Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 15 :

Fitted Values :
Estimates of Conditional Means
Predictors for Individuals

Linear Regression Models look like...

...

$$E(y) = \sum_{i=0}^{k} \beta_i x_i$$

Once a fit is obtained, we have

$$\mathrm{Y} = \hat{\mathrm{y}} = \sum_{i=0}^{k} \hat{\beta}_i x_i = \sum_{i=0}^{k} b_i x_i$$

For the data at hand, we can compute:

for each set of the observed variables $x_{i\alpha}$:

$$\mathrm{Y}_{\alpha} = \sum_{i=0}^{k} b_i x_{i\alpha}$$

and then compare this value with the actual value observed with that set of variables: y_{α}

The residuals can be studied:

$$\mathrm{r}_{\alpha} = \mathrm{y}_{\alpha} - \mathrm{Y}_{\alpha}$$

Estimates of a Conditional Mean

Indeed, one can consider: $$\mathrm{Y} = \sum_{i=0}^{k} b_i x_i$$

for any potential set of values of the variables. One then gets an estimates of the mean for that set of values.

Further, one can determine the standard error of this estimate (the estimate is just a linear combination of the variables)

LDL, BMI and Statins

Lets us consider a model for expected LDL:

$$E(LDL) = \beta_0 + \beta_1 B + \beta_2 S + \beta_3 B * S$$

where B= BMI and S= indicator for Statin use

```
. use hers.dta
. regr ldl b s bs

      Source |       SS       df       MS              Number of obs =    2747
-------------+------------------------------           F(  3,  2743) =   49.33
       Model |  200927.514     3  66975.8381            Prob > F      =  0.0000
    Residual |  3724447.14  2743  1357.80063            R-squared     =  0.0512
-------------+------------------------------           Adj R-squared =  0.0501
       Total |  3925374.66  2746  1429.48822            Root MSE      =  36.848

------------------------------------------------------------------------------
         ldl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           b |   .6396803   .1563539     4.09   0.000     .3330971    .9462635
           s |   3.873074   7.817708     0.50   0.620    -11.45612    19.20226
          bs |  -.7206643   .2691888    -2.68   0.007    -1.248497    -.192831
       _cons |    132.812   4.561636    29.11   0.000     123.8674    141.7566
------------------------------------------------------------------------------
```

