Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Two Measured Explanatory Variables

There are a number of options for interpreting models with two measured explanatory variables.

Now we have a function of two variables: f(x,y)

One can consider sections by, say, fixing some values for y and graphing the function of x [or vice versa].

One can also consider contour plots z=f(x,y) for various values of z.

The simplest function f would be the additive function:

f(x,y) = a + bx + cy
The graph of this function is a plane.

Fixing y, we get a series of lines in x with intercept 'a+cy' and slope 'b': f(x,y) = a+cy + bx
Fixing x, f(x,y) = a + bx +cy with intercept 'a+bx' and slope 'c'.

Or we can graph the contours z = a + bx + cy
which will be lines:

y = (z-a)/c - (b/c)x

The next simplest function would be:

f(x,y) = axy + bx + cy +d
The graph of this function is not a plane but is an example of a ruled surface. A surface S is said to be ruled if through every point of S there is a line that lies on S.

Fixing y, we again get a series of lines in x now with intercept 'd+cy' and slope 'b+ay':
f(x,y) = d + cy + (b+ay)x.
Or fixing x, f(x,y) = d + bx + (c+ax)y with intercept 'd +bx' and slope 'c+ax'
So we get a [doubly] ruled surface : lines for given x and lines for given y

Or we can graph the contours z= axy + bx + cy + d
which will be curves:

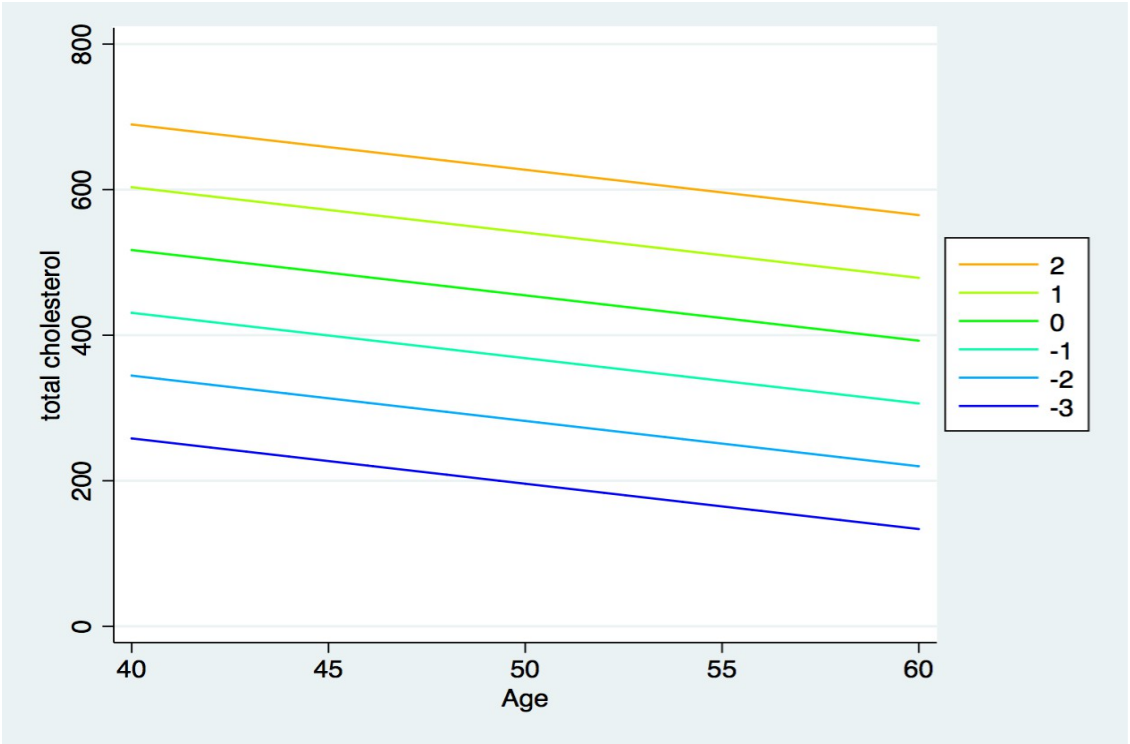
$$y=\frac{z-(d+bx)}{c+ax}$$

Notice that when a=0 in these curves, we get lines again.

