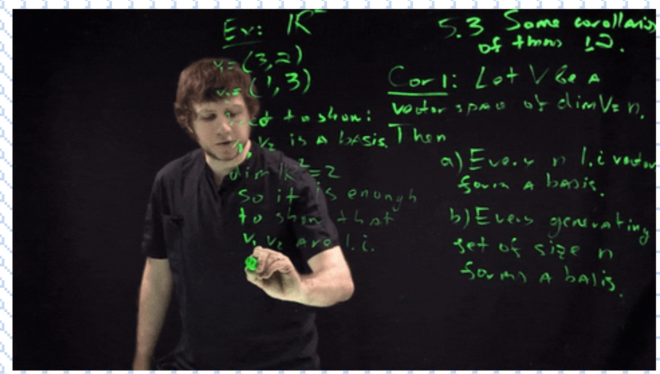
Biostatistics 200

Linear and Logistic Models

Multiple Regression & Logistic Regression



Multiple regression

- Additional explanatory variables might increase our understanding of a dependent variable
- We can posit the population equation

  $\mu_{y|x1,x2,\ldots,xq} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q$
- α is the mean of y when all the explanatory variables are 0
- β_i is the change in the mean value of y the corresponds to a 1 unit change in x_i when all the other explanatory variables are held constant
- $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon$

Multiple regression

- $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon$
- Assumptions
  - $x_1, x_2, \ldots, x_q$ are measured without error
  - The distribution of y is normal with mean $\mu_{y|x1,x2,\ldots,xq}$ and standard deviation $\sigma_{y|x1,x2,\ldots,xq}$
  - The population regression model holds
  - For any set of values of the explanatory variables, $x_1, x_2, \ldots, x_q$, $\sigma_{y|x1,x2,\ldots,xq}$ is constant – homoscedasticity
  - The y outcomes are independent
- We estimate the regression line

$\hat{y} = \widehat{\alpha} + \widehat{\beta_1}\, x_1 + \widehat{\beta_2}\, x_2 + \ldots \widehat{\beta_q}\, x_q$ using the method of least squares to minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_q x_{qi})]^2$$

Multiple regression

- For one explanatory variable – the regression model represents a straight line through a cloud of points in 2 dimensions
- With 2 explanatory variables, the model is a plane in 3 dimensional space (one for each variable), etc
- In Stata we just add explanatory variables to the regress statement
- regress fev age ht

Now the F-test has 2 degrees of freedom in the numerator because there are 2 explanatory variables

R^2 will always increase as you add more variables into the model

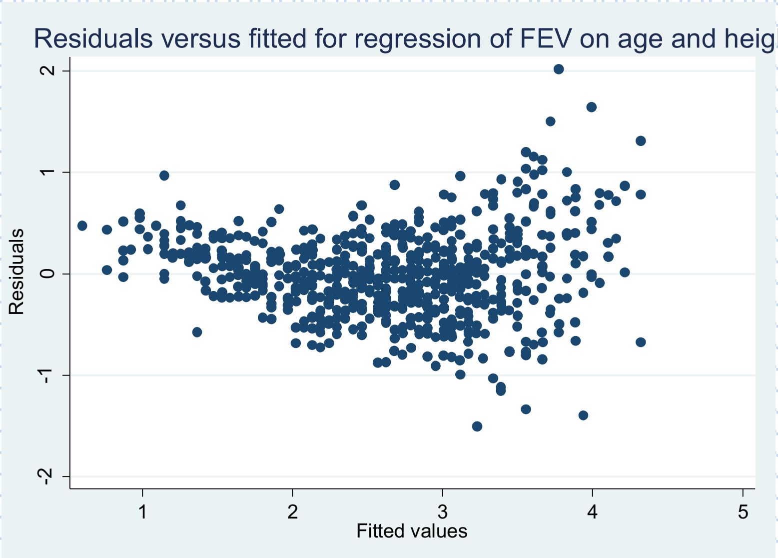
The Adj R-squared accounts for the addition of variables and is comparable across models with different numbers of parameters

