Research & Reviews: Journal of Statistics and Mathematical Sciences

The Usefulness of Implicit Regression in Model Quality

Rebecca D Wooten*
Department of Mathematics and Statistics, University of South Florida, USA

Letter

Received date: 07/07/2016
Accepted date: 20/07/2016
Published date: 27/07/2016

***For Correspondence**

Rebecca WD, Department of Mathematics and Statistics, University of South Florida, USA

E-mail: rwooten@usf.edu

Keywords: Implicit regression, Bivariate, Non-linear regression, Coefficient of variation, Model quality.

IMPLICIT REGRESSION

Implicit Regression is useful in measuring the constant nature of a measured variable [1]; it helps detect bivariate and multivariate random error; it is sensitive to incorrect modeling; and can handle co-dependent relationships more readily than standard non-linear regression [2,3].

Implicit Regression [4,5] using ordinary least squares to determine parameter estimates for models of the form

$$g\left(x_1, x_2, ..., x_p\right) = h\left(x_1, x_2, ..., x_p \mid \theta\right)$$

Where $g\left(x_1, x_2, \ldots, x_p\right)$ is a fixed function with well-defined constant coefficients and $h\left(x_1, x_2, \ldots, x_p\right)$ is defined in terms of the unknown coefficients $\theta = \left\{\alpha_1, \alpha_2, \ldots, \alpha_m\right\}$.

The only restriction is uniqueness of the terms; that is, given the terms of interest $\left\{T_i\left(x_1, x_2, \ldots, x_p\right) \mid i = 1, 2, \ldots, N\right\}$ where N is the number of terms to be considered, then terms contained in the function $g\left(x_1, x_2, \ldots, x_p\right)$ may not be contained in the function $h\left(x_1, x_2, \ldots, x_p\right)$. Common terms include unity (a column vector of 1), individual measures, higher order terms and interaction terms; that is $\left\{1, x_1, \ldots, x_p, x_1^2, x_2^2, \ldots x_p^2, x_1 x_2, \ldots\right\}$.

There are two types of resulting analysis: Non-response Analysis and Rotational Analysis. Non-response analysis is where $g\left(x_1, x_2, \ldots, x_p\right) = 1$; whereas Rotational Analysis allows for each term other than unity to be taken as the response variable in turn.

The first use is the same in both standard regression and the non-response model defined under Implicit Regression which is the explanatory power otherwise referred to as the coefficient of determination, R^2.

In univariate, the explanatory power is a function of the sample size, the sum of the data and the sum of squares and is given by

$$R^2 = \frac{n\bar{x}^2}{\sum x^2} = h\left(n, \sum x, \sum x^2\right).$$

This is the percent variance explained by the mean. This is a measure of the constant nature of a "variable". Variable here is placed in quotations because as $R^2 \to 1$, the smaller the variance and the measure is less variant and may be consider a constant, relatively speaking.

The measure the explanatory power is the same using standard regression and non-response analysis with analytical forms. $x_i = \beta + \varepsilon_i$ and $1 = \alpha x_i + \omega_i$ respectively. The first of which minimize the variance and the solution is

$$\hat{\beta} = \frac{\sum x}{n} = f\left(n, \sum x\right)$$

The second met $\hat{\alpha} = \frac{\sum x^2}{\sum x} = h\left(\sum x, \sum x^2\right)$ hod minimizes the coefficient of variation with solution

Notice that the explanatory power is the ratio of these two point estimates:

$$R^2 = \frac{\hat{\beta}}{\hat{\alpha}}.$$

In bivariate, there are three rotations in simple linear relationships:

$y = \beta_0 + \beta_1 x$

$x = \gamma_0 + \gamma_1 y$

$1 = \alpha_1 x + \alpha_2 y$

These models are such that the parameter estimates are functions of the summary statistics: the sample size, the sums, the sums of the products and the sum of the squares:

$$\hat{\beta} = f\left(n, \sum x, \sum xy, \sum y, \sum x^2\right),$$
$$\hat{\gamma} = g\left(n, \sum x, \sum xy, \sum y, \sum y^2\right),$$
$$\hat{\alpha} = h\left(\sum x, \sum xy, \sum y, \sum x^2, \sum y^2\right)$$

where the solutions to the non-response model does not depend on the sample size, but rather the base variance (Wooten R. D., 2016). Like turning the end of a kaleidoscope, these rotational views give more insight to the relationship that exist between the measures including the possibility of a co-dependent relationship that does not depend on the sample size.

