Model Based Statistics in Biology.
Part IV. The General Linear Model. Multiple Explanatory Variables.
Chapter 14.3 ANCOVA - Model Revision

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11)
ReCap Multiple Regression (Ch 12)
ReCap Multiple Categorical Variables (Ch 13)
14.1 Comparing Regression Lines
14.2 Statistical Control
14.3 Model Revision
14.4 More than two explanatory variables (to be written)

SRBx14_9.xls
Ch14.xls

on chalk board

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning is based on models, including statistical analysis based on models.
ReCap Part II (Chapters 5,6,7)
Hypothesis testing uses the logic of the null hypothesis to declare a decision.
Estimation is concerned with the specific value of an unknown population parameter.
ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.
ReCap (Ch 12) GLM with more than one regression variable (multiple regression)
ReCap (Ch 13) GLM with more than one categorical variable (ANOVA).
ReCap (Ch 14) ANCOVA with GLM
Comparing regression lines.
Statistical control

Today: ANCOVA - Model revision, based on analysis of residuals.

Wrap-up.

Introduction.
Revision of model is a logical outgrowth of the model based approach to hypothesis testing with the GLM. If our model includes a regression variable (ANCOVA, multiple and simple linear regression) then we need to make sure that fitting a straight line represents the relation of the response to the explanatory (regression) variable. If there are bowls or arches, then our straight line model does not represent the data. The model is inappropriate and hence out conclusions suspect.

Model revision is an example of quantitative reasoning about biological data, rather than learning "the right statistical method."

Data from: Yamauchi, A. H. Kimizuka. 1971. Study of bio-ionic potentials. *Journal of Theoretical Biology* 30: 285-295.

Purpose of study was to estimate the rate of increase in membrane potential with increase in the logarithm of activity ratio. It is interest to investigate whether the rate of increase depends on ionic composition. Sokal and Rohlf (1995) use the data to illustrate ANCOVA for statistical control: comparison among cation systems, controlled for a regression variable (Box 14.9 on page 504)

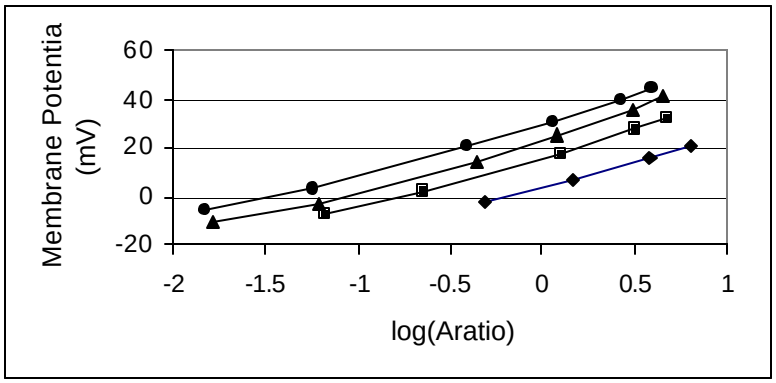
What is the rate of change in membrane potential with change in cation concentration, as measured by the activity ratio ?
Does this rate depend on the cation system?

mV	Aratio	Gr
-2.4	-0.31	1
6.3	0.17	1
15.8	0.58	1
20.5	0.81	1
-7	-1.18	2
2.1	-0.65	2
17.8	0.1	2
27.3	0.5	2
32	0.67	2
-10.8	-1.79	3
-2.8	-1.21	3
14.2	-0.35	3
25.5	0.08	3
35.7	0.49	3
41.2	0.65	3
-5.4	-1.83	4
3	-1.25	4
20.7	-0.41	4
30.5	0.05	4
39.9	0.43	4
45	0.59	4

1. Construct model

Verbal model.
Membrane potential increases with activity ratio. The rate of increase may depend on the cation system

Graphical model.



Response variable is membrane potential V = millivolts

Explanatory variable. Group = cation system
The four cation systems were strontium-sodium, calcium-potassium, calcium-sodium, and calcium-lithium, for which the symbols are
Gr = Sr-Na Ca-K Ca-Na Ca-Li

The other explanatory variable, which needs to be taken into account, is:
logarithm of the activity ratio logAr (no units).