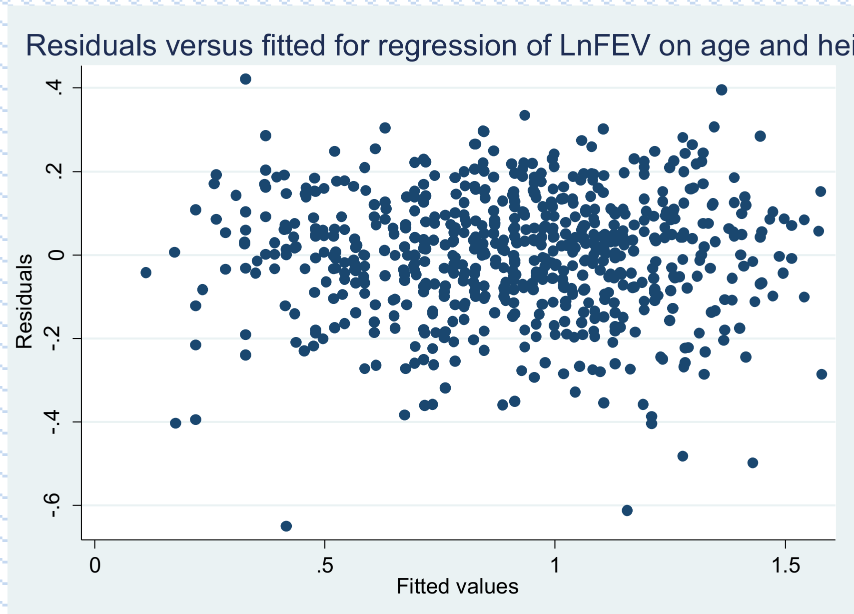
```
. regress fev age ht

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  2,   651) = 1067.96
       Model |  376.244941     2  188.122471           Prob > F      =  0.0000
    Residual |  114.674892   651  .176151908           R-squared     =  0.7664
-------------+------------------------------           Adj R-squared =  0.7657
       Total |  490.919833   653  .751791475           Root MSE      =   .4197

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0542807   .0091061     5.96   0.000     .0363998    .0721616
          ht |   .1097118   .0047162    23.26   0.000      .100451    .1189726
       _cons |  -4.610466   .2242706   -20.56   0.000    -5.050847   -4.170085
------------------------------------------------------------------------------
```

Examine the residuals…





Logistic regression

- For linear regression
  - The predicted values can be from $-\infty$ to $+\infty$
  - Y values are assumed to follow a normal distribution
- For many situations we want to model a dichotomous outcome, such as disease/no disease and cannot use linear regression
- We want to model p, the probability of disease

Logistic regression

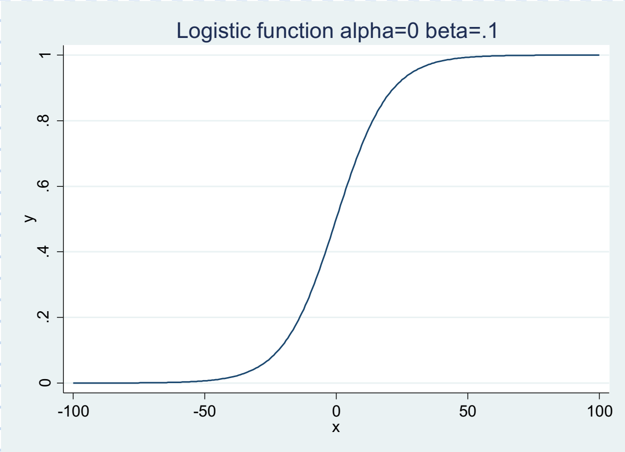
- $p=P(Y=1)$ i.e. P(outcome=disease)
- A model of the form $p= \alpha + \beta x$ would be able to take on negative values or values more than 1
- $p=e^{\alpha + \beta x}$ is an improvement because it cannot be negative, but it still could be greater than 1

- $p=P(Y=1)$ i.e. P(outcome=disease)
- A model of the form $p= \alpha + \beta x$ would be able to take on negative values or values more than 1
- $p=e^{\alpha + \beta x}$ is an improvement because it cannot be negative, but it still could be greater than 1

Logistic regression

- How about the function?

$$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$



- When $\alpha + \beta x = 0$, $p=1/(1+1) = 0.5$
- When x=big negative number, p nears 0/(1+0) = 0
- When x=big positive number, p nears (big number)/(1+big number) = 1
- The function models the probability slowly increasing over the values of X, until there is a steep rise, and another leveling off

Logistic regression

- The logistic function

$$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

- Now check this out:

$$1-p = 1-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1+e^{\alpha+\beta x}-e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{\alpha+\beta x}}$$

- So odds of disease/success are

$$\frac{p}{1-p} = \frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{\frac{1}{1+e^{\alpha+\beta x}}} = e^{\alpha+\beta x}$$

- And therefore $\ln(p/(1-p)) = \ln(e^{\alpha+\beta x}) = \alpha + bx$

Logistic regression

- So instead of assuming that the relationship between p and X is linear (as in linear regression), we are assuming that the relationship between ln(p/(1-p)) and X is linear.
- ln(p/(1-p)) is called the logit function, Logit(p) = ln(p/(1-p))
- In generalized linear models, some function of the outcome variable is assumed to be linear

  e.g. $f(y) = \alpha + \beta x$ or $f(p) = \alpha + \beta x$

  e.g. $\text{logit}(p) = \alpha + \beta x$

Interpretation of coefficients

- One dichotomous explanatory variable
- For x=0 $\quad \text{logit}(p_0) = \ln(p_0/(1-p_0))$

