Statistical Science.

Part III. The General Linear Model.

Chapter 9.1 Regression. Explanatory Variable Fixed by Experiment

ReCap.	Part I (Chapters 1,2,3,4)
ReCap	Part II (Ch 5, 6, 7)
ReCap	Part III (Ch 8 GLM components)
9.1	Explanatory Variable Fixed by Experiment
9.2	Explanatory Variable Fixed into Classes
9.3	Explanatory Variable Measured with Error
9.4	Exponential Functions
9.5	Power Laws. Linear Regression
9.6	Model Revision

Data files & analysis
SC_9_3_1.out
Ch9.xls

on chalk board

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops,
which combined models (what is the relation of scallop density to substrate?)
with statistics (how certain can we be?)

ReCap Part II (Chapters 5,6,7)
Data equations summarize pattern in data as a series of parameters (means, slopes).
Frequency distributions, a key concept in statistics, are used to quantify uncertainty.
Hypothesis testing uses the logic of the null hypothesis to make a decision about an unknown population parameter.
Estimation is concerned with the specific value of an unknown population parameter.

Today: The General Linear Model
Regression: Single explanatory variable fixed by experiment.
Work through this example, using a generic recipe.

Wrap-up

Regression a special case of the general linear model.
Response variable as a function of a single explanatory variable.
Relation between variables is expressed as a slope.
H_A/H_o pair about the slope parameter.
The variance in the response variable is partitioned according to the model.
The partitioning is used to evaluated the fit of the model to the data.
The inverse of the unexplained variance is used to calculate the likelihood ratio.
The ratio of explained to unexplained variance is used to calculate the F-ratio.
The F-ratio is used to calculate the p-value.
This is a measure of uncertainty on the measure of evidence, the likelihood ratio.
This measure of uncertainty is used to declare a descision about the null hypothesis.

GLM regression. Example: An experimental study with single fixed variable.

Example 9.3.1 from Snedecor and Cochran (1989). The quantity of interest is the phosphorus content of corn (*Pcorn* in ppm), in relation to the phosphorus levels in samples of soils with experimentally fixed levels of phosphorus (*Psoil* in ppm).

Psoil	Pcorn
1	64
4	71
5	54
9	81
13	93
11	76
23	77
23	95

Does the phosphorus content of corn increase when soil phosphorus is increased ?

1. Construct model

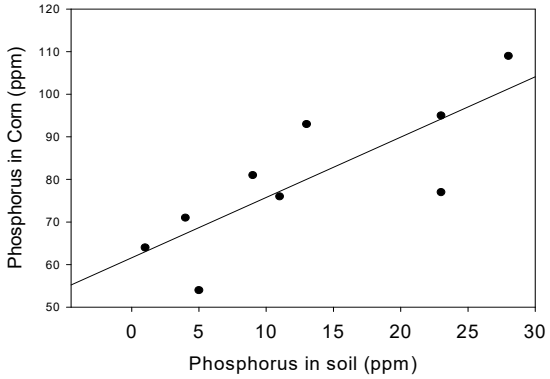
Constructing a model is unfamiliar to most students so we do it step by step. We will start with a verbal model, proceed to graphical, and from there write the formal model.

We begin by listing the variables with a symbol and type of units.

Pcorn = Phosphorus content of corn (ppm).
Continuous variable on a ratio type of scale.
Psoil = phosphorus content of prepared sample of soil (ppm).
Continuous variable on a ratio type of scale.

State the verbal model: Phosphorus content of corn depends on phosphorus content of soil.

Sketch a graphical model: A straight line relation.



The sketch distinguishes the response (Y-axis) from the explanatory variable (the X axis).

Next, table the variables by role

Name	Units	Type	Role
Pcorn	ppm	Ratio	Response
Psoil	ppm	Ratio	Explanatory

The response variable is called the dependent variable.
The explanatory variable is called an independent variable.
This defines the mathematical relation.
This can be confusing because “independent” variables can depend on each other.
So response/explanatory is preferable. It arises from the scientific question.

Separating the response from explanatory variables is the first and most important step in statistical analysis. If someone comes to you for statistical advice, the best way to help is often to ask them to state the response variables, and to separate these from the explanatory variables. This clears away the fog that surrounds the search for the “right statistical test.”

1. Construct model

Use the table of variables to construct the model.

