Additive and Multiplicative Linear Regression Models on a Quantum Dot Database, for the Molecular Beam Epitaxy Technique

Zoltán Bertalan Farkas

Obuda University, Kandó Kálmán Faculty of Electric Engineering, Tavaszmező street, 17, 1084, Budapest, Hungary

farkas.zoltan@uni-obuda.hu

Abstract: *Molecular beam epitaxy is a promising technique for the fabrication of composite semiconductor nanostructures, on a substrate of similar material. The author and his team have a database of the technological input and output parameters (or geometric parameters) for such structures. The database contains 25 quantum dot data vectors, although not every vector has value in all fields. The technological parameters are: growth temperature, surface coverage (of group III droplet material), droplet flux, As pressure, annealing time and temperature. Geometric data are: base circle diameter, structure height and number of structures per unit area. Although Molecular Beam Epitaxy is a stochastic process, therefore its results are best described with distributions, the publications we used as literature sources contained diagrams of specific nanostructures, and our database was built upon those. Focusing on quantum dots, two multiple linear regression models are specified. The first model is defined on the raw data (lin-lin regression or additive model) and the second is defined on the logarithms of the raw data (log-log regression or multiplicative model). This paper contains the results of the regression processes.*

Keywords: molecular beam epitaxy; composite semiconductor quantum dots; multiple linear regression; cross-validation

1 Introduction

The field of nanostructures is currently a hot topic in science and technology. This research focus is the composite semiconductor (GaAs, GaAlAs) subfield. When the idea of creating nanostructures emerged, lattice mismatch technique was the only perceived possible way to do it. Madhukar and Franchi et al comprehensively details the physical processes behind the lattice mismatched system (e.g., Stranski-Krastanov growth mode) [1][2]. Yastrubchak et al. details the effect of dislocation in deeper layers on the surface of lattice mismatched heterostructures and describes the resulting anisotropic growth patters [3]. However, in the 1990s the, then-novel,

droplet epitaxy was developed [4]. This technique offered advantages for the user over lattice mismatching technique in terms of size and topology of nanostructures and the choice the materials for the process. The fabrication of droplet epitaxial nanostrutures belongs to both, semiconductor science/technology (a part of solid-state physics) and materials science. Since our research in recent years, and this paper as well, uses statistical methods, it represents an interdisciplinary approach. The droplet epitaxy technique was described in a former publication of the author and his associates [5].

Over the years several scientific results on droplet epitaxial method have been published. A database has been created on these sources. The database was divided into quantum dot (QD) and nanohole (NH) parts, as there has been too few quantum ring (QR) data vectors to obtain meaningful multiple regression results on them. Two kinds of models were specified over the mentioned two sections of the partitioned database. *Multiple linear regression (MLR)* model was calculated on raw data (additive model or lin-lin regression) and on the logarithms of raw data (multiplicative model or log-log regression). We have sound results even with this simple statistical method; many regression coefficients proved to be significant.

In this work, the ***technological input variables are regressors (estimators, independent variables, explanatory variables)*** and the ***technological output variables are regressands (dependent variables, response variables)***. Therefore, with quantum dots in focus now, the estimators are growth temperature (Kelvin, K), surface coverage [of group III material] (monolayers, ML), droplet [group III material] flux (monolayers per second, ML/s), Arsenic flux (background pressure torr), annealing temperature (Kelvin, K), annealing time (second, s). The regressands are base circle diameter (dimension A, nanometer, nm), structure height (dimension C, nm) and surficial density (number of structures per unit area (SD or S), cm^{-2}).

Unlike prior MBE process studies that report qualitative trends between individual growth parameters and QD morphology, this work derives additive and multiplicative MLR models over a database of 25 QD growth runs. The six technological parameters, four of which was used in the MLR models, varied simultaneously. The regression coefficients can be interpreted directly by future users of droplet epitaxial technique. The alternative analysis tools are more complicated, generally require more data and do not necessarily provide information comparable to pValues of MLR hypothesis testing. The novelty of our research lies in the compilation of the database, and obtaining significant, physically directly interpretable regression coefficients for each model.

In the recent years there has been a huge number of papers, especially machine learning (ML) related research and review papers, on quantum dots. G. Matyszczak et al. mention several application fields of QD's and material science and numerous simulations, data processing and machine learning tools/method in their respective review papers [6].

Both, Hao et al. [7] and Höglund et al. [8] discussed InAs QD growth results with the goal of obtaining uniform QD's after applying Stranski-Krastanov growth mode. Shen et al. [9] applies machine learning to control surface density of InAs QD's (in Stranski-Krastanov growth mode).

Section 2 details the inputs (the evolution of the database, linear regression and descriptive statistics of the data used in linear regression) of the part of our research described in this paper. The test results of multicollinearity of the independent variables are shown in section 2 as well. Section 3 highlights the results of the models, the regression coefficients with their respective significances can be found there. Section 3 also includes auxiliary statistics that are important for the statistical correctness of the application of MLR, as well as metrics for comparing lin-lin and log-log models. This paper ends with conclusions (Section 4) and references. The author extensively used both, perplexity.ai and gemini.google.com. Youtube videos (for example "StatQuest" and "Very Normal") were also used for orienting in both, statistical methodologies and practical calculations.

