Chapter 17

Linear Regression

Linear regression is a statistical method for examining the relationship between two continuous variables, typically called Y and X. It is usually assumed there is a causal relationship between Y and X, with different values of X causing changes in Y. For this reason, Y is often called the **dependent variable** while X is the **independent variable** in the analysis. The variable X is sometimes under the control of the investigator, similar to a fixed effect in ANOVA, but can also be a random variable. For example, we might be interested in the effect of temperature on the growth rate of fish. Temperature might cause an increased growth rate, but clearly growth rate cannot influence temperature. This causal relationship is a distinguishing feature of regression as opposed to **correlation** analysis. Correlation is used to examine the **association** between two continuous variables and no causal direction is assumed (see Chapter 18). For example, we might be interested in the relationship between fish length and weight but there is no obvious causal relationship between the two variables.

Although linear regression assumes a different statistical model than ANOVA, there are a number of similarities in the estimation process and statistical tests for the two types. For example, both ANOVA and linear regression models use likelihood methods for parameter estimation and test construction, and employ F statistics to test various hypotheses. Both are examples of **general linear models**, in which the model parameters and error terms enter the model in an additive (linear) fashion.

What do the data look like for linear regression? As an example, we will use data from study on the southern pine beetle, *Dendroctonus frontalis* (Reeve et al. 1998). The study used cages to experimentally manipulate the

density of *D. frontalis* attacking pine trees. The independent or X variable in the study was the number of beetles added to the cages, while the dependent or Y variable was the number of attacks the beetles made through the bark into the tree (Table 17.1). Besides establishing the relationship between the two variables, there was also some interest in predicting the attack density as a function of the number of beetles added to the cage, for use in future studies. The notation Y_i and X_i refers to the values for the ith pair of numbers. For example, $Y_2 = 2.660$ and $X_2 = 1.000$. Fig. 17.2 shows there is a positive relationship between the two variables, with attack density (Y) increasing as more beetles are added to the cages (X).

Table 17.1: Example 1 - Observations from an experiment in which different numbers of the bark beetle *D. frontalis* were introduced into cages and the resulting attack density recorded (Reeve et al. 1998). Here Y is the attack density (attacks per 100 cm^2 of bark) while X is the number of beetles added ($\times 10^3$). Also shown are some preliminary calculations for the regression analysis.

i	Y_i	X_i	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})(X_i - \bar{X})$	$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \bar{Y})^2$	$(Y_i - \bar{Y})^2$
1	1.250	0.100	0.740	2.779	2.206	-0.956	0.914	5.176	10.440
2	2.660	1.000	0.002	-0.073	4.586	-1.926	3.711	0.011	3.316
3	7.330	2.000	1.081	2.962	7.231	0.099	0.010	7.563	8.116
4	1.600	1.250	0.084	-0.835	5.248	-3.648	13.305	0.588	8.301
5	2.620	0.500	0.212	0.856	3.264	-0.644	0.415	1.481	3.464
6	1.000	0.200	0.578	2.646	2.471	-1.471	2.162	4.042	12.118
7	4.340	1.500	0.291	-0.076	5.909	-1.569	2.461	2.038	0.020
8	5.230	0.750	0.044	-0.157	3.925	1.305	1.702	0.309	0.561
9	2.500	0.250	0.504	1.407	2.603	-0.103	0.011	3.528	3.925
10	3.250	0.500	0.212	0.567	3.264	-0.014	0.000	1.481	1.516
11	6.000	2.000	1.081	1.579	7.231	-1.231	1.516	7.563	2.307
12	4.750	1.500	0.291	0.145	5.909	-1.159	1.343	2.038	0.072
13	2.500	0.250	0.504	1.407	2.603	-0.103	0.011	3.528	3.925
14	8.750	2.000	1.081	4.439	7.231	1.519	2.307	7.563	18.223
15	6.000	1.000	0.002	0.060	4.586	1.414	1.998	0.011	2.307
16	5.000	0.500	0.212	-0.239	3.264	1.736	3.014	1.481	0.269
17	7.150	1.000	0.002	0.106	4.586	2.564	6.572	0.011	7.123
18	6.750	1.500	0.291	1.225	5.909	0.841	0.708	2.038	5.158
19	7.500	1.500	0.291	1.630	5.909	1.591	2.532	2.038	9.114
20	2.500	0.500	0.212	0.912	3.264	-0.764	0.584	1.481	3.925
21	5.000	2.000	1.081	0.540	7.231	-2.231	4.979	7.563	0.269
22	2.250	0.250	0.504	1.585	2.603	-0.353	0.124	3.528	4.978

i	Y_i	X_i	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})(X_i - \bar{X})$	$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \bar{Y})^2$	$(Y_i - \bar{Y})^2$
23	1.250	0.125	0.698	2.699	2.272	-1.022	1.045	4.879	10.440
24	4.750	1.000	0.002	0.011	4.586	0.164	0.027	0.011	0.072
25	4.500	0.250	0.504	-0.013	2.603	1.897	3.599	3.528	0.000
26	9.560	2.000	1.081	5.281	7.231	2.329	5.423	7.563	25.795
27	5.000	0.500	0.212	-0.239	3.264	1.736	3.014	1.481	0.269
$\sum$			11.798	31.203			63.486	82.528	146.014

17.1 Linear regression model

Suppose that we want to model the observations in studies like Example 1, where Y is observed for a number of X values. Let Y_i and X_i stand for the ith pair of values. The linear regression model takes the form

$$Y_i = \alpha + \beta X_i + \epsilon_i, \tag{17.1}$$

where α is the intercept and β the slope of a line, while $\epsilon_i \sim N(0, \sigma^2)$ (Searle 1971). Thus, the linear regression model represents the relationship between Y_i and X_i as a line on which random deviations due to natural variability (ϵ_i) are imposed.

For the ith pair of values, we have $E[Y_i] = \alpha + \beta X_i$ and $Var[Y_i] = \sigma^2$ using the rules for expected values and variances. Thus, $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$ for any X_i value. The behavior of the linear regression model can be illustrated by plotting this distribution across a range of X_i values . When β is positive, the mean of Y_i will increase as X_i increases (Fig. 17.1), while if β is negative the mean would decrease (not shown). The variance remains the same for all X_i. Note that the linear regression model has assumptions similar to the ANOVA models – the observations are assumed be normal and have the same variance.

The usual objectives in linear regression are to estimate the model parameters, especially the slope β, and then test whether the slope is different from zero. In particular, we will be interested in testing $H_0 : \beta = 0$. If a test of this hypothesis is significant this suggests there is some relationship (positive or negative) between Y and X. The alternative hypothesis can be written as $H_1 : \beta \neq 0$. It is also possible to test whether the intercept differs from zero although this is less common. We will discuss how these parameters are estimated and hypotheses tested in the next section.

17.2 Linear regression and likelihood

The maximum likelihood method can be used to estimate the parameters for regression models, similar to ANOVA models. Suppose we have n observations conforming to the linear regression model

$$Y_i = \alpha + \beta X_i + \epsilon_i. \tag{17.2}$$

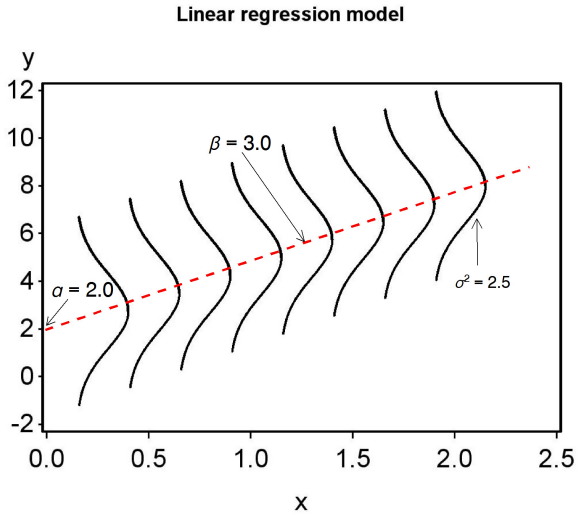

Figure 17.1: The linear regression model plotted across a range of X values, with $\alpha = 2.0$, $\beta = 3.0$, and $\sigma^2 = 2.5$.

This model has three parameters to estimate, namely α, β, and σ^2 (the variance of ϵ_i). What would the likelihood function be for these data? Consider the first observation in the *D. frontalis* cage experiment, for which $Y_1 = 1.250$ and $X_1 = 0.100$. For this observation, the model states that $Y_1 \sim N(\alpha + \beta X_1, \sigma^2)$, and so the likelihood would be

$$L_1 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(Y_1-(\alpha+\beta X_1))^2}{\sigma^2}} \qquad = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(1.250-(\alpha+\beta 0.100))^2}{\sigma^2}} \tag{17.3}$$

The likelihood L_i for the ith observation would be similar, and the overall likelihood is defined as their product:

$$L(\alpha, \beta, \sigma^2) = L_1 \times L_2 \times \ldots \times L_n. \tag{17.4}$$

Finding the maximum likelihood estimates involves maximizing this quantity with respect to the parameters α, β, and σ^2. Using some calculus to find the maximum, it can be shown that estimators of these parameters are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{17.5}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \tag{17.6}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - (\hat{\alpha} + \hat{\beta}X_i))^2}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}. \tag{17.7}$$

Here $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, the value of Y_i predicted by the model at X_i.