		Units	Dimensions	Type of measurement scale
Response	*Pcorn*	(ppm)	dimensionless	ratio
Explanatory	*Psoil*	(ppm)	dimensionless	ratio

The scale type of the response variable determines the analytic procedure.
- Response variable
  - if nominal–> frequencies, then Generalized Linear Model GzLM
  - if nominal then median test, binomial tests, etc
  - if ordinal then "non-parametric" tests. e.g. Kruskal-Wallis
  - if ratio then GLM

The scale of the explanatory variables determine type of model.
- Explanatory variable in GLM (normal error)
  - if nominal then ANOVA
  - if ordinal then ANOVA
  - if ratio then regression
  - if ratio and otherwise then ANCOVA
- Explanatory variable in GzLM (normal and non-normal errors)
  - if nominal then ANODEV, G-tests, and extensions
  - if ordinal then ANODEV, G-tests and extensions
  - if ratio then logistic regression
  - if ratio and otherwise then ANODEV, likelihood ratio tests

Now write the model using names of quantities.

$Pcorn = f(Psoil)$.

"Phosphorus in soil depends on phosphorus in the soil"

Finally, write the model in more abstract form, which is what the computer will use to carry out the analysis.

$$Pcorn = \alpha + \beta_{PSoil} \cdot Psoil + \varepsilon$$

With practice, this sequence becomes practically automatic.

The model is for the population. α and β_{PSoil} are the parameters for the population.

α is the Y-axis intercept.

β_{PSoil} is the slope parameter.

	Units	Dimensions	Type of measurement scale
α	same as *Pcorn*	none (mass/mass)	ratio
β_{PSoil}	(ppm/ppm)	none (ppm/ppm)	ratio

1. Construct model

Here is the model for the sample. It uses parameter estimates from data.

$$Pcorn = \hat{\alpha} + \hat{\beta}_{Psoil} \cdot Psoil + residual$$

$\hat{\alpha}$ and $\hat{\beta}_{PSoil}$ are estimates of the parameters α and β_{PSoil}

An alternative convention is to use greek letters for population parameters, roman letters for estimates. Using this convention, the model for the sample is written with roman letters:

$Pcorn = \alpha + \beta_{PSoil} \cdot Psoil + \varepsilon$ Population

$Pcorn = a + b_{Psoil} \cdot Psoil + residual$ Estimates from sample

This convention falls apart after 2 letters. α, β, γ to a, b, g ?
So we will use hats on top of the greek symbols for estimates of these parameters.

2. Execute analysis. Place data in model format:

Column of data for response variable labelled *Pcorn*
Column of data for explanatory variable labelled *Psoil*

Code model statement in statistical package according to the GLM

$Pcorn = \alpha + \beta_{PSoil} \cdot Psoil + \varepsilon$

```
MTB> regress 'Pcorn' 1 'Psoil'
MTB> GLM Pcorn = Psoil.
```

All packages use a model statement to code the GLM. In some packages this model statement is typed (SAS, Minitab, R). In other packages it is present but not obvious. The example in the box shows the coding in Minitab for two different commands that produce the same result. Many packages have a graphics interface that allows you to code this model (*e.g.,* SPlus, SPSS). If you are using the graphics interface, it helps to look at the code produced, so that you understand how the model you produce with the graphic interface translates into a model statement.

2. Execute analysis. Compute fitted values and residuals.

The model statement is similar across statistical packages.
Extracting the residuals and fits differs among packages.

```
MTB > GLM Pcorn = Psoil;
SUBC> fits fits;
SUBC> res resisuals.    Minitab
```

```
Proc GLM;
   Model Pcorn = Psoil      SAS
```

```
CornModel <- lm(Pcorn ~ Psoil)
                            R
```

```
Regress Pcorn Psoil,
                        Stata
```

```
MTB > print 'Pcorn' 'fits' 'residuals'
  ROW  Pcorn     fits       res
    1     64   62.997    1.0031
    2     71    67.248    3.7524
    3     54    68.665  _14.6645
    4     81    74.332    6.6679
    5     93    80.000   13.0003
    6     76    77.166   _1.1659
    7     77    94.169  _17.1687
    8     95    94.169    0.8313
    9    109   101.253    7.7468
```

2. Execute analysis. Compute fitted values and residuals.

Here is the computional sequence:

1. Estimate the parameters
2. Calculate fitted values from the parameter estimates, for each data equation.
3. Calculate the residuals (response variable – fitted value).

These are readily obtained in from model based routines in statistical packages. We depend on the package to make these calculations correctly. Here is a brief tour of the machinery, for those who are interested.

