To investigate possible regression models for an ordinal predictor with three possible values.

ORIGIN := 1 Seed(1953) Set system parameters: origin for matrices and seed for random numbers

Introduction

Suppose that we have a predictor X that can take on one of three possible values X_1, X_2, X_3. We assume without loss of generality that X_3 may be encoded as 1 and X_1 as 0. If we have an assumption of monotonicty then we can encode X_2 as $0<\theta<1$.

We also assume that we have outcome values Y and that the frequencies of the three possible predictor values are f_1, f_2, f_3 with $f_1+f_2+f_3 = n$.

Note that the set-up above will be encountered commonly in genetics. Suppose we take a given locus and the possibles alleles are *a* and *A*. We then have three possible groups, the two homozygous groups X_1= *aa* and X_3= *AA* and the heterozygous group X_2= *aA*. In that case $\theta = 0$ is the case where *a* is dominant, $\theta=1$ is the case where *A* is dominant and $\theta=1/2$ is the case where there is a linear phenotypic response on alleles.

See chapter 25 of Senn, *Statistical Issues in Drug Development*, 2007 for a discussion

Formulation in terms of a linear combination

Consider the means for the three groups,
Y_{m1}, Y_{m2} and Y_{m3} (say)
and suppose that they have expectation

μ, $\mu+\theta\tau$ and $\mu+t$.

Now consider a linear estimator of this form:

$w_1 Y_{m1} + w_2 Y_{m2} + w_3 Y_{m3}$.

We now investigate what conditions the weights, w_1, w_2, w_3 have to satisfy.

Solutions for weights

We require this estimator to have an expectation τ.

We thus require that the sum of the weights is 1

$$w_1 + w_2 + w_3 - 0 \text{ solve}, w_1 \rightarrow -w_2 - w_3$$

However we also clearly require that

$$-(w_2 + w_3)\cdot\mu + w_2(\mu + \theta\cdot\tau) + w_3\cdot(\mu + \tau) = \tau$$

$$-(w_2 + w_3)\cdot\mu + w_2\cdot(\mu + \theta\cdot\tau) + w_3\cdot(\mu + \tau) - \tau \text{ solve}, w_2 \rightarrow -\frac{w_3 - 1}{\theta}$$

Hence $w_2 = \frac{1 - w_3}{\theta}$

and we can solve fully for w_1 in terms of w_3.

$$-\left(w_2 + w_3\right) \text{ substitute}, w_2 = \frac{1 - w_3}{\theta} \rightarrow -\frac{\theta \cdot w_3 - w_3 + 1}{\theta}$$

$$-\frac{\theta \cdot w_3 - w_3 + 1}{\theta} \text{ simplify} \rightarrow \frac{w_3 - 1}{\theta} - w_3$$

Hence must we have an estimator of the form

$$\left(\frac{w_3 - 1}{\theta} - w_3\right) \cdot Y_{m1} + \left(\frac{1 - w_3}{\theta}\right) \cdot Y_{m2} + w_3 \cdot Y_{m3}$$

if it is to be unbiased

The least squares estimator

Now consider the variance of this estimator. It is proportional to

$$\frac{\left(\frac{w_3 - 1}{\theta} - w_3\right)^2}{f_1} + \frac{\left(\frac{1 - w_3}{\theta}\right)^2}{f_2} + \frac{{w_3}^2}{f_3}$$

We wish to minimise this, so we obtain the derivative

Derivative

$$\frac{d}{dw_3}\left[\frac{\left(\frac{w_3-1}{\theta}-w_3\right)^2}{f_1}+\frac{\left(\frac{1-w_3}{\theta}\right)^2}{f_2}+\frac{{w_3}^2}{f_3}\right]\rightarrow\frac{2\cdot w_3}{f_3}-\frac{2\cdot\left(w_3-\frac{w_3-1}{\theta}\right)\cdot\left(\frac{1}{\theta}-1\right)}{f_1}+\frac{2\cdot w_3-2}{\theta^2\cdot f_2}$$

