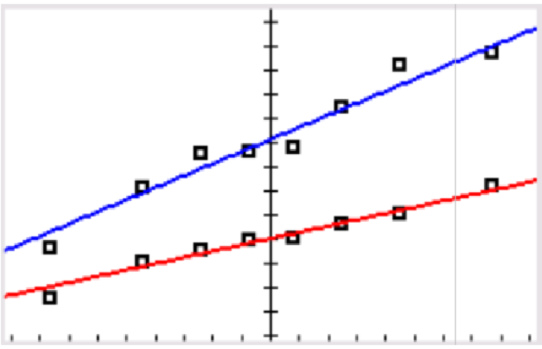
Creating Pseudo-Flies

Suppose we have two variables x and y that are linearly related. Suppose further, that either by evaluating the normal probability plots or by assumption, the observed values can be considered a random sample from some normally distributed population. We can create a set of values that mimic the population from which x and y were selected.

In this example, we assume x is approximately normally distributed with a mean of 9.063 and a standard deviation of 1.309. We also assume that y is approximately normally distributed with a mean of 13.212 and a standard deviation of 2.48.

The regression line for the observed data is $y = -2.403 + 1.723x$. The mean of the residuals is always 0 and the standard deviation is 1.033.



Normal Probability Plots

$x_i :=$	$y_i :=$
6.6	8.7
8.1	12.6
8.6	11.2
9	12.9
9.1	14.5
9.7	12.7
10.1	16.3
11.3	16.8

Observed Data

mean(x) = 9.063

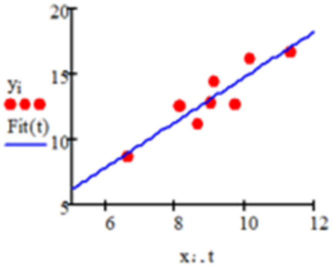
stdev(x) = 1.309

mean(y) = 13.212

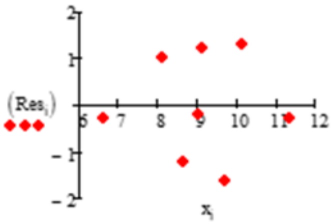
stdev(y) = 2.48

Fit(x) := −2.403 + 1.723·x

Stats



Regression Model



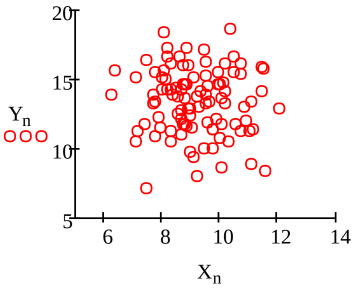
stdev(Res) = 1.033

Residual Plot

We want to create 100 flies that match these characteristics. We would like to believe that our 8 observations could be considered a random sample from this population of pseudo-flies. They will represent the possible future captured flies and we can see how well our boundary works with them (how often we misclassify the flies and which species is most likely to be misclassified.

First, we need to recognize that we can' just create of population with x chosen from the normal distribution with mean 9.063 and standard deviation 1.309, denoted $N(9.063, 1.309)$, and y chosen from $N(13.212, 2.48)$.

This would create a circle of data which has examples with large values of x and small values of y as well as small values of x and large values of y which don't exist in our population. These created data are uncorrelated, while our observations are correlated .

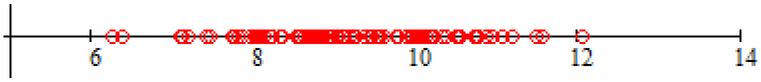


X := rnorm (100, 9.063, 1.309)

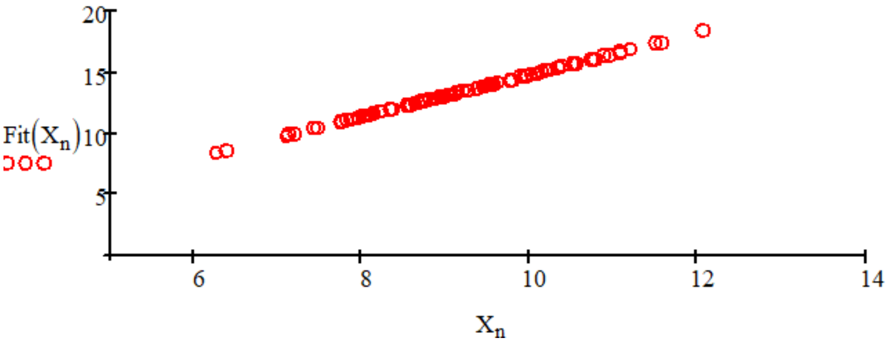
Y := rnorm (100, 13.212, 2.48)

One way to create pseudo-flies using the distribution of the independent variable and the distribution of the residuals from our regression model is shown below:

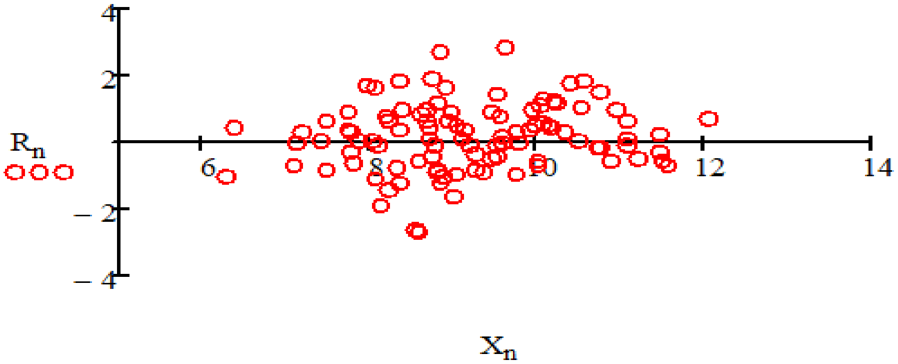
1. Create a random collection of x-values using $N(9.063, 1.309))$ and place them on the horizontal axis.