1. Estimate parameters α and β_{PSoil} from the sample. The least squares estimate of β_{PSoil} minimizes the sum of the squared residuals (deviations of the data from the line).

$$SSE = \sum_{i=1}^{n} \left(Y - \hat{Y}\right)^2 = \sum_{i=1}^{n} \left(Y - \hat{\alpha} - \hat{\beta} X\right)^2$$

where there are n observations indexed by i. Some routines use iterative search: make a guess, compute the SS_{error}, make another guess, compute SS_{error} for this guess, compare to previous SS_{error}, continue until SS_{error} is as small as possible.

For a simple straight line model most routines obtain the estimate of $\hat{\beta}$ from the following formula.

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

For the Corn data the estimate is $\hat{\beta}_{Psoil} = 1.4169$

To estimate the y-intercept α we use the mean values of the response and explanatory variable.

mean(*Pcorn*) = $\hat{\beta}_0$ = 80 ppm

mean(*Psoil*) = 13.0 ppm

$\hat{\alpha} = \hat{\beta}_0 - \hat{\beta}_{Psoil}$ (mean(*Psoil*) = 80 – 1.42(13)

$\hat{\alpha}$ = 61.58 ppm

Here is the relation of the regression equation (which has the y-intercept α) to the GLM (which has β_o the mean value of the response variable).

GLM:	*Pcorn*	– 80.0 =		1.42 (*Psoil* – 13.0)		+ res
Regression Eq:	*Pcorn*	=	61.58	+ 1.42 *Psoil*		+ res

2. Calculate the fitted values from the parameter estimates,

Fitted values: Fits = E[*Pcorn*] = $\hat{\alpha} + \hat{\beta}_{Psoil} \cdot Psoil$

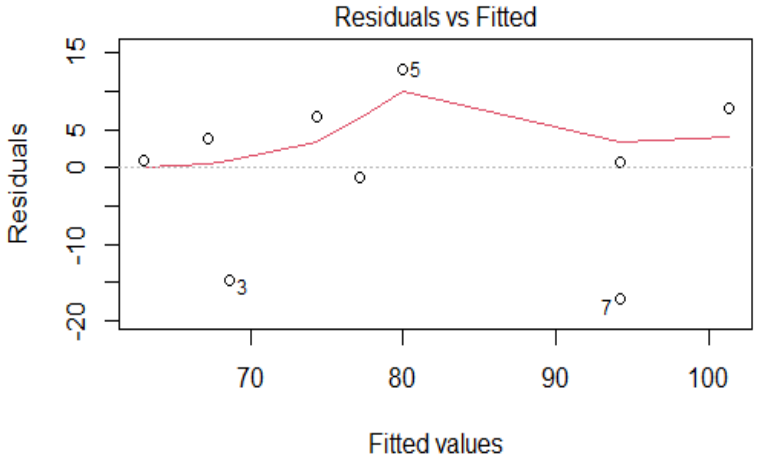
3. Calculate the residuasl from the fitted values:

Residuals: Res = *Pcorn* – Fits

3. Evaluate Model.

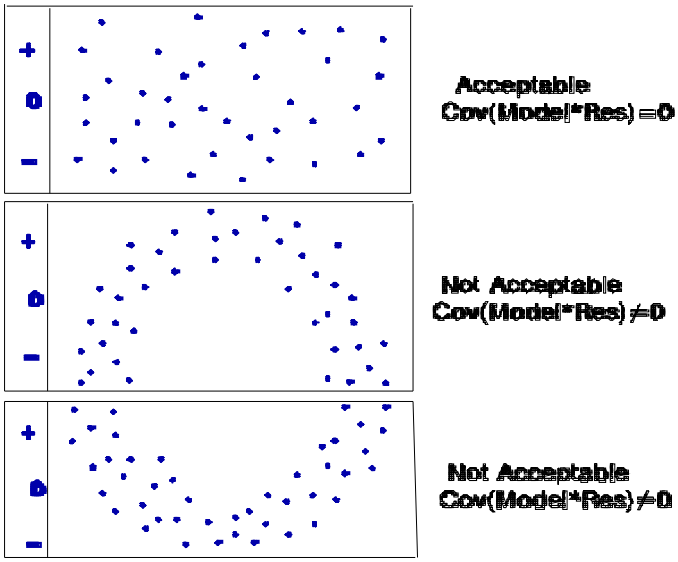
We use the residuals to evaluate the model.

We begin by examining the straight line assumption. Is this valid?
We use the residual versus fit plot to evaluate this assumption.
If the assumption is valid, the plot will show a band from left to right.
If the assumption is not valid, the plot will show either a bowl or an arch pattern. Looking at the residual fit plot we see that the straight line model is acceptable for the corn data.



