Model Based Statistics in Biology.
Part IV. The General Linear Model. Multiple Explanatory Variables.
Chapter 13.2 Fixed Effects ANOVA (Interactive effects)

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)
13.1 Fixed Effects ANOVA (no interactive effects)
13.2 Fixed Effects ANOVA (interactive effects)
13.3 Fixed*Random Effects (Paired t-test)
13.4 Fixed*Random Effects (Randomized Block)
13.5 Fixed*Random Effects (Repeated Measures)
13.6 Nested Random Effects (Hierarchical ANOVA)
13.7 Random within Fixed (Hierarchical ANOVA)
13.8 More Than Two Factors (to be written)

SC16_6_1.xls
Ch13.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).
New concept, the interaction term.

Today: Two-way ANOVA.
One response variable Y as a function of two explanatory variables X_1 X_2.
Both explanatory variables are categorical, on a nominal scale.
Significant interaction term.

Wrap-up. General Linear Model with two classification variables,
i.e. two explanatory variables on a nominal scale.

We used judgement to decide whether interactive effects were important.
When interactive effects are important, we analyze one factor within each level of another factor because the effects of one factor differ across the other factor.

Example. GLM, applied to 2-way ANOVA factorial design, gains in weight (g) of male rats under six diets. Data from Table 16.6.1 (p304) in Snedecor and Cochran 1989.

Does weight gain depend on protein source and protein level ?

1. Construct model

Data are: weight gains in rats (grams) fed 6 diets classified by source of protein (cereal, beef, or pork) and by level of protein (low or high).

Response variable

Weight gain: ΔM in grams (ratio scale)

Explanatory variables are protein source and protein level.
Source: X_S = source of protein in three categories: cereal, beef, or pork
Level. X_L = protein level in two categories, high or low.

Both explanatory variables are on nominal scales

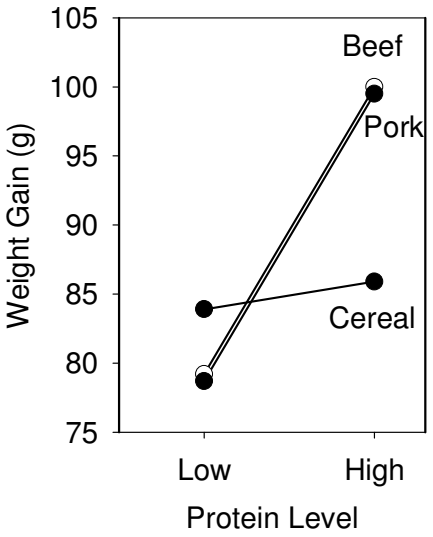
Verbal model.

Weight gain depends on protein source and level.

Graphical model.

Y-axis = Weight gain
X-axis = Low or high levels of protein
Connect two means of each of three sources

Graph suggests that growth depends on level for beef in pork more than for cereal.



Formal Model

Write GLM:	ΔM	=	β_o	+	$\beta_S \cdot X_S$	+	$\beta_L \cdot X_L$	+	$\beta_{S\,x\,L} \cdot X_S \cdot X_L$	+	residual
S&R95	V_{ijk}	=	μ	+	α_i	+	β_j	+	$(\alpha \cdot \beta)_{ij}$	+	ϵ_{ijk}

The model has been written in two forms. One is typical notation for the GLM, the other is from the tradition of experimental design (e.g. Snedecor and Cochran 1989, Sokal and Rohlf 1995) in which fixed factors are represented by greek letters, random factors by roman letters. In this example both factors are fixed by experimental design.

There are three explanatory terms, one for protein source, one for protein level, and one for interactive effects--the dependence of level effects on source. Graphical interpretation is that the relation of weight gain to protein level depends on source of protein.

2. Execute analysis.

Place data in model format:

Column labelled M, with response variable weight gain.

Column labelled XS with explanatory variable, $X_S = -1, 0,$ or 1

-1 = cereal, 0=beef, 1 = pork

Column labelled XL with explanatory variable, $X_L = 0$ (low) or 1 (high)

These are labels (categories), not numbers on ratio scale.

