May 10, 2007

Chapter 3: Regression Models

Statistical Models

The chapter on sampling dealt with the statistical problem of collecting data for observational studies.

This chapter introduces the statistical problem of modeling data.

Research questions generally deal with a population of interest that is defined in terms of population parameters.

Statistical models can be used to describe populations of interest.

Like the population, the models are defined in terms of parameters and they provide relations between variables on interest.

A very simple model for a variable y is

$$y = \mu + \epsilon, \tag{1}$$

where μ can be regarded as the population mean or average and ϵ is a random error.

Suppose y represents the shell length of a random selected zebra mussel from a stream or lake in Michigan.

Then the model states that the length of a randomly selected zebra mussel is equal to some overall average value plus some random error.

The random error is needed since no two mussels have exactly the same shell length.

In other words, there is variability in shell lengths of the zebra mussels.

Notation for an Estimator

Once a sample of data is collected, the mean shell length μ of the population of zebra mussels can be estimated, which we would typically denote by

$$\hat{\mu}.$$

An important point to keep in mind is that if we had obtained a different sample of zebra mussels, then our estimate of μ would come out differently.

However, using results from statistical inference, we can quantify how much our estimator of μ will vary from sample to sample.

The simple model given in (1) is defined in terms of the model parameter μ.

However, there is another very important parameter not explicitly noted in (1) – the variance of the error term ϵ.

The variance is typically denoted by the Greek letter sigma-squared:

$$\sigma^2$$

which is a measure of "spread" of the distribution.

If $\sigma^2 = 0$, then there would be no variability of shell lengths which means all the zebra mussel would have exactly the same shell length.

Of course, this is not the case.

Mathematically, the variance σ^2 is the average over the entire population of the squared deviations: $(y - \mu)^2$.

The formal definition involves integral calculus.

Another issue of importance when specifying the model in (1) concerns the probability distribution of the random error ϵ.

Often in practice it is assumed that ϵ has a normal distribution when the variable y is continuous.

Although this normality assumption is often approximately valid, many times it is not valid.

The data should always be examined (usually graphically) to access if the distribution of the error (and hence y) is approximately normal or skewed, or some other shape.

The primary focus of this chapter is to introduce more complicated models. In particular, we shall focus on

- simple and multiple regression models,
- Spline models
- Nonlinear regression models,
- Logistic regression models, and
- Generalized linear models.

We shall start with an example to introduce the simple linear regression model.

The Simple Linear Regression Model.

Many statistical applications deal with modeling how one variable y, called a response or dependent variable, depends on another variable x which is called the independent or predictor variable (also called the regressor variable).

The simplest way to model a relation between the two variables is via a linear function:

$$y = \beta_0 + \beta_1 x,$$

where β_0 and β_1 are the y-intercept and slope of the line.

The problem with this model is that it is completely deterministic.

If one were to collect data on any two variables from an experiment, even if there is a linear relationship between the variables, the data points will not fall exactly on a line.

Thus, a probabilistic model is needed to account for the variability of points about the line.

This can be achieved by adding a random error the linear relationship above.

There are many reasons for the vast popularity of regression models in statistical practice.

One of the leading reasons is that regression models allow us to relate variables together in a mathematical form which can provide insight into the relationships between variables of interest.

Related to this reason, regression models allow us to determine statistically if a response variable is related to one or more other explanatory variables.

For instance, we may want to determine the effect of increasing levels of DDT on eggshell thicknesses – how does increasing levels of DDT effect eggshell thickness?. Regression models can help answer these sorts of questions.

Prediction

Another common use of regression models is to *predict* a response. For instance, if water is contaminated with a certain level of toxin, can we predict the amount of accumulation of this toxin in a fish that lives in the water?

Once again, regression models can be used to answer such questions.

In order to illustrate the basic principals and ideas, we provide an example of a simple linear regression model next.

Example. The population of the shrimp-like crustacean called *Diporeia* has been virtually wiped out in large areas of the lakes. The cause of this elimination is thought to be due to the relatively recent introduction of the Zebra mussel into these lakes. In studying this problem Nalepa et al (2000) examine the effect of the depth (in meters) of water in southern Lake Michigan where the diporeia were found on the dry weight (in mg) of this crustacean. Figure 1 shows a scatterplot of the data similar to the data found in the article.

The data below are similar to the data in the article:

Depth	Weight
35	0.59
35	0.66
38	0.82
45	0.63
50	0.60
52	0.68
54	0.62
62	0.47
75	0.40
75	0.50
95	0.49

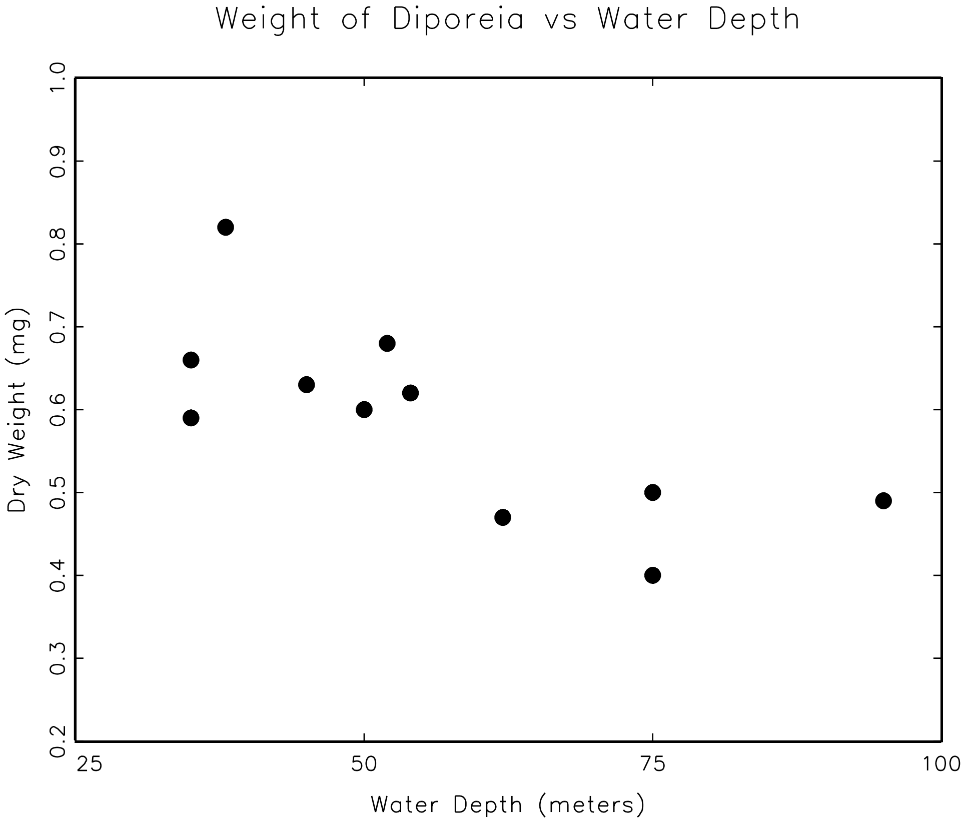

Figure 1: Scatterplot of the dry weight (in mg) of a sample of *Diporeia* versus the depth of water (in meters) in the southern portion of Lake Michigan where they were sampled.

From Figure 1, one can see a fairly strong relationship between between the weight of the diporeia and the depth of water where the diporeia are found.

From the plot, it appears that a straight line relation between weight of diporeia y and water depth x may be a reasonable way to model the data:

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{2}$$

where β_0 and β_1 are y-intercept and slope of the line respectively.

The points in Figure 1 appear to cluster about an imaginary line but they do not all lie on a line and that is why the error term ϵ is needed. The error ϵ is a random variable with mean zero.

If a sample of diporeia is obtained, then it is assumed that the random errors from (2) will be independent of each other and that they all have the same variance.

The model given in (2) is called the **simple linear regression model**.

The first statistical problem when analyzing the data is to estimate the parameters β_0 and β_1 in (2).

The problem of parameter estimation can be motivated by the simple model in (1) with $y_i = \mu + \epsilon_i$, for observations $i = 1, 2, \ldots, n$. One criterion for estimating μ from the data is to determine the value of μ that minimizes the sum of squares:

$$\sum_{i=1}^{n} (y_i - \mu)^2.$$

The value of μ that minimizes this sum of squares is $\hat{\mu} = \bar{y}$, the sample mean which is what is typically used to estimate the mean.

In the simple linear regression framework, the same criterion is also used. That is, find the values of β_0 and β_1 that minimize the sum of squares:

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

This method is known as *least-squares* since we are finding the estimates of the parameters that make the sum of squares take the least value possible.

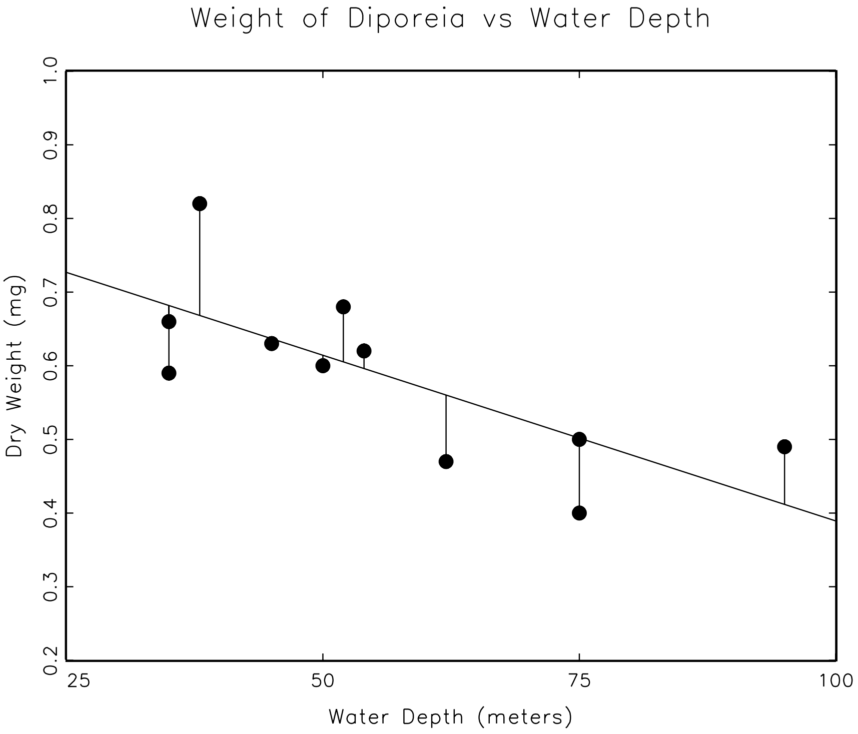

Figure 2: Illustration of the least-squares criterial for fitting a line to data. The figure shows the scatterplot of the dry weight of *Diporeia* versus the depth of water as in Figure 2 along with the least-squares regression line. The line is found by determining the best fitting line in terms of minimizing the sum of squared vertical distances between the points and the corresponding point on the line.

Geometrically, finding the least-squares estimates corresponds to determining the best fitting line to the data where “best” means the line minimizing the sum of squared vertical differences between the actual data points and the corresponding points on the line as illustrated in Figure 2.

The solution to the least-squares problem requires multivariate differential calculus which we will not go into here. The least squares estimators are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and are given by the following formulas:

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\\ \hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}}\end{aligned}$$

where

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

and

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Statistical software programs typically perform these computations for us. In SAS, one can use Proc Reg using the following syntax:

```
proc reg;
     model y = x;
run;
```

Below is the SAS output from Proc Reg for the *Diporeia*:

```
                              The REG Procedure
                                Model: MODEL1
                            Dependent Variable: y

                             Analysis of Variance

                                        Sum of           Mean
Source                   DF           Squares         Square    F Value    Pr > F

Model                     1           0.07407        0.07407      10.59    0.0099
Error                     9           0.06294        0.00699
Corrected Total          10           0.13702


           Root MSE                   0.08363    R-Square     0.5406
           Dependent Mean             0.58727    Adj R-Sq     0.4896
           Coeff Var                 14.24016


                             Parameter Estimates

                            Parameter       Standard
   Variable      DF          Estimate          Error     t Value     Pr > |t|

   Intercept      1           0.83914        0.08139       10.31       <.0001
   x              1          -0.00450        0.00138       -3.25       0.0099
```

The last part of the output gives the least-squares estimates for the intercept and slope: $\hat{\beta}_0 = 0.83914$ and $\hat{\beta}_1 = 0 - .00450$. The estimated regression line can be expressed as

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= 0.83914 - 0.0045x,\end{aligned}$$

where $\hat{y}$ is known as the *predicted value* from the regression line.

Hypothesis Testing for Simple Linear Regression.

Typically, the parameter of primary importance in a simple linear regression is the slope β_1.

The slope measures the average rate of change in the response variable relative to the predictor variable.

Occasionally interest also lies in the y-intercept β_0, but not nearly as often. From the fitted regression line, the slope is estimated to be $\hat{\beta}_1 = -0.0045$ which can be interpreted as following: the average weight of the diporeia decreases by 0.0045 mg. at each additional meter of depth in the lake.

Hypothesis Testing continued ...

It is natural to ask at this point if the slope differs significantly from zero.

If the slope β_1 does indeed equal zero, then the weight of the diporeia will not depend on the depth of the water (this assumes that the relationship between size and water depth is indeed linear in the range that is being studied).

Assuming the distribution of the error ϵ is normal, then the estimated regression parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ will also have normal sampling distributions.

The formulas for the estimated standard errors for these estimators are messy, but usually statistical software programs give the estimated standard errors automatically.

From the SAS output above, the estimated standard errors for the intercept and slope are

$$\hat{se}(\hat{\beta}_0) = 0.08139 \text{ and } \hat{se}(\hat{\beta}_1) = 0.00138.$$

One can test the following hypotheses:

$$H_0 : \beta_0 = 0 \text{ and } H_0 : \beta_1 = 0,$$

using a t-test statistic:

$$t = \frac{\hat{\beta}_0}{\hat{se}(\hat{\beta}_0)} \text{ and } t = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)}.$$

If the null hypotheses are true, then the t-test statistics follow t-distributions on $n-2$ degrees of freedom.

As long as the error distribution does not deviate to much from normality, the inference based on the t-distribution should be approximately correct. Note that two degrees of freedom are lost by estimating both the intercept and slope. SAS automatically tests these hypotheses and reports *two*-tailed p-values in the output.

Hypothesis Testing for the Slope

For example, to test if the slope in the regression model for diporeia is zero or not, the t-test statistic is

$$t = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} = \frac{-0.00450}{0.00138} = -3.25.$$

There were $n = 11$ observations in this data set and the degrees of freedom for the t-test statistic is $n - 2 = 11 - 2 = 9$ which can be seen under "DF" in the ANOVA table at the top of the output for the Error.

The resulting two-tailed p-value is $p = 0.0099$ indicating strong evidence that the slope is not zero.

In other words, we have strong evidence that the average weight of the diporeia does depend on the depth at which they are found.

Although SAS reports a p-value for a two-tailed test, we can use the statistics to test more general hypotheses about the slope.

For instance, we can test the hypothesis $H_0 : \beta_1 = \beta_1^*$ where β_1^* is some hypothesized value of the slope.

The test statistic then becomes the standardized difference between the estimated slope and the hypothetical slope:

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{se}(\hat{\beta}_1)},$$

which should be compared to a t-distribution on $n-2$ degrees of freedom. The testing procedure of $H_0 : \beta_1 = \beta_1^*$ at a significance level α is given as follows: Reject H_0 and accept

$$\begin{array}{lll} H_a : \beta_1 > \beta_1^* & \text{if} & t > t_{\alpha,n-2} \\ H_a : \beta_1 < \beta_1^* & \text{if} & t < -t_{\alpha,n-2} \\ H_a : \beta_1 \neq \beta_1^* & \text{if} & t > t_{\alpha/2,n-2} \text{ or } t < -t_{\alpha/2,n-2} \end{array}$$

Confidence intervals for Regression Coefficients. Confidence intervals can be computed for the y-intercept and the slope coefficients. A $(1-\alpha)100\%$ confidence interval for the slope β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{se}(\hat{\beta}_1).$$

A confidence interval for the intercept can be computed similarly.

The confidence intervals have the same interpretation as confidence intervals for any other parameter.

If we compute a 95% confidence interval for β_1, then we can interpret this as a procedure that will generate an interval containing the true value of β_1 95% of the time when repeating the experiment over and over again.

We shall ignore for now the ANOVA table given in the SAS output – it becomes more important for multiple regression problems.

The R^2 value is a very important statistic in regression applications.

From the SAS output, we see that the $R^2 = 0.5406$.

The value of R^2 is always between zero and one and it represents the proportion of variability in the response that is explained by the regression model.

We shall discuss this in more detail in the multiple regression section.

Accessing the Fit of the Model. We assumed that the diporeia data could be modeled adequately by a simple linear regression. In other words, the functional relationship between the mean diporiea weight and water depth is linear. At this point, it may be useful to recall the famous quote of George Box (1980):

> "All models are wrong, but some are useful."

The relationship between the weight of diporeia and water depth may be very complicated. However, as Figure 1 shows, it appears that the relationship can probably be modeled fairly well by a straight line regression.

Of course, it is not always the case that the relationship between a response and a predictor variable will be approximated well by a straight line.

One of the essential components of performing a regression analysis is to access how well the model fits the data.

This is often done using *residual plots.*

Residuals

Once the model is fitted, we can compute predicted values $\hat{y}_i$ for the ith observation.

Ideally, we would like the straight line to provide a good fit to the data and consequently we would like to see small differences between the observed response y_i and the corresponding fitted value $\hat{y}_i$.

The ith residual r_i is defined as the difference between these two quantities:

$$r_i = y_i - \hat{y}_i.$$

Note that the error ϵ in (2) can be expressed as $\epsilon = y - \beta_0 - \beta_1 x$.

Thus, one can think of the ith residual as an estimate of the ith error term. The error in the model is suppose to be random, and consequently, the residuals from the fitted model should look like a random scatter of points.

Residual Plots

A common model diagnostic tool is to plot the residuals versus the fitted values (r_i vs $\hat{y}_i$) or the residuals versus the x_i values.

If the model is specified correctly, then these plots should show no structure.

If there is some obvious structure in a residual plot, then there is something wrong with the specification of the model (perhaps the straight line relationship is inadequate).

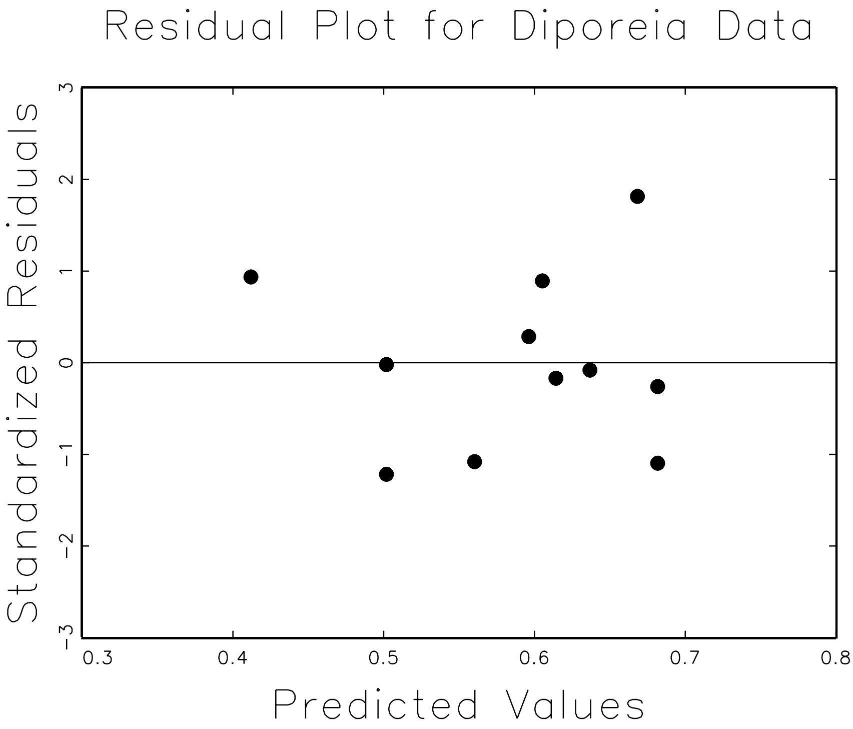

Figure 3: Standardized residual plot for the diporeia data versus fitted values $\hat{y}_i$.

Figure 3 shows a plot of the standardized residuals versus predicted values. The standardized residuals are simply the usual residuals divided by the square root of the mean squared error (MSE) of the regression fit.

The MSE is an estimate of the variance of the model error ϵ and is given by the formula:

$$\text{MSE} = \sum_{i=1}^{n} r_i^2/(n-2).$$

Recall that for the normal distribution, about 95% of the distribution lies within two standard deviations of the mean and almost all the observations will lie within three standard deviations of the mean. Consequently, almost all of the standardized residuals should take values between ±3.

Also recall that the random error ϵ has mean zero. Corresponding, the average values of the residuals is always zero. Because the residuals should show a random scatter about zero, it is helpful in residual plots to include a horizontal line at $r = 0$ as can be seen in Figure 3.

Figure 3 does not show any apparent structure which indicates that there are no obvious problems with the fit of the straight line model.

The next example illustrates a case where the straight line model is highly suspect.

Example. The *Adenosine triphosphate* (ATP) molecule is the universal currency of energy exchange in biological systems. ATP bioluminescence assay is routinely used for microbiology quality assurance purposes.

It provides rapid quantification of bacteria in biological samples.

A test involving changes in ATP content used to determine toxicity of industrial waste water discharged to a sewer was conducted.

ATP changes in the organism *P. Phosphoreum* in industrial sludge was measured for different levels of the pH (Arretxe et al 1997).

The data in the following table were derived from plots in the Arretxe (1997) publication:

pH	ATP (mg/g)
2	.16
5	.31
8	.32
9	.26
11	.07

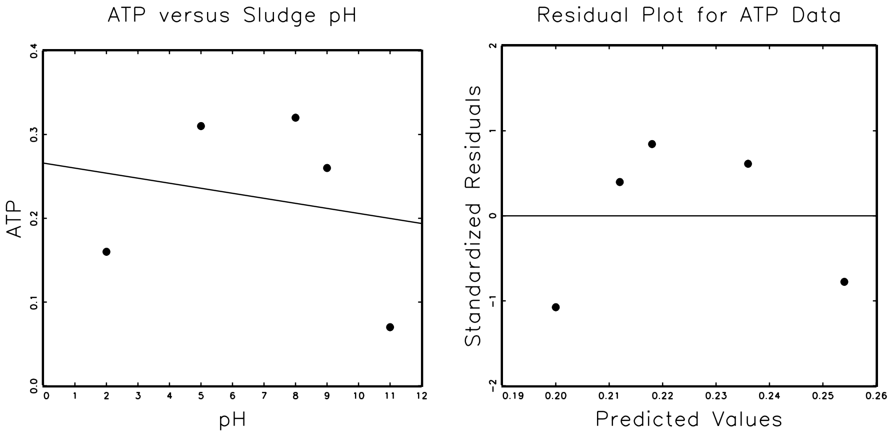

Figure 4: Standardized residual plot for the ATP sludge data versus fitted values $\hat{y}_i$.

The data is plotted in the left frame of Figure 4 along with the estimated regression line.

The right frame shows the standardized residual plot.

The estimated regression function is given by

$$\hat{\text{ATP}} = 0.266 - 0.006\text{pH}.$$

The estimated standard error for the estimated slope $\hat{\beta}_1$ is 0.0171 which gives a t-test statistic value of $t = -0.006/0.0171 = -.3506$ and the corresponding two-tailed p-value for testing if the slope differs from zero is $p = 0.75$.

In other words, there does not appear to be a linear relationship between the ATP level in the sludge and the pH level.

However, the residual plot in the right frame of Figure 4 clearly shows a strong structure that indicates that the straight line simple linear regression model does not fit the data very well in this example.

The residuals are negative for low and high predicted values and positive for intermediate predicted values.

This is a clear example of a case where a zero slope does not necessarily mean there is no relationship between the predictor and the response variables. It is clear from the scatterplot in Figure 4 that there is a relationship between ATP and pH.

However, the relationship appears to be quadratic and not linear. We shall return to this example when we discuss polynomial regression later.

Estimating a Mean Response and Predicting a New Response.

Regression models are often used to predict a new response or estimate a mean response for a given value of the predictor x.

We have seen how to compute a predicted value $\hat{y}$ as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

However, as with parameter estimates, we need a measure of reliability associated with $\hat{y}$.

Returning to the *Diporeia* example, the fitted regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.83914 - 0.0045x,$$

where x is the water depth and y is the weight of the diporeia.

To illustrate, suppose we want to predict the weight of a diporeia at a depth of 40 meters. Then we would simply plug $x = 40$ into the estimated regression equation to get a predicted value of

$$\hat{y} = 0.83914 - 0.0045(40) = 0.65914mg.$$

On the other hand, suppose we wanted to estimate the mean weight of the diporeia found at a depth of 40 meters. For the prediction problem, we want to predict the weight of a single diporeia.

For the estimation problem, we want to estimate the mean of a *conditional* population, i.e. the population of diporeia found at a depth of 40 meters.

In both cases, we shall use $\hat{y} = 0.65914$ as the predicted weight and as the estimate of the mean weight of diporeia found at a depth of 40 meters.

In both cases, we need to associate a measure of uncertainty to these predicted and estimated values since they are based on a random sample from the population.

Estimating a Mean Response and Predicting a New Response continued ...

It is customary to compute confidence intervals for an estimated mean response for a given value of x and a *prediction* interval for a predicted value.

The idea of a prediction interval is to determine an interval that will contain a certain percentage of the population.

Because a prediction interval is attempting to capture a single, random future response as opposed to the mean of the conditional population, the prediction interval must be wider than the confidence interval.

