Model Based Statistics in Biology.

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

Going to use the General Linear Model.
The GLM consists of familiar and new components.
Regression a special case of the general linear model.
Response variable as a function of a single explanatory variable.
Relation between variables expressed as a slope.
H_A/H_o pair about this parameter.
Model evaluated by partitioning variance in the data according to the model.
Decision declared about F = ratio of explained (model)
to unexplained (residual) variance.

GLM, applied to regression X variable fixed. Experimental studies.

Example 9.3.1 from Snedecor and Cochran (1989). 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 levesl of phosphorus (*Psoil* in ppm).

Psoil	Pcorn
1	64
4	71
5	54
9	81
13	93
11	76
23	77
23	95
28	109

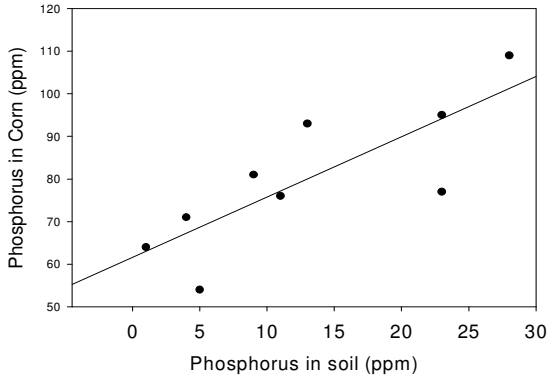
Does the phosphorus content of corn increase when soil phosphorus is increased ?

1. Construct model

Constructing a model can appear formidable, but it is not, if taken step by step. We will proceed around the triangle, starting with verbal model, proceeding to graphical, and finally formal model.

Verbal model: Phosphorus content of corn depends on Phosphorus content of soil.

Graphical model.



The verbal and graphic model help us to distinguish the response from the explanatory variables.

First, distinguish response variable from the explanatory variable or variables.
Response variable is

Pcorn = Phosphorus content of corn (ppm).
Continuous variable on a ratio type of scale.
This is also called the dependent variable.

The Explanatory variable is

Psoil = phosphorus content of prepared sample of soil (ppm).
The levels of *Psoil* are fixed in this experimental study.

This is an important step. 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 closest known statistical test.

1. Construct model

State type of measurement scale for response variable and explanatory variable(s).
Assign symbols, state units and type of measurement scale.

	Units	Dimensions	Type of measurement scale
Response	*Pcorn*	none (ppm)	ratio
Explanatory	*Psoil*	none (ppm)	ratio

This helps in deciding on the statistical analysis to use.

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

Explanatory variable in GLM
- if nominal then ANOVA
- if ordinal then ANOVA
- if ratio then regression
- if ratio and otherwise then ANCOVA

Explanatory variable in GzLM
- if nominal then G-tests and extensions
- if ordinal then G-tests and extensions
- if ratio then logistic regression etc

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.

This is the model for the population

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

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

1. Construct model

Here is the model for the population. It uses estimates of the parameters, 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}

We will use the greek letters for the population parameters, and the hats on top of the greek symbols for estimates of these parameters. An alternative convention is to use greek letters for population parameters, roman letters for estimates.

Using this convention, the model for the sample is also written as:

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

2. Execute analysis. Place data in model format:

Column with response variable, *Pcorn*.

Column with explanatory variable, *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.
```

Most packages have a model statement that codes the GLM directly.
The example in the box shows the coding in Minitab for two different commands that produce the same result. Most packages have a graphics interface that allows you to code this model. If you are using the graphics interface, be sure to look at the code produced, so that you understand how the model you wrote translates into a model statement in your package.

2. Execute analysis. Compute fitted values and residuals.

Most routines produce residuals and fits as output.

```
MTB > GLM Pcorn = Psoil;
SUBC> fits c3;
SUBC> res c4.
```

```
MTB > print 'Pcorn' 'fits' 'res'
  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.