Here is a diagram that contrasts an acceptable plot with two unacceptable plots. The acceptable plot shows a uniform band. The unacceptable plots show either an arch or a bowl. Bowls or arches result if the relation of the response to explanatory variable is curvilinear.

If this assumption is violated we go back to step 1 and reformulate the model to something other than a straight line.



3. Evaluate error model.

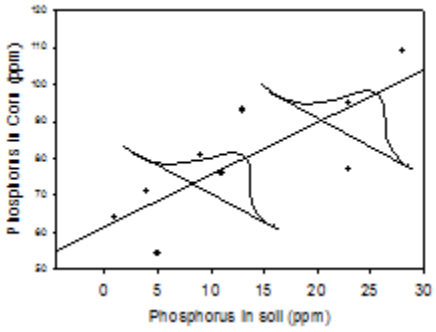
Next, we evaluate the error model that was used to estimate the parameters and likelihood ratios that will be used to calculate p-values (Type I error) from a statistical distribution (chisquare or *t* or *F*).

These distributions assume:

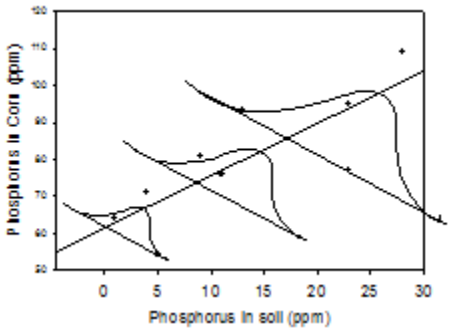
Fixed variance (errors homogeneous)
Normally distributed errors.
Independent errors
Unbiased estimate (errors sum to zero)

In practice the first assumption is the most important. The second assumption is more often mentioned and often incorrectly diagnosed (*i.e.* diagnosed before estimating parameters and computing residuals). The third assumption is best diagnosed if we know the order in which samples were gathered or we know the spatial arrangement of samples. The fourth assumption does not need to be checked when parameters are estimated by statistical packages that automatically produce unbiased estimates. We will focus on the first two assumptions, unless we have information allowing us to diagnose the independent error assumption.

To evaluate the fixed error assumption we again examine the residual versus fit plot. If the assumption is valid the plot will show a horizontal band. The dispersion around zero will be uniform across the plot. If the assumption is not valid the plot will show vertical dispersion that changes from left to right in the plot, usually with an obvious cone, fan, or spindle pattern. The residual fit plot for the corn data shows an acceptable band, with no evidence of change in dispersion going from left to right. The assumption of homogeneous error is acceptable for the corn data.



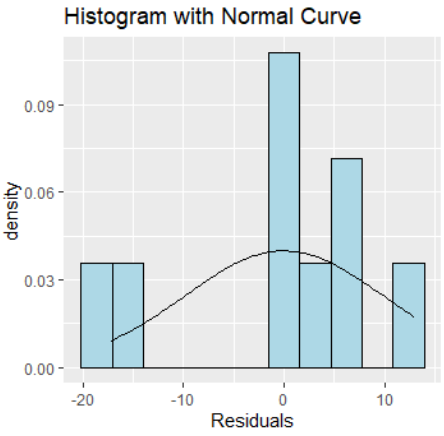
Here is a diagram that shows acceptably homogenous residuals in idealized form, superimposed on the corn data. The dispersion around the regression line is equal all along the line. As a result, the residual versus fit plot shows a uniform band running from left to right.



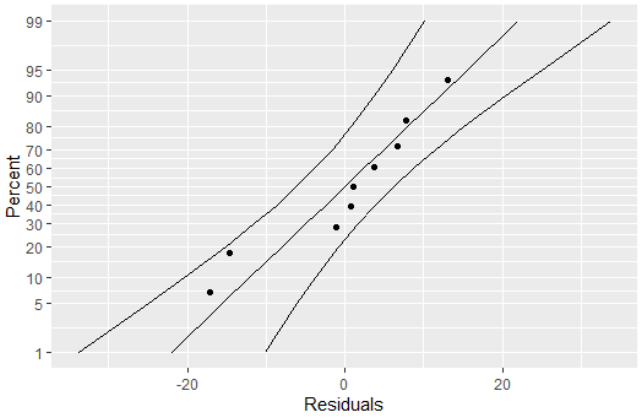
This diagram shows a pattern of heterogeneity in which the dispersion increases from left to right along the regression line. This pattern, which is common, will result in a cone opening out to the right in the plot of residuals versus fitted values. The opposite pattern, of a cone opening out to the left due to high variance at small fitted value, is rare.

