Ch. 6

Linear models[1]

"The shortest distance between two points is a straight line."
-- Archimedes

Questions to ponder:

- *How is a linear model used to describe variation in the value of a parameter?*
- *How do you put a discrete covariate in a linear model?*
- *What is a link function?*
- *How can we calculate the probability of survival from a set of estimated coefficients: B_0 and B_1?*

Thinking about parameters in linear models

You may be able to visualize the estimation of two survival rates for individuals in a sample—males and females, for example. It seems clear that we could divide the data into two groups—one for males and another for females. We could then estimate the survival rate for each of these sub-samples of the data.

In the language of parameter estimation, we would label "gender" in this example as a "covariate" of survival—we hypothesize that survival varies as gender varies. And, we note that "gender" is a discrete covariate—individuals are either male or female. We can place individuals into categories, based on the "gender" covariate.

In Chapter 4, we noted that modern software platforms for the estimation of parameters (e.g., MARK, SAS procs, R modules) provide the capability to assess continuous covariates such as mass, vegetation height, age (in days), body length, etc.. As such, we need to explore how this process works—as it is less intuitive than estimating survival for discrete groups, like males and females. *How can you estimate the probability of survival for individuals of a certain mass?*

To implement such analyses, we need to learn to view parameter estimation from the context of a **linear model**.

[1] *With thanks for content to Evan Cooch and Gary White*

Starting with the basics

If we return to what we know of the simplest linear model—the equation of a straight line—we know that models are often described, mathematically, as:

$$y = b + mx$$

This line describes how the value of y (our dependent variable) varies as the value of x changes. We may remember that a simple regression analysis would provide us with estimates of b, the intercept (where the line crosses the y-axis), and m, the slope (describes how steeply y changes with a change in x).

We need to modify the symbology used in this equation to match the manner in which quantitative biologists describe their models. So, we can write the same model using different symbols as:

$$\Upsilon = B_0 + B_1 x$$

Where

Υ is the response variable, such as the probability of capture, occupancy, or survival
B_0 = the intercept
B_1 = the slope, or the coefficient for x

We can then imagine a scenario for which the estimated value of the parameter increases as mass increases—it might look like Figure 6.1. We know that the slope, B_1, is positive, and it seems that mass is an important covariate for the parameter. This model has one covariate (mass) and we will estimate two parameters for this analysis (the intercept, B_0, and the covariate for mass, B_1). Hence, K=2.

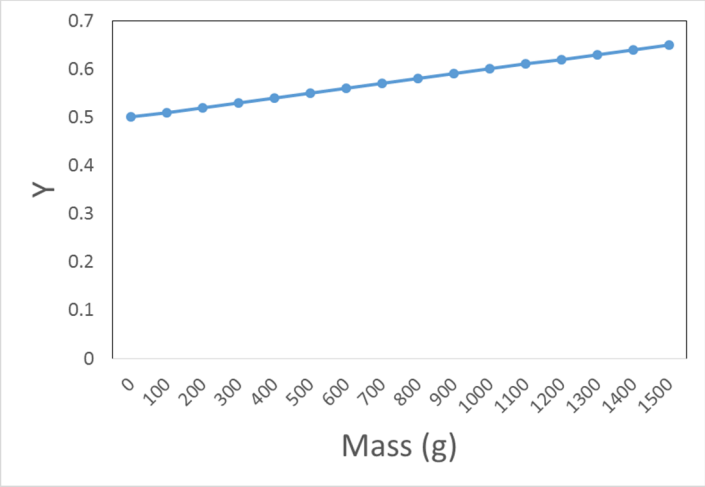

Figure 6.1: *Graphical illustration of the predictions of a hypothetical linear model showing a positive relationship between body mass and a response variable, Y.*

But, let's start even more simply: what does the linear model look like if mass does not affect the parameter value? The simplest linear model has only the intercept:

$$\Upsilon = B_0$$

And, if we plot this type of a model, we would see something like Figure 6.2 (at right)—showing that the value of the parameter is the same, regardless of mass. This model has no covariates, and there is one parameter estimated (the intercept): $K = 1$.