2. The inputs of this research paper

2.1. The evolution of the database

As mentioned before, a database was created on literary sources. The database was cleaned and checked for consistency with the literature. The database in its current form is the intellectual product to substantial extent of the author. Now, at the time of writing, there is 13 Ga droplet nanohole (NH) vectors, 10 Al droplet NH vectors, 26 quantum dot (QD) data vectors (of which 1 is InAs nanostructure and the rest is GaAs nanostructure) and 9 quantum ring (QR) data vectors. However, the actual data vectors are not of uniform strength as some of them contain missing fields.

2.2. Multiple Linear Regression (MLR)

The MLR is a simple statistical model. It can be referred to in many textbooks in statistics (for example [10]). Among the several prerequisites of MLR (the tool used in this paper is also termed Ordinary Least Squares (OLS) method) the most important ones are that *estimators cannot be collinear variables* and the *residuals (difference of estimated and actual values of the dependent variables) must have normal distribution* and *finally residuals must have identical variance over the regression interval (termed homoscedasticity)*.

Among these prerequisites, multicollinearity is checked separately (section 2.3), since this test only uses the data to be used in the regression calculations. Both,

normality and homoscedasticity of residuals are checked after the regression results are presented, since these tests use the regression residuals (see table 3 for detailed test locations).

In this model one has an output variable which is estimated by the linear combination of the input variables. The regression equation for six independent variables (1).

$$y = c + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 \quad (1)$$

where y is the dependent (regressed, response) variable, c is the constant (y axis intersection, intercept), b_i's are regression coefficients and x_i's are the independent (regressor, estimator) variables.

MLR results do not indicate causality, and with this statistical tool the researcher can never be sure of causal relationships. However, the statistical significance helps the researcher to identify important variables. To explain the significance, the concept of null hypothesis must be introduced. In the context of MLR, the null hypothesis means that a given coefficient does not contribute to the estimation of the dependent variable, its value is effectively zero, and it is only due to statistical fluctuations that its value differs from zero. The probability of the null hypothesis actually being true is based on the ratio of the calculated ("estimated") value of the coefficient to its standard error. This ratio is used as an input of Student's distribution and a "pValue" (probability value) is generated as output. The pValue indicates the probability of rejecting the actually true null hypothesis. Consequently, the smaller the pValue, the more significant the statistical result. In general, a threshold value of 0.05 (5%) is widely accepted in both, natural sciences/engineering and quantitative social sciences (economics, sociology, etc.).

We termed "multiplicative model" the MLR model which is applied to ten based logarithms of the "raw" data as the regression takes the form:

$$\log y = k + d_1 \log x_1 + d_2 \log x_2 + d_3 \log x_3 + d_4 \log x_4 + d_5 \log x_5 + d_6 \log x_6 \quad (2)$$

Taking the antilogarithm of both side (in other words, raising 10 as the base of power to both sides of the equation) equation (3) is obtained:

$$y = 10^k \cdot {x_1}^{d_1} \cdot {x_2}^{d_2} \cdot {x_3}^{d_3} \cdot {x_4}^{d_4} \cdot {x_5}^{d_5} \cdot {x_6}^{d_6} \quad (3)$$

For this reason, it is called multiplicative (or log-log) model.

R^2 is called the coefficient of determination because it explains the proportion of the variance of the dependent variable that can be attributed to the change in the independent variables. In a MLR setting, including additional variables increases the R^2, therefore the adjusted R^2, is calculated to take this effect into account.

The F significance value expresses the overall significance of a MLR model. For this parameter, the null hypothesis is that *all* regression coefficients are zero, so the

dependent variable is determined only by the intercept. The F significance value is the probability that this null hypothesis holds, given the current values of the dependent and independent data used in the regression model. Consequently, as with the pValues for individual variables, the lower the F significance value, the more likely the model is to reveal true relationships.

The additive and multiplicative model forms were both tested because the underlying growth kinetics of GaAs quantum dots are not, a priori, expected to combine additively. Surface diffusion of adatoms follows an Arrhenius-type dependence on substrate temperature (4), and classical nucleation theory predicts that the areal island/dot density scales as a power law of the deposition flux and diffusion coefficient (5) [11].

$$D(T) = D_0 e^{-\frac{E_a}{kT}} \tag{4}$$

$$S \propto (F/D)^p \tag{5}$$

D and D_0 stands for diffusion and coefficient of surface diffusion respectively, F stands for [Ga] flux.

Both relationships combine the governing growth parameters multiplicatively — as products of exponential and power-law terms — rather than additively. Taking the logarithm linearizes them into a sum of log-transformed variables, (2), exactly the multiplicative model considered here. Under this model, each exponent acquires direct physical meaning: for flux/rate predictors, is the coefficient in multiplicative model is power-law scaling exponent; for temperature, the coefficient in multiplicative model reflects an effective activation-type sensitivity local to the sampled range. The additive model instead assumes each parameter contributes an independent, constant marginal effect, a reasonable local approximation over narrow ranges, but less exact than the log-log form when ranges are wide or mechanisms interact. We compared the two families using:

(i) Adjusted R^2

(ii) AIC/BIC, and report the better-performing family per response variable, favoring the multiplicative form on ties given its stronger theoretical grounding.

