

Statistical Science.

Part IV. The General Linear Model. Multiple Explanatory Variables.

Chapter 12.1 Multiple Regression. Two Explanatory Variables.

ReCap.	Part I (Chapters 1,2,3,4)
ReCap	Part II (Ch 5, 6, 7)
ReCap	Part III (Ch 9, 10, 11)
12	Multiple Regression. Introduction
12.1	Two Explanatory Variables
12.2	Three Explanatory Variables
13	GLM multiway ANOVA
14	GLM ANCOVA
15	Review - GLM with multiple explanatory variables.

on chalk board

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning based on models.

ReCap Part II (Chapters 5,6,7)

Data Equation

Frequency distributions

Three modes of inference.

ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.

Unifying concepts rather than list of statistical tests.

GLM is more useful and flexible than a collection of special cases.

Today: Introduction to GLM, Multiple Explanatory Variable.
Distinction among Multiple regression, Multiway ANOVA, ANCOVA
Example: Multiple Regression

Wrap-up.

Multiple regression is a special case of the General Linear Model in which there are two or more explanatory variables on a ratio scale.

The regression coefficients estimated by many statistical packages are by default either sequential (Type I) sum of squares and adjusted (Type III) sum of squares.

Type I are adjusted for those coefficients that occur earlier in the list of coefficients.

Type III are adjusted for all other coefficients in the model.

The choice between sequential and adjusted depends on the analysis. In practice, multiple regression estimates are adjusted (Type III) in the absence of any rationale for ordering more than 2 explanatory covariates.

Introduction Analysis of data from Snedecor and Cochran 1980 Table 17.2.1

Do corn plants use organic or inorganic phosphorus in soil? Phosphorus in corn (ppm) was measured in 17 Iowa soils at 20 deg C in relation to inorganic and organic phosphorus in the soil.

1. Construct model

Verbal model. Plant available phosphorus depends on the amount of both organic and inorganic soil phosphorus.

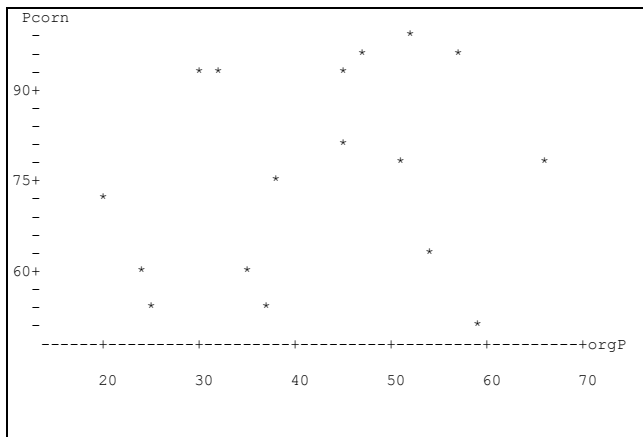
Response variable is phosphorus content of corn. P_{corn} = ppm

Explanatory variables is organic phosphorus in soil. oP = ppm

Explanatory variables is inorganic phosphorus in soil. ioP = ppm

All variables are on a ratio type of scale.

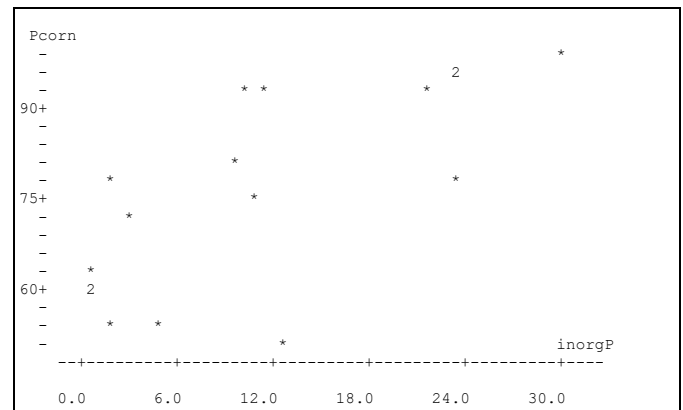
Graphical model.



