Regression

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

December 8–15, 2011

Example

Case Study

- The proportion of blackness in a male lion's nose increases as the lion ages.
- This proportion can be used to predict the age of a lion with unknown age.
- To find a predictive equation, researchers determined the proportion of blackness in 32 male lions of known age.
- The data is displayed in a scatter plot, and a good-fitting line is found for the data.

$$(\text{age in years}) = a + b \times (\text{proportion of blackness in the nose})$$

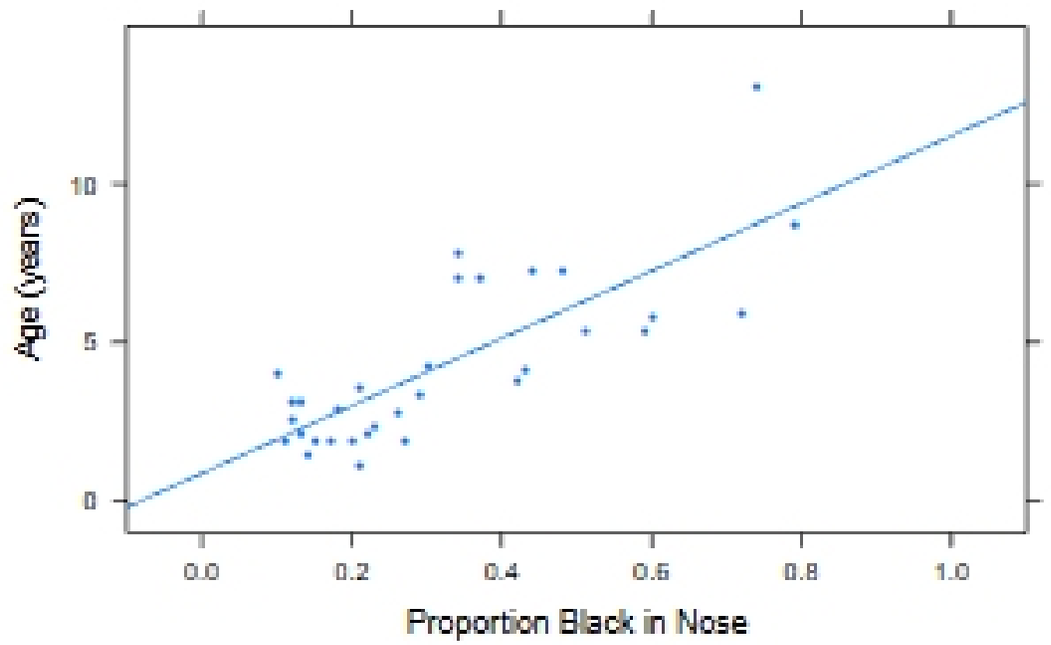
- The line may be interpreted as a *conditional expectation*: the expected age of a lion given a proportion of blackness in its nose.

The Data

- age is the age of a male lion in years;
- proportion.black is the proportion of a lion's nose that is black.

age	proportion.black
1.1	0.21
1.5	0.14
1.9	0.11
2.2	0.13
2.6	0.12
3.2	0.13
3.2	0.12
...	

Lion Data Scatter Plot



Observations

- We see that *age and blackness in the nose are positively associated*.
- The points do not fall exactly on a line.
- How do we find a good-fitting line?
- How do we decide if a line is a sufficient summary of the relationship between the variables?

Simple Linear Regression

Definition

Simple linear regression is the statistical procedure for describing the relationship between an quantitative explanatory variable X and a quantitative response variable Y with a straight line;

$$Y = a + bX$$

- The value a is the Y-intercept of the estimated line.
- It is the location where the line crosses the Y-axis, and may be interpreted as an estimate of $\mathrm{E}(Y \mid X = 0)$, the expected value of Y given X is zero, which may or may not be meaningful in the context.
- The slope b is the estimated change in Y per unit change in X.

A Model

- The *simple linear regression model* for the data is

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where

  - i varies from 1 to n, the sample size;
  - α and β are fixed population parameters;
  - ε_i is the random vertical deviation between the line and the ith observed data point; and
  - the deviations are assumed to be independent and normally distributed with standard deviation σ.

$$\varepsilon_i \sim \text{ i.i.d } \mathrm{N}(0, \sigma^2)$$

- Notice that in this model, there is a *common variance* for all observations.
- This means that we should expect the size of a typical deviation from the line to be the same size at all locations.

Fitted Values and Residuals

- The estimated regression line takes the form

$$Y = a + bX$$

  where a is an estimate of α and b is an estimate of β.
- The height of the point on the line at $X = X_i$ is called the ith *fitted value* or *predicted value*.

$$\hat{Y}_i = a + bX_i$$

- The difference between the ith data point Y_i and the ith predicted value is called the ith *residual*, $Y_i - \hat{Y}_i$.

Estimation

- The parameters of the model may be estimated either by the criteria of *least squares*, which *minimizes the sum of squared residuals*.
- The parameters may also be estimated by *maximum likelihood*, which makes the probability density of the observed data as large as possible.
- In simple linear regression, the least squares and maximum likelihood estimates of α and β are identical.
- The maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\text{(sum of squared residuals)}}{n}$$

is slightly different than the conventional unbiased estimate.

$$s^2 = \frac{\text{(sum of squared residuals)}}{n-2}$$

- Note that there are 2 parameters used to describe all means (α and β) as two points determine a line, and so there are $n-2$ remaining pieces of information remaining to estimate variation around the line.

Formulas for Estimation

- It is an exercise in calculus (or inspired algebra) to show that

$$b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

is the least squares (and maximum likelihood) estimate of the slope.

- There is a simple formula for the estimated intercept given the estimated slope.

$$a = \bar{Y} - b\bar{X}$$

- The *residual sum of squares* is

$$\text{RSS} = \sum_{i=1}^{n}\left(Y_i - (a + bX_i)\right)^2$$

is used to estimate σ^2 by dividing by either $n-2$ (for an unbiased estimate) or n (for the maximum likelihood estimate).

An alternative formulation

- The estimated parameters may also be described in an alternative manner based on the means and standard deviations of X and Y and the correlation between them.
- The formulas are based on this idea:

> *When X is $z = \frac{X-\bar{X}}{s_X}$ standard deviations above the mean, ($z < 0$ when X is less than $\bar{X}$), the predicted Y is rz standard deviations above its mean, or*

$$\hat{Y} = \bar{Y} + r\left(\frac{X-\bar{X}}{s_X}\right)s_Y = \left(\bar{Y} - \left(r\frac{s_Y}{s_X}\right)\bar{X}\right) + \left(r\frac{s_Y}{s_X}\right)X$$

- The slope is $b = rs_Y/s_X$.
- When $X = \bar{X}$, then $z = 0$ and $\hat{Y} = \bar{Y}$, so the regression line goes through the point $(\bar{X}, \bar{Y})$.
- When $X = \bar{X} + s_X$ is one standard deviation above the mean, the predicted value is $\hat{Y} = \bar{Y} + rs_Y$ r standard deviations above the mean.

Regression Toward the Mean

- Heights of fathers and sons in human populations often have a correlation close to $r = 0.5$.
- If one uses the height of the father to predict the height of the son, the average heights of all sons of fathers whose height is one standard deviation above the mean is only about one half of a standard deviation above the mean.
- Similarly, the heights of sons of fathers that are one standard deviation below the mean are expected to be only half a standard deviation below the mean.
- This general phenomenon is called *regression toward the mean*.