The computation sequence is worth knowing:

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*] = $\hat{\alpha}$ + $\hat{\beta}_{Psoil} \cdot Psoil$

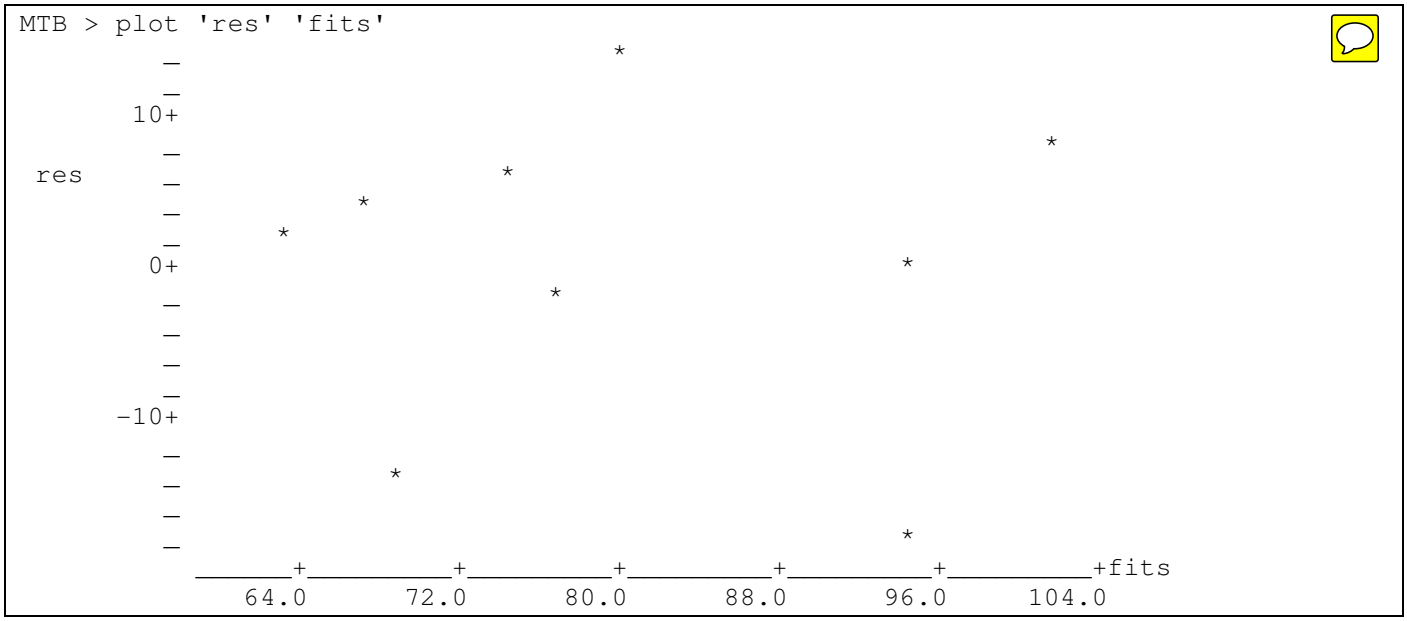
3. Calculate the residuasl from the fitted values:

Residuals: Res = *Pcorn* – Fits

3. Evaluate Model. Plot residuals against fitted values

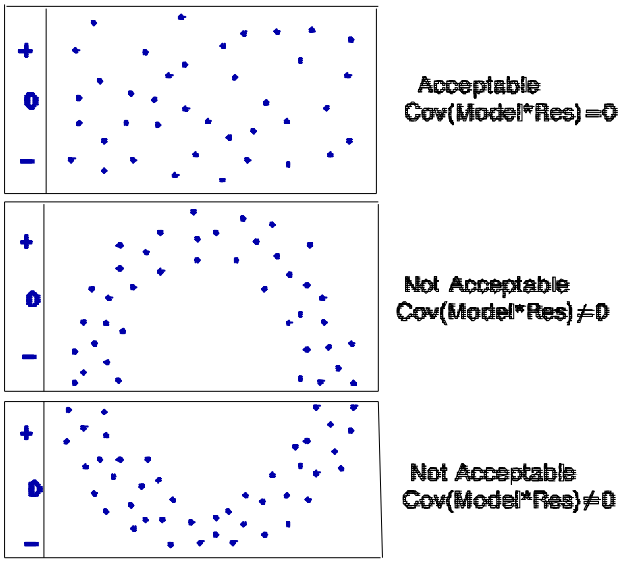
We use residuals to evaluate the model.

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. The residual fit plot shows that the straight line model is acceptable for the corn data.



