BIOMETRICS INFORMATION

(You're 95% likely to need this information)

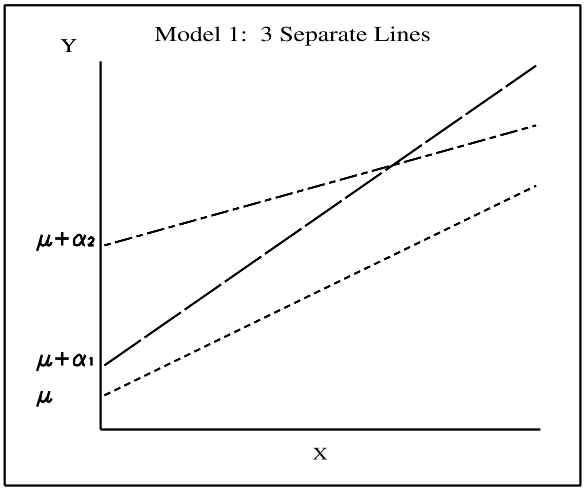
PAMPHLET NO. # 31 DATE: June 24, 1991

SUBJECT: ANCOVA: The Linear Models behind the F-tests

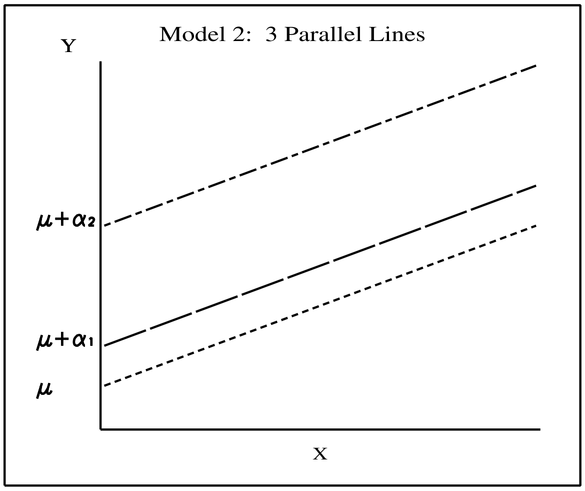
A common data set might have k = 3 treatment groups with a response, Y_i, and a covariate, X_i, recorded for each treatment unit. One-way Analysis of Covariance (ANCOVA) would be typically used to analyse this data. But there are also four other linear models which could be fitted. Tests between appropriate pairs of these models provide the F-tests of ANCOVA and can be used to select the model that best fits the data (which may not be ANCOVA).

The ANCOVA model is that of k = 3 parallel lines, each with the same slope, β, but with different intercepts which can be written as: $\mu + \alpha_1$, $\mu + \alpha_2$, and μ. Hence group 1 has intercept $\mu + \alpha_1$, group 2 has intercept $\mu + \alpha_2$, and group 3 has intercept μ (see graph for model 2 below). The difference in intercept between groups 1 and 3 is α_1 and between groups 2 and 3 is α_2. This notation for the intercepts may seem convoluted but corresponds closely to that used by SAS and to the parameter[1] estimates which SAS will output if the `SOLUTION` option is added to the `MODEL` statement in `PROC GLM`.

For 3 groups, the ANCOVA model has four parameters, and restricts the slopes to the same value. A less restricted model would fit 3 separate lines to the data, allowing each group to have different intercepts: $\mu + \alpha_1$, $\mu + \alpha_2$, and μ; and different slopes: $\beta + \beta_1$, $\beta + \beta_2$, and β. Note that the slope parameters are symbolized or parameterized in the same manner as the intercepts, so that β_1 and β_2 are the differences in slope between groups 1 and 3 and groups 2 and 3 respectively. This model has six parameters, instead of four. These two models are pictured below:



$$Y_i = \mu + \alpha_j + (\beta + \beta_j)\,X_i$$



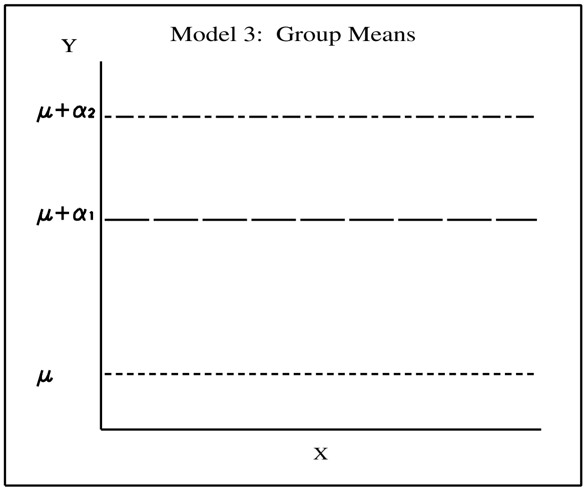
$$Y_i = \mu + \alpha_j + \beta\,X_i$$

[1] Note that β, μ, α_1, and α_2 are parameters.

