Lecture 7 – Linear relationships

Introduction

Often in biology we are faced with data that are linearly related to one another, where the observed value is dependent on another. Therefore if we know the relationship between the values a can predict one value if we know the other.

For example, you go fishing, catch a fish and then measure it. When you get home, you are interested in its weight to brag to your mates. It is known that fish growth in weight as a function of its length, so given the length and some length-weight relationship you can estimate its weight.

The simplest way to investigate linearly correlated values would be with the equation $y = \beta x$, where y is simply linearly related to an x by some sort of parameter β - a slope parameter. A slightly more complicated model is $y = \beta_0 + \beta_1 x$ where β_0 and β_1 is the intercept and slope, respectively. In reality, the first model simply a subset of the second model as it merely has an intercept of zero.

In the real world, these slope and intercept parameters are not known and to further complicate issues there tends to be some form of sampling error in collecting the data. We therefore require a statistical method to estimate these parameters such that we may make additional inferences about our confidence in their values or draw inference about any prediction we wish to make.

This method is known as regression, and in our specific case linear regression as there are non-linear[1] forms around. The term "regression" was coined when the heights of children were plotted against the difference in the heights of their parents where it was shown that children height "regressed" to the average of their parents heights.

Regression is therefore a method to find the average value of a *y*-value given some known *x*-value.

Method of least squares

The simplest method is to find the best parameter estimates that minimise the variance, or simply the sum-of-squared differences as in ANOVA, between the observed data and the fitted line.

---

[1] A linear model can be defined as having all exponents for any dependent value equal to one. For example, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is linear as is $y = \beta_0 x_1^{\beta_1}$ (because $\ln y = \ln \beta_0 + \beta_1 \ln x_1$ through a logarithmic transformation). Unfortunately, $y = \beta_0 \left(1 - e^{-\beta_1 (x - \beta_2)}\right)$ is non-linear. This model is the Von Bertalanffy growth model used commonly to model fish length *y* as a function of age *x*.

Under the assumption that there is no error in the independent data (the *x*'s) then the sum-of-squared error is calculated from the error-prone dependent data (*y*'s) as

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Note that the "hat" symbol means "estimate".

To solve for the optimal $\hat{\beta}_0$ and $\hat{\beta}_1$ that provides the least squared error we use differential calculus. The solution is found by taking the first derivative of *SSE* with respect to each parameter of interest, that is $\hat{\beta}_0$ and $\hat{\beta}_1$, setting each resultant equation to zero and then solving the equation simultaneously.

To spare you the mathematical schlep, the least squares estimates of the two parameters are

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \text{ and } \beta_0 = \bar{y} - \beta_1 \bar{x}$$

The estimate of $\hat{\beta}_1$ is a touch unwieldly, so a little mathematics provides us with the computationally simpler and more intuitive solution of

$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}}$ which the ratio of the deviations between the *x*'s and *y*'s and the *x*'s (a statistical "run over rise" that you learnt in high school for the slope)

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

and $SS_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$. Note that SS_{YY} is found in a similar fashion such that $SS_{YY} = \sum y_i^2 - n\bar{y}^2$.

Therefore, $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$.

The intercept is calculated from the slope and the average *x* and *y* values.

We know that the regression line will pass through the average of both the dependent (y) and independent data because the best fit is the average *y* at the average *x*.

Therefore If we know the slope then $\hat{\beta}_1 = \frac{\bar{y} - \hat{\beta}_0}{\bar{x} - 0}$ such that $\hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_0$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Model variance

The deviations (or technically residuals) between observed and predicted *y*'s are centred around zero with a regression model variance of s_ε^2 calculated as

$$s_\varepsilon^2 = \frac{SSE}{n-2}$$

Partitioning the different sources of error

An with ANOVA, which is in fact a form of linear regression as are other multivariate statistical methods such as Principle Component Analysis and Discriminant Function Analysis, we need to partition our sums-of-squares such that we can make inferential statements about the linear model.

The total error in the model can be divided into two components; error associated with the regression, and error associated with the sampling error.

Therefore as

$$SST = SSE + SSREG$$

then we can calculate that

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}_i)]^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y}_i)^2$$

Determining the model's goodness-of-fit

Questions arise as to how much variance does our linear model describe. If our model described absolutely nothing because the data looked like the scatter from a shotgun

blast then our goodness-of-fit criterion, also known as the Coefficient of Determination, would be 0%. If on the other hand, all our data fitted the linear model exactly then the model would have described 100% or 1.

The ratio $\frac{SSE}{SST}$ holds the key. If SSE is the same as the SST then the data are random and the ratio would be 1. Alternatively, if the data fitted the model exactly then the ratio would be 0 (because SSE = 0).

Therefore, $R^2 = 1 - \frac{SSE}{SST} = \frac{SSREG}{SST}, \quad 0 \le R^2 \le 1$

Parameter correlation

The Correlation Coefficient, r, is used to determine the degree of correlation between parameters (not variables x and y). For an increase of one unit of β_0 we can expect to get an r increase in parameter β_1 and *visa versa*.