3. Evaluate error model. Normal Errors

The next assumption is that the errors are normal. The histogram of residuals for the corn data is roughly symmetrical around zero, the mean value of the residuals. When we superimpose the normal distribution with the same mean and dispersion, we see deviations.



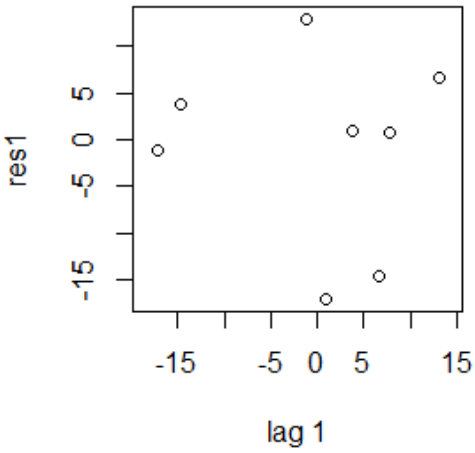
There is an excess near the center of the distribution, with skew to low values



A more detail plot for diagnosing normality shows each residual plotted on a normal probability scale. If the residuals are normal, the residuals will fall on a straight line rising diagonally from left to right.
The normal probability plot for the *Pcorn* analysis shows residuals that deviate from the diagonal. However, all the residuals fall within the bounds of the first (25%) and third (75%) quantiles of a normal distribution.

3. Evaluate error model. Independent Errors

Because this is a text example, we do not have information on spatial layout of samples, or on collection sequence. The data are ordered by magnitude of the explanatory variable, *Psoil*, so we expect the response variable to show positive autocorrelation if *Pcorn* is positively related to *Psoil*. We further expect this to be removed by the regression, leaving the residuals independent of one another. To check this we plot the each residual against its neighboring value. As expected, the plot shows independent residuals when ordered by the explanatory variable. There is no positive or negative trend in the graph. It is a fallacy to assume that data are 'not independent' because they were taken at a single location. See Hurlbert (1984) for an example of this fallacy.



3. Evaluate error model. Conclusion.

Based on graphical evaluation we judge the residuals acceptable. To check this judgement we will compare the p-value assuming normal error to the randomizaed p-value (no assumption of normal error).

4. Partition df and SS according to model.

Compute total degrees of freedom $df_{total} = n - 1 = 9 - 1 = 8$

Partition df_{total} according to model, using rules $df_{model} = 1$ regression line

$df_{res} = df_{total} - df_{model}$ $df_{res} = 8 - 1 = 7$

Model	$Pcorn - \beta_o$	=	$\beta_{PSoil} \cdot Psoil$	+	ε
Source	Total	=	$Psoil$	+	Resid
df	9 – 1	=	1	+	7

General linear model routines estimate the total sum of squares, partitioned according to the model into two parts, the SS due to regression, and the remainder SS for the residuals.

GLM	$Pcorn - \beta_o$	=	$\beta_{PSoil} \cdot P_{Soil}$	+	ε
Source	Total	=	$PSoil$	+	Resid
df	9 – 1	=	1	+	7
SS	2274	=	1473	+	800.4
	SS_{tot}	=	SS_{regr}	+	SS_{res}

The sums of squares are calculated from the data equations.

Data Equations for null model, Pcorn = mean(Pcorn)

	Data =	Model	+ Res	Res^2	
	64.00	80.00	-16.00	256.0000	
	71.00	80.00	-9.00	81.0000	
	54.00	80.00	-26.00	676.0000	
	81.00	80.00	1.00	1.0000	
	93.00	80.00	13.00	169.0000	
	76.00	80.00	-4.00	16.0000	
	77.00	80.00	-3.00	9.0000	
	95.00	80.00	15.00	225.0000	
	109.00	80.00	29.00	841.0000	
Sums	720.00	720.00	0.00	2274.0000	= sum(res^2)
	720/9 = 80				

Data Equations for regression model
Pcorn = 61.58 + 1.417*Psoil

Psoil	Data =	Model	+ Res	Res^2			
1	64.00	63.00	1.00	1.0055			
4	71.00	67.25	3.75	14.0778			
5	54.00	68.66	-14.66	215.0578			
9	81.00	74.33	6.67	44.4566			
13	93.00	80.00	13.00	169.0000			
11	76.00	77.17	-1.17	1.3601			
23	77.00	94.17	-17.17	294.7724			
23	95.00	94.17	0.83	0.6907			
28	109.00	101.25	7.75	60.0097			
						2274.00	SS total, from above
intercept	61.5804		0.00	800.4305	= sum(res^2)	800.43	SS residual
slope	1.4169					1473.57	SS improvement

4. What is the evidence? Calculate likelihood ratio for the overall model

The likelihood ratio compares the likelihood of the full model (fit to a constant) to the likelihood of the alternative model (constant + plus slope parameter)

To calculate the likelihood ratio, we take the ratio of the total variance to the unexplained variance, then raise it to a power of sample size/2.