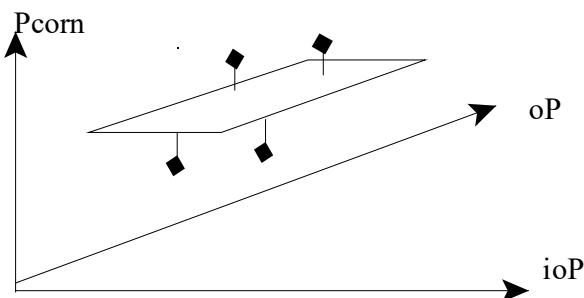
<--- P_{corn} versus oP

Cloud of points.
No clear trend
to describe as a line

P_{corn} vs ioP



A line can be fit through the points - - >
Then fit two perpendicular lines



Formal model. We begin with one explanatory variable ioP

GLM: $P_{corn} = \beta_o + \beta_{ioP} \cdot ioP + \text{res}$

The parameter β_{ioP} stands for rate of change in phosphorus content of corn, with respect to rate of change of inorganic phosphorus. It is represented as a line through the cloud of points in a graph of P_{corn} versus inorganic phosphorus.

$$P_{corn} = \beta_o + \beta_{oP} \cdot oP + \text{res}$$

1. Construct model

Next, a model for the other explanatory variable, oP .

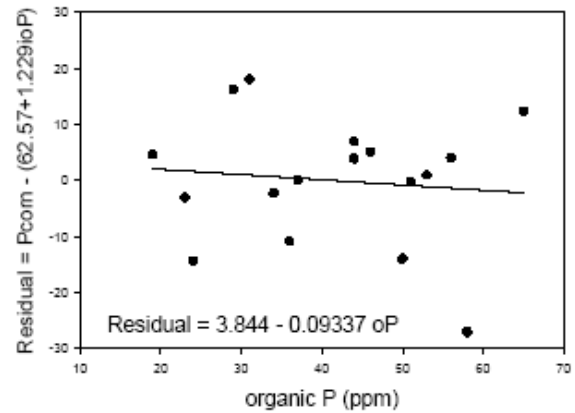
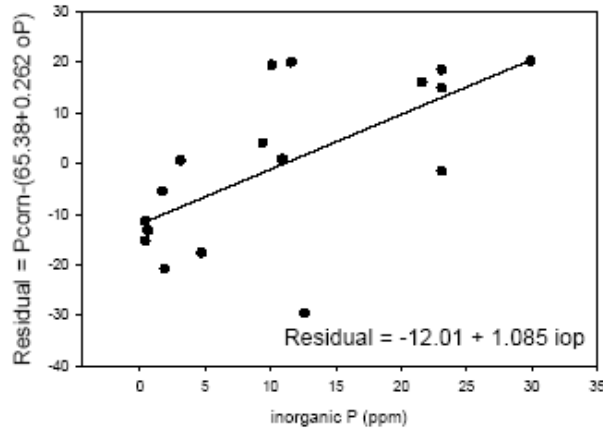
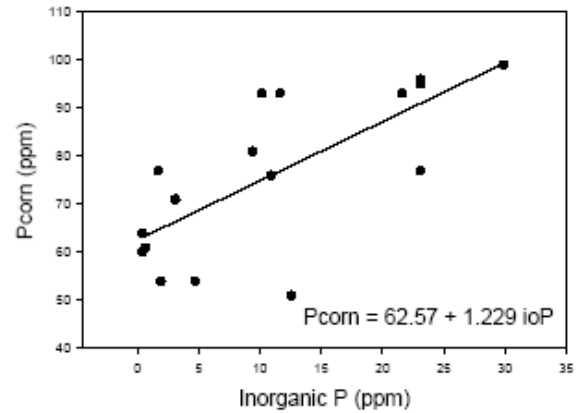
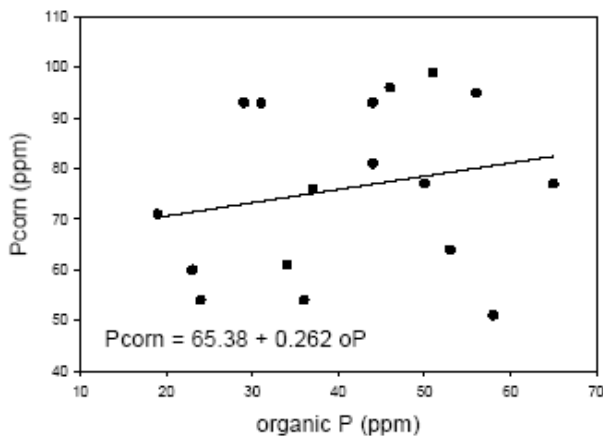
The parameter β_{oP} stands for rate of change in phosphorus content of corn, with respect to rate of change of organic phosphorus.