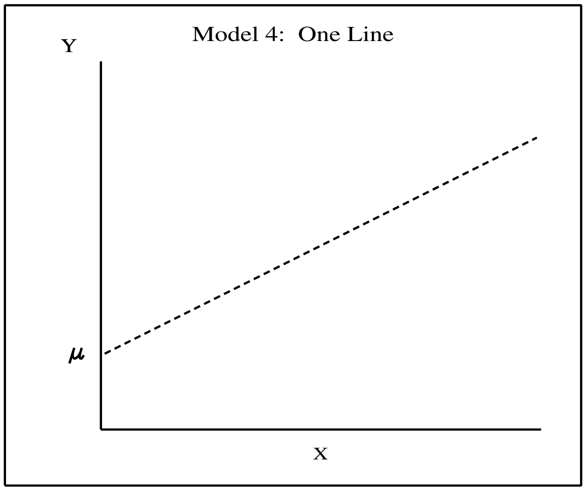
BRITISH COLUMBIA
Ministry of Forests
Research Program

The first logical step is to test an assumption of ANCOVA: that of homogeneity of regression or parallel lines (same slope). This test is accomplished by first fitting models 1 and 2 to the data and calculating a residual sums of squares (SSR) for each, denoted by SSR**KL** and SSR**KLP** respectively. Each SSR is the sum of the squared differences between the observed data and the values predicted by the model and has an associated degrees of freedom (df). This df is the number of observations in the dataset minus the number of parameters in the model. The difference of SSR**KL** and SSR**KLP** is also a sums of squares (SS) with a df that is equal to the number of parameters in model 1 but not in model 2. These parameters are β_1 and β_2, so the df is 2. This SS and df are used in an F-test to test the null hypothesis that β_1 and β_2 are zero and unnecessary in model 1. If this is the case, then model 2 provides a adequate fit to the data making the more complicated model 1 unnecessary.

Other models can also be fit to this data for tests of the treatment groups (K) and the covariate (X). These models are:



$Y_i = \mu + \alpha_j$



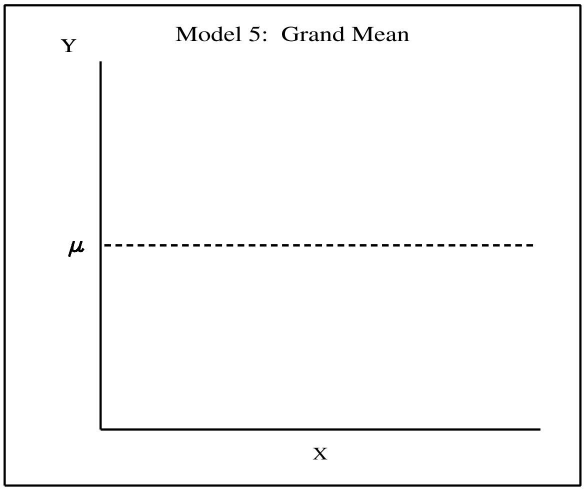
$Y_i = \mu + \beta X_i$

with the simplest model of all, the grand mean, shown on the next page.

The five models used for the F-tests are mathematically described by:

1) 3 separate lines: $Y_i = \mu + \alpha_j + (\beta + \beta_j)X_i$, for j = 1, 2, 3 (with α_3 and β_3 set to 0)
2) 3 parallel lines: $Y_i = \mu + \alpha_j + \beta X_i$, for j = 1, 2, 3 (with α_3 set to 0)
3) 3 group means: $Y_i = \mu + \alpha_j$, for j = 1, 2, 3 (with α_3 set to 0)
4) one line: $Y_i = \mu + \beta X_i$
and 5) the grand mean: $Y_i = \mu$

where μ is the parameter for a mean in models 3 and 5, and an intercept in models 1, 2, and 4. In model 3, the parameters, α_1 and α_2, are the differences of the means for treatments 1 and 2 with the mean, μ, of treatment 3. In models 1 and 2, α_1 and α_2 are the differences between the intercepts for treatments 1 and 2 with the intercept, μ, for treatment 3. For models 1, 2 and 3, μ is the mean or intercept for the third treatment. The slope of the covariate, X_i, is given by β in models 2 and 4. For model 1, β_1 and β_2 are the differences in slope between treatments 1 and 2 with the slope, β, of treatment 3.



$Y_i = \mu$

Each of these models is fit to the data and a residual sums of squares (SSR) calculated. Appropriate differences of SSR's are calculated between models where one contains a subset of the parameters of the other. For instance, models 2 to 5 are all subsets of model 1. These SS are associated with one or more parameters in the longer model and are used to test the null hypothesis that those parameter(s) are zero, and unnecessary in the longer model.