The regression results and statistics were computed by Microsoft Excel 2016 software ("Data" in primary menu, then "Data analysis" and "Regression"). The diagrams were created also by MS Excel 2016.

Since tests for applicability of MLR are mostly positive (MLR *can be* applied), this version of the article does not include handling of outliers, although attempts were made to do so during Excel operations.

2.3. The descriptive statistics of the database

As described below, all data in the additive model except for arsenic pressure and surface density are in "raw" form. The reason for representing arsenic pressure and surface density using logarithms is that these data span several decimal orders of magnitude, so we can only obtain meaningful regression coefficients by using their logarithms. We emphasize again that in this paper both arsenic pressure and surface density appears only as the logarithm of raw data. Descriptive statistics of the QD database are shown on Table 1 a-b.

Table 1 a

Descriptive statistics of the technological input parameters of the quantum dot database (ML stands for monolayer, s and sec stands for secundum)

	Growth temp. (Kelvin)	**Droplet flux (ML/s)**	**Surf. Cov. (ML)**	**LOG10 (Arsenic flux gauge (Torr))**	**Annealing time (sec)**	**Annealing temp. (Kelvin)**
average	517.8	0.571	2.77	-4.501	528	617.8
std. Dev.	63.6	0.864	0.965	1.122	445	101.6
No. of data	25	14	25	25	25	24
min	473	0.025	1.5	-9	60	423
max	623	3	3.75	-3.61	2400	773

Table 1 b

Descriptive statistics of the technological output parameters of the quantum dot database

	Base circle diameter (size A) (nm)	**Struct. height (size C) (nm)**	**LOG10 (Surface density (1/cm^2))**
average	95.6	18.61	9.16
std. Dev.	49.77	16.76	0.740
No. of data	25	25	25
min	40	2	8.204
max	250	51	10.560

2.3 Variance Inflation Factors of the independent variables

MLR models have prerequisites for being statistically adequate. The predictors of a linear regression model must be checked for collinearity (the predictors themselves must not be linearly connected). A useful metric for collinearity is the Variance

Inflation Factors (VIF). Each predictor (regressor) x_i in the model is estimated by a MLR of the *other* predictors. Then the coefficient of determination is calculated for the regressed predictor (R_i^2). The VIF_i can be obtained by the formula (4):

$$VIF_i = \frac{1}{1-R_i^2} \tag{4}$$

The values of VIF below 5 (R^2 under 0.8) are considered to indicate weak collinearity.

Below, Table 2a shows R^2 and adjusted R^2 values for each predictor are shown together with the respective VIF's.

Table 2a

Coefficient of determination (R^2), adjusted coefficient of determination (adj R^2), VIF values calculated by each R^2, and number of observations (number of data vectors used to calculate R^2 s)

	Growth.T	**Dropl.Flx.**	**Surf.cov.**	**Log As**	**Ann.time**	**Ann, temp.**
R^2	0.815	0.973	0.010	0.615	0.739	0.860
adj R^2	0.777	0.954	-0.139	0.533	0.684	0.831
VIF	5.415	37.584	1.009	2.596	3.829	7.163
adj VIF	4.473	21.924	0.878	2.145	3.163	5.917
# of obs.	24	13	24	24	24	24

As it can be seen both, droplet flux and annealing temperature has high VIF score, that of former is extreme high, therefore these variables will be removed from further calculations.

As multiplicative model (MLR on logarithms of raw data) is a part of this paper, VIF's calculated on logarithms of raw data are presented in Table 2b.

Table 2b

R^2, adj R^2, VIF values calculated by each R^2, and number of observations calculated on logarithms of all technological variables

	LOG10 Growth.T	**LOG10 Dropl.Flx.**	**LOG10 Surf.cov.**	**Log As**	**LOG10 Ann.time**	**LOG 10 Ann, temp.**
R^2	0.300	0.963	0.038	0.127	0.304	0.854
adj R^2	0.200	0.938	-0.010	-0.010	0.205	0.824
VIF	1.428	27.52	1.039	1.131	1.437	6.867
adj VIF	1.250	16.05	0.909	0.990	1.258	5.673
# of obs.	25	13	25	25	25	24

As with VIF calculations of raw data, LOG droplet flux and LOG annealing temperature is strongly collinear with the other explanatory variables, thus these two variables will not be used in the multiplicative model.

In the following section, I present the results (coefficients) of the regression models and further tests of the statistical correctness of the use of MLR.

2.4. Theory-results correspondence table

Table 3 shows our results in the context of theoretical background.

