Relationship between Sums of Squares in Linear Regression and Semi-parametric Regression

Dursun Aydin, and Bilgin Senel

Abstract—In this paper, the sum of squares in linear regression is reduced to sum of squares in semi-parametric regression. We indicated that different sums of squares in the linear regression are similar to various deviance statements in semi-parametric regression. In addition to, coefficient of the determination derived in linear regression model is easily generalized to coefficient of the determination of the semi-parametric regression model. Then, it is made an application in order to support the theory of the linear regression and semi-parametric regression. In this way, study is supported with a simulated data example.

Keywords—Semi-parametric regression, Penalized Least Squares, Residuals, Deviance, Smoothing Spline.

I. INTRODUCTION

REGRESSION analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable $y = \{y_1, y_2, ..., y_n\}^T$ and independent variables $z_1, z_2, ..., z_k$. Generally, regression models can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships [1]; [2]. It is frequently encountered to these models in many application areas. Most used models can be given in the following way:

**Linear regression model (LRM):** Linear regression model attempts to model the relationship among a dependent variable and $k$ explanatory variables. LRM is given as following:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j z_{ij} + \epsilon_i, i = 1, 2, ..., n$$

(1)

where $\beta = [\beta_0, \beta_1, ..., \beta_k]$ is a vector of unknown regression coefficients and $\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_n]^T$ is a vector of random errors, assumed to follow normal distribution with zero mean and constant variance $\sigma^2$.

**Generalized linear regression model (GLRM):** Generalized linear models extend the concept of the widely used linear regression model. GLRM is assumed to have the form:

$$g(y_i) = \beta_0 + \sum_{j=1}^{k} \beta_j z_{ij} + \epsilon_i, i = 1, 2, ..., n$$

(2)

where $g(.)$ is called a link function, and $\epsilon$ is a vector of random error with a suit distribution.

**Semi-parametric regression model (SPRM):** A semi-parametric regression model (SPRM) consists of two additive components, a linear parametric and a nonparametric part:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j (z_{ij}) + f(x_i) + \epsilon_i, i = 1, 2, ..., n$$

(3)

where $\beta$ is a vector of finite dimensional parameter (or the vector of unknown regression coefficients), and $f(.)$ is a smooth function of explanatory variable $x$, and $\epsilon$ is denote an error term with zero mean and common variance $\sigma^2$.

**Generalized semi-parametric regression model (GSPRM):** Introducing a link $g(.)$ for a semi-parametric model in (3) yields the generalized semi-parametric regression model:

$$g(y_i) = \beta_0 + \sum_{j=1}^{k} \beta_j (z_{ij}) + f(x_i) + \epsilon_i, i = 1, 2, ..., n$$

(4)

g denotes a known link function as in generalized additive model, and $\epsilon$ is a vector of random error with a suit distribution, and with zero mean and common variance $\sigma^2$. In the case of an identity link function $g$ given in Eq. (4), GSPRM reduces to SPRM. [3]

In the section II, least square estimation of the linear regression model and analysis of variability in response are discussed. Section III reviews smoothing spline estimation of the semi-parametric regression model. Section IV discusses an application on simulated data set, while conclusions and discussion are offered in the section V.

II. LEAST SQUARES ESTIMATION OF THE LRM

One important goal of a regression analysis is to estimate the vector of unknown regression coefficients in model Eq. (1). The method of least squares is used more extensively than any other estimation procedure for building regression models. The method of least squares is designed to provide estimator $\hat{\beta}$ of the $\beta$ in Eq (1). Not that there are $p = k + 1$ regression coefficients. (1). It is suitable at this point to reintroduce the model Eq. (1) in matrix notation. The model can be written as:

$$y = Z\beta + \epsilon$$

(5)
In general, \( y \) is a \((n \times 1)\) vector of the observations, \( Z \) is an \((p \times 1)\) matrix of the levels of the independent variables, \( \beta \) is a \((p \times 1)\) vector of the regression coefficients, and \( \varepsilon \) is an \((n \times 1)\) vector of the random errors.

