CURVE-FITTING MODELS FOR IMMUNE MEDIATED CELL DISTRUCTION - COMPARISON

Bogdan Timar [1,2], Corina Vernic [1], Simona Apostol [3], Viorel Şerban [1]

[1]"Victor Babes" University of Medicine and Pharmacy Timisoara
[2] Emergency County Clinical Hospital Timisoara
[3] "Tibiscus" University of Timisoara

ABSTRACT:
Background and aims: Variables in biology and medicine have a series of particularities when fitted to a curve, especially when their value is conditioned by time. Many curve fitting comparisons are based on the value of R squared, indicator which is not accurate in describing not-nested non-linear models.
Material and method: We fitted a model of remaining quantity of viable cells affected by an immune-mediated destruction process in relation with disease duration on different types of curves, and we assessed the evidence ratio for each one with the proper indicator, the Akaike's Information Criterion, then we investigated the relationship between this and value of R squared.
Results: Linear curve fitting was not appropriate for our model. The values of R squared were disconcordant in relation to the proper indicator, the Aikake's Information Criterion and Sum of Squares.
Conclusion: Best fitting curve model for our example is the logarithmic one. In comparing not-nested, non-linear models in medicine Aikake's Information Criterion should always be used in detriment of R squared.
KEYWORDS: curve fitting models, Akaike's Information Criterion, R squared, Diabetes Mellitus

1. BACKGROUND AND AIMS

Due to their particularities interdependent biological in general and medical in special variables are often mistaken correlated using a classical regression model, more frequently using linear regression. This wrong approach leads to significant biases which are compromising the results of the research, otherwise in many cases valid and valuable. Knowing these facts, we wanted to compare different curve fitting-models on some hands-on examples. We compared different regression models applied on the correlation between the intensity of immune mediated cell destruction in relation to disease duration. For this model we had chosen the relationship between C-peptide and Type 1 Diabetes Mellitus duration, considering this model as a representative one for this instance.
Diabetes Mellitus is a metabolic disorder of multiple etiologies, characterized by chronic hyperglycemia together with disturbances of carbohydrate, fat and protein metabolism resulting from defects of insulin secretion, insulin action or both [WHO99]. Type 1 Diabetes Mellitus is primarily caused by β-pancreatic cell destruction, in more than 90% of the cases, this process being immune-mediated [MYE07]. This destruction is progressive after the debut of the disease, having a series of particularities (it is broad in the early phases of the disease, decreasing in intensity with time and is always leading to complete or near complete destruction), similar to the rest of immune-mediated cell destructions in humans [KPB93]. The destruction extensity is reverse correlated with residual insulin secretion, which can be measured indirectly, by measuring the serum levels of C-peptide, a component of proinsulin which is not found in exogenous insulin (delivered as a treatment for Type 1 Diabetes, being mandatory for survival), released in the blood stream after the conversion of endogenous proinsulin to active insulin [H94]. Considering this we can agree to further use the C-peptide levels as an accurate indicator of residual, viable, β-pancreatic cells.

2. MATERIAL AND METHOD

2.1 Enrollment and blood tests

We enrolled 367 individuals, children and young adults aged between 1 and 35 years, all diagnosed with Type 1 Diabetes Mellitus, and with a disease duration no longer than 10 years. These individuals were admitted in "Cristian Serban" Medical Center, Buzias, the only hospital in Romania focused on Type 1 diagnosis, treatment and patient education. The disease start date was considered the day of the first insulin injection (hormone which exogenous administration is mandatory for survival in this patients), and was determined from the medical history of the patient. All blood tests were assessed in the second day of admission. To estimate the remaining viable β-pancreatic cell mass we assessed the fasting C-peptide serum levels, using standardized chemiluminescent enzyme immunoassay method.