1. Construct model

Formal model

Sketch a graph above each term

Write model $V_{mem} = \beta_o + \beta_{logAr} logAr + \beta_{Gr} Gr + \beta_{logAr \times Gr} logAr Gr + \varepsilon$

The parameter β_{logAr} stands for the rate of change in membrane potential with respect to the rate of change in the logarithm of activity ratio.

The parameter β_{Gr} stands for a set of means: one for each cation group.

The parameter $\beta_{logAr \times Gr}$ stands for degree to which slope in each group varies from the overall slope β_{logAr}

2. Execute analysis.

Place data in model format:

Column labelled V, with response variable membrane potential
Column labelled Gr, with nominal scale explanatory variable, one category for each cation group.
Column labelled A, with ratio scale explanatory variable, log activity ratio.

Code the model statement in statistical package according to the GLM

$$V = \beta_o + \beta_A A + \beta_{Gr} Gr + \beta_{A \times Gr} A Gr + \varepsilon$$

```
MTB > glm 'V' =          'A'   'Gr'    'A'*'Gr';
SUBC> covariate  'A';
SUBC> fits c4;
SUBC> residuals c5.
```

Fits and residuals from:

model statement output of fitted values and residuals (as above)
or parameters reported by GLM routine
or direct calculation of parameters

Here are the parameter estimates.

The overall mean is $\bar{y} = 349 / 21 = 16.62$ mV $= \beta_o$

The mean for each cation system is expressed as a deviation from β_o

$\beta_{Gr} =$

− 6.57 mV	on board
− 2.18 mV	beneath β_{Gr}
+0.55 mV	in model
+5.66 mV	

The slope parameter for all cation systems together is $\beta_A = 20.999$ mV

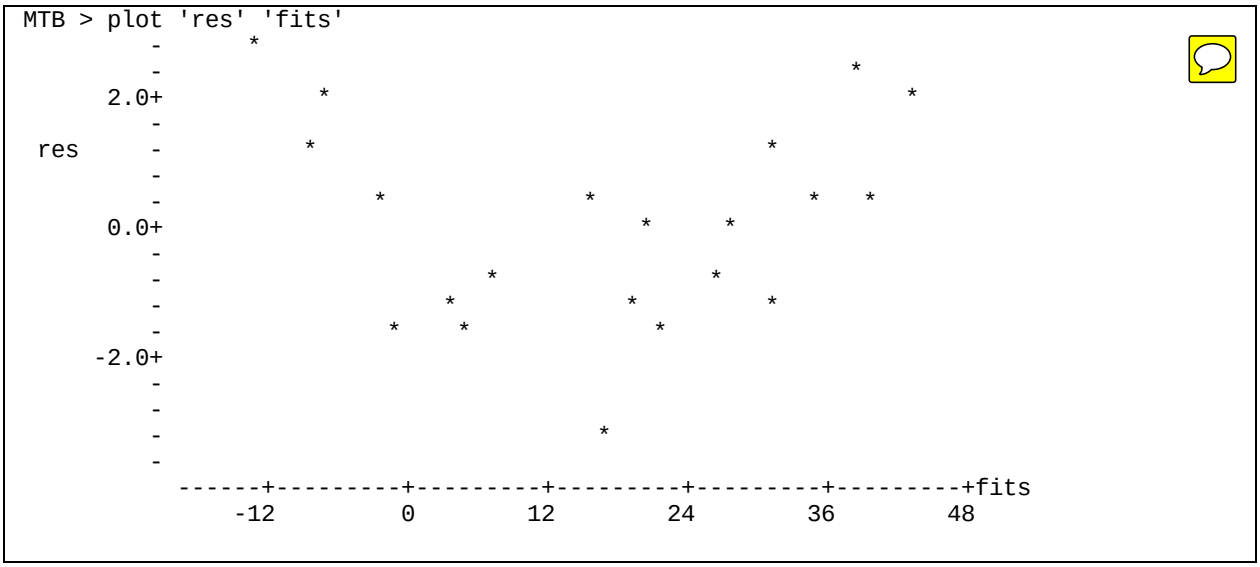
2. Execute analysis.