The formulas for a $(1-\alpha)100\%$ confidence interval for a mean response at $x = x_0$ and a $(1-\alpha)100\%$ prediction interval for predicting a new response for $x = x_0$ are given by:

Confidence Interval for a Mean Response: $\hat{y} \pm t_{\alpha/2,n-2}\sqrt{\text{MSE}(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}})}$,

and

Prediction Interval for a New Response: $\hat{y} \pm t_{\alpha/2,n-2}\sqrt{\text{MSE}(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}})}$.

Notice that the only difference between the confidence interval and the prediction interval is that the prediction interval has an extra term of 1 under the radical sign. This extra 1 is needed to account for the fact that we are predicting a random response y at $x = x_0$ as opposed to a fixed population mean.

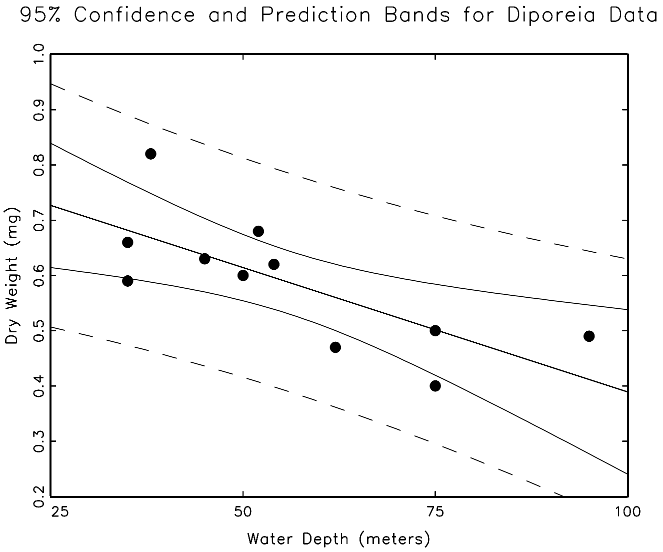

Figure 5: Plot of 95% confidence bands (solid curve) for the estimated mean response and 95% prediction bands (dashed curve) for predicted responses for the diporeia data.

Note that both the confidence interval for the mean response and the prediction interval for a new response are both narrowest at $x_0 = \bar{x}$. Figure 5 shows the estimated regression line for the diporeia data along with 95% confidence bands for estimated mean responses (solid curve) and 95% prediction bands for predicted responses (dashed curve).

Figure 5 illustrates that the confidence bands are narrower than the prediction bands. Note also that all of the observations fall within the 95% prediction bands.

Correlation.

One of the most used statistics in common practice is the correlation between two variables.

Suppose we have data on pairs $(x_1, y_1), \ldots, (x_n, y_n)$. The sample *correlation*, which we shall denote by r, is a measure of the strength of the linear relation between x and y. The formula for the sample correlation is:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \tag{3}$$

Thus, the correlation is the covariance between x and y divided by the standard deviations of x and y.

The sample correlation r is an estimate of the population correlation, which is typically denoted by the Greek letter ρ.

A formal definition of the population correlation requires integral calculus for continuous variables. However, one can think of the population correlation as computed by formula (3) by summing over the entire population of values.

Correlation continued ...

Two properties for the correlation r follow:

1) $-1 \leq r \leq 1$.
2) If $r = \pm 1$, then x and y are perfectly related by a linear transformation, that is, there exists constants a and $b \neq 0$ so that $y = a + bx$.

Property (1) highlights the fact that the correlation is a unitless quantity.

Property (2) highlights the fact that the correlation is a measure of the strength of the *linear* relation between x and y.

A perfect linear relation produces a correlation of 1 or -1. A correlation of zero indicates no linear relation between the two random variables.

Figure 6 shows scatterplots of simulated data obtained from bivariate distributions with different correlations.

The distribution for the top-left panel had a correlation of $\rho = 0.95$. The plot shows a strong positive relation between x and y with the points tightly clustered together in a linear pattern.

The correlation for the top-right panel is also positive with $\rho = 0.50$ and again we see a positive relation between the two variables, but not as strong as in the top-right panel.

The bottom-left panel corresponds to a correlation of $\rho = 0$ and consequently, we see no relationship evident between x and y in this plot.

Finally, the bottom-right panel shows a negative linear relation with a correlation of $\rho = -0.50$.

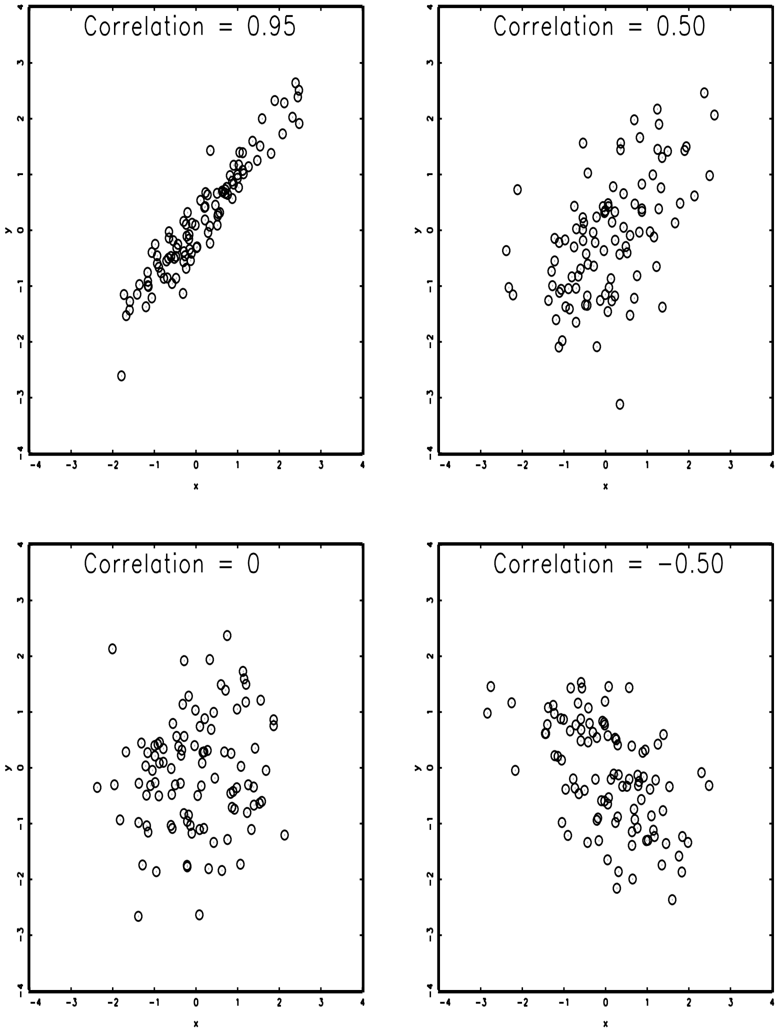

Figure 6: Scatterplots of data obtained from bivariate distributions with different correlations.

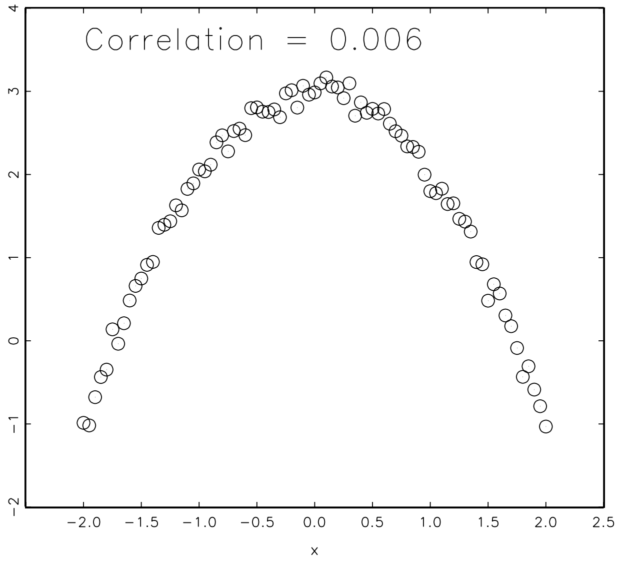

Figure 7: A scatterplot showing a very strong but nonlinear relationship between x and y. The correlation is nearly zero.

A note of caution is in order:

Two variables x and y can be strongly related, but the relation may be nonlinear in which case the correlation may not be a reasonable measure of association.

Figure 7 shows a scatterplot of data from a bivariate distribution. There is clearly a very strong relation between x and y, but the relation is nonlinear. The correlation is not an appropriate measure of association for this data. In fact, the correlation is nearly zero. To say x and y are unrelated because they are uncorrelated can be misleading if the relation is nonlinear. This was the case in the ATP example above.

Here are a couple notes about correlation:

- Correlation does not necessarily imply causation.

  In some examples, a strong correlation is indicative of a causal relationship. For instance, the height y of a plant may be highly correlated with the amount of fertilizer x and it is reasonable in this example to assume that higher levels of fertilizer cause the plant to grow higher. On the other hand, consider an example of the relationship between y, the dollar amount of damage from a fire and x, the number of fire fighters that tried to put the fire out. There will almost surely be a positive correlation in this example but it would be silly to conclude that sending more fire fighters to a fire will increase the dollar amount of the damage. The reason one sees more damage at fires where there were more fire fighters is that big fires require more fire fighters and big fires cause more damage.

- In a simple linear regression, the slope is zero if and only if the correlation is zero. Thus, one can test for a correlation of zero by simply testing if the slope of the regression line is zero. If we want to test if the population correlation ρ takes a value other than zero, then the testing procedure is a bit more complicated (e.g. see Montgomery and Peck 1992, page 56).

We shall finish our discussion of the simple linear regression model with some cautionary notes on the use of regression models and a famous simulated example:

Example: Anscombe's Regression Data. Anscombe (1973) simulated 4 very different data sets that each produce an identical least-square regression line. One of the benefits of this example is to illustrate the importance of plotting your data. Figure 8 shows scatterplots of the 4 data sets along with the fitted regression line. The top-left panel shows a nice scatter of points with a linear trend and the regression line provides a nice fit to the data. The data in the top-right panel shows a very distinct non-linear pattern. Although one can fit a straight line to such data, the straight line model is clearly wrong. Instead one could try to fit a quadratic curve (see polynomial regression). The points in the bottom left plot all lie in a line except a single point. The least squares regression line is pulled towards this single outlying point. In a simple linear regression it is fairly easy to detect a highly unusual point as in this plot. However, in multiple regression (see next section) with several regressor variables, it can be difficult to detect extreme points graphically. There exist many diagnostic tools for accessing how influential individual points are when fitting a model. There also exist *robust* regression techniques that prevent the fit of the line to be unduly influenced by a small number of observations. The bottom-right panel shows data from a very poorly designed experiment where all but one observation was obtained at one level of the x variable. The single point on the right determines the slope of the fitted regression line. The bottom-right panel demonstrates how a single point can be very *influential* when a least-squares line is fit to the data.

Cautionary Notes on the Use of Linear Regression.

Prediction. It is dangerous to use a regression model to predict responses outside the range of data that was used to predict the model.

We may see a nice linear relationship between the response y and the predictor x in our data, but we may not know if this linear relationship persists for values of x and y outside the range we have observed.

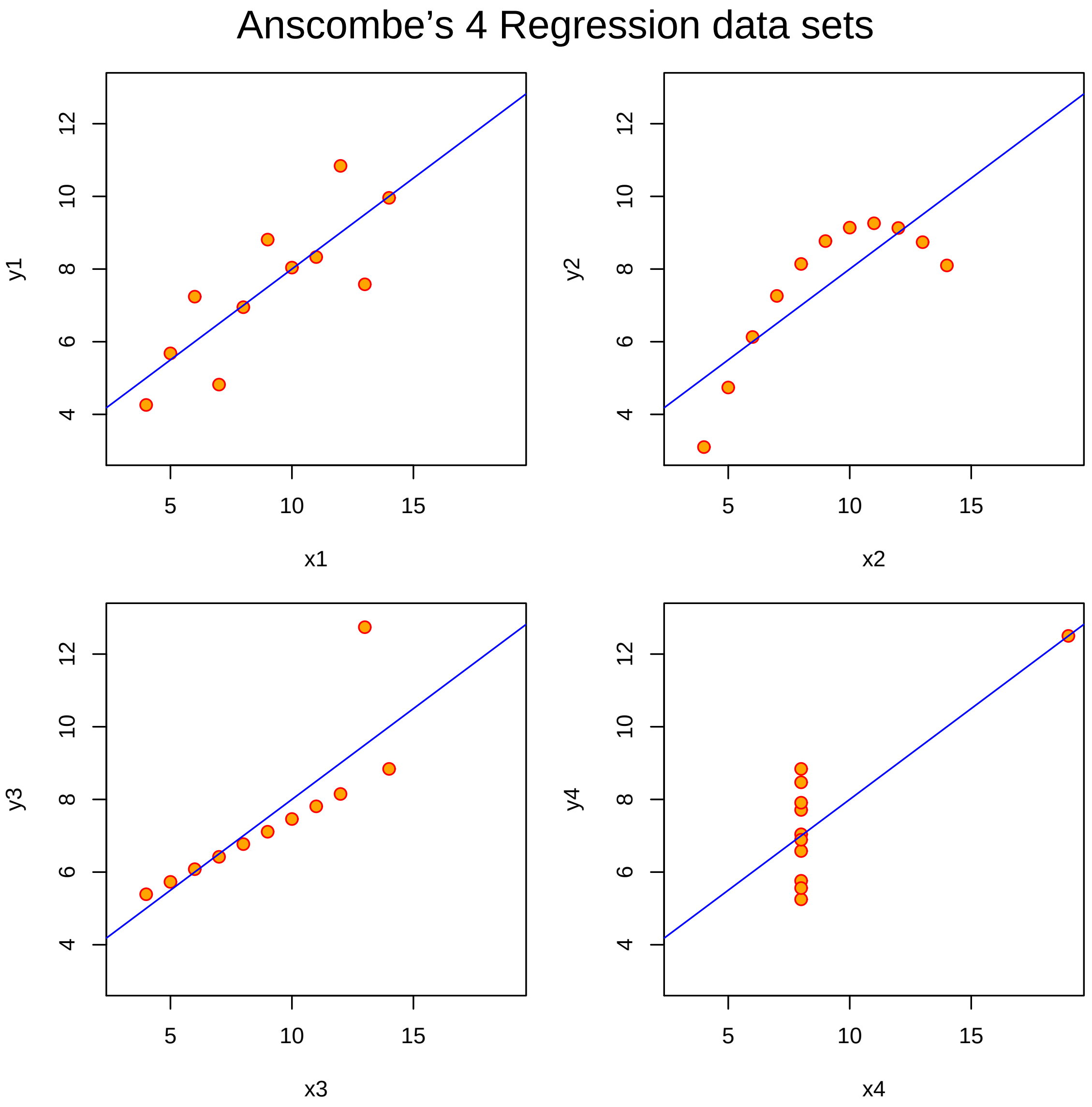

Figure 8: Anscombe simple linear regression data. Four very different data sets yielding exactly the same least squares regression line.

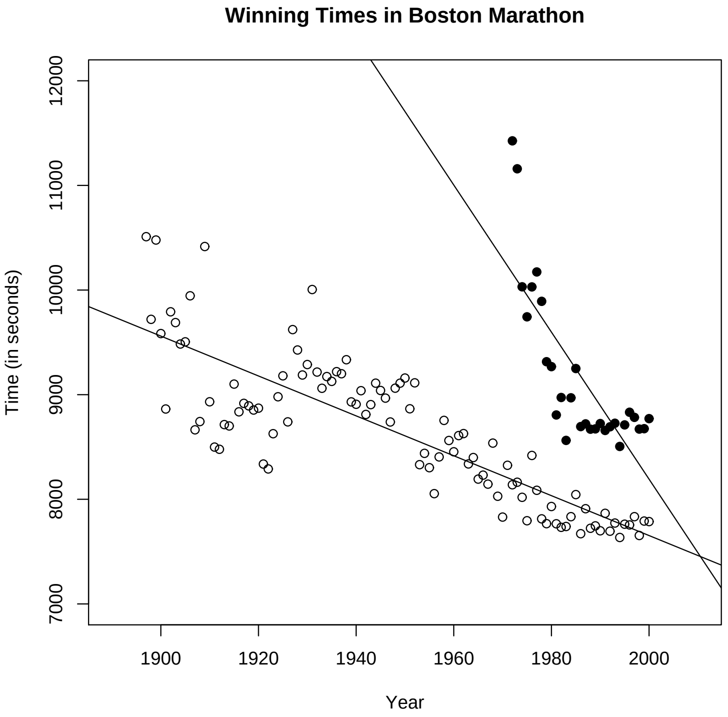

Figure 9: Winning times in the Boston Marathon versus year for men (open circles) and women (solid circles). Also plotted are the least-squares regression lines for the men and women champions.

A stark illustration of this danger is shown in Figure 9.

This plot shows the winning times for the men and women runners in the Boston marathon over the last century (note that women were not allowed into the race until the mid-70's). From the plot we see that there was a dramatic improvement in the winning times for women in the first several years and then the rate of improvement levels off. If the rate of improvement changes, then a linear function is not appropriate. The estimated regression lines are extended beyond the range of the data and we see that the lines eventually cross. If we were to use the models to extrapolate, then we would conclude that top woman runner will eventually beat the top male runner. This may happen, but we cannot use these regression models to make this claim because the models are not valid for predicting future winning times far out into the future. In fact, suppose we extrapolated further into the future – then eventually both lines would cross the x-axis giving negative winning times in the marathon! This is clearly impossible.

The appropriate model for the Boston marathon data probably requires a horizontal asymptote for both the men and women's data. Recalling George Box's quote, the straight line model is wrong. However, even for complex nonlinear relationships, a straight line model may provide a very good approximation over a limited range of the regressor variable.

Influential Points

The least-squares regression line can be highly influenced by one or more points. For instance, in the bottom right frame of Figure 8, there is a single point determining the slope of the regression line.

This is a very unstable situation and very undesirable. If we moved that point down, then the slope would change from positive to negative.

Because the estimated regression line can be strongly influenced by just a few points, it is very important to plot your data to determine if this could be a problem. A residual plot is very useful for identifying some outliers – if a point has a large residual, then there could be problem with that point (e.g. maybe its value was recorded incorrectly).

However, note that the influential point in the bottom left frame of Figure 8 will have residual of zero. This illustrates that the residual plot by itself will not necessarily identify all problematic points from your data. Influential points have the tendency to *pull the regression line* towards themselves which in turn makes their residuals small.

There are numerous regression diagnostic tools for identifying influential points (see for example Montgomery and Peck (1992)).

In a simple linear regression, a simple plot of the data will often suffice for identifying problematic points, but in multiple regression (next section), the data is higher dimensional and thus it is more difficult to find influential points graphically.

One simple and popular idea is the *jackknife* idea. The idea behind the jackknife is to leave out a single observation and fit the regression equation with out this single point and see how much the estimated coefficients change. Now repeat this process for all the data points and see which data points cause a big change in the fit of the model when left out.

There also exist *robust* regression fitting techniques that were developed in the hopes that the resulting regression line will not be unduly influenced by a small number of points.

Multiple Regression Models

The simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

with a single predictor variable can be easily generalized to handle models with $p > 1$ predictor variables $x_1, \ldots, x_p$ by

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon. \tag{4}$$

In order perform statistical inference using a multiple regression model, several assumptions are made. We require that the n observations are independent and that the variability of the error ϵ is constant for all values of the regressors.

In addition, many of the statistical tests require that the error distribution is normal.

If the error distribution deviates somewhat from normality, the inference procedures will remain approximately valid.

The slope parameters β_j are estimated using least-squares, as was done in simple linear regression, by determining the values of the β_j's that minimize

$$\text{SSE} = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\}^2.$$

Here, the index i represents the observation number, $i = 1, \ldots, n$.

The formulas for obtaining the least-squares estimators are derived from multivariate calculus and and closed form formulas are obtained using linear algebra.

The calculations for determining the estimated slope parameters are built into most statistical software packages.

ANOVA in Multiple Regression

Similar to an Analysis of Variance (ANOVA), the total variability in the response variable y is partitioned into an error component and the component that is explained by the regressor variables.

The total sum of squares is

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

where $\bar{y}$ is the overall mean value of the response variable.

Writing

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

shows how to partition the variability in the response into the part due to the regression and the part due to error. The $r_i = (y_i - \hat{y}_i)$ is just the ith residual. If we square both sides of the previous equation and sum over all $i = 1, \ldots, n$ observations, we get

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE},$$

where SSR stands for the *regression sum of squares* and is given by

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

and the error sum of squares (SSE) is

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The statistic most often reported in the multiple regression examples is probably the R^2 ("R-squared") which represents the proportion of variability in the response explained by the regressor variables. R^2 is known as the *coefficient of determination* and is defined as

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}.$$

Equivalently, one can write

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

- By definition, R^2 can take values only between 0 and 1.
- If $R^2 = 1$, then all the points lie exactly on a plane and SSE $= 0$.

Often one would like the regression model to explain a large proportion of the variability in the response y.

However, the criteria for what is considered a large R^2 varies across different fields of study.

In many engineering applications, high R^2's will be in the 90% or higher range whereas in many social science applications, an $R^2 = 0.5$ may be considered high.

There are some **dangers** on relying on R^2 to heavily when interpreting multiple regression output.

- For instance, a high R^2 in a simple linear regression model does not necessarily mean the relationship between x and y is linear.
- Also, if the model is to be used for prediction, a high R^2 will not necessarily indicate that the model can provide predictions with the needed precision.
- Nonetheless, R^2's are reported very frequently in regression settings.
- R^2's can also be useful when attempting to determine a parsimonious model for predicting y from several regressor variables – various models can be compared in terms of R^2 where models with higher R^2 may be preferable.

One very important point to note is that R^2 can always be made larger (or at least not smaller) by including more regressor variables.

Therefore, picking a model based on the largest R^2 is a bad idea because it will always lead to the most complicated model.

If a simpler model (i.e. fewer regressor variables) explains practically the same amount of variability as a more complicated model, then the simpler model is to be preferred.

ANOVA Table in Multiple Regression

As in an ANOVA, an F-test can be carried out by comparing the variability explained by the regression relationship (SSR) with the error variability SSE. Formally, the F-test is a test of the null hypothesis:

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

verus

$$H_a : \text{ not all } \beta_j \text{ are zero.}$$

- If H_0 is true, then the observed variability in y due to the regressors should be comparable to the error variability.
- If H_0 is false, then the variability due to the regressors will exceed the error variability.

In order to compare these two sources of variability, the respective sum of squares are normalized by dividing each by their respective degrees of freedom which yields the mean squares. The degrees of freedom for SSR is p for the p regressor variables and the degrees of freedom for the error SSE is $n-p-1$. Thus,

$$\text{Mean Square for Regression} = \text{MSR} = \frac{\text{SSR}}{p},$$

and

$$\text{Mean Square Error} = \text{MSE} = \frac{\text{SSE}}{n-p-1}.$$

F-test

In order to test H_0, an F-test statistic is computed as

$$F = \frac{\text{MSR}}{\text{MSE}}.$$

H_0 is rejected at level of significance α if F exceeds the α critical value of the F-distribution on p numerator degrees of freedom and $n-p-1$ denominator degrees of freedom.

Typically, statistical software packages will simply report the p-value of this test:

$$p - \text{value} = P(F > \text{observed } F \text{ value}),$$

where F in the probability statement represents an F random variable on p and $n-p-1$ numerator and denominator degrees of freedom.

If H_0 is rejected, then one can conclude that at least one of the regression coefficients is nonzero. If a regression coefficient $\beta_j \neq 0$, then it seems reasonable to assume that the corresponding regressor x_j influences the response y. However, care must be taken when interpreting regression coefficients in a multiple regression framework as we shall see.

If the overall F test just described rejects H_0, then a natural question to ask is:

> which β_j's differ significantly from zero?

To answer this question, t-test statistics are computed individually for the hypotheses

$$H_0 : \beta_j = 0,\ j = 1, \ldots, p$$

using

$$t_j = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)},$$

which is to be compared to the t critical value on $n - p - 1$ degrees of freedom.

Typically software programs will report the value of the estimated parameter, the estimated standard error, the t-test statistic for each of the β_j's along with the corresponding p-values.

The alternative hypothesis concerning β_j can be one- or two-sided.

In a multiple regression setup, the formulas for the estimated standard errors $\hat{se}(\hat{\beta}_j)$ are quite complicated to write out without the aid of matrix notation.

Confidence intervals for individual slope parameters can also be computed using the estimate and its standard error, as was done in the simple linear regression setup.

Confidence Ellipsoids.

The estimated slope parameters $\hat{\beta}_j$'s are random variables (or realizations of random variables if the data is already collected). Unless a carefully designed experiment is performed, the estimated slope parameters will be correlated with one another.

If the coefficient estimators are correlated, then computing confidence intervals for individual coefficients can be misleading. For instance, when considering confidence intervals for coefficients β_1 and β_2 jointly, the joint region determined by two intervals is their **Cartesian product** which forms a rectangle.

However, if $\hat{\beta}_1$ and $\hat{\beta}_2$ are correlated, then their values tend to vary jointly in an *elliptical* region in most cases. This suggests that an **elliptical confidence region** should be used for (β_1, β_2) instead of a rectangular region. The details of elliptical confidence regions can be found in many multivariate statistics textbooks.

Multiplicity

Another point of importance is that if confidence intervals (or hypothesis tests) are considered for more than one slope coefficient, some sort of multiplicity correction should be considered, like the Bonferroni correction.

Partial F-tests.

In multiple regression, we have seen how to use an F-test to test the null hypothesis that all the regression coefficients are zero versus the alternative that at least one of the coefficients differs from zero.

One of the important principals in estimating statistical models is that the simplest model that adequately explains the data should be used. This is sometimes known as the *principal of parsimony.* If a regression model has redundant regressors, they should be dropped out of the model. We can use an F-test for this purpose to test if a subset of regression coefficients are all zero.

Partial F-tests continued...