We can gain some insight into the estimation process by rearranging the likelihood function. It can be written in the form

$$L(\alpha, \beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(Y_i-(\alpha+\beta X_i))^2}{\sigma^2}}. \tag{17.8}$$

Now examine the terms in the sum, which are of the form $(Y_i - (\alpha + \beta X_i))^2$. Values of α and β that minimize these terms will make the overall likelihood larger, because of the negative sign in the exponent. The likelihood will reach its maximum when this sum is smallest. Thus, values of α and β that minimize

$$\sum_{i=1}^{n}(Y_i - (\alpha + \beta X_i))^2 \tag{17.9}$$

are the maximum likelihood estimates. These estimates are also called **least squares** estimates because they minimize the sum of these squared terms. In fact, we could directly estimate α and β using this method without recourse to likelihood (Searle 1971). The two methods yield the same results when the data have a normal distribution.

A likelihood ratio test for linear regression can be constructed as follows. Suppose we want to test $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, the latter implying a linear relationship between Y and X. The statistical model under H_0 would be

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{17.10}$$

$$= \alpha + \epsilon_i \tag{17.11}$$

because $\beta = 0$ under H_0. The statistical model under H_1 would be the full model including a slope term, namely

$$Y_i = \alpha + \beta X_i + \epsilon_i. \tag{17.12}$$

We would need to find the maximum likelihood estimates under both H_1 (see previous section) and H_0, as well as L_{H_0} and L_{H_1}, the maximum height of

the likelihood function under H_0 and H_1. We would then use the likelihood ratio test statistic

$$\lambda = \frac{L_{H_0}}{L_{H_1}}. \tag{17.13}$$

There is a one-to-one correspondence between $-2\ln(\lambda)$ and the statistic F_s used to test this null hypothesis (McCulloch & Searle 2001).

We can gain further insight into this test by defining various sum of squares and mean squares used to calculate F_s. In particular, we will define SS_{error}, $SS_{regression}$, and SS_{total} and their associated mean squares, which have functions similar to those in ANOVA. We will also summarize the calculations in an ANOVA table.

SS_{error} describes variation in the data around the regression line, or variation not explained by the model. It is defined as

$$SS_{error} = \sum_{i=1}^{n} \left(Y_i - (\hat{\alpha} + \hat{\beta}X_i)\right)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{17.14}$$

SS_{error} has $n-2$ degrees of freedom. We can therefore define

$$MS_{error} = \frac{SS_{error}}{n-2} = \hat{\sigma}^2. \tag{17.15}$$

Thus, MS_{error} is equivalent to $\hat{\sigma}^2$, the maximum likelihood estimate of σ^2, the same relationship as found in ANOVA. SS_{error} and MS_{error} will be small if the data lie on a straight line and large if the data are scattered around the line.

$SS_{regression}$ describes variation in the data explained by the regression model. It is defined as

$$SS_{regression} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{17.16}$$

and has one degree of freedom. We therefore have

$$MS_{regression} = \frac{SS_{regression}}{1} = SS_{regression}. \tag{17.17}$$

$SS_{regression}$ and $MS_{regression}$ will be large if the data have a strong positive or negative slope. To see this, recall that $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$. If the estimated slope

$\hat{\beta}$ is large, the values of $\hat{Y}_i$ will vary strongly as X_i changes and so generate a large sum of squares.

The total sum of squares is defined (as in ANOVA) to be

$$SS_{total} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{17.18}$$

and has $n-1$ degrees of freedom. There is also a familiar relationship among the different sums of squares, namely

$$SS_{regression} + SS_{error} = SS_{total}. \tag{17.19}$$

The likelihood ratio statistic used to test $H_0 : \beta = 0$ is defined as

$$F_s = \frac{MS_{regression}}{MS_{error}}. \tag{17.20}$$

Under H_0, F_s has an F distribution with $df_1 = 1$ and $df_2 = n-2$ the degrees of freedom. Given the definitions of $MS_{regression}$ and MS_{error}, we can see that F_s tends to be large when the data have a strong slope (the numerator of this expression) relative to the amount of scatter in the data (the denominator).

We can organize the different sum of squares and mean squares into an ANOVA table for linear regression. It lists the different sources of variation in the data (regression, error, and total), their degrees of freedom, as well as the F test. Table 17.2 shows the general layout for linear regression.

Table 17.2: General ANOVA table for linear regression, showing formulas for different mean squares and the F test.

Source	*df*	Sum of squares	Mean square	F_s
Regression	1	$SS_{regression}$	$MS_{regression} = SS_{regression}/1$	$MS_{regression}/MS_{error}$
Error	$n-2$	SS_{error}	$MS_{error} = SS_{within}/(n-2)$	
Total	$n-1$	SS_{total}		

Table 17.3: ANOVA table for the Example 1 data set, including a P value for the test.

Source	*df*	Sum of squares	Mean square	F_s	P
Regression	1	82.528	82.528	32.504	< 0.001
Error	25	63.486	2.539		
Total	26	146.014			

17.2.1 Sample calculation - $\hat{\beta}$, $\hat{\alpha}$, and F test

We will illustrate the above calculations using the Example 1 data set, where Y is *D. frontalis* attack density and X is the number of beetles added to the cage. We are interested in estimating the slope and intercept (β and α) of the relationship between the two variables, and then testing whether the slope is significantly different from zero ($H_0 : \beta = 0$).

The first step is to calculate the sample mean for both Y and X, and we obtain $\bar{Y} = 4.481$ and $\bar{X} = 0.960$. We then calculate $(X_i - \bar{X})^2$ for each value of X_i (see Table 17.1) and sum these values to obtain

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 11.798. \tag{17.21}$$

We then calculate the $(Y_i - \bar{Y})(X_i - \bar{X})$ for each pair of numbers and sum these to obtain

$$\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = 31.203. \tag{17.22}$$

The estimate of β can then be calculated, and we find

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{31.203}{11.798} = 2.645. \tag{17.23}$$

We can then estimate α using the formula

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 4.481 - 2.645(0.960) = 1.942. \tag{17.24}$$

The next step is to calculate the predicted values of Y_i using the formula $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, for each value of X_i (see Table 17.1). We then calculate $Y_i - \hat{Y}_i$ in another column, which contains the residuals for each observation. Squaring and summing the residuals, we find

$$SS_{error} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 63.486, \tag{17.25}$$

and

$$MS_{error} = \frac{SS_{error}}{n-2} = \frac{63.486}{27-2} = 2.539. \tag{17.26}$$

We next calculate a column consisting of $(\hat{Y}_i - \bar{Y})^2$ for each observation, then sum these values to obtain

$$SS_{regression} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = 82.528, \tag{17.27}$$

and so

$$MS_{regression} = SS_{regression}/1 = 82.528. \tag{17.28}$$

We are now in a position to calculate F_s, the statistic used to test $H_0 : \beta = 0$. We have

$$F_s = \frac{MS_{regression}}{MS_{error}} = \frac{82.528}{2.539} = 32.504. \tag{17.29}$$

Under H_0, F_s has an F distribution with $df_1 = 1$ and $df_2 = 27 - 2 = 25$ degrees of freedom. Using Table F, we find the $P < 0.001$. There is a highly significant effect of beetles numbers on the attack density of *D. frontalis* ($F_{1,25} = 32.504, P < 0.001$).

The last column in Table 17.1 calculates $(Y_i - \bar{Y})^2$, the components of SS_{total}. Summing these components we obtain $SS_{total} = 146.014$. It can also be calculated using the formula $SS_{regression} + SS_{error} = SS_{total}$. Table 17.3 shows the completed ANOVA table.

The observations for Example 1 and the fitted linear regression model are shown in Fig. 17.2. The estimation procedure (maximum likelihood or least squares) finds values of α and β that minimize the sum of the squared differences between the data points and the line. In particular, it minimizes the sum of the squared residuals, where the residuals are $Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$.

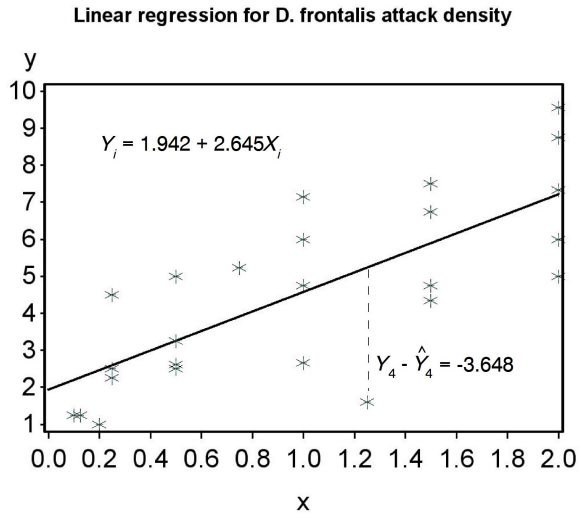

Figure 17.2: Linear regression model fitted to the Example 1 data, where Y is attack density and X is beetles added to the cages. The vertical dashed line shows the residual $Y_4 - \hat{Y}_4 = -3.648$ for the $i = 4$ observation.

