Reading Guide - Chapter 17 Analysis of Biological Data 2nd ed.

brouwern@gmail.com

November 4, 2017

CHAPTER 17: REGRESSION

aka "linear regression", "linear model", "line of best fit", "least squares regression"

Done with the function lm() in R, which stands for linear model. lm() is used for regression and for ANOVA, which points towards the underlying similarity of these two methods.

Regression "is a method that predicts values of one numerical variable from values of another numerical variable." (pg 541). That is, the response AND the predictor are both numeric. This is in contrast to t-tests and ANOVA, where the predictors are categorical/factors.

In R-ish notation:

lm(continuous.response ~ continous.predictor.)

17.1 Linear Regression

Ex 17.1 The lion's nose

This is an interesting example of how to use regression for **prediction.**

The method of least squares (pg 542)

Figures 17.1-2

17.1-2 is a very important figure showin how the residuals change when different lines are fit to the data.

On page 543 they give the formula for the line as: Y = a + bX

Notation can vary between authors. Here, "a" is the intercept and "b" is the slope. It doesn't matter that they put the intercept first; it means the same thign to say Y = a + bX as to say Y = bX + a.

In words, this would be Y = intercept + slope*X

Slope "The slope of a lienar regresion is the rate of change in Y per unit of X" (pg 543). That is, how much Y changes as X changes.

FIgure 17.1-3

17.1-3 shows the difference between positive, negative, and flat slopes.

Calculating the slope and intercept (pg 544)

SKip the equations on page 544.

On page 545 they present the regression equation for the lion example: Age = 0.88 + 10.65(proportion black)

which could also be written as Age = 0.88 + 10.65*proportion black

and

Age = + 10.65*proportion black + 0.88

Where 10.65 is the slope and 0.88 is the intercept. This equation allows you to predict Age (the response variable) based on the proportion of the nose that is black.

Populations and samples (pg 545)

Predicted values (aka "Y hat", pg 546)

Key idea: "The predicted value of Y [Y hat] from a regression line estiamte the mean value of Y for all individuals having a given value of X" (pg 546). This is if you sample many many lions that had, say, a proption black on their noses of 0.5, their would be variation in their ages. However, the mean of all those many lions with proportion black of 0.05 would be 6.2, because according to the estimated regression equation

- age = 0.88 + 10.65*proportion black
- age = 0.88 + 10.65*0.5 = 6.2

A regression line can be thought of as a continous set of predictions. That is, a continous serioes of Y hats

Residuals (pg 546)

Residuals are the distance from the regression lion to each data pont. That is, the distance from Y hat (the regression line) to a real data point.

[Skipe the equations on pag 547]

Standard error of the slope (pg 547)

We won't worry about the precise definition, but its very important to know that the slope (and intercept) are estimated **with error** and that error is characterized by the standard error.

Confidence interval for the slope (pg 548)

Slopes (and intercepts) also have confidence intervals (CI). As before, 1.96*SE will give you an approximate confidence interval around the parameter.

Note that there is a difference between the SE and CI for a parameter, and the confidence interval the surround an entire line. To put a confidence intervals (or Confidence band) around an entire regression line invovles combining the uncertainty in the slope and in the intercept.

17.2 Confidence in predictions

There is a subtle differene between a **confidence band** and a **prediction interval**. Prediction intervals will always be bigger than confidence bands. I almost always focus on confidence bands because as the author states I am usually "interested in the overall trend" of the data. When you are interesetd in a particular value, such as a single lion, how black its nose is and what its age might be, things change.

Confidence intervals for predictions (pg 549)

Figure 17.2-1

17.2-1 Shows how CIs and Prediction INtervals difer.

The authors summarize: "**Confidence bands** measure the precision of the prediction *mean* for each value of X. **Prediction intrvals** measure the precision of the prediction single Y-values for each X" (pg 550)

Extrapolation (pg 550)

Extrapolation means making predictions beyond the range of the origianl data.

Figure 17.2-2

17.2-2 shows a dataset where ear length was only measured for people between 30 and 100 years. If you used the regression equation to estimate ear length of infants, you would be extrapolating to ages not in the originaly data. This is generally a bad idea!

