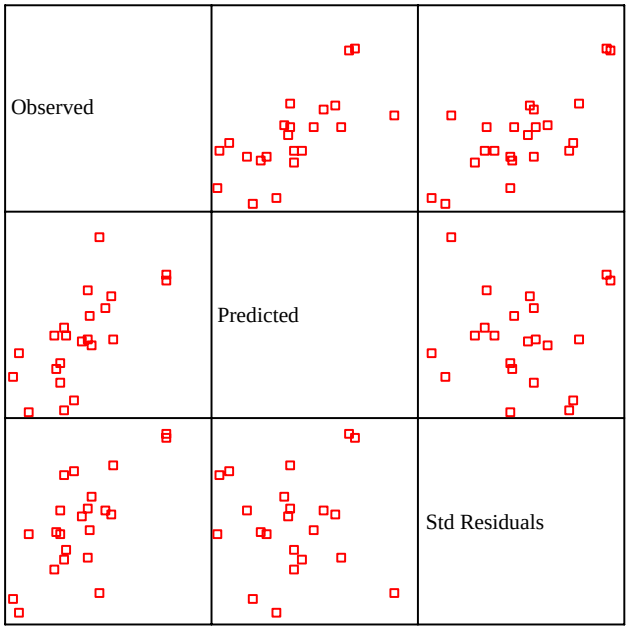
The ANOVA table indicates the covariate (REGRESSION) age was significant, $F(1,19) = 17.60$, $p<.001$. The differences between states was not significant, $F(1,19) = 2.19$, $p = .15$. Although the covariate significantly reduced error, the treatment effects were not significant. This result may be due to the small sample size.

Tests of Significance for CHOL using UNIQUE sums of squares

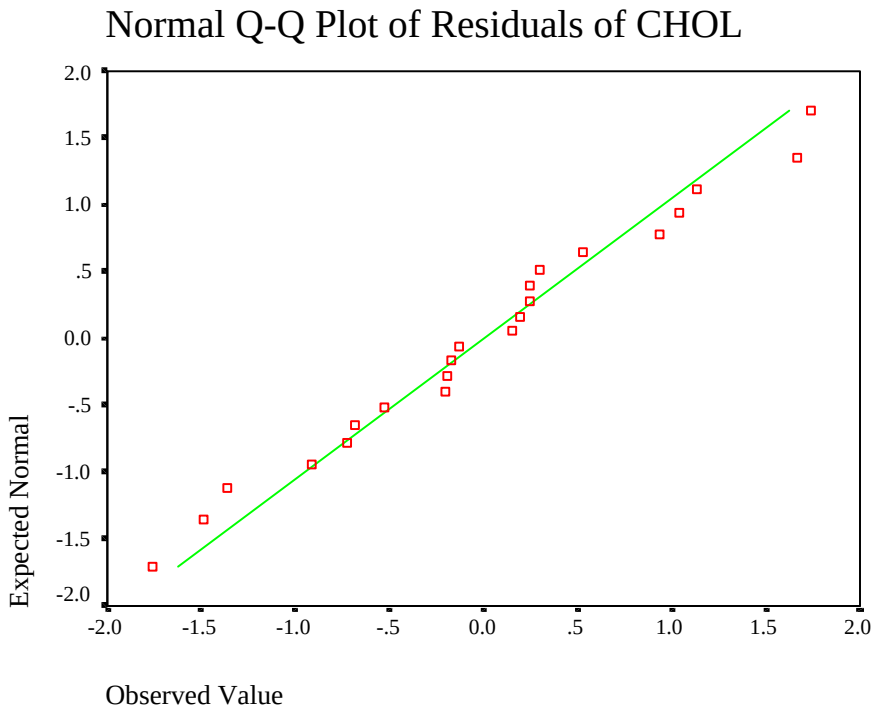
Source of Variation	SS	DF	MS	F	Sig of F
WITHIN+RESIDUAL	35702.08	19	1879.06		
REGRESSION	33077.74	1	33077.74	17.60	.000
STATE	4110.60	1	4110.60	2.19	.156

The residuals in both cases look quite reasonable. The scatterplot below displays the predicted values versus the standardized residuals.

Dependent variable: CHOL



The graph below is a Normal Probability Plot which plots the observed, predicted, and residual case values. As mentioned above, this graph also looks reasonable.