Finally, a model that includes both explanatory variables

$$P_{corn} = \beta_o + \beta_{oP:ioP} \cdot oP + \beta_{ioP:oP} \cdot ioP + \text{res}$$

The parameter $\beta_{ioP:oP}$ stands for rate of change in phosphorus content of corn relative to rate of change of inorganic phosphorus, adjusted for effects of organic phosphorus. It is read as ‘the rate of change in available phosphorus with change in inorganic phosphorus, controlled for organic phosphorus.’

The parameter $\beta_{oP:ioP}$ stands for rate of change in phosphorus content of corn, relative to rate of change in organic phosphorus, adjusted for effects of inorganic phosphorus.



Together, these two parameters describe a plane through the data points (see Fig 1 above). These parameters are partial derivatives, for those who have had this in calculus. The next figure distinguishes the simple regression coefficients (β_{ioP} β_{oP}) from the partial coefficients ($\beta_{ioP:oP}$ $\beta_{oP:ioP}$)

1. Construct the model (continued)

Partial regression is the same as regression of the residuals on the remaining variable.

Regress *Pcorn* against *ioP*: $Pcorn = 62.6 + 1.23 \cdot ioP$

Take residuals from this model

Regress these against another variable, *oP*

$$Res = 3.84 - 0.09337 \cdot oP$$
$$\beta_{oP:ioP}$$

This estimate is close to the original estimate of $\beta_{oP:ioP} = -0.111$

Because we have two explanatory variables we can investigate their interactive effects on the response variable. Does the effect of one variable on the response variable depend on the other explanatory variable? This interactive effect is described by the interaction term, $\beta_{oP*ioP} \cdot ioP \cdot oP$

$$Pcorn = \beta_o + \beta_{ioP:oP} \cdot ioP + \beta_{oP:ioP} \cdot oP + \beta_{oP*ioP} \cdot ioP \cdot oP + res$$

The interaction term can be visualized as the degree of curvature of the surface fitted to the data. If there is no curvature, a flat plane describes the phosphorus content of corn relative to the two measures of soil phosphorus. If there is interaction (curvature) then a flat plane will not suffice.

2. Execute analysis.

Place data in model format:

Column labelled *Pcorn* with response variable phosphorus content of corn (ppm)

Column labelled *ioP*, with explanatory variable inorganic phosphorus (ppm)

Column labelled *oP*, with explanatory variable organic phosphorus (ppm)

Code the model statement in statistical package according to the GLM

$$Pcorn = \beta_o + \beta_{ioP:oP} \cdot ioP + \beta_{oP:ioP} \cdot oP + \beta_{oP*ioP} \cdot ioP \cdot oP + res$$

```
MTB > glm 'Pcorn' = 'ioP' 'oP' 'ioP'*'oP' ;  
SUBC> covariate 'ioP' 'oP' ;  
SUBC> fits c4;  
SUBC> residuals c5.
```

Fits and residuals from:

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

2. Execute analysis.

The overall mean is

$$\text{mean}(P_{\text{corn}}) = \beta_o = 76.18 \text{ ppm}$$

The regression equation for ioP is

$$P_{\text{corn}} = 62.6 + 1.23 \text{ } ioP$$

The regression equation for oP is

$$P_{\text{corn}} = 65.4 + 0.262 \text{ } oP$$

These are the simple regression coefficients. The equations have been written in slope/intercept form, rather than in GLM form. GLM form uses the grand mean \bar{y}_o rather than the Y-intercept. The Y-intercept is calculated from the grand mean and the slope estimate. The Y-intercept is not itself estimated because the estimate of the grand mean will be better. This is because the grand mean will, by definition, be at the centre of the cloud of data points. The Y-intercept will rarely be at the centre. In many cases the Y-intercept will be completely outside the data points, and so cannot be estimated directly.

The regression equation for both variables:

$$P_{\text{corn}} = 45.92 + 0.3278 \text{ } oP + 5.304 \text{ } ioP - 0.0830 \text{ } ioP \cdot oP$$

These are the estimates of the partial regression coefficients. Notice that they are not the same as the estimates of the simple regression coefficients.