Extrapolation (pg 551)

17.3 Testing hypotheses about a slope

On page 551 they show the equation for the t-statistic associated with the slope of a regression line. You don't need to know the equation, but do need to know that t-statistics do play a roll in regression.

Example 17.3 Prarie Home Campion

Figure 17.3-1

Note that in this example the data are from an experiment, but regression analysis is being used. Also, the y-axis is log transformed.

The t-test of regression slope (pg 552)

Don't worry about the equations; you do need to know that you can do a t-test to test the hypothesis that the slope of the line is different than zero. Note that this is very general hypothesis: the slope or even whether its positive or negative doesn't factor into the hypothesis, just that its *not* exactly 0.0.

The ANOVA approach (pg 554)

The details aren't important. YOu do need to know that regression models can be tested by comparing two models and an F-statistics generated, just like in ANOVA.

Using R2 to measure the fit of the line to data (pg 555)

R^2 tells you what "fraction of variation in Y... is 'explain' by x." That is, how much scatter there is around the regression line. If all the points fall on the line, then R^2 would be 1.0 (100% of variation explained.)

17.4 Regression toward the mean

[skipped, but very interesting topic]

17.5 Assumptions of regression

The author's list 4, I tend to focus on the middle 2:

- "At each value of X, the distribution of possible Y-values is normal" (pg 557)
- This relates to the concept discussed below about the "normality of the residuals."

- Figure 17.5-1 Shows what this would look like.
- This can be assessed using resdiual analysis using a qqplot, or a histogram of the residuals. A qqplot is preferred.
- Log transformation and frequently reduce the impact of this violation
- Biologist have frequently focused on this issue, thought its been shown that its not as much of a deal breaker as the next one...
- "The variance of the Y-values is the same at all values of X"
- This is easiest to show in a picture, which the author's don't do. See lecture notes.
- This is often called "unequal variance" or "heteroskedasticity" (dont worry, I won't make you spell it.)

Figure 17.5-1

Outliers

Outliers can have a strong influence on regression. This is because the line is fit by calculating the residuals and squaring them. An outlier will have a large residual because it is far from the line. Square this large residual, and it increases the sum of squared residual (sum of all the residuals after they have been squared.) So regression model fitting will try to reduce how large that residual is, throwing everything off.

Figure 17.5-2

Note the outlier in the lower left hand corner of the plant at around x = 33, y = 36. Imagine how large the residual would be.

Detecting nonlinearity (pg 559)

The discuss **smoothing** here, or what they more specifically call "scatterplot smoothing".

Detecting non-normality and unequal variance (pg 559)

Here they discuss regression diagnostics and residual analysis using a **residual plot**.

Residual plot

17.6 Transformations

Figure 17.6-1

THey show how curving lines (the 2 left-hand panels) can be straightened using logs. Note that they explicitly using ln for the natural log (but R uses log for the ln and log10 for the log)

Figure 17.6-2

Compare 17.5-3, which plots the raw data, to 17.6-2, where both the x and y axes are transformed.

Figure 17.6-3

These graphs show the impact of transformation on the residuals when there are problems with **unqual variance / heteroskedasticity** (sometiems called non-constant variance)

[17.7 The effects of measurement error on regression]

important topic, but skipped

17.8 Nonlinear regression

[A curve with an asymptote]

skip

Quadratic curves (pg 565)

NLB note: quadratic curves are AKA "x^2 terms", "squared terms", "quadratic terms", "squared effect", "quadratic effects"; I will try to be consistent but will probably fail...

Figure 17.8-2

Formula-free curve fitting (pg 566)]

The go into the details of **smoothing** here. You don't need to know the different kinds of smoothers, just what the overall goal is.

Ex 17.8 The incredible shrinking seal

Figure 17.8-3

The line through 17.8-3 is a smoothed curve.

17.9 Logistic regression: fitting a binary response variable (pg 567)

You should be familiar with the concept that logistic regression is used when you have a categorical response variable, like mortality, and a numeric predictor varialble, like time. You don't need to know the math.

Figure 17.9-1

In this example, the guppies are categorized as alive or dead. This is their categorical "response" to the treatment. The predictor variable is a continous variable: how long they were exposed to a cold treatment.

You don't need to know the math.

17.10 Summary