Confidence interval for a regression coefficient

A $(1-\alpha)100\%$ confidence interval for β_j has the old familiar form:

$$\widehat{\beta}_j \pm t_{n-m-1,\alpha/2}\,\widehat{SE},$$

where $\widehat{SE}$ is the estimated standard error of $\widehat{\beta}_j$. The value of $\widehat{SE}$ is given in SPSS in the "Coefficients" section of the linear regression output. It's on the same line as $\widehat{\beta}_j$ under "Std. Error" and "Unstandardized Coefficients."

For the L.A. Heart Study data, a 95% confidence interval for β_1, the regression coefficient for age, is

$$0.888 \pm 2.074(0.247) = 0.888 \pm 0.512,$$

or $(0.376, 1.400)$. So, we're 95% sure that the true regression coefficient for age is between 0.376 and 1.400.

Confidence interval for average response

Our estimate of $\mu_x = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$, the average response when the values of the independent variables are $x_1, x_2, \ldots, x_m$, is

$$\widehat{\mu}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_m x_m.$$

Of course, this estimate is subject to error, and so we'd like to have a confidence interval. A $(1-\alpha)100\%$ confidence interval for μ_x is

$$\widehat{\mu}_x \pm t_{n-m-1,\alpha/2}\, \widehat{SE}_x,$$

where $\widehat{SE}_x$ is the estimated standard error of $\widehat{\mu}_x$.

Prediction interval for future response

Suppose we want to predict a future response at values $x_1, \ldots, x_m$ of the independent variables. A $(1-\alpha)100\%$ prediction interval is

$$\widehat{\mu}_x \pm t_{n-m-1,\alpha/2} \sqrt{\widehat{\sigma}^2 + \widehat{SE_x}^2}.$$

This has exactly the same form as we encountered in the case of simple linear regression.

95% confidence intervals for μ_x and 95% prediction intervals are obtained in SPSS by clicking on "Save" (after you have defined the variables in your regression analysis) and then clicking on "Mean" and/or "Individual" prediction intervals.

For the heart study data, suppose we are interested in 40 year old men who weigh 160 lbs. and have cholesterol levels of 210. A 95% confidence interval for the average systolic blood pressure of such men is $(105.80, 128.93)$.

Now, suppose your cousin is a 40 year old man weighing 160 lbs. and having cholesterol level of 210. Then we are 95% sure that his systolic blood pressure is between 95.11 and 139.62. (Of course, this assumes that the 26 men in the study are a random sample from a population of which your cousin is a member.)

Polynomial Regression

The model with m independent variables is useful in *simple* regression as well as multiple regression. If the curve of averages in simple regression is not a straight line, we can model it by a *polynomial* of the form

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m.$$

We simply take $x_1 = x$, $x_2 = x^2, \ldots, x_m = x^m$ and use the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \epsilon,$$

as we have been doing.

Suppose we have decided to use a third degree polynomial, i.e., one with $m = 3$. In SPSS, we may fit this model as follows.

- Create the appropriate variables by selecting "Transform" and then clicking on "Compute Variables."

- Then obtain the third degree model exactly as we did in the multiple regression case by using the variables created in the previous step.

Example *Rabbit jawbone data*

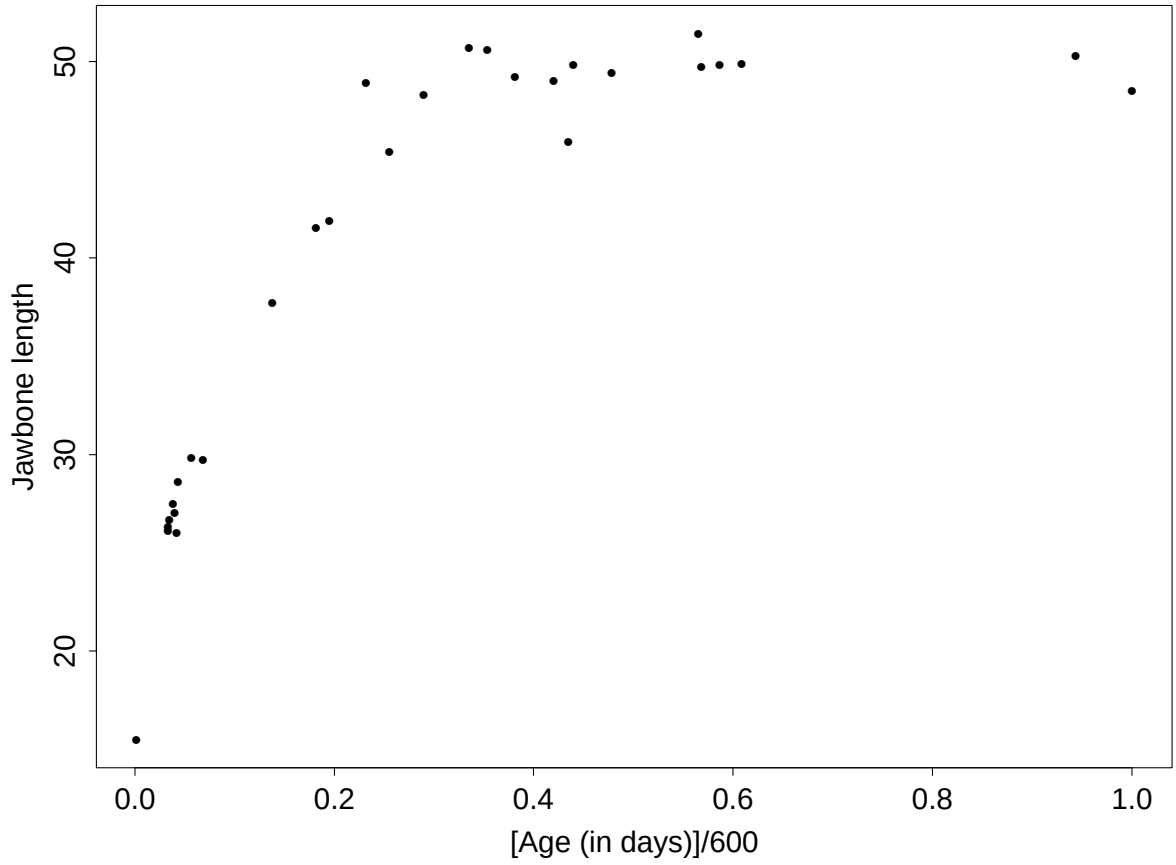
Independent variable: age of rabbit
Response: rabbit's jawbone length