Here is the setup: consider the usual multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_p x_p + \epsilon,$$

where $q < p$. Our interest now is to test the null hypothesis

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0,$$

versus H_a : that at least one of the $\beta_j \neq 0, q+1 \leq j \leq p$. If H_0 is true, then the model reduces to the *reduced model*:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \epsilon \qquad \text{(REDUCED MODEL)}.$$

In other words, we want to test if we can safely drop the regressors x_{q+1} to x_p from the model.

If these regressors do not add significantly to the model, then dropping them will make the model simpler and thus preferable.

Partial F-tests continued...

An F-test can be used to test the hypothesis that some of the regressors can be dropped from the model.

We call the model with all the regressors the FULL model and the model with only regressors $x_1, \ldots, x_q$, the REDUCED model.

The F-test statistic is computed as

$$F = \frac{(\text{SSR(Full) - SSR(Reduced)})/(p-q)}{\text{MSE(full)}}. \tag{5}$$

If the null hypothesis is true, then this F-test statistic follows an F distribution on $p-q$ numerator degrees of freedom and $n-p-1$ denominator degrees of freedom.

SSR(Full) is the regression sum of squares from the full model and SSR(Reduced) is the regression sum of squares from the reduced model.

The denominator of this F-statistic is the MSE from the full model.

In order to perform this test in practice, simply run the full and reduced models and plug the resulting sum of squares and MSE into the F-test statistic formula given in (5).

Example. The next example illustrates the partial F testing procedure.

Mercury contamination is a serious health problem for humans. A study of mercury contamination in Largemouth bass from 53 different Florida lakes was undertaken to examine the factors that are associated with the level of mercury contamination.

Water samples were collected from the surface of the middle of each lake in the early 1990's.

The pH level, the amount of chlorophyll, calcium, and alkalinity were measured in each sample. From fish samples, the average mercury concentration in a three year old fish was estimated for each lake.

(This data was made available on the Data and Story Library (DASL) found at http://lib.stat.cmu.edu/DASL/DataArchive.html.)

The data are below:

Observation	alkalinity	ph	calcium	chlorophyll	mercury
1	5.9	6.1	3.0	0.7	1.53
2	3.5	5.1	1.9	3.2	1.33
3	116.0	9.1	44.1	128.3	0.04
4	39.4	6.9	16.4	3.5	0.44
5	2.5	4.6	2.9	1.8	1.33
6	19.6	7.3	4.5	44.1	0.25
7	5.2	5.4	2.8	3.4	0.45
8	71.4	8.1	55.2	33.7	0.16
9	26.4	5.8	9.2	1.6	0.72
10	4.8	6.4	4.6	22.5	0.81
11	6.6	5.4	2.7	14.9	0.71
12	16.5	7.2	13.8	4.0	0.51
13	25.4	7.2	25.2	11.6	0.54
14	7.1	5.8	5.2	5.8	1.00
15	128.0	7.6	86.5	71.1	0.05
16	83.7	8.2	66.5	78.6	0.15
17	108.5	8.7	35.6	80.1	0.19
18	61.3	7.8	57.4	13.9	0.49
19	6.4	5.8	4.0	4.6	1.02
20	31.0	6.7	15.0	17.0	0.70
21	7.5	4.4	2.0	9.6	0.45
22	17.3	6.7	10.7	9.5	0.59
23	12.6	6.1	3.7	21.0	0.41
24	7.0	6.9	6.3	32.1	0.81
25	10.5	5.5	6.3	1.6	0.42
26	30.0	6.9	13.9	21.5	0.53
27	55.4	7.3	15.9	24.7	0.31
28	3.9	4.5	3.3	7.0	0.87
29	5.5	4.8	1.7	14.8	0.50
30	6.3	5.8	3.3	0.7	0.47
31	67.0	7.8	58.6	43.8	0.25
32	28.8	7.4	10.2	32.7	0.41
33	5.8	3.6	1.6	3.2	0.87
34	4.5	4.4	1.1	3.2	0.56
35	119.1	7.9	38.4	16.1	0.16
36	25.4	7.1	8.8	45.2	0.16
37	106.5	6.8	90.7	16.5	0.23
38	53.0	8.4	45.6	152.4	0.04
39	8.5	7.0	2.5	12.8	0.56
40	87.6	7.5	85.5	20.1	0.89
41	114.0	7.0	72.6	6.4	0.18
42	97.5	6.8	45.5	6.2	0.19
43	11.8	5.9	24.2	1.6	0.44
44	66.5	8.3	26.0	68.2	0.16
45	16.0	6.7	41.2	24.1	0.67
46	5.0	6.2	23.6	9.6	0.55
47	25.6	6.2	12.6	27.7	0.58

```
48          81.5       8.9       20.5            9.6         0.27
49           1.2       4.3        2.1            6.4         0.98
50          34.0       7.0       13.1            4.6         0.31
51          15.5       6.9        5.2           16.5         0.43
52          17.3       5.2        3.0            2.6         0.28
53          71.8       7.9       20.5            8.8         0.25
```

All the variables measured (except pH) are quite strongly skewed to the right, so the log-transformed variables were used in the analysis.

The following full model was fit to the data:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon,$$

where x_1 to x_4 are the logarithms of alkalinity (mg/L as Calcium Carbonate), pH, calcium (mg/l), and chlorophyll (mg/l) respectively. The response y is the average mercury concentration in a three-year old fish (parts per million).

The SAS output from fitting the full model follows:

```
                          The REG Procedure
                            Model: MODEL1
                      Dependent Variable: lmer

                        Analysis of Variance

                                Sum of          Mean
Source                 DF      Squares        Square   F Value   Pr > F

Model                   4     21.32066       5.33016     19.14   <.0001
Error                  48     13.36713       0.27848
Corrected Total        52     34.68778


           Root MSE              0.52771    R-Square     0.6146
           Dependent Mean       -0.92734    Adj R-Sq     0.5825
           Coeff Var           -56.90592

                        Parameter Estimates

                        Parameter       Standard
  Variable     DF        Estimate          Error    t Value    Pr > |t|

  Intercept     1         0.71590        0.48649       1.47      0.1477
  lalk          1        -0.47148        0.13673      -3.45      0.0012
  ph            1         0.00349        0.11574       0.03      0.9761
  lcalcium      1         0.10448        0.11809       0.88      0.3807
  lchlor        1        -0.20770        0.07530      -2.76      0.0082
```

In this output, “lalk” stands for log(alkalinity), “lcalcium” stands for log(calcium) etc.

Note that the overall F-test statistic if $F = 19.14$ with corresponding p-value $p < 0.0001$ giving strong evidence that at least one of the slope coefficients is non-zero.

However, from the partial t-test statistics for testing if individual coefficients differ from zero or not, we see that the coefficient of pH and log(calcium) give p-values of $p = 0.9761$ and 0.3807 respectively.

Mercury example continued ...

We can use a partial F-test to test $H_0 : \beta_2 = \beta_3 = 0$ versus the alternative that at least one of these coefficients differs from zero. To perform this test, we run a reduced model using only log(alkalinity) and log(chlorophyll) which is given in the following SAS output:

```
                              The REG Procedure
                                Model: MODEL1
                          Dependent Variable: lmer

                            Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     2       21.08907       10.54454      38.77    <.0001
Error                    50       13.59871        0.27197
Corrected Total          52       34.68778


           Root MSE              0.52151    R-Square     0.6080
           Dependent Mean       -0.92734    Adj R-Sq     0.5923
           Coeff Var           -56.23710
                             Parameter Estimates

                         Parameter       Standard
  Variable      DF        Estimate          Error     t Value     Pr > |t|

  Intercept      1         0.69449        0.19809        3.51       0.0010
  lalk           1        -0.37877        0.07002       -5.41       <.0001
  lchlor         1        -0.20049        0.06688       -3.00       0.0042
```

The regression sum of squares for the full and reduced models are 21.32066 and 21.08907 respectively.

The partial F-test statistic is computed as

$$F = \frac{(21.32066 - 21.08907)/(4-2)}{0.27848} = 0.4158$$

where the denominator 0.27848 is the MSE from the full model with all four regressors.

The p-value for this test is $p = 0.6621$ indicating that the coefficients for pH and log(calcium) do not differ significantly from zero.

This suggests that we can drop the regressors pH and log(calcium) from our model and use the estimated model:

$$\hat{y} = 0.69449 - 0.37877\text{log(alkalinity)} - 0.20049\text{log(chlorophyll)}.$$

Note that the R^2 from the reduced model is 0.6080 which is only slightly less than the R^2 from the full model ($R^2 = 0.6146$).

Coefficient Interpretation in Multiple Regression.

Unlike the simple linear regression model, the modeling strategy for multiple regression is often much more complex.

One of the reasons for the added complexity is that the regression variables (the predictors) are often correlated with one another.

In a simple linear regression, the slope of the model corresponds to the average rate of change in the response variable corresponding to changes in the predictor variable.

The same interpretation exists in the multiple regression setting with one important caveat: the slope β_j for the jth regressor variable x_j represents the average change in the the response y for a unit change in x_j *provided all other regressor variables are held fixed.*

The problem with slope interpretation in multiple regression is that if one regressor changes, then other regressor variables tend to change as well when the regressors are correlated.

Therefore, it is very difficult to access the effect of a regressor on the response by examining an individual slope coefficient β_j by itself.

Model Fitting in Multiple Regression

Often the problem of fitting a multiple linear regression model is a trial and error process.

One may begin by postulating a model only to find that it is inadequate because some of the estimated slope coefficients are unstable (i.e. large p-values when testing if the coefficient is zero) and/or because the residual plots show structure indicating the model is not appropriate.

This next example illustrates some of the trial-and-error approach to model fitting.

Example. The ecological balance of severely nutrient-limited areas such as the Florida Everglades can be jeopardized due to anthropogenic nutrient enrichment.

In order to detect and monitor the presence of nutrient contamination in the Everglades, a study was conducted to examine variation in plant morphology that results from soil characteristics.

In particular, the negative effects of phosphorus enrichment in the Everglades are a concern.

Determining the effect of soil characteristics on the size and/or the shape of plants can be valuable to Everglades managers.

Motivation for this model comes from a study of the plant *Sagittaria lancifolia* which is common throughout the wetlands of the southeastern United States.

Survey sites were randomly located throughout the Florida Everglades and a sample of plants and soil measurements were obtained at each site.

Sagittaria lancifolia Example continued ...

Two regressor variables related to the soil were considered: x_1 = total phosphorus in soil, and x_2 = percent ash-free dry weight of soil, that indicates the amount of organic matter in the soil which can affect accessibility of nutrients to plants.

For this example, the response variable will be y = leaf-width.

Sagittaria lancifolia **example continued ...**

The SAS code for running the multiple regression model

$$y = \beta_0 + \beta_1 \text{PHOS} + \beta_2 \text{ASH} + \epsilon,$$

is:

```
proc reg;
     model width=phos ash;
     output out=a p=yhat r=res;
run;
```

The output statement tells SAS to create an internal data set (called "a") which will store the predicted values (called "yhat") as well as the residuals (called "res").

Sagittaria lancifolia **example continued ...**

The SAS output from proc reg using total phosphorus and ash free dry weight is given below:

```
                           The REG Procedure
                              Model: MODEL1
                        Dependent Variable: width

                          Analysis of Variance

                                  Sum of          Mean
Source                  DF        Squares        Square    F Value    Pr > F

Model                    2          23174         11587      34.14    <.0001
Error                  284          96375     339.34742
Corrected Total        286         119548


           Root MSE             18.42138    R-Square     0.1938
           Dependent Mean       22.93833    Adj R-Sq     0.1882
           Coeff Var            80.30832


                           Parameter Estimates

                        Parameter       Standard
   Variable     DF       Estimate          Error    t Value    Pr > |t|

   Intercept     1        3.53875        3.54275       1.00      0.3187
   phos          1        0.05259        0.00734       7.16      <.0001
   ash           1        0.05137        0.05192       0.99      0.3233
```

- From the ANOVA table at the top of the output, we see the total degrees of freedom is 286 indicating that there were $n = 287$ observations in this data set.

- The sum of squares for regression and error are 23174 and 96375 respectively.

- Dividing these sum of squares by their respective degrees of freedom ($p = 2$ and $n - p - 1 = 284$) gives the mean squares.

- The observed F-test statistic is $F = 34.14$.

- The probability of observing an F value this big or bigger when $\beta_1 = \beta_2 = 0$ is practically zero ($p < 0.0001$).

- Therefore, we reject the null hypothesis and conclude that either phosphorus and/or ash free weight have an effect on the width of the leafs.

Note that $R^2 = 0.1938$. Thus, phosphorus and ash free dry weight of the soil explain about 20% of the variability in the leaf widths.

Sagittaria lancifolia **example continued ...**

As in a simple linear regression, it is important to access the fit of the model by examining residual plots.

In a multiple regression setting, it is not as easy to do a residual analysis due to the different regressor variables.

Typically, one will want to look at plots of the residuals versus the fitted values as well as residuals versus each of the individual repressor variables.

If problems show up in these plots such as indications of nonlinearities or non-constant variance, then the analyst should investigate adding higher order polynomial terms (see polynomial regression) and/or transformations of the variables.

Sagittaria lancifolia **example continued ...**

From the parameter estimates part of the output, we see that the estimated regression equation is

$$\hat{y} = 3.539 + 0.052\text{PHOS} + 0.051\text{ASH}.$$

Both of these estimated coefficients are positive which appears to indicate that by increasing the amount of phosphorus and percentage ash free dry weight of the soil will lead to larger leaf widths.

However, from the t-test statistics to test if β_1 and β_2 differ significantly from zero, we find that only phosphorus appears significant. The t-test statistic for phosphorus is

$$t_1 = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} = \frac{0.05259}{0.00734} = 7.16.$$

The probability that a t random variable on 284 degrees of freedom (the error d.f.) takes a value of 7.16 or larger in magnitude is very small ($p < 0.0001$).

In other words, we have strong statistical evidence that $\beta_1 \neq 0$ and consequently that the leaf width is effected by phosphorus.

On the other hand, the t-test statistic for ash free dry weight is $t = 0.05137/0.05192 = 0.99$ indicating that this coefficient is very unstable. The p-value is $p = 0.3233$ indicating that β_2 does not differ significantly from zero.

***Sagittaria lancifolia* example continued ...**

At this point, it is tempting to state that the width of the *Sagittaria lancifolia* plants are not effected by the ash free dry weight of the soil since the coefficient for ash-free dry weight does not differ significantly from zero. However, if we run a regression using only ash-free dry weight as a regressor, we get the following results:

```
                           The REG Procedure
                             Model: MODEL2
                       Dependent Variable: width

                         Analysis of Variance

                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F

Model                     1      5764.51085     5764.51085      14.44    0.0002
Error                   285          113784      399.24136
Corrected Total         286          119548


           Root MSE             19.98102    R-Square     0.0482
           Dependent Mean       22.93833    Adj R-Sq     0.0449
           Coeff Var            87.10759


                          Parameter Estimates

                        Parameter       Standard
   Variable     DF       Estimate          Error    t Value    Pr > |t|

   Intercept     1        9.46669        3.73637       2.53      0.0118
   ash           1        0.19694        0.05183       3.80      0.0002
```

Sagittaria lancifolia example continued ...

Note that the R^2 has dropped to $R^2 = 0.0482$ using only ash-free dry weight as a regressor from $R^2 = 0.1938$ from the model using both ash-free dry weight and phosphorus as regressors.

Nonetheless, the coefficient for ash-free dry weight (0.19694) is highly significant ($p = 0.0002$) even though the coefficient did not differ significantly from zero in the model that also included phosphorus as a regressor.

***Sagittaria lancifolia* example continued ...**

Thus, ash-free dry weight *does* have an impact on the leaf width.

How then do we explain this apparent paradox?

The answer is that in the model containing both phosphorus and ash-free dry weight as regressors, ash-free dry weight does not provide any significant additional information about the response given that the phosphorus variable is in the model.

In other words, in the model containing both phosphorus and ash-free dry weight as regressors, ash-free dry weight is essentially a redundant variable.

The following is the SAS output from proc reg using only phosphorus as a regressor:

```
                          Analysis of Variance

                                  Sum of          Mean
Source                  DF       Squares        Square   F Value   Pr > F

Model                    1         22841         22841     67.31   <.0001
Error                  285         96707     339.32224
Corrected Total        286        119548


           Root MSE             18.42070    R-Square     0.1911
           Dependent Mean       22.93833    Adj R-Sq     0.1882
           Coeff Var            80.30534


                           Parameter Estimates

                        Parameter       Standard
   Variable     DF       Estimate          Error    t Value    Pr > |t|

   Intercept     1        6.19413        2.31243       2.68      0.0078
   phos          1        0.05543        0.00676       8.20      <.0001
```

Sagittaria lancifolia **example continued ...**

Again, phosphorus is highly significant ($p < 0.0001$) and the proportion of variability in leaf widths explained by phosphorus is $R^2 = 0.1911$ which is only slightly less than the R^2 using both phosphorus and ash-free dry weight as regressors ($R^2 = 0.1938$).

From this discussion, it appears as if we can model leaf width using only the phosphorus regressor.

Sagittaria lancifolia **example continued ...**

Before we settle on the above regression model, we need to check that the model is correctly specified. Figure 10 shows a plot of residuals versus predicted values for the full model of width regressed on phosphorus and ash-free dry weight.

The residual plot does not show a random scatter of points.

In particular, it appears that the variability in the error increases with increases leaf widths and this is a problem (a similar looking residual plot is obtained using only phosphorus as a regressor).

Part of the problem lies in the leaf width distribution.

Figure 11 shows a histogram of the leaf widths from this example that illustrates a distribution that is strongly skewed to the right and non-normal.

Skewness is not unusual for data of this sort.

In such cases, a logarithm transformation will often yield a distribution that is more consistent with a normal distribution.

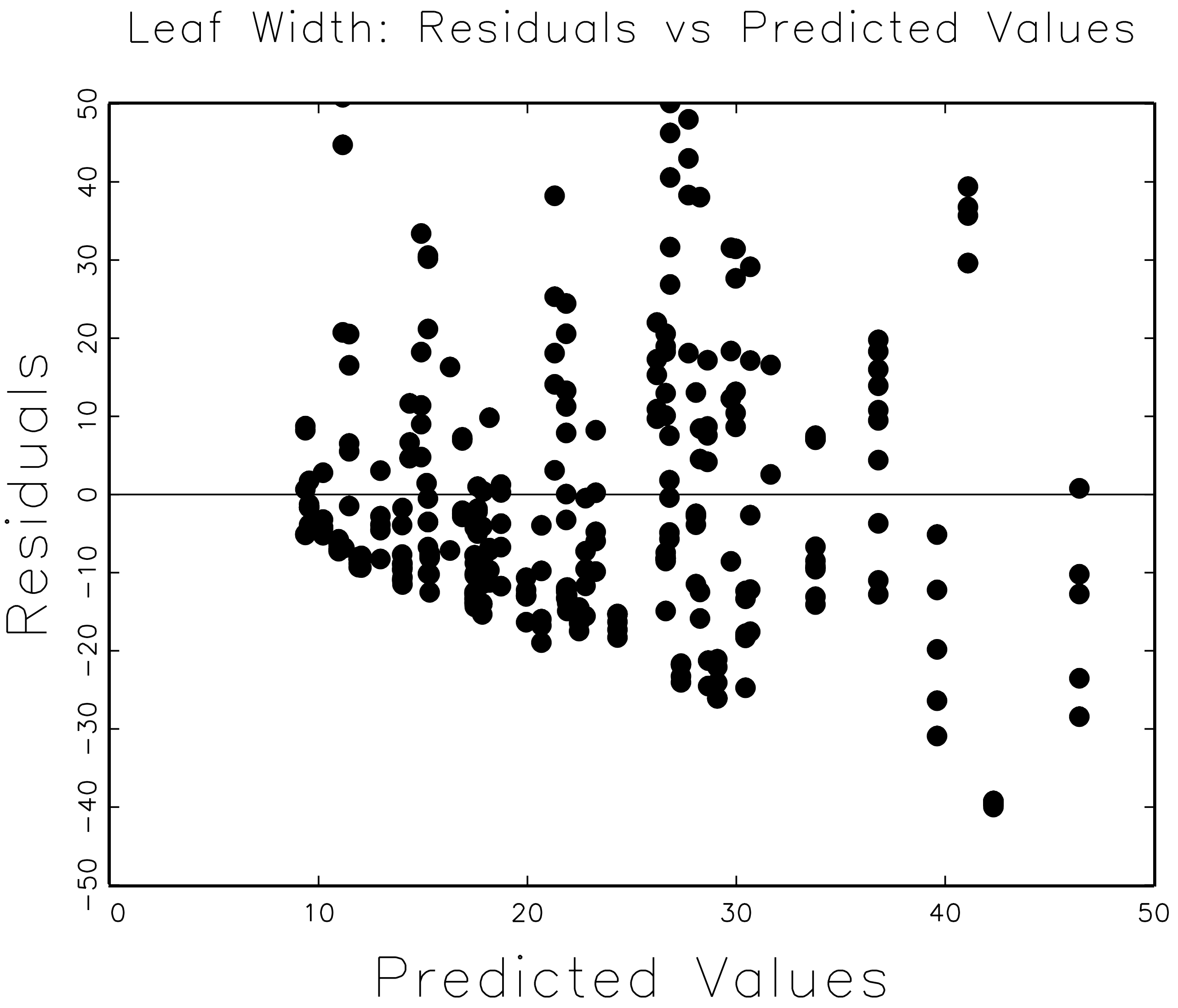

Figure 10: Residual plot from the model Leaf Width $= \beta_0 + \beta_1\text{PHOS} + \beta_2\text{ASH} + \epsilon$.

Sagittaria lancifolia **example continued ...**

We shall now fit a regression of log(leaf width) versus phosphorus and ash-free dry weight.

Figure 13 shows a 3-dimensional plot of log(leaf width) versus phosphorus and ash-free dry weight as well as plots of ln(width) versus phosphorus and ash-free dry weight individually.

These plots in Figure 13 indicate that the relation may not be linear, but that additional terms involving phosphorus and ash-free weight raised to higher powers may be needed.

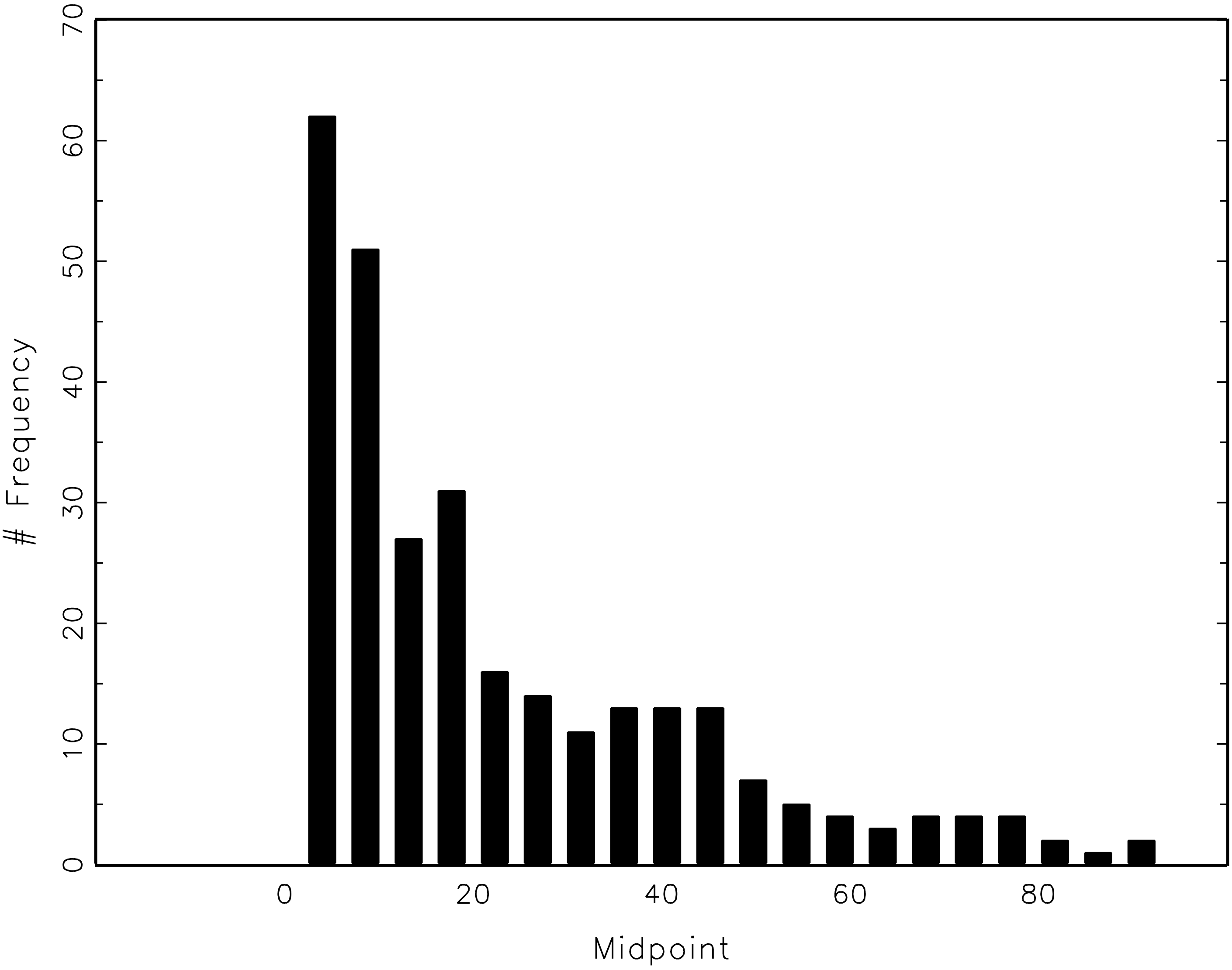

Figure 11: Histogram of the *Sagittaria lancifolia* leaf widths indicating a strongly skewed distribution.