Table 3

Concept	Instantiation
Additive (1) and multiplicative model (2)	Table 4 and table 6
OLS method	pValues
Model selection	AIC, BIC, adjusted R^2 (table 8), cross-validation MAE, RMSE, R^2 (diagram 2)
Diagnostic tests	VIF (table 2 a-b), skewness and kurtosis (section 3.1 and table 4, section 3.2 and table 6), Shapiro-Wilk test (section 3.1 and 3.2), White test (table 5), Q-Q plots (diagram 1 A, C, S)
Physical explanation of coefficients	Section 3.1

3. Multiple linear regression results

Both regression families introduced in section 2, the additive model (1) and the multiplicative (log-log) model (2), are fitted independently to each of the three geometric response variables using the N = 25 MBE growth runs described in section 2.3 (Table 1a-b). Coefficients are estimated by ordinary least squares, model adequacy is then assessed using the criteria (adjusted R^2, AIC/BIC, cross-validated RMSE/MAE) and the diagnostic tests (Shapiro–Wilk, White, VIF).

3.1 Additive model results

If the regression coefficients are calculated on raw data, an additive model is defined as the regressed variable is approximated by the *sum* of the products of regression coefficients and independent variable values. In this case the dimension of the regression coefficients is either size (in nanometers) over the dimension of the estimator or inverse area (number of nanostructures on one square centimeter) over the dimension of estimator. If the independent (regressor) variable is growth temperature and the dependent (regressed) variable is nanostructure diameter (base circle diameter), the regression coefficient has the dimension that converts Kelvin to nanometers, therefore the corresponding dimension is nm/K. All the coefficients should be interpreted in accordingly.

Table 4 shows the regression coefficients of lin-lin model and the respective significances. The significance legend can be found under Table 4. Please note, if

no asterisks can be found at the upper right index of the coefficient, then the coefficient is not significant even at 0.05.

Table 4

Additive model QD- regression coefficients and their significance levels

	Coeff. "A"	***Coeff. "C"***	***Coeff. "logS"***
Intercept	-242.7**	-107.0***	12.27****
Growth temperature (Kelvin)	0.825****	0.276****	-0.0108****
Surface coverage (ML)	8.162	3.673	0.365**
LOG10 (Arsenic flux gauge (Torr))	34.33**	8.681*	-0.479**
Annealing time (sec)	0.0818**	0.0215*	-0.00125**
Number of data vectors	25	25	25
Adjusted R square	0.504	0.557	0.614
Significance F	0.00099	0.000344	9.21E-05
Kurtosis of residuals	3.484	1.645	-0.863
Skewness of residuals	-1.451	0.900	-0.314

Significance levels: * significant @ 0.05 (pValue<0.05);

** significant @ 0.01; ***significant @0.001; **** significant @0.0001

In the subsequent paragraphs, the explanation of how to interpret i. raw data - raw data, ii. logarithm – raw data iii. raw data – logarithm and iv. logarithm – logarithm coefficients is presented.

1. For example, the 0.825 (nm/K) value in the “growth temperature” row and Coeff. “A” of table 3 column is interpreted, that statistically every additional Kelvin raise in growth temperature coincides a 0.825 nm increase in dimension “A” (base circle diameter).
2. The 34.33 entry in Log 10 Arsenic flux row is interpreted, that every 10-fold increase in Arsenic flux statistically causes additional 34.33 nm increase in base circle diameter. Alternatively, this entry can be interpreted that each 1% increase in Arsenic flux statistically increases the base circle diameter by $0.148 = 34.33 \log_{10}(1.01)$ nanometers.
3. The -0.0108 entry of growth temperature row and coeff. “logS” column of table 2 is interpreted that every additional Kelvin increment in growth

temperature statistically causes $10^{-0.0108}$=0.975-fold increase (about a 2.5% decline) in surficial density.

4. Finally, the -0.479 entry in log10 As flux gauge row and coeff. "logS" column of table 2 implies that every 10-fold increase in As pressure statistically coincides a factor of $10^{-0.479}$=0.332-fold increase (decrease by about 66.8%) in surface density. Alternatively, this entry means, that every 1% increase in As pressure statistically coincides a $1.01^{-0.479}$ = 0.9952-fold increase (a 0.48% decrease) in surface density.

All four technological variables were significant for surface density (S), and all technological variables except surface coverage were significant for base diameter (A) and structure height (C). The sign of the regression coefficients is positive for base diameter and structure height. This means that the increase in growth temperature, Arsenic pressure and annealing time has a statistically positive (increasing) effect on both, base diameter and structure height. For surface density, the surface coverage coefficient is positive, while the other three coefficients are negative. All these results can be interpreted as the amount of group III droplets positively correlates with the number of structures per unit area, and the increase in the other three technological parameters results in fewer structures per unit area, but the structures become larger. The adjusted coefficients of determination are in the range 0.5-0.6, which means that about 50-60% of the variation in the technological output variables is attributable to the estimator variables. The significance F indicates the probability that an actually valid null hypothesis for the *entire model* is rejected (H_0: *none* of the explanatory variables contributes to the estimation of the regressand). The F values are very small, therefore, our results are significant.

The high value of growth temperature contributes to longer average diffusion length of adatoms therefore the positive coefficient of growth temperature for A and C regression, and the negative coefficient for S regression is completely explained by physics. Longer diffusion length results in less but larger nanostructure.

The high Arsenic pressure increases the speed of nucleation (less time for Ga atom, to move), theoretically high As pressure should correspond with more and smaller dots. For this reason, positive coefficient for S parameter and negative coefficient for A and C parameters should be obtained. What we have is just the opposite, moreover all the Arsenic coefficients are significant. Further examination is required to find explanation for this result.