17.3 Confidence and prediction intervals

In this section, we will examine confidence intervals for the parameters of the regression model, and for the mean value of Y_i at a given value of X_i. Like other confidence intervals, they provide a measure of the accuracy or reliability of an estimate, with wider intervals indicating lower accuracy (Chapter 9). Another type of interval for linear regression are **prediction intervals**. These are used to set limits for future Y_i values given some value of X_i. See Draper & Smith (1981) for further details.

The confidence interval for the slope β is based on $\hat{\beta}$, the maximum likelihood estimate of β, and the standard error of this estimate $s_{\hat{\beta}}$, given by the formula

$$s_{\hat{\beta}} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \tag{17.30}$$

where $\hat{\sigma}^2 = MS_{error}$. Note that $s_{\hat{\beta}}$ depends on the scatter of the data around the line ($\hat{\sigma}^2$) as well as the amount of variability in X_i. **A study using a larger range of X_i values will thus provide a more accurate estimate of** β, because it reduces $s_{\hat{\beta}}$. **Increasing the sample size n would also increase the accuracy**, by increasing the sum of squares in the denominator for $s_{\hat{\beta}}$.

It can be shown that the quantity

$$\frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} \tag{17.31}$$

has a t distribution with $n - 2$ degrees of freedom, the same as for MS_{error}. This fact can be used to derive a confidence interval for β. Using Table T, we first find a value of $c_{\alpha,n-2}$ for $n-2$ degrees of freedom such that the following equation is true:

$$P\left[-c_{\alpha,n-2} < \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} < c_{\alpha,n-2}\right] = 1 - \alpha. \tag{17.32}$$

Rearranging this equation we obtain

$$P\left[\hat{\beta} - c_{\alpha,n-2}s_{\hat{\beta}} < \beta < \hat{\beta} + c_{\alpha,n-2}s_{\hat{\beta}}\right] = 1 - \alpha. \tag{17.33}$$

It follows that the interval

$$(\hat{\beta} - c_{\alpha,n-2}s_{\hat{\beta}}, \hat{\beta} + c_{\alpha,n-2}s_{\hat{\beta}}) \tag{17.34}$$

is a $100(1-\alpha)\%$ confidence interval for β. The center of the confidence interval would be $\hat{\beta}$.

We may also want to test various null hypotheses concerning β. For example, we may want to test $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$, where β_0 takes some value of interest. Similar to the approach in Chapter 10, we would use the test statistic

$$T_s = \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}}. \tag{17.35}$$

Under H_0, T_s has a t distribution with $n-2$ degrees of freedom, and we would reject H_0 for sufficiently large values of this statistic. For $\beta_0 = 0$, this test is equivalent to the F test we developed earlier for $H_0 : \beta = 0$, and in fact $T_s^2 = F_s$. The t test is more general, however, because we can also test $H_0 : \beta = \beta_0$ for any value of β_0.

It is possible to derive similar t tests and confidence intervals for the intercept parameter α. The t test is most commonly used to test $H_0 : \alpha = 0$. If the test is significant this implies an intercept different from zero. We will let SAS handle the calculations here.

We can also derive a confidence interval for the theoretical mean of Y_i at a given X_i value. Recall that according to the linear regression model, $E[Y_i] = \alpha + \beta X_i$. Thus, Y_i has a mean of $\mu = \alpha + \beta X_i$ for any X_i value. The confidence interval is based on $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, the predicted value of Y_i at X_i. It also depends on the standard error $s_{\hat{Y}}$ of $\hat{Y}$, which is given by the formula

$$s_{\hat{Y}} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right]}. \tag{17.36}$$

Note that the standard error $s_{\hat{Y}}$ depends on the value of $(X_i - \bar{X})^2$, which is the squared distance of X_i from $\bar{X}$. The farther X_i is from $\bar{X}$, the larger the value of $s_{\hat{Y}}$.

Using methods similar to the confidence interval for β, it can be shown that a $100(1-\alpha)$ confidence interval for $\mu = \alpha + \beta X_i$ has the form

$$(\hat{Y}_i - c_{\alpha,n-2} s_{\hat{Y}}, \hat{Y}_i + c_{\alpha,n-2} s_{\hat{Y}}). \tag{17.37}$$

The interval will be broader for values of X_i far from $\bar{X}$ because $s_{\hat{Y}}$ will be larger. In other words, the precision of the confidence interval decreases with the distance from $\bar{X}$.

Another type of interval associated with regression are **prediction intervals**. Here, we are trying to find an interval that contains a defined percentage of future Y_i values for a given value of X_i, hence the name prediction interval. These are similar in form to the intervals for the theoretical mean $\mu = \alpha + \beta X_i$, but are always wider because you are trying to enclose a single future observation rather than a mean value.

The prediction interval is based on $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, the predicted value of Y_i at X_i, and the standard error $s_{\hat{Y}(1)}$ of $\hat{Y}_i$, which is given by the formula

$$s_{\hat{Y}(1)} = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right]}. \tag{17.38}$$

Note the additional term (1+) within the square brackets, which makes this standard error larger than $s_{\hat{Y}}$). It also depends on the value of $(X_i - \bar{X})^2$, and so the farther X_i is from $\bar{X}$, the larger the value of $s_{\hat{Y}(1)}$. It can be shown that a $100(1-\alpha)$ prediction interval for a single future Y_i has the form

$$(\hat{Y}_i - c_{\alpha,n-2} s_{\hat{Y}(1)}, \hat{Y}_i + c_{\alpha,n-2} s_{\hat{Y}(1)}). \tag{17.39}$$

17.3.1 Sample calculation - confidence and prediction intervals

We now illustrate the calculations for confidence intervals using the Example 1 data. We earlier found that $\hat{\beta} = 2.645$ and $\hat{\alpha} = 1.942$. To find a confidence interval for β, we first need to calculate $s_{\hat{\beta}}$. From Table 17.1, we see that $\sum_{i=1}^{n} (X_i - \bar{X})^2 = 11.798$, and we earlier calculated that $\hat{\sigma}^2 = MS_{error} = 2.539$. Inserting these quantities into the formula for $s_{\hat{\beta}}$, we find

$$s_{\hat{\beta}} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} = \sqrt{\frac{2.539}{11.798}} = 0.464. \tag{17.40}$$

For a 95% confidence interval and $\alpha = 0.05$, the confidence interval for β has the form

$$(\hat{\beta} - c_{0.05,n-2} s_{\hat{\beta}}, \hat{\beta} + c_{0.05,n-2} s_{\hat{\beta}}) \tag{17.41}$$

From Table T, with $\alpha = 0.05$ and $df = n - 2 = 27 - 2 = 25$, we find that $c_{0.05,25} = 2.060$. Inserting this value, $\hat{\beta} = 2.645$, and $s_{\hat{\beta}} = 0.464$ in this formula, we obtain

$$(2.645 - 2.060(0.464), 2.645 + 2.060(0.464)) \tag{17.42}$$

or

$$(1.689, 3.601). \tag{17.43}$$

We next find a confidence interval for the theoretical mean $\mu = \alpha + \beta X_i$ at $X_i = 0.5$. For this value of X_i, we have

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i = 1.942 + 2.645(0.5) = 3.265. \tag{17.44}$$

From Table 17.1 we have $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 11.798$, and earlier found that $\bar{X} = 0.960$ and $\hat{\sigma}^2 = MS_{error} = 2.539$. Inserting these quantities into the formula for $s_{\hat{Y}}$, we find that

$$s_{\hat{Y}} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right]} \tag{17.45}$$

$$= \sqrt{2.539 \left[\frac{1}{27} + \frac{(0.5 - 0.960)^2}{11.798} \right]} \tag{17.46}$$

$$= \sqrt{2.539 \left[0.037 + \frac{0.212}{11.798} \right]} \tag{17.47}$$

$$= 0.374. \tag{17.48}$$

For a 95% confidence interval and $\alpha = 0.05$, the confidence interval for the theoretical mean $\mu = \alpha + \beta X_i$ has the form

$$(\hat{Y} - c_{0.05,n-2} s_{\hat{Y}}, \hat{Y} + c_{0.05,n-2} s_{\hat{Y}}) \tag{17.49}$$