As you explore linear models, we suggest visualizing your analyses as figures to conceptualize the models you are using and creating.

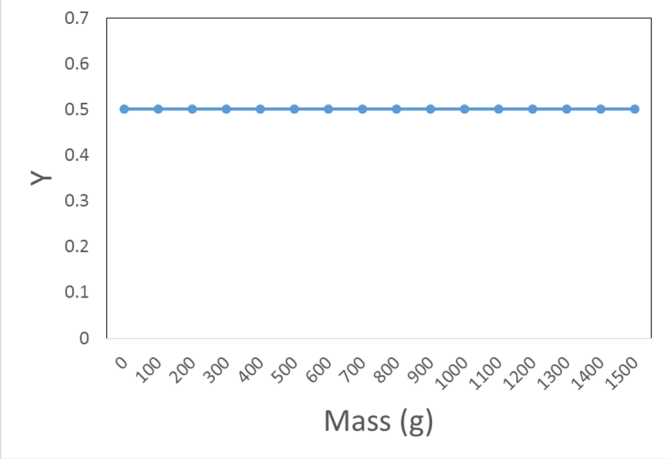

Figure 6.2: *Model predictions from a linear model with only an intercept ($Y=B_0$). Here, the response variable, Y, does not change with body mass. Such a model is often used as a 'null model' or 'intercept model' for comparison to more complicated models.*

Let's take a more complicated linear model—this time with two covariates: mass and age. Both covariates are continuous, ranging from 0 to some potentially infinite value.

The equation for this model would be:

$$Y = B_0 + B_1(\text{mass}) + B_2(\text{age})$$

We will be estimating three parameters, the intercept and the two coefficients for our covariates. If we were to plug in values for mass and age, we might expect to be able to provide the following summary of our predictions for the parameter value.

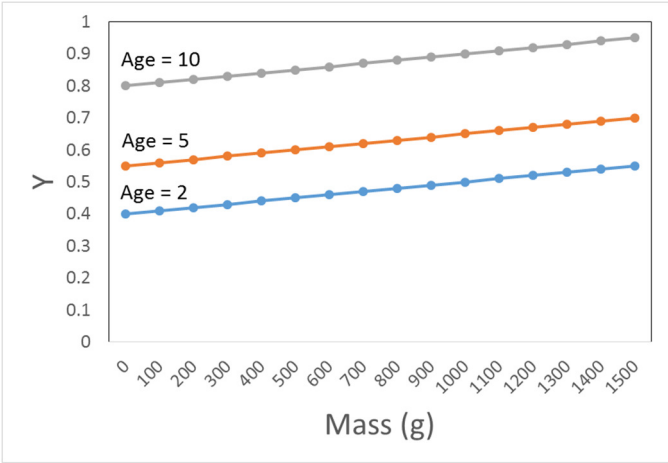

Figure 6.3: *Model predictions for the value of a response variable, Y, as body mass increases. Predictions are shown for 3 different ages (age 2, 5, and 10).*

Now, we can see that both mass and age of the individual are important. Our estimate of the parameter value, Y, is going to vary depending on the mass and age. You should be able to see that for individuals with a mass of 800 g, the oldest individuals would be predicted to have higher survival than younger individuals.

Dummy variables: for discrete variables

Now that we have conquered understanding of continuous covariates in linear models for parameter estimation, we need to come back and think about our simpler discrete models. How do we put a discrete covariate into a linear model?

Let's use the example of an effect of gender, as before. If we follow the general format of our continuous models, the model would look something like:

$$\Upsilon = B_0 + B_1(\text{gender})$$

And, that is correct! However…how do we put "male" and "female" into the model to calculate a value for the parameter, Y? The answer is—we create a "dummy variable".