An example should help here.

First an 'additive' function :

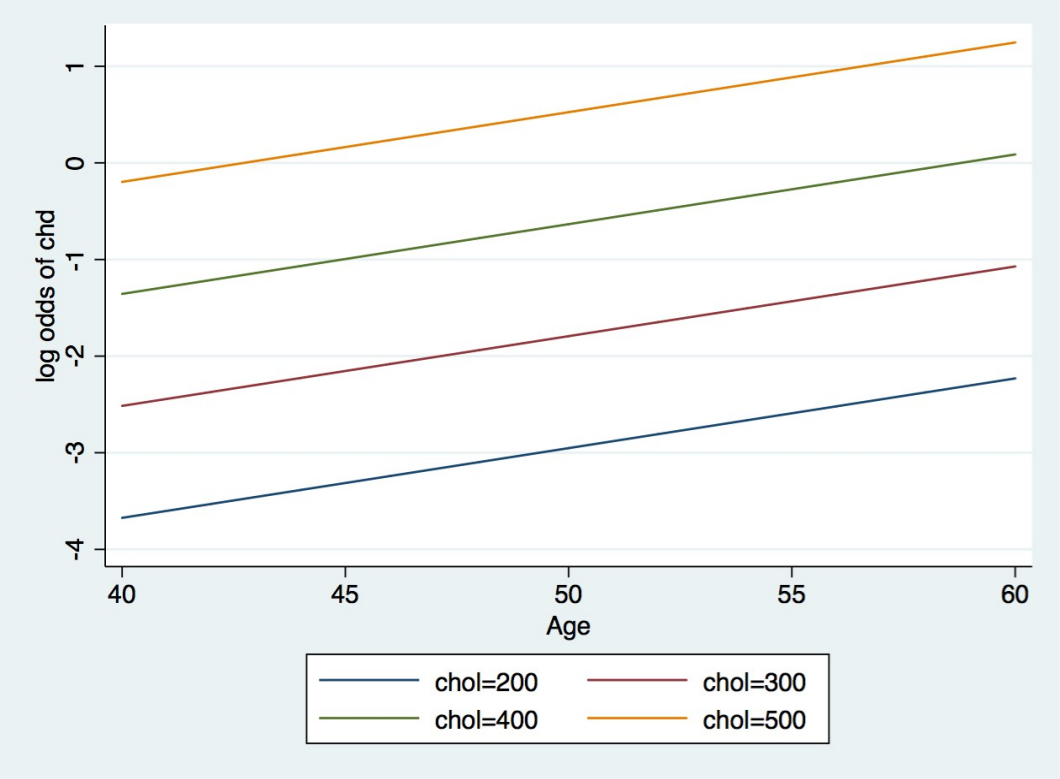
```
. use wcgs.dta
. logit chd chol age smoke

Logistic regression                                     Number of obs     =      3,142
                                                        LR chi2(3)        =     130.85
                                                        Prob > chi2       =     0.0000
Log likelihood = -824.17584                             Pseudo R2         =     0.0735

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   .0115898   .0014588     7.94   0.000     .0087306     .014449
         age |   .0721628     .01163     6.20   0.000     .0493683    .0949572
       smoke |   .5646608   .1370408     4.12   0.000     .2960659    .8332558
       _cons |  -8.878352   .6732304   -13.19   0.000    -10.19786   -7.558845
------------------------------------------------------------------------------

. summ age chol

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      3,154    46.27869    5.524045        39         59
        chol |      3,142    226.3724    43.42043       103        645

. quietly margins, predict(xb) at(age=(40(2)60) chol=(100(20)700) smoke=0)
saving(for_cl, replace)

. use for_cl.dta,clear
(Created by command margins; also see char list)

. twoway contourline _margin _at1 _at2, colorlines ccuts(-3(1)2)
```