The accompanying diagram contrasts an acceptable plot with two unacceptable plot. One shows an arch, where residuals are too low at high or low fitted values. The other diagram shows a bowl, where residuals are too high at high or low fitted values. 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.



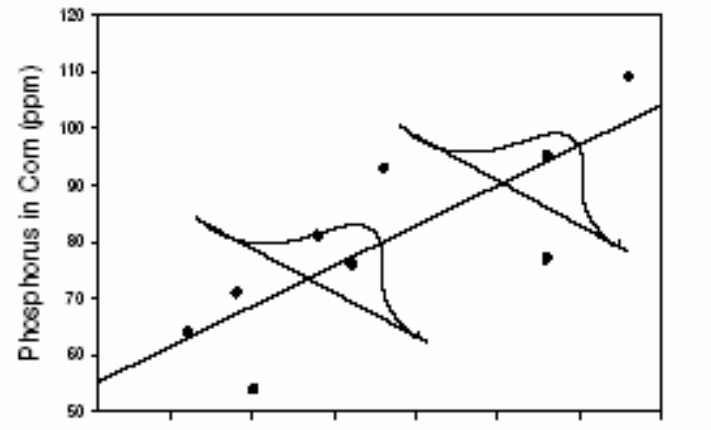
3. Evaluate model.

Next, we evaluate the error model that was used to estimate the parameters and that will be used to calculate p-values (Type I error) from a statistical distribution (chisquare or t or F) The four assumptions for using these distributions are:

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.

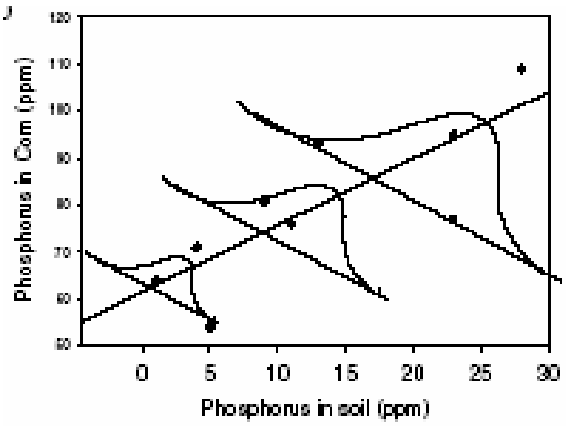
The fixed error assumption is often violated, especially when analyzing count data. To evaluate this assumption we again examine the residual versus fit plot. If the assumption is valid the plot will show a horizontal band. The dispersion around zero will be uniform across the plot. If the assumption is not valid the plot will show dispersion around zero changes from left to right in the plot, usually with an obvious cone 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.



The accompanying diagram 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.

3. Evaluate model. Homogeneous Errors

The next 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 model. Normal Errors

The next assumption, which is readily diagnosed, as that the errors are normal. The histogram of residuals for the corn data look "somewhat" normal, but there is evidence of skew to low values.

```
MTB > hist 'res'
 Histogram of res   N = 9

 Midpoint   Count
      -15       2  **
      -10       0
       -5       0
        0       3  ***
        5       2  **
       10       1  *
       15       1  *
```