The deviations from this slope are

$$\$_{A \times Gr} = \begin{matrix} -0.333 \text{ mV} \\ +0.070 \text{ mV} \\ +0.395 \text{ mV} \\ +0.132 \text{ mV} \end{matrix}$$ on board beneath $\$_{A * Gr}$ in model

The GLM routine computes fitted and residual values.

3. Evaluate the model Plot residuals versus fitted values.



a. Straight line assumption

Is a straight line model (for effects of activity ratio) appropriate ?

No. A very clear bowl from left to right

Back to step 1.

It is of interest to note that if we were to continue using this model rather than revising it, we would conclude that the slopes are homogeneous, *i.e.* relation of *V* to logAr does not depend on group.

```
Analysis of Variance for mempt
  Source         DF     Seq SS     Adj SS     Adj MS       F      P
  A               1    4197.01    3192.09    3192.09  876.71  0.000
  Gr              3    1768.58    1413.83     471.28  129.44  0.000
  Gr*A            3       0.80       0.80       0.27    0.07  0.973
  Error          13      47.33      47.33       3.64
  Total          20    6013.72
```

This conclusion would, however, be based on a model that deviates systematically from the data.

Model Selection

The following models were then tried because logarithmic scaling of variables will often straighten out simple bowl or arch patterns.

$V = \log(A) + Gr + Gr*\log(A)$

$\log(V) = \log(A) + Gr + Gr*\log(A)$

$\log(V) = A + Gr + A*Gr$

The variable ($A = \log(Ar)$ is already on a logarithmic scale so the first two models are double logarithmic scaling, and hence biologically uninterpretable, as follows.

$V = \log(\log(Ar)) + Gr + Gr*\log(\log(Ar))$

$\log(V) = \log(\log(Ar)) + Gr + Gr*\log(\log((Ar))$

The third model is a power law (V scales as Ar^{β}) because A was already on log scale.

$\log(V) = \beta_o + (\beta_A + \beta_{A*Grp}\, Grp)\log(A) + \beta_{gGr}\, Gr$

$V = 10^{\beta o}\, A^{(\beta A + \beta Gr*Gr)}\, 10^{\beta Gr*Gr}$

All three models resulted in bowl shaped residual plots and so were discarded.

The following models were then tried.

$1/V = 1/A + Gr + Gr*(1/A)$

$V^2 = A + Gr + Gr*A$

Both models resulted in bowl shaped residual plots and so were discarded.

To develop an acceptable model, the residual plot was examined more closely by comparing it to the plot of *V* against activity ratio. Taken together, these two graphs suggest that the relation of *V* to activity ratio is linear at activity ratios above 0.7, with a different relation at smaller levels of A = log(*aR*). A new variable, called level (factor with two levels), was introduced to control for this.