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX} \cdot SS_{YY}}}, \quad -1 \le r \le 1$$

Inference about β_0 and β_1

The parameters β_0 and β_1 are unknown and have been replaced by their least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

We can therefore construct $(1-\alpha)$% confidence intervals around the parameters using the same method used in constructing confidence intervals of μ using the *t*-distribution.

For the slope parameter as

$$\beta_1 = \hat{\beta}_1 \pm t_{\alpha/2,n-2} s_\varepsilon / \sqrt{SS_{XX}}$$

and for the intercept parameter as

$$\beta_0 = \hat{\beta}_0 \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{\frac{\sum x_i^2}{n \cdot SS_{XX}}}$$

Note that the standard error of the parameter estimates have n - 2 degrees of freedom. This is because we have estimated two parameters - $\hat{\beta}_0$ and $\hat{\beta}_1$. For example, in the case of a more complicated model with 5 parameters then there would be $n - 5$ degrees of freedom.

Hypotheses about β_0 and β_1

Hypotheses can be made about parameters. Examples include: "Is the slope different from zero?" or "Is there a one-to-one relationship between the observed and predicted data (in effect is the slope equal to 1)?"

The same methods employed in the one-sample *t*-tests apply.

The test statistics are simply

$T_1 = \dfrac{\hat{\beta}_1 - \beta_1}{s_\varepsilon / \sqrt{SS_{XX}}} \sim t_{n-2}$ for the slope parameter, and

$T_0 = \dfrac{\hat{\beta}_0 - \beta_0}{s_\varepsilon \sqrt{\dfrac{\sum x_i^2}{n \cdot SS_{XX}}}} \sim t_{n-2}$ for the intercept parameter.

For example, under $H_0 : \beta_1 = 0$ then $T_1 = \dfrac{\hat{\beta}_1 - 0}{s_\varepsilon / \sqrt{SS_{XX}}}$

Presenting results

As mentioned earlier, ANOVA is a type of linear regression. Therefore, results from a linear regression with n data points and p parameters is summarised in ANOVA form for its ease of interpretation as:

	SS	*df*	*MSE*	*F*
Regression	*SSREG*	$p-1$	$\dfrac{SSREG}{df_{REG}}$	$\dfrac{MSE_{REG}}{MSE_{ERROR}}$
Error	*SSE*	$n-p$	$\dfrac{SSE}{df_{ERROR}}$	

Total	*SST*	

The overall significance of the regression is summarised by the *F* test-statistic, which is *F*-distributed with p - 1 numerator and n - p denominator degrees of freedom. If the statistic is greater than the specified critical *F*-value then we can reject the null hypothesis that there is no linear relationship between the x and y data.

Hypothesis tests are also summarised in table form as follows, where the *T* test-statistic is t-distributed with $n - 2$ degrees of freedom as follows:

Parameter	Estimate	*SE*	*T*
β_0	$\hat{\beta}_0$	$s_\varepsilon \cdot \sqrt{\frac{\sum x_i^2}{n \cdot SS_{XX}}}$	$\frac{\hat{\beta}_0 - \beta_0}{SE_{\beta_0}}$
β_1	$\hat{\beta}_1$	$s_\varepsilon / \sqrt{SS_{XX}}$	$\frac{\hat{\beta}_1 - \beta_1}{SE_{\beta_1}}$

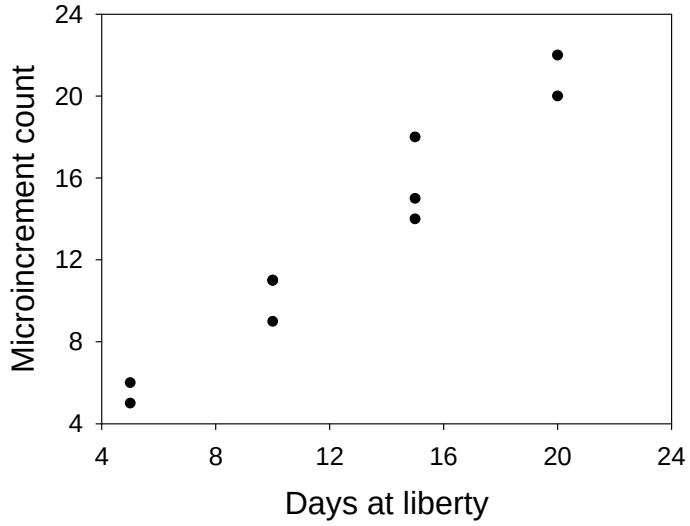
An example:

You are given the data from a daily growth ring validation study. In this study, you used a fluorochrome marker, such as the antibiotic oxytetracycline, to deposit a fluorescent band into the otolith of each fish at the date of marking. The fish were then released before being recaptured later and sacrificed. The otoliths of each fish were removed and the number of growth microincrements counted from the fluorescent band to the edge of the growing margin of the otolith. The number of days at liberty as therefore assumed known without error.