This can be calculated directly: $LR = (2274/800.4)^{9/2} = 110$

This can also be calculated from the explained variance r^2.

$r^2 = SS_{explained} / SS_{total}$

$1 - r^2 = SS_{res} / SS_{total}$

The likelihood ratio is then calculated from the inverse of the unexplained variance.

$LR = (1 - r^2)^{-n/2}$

$LR = (SS_{res} / SS_{total})^{-n/2}$

$LR = (800.4 / 2274)^{-9/2} = 110$

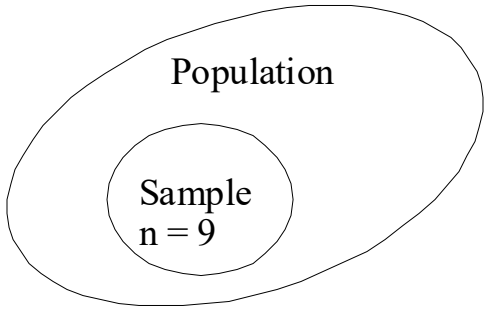
The alternative model is 110 times more likely than the full (null) model.

We have good evidence in favor of the alternative over the null model.

5. State population and whether sample is representative.

The population is a very large number of repeats of the experiment. We assume that our run of the experiment is representative of the population, for which the model is:

$$Pcorn = \alpha + \beta_{PSoil} \cdot PSoil + \varepsilon$$



Running the experiment thousands of times, while conceivable, is hardly feasible. We will use the law of large number to infer from the measurements we have (the sample) to the larger population (thousands of runs). We will infer from our sample to the true value of the relation of phosphorus in corn to phosphorus in soil, as represented by the parameters β_{PSoil} and α. From this point of view, the population is all possible measurements, given the experimental protocol. This view of inference emphasizes the importance of the experimental protocol.

5. State population – Infinite vs Enumerable populations.

The population is not enumerable. It is not all corn plants in the world nor is it all corn plants in Iowa, where the experiment was conducted. If it were an enumerable population, which it could be, we would need to state the enumerable population and justify the sampling protocol as taking a representative sample from that population.

Looking beyond the results from a single run at a single location, can we infer beyond the location at the experimental agriculture station? To do this, we need more information. For example, would the result depend on soils types? If we thought so, we would need to investigate this dependence by running the experiment across a range of the 17 soil types in Iowa.

5. Decide on mode of inference. Is hypothesis testing appropriate? If yes step 6, otherwise, calculate and report the likelihood ratio with parameter estimates.

We expect the results to vary from run to run due to measurement error and other sources. From the law of large numbers (Bernoulli 1713) we expect better and better estimates of the true value of the parameters as we increase the number of repeats. We will infer from our data to this large number of repeats. This is called frequentist inference.

With frequentist inference we use the likelihood ratio (a measure of strength of evidence) to calculate the uncertainty on this measure. Equivalently we ask: what is the cost (or risk) of drawing a false conclusion? In an applied context, such as this, one cost is recommending an agricultural practice, such as adding phosphorus to soil. In particular we wish to avoid recommending a practice with little or no effect. In such a context, we control our false conclusion rate at a fixed value. By convention this is 5%.

6. State H_A / H_o

The term of interest in the model is $\beta_{Psoil} \cdot Psoil$.
The research hypothesis for this term is that phosphorus in corn depends on phosphorus in the soil, and hence variation in $\beta_{Psoil} \cdot Psoil$ term. With hypothesis testing the goal is to reject the null hypothesis, that of no relation and hence zero slope. Phosphorus is an essential nutrient so we do not expect a negative relation. The test will be one-tailed.

H_A: $\beta_{Psoil} > 0$ The alternative hypothesis
H_o: $\beta_{Psoil} = 0$ The null hypothesis

hypotheses about the parameters in the model are equivalent to hypotheses about terms in the model.

