International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:2, No:6, 2008

A Comparison of the Sum of Squares in Linear and Partial Linear Regression Models

Dursun Aydın

***Abstract*—**In this paper, estimation of the linear regression model is made by ordinary least squares method and the partially linear regression model is estimated by penalized least squares method using smoothing spline. Then, it is investigated that differences and similarity in the sum of squares related for linear regression and partial linear regression models (semi-parametric regression models). It is denoted that the sum of squares in linear regression is reduced to sum of squares in partial linear regression models. Furthermore, we indicated that various sums of squares in the linear regression are similar to different deviance statements in partial linear regression. In addition to, coefficient of the determination derived in linear regression model is easily generalized to coefficient of the determination of the partial linear regression model. For this aim, it is made two different applications. A simulated and a real data set are considered to prove the claim mentioned here. In this way, this study is supported with a simulation and a real data example.

***Keywords*—**Partial Linear Regression Model, Linear Regression Model, 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 $\mathbf{y} = \{y_1, y_2, ..., y_n\}^T$ 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 encounter to these models in many application areas. Most used models can be given in the following way:

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

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

where $\boldsymbol{\beta} = \{\beta_0, \beta_1, ..., \beta_k\}$ is a vector of unknown regression coefficients and $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\}^T$ is a vector of random errors, assumed to follow normal distributed with zero mean and constant variance σ^2.

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

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

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

Partial linear regression model (PLRM): A partial linear regression model is consists of two additive components, a linear parametric and a nonparametric part:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j (z_{ij}) + f(x_i) + \varepsilon_i, i = 1, 2, ..., n \quad (3)$$

where $\boldsymbol{\beta}$ 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 $\boldsymbol{\varepsilon}$ is denote an error term with zero mean and common variance σ^2.

Generalized partial linear regression model (GPLRM): Introducing a link $g(.)$ for a partial linear model in (3) yields the generalized partial linear regression model:

$$g(y_i) = \beta_0 + \sum_{j=1}^{k} \beta_j (z_{ij}) + f(x_i) + \varepsilon_i, i = 1, 2, ..., n \quad (4)$$

g denotes a known link function as in GLRM, and $\boldsymbol{\varepsilon}$ is a vector of random error with a suit distribution, and with zero mean and common variance σ^2. In the case of an identity link function g given in Eq. (4), GPLRM reduces to PLRM. [3]

In the section 2, least square estimation of the linear regression model and analysis of variability in response are discussed. Section 3 reviews smoothing spline estimation of the partial linear regression model. Section 4 discusses an application on simulated data set, while conclusions and discussion are offered in the section 5.

II.LEAST SQUARES ESTIMATION OF THE LINEAR REGRESSION MODEL

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 $\hat{\boldsymbol{\beta}}$ of the $\boldsymbol{\beta}$ in Eq. 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

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (5)$$

In general, $\mathbf{y}$ is a $(n \times 1)$ vector of the observations, $\mathbf{Z}$ is an $(p \times 1)$ matrix of the levels of the independent variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ 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, $\hat{\boldsymbol{\beta}}$, minimize the sum of squares of the residuals: $\sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})$. The least squares estimators that provide this minimum, defined as follows:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{Z}^T\mathbf{y} \qquad (6)$$

The fitted values and the residuals in Eq. (5) are defined as $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}}$ 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:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (7)$$

Thus, as indicated in Equations (7), the total sum of squares $\mathbf{SS_T}$ is partitioned into a regression sum of squares $\mathbf{SS_R}$ and a residual sum of squares $\mathbf{SS_{Res}}$:

$$\mathbf{SS_T} = \mathbf{SS_R} + \mathbf{SS_{Res}}$$

It can be arranged a variance of analysis (ANOVA) table used for testing the significant of the model in Eq. (1) via these important sums of squares. ANOVA is defined as Table1.

A. Analysis of Variability in the Response

TABLE I ANALYSIS OF VARIANCE

Source of Variation	**Degrees of Freedom** (DF)	**Sum of Squares** (SS)	**Mean Square** (MS)	F - *statistic*
Regression	k	$\mathbf{SS_R} = \hat{\boldsymbol{\beta}}^T\mathbf{Z}^T\mathbf{y} - n\bar{\mathbf{y}}^2$	$\mathbf{MS_R} = \mathbf{SS_R} / k - 1$	$\dfrac{\mathbf{MS_R}}{\mathbf{MS_{Res}}}$
Residual	$n - k - 1$	$\mathbf{SS_{Res}} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{Z}^T\mathbf{y}$	$\mathbf{MS_{Res}} = \mathbf{SS_{Res}} / n - k - 1$	
Total	$n - 1$	$\mathbf{SS_T} = \mathbf{y}^T\mathbf{y} - n\bar{\mathbf{y}}^2$		

