Mathematics 231

Lecture 30

Liam O'Brien

Announcements

- Reading
  - Today M&M 10.1 559-576
  - Next M&M 11 607-627

Review: Simple Linear Regression

- The term "simple" may be a little misleading.
- Basic idea: We have n pairs of values – a set of n responses, and a set of n predictors.
- We attempt to describe the relationship between the response and predictor with a straight line.
- We fit the straight line in such a way that the distance between our observed responses, and our predicted responses, are as small as possible.

Response and Explanatory

- The y-variable is called the response, or dependent, variable.
- The x-variable is called the predictor, explanatory, or independent variable.
- It is of interest to use the explanatory variable to help predict the response.
- This relationship won’t be perfect, but hopefully the explanatory variable will explain a lot of the response variable’s behavior.

Sample vs. Population

- Just as in all statistical procedures, we can draw a distinction between the sample and population.

  The line relating the observed response, y_i, and the explanatory variable, x_i, for the sample is given by,

  $$y_i = b_0 + b_1 x_i + e_i$$

  For the population, this line is represented by,

  $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  The errors, represented by e_i and ε_i, have a mean (or expected value) of 0.

Terminology

The letters/symbols are fairly standard in regression:

$\hat{y}_i$ is the predicted value of y_i for subject i.

e_i (or ε_i) is the residual ($y_i - \hat{y}_i$).

b_0 (or β_0) is the intercept.

b_1 (or β_1) is the slope.

The least squares regression line is the line such that

$\text{SSE}=\sum_{i=1}^{n} e_i^2$ is a minimum (and $\sum_{i=1}^{n} e_i = 0$).

Assumptions

- There are 3 assumptions that must hold for a linear regression to be valid.

1. The relationship between the response and predictor must be linear.
2. The amount of variation in the response must be the same for all values of the predictor.
3. For any given value of the predictor, the response must have a bell-shaped distribution.

Standard Deviation of Regression

- The regression line has a standard deviation.

  To calculate this, you need the sum of squared residuals:

  $$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

  The standard deviation from the regression line is,

  $s = \sqrt{\dfrac{SSE}{n-2}}$ for the sample. This estimates the SD of the deviations about the population line.

Explaining the Variance in Y

- We want to use the explanatory variable (x) to explain the variability in the response (y).
- It's possible that the explanatory variable isn't good at this.
- The measure of how much of the variability in the response is explained by the variability in the explanatory variable is called R^2.
- This value is simply the correlation coefficient, r, squared.

Example: Birthweights

- Birthweight data were gathered from several Boston area hospitals.
- Birthweight is our response variable and infant length is our explanatory variable.
- We may want to know whether the relationship between birthweight and length is strong.
- We check the correlation before running the regression.

