

BIOMETRICS INFORMATION

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

PAMPHLET NO. # 46

DATE: April 6, 1994

SUBJECT: GLM: Comparing regression lines

General Linear Models (GLM) are often used to test for differences between means as, for example, when ANOVA is used for data analysis. But GLM's have a wider applicability than this. In this pamphlet I will describe how we can use GLM's to compare regression lines. The methods used are the same as Analysis of Covariance (ANCOVA) but the emphasis is different. In ANCOVA, we are primarily interested in the treatment means and the covariate (X) is an extra variable which explains some of the variability in the response variable (Y)¹. When comparing regressions, we are primarily interested in the relationship between Y and X and how it changes with treatments. Testing for differences between slopes is of primary interest and not a test of an assumption. This pamphlet is a continuation of #31: ANCOVA: The Linear Models behind the F-tests.

Suppose that water quality measurements were taken in three nearby tributaries once a month. Two of the watersheds have been treated in some way, while the third has not been and so is a control. The measurements are taken far enough apart in time that any autocorrelation is negligible and the responses can be considered independent². Suppose that we have the following data:

<u>Week</u>	<u>Stream 1</u>	<u>Stream 2</u>	<u>Stream 3</u>
0	139	539	654
2	209	623	523
4	217	578	545
6	142	409	432
8	122	419	437
10	89	313	438
12	154	253	345
14	224	261	230
16	169	189	302
18	129	137	184

These data are shown in the plot on page 4 with the best fit line for each stream.

The first step in comparing regressions is to test whether the slopes are parallel (known as the heterogeneity of regression test). If the overall test does not reject this null hypothesis, then we might assume that they are parallel and move on to test for differences in the intercepts (where

¹ If the covariate has a strong linear relationship to the response then including the covariate provides a more powerful test for the treatment differences.

² The methods described here require this assumption of no autocorrelation. Otherwise methods such as time series should be used.

$(X = 0)$ ³. If the null hypothesis of parallelism is rejected, then we might want to ask which slopes are different from which. This can be done with contrasts just as in ANOVA.

The contrasts are designed in the same way as in ANOVA⁴, but now they will be weighted sums of slopes instead of means. **These tests can be set up by recognizing that the test for heterogeneity of regression is an interaction between the treatment factor and the covariate, and that contrasts can be done on this interaction.** For the example, there are two obvious contrasts of interest: 1) Is there a difference between the two treated streams? and 2) Is there a difference between the two treated streams and the control? The following SAS program would do a basic analysis (the estimate statements are not really necessary).

```

title 'Comparing Regression Lines';
data stream;  infile 'stream.asc';
  input week @;
  do stream = 1 to 3;
    input y @;  output;
  end; run;
proc glm data=stream;
  class stream;
  model y = stream week stream*week / solution;
  contrast 'Streams 2 vs 3'  stream*week  0  1 -1;
  contrast 'Stream 1 vs 2&3' stream*week -2  1  1;
  estimate 'Stream 1: intercept' intercept 1 stream 1 0 0;
  estimate 'Stream 1: slope'      week 1 week*stream 1 0 0;
  estimate 'Stream 2: intercept' intercept 1 stream 0 1 0;
  estimate 'Stream 2: slope'      week 1 week*stream 0 1 0;
  estimate 'Stream 3: intercept' intercept 1 stream 0 0 1;
  estimate 'Stream 3: slope'      week 1 week*stream 0 0 1;
  output out = resid r = resid p=pred;
title2 'Comparing regression slopes with contrasts';
run;
proc univariate data=resid plot normal;  var resid;
title2 'Testing the residuals for normality and reasonable appearance';
run;

```

Some of the output from this program is shown on the next page. The Week*Stream test ($p \leq 0.0001$) indicates strong evidence for different slopes between the three streams. Notice that the test for the contrasts show that streams 2 and 3 have similar slopes ($p = 0.33$) while their average slope is different from that of stream 1 ($p \leq 0.0001$). This matches the pattern that we see in the plot on the last page.

³ This is the usual ANCOVA test and can also be thought of as testing the adjusted means, where the adjusted means are the predicted values for each line at the overall mean of the covariate values. Since the lines are parallel, the differences between the lines is constant regardless of which X-value is used for comparison purposes.

⁴ Contrasts have also been discussed in BI #12, 16, 23, 26 and 45.

Output showing tests of the sources:⁵

Source	DF	Type III SS	Mean Square	F Value	Pr > F
STREAM	2	384069.36459330	192034.68229665	85.11	0.0001
WEEK	1	283701.09191919	283701.09191919	125.74	0.0001
WEEK*STREAM	2	127683.73535354	63841.86767677	28.30	0.0001

Output from the Contrast statements:

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Streams 2 vs 3	1	2192.74090909	2192.74090909	0.97	0.3340
Stream 1 vs 2&3	1	125490.99444444	125490.99444444	55.62	0.0001

Output from the Estimate statements:

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
Stream 1: intercept	168.45454545	6.03	0.0001	27.9179124
Stream 1: slope	-1.00606061	-0.38	0.7038	2.6147521
Stream 2: intercept	612.50909091	21.94	0.0001	27.9179124
Stream 2: slope	-26.71212121	-10.22	0.0001	2.6147521
Stream 3: intercept	616.60000000	22.09	0.0001	27.9179124
Stream 3: slope	-23.06666667	-8.82	0.0001	2.6147521

Output from the solution option on the model statement:

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	616.6000000 B	22.09	0.0001	27.91791243
STREAM 1	-448.1454545 B	-11.35	0.0001	39.48189040
2	-4.0909091 B	-0.10	0.9183	39.48189040
3	0.0000000 B	.	.	.
WEEK	-23.0666667 B	-8.82	0.0001	2.61475207
WEEK*STREAM 1	22.0606061 B	5.97	0.0001	3.69781784
2	-3.6454545 B	-0.99	0.3340	3.69781784
3	0.0000000 B	.	.	.