2.2 Curve fitting and comparison

After assessing and pairing values, both of the previous instances were fitted with linear and non-

linear models (exponential, logarithmic and power type), using Graph Pad Prism 5 software suite and Microsoft Excel 2010. We didn't research the polynomial model because it is known that few biological or chemical models are described by a polynomial equation. Even in most of cases polynomial models are fitting the curve best, the coefficients of the equation almost never can be interpreted in terms of biology or chemistry [MC04]. For non-linear regressions the values were curvilinear fitted using smallest squares method. We computed the Pearson correlation coefficient, line and curve fitting equation and the sum of squares. To assess the significance of regression we used the t-distribution method which was interpreted according the studied lot size (N); in this case, p values lower than 0.05 were considered to be significant and so we accepted the regression model. Regressions with p value higher than 0.05 were not considered significant. Since our compared models were not nested, for quality of the models comparison we used the Akaike's Information Criterion (AICc) method. This method doesn't return the significance value of the regression, it doesn't state a null hypothesis, which was the reason why we computed the p-value using t-distribution individually for each model, but lets us to determine which model is most likely to be correct and quantifies how much more likely. This method has the advantage that it can be used to compare either nested or nonnested models. It combines maximum likehood theory, information theory and the concept of the entropy of information [BA02]. If we accept the usual assumptions of nonlinear regression, the AIC is defined by the corrected AICc equation below, where N is the number of data points, K is the number of parameters fit by the regression (in case of nonlinear regression the number of fitted parameters plus one, because regression is estimating the sum-of-squares as well as the values of parameters), and SS is the sum of square of the vertical distances of the points from the curve.

$$AICc = N * \ln\left(\frac{SS}{N}\right) + 2K + \frac{2K(K+1)}{N-K-1}$$

Since in our studied examples we have more different y values corresponding to one x value, we considered each replicated value as a separated one and we incremented the N value accordingly [GI09].
The model with the lower AICc score is the model more likely to be correct. If the AICc scores are very close, there isn't much evidence to choose one model over another. If the AICc scores are far apart, then the evidence is overwhelming. The probability that we have chosen the correct model is computed by the following equation, where Δ is the difference between AICc scores

$$probability = \frac{e^{-0.5\Delta}}{1 + e^{-0.5\Delta}}$$

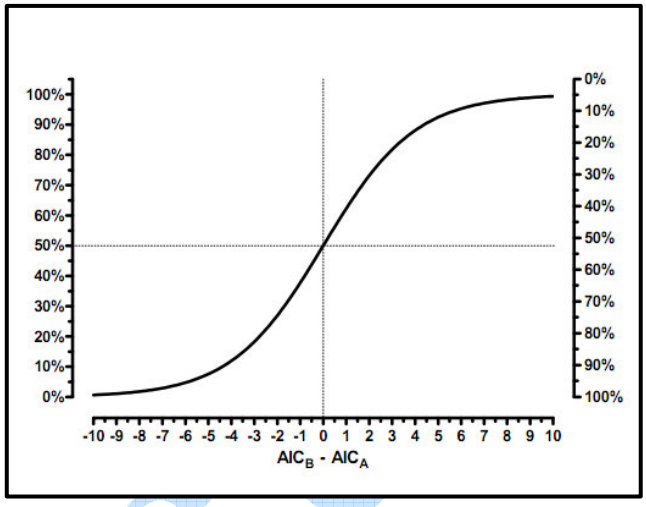

Figure 1. The relationship between the difference in AICc scores and the probability that each model is true (in left model A and right model B)

To assess the probability of correct model for curve fitting only absolute difference between AICc scores matters, not the relative difference. When comparing two models we can divide the probabilities that one model is correct by the probability the other model is correct to obtain the evidence ratio, defined by the following equation:

$$\textbf{Evidence Ratio} = \frac{\textbf{Probability that model 1 is correct}}{\textbf{Probability that model 2 is correct}} = \frac{1}{e^{-0.5\Delta AICc}}$$

The evidence ratio doesn't tell us anything about the significance of any regression, but only compares the relative likehood of the two models being correct.
Given these a complete and correct analysis for comparing regression models should always contain both significance level analysis (using t-distribution analysis) and evidence ratio compared to the other regression studied using Akaike's Information Criterion for nonnested models or the extra sum of squares F-test for nested models [SN10].