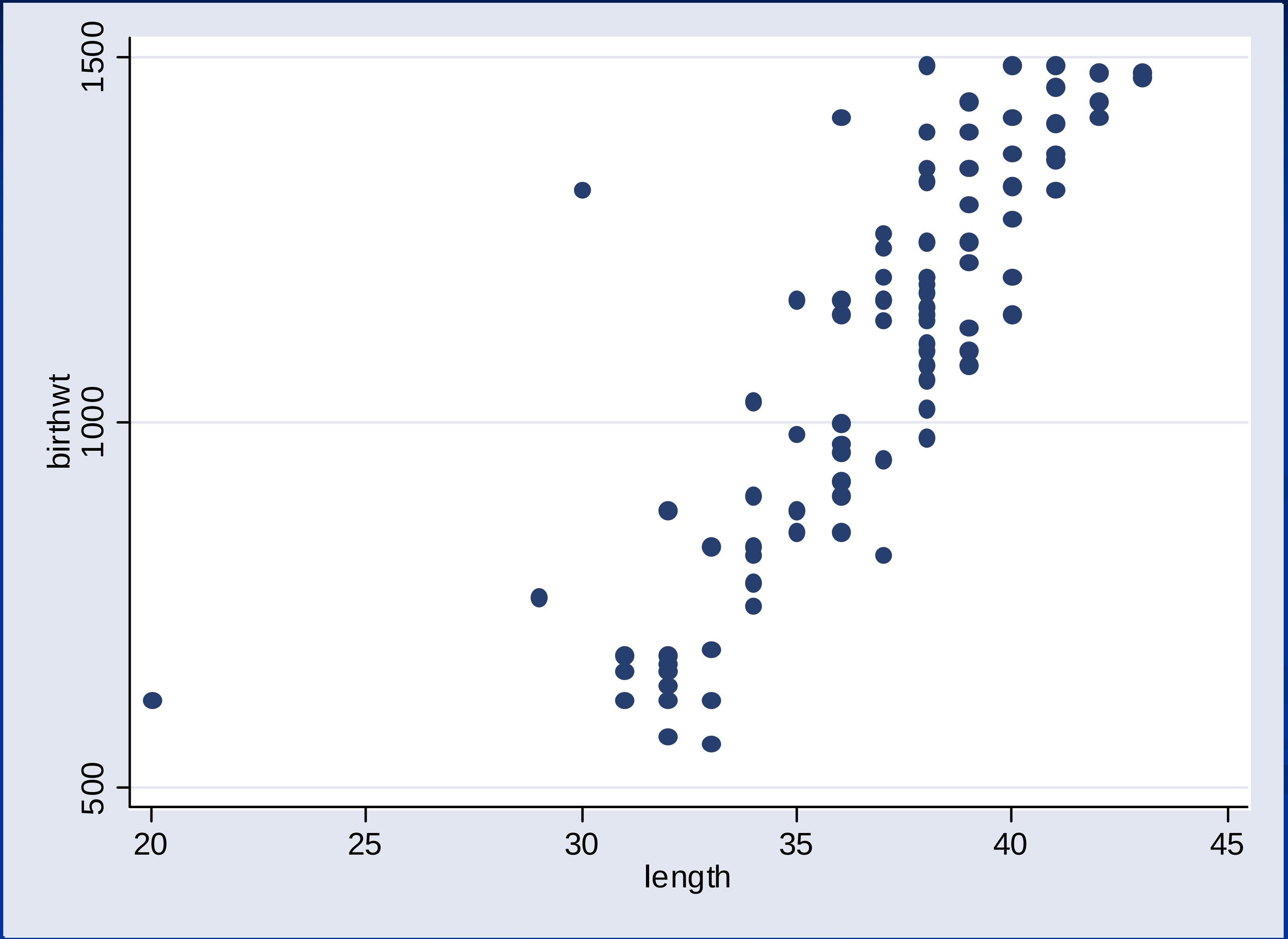
Correlations in Stata

- Click on **Statistics > Summaries, tables & tests > Pairwise correlations**
- Enter the variable you want correlations between in the "variables" box.
- Click on "Print significance level for each entry" box to get p-values for each correlation.
- These p-values tell you whether or not the correlation is significantly different from 0.

Example: Birthweights

```
. pwcorr birthwt length, sig

             |  birthwt   length
-------------+------------------
     birthwt |   1.0000
             |
             |
      length |   0.8156   1.0000
             |   0.0000
             |
```

Birthweight versus Length



Example: Birthweights

```
. regress birthwt length

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  194.65
       Model |  4800034.87     1  4800034.87           Prob > F      =  0.0000
    Residual |  2416707.88    98  24660.2845           R-squared     =  0.6651
-------------+------------------------------           Adj R-squared =  0.6617
       Total |  7216742.75    99  72896.3914           Root MSE      =  157.04

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   61.65408   4.419149    13.95   0.000     52.88442    70.42373
       _cons |  -1171.253   163.4691    -7.16   0.000    -1495.652    -846.854
------------------------------------------------------------------------------
```

Example: Birthweights

- So we can say that 66.5% of the variability in birthweight is explained by the variability in infant length.
- The correlation between birthweight and infant length is 0.8156.
- Note that even though the correlation is the square root of R^2, the correlation may be negative (although it isn't here).
- How do I know if my explanatory variable is telling me anything useful?

Testing the Slope

- What would happen if your explanatory variable told you nothing about the response?
- The association between the explanatory and response is not significant.
- If the explanatory variable is a important predictor of the response, the slope of the line will be nonzero.
- How could we tell whether it was nonzero?

Testing the Slope

We perform a formal hypothesis test of:

$\mathrm{H}_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$

The test statistic is given by,

$$t = \frac{\text{sample statistic-null value}}{\text{standard error}} = \frac{b_1 - 0}{s.e.(b_1)}$$

This test statistic has a t-distribution with n-2 df.

Stata (and all statistical software) does this test for you.