Find least squares solution

$$\frac{2\cdot w_3}{f_3}-\frac{2\cdot\left(w_3-\frac{w_3-1}{\theta}\right)\cdot\left(\frac{1}{\theta}-1\right)}{f_1}+\frac{2\cdot w_3-2}{\theta^2\cdot f_2}\ \text{solve},w_3\rightarrow\frac{f_1\cdot f_3+f_2\cdot f_3-\theta\cdot f_2\cdot f_3}{f_1\cdot f_3+f_2\cdot f_3-2\cdot\theta\cdot f_2\cdot f_3+\theta^2\cdot f_1\cdot f_2+\theta^2\cdot f_2\cdot f_3}$$

This is the solution for the weight *w3*. The other weights can be calculated from this one.

Define function for the weights of the various groups

First w_3

$$w_3\left(f_1,f_2,f_3,\theta\right):=\frac{f_1\cdot f_3+f_2\cdot f_3-\theta\cdot f_2\cdot f_3}{f_1\cdot f_3+f_2\cdot f_3-2\cdot\theta\cdot f_2\cdot f_3+\theta^2\cdot f_1\cdot f_2+\theta^2\cdot f_2\cdot f_3}$$

Then w_2

$$\frac{1 - w_3}{\theta} \text{ substitute}, w_3 = \frac{f_1 \cdot f_3 + f_2 \cdot f_3 - \theta \cdot f_2 \cdot f_3}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3} \rightarrow \frac{\theta \cdot f_1 \cdot f_2 - f_2 \cdot f_3 + \theta \cdot f_2 \cdot f_3}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3}$$

$$w_2(f_1, f_2, f_3, \theta) := \frac{\theta \cdot f_1 \cdot f_2 - f_2 \cdot f_3 + \theta \cdot f_2 \cdot f_3}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3}$$

Finally w_1

$$\frac{w_3 - 1}{\theta} - w_3 \text{ substitute}, w_3 = \frac{f_1 \cdot f_3 + f_2 \cdot f_3 - \theta \cdot f_2 \cdot f_3}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3} \rightarrow -\frac{f_1 \cdot (f_3 + \theta \cdot f_2)}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3}$$

$$w_1(f_1, f_2, f_3, \theta) := -\frac{f_1 \cdot (f_3 + \theta \cdot f_2)}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3}$$

Some special solutions

In the following three examples we illustrate the weight, w_2, for the middle group

Group 2 like group 3

$$\frac{\theta \cdot f_1 \cdot f_2 - f_2 \cdot f_3 + \theta \cdot f_2 \cdot f_3}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3} \text{ substitute}, \theta = 1 \rightarrow \frac{f_2}{f_2 + f_3}$$

Note how the weight is simply the fraction that number of subjects in the second group represent of the total of the number in group 2 and 3 which is logical in view of the fact that the response in these two groups is identical.

Group 2 like group 1

$$\frac{\theta \cdot f_1 \cdot f_2 - f_2 \cdot f_3 + \theta \cdot f_2 \cdot f_3}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3} \text{ substitute}, \theta = 0 \rightarrow -\frac{f_2}{f_1 + f_2}$$

Note how the weight is simply the fraction that the number of subjects in the second group represents of the total of the number in group 2 and 1 (but with a minus sign) which is logical in view of the fact that the response in these two groups is identical and they will therefore be subtracted from the group 3 response to give the estimate of the contrast.

Linear response case

$$\frac{\theta \cdot f_1 \cdot f_2 - f_2 \cdot f_3 + \theta \cdot f_2 \cdot f_3}{f_1 \cdot f_3 + f_2 \cdot f_3 - 2 \cdot \theta \cdot f_2 \cdot f_3 + \theta^2 \cdot f_1 \cdot f_2 + \theta^2 \cdot f_2 \cdot f_3} \text{ substitute}, \theta = \frac{1}{2} \rightarrow \frac{2 \cdot f_1 \cdot f_2 - 2 \cdot f_2 \cdot f_3}{f_1 \cdot f_2 + 4 \cdot f_1 \cdot f_3 + f_2 \cdot f_3}$$

Now it is much more difficult to see what is going on. However, consider the special case where, $f_1 = f_3$.

$$\frac{2 \cdot f_1 \cdot f_2 - 2 \cdot f_2 \cdot f_3}{f_1 \cdot f_2 + 4 \cdot f_1 \cdot f_3 + f_2 \cdot f_3} \text{ substitute}, f_1 = f_3 \rightarrow 0$$

Then in that case the middle group plays no part at all, which is logical.