From Table T with $\alpha = 0.05$ and $df = n - 2 = 27 - 2 = 25$, we find that $c_{0.05,25} = 2.060$. Inserting this value, $\hat{Y} = 3.265$, and $s_{\hat{Y}} = 0.374$ in this formula, we find

$$(3.265 - 2.060(0.374), 3.265 + 2.060(0.374)) \tag{17.50}$$

or

$$(2.495, 4.035). \tag{17.51}$$

Lastly, we calculate a prediction interval for a single future observation Y_i at $X_i = 0.5$. For this value of X_i, we earlier calculated that

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i = 1.942 + 2.645(0.5) = 3.265. \tag{17.52}$$

We again have $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 11.798$, $\bar{X} = 0.960$ and $\hat{\sigma}^2 = MS_{error} = 2.539$. Inserting these quantities into the formula for $s_{\hat{Y}(1)}$, we obtain

$$s_{\hat{Y}(1)} = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} \tag{17.53}$$

$$= \sqrt{2.539 \left[1 + \frac{1}{27} + \frac{(0.5 - 0.960)^2}{11.798}\right]} \tag{17.54}$$

$$= \sqrt{2.539 \left[1 + 0.037 + \frac{0.212}{11.798}\right]} \tag{17.55}$$

$$= 1.637. \tag{17.56}$$

For a 95% prediction interval and $\alpha = 0.05$, the interval has the form

$$(\hat{Y} - c_{0.05,n-2}s_{\hat{Y}(1)}, \hat{Y} + c_{0.05,n-2}s_{\hat{Y}(1)}) \tag{17.57}$$

From Table T with we have $c_{0.05,25} = 2.060$. Inserting $c_{0.05,25} = 2.060$, $\hat{Y} = 3.265$, and $s_{\hat{Y}(1)} = 1.637$ in this formula, we obtain

$$(3.265 - 2.060(1.637), 3.265 + 2.060(1.637)) \tag{17.58}$$

or

$$(-0.107, 6.637). \tag{17.59}$$

Note this interval is much wider than the interval for the theoretical mean $\mu = \alpha + \beta X_i$, which was $(2.495, 4.035)$. This is because you are trying to enclose a single future observation, a random variable Y_i, rather than a theoretical mean.

17.4 R^2 values

R^2 values are a measure of how well a statistical model explains the data. Recall that the following relationship holds among the sum of squares in linear regression:

$$SS_{regression} + SS_{error} = SS_{total}. \tag{17.60}$$

We can think of the different sum of squares as partitioning the variability in the data into different sources. $SS_{regression}$ represents variability explained by

the regression line, SS_{error} represents variability of the observations around the regression line, while SS_{total} is the total amount of variability in the data. The R^2 value for a linear regression model is the proportion of total variability explained by the model, or

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{SS_{regression}}{SS_{regression} + SS_{error}}. \tag{17.61}$$

It is clear from this formula that R^2 must range between 0 and 1 ($0 \leq R^2 \leq 1$). For the Example 1 data, we have

$$R^2 = 82.528/146.014 = 0.565. \tag{17.62}$$

Thus, 56.5% of the variation is explained by the regression model for these data. Small R^2 values indicate there is substantial variability in the data not explained by the model, while large ones indicate the model explains most of the variation.

More generally, we can define an R^2 value for both ANOVA and regression models as

$$R^2 = \frac{SS_{model}}{SS_{total}} = \frac{SS_{model}}{SS_{model} + SS_{error}}. \tag{17.63}$$

For example, we have $SS_{model} = SS_{among}$ for one-way ANOVA while $SS_{error} = SS_{within}$. The R^2 value here is the proportion of the variation explained by the one-way ANOVA model, in particular the variation among the group means. The SAS output for `proc glm` provides an R^2 for ANOVA models of this form.

17.5 Linear regression for Example 1 - SAS demo

The linear regression analysis can be conducted using `proc glm` and a program similar in structure to ANOVA ones (see SAS program and output below). We first input the observations using a `data` step, applying transformations if necessary. The dependent variable Y is defined as the SAS variable `y`, while the independent variable X is defined as `x`. **It is important to realize that the actual names of these variables are not important - it is their position in `proc gplot` and `proc glm` that determines which one is the dependent variable, and which is the independent one.**

The dependent variable always goes first. Note also the additional observation at end of the data set, for which `x` = 0.5 but `y` is a missing value. The purpose of this observation is to make `proc glm` calculate a confidence interval for the mean, as well as a prediction interval, at that particular value of `x`.

The data are then plotted along with the fitted line plus confidence and prediction intervals. This accomplished using the following `proc gplot` code (SAS Institute Inc. 2014a). The three `y*x` statements in the `plot` command plot the same data in three different ways, which are then combined into one graph using the `overlay` option. The first plot, using the `symbol1` command, draws the data points. The second plot, using the `symbol2` command, draws a regression line through the points and also plots 95% confidence intervals for the mean of Y_i at X_i, or $\mu = \alpha + \beta X_i$, across the range of X_i values. The third plot, using the `symbol3` command, plots 95% prediction intervals for a single future observation, again across the range of X_i values.

The regression analysis is conducted using `proc glm` as shown below (SAS Institute Inc. 2014b). There is no `class` statement because the independent variable `X` is a continuous variable and does not fall into discrete groups like ANOVA. Note the similarity of the `model` statement to the linear regression model. The option `clparm` is used to generate 95% confidence intervals for α and β, while `clm` generates a 95% confidence interval for the mean of Y_i at each value of X_i. If we want prediction intervals it is necessary to run `proc glm` a second time using the `cli` option in the `model` statement (see below). This is necessary because `proc glm` cannot generate both types of intervals at the same time.

The data points, regression line, and confidence or prediction intervals are shown in Fig. 17.3. The prediction intervals are much wider than the confidence intervals, because the prediction intervals are for single future Y_i while the confidence intervals enclose a mean. Note that both types of interval increase in width as you move away from the center of the X values. This follows from the fact that the standard errors involved in these calculations are a function of $(X_i - \bar{X})^2$, which increases as X_i moves away from $\bar{X}$.

Examining the output for `proc glm`, first note that the slope β is labeled as `x` while the intercept α is `Intercept`. We see that attack density `y` increases with beetle numbers `x`, because $\hat{\beta} = 2.645$ and is positive. The effect of beetle numbers on attack density was highly significant ($F_{1,25} = 32.5, P < 0.0001$). There are several F tests to chose from in the output, but all give the same result for simple linear regression. Alternately, we could report the t test for

β ($t_{25} = 5.70, P < 0.0001$), which also tests $H_0 : \beta = 0$. We see that $R^2 = 0.565$, indicating that 56.5% of the variation is explained by the regression model.

The `proc glm` output also provides 95% confidence intervals for α and β. A 95% confidence interval for the mean of Y_i at $X_i = 0.5$ is also given, and labeled as `95% Confidence Limits for Mean Predicted Value`. The second set of output for `proc glm` contains a 95% prediction interval for a single future Y_i at $X_i = 0.5$, labeled as `95% Confidence Limits for Individual Predicted Value`.

Note that the estimated intercept is some distance from zero ($\hat{\alpha} = 1.942$), and in fact the t test of $H_0 : \alpha = 0$ reported by SAS is highly significant ($t_{25} = 3.59, P = 0.0014$). This cannot really be true because the addition of zero beetles should give you an attack density of zero. A more accurate (and possibly non-linear) model would require that the intercept be zero.

This is a potential pitfall when using linear regression. Many biological phenomenon are approximately linear over some range of the data but the approximation breaks down for more extreme values. A linear regression does not take this possibility into account and so cannot provide a general explanation of some phenomena.

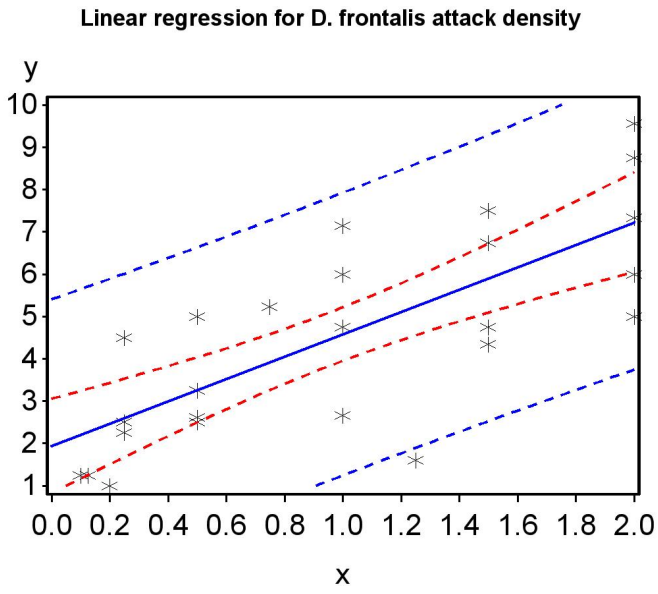

Figure 17.3: Linear regression model fitted to the Example 1 data, where Y is attack density and X is beetles added to the cages. Also shown are 95% confidence intervals for the mean, and prediction intervals for a single future observation.