Example: Birthweights

```
. regress birthwt length

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  194.65
       Model |  4800034.87     1  4800034.87           Prob > F      =  0.0000
    Residual |  2416707.88    98  24660.2845           R-squared     =  0.6651
-------------+------------------------------           Adj R-squared =  0.6617
       Total |  7216742.75    99  72896.3914           Root MSE      =  157.04

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   61.65408   4.419149    13.95   0.000     52.88442    70.42373
       _cons |  -1171.253   163.4691    -7.16   0.000    -1495.652    -846.854
------------------------------------------------------------------------------
```

Example: Birthweights

- So the test statistic equals 13.95 ($p < 0.001$), which means that the population slope is significantly different from 0 (and is greater than 0).
- There is a positive association between gestation length and infant length.
- If this value were not significant, then the coefficient would not be significantly different from 0.
- We would do just as well using a horizontal line at the mean value of the response (birthweight) in predicting the response.

Example: Birthweights

- Note: The F-statistic in the upper right tests our model against an intercept-only model.
- In this case, the intercept-only model would occur if the slope for infant length were not significantly different from 0.
- The F-statistic, in the case where we have one predictor, is the square of the t-statistic.
- We'll be more concerned with the F-statistic when we consider more than one predictor.
- Let's consider prediction of birthweight from length.

Predicting the Response for an Individual

- We can predict the response for an individual "subject" with a given value of the explanatory variable.
- The "best guess" is the same as before – just use the regression line equation.
- So, for an infant length of 38 cm, the predicted value for the birthweight is 1172 grams.
- How do we quantify the uncertainty in this estimate?

$$se(\text{prediction}) = \sqrt{\frac{SSE}{n-2}}\sqrt{1+\frac{1}{n}+\frac{x_i-\overline{x}}{n}}$$

Generating PIs in Stata

- After running the regression, type "predict yhat" on the command line.
- Then type "predict sef, stdf" to get the SE.
- To get the t-multiplier use the "display invttail(df,p) command for a t-distribution with df=n-2, and p in the upper tail.
- Generate the upper and lower confidence bounds by:
  - "generate upi=yhat+t*sef"
  - "generate lpi=yhat-t*sef"

Generating PIs in Stata

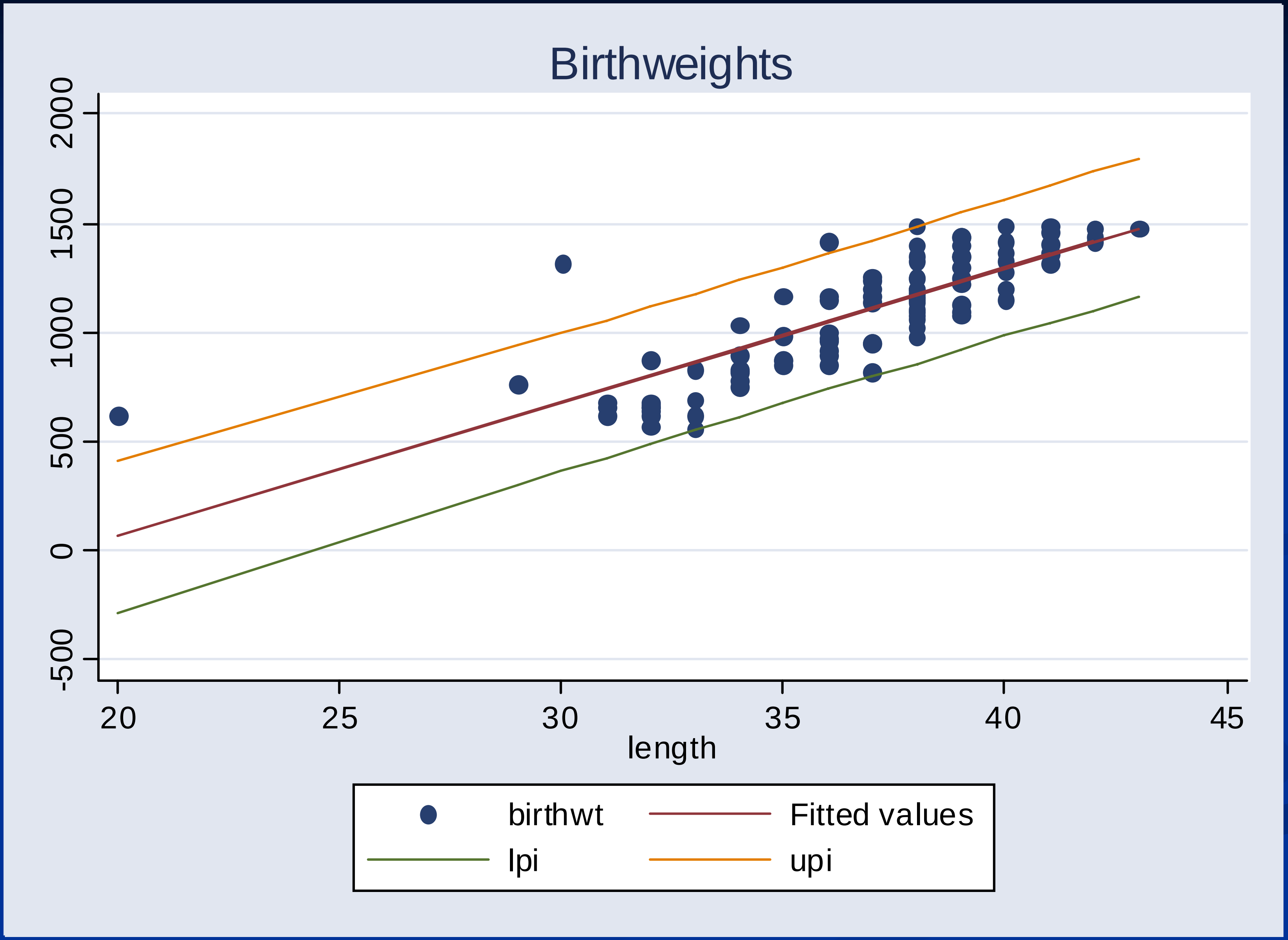
- The new variables *lpi* and *upi* represent the lower and upper prediction bounds.
- You can find the bounds by typing: "list lpi yhat upi if x==*x-value*" where *x-value* is the value of x for the individual subject.
- You can also plot the data, regression line, and prediction interval using the "twoway plots" menu.