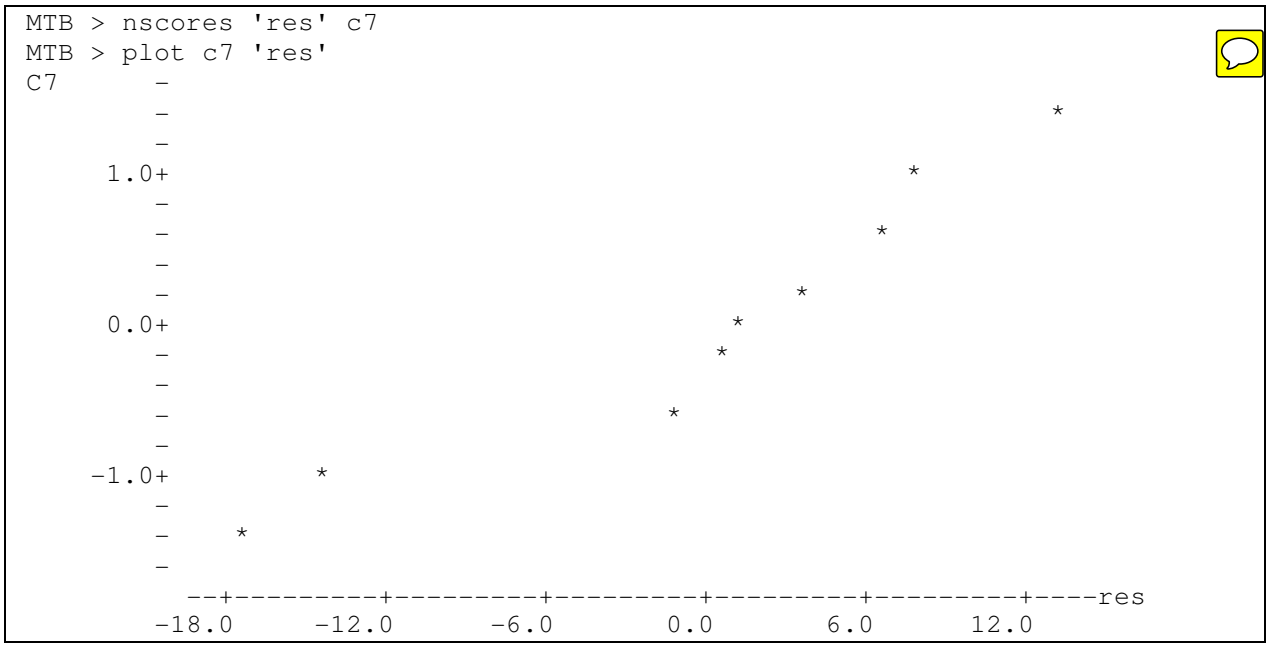
Another diagnostic is to compare the observed distribution to a normal distribution with the same mean and standard error.

Add diagram

In this plot there is again some evidence of deviation from normal distribution.

3. Evaluate model. Normal Errors

Perhaps the most useful plot for diagnosing normality shows each residual, transformed to a normal score, plotted against the (untransformed) residual. If the residuals are normal, this normal probability plot will exhibit a straight line rising from left to right.



The normal probability plot for the Pcorn analysis shows deviation from a straight line for strongly negative residuals.

3. Evaluate model. Independent Errors

This is a text example, we do not have information on spatial layout of samples, or on collection sequence. We will assume independence

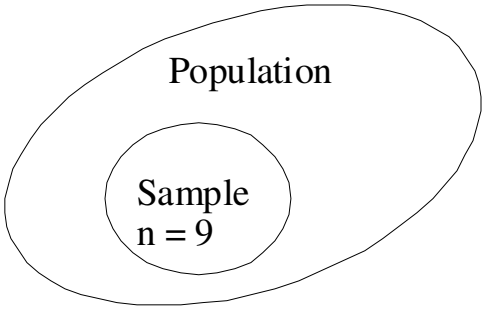
3. Evaluate error model. Conclusion.

Residuals appear to homogeneous, but not normal. We assume independence, we do not have enough information to evaluate this assumption.

We may need to use an empirical distribution to compute p-values or confidence limits.

4. State population and whether sample is representative.

The population is not an enumerable population. It is not all corn plants in the world nor is it all corn plants in Iowa, where the experiment was conducted. It is not even all corn plants at the agricultural station where the experiment was conducted.



The population is all values of phosphorus in corn, given knowledge of phosphorus in the soil. The population is represented by the model

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

To obtain the true value of phosphorus in corn given phosphorus in soil we would run the experiment repeatedly. We estimate the parameters β_{PSoil} and α from each experiment, then average across experiments to obtain the true value of the parameters. In this manner we obtain the expected value of the parameters, in the sense that statisticians use this word. The more times we run the experiment, the better our estimate.

Rather than running the experiment thousands of times (!!!), we are going 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.
This view of inference emphasizes the importance of the experimental protocol. From this point of view, the population is all possible measurements, given the experimental protocol.

If the goal is to estimate the relation between phosphorus in corn and in soil, then we want as wide a range as possible of soil types. If these 17 soils represent the complete range of soil types, then this sample will be representative for the purposes of establishing the relation. Thus, for the purposes of investigating the relation of phosphorus content to soil phosphorus, this sample is representative of similar experiments on corn plants from Iowa, provided it includes the range of possible soil types.