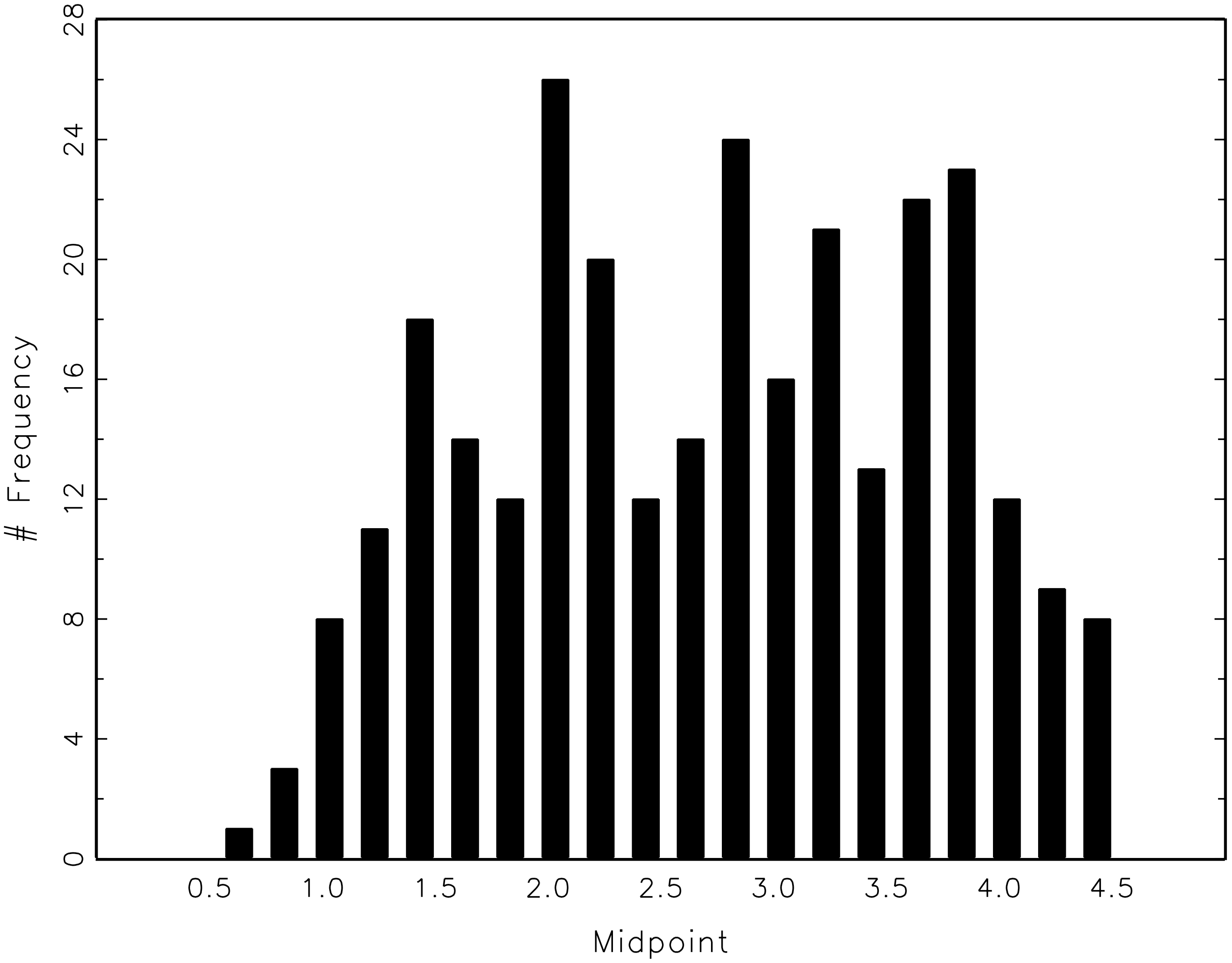

Figure 12: Histogram of the *Sagittaria lancifolia* log(leaf widths) that appears roughly symmetric.

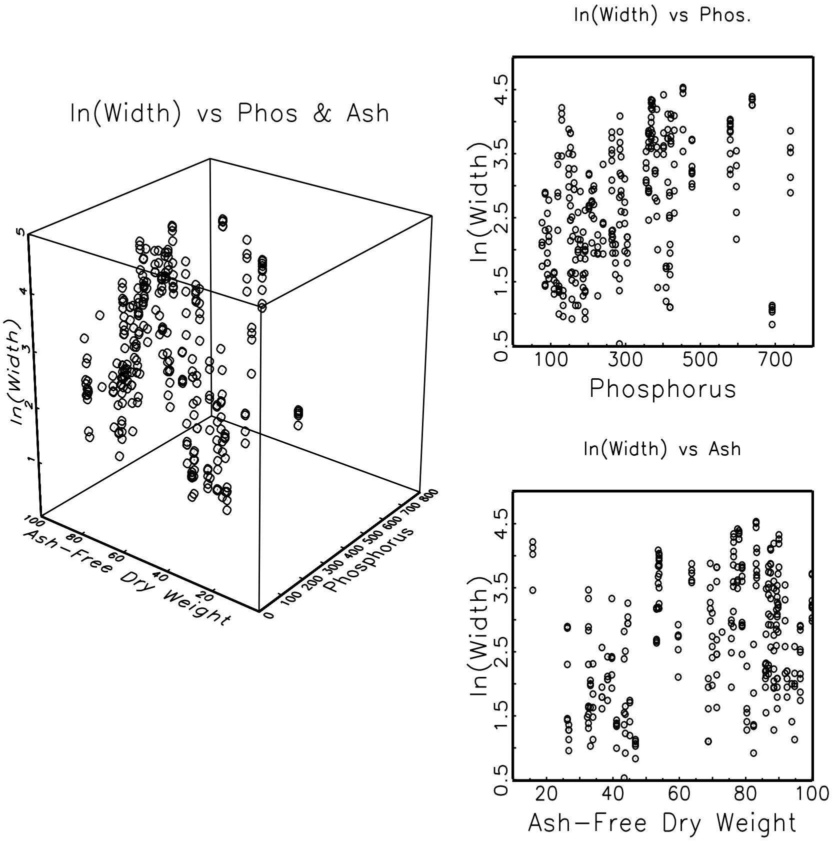

Figure 13: 3-dimensional plot of ln(Width) versus phosphorus and ash-free dry weight in the left frame and the two frames on the right show scatterplots of ln(width) versus phosphorus (top) and ln(width) vs ash-free dry weight (bottom)

Sagittaria lancifolia **example continued ...**

A second order multiple regression model with terms $\text{PHOS}^2, \text{ASH}^2$, and an interaction term $\text{PHOS} \times \text{ASH}$ was fit to the data which yielded an $R^2 = 0.2962$.

The residual plot from this fit shown in Figure 14 looks pretty good except that the predicted values appear to be too small for short widths and too big for large widths.

If structure of this sort is evident in a residual plot, then that suggests that even higher order terms may be needed.

Sagittaria lancifolia example continued ...

Cubic Model The following model with cubic terms was considered where x_1 denotes phosphorus and x_2 denotes ash-free dry weight.

$$\log(\text{leaf width}) = \beta_0+\beta_1 x_1+\beta_2 x_1^2+\beta_3 x_1^3+\beta_4 x_2+\beta_5 x_2^2+\beta_6 x_2^3+\beta_6 x_1 x_2+\epsilon, \tag{6}$$

There does not appear to be any structure in the residual plot for the cubic model (6) shown in Figure 15 indicating that the cubic model is a good candidate model.

Sagittaria lancifolia example continued ...

Additionally, as the SAS output below shows, each of the regression terms in (6) are stable in that each of the coefficients differs significantly from zero (as indicated by the p-values of the t-tests). The SAS output from fitting the model (6) is given below:

```
                                   The REG Procedure
                                     Model: MODEL1
                                Dependent Variable: lw

                                 Analysis of Variance

                                         Sum of           Mean
Source                      DF          Squares         Square    F Value    Pr > F

Model                        7        107.51511       15.35930      27.01    <.0001
Error                      279        158.65373        0.56865
Corrected Total            286        266.16884


           Root MSE                   0.75409    R-Square     0.4039
           Dependent Mean             2.71111    Adj R-Sq     0.3890
           Coeff Var                 27.81482


                                  Parameter Estimates

                          Parameter       Standard
   Variable      DF        Estimate          Error    t Value    Pr > |t|

   Intercept      1         6.65334        0.89043       7.47      <.0001
   phos           1        -0.01200        0.00398      -3.02      0.0028
   ash            1        -0.24420        0.04712      -5.18      <.0001
   p2             1      0.00003342     0.00001158       2.89      0.0042
   a2             1         0.00474     0.00080618       5.88      <.0001
   p3             1     -3.33403E-8     9.69351E-9      -3.44      0.0007
   a3             1     -0.00002877     0.00000431      -6.67      <.0001
   pa             1      0.00007792     0.00001779       4.38      <.0001
```

Here p2, a2 and p3, a3 are the quadratic and cubic phosphorus and ash-free dry weight terms and “pa” is the interaction regressor term phos×ash.

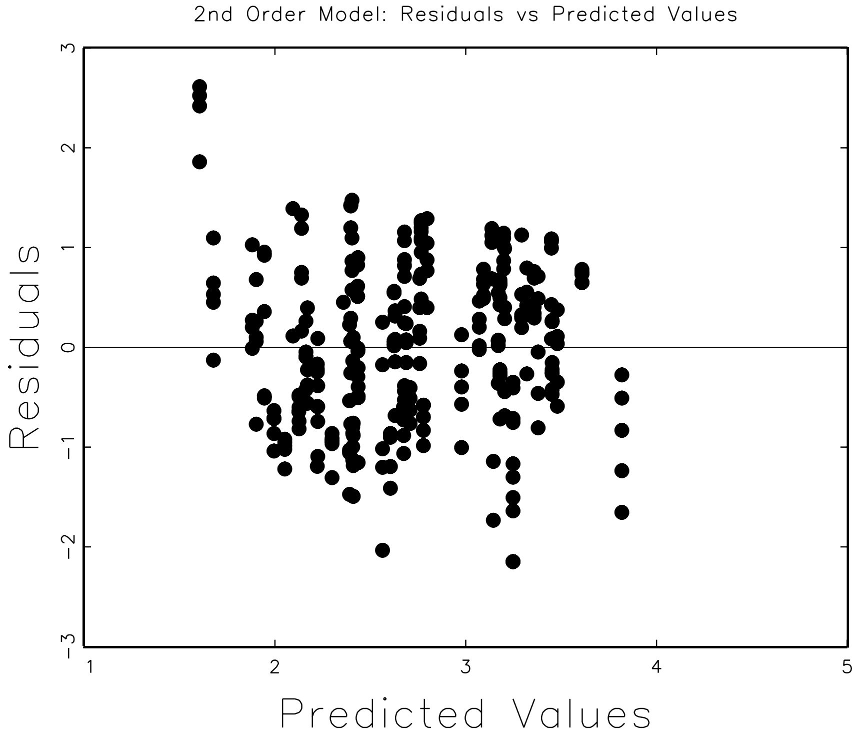

Figure 14: Residual plot for the quadratic model for the *Sagittaria lancifolia* data.

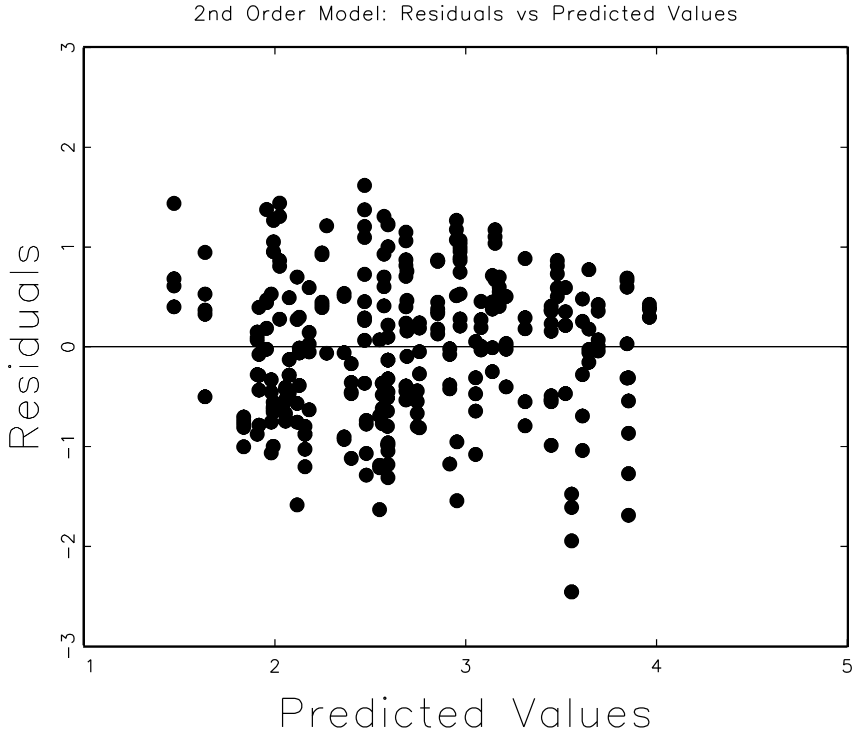

Figure 15: Residual plot for the cubic model (6) for the *Sagittaria lancifolia* data.

Sagittaria lancifolia **example continued ...**

Models with higher order interaction terms were also fit but these terms were not significant.

Note that the R^2 for the cubic model is $R^2 = 0.4039$ which is quite an improvement over the quadratic model.

The problem with this cubic model is that it is difficult to interpret the estimated regression coefficients.

The interpretation is made easier by plotting the regression surface which is shown in Figure 16.

This three-dimensional plot shows that the mean log(leaf width) increases with increasing percentage of ash-free dry weight and also with increasing phosphorus. However, it appears that the log(leaf-width) begins to decrease when phosphorus values become too large.

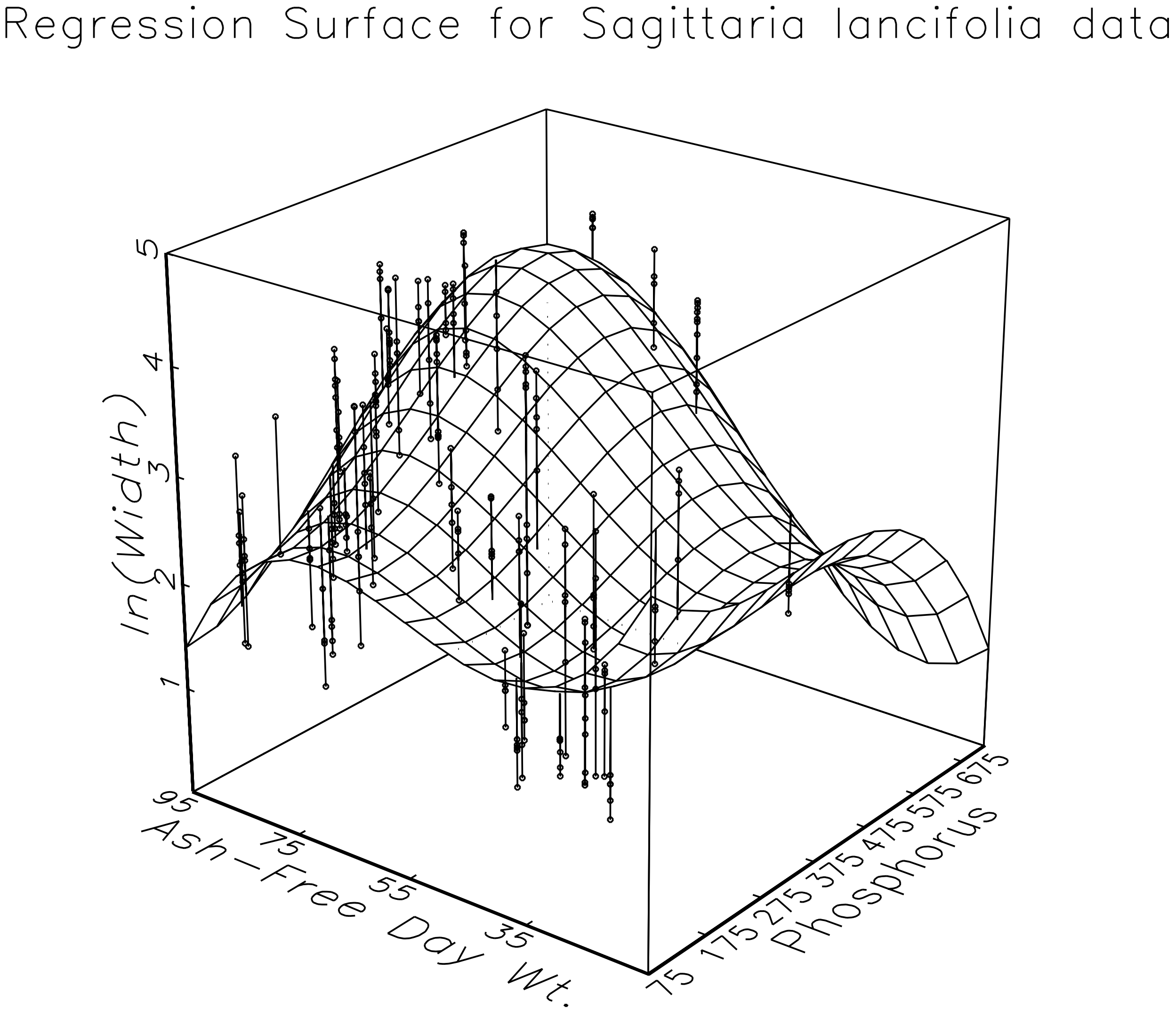

Figure 16: The cubic regression surface for the *Sagittaria lancifolia* data. Also shown are the actual data points connected to their respective predicted values by line segments.

Polynomial Regression

A very important special case of the multivariate regression model useful for situations where the relation between a response y and a predictor x appears nonlinear is a polynomial regression model.

Suppose we are interested in a situation where a response y is related to a single regressor variable x, but the relationship is nonlinear:

$$y = f(x) + \epsilon,$$

for some nonlinear function $f(x)$. Often in practice the functional relationship between x and y is unknown.

However, if the function f is "well-behaved" (for example continuous and smooth), then $f(x)$ can be approximated fairly well by a polynomial due to Taylor's theorem.

Therefore, the following polynomial model often works well in practice:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \epsilon.$$

Note that this is just a special case of a multiple regression model with regressors $x, x^2, \ldots, x^p$.

For this reason, the model is still called a *linear model* because the model is linear in the parameters $\beta_0, \ldots, \beta_p$, even though the relationship between y and x may be nonlinear.

In other words, because we can express the polynomial relationship as a special case of a multiple regression model, polynomial regression models are linear models.

ATP Sludge Example continued. Recall that in the ATP example that the relationship between the ATP in the sludge and the pH level appeared nonlinear. In fact, from the data in Figure 4, it appears as if a quadratic model may fit the data well:

$$\text{ATP} = \beta_0 + \beta_1(\text{pH}) + \beta_2(\text{pH})^2 + \epsilon.$$

In order to fit this model in SAS, the quadratic term needs to be defined, as in the following SAS code:

```
options ls=76 nodate;
data atp;
input ph atp;
ph2=ph**2;
datalines;
2 .16
5 .31
8 .32
9 .26
11 .07
;
run;
proc reg;
     model atp=ph ph2;
run;
```

The variable “ph2” in the SAS code corresponds to $(\text{pH})^2$.

Sludge example continued...

When we first analyzed this data using only a simple linear regression, the regressor pH was not significant, but Figure 4 clearly shows a relationship between pH level and ATP. The output from fitting the quadratic model from the above SAS code gives:

```
                           Analysis of Variance

                                       Sum of           Mean
Source                   DF           Squares         Square    F Value    Pr > F

Model                     2           0.04492        0.02246      55.85    0.0176
Error                     2        0.00080421     0.00040211
Corrected Total           4           0.04572


           Root MSE               0.02005    R-Square     0.9824
           Dependent Mean         0.22400    Adj R-Sq     0.9648
           Coeff Var              8.95204

                           Parameter Estimates

                            Parameter       Standard
  Variable      DF           Estimate          Error    t Value    Pr > |t|

  Intercept      1           -0.07758        0.03969      -1.95      0.1898
  ph             1            0.13772        0.01417       9.72      0.0104
  ph2            1           -0.01123        0.00108     -10.35      0.0092
```

Note that the coefficients for the linear and quadratic terms are now both highly significant ($p = 0.0104$ and $p = 0.0092$ respectively). The coefficient of determination is $R^2 = 0.9824$ indicating that the quadratic regression explains almost all the variability in ATP response. Figure 17 shows the fitted quadratic response as well as a scatterplot of the data.

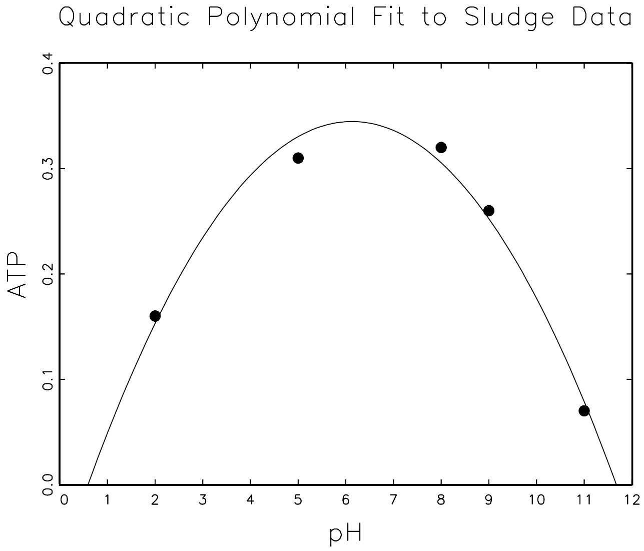

Figure 17: Scatterplot of the ATP sludge data versus pH level and the fitted quadratic polynomial curve.

Once the quadratic model is fit to the data, it can be used to estimate the value of pH that leads to the highest level of ATP.

Cautionary Notes on the Use of Polynomial Regression.

One must use extreme care when fitting polynomial models. In the ATP example above, fitting a quadratic model is certainly a reasonable approach.

However, it is tempting in practice to *overfit* models using polynomial regression. If, for example, we added a cubic term to the quadratic model (i.e. $\text{ATP} = \beta_0 + \beta_1(\text{pH}) + \beta_2(\text{pH})^2 + \beta_3(\text{pH})^3 + \epsilon$), that will increase the coefficient of determination to $R^2 = 0.9987$ which is usually considered desirable. However, this is only a small increase in R^2 compared to the quadratic model.

The SAS output for the estimated parameters of the cubic fit is given below:

```
                           Parameter Estimates

                          Parameter        Standard
Variable     DF            Estimate           Error    t Value    Pr > |t|

Intercept     1             0.02561         0.03325       0.77      0.5821
ph            1             0.06749         0.02079       3.25      0.1902
ph2           1             0.00102         0.00352       0.29      0.8202
ph3           1         -0.00061804      0.00017645      -3.50      0.1770
```

Note that none of the estimated coefficients are stable (all the p-values are large). This is clearly an undesirable model. The quadratic model fits the data very well and there is no need to further complicate the model by the inclusion of higher order polynomial terms.

In fact, higher order terms of the form x^2, x^3, x^4 and so on tend to be highly correlated with one another which leads to very unstable estimated models. High correlations between regressor variables is the topic of the next section.

Collinearity.

One of the most serious problems in a multiple regression setting is (multi)collinearity which occurs when the regressor variables are correlated with one another. Collinearity causes many problems in a multiple regression model.

One of the main problems is that if collinearity is severe, the estimated regression model becomes very unstable and the resulting coefficient estimates become difficult to impossible to interpret. In order to illustrate the problem of collinearity, we give an example:

Heart Catheter Example. A study was conducted and data collected to fit a regression model to predict the length of a catheter needed to pass from a major artery at the femoral region and moved into the heart for children (Weisberg 1980). For 12 children, the proper catheter length was determined by checking with a fluoroscope that the catheter tip had reached the right position. The goal is to determine a model where a child's height and weight could be used to predict the proper catheter length. The data are given in the following table:

Height	Weight	Length
42.8	40.0	37
63.5	93.5	50
37.5	35.5	34
39.5	30.0	36
45.5	52.0	43
38.5	17.0	28
43.0	38.5	37
22.5	8.5	20
37.0	33.0	34
23.5	9.5	30
33.0	21.0	38
58.0	79.0	47

Heart Catheter Example continued...

After fitting a multiple regression model using height and weight as regressors, how do we interpret the resulting coefficients? The coefficient for height tells us how much longer the catheter needs to be for each additional inch of height of the child *provided the weight of the child stays constant*. But the taller the child, the heavier the child tends to be.

Figure 18 shows a scatterplot of weight versus height for the $n = 12$ children from this experiment. The plot shows a very strong linear relationship between height and weight. The correlation between height and weight is $r = 0.9611$. This large correlation complicates the interpretation of the regression coefficients.

Heart Catheter Example continued...

Figure 19 shows a 3-dimensional plot of the height and weight as well as the response variable $y =$ length.

The goal of the least-squares fitting procedure is to determine the best-fitting plane through these points.

However, the points lie roughly along a straight line in the height-weight plane.

Consequently, fitting the regression plane is analogous to trying to build a table when all the legs of the table lie roughly in a straight line.

The result is a very wobbly table.

Ordinarily, tables are designed so that the legs are far apart and spread out over the surface of the table.

When fitting a regression surface to highly correlated regressors, the resulting fit is very unstable.

Slight changes in the values of the regressor variables can lead to dramatic differences in the estimated parameters.

Consequently, the standard errors of the estimated regression coefficients tend to be inflated.

In fact, it is quite common for none of the regression coefficients to differ significantly from zero when individual t-tests are computed for regression coefficients.

Additionally, the regression coefficients can have the wrong sign – one may obtain a negative slope coefficient when instead a positive coefficient is expected.