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

$$R^2 = \frac{\mathbf{SS_R}}{\mathbf{SS_T}} = 1 - \frac{\mathbf{SS_{Res}}}{\mathbf{SS_T}} \qquad (8)$$

Another way to represent the proportion of variation in the response is adjusted R^2, denoted as $R^2_{Adj.}$ Some analyst prefer to use an adjusted R^2 statistic, defined as

$$R^2_{Adj.} = 1 - \frac{\mathbf{MS_{Res}}}{\mathbf{MS_T}} = 1 - \frac{\mathbf{SS_{Res}} / (\mathbf{DF_{Res}})}{\mathbf{SS_T} / (\mathbf{DF_T})}. \qquad (9)$$

III. SMOOTHING SPLINE ESTIMATION OF THE PARTIL LINEAR REGRESSION MODEL

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

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \qquad (10)$$

where $\mathbf{Z}$ is the $(n \times n)$ matrix of the predictors z_i, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$.

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

$$PSS\,(\boldsymbol{\beta}, \mathbf{f}) = \sum_{i=1}^{n}\left\{y_i - z_i^T\boldsymbol{\beta} - f(x_i)\right\}^2 + \lambda \int_0^1 \left(f^{(m)}(x)\right)^2 dx \qquad (11)$$

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

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

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}^T\tilde{\mathbf{Z}}\right)^{-1} \tilde{\mathbf{Z}}^T\,\mathbf{y} \qquad (12)$$

$$\hat{\mathbf{f}} = S_\lambda\left(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}\right) \qquad (13)$$

A. Generalization of the Sum of Squares to Different Deviances

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

$$D = 2\left\{ l(\hat{\beta}_{\max}) - l(\hat{\beta}) \right\}\Phi \qquad (14)$$

Where $l(\hat{\beta}_{\max})$ denotes the maximized likelihood of the saturated model that have one parameter per data point. $\hat{\beta}_{\max}$ is parameter value of β which maximizes $l(\beta)$, and $l(\beta)$ is a log-likelihood function of a sample n observation (i.e., $l(\hat{\beta}) = \sum_{i=1}^{n} \log f(y_i)$), and Φ is a dispersion parameter [8]; [9].

In the Gaussian family of distributions (for example, in PLRM), Φ is just standard variance σ^2 and ***the residual deviance in PLRM*** reduces to ***the residual sum of squares in LRM***. ***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 in PLRM*** reduces to the ***total sum of squares in LRM***. Then, analogously to the equations (7), regression deviance for PLRM is defined as

$$\text{Regression Dev.} = \text{Null Dev.} - \text{Res. Dev.} \qquad (15)$$

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

$$\begin{aligned} R^2_{PLRM} &= \frac{\text{Regresion Deviance}}{\text{Null Deviance}} \\ &= \frac{(\text{Null Deviance - Residual Deviance})}{(\text{Null Deviance})} \end{aligned} \qquad (16)$$

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

$$\begin{aligned} R^2_{Adj-PLRM} &= \frac{(\text{Mean Null Dev. - Mean Res. Dev.})}{(\text{Mean Null Dev.})} \\ &= \frac{\left(\frac{\text{Null Dev.}}{\text{DF Null Dev.}}\right) - \left(\frac{\text{Res. Dev.}}{\text{DF Res. Dev.}}\right)}{\left(\frac{\text{Null Dev.}}{\text{DF Null Dev.}}\right)} \end{aligned} \qquad (17)$$

For assessment of the PLRM, it is necessary to perform test on both the parametric and the nonparametric component. For the parametric component of the PLRM, we can generalize such as $F-statistic$ given Table 1. The $F-statistic$ can be defined as:

$$\begin{aligned} \mathbf{F}_{Par.} &= \frac{\frac{(\text{Regression Deviance})}{(\text{DF Regression Deviance})}}{\frac{(\text{Residual Deviance})}{(\text{DF Residual Deviance})}} \\ &= \frac{(\text{Mean Regression Deviance})}{(\text{Mean Residual Deviance})} \end{aligned} \qquad (18)$$