SAS Program

```
* SPBattack.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'Linear regression for D. frontalis attack density';
data frontalis;
    input attacks beetles;
    * Apply transformations here;
    y = attacks;
    x = beetles;
    datalines;
1.25 0.100
2.66 1.000
7.33 2.000
1.60 1.250
2.62 0.500

etc.

5.00 0.500
.    0.500
;
run;
* Print data set;
proc print data=frontalis;
run;
* Plot data and regression line;
proc gplot data=frontalis;
    plot y*x y*x y*x / overlay vaxis=axis1 haxis=axis1;
    symbol1 i=none v=star c=black height=2 width=3;
    symbol2 i=rlclm v=none c=red height=2 width=3;
    symbol3 i=rlcli v=none c=blue height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Regression analysis with confidence intervals;
proc glm data=frontalis;
    model y = x / clparm clm;
    output out=resids p=pred r=resid;
run;
* Regression analysis with prediction intervals;
proc glm data=frontalis;
    model y = x / clparm cli;
run;
goptions reset=all;
```

```
title "Diagnostic plots to check ANOVA assumptions";
* Plot residuals vs. predicted values;
proc gplot data=resids;
    plot resid*pred=1 / vaxis=axis1 haxis=axis1;
    symbol1 v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Normal quantile plot of residuals;
proc univariate noprint data=resids;
    qqplot resid / normal waxis=3 height=4;
run;
quit;
```

SAS Output

Linear regression for D. frontalis attack density 1
08:45 Sunday, November 14, 2010

Obs	attacks	beetles	y	x
1	1.25	0.100	1.25	0.100
2	2.66	1.000	2.66	1.000
3	7.33	2.000	7.33	2.000
4	1.60	1.250	1.60	1.250
5	2.62	0.500	2.62	0.500

etc.

27	5.00	0.500	5.00	0.500
28	.	0.500	.	0.500

Linear regression for D. frontalis attack density 2
08:45 Sunday, November 14, 2010

The GLM Procedure

Number of Observations Read	28
Number of Observations Used	27

Linear regression for D. frontalis attack density 3
08:45 Sunday, November 14, 2010

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	82.5283492	82.5283492	32.50	<.0001
Error	25	63.4855174	2.5394207		
Corrected Total	26	146.0138667			

R-Square	Coeff Var	Root MSE	y Mean
0.565209	35.56163	1.593556	4.481111

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	82.52834922	82.52834922	32.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	82.52834922	82.52834922	32.50	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.941567811	0.54083158	3.59	0.0014
x	2.644847410	0.46394486	5.70	<.0001

Parameter	95% Confidence Limits	
Intercept	0.827704323	3.055431300
x	1.689335080	3.600359740

The GLM Procedure

Observation	Observed	Predicted	Residual
1	1.25000000	2.20605255	-0.95605255
2	2.66000000	4.58641522	-1.92641522
3	7.33000000	7.23126263	0.09873737
4	1.60000000	5.24762707	-3.64762707
5	2.62000000	3.26399152	-0.64399152

etc.

27	5.00000000	3.26399152	1.73600848
28 *	.	3.26399152	.

Observation	95% Confidence Limits for Mean Predicted Value	
1	1.16947580	3.24262930
2	3.95365127	5.21917917
3	6.05393677	8.40858849
4	4.55796883	5.93728532
5	2.49438766	4.03359537

etc.

27	2.49438766	4.03359537
28 *	2.49438766	4.03359537

* Observation was not used in this analysis

Sum of Residuals	-0.00000000
Sum of Squared Residuals	63.48551745
Sum of Squared Residuals - Error SS	0.00000000
PRESS Statistic	73.72506348
First Order Autocorrelation	0.45535896
Durbin-Watson D	1.02741345

etc.

Linear regression for D. frontalis attack density 8
08:45 Sunday, November 14, 2010

The GLM Procedure

Observation	Observed	Predicted	Residual
1	1.25000000	2.20605255	-0.95605255
2	2.66000000	4.58641522	-1.92641522
3	7.33000000	7.23126263	0.09873737
4	1.60000000	5.24762707	-3.64762707
5	2.62000000	3.26399152	-0.64399152

etc.

```
27          5.00000000        3.26399152        1.73600848
28 *         .                3.26399152         .

                          95% Confidence Limits for
  Observation             Individual Predicted Value

            1             -1.23574200       5.64784710
            2              1.24398368       7.92884676
            3              3.74449413      10.71803113
            4              1.89395940       8.60129475
            5             -0.10702442       6.63500745

etc.

           27             -0.10702442       6.63500745
           28 *           -0.10702442       6.63500745


* Observation was not used in this analysis


                Sum of Residuals                        -0.00000000
                Sum of Squared Residuals                63.48551745
                Sum of Squared Residuals - Error SS      0.00000000
                PRESS Statistic                         73.72506348
                First Order Autocorrelation              0.45535896
                Durbin-Watson D                          1.02741345
```

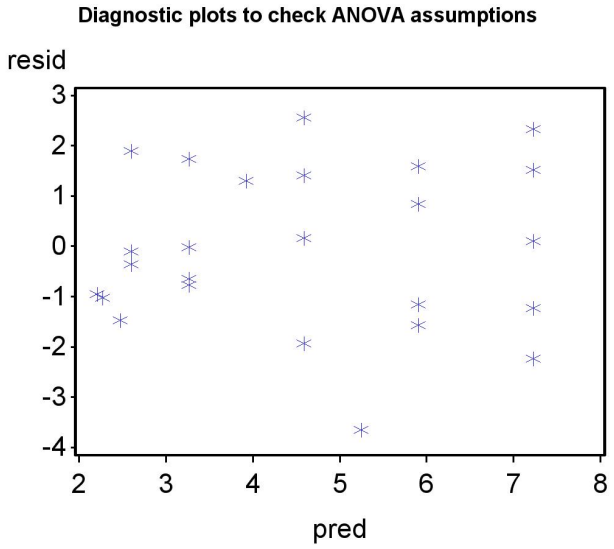

Figure 17.4: Residual vs. predicted plot for the Example 1 analysis.

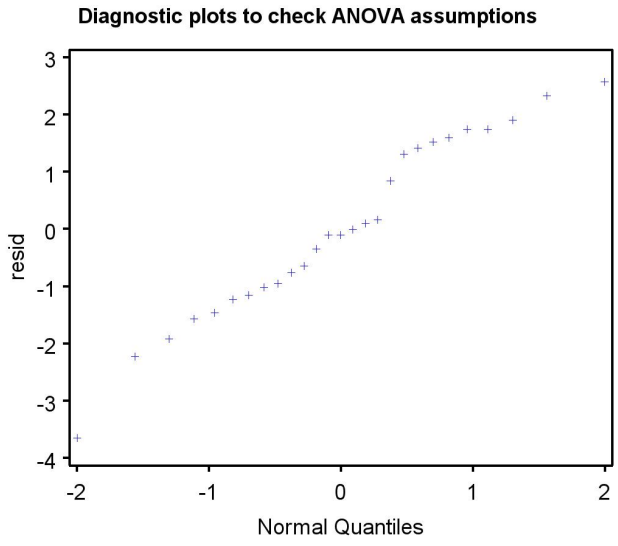

Figure 17.5: Normal quantile plot for the Example 1 analysis.

17.6 Assumptions and transformations

Linear regression makes the same assumptions as ANOVA, including homogeneity of variances and normality, and the same types of plots can be used to assess them. If the homogeneity of variances assumption is satisfied, the points in a residual vs. predicted plot should be equally scattered across the range of predicted values. Outliers can also be identified using this plot. The normality assumption can be evaluated using a normal quantile plot of the residuals, with a straight diagonal line indicating this assumption is satisfied.

Examining the residuals from the Example 1 analysis, we see no obvious pattern in the residual vs. predicted plot, suggesting the homogeneity of variances assumption is satisfied (Fig. 17.4). No outliers were present. The normal quantile plot suggests the normality assumption is satisfied (Fig. 17.5).

Linear regression makes another key assumption, namely that the relationship between Y and X is linear. This assumption can be checked by examining a plot of Y vs. X as well as the residual vs. predicted plot (see examples below). What can be done if the relationship seems nonlinear? We can sometimes fix this problem by applying a transformation to Y, X, or both Y and X, so that linear regression can be applied to the transformed data. **This use of transformations greatly extends the utility of linear regression.** Some commonly used transformations are $\log Y$ vs. X, $\log Y$ vs. $\log X$, Y vs. $\log X$, and $1/Y$ vs. X. A transformation that linearizes the data sometimes corrects for problems with the homogeneity of variances and normality assumptions.

A transformation may be selected based on prior information about the data and system. For example, a conservation biologist may be interested in the relationship between island area A and the number of species S on the island, and previous studies suggest the relationship between $\log_{10} S$ and $\log_{10} A$ will be linear (MacArthur & Wilson 1967). Another approach is to try a number of transformations and chose the one that makes the data most linear. We will illustrate each approach with an example below.