In the method of least squares, we wish to find the vector of least squares estimators, \( \hat{\beta} \), minimize the sum of squares of the residuals: \( \sum_{i=1}^{n} \varepsilon_i^2 = (y - Z\beta)^T (y - Z\beta) \). The least squares estimators that provide this minimum, defined as follows:

\[
\hat{\beta} = (Z^T Z)^{-1} Z^T y
\]  

A. Analysis of Variability in the Response

The fitted values and the residuals in Eq. (5) are defined as \( \hat{y} = Z\hat{\beta} \) and \( e = y - \hat{y} \) respectively. In any regression problem, it will be observed that variation in response variable. Of course, it is wanted that fitted values follow the real values closely. It is natural to consider the sources of variation, the total sum of squares, and the regression sum of squares:

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]  

Thus, as indicated in Equations (7), the total sum of squares \( S_{S_y} \) is partitioned into a regression sum of squares \( SS_{R} \) and a residual sum of squares \( SS_{Res} \):

\[
SS_{y} = SS_{R} + SS_{Res}
\]

It can be arranged analysis of variance (ANOVA) table used for testing the significant of the model in Eq. (1) via these important sums of squares. ANOVA is defined as Table I.

### Table I

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of Freedom (DF)</th>
<th>Sum of Squares</th>
<th>Mean Square (MS)</th>
<th>F-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( k )</td>
<td>( SS_{R} = \hat{\beta}^T Z^T y - n\bar{y}^2 )</td>
<td>( MS_{R} = SS_{R} / k - 1 )</td>
<td>( MS_{R} )</td>
</tr>
<tr>
<td>Residual</td>
<td>( n - k - 1 )</td>
<td>( SS_{Res} = y^T y - \hat{\beta}^T Z^T y )</td>
<td>( MS_{Res} = SS_{Res} / n - k - 1 )</td>
<td>( MS_{Res} )</td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td>( SS_{y} = y^T y - n\bar{y}^2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here \( F \)-statistic may be viewed as ratio that states variance explained by the model divided by variance due to model error. As a result, large values of \( F \)-statistic are state the signification of model. The coefficient of determination denoted as \( R^2 \) is represent the proportion of variation in the response data that is explained by model. \( R^2 \) is denoted as

\[
R^2 = \frac{SS_{R}}{SS_{y}} = 1 - \frac{SS_{Res}}{SS_{y}}
\]  

Another way to represent the proportion of variation in the response is adjusted \( R^2 \), denoted as \( R_{adj}^2 \). Some analyst prefer to use an adjusted \( R^2 \) statistic, defined as:

\[
R_{adj}^2 = 1 - \frac{SS_{Res}}{MS_{Res} / (DF_{Res})} = 1 - \frac{SS_{Res}}{SS_{y} / (DF_{y})}
\]  

III. SMOOTHING SPLINE ESTIMATION OF THE SPRM

We consider the estimation of the SPRM in (3). In the matrix notation, Eq. (3) can be written as following way:

\[
y = Z\beta + f + \varepsilon \quad (10)
\]

where \( Z \) is the \((n \times n)\) matrix, \( \beta = (\beta_1, ..., \beta_p)^T \), \( y = (y_1, ..., y_n)^T \), \( f = (f(x_1), ..., f(x_n))^T \), and \( \varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T \).

Estimation of the parameters of interest in equation (10) can be performed using smoothing spline. Mentioned here the vector parameter \( \beta \) and the values of function \( f \) at sample points \( x_1, x_2, ..., x_n \) are estimated by minimizing the penalized residual sum of squares:

\[
PSS(\beta, f) = \sum_{i=1}^{n} (y_i - z_i^T \beta - f(x_i))^2 + \lambda \int_{0}^{1} (f'(x))^2 \, dx \quad (11)
\]

Here, \( f \in C^2[0,1] \) and \( z_i \) is the \( ith \) row of the matrix \( Z \). When the \( \beta = 0 \), resulting estimator has the form \( \hat{f} = (\hat{f}(x_1), ..., \hat{f}(x_n))^T = S_p y \), where \( S_p \) a known positive-definite smoother matrix that depends on \( \lambda \) called as smoothing parameter, and the knots \( x_1, ..., x_n \) (see, [4];[5];[6];[7]).