Illustration of the solution

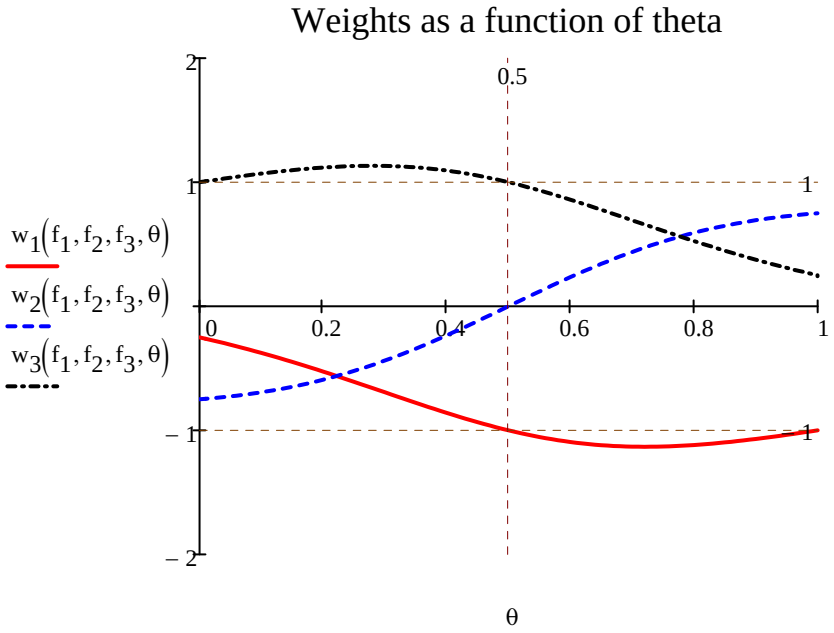
Now let us plot some possible weights

Set possible parameters

$n := 100$

$f_1 := 20 \qquad f_3 := 20 \qquad f_2 := n - f_1 - f_3$

$\theta := 0, 0.01 .. 1$



Discussion of a possible use

We can make a possible use of this as follows.

We take a plausible range of values for θ. We estimate the treatment effect, τ,
and its variance for each possible value. We consider our estimates in consequence. (Of course,
this suggests that we could include a profile likelihood approach but this will not be considered here.)

A Simulated Example

Set parameters

$f_1 := 30$ $f_2 := 20$ $f_3 := 50$ Number of observations per group

$\mu := 0$ $\theta := 0.3$ $\tau := 1.2$ Parameters determining expected response. Note that because θ is not equal to 0.5 the response is not perfectly linear.

$\sigma := 1.5$ Standard deviation of error term

Generate vectors

$Y_1 := \text{rnorm}(f_1, \mu, \sigma)$ $Y_2 := \text{rnorm}(f_2, \mu + \theta \cdot \tau, \sigma)$ $Y_3 := \text{rnorm}(f_3, \mu + \tau, \sigma)$

Calculate means

$$Y_{m1} := \sum_{i=1}^{f_1} \frac{Y_{1_i}}{f_1} \qquad Y_{m1} = 0.426 \qquad Y_{m2} := \sum_{i=1}^{f_2} \frac{Y_{2_i}}{f_2} \qquad Y_{m2} = 0.471$$

$$Y_{m3} := \sum_{i=1}^{f_3} \frac{Y_{3_i}}{f_3} \qquad Y_{m3} = 1.29$$

Calculate variance

$$\mathrm{Var}_{pooled} := \frac{\sum_{i=1}^{f_1}\left(Y1_i - Ym1\right)^2 + \sum_{i=1}^{f_2}\left(Y2_i - Ym2\right)^2 + \sum_{i=1}^{f_3}\left(Y3_i - Ym3\right)^2}{f_1 + f_2 + f_3 - 3} \qquad \mathrm{Var}_{pooled} = 2.48$$