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

Once we have an acceptable model, we use the model to draw conclusions. If we already knew there was an effect, and we were interested the rate at which phosphorus content of corn increases as a function of phosphorus content of soil, we would skip hypothesis testing and report parameter estimates with confidence limits (skip to step 10). In this case we have no idea whether phosphorus in corn will be related to phosphorus in soil so we test for the presence of this effect. To do this we use the machinery of statistical hypothesis testing (step 6).

6. State H_A / H_o , test statistic and α

There is one term in the model, $\beta_{Psoil} \cdot Psoil$

The research hypothesis for this term is that there phosphorus in corn depends on phosphorus in the soil, and hence β_{Psoil} is not zero. variance in corn phosphorus due to soil phosphorus. Hence there will be variation in the fitted values, $\beta_{Psoil} \cdot Psoil$

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

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

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

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

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

The distribution will be the F-distribution, which is readily calculated.

There is some evidence that the errors are not normal, so later we will evaluate whether we should trust the p-value from the F-distribution.

Tolerance for Type I error is 5%

7. ANOVA: partition df according to model.

```
Establish relation of model to ANOVA table.
GLM on left side of chalk board.
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, p.
```

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

Partition df_{total} according to model, using rules

regression line $df_{model} = 1$

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

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

7. ANOVA: Calculate SS, partition according to model.

Here is SS from 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

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

General linear model routines estimate the total sum of squares, partition it, and produce an ANOVA table with the partitioned SS_{tot}

Move Source, df, and SS to 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. ANOVA: Partition df, SS according to model. Complete 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 for the regression

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 8.
MTB > print k3
K3          0.991146
MTB > let k3 = 1-k3
MTB > print k3
K3          0.00885403
```

8. Recompute p-value if necessary.

Confidence limits and p-values are computed from statistical distributions (chisquare, F, t, and normal). However, these calculations can be inaccurate if assumptions for these distributions are violated. The distortion depends on sample size. As a rule of thumb distortion is greatest if $n < 30$, it is less serious if $30 \le n \le 100$, and usually not serious if $n > 100$. In this example $n = 9$, so violations are potentially important.

When assumptions are not met, recomputed Type I error if two conditions are met:

n small (Yes, n = 9)

p near α (No, p = 0.00885)

The p-value is far from α and hence recomputing the p-value will not 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. This dicision is a matter of judgement, through experience with randomized p-values. 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. But 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 publish the result then we may well decide to compute a p-value free of assumptions, rather than appealing to personal judgement. If we had to defend our conclusion in court, then we would certainly use a p-value free of assumptions rather than appealing to judgement.

For the example at hand, it is of interest to find out whether our judgement was correct, that the p-value is so small that a better estimate won't change the decision at a 5% level.

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 theoretical distribution was high by a factor of

0.008854 / 0.00675 = 1.3

```
MTB > let k2 = 27/4000
MTB > print k2
K2        0.00675000
```

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).

We will report the randomized p-value because it is free of assumptions. But in the future we will be less likely to undertake randomization unless p is near α.

9. Declare and report decision about model terms.

p = 0.006750 (via randomization, hence no assumptions)

$p < \alpha = 5\%$ so reject H_o and accept H_A: $\text{var}(\beta_{PSoil} \cdot Psoil) > 0$

Hence accept equivalent hypothesis concerning parameter: H_A: $\beta_{PSoil} \neq 0$

Report decision with evidence:

There was a significant increase in available phosphorus with increase in soil phosphorus ($F_{1,7} = 12.89$, p = 0.0085 by randomization)

10. Report and 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 about the relation.

We begin by interpreting the slope parameter, for which the estimate is: $\hat{\beta}_{Psoil} = 1.4$
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?

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 model based routines in statistical packages.

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 do not exclude the hypothesis ($\beta_{PSoil} = 1$) that available phosphorus increases in direct proportion to soil phosphorus. We cannot conclude that increases in soil phosphorus will 'boost' available phosphorus disproportionately ($\beta_{PSoil} > 1$).

Note also that the confidence limits exclude the null hypothesis ($\beta_{PSoil} = 0$) of no relation. Confidence limits allow us to exclude several hypotheses, not just the null of no relation.

For the same of completeness, we could also report the regression equation:

$$P_{corn} = 61.58 + 1.42\ P_{Soil}$$