Heart Catheter Example continued ...

A multiple regression was used to model the length of the catheter (y) with regressors height (x_1) and weight (x_2):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

The ANOVA table (from SAS) is below:

```
                        Analysis of Variance

                                    Sum of           Mean
Source                     DF      Squares         Square    F Value    Pr > F

Model                       2    607.18780      303.59390      21.27    0.0004
Error                       9    128.47887       14.27543
Corrected Total            11    735.66667
```

The overall F-test indicates that at least one of the slope parameters β_1 and/or β_2 differs significantly from zero (p-value $= 0.0004$).

Furthermore, the coefficient of determination is $R^2 = 0.8254$ indicating that height and weight explain a great deal of the variability in the catheter lengths.

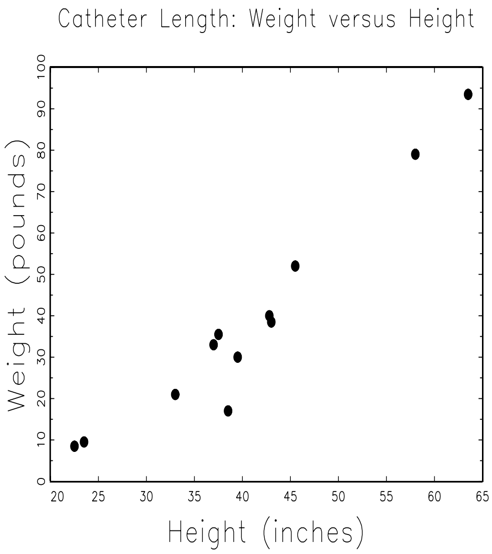

Figure 18: A scatterplot of weight versus height for $n = 12$ children in an experiment used to predict the required length of a catheter to the heart based on the child's height and weight.

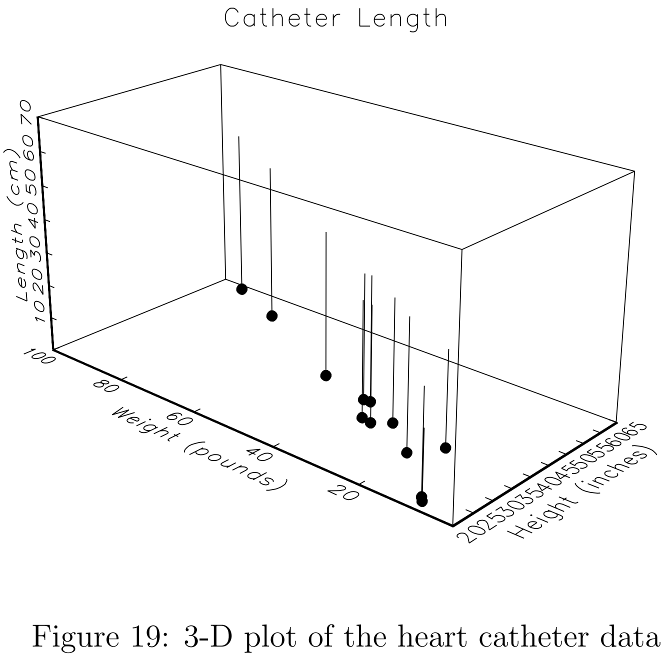

Figure 19: 3-D plot of the heart catheter data.

Catheter Example continued ...

The parameter estimates, standard errors, t-test statistics and p-values shown below indicate that neither β_1 nor β_2 differ significantly from zero.

```
                        Parameter Estimates

                     Parameter       Standard
Variable     DF        Estimate           Error    t Value    Pr > |t|

Intercept     1        20.37576         8.38595       2.43      0.0380
height        1         0.21075         0.34554       0.61      0.5570
weight        1         0.19109         0.15827       1.21      0.2581
```

The apparent paradox is that the overall F-test says at least one of the coefficients differs from zero whereas the individual t-tests says neither differ from zero is due to the collinearity problem. This phenomenon is quite common when collinearity is present.

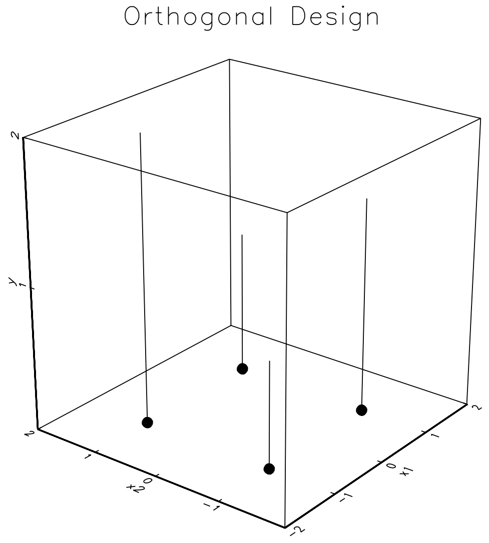

Figure 20: 3-D plot of orthogonal regressor variables.

Collinearity continued ...

Another major problem when collinearity is present is that the fitted model will not be able to produce reliable predictions for values of the regressors away from the range of the regressor values.

If the data changes just slightly, then the predicted values outside the range of the data can change dramatically when collinearity is a present – just think of the wobbly table analogy.

In a well designed experiment, the values of the regressor variables will be orthogonal which means that the covariances between estimated coefficients will be zero. Figure 20 illustrates an orthogonal design akin to the usual placement of legs on a table.

There are many **solutions** to the collinearity problem:

- The easiest solution is to simply drop regressors from the model.

  In the heart catheter example, height and weight are highly correlated.

  Therefore, if we fit a model using only height as a regressor, then we do not gain much additional information by adding weight to the model.

  In fact, if we fit a simple linear regression using only height as a regressor, the coefficient of height is highly significant (p-value < 0.0001) and the $R^2 = 0.7971$ is just slightly less than if we had used both height and weight as regressors.

  (Fitting a regression using only weight as a regressor yields an $R^2 = 0.8181$.)

  When there are several regressors, the problem becomes one of trying to decide upon which subset of regressors works best and which regressors to throw out.

  Often times, there may be several different subsets of regressors that work reasonably well.

  The problem of determining an appropriate set of regressor variables to use in the final model involves both statistical considerations (i.e. minimizing the collinearity problem, stable coefficient estimates, etc.) as well as input from the scientists who are knowledgable about which regressors are most important to the problem at hand.

- Collect more data with the goal of spreading around the values of the regressors so they do not form the “picket fence” type pattern as seen in Figure 19.

  Collecting more data may not solve the problem in situations where the experimenter has no control over the relationships between the regressors, as in the heart catheter example.

- Biased regression techniques such as *ridge regression* and *principal component regression.*

  Details can be found in advanced textbooks on regression.

  These methods are called biased because the resulting slope estimators are biased.

  The tradeoff is that the resulting estimators will be more stable.

Centering Variables

In polynomial regression, collinearity is a very common problem when fitting higher order polynomials and the resulting least-squares estimators tend to be very unstable.

Centering the regressor variable x at its mean (i.e. using $x_i - \bar{x}$ instead of x_i) helps alleviate this problem to some extent. Another solution is to fit a polynomial regression using *orthogonal polynomials* (details of this method can also be found in many advanced texts on regression analysis).

Regression with Indicator Variables (or Dummy Variables)

In many statistical applications, the goal is to compare two or more groups in terms of a continuous response variable y. The statistical tests for making such comparisons are the two-sample t-test or an analysis of variance (ANOVA) for comparing more than two population means.

In many such cases, there may be additional information available that can help in the comparison of the populations.

This additional information may exist in the form of a continuous variable x that is correlated with the response variable y.

Thus, we could model the response y using a regression relation with x, but we would also like to build into the model a way of differentiating between the different populations.

To make matters clearer, we illustrate with an example:

Example. Greenfield et al (2003) studied chemical toxins in several fish in and around the San Francisco Bay. The data below show data collected on mercury (HG) contamination in the Jacksmelt and White Croaker. The length of the fish was also recorded. Older fish tend to be longer and one may suspect that older fish would have higher levels of mercury contamination due to longer exposure time.

Jacksmelt		White Croaker	
Length (cm)	HG (mg/g)	Length (cm)	HG (mg/g)
26.8	0.297	26.4	1.030
27.0	0.271	27.4	1.150
27.4	0.243	27.6	1.050
25.8	0.202	24.0	0.587
26.4	0.249	24.8	0.933
26.6	0.234	27.4	0.645
27.0	0.310	24.8	0.782
27.2	0.353	26.6	0.853
27.6	0.299	27.6	0.870
27.4	0.220	28.0	1.040
27.8	0.478	28.0	0.820
		28.6	0.778
		26.4	0.828
		27.8	0.858
		27.8	1.080

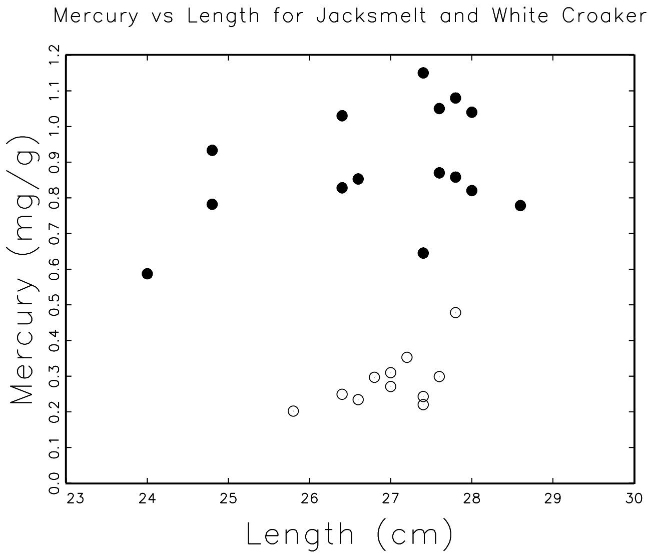

Figure 21: Scatterplot of mercury levels (mg/g) versus length (cm) for Jacksmelt (open circles) and white croaker (solid circles).

Figure 21 shows a scatterplot of mercury level (mg/g) in the fish versus length of the fish (cm.) for Jacksmelt (open circles) and white croaker (solid circles).

From Figure 21 it appears that the jacksmelt fish have lower levels of mercury concentration than the white croaker fish.

We can also see that the mercury concentration appears to increase with increasing size of the fish.

A goal of this example is to compare the mercury levels in the two species of fish. Since the mercury level also depends on length, we can use a regression model to compare mercury levels while taking the length of the fish into consideration.

The variable length in this type of situation is known as a *covariate* and the analysis is sometimes referred to as *analysis of covariance* (ANCOVA).

We can model the relation of mercury contamination with fish length and fish species by incorporating an indicator variable (or a "dummy" variable) that takes only the values zero or one as follows:

$$d_i = \begin{cases} 0, & \text{if } i\text{th fish is jacksmelt} \\ 1, & \text{if } i\text{th fish is white croaker} \end{cases}.$$

Consider the following multiple regression model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 (x_i d_i) + \epsilon_i. \tag{7}$$

Note that if the ith fish is a jacksmelt, then $d_i = 0$ and consequently, the terms in (7) involving d_i are zero. Hence, (7) can be broken down as:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i, & \text{if } i\text{th fish is jacksmelt} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i + \epsilon_i, & \text{if the } i\text{th fish is white croaker} \end{cases}.$$

Therefore, (7) yields two distinct regression lines for the two different species. The coefficients β_2 and β_3 distinguish the two species.

If $\beta_2 = \beta_3 = 0$, then the two species have the same regression line and hence there would be no difference in mean mercury levels based on the type of fish for any given length.

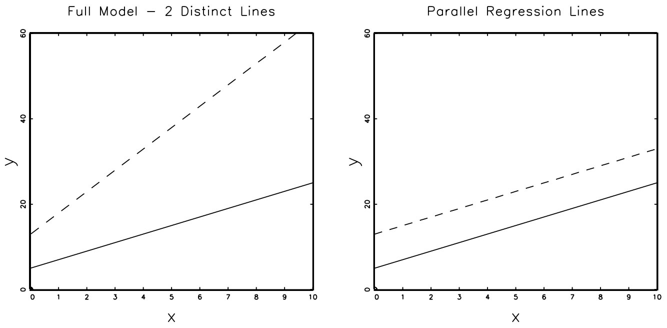

Figure 22: **Left Panel**– Two distinct regression lines when β_2 and β_3 are both non-zero in (7). **Right Panel**–Parallel regression lines when $\beta_3 = 0$ and $\beta_2 \neq 0$ in (7)

We can test this hypothesis using the partial F-test described above.

On the other hand, if $\beta_3 = 0$ but $\beta_2 \neq 0$, then the two regression lines have the same slope but different y-intercepts, i.e. the two lines are parallel.

The left panel of Figure 22 illustrates the case where β_2 and β_3 are both non-zero resulting in two distinct regression lines.

The right panel of Figure 22 illustrates the case where $\beta_3 = 0$ in (7) but $\beta_2 \neq 0$ which yields two parallel regression lines.

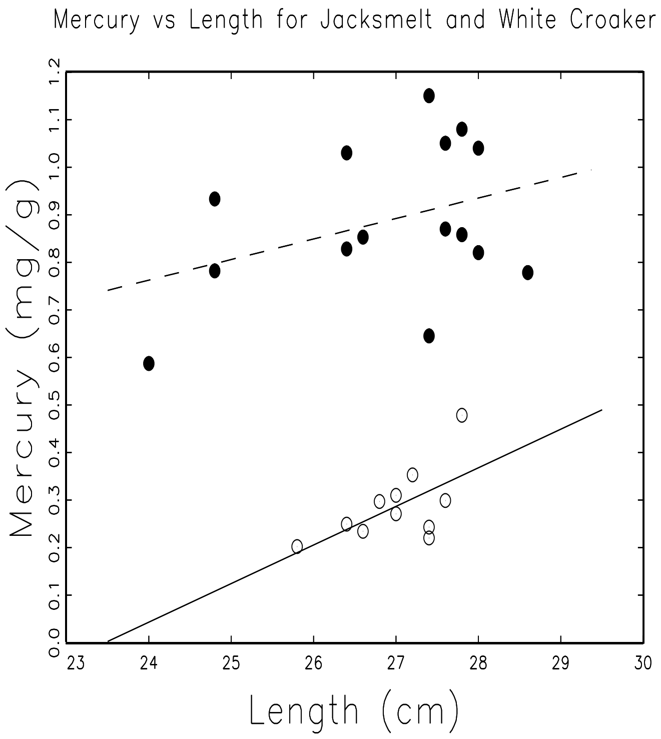

Figure 23: Scatterplot of mercury levels (mg/g) versus length (cm) for Jacksmelt (open circles) and white croaker (solid circles). Also plotted are the regression lines from (7).

The model (7) was fit to the fish data using SAS and the fitted lines are shown in Figure 23. Also shown is the fitted model (7).

The SAS code needed to fit the model and test hypotheses is below:

```
options ls=76;
data hg;
input length hg fish;
loghg=log(hg);
lf=length*fish;
datalines;
26.8 0.297 0
27.0 0.271 0
27.4 0.243 0
25.8 0.202 0
26.4 0.249 0
26.6 0.234 0
27.0 0.310 0
27.2 0.353 0
27.6 0.299 0
27.4 0.220 0
27.8 0.478 0
26.4 1.030 1
27.4 1.150 1
27.6 1.050 1
24.0 0.587 1
24.8 0.933 1
27.4 0.645 1
24.8 0.782 1
26.6 0.853 1
27.6 0.870 1
28.0 1.040 1
28.0 0.820 1
28.6 0.778 1
26.4 0.828 1
27.8 0.858 1
27.8 1.080 1
;
run;
proc reg;
     model hg=length fish lf;
     model hg=length fish;
     model hg=length;
run;
```

Note that this program fits three regression models whose output we will use to test hypotheses about the relation between the two regression lines. The variable “lf” in the SAS code is the *interaction* term which is defined as the product length × fish where fish is the indicator variable for the type of fish.

We can regard (7) as the FULL model with all the regression terms. The simplest model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

This simple model does not have any terms involving the indicator variable and hence yields a single regression line for both species of fish.

The "**parallel**" regression line model is intermediate between the full and simple model and is given by $y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i$. This parallel regression line model distinguishes between the two species of fish only in the y-intercepts of the two lines. The SAS output from the full model is given below:

```
                               Analysis of Variance

                                       Sum of           Mean
Source                   DF           Squares         Square    F Value    Pr > F

Model                     3           2.35579        0.78526      49.43    <.0001
Error                    22           0.34947        0.01589
Corrected Total          25           2.70526


           Root MSE              0.12604    R-Square     0.8708
           Dependent Mean        0.63308    Adj R-Sq     0.8532
           Coeff Var            19.90844


                                Parameter Estimates

                        Parameter       Standard
   Variable     DF       Estimate          Error    t Value    Pr > |t|

   Intercept     1       -1.90523        1.85686      -1.03      0.3160
   length        1        0.08119        0.06876       1.18      0.2503
   fish          1        1.63218        1.97146       0.83      0.4166
   lf            1       -0.03804        0.07303      -0.52      0.6077
```

The fitted model is given by

$$\hat{y}_i = -1.905 + 0.081x_i + 1.632d_i - 0.038x_i d_i,$$

where x_i is the length and y_i is the mercury level of the ith fish.

Notice that the p-values associated with the t-tests for each of the coefficients are all large which appears to indicate that none of the regression coefficients differ significantly from zero.

However, Figure 21 appears to show differences between the jacksmelt and white croaker and also increasing mercury levels with increasing lengths. This model fit is unstable.

For the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i$ the SAS output is:

```
                          Analysis of Variance

                                  Sum of          Mean
Source                  DF       Squares        Square   F Value    Pr > F

Model                    2       2.35148       1.17574     76.44    <.0001
Error                   23       0.35378       0.01538
Corrected Total         25       2.70526


           Root MSE              0.12402    R-Square     0.8692
           Dependent Mean        0.63308    Adj R-Sq     0.8579
           Coeff Var            19.59051


                           Parameter Estimates

                        Parameter       Standard
   Variable     DF       Estimate          Error    t Value    Pr > |t|

   Intercept     1       -0.99490        0.61679      -1.61      0.1204
   length        1        0.04747        0.02280       2.08      0.0487
   fish          1        0.60572        0.04931      12.28      <.0001
```

Finally, for the simplest model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ that does not differentiate between the two sites, we obtain:

```
                          Analysis of Variance

                                      Sum of           Mean
Source                   DF          Squares         Square    F Value    Pr > F

Model                     1          0.03026        0.03026       0.27    0.6071
Error                    24          2.67500        0.11146
Corrected Total          25          2.70526


           Root MSE              0.33385    R-Square     0.0112
           Dependent Mean        0.63308    Adj R-Sq    -0.0300
           Coeff Var            52.73511


                           Parameter Estimates

                        Parameter       Standard
   Variable     DF       Estimate          Error    t Value    Pr > |t|

   Intercept     1       -0.22683        1.65177      -0.14      0.8919
   length        1        0.03193        0.06129       0.52      0.6071
```

The process of determining the appropriate model begins with a test of the null hypothesis

$$H_0 : \beta_2 = \beta_3 = 0,$$

versus the alternative that at least one of these two coefficients differ from zero in the model (7).

This null hypothesis states that the regression relationships of mercury levels versus length for the two species of fish are the same.

We can test this using the partial F-test statistic introduced earlier:

$$F = \frac{(\text{SSR(full)} - \text{SSR(reduced)})/2}{\text{MSE(full)}} = \frac{(2.35579 - 0.03026)/2}{0.01589} = 73.1759,$$

which we compare to the F-distribution on two numerator degrees of freedom (since the full and reduced models differ by two parameters) and 22 denominator degrees of freedom (i.e., the degrees of freedom for MSE of the full model).

The resulting p-value for this test is $p < 0.0001$. Thus, we have very strong evidence that either β_2 and/or β_3 differ from zero.

The next step in the analysis is generally to test

$$H_0 : \beta_3 = 0,$$

in (7) versus the alternative that β_3 differs from zero.

If we can demonstrate that $\beta_3 = 0$ is consistent with the data, then the resulting model yields two parallel regression lines. Using the partial F-test again we obtain

$$F = \frac{(\text{SSR(full)} - \text{SSR(parallel)})/1}{\text{MSE(full)}} = \frac{(2.35579 - 2.35148)}{0.01589} = 0.1356.$$

The p-value for this test is $P(F > 0.1356) \approx 0.7162$.

Because this p-value is quite large, there is insufficient evidence that β_3 differs from zero.

Note that the regression sum of squares for the full model and the parallel regression line model are almost identical indicating that the addition of the interaction terms adds little to the model.

In addition, the coefficients of determination for the full model and the parallel regression model are $R^2 = 0.8708$ and $R^2 = 0.8692$ respectively indicating that the parallel regression model explains almost as much variability in the mercury levels as the full model.

Fish Example continued ...

We shall settle upon the parallel regression model for this data which gives the fitted equation:

$$\hat{y}_i = -0.9949 + 0.04747x_i + 0.6052d_i.$$

Breaking this down for the individual fish species, we get:

$$\hat{y} = \begin{cases} -0.9949 + 0.04747(\text{length}) & \text{for jacksmelt} \\ -0.3892 + 0.04747(\text{length}) & \text{for white croaker} \end{cases}.$$

This parallel regression line model is plotted in Figure 24.

Coefficient Interpretation. Recall that in a simple linear regression, the slope represents the rate of change in the mean response for changes in the regressor.

In the fish example, both species of fish share the same slope $\hat{\beta}_1 = 0.04747$ (mg/g)/cm with estimated standard error $\hat{\text{se}}(\hat{\beta}_1) = 0.02280$. We can use these statistics to form a confidence interval for the slope.

For 95% confidence, our t-critical value is $t_{\alpha/2,n-3} = t_{.025,23} = 2.0687$.

The 95% confidence interval for the common slope is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-3}\hat{\text{se}}(\hat{\beta}_1) = 0.04747 \pm 2.0687(0.02280) = 0.04747 \pm 0.0228,$$

which gives an interval of $(0.0247, 0.0703)$.

Thus, with 95% confidence, we estimate that for every additional centimeter in length, the mean mercury level in white croaker and jacksmelt fish increases by 0.0247 to 0.0703 mg/g.

How do we interpret β_2, the coefficient of the indicator variable d_i?

The coefficient β_2 is the vertical distance between the two regression lines.

For a given value of fish length x, the difference in the mean response between the two fish species is β_2.

From the SAS output we find that $\hat{\beta}_2 = 0.60572$ with estimated standard error 0.04931.

Since we coded the indicator variable to take the value 1 for white croaker, it follows that for any given fish length, we estimate that the mean concentration of mercury in white croaker is 0.60572 mg/g higher than for jacksmelt.

Using a 95% confidence interval for β_2 we obtain

$$\hat{\beta}_2 \pm t_{\alpha/2,n-3}\hat{\text{se}}(\hat{\beta}_2) = 0.60572 \pm 2.0687(0.04931) = 0.60572 \pm 0.1020,$$

which gives an interval of $(0.5037, 0.7077)$.

Thus, with 95% confidence we estimate that the mean mercury contamination in white croaker is 0.5037 to 0.7077 mg/g higher than for jacksmelt regardless of the length of the fish.

Note that we have formed two confidence intervals (for β_1 and β_2) but we have not corrected for multiplicity.

A Bonferroni correction could be made if we want to have a joint level of confidence for both intervals simultaneously.

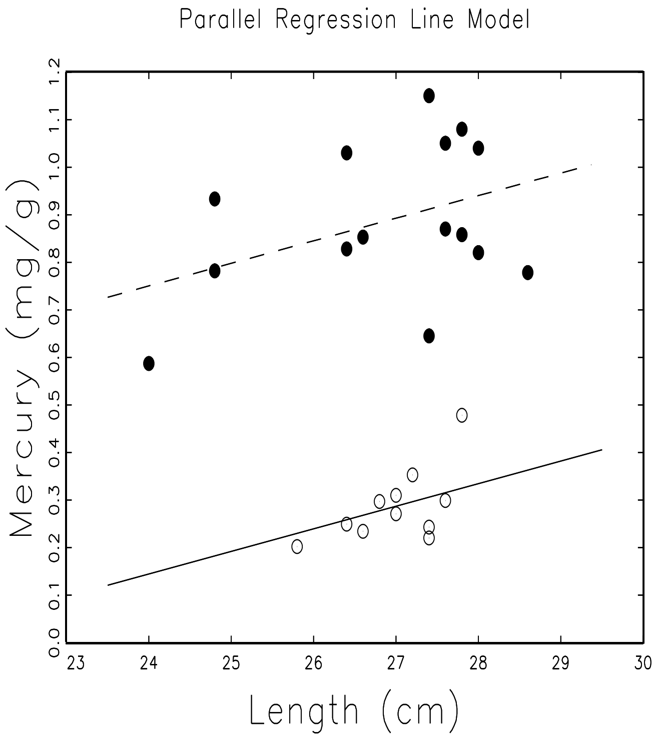

Figure 24: **Parallel regression line model:** Scatterplot of mercury levels (mg/g) versus length (cm) for Jacksmelt (open circles) and white croaker (solid circles). Parallel regression lines from from the model (7) with $\beta_3 = 0$ are also plotted.

In the modeling above, we used a single regression model for both species of fish.

One could also simply fit regression lines individually for each species.

However, doing the fitting using a single model makes it easier to test hypotheses about common slopes etc.

Also, fitting a single model allows us to pool all the data to estimate the error variance.

Pooling the data in this fashion is preferred if the regression relationships for the two populations share a common error variance. However, from Figure 21, it appears that the error variance is probably higher for the white croaker fish than for the jacksmelt.

More than two populations. In the example above, suppose we wanted to compare mercury contamination among three species of fish instead of just two – how could we accomplish this?

The answer once again is to use indicator variables. Generally speaking, if there are k groups, then $k-1$ indicator variables will be needed.