A rabbit's age is given in days. The 30 rabbits in the data set range in age from 1 to 600 days. We'll define

$$x = \frac{\text{age}}{600}.$$

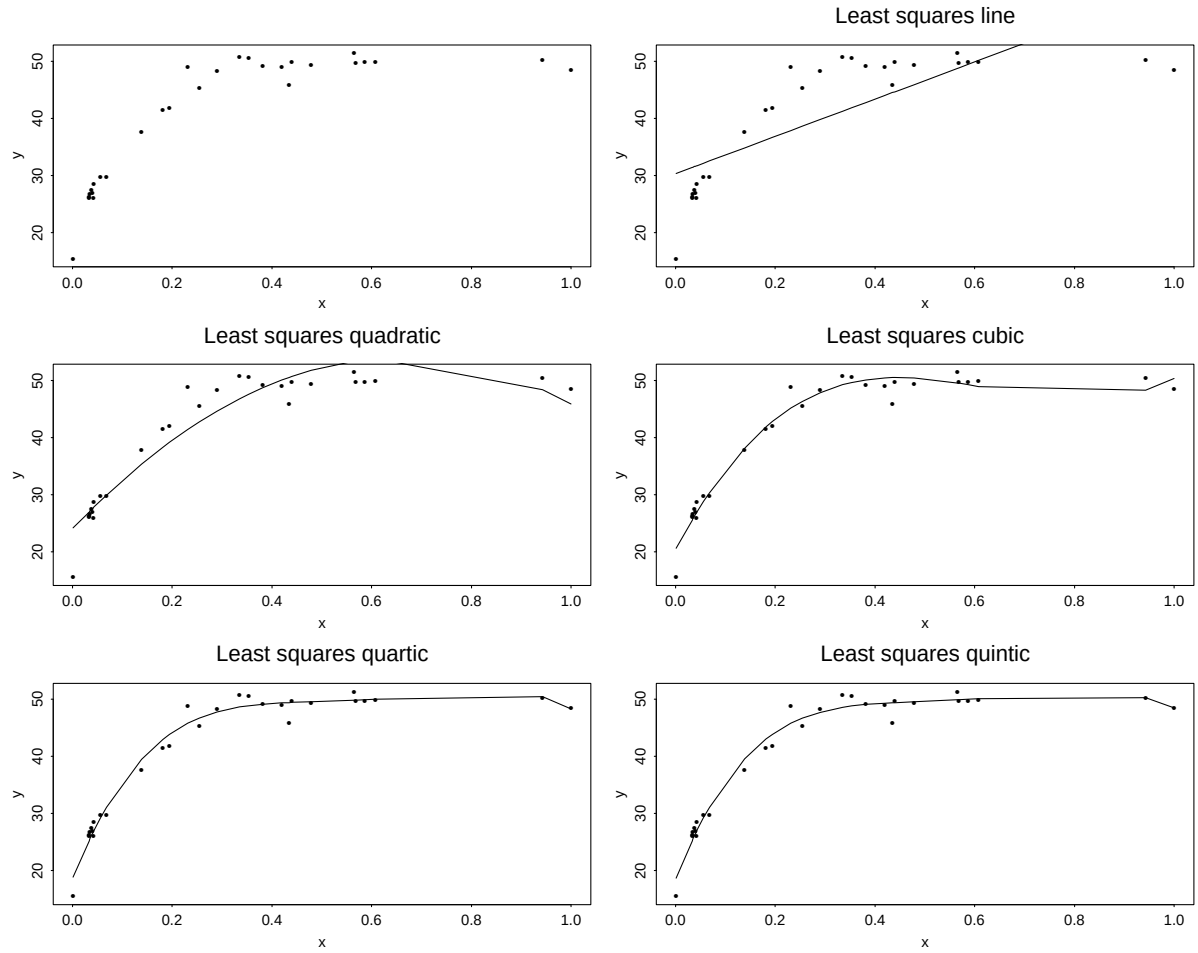
This makes the process of finding least squares estimates a bit more numerically stable.

Rabbit Jawbone Data



Note the rapid increase in jawbone length for young rabbits and the leveling off when they reach maturity.

Various fitted curves for rabbit data



In a visual sense, the 4th or 5th degree polynomial seems best. From a parsimony standpoint, the 4th is preferable to 5th.