Time at liberty	Number of increments
5	6
5	5
10	9
10	11
10	11
15	14
15	18
15	15
20	22
20	20

1. Is there a significant relationship between the number of daily growth rings and the time of fish at liberty at the 5% level of significance?
2. How much variance is explained by the regression?
3. What is the confidence interval of the slope at the 5% level of significance?
4. Is the slope equal to 1 – i.e., one growth increment is equivalent to one day at the 5% level of significance?

The first step in the analysis process is to plot the data. Here it is below and the relationship appears to be linear. We then proceed further.



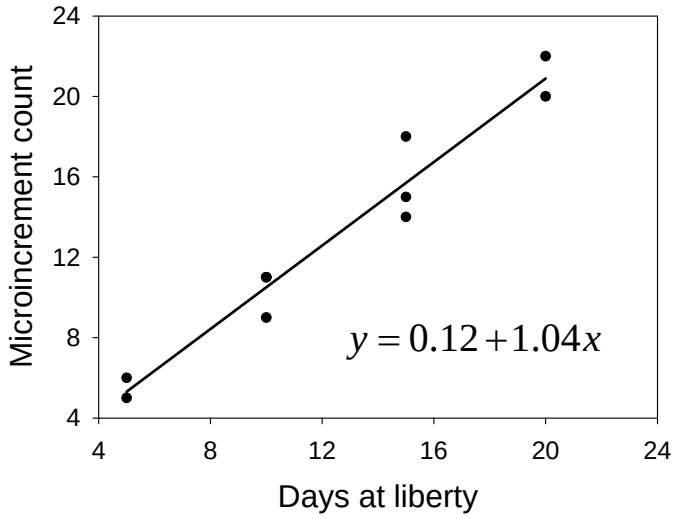
We now wish to estimate the slope and the intercept where.

As $SS_{XY} = \sum x_i y_i - n\overline{x}\overline{y} = 1910 - 10 \cdot 12.5 \cdot 13.1 = 272.5$
and $SS_{XX} = \sum x_i^2 - n\overline{x}^2 = 1825 - 10 \cdot 12.5^2 = 262.5$ then

Then the parameter estimates are

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{272.5}{262.5} = 1.04 \text{ and } \beta_0 = \overline{y} - \beta_1 \overline{x} = 13.1 - 1.04 \cdot 12.5 = 0.12$$

This can be plotted as



The next step is answer the first question pertaining to the overall significance of the model. We then calculate *SSREG* and *SSE* to fill in the ANOVA table under the null hypothesis that there is no relationship – or simply the variance of the regression is the same as the variance of the error.

The hypothesis is rejected if the ANOVA table F test-statistic is greater than the critical F statistic of $F_{0.05,1,8} = 5.31$.

	SS	*df*	*MSE*	*F*
Regression	283.92	1	283.92	161.96
Error	14.02	8	1.75	
Total	296.94			

As our test statistic is larger, we can reject the null hypothesis and conclude that there is a linear relationship between the days at liberty and the number of growth rings on the otoliths.

The second part of the question requires us to calculate the Coefficient of Determination, the R^2.

Therefore, as $R^2 = \frac{283.92}{296.94} = 0.96$, we can conclude that the linear regression model explained 96% of the total data variance.

The next two parts of the question pertain to inference about the slope parameter, $\hat{\beta}_1$, using either a confidence interval approach or a hypothesis testing approach.

The confidence interval of the slope is calculated from $\beta_1 = \hat{\beta}_1 \pm t_{\alpha/2,n-2} s_\varepsilon / \sqrt{SS_{XX}}$.

As $s_\varepsilon^2 = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{14.02}{10-2}} = 1.32$ and $t_{0.025,8} = 2.306$ from the inverse *t*-tables then

$\beta_1 = 1.04 \pm 2.306 \cdot 1.32 / \sqrt{262.5}$ or $0.85 \le \beta_1 \le 1.23$

The 95% confidence interval includes the value of 1 and points us towards noting that there is indeed a 1-to-1 relationship between days at liberty and increment counts

The third question is a hypothesis test, with the null hypothesis stating $H_0 : \beta_1 = 1.$

Using $\alpha = 0.05$ then we would reject the null hypothesis if of *T* test-statistic was greater than $t_{0.025,8} = 2.306$.

The test-statistic is calculated as $T_1 = \frac{\hat{\beta}_1 - \beta_1}{s_\varepsilon / \sqrt{SS_{XX}}} = \frac{1.04 - 1}{1.32 / \sqrt{262.5}} = 0.49$.

As this value is less than the critical *t*-statistic we fail to reject the null hypothesis and conclude that there is a 1-to-1 relationship between days at liberty and the number of growth increments deposited on the otoliths.