For a pre-specified value of \( \lambda \) the corresponding estimators for \( f \) and \( \beta \) based on Eq. (11) can be obtained as follows [4]: Given a smoother matrix \( S_p \), depending on a smoothing parameter \( \lambda \), construct \( \hat{Z} = (I - S_p)Z \). Then, by using penalized least squares, mentioned here estimator are given by:

\[
\hat{\beta} = (Z^T \hat{Z})^{-1} \hat{Z}^T y
\]  

\( \hat{Z} \) is a \((m \times 1)\) vector of the observations, \( Z \) is an \((p \times 1)\) matrix of the levels of the independent variables, \( \beta \) is a \((p \times 1)\) vector of the regression coefficients, and \( \varepsilon \) is an \((n \times 1)\) vector of the random errors.
\[ \hat{f} = S_n \left( y - Z \hat{\beta} \right) \]  

**A. Relationship between Deviance and Sum of Squares**

The deviance plays the role of the residual sum of squares for generalized models, and can be used for assessing goodness of fit and comparing models. The deviance or likelihood ratio statistic of a fitted model is defined as

\[ D = 2 \left[ l(\hat{\beta}_{\text{max}}) - l(\hat{\beta}) \right] \Phi \]  

(14)

Where \( l(\hat{\beta}_{\text{max}}) \) denotes the maximized likelihood of the saturated model that have one parameter per data point, \( \hat{\beta}_{\text{max}} \) is parameter value of \( \beta \) which maximizes \( l(\beta) \), and \( l(\hat{\beta}) \) is a log-likelihood function of a sample \( n \) observation (i.e.,

\[ l(\hat{\beta}) = \sum_{i=1}^{n} \log f(y_i) \], and \( \Phi \) is a dispersion parameter [8]; [9].

In the Gaussian family of distributions (for example, in SPRM), \( \Phi \) is just standard variance \( \sigma^2 \) and the **residual deviance** reduces to the **residual sum of squares**. The **residual deviance** is the deviance of fitted model, while the deviance for a model which includes the offset and possible an intercept term is called as **null deviance**. In this case, the **null deviance** reduces to the **total sum of squares**. Then, analogously to the equations (7), regression deviance for SPRM is defined as

\[ \text{Regression Dev.} = \text{Null Dev.} - \text{Res. Dev.} \]  

(15)

These can be combined to give the **proportion deviance explained**, a generalization of the \( R^2 \) value given in Eq. (8), as following way:

\[ R^2_{\text{SPRM}} = \frac{\text{Regression Deviance}}{\text{Null Deviance}} = \frac{(\text{Null Deviance} - \text{Residual Deviance})}{(\text{Null Deviance})} \]  

(16)

Similarly, we can generalize adjusted coefficient of determination given in Eq. (9), as follow:

\[ R^2_{\text{Adj.-SPRM}} = \frac{\text{(Mean Null Dev.) - (Mean Res. Dev.)}}{\text{(Mean Null Dev.)}} \frac{\text{Res. Dev.}}{\text{DF Res. Dev.}} = \frac{\text{Res. Dev.}}{\text{DF Res. Dev.}} \]  

(17)

For assessment of the SPRM, it is necessary to perform test on both the parametric and the nonparametric component. For the parametric component of the SPRM, we can generalize such as **F-statistic** given Table I. The **F-statistic** can be defined as:

\[ F_{\text{Par}} = \frac{\left( \text{Regression Deviance} \right)}{\left( \text{Residual Deviance} \right)} \]  

\[ F_{\text{Nosp}} = \frac{\left( \text{Null Dev.} \right) - \text{Residual Deviance}}{\left( \text{DF Residual Deviance} \right)} \]  

(18)

(19)

By considering the deviances in SPRM and residual sum of squares in LRM, it can be performed by an approximate **F-statistic** whether the nonparametric component of model is linear or whether SPRM provides a significantly better fit. The test is based on the differences of residual deviances and residual sum of squares for SPRM and LRM respectively. The **F-statistic** can be given by

\[ F = \frac{\left( \text{SS}_{\text{null}} - \text{Residual Deviance} \right)}{\left( \text{DFSS}_{\text{null}} - \text{DF Residual Deviance} \right)} \]  

(19)

**IV. HELPFUL HINTS**

A semi-parametric regression model is basically a multiple linear regression model in which some of the linear predictors are replaced with additive smooth functions. It is used that **S-plus** and **R** programs based on penalized least square to estimate the semi-parametric regression model. These programs use “**gam package**” for estimation [10]. To estimate unknown functions \( f \), **S-plus** and **R** programs use mainly smoothing splines denoted by \( s(.) \). It is considered here only smoothing spline. The **gam package** provides model fitting for different family types (Normal, Poisson, Binomial, Gamma and inverse Gaussian) with the suitable link functions. Here it is only used identity link function. Analogously to analysis of variance table which provides summary statistics in an ordinary regression analysis, the **gam package** provides an analysis of deviance table. A simple simulated data set used to analysis the relation between sums of squares in linear regression and the deviances obtained via the SPRM. The variables related with data are defined as follows:

