Model Based Statistics in Biology.
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 combined with statistics.
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.
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 most statistical packages are partial regressions. They express the rate of change in the response variable with respect to change in the explanatory variable, controlling for other variables.

The sum of squares that correspond to these partial regression coefficients are the adjusted (Type III) sum of squares. In most situations these are tested, rather than the sequential (Type I) sum of squares.

Regression coefficients express the rate of change of one variable with respect to another. Because of this relative quality, estimates can often be inferred to far larger populations than can means.

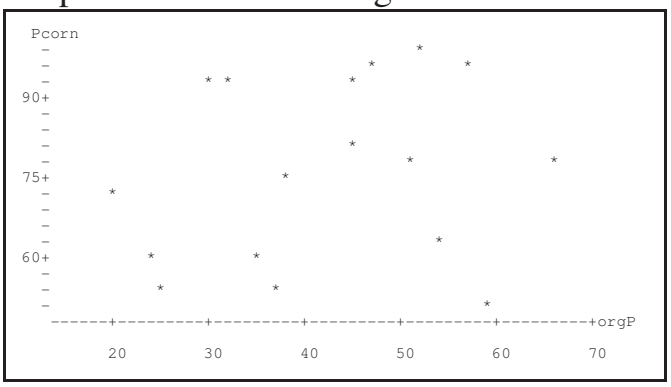
Introduction Analysis of data from Snedecor and Cochrane 1980 Table 17.2.1
Does phosphorus content of corn (ppm) from 17 Iowa soils at 20 deg C depend on inorganic and organic phosphorus in the soil?

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. *Pcorn* = ppm
Explanatory variables is organic phosphorus in soil. *oP* = ppm
Explanatory variables is inorganic phosphorus in soil. *ioP* = ppm

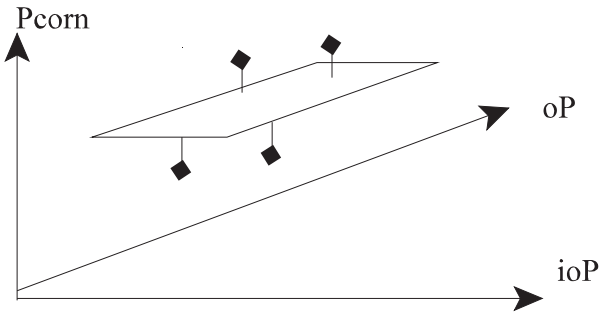
All variables are on a ratio type of scale.

Graphical model. L19Fig1

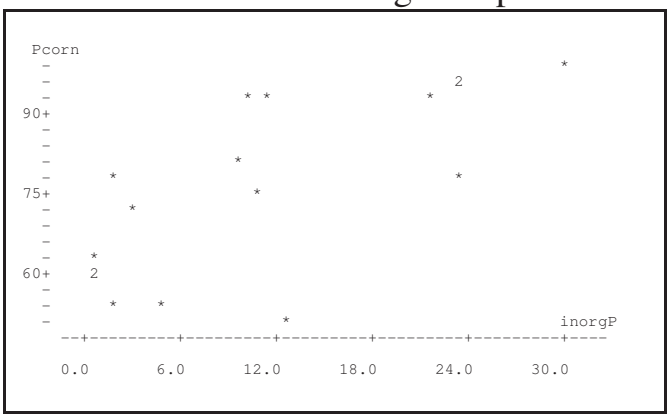


<---*Pcorn* versus *oP*

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



Pcorn vs *ioP*
Can fit a line through the points



To begin, a model with just a single explanatory variable, *ioP*

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

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 *Pcorn* versus *ioP*.

Next, a model with the other explanatory variable, *oP*.