"Dummy variables" are used to code discrete, or categorical, covariates. Once a covariate is selected, such as gender, we establish a binary code for the states of the covariate. For example, **we could use "0" to code for male and "1" to code for female.**

Thus, if we insert the value of the code for gender in our linear model, we get the following models:

Male: $\Upsilon = B_0 + B_1(0)$
Female: $\Upsilon = B_0 + B_1(1)$

You will note that because the "0" we insert for males eliminates the value of the covariate for gender ($0*B_1 = 0$), we end up with the intercept of the model (B_0) as the estimate for the parameter. Further, the value of the parameter for females is the value for males, plus or minus an amount controlled by the coefficient for gender. In essence, we are estimating a "female effect." Is the parameter estimate higher for females (B_1 for females >0) or lower (B_1 for females <0)?

We've used one dummy variable and the "0" and "1" coding to create a linear model for a discrete covariate with two states, such as male and female for gender. But, how do you treat a categorical covariate such as month—perhaps we have three values in our study: May, June, and July.

Use dummies! Don't be a dummy!

We stress that some discrete covariates "appear" to have numerical values, and some readers may be tempted to use the values as a continuous covariate. For example, **months** of the year can be given values from 1-12. However, **under no circumstances** should you create a model with a continuous covariate using the numbers associated with each month, such as:

$$\Upsilon = B_0 + B_1(\text{month})$$

Months are, of course, discrete units. June does not really equal "6", nor does July really equal "7". Further, July is not 1 unit larger than June—and December (month #12) is not twice as large in magnitude as June (month #6). Most importantly, the probability of survival should not be expected to increase in proportion to an increase of 1 'month unit'. Rather, **we should use dummy variables to create an appropriate linear model.**

In our previous example with gender, we had two states of the discrete covariate—and thus, we were able to use "1" and "0" to code for both states with a single covariate. Now, we have three states—the three months. We solve the problem by using two dummy variables as such:

	Covariates used	
State of discrete variable: *Month*	**June**	**July**
"May"	0	0
"June"	1	0
"July"	0	1

We note that it appears to be a **general rule for dummy variables** (and, yes—it's true!):

If we have n *states of a discrete covariate, we must have* n-*1 dummy variables to code for the states.*

We suggest using values of 0's for the first state—which serves as a baseline for the other states. Here, we establish two covariates "June" and "July". When we code the June column, we ask the question, "Is this state June"? If not, we place a "0". If yes, we place a "1". Similarly for the July column, we ask, "Is this state July?" If so, "1". If not, "0". "May", as indicated by the two 0's in the two columns for June and July is neither "June" nor "July".

The linear model, then, has 3 parameters—an intercept and the two covariates (June and July):

$$\Upsilon = B_0 + B_1(\text{June}) + B_2(\text{July})$$

To predict the value of the parameter for May, we insert the 0's as such:

$$\Upsilon_{\text{May}} = B_0 + B_1(0) + B_2(0)$$

Again, we note that the value of the parameter is the value of the intercept, as the June and July coefficients become 0—leaving us only with the intercept.

Similarly, for June and July:

$$\Upsilon_{\text{June}} = B_0 + B_1(1) + B_2(0)$$
$$\Upsilon_{\text{July}} = B_0 + B_1(0) + B_2(1)$$

And, we can again see that we are essentially estimating a "June" effect and a "July" effect for our parameter. How much higher or lower is the estimate, relative to our **baseline estimate** (the intercept) for May?

You try it

Suppose you are interested in estimating survival as a function of the land cover in which animals were living. You have 5 states for land cover: wetland, prairie, forest, crop fields, and urban. How many dummy variables would you use? Can you create a table similar to the table above for this problem? Can you write out the linear model that predicts survival in forest?

Link functions

We should now be comfortable thinking about parameter estimation in the context of a linear model. As an example, let us return to the interest in describing variation in survival by the mass of the animal and the age of the animal (both continuous covariates):

$$S = B_0 + B_1(\text{mass}) + B_2(\text{age})$$

As we prepare to estimate the coefficients, B_0, B_1 and B_2, we find that we are not performing a standard regression analysis. The first clue that we will need some special treatment is that we can see that our response variable, **S, or survival, needs to be bounded between 0 and 1**—that is, survival is a probability.