Consider the following simulation: let $x \sim U(10,100)$ and y = x^a, and the observed value (x_i, y_i) contain random error $\delta_i \sim N\left(0, \sigma^2\right)$ and $\varepsilon_i \sim N\left(0, \sigma^2\right)$, respectively. That is, bivariate error:

$x_i = x + \delta_i$ and $y_i = y + \varepsilon_i$,

Consider the following four cases:

$x_i = x$ and $y_i = y + \varepsilon_i$ where $y = \beta_0 + \beta_1 x$

$x_i = x + \delta_i$ and $y_i = y + \varepsilon_i$ where $y = \beta_0 + \beta_1 x$

$x_i = x$ and $y_i = y + \varepsilon_i$ where $y = \beta_0 + \beta_1 x^{1.25}$

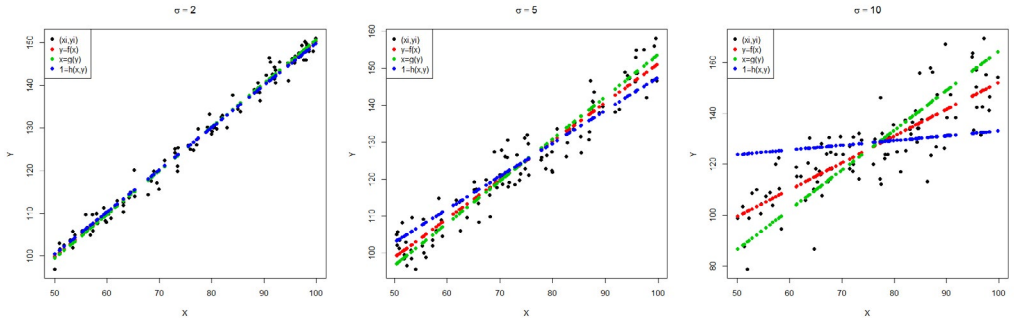
$x_i = x + \delta_i$ and $y_i = y + \varepsilon_i$ where $y = \beta_0 + \beta_1 x^{1.25}$

Consider the three rotations with σ = 2,5,10 in the four outlined cases, illustrated on the next page. Standard regression is shown in red and green; and implicit regression is shown in green. The larger the variance, the larger the pin wheel effect; however, when the relationship is non-linear, the effect is more pronounced. That is, comparing the graphics when there is small variance in a simple linear relationship, then the three rotations converge. However, when there is large variance or a non-linear relationship, the graphics show a pin-wheel effect.

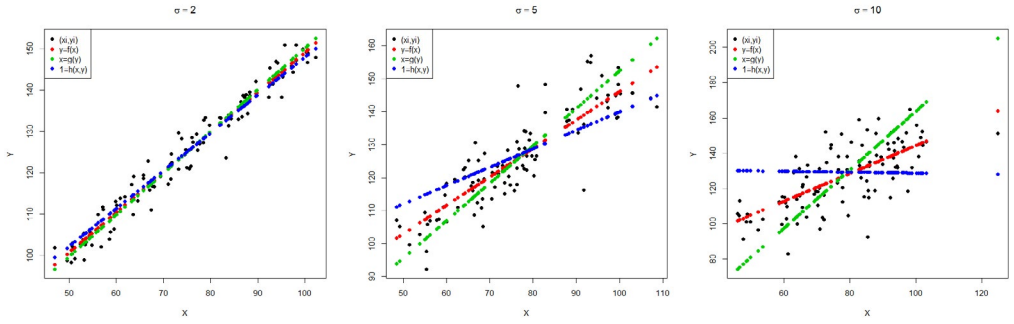
Implicit Regression enables the researcher to detect bivariate random error and model non-linear co-dependent relationships in multivariate analysis **(Figure 1).**

