Predictions by Regression

The regression analysis has three goals: predicting, modeling, and characterization. What would be the logical order in which to tackle these three goals such that one task leads to and /or and justifies the other tasks? Clearly, it depends on what the prime objective is. Sometimes you wish to model in order to get better prediction. Then the order is obvious. Sometimes, you just want to understand and explain what is going on. Then modeling is again the key, though out-of-sample predicting may be used to test any model. Often modeling and predicting proceed in an iterative way and there is no 'logical order' in the broadest sense. You may model to get predictions, which enable better control, but iteration is again likely to be present and there are sometimes special approaches to control problems. The following contains the main essential steps during modeling and analysis of regression model building, presented in the context of an applied numerical example.

**Formulas and Notations:**

* xbaru= x /n This is just the mean of the x values.
* ybar= y /n This is just the mean of the y values.
* Sxx = SSxx = (x(i) - xbaru)2 = x2 - (x)2 / n
* Syy = SSyy = (y(i) - ybar)2 = y2 - (y) 2 / n
* Sxy = SSxy = (x(i) - xbaru)(y(i) - ybar) = x y – (x)  (y) / n
* Slope m = SSxy / SSxx
* Intercept, b = ybar- m . xbaru
* y-predicted = yhat(i) = mx(i) + b.
* Residual(i) = Error(i) = y – yhat(i).
* SSE = Sres = SSres = SSerrors = [y(i) – yhat(i)]2.
* Standard deviation of residuals = s = Sres = Serrors = [SSres / (n-2)]1/2.
* Standard error of the slope (m) = Sres / SSxx1/2.
* Standard error of the intercept (b) = Sres[(SSxx + n. xbaru2) /(n  SSxx] /2.

**Page1/6**

**An Application:** A taxicab company manager believes that the monthly repair costs (Y) of cabs are related to age (X) of the cabs. Five cabs are selected randomly and from their records we obtained the following data: (x, y) = {(2, 2), (3, 5), (4, 7), (5, 10), (6, 11)}. Based on our practical knowledge and the scattered diagram of the data, we hypothesize a linear relationship between predictor X, and the cost Y.

Now the question is how we can best (i.e., least square) use the sample information to estimate the unknown slope (m) and the intercept (b)? The first step in finding the least square line is to construct a sum of squares table to find the sums of x values (x), y values (y), the squares of the x values (x2), the squares of the x values (y2), and the cross-product of the corresponding x and y values (xy), as shown in the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **x** | **y** | **x**2 | **xy** | **y**2 |
| 2 | 2 | 4 | 4 | 4 |
| 3 | 5 | 9 | 15 | 25 |
| 4 | 7 | 16 | 28 | 49 |
| 5 | 10 | 25 | 50 | 100 |
| 6 | 11 | 36 | 66 | 121 |
| **SUM** | **20** | **35** | **90** | **163** | **299** |

The second step is to substitute the values of x, y, x2, xy, and y2 into the following formulas:

SSxy = xy – (x)(y)/n = 163 - (20)(35)/5 = 163 - 140 = 23

SSxx = x2 – (x)2/n = 90 - (20)2/5 = 90- 80 = 10

SSyy = y2 – (y)2/n = 299 - 245 = 54

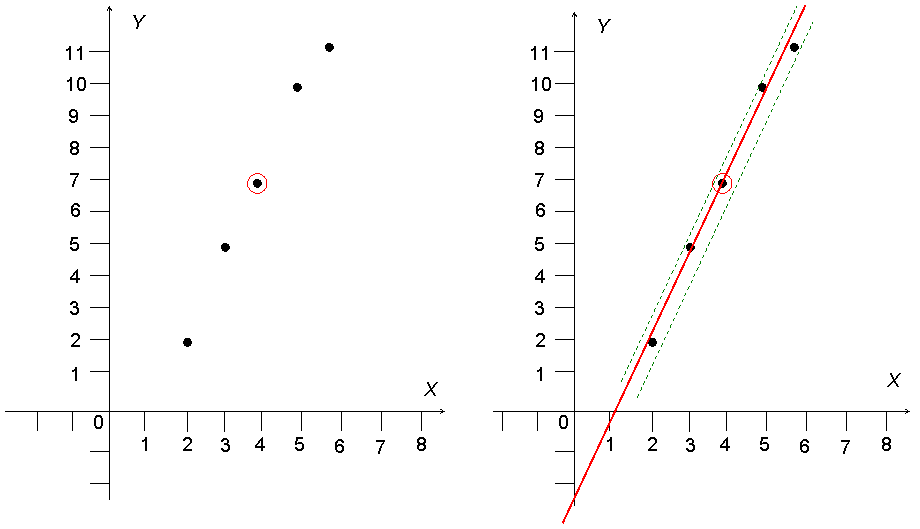
Use the first two values to compute the estimated slope:

Slope = m = SSxy / SSxx = 23 / 10 = 2.3

To estimate the intercept of the least square line, use the fact that the graph of the least square line always pass through (xbaru, ybar) point, therefore,

The intercept = b = ybar– (m)(xbaru) = (y)/ 5 – (2.3) (x/5) = 35/5 – (2.3)(20/5) = -2.2

**Page 2/6**



Is There Any Linear Relationship? How to Test It from the Scatter Diagram?

Therefore the least square line is:

y-predicted = yhat = mx + b = -2.2 + 2.3x.

After estimating the slope and the intercept the question is how we determine statistically if the model is good enough, say for prediction. The standard error of slope is:

Standard error of the slope (m)= Sm = Sres / Sxx1/2,

and its relative precision is measured by statistic

tslope = m / Sm.

For our numerical example, it is:

t slope = 2.3 / [(0.6055)/ (101/2)] = 12.01,

which is large enough, indication that the fitted model is a “good” one.

