1-2014

Generalized Least-Squares Regressions III: Further Theory and Classification

Nataniel Greene
CUNY Kingsborough Community College

Generalized Least-Squares Regressions III: Further Theory and Classification

NATANIEL GREENE
Department of Mathematics and Computer Science
Kingsborough Community College, CUNY
2001 Oriental Boulevard, Brooklyn, NY 11235
UNITED STATES
ngreene.math@gmail.com

Abstract: This paper continues the work of this series with two results. The first is an exponential equivalence theorem which states that every generalized least-squares regression line can be generated by an equivalent exponential regression. It follows that every generalized least-squares line has an effective normalized exponential parameter γ between 0 and 1 which classifies the line on the spectrum between ordinary least-squares and the extremal line for a given set of data. The second result is a fundamental formula for the generalized least-squares slope: $b = (\rho + \lambda\kappa)(\sigma_y/\sigma_x)$ where κ is a measure of scatter and λ is an increasing function of γ ranging from 0 to 1. According to the formula it is the nature of generalized least-squares regression to incorporate the scatter of the data into the slope. The regression methods differ in how much the scatter should be weighted. The parameter λ measures this and varies from one regression method to another.

Key–Words: Least-squares, generalized least-squares, symmetric least-squares, weighted ordinary least-squares, orthogonal regression, geometric mean regression.

1 Overview

In the first two papers of this series [2, 3], the known cases of symmetric least-squares regression, as well as a variety of least-squares regression methods which may not have been known or fully explicated were derived by this author. The derivation of each method was made efficient through the use of Ehrenberg's formula for the ordinary least-squares error [1] and through the extraction of a weight function $g(b)$ which characterizes the regression. For every case of generalized least-squares, the error between the line and the data was shown to be a product of the weight function $g(b)$ and Ehrenberg's error formula. The re-derivation of orthogonal regression was notable in showing that orthogonal regression minimizes the average harmonic mean of the square x and y deviations, allowing it to also be categorized alternatively as harmonic mean regression.

This work was then generalized into a theory for deriving, analyzing, and classifying all symmetric and weighted least-squares regression methods. All symmetric least-squares regressions were categorized by a generating function $\psi(x, y)$ that is positive, even, and homogenous of degree two in x and y, a weight function $g(b) = \psi(1, 1/b)$, and an indicative function $G(b) = 2g'(b)/g(b) - g''(b)/g'(b)$. All weighted ordinary least-squares regressions, of which symmetric regressions are a part, were categorized by grouping them into classes with the same general indicative function $G(b)$ and the same general weight function $g(b) = 1/(c + k\int \exp(-\int G(b)db)db)$. This paper continues the development of this theory.

It is shown here that every generalized least-squares regression line can be generated by an equivalent exponential regression. All generalized least-squares regression lines fan out from the mean point and are bounded between the OLS $y|x$ line and the extremal exponential line. The normalized parameter γ in the weight function $g_0(b) = \exp(-\gamma p_0 |b|)$, $0 \leq \gamma \leq 1$ is then used to numerically classify the regression lines for a given set of data on the spectrum between the ordinary least-squares line and the extremal line.

A fundamental formula for the generalized least-squares slope is then derived based on γ. It is clear from the formula that every generalized least-squares method incorporates the scatter of the data into the slope. The degree to which they do so is measured by a parameter λ that is a function of γ and also classifies the lines on the spectrum between the ordinary least-squares line and the extremal line.

ISBN: 978-960-474-366-7

2 The Equivalence of Generalized Regression and Exponential Regression

A variety of symmetric and weighted least-squares regression methods were derived and the various regression lines were observed empirically to fan out between the OLS $y|x$ line and the OLS $x|y$ line. Exponential regressions were also defined with weight function given by $g(b) = \exp(-p|b|)$. Exponential regression lines fan out between the OLS $y|x$ line and the extremal line $y = a_0 + b_0 x$ where $a_0 = \mu_y - b_0\mu_x$ and $b_0 = \rho\frac{\sigma_y}{\sigma_x} + \operatorname{sgn}\rho\sqrt{1-\rho^2}\frac{\sigma_y}{\sigma_x}$. Exponential regressions were treated as just another interesting case of weighted ordinary least-squares regressions. Here their importance is made clear in that exponential regressions subsume all possible generalized least-squares lines.