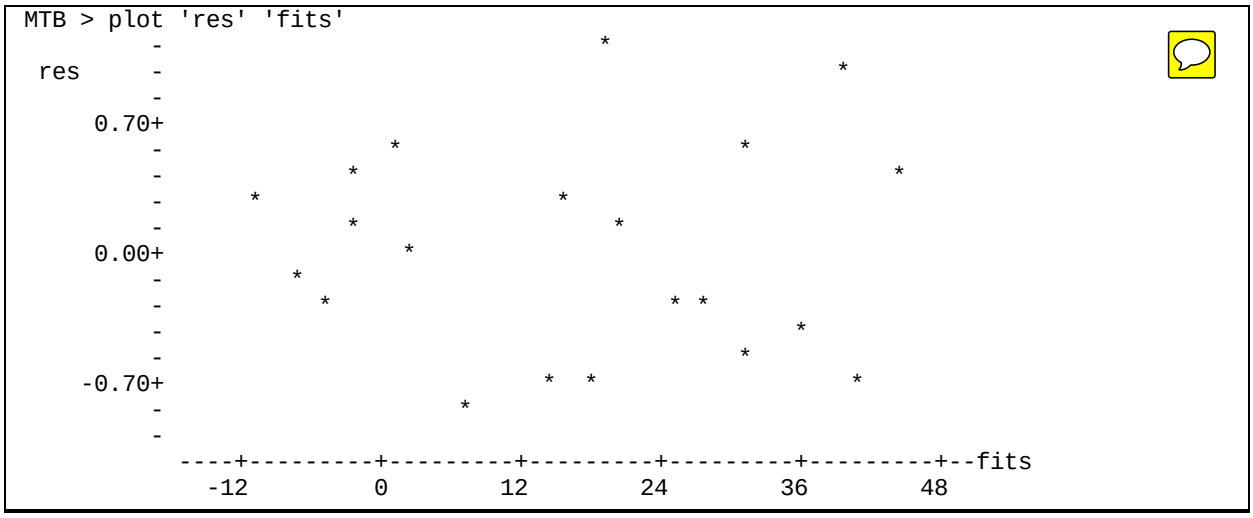
1. Construct Model

$V = \beta_o + \beta_A A + \beta_{Gr} Gr + \beta_{Lvl} Lvl + \beta_{A*Lvl} A Lvl + \beta_{A*Gr} A Gr + \varepsilon$

2. Execute Analysis

```
MTB > glm 'V' =   'A'   'Gr'     'Lvl'     'A'*'Lvl'   'A'*'Gr';
SUBC> covariate  'A'   'A2';
SUBC> fits c4;
SUBC> residuals c5.
```

3. Evaluate Model.



Straight line assumption for regression variable log(*aR*) now acceptable.

7. ANOVA Table.

Analysis of Variance for V

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Lvl	1	2956.08	14.08	14.08	24.25	0.000
A	1	1255.08	360.51	360.51	620.81	0.000
Gr	3	1760.80	1254.03	418.01	719.83	0.000
Gr*A	3	3.11	15.59	5.20	8.95	0.003
Lvl*A	1	32.27	32.27	32.27	55.56	0.000
Error	11	6.39	6.39	0.58		
Total	20	6013.72				

Note that the p-values are several orders of magnitude less than 5% and so we can forego evaluation of the assumptions for these p-values because a randomized p-value will not change our decision. We skip to step 9.

9. Declare decision.

The ANOVA table showed that showed that both interaction terms were significant. Slopes were heterogeneous across classes (*A*Gr* was significant).
This conclusion differs from that arising from the initial analysis (based on an inappropriate model), which would have been that rate of increase in *V* with increase in A is uniform across groups (F = 0.07, p = 0.973).

10. Evaluate parameters.

Because the interaction is significant we cannot move on to testing the group effects. Further, we must estimate a slope for each group within each level, a total of 8 different slopes. We only have 21 observations, so estimating 8 slopes will be impractical. We have arrived at a statistically acceptable model, but cannot use it to undertake analysis of the parameters of biological interest.

1. Construct Model

Next, a quadratic expression was tried.

$V = Gr + A + A^2 + Gr*A + Gr*A^2$

where $A = \log(Ar)$

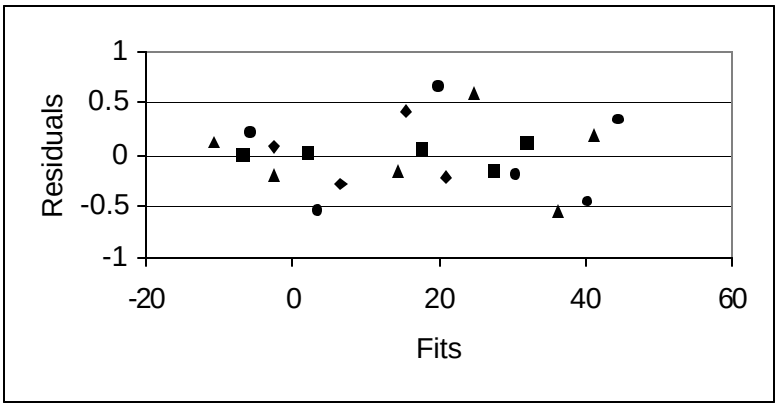
$V = \beta_o + \beta_A A + \beta_{A2} A^2 + \beta_{Gr} Gr + \beta_{A:Gr} A Gr + \beta_{A2:Gr} A^2 Gr + \varepsilon$

2. Execute Analysis

```
MTB > glm 'V' =   'A'   'A2'     'Gr'     'A'*'Gr'   'A2'*'Gr';
SUBC> covariate  'A'   'A2';
SUBC> fits c4;
SUBC> residuals c5.
```

3. Evaluate Model

a. Straight line assumption
There is no prominent bowl or arch in this graph, although there some indication that the relation of V to log(Ar) at low values (eight smallest fitted values) differs from that at higher values of log(Ar).