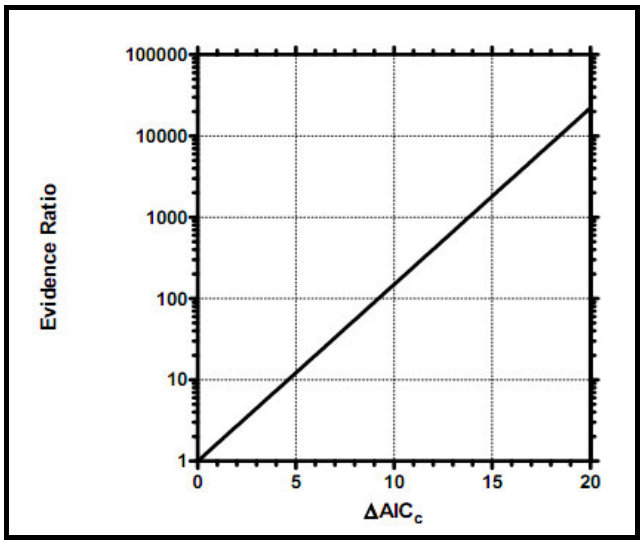

Figure 2. Evidence ratio for the lowest AICc score regression in relation to ΔAICc

3. RESULTS

We first researched the regression between disease duration and remaining β-cells according that moment, estimated by the serum value of fasting C-peptide, fitting the values on a linear regression and nonlinear ones: exponential, logarithmic and power-type. Both values, diabetes duration and serum C peptide proved to be normal distributed, D'Agostino & Pearson omnibus normality test revealed a K2 of 211.5 and p=0.08 for the first group and a K2 of 41.45 and p=0.096 for the values of C peptide.
When fitted to a linear regression line, we found a weak but significant correlation coefficient between the values of C peptide and disease duration, as shown in Graph 3. In this case, Pearson r coefficient was -0.39, r^2=0.15, absolute sum of squares 143.1, t=-7.752 (335 df), with a significant p for correlation (p<0.001).

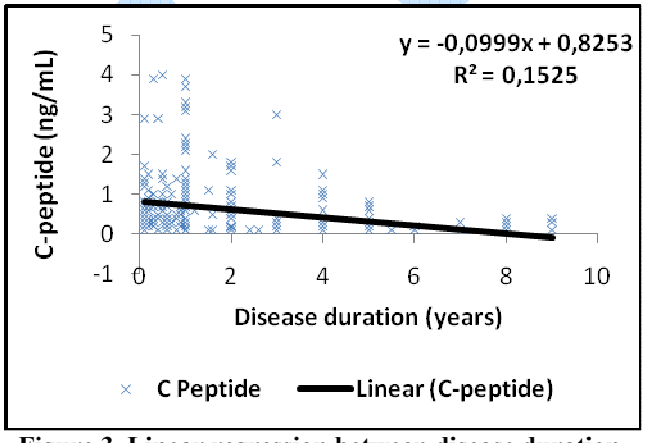

Figure 3. Linear regression between disease duration and C-peptide levels

The main disadvantage of regression is that always the line crosses somewhere the x-axis and becomes negative at x levels where always measurements are positive or zero, this model having no correspondence in biology (since values cannot be negative), the residual plot for this regression (Graph 4) emphasizeing this statement, since after 7 years of disease out of 10 studied, in our example, we have only positive residual values from model line.

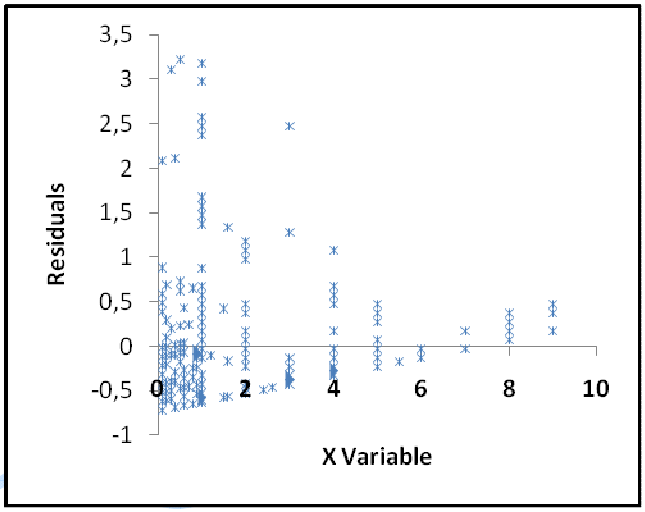

Figure 4. X variable residual plot for linear regression

Next regression model tested was the logarithmic one. Here we also found a significant correlation as shown in Graph 5, with a Pearson r of -0.35, r^2=0.13, absolute sum of squares 147.5, t=-7.81 (335 df), with a significant p for correlation (p<0.001).