2.1 Coefficient of Scatter

Recall that $1-\rho^2$ measures the scatter of the data cloud away from the ordinary least-squares regression line. It follows that the square root of this is also a measure of scatter. It is useful here to have a notation for this quantity and for its sign to agree with ρ.

Definition 1 *Define*

$$\kappa = \operatorname{sgn}\rho\sqrt{1-\rho^2} \tag{1}$$

to be the coefficient of scatter satisfying $\rho^2 + \kappa^2 = 1$.

When the correlation coefficient ρ is parametrized as $\cos\theta$ then κ is parameterized as $\operatorname{sgn}\rho\sin\theta$. The κ notation makes the expressions in the formulas developed earlier simpler and more more intuitive. In κ notation the slope of the extremal line can be rewritten as $b_0 = (\rho+\kappa)\frac{\sigma_y}{\sigma_x}$.

2.2 The Exponential Parameter Classifies Non-exponential Regressions

An explicit formula for the exponential regression slope b was already given [2] and is quoted in the next theorem.

Theorem 2 *(Exponential Slope Formula) Let* b *be the slope of an exponentially weighted least-squares regression with parameter* p*. Then*

$$b = \rho\frac{\sigma_y}{\sigma_x} + \frac{\operatorname{sgn}\rho}{p}\left(1 - \sqrt{1-\kappa^2\left(\frac{\sigma_y}{\sigma_x}\right)^2 p^2}\right) \tag{2}$$

where $0 \le p \le p_0$ *and* $p_0 = 1/\left(|\kappa|\frac{\sigma_y}{\sigma_x}\right)$.

The exponential regression lines fan out from the mean point (μ_x, μ_y) and vary continuously from the ordinary least-squares $y|x$ line with $p = 0$ to the extremal line with $p = p_0$. The negative sign in front of the radical was chosen in order that the Hessian determinant be positive.

In the second part of the series [3] the Second Discrepancy Formula was derived. It is re-written here in parallel form in the next theorem.

Theorem 3 *(Second Discrepancy Formula) The slope of a generalized least-squares regression line is given by*

$$b = \rho\frac{\sigma_y}{\sigma_x} - \frac{g(b)}{g'(b)}\left(1 - \sqrt{1-\kappa^2\left(\frac{\sigma_y}{\sigma_x}\right)^2\left(\frac{g'(b)}{g(b)}\right)^2}\right). \tag{3}$$

The similarity between the two formulas is striking. The two formulas become the same when one sets

$$\frac{g'(b)}{g(b)} = -p\operatorname{sgn} b \tag{4}$$

where again $\operatorname{sgn} b = \operatorname{sgn}\rho$. Denote the weight function satisfying this equation by $g_0(b) = \exp(-p|b|)$. It is now clear that every least-squares regression line with arbitrary weight function $g(b)$ is the solution to a corresponding exponential regression problem with weight function $g_0(b) = \exp(-p|b|)$ and parameter p.

Since p_0 varies from problem to problem it is useful to have a normalized exponential regression parameter γ satisfying $p = \gamma p_0$ where $0 \le \gamma \le 1$.

Definition 4 *Define*

$$\gamma = \frac{p}{p_0} \tag{5}$$

to be the normalized exponential parameter and write

$$g_0(b) = \exp(-\gamma p_0|b|). \tag{6}$$

As γ varies between 0 and 1 the weight function $g_0(b) = \exp(-\gamma p_0|b|)$ generates all the possible regression lines. The next theorem gives the formula for the parameter p and the normalized parameter γ once the slope is known.