3. Evaluate model.

Plot residuals versus fitted values

Structural Model. No bowls or arches are evident in a plot of residuals against fitted values, so straight line assumption acceptable.

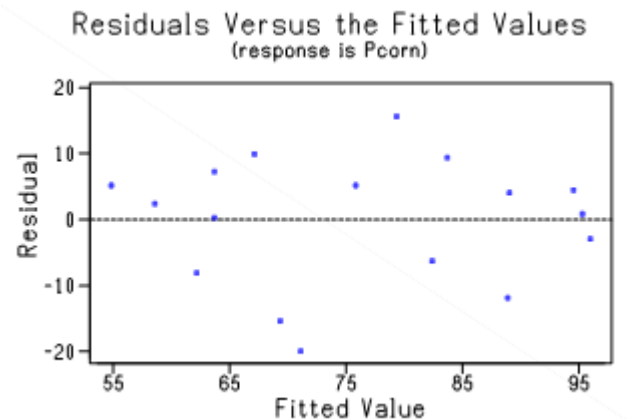
Error model

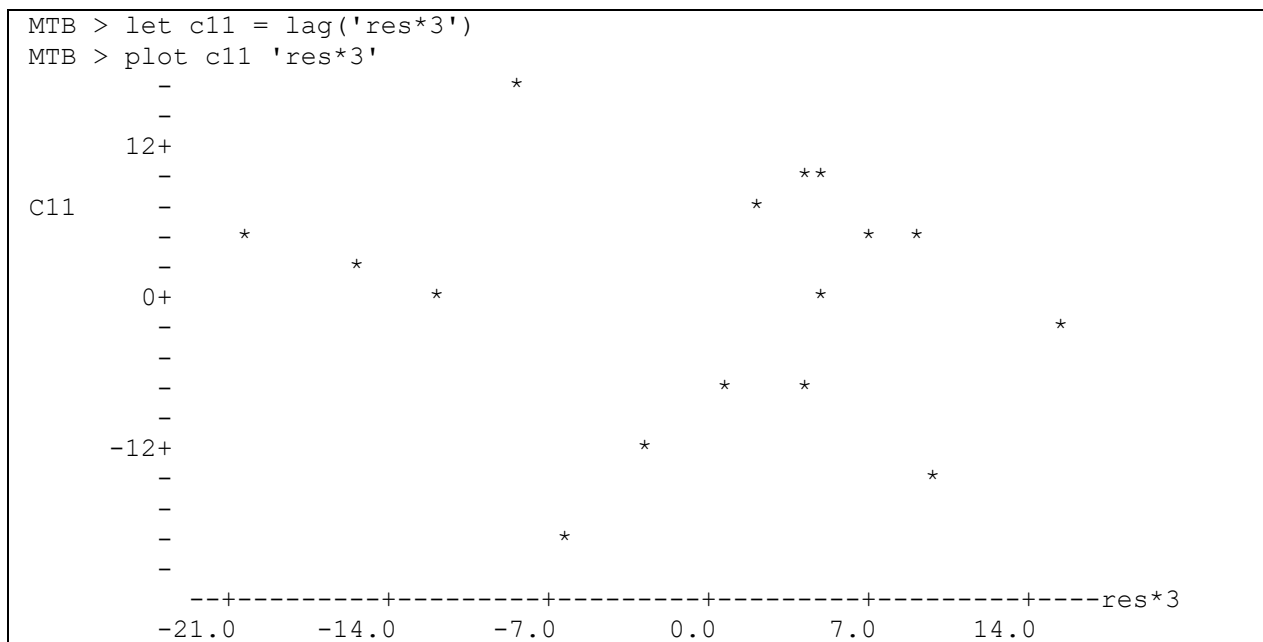
Residuals homogeneous ?

No. Spindles present.

Residuals independent ?

We have no information on temporal order or on spatial arrangement. So diagnosis is limited. If we take the observations in the order presented we find no trends upward or downward.





The plot of residuals versus themselves (at lag 1) shows no positive or negative trends.

Normal ?