First consider the bivariate example of a circle with unit deviation and sinusoidal frequency; that is, the equation $(x-100)^2+(y-100)^2=6.25$, $x_i = x + \delta_i$, $y_i = y + \varepsilon$, and $\varepsilon, \delta \sim N(0,1)$ that both accelerate and decelarate based on time. The data was simulated using the following algorithm over 10 periods:

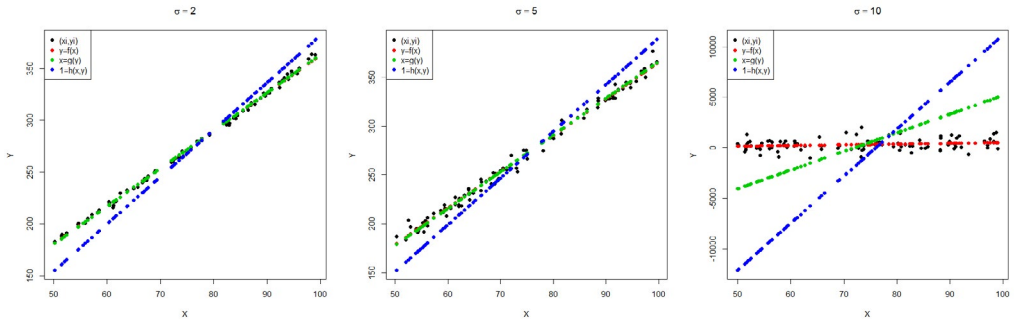
```
t <- seq(from=0,to=10*pi,by=10*pi/999)        #time frame t ∈ U(0,10π)
e <- rnorm(1000,0,1)                          #residual error in y
d <- rnorm(1000,0,1)                          #residual error in x
u <- rep(1,1000)                              #unity – a column of ones
x <- 100+2.5*cos(10*cos(t))                   #movement in x over time
y <- 100+2.5*sin(10*cos(t))                   #movement in y over time
```



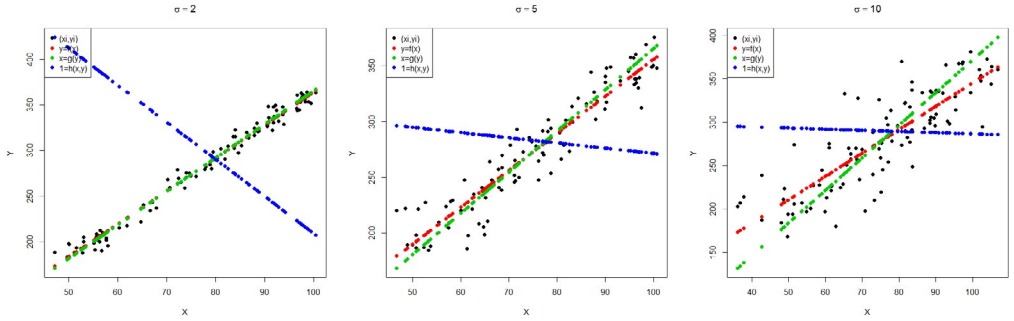
(a) Scatter plots of the developed linear model in case 1 assuming variance in the *y* direction only



(b) Scatter plot of the developed linear model in case 2 assuming variance in both directions



(c) Scatter plot of the developed model to the non-linear data in case 3 assuming variance in the *y* direction only



(d) Scatter plot of the developed model to the non-linear data in case 4 assuming variance in both directions