Estimates and confidence intervals

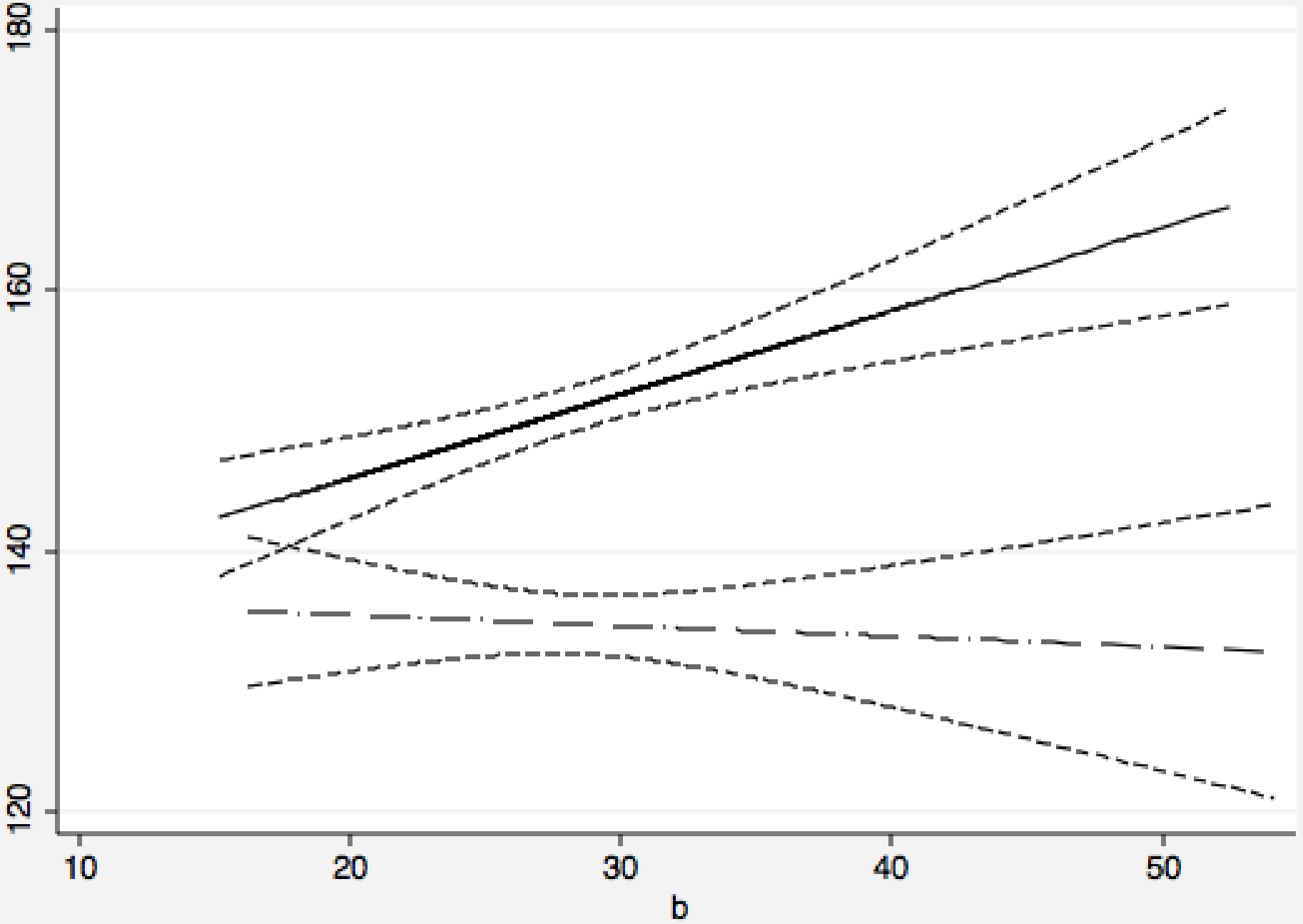
the code:

```
. predict Y,xb
. predict see,stdp

. gen cil=Y-1.96*see
(5 missing values generated)

. gen ciu=Y+1.96*see
(5 missing values generated)

. twoway (line Y b if s==0)(line cil b if s==0,lpattern(-)) (line ciu b if
s==0,lpattern(-))(line Y b if s==1)(line cil b if s==1,lpattern(-)) (line ciu b if
s==1,lpattern(-)), legend(off) scheme(s2mono)
```


180
160
140
120
10
20
30
40
50
b

Vertical Scale is now...

...the mean (or centre of mass) of a conditional probability distribution. The functional form of the distribution is assumed fixed and the variance of said distribution does not depend on the condition(s).

The magnitude of this variance (or equivalently this standard deviation) is viewed to be of secondary importance.

Previously, with logistic regression...

...these now familiar graphs have the vertical scale as the log odds of an outcome (0=absence 1=presence).

Here, the distributions were just Bernouilli (or Binomial). The outcome was a zero or a one.

The 'form' of this distribution being completed determined by the log odds (or the probability) of outcome =1

With linear regression,

we have focussed our attention on the mean of the distribution: a single piece of the distribution.

Let us now consider the entire distribution and reexpress the model as:

$$y = \sum_{i=0}^{k} \beta_i x_i + \epsilon$$

An actual response differs from the expected value by a term ϵ

The 'error' distribution

The difference between a response and its mean is often called the error.

The error distribution has mean zero (in principle) and [constant] variance σ^2

The consideration of E(y) + ϵ is, in a real sense, a statement about a prediction of a response for an individual with set of variables x_i

Notice that, unlike logistic regression, one can consider a prediction, here, directly without the specification of a classification rule.

The variance of prediction

Using the assumption the estimate of the conditional mean is independent of the error, we can see that:

$$Var(Y+e)=Var\,Y+Var(\epsilon)$$

Or, the variance of prediction is variance of estimation plus the variance of the error.

The catch...

We see that the standard error of prediction is:

$$\sqrt{Var\,Y + \sigma^2}$$

Often times, the variance σ^2 is much larger than Var Y.

This variance σ^2 is estimated by the mean square error (MSE).

When prediction is being assessed, attention can focus on the root(MSE). The magnitude of the root MSE directly informs the utility of a candidate model for the purposes of prediction.

LDL & BMI relationship modified by Statin use

Here the Root MSE is 36.848.

This number is not materially different from the SD of the LDL measurements:

```
. summ ldl

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         ldl |      2752    145.0385    37.80322        36.8      393.4
```

so... in what sense does knowledge of a person's BMI and their Statin use, inform us about their LDL level?