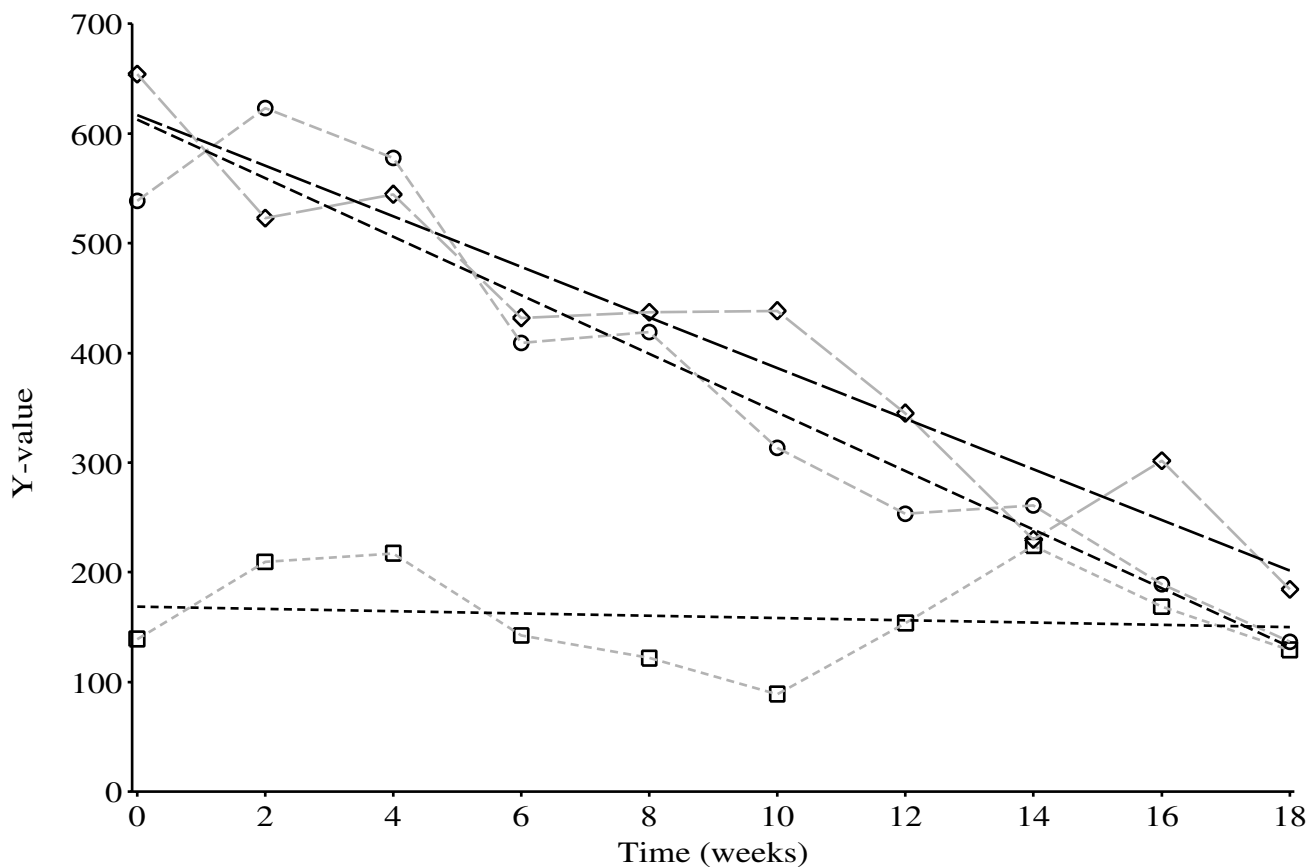
The intercepts and slopes of the three lines (their parameters) can be calculated using the output from the solutions option on the model statement (shown above) as follows.

Stream 1: $(616.600 - 448.145) + (-23.067 + 22.061) * \text{week} = 168 - 1.0 * \text{week}$
 Stream 2: $(616.600 - 4.091) + (-23.067 - 3.645) * \text{week} = 613 - 26.7 * \text{week}$
 Stream 3: $(616.600 - 0.000) + (-23.067 + 0.000) * \text{week} = 617 - 23.1 * \text{week}$

⁵ Note that only the week*stream test is meaningful. If homogeneity of regression is reasonable then the normal ANCOVA tests must be done using a model **without** the week*stream source.

The `Estimate` statement can also be used to calculate the parameters of the six lines, since it calculates the value of a specified weighted sum of the parameters. These parameters can be determined from the `solutions` output above. For example, the intercept for stream 1 is the sum of the intercept and the first stream parameter in the `solutions` table. Similarly, the slope for stream 2 is the parameter for week plus the second week*stream parameter (notice that in these cases, all the weights are either zero or one). The output matches the hand calculations above and provides additional information. For instance, the slope for stream 1 can be tested. Since $p = 0.70$, the null hypothesis that stream 1 has a slope of zero cannot be rejected.

The data with the accompanying best fit lines are shown below:



CONTACT: Wendy Bergerud
387-5676

NEW PROBLEM

Check that the individual equations obtained from the `PROC GLM` are the same as would be obtained from simple regressions on the data for each stream.