Figure 1. Fitting a linear model to linear and non-linear relationships. The four cases are illustrated by increasing deviations

xi <- x+d #observed value of x

yi <- y+e #observed value of y

The resulting data creates the random pattern shown in **Figure 2.**

$$\alpha_1 x^2 + \alpha_2 y^2 + \alpha_3 x + \alpha_4 y + \alpha_5 xy = 1$$

Where,

$$\alpha_1 x_i^2 + \alpha_2 y_i^2 + \alpha_3 x_i + \alpha_4 y_i + \alpha_5 x_i y_i = 1 + \omega_i$$

The resulting solution shown in green, **(Figure 3)**, is representative of the underlying relationship shown in red and the

observed data in blue. Implicit regression also detected that the term *xy* was insignificant.

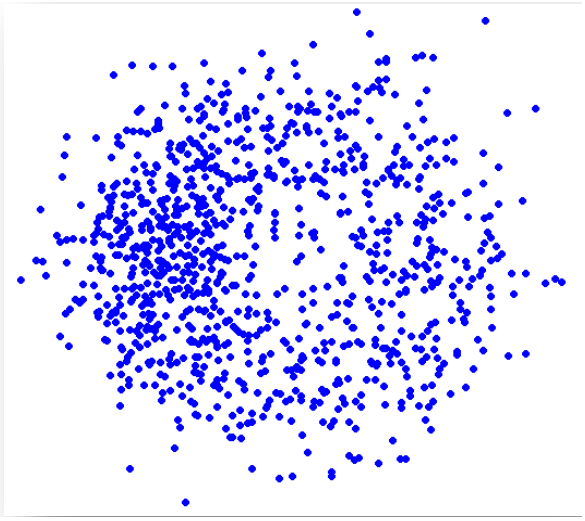


Figure 2. Scatter plot of simulated data as matched pairs (x_i,y_i)

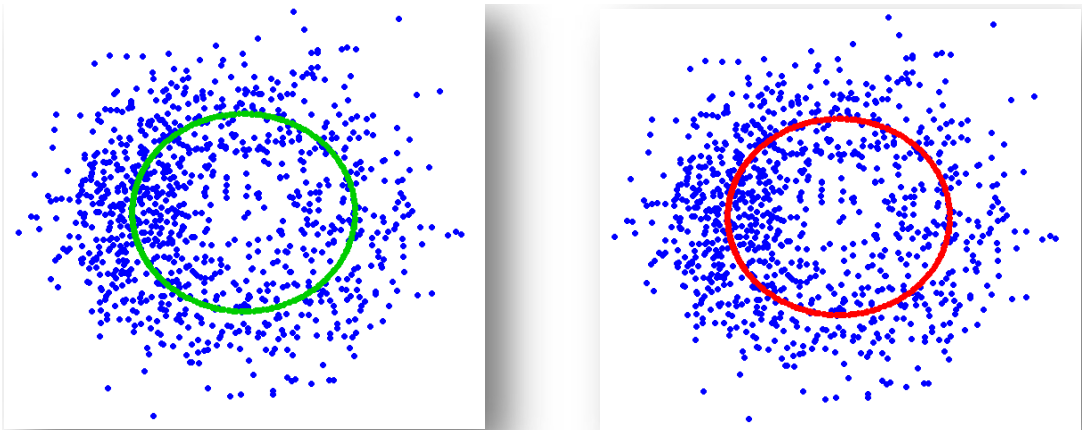


Figure 3. Scatter plot of simulated data (x_i,y_i) in blue, the solutions to the developed model $\{(x,y)|\hat{\alpha}_1x^2+\hat{\alpha}_2y^2+\hat{\alpha}_3x+\hat{\alpha}_4y+\hat{\alpha}_5xy=1\}$ in green, and the underlying variables. (x,y)

As the assumption of independence is not required, Implicit Regression does not have the same measure of explanatory power and is subject to the tractability of the individual variables. There are three measures of error in modeling between the following measures: the data $(x_{1i},x_{2i},\ldots,x_{pi})$, the central tendencies $(\overline{x}_1,\overline{x}_2\ldots,\overline{x}_p)$; and the resulting surface obtained by the developed model $(\hat{x}_{1i},\hat{x}_{2i},\ldots,\hat{x}_{pi})$, The total errors are $T_{ij}=x_{ji}-\overline{x}_j$ and the errors explained by the model are $M_{ij}=\hat{x}_{ji}-\overline{x}_j$ and the residual errors are $E_{ij}=x_{ji}-\hat{x}_{ji}$.