Higher Gallium volume (surface coverage) contributes to both, getting bigger nanostructures (A and C coefficients are positive) and getting more nanostructures (S coefficient is positive).

Finally, longer annealing time contributes to processes like Ostwald ripening, which results in getting fewer large nanostructures instead of many large ones. The findings coincide with the theory, as the coefficients for A and C dimensions are positive and coefficient for S parameter is negative.

The skewness and kurtosis are higher moments of a distribution (the first moment is the average, the second moment is the standard deviation). The values of skewness and kurtosis indicate if a distribution is a normal distribution. One prerequisite of linear regression that the residuals must be normally distributed. The skewness and kurtosis of the residuals marginally exceed the z score limits of 5% significance in the case of base circle regression, but both moments of residuals of the structure height and surface density regression are within the limits.

Another condition to be met when using linear regression is the homoscedasticity of the variables, in other words, the residuals must have the same variance across all values of the predictor variables. A widely used test for homoscedasticity is the White test. In the White test, an auxiliary regression is calculated for the square of the residuals of each regression (in our case, for the regressions of the base circle diameter, the structure height and the surface density). In the auxiliary regression, the square of the residuals is regressed on three groups of explanatory variables. These groups are the explanatory variables of the "original" regression, the squares of these explanatory variables and the cross products of the explanatory variables. Given that the "original" regressions had four explanatory variables, the auxiliary regression contains fourteen explanatory variables (four technical parameters, the squares of the latter (four elements) and six two-factor cross products). The null hypothesis of the White test is that these fourteen explanatory variables explain the change in variance over the regression interval (of the original regression). The metric of the White test is the R-squared of the auxiliary regressions multiplied by the number of data vectors. This metric is compared to the critical value of the χ^2 (chi-squared) distribution, whose degrees of freedom are equal to the nonzero coefficients of the auxiliary regression and the significance level (which, as in the previous cases, was chosen to be 0.05). The results of the auxiliary regressions are shown in Table 5.

Table 5

Auxiliary regression results; regressions were computed on the squares of residuals with using all the four technological variables, obtained in the regressions presented in Table 3

	aux. regr. res. "A"	**aux. regr. res. "C"**	**aux. regr. res. "S"**
R^2	0.419	0.3964	0.7426
number of data vectors (n)	25	25	25
nR^2	10.48	9.911	18.57
crit. value $\chi^2_{0.05,11}$	19.675	19.675	19.675
crit. value $\chi^2_{0.05,14}$	23.685	23.685	23.685

Since all nR^2 values are below the critical chi-square value for both degrees of freedom (11 and 14), the null hypotheses are rejected at the 5% significance level in all three cases, so homoscedasticity holds for the residuals in all three additive model regressions. As mentioned, there were fourteen explanatory variables in the auxiliary regressions (hence the degrees of freedom are 14), three of which had regression coefficients of zero (hence the degrees of freedom are 11).

Another tool to check the normal distribution of residuals is the quantile-quantile plot (or Q-Q plot for short). Q-quantile means to decompose a finite set to equal q subsets, which are ordered according the magnitude of an attribute. In this research we have 25 data vectors, therefore quantiles can be defined (and each quantile represents a single data vector) that represent 4% of the "population". As the name suggests, a Q-Q plot compares quantiles to each other. The horizontal axis shows the quantiles of the cumulative distribution function of the distribution under study (in this case the normal distribution), and the vertical axis shows the quantiles of the values of the attribute to be compared with the distribution (in this case the residuals of the lin-lin regressions of each technological output variable). As reference, the trend line is also shown in each diagram. If the residuals fit the normal distribution, then the points of the Q-Q plot would fit the trend line. The Q-Q plots for the residuals of A, C, S regressions are shown in Diagram 1 A, C, S respectively.