Data from Huitema (pg 38) will serve as an example (the data are in a SAS program in the Appendix). There are k = 3 groups. The most complicated linear model to fit to this data is 3 separate lines for a total of 6 parameters. The results of fitting the 5 different models is shown below:

	Number of		SSR[2]	SSR		Proportion
Model	Parameters	df	Name	Value	Difference	of SST (R^2)
1. K Separate Lines	2k = 6	24	SSR**KL**	1626.69		0.41
					SSHR = 53.96	0.01
2. K Parallel Lines	k+1 = 4	26	SSR**KLP**	1680.65		0.42
					SSCov = 1855.35	0.47
3. Group Means	k = 3	27	SSR**WG**	3536.00		0.89
					not meaningful	--
4. One Line	2	28	SSRL	2388.64		0.60
					SSL = 1567.36	0.40
5. Grand Mean	1	29	SST	3956.00		1.00

[2] This column identifies a unique name for the residual sums of squares from each of the 5 models.

The testing should proceed in the following logical order:

1. Test for parallel lines or homogeneity of regresssion, i.e., **H_o: Is it reasonable to decide that the k lines have the same slope? or Is it reasonable that β_1 and β_2 are zero?** Models 1 and 2 are compared by:

   SS**HR** = SSR**KLP** - SSR**KL** = 1680.65 - 1626.60 = 53.96, df = 2.

   and so $F = \frac{53.96/2}{1626.69/24} = 0.40$, df = 2, 24, p = 0.68

   hence, there the evidence against the hypothesis of parallel lines is very weak.

2. Given homogeneity of regression, test the covariate, i.e., **H_o: Is the slope of the covariate (β) zero?** Models 2 and 3 are compared by:

   SS**Cov** = SSR**WG** - SSR**KLP** = 3536.00 - 1680.65 = 1855.35, df = 1.

   and so $F = \frac{1855.35/1}{1680.65/26} = 28.70$, df = 1, 26, p = 0.0001

   hence, there is very strong evidence against the hypothesis that β in model 2 is zero.

3. Given homogeneity of regression, test for group differences, i.e., **H_o: Are the k lines coincident? i.e. Are α_1 and α_2 in model 4 zero?** If there are no group differences then the k lines should be the same line, hence models 2 and 4 are compared by:

   SS**BG**$_1$ = SSRL - SSR**KLP** = 2388.64 - 1680.65 = 707.99, df = 2.

   and so $F = \frac{707.99/2}{1680.65/26} = 5.48$, df = 2, 26, p = 0.010

   hence, there is strong evidence against the hypothesis of no group differences.

4. Given homogeneity of regression and an ineffective covariate (tests 1 & 2 above), test for group differences, i.e. **H_o: Are the k means different? i.e., Are α_1 and α_2 in model 3 zero?** This is the usual ANOVA test (with the usual error term). Models 3 and 5 are compared by:

   SS**BG**$_2$ = SST - SSR**WG** = 3956.00 - 3536.00 = 420, df = 2.

   and so $F = \frac{420/2}{3536.00/27} = 1.60$, df = 2, 27, p = 0.22

   hence, there is little evidence against the hypothesis of no differences between the means.

5. Given homogeneity of regression and coincident lines (tests 1 & 3 above), test the slope of the one line model, i.e. **H_o: Is the slope of the one line zero? or Is β in model 2 zero?** This is the usual simple regression test. Models 4 and 5 are compared for this test by:

   SSL = SST - SSRL = 3956.00 - 2388.64 = 1567.37, df = 1.

   and so $F = \frac{1567.37/1}{2388.64/28} = 18.37$, df = 1, 28, p = 0.0002

   hence, there is very strong evidence against the hypothesis of zero slope for one line.

Assuming homogeneity of regression (test 1 above) the following is the final form of the ANCOVA table.

Source of Variation	Sums of Squares Notation	df	Sums of Squares	Mean Square	F-value	p-value
Between Groups	SS**BG**$_1$ = SSR**L**-SSR**KLP**	2	707.99	354.00	5.48	0.010
Covariate	SS**Cov** = SSR**WG**-SSR**KLP**	1	1855.35	1855.35	28.70	0.0001
Error	SSR**KLP**	26	1680.65	64.64		
Total	SST	29	3956.00			