Plotting PIs in Stata

- Click on **Graphics > Twoway graphs**.
- Click “Create” and let the plot default to “scatter” and enter the response in the Y-variable box and the explanatory variable in the X-variable box. Click “accept”.
- Click “Create” and select “line” for the graph type, enter the explanatory variable in the X-variable box, and *yhat* the Y-box. Click “accept”.
- Click “Create” select “line” for the graph type, enter the explanatory variable in the X-variable box, and *lpi* the Y-box. Click “accept”.
- Click “Create” and select “line” for the graph type, enter the explanatory variable in the X-variable box, and *upi* the Y-box.

Example: Birthweights

```
. quietly regress birthwt length

. predict yhat
(option xb assumed; fitted values)

. display invttail(98, .025)
1.9844675

. predict sef,stdf

. gen lpi=yhat-1.984*sef

. gen upi=yhat+1.984*sef

. list lpi yhat upi if length == 38

     +--------------------------------+
     |      lpi       yhat        upi |
     |--------------------------------|
  3. | 858.3177   1171.602   1484.886 |
```



Birthweights
2000
1500
1000
500
0
-500
20
25
30
35
40
45
length
birthwt
Fitted values
lpi
upi

Predicting the Average Response

- We can also predict the average response for *all* "subjects" with a given value of the explanatory variable.
- The "best guess" is the same as before – just use the regression line equation.
- So, for an infant 38 cm long, the predicted weight is 1172 grams.
- What about the confidence interval for this value (NOT the same as the prediction interval).

Finding the CI for the Average Response

- The procedure is identical to that for the prediction interval except for one part…
- The standard error is smaller than it is for the PI.

$$se\left[E(Y\,|\,X)\right] = \sqrt{\frac{SSE}{n-2}}\sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})}{n}}$$

- We don't use se(fit), but instead ask Stata to generate a different set of standard errors.
- The generation of the CI is the same general procedure as that for generating the PI though.

Generating CIs in Stata

- After running the regression, type "predict yhat" on the command line.
- Then type "predict sea, stdp" to get the SE.
- To get the t-multiplier use the "display invttail(df,p) command for a t-distribution with df=n-2, and p in the upper tail.
- Generate the upper and lower confidence bounds by:
  - "generate uci=yhat+t*sea"
  - "generate lci=yhat-t*sea"