H_A: $\text{var}(\beta_{Psoil} \cdot Psoil) > 0$ Equivalent to H_A: $\beta_{Psoil} > 0$
H_o: $\text{var}(\beta_{Psoil} \cdot Psoil) = 0$ Equivalent to H_o: $\beta_{Psoil} = 0$

6. State test statistic, and tolerance for Type I error, α

The test statistic will be the F-statistic. This is a variance ratio, which we will construct in the next step.

The distribution will be the F-distribution, which is readily obtained from a cumulative frequency distribution..
We judged the errors to be acceptably homogeneous and normal, so later we will check our judgement against a randomized p-value, free of assumptions.

Tolerance for Type I error is $\alpha = 5\%$

7. ANOVA. Table Source, SS, df, MS, F-ratio.

```
Establish relation of model to ANOVA table.
GLM on left side of chalk board, with df and SS
ANOVA table headings at top, to right.
Fill in below GLM, then move Source, df, SS to table.
Move Source, df, and SS to ANOVA table
Complete calculations of MS, F,
```

Having used the model to partition the degrees of freedom and sums of squares, we move this information to an ANOVA table. An ANOVA table is a GLM turned on its side.

Source	df	SS	MS	F	---->	p
PSoil	1	1473.6				
Res	7	800.4				
Total	8	2274.0				

7. Complete the ANOVA table

```
Computational flow is left to right,
  compute MS from SS and df in ANOVA table
  compute F from MS
```

MS = SS/df
$MS_{model} = 1473.6/1 = 1473.6$
$MS_{res} = 800.4/7 = 114.34$
$F = MS_{model} / MS_{res} = 12.89$

Source	df	SS	MS	F	---->	p
PSoil	1	1473.6	1473.6	12.89		
Res	7	800.4	114.34			
Total	8	2274.0				

Here are some additional statistics.

r^2 = explained variance = SS_{model} / SS_{tot}
r is the correlation coefficient
SS_{model} / SS_{tot} = coefficient of determination, for any GLM
1 – coefficient of determination = coefficient of non-determination.

Here is a summary, before moving on to the next step in the recipe.

The ANOVA table represents a sequence of computations from left to right.

Factors are listed, based on the model that was written.

df are listed, for each factor, and for the residuals. $df_{tot} = df_{model} + df_{res}$

SS_{tot} is estimated as the sum of the squared deviations of the response variable from the grand mean of the reponse variable. This is partitioned into a component for each factor in the model, and one residual component.

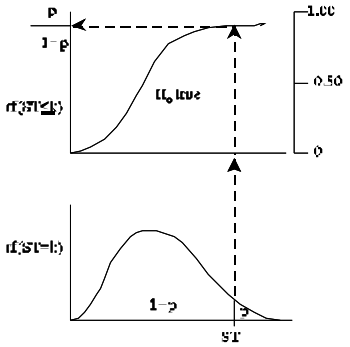
MS means squares are computed for each source (model factors + residuals).

F ratios are formed as ratios of mean squares MS

7. ANOVA: Calculate Type I error from F distribution.

Packages compute and place the p-value in the ANOVA table. p = 0.00885

```
MTB > cdf 12.89 k3;
SUBC> f 1 7.
MTB > print k3
K3        0.991142
MTB > let k3 = 1-k3
MTB > print k3
K3        0.008858
```



8. Recomputing p-value if residuals judged unacceptable.

p-values and confidence limits are computed from statistical distributions (χ^2, *F, t*, and normal). These calculations can be inaccurate if assumptions for these distributions are violated. The distortion depends on sample size. As a rule of thumb distortion can be serious if n < 30, less serious if 30 ≤ n ≤ 100, and usually not serious if n > 100. In this example n = 9, so violations are potentially serious.

When assumptions are not met, a recomputed Type I error is worth considering if:

n small	In this case Yes, n = 9
p near α	In this case No, p = 0.00886

Randomization can easily change our estimate of Type I error. However, the p-value is far from α and hence recomputing the p-value cannot easily change our decision. Randomized p-values generally differ from the theoretical p-values by less than a factor of two, rarely by a factor of 5 or more. In this case, a factor of 5 would leave the p-value less than 5%, and hence not alter the decision. So we might well judge at this point not to recompute the p-value.

8. Recompute p-value to check our judgement concerning the residuals.

While good judgement will suffice in day to day practice, it is not adequate when we must defend our conclusion from critical scrutiny. If we had to defend our conclusion to a journal referee or in court, then we would need to be prepared to defend our judgement with randomized p-value.