You may ask, in what sense is the least squares line the “best-fitting” straight line to 5 data points. The least squares criterion chooses the line that minimizes the sum of square vertical deviations, i.e., residual = error = y - yhat:

SSE =  (y – yhat)2 = (error)2 = 1.1

The numerical value of SSE is obtained from the following computational table for our numerical example.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **x Predictor** | **-2.2+2.3x y-predicted** | **y observed** | **error y** | **squared errors** |
| 2 | 2.4 | 2 | -0.4 | 0.16 |
| 3 | 4.7 | 5 | 0.3 | 0.09 |
| 4 | 7 | 7 | 0 | 0 |
| 5 | 9.3 | 10 | 0.7 | 0.49 |
| 6 | 11.6 | 11 | -0.6 | 0.36 |
|  | | | | **Sum=0** | **Sum=1.1** |

Alternately, one may compute SSE by:

SSE = SSyy – m SSxy = 54 – (2.3)(23) = 54 - 52.9 = 1.1,

**Page 3/6**

as expected

Notice that this value of SSE agrees with the value directly computed from the above table. The numerical value of SSE gives the estimate of variation of the errors s2:

s2 = SSE / (n -2) = 1.1 / (5 - 2) = 0.36667

The estimate the value of the error variance is a measure of variability of the y values about the estimated line. Clearly, we could also compute the estimated standard deviation s of the residuals by taking the square roots of the variance s2.

As the last step in the model building, the following Analysis of Variance (ANOVA) table is then constructed to assess the overall goodness-of-fit using the F-statistics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance Components** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Prob > F** |
| Model | 1 | 52.90000 | 52.90000 | 144.273 | 0.0012 |
| Error | 3 | SSE = 1.1 | 0.36667 |  |  |
| Total | 4 | SSyy = 54 |  |  |  |

For practical proposes, the fit is considered acceptable if the F-statistic is more than five-times the F-value from the F distribution tables at the back of your textbook. Note that, the criterion that the F-statistic must be more than five-times the F-value from the F distribution tables is independent of the sample size.

Notice also that there is a relationship between the two statistics that assess the quality of the fitted line, namely the T-statistics of the slope and the F-statistics in the ANOVA table. The relationship is:

t2slope = F

This relationship can be verified for our computational example.

**Page 4/6**

**Predictions by Regression:** After we have statistically checked the goodness of-fit of the model and [the residuals conditions](http://home.ubalt.edu/ntsbarsh/Business-stat/opre504.htm#rregconditions) are satisfied, we are ready to use the model for prediction with confidence. Confidence interval provides a useful way of assessing the quality of prediction. In prediction by regression often one or more of the following constructions are of interest:

1. A confidence interval for a single future value of Y corresponding to a chosen value of X.
2. A confidence interval for a single pint on the line.
3. A confidence region for the line as a whole.

**Confidence Interval Estimate for a Future Value:** A confidence interval of interest can be used to evaluate the accuracy of a single (future) value of y corresponding to a chosen value of X (say, X0). This JavaScript provides confidence interval for an estimated value Y corresponding to X0 with a desirable confidence level 1 - .

Yp  Se . tn-2, /2 {1/n + (X0 – xbaru)2/ Sx}1/2

**Confidence Interval Estimate for a Single Point on the Line:** If a particular value of the predictor variable (say, X0) is of special importance, a confidence interval on the value of the criterion variable (i.e. average Y at X0) corresponding to X0 may be of interest. This JavaScript provides confidence interval on the estimated value of Y corresponding to X0 with a desirable confidence level 1 - .

Yp  Se . tn-2, /2 { 1 + 1/n + (X0 – xbaru)2/ Sx}1/2

It is of interest to compare the above two different kinds of confidence interval. The first kind has larger confidence interval that reflects the less accuracy resulting from the estimation of a single future value of y rather than the mean value computed for the second kind confidence interval. The second kind of confidence interval can also be used to identify any outliers in the data.

**Confidence Region the Regression Line as the Whole:** When the entire line is of interest, a confidence region permits one to simultaneously make confidence statements about estimates of Y for a number of values of the predictor variable X. In order that region adequately covers the range of interest of the predictor variable X; usually, data size must be more than 10 pairs of observations.

Yp  Se { (2 F2, n-2, ) . [1/n + (X0 – xbaru)2/ Sx]}1/2

In all cases the [JavaScript](http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Regression.htm) provides the results for the nominal (x) values. For other values of X one may use computational methods directly, graphical method, or using linear interpolations to obtain approximated results. These approximation are in the safe directions i.e., they are slightly wider that the exact values.

**Page 5/6**

**Formulas and Notations:**

* xbaru= x /n

This is just the mean of the x values.

* ybar= y /n

This is just the mean of the y values.

* Sxx = SSxx = (x(i) - xbaru)2 = x2 - (x)2 / n
* Syy = SSyy = (y(i) - ybar)2 = y2 - (y) 2 / n
* Sxy = SSxy = (x(i) - xbaru)(y(i) - ybar) = x y – (x)  (y) / n
* Slope m = SSxy / SSxx
* Intercept, b = ybar- m . xbaru
* y-predicted = yhat(i) = mx(i) + b.
* Residual(i) = Error(i) = y – yhat(i).
* SSE = Sres = SSres = SSerrors = [y(i) – yhat(i)]2.
* Standard deviation of residuals = s = Sres = Serrors = [SSres / (n-2)]1/2.
* Standard error of the slope (m) = Sres / SSxx1/2.
* Standard error of the intercept (b) = Sres[(SSxx + n. xbaru2) /(n  SSxx] /2.

**Page 6/6**