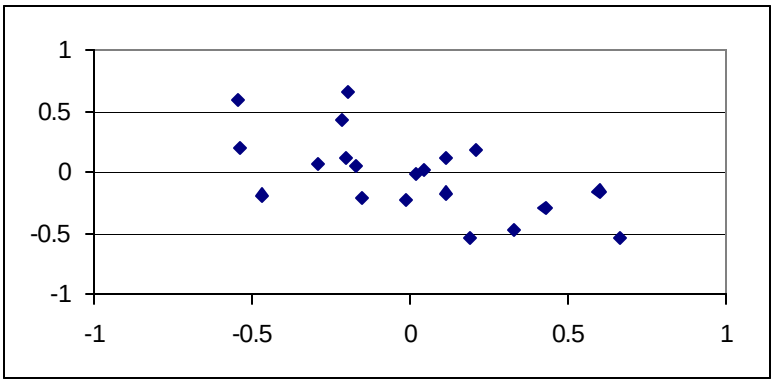


b. Homogeneous error assumption (used in estimating parameters) acceptable.
Residuals do not change in any systematic way with fitted values (no cones).

c. If n small, evaluate assumptions for p-values from chisquare (t, F) distributions.
n = 21

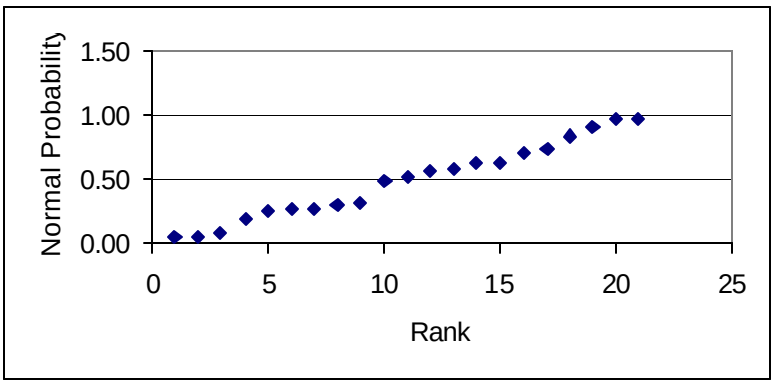
Homogeneous? Yes
Sum(res) = 0? Yes
Independent?

Each residual plotted against its neighbor, in order of low to high log(Ar) values within each bath group.



Clear pattern of negative dependence across all data and within each group.
Residuals tend to pattern of + ! + ! + within each group.

Normal? Acceptably normal



4. State population and whether sample is representative.

Membrane potential of all nerves in all electrolytic baths ? Not likely.
Membrane potential of all nerves, if placed in these four baths ? Probably not.
Membrane potential of nerves from all squid in these four baths ? Possibly.
Membrane potential of all nerves from one squid in these four baths ? Maybe.
All possible measurements of membrane potential of one nerve from one squid in these four baths ? Yes.

5. Decide on mode of inference. Is hypothesis testing appropriate?

Yes. We wish to test whether the slopes differ among cation baths.

6. State H_A H_o pairs, test statistic, distribution, tolerance for Type I error.

Here are the HA / Ho pairs, listed in the order in which they will be tested. Hypotheses concerning heterogeneity of slopes are examined first.

Interaction term *A*Gr*. Are the slopes different among the groups ?
Does change in membrane potential *V* with change in activity ratio $A = \log(Ar)$ depend on electrolytic bath group *Gr*?