In cases where no transformation can linearize the data, another possibility would be **nonlinear regression** (Juliano 1993). This type of analysis requires that the user specify a model $Y = f(X, \theta_1, \theta_2, \ldots) + \epsilon$ for the data, where f is a function with parameters $\theta_1, \theta_2, \ldots$ to be estimated. SAS implements this type of nonlinear regression in `proc nlin`, while `proc nlmixed` allows

for nonlinear functions as well as random effects and nonnormal distributions.

17.6.1 Species-area data - SAS demo

For many organisms there is a relationship between a defined area of habitat, such as an island, and the number of species found there. If S is the number of species, and A the area of habitat, then the model $S = cA^z$ seems to describe many data sets (MacArthur & Wilson 1967). Taking the $\log_{10}$ of both sides of this equation, we obtain

$$\log_{10} S = \log_{10} c + z \log_{10} A. \tag{17.64}$$

This form of the model is linear and suggests linear regression could be used to analyze species-area data. The SAS program listed below shows how these transformations can be applied to the bird fauna on archipelagos and islands of varying areas. The data are the number of species vs. island area (square miles) for 23 islands. The data were simulated to resemble Fig. 9 in MacArthur & Wilson (1967). An extra observation is included with a missing value for the number of species, but an island area of 5000 square miles, to make `proc glm` calculate a confidence interval for the mean of this island.

We first conduct the analysis without any transformation and examine the `gplot` graph of Y vs. X, where Y is the number of species and X is island area (Fig. 17.6). Note the nonlinear nature of the relationship between the number of species and island area. This pattern is also reflected in the residual vs. predicted plot (Fig. 17.7), which appears to be hump-shaped. Both plots suggest that a transformation is required for these data in order to linearize the relationship between the two variables.

The picture improves after a $\log_{10}$ transformation is applied to both species and area. We see that the graph of the transformed variables is linear (Fig. 17.8) and residual vs. predicted plot is featureless (Fig. 17.9). The normal quantile plot is also well-behaved (Fig. 17.10). Now that the various assumptions are satisfied we can interpret the rest of the SAS output (see below). We see that the number of species increases with island area ($\hat{\beta} = 0.241$) and the effect is highly significant ($F_{1,21} = 148.16, P < 0.0001$). In terms of the original model, where $S = cA^z$, we see that $\hat{\beta} = 0.241$ is also an estimate of z. The R^2 value is 0.876, indicating that 87.6% of the variation is explained by the regression model. Confidence intervals are also provided for the intercept and slope.

The `proc glm` output also generates a predicted value $\hat{Y}_i = 1.800$ at $X_i = 3.699$ ($\log_{10} 5000 = 3.699$). We need to convert this to the original scale measurement using antilogs. We have $\hat{S}_i = 10^{\hat{Y}_i} = 10^{1.800} = 63.10$ species. So, we predict there would be 63 species on an island of 5000 square miles. The confidence interval for the mean is $(1.746, 1.855)$, which we can similarly convert to $(10^{1.745}, 10^{1.855})$ or $(55.72, 71.61)$.

SAS Program

```
* SAprob2.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'Linear regression for species-area data';
data sa;
    input species area;
    * Apply transformations here;
    y = log10(species);
    x = log10(area);
    datalines;
 15      28
104  113480
165  380358
116   33252
 35    1010
 33     305
 78   37620
 93    4762
 50     213
 76    2976
 18      23
 28     186
 20     423
121  108512
 53     364
 22     269
102   11163
 28     487
158  445409
 19      70
111   38309
152  100873
 55    1354
  .    5000
;
run;
* Print data set;
proc print data=sa;
run;
* Plot data and regression line;
proc gplot data=sa;
    plot y*x=1 y*x=2 y*x=3 / overlay vaxis=axis1 haxis=axis1;
    symbol1 i=none v=star c=black height=2 width=3;
```

```
    symbol2 i=rlclm v=none c=red height=2 width=3;
    symbol3 i=rlcli v=none c=blue height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Regression analysis with confidence intervals;
proc glm data=sa;
    model y = x / clparm clm;
    output out=resids p=pred r=resid;
run;
* Regression analysis with prediction intervals;
proc glm data=sa;
    model y = x / clparm cli;
run;
goptions reset=all;
title "Diagnostic plots to check ANOVA assumptions";
* Plot residuals vs. predicted values;
proc gplot data=resids;
    plot resid*pred=1 / vaxis=axis1 haxis=axis1;
    symbol1 v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Normal quantile plot of residuals;
proc univariate noprint data=resids;
    qqplot resid / normal waxis=3 height=4;
run;
quit;
```

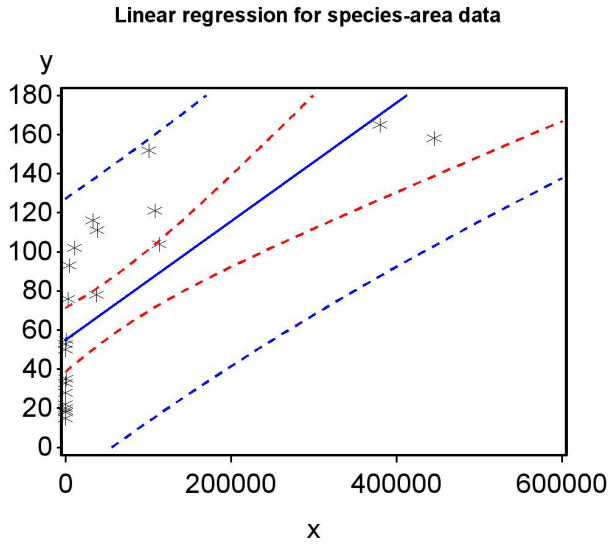

Figure 17.6: Linear regression model fitted to the species-area data, where Y is the number of species and X is island area.

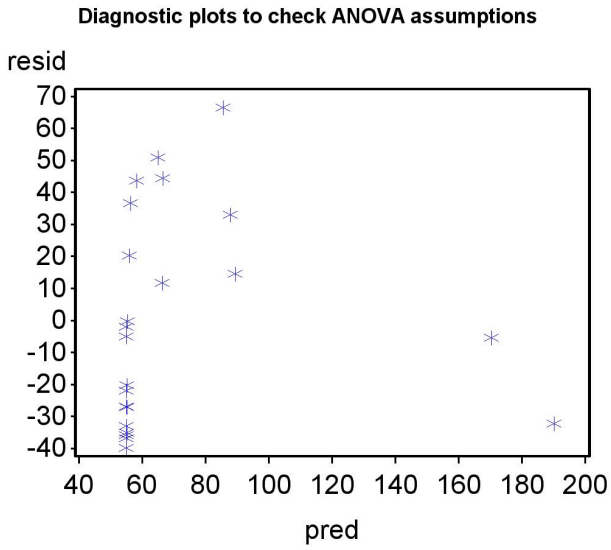

Figure 17.7: Residual vs. predicted plot for the species-area data.

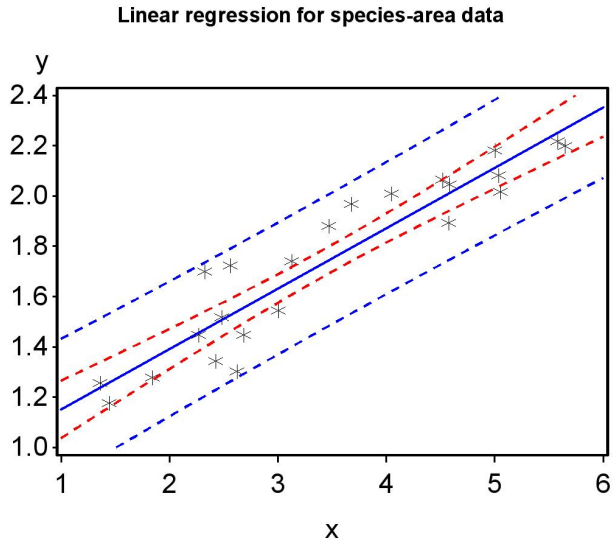

Figure 17.8: Linear regression model fitted to the species-area data, where Y is log-transformed species and X is log-transformed area.

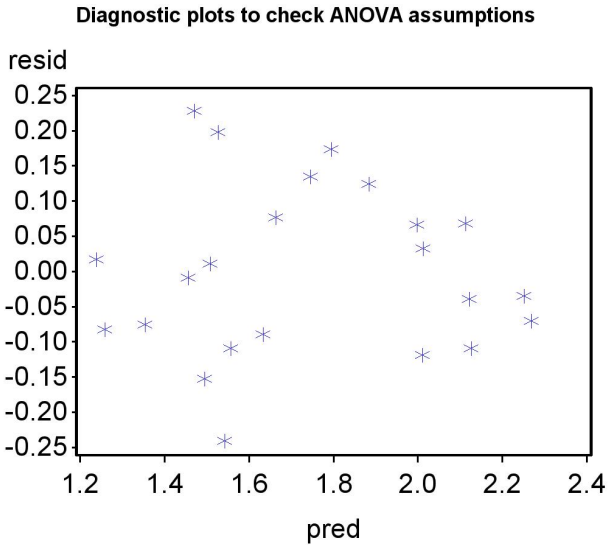

Figure 17.9: Residual vs. predicted plot for the log-transformed species-area data.