For $k = 3$ species of fish, we can define indicator variables as follows:

$$d1_i = \begin{cases} 0 & \text{if } i\text{th fish is species 1} \\ 1 & \text{otherwise.} \end{cases} \quad \text{and} \quad d2_i = \begin{cases} 0 & \text{if } i\text{th fish is species 1 or 2} \\ 1 & \text{otherwise.} \end{cases}.$$

Then the full model can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d1_i + \beta_3 d2_i + \beta_4 (x_i d1_i) + \beta_5 (x_i d2_i) + \epsilon_i.$$

One can readily verify that this model yields three distinct regression lines for the three different species.

As in the previous example, partial F-tests can be conducted to see if a suitably reduced model is consistent with the data.

Connection with ANOVA and t-tests.

In the above examples, we have considered regression models with regressors that are both continuous and indicator variables.

We can also consider regression models using only indicator variables.

For instance, in the fish-mercury contamination example, suppose we deleted the regressor length from the model and use the model

$$y_i = \beta_0 + \beta_1 d_i + \epsilon_i,$$

where d_i is the indicator distinguishing between jacksmelt and white croaker fish.

One can show with some algebra that the test of $H_0 : \beta_1 = 0$ is equivalent to the two-sample t-test for a difference in the mean mercury levels for the two fish species.

Similarly, if we had three groups in our data and defined two indicator variables to distinguish between the groups, then the regression ANOVA table from this model would coincide with the ANOVA table when testing equality of the three population means.

One can generalize this to handle factorial experiments as well (two or more factors) by defining indicator variables appropriately.

Spline Models

This section provides a brief introduction to regression spline models.

Regression models are used to model the functional relationship between a response variable y and one or more predictor variables.

Often in regression examples, the functional relationship between the response the predictors may vary depending on the value of the predictors.

In order to accommodate these changes, spline models are often useful.

The basic idea of spline models is to "piece" together different functions.

For example, consider the growth of a child from birth. The child may grow very rapidly during the first few years of life and then the growth will slow down.

In order to model the relationship between the child's height and age, it may be necessary to piece together one or more different functions for the different periods of growth in a child's life.

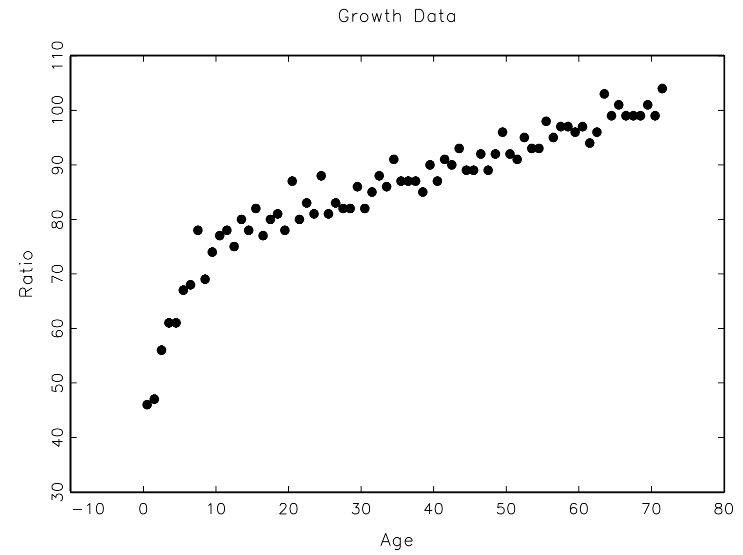

Figure 25: Scatterplot of children's height to weight ratio versus age.

The following example will help illustrate the ideas.

Example. Eppright et al (1972) looked at the growth of 72 children from birth to 70 months of age. Figure 25 shows a scatterplot of the children's height to weight ratio versus age.

The plot shows a rapid increase in the ratio during the first several months of age indicating that the children's height is increasing at a much faster pace than the children's weight in the first few months of age, followed by what appears to be roughly a linear increase in the ratio beginning around 15 months of age.

In order to model the data shown in Figure 25, a multiple regression model will be need to be used.

In the above example, let x denote the age of the child in months and let y denote the height to weight ratio. Clearly, the response y depends on x in a nonlinear fashion.

In order to define the spline model, a new type of regressor variable needs to be defined to account for the change in the functional response.

Spline Models continued ...

For a fixed value x_0, define a new regressor variable as:

$$(x - x_0)_+ = \begin{cases} 0 & \text{if } x < x_0 \\ x - x_0 & \text{if } x \geq x_0. \end{cases}$$

The point x_0 is known as a *knot point.*

Consider the following multiple regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 (x - x_0)_+^0 + \beta_3 (x - x_0)_+ + \epsilon.$$

From the definition of $(x - x_0)_+$, this regression model can be written as

$$y = \begin{cases} \beta_0 + \beta_1 x + \epsilon & \text{if } x < x_0 \\ \beta_0 + \beta_1 x + \beta_2 + \beta_3 (x - x_0) + \epsilon & \text{if } x \geq x_0, \end{cases}$$

which yields two distinct regression lines for values of $x < x_0$ and $x \geq x_0$.

Note that if $\beta_2 = 0$ in the above model, then the resulting regression model is

$$y = \begin{cases} \beta_0 + \beta_1 x + \epsilon & \text{if } x < x_0 \\ \beta_0 + \beta_1 x + \beta_3 (x - x_0) + \epsilon & \text{if } x \geq x_0, \end{cases}.$$

The nice aspect of this model is that the response curve is piecewise linear *and* it is continuous at the knot point. To see that the response function is continuous, evaluate it for $x < x_0$ and for $x \geq x_0$ and note that as x gets closer to x_0 from the right, the β_3 term goes to zero. In other words, the response function consists of two lines connected at the knot point.

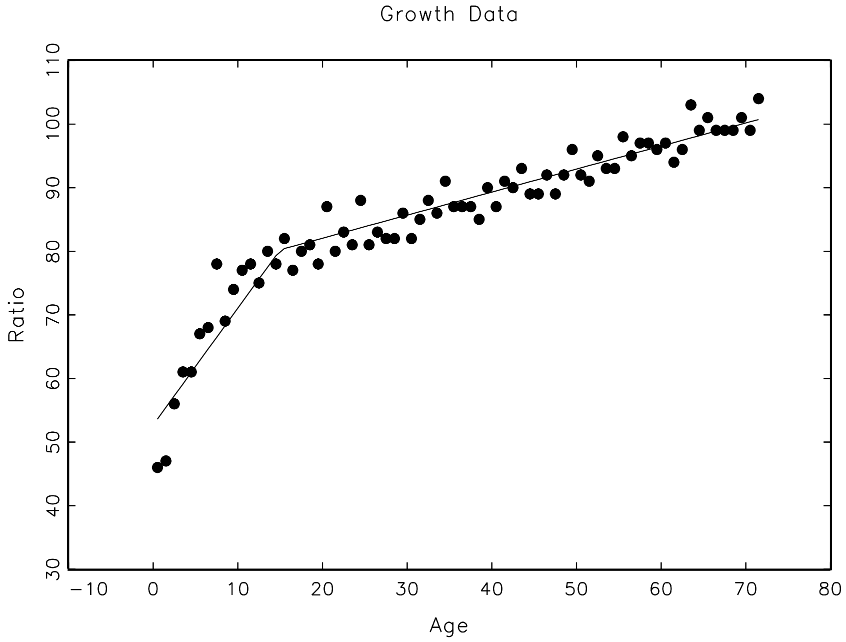

Figure 26: Scatterplot of children's height to weight ratio versus age with a fitted piecewise linear spline.

Spline Models continued ...

Figure 26 shows a scatterplot of the data once again along with the fitted piecewise linear spline

$$\hat{y} = 52.70483 + 1.836x - 1.474(x - 15)_+.$$

We can rewrite this equation using the definition of the $(x - 15)_+$ regressor as

$$\hat{y} = \begin{cases} 52.70483 + 1.836x & \text{for } x < 15 \\ 74.81483 + 0.362x & \text{for } x \geq 15 \end{cases}$$

Cubic Splines.

The piecewise linear fit shown in Figure 26 has the undesirable property of having a sharp "bend" at the knot point.

Also, the straight line fit for the first 15 months does not seem to be adequate.

In order to fit a more flexible curve that fits the data better, we could try instead to fit a *cubic spline* which will also use the $(x-x_0)_+$ regressor variable:

$$y = \beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4(x-x_0)_+^0+\beta_5(x-x_0)_++\beta_6(x-x_0)_+^2+\beta_7(x-x_0)_+^3+\epsilon. \tag{8}$$

We can consider this "full" model or some "reduced" models be dropping some of the terms.

- The full model (8) produces two completely distinct cubic polynomials to the left and to the right of the knot point x_0. If the biological setting for the model requires a continuous response, then the full model will not be appropriate because it will generally be discontinuous at the knot point.

- Dropping the β_4 term will produce two cubic polynomials that join together at the knot point. However, the response at the knot point will generally be "sharp" and not smooth.

- Dropping the β_4 and β_5 terms will produce two cubic polynomials that join together at the knot point x_0 and the function will be continuous at the knot point and the first derivative will also be continuous making the join point a little smoother.

- Dropping the β_4, β_5 and β_6 terms will produce a spline of two cubic polynomials that join together at the knot point and the function and its first two derivatives will be continuous at the knot point producing a smooth response curve.

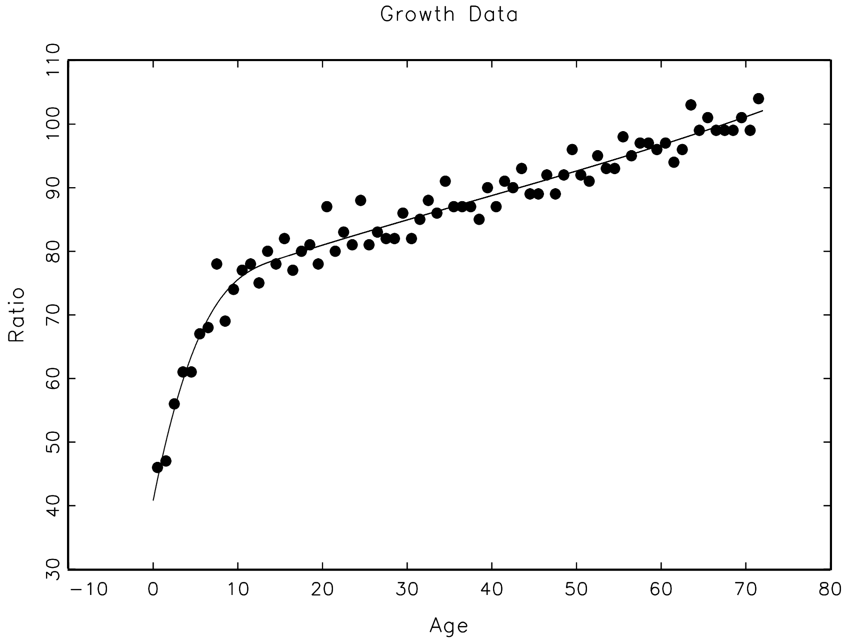

Figure 27: Scatterplot of children's height to weight ratio versus age with a fitted cubic spline.

Figure 27 shows the height to weight ratio data versus age along with a fitted cubic spline (8) with the β_4, β_5 and β_6 terms dropped from the model. As you can see, the fitted response curve is quite smooth at the knot point of $x_0 = 15$ months.

Note that if we drop β_4, β_5 and β_6 from (8), not only is the fitted curve smoother at the knot point, but the model is more parsimonious due to fewer parameters.

In order to test if one can drop the β_4, β_5 and β_6 terms from the cubic spline model (8), the usual F-test from a multiple regression can be performed by fitting the full and reduced model and comparing the regression (or error) sum of squares.

Here is a SAS program for fitting the cubic spline to the ratio versus age data set:

```
/*******************************************************************
Growth data for children from 72 children from birth
to 70 months (Eppright et al, 1972).
Response is the height to weight ratio (in column 1) and the
regressor is age in months (column 2).
Fit a cubic spline model with a knot at age=15 months.
********************************************************************/
options ls=76 nodate;
data growth;
infile 'c:\es714\notes\growth.dat';
input ratio age;
ratio=ratio*100; * rescale the ratio;
age2=age*age;
age3=age2*age;
if age >= 15 then x11=(age-15);
else x11=0;
x12=x11*x11;
x13=x12*x11;
run;
proc reg; model ratio=age x11;run;
proc plot;
     plot ratio*age;
run;
proc reg;
     model ratio=age age2 age3 x13;
     output out=a p=p r=r;
run;
proc plot;
     plot r*p/vref=0; * residual plot;
run;
quit;
```

Part of the output produced by this SAS program is shown below:

```
                          Analysis of Variance

                                Sum of           Mean
Source                   DF    Squares         Square   F Value   Pr > F

Model                     4      10155     2538.81491    461.58   <.0001
Error                    67  368.51812        5.50027
Corrected Total          71      10524


           Root MSE              2.34527    R-Square     0.9650
           Dependent Mean       85.55556    Adj R-Sq     0.9629
           Coeff Var             2.74122


                           Parameter Estimates

                         Parameter       Standard
   Variable     DF        Estimate          Error     t Value     Pr > |t|

   Intercept     1        40.80249        1.81603       22.47       <.0001
   age           1         6.73817        0.54298       12.41       <.0001
   age2          1        -0.41891        0.04358       -9.61       <.0001
   age3          1         0.00927        0.00104        8.90       <.0001
   x13           1        -0.00924        0.00108       -8.54       <.0001
```

We can write the estimated spline model as

$$\hat{y} = 40.8025 + 6.738x - 0.4189x^2 + 0.00927x^3 - 0.00924(x-15)_+^3.$$

A plot of the raw data and this fitted cubic spline is shown in Figure 27

The cubic spline model of (8) can easily be generalized to models with more than one knot point.

Nonlinear Regression

All of the regression models we have considered so far have been linear models. By “linear” we mean that the model is linear in the parameters (i.e. the β’s).

Even the polynomial models are considered linear models even though the response is a nonlinear function of the regressor x.

Many regression applications involve nonlinear models – that is, models that are not linear in the parameters.

A common example is an exponential growth model. Let μ_y denote the mean size of a population that grows (or decays) with time t. Then an exponential model for growth (or decay) is

$$\mu_y = \alpha_0 e^{-\alpha_1 t}. \tag{9}$$

This is a nonlinear model because the functional relationship is not a linear function of the parameters α_0 and α_1. The model (9) can easily be linearized by taking the natural logarithm on both sides giving:

$$\ln(\mu_y) = \beta_0 + \beta_1 t,$$

where $\beta_0 = \ln(\alpha_0)$ and $\beta_1 = -\alpha_1$. The logarithm transformation often works well in such cases, but it assumes the error in (9) is multiplicative instead of additive. If the error in (9) is additive, then the logarithm transformation may not yield a very good fit.

Different applications may yield other types of nonlinear models such as the logistic growth model:

$$y = \frac{\theta_1}{1 + \theta_2 e^{-\theta_3 x}} + \epsilon,$$

where $\theta_1, \theta_2, \theta_3$ are the parameters. This logistic growth model is also a nonlinear model. Note that we cannot linearize this logistic growth model as we did with the exponential growth model.

For linear models, there exist closed form solutions for the least-squares estimators of the parameters.

However, for most nonlinear models, there do not exist closed form solutions. Instead, for nonlinear models, the least-square estimators must be determined by iterative algorithms.

Consider the following general setup. Suppose the response variable y is related to the regressor variable x via a nonlinear function f that depends on unknown parameters $\theta_1, \ldots, \theta_k$ (in the logistic growth model, $k = 3$). Then we can write the model as

$$y = f(x; \theta_1, \ldots, \theta_k) + \epsilon.$$

As with any regression model setup, we obtain data $(x_1, y_1), \ldots, (x_n, y_n)$, and use this data to estimate the parameters $\theta_1, \ldots, \theta_k$.

Least-squares is used most frequently for parameter estimation: find the values of the θ_j's that minimize the sum of squares:

$$\sum_{i=1}^{n} (y_i - f(x_i; \theta_1, \ldots, \theta_k))^2.$$

The least squares solutions will be denoted by $\hat{\theta}_j$ for $j = 1, \ldots, k$.

It turns out that if the error ϵ is approximately normal, then the least squares estimators will have approximately normal distributions and they will be approximately unbiased.

The approximation becomes better as the sample size gets larger.

Most statistical software programs for nonlinear least-squares will provide the parameter estimates as well as estimated standard errors that are based on asymptotic results and the assumption that the error is normal.

An alternative to the typical standard error estimates is to use a method called the *bootstrap* which is a computer intensive technique (see Efron and Tibshirani (1993) for a good introduction to the bootstrap and its uses).

Gauss-Newton Algorithm

The most well-known method of finding the least-squares estimators is the *Gauss-Newton Algorithm.*

Most of the common statistical software programs use the Gauss-Newton algorithm or variants of it.

The idea behind the Gauss-Newton algorithm involves estimating the function f by the linear portion of its Taylor series expansion.

From this linear approximation, the usual least-squares fit can be obtained, which in turn can be used to update the parameter estimates.

The Taylor series expansion involves partial differentiation of the function f with respect to the parameters, the details of which we shall skip.

Gauss-Newton Algorithm continued...

To begin the algorithm, initial values need to be proposed for the parameters.

- One must be careful when providing initial values because the final solution once the algorithm iterates can often depend quite heavily on these initial guesses.
- If the initial values provided are way off, then it is possible for the algorithm to wander off in the wrong direction. If this happens, then the algorithm may not converge or it may yield very poor solutions for parameter estimates.

Gauss-Newton Algorithm continued...

At each iteration of the algorithm, the parameter estimates are updated to a new value.

After the algorithm has gone through several iterations, the change in the parameter estimates from iteration to iteration will be very small.

Typically software programs have some cut-off value for the change in value of parameter estimates at which point the algorithm stops.

With most software packages, the user can also specify the cut-off value for determining if the algorithm has converged.

Gauss-Newton Algorithm continued...

It is quite common to run into **problems** when fitting a nonlinear regression model.

The most common problem is that the algorithm will not converge or it produces nonsense solutions, i.e. parameter estimates that are not allowed.

- This problem is usually due to poor starting values.
- One possible solution is to try different starting values. Sometimes good starting values can be determined by looking at a plot of the data and using properties of the function such noting where a horizontal asymptote should occur.
- Another solution is to perform a grid search over different parameter values in order to find good starting values.

Nonlinear Regression Example

In order to illustrate nonlinear regression models, we provide an example.

Bioconcentration Factors (BCF) using Compartment Models. Compartment models (e.g. see Bates and Watts, Chapter 5) are used to describe outcome measures from physical systems of compartments (e.g., in chemical kinetics, pharmacokinetics, and ecological and environmental systems) in which material flows to and from various compartments.

The parameters of primary interest in compartment models are the rate parameters which specify the rate of transfer of material between compartments.

Bioconcentration factors (BCF's) are indices that quantify how much chemical accumulates in an organism relative to the environmental exposures.

The environment is considered one compartment and the organism is considered the other compartment.

Nonlinear Regression Example continued...

BCF's can be estimated using compartment models as the ratio of the organism's uptake rate to its discharge rate.

If we let k_u denote the uptake rate and k_d denote the rate of depuration (the removal of impurities from the organism), then

$$\text{BCF} = \frac{k_u}{k_d}. \tag{10}$$

Let C_e denote the environmental concentration and $C_o(t)$ denote the organism concentration at time t.

Assuming the rates of chemical uptake and depuration for an organism remain constant, then we have the following equation:

$$\text{Change in concentration w.r.t time} = k_u C_e - k_d C_o(t).$$

This equation can be expressed as a differential equation:

$$dC_o(t)/dt = k_u C_e - k_d C_o(t).$$

The BCF is the ratio of the organism concentration to the environmental concentration at steady state (i.e. when $dC_o(t)/dt = 0$) which corresponds to the definition given in (10).

The function $C_o(t)$ that satisfies the differential equation is

$$C_o(t) = \frac{k_u}{k_d} C_e[1 - e^{-k_d t}]. \tag{11}$$

Note that this is a nonlinear function of the rate parameters k_u and k_d and therefore, nonlinear regression techniques are needed to find estimators of the parameters.

Nonlinear Regression Example...

A study on BCF of the chemical pyrene in *Diporeia* from Lake Erie was undertaken (data provided courtesy of Peter Landrum at NOAA). The data in the following table gives the time (in hours) and the corresponding pyrene concentration (in DPMq/g):

2973.94	44.92
2368.59	46.82
3013.45	46.82
3483.96	46.90
2327.23	46.90
6170.50	92.75
6722.54	92.75
5969.43	99.75
5396.02	99.75
5657.01	99.83
5649.33	99.83
8892.11	168.08
9341.05	168.08
8143.53	173.35
7435.38	173.35
9672.66	173.40
8210.14	173.40
16295.88	310.90
17020.00	310.90
18362.03	339.80
16293.29	339.80
17064.71	339.88
17563.29	339.88
12292.31	479.00
17842.22	479.00
16095.87	507.60
16521.65	507.60
18379.32	507.72
18846.31	507.72
23298.64	672.48
25863.87	672.48
20820.86	676.57
21723.71	676.57
18174.71	676.68
20978.90	676.68

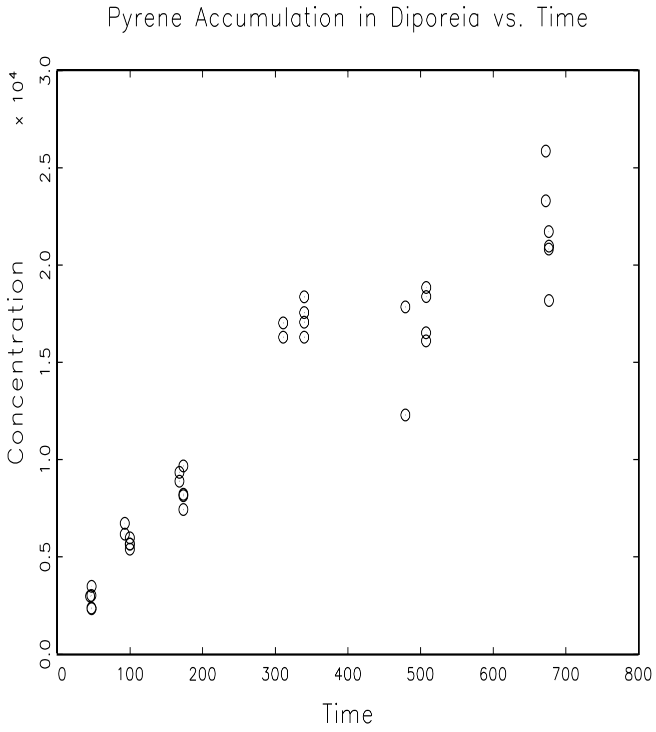

Figure 28: Pyrene accumulation in *Diporeia* versus time (in hours)

Nonlinear Regression Example continued...

A plot of the data is shown in Figure 28.

Nonlinear Regression Example...

SAS was used to fit the model (11) using this data. In SAS, nonlinear regression models are fit using the procedure “proc nlin”. Here is the SAS code for this example:

```
options ls=76 nodate;
data pyrene;
infile 'pyrene.dat';
input pyrene tetra time;
ce=17970.3017;
run;
proc nlin;
parameters ku=0.01
           kd=0.01;
   model pyrene = (ku/kd)*ce*(1-exp(-kd*time));
run;
```

Note that the data is in a file called “pyrene.dat’.

The model statement in SAS’s proc nlin is needed to specify the nonlinear function that is to be fit.

In our case, we are using the function given in (11).

The “parameters” statement gives the initial values for starting the algorithm.

These initial values need to be supplied by the user.

The “ce=17970.3017” is an estimate of the environmental concentration of pyrene that was estimated separately from another data set from sediment.

Nonlinear Regression Example continued...

The SAS output from running this data is given below:

```
                  The NLIN Procedure
               Dependent Variable pyrene
                 Method: Gauss-Newton

                   Iterative Phase
                                              Sum of
             Iter          ku          kd    Squares

                0      0.0100      0.0100   6.3482E8
                1     0.00519     0.00472   1.9193E8
                2     0.00368     0.00258   1.3003E8
                3     0.00385     0.00278   1.2847E8
                4     0.00386     0.00278   1.2847E8
                5     0.00386     0.00278   1.2847E8


NOTE: Convergence criterion met.


                  Estimation Summary

              Method                    Gauss-Newton
              Iterations                           5
              Subiterations                        1
              Average Subiterations              0.2
              R                             5.988E-8
              PPC(kd)                       5.239E-8
              RPC(kd)                       8.896E-6
              Object                        2.43E-10
              Objective                     1.2847E8
              Observations Read                   35
              Observations Used                   35
              Observations Missing                 0
              NOTE: An intercept was not specified for this model.


                                          Sum of        Mean
   Source                         DF     Squares      Square    F Value

   Model                           2    6.9131E9    3.4565E9     887.91
   Error                          33    1.2847E8     3892894
   Uncorrected Total              35    7.0416E9


                     The NLIN Procedure

                                Approx     Approximate 95% Confidence
Parameter      Estimate      Std Error               Limits

ku              0.00386       0.000318     0.00321      0.00451
kd              0.00278       0.000425     0.00192      0.00365

               Approximate Correlation Matrix
                             ku              kd


             ku        1.0000000
             kd        0.9576051       1.0000000
```

From the SAS output, we see that the estimated rate parameters are $\hat{k}_u = 0.00386$ and $\hat{k}_d = 0.00278$. SAS also gives approximate 95% confidence intervals for the estimated parameters.

Note that neither confidence interval contains zero which indicates that the estimated rate parameters differ significantly from zero (using significance level $\alpha = 0.05$).