By considering the deviances in PLRM 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 PLRM provides a significantly better fit. The test is based on the differences of residual deviances and residual sum of squares for PLRM and LRM respectively. The $\boldsymbol{F}$ *- statistic* can be given by

$$\mathbf{F}_{Nonp.} = \frac{\frac{(\text{SS}_{\text{Res}}\text{ - Residual Deviance})}{\text{DF SS}_{\text{Res}}\text{ - DF Residual Deviance}}}{\frac{(\text{Residual Deviance})}{(\text{DF Residual Deviance})}} \qquad (19)$$

IV. APPLICATIONS

A partial linear 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 partial linear 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. Simulated Data Example

A simple simulated data set used to analysis the relation between sums of squares in linear regression and the deviances obtained via the PLRM. The variables related with data are defined as fallows:

y *is a numeric vector with sized* $n = 100$ *that made by random - the response*
z *is a numeric vector with sized* $n = 100$ *that made by random – predictor*
x *is a numeric vector with sized* $n = 100$ *that made by random - noise predictor.*

Empirical Results for Simulated Data

A partial linear additive model relates y called as response or dependent variable to the independents variables given in previous section. As shown Table 2, the parametric coefficients of the PLRM appear, while nonparametric coefficient doesn't appear. It can be only displayed graphically because it can't be expressed as parametric.

Figure 1 shows the estimates (solid) and the 95% confidence intervals (dashed) for PLRM 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.

According to the simulated data set, the PLRM in ***gam package*** is appeared as follows:

`Call: gam(formula = y ~ s(x) + z, data = gam.data)`

(Dispersion Parameter for gaussian family taken to be 0.0841)

Null Deviance: 57.7496 on 99 degrees of freedom

Residual Deviance: 7.9077 on 94 degrees of freedom

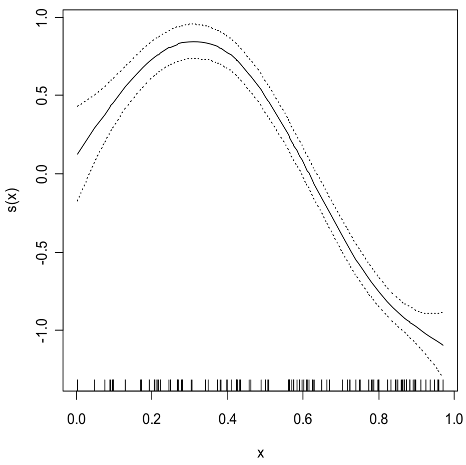

Fig. 1 Estimates (solid) and the 95 % confidence intervals (dashed) of the nonparametric components for PLRM

TABLE II DF FOR TERMS AND F-VALUES FOR NONPARAMETRIC EFFECTS AND T-VALUES FOR PARAMETRIC PART.

	Nonparametric Part				Parametric part			
Variables	Df	Npar Df	Npar F	Pr(F)	Estimate	Std.Error	t-val	Pr(>$\|t\|$)
(Const.)	1				1.987	0.087	23.05	1.85e-40
s(x)	1	3	45.485	2.2e-16				
Z	1				-0.125	0.108	-1.121	2.65e-01
	Response: y							

By using the variables in above, the LRM in ***gam package*** is appeared as follows:

Call:`lm(formula = y ~ x + z, data = gam.data)`

The summary of the results obtained by LRM is giving following in Table 3-4.

TABLE III COEFFICIENTS OF LINEAR REGRESSION

	Estimate	Std. Error	t value	Pr(>|t|)
(Constant)	1.9944	0.1325	15.047	2e-16
	-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

Source of variation	**DF**	**Sum Sq**	**Mean Sq**
Regression	2	38.362	19.181
Residual	97	19.387	0.200
Total	99	57.749	0.583
$\mathbf{R}^2$	0.664	**F-stat**: 95.905 **p-value**: < 2.2e-16	
$\mathbf{R}^2_{\mathbf{Adj}}$	0.657		

B. Real Data Example

A real data set used to analysis the relation between sums of squares in linear regression and the deviances obtained via the PLRM. For the purpose of illustration let us analyze a data set, known as the GDP for Turkey. Data related to variables used in this study consists of monthly time series which starts January, 1984 and ends December 2001, comprising

$n = 216$ observations. Mentioned here variables are defined as follow:

gdp : Gross Domestic Product (TL)
time: Data monthly from January 1984 up to December 2001

$D_{k=1}^{r-1}$: Dummy variables that denotes the effects seasonality

Empirical Results for Real Data

According to the real data set, the PLRM in ***gam package*** is appeared as follows:

Call: gam(formula = log(gdp)~ s(time1, 15)+D1+D2+D3+D4+D5+D6+D7 +D8+D9+D10+D11, data = gdp)
(Dispersion Parameter for gaussian family taken to be 0)