Yes. The histogram looks close to normal.

```
MTB > hist 'res';
SUBC> increment 1.
Histogram of res      N = 17
Midpoint    Count
-2.00         1  *
-1.00         3  ***
 0.00         9  *****
 1.00         3  ***
 2.00         1  *
```

4. Partition df and SS according to model.

$$df_{\text{tot}} = 17 - 1$$

$$df_{\text{ioP}} = 1 \quad \text{Because relation is expressed by regression line}$$

$$df_{\text{oP}} = 1 \quad \text{Because relation is expressed by another regression line}$$

$$df_{\text{oP} \times \text{ioP}} = 1 \quad \text{This is the product of } df_{\text{ioP}} \text{ and } df_{\text{oP}}$$

$$df_{\text{res}} = 13 \quad \text{This is what is left over}$$

Calculate SS_{total}

$$SS_{\text{total}} = \sum Y^2 - n^{-1}(\sum Y)^2 = 4426.47$$

In Minitab:

```
MTB> let k1 = stdev('Pcorn')*stdev('Pcorn')*16
MTB> print k1
      k1      4426.47
```

4. Partition df and SS according to model

GLM:	$P_{corn} - \beta_o$	=	$\beta_{ioP:oP} \cdot ioP$	+	$\beta_{oP:ioP} \cdot oP$	+	$\beta_{oP*ioP} \cdot oP \cdot ioP$	+	res
Source:	Total	=	ioP		oP		ioP * oP		+ res
df	17 - 1	=	1	+	1	+	1		+ 13
SS	4426.5	+	2295.2	+	29.9	+	626.6		+ 1474.7
	Null model								Reduced model

4. Calculate likelihood ratio for overall (omnibus) model.

$$LR = (1474.7/4426.5)^{-17/2} = 11416$$

The regression model is 11,000 times more likely than the unreduced (null) model.

The omnibus model is far more likely than the null model so we proceed to an analysis of each term in the model.

5. Choose mode of inference.

If we had prior estimates of the three model parameters, priorist inference to a revised belief (posterior probability) could be used. We have no estimates. The agronomists who made these measurements may well have had no estimates at the time. The agronomists presumably used a fixed measurement protocol, which lends itself to frequentist inference. In a research setting such as an agricultural station there is no compelling reason to control Type I error at a fixed value in an observational study. An evidentialist approach (Edwards 1972, Royal 1997, Vieland and Hodge 1998) is appropriate.

Moving to a World Beyond “ $p < 0.05$ ”

In 2016 the American Statistical Association published a statement on the use and abuse of p-values (Wasserstein, R and N. Lazar 2016 *The American Statistician* 70, 129–133). Of particular note was the use of $p < 0.05$ as “significant,” and the tendency to treat this as yes/no conclusion as evidence. The ASA Statement stopped just short of recommending that declarations of “statistical significance” be abandoned. In 2019 that step was taken (Wasserstein et al 2019 *The American Statistician* 73- S1, 1–19: Editorial). The statement concluded, based on a review of the broader literature and on the 44 articles in the special issue, that it is time to stop using the term “statistically significant” entirely. Each of the articles in the special issue offered recommendations for change, including a wide variety of replacements. Only one article addressed teaching statistics, offering generalities. None of the articles mention the evidentialist approach. This is curious. Likelihood ratios are fundamental to both priorist and frequentist inference. They do what frequentist or priorist inference cannot do—say which model has better evidential support. They avoid the problems that attend declaring statistical significance at a fixed Type I error. They certainly have a place in Moving to a World Beyond ‘ $p < 0.05$,’ the title of the 2019 ASA statement.

In a subsequent article Benjamini et al. 2021 (Harvard Data Science Review, 3(3) <https://doi.org/10.1162/99608f92.f0ad0287>) concluded that “P-values and significance tests, when properly applied and interpreted, increase the rigor of the conclusions drawn from data.”

Moving to a World Beyond “ $p < 0.05$ ”

Likelihood ratios are readily grasped by university students in the sciences. At the same time evidential inference goes unknown by most thesis supervisors, thesis examiners, and journal referees. One way of moving beyond “ $p < 0.05$ ” is to compare an evidentialist approach relative to standard frequentist inference that stops at “statistically significant.” The two approaches often result in similar conclusions, with the exception that Type I error and “statistical significance” are absent from evidentialist inference. Multiple regression will be presented first with a standard frequentist approach, then with an evidentialist approach for comparison.