Code model statement in statistical package according to the GLM

$$\Delta M - \beta_o = \beta_S \cdot X_s + \beta_L \cdot X_L + \beta_{S\,x\,L} \cdot X_S \cdot X_L + \epsilon$$

```
MTB> ANOVA 'M' = 'XS'  'XL'  'XS'*'XL'
MTB> GLM   'M' = 'XS'  'XL'  'XS'*'XL'
SUBC> fits c4;
SUBC> res c5.
```

The grand mean. $\hat{\beta}_0 = 87.9$ g

The fitted values are the means for each of the 6 cells.

	Low	High
Cereal	83.9	85.9
Beef	79.2	100
Pork	78.7	99.5

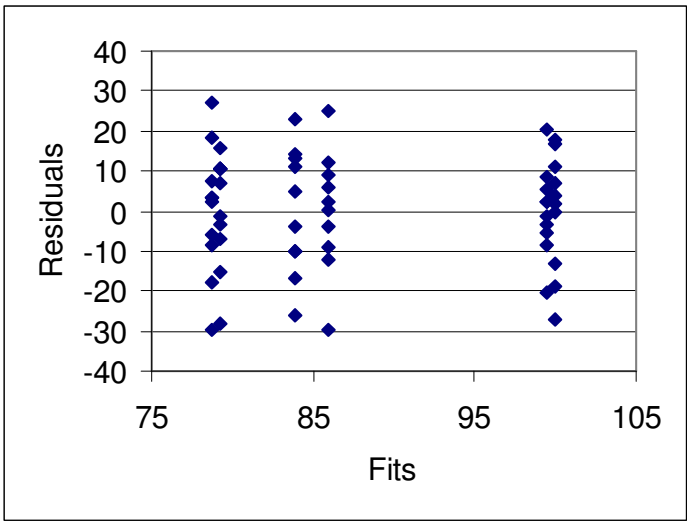
These values are used to compute a residual value, one for each of 60 observations.

Residuals can also be calculated from the 6 parameter estimates produced the GLM command. Here are the GLM parameters produced by the SPlus statistical package.

	Value	Std.Error	t value	Pr(>|t|)
(Intercept)	87.867	1.891	46.4654	0.000
Level	7.2667	1.891	3.8427	0.000
Source1	2.35	2.316	1.0147	0.315
Source2	0.6167	1.3371	0.4612	0.647
LevelSource1	4.7	2.316	2.0294	0.047
LevelSource2	1.5667	1.3371	1.1716	0.247

Plot residual versus fitted values.

There are six stacks of values, one stack for each of six cell means (fitted values).



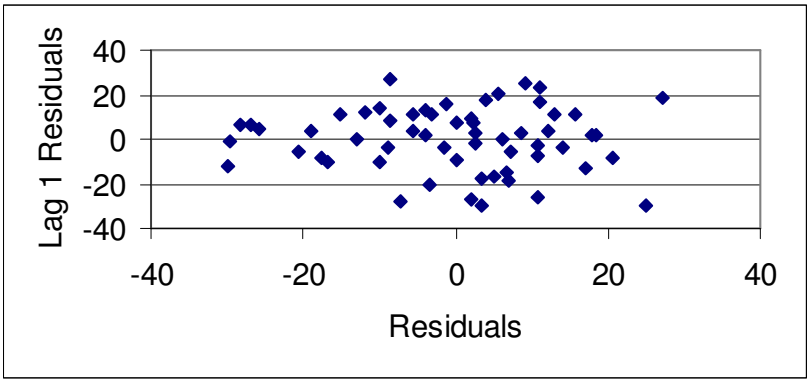
3. Evaluate the model.

a. No line fitted in model, so skip this evaluation.

b. No systematic change in residuals with increase in fitted values (*i.e.* no cones) so residual homogeneous, no need to revise error structure of model.
The six stacks in the plot should be similar in vertical dispersion.

c. If n small, evaluate assumptions for computing p-values.
n = 60 so only large violations will distort p-values or confidence limits.
Homogeneous? Yes
Sum(res) = 0? Yes
Independent? Yes

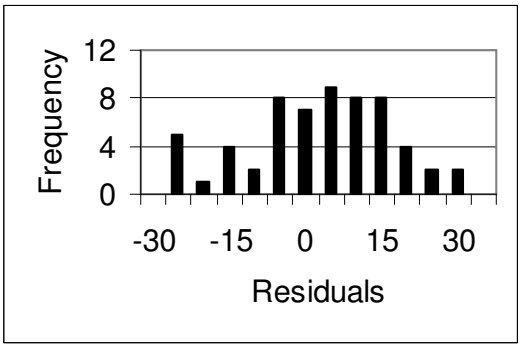
Graph shows no evidence of upward or downward trend.



c. If n small, evaluate distributional assumptions.
Because n is substantially greater than 30, any deviations from normal errors will have little effect on computation of p-values. As a matter of interest, we examine the histogram of the residuals, to see if it looks normal.