To ensure that $0 \leq S \leq 1$, we will need to use a link function. A link function literally links the response variable (alive or dead) with the explanatory variable, such as mass or age in our example. We will estimate our parameter as a transformed variable, and then we will back-transform to obtain our survival estimate.

Some **common link functions** used for parameter estimation include the following:

sin
logistic (or logit)
log

Note that the sin and logit link functions act to constrain the parameter value within the space 0.0 to 1.0. The log link function does not. For that reason, you may find that software such as program MARK suggest the log function when estimating population size (which does not have to be constrained between 0-1).

You should notice that we have been using the form of the linear model, for our previous discussions:

$$\Upsilon = B_0 + B_1(\text{mass}) + B_2(\text{age})$$

But, when using the logit link function, the linear model is actually:

$$\log\left(\frac{\gamma}{1-\gamma}\right) = \beta_0 + \beta_1(mass) + \beta_2(age)$$

Hence, the coefficients B_0, B_1, and B_2 are estimated as logit-scale coefficients. To make sense of them, and to use a linear model to predict the probability of survival under various values of the covariates, we must back-transform the logit-scale function to the "real world". We do this with the equation:

$$\gamma = \frac{\exp^{(\beta_0+\beta_1(mass)+\beta_2(age))}}{1+\exp^{(\beta_0+\beta_1(mass)+\beta_2(age))}}$$

Dipper example

Let's use a simple example, provided by Cooch and White (2014)—the infamous European dipper (*Cinclus cinclus*) data set used by Lebreton et al. (1992). Cooch and White (2014) used this example to show beginning users of program MARK how to perform an analysis of CJS-type data (Chapter 10), and the software provides an estimate of annual survival under a null, constant-survival model, S = 0.5658.

Figure 6.4: *European dipper (Cinclus cinclus), Kirkcudbright, Scotland. Photo by Mark Medcalf, and available in the public domain.*

But, how did the software obtain that estimate? Program MARK has estimated the survival estimate through the use of a simple linear model with a "sin" link function and no covariates:

$$sin\ S = B_0$$

Our software provides the maximum likelihood estimate of B_0, using the sin link function:

$$B_0=0.13203.$$

Because the coefficient, B_0, is a sin-scale coefficient, so we must back-transform it to make sense of it as a survival probability.

We do this with the general equation, which transforms sin-scale linear models to real-life survival probabilities that are constrained between 0.0 and 1.0:

$$S = \frac{\sin(B_0)+1}{2}$$

Hence:

$$S = \frac{\sin(0.13203)+1}{2} = 0.56582$$

That survival probability of S=0.56582 matches the earlier estimate that the software provided. And, we can now see how we can work back from a sin-scale coefficient to predict the probability of survival. This will be important when we begin our analyses and find coefficients (e.g., B_1) that are on a logit- or sin-scale.

Ring-necked pheasant nest survival example

The previous dipper example allowed us to practice back-transforming a linear model from the sin scale to a survival probability. But, it was the simplest type of linear model that only had an intercept and no covariate. The survival rate, then, was constant across space and time. What if survival varies?

To illustrate how to work with a more complex linear model, we will add one covariate. The ring-necked pheasant (*Phasianus colchicus*) is a species of economic importance in the central United States, although it is an introduced species. Landowners have interest in managing the species for hunting opportunities.

Matthews et al (2012) conducted a nest survival study, so the data were considered "known fate" (Chapter 9). In some initial investigations, the authors considered a model to evaluate the effects of bare ground on the daily probability of nest survival. Bare ground is measured as the "percent cover" of bare ground in a quadrat surrounding the nest site, and the average value for percent bare ground in their study was 8%.