Define weighted estimate

$$\tau_{weight}(\theta) := w_1\left(f_1, f_2, f_3, \theta\right)\cdot Ym1 + w_2\left(f_1, f_2, f_3, \theta\right)\cdot Ym2 + w_3\left(f_1, f_2, f_3, \theta\right)\cdot Ym3$$

Define weighted standard error

$$SE_{weight}(\theta) := \sqrt{\left(\frac{w_1\left(f_1, f_2, f_3, \theta\right)^2}{f_1} + \frac{w_2\left(f_1, f_2, f_3, \theta\right)^2}{f_2} + \frac{w_3\left(f_1, f_2, f_3, \theta\right)^2}{f_3}\right)\cdot \mathrm{Var}_{pooled}}$$

Define critical value of t

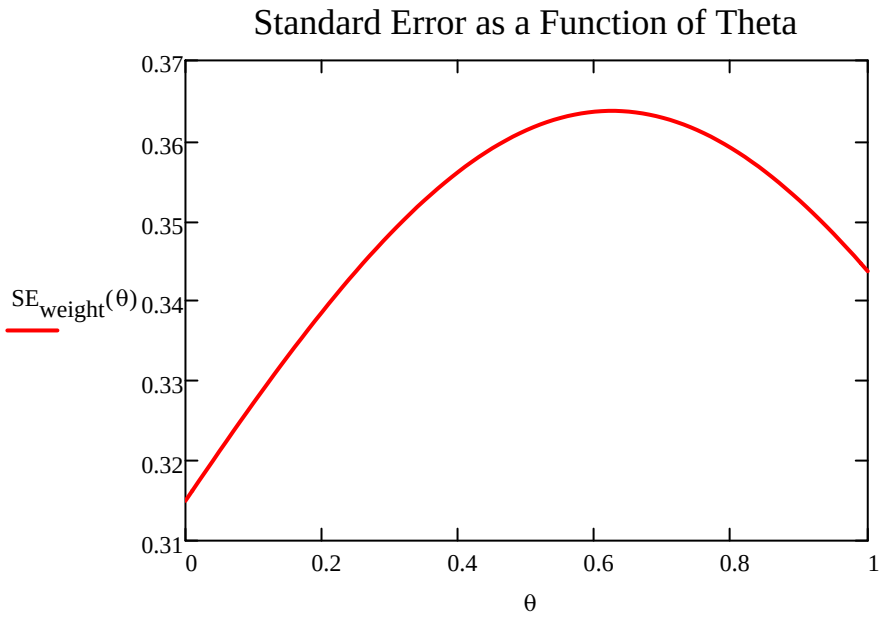
$$t_{crit} := qt\left(0.975, f_1 + f_2 + f_3 - 3\right) \qquad t_{crit} = 1.985$$

Define confidence limits as a function of θ

$$LCL(\theta) := \tau_{weight}(\theta) - t_{crit} \cdot SE_{weight}(\theta) \qquad UCL(\theta) := \tau_{weight}(\theta) + t_{crit} \cdot SE_{weight}(\theta)$$

$$\theta := 0, 0.01 .. 1$$

Plot standard error as a function of θ



Now illustrate linear regression solution

Create matrices and vectors

$$X_1 := \left| \begin{array}{l} \text{for } i \in 1..f_1 \\ \quad X_{1_i} \leftarrow 0 \\ X_1 \end{array} \right. \qquad X_2 := \left| \begin{array}{l} \text{for } i \in 1..f_2 \\ \quad X_{2_i} \leftarrow \frac{1}{2} \\ X_2 \end{array} \right. \qquad X_3 := \left| \begin{array}{l} \text{for } i \in 1..f_3 \\ \quad X_{3_i} \leftarrow 1 \\ X_3 \end{array} \right.$$

$$n := f_1 + f_2 + f_3$$

$$\text{intercept} := \left| \begin{array}{l} \text{for } i \in 1..n \\ \quad \text{intercept}_i \leftarrow 1 \\ \text{intercept} \end{array} \right.$$