Normal ? The residuals are normal except for skewness to the left due to a relatively large number of residuals in the smallest size class, at a value of −25.

In this analysis, n=60, the residuals are homogeneous, and the deviations from normality are minor, so we judge that the p-value from the cumulative distribution function will be accurate.



4. State population and whether sample is representative.

When we draw conclusions from this sample, what is the population we are prepared to discuss? The protein levels and sources were chosen by experimental design. We will view these as fixed factors and hence infer only to these three sources and two levels of protein. We assume that rats were assigned randomly to treatments, so that the results are representative of any other experiment with the same design.

We can safely infer to a population of all possible measurements of weight gain in male rates for this experimental design. We will be considering a hypothetical population of possible measurements, not an enumerable biological population.

5. Decide on mode of inference. Is hypothesis testing appropriate?
Hypothesis testing is appropriate (step 6) because research question is binary: do protein source and level affect weight gain?

6. State H_A / H_o pairs, test statistic, distribution, tolerance for Type I error.
Analysis will focus first on the interaction term $\beta_{S\,x\,L} \cdot X_S \cdot X_L$
If the factors have interactive effects on the response variable, then the observed difference in weight gain due to one factor (protein source) will depend on the other factor (protein level). If there are significant interactive effects then the weight gains among the three sources cannot be interpreted unless we know the protein level.

The symbol $\beta_{S\,x\,L}$ stands for two parameters, which quantify the degree to which the effects of protein source on weight gain depends on level. One parameter measures the change in change in weight from low to high level in beef, relative to cereal. The second parameter measures the change in weight from low to high in pork, relative to beef.

Hypotheses for the interaction term.
The research hypothesis H_A is that $\beta_{S\,x\,L} = 0$
H_A: $\beta_{S\,x\,L} \neq 0$
H_o: $\beta_{S\,x\,L} = 0$

Are there more specific hypotheses about the interaction term?
Yes. The experimenters expected the weight gain to be greater from high to low for animal protein sources (beef and pork) than for cereal.

If the parameter values are not zero, then there will be variance.
The H_A / H_o pairs equivalent to those listed above are:
H_A: $\mathrm{Var}(\beta_{S\,x\,L}) > 0$ or equivalently H_A: $\mathrm{Var}(\beta_{S\,x\,L} \cdot X_S \cdot X_L) > 0$
H_o: $\mathrm{Var}(\beta_{S\,x\,L}) = 0$ or equivalently H_o: $\mathrm{Var}(\beta_{S\,x\,L} \cdot X_S \cdot X_L) = 0$

If the interaction term is not significant, then research hypotheses concerning each of the other terms in the model become of interest because we can interpret the effects of one factor (*e.g.* protein source) regardless of the effects of the other factor (*e.g.* protein level).

Hypotheses for the protein level term.
Fixed effects term so the contrast in means will be of interest.
H_A: $\mathrm{PopMean}(\Delta M_{L=low}) < \mathrm{PopMean}(\Delta M_{L=high})$ The population means differ
H_o: $\mathrm{PopMean}(\Delta M_{L=low}) = \mathrm{PopMean}(\Delta M_{L=high})$ The population means do not differ
These hypotheses are equivalent to following H_A / H_o for parameters.
H_A: $\mathrm{Var}(\beta_L) > 0$ There is variance present, due to level
H_o: $\mathrm{Var}(\beta_L) = 0$ The is no variance due to level.
Are there more specific hypotheses about parameters ? Yes. We might expect that weight gain is greater for high level than low levels of protein.

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

Hypotheses for the protein source term. Fixed effects term so the contrast in means will be of interest.

H_A: PopMean($\Delta M_{S=cereal}$) ≠ PopMean($\Delta M_{S=beef}$) ≠ PopMean($\Delta M_{S=pork}$)

H_A: PopMean($\Delta M_{S=cereal}$) ≠ PopMean($\Delta M_{S=beef}$) ≠ PopMean($\Delta M_{S=pork}$)

The population means differ among protein source.