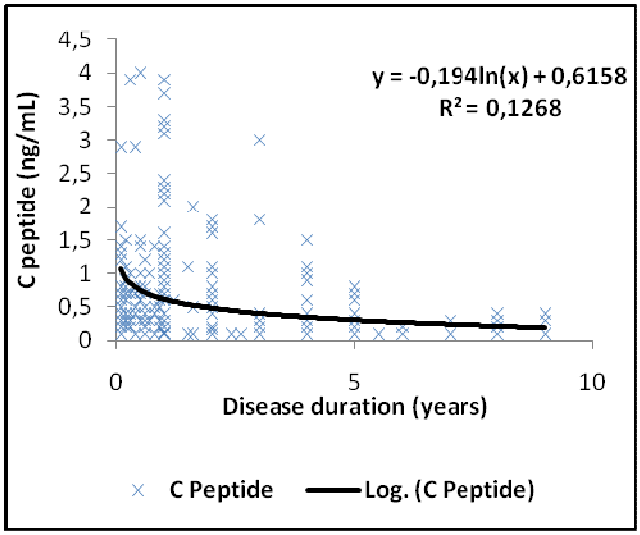

Figure 5. Logarithmic model regression between disease duration and C-peptide levels

Table 1 - Comparison of regression parameters

Regression model	r	r^2	Sum of squares	t-distribution	dF	p for correlation	AICc
Linear	-0.39	0.15	143.1	-7.75	335	<0.001	n/a
Logarithmic (x-log, y-normal)	-0.35	0.13	147.5	-7.81	335	<0.001	-270.55
Exponential	-0.56	0.31	174.7	-12.37	335	<0.001	-215.72
Power-type	-0.55	0.30	176.8	-12.42	335	<0.001	-211.71

For the exponential regression model (Graph 6) we found a significant correlation between the values of disease duration measured in years and serum C-peptide levels ($p<0.001$), together with a Person r of -0.56, $r^2=0.31$, absolute sum of squares 274.7, $t=-12.371$.
The power-type regression (Graph 7) had a similar correlation between the values of disease duration and the values o f C-peptide compared to the exponential model. The correlation was significant: ($p<0.001$), $r=0.55$, $r^2=0.304$ and absolute sum of squares 276.78.
Even if all regression models shown significant correlation between the two studied parameters, we found important variance between the curve fitting qualities among this models. As we expected, there was a huge disconcordance between the r and r^2 values and the proper indicator of nonlinear curve-fitting quality in non nested models, the Akaike's Information Criterion. Even if exponential and power-type models had best r and r^2, with similar values (-0.55 and 0.3 accordingly) the sum of squares and sum of residuals were significantly higher compared to the logarithmic model, and so their curve fitting quality is questionable.