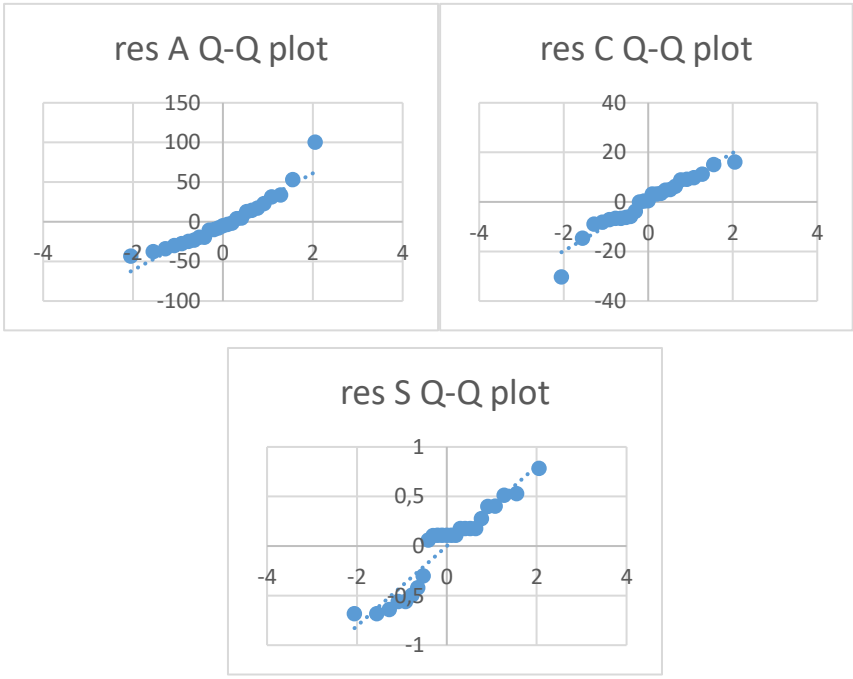


Diagram 1 A, C, S

Qunatile-Quantile plots of the residuals of A, C and S regressions

The Q-Q plots deviate from the trend lines which fits our findings on kurtosis and skewness, the residuals do not have perfect normal distribution.

As a part of the discussion of the result, the cross-validation results are presented next. Cross validation means that a part of the database is used to "train" the statistical model, the rest of the database is used to test the "trained" model. In order to efficiently use data, the roles of "training data" and "test data" are swapped between the parts. Specifically, we divide the data into three parts, 3-fold cross validation is carried out. Two of the three parts are used to train the model (the regression coefficients are calculated on the union of the data vectors of the two parts) and the third part is used to test it, by calculating metrics to check the accuracy of the model. Two approaches are used for the division into three parts. The QD database was decomposed into three parts in two ways. The first method was to create "blocks" according to the consecutive ordinal numbers of the vectors. The second method was to create sections by selecting every third vector.

Diagram 2 summarizes the metrics of testing the model. Two three-fold cross validation runs, result in six markers on each diagram. MAE stands for mean absolute error, RMSE stands for Root Mean Square Error.

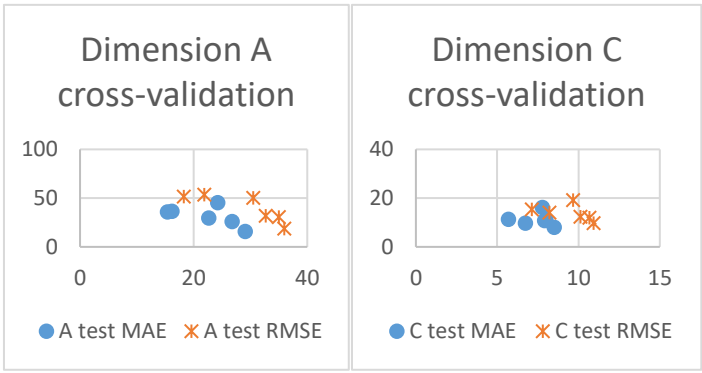


Diagram 2 A, C

Cross-validation diagram of dimensions A and C, horizontal axis: training data MAE and RMSE values, vertical axis: test data values

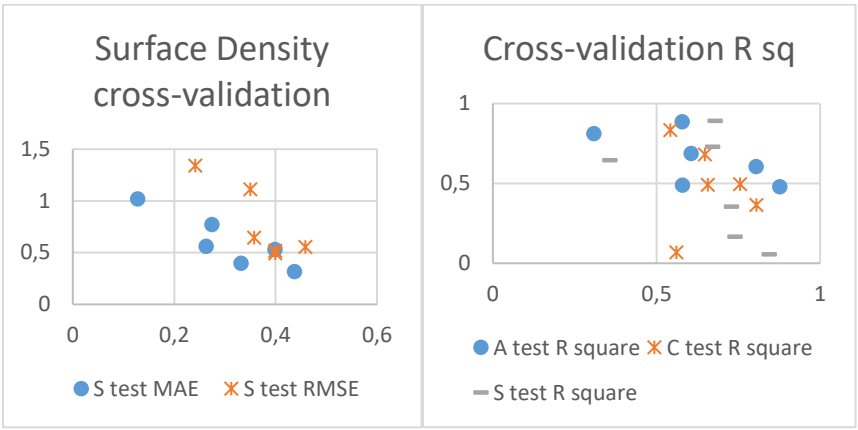


Diagram 2 S, R

Diagram S: MAE and RMSE values as with diagram 2 A and C, diagram R: horizontal axis: training data R square values, vertical axis: test data R square values of all the three cross-validation runs

To check once more whether the residuals are normally distributed, Shapiro-Wilk tests were calculated. The W score of the test was 0.025 for A regression, 0.008 for C regression and 0.004 for LOG S regression (all regressions are in lin-lin terms). The critical value for 5% significance and N=25 residuals is W=0.918 *above which* the residuals would be assumed to fit a normal distribution, therefore the non-normality of residuals holds with much higher than 5% significance (lower pValue).

3.2. Multiplicative model results

If the MLR is calculated on the logarithms of the raw values an additive model is obtained as the dependent variable is obtained as the *product* of the powers of the independent variables. Each independent variable is raised to the power of the respective regression coefficient. Finally, the product is multiplied by ten raised to the power of constant “k” obtained in the same regression.