Predictions and prediction Intervals

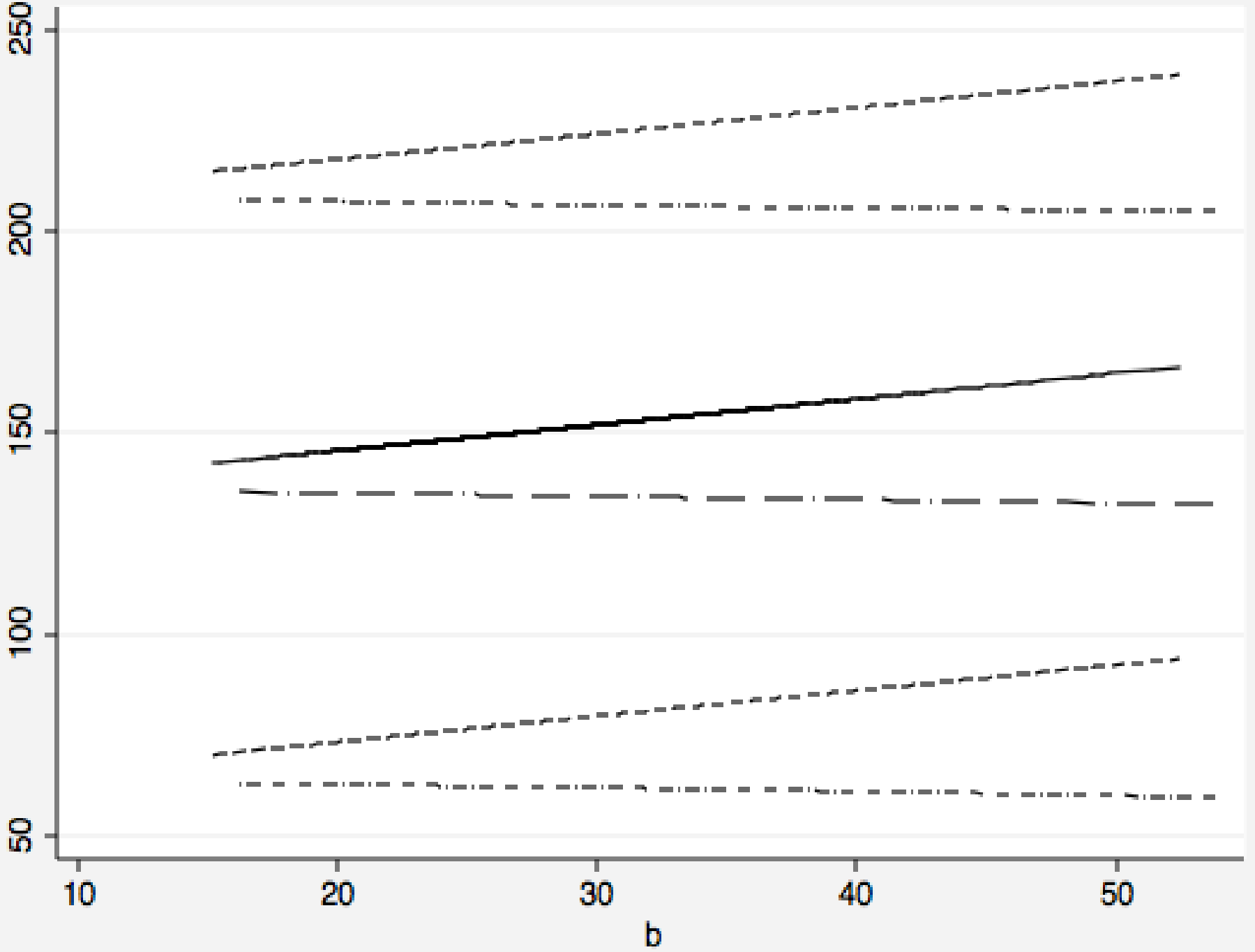
the code:

```
. predict sep,stdf

. gen pil=Y-1.96*sep
(5 missing values generated)

. gen piu=Y+1.96*sep
(5 missing values generated)

. twoway (line Y b if s==0)(line pil b if s==0,lpattern(-)) (line piu b if
s==0,lpattern(-))(line Y b if s==1)(line pil b if s==1,lpattern(-)) (line piu b if
s==1,lpattern(-)), legend(off) scheme(s2mono)
```


250
200
150
100
50
10
20
30
40
50
b

Identify the curves

The graph of the predictions and the prediction intervals illustrates the massive difference here between estimation and prediction.

This model has little utility in the prediction of an individual's LDL.