5. Define target of inference and whether sample is representative.

The population is not enumerable (e.g. all corn plants in Iowa). With frequentist inference the population is the result of a data generating mechanism defined by the experimental protocol and by the procedural statements for the response and explanatory variables. The population is hypothetical, generated by repeating the experimental protocol. Thus, for the purposes of investigating the relation of phosphorus content to soil phosphorus, this sample is representative of the similar experiments on the same variety of corn plants for the range of inorganic and organic phosphorus in this experiment. We cannot infer beyond these ranges.

With an evidentialist approach inference rests on the choice of the probability model to calculate the likelihoods.

The population is represented by the model

$$P_{corn} = \beta_o + \beta_{ioP:oP} \cdot ioP + \beta_{oP:ioP} \cdot oP + \beta_{oP:ioP} \cdot ioP \cdot oP + \varepsilon$$

where ε represents a normally distributed error. The justification rests on two arguments. First, the residuals conform satisfactorily to a normal error distribution. Second, the process that generates a normal distribution--errors generated by large number of contributing causes-- are a reasonable supposition in this case. Inference beyond the data at hand rests on the probability model and whether it applies to measurement of phosphorus content of corn relative to the soil. The supposition rests on the propensity of corn to take up phosphorus from soil, a propensity subject to enough sources of perturbation to result in a normal distribution.

6. State statistic and sampling distribution.

For evidentialist inference we will calculate a likelihood ratio for each parameter. If we judge that the deviation from homogeneity was serious we could generate a likelihood ratio by Monte Carlo methods. (Owen, A.B. 2001. *Empirical Likelihood*. CRC Press). For frequentist inference the statistic for calculating Type I from the sum of squares is the F-ratio. Its sampling distribution is the F-distribution. If we judge that the deviation from homogeneity was serious, then we could generate the sampling distribution by Monte Carlo methods.

6. State H_A H_o pairs for parameters.

Here are the hypothesis pairs listed in the order in which they appear in the model.

The first term concerns the effect of inorganic phosphorus, controlled for organic phosphorus.

$$H_A: \beta_{ioP:oP} \neq 0$$

$$H_o: \beta_{ioP:oP} = 0$$

This is equivalent to the following hypotheses concerning parameter.

$$H_A: \text{var}(\beta_{ioP:oP} \cdot ioP) > 0$$

$$H_o: \text{var}(\beta_{ioP:oP} \cdot ioP) = 0$$

The second term concerns the effect of organic phosphorus, controlled for inorganic phosphorus.

$$H_A: \beta_{oP:ioP} \neq 0$$

$$H_o: \beta_{oP:ioP} = 0$$

This is equivalent to the following hypotheses concerning the variance.

$$H_A: \text{var}(\beta_{oP:ioP} \cdot oP) > 0$$

$$H_o: \text{var}(\beta_{oP:ioP} \cdot oP) = 0$$

The third term concerns the interactive effect of organic phosphorus and inorganic phosphorus on phosphorus content of corn.

$$H_A: \beta_{oP*ioP} \neq 0$$

$$H_o: \beta_{oP*ioP} = 0$$

This is equivalent to the following hypotheses concerning the variance.

$$H_A: \text{var}(\beta_{oP*ioP} \cdot oP) > 0$$

$$H_o: \text{var}(\beta_{oP*ioP} \cdot oP) = 0$$

7. ANOVA – Table source, df, SS.

Source	df	Seq SS.	SS _{adj}	MS _{adj}	F---->p
ioP	1	2295.2			
oP	1	29.9			
ioP·oP	1	626.6			
Error	<u>13</u>	<u>1474.7</u>			
Total	16	4426.5			

This partitioning is in the order in which the variables are listed in the model.

7. ANOVA - partition variance according to model.

If we change the order of the variables, the partitioning will change.

To take this into account, we use the adjusted Sums of Squares.

That is, we use the SS for each explanatory variable when it is entered last into the GLM.

Source	df	Seq SS.	adjSS	MSadj.	F---->	p
ioP	1	2295.2	1061.8			
oP	1	29.9	149.4			
ioP*oP	1	626.6	626.6			
Error	<u>13</u>	<u>1474.7</u>	1474.7			
Total	16	4426.5				

The SS for $oP*ioP$ remained the same as the previous partitioning. This is because it was the last SS in the sequential partitioning.