Table 6

Multiplicative model regression coefficients and their significance levels

	Coeff "LOG A"	Coeff "LOG C"	Coeff "LOG S"
Intercept	-7.99***	-17.96***	41.23
LOG Growth temperature (Kelvin)	3.563****	6.175***	-11.54****
LOG Surface coverage (ML)	0.123	0.717	2.043**
LOG (Arsenic flux gauge (Torr))	0.062*	0.0867	-0.189*
LOG Annealing time (sec)	0.189*	0.359	-0.950**
Number of data vectors	25	25	25
Adjusted R square	0.512	0.454	0.610
Significance F	0.00086	0.00245	0.0001
Kurtosis of residuals	0.804	1.170	-0.198
Skewness of residuals	0.584	-0.952	-0.160

Of the significant coefficients the following observations can be made. The LOG growth temperature regressor of LOG base circle diameter is close to 3.5, the square root of the seventh power, thus it can be interpreted that base circle diameter depends on the square root of seventh power of growth temperature. Though the regression coefficients are not significant at 5%, base circle diameter seem to depend on the eighth root of the surface coverage. Similarly, base circle diameter seems to depend on sixteenth root of Arsenic pressure, and fifth root of annealing time (both coefficients are significant at 5%). Structure height statistically significantly depends on the sixth power of temperature, the pValue of this coefficient is under 0.01%. No other coefficient proved to be significant at 5% level for LOG structure height. All the four coefficients were significant with respect to LOG surficial density (LOG S). Interestingly, surficial density proportional with minus eleventh power of growth temperature (surficial density decreases very steeply as growth temperature increases). Surficial density increases proportionally with the second power of droplet material quantity (surficial coverage). Surficial density depends inversely on both Arsenic pressure and annealing time. Surficial density depends on the reciprocal of fifth root of Arsenic pressure and the reciprocal of annealing time. Of the coefficients of the multiplicative model, the coefficients for LOG growth temperature were the most significant ones.

The results of White test of the log-log regression residuals are shown in Table 7.

Table 7

Auxiliary regression results; regressions were computed on the squares of residuals with using all the logarithms of four technological variables, obtained in the regressions presented in Table 6

	aux. regr. res. "LOG10 A"	**aux. regr. res. "LOG10 C"**	**aux. regr. res. "LOG10 S"**
R^2	0.293	0.2061	0.7488
number of data vectors (n)	25	25	25
nR^2	7.346	5.153	18.72
crit. value $\chi^2_{0.05,11}$	19.675	19.675	19.675
crit. value $\chi^2_{0.05,14}$	23.685	23.685	23.685

It was found that in the auxiliary regressions the coefficients of three estimator were zero, therefore, similar to the White test of the residuals of the additive models, the degrees of freedom of the chi square distribution are 11 and 14. The residuals of all three log-log regressions are proved to be homoscedastic at the 5% confidence level.

The Shapiro-Wilk test was taken for log-log regression residuals as well. The W statistics was 0.007 for LOG A regression, 0.017 for LOG C regression and 0.0007 for LOG S regression. As with their lin-lin W statistics counterparts these values are very close to zero, thus indicate a depart from normality with much higher than 5% significance (critical value for 5% significance and N=25 residuals is W=0.918). This result is consistent with the Q-Q plots of the multiplicative model (not shown).

3.3 Comparing additive and multiplicative models

The general tools for comparing statistical models, are coefficient of determination, the Akaike Information Criterion corrected for small samples (AICc) and Bayesian Information Criterion (BIC).

The formulae for AICc (5) and BIC (6) for MLR and supposing the normality of residual are as follows:

$$AICc = n\, ln\left(\frac{SSE}{n}\right) + 2k + \frac{2k(k+1)}{n-k-1} + n\ln(2\pi) + n \tag{5}$$

$$BIC = n\, ln\left(\frac{SSE}{n}\right) + k\ln(n) + n\ln(2\pi) + n \tag{6}$$

Table 8 summarizes the results obtained for AICc, BIC and adjusted R^2.

Table 8

AICc, BIC and adjusted R^2 for lin-lin and log-log regressions

	A	C	S	LOG A	LOG C	LOG S
AICc	259.8	202.6	43.19	-15.4	24.5	43.40
BIC	262.5	205.3	45.83	-12.75	27.14	46.04
adj R^2	0.504	0.557	0.614	0.512	0.454	0.610

In terms of AICc and BIC, the log-log regressions were better models (smaller metrics was obtained) for dimensions A and C, whereas lin-lin and log-log surface density (S) regressions resulted nearly equal metrics. In terms of adjusted R^2 lin-lin and log-log regression resulted nearly equal metrics for dimension "A" and surface density "S", and lin-lin regression resulted somewhat higher coefficient of determination for dimension "C".

4. Conclusions

In this paper, multiple linear regression models were applied on a database of 25 Quantum Dot (QD) data vectors. The QD database consisted six technological parameters (estimators/explanatory variables) and three geometric parameters (regressands/response variables). The regression models were reduced to four explanatory variables due to multicollinearity problems. The explanatory variables used in this paper were growth temperature, surface coverage (group III material quantity), Arsenic pressure and annealing time. A basic model of statistics proved to yield meaningful results on the QD database. Both, additive and multiplicative model yielded positive regression coefficients when the dependent variables were the base circle diameter and structure height. The regression of surface density resulted in positive coefficient for surface coverage and negative coefficients for the other three explanatory variables.