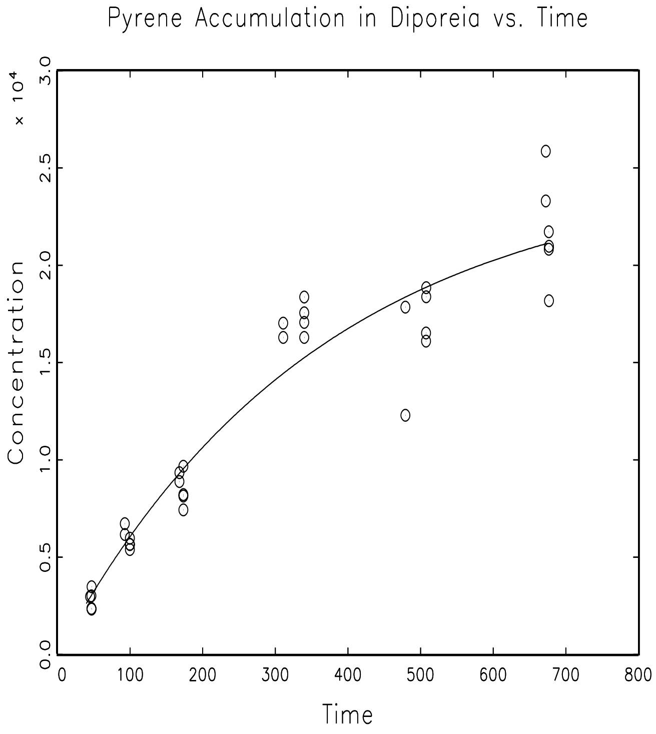

Figure 29: The fitted nonlinear least squares function along with the raw data of pyrene accumulation in *Diporeia* versus time (in hours)

Nonlinear Regression Example continued ...

Figure 29 shows the scatterplot of the raw pyrene data along with the nonlinear least-squares fitted function estimate.

As the figure shows, the curve provides a pretty good fit to the data. The estimated BCF is given by

$$\widehat{\text{BCF}} = \frac{\hat{k}_u}{\hat{k}_d} = \frac{0.00386}{0.00278} = 1.3885.$$

Nonlinear Regression Example continued ...

It would be useful to obtain a standard error for this estimated ratio.

Bailer et al (2000) suggest using the *delta method* (based on a Taylor series approximation to the ratio which defines the BCF) to obtain an estimated standard error.

Using the delta method, the approximate standard error for the estimated BCF is (see equation (4) in Bailer et al (2000)):

$$\hat{\text{se}}(\hat{\text{BCF}}) = \sqrt{\frac{\hat{\sigma}_u^2}{\hat{k}_d^2} - \frac{2\hat{k}_u\hat{\sigma}_{ud}}{\hat{k}_d^3} + \frac{\hat{k}_u^2\hat{\sigma}_d^2}{\hat{k}_d^4}}, \tag{12}$$

where $\hat{\sigma}_u$ and $\hat{\sigma}_d$ are the estimated standard errors for the two rate parameters. The term $\hat{\sigma}_{ud}$ is the estimated *covariance* between the two estimated rate parameters.

SAS gives the estimated correlation between $\hat{k}_d$ and $\hat{k}_u$ (found in the SAS output under " Approximate Correlation Matrix").

For this example, the estimated correlation is $r_{ud} = 0.9576051$.

Recall that the correlation is defined as

$$r_{ud} = \frac{\hat{\sigma}_{ud}}{\hat{\sigma}_u\hat{\sigma}_d}.$$

From this formula, it follows that

$$\hat{\sigma}_{ud} = r_{ud}\hat{\sigma}_u\hat{\sigma}_d.$$

Once this value is computed, it can be plugged into (12) to obtain the approximate standard error for the estimated BCF.

Bootstrap Standard Errors

As an alternative to the delta method, one could use the bootstrap to obtain a standard error for BCF as well.

- To obtain bootstrap standard errors, one can resample the data *with replacement* to obtain a bootstrap sample.

- From the bootstrap sample, estimate the parameters using nonlinear regression.

- Repeat this process for 100's of bootstrap samples.

- The standard error of the estimated parameter can be obtained by taking the standard deviation of the 100's of bootstrap estimates.
- **Bootstrap Confidence Interval**: Arrange the bootstrap parameter estimates from smallest to largest and a 95% bootstrap confidence interval will be given by the range of the middle 95% of the bootstrap parameter estimates.

Another issue that has been ignored is that the environmental concentration C_e is considered a known quantity, but in reality, the environmental concentration needs to be estimated as well.

If the variability in the estimate of C_e is large, then accounting for the variability in estimating C_e will inflate the standard error for the estimated BCF.

Logistic Regression

In the classical regression framework, we are interested in modeling a continuous response variable y as a function of one or more predictor variables.

Most regression problems are of this type.

However, there are numerous examples where the response of interest is not continuous, but binary.

Logistic Regression continued ...

Consider an experiment where the measured outcome of interest is either a success or failure, which we can code as a 1 or a 0.

The probability of a success or failure may depend on a set of predictor variables.

One idea on how to model such data is to simply fit a regression with the goal of estimating the probability of success given some values of the predictor.

However, this approach will not work because probabilities are constrained to fall between 0 and 1.

In the classical regression setup with a continuous response, the predicted values can range over all real numbers. Therefore, a different modeling technique is needed.

Logistic Regression continued ...

In risk assessment studies, a *dose–response curve* is used to describe the relationship between exposure to a given dose of a drug or toxin say and its effect on humans or animals. Let us illustrate matters with an example.

Example. Beetles were exposed to gaseous carbon disulphide at various concentrations (in mf/L) for five hours (Bliss 1935) and the number of beetles killed were noted. The data are in the following table:

Dose	# Exposed	# Killed	Proportion
49.1	59	6	0.102
53.0	60	13	0.217
56.9	62	18	0.290
60.8	56	28	0.500
64.8	63	52	0.825
68.7	59	53	0.898
72.6	62	61	0.984
76.5	60	60	1.000

If we let x denote the concentration of CS_2, then from the table it appears that the probability of death increases with increasing levels of x.

Note that the mortality rate increases with increasing dose.

Our goal is to model the probability of mortality as a function of dose x using a regression model.

Logistic Regression Example continued...

If we consider for a moment the beetles exposed to the lowest concentration of 49.1 mf/L of CS_2, we see that a total of $n = 59$ beetles were exposed and 6 of them died.

The *binomial probability* model is used to model outcomes of this type.

A **Binomial Experiment** is one that consists of

(i) n independent trials where

(ii) each trial results in one of two possible outcomes (typically called success or failure which are recorded as 1 or 0 respectively), and

(iii) the probability p of a success stays the same for each trial.

Binomial Distribution and Beetle Example.

- In the beetle example at the lowest dose, we have $n = 59$ beetles. For each beetle, the outcome is either death (success) or alive (failure).
- We can let p denote the probability an individual beetle will die after five hours at this dose.
- If we let y denote the number of beetles that die, then y is a random variable with a binomial distribution.
- One can show that the probability that y assumes a value k where $k = 0, 1, \ldots, 59$, is given by the following formula:

$$P(y = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Back to Logistic Regression and Beetle example continued...

In the current setup, the situation is more complicated than this because for each dose, we have a corresponding binomial distribution where the success probability p depends on x, the concentration of CS_2.

Thus, our goal then becomes to model $p(x)$, the probability of death given an exposure to a CS_2 concentration equal to x.

As mentioned above, the model

$$p(x) = \beta_0 + \beta_1 x$$

will NOT work since $p(x)$ must take values between 0 and 1.

The Logistic Regression Model

One of the standard ways of modeling the data in this situation is to use the *logistic regression function*:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}, \tag{13}$$

where β_0 and β_1 are the regression parameters of interest (similar to the intercept and slope parameters of a simple linear regression).

Note that by construction, the values of $p(x)$ in (13) are constrained to lie between 0 and 1 as required.

Some notes about the logistic regression function

1. If $\beta_1 = 0$ then $p(x)$ is constant. That is, the probability of success will not depend on x.
2. $p(x)$ is an increasing function if $\beta_1 > 0$ and a decreasing function if $\beta_1 < 0$.
3. For a probability p, the function $p/(1-p)$ is called the *odds ratio*
4. One can show that

   $$\ln[p(x)/(1-p(x))] = \beta_0 + \beta_1 x.$$

   The function $\ln[p(x)/(1-p(x))]$ is called the *logit* of $p(x)$:

   $$\text{logit}(p(x)) = \ln[p(x)/(1-p(x))] = \beta_0 + \beta_1 x,$$

   and is known as the *link* function for logistic regression. Note that the logit of $p(x)$ yields the usual linear regression expression.

Estimation in Logistic Regression – Maximum Likelihood.

The natural question now becomes – how do we use the data to estimate the parameters β_0 and β_1 in the logistic regression model?

The answer is to use the method of *maximum likelihood estimation.*

- The logic behind maximum likelihood estimation is to determine the values of β_0 and β_1 what make the observed data most likely to have occurred.
- The method of maximum likelihood estimation is used very broadly in many statistical applications besides logistic regression.
- Maximum likelihood estimators often perform better than other types of estimation procedures in terms of being the most efficient use of data. Hence, maximum likelihood estimation is a very popular method of estimation in statistical practice.

Maximum Likelihood Estimation continued ...

There do not exist formulas that give the estimates of β_0 and β_1 from a logistic regression in closed form as was the case in simple linear regression.

Instead, iterative algorithms are needed to determine the maximum likelihood estimates of β_0 and β_1.

Many software packages have the ability to fit logistic regression models.

Maximum Likelihood Estimation continued ...

The two most popular algorithms for finding the maximum likelihood estimates are the

- *Newton–Ralphson algorithm* and the
- *Iterated Re–Weighted Least Squares algorithm.*

Logistic Regression in SAS

We will illustrate the fitting of a logistic regression model using Proc logistic in SAS.

The SAS program for estimating the logistic regression model for the beetle mortality data is given below:

```
/**********************************************************************
Beetle Mortality by exposure to CS_2 (Bliss 1935)
Data: x = CS_2 concentration
      n = number of beetles exposed at each concentration.
      y = number of beetles killed
**********************************************************************/
options ls=76 nodate;
data beetles;
   title 'Beetle Mortality Data';
input concentration n deaths;
datalines;
49.1 59 6
53.0 60 13
56.9 62 18
60.8 56 28
64.8 63 52
68.7 59 53
72.6 62 61
76.5 60 60
;
run;
proc logistic data=beetles descending;
   model deaths/n=concentration;
run;
```

Logistic Regression in SAS continued ...

Here are some notes on the SAS program:

1. The model statement is very similar to the model statement using proc reg. The only difference is the "deaths/n". We need to let SAS know the number of trials n at each dose level. In some logistic regression applications, there may be only a single trial at each value of x in which case the a model statement of the form: "model y = x;" will work.

2. SAS is a bit unusual because it automatically predicts the odds for the lower value of the response $y = 0$ instead of $y = 1$. To get around this problem, we insert the word "descending" in the proc logistic statement. If "descending" was left off, then SAS would end up predicting the probability of survival instead of the probability of death in the beetle example.

Logistic Regression in SAS continued ...

Here is selected output from this SAS program :

```
             Beetle Mortality Data                           1

             The LOGISTIC Procedure

               Model Information

Data Set                        WORK.BEETLES
Response Variable (Events)      deaths
Response Variable (Trials)      n
Number of Observations          8
Model                           binary logit
Optimization Technique          Fisher's scoring

                Response Profile

       Ordered     Binary          Total
         Value     Outcome     Frequency

             1     Event             291
             2     Nonevent          190

            Model Convergence Status

 Convergence criterion (GCONV=1E-8) satisfied.

             Model Fit Statistics

                                  Intercept
                   Intercept         and
    Criterion        Only        Covariates

    AIC              647.441        372.623
    SC               651.617        380.975
    -2 Log L         645.441        368.623
```

```
                Testing Global Null Hypothesis: BETA=0

        Test                 Chi-Square       DF     Pr > ChiSq

        Likelihood Ratio       276.8175        1         <.0001
        Score                  226.1084        1         <.0001
        Wald                   136.0109        1         <.0001


               Analysis of Maximum Likelihood Estimates

                                        Standard          Wald
Parameter        DF     Estimate           Error    Chi-Square    Pr > ChiSq

Intercept         1     -14.8230          1.2896      132.1210        <.0001
concentration     1       0.2494          0.0214      136.0109        <.0001


                          Beetle Mortality Data

                         The LOGISTIC Procedure

                         Odds Ratio Estimates

                                 Point          95% Wald
           Effect             Estimate      Confidence Limits

           concentration         1.283       1.231       1.338


     Association of Predicted Probabilities and Observed Responses

           Percent Concordant     87.0    Somers' D    0.802
           Percent Discordant      6.8    Gamma        0.856
           Percent Tied            6.3    Tau-a        0.384
           Pairs                 55290    c            0.901
```

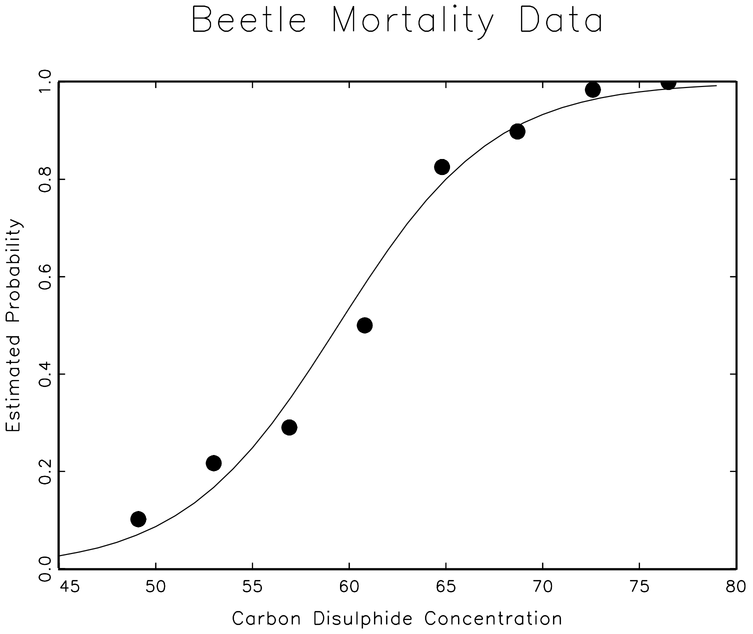

Figure 30: Proportion of beetle deaths versus Carbon disulphide concentration along with the estimated logistic regression curve.

SAS Output for Proc Logistic

SAS gives the following maximum likelihood estimates for the intercept and slope: $\hat{\beta}_0 = -14.8230$ and $\hat{\beta}_1 = 0.2494$, which yields the following estimated logistic regression model:

$$\hat{p}(x) = \frac{e^{-14.8+0.25x}}{1+e^{-14.8+0.25x}}.$$

Plugging in a value for x, the CS_2 concentration, yields the estimated proportion of beetles that would die after exposure for five hours.

A plot of the actual proportions and the estimated logistic regression curve is plotted in Figure 30.

SAS Output for Proc Logistic continued ...

The “percent concordant” and “percent discordant” in the SAS output is explained by the following:

These statistics are computed by first looking at all possible pairs of observations (beetles) in which the outcome differs (died or survived).

For each possible pair of beetles, the probability of death is estimated.

If the beetle that actually died has the higher estimated probability of death, then the pair is called *concordant*.

Otherwise, it is called *discordant*. We desire logistic regression models with a high percentage of concordance and a low percentage of discordance.

Inference for the Slope.

The primary question of interest in a simple logistic regression (i.e. a logistic regression with only one predictor variable) is if the slope β_1 differs from zero.

Recall that if $\beta_1 = 0$, then the predictor has no bearing on the probability of success or failure.

In the beetle example, it is quite clear that as the CS_2 concentration increases, the probability of mortality increases as well.

A simple method to test significance of β_1 is to use *Wald's* test.

For large sample sizes, maximum likelihood estimators such as $\hat{\beta}_1$ from logistic regression typically follow normal distributions approximately.

If we want to test the hypothesis

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0,$$

then the statistic

$$z = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)},$$

will follow a standard normal distribution approximately when H_0 is true and the sample size is fairly large.

If H_0 is false, then $|z|$ will tend to be large and we can compare z to critical values of the standard normal distribution.

Note that if you square a standard normal random variable, you obtain a chi-squared random variable on one degree of freedom.

Thus, alternatively, we could compare z^2 to critical values of the chi-squared distribution on one degree of freedom.

If H_0 is false (i.e. $\beta_1 \neq 0$), the z^2 will tend to be large.

SAS automatically computes the p-value using Wald's test comparing z^2 to the chi-square distribution on one degree of freedom.

Inference for the Slope continued...

In the beetle example, $\hat{\beta}_1 = 0.2494$ and $\hat{se}(\hat{\beta}_1) = 0.0214$.

This standard error is found using the theory of maximum likelihood (the theory of which is not covered in these notes).

The z test statistic then is

$$z = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} = \frac{0.2492}{0.0214} = 11.6449,$$

and $z^2 = 135.603$.

The probability of observing a value of $z^2 \geq 135.603$ when the true value of β_1 is zero, is less than 0.0001.

In other words, the p-value from Wald's test is $p < 0.0001$ indicating that there is very strong evidence that the probability of beetle mortality increases with increasing dosage of CS_2.

Interpreting the Logistic Regression Coefficients.

The estimated slope in the beetle mortality example was $\hat{\beta}_1 = .2492$.

What does this tell us about the relationship between dose of CS_2 and mortality?

Recall that in a simple *linear* regression, the slope β_1 measures the change in mean response y for a one unit change in x.

The interpretation of the slope β_1 in a logistic regression however is not so straight forward.

To help understand and interpret a logistic regression slope, we will first look at a very simple example where the predictor variable x is dichotomous, only two levels.

Prostate Cancer Example. Data on $n = 53$ prostate cancer patients (Collet 1991) was collected. A laparectomy was performed on each patient to determine if the cancer had spread to surrounding lymph nodes.

The goal is to determine if the size of the tumor can predict whether or not the cancer has spread to the lymph nodes. Define a indicator regressor variable x as

$$x = \begin{cases} 0 & \text{for small tumors} \\ 1 & \text{for large tumors} \end{cases}$$

and

$$y = \begin{cases} 0 & \text{lymph nodes not involved} \\ 1 & \text{lymph nodes involved} \end{cases}.$$

The maximum likelihood estimators from fitting a logistic regression of y on x are

$$\hat{\beta}_0 = -1.4351 \quad \text{and} \quad \hat{\beta}_1 = 1.6582.$$

The estimated probability that the cancer will spread to the lymph nodes is

$$\begin{aligned} \hat{p}(1) &= \frac{e^{-1.435+1.658}}{1 + e^{-1.435+1.658}} = .5555 \text{ for large tumor patients} \\ \hat{p}(0) &= \frac{e^{-1.435}}{1 + e^{-1.435}} = .1923 \text{ for small tumor patients} \end{aligned}$$

ODDS: For large tumor patients (i.e. $x = 1$), the (estimated) *ODDS* that the cancer will spread to the lymph nodes is

$$\frac{\hat{p}(1)}{1 - \hat{p}(1)} = .5555/(1 - .5555) = 1.25.$$

The interpretation of the odds is: For *large tumor patients* the probability the cancer spreads to lymph nodes is about one and a quarter times higher than the probability the cancer will not spread to the lymph nodes.

For small tumor patients (i.e. $x = 0$), The (estimated) odds that the cancer has spread to the lymph nodes is

$$\frac{\hat{p}(0)}{1 - \hat{p}(0)} = .1923/(1 - .1923) = .238.$$

Thus, for small tumor patients, the probability the cancer spreads is only about 1/4 ($\approx .238$) of the probability the cancer will not spread.

Or, inverting things, we could say that the probability the cancer will not spread is about 4 times higher than the probability it will spread for small tumor patients.

The **ODDS RATIO** is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$:

$$\text{ODDS RATIO } = \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}.$$

One can do some algebra to show:

$$\text{ODDS RATIO } = e^{\beta_1}. \tag{14}$$

The logistic regression slope β_1 is related to the odds ratio by:

$$\textbf{LOG–ODDS } = \ln(\text{odds ratio}) \ = \beta_1.$$

For the prostate cancer example we found that

$$\hat{\beta}_1 = 1.658.$$

Thus, the odds–ratio (of those with large tumor to those with small tumors) is estimated to be

$$e^{\hat{\beta}_1} = e^{1.658} = 5.25.$$

The interpretation of the odds–ratio in this example is that the odds the cancer spreads is about 5 times greater for those with large tumors compared to those with small tumors.

Caution: This does not mean that those with large tumors are five times more likely to have the cancer spread than those with small tumors (see *relative risk*).

Understanding the Odds Ratio.

To help understand the odds ratio, suppose there are 100 patients with small tumors.

The odds that the cancer spreads for small tumor patients is approximately 1/4.

Thus, we would expect that for every 20 patients who have had the cancer spread, there will be 80 patients for whom the cancer has not spread: $20/80 = 1/4$.

Now the odds ratio is approximately equal to 5. That is, the odds the cancer has spread is about 5 times higher for large tumor patients than for small tumor patients:

$$\begin{aligned}\text{Odds for Large Tumor Patients} &= 5 \times (\text{Odds for Small Tumor patients}) \\ &= 5 \times \frac{20}{80} \\ &= \frac{100}{80}\end{aligned}$$

which can be interpreted as: out of 180 patients with large tumors, we would expect to see 100 patients who have had the cancer spread compared to only 80 patients who have not had the cancer spread.

If we interpret this for a total of 100 patients with large tumors (instead of 180 patients) we would expect to see about 55 patients where the cancer has spread compared to about 45 for whom the cancer has not spread.

Independence. Note that if the odds ratio equals 1, then the odds of the cancer spreading is the same for large and small tumor patients.

In other words, the odds of the cancer spreading does not depend on the size of the tumor.

Recall

$$\text{ODDS RATIO } = e^{\beta_1}.$$

Thus, if the odds ratio equals 1, this implies that the logistic regression slope $\beta_1 = 0$ since $e^0 = 1$.

Relative Risk.

The relative risk is defined as:

$$\text{Relative Risk} = \frac{p(1)}{p(0)}.$$

In the prostate example, the relative risk is estimated to be

$$\frac{\hat{p}(1)}{\hat{p}(0)} = \frac{0.5555}{0.1923} = 2.89$$

which is quite a bit less than the odds ratio of 5.25.

The relative risk (2.89) means that the probability of the cancer spreading for the large tumor patients is almost 3 times higher compared to the small tumor patients.

Interpreting the slope when x is continuous.

Returning to the beetle mortality example, recall that $\hat{\beta}_1 = 0.2494$.

If we increase the dose by one unit, then the odds ratio for a unit change in x is:

$$\frac{p(x+1)/(1-p(x+1))}{p(x)/(1-p(x))} = e^{\beta_1}.$$

The estimated odds ratio for an increase of one unit of CF_2 concentration is

$$e^{\hat{\beta}_1} = e^{0.2494} = 1.283,$$

indicating that the odds of dying is about 1.283 times greater for each additional unit increase in CF_2.

Multiple Logistic Regression.

In the classical regression setting we have seen how to include more than one regressor variable in the regression model.

Multiple regressors can also be incorporated into the logistic regression model as well.

Suppose we have p regressor variables $x_1, x_2, \ldots, x_p$.

Then we can generalize (13) and define a multiple logistic regression function:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \tag{15}$$

and the logit of $p(x)$ is

$$\text{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Maximum likelihood is generally used to estimate the β_j's and their standard errors for the multiple logistic regression model as was done for the *simple* logistic regression.

Tests for Significance for Multiple Logistic Regression.

Just as in the case for multiple regression, we can also perform statistical tests to determine if subsets of the the regression coefficients differ from zero.

The testing procedures for both multiple regression and multiple logistic regression are based on the same principal: fit the full model and the reduced model and compare the two fits.

If the reduced model does nearly as good a job as the full model, then the reduced model is preferred.

The actual mechanics of the testing procedure in multiple logistic regression differ from that of multiple regression though which we now discuss.

Generalized Linear Model and Deviance.

The logistic regression model is a special case of a *generalized linear model.*

For generalized linear models, a statistic called the **deviance** is computed which measures how close the predicted values from the fitted model match the actual values from the raw data.

Maximum likelihood is generally used to estimate the parameters for generalized linear models.

The *likelihood* is simply the probability density computed from the observed data values with the parameters replaced by their estimates.

An extreme case is to fit a *saturated* model where the number of parameters equals the number of observations.

One of the fundamental goals of statistics is to determine a simple model with as few parameters as possible.

The saturated model has as many parameters as observations and hence it provides no simplification at all.

However, we can compare any proposed model to the saturated model to determine how well the proposed model fits the data.

The deviance D is defined as

$D = 2[\text{log-likelihood(saturated model)} - \text{log-likelihood(proposed model)}]$.

If the proposed model is correct, then the sampling distribution of the deviance D follows a chi-squared distribution on $n - p - 1$ degrees of freedom approximately.

This is an asymptotic result meaning that it is valid as the sample size n goes to infinity.

Hypothesis Tests for Regression Coefficients in Logistic Regression

However, this asymptotic result is usually not of much use.

Instead, interest lies in comparing *nested* models – comparing reduced models to a full model. The principal is the same as it was for multiple regression.

Consider the full model

$\text{logit}(p(x_1, x_2, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_p x_p,$

and we want to test the null hypothesis

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0,$$

versus the alternative hypothesis that at least one of these coefficients differs from zero.

If H_0 is true, then the regressor variables $x_{q+1}, \ldots, x_p$ are redundant in the full model and can be dropped from the model.

Hypothesis Tests for Regression Coefficients in Logistic Regression continued ...

In order to test H_0 in practice, all one needs to do is fit the full model and the reduced model and compare their respective deviances.

The test statistic is:

$$X^2 = D_{\text{Reduced}} - D_{\text{Full}}.$$

- If H_0 is true, then the test statistic X^2 has an approximate chi-squared distribution (provided the sample size is sufficiently large) whose degrees of freedom is equal to the difference in the number of parameters between the full and reduced models: $p - q$.
- If H_0 is false, then the test statistic tends to be too large to be considered as deriving from the chi-squared distribution on $p - q$ degrees of freedom. That is, if X^2 is too big, reject H_0.
- If we are testing at a level of significance α, then we reject H_0 if $X^2 > \chi_{\alpha,p-q}$, the α critical value of the chi-squared distribution on $p - q$ degrees of freedom.