The authors used the logit link function for their analysis, so the model was:

$$\log\left(\frac{S}{1-S}\right) = \beta_0 + \beta_1(\%\_bareground)$$

The estimates for the coefficients, B_0 and B_1, were:

B_0 = 3.4200653 (SE = 0.27)
B_1 = 0.0336702 (SE = 0.03)

How do we obtain estimates of survival along a gradient of bare ground conditions? For this example, we will ignore the standard error so that we can focus on our objective to show how to estimate nest survival with a single covariate. An astute reader may see that the standard error for the bare ground coefficient (B_1) is very close to the value of the coefficient (both are equal to approximately 0.03), which suggests that this effect may not be the most important factor to explain variation in survival of pheasant nests—but stay with us as we follow this as a simple example of a covariate in a linear model.

The first step to estimating survival at various levels of bare ground is to create a table, and we suggest using a spreadsheet to help with the calculations—although you can certainly do them by hand or with a calculator if needed. Our table will have three important columns. The first is the values of bare ground that we wish to consider (we will use 0% to 12%). The second, is our intercept, B_0. And, the third is the coefficient for our variable of interest (bare ground, in our example), B_1. We can then fill those columns with the values that will be used in our equations. Our coefficients, B_0 and B_1 will be constant in all rows of the table, and the values for bare ground that we wish to consider are selected along a suitable range for the data.

% bare ground	B_0	B_1	logit(S)	S
0	3.420065	0.03367	3.420065	0.968326
2	3.420065	0.03367	3.487406	0.970327
4	3.420065	0.03367	3.554746	0.972206
6	3.420065	0.03367	3.622087	0.973969
8	3.420065	0.03367	3.689427	0.975623
10	3.420065	0.03367	3.756767	0.977174
12	3.420065	0.03367	3.824108	0.978629

Now that we have our first three columns populated with numbers, we can start our calculations for *S*. We have highlighted four cells in the table above, logit(S) and S for the bare ground level of 0%, and logit(*S*) and *S* for the bare ground level of 10%. How are those values calculated?

We start with the logit(*S*) column, which is the logit-scale estimate of survival, and we remember our equation:

$$\log\left(\frac{S}{1-S}\right) = \beta_0 + \beta_1(\%\_bareground)$$

Therefore, for the level of bare ground of 0%, we calculate:

$$\log\left(\frac{S}{1-S}\right) = 3.420065 + (0.03367 * 0) = 3.420065$$

And, for the level of 10%, we calculate:

$$\log\left(\frac{S}{1-S}\right) = 3.420065 + (0.03367 * 10) = 3.756767$$

In similar fashion, we can fill in all other values of logit-scale survival for our range of values for bare ground. But, obviously, survival probability must be constrained between 0 and 1, so our next step is to back-transform our logit-scale survival to a probability by adapting the formula we used earlier in this chapter:

$$S = \frac{\exp^{(\beta_0+\beta_1(\%\_bareground))}}{1+\exp^{(\beta_0+\beta_1(\%\_bareground))}}$$

For our values of 0 and 10, we would calculate daily survival as:

$$S = \frac{\exp^{(3.420065)}}{1+\exp^{(3.420065)}} = 0.968326 \qquad S = \frac{\exp^{(3.756767)}}{1+\exp^{(3.756767)}} = 0.977174$$

And, we can finish the table with values for all other levels of bare ground. At this point, we can see the daily nest survival does increase, gradually, as bare ground increases (Figure 6.5). But,

survival only increases by approximately 1% as bare ground increases from 0% to 12%, which confirms our initial assessment from inspection of our model coefficients and standard errors.

You may also detect a faint curve in the line. You might ask, "How that is possible if this is a linear model?" And, it is true that we used a linear model to estimate our coefficients—but we did that in the logit-scale. When we back-transform, our linear model often appears non-linear with a slight curve because of the transformation.