The SS for ioP is smaller in this partitioning, because now it is last instead of first.

This partitioning (each variable last) is called the adjusted SS.

The adjusted SS no longer add up to the total $SS_{tot} = 4426.5$

so the total SS is not shown. The SS_{error} remains the same.

Statistical packages produce both the sequential and adjusted SS. The sequential SS is the default partitioning in some packages (e.g. R). The adjusted SS is the default in other packages (e.g. Minitab). Adequate packages allow the user to choose the partitioning.

7. ANOVA- Frequentist. Calculate MS, F, Type I error

Source	df	Seq SS.	SS_{adj}	MS_{adj}	F---->	p
ioP	1	2295.2	1061.8	1061.8	9.36	0.009
oP	1	29.9	149.4	149.4	1.32	0.272
ioP·oP	1	626.6	626.6	626.6	5.52	0.035
Error	13	<u>1474.7</u>	1474.7	113.4		
Total	16	4426.5				

$MS = SS/df$ for each term.

Calculate the correct F-ratios.

All terms in the model are regressions and so are taken as fixed. The correct F-ratio is relative to the residual MS.

$$F_{ioP:oP} = 1061.8 / 113.4 = 9.36$$

$$F_{oP:ioP} = 149.4 / 113.4 = 1.32$$

$$F_{op*ioP} = 626.6 / 113.4 = 5.52$$

```
MTB> cdf 5.52;
SUBC> F 1 13.
```

```
5.52      0.965
```

Statistical routines automatically report the Type I error from the F-distribution for each F-ratio in the ANOVA table.

The probability of obtaining $F_{ioP:oP}$ this large from our sampling distribution is $p = 0.035$

8. Recompute Type I error if necessary.

Assumptions judged to be met.

Monte Carlo calculations are available if assumptions judged not met.

9. Report frequentist statistical conclusions about model terms.

$$H_o: \text{var}(\beta_{ioP*oP} \cdot ioP) = 0 \quad 0.035 = p$$

We reject H_o that there is no interactive effect of the two forms of soil phosphorus on phosphorus content of corn.

For an analysis with an interactive term it is best to report the entire table, showing Sources of variance, df, SS, MS, F and Type I error.

The SS should be clearly labelled as adjusted (Type III) SS.

Similarly, we reject the null hypothesis of no effect of inorganic phosphorus.

$$F_{1,13} = 9.36 \quad p = 0.009$$

We cannot reject the null hypothesis of an effect of organic phosphorus.

$$F_{1,13} = 1.32 \quad p = 0.272$$

Critique of the analysis. We rejected the null hypothesis twice. In one case the rejection was marginal, with a Type I error of 3.5%. The other rejection was at a far smaller error rate, less than a 0.1%. The reject/not reject convention leaves aside the magnitude of the differences. In this study there was no cost-based reason for limiting Type I error to a fixed level. An evidentialist approach is sufficient—report the likelihood ratios.

7. ANOVA- Evidentialist. Calculate SS_{full} , $SS_{reduced}$, LR

Here is an ANOVA table showing calculation of the likelihood ratio.

Source	df	SS_{adj}	$SS_{reduced}$	SS_{full}	SS_{full}/SS_{red}	LR	Evidence
ioP	1	1061.8	1474.7	2536.5	1.720	100	Good
oP	1	149.4	1474.7	1624.1	1.101	2.3	Inadequate
ioP*oP	1	626.6	1474.7	2101.3	1.425	20	Some
Residual	13	1474.7					
Total	16	4426.5					

The flow of calculation begins with the adjusted sum of squares in the previous table.

The flow proceeds from left to right as with the pervious table.

The reduced model is the same for all three terms in the model.

This will not always be so. It will vary in more complex models.

The full model is calculated as

$$SS_{full} = SS_{adj} + SS_{reduced}$$

The full (null) relative to the reduced model is

This ratio decreases to 1 as the variance SS_{adj} decreases to zero.

The likelihood ratio is

$$LR = (SS_{full}/SS_{reduced})^{(n/2)}.$$

7. ANOVA- Evidentialist.

To avoid the loss of information that accompanies yes/no decisions we sort the likelihood ratios into categories. An easily remembered ranking is

Some evidence $LR \geq 20$
Good evidence $LR \geq 100$
Strong evidence $LR \geq 1000$

The categories were chosen from the point of view of an odds ratio.