2. Next, lift these points onto the regression line $y = -2.403 + 1.723x$.



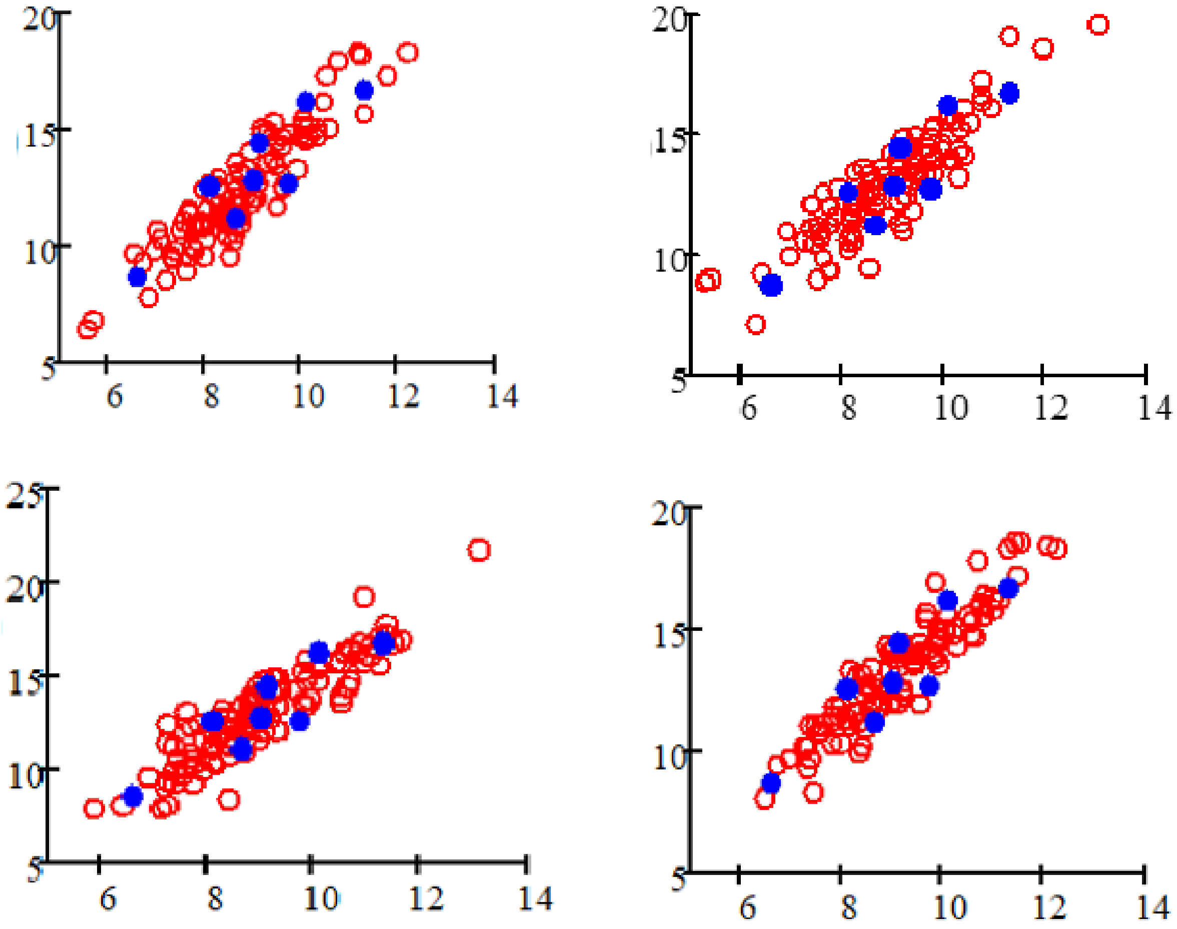
3. Create a random collection of *Residuals* using $N(0, 1.033)$ since the residuals are always assumed to be approximately normally distributed.



4. Add the residuals to the values on the Fit in step 2. So, the pseudo-flies are created by

$$P=\overbrace{\left(\underbrace{-2.403+1.723\cdot\left(\underbrace{Random\ x}_{Step\ 1}\right)}_{Step\ 2}\right)+\left(\underbrace{Random\ Residual}_{Step\ 3}\right)}^{Step\ 4}.$$

Four examples are shown below with the observations shown in blue.



This process prioritizes the independent variable. Students might want to compare results if we fit *y* onto *x*. In the Powerpoint slides, I show an example of combining 50 pseudo-flies from *x*-on-*y* and 50 from *y*-on-*x* (then reflected to match the chosen orientation).