Notice here that chol is vertical (y) and age is horizontal (x) and smoke = 0. We are seeing estimated log odds of chd (z) for -3, -2, -1, 0, 1, 2. The graphed lines are:
chol = z/0.0116 + 8.8784/0.0116 - 0.0721/0.0116 age
chol = 86.2069z + 516.7593 - 6.2155 (age - 40)

Or we can graph the sections: estimated log odds of chd by age for selected values of chol [smoke=0]

```
. twoway (line _margin _at2 if _at1==200)(line _margin _at2 if _at1==300)(line
_margin _at2 if _at1==400)(line _margin _at2 if _at1==500),legend(order (1
"chol=200" 2 "chol=300" 3 "chol=400" 4 "chol=500")) ytitle("log odds of chd")
```



The graphed lines are:
estimated log odds of chd = -3.6904 + 0.0115(chol-200) + 0.0722(age-40)

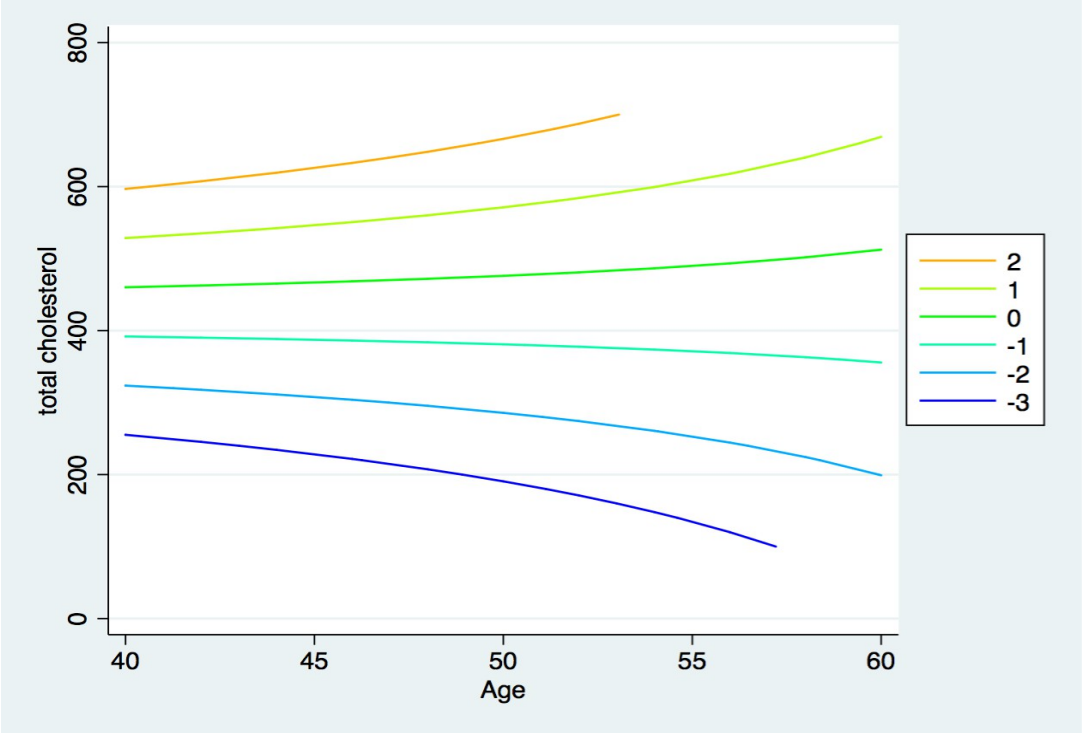
As well, one could graph the estimated log odds of chd versus chol for selected values of age.