H_o: PopMean($\Delta M_{S=cereal}$) = PopMean($\Delta M_{S=beef}$) = PopMean($\Delta M_{S=pork}$)

The population means do not differ

These hypotheses are equivalent to

H_A: $\beta_S \neq 0$

H_o: $\beta_S = 0$

The H_A H_o pair above is equivalent to the following hypotheses.

H_A: $Var(\beta_S) > 0$ There is variance present, due to protein source.

H_o: $Var(\beta_S) = 0$ There is no variance present, due to protein source.

Additional hypotheses for parameters in the source term ? Yes

H_A: $\Delta M_{cereal} < ((1/2)(\Delta M_{beef} + \Delta M_{pork})$ Growth rates for cereal less than those for animal sources of protein.

H_o: $\Delta M_{cereal} = ((1/2)(\Delta M_{beef} + \Delta M_{pork})$

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

7. ANOVA - Calculate then partition df according to model.

```
Model at top of board on left.
ANOVA table at top of board on right.
```

GLM	ΔM	=	β_o	+	$\beta_S \cdot X_S$	+	$\beta_L \cdot X_L$	+	$\beta_{S*L} \cdot X_S \cdot X_L$	+	ϵ
Source	Total	=			Source	+	Level	+	Source ·Level	+	Resid

Take df from beneath each term and place in table.

Source	df	SS	MS	F	---->	p
Source	2					
Level	1					
Source ·Level	2					
Res	?		? = 59 − 1 − 2 − 2 = 54			
Total	60 − 1					

Compute total degrees of freedom $df_{total} = n - 1 = 60 - 1 = 59$

Partition df_{total} according to model, using rules

two species, hence 2-1 = 1 df $df_s = 2 - 1 = 1$

three salinities, hence 3-1 = 2 df $df_L = 3 - 1 = 2$

$df_{s*L} = df_s \cdot df_L$ $df_{s*L} = 1 \cdot 2 = 2$

$df_{res} = df_{total} - df_s - df_L - df_{s*L}$ $df_{total} = 59 - 1 - 2 - 2 = 54$

7. ANOVA - Calculate then partition variance according to model.

Compute $SS_{tot} = Var(\Delta M) \cdot df_{total}$

By hand

$$SS_{tot} = \Sigma \Delta M^2 - n^{-1}(\Sigma \Delta M)^2 = 479432 - 60^{-1} \cdot 5272^2 = 16198.93$$

By Minitab

```
MTB> let k1=ssq('M'-mean(M))
MTB> print k1
```

In a spreadsheet (Excel) the command is =DEVSQ(A5:A64)
where the data are in rows 5 through 64 of column A.

Partition SS_{tot} using a statistical package such as Minitab

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_{1x2} X_{1x2} + \epsilon$$

```
MTB> ANOVA 'Response' = 'X1'   'X2'   'X1'*'X2'
MTB> ANOVA      c1    =  c2     c3     c2*c3
MTB> GLM        c1    =  c2     c3     c2*c3
MTB> GLM       'M'    = 'XS'   'XL'   'XS'*'XL'
```

Line up the GLM model with the ANOVA and GLM commands

$Y - \beta_o$	=	$\beta_S X_S$	+	$\beta_L X_L$	+	$\beta_{s*L} X_{s*L}$	+	ϵ
16198.93	=	266.5	+	3168.3	+	1178.2	+	11585.7

Most statistical packages have a GLM command.

Bring SS components from beneath model to table
Start with SS_{tot} at bottom,
then add partitioned components SS_S, SS_L, $SS_{S \times L}$, SS_{res} from Handout

Source	df	SS	MS	F	---->	p
Source	2	266.5				
Level	1	3168.3				
Source ·Level	2	1178.2				
Res	54	11585.7				
Total	59	16198.9				

r^2 = explained variance = SS_{model} / SS_{tot} = (16198.9 − 11585.7) / 16198.9 = 28%

Compute MS $MS_S = SS_S / df_S = 133.3$ *etc*

Add to table

Compute F-ratios.
Fixed effects for source and level, so all variance ratios taken relative to MS_{res}