Theorem 5 *(Exponential Equivalence Theorem) Let* b *be the slope of a generalized least-squares regression line with associated weight function* $g(b)$*.*

ISBN: 978-960-474-366-7

Then this line can be generated from an equivalent exponentially weighted least-squares regression with weight function $g_0(b) = \exp(-p|b|) = \exp(-\gamma p_0|b|)$ *and effective parameters given by*

$$p = \frac{2\left(b - \rho\frac{\sigma_y}{\sigma_x}\right)}{\left(b - \rho\frac{\sigma_y}{\sigma_x}\right)^2 + \kappa^2\left(\frac{\sigma_y}{\sigma_x}\right)^2} \tag{7}$$

and

$$\gamma = p\kappa\frac{\sigma_y}{\sigma_x}. \tag{8}$$

Proof. Solve for p in terms of b in the exponential slope formula. ■

Once a generalized regression line has been computed, one can always go back and compute the effective parameters p or γ corresponding to the equivalent exponential regression line. Every generalized regression line can now be assigned an effective γ value between 0 and 1. The next formula reveals the simple form which all generalized least-squares regression lines have.

Theorem 6 *(General Slope and Intercept Formulas) Let* b *be the slope of a generalized least-squares regression line with effective normalized exponential parameter* γ*, then*

$$b = (\rho + \lambda\kappa)\frac{\sigma_y}{\sigma_x} \tag{9}$$

where $0 \leq \lambda \leq 1$ *and* λ *and* γ *are related by the equations*

$$\lambda = \frac{\gamma}{1 + \sqrt{1 - \gamma^2}} \tag{10}$$

and

$$\gamma = \frac{2\lambda}{\lambda^2 + 1}. \tag{11}$$

The y*-intercept is*

$$a = \mu_y - b\mu_x. \tag{12}$$

Proof. Begin with the exponential slope formula, substitute $p = \gamma/(\kappa\sigma_y/\sigma_x)$ and simplify. ■

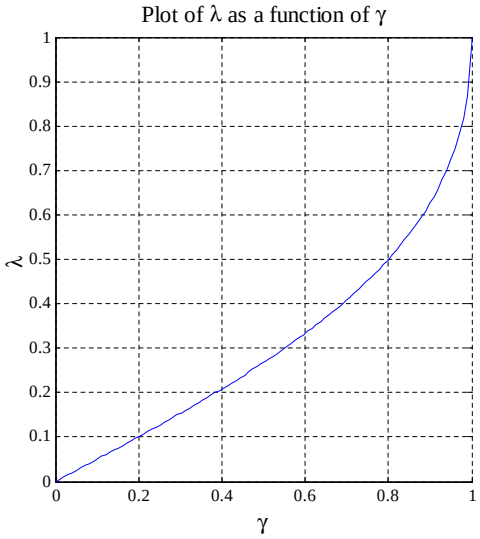
This fundamental formula makes clear that every generalized least-squares regression method seeks to incorporate the scatter of the data into the slope. The regression methods differ in how much the scatter should be weighted. The scatter parameter λ measures this and varies from one regression method to another.

Corollary 7 *(Third Discrepancy Formula) For any generalized least-squares regression*

$$b - b_{OLS} = \lambda\kappa\frac{\sigma_y}{\sigma_x}. \tag{13}$$

In words: the discrepancy between the generalized least-squares slope and the ordinary least-squares slope is given by a weighted scatter term. Equivalently, the generalized least-squares slope is the sum of the ordinary least-squares slope plus a weighted scatter term

The functions $\lambda = \lambda(\gamma)$ and $\gamma = \gamma(\lambda)$ are both increasing over $[0, 1]$. When $\gamma = 0$ then $\lambda = 0$ and the line is the OLS $y|x$ line. When $\gamma = 1$ then $\lambda = 1$ and the line is the extremal line. The graph of λ as a function of γ is displayed, allowing one to visually estimate the normalized exponential parameter corresponding to a particular λ value and vice versa. It is clear from the graph that λ is always less than or equal to γ with equality only at the endpoints.



In the next table, quarter values and the corresponding exact values for γ and λ are calculated here as a reference. These reference values are used for comparison purposes in the numerical examples below.