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

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

1. Construct the model

Now a model that includes both explanatory variables

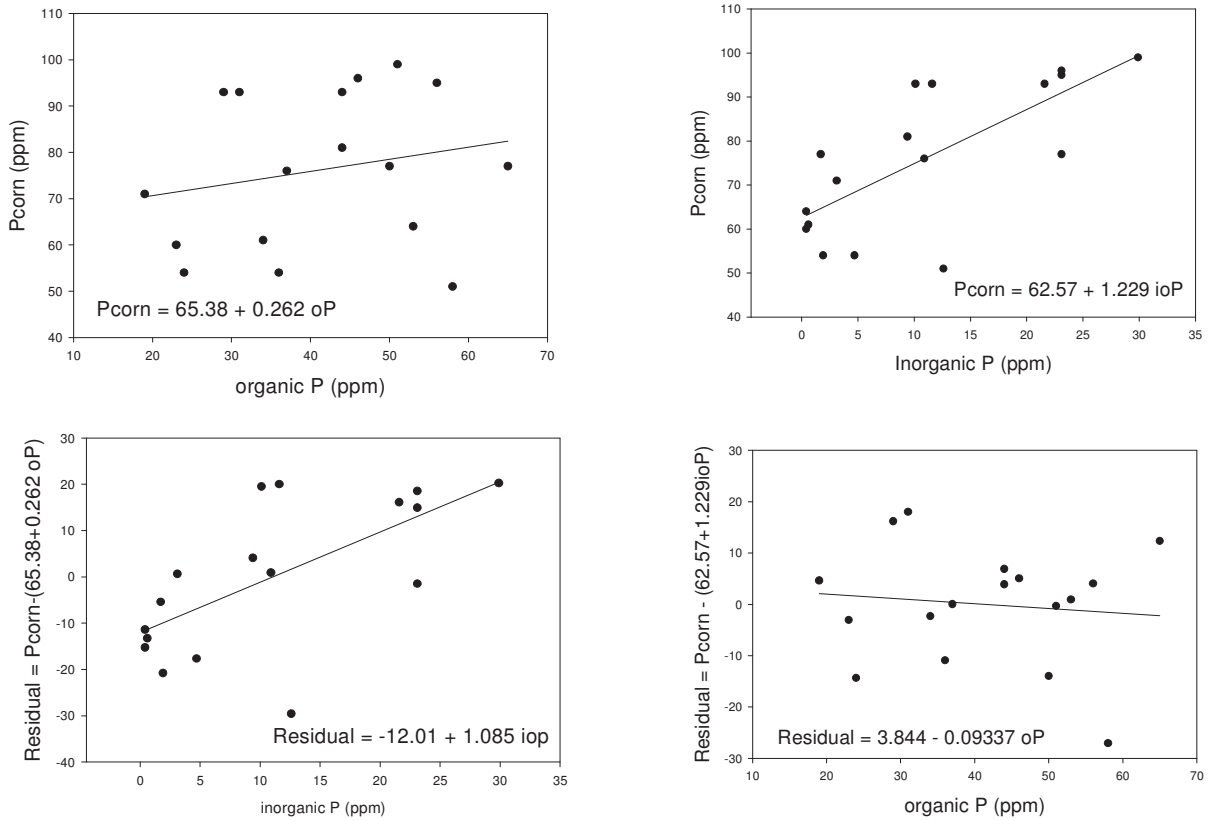
$$Pcorn = \beta_o + \beta_{oP:ioP} \cdot oP + \beta_{ioP:oP} \cdot ioP + \text{res}$$

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 ($\beta_{ioP:oP}$ $\beta_{oP:ioP}$)

L19Fig2



1. Construct the model (continued)

Partial regression is the same are regressing the residuals on the remaining variable.

Regress Pcorn against ioP: Pcorn = 62.6 + 1.23 ioP

Take residuals from this model.

Regress these against the other variable, oP.

Res = 3.84 -0.09337 oP

$\hat{\beta}_{oP:ioP}$

This estimate is close to the multiple regression estimate of $\hat{\beta}_{oP:ioP} = -0.111$

Next an interaction term. If 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, $\beta_{oP*ioP} \cdot ioP \cdot oP$

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

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

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

```
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

mean(Pcorn) = $\hat{\beta}_o$ = 76.18 ppm

The regression equation for *ioP* is

Pcorn = 62.6 + 1.23 *ioP*

The regression equation for *oP* is

Pcorn = 65.4 + 0.262 *oP*

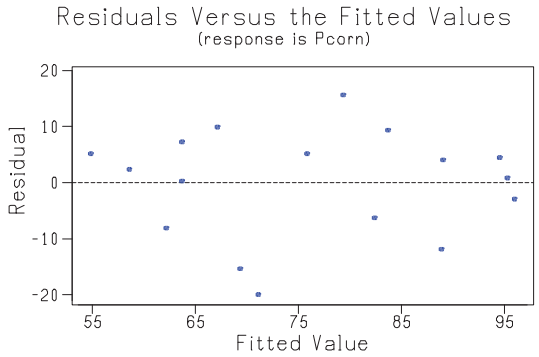
These are the simple regression coefficients. The equations have been written in slope/intercept form, rather than in GLM form. GLM form uses the grand mean β_o rather than the Y-intercept. The Y-intercept is calculated from the grand mean and the slope estimate. The Y-intercept is not itself estimated 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 will be completely outside the data points, and so very poorly estimated directly.

The regression equation for both variables:

Pcorn = 45.92 + 0.3278 *oP* + 5.304 *ioP* −0.0830 ioP*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.

Plot residuals versus fitted values.



3. Evaluate model.

a. No bowls or arches are evident in plot of residuals against fitted values, so straight line assumption valid
b. Residuals fixed ? Yes, no cones.
c. Distributional assumptions
<u>Homogeneous</u>? Yes.
<u>Independent</u> ? Yes.