The following is the SAS output for the ANCOVA (see program in the Appendix). Note that the correct SS's and tests are from the TYPE III Sums of Squares. K is the variable denoting the different groups, with X denoting the covariate.

```
                       General Linear Models Procedure

Dependent Variable: Y
                                    Sum of           Mean
Source                  DF         Squares         Square   F Value      Pr > F

Model                    3     2275.351579     758.450526     11.73      0.0001
Error                   26     1680.648421      64.640324
Corrected Total         29     3956.000000

                  R-Square            C.V.       Root MSE                Y Mean
                  0.575165        22.97120       8.039921            35.0000000

Source                  DF       Type I SS    Mean Square   F Value      Pr > F

K                        2      420.000000     210.000000      3.25      0.0550
X                        1     1855.351579    1855.351579     28.70      0.0001

Source                  DF     Type III SS    Mean Square   F Value      Pr > F

K                        2      707.991625     353.995812      5.48      0.0104
X                        1     1855.351579    1855.351579     28.70      0.0001
```

It is important to note that there are two Between Group SS's, SS**BG**$_1$ and SS**BG**$_2$, that could be used to test for group differences. These two values will only be the same when each group has exactly the same covariate values (X_i). They correspond to two different hypotheses about the data, namely:

1) SS**BG**$_1$ tests for group differences **after** including the covariate in the model (i.e. the differences between models 2 and 4). The Type III SS in the SAS output above gave this value.

2) SS**BG**$_2$ tests for group differences **before** including the covariate in the model (i.e. the differences between models 3 and 5). This is the usual ANOVA SS for between groups and is given by the Type I SS in the above output (since K occurrs **before** X in the `MODEL` statement).

Similarly, there are two SS's that could be used to test the covariate, SS**Cov** and SS**L**.

1) SS**Cov** tests for the covariate **after** groups have been included in the model (i.e. the differences between models 2 and 3). The Type III SS in the SAS output above gave this value.

2) SS**L** tests for the covariate **before** groups are added to the model (i.e. the differences between models 4 and 5). The Type I SS would provide this value **if** X had been put in the `MODEL` **before** K.

The correct SS's for the ANCOVA are the **last-in** SS's obtained from the Type III SS's output by SAS and corresponding to SS**BG**$_1$ and SS**Cov**.

References:

Draper, N.R. and H. Smith. 1981. *Applied Regression Analysis*. John Wiley and Sons, New York, New York.

Huitema, B.E., 1980. *The Analysis of Covariance and Alternatives*. John Wiley and Sons, New York, New York.

Winer, B.J., 1971. *Statistical Principles in Experimental Design*. 2nd ed., McGraw-Hill Book Co., New York, New York.

CONTACT: Wendy Bergerud or Vera Sit
387-5676 356-0435

NEW PROBLEM

Do an ANCOVA analysis on the following data taken from Winer, page 794.

Group:	1		2		3		4		5		6	
	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
	3	8	2	14	3	16	7	18	0	8	0	10
	5	16	1	11	2	10	0	7	4	16	1	15
	1	10	8	20	1	14	4	10	8	20	9	26
	9	24	7	15	2	14	6	15	5	18	4	18
			4	12	6	22	9	23			4	18
					2	16					7	26
											8	24

PROBLEM FROM BI#30

The null hypotheses for the three contrasts are:

1. H_o: The response to the control is no different from the response due to the treatments.
2. H_o: The response to the new fertilizer is no different from that of the standard fertilizer.
3. H_o: The response to the two different levels of new fertilizer used is not different.

Appendix: The SAS program required to do the analyses described in the text.

```
/* Huitema.SAS */

title 'ANCOVA Example:  Data taken from Huitema, page 38';
data ancova;
  do k = 1 to 3;
   input x y @@;
   output;
  end;
cards;
29 15 22 20 33 14
49 19 24 34 45 20
48 21 49 28 35 30
35 27 46 35 39 32
53 35 52 42 36 34
47 39 43 44 48 42
46 23 64 46 63 40
74 38 61 47 57 38
72 33 55 40 56 54
67 50 54 54 78 56
run;
```

```
proc glm;
 class k;
 model y = k|x / solution;            * See note below;
title2 'Test for Homogeneity of Regression (parallel lines)';
title3 'This output is good ONLY for this test';
title4 'If lines are parallel then use next output for ANCOVA tests';
run;
proc glm;
 class k;
 model y = k x / solution;
 means k;            /* this provides the ordinary or unadjusted means        */
 lsmeans k / stderr; /* this provides the adjusted means with standard errors */
title2 'Results for the ordinary ANCOVA';
run;
proc reg;
 model y = x;
title2 'This output required to obtain the SS for the One Line Model';
run;
```

*   `k|x` is shorthand for `k x k*x`.

*   The homogeneity of regression sums of squres, SS**HR**, is obtained from the Type III SS for `k*x`. It could also be obtained by fitting separate regressions to each line, adding the resulting residual sums of squares to get SSR**KL**, and then subtracting SSR**KLP** from SSR**KL**.