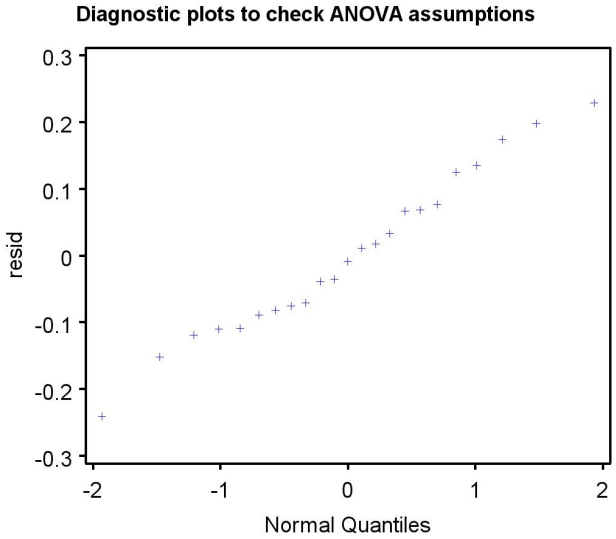

Figure 17.10: Normal quantile plot for the log-transformed species-area data.

SAS Output

Linear regression for species-area data 1
11:32 Tuesday, November 16, 2010

Obs	species	area	y	x
1	15	28	1.17609	1.44716
2	104	113480	2.01703	5.05492
3	165	380358	2.21748	5.58019
4	116	33252	2.06446	4.52182
5	35	1010	1.54407	3.00432

etc.

23	55	1354	1.74036	3.13162
24	.	5000	.	3.69897

Linear regression for species-area data 2
11:32 Tuesday, November 16, 2010

The GLM Procedure

Number of Observations Read	24
Number of Observations Used	23

Linear regression for species-area data 3
11:32 Tuesday, November 16, 2010

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.25182542	2.25182542	148.16	<.0001
Error	21	0.31916133	0.01519816		
Corrected Total	22	2.57098675			

R-Square	Coeff Var	Root MSE	y Mean
0.875860	7.083042	0.123281	1.740507

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	2.25182542	2.25182542	148.16	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	2.25182542	2.25182542	148.16	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.9102215097	0.07289411	12.49	<.0001
x	0.2405722961	0.01976395	12.17	<.0001

Parameter	95% Confidence Limits	
Intercept	0.7586299190	1.0618131004
x	0.1994709127	0.2816736795

Linear regression for species-area data 4
11:32 Tuesday, November 16, 2010

The GLM Procedure

Observation		Observed	Predicted	Residual
1		1.17609126	1.25836764	-0.08227638
2		2.01703334	2.12629506	-0.10926172
3		2.21748394	2.25266125	-0.03517730
4		2.06445799	1.99804559	0.06641240
5		1.54406804	1.63297800	-0.08890996

etc.

23		1.74036269	1.66360220	0.07676049
24	*	.	1.80009122	.

Observation	95% Confidence Limits for Mean Predicted Value	
1	1.16016869	1.35656659
2	2.04142998	2.21116013
3	2.15012264	2.35519985
4	1.92880841	2.06728278
5	1.57645124	1.68950477

etc.

23	1.60855303	1.71865138
24 *	1.74567238	1.85451005

* Observation was not used in this analysis

Sum of Residuals	-0.00000000
Sum of Squared Residuals	0.31916133
Sum of Squared Residuals - Error SS	0.00000000
PRESS Statistic	0.36922092
First Order Autocorrelation	0.04242134
Durbin-Watson D	1.87548592

etc.

17.6.2 Population growth rates - SAS demo

As another example of transformations, consider a study of the population growth of phytophagous mites on leaf sections. An experiment is conducted in which leaf sections are inoculated with a range of mite densities and the number of offspring recorded one generation later. The number of offspring per initial mite is the finite growth of the population, usually symbolized as λ. The SAS program listed below gives the mite densities and the λ values for this experiment.

We first conduct the analysis without any transformation. Looking at the plot of Y (λ) vs. X (density), we see a curvilinear relationship (Fig. 17.11) that also appears in the residual vs. predicted plot (Fig. 17.12). A transformation is clearly needed, but which one? A natural log transformation usually a good starting point for population data, both for growth rates and numbers. We begin by log-tranforming the dependent variable λ and examining the plots (see program below). The graph after transformation is linear (Fig. 17.13) and the residual vs. predicted plot shows no pattern (Fig. 17.14). The normal quantile plot is also adequate (Fig. 17.15).

Interpreting the SAS output (see below), we see that λ decreases with mite density ($\hat{\beta} = -0.020$) and the effect is highly significant ($F_{1,15} = 1695.22, P < 0.0001$). The R^2 value is 0.991, indicating that almost all the variation in the data is explained by the regression line. It appears that the growth rate of the mites is adversely affected by their density, probably through competition for resources or other intraspecific interactions.

SAS Program

```
* logistic.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'Linear regression for growth rate-density data';
data grd;
    input lambda density;
    * Apply transformations here;
    y = log(lambda);
    x = density;
    datalines;
  7.32   5
  4.82  15
  4.69  25
  3.90  35
  2.65  45
  2.52  55
  1.70  65
  1.68  75
  1.43  85
  1.07  95
  0.74 105
  0.72 115
  0.64 125
  0.47 135
  0.40 145
  0.38 155
  0.25 165
;
run;
* Print data set;
proc print data=grd;
run;
* Plot data and regression line;
proc gplot data=grd;
    plot y*x=1 y*x=2 y*x=3 / overlay vaxis=axis1 haxis=axis1;
    symbol1 i=none v=star c=black height=2 width=3;
    symbol2 i=rlclm v=none c=red height=2 width=3;
    symbol3 i=rlcli v=none c=blue height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Regression analysis with confidence intervals;
proc glm data=grd;
    model y = x / clparm clm;
```

```
    output out=resids p=pred r=resid;
run;
* Regression analysis with prediction intervals;
proc glm data=grd;
    model y = x / clparm cli;
run;
goptions reset=all;
title "Diagnostic plots to check ANOVA assumptions";
* Plot residuals vs. predicted values;
proc gplot data=resids;
    plot resid*pred=1 / vaxis=axis1 haxis=axis1;
    symbol1 v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Normal quantile plot of residuals;
proc univariate noprint data=resids;
    qqplot resid / normal waxis=3 height=4;
run;
quit;
```

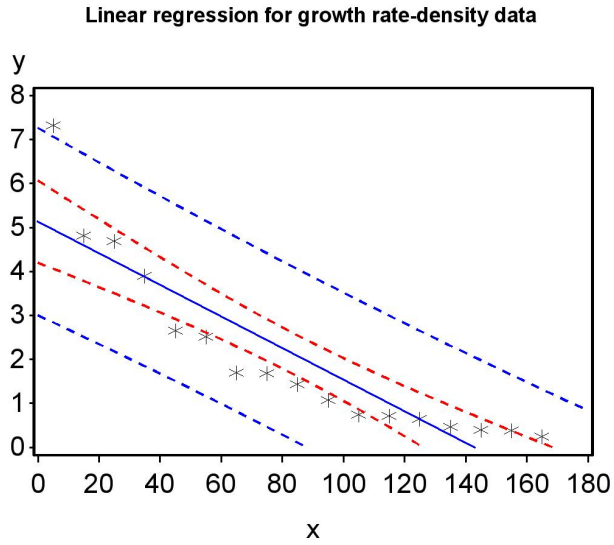

Figure 17.11: Linear regression model fitted to the λ-density data, where Y is λ and X is initial mite density.

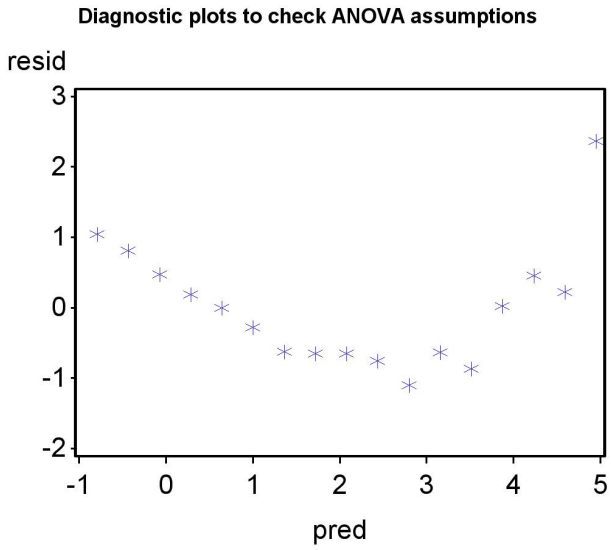

Figure 17.12: Residual vs. predicted plot for the λ-density data.