```
MTB > let c11 = lag('res*3')
MTB > plot c11 'res*3'
        -                     *
        -
      12+
        -                                       **
C11     -                                    *
        -    *                                      *  *
        -           *
       0+                *                       *
        -                                                       *
        -
        -                                  *    *
        -
     -12+                             *
        -                                               *
        -
        -                        *
        -
          --+---------+---------+---------+---------+---------+----res*3
        -21.0     -14.0      -7.0       0.0       7.0      14.0
```

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

<u>Normal</u> ? 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. State population and whether sample is representative.

The population is not enumerable (e.g. all corn plants in Iowa).
The population is all values of phosphorus in corn, given knowledge of inorganic and organic phosphorus in the soil. The population is represented by the model

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

The population is all possible measured values, give 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 inorganic and organic phosphorus in this experiment. We cannot extrapolate beyond these ranges.

5. Decide on mode of inference. Is hypothesis testing appropriate?
The goal of the experiment was to decide whether phosphorus in corn depends on inorganic and organic phosphorus in soil. Hypothesis testing is an appropriate way of investigating whether either type of soil phosphorus is important, controlled for any effects of the other type.

6. State H_A / H_o with tolerance for Type I error
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 parameter.

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 parameter.

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

Test statistic will be F-ratio
Distribution will be F-distribution
Tolerance for Type I error. $\alpha = 5\%$

7. ANOVA - Calculate df and variance, partition according to model.

Compute total df, partition according to model.

```
GLM model at top of board, on left
ANOVA table at top, on right.
```

GLM:	$Pcorn - \beta_o$	=	$\beta_{ioP:oP} \cdot ioP$	+	$\beta_{oP:ioP} \cdot oP$	+	$\beta_{oP*ioP} \cdot oP \cdot ioP$	+	res
Source:	Total	=	ioP		oP		ioP * oP	+	res
df	17 − 1	=	1	+	1	+	1	+	13

df_{tot} = 17 − 1
df_{ioP} = 1 (because relation is expressed by slope)
df_{oP} = 1 (because relation is expressed by slope)
df_{oP*ioP} = 1 (because product of df_{ioP} and df_{oP})
df_{res} = 13 (what is left over)

Source	df	SS	MS	F ---->	p
oP	1				
ioP	1				
oP*ioP	1				
residual	?			? = 16−1 − 1 − 1	
Total	17 − 1				

Fill in df.

Calculate SS_{tot}

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

In Minitab:

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

Partition SS_{tot} according to model.
Model statement in any package will partition the variance

Add SS_{tot} to bottom of table.

Source	df	Seq SS.	adjSS	MSadj.	F---->	p
ioP	1	2295.2				
oP	1	29.9				
ioP*oP	1	626.6				
Error	13	1474.7				
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 correct for this, 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	13	1474.7	1474.7			
Total	16	4426.5				

In this example, the SS for *oP*ioP* remained the same as the previous partitioning (because it was last SS seq in the last partitioning).
The SS for ioP is smaller in this partitioning, because now it is last.
It was larger in the previous partitioning because it was 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.

Some regression routines do not calculate the Adj (Type III) SS, which are shown in this table. The GLM command will always do this.

Calculate the correct MS = SS/df if not already done by computer

Source	df	Seq SS.	adjSS	MSadj.	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	1474.7	1474.7	113.4		
Total	16	4426.5				

Calculate F-ratios

$F_{iop:oP} = 1061.8 / 113.4 = 9.36$
$F_{op:ioP} = 149.4 / 113.4 = 1.32$
$F_{op*ioP} = 626.6 / 113.4 = 5.52$

Calculate Type I error.
Statistical packages automatically compute the p-value from the F-distribution for each F-ratio, then places it in the table.

The p-value can be calculated from the F-distribution. For the interaction term.

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

  5.52     0.965
```

$p = 1 - 0.965 = 0.035$

The chance of obtaining an F_{ioP*oP} this large from our population of all possible measurements is p = 0.035

8. Recompute p-value if necessary.
Assumptions met, not necessary.

9. Declare decision about model terms, with evidence
Reject H_o: $\text{var}(\beta_{ioP*oP} \cdot ioP) = 0$ $\quad 0.009 = p < \alpha = 0.05$
Accept H_A: $\text{var}(\beta_{oP*ioP} \cdot oP) > 0$

Decision is to accept H_A that there are interactive effects of the two forms of soil phosphorus on phosphorus content of corn.

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.

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

The inference regarding relation of available phosphorus in corn to soil phosphorus is to all 17 soil types used in the analysis. If these 17 include the range of soil types in Iowa, then inference concerning the relation is to all corn plants in Iowa, not just to all possible measurements on this sample.

10. Analysis of parameters of biological interest.
Report the parameter estimates in an equation, then report some measure of uncertainty (standard deviation, standard error, etc) for each parameter.

$Pcorn = 45.92 + 0.3278\ oP + 5.304\ ioP - 0.0830\ ioP*oP$

Most GLM routines will report standard errors or confidence limits for each parameter.

Term	Coef	SE Coef	T	P
Constant	45.92	12.24	3.75	0.002
ioP	5.304	1.734	3.06	0.009
oP	0.3278	0.2856	1.15	0.272
ioP*oP	-0.08309	0.03536	-2.35	0.035

Organic and inorganic soil phosphorus have interactive effects on phosphorus content of corn. If we wish to look at the effects of soil phosphorus on corn phosphorus content we need to know both organic and inorganic concentrations in the soil. We need to use the interaction term to compute the expected levels of corn phosphorus.

β_{oP} was not significant so should we simplify the model by dropping the *oP* term? The revised model would be:

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

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.