$F = MS_S / MS_{res} = 133.3 / 214.56 = 0.62$ *etc*

Add to table

Calculate Type I error from F-distribution.
$F_{2,54} = 0.62$ p = 0.54 *etc*

Draw picture of computational flow. Add p-values to table

Source	df	SS	MS	F	p
Source	2	266.5333	133.2667	0.62	0.541132
Level	1	3168.267	3168.267	14.77	0.000322
S * L	2	1178.133	589.0667	2.75	0.073188
Error	54	11586.00	214.5556		
Total	59				

8. Decide whether to recompute p-value.

30 < n < 100, residuals homogeneous, only slight deviation from normal distribution of errors, so p-value judged to be accurate. We check that judgement.

	p-value	p random	p-value/prandom
Source	0.541	2688/5000 = 0.538	1.006
Level	0.000322	1/5000 = 0.0002	1.56 (poor estimate)
Source x Level	0.0732	351/5000 = 0.070	1.038

Note that the p-value for Level was poorly estimated because only 5000 randomizations were run. The p-value from the F-distribution was 3 in 1000 (or 15 in 5000), so our estimate can be no better than 1 part in 15. If we had used 10,000 randomizations our estimate the p-value would improve to 1 part in 30.

Our judgement that the p-value from the F-distribution would be acceptable was correct.

9. Declare decision about terms.

Start with interaction term.

$p = 0.073 > \alpha = 5\%$

Cochran and Cox (1989 p305) comment as follows.

In factorial designs 'it often happens that a few comparisons comprising the main effects have substantial interactions, while the majority of the comparisons have negligible interactions. Consequently, the F-test of the AB interaction sum of squares as a whole is not a good guide as to whether interactions can be ignored. It is wise to look over the two-way table of treatment totals or means before concluding that there are no interactions, particularly if the F is larger then 1.'

Here is the table of cell means. To examine the interactive effects, we subtract the we take the difference from low to high, within each level of the other factor, protein source.

	Low	High	mean	Difference
cereal	83.9	85.9	84.9	2.00
beef	79.2	100.0	89.6	20.80
pork	78.7	99.5	89.1	20.80
mean	80.6	95.1	87.87	

The difference from low to high in cereal (2.0 g) is one-tenth the magnitude of the contrast in beef and pork (20.8 g). The interactive effect due to cereal is substantial relative to the other two sources.

9. Declare decision about terms.

Following the advice of Snedecor and Cochran (1989) we use judgement to temper our conclusion about the presence of an interactive effect.
Instead of ignoring the interaction term because the p-value was just short of statistical significance at the 5% level, we judge that interactive effects do need to be considered in this experiment, for which the F-ratio was substantially greater than 1 (F = 2.75) and the one of the interaction contrasts was large compared to the other.

Interactive effects mean that we cannot interpret one main effect independently of the other. Thus, we do nto proceed to the analysis of main effects in the ANOVA table. Instead, we break down the two-way table by comparing means across one factor within in each level of the other factor. We can examine the differences from low to high within each level of protein source, as in the table above. Alternatively, we could examine the differences across sources within each of the two levels.

10. Report and interpret parameters of biological interest.

Interactive effects cannot be ignored in this experiment. Thus we report contrasts across one factor (protein level) in each class of the other factor (protein source).

Report conclusions based on parameters and some measure of uncertainty.

Here is the output from the Minitab package, which shows the coefficients calculated above, from the cell means. The interactive effect due to cereal is significant (p = 0.023).

Term	Coef	SE Coef	T	P
Constant	87.867	1.891	46.47	0.000
Level				
0	-7.267	1.891	-3.84	0.000
Source				
-1	-2.967	2.674	-1.11	0.272
0	1.733	2.674	0.65	0.520
Level*Source				
0 -1	6.267	2.674	2.34	0.023
0 0	-3.133	2.674	-1.17	0.246

The results of this experiment are conveniently summarized as differences of means within each protein source, along with confidence limits on the difference.

				SS		MS=			
	Low	High	Diff	Low	High	Sum(SS)/18	sterr	Lower	Upper
cereal	83.9	85.9	2.00	2220.9	2030.9	236.2	3.437	-5.22	9.22
beef	79.2	100.0	20.80	1735.6	2062.0	211.0	3.248	13.98	27.62
pork	78.7	99.5	20.80	2464.1	1072.5	196.5	3.134	14.22	27.38

The confidence limits for cereal include zero (no difference in growth at low and high levels of protein). The confidence limits for beef and pork exclude zero (no difference in growth).