The adjusted R^2 values were in the 0.45-0.61 range, which means that the independent variables account for a significant part of the observed variance of dependent variables, while realistically accounting for both, the stochastic character of growth process and unobserved factors. The difference between the R^2 values (not presented) and the adjusted R^2 values are within the 0.1 mark which implies a balance between goodness-of-fit and model complexity. The use of the four independent variables contributes to explanatory power of the model without compromising model parsimony. The highest F statistics of the six MLR is 0.00245, the other five are well below this level, therefore our model calculations are robust.

Both, the database and regression results were subjected to three tests to determine whether the prerequisites of OLS (Ordinary Least Squares) method are fulfilled. As

mentioned, the multicollinearity test resulted in reduction the number explanatory variables from six to four. The normal distribution of residuals was checked through excess kurtosis and skewness of the distribution of residuals. Here, the normality of residuals marginally failed to pass the 5% significance test in case of base circle regression of the additive model. Shapiro-Wilk test indicated in the case of both, additive and multiplicative models, at high significance that residuals are do not have a normal distribution. No action was taken in this case, as normality of residuals is the "softest" prerequisite of the three.

Finally, the homoscedasticity (identical variance) of the residuals has been checked with White's test, all the six-regression proved to be homoscedastic with a 5% significance. The regression results obtained by significant coefficients were compared to database values (not shown in this paper).

Here, the coefficients of correlation were in the 0.6-0.85 range which means that about 40%-70% of the variance of the geometrical parameters is explained by the four technological parameters.

The effectivity of estimating the response variables with regressions on the explanatory variables additive and multiplicative models were compared by using Akaike Information Criterion (corrected for small samples), Bayesian Information Criterion and adjusted coefficient of determination. The AICc and BIC score of log-log model was less than that of the lin-lin model in the case of dimensions A and C, therefore the log-log estimation is more effective for these parameters.

In the case of LOG S, these criteria resulted in nearly identical scores. The adjusted R^2 values were nearly identical for lin-lin and log-log model in the case of dimension A and LOG S values. The lin-lin estimation of dimension C was better, in terms of adjusted R^2, than its log-log counterpart.

References

[1] A. Madhukar: Some novel in situ studies of strained semiconductor epitaxy on patterned/compliant substrates. (Final technical report) University of Southern California. 1999

[2] S. Franchi et al.: Quantum dot nanostructures and molecular beam epitaxy. Progress in Crystal Growth and Characterization of Materials. Vol. 47 Issues 2-3 (2003) pp 166-195

[3] O. Yastrubchak et al.:Misfit dislocations and surface morphology of lattice mismatched GaAs/InGaAs heterostructures. Physica E: Low-dimensional ystems and Nanostructures. Vol 17, (April 2003) 561-563

[4] T. Suzuki – T. Nishinaga: First real time observation of reconstruction transition associated with 0Ga droplet transition and annihilation during

molecuar beam epitaxy of GaAs. Journal of Crystal Growth. Vol 142 Issues 1-2 (1994) pp. 49-60

[5] Z. Farkas, A. Ürmös, Á. Nemcsics: Regression study on logarithms of geometrical and technological parameters of zero-dimensional nanostructures. 2021 IEEE 4th International Conference and Workshop… (CANDO-EPE) pp 163-166

[6] J. J. Mim et al: Machine learning-driven advances in nanotechnology: from materials design to process optimization – A review. Materialstoday Communications. (2026) 114485. DOI:10.1016/j.mtcomm.2025.114485

[7] Hao Hui-Ming, Su Xiang-Bin, Zhang Jing, Ni Hai-Qiao, Niu Zhi-Chuan, "Molecular beam epitaxial growth of high quality InAs/GaAs quantum dots," *Chinese Physics B*, Vol. 28, No. 7, Article 078104 (2019). DOI: 10.1088/1674-1056/28/7/078104.

[8] Linda Höglund, E. Petrini, C. Asplund, H. Malm, J. Y. Andersson, Per-Olof Holtz, "Optimising uniformity of InAs/(InGaAs)/GaAs quantum dots grown by metal organic vapor phase epitaxy," *Applied Surface Science*, Vol. 252, Issue 15, pp. 5525–5529 (2006). DOI: 10.1016/j.apsusc.2005.12.128.

[9] Shen, C., Zhan, W., Xin, K. *et al.* Machine-learning-assisted and real-time-feedback-controlled growth of InAs/GaAs quantum dots. *Nat Commun* **15**, 2724 (2024). https://doi.org/10.1038/s41467-024-47087-w

[10] J.D. Jobson: Multiple Linear Regression. In: Applied Multivariate Data Analysis. (Springer Texts in Statistics. Springer, New York, NY.,1991), pp. 219-398

[11] J. A. Venables, G. D. T. Spiller, and M. Hanbücken, "Nucleation and growth of thin films," Reports on Progress in Physics, vol. 47, no. 4, pp. 399–459, 1984, DOI: 10.1088/0034-4885/47/4/002