γ	0	$\frac{1}{4}$	$\frac{8}{17}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{4}{5}$	$\frac{24}{25}$	1
λ	0	$4-\sqrt{15}$	$\frac{1}{4}$	$2-\sqrt{3}$	$\frac{4-\sqrt{17}}{3}$	$\frac{1}{2}$	$\frac{3}{4}$	1

In those cases where an explicit formula for b was derived one can substitute and write specific formulas for γ and λ pertaining to that regression. This is done now for geometric mean regression and OLS $x|y$ regression since their γ and λ parameters are compact expressions in ρ and κ.

	OLS y | x	GMR	OLS x | y ($\rho^2 \geq \frac{1}{2}$)	Extremal
γ	0	$\lvert\kappa\rvert$	$2\kappa\rho$	1
λ	0	$\frac{\lvert\kappa\rvert}{1+\lvert\rho\rvert}$	$\frac{\kappa}{\rho}$	1

ISBN: 978-960-474-366-7

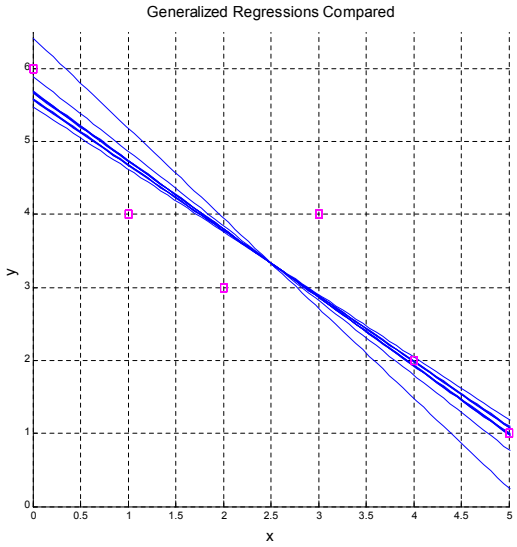
In all cases, including those cases where the slope was a solution to a cubic or higher-degree polynomial equation, the effective exponential parameter γ can be found numerically once the slope is known. This is done in the examples in the next section.

3 Numerical Examples

This section revisits the examples explored in the previous work [2] taking into account the exponential equivalence theorem. In addition to placing symmetric, hybrid symmetric and exponential regressions on the same tables, the corresponding numerical values for γ, λ are now computed for each regression line.

Example 1 Six data values are given: $(0, 6)$, $(1, 4)$, $(2, 3)$, $(3, 4)$, $(4, 2)$, and $(5, 1)$. The reader can verify that $\rho = -0.9157$, $\kappa = -0.4019$, $\mu_x = 2.5000$, $\mu_y = 3.3333$, $\sigma_x = 1.7078$, and $\sigma_y = 1.5986$. The effective exponential parameter γ and the corresponding scatter parameter λ are computed along with the equation of each line. The reader can verify in each case that $b = (\rho + \lambda\kappa)\,\frac{\sigma_y}{\sigma_x}$ and $a = \mu_y - b\mu_x$.

The graph shows the generalized regression lines together with the extremal line thereby displaying the region containing all possible generalized regression lines.



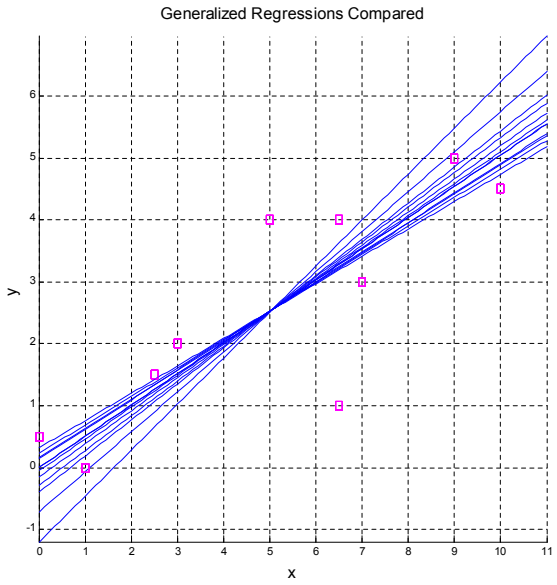
Additional exponential regressions with $\gamma = \frac{1}{4}, \frac{1}{2}$ and $\frac{3}{4}$ are included in the table. The exponential regressions with $\gamma = \frac{8}{17}, \frac{4}{5}$ and $\frac{24}{25}$ corresponding to $\lambda = \frac{1}{4}, \frac{1}{2}$ and $\frac{3}{4}$ respectively are also included here for comparison purposes. They are not shown in the graph.