For the example at hand, it is of interest to find out whether our judgement was correct, that assumptions were met using the F-distribution to calculate a p-value.

In 4000 randomizations there were 27 instances of an F-ratio greater than 12.89. The randomized p-value is somewhat smaller than the p-value from the F-distribution. The p-value from the F- distribution was high by a factor of

```
MTB > let k2 = 27/4000
MTB > print k2
K2        0.00675
```

0.008858 / 0.00675 = 1.3

In 10,000 randomizations there were 79 instances of F > 12.89. p = 0.0079

0.008858 / 0.00795 = 1.1 (closer to p-value from F-distribution)

8. **Recompute p-value to check our judgement concerning the residuals .**
   The randomized p-value confirmed out judgement that the residuals, while not perfectly normal, were acceptably normal. The ratio of the p-values (F distribution / randomized) was less then 2. This is consistent with experience with randomization, which shows that the recalculated p-value will generally differ from that from the F-distribution by less than a factor of 2 (and rarely by more than a factor of 5).

   Having calculated a randomized p-value, we will report it. The more randomizations the better, so we will report the empirical value from 10,000 randomization: p = 0.0079. In the future, as our experience with judging assumptions from graphical displays grows we will have increased confidence in our judgement.

9. **Report the measure of evidence and degree of certainty**

LR = 110	The alternative hypothesis was 110 times more likely than the null.
p = 0.0079	The null hypothesis was rejected with a high degree of certainty. $p < \alpha = 5\%$ so reject H_o : var($\beta_{PSoil} \cdot Psoil$) = 0

Rejecting the null on the variance is equivalent to rejecting the null on the slope parameter, $\beta_{PSoil} = 0$.

We report the test statistic, the sample size (or equivalent) and the p-value
$F_{1,7} = 12.89$, p = 0.0079 by randomization

10. Report science conclusion. Interpret parameters of biological interest.

In this example our interest was in whether or not phosphorus content in corn was related to phosphorus content in the soil. However, we can learn more from the analysis than just this yes/no decision.

We begin by interpreting the slope parameter, estimated at $\hat{\beta}_{Psoil} = 1.4$ ppm/ppm.
The model with this parameter was 110 times more likely than the null model.
The p-value from the F-distribution was close to that by randomization. The deviation from normality in the residual plot did not have a large effect on the p-value estimate. So we can be confident that the deviation from normal error did not have a large effect on the parameter estimate.

The observed phosphorus content in corn increases by 1.4 units for each unit increase in soil phosphorus. Are the results consistent with amplification in corn relative to the soil? In other words, can we exclude the hypothesis ($\beta_{Psoil} = 1$) that there was no amplification?

To avoid repeated testing, we compute the confidence limits using the standard error of the estimate of the slope parameter $s_b = 0.3947$. This estimate is provided by the statistical package. It differs from the standard error of the mean.

95% confidence limits for parameters.

$P\{\text{Lower} < \beta_{Psoil} < \text{Upper}\} = 1 - \alpha = 95\%$

Lower =	$\hat{\beta}_{Psoil}$	–	$t_{0.025[7]} * s_b$	
Lower =	1.4169	–	2.3646 * 0.3947	= 0.484
Upper =	$\hat{\beta}_{Psoil}$	+	$t_{0.025[7]} * s_b$	
Upper =	1.4169	+	2.3646 * 0.3947	= 2.35

$0.484 \leq \beta_{PSoil} \leq 2.35$ for $\alpha = 5\%$.

These limits exclude the $\beta_{PSoil} = 0$ hypothesis. They do not exclude the hypothesis $\beta_{PSoil} = 1$ hypothesis, that available phosphorus increases in direct proportion to soil phosphorus. We cannot reject the hypothesis that phosphorus in corn increases in direction proportion to phosphorus in soil $\beta_{PSoil} = 1$.

Confidence limits allow us to exclude several hypotheses, not just the null of no relation.

For the sake of completeness, we report the regression equation with standard error of the slope parameter and sample size.

$$P_{corn} = 61.58 + 1.42\ P_{Soil} \quad s_b = 0.3947,\ n = 9$$

References

Bernoulli, Jakob (2005) [1713], The Art of Conjecturing, together with Letter to a Friend on Sets in Court Tennis (English translation), translated by Edith Sylla. Baltimore: Johns Hopkins Univ Press.

Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187-211

Snedecor, G.W. and Cochran, W.G. (1989) *Statistical Methods*. 8th Edition. Ames: Iowa State University Press.