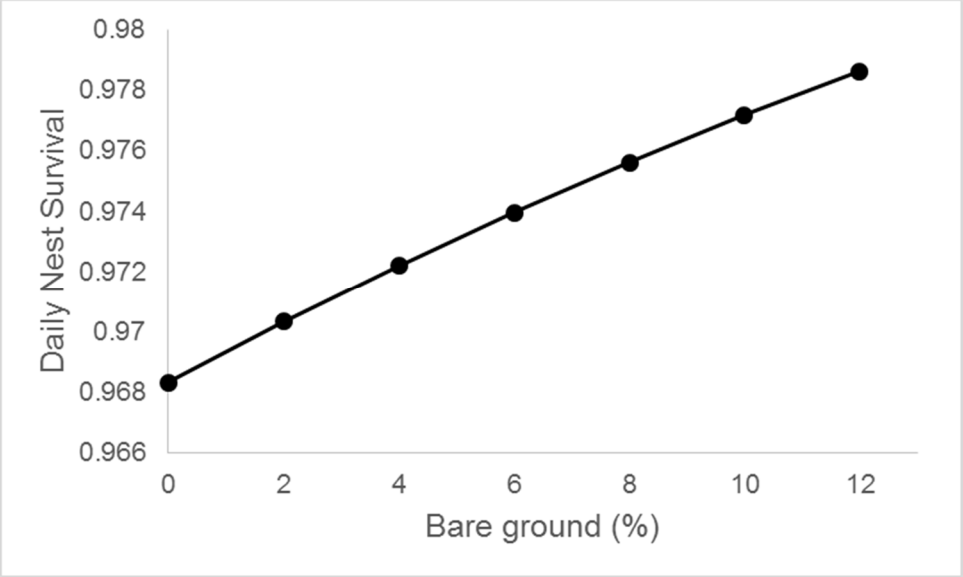

Figure 6.5: *Change in estimate of daily nest survival for ring-necked pheasant across a gradient of cover (%) of bare ground.*

Conclusion

The ability to conceptualize parameter estimation in the context of linear models—and the ability to write out the linear models for our analyses—takes our skill-set to a new level. To use linear models to estimate parameters such as probability of survival or capture, link functions are needed to constrain the resulting estimates between 0 and 1. We must use a back-transformation to predict the probability of survival rate after we are provided estimates of logit- or sin-scale coefficients for our linear models.

References

Cooch, E., and G. White. 2014. Chapter 6: Adding constraints: MARK and linear models. *In* Program MARK: a gentle introduction, 12th edition, Cooch, E. and G. White, eds. Online: http://www.phidot.org/software/mark/docs/book/pdf/chap6.pdf

Matthews, T. W., J. S. Taylor, J. S., and L. A. Powell. 2012. Mid-contract management of Conservation Reserve Program grasslands provides benefits for ring-necked pheasant nest and brood survival. Journal of Wildlife Management 76:1643-1652.

Lebreton, J. D., K. P. Burnham, J. Clobert, and D. R. Anderson. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs 62: 67-118.

For more information on topics in this chapter

Conroy, M. J., and J. P. Carroll. 2009. Quantitative Conservation of Vertebrates. Wiley-Blackwell: Sussex, UK.

Donovan, T. M., and J. Hines. 2007. Exercise 4: Single-species, single-season model with site level covariates. *In* Donovan, T. M., and J. Hines. Exercises in occupancy modeling and estimation. On-line: http://www.uvm.edu/envnr/vtcfwru/spreadsheets/occupancy/occupancy.htm

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and management of animal populations. Academic Press, San Diego.

Answers: You try it!

	covariates used				
State of discrete variable: ***land cover***	**prairie**	**forest**	**crop**	**fields**	**urban**
wetland	0	0	0	0	0
prairie	1	0	0	0	0
forest	0	1	0	0	0
crop	0	0	1	0	0
fields	0	0	0	1	0
urban	0	0	0	0	1

$\Upsilon_{Forest} = B_0 + B_1(0) + B_2(1)+B3(0)+B4(0)+B5(0)$

An aluminum leg band is used to tag a cliff swallow (Petrochelidon pyrrhonota) in a large mark-recapture study in Nebraska, USA. Biologists have conducted many analyses that incorporate linear models to detect sources of variation in probabilities of movement and survival for this species. Photo provided by Mary Bomberger Brown and used with permission.