MA397B – Applied Regression Modeling
Simple Linear Regression Using Stata

Goals

The objective of this exercise is to demonstrate how to conduct a simple linear regression using Stata. The data are contained in the *birthwt.dta* file on the course webpage: http://www.colby.edu/personal/l/lobrien/ma397.html.

Description of Data

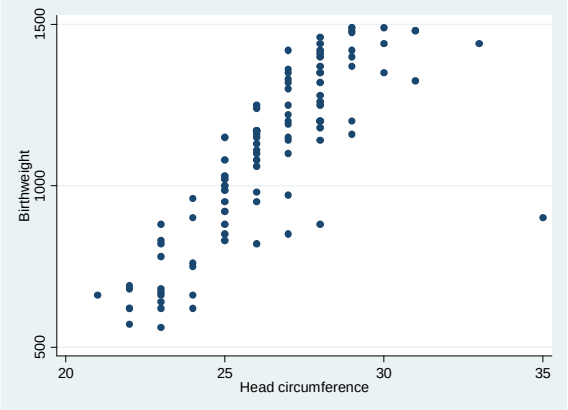
The data were obtained from birth records in Massachusetts and Rhode Island. There are 100 subjects and 6 variables including:

- Headcric: head circumference in cm
- Length: infant length in cm
- Gestage: gestational age in weeks
- Birthwt: birthweight in grams
- Momage: mother's age in years
- Toxemia: Indicator of toxemia presence

In this exercise, we will focus on the ability of head circumference to explain birthweight. Although note that any number of response and predictor variables could be selected.

Simple Linear Regression in Stata

Before doing anything, we should plot the data to see if we can determine the best model for E(Y). To generate a scatterplot, select **Graphics > Twoway graph**. Select "Create" It defaults to a simple scatter plot, so just enter the response (birthwt) in the Y-variable box, and the predictor (headcirc) in the X-variable box. You may add a title if you wish. Click "Accept" and "Ok".



A linear model seems to make sense. To run a simple linear regression in Stata, go to **Statistics > Linear models and related > Linear regression**. Although there are many options, we will begin with a basic regression model. Enter "birthwt" in the dependent variable box, and "headcirc" in the independent variable box. Click ok.

```
. regress birthwt headcirc

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  172.82
       Model |  4605298.87     1  4605298.87           Prob > F      =  0.0000
    Residual |  2611443.88    98  26647.3866           R-squared     =  0.6381
-------------+------------------------------           Adj R-squared =  0.6344
       Total |  7216742.75    99  72896.3914           Root MSE      =  163.24

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    headcirc |   85.17802   6.479268    13.15   0.000     72.32013    98.03592
       _cons |  -1154.109   172.1523    -6.70   0.000    -1495.739    -812.478
------------------------------------------------------------------------------
```

- There are 100 observations.
- The intercept = -1154 and represents the expected birthweight for an infant with a head circumference of 0 (which of course cannot happen).
- The slope = 85.2 and represents the expected increase in birthweight for a 1 cm increase in head circumference.
- The standard error of the intercept = 172.1 and represents the standard deviation of the estimate of β_0.
- The standard error of the slope = 172.1 and represents the standard deviation of the estimate of β_1.
- The test statistic for testing the hypothesis H_0: $\beta_0 = 0$ vs. H_A: $\beta_0 \neq 0$ is -6.70 and has a p-value < 0.001. Thus we reject H_0.
- The test statistic for testing the hypothesis H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 \neq 0$ is -6.70 and has a p-value < 0.001. Thus we reject H_0.
- The 95% confidence interval for the intercept is **(-1495.739, -812.478)**.
- The 95% confidence interval for the slope is **(72.32013, 98.03592)**.
- The sum of squared residuals (or errors) is SSE = 2611443.88. This value is as small as possible for any line that we place into the data.
- The conditional variance of birthweights given head circumference is the MSE = 26647.3866. There are 98 degrees of freedom associated with this error.
- The conditional standard deviation of birthweight given head circumference is the root MSE = 163.24.
- The R^2 = 0.6381, telling us that about 64% of the variance in birthweights is explained by the model.

Since the slope is significantly larger than 0, head circumference is a significantly important predictor of birthweight. However, over 35% of the variance in birthweights remains unexplained.

In order to see if the model assumptions are satisfied, recall the four assumptions about the distribution of ε.