- **y** is a numeric vector with sized \( n = 100 \) that made by random
- **z** is a numeric vector with sized \( n = 100 \) that made by random
- **x** is a numeric vector with sized \( n = 100 \) that made by random

**A. Empirical Results**

According to the variables in above, the SPRM in **gam package** is appeared as follows:

**Call**: gam(formula = \( y \sim s(x) + z \), data = gam.data)

Deviance Residuals:
Min 1Q Median 3Q Max
-0.681 -0.214 0.029 0.245 0.531

(Dispersion Parameter for gaussian family taken to be 0.0841)

Null Deviance: 57.7496 on 99 degrees of freedom

A partial linear additive model relates $y$ called as response or dependent variable to the independents variables given in previous section. As shown Table II, the parametric coefficients of the SPRM appear, while nonparametric coefficient doesn’t appear. It can be only displayed graphically because it can’t be expressed as parametric.

Fig. 1 shows the estimates (solid) and the 95% confidence intervals (dashed) for SPRM using smoothing spline. The plotted curve is a contribution of a term to the additive predictor. The effects of $x$ called as noise predictor is very strong on the response variable. Firstly, as $x$ is increasing, $y$ is increasing too. Then, as $x$ is again increasing, $y$ is decreasing.

B. Comparison of the Performances of the LRM and SPRM

To compare performances of the SPRM and LRM, it is performed an analysis of deviance table by using formula given in section 3.A. In summary, these results are given in the Table V. The residual deviance (9.9077) in Table V is smaller than residual sum of squares (19.387) in Table IV. Similarly both coefficient of determination and adjusted coefficient of determination given in the Table V are bigger than those of the Table IV. It can be said that SMPR provides a better fit than LRM. However, the difference between the adjusted coefficients of determination for SPRM and LRM are smaller than the difference between non-adjusted coefficients of determination. Thus, it can be said that adjusted coefficients of determination are more realistic in assessing the overall model performance. As shown Table V, it can be said that all of parametric coefficients are also
### Table III: Coefficients of Linear Regression

|         | Estimate | Std. Error | t value | Pr(>|t|) |
|---------|----------|------------|---------|----------|
| (Constant) | 1.9944   | 0.1325     | 15.047  | 2e-16    |
| x       | -2.3278  | 0.1680     | -13.854 | 2e-16    |
| z       | -0.1460  | 0.1672     | -0.873  | 0.385    |

### Table IV: Analysis of Variance Table for LRM

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>DF</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>38.362</td>
<td>19.181</td>
</tr>
<tr>
<td>Residual</td>
<td>97</td>
<td>19.387</td>
<td>0.200</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>57.749</td>
<td>0.583</td>
</tr>
</tbody>
</table>

F-stat: 95.905  
**p-value:** < 2.2e-16

### Table V: Analysis of Deviance Table for SPRM

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>DF</th>
<th>Deviance</th>
<th>Mean Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>5</td>
<td>49.8419</td>
<td>9.96982</td>
</tr>
<tr>
<td>Residual</td>
<td>94</td>
<td>7.9077</td>
<td>0.08412</td>
</tr>
<tr>
<td>Null</td>
<td>99</td>
<td>57.7496</td>
<td>0.5833</td>
</tr>
</tbody>
</table>

F-stat (Parametric) = 118.519  
F-stat (Nonparametric) = 45.485

### Table VI: Analysis of Variance Table

| Model | Res. Df | Res.Sum Sq | DF  | Sum Sq | F          | Pr(>|F|) |
|-------|---------|------------|-----|---------|------------|---------|
| LRM   | 97      | 19.387     | 3   | 11.4794 | 48.485     | 2.2e-16 |
| PLAM  | 94      | 7.9077     | 3   | 48.485  | 2.2e-16    |         |

**IV. Conclusion and Discussion**

In the Gaussian family of distributions, we have demonstrated that the residual deviance can be easily reduced to the residual sum of squares. Besides, it is shown that the null deviance can be also reduced to the total sum of squares.

Furthermore, coefficient of determination and adjusted coefficient of determination play quite important role in assessing the goodness of fit of the regression models. We have indicated that these coefficients obtained by using LRM can be easily generalized to SPRM. Especially, adjusted coefficient of determination in SPRM is very proper for assessment of the model goodness of fit because it detects the degree of complexity of the SPRM.

**REFERENCES**