In Standard Regression, under the assumption of independence and limitations placed on terms, the Pythagorean Theorem holds for the measured response:

$$SST=SSM+SSE.$$

That is, when given $(x_{1i},x_{2i},\ldots,x_{pi},y)$, where the model is of the form $y=f(x_{1i},x_{2i},\ldots,x_{pi})$, and the error terms are only taken for this single variable, does this relationship hold true.

Using Implicit Regression, we can consider all variables individually or in combinations using the extended formula for law of cosines which is given by

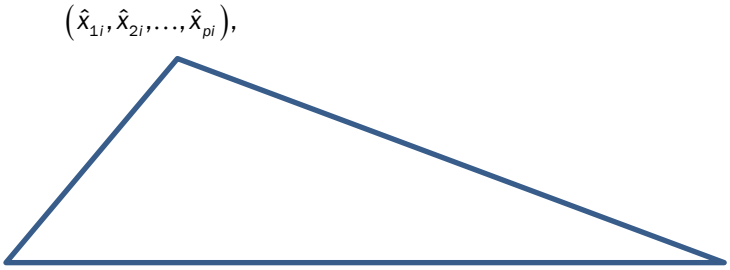
$$SST=SSM+SSE-2\sqrt{SSM\times SSE}\cos\theta$$

The angle is the degree of separation between M and E in the vector space created by *M*, *E* and *T*. The degree of separation can be measured one variable at a time with $T_{ij}=x_{ji}-\overline{x}_j$, $M_{ij}=\hat{x}_{ji}-\overline{x}_j$ and $E_{ij}=x_{ji}-\hat{x}_{ji}$ for the selected variable x_j with

$$SST=\sum_{i=1}^{n}\left(x_{ji}-\overline{x}_j\right)^2;SSM=\sum_{i=1}^{n}\left(\hat{x}_{ji}-\overline{x}_j\right)^2;\text{and } SSE=\sum_{i=1}^{n}\left(x_{ji}-\hat{x}_{ji}\right)^2.$$

Moreover, you can assess the degree of separation between subsets of variables or all the variables as illustrated in the

vector space below.



$$SST = \sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ji} - \overline{x}_j\right)^2; SSM = \sum_{j=1}^{p}\sum_{i=1}^{n}\left(\hat{x}_{ji} - \overline{x}_j\right)^2; \text{and } SSE = \sum_{j=1}^{n}\sum_{i=1}^{n}\left(x_{ji} - \hat{x}_{ji}\right)^2.$$

The closer the degree of separation is to 90°, the stronger the independent relationship between the three measured errors.

REFERENCES

1. Wooten RD, et al. Implicit regression: Detecting constants and inverse relationships with bivariate random error. Cornell University Library, 2015.
2. Wooten RD. Statistical analysis of the relationship between wind speed, pressure and temperature. Journal of Applied Sciences. 2011;11:2712-2722.
3. Wooten RD and D'Andrea J. Modeling hurricanes using principle component analysis in conjunction with non-response analysis. Cornell University Library. 2016.
4. Wooten RD. An introduction to implicit regression: Extending standard regression to rotational analysis and non-response analysis. Cornell University Library. 2016.
5. Wooten RD. Lattice designs in standard and simple implicit multi-linear regression. Cornell University Library. 2016.

This article was published under the special issue, **Mathematics Trends and Multivariate Statistical Analysis** handled by Editor. Dr. Huseyin BOR, Ankara University, Turkey