20.3 Exercises

1. Cochran (1958) gives data on how well patients respond to different drugs in the treatment of leprosy. X is the score before treatment and Y is the score after 6 months of treatment. Drugs A and B are antibiotics whereas drug C is a control. The data are given below.

SCORES FOR LEPROSY BACILLI BEFORE (X) AND AFTER (Y) TREATMENT

Drugs					
A		B		C	
X	Y	X	Y	X	Y
11	6	6	0	16	13
8	0	6	2	13	10
5	2	7	3	11	18
14	8	8	1	9	5
19	11	18	18	21	23
6	4	8	4	16	12
10	13	19	14	12	5
6	1	8	9	12	16

11	8	5	1	7	1
3	0	15	9	12	20

a. Plot the data.

b. Perform the appropriate ANOVA.

c. Compare the two drugs and the two drugs vs. the control. Are the contrasts orthogonal?

d. Analyse the data as fully as possible.

2. Smith, G. and Beecher, H., *J. of Pharm and Exp. Therapy*, 1962, 136, 47-56, examined how different drugs affect mental activity. The treatments were morphine, heroin, and placebo. Scores on participants were taken before and after treatment. The data are given below.

MENTAL ACTIVITY SCORES BEFORE (X) AND TWO HOURS AFTER (Y) A DRUG

Subject	Morphine		Heroin		Placebo	
	X	Y	X	Y	X	Y
1	7	4	0	2	0	7
2	2	2	4	0	2	1
3	14	14	14	13	14	10
4	14	0	10	0	5	10
5	1	2	4	0	5	6
6	2	0	5	0	4	2
7	5	6	6	1	8	7
8	6	0	6	2	6	5
9	5	1	4	0	6	6
10	6	6	10	0	8	6
11	7	5	7	2	6	3
12	1	3	4	1	3	8
13	0	0	1	0	1	0
14	8	10	9	1	10	11
15	8	0	4	13	10	10
16	0	0	0	0	0	0
17	11	1	11	0	10	8
18	6	2	6	4	6	6
19	7	9	0	0	8	7
20	5	0	6	1	5	1
21	4	2	11	5	10	8
22	7	7	7	7	6	5
23	0	2	0	0	0	1
24	12	12	12	0	11	5

a. Plot the data.

b. Perform the ANOVA analysis. State your conclusion clearly.

c. Comment on the residuals.

3. A clinical psychologist would like to study weight gain under four different treatment conditions. Initial age and weight were also recorded since it is known they affect final weight. The data for 40 individuals is given below. The means for both initial age and weight across treatments are very similar. This was by design and is called *balancing*. This balancing produces a more accurate comparison of Y. The data are given below.

INITIAL AGE (X_1), INITIAL WEIGHT (X_2) AND RATE OF GAIN (Y)

Treatment 1			Treatment 2		
Initial Age, X_1 (days)	Weight X_2 (pounds)	Gain Y (pounds/day)	Initial Age X_1 (days)	Weight X_2 (pounds)	Gain Y (pounds/day)
78	61	1.40	78	74	1.61
90	59	1.79	99	75	1.31
94	76	1.72	80	64	1.12
71	50	1.47	75	48	1.35
99	61	1.26	94	62	1.29
80	54	1.28	91	42	1.24
83	57	1.34	75	52	1.29
75	45	1.55	63	43	1.43
62	41	1.57	62	50	1.29
67	40	1.26	67	40	1.26
Treatment 3			Treatment 4		
78	80	1.67	77	62	1.40
83	61	1.41	71	55	1.47
79	62	1.73	78	62	1.37
70	47	1.23	70	43	1.15
85	59	1.49	95	57	1.22
83	42	1.22	96	51	1.48
71	47	1.39	71	41	1.31
66	42	1.39	63	40	1.27
67	40	1.56	62	45	1.22
67	40	1.36	67	39	1.36

a. Perform the appropriate ANOVA analysis.

b. Comment on the residuals.

c. Compare $T_1 + T_2$ with $T_3 + T_4$.

Compare T_1 with T_2.
Compare T_3 with T_4.

Using contrasts.

d. Assume the 4 treatments came from a 2 x 2 factorial.

	B_1	B_2
A_1	T_1	T_2
A_2	T_3	T_4

Perform the appropriate ANOVA analysis.