1. The observations are independent.
2. The $E(\varepsilon) = 0$.
3. The distribution of ε is approximately normal.
4. The conditional variance of Y (birthweight) given X (head circumference) is constant for all values of X.

We will learn methods to check these assumptions soon, but for now we will assume that they are tenable.

Using the Model for Prediction and Estimation

We can estimate the expected birthweight for all infants with a head circumference of 21 cm using this model. We can also predict an individual infant's birthweight given their head circumference is 21 cm. In both cases, we will just use the LS estimates for the intercept and slope to obtain y-hat. We can also have Stata do this for us. On the command line, type "predict yhat". This generates a new variable called "yhat" that contains the point estimates of y for all values of x in our dataset. To see the predicted value for an infant with a head circumference of 21 cm, type "list yhat if headcirc==21".

Recall that the standard errors differ, however, depending on whether you are estimating E(Y) for all subjects with a certain condition, or whether you are predicting an individual Y for a subject with a certain condition.

Stata will generate these standard errors for you so that you can "manually" construct confidence and prediction intervals.

- To generate standard errors for the CI – the E(Y) for all subjects with a given value of X – type "predict sea, stdp" on the command line. This generates a variable called "sea" that contains the standard errors.
- To generate standard errors for the PI – the predicted value of Y for an individual with a given value of X – type "predict sef, stdf" on the command line. This generates a variable called "sef" that contains the standard errors.

We can now generate, and eventually plot, the lower and upper bounds for both the CI and the PI. Recall that each follows the simple formula:

$$estimate \pm (multiplier)(s.e.of\ estimate)$$

You already have the estimates and their standard errors, so now all you need is to find the multiplier. For simple linear regression, this will be a t-multiplier with n – 1 degrees of freedom. You can find this value in Stata by typing "display innvttail(df, alpha/2)" in the command line. For this example, that is "display invttail(98, .025)" is we are looking for a 95% interval. The obtained multiplier is 1.984.

Once we have the multiplier, we simply generate four new variables – two for the bounds on the CI and two for the bounds on the PI. To get these follow these steps:

- Type "generate lci = yhat – 1.984*sea". This gives the lower CI bound.
- Type "generate uci = yhat + 1.984*sea". This gives the upper CI bound.
- Type "generate lpi = yhat – 1.984*sef". This gives the lower PI bound.
- Type "generate upi = yhat + 1.984*sef". This gives the lower PI bound.

If you want to see the bounds for an infant with a head circumference of 21cm, type "list lci uci if headcirc == 21" to get the CI and "list lpi upi if headcirc == 21" to get the PI. The 95% CI is (557.447, 711.8125) and the 95% PI is (301.6912, 967.5684). As expected the PI is larger than the CI.

We can also plot these bounds to see the curvature in the intervals. We will essentially be plotting 6 graphs:

1. The scatterplot of birthweight versus head circumference
2. The fitted least squares regression line
3. The lower 95% CI
4. The upper 95% CI
5. The lower 95% PI
6. The upper 95% PI

To accomplish this in Stata, go to **Graphcs > Twoway graphs**.

1. Click on "Create" and it will default to a scatterplot. Place *birthwt* in the Y-variable box and *headcirc* in the X-variable box. Click on "Sort on X variable" and press "Accept".
2. Click on "Create" and change the graph type to "Line". Place *yhat* in the Y-variable box and *headcirc* in the X-variable box. Click on "Sort on X variable" and press "Accept".
3. Click on "Create" and change the graph type to "Line". Place *lci* in the Y-variable box and *headcirc* in the X-variable box. Click on "Sort on X variable" and press "Accept".
4. Click on "Create" and change the graph type to "Line". Place *uci* in the Y-variable box and *headcirc* in the X-variable box. Click on "Sort on X variable" and press "Accept".
5. Click on "Create" and change the graph type to "Line". Place *lpi* in the Y-variable box and *headcirc* in the X-variable box. Click on "Sort on X variable" and press "Accept".

6. Click on “Create” and change the graph type to “Line”. Place *upi* in the Y-variable box and *headcirc* in the X-variable box. Click on “Sort on X variable” and press “Accept”.
7. Click “Ok”.

You can play with a lot of graphics options in Stata both before and after the graph is produced. You can clearly see the PI is wider than the CI. In fact, it’s so wide that it is hard to see the curvature in the bounds, but it is there. Note that our infant with a head circumference of 21cm is at the far left side, and hence, has the widest PI and CI in our data set.