Wald Statistics.

Alternatively, one can also test for significance of individual regression coefficients (i.e. $H_0 : \beta_j = 0$) by computing *Wald* statistics which are similar to the partial t-statistics from classical regression:

$$w_j = \frac{\hat{\beta}_j}{\hat{\text{se}}(\hat{\beta}_j)}.$$

Under the null hypothesis that $\beta_j = 0$, the Wald test statistic w_j follows approximately a standard normal distribution (and its square is approximately a chi-square on one-degree of freedom).

In order to illustrate these testing ideas, we now illustrate with an example.

A Multiple Logistic Regression Example.

Example. Inbreeding occurs naturally within plant populations. A study was conducted to study the effect of plant inbreeding on the resistance and tolerance of the plant to native herbivores in Napa County, California using the plant Yellow Monkeyflower.

The response variable y of interest in this study was whether or not the plant produced flowers (0 for no and 1 for yes).

Flower production is needed for reproduction.

The primary explanatory variable x_1 of interest was the indicator variable indicating whether or not the plant was inbreed ($x_1 = 1$) or crossbreed ($x_1 = 0$).

Two other covariates were also recorded: Herbivore damage to the plants due to spittlebugs, adult and larval beetles, slugs, deer, etc. was recorded as a percentage on a discretized scale (12 categories); and the dried above ground biomass of the plant (in grams).

Example continued ...

The data (compliments of C. Ivey) from one part of the study is shown in the following table.

Damage	x_1	y	Biomass
0.2138	0	1	0.0525
0.2138	0	1	0.0920
0.2138	1	1	0.0239
0.2138	0	0	0.0149
0	0	0	0.0410
0.3907	0	0	0.0264
0.2138	0	1	0.1370
0.2138	1	1	0.0280
0.2138	1	1	0.0248
0.2138	1	1	0.0892
0.2138	1	0	0.0123
0	0	0	0.0105
0	0	0	0.0081
0.2138	1	0	0.0234
0.2138	0	1	0.0349
0.2138	0	1	0.0251
0.2138	0	1	0.1199
0.2138	0	0	0.0187
0	0	1	0.0057
0.2138	1	0	0.0219
0	0	0	0.0062
0.6278	1	0	0.0093
0.3907	0	1	0.0123
0.2138	0	1	0.4092
0	0	1	0.0111
0	1	0	0.0185
0.3907	1	0	0.0158
0	1	0	0.0059
0.2138	0	1	0.0874
0.2138	0	0	0.0277
0.2138	0	1	0.0139
0	1	1	0.0115
0	0	0	0.0023
0.2138	1	1	0.0630
0.2138	0	1	0.2353
0	1	0	0.0040
0	1	0	0.0201
0.2138	1	1	0.0876
0.7303	0	0	0.0075
0.5178	0	0	0.0568
0.3907	0	0	0.0152
0.2138	0	1	0.0375
0.2138	1	1	0.1454
0	0	0	0.0104
0.3907	0	1	0.0415
0.2138	1	1	0.0219
0.3907	0	1	0.3934
0.2138	0	0	0.0210
0	1	1	0.0506
0.3907	0	0	0.0934

Example continued ...

The regressor variables biomass and damage are thought to explain much of the variability in the response and they are included as covariates.

The regressor of primary interest is x_1, the indicator for whether or not the plant is inbred or not.

The logit for the full model is

$$\text{logit}(p(x_1, \text{DAMAGE}, \text{BIOMASS})) = \beta_0 + \beta_1 x_1 + \beta_2 \text{DAMAGE} + \beta_3 \text{BIOMASS}.$$

(Note that this model does not contain any interaction terms – none of them were significant.)

The SAS output from running proc logistic for the full model is given below:

```
                       Model Fit Statistics
                                           Intercept
                             Intercept         and
              Criterion        Only        Covariates

              AIC               71.235         60.463
              SC                73.147         68.111
              -2 Log L          69.235         52.463

              Testing Global Null Hypothesis: BETA=0

      Test                 Chi-Square       DF      Pr > ChiSq

      Likelihood Ratio        16.7718        3          0.0008
      Score                    8.4458        3          0.0376
      Wald                     6.8681        3          0.0762

                       The LOGISTIC Procedure

             Analysis of Maximum Likelihood Estimates

                               Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -1.1230      0.6796        2.7308        0.0984
x1            1      0.3110      0.6821        0.2079        0.6484
biomass       1     41.3349     15.9402        6.7243        0.0095
damage        1     -1.8555      2.1421        0.7503        0.3864
```

The deviance for this full model fit is $D_{\text{full}} = 52.463$.

Note that the Wald test for significance of the coefficients for x_1 and Damage yield p-values of $p = 0.6484$ and $p = 0.3864$ indicating that both of these regressors appear to be redundant in the full model.

A reduced logistic regression model was fit using only Biomass to the data yielding the following SAS output:

```
                           Model Fit Statistics

                                                   Intercept
                                    Intercept         and
                  Criterion           Only        Covariates

                  AIC                 71.235          57.579
                  SC                  73.147          61.403
                  -2 Log L            69.235          53.579

                  Testing Global Null Hypothesis: BETA=0

          Test                   Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio          15.6555        1         <.0001
          Score                      7.8423        1         0.0051
          Wald                       6.6165        1         0.0101

                 Analysis of Maximum Likelihood Estimates

                                         Standard          Wald
    Parameter     DF      Estimate          Error    Chi-Square     Pr > ChiSq

    Intercept      1       -1.2556         0.5129        5.9934         0.0144
    biomass        1       37.9157        14.7402        6.6165         0.0101
```

The deviance for this reduced model is $D_{\text{reduced}} = 53.579$.

A likelihood ratio test of

$$H_0 : \beta_1 = \beta_2 = 0$$

versus the alternative that at least one of these two coefficients differs from zero yields a test statistic of

$$X^2 = D_{\text{reduced}} - D_{\text{full}} = 53.579 - 52.463 = 1.117.$$

This value of the test statistic should be compared to the chi-squared distribution on two degrees of freedom because the difference in the number of parameters between the full and reduced models is two.

The probability that a chi-square random variable on two degrees of freedom takes the value 1.117 or greater is approximately 0.5721.

That is, the p-value for the likelihood ratio test is $p = 0.5721$.

Thus we conclude that there is insufficient evidence that the coefficients β_1 and β_3 differ from zero.

This allows us to settle on a logistic regression model involving only the biomass as a predictor.

The estimated probability of a plant producing a flower for a given biomass is

$$\frac{\exp\{-1.2556 + 37.9157\text{Biomass}\}}{1 + \exp\{-1.2556 + 37.9157\text{Biomass}\}}.$$

From this analysis, it appears that whether or not this particular plant species was inbred does not effect its ability to reproduce, even when accounting for the plant's size and the damage sustained from herbivores.

It is useful to note that when a logistic regression is run using only the indicator for inbreeding (x_1) in the model, the regressor x_1 is still not significant ($p = 0.9442$).

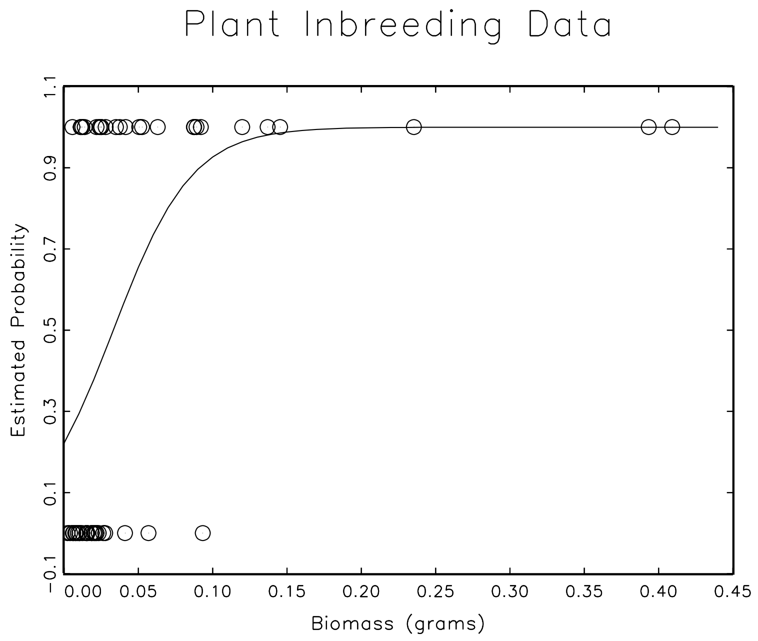

Figure 31: A logistic regression plot for the Yellow Monkey flower plant. The response is 1 if the plant produced at least one flower and 0 if it did not produce a flower. The response is plotted against biomass plant. Also shown is the estimated logistic regression curve.

A plot of the response y versus biomass along with the fitted logistic regression curve is shown in Figure 31

Generalized Linear Models

The classical linear model which encompasses linear regression models as well as the usual analysis of variance models for comparing treatment means can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon. \tag{16}$$

Factorial ANOVA models result when the regressor variables are indicator variables and we obtain regression models when the regressors are continuous or they are a mix of continuous/indicator variables.

The response y in (16) is a linear function of the model coefficients, the β_j's and hence the name *linear model.*

The error in (16) is often assumed to have a normal distribution. (16) can be generalized to handle situations where the error is non-normal.

In the generalized setting, the response variable y is still related to the regressors through a linear combination $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, except that the relation may not be direct as in (16). Instead, the response is linked to the linear combination $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ by a *link* function.

In the logistic regression model, the link function was the logit.

Recall that in the logistic regression setup, the response y is a zero-one Bernoulli random variable whose success probability $p(x_1, \ldots, x_p)$ depends on a set of regressors $\boldsymbol{x} = (x_1, \ldots, x_p)$ by means of

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}, \tag{17}$$

and the logit link function is defined as:

$$\text{logit}(p(x_1, \ldots, x_p)) = \ln[p(x_1, \ldots, x_p)/(1-p(x_1, \ldots, x_p))] = \beta_0+\beta_1 x_1+\cdots+\beta_p x_p.$$

In the classical regression model (16), the link function is simply the identity function.

Logistic regression and (16) are two examples of *generalized linear models.*

Other examples of generalized linear models are Poisson regression, log-linear models, survival analysis models. The three main principals of generalized linear models are:

1. We have a sample $Y_1, \ldots, Y_n$ of independent response variables from an *exponential family.* Basically, the exponential family of probability distributions are those whose density function can be expressed as an exponential function and the *support* of the density (i.e. the points where the density is greater than zero) does not depend on the parameters of the model (the normal and binomial distributions are both in the exponential family).

2. There exist a linear predictor $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ of the response variable Y.

3. The mean response depends on the linear predictor through a link function g. This mean response that depends on the values of x_j's is known as a conditional expectation: $\mu_{y|\boldsymbol{x}} = E[Y|x_1, \ldots, x_p]$ and

   $$g(\mu_{y|\boldsymbol{x}}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

   For (16), the function g is the identity $g(x) = x$ and for logistic regression, $p(x)$ is the conditional mean of the Bernoulli response Y given values of the regressor variables $\boldsymbol{x}$.

Maximum Likelihood Estimation for Generalized Linear Models

The statistical estimation and testing for generalized linear models is usually done using likelihood theory, as was the case for logistic regression.

Once a distribution for the response is decided upon, the likelihood equations can be written out and the usual maximum likelihood estimation procedure is then followed.

Additionally, model testing is typically carried out using likelihood ratio tests based on differences in deviances as we saw with the multiple logistic regression example.

Using the theory of maximum likelihood estimation, the test statistics under the null hypotheses based on deviances follow chi-square distributions for large sample sizes.

Further Topics for Generalized Linear Models

There exist many additional issues for generalized linear models including topics on

- model building,
- goodness of fit,
- over-dispersion.

Details on these topics can be found in books that focus on generalized linear models.

The classic text on the subject is McCullagh and Nelder's 1989 book *Generalized Linear Models.*

Poisson Regression.

The final generalized linear model we shall consider is the Poisson Regression Model.

The Poisson probability distribution is one of the most important probability distributions in practice. First, we review the Poisson probability distribution.

The Poisson Probability Distribution.

The Poisson distribution is useful for modeling count data.

Example. A study was done on occurrences of endemic gastroenteritis as measured by hospitalizations, emergency room, physician visits, and long-term care visits.

One of the interests lies in associating gastroenteritis with water quality measures. If we let Y denote the number of occurrences of this illness over a specific period of time, then a Poisson distribution may be a reasonable way of modeling the distribution of Y.

Note that if Y equals the number of cases of gastroenteritis, then Y can assume values $0, 1, 2, \ldots$.

The Poisson distribution is parameterized by a rate parameter $\mu > 0$ and the probability density function $f(y)$ is given by

$$f(y) = P(Y = k) = e^{-\mu}\frac{\mu^k}{k!}, \text{ for } k = 0, 1, \ldots.$$

The Poisson Probability Distribution continued ...

The mean and variance of a Poisson random variable equals μ.

The Poisson probability model can be derived theoretically in situations where the following three conditions hold:

1. The occurrences of the event of interest in non-overlapping "time" intervals are independent.
2. The probability two or more events in a small time interval is small, and
3. The probability that an event occurs in a short interval of time is proportional to the length of the time interval.

One can show that the Poisson distribution belongs to the exponential class of probability distributions and hence, generalized linear model approach can be used to relate a Poisson response to predictor variables.

Poisson Regression

Suppose that a Poisson response y depends on a set of regressor variables $x_1, \ldots, x_p$. Because the mean μ must be positive, a natural way of modeling the conditional mean response of y is

$$\mu = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}. \tag{18}$$

If we take the natural logarithm of each side of this equation we obtain

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

which indicates that the natural logarithm is the link function for a Poisson regression. This is an example of a *log-linear* model.

If a data set consists of measurements on a response variable y that correspond to counts as well as measurements on a set of regressor variables, one may want to simply model this using the classical regression model.

However, if a standard regression is used to model count data, one will often encounter problems with *heteroscedasticity* which means unequal error variances.

Recall that one of the assumptions in the standard regression setup is that the variance of the error distribution is constant, regardless of the values of the regressors.

However, as noted above, the variance of a Poisson random variable is μ.

The Poisson distribution is unique in that its mean and variance are equal to each other. If the Poisson mean is related to regressors as in (18), then

$$\text{var}(Y) = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p},$$

clearly indicating that the variance depends on the values of the regressors as well and hence the equal variance assumption will not hold for Poisson count data.

Poisson Regression Example.

The following example illustrates the Poisson regression model.

Example. The Center for Disease Control (CDC) has recorded data on the number of lyme disease cases in the United States since the early 1980's (as reported on their website). The data is in the following table:

Year	Cases
1982	491
1983	595
1984	1518
1985	2748
1986	1387
1987	2392
1988	4882
1989	8803
1990	7943
1991	9470
1992	9908
1993	8257
1994	13043
1995	11700
1996	16455
1997	12801

Because this is count data, a natural way to model the trend in the number of cases over this period of time is a Poisson regression. The model we shall consider is

$$y_i = e^{\beta_0+\beta_1 x_i} + \epsilon_i,$$

where y_i is the number of cases in the ith year and x_i is the year.

Poisson Regression Example continued ...

In order to fit this model, PROC GENMOD in SAS was used. This SAS procedure can be used to fit and test a wide variety of generalized linear models. The SAS code for fitting the Poisson regression model using Proc Genmod is

```
options ls = 76 nodate;
data lyme;
infile 'lyme.dat';
input year y;
title 'Lyme Disease data';
run;
proc genmod;
     model y=year/dist=poisson
                  link=log;
run;
```

In order to use proc genmod to fit a generalized linear model, the user needs to specify the link function.

For Poisson regression, the link function is the logarithm. Other link functions that can be implemented in SAS are *identity* for normal errors, *logit* for logistic regression models, *probit* for an alternative link to the logit for a binomial response, *power*: $g(\mu) = \mu^{\lambda}$, for $\lambda \neq 0$, and *complementary log-log*: $g(\mu) = \log(-\log(1 - \mu))$.

We need to also specify the distribution using "dist = poisson." Other possible distributions the user can specify are: normal, binomial, gamma, inverse gamma.

Poisson Regression Example continued ...

The output from the proc genmod for the lyme disease data follows:

```
                         Lyme Disease data

                       The GENMOD Procedure

                         Model Information

                  Data Set              WORK.LYME
                  Distribution            Poisson
                  Link Function               Log
                  Dependent Variable            y
                  Observations Used            16

              Criteria For Assessing Goodness Of Fit

       Criterion                 DF          Value        Value/DF

       Deviance                  14     11009.9046        786.4218
       Scaled Deviance           14     11009.9046        786.4218
       Pearson Chi-Square        14     10624.5111        758.8937
       Scaled Pearson X2         14     10624.5111        758.8937
       Log Likelihood                  909565.7562

Algorithm converged.

                    Analysis Of Parameter Estimates

                          Standard       Wald 95%          Chi-
 Parameter  DF  Estimate     Error   Confidence Limits   Square  Pr > ChiSq

 Intercept   1  -311.050    1.5001  -313.990  -308.109  42994.5      <.0001
 year        1    0.1607    0.0008    0.1592    0.1621  45545.1      <.0001
 Scale       0    1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.
```

There were $n = 16$ observations and we are estimating two β parameters and therefore, there are $16 - 2$ degrees of freedom associated with the deviance. The estimated Poisson regression model is

$$\hat{y} = e^{-311.05+0.1607x}. \tag{19}$$

Poisson Regression Example continued ...

The coefficient β_1 for the year is highly significant. The estimated slope is $\hat{\beta}_1 = 0.1607$ with standard error 0.0008.

The Wald test statistic for testing if the slope differs from zero is $z = \hat{\beta}/\hat{\text{se}}(\hat{\beta}_1) = 0.1607/.0008 = 200.8$ which is off the chart when compared to the standard normal distribution indicating very strong evidence that the slope differs from zero.

SAS gives automatically a 95% confidence interval for β_1 as $(0.1592, 0.1621)$.

How do we interpret this confidence interval in the context of the lyme disease example? Recall that we are modeling the number of cases of lyme disease as an exponential function.

In the classic regression setup, the slope is interpreted as the mean change in the response for a unit change in the regressor. Because of the exponential relationship, it makes more sense to look at the ratio of the estimated response from one year to the next year.

Evaluating the ratio of (19) for year $x + 1$ to year x we obtain

$$e^{-311.05+0.1607(x+1)}/e^{-311.05+0.1607x} = e^{0.1607} \approx 1.174,$$

indicating that the model predicts that the number of cases increases by a factor of 1.174 per year.

With 95% confidence we estimate that the number of new lyme disease cases over this period of time increases by a factor of $e^{0.1592} = 1.1726$ to $e^{0.1621} = 1.1760$ from year to year.

Poisson Regression Example continued ...

Note that the fit of the Poisson regression curve does not appear to be very good, particularly for the earlier years where the fitted model is over-estimating the number of lyme cases.

Also note that (19) postulates an exponential growth rate in the number of new lyme cases which is certainly cause for concern.

An alternative approach to modeling the lyme disease data is to fit a Poisson regression model using the logarithm of the year as the regressor:

$$y_i = e^{\beta_0+\beta_1 \ln(x_i)} + \epsilon_i = \alpha x_i^{\beta_1} + \epsilon_i,$$

where $\alpha = e^{\beta_0}$.

Thus, a Poisson regression with $\ln(x)$ as a regressor results in a model where the predicted number of new lyme disease cases increases according to a polynomial trend as opposed to the much more scary prospect of an exponential growth rate.

The fitted model obtained by regressing the number of new cases on the logarithm of the year is

$$\hat{y} = 327.01x^{1.3945},$$

and is shown in Figure 33. Figure 33 shows a much better fit to the data than the original model (19).

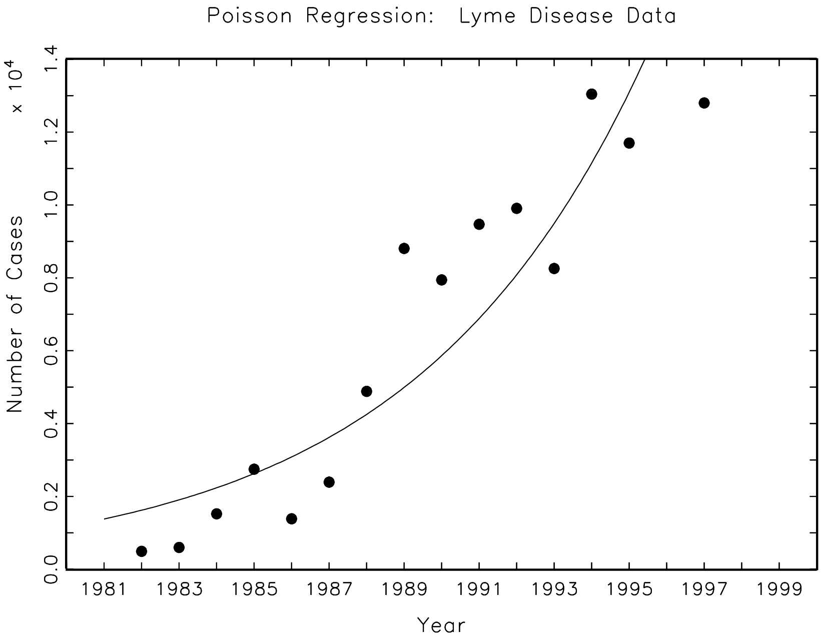

Figure 32: Number of cases of lyme disease for the years 1982 to 1997 as reported by the Centers for Disease Control. Also plotted is the Poisson regression estimated curve.

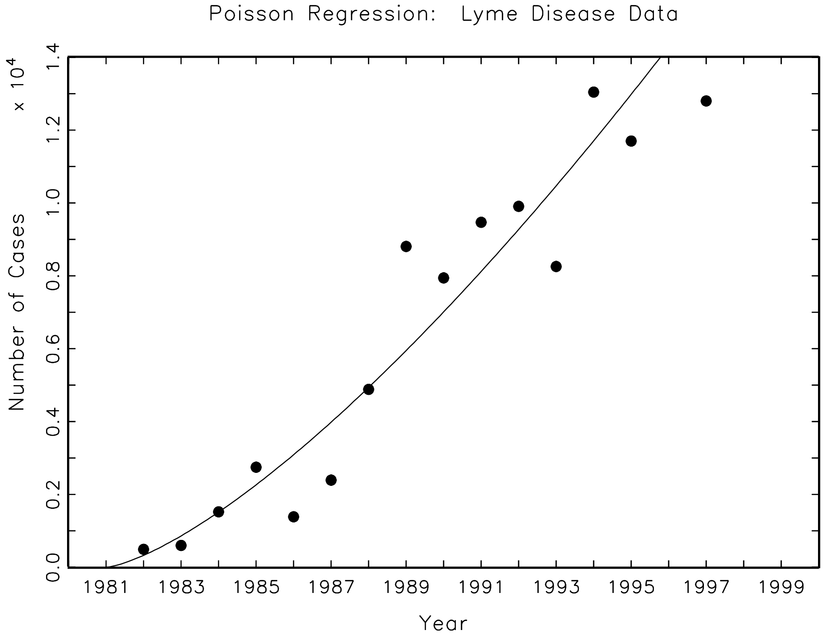

Figure 33: Number of cases of lyme disease for the years 1982 to 1997 as reported by the Centers for Disease Control. Also plotted is the Poisson regression estimated curve using logarithm of the year as the regressor variable.

References

Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17-21.

Arretxe, M., Heap, J. M., and Christofi, N. (1997), "The Effect of Toxic Discharges on ATP Content in Activated Sludge," *Environmental Toxicology and Water Quality*, **12**, 23–29.

Bailer, A. J., Walker, S. E., and Venis, K. J. (2000), "Estimating and Testing Bioconcentration Factors," *Environmental Toxicology and Chemistry*, **19**, 2338–2340.

Baird, S. J. (1996), "Nonfish species and fisheries interactions working group report," May 1996, New Zealand Fisheries Assessment Working Group Report 96/1, Ministry of Fisheries, Wellington, New Zealand.

Bates, D. M. and Watts, D. G., (1988), *Nonlinear Regression Analysis and Its Applications*, Wiley: New York.

Bliss, C. I. (1935) "The Calculation of the Dosage–Mortality Curve," *Annals of Applied Biology*, **22**, 134–167.

Box, G. E. P. (1980), “Sampling and Bayes’ Inference in Scientific Modeling and Robustness,” *JRSS A*, **143**, 383–430.

Collet, D. R. (1991), *Modeling Binary Data*, London: Chapman & Hall.

Efron, B. and Tibshirani, R. (1993), *An introduction to the bootstrap*, Chapman & Hall.

Eppright, E. S., Fox, H. M., Fryer, B. A., Lamkin, G. H., Vivian, V. M., and Fuller, E. S. (1972), “Nutrition of infants and preschool children in the north central region of the United States of America,” *World Review of Nutrition and Dietetics*, **14**, 269–332.

Greenfield , B.K., Davis, J.A., Fairey, R., Roberts, C., Crane, D.B., Ichikawa, G., and Petreas, M., (2003), ”Contaminant Concentrations in Fish from San Francisco Bay, 2000,” RMP Technical Report: SFEI Contribution 77. San Francisco Estuary Institute, Oakland, CA.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman & Hall.

Montgomery, D. C. and Peck, E. A. (1992), *Introduction to Linear Regression Analysis, 2nd Edition*, Wiley: New York.

Nalepa, T., Hartson, D., Buchanan, J., Covaletto, J., Lang, G., and Lozano, S. (2000), “Spatial variation in density, mean size and physiological condition of holartic amphipod *Diporeia* spp. in Lake Michigan,” *Freshwater Biology*, **43**, 107–119.