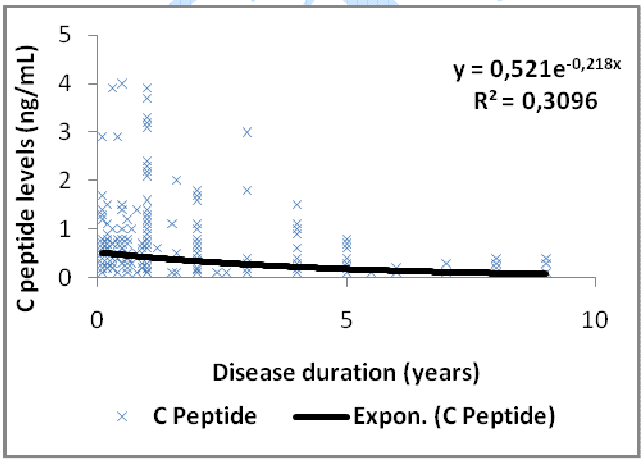

Figure 6. Exponential regression between disease duration and C-peptide levels

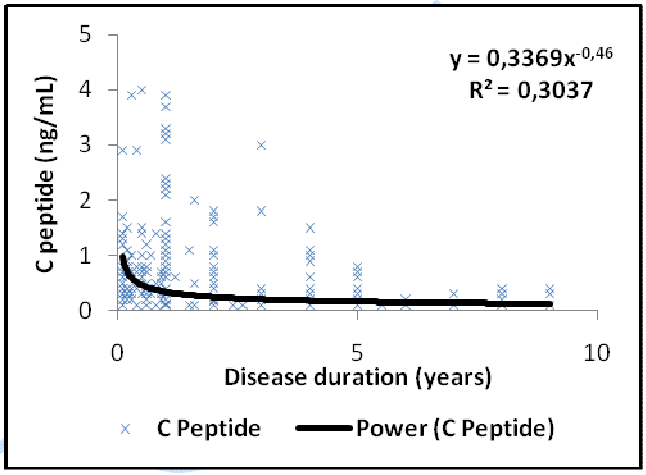

Figure 7. Power-type model regression between disease duration and C-peptide levels

These observations were validated by the analysis of Akaike's Information Criterion, shown in Table 1. The AICc coefficient was far lower in the logarithmic model compared to those of exponential and power-type and therefore the logarithmic model has a far better chance to be the correct one.
For comparing the exponential versus power-type correlation, the difference in the AICc coefficient allowed us to make a larger difference between them, differences between r and r^2 coefficients having no significance.

4. CONCLUSIONS AND DISCUSSIONS

Even it is wide used and accepted, the linear model for fitting biological variables, in many cases is not reliable, having major issues in interpreting and extrapolation of mathematical values to biological significance. In our model, the linear regression line had a good fitting of values, but it extrapolated values in an unacceptable way. At some point, in our model the line coresses the x-axis and so becoming negative, which is a nonsense when when interpreting the results, and as a conclusion this model was rejected despite its good fitting indicators.
For nonlinear models we found a great inconsistency between the r^2 coefficient and the more appropriate indicator, the Akaike's Information Criterion. Models

with lower r and r^2 had a much higher evidence ratio (assessed using the corrected Akaike's Information Criterion method) compared to those with higher r^2. Unfortunately the r^2 is widely missused in comparing nonlinear regression models, many scientists and reviewers insist on it being supplied in research papers even they are dealing with nonlinear data analysis [SN10]. R^2 is not an optimal choice in a nonlinear regression as the total sum-of-squares is not equal to the regression sum-of-squares plus the residual sum-of-squares as is in the case of linear regression and so lacks the above interpretion. As a conclusion we recommend that r^2 should not be used in comparing non-linear regression in biology and especially medical models.

Table 2 - Evidence ratio in comparing regression models

Regression model	Evidence Ratio
Logarithmic over Exponential	99.6%
Logarithmic over Power-type	99.8%
Exponential over Power-type	72.89%

For curve-fitting the relationship between the remaining viable cellular mass in the process of immune mediated cell distruction, the most reliable model is the logarithmic one, more precisely an inverse-logarithmic model. This can be translated in medicine with the important cell distruction from the early phases of the disease, resulting in rapid decrease of residual viable cells, followed by a decrease in progression speed but continuous in time, reaching in late stages of disease values which are tending to zero.

REFERENCES

[BA02] **K. P. Burnham, D. R. Anderson** – *Model Selection and Multi-model Interference; A practical Information-theoretic approach*, Oxford University Press 2002

[GI09] **E. Gayawan, R. Ipinyomi** – *A Comparison of Akaike, Schwartz and R Square Criteria for Model Selectioning Using Some Fertility Models*, in Australian Journal of Basic and Applied Sciences, 3(4): 3524-3530, 2009

[Hut94] **J.C. Hutton** – *Insulin secretory granule biogenesis and the proinsulin processing endopeptidase*, in Diabetologia, 1994; 37:S48-S56

[KPB93] **SE Kahn, RL Prigeon, RN Bergman** – *Quantification of the relationship between insulin sensitivity and beta-cell function in human subjects*, in Diabetes, 1993; 42:1663-1672

[MC04] **H. Motulsky, A. Christopulos** – *Fitting models in biological data using linear and nonlinear models*, 2004

[MYE07] **D. Miao, L. Yu, GS Eisenbarth** – *Role of autoantibodies in type 1 diabetes*, in Front Biosci, 2007; 12:1889-1898

[SN10] **A. N. Spiess, N. Neumeyer** – *An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach*, in BMC Pharmacology, 2010

[WHO99] **World Health Organisation** – *Definition, diagnosis and classification of diabetes mellitus*, in WHO/NCD/NCS/99.2, Geneve, 1999