H_A: $\beta aR[Sr\text{-}Na] \neq \beta aR[Ca\text{-}K] \neq \beta aR[Ca\text{-}Na] \neq \beta aR[Ca\text{-}Li]$
The slopes differ
H_o: $\beta aR[Sr\text{-}Na] = \beta aR[Ca\text{-}K] = \beta aR[Ca\text{-}Na] = \beta aR[Ca\text{-}Li]$
The slopes do not differ

This pair is equivalent to the following hypotheses concerning variances.

H_A: $\mathrm{Var}(\beta_{A*Gr} * A*Gr) > 0$ Slopes differ hence variance present.
H_o: $\mathrm{Var}(\beta_{A*Gr} * A*Gr) = 0$ Slopes same so no variance among slopes.

Interaction term $A^2 * Gr$. Do the curvature of the slope differ among cation groups?
Does change in membrane potential *V* with change in activity ratio *aR* depend on whether activity ratio is high or low (*Lvl* greater or less than 0.7).

H_A: $\mathrm{Var}(\beta_{A2*Gr} * A^2 * Gr) > 0$
H_o: $\mathrm{Var}(\beta_{A2*Gr} * A^2 * Gr) = 0$

Categorical term (*Gr* = type of electrolytic bath). This is a fixed factor.
Goal of the analysis was to compare rate of change in potential with change in activity ratio across groups. If rate of change is the same (interaction terms not significant) then we could examine this factor.

We expect membrane potential to differ among electrolytic baths

H_A: $\beta Gr[Sr\text{-}Na] \neq \beta Gr[Ca\text{-}K] \neq \beta Gr[Ca\text{-}Na] \neq \beta Gr[Ca\text{-}Li]$
H_o: $\beta Gr[Sr\text{-}Na] = \beta Gr[Ca\text{-}K] = \beta Gr[Ca\text{-}Na] = \beta Gr[Ca\text{-}Li]$

This pair is equivalent to the following hypothesis concerning variances

H_A: $\mathrm{Var}(\beta Gr) > 0$ Means differ hence there is variance in membrane potential due to bath type.
H_o: $\mathrm{Var}(\beta Gr) = 0$ No variance due to bath type.

Regression term $A = \log(Ar)$. (relation of membrane potential to activity ratio)
If the rates are the same across cation groups (no interaction) then it is of interest to examine this term.

6. State H_A H_o pairs, test statistic, distribution, tolerance for Type I error.

State test statistic	F-ratio
Distribution of test statistic	F-distribution
Tolerance for Type I error	5% (conventional level)

7. ANOVA

Source	DF	Seq SS	Adj SS	Adj MS	F	P
A=log(Ar)	1	4197.01	1122.52	1122.52	4447.74	0
Group	3	1768.58	801.33	267.11	1058.36	0
A^2	1	26.41	8.76	8.76	34.7	0
Group*A	3	18.44	9.84	3.28	13	0.001
Group*A^2	3	1.01	1.01	0.34	1.33	0.325
Error	9	2.27	2.27	0.25		
Total	20	6013.72				

8. When assumptions not met, decide whether to recompute p-values.

Assumptions not strongly violated and p-values far from 5%, hence recomputing p-values will not change decision.

9. Declare decision, with evidence.

As in any ANCOVA, we start with the interaction term. If the slopes are heterogeneous there is little point in trying to interpret the variable of interest, the effect of cation system on membrane potential.

reject H_o: $Var(\$_{A*Gr}) = 0$ $0.001 = p < \alpha = 0.05$ $F_{3,9} = 13$
accept H_A: $Var(\$_{A*Gr}) = 0$ the slopes differ across baths.
accept H_o: $Var(\$_{A2*Gr}) = 0$ $0.325 = p > \alpha = 0.05$ $F_{3,9} = 1.33$
the shape of the quadratic is unchanged across groups

For a complex analysis of this sort, it is best to report the entire table, showing Sources of variance, df, SS, MS, F and p values.

Before concluding that the shape of the quadratic expression is constant across cations groups, we examine the individual contrasts.