Let us now consider the slightly more complicated function:

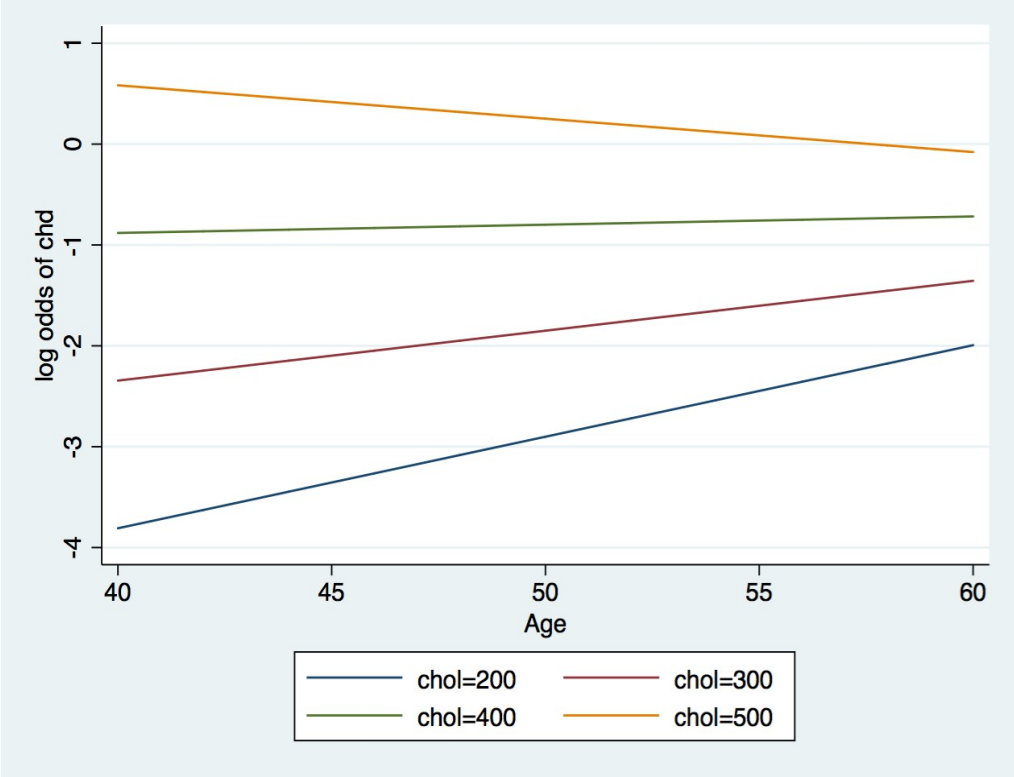
f(x,y) = axy + bx + cy +d : with an additive piece.

```
. logit chd c.chol##c.age smoke

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   .0311532   .0118724     2.62   0.009     .0078837    .0544228
         age |   .1733368   .0619082     2.80   0.005     .0519991    .2946746
             |
```

```
c.chol#c.age |  -.0004129   .0002483    -1.66   0.096    -.0008995    .0000737
             |
       smoke |   .5549429    .137184     4.05   0.000     .2860672    .8238186
       _cons |  -13.66999   2.969042    -4.60   0.000    -19.48921   -7.850777
------------------------------------------------------------------------------

. quietly margins, predict(xb) at(age=(40(2)60) chol=(100(20)700) smoke=0)
saving(for_cl, replace)

. use for_cl.dta

. twoway contourline _margin _at1 _at2,colorlines ccuts(-3(1)2)
```




```
. twoway (line _margin _at2 if _at1==200)(line _margin _at2 if _at1==300)(line
_margin _at2 if _at1==400)(line _margin _at2 if _at1==500),legend(order (1
"chol=200" 2 "chol=300" 3 "chol=400" 4 "chol=500")) ytitle("log odds of chd")
```



Let us now proceed like we did with parabolae:

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

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

and then:

$f(x,y)=axy+bx+cy+d=a(x+\frac{c}{a})(y+\frac{b}{a})+d-\frac{bc}{a}$ 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. It is called a hyperbolic paraboloid.

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

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

This can be written as:

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

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 is $x=-\frac{\beta_2}{\beta_3}$ 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 is $E=-\frac{\beta_1}{\beta_3}$ 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]

```
. 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.
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.