$X := \text{augment}(\text{intercept}, \text{stack}(X_1, X_2, X_3))$ Matrix of predictors

$Y := \text{stack}(Y_1, Y_2, Y_3)$ Matrix of outcomes

$\beta := (X^T X)^{-1} \cdot X^T Y$ $\beta = \begin{pmatrix} 0.324 \\ 0.904 \end{pmatrix}$ $\tau_{\text{weight}}\left(\frac{1}{2}\right) = 0.904$

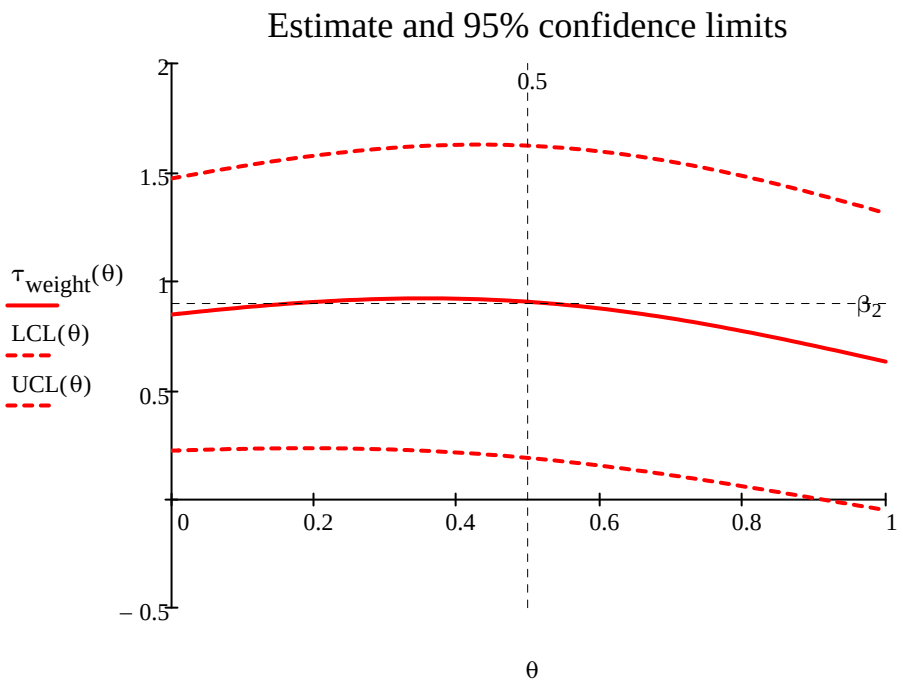
This shows that the value of the weighted estimator with θ=1/2 is just the ordinary regression estimator

$$Y_{hat} := X \cdot \beta \qquad resid := Y - Y_{hat} \qquad Var_{res} := \frac{resid^{T} \cdot resid}{n - 2}$$

$Var_{res} = 2.4787$ $Var_{pooled} = 2.4799$

Note the slight difference. This is because one is based on *n*-2 degrees of freedom and the other on *n*-3 degrees of freedom. One could argue that the latter is better

Plot confidence interval as a function of θ



Pragmatic suggestion

A pragmatic suggestion is to fit a linear regression on the assumption that the response is approximately linear and illustrate its sensitivity by using the sort of confidence interval plot as a function of assumed known θ.

Another roughly equivalent approach would be to calculate orthogonal polynomials, fit the linear polynomial as a primary approach and then examine the quadratic term for departure from linearity. This approach could be easily generalised to predictors with more than two categories.

Yes another approach is to implement ***isotonic regression***, but this is not illustrated here.