Term	Coef	SE Coef	T	P
A^2*Group				
1	-0.21	1.44	-0.14	0.89
2	-0.11	0.72	-0.15	0.88
3	0.67	0.58	1.15	0.28

There is no evidence of change in the shape of the quadratic term across groups in the components of the overall interaction term.

No conclusion will be drawn about all nerves when placed in these 4 cation systems. Inference was made to a more restricted population, all possible measurements on small number of one type of nerve. The conclusion about this statistical population can then be used to form expectations about membrane potential in other situations, keeping in mind that only a limited number of nerves were measured.

10. Analysis of parameters of biological interest.

Our conclusion is that there is a difference in the linear component of change in membrane potential with activity ratio, but no change in the shape of relation, as expressed by the quadratic term. Here are the coefficients estimated by the general linear model.

Term	Coef	SE Coef	T	P
Constant	17.75	0.17	103.64	0.000
A=log(Ar)	22.87	0.34	66.69	0.000
Group				
1	-14.53	0.33	-44.40	0.000
2	-2.28	0.32	-7.05	0.000
3	5.08	0.27	19.02	0.000
A^2	3.07	0.52	5.89	0.000
A*Group				
1	-3.59	0.84	-4.26	0.002
2	-0.35	0.48	-0.74	0.481
3	2.67	0.48	5.61	0.000

The coefficient for the quadratic term is 3.0668 mV. To estimate the expression for each cation group we write the GLM for each group, then rearrange it as follows.

$$V = \$_o + \$_A A + \$_{A2} A^2 + ,$$
$$V - \$_{A2} A^2 = \$_o + \$_A A + ,$$
$$V - 3.0668 A^2 = \$_o + \$_A A + ,$$

We compute the value of the expression on the left for each observation in a group, then regress this new variable against $A = \log(Ar)$
Here are the results.

```
The regression equation is
mV1-3.07A^2 = 3.20 + 19.2 A

Predictor        Coef     SE Coef          T        P
Constant       3.1998      0.2492      12.84    0.006
Ar1           19.1788      0.4716      40.67    0.001

The regression equation is
mV2-3.07A^2 = 15.4 + 22.6 A

Predictor        Coef     SE Coef          T        P
Constant      15.4203      0.0590     261.38    0.000
AR2           22.5742      0.0830     271.85    0.000

The regression equation is
mV3-3.07A^2 = 23.2 + 24.8 A

Predictor        Coef     SE Coef          T        P
Constant      23.1789      0.2794      82.96    0.000
AR3           24.7968      0.2929      84.66    0.000

The regression equation is
mV4-3.07A^2 = 29.3 + 24.6 A

Predictor        Coef     SE Coef          T        P
Constant      29.3233      0.2621     111.89    0.000
AR4           24.5701      0.2709      90.69    0.000
```

Each equation is then rewritten to standard form.

```
Group 1     mV1 = 3.20 + 19.2 A + 3.07A^2          Sr-Na
Group 2     mV2 = 15.4 + 22.6 A + 3.07A^2          Ca-K
Group 3     mV3 = 23.2 + 24.8 A + 3.07A^2          Ca-Na
Group 4     mV4 = 29.3 + 24.6 A + 3.07A^2          Ca-Li
```

The rate of change in membrane potential with change in the log of the activity ratio is greater for the calcium baths than for the strontium bath. The anion (K, Na, or Li) has no significant effect.

Summary
This example of ANCOVA illustrated:

(1) Revision of model to obtain appropriate structural model with acceptable residuals (no bowls or arches).

(2) Dependence of the statistical conclusion on the model structure. In this case the rate of change in membrane potential appeared to be uniform across the four cation groups, when a linear model was used. However, the analysis of residuals showed that a linear model was not consistent with the data. When a model was adopted that was consistent with the data (no bowl or arch in residual plot), there was a heterogeneity in rate of change across the four groups.

This example shows how model revision can improve the analysis of data.
This is an example of reasoning about biological relationships, using a model. In this case we learned something about this data, by revising the model. We learned the relation between membrane potential in nerves depends on activity ratio, but that this relationship changes in going from low to high activity ratios. We could use this information in further work with nerves. We would want to keep this change in mind in designing further experiments, or in working out how nerves function.