Regression Type	$y = a + bx$	Exponential Parameter γ	Scatter Parameter λ
Extremal	$y = 6.4166 - 1.2333x$	1.0000	1.0000
Exponential	$y = 6.1815 - 1.1393x$	0.9600	0.7500
Exponential	$y = 5.9464 - 1.0452x$	0.8000	0.5000
Exponential	$y = 5.9007 - 1.0269x$	0.7500	0.4514
OLS $x \mid y$	$y = 5.8889 - 1.0222x$	0.7360	0.4389
Exponential	$y = 5.7282 - 0.9579x$	0.5000	0.2680
Exponential	$y = 5.7113 - 0.9512x$	0.4706	0.2500
Pythagorean	$y = 5.6855 - 0.9409x$	0.4241	0.2226
Least Perimeter Squared	$y = 5.6797 - 0.9386x$	0.4135	0.2164
GMR	$y = 5.6735 - 0.9361x$	0.4019	0.2098
Squared Harmonic Mean	$y = 5.6667 - 0.9333x$	0.3891	0.2026
Orthogonal	$y = 5.6593 - 0.9304x$	0.3752	0.1947
Exponential	$y = 5.5956 - 0.9049x$	0.2500	0.1270
Hybrid Pythagorean	$y = 5.5860 - 0.9011x$	0.2304	0.1168
Hybrid Least Perimeter	$y = 5.5811 - 0.8991x$	0.2203	0.1115
Hybrid Harmonic Mean	$y = 5.5705 - 0.8949x$	0.1985	0.1003
Hybrid Orthogonal	$y = 5.5648 - 0.8926x$	0.1869	0.0943
OLS $y \mid x$	$y = 5.4762 - 0.8571x$	0.0000	0.0000

In the last column, λ is again the fraction of the scatter that regression method contributes to the slope. It is seen, for example, that OLS $x|y$ contributes approximately 44% of the scatter to the slope whereas GMR contributes 21% and orthogonal regression contributes 19%. In this example all the non-exponential methods contribute less than 50% of the scatter to the slope.

Example 2 Ten data values are given: $(0, 0.5)$, $(1, 0)$, $(2.5, 1.5)$, $(3, 2)$, $(5, 4)$, $(6.5, 4)$, $(6.5, 1)$, $(7, 3)$, $(9, 5)$, and $(10, 4.5)$. The reader can verify that $\rho = 0.8268$, $\kappa = 0.5625$, $\mu_x = 5.0500$, $\mu_y = 2.5500$, $\sigma_x = 3.1737$, and $\sigma_y = 1.6948$. The effective exponential parameter γ and the corresponding scatter parameter λ are again computed along with the equation of each line. The reader can verify in each case that $b = (\rho + \lambda\kappa)\,\frac{\sigma_y}{\sigma_x}$ and $a = \mu_y - b\mu_x$. Again, in the graph, the generalized regression lines are plotted together with the extremal line thereby displaying the region containing all possible generalized regression lines.

ISBN: 978-960-474-366-7



Additional exponential regressions are included in the table for comparison purposes.