Generating CIs in Stata

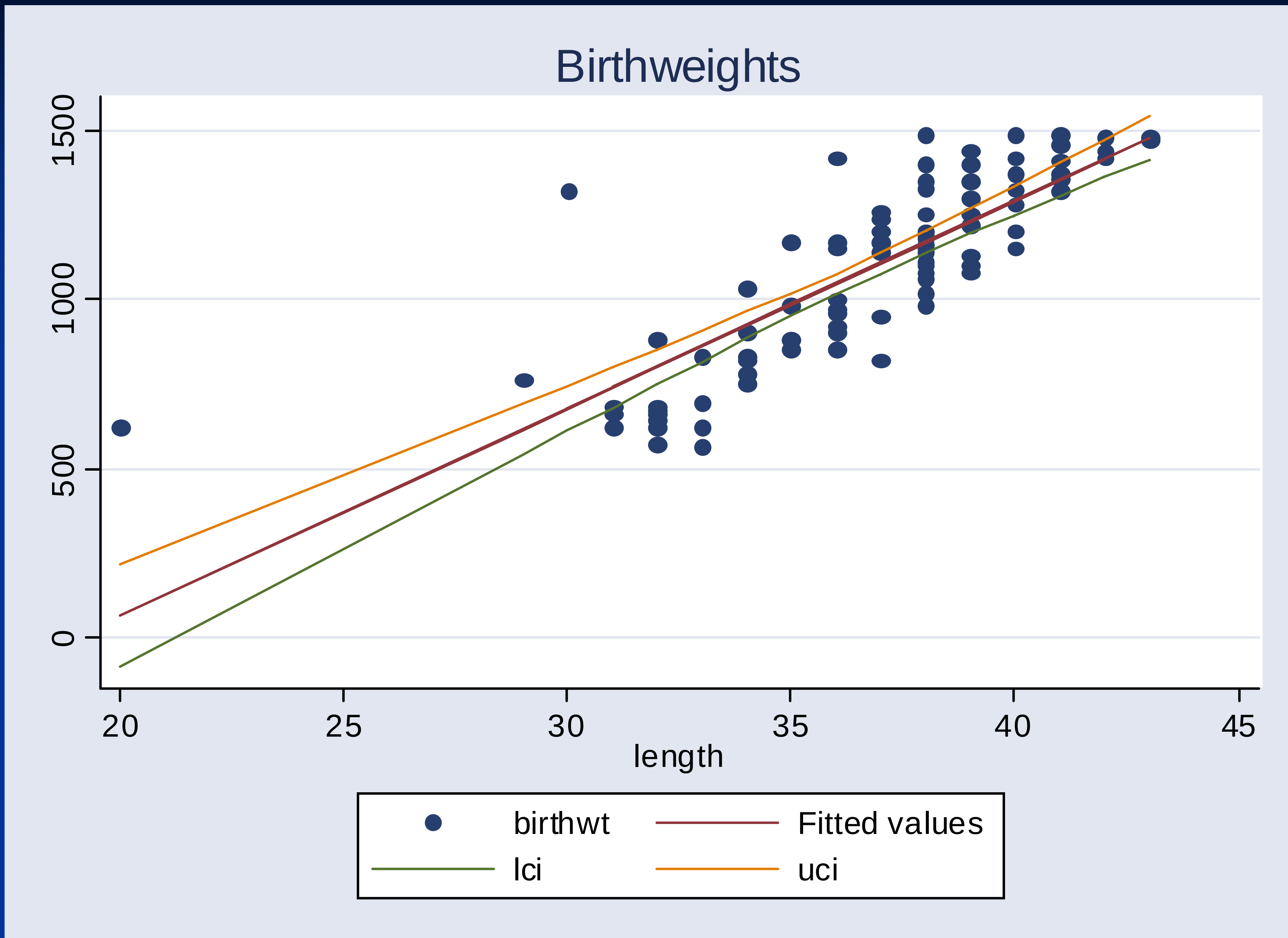
- The new variables *lci* and *uci* represent the lower and upper confidence bounds.
- You can find the bounds by typing: "list lci yhat uci if x==*x-value*" where *x-value* is the value of x for the individual subject.
- You can also plot the data, regression line, and prediction interval using the "twoway plots" menu.

Plotting CIs in Stata

- Click on **Graphics > Twoway graphs**.
- Click "Create" and let the plot default to "scatter" and enter the response in the Y-variable box and the explanatory variable in the X-variable box. Click "accept".
- Click "Create" and select "line" for the graph type, enter the explanatory variable in the X-variable box, and *yhat* the Y-box. Click "accept".
- Click "Create" select "line" for the graph type, enter the explanatory variable in the X-variable box, and *lci* the Y-box. Click "accept".
- Click "Create" and select "line" for the graph type, enter the explanatory variable in the X-variable box, and *uci* the Y-box.

Example: Birthweights

```
. quietly regress birthwt length

. display invttail(98, .025)
1.9844675

. predict sep,stdp

. gen lci=yhat-1.984*sep

. gen uci=yhat+1.984*sep

. list lci yhat uci if length == 38

     +---------------------------------+
     |      lci       yhat         uci |
     |---------------------------------|
  3. | 1138.773   1171.602    1204.431 |
```


Birthweights
0
500
1000
1500
20
25
30
35
40
45
length
birthwt
Fitted values
lci
uci