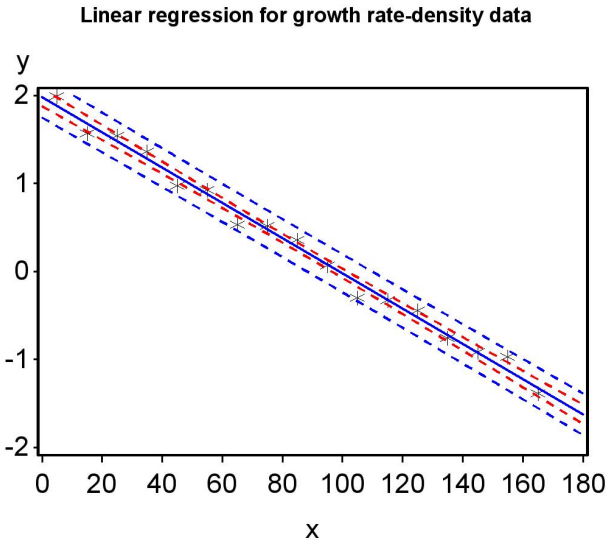

Figure 17.13: Linear regression model fitted to the λ-density data, where Y is $\log \lambda$ and X is initial mite density.

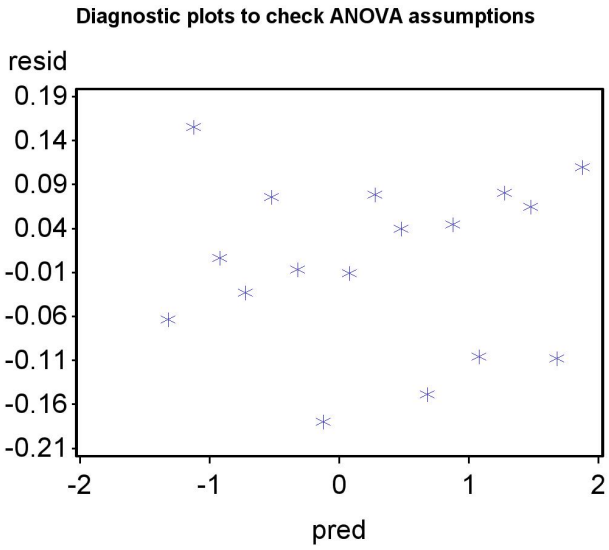

Figure 17.14: Residual vs. predicted plot for the transformed λ-density data.

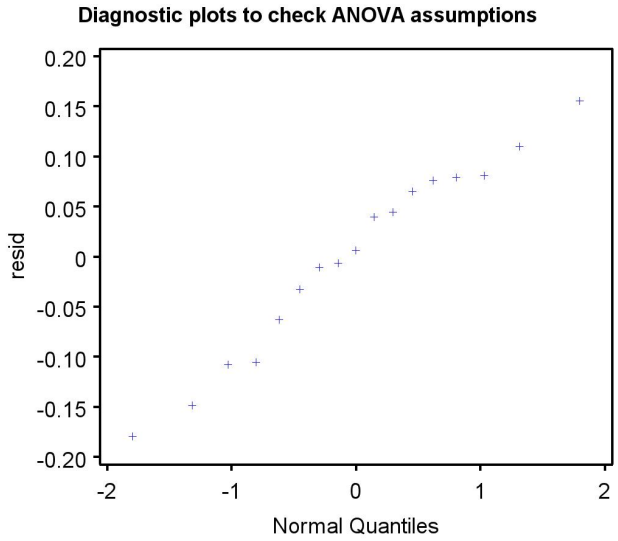

Figure 17.15: Normal quantile plot for the transformed λ-density data.

SAS Output

Linear regression for growth rate-density data 1
18:39 Tuesday, November 16, 2010

Obs	lambda	density	y	x
1	7.32	5	1.99061	5
2	4.82	15	1.57277	15
3	4.69	25	1.54543	25
4	3.90	35	1.36098	35
5	2.65	45	0.97456	45

etc.

Linear regression for growth rate-density data 2
18:39 Tuesday, November 16, 2010

The GLM Procedure

Number of Observations Read	17
Number of Observations Used	17

Linear regression for growth rate-density data 3
18:39 Tuesday, November 16, 2010

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	16.36176928	16.36176928	1695.22	<.0001
Error	15	0.14477544	0.00965170		
Corrected Total	16	16.50654472			

R-Square	Coeff Var	Root MSE	y Mean
0.991229	35.21791	0.098243	0.278958

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	16.36176928	16.36176928	1695.22	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	16.36176928	16.36176928	1695.22	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.981131688	0.04771689	41.52	<.0001
x	-0.020025578	0.00048638	-41.17	<.0001

Parameter	95% Confidence Limits	
Intercept	1.879425551	2.082837825
x	-0.021062263	-0.018988893

Linear regression for growth rate-density data 4
18:39 Tuesday, November 16, 2010

The GLM Procedure

Observation	Observed	Predicted	Residual
1	1.99061033	1.88100380	0.10960653
2	1.57277393	1.68074802	-0.10797410
3	1.54543258	1.48049225	0.06494033
4	1.36097655	1.28023647	0.08074008
5	0.97455964	1.07998070	-0.10542106

etc.

Observation	95% Confidence Limits for Mean Predicted Value	
1	1.78375413	1.97825347
2	1.59217362	1.76932242
3	1.40019098	1.56079352

```
                    4          1.20766853      1.35280442
                    5          1.01441498      1.14554641

etc.


          Linear regression for growth rate-density data                5
                                        18:39 Tuesday, November 16, 2010

                              The GLM Procedure

          Sum of Residuals                          -0.00000000
          Sum of Squared Residuals                   0.14477544
          Sum of Squared Residuals - Error SS       -0.00000000
          PRESS Statistic                            0.18945485
          First Order Autocorrelation               -0.31722141
          Durbin-Watson D                            2.52386773

etc.
```

17.7 Problems

1. An experiment was conducted to measure the effect of density on the rate of egg laying in cowpea weevils. Ten different densities were used in the experiment, and the rate of egg laying determined for each density. The following data were obtained:

Density	Eggs per day
100	7.629
200	4.530
500	3.820
700	2.718
1200	2.403
1500	1.756
1700	1.772
2000	1.508
2200	1.518
2500	1.359

   (a) Plot the rate of egg laying (Y) vs. density (X), and observe the nonlinear relationship between Y and X. Find a transformation of Y and/or X that linearizes this relationship using SAS.

   (b) For the transformed data, use SAS to plot a 95% confidence interval for the mean of Y_i and a 95% prediction interval for a single value of Y_i. Label the intervals (confidence or prediction) on the `gplot` graph.

   (c) Analyze the transformed data set using linear regression and SAS. In your SAS output, label the 95% confidence intervals for the intercept (α) and slope (β) in your SAS printout.

   (d) Interpret the results of the regression analysis. Is there a significant effect of density on the rate of egg production? What direction is the effect?

2. A zoologist wants to establish the relationship between the length of an animal and its weight. He wants to use length to predict weight in future studies, because length is easier to measure. The lengths and weights of a random sample of 20 animals were determined, yielding the following data:

Length (mm)	Weight (g)
14.7	1.65
19.9	4.86
15.8	2.04
19.0	3.53
8.4	0.32
10.2	0.46
13.5	1.68
22.1	6.24
16.2	1.85
8.2	0.28
10.1	0.48
19.8	4.18
20.6	4.77
22.0	6.10
18.1	2.78
22.4	5.26
10.5	0.55
14.5	1.56
11.9	1.07
14.7	1.74

(a) Plot the weight (Y) vs. length (X) using SAS, and observe the nonlinear rela-tionship between Y and X. Attach your graph of this relationship. Then, find a transformation of Y and/or X that linearizes this relationship using SAS or R. What transformation did you use? Attach your graph showing the transformed relationship.

(b) Analyze the transformed data using linear regression and SAS. Briefly interpret your results using P values. Is there a significant effect of length on weight? What direction is the effect? Attach your program and output.

(c) For animals that are 21 mm long, find a 95% confidence interval for the mean weight.

17.8 References

MacArthur, R. H. & Wilson, E. O. (1967) *The Theory of Island Biogeography.* Princeton University Press, Princeton, NJ.

McCulloch, C. E. & Searle, S. R. (2001) *Generalized, Linear, and Mixed Models.* John Wiley & Sons, Inc., New York, NY.

Reeve, J. D., Rhodes, D. J. & Turchin, P. (1998) Scramble competition in southern pine beetle (Coleoptera: Scolytidae). *Ecological Entomology* 23: 433-443.

SAS Institute Inc. (2014a) *SAS/GRAPH 9.4: Reference, Third Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2014b) *SAS/STAT 13.2 Users Guide.* SAS Institute Inc., Cary, NC.

Searle, S. R. (1971) *Linear Models.* John Wiley & Sons, Inc., New York, NY.