Regression Type	$y = a + bx$	Exponential Parameter γ	Scatter Parameter λ
Extremal	$y = -1.1966 + 0.7419x$	1.0000	1.0000
Exponential	$y = -0.8174 + 0.6668x$	0.9600	0.7500
OLS $x \mid y$	$y = -0.7116 + 0.6459x$	0.9301	0.6803
Exponential	$y = -0.4382 + 0.5917x$	0.8000	0.5000
Pythagorean	$y = -0.3924 + 0.5827x$	0.7697	0.4698
Exponential	$y = -0.3645 + 0.5771x$	0.7500	0.4514
Least Perimeter Squared	$y = -0.2824 + 0.5609x$	0.6862	0.3972
GMR	$y = -0.1468 + 0.5340x$	0.5625	0.3079
Exponential	$y = -0.0863 + 0.5220x$	0.5000	0.2680
Exponential	$y = -0.0590 + 0.5166x$	0.4706	0.2500
Hybrid Pythagorean	$y = -0.0504 + 0.5149x$	0.4611	0.2443
Squared Harmonic Mean	$y = 0.0041 + 0.5041x$	0.3994	0.2084
Hybrid Least Perimeter	$y = 0.0065 + 0.5037x$	0.3966	0.2068
Exponential	$y = 0.1275 + 0.4797x$	0.2500	0.1270
Orthogonal	$y = 0.1406 + 0.4771x$	0.2335	0.1184
Hybrid Harmonic Mean	$y = 0.1638 + 0.4725x$	0.2041	0.1031
Hybrid Orthogonal	$y = 0.2336 + 0.4587x$	0.1138	0.0571
OLS $y \mid x$	$y = 0.3202 + 0.4416x$	0.0000	0.0000

Here OLS $x|y$ contributes approximately 68% of the scatter to the slope whereas GMR contributes 31% and orthogonal regression contributes 12%. In this example all the non-exponential methods contribute less than 75% of the scatter to the slope.

4 Summary

Two fundamental results on generalized least-squares regression have been presented. The first is the exponential equivalence theorem, which states that any generalized least-squares regression line with arbitrary weight function $g(b)$ can be generated by an equivalent exponentially weighted least-squares problem. It follows that every generalized least-squares line can be assigned an effective exponential parameter γ which classifies it on the spectrum between the ordinary least-squares line $(\gamma = 0)$ and the extremal line $(\gamma = 1)$.

The second result is a fundamental formula for the generalized least-squares slope: $b = (\rho + \lambda\kappa)\frac{\sigma_y}{\sigma_x}$ where $0 \leq \lambda \leq 1$ and $\lambda = \frac{\gamma}{1+\sqrt{1-\gamma^2}}$. The formula explains what all generalized least-squares methods do and how they differ from one another. Every generalized least-squares method incorporates the scatter of the data κ into the slope. The degree to which it does this is measured by the parameter λ, again classifying every generalized regression line on the spectrum between the ordinary least-squares line and the extremal line.

References

[1] S. C. Ehrenberg, *Deriving the Least-Squares Regression Equation*, The American Statistician, Vol. 37, No. 3 (Aug. 1983), p.232.

[2] N. Greene, *Generalized Least-Squares Regressions I: Efficient Derivations*, in: Recent Advances in Intelligent Control, Modelling and Computational Science, Proceedings of the 1st International Conference on Computational Science and Engineering (CSE'13), Valencia, Spain, August 6-8, 2013, pp. 145-158.

[3] N. Greene, *Generalized Least-Squares Regressions II: Theory and Classification*, in: Recent Advances in Intelligent Control, Modelling and Computational Science, Proceedings of the 1st International Conference on Computational Science and Engineering (CSE'13), Valencia, Spain, August 6-8, 2013, pp. 159-166.[1]

[4] R. Taagepera, *Making Social Sciences More Scientific: The Need for Predictive Models*, Oxford University Press, New York, 2008.

[1] In Part II of this series [3], page 159, Definition 1, Part (iv) should read as follows: Non-decreasing in x and y for $x \geq 0$ and $y \geq 0$. For ψ differentiable this means $\psi_x \geq 0$ and $\psi_y \geq 0$ when $x \geq 0$ and $y \geq 0$. On page 162, the second line of the proof of Theorem 11 should read $\frac{d}{db}\ln g(b)$. In the chart on page 165, Row 4 should read $g(b) = \sqrt{1 + 1/b^2}$.