Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Saddles and Crossings

We can explore functions of two variables like we did with parabolae:

$$f(x,y)=axy+bx+cy+d=a(x+A)(y+B)+C=axy+aBx+aAy+C$$

and so:

$$A=\frac{c}{a} \quad B=\frac{b}{a} \text{ and } C=d-\frac{bc}{a}$$

and then:

$$f(x,y)=axy+bx+cy+d=a(x+\frac{c}{a})(y+\frac{b}{a})+d-\frac{bc}{a} \text{ where } a\neq 0$$

Notice that when either $x=-\frac{c}{a}$ or $y=-\frac{b}{a}$ we see that $f(x,y)=d-\frac{bc}{a}$

If $x>-\frac{c}{a}$ then f as a function of y is a line with slope $a(x+\frac{c}{a})$ This line has a slope with the same sign as a. This slope increases as x increases.

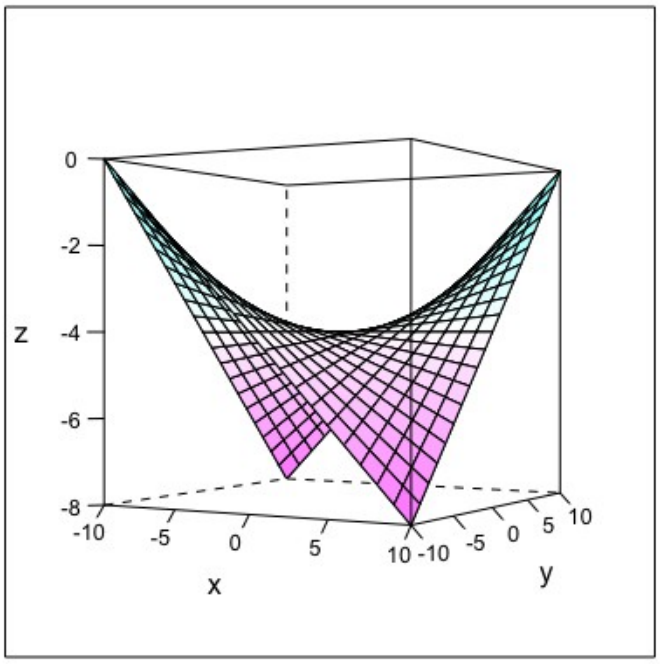
If $x<-\frac{c}{a}$ then f as a function of y is a line with slope $a(x+\frac{c}{a})$ This line has a slope with the negative sign times a. This slope decreases as x decreases.

If $y>-\frac{b}{a}$ then f as a function of x is a line with slope $a(y+\frac{b}{a})$ This line has a slope with the same sign as a. This slope increases as y increases.

If $y<-\frac{b}{a}$ then f as a function of x is a line with slope $a(y+\frac{b}{a})$ This line has a slope with the negative sign times a. This slope decreases as y decreases.

Sometimes, $x=-\frac{c}{a}$ and $y=-\frac{b}{a}$ are called crossings. This function is called a hyperbolic paraboloid and, more colloquially, it is called a saddle. This function is, indeed, the shape of a saddle.

See the picture below : z = f(x,y) = x*y/25 - 4
Notice the crossings. If x = 0, then f = - 4 no matter what y is. If y=0, the f = - 4 no matter what x is. Observe how the slopes change as one looks at lines parallel to x=0 and parallel to y=0.



Lets consider a measured exposure E and a measured potential confounder/modifier x. Assuming linearity :

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 x+\beta_2 E+\beta_3 Ex$$

This can be written as:

$$\log\left(\frac{p}{1-p}\right)=\beta_3\left(x+\frac{\beta_2}{\beta_3}\right)\left(E+\frac{\beta_1}{\beta_3}\right)+\beta_0-\frac{\beta_1\beta_2}{\beta_3} \quad \text{where} \quad \beta_3\neq 0$$

When $x=-\frac{\beta_2}{\beta_3}$, then as a function of E, we get a horizontal line : the log of odds does not depend on E.

When $x>-\frac{\beta_2}{\beta_3}$, then as a function of E, we get a line with the same sign as β_3

When $x<-\frac{\beta_2}{\beta_3}$, we get a line with the sign changed.

So $x=-\frac{\beta_2}{\beta_3}$ is the value that leads to a change in sign. This is the notion of a crossing.
In one range of x, the log of odds increases with increasing values of E. In the other range, the log of odds decreases with increasing values of E. Depending on the context, this may be an unattractive feature of this model. It may be that $x=-\frac{\beta_2}{\beta_3}$ is outside the range of possible values of x and so this feature is not a concern.

We also have a crossing of $E=-\frac{\beta_1}{\beta_3}$ In one range of E, the log of odds increases with increasing values of x. In the other range, the log of odds decreases with increasing values of x. Depending on the context, this [again] may be an unattractive feature of this model. It may be that $E=-\frac{\beta_1}{\beta_3}$ is outside the range of possible values of x and so this feature is not a concern.