20:1 odds approximate the odds at $p = 5\%$: $Odds = (1-5\%)/5\% = 0.95/0.05 = 19:1$

100:1 odds approximate the odds at $p = 1\%$: $Odds = (1-1\%)/1\% = 0.99/0.01 = 99:1$

1000:1 odds approximate the odds at $p = 0.1\%$: $Odds = (1-0.1\%)/0.1\% = 999:1$

These categories, while convenient, are not a recasting of $p < 5\%$, $p < 1\%$, *etc.*

We are no longer evaluating our results based on Type I error. We are evaluating based on a measure of relative evidence, the likelihood ratio.

8. Recompute Type I error if necessary.

Assumptions were judged to be met. If not met, we can estimate the likelihood ratio with Monte Carlo methods. In this case, inference no longer rests on the normal error model. Consequently inference applies only to the data at hand.

9. Report evidentialist statistical conclusions about model terms.

H_o / H_A : The likelihood of an interactive effect relative to no interactive effect.

$$LR = L(\beta_{ioP*oP} \neq 0) / L(\beta_{ioP*oP} = 0)$$

$LR = 20$ Some evidence for interactive effect

H_o / H_A : The likelihood of an organic phosphorus effect relative to no effect.

$$LR = L(\beta_{oP} \neq 0) / L(\beta_{oP} = 0)$$

$LR = 2.3$ Inadequate evidence for an effect of organic phosphorus

H_o / H_A : The likelihood of an inorganic phosphorus effect relative to no effect.

$$LR = L(\beta_{ioP} \neq 0) / L(\beta_{ioP} = 0)$$

$LR = 100$ Good evidence for an effect of inorganic phosphorus

Comparison of the analyses.

Source	df	SS _{adj}	p-value	Criterion	Category	LR	Evidence
ioP	1	1061.8	0.009	< 0.05	$p < 0.01$	100	Good
oP	1	149.4	0.272	> 0.05	$p > 0.1$	2.3	Inadequate
ioP*oP	1	626.6	0.035	< 0.05	$P < 0.05$	20	Some
Residual	13	1474.7					
Total	16	4426.5					

The Neyman-Pearson decision-theoretic criterion is less informative than Fisher sorting into categories. Fisher sorting of degree of certainty gave the same results as evidence categories. This may not always be the case.

The results are adequately reported by displaying each parameter estimate with its likelihood ratio from which measures of uncertainty (standard errors and p-values) can be calculated. The calculation of uncertainty from a likelihood ratio depends on the ratio of the degrees of freedom so in practice a measure of uncertainty needs to be displayed.

10. Report science conclusions.

There is good evidence ($LR = 100$) for uptake of inorganic soil phosphorus, compared to inadequate evidence ($LR = 2.3$) for uptake of organic soil phosphorus. This is consistent with the conclusions in the text example. However there is some evidence ($LR = 20$) for an interactive effect. This differs from the text example, which did not consider the interactive effect. Our best estimate of phosphorus content of corn is:

$$P_{corn} = 45.92 + 0.3278 \, oP + 5.304 \, ioP - 0.0830 \, ioP \cdot oP$$

A measure of uncertainty can be displayed directly (as a p-value) or indirectly (as a standard error). Statistical packages routinely report parameters with a standard error, from which the p-value measure of uncertainty can be calculated exactly.

Term	Coef	SE Coef
Constant	45.92	12.24
ioP	5.304	1.734
oP	0.3278	0.2856
ioP*oP	-0.08309	0.03536

For a reader with some knowledge of statistics, twice the standard error yields an approximate estimate of the uncertainty relative to a 5% cutoff. For example, $5.304 - 2(1.734) > 0$ hence significant at 5%.

The insufficient evidence for uptake of organic soil phosphorus suggests that we should simplify the model by dropping the oP term. The revised model would become:

$$P_{corn} = \beta_o + \beta_{ioP:oP} \cdot ioP + \beta_{oP*ioP} \cdot ioP \cdot oP + \varepsilon$$

However, the oP term must be retained in the model in order to estimate the interaction term. Because oP appears in the interaction term we would still need to know the value of oP to predict phosphorus in corn.