  $= \ln(P(Y=1|X=0)/(1- P(Y=1|X=0)))$

  $=$ ln(odds of the outcome when x=0)

  $= \alpha + \beta*x = \alpha + \beta*0 = \alpha$
- For x=1 $\quad \text{logit}(p_1) = \ln(p_1/(1-p_1))$

  $= \ln(P(Y=1|X=1)/(1- P(Y=1|X=1)))$

  $=$ odds of the outcome when x=1

  $= \alpha + \beta*1 = \alpha + \beta$

Difference between treatment arms in PEth outcome (>50 ng/ml)*

```
logistic peth_audit_50 i.studyarm_n

Logistic regression                             Number of obs     =        324
                                                LR chi2(1)        =       1.14
                                                Prob > chi2       =     0.2858
Log likelihood = -223.91123                     Pseudo R2         =     0.0025

------------------------------------------------------------------------------
peth_audit_50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
 1.studyarm_n |   .7869577   .1768227    -1.07   0.286     .5066332    1.222388
        _cons |    1.05618   .1562084     0.37   0.712     .7903976    1.411335
------------------------------------------------------------------------------
```

*BREATH cohort and comparisons_v3.dta

Difference between treatment arms in PEth outcome (>50 ng/ml)

logistic depvar indepvar

```
logistic peth_audit_50 i.studyarm_n

Logistic regression                               Number of obs    =        324
                                                  LR chi2(1)       =       1.14
                                                  Prob > chi2      =     0.2858
Log likelihood = -223.91123                       Pseudo R2        =     0.0025

------------------------------------------------------------------------------
peth_audit_50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
 1.studyarm_n |   .7869577   .1768227    -1.07   0.286     .5066332    1.222388
        _cons |    1.05618   .1562084     0.37   0.712     .7903976    1.411335
------------------------------------------------------------------------------
```

$=e^{\beta}$

This is used in comparing models of the same data set

Test of whether the model fit differs from the model with only α and no covariates

The Pseudo R^2 is an attempt to be analogous with linear regression but it is not

Stata logistic notes

- logistic depvar indepvars
  - gives you odds ratios
- logistic depvar indepvars, coef
  - gives you coefficients (alphas and betas)
- logit depvar indepvars
  - gives you coefficients
- logit depvar indepvars, or
  - gives you odds ratios

```
logit peth_50  i.studyarm_n

Iteration 0:   log likelihood = -198.01065
Iteration 1:   log likelihood = -197.99359
Iteration 2:   log likelihood = -197.99359

Logistic regression                             Number of obs     =        287
                                                LR chi2(1)        =       0.03
                                                Prob > chi2       =     0.8534
Log likelihood = -197.99359                     Pseudo R2         =     0.0001

------------------------------------------------------------------------------
     peth_50 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.studyarm_n |  -.0453478   .2455377    -0.18   0.853    -.5265928    .4358972
       _cons |  -.1438942   .1490438    -0.97   0.334    -.4360146    .1482262
-------------------------------------------------------------------------------
-
```

Odds ratio of PEth>=50 for comparison vs. standard group

```
 tabodds peth_audit_50 studyarm_n, or

---------------------------------------------------------------------------
  studyarm_n |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
           0 |    1.000000          .            .             .          .
           1 |    0.786958       1.13       0.2867      0.505895   1.224172
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     1.13
                                  Pr>chi2  =   0.2867

Score test for trend of odds:     chi2(1)  =     1.13
                                  Pr>chi2  =   0.2867
```

For <u>one independent variable that is categorical,</u> the odds ratio estimate is the same for logistic regression as for tabular methods.

What is the odds ratio?

- The odds ratio represents the change in the odds of membership in the "target" category (e.g., the category coded as 1.0), given a 1-unit increase in the predictor variable.
- Odds ratio is the increase (or decrease if negative) in odds of being in one outcome category when the value of the predictor variable increases by one unit.
- The odds ratio is calculated by the formula: e^g (where e = 2.718, and g = the b coefficient from the logistic regression output).
- For example:
  - An odds ratio of 1.5 indicates that that outcome of "1" is 1.5 times as likely (or 50% more likely) with a 1-unit increase in the predictor. The odds increase by 50%.
  - An odds ratio of 0.80 indicates that an outcome of "1" is 0.8 times as likely (or 20% less likely [1 – 0.8 = 0.2] with a 1-unit increase in the predictor; the odds are decreased by 20%.

How do you compute the probability of being in the "target" group, for a specific individual?

- Multiply each predictor value by its "b" coefficient, and add them to the constant. That sum is called "g."
- Use "g" in the following formula:
- $e^g / 1 + e^g$