In both cases, the crossing : $E=-\frac{\beta_1}{\beta_3}$ or $x=-\frac{\beta_2}{\beta_3}$ and the constant value at the crossing : $\log(\frac{p}{1-p})=\beta_0-\frac{\beta_1\beta_2}{\beta_3}$ can be estimated and confidence intervals can be determined using nlcom [with the Delta method]

```
. use wcgs.dta
. gen chola = chol*age
. logit chd age chol chola

Logistic regression                             Number of obs     =      3,142
                                                LR chi2(3)        =     116.82
                                                Prob > chi2       =     0.0000
Log likelihood = -831.18785                     Pseudo R2         =     0.0657

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1825711     .06192     2.95   0.003     .0612102     .303932
        chol |   .0338578   .0118737     2.85   0.004     .0105858    .0571298
       chola |  -.0004583   .0002481    -1.85   0.065    -.0009446    .0000281
       _cons |  -13.93615   2.973337    -4.69   0.000    -19.76378   -8.108518
------------------------------------------------------------------------------

. nlcom - _b[chol]/_b[chola]

       _nl_1:  - _b[chol]/_b[chola]

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   73.87952   14.64455     5.04   0.000     45.17672    102.5823
------------------------------------------------------------------------------

. nlcom - _b[age]/_b[chola]

       _nl_1:  - _b[age]/_b[chola]

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   398.3795   86.69219     4.60   0.000     228.4659    568.2931
------------------------------------------------------------------------------
. nlcom _b[_cons] - _b[age]*_b[chol]/_b[chola]

       _nl_1:  _b[_cons] - _b[age]*_b[chol]/_b[chola]

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |  -.4478859    1.07958    -0.41   0.678    -2.563823    1.668052
------------------------------------------------------------------------------
```

The fit is :

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.0004583(A-73.87952)(C-398.3795) - 0.4478859$$

The range of ages is 39 to 59 years so the age crossing [A=73.87952] is outside this range.
The range of cholesterol is 103 to 645 mg/dL so this crossing [C=398.3795] is in the range.
You may find it instructive to change cholesterol to mmol/L by multiplying by 0.02586.
Also try out centring age and cholesterol.

We can now add a dichotomous variable, say gender : G. This addition takes us back to :

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 G + \beta_2 A + \beta_3 GA + \beta_4 E + \beta_5 GE + \beta_6 AE + \beta_7 GAE$$

Now, with a different rearranging, we have :

$$\log\left(\frac{p}{1-p}\right) = (\beta_6 + \beta_7 G)AE + (\beta_2 + \beta_3 G)A + (\beta_4 + \beta_5 G)E + (\beta_0 + \beta_1 G)$$

$$(\beta_6 + \beta_7 G)\ \left[A + \frac{\beta_4 + \beta_5 G}{\beta_6 + \beta_7 G}\right]\ \left[E + \frac{\beta_2 + \beta_3 G}{\beta_6 + \beta_7 G}\right]\ +\ (\beta_0 + \beta_1 G)\ -\ \frac{(\beta_2 + \beta_3 G)(\beta_4 + \beta_5 G)}{\beta_6 + \beta_7 G}$$

So we can see that there can be two saddles. One saddle for G=0 [when $\beta_6 \neq 0$] and one for G=1 [when $\beta_6 + \beta_7 \neq 0$]
One could study how the estimated crossings depend on gender. There may be some insights as to the nature of the modification and again some potential criticism of the model if the crossings make no sense.

One could have two measured exposures and then notions of interaction.

Three Measured Variables : Lots of Saddles

Now lets move to three measured explanatory variables. The background needed is just a bit more elaborate than two measured.

$$f(x,y,z)=axyz+bxy+cxz+dyz+ex+fy+gz+h$$

Now think of this function for given values of z

$$f(x,y,z)=(az+b)xy+(cz+e)x+(dz+f)y+(gz+h)$$

Now, as we did for two variables,

$$f(x,y,z)=(az+b)(x+A)(y+B)+C$$

so

$$A=-\frac{cz+e}{az+b}\quad B=-\frac{dz+f}{az+b}\quad C=(gz+h)-\frac{(cz+e)(dz+f)}{az+b}$$

For each value of z, we get a saddle [a hyperbolic paraboloid]. The crossings are now curves. [setting x=A or y=B]

This process could have been done for given x of for given y.

Parabolae and Lines

We start with:

$$f(x,y)=a\,x^2y+bxy+cx^2+dx+ey+f$$

Now notice that, for given x, we get lines in y

$$f(x,y)=a\,x^2y+bxy+cx^2+dx+ey+f=(ax^2+bx+c)y+(cx^2+dx+f)$$

Now, for given y, we get parabolae in x.

$$f(x,y)=(ay+c)x^2+(by+d)x+(ey+f)=(ay+c)\left(x+\frac{(by+d)}{2(ay+c)}\right)^2+(ey+f)-\frac{(by+d)^2}{4(ay+c)}$$

Parabolae for Both

$$f(x,y)=ax^2y^2+bxy^2+cy^2+dx^2y+exy+fy+gx^2+hx+i$$

Notice that, for given x, we get parabolae in y :

$$f(x,y)=(ax^2+bx+c)y^2+(dx^2+ex+f)y+(gx^2+hx+i)$$

Similarly for given y, we get parabolae in x.