Null Deviance: 3.113 on 215 degrees of freedom
Residual Deviance: 0.0094 on 189 degrees of freedom

The summary of the results obtained by SPRM is given in Table 5. According to Table 5, all of the parametric coefficients of the PLRM are appear, while nonparametric coefficient doesn't appear. Because it can't be expressed as parametric, it can be only displayed graphically as before example. Figure 2 shows this graphic. The plotted curve is a contribution of a term to the additive predictor. The effect of **time** called as noise predictor is very strong on the **gdp**. Firstly, as **time** is increasing, **gdp** is increasing too. Then, as **time** is again increasing, **gdp** is decreasing

TABLE V DF FOR TERMS AND F-VALUES FOR NONPARAMETRIC EFFECTS AND T-VALUES FOR PARAMETRIC PART

	Parametric Part			
	Est.	St. Error	t value	Pr(>|t|)
(Intercept)	-17.352	1.84e-01	-94.35	6.38e-16
S(time,15)	0.020	9.23e-05	220.90	4.74e-23
D1	0.019	1.59e-03	11.711	3.61e-24
D2	-0.073	1.59e-03	-46.235	5.60e-10
D3	0.019	1.59e-03	11.711	3.61e-24
D4	-0.014	1.59e-03	-8.935	3.65e-16
D5	0.019	1.59e-03	11.711	3.61e-24
D6	-0.014	1.59e-03	-8.935	3.65e-16
D7	0.019	1.59e-03	11.711	3.61e-24
D8	0.019	1.59e-03	11.711	3.61e-24
D9	-0.014	1.59e-03	-8.935	3.65e-16
D10	0.019	1.59e-03	11.711	3.61e-24
D11	-0.014	1.59e-03	-8.935	3.65e-16
	Nonparametric Part			
	Df Npar	Df	Npar F	Pr(F)
S(time1)	1	14	766.07	2.2e-16
Response: log(gdp);				

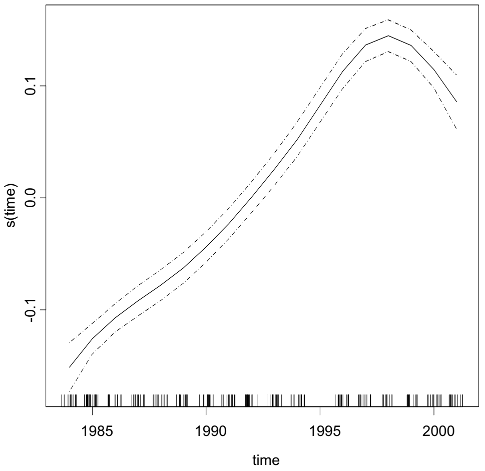

Fig. 2 Estimates (solid) and the 95 % confidence intervals (dashed) of the nonparametric components for PLRM

By using the variables in above, the LRM in ***gam package*** is appeared as follows:
Call:lm(formula=log(gdp)~(time)+D1+D2+D3+D4+D5+D6+ D7+D8+D9+D10+D11, data = gdp)

The summary of the results obtained by LRM is giving following in Table 6-7.

TABLE VI ANALYSIS OF VARIANCE TABLE FOR LRM

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.735e+01	1.349e+00	-12.868	2e-16 ***
time1	2.039e-02	6.768e-04	30.128	2e-16 ***
D1	1.860e-02	1.165e-02	1.597	0.112
D2	-7.344e-02	1.165e-02	-6.306	1.76e-09 ***
D3	1.860e-02	1.165e-02	1.597	0.112
D4	-1.419e-02	1.165e-02	-1.219	0.224
D5	1.860e-02	1.165e-02	1.597	0.112
D6	-1.419e-02	1.165e-02	-1.219	0.224
D7	1.860e-02	1.165e-02	1.597	0.112
D8	1.860e-02	1.165e-02	1.597	0.112
D9	-1.419e-02	1.165e-02	-1.219	0.224
D10	1.860e-02	1.165e-02	1.597	0.112
D11	-1.419e-02	1.165e-02	-1.219	0.224

TABLE VII ANALYSIS OF VARIANCE TABLE FOR LRM

Source of variation	DF	Sum Sq	Mean Sq
Regression	12	2.57241	0.21437
Residual	203	0.54061	0.00266
Total	215	3.11302	3.76742
R^2	0.8263	**F-stat**: 80.59 **p-value**: < 2.2e-16	
R^2_{Adj}	0.8161		

C. Relationship between Deviance and Sum of Squares

To indicate the relations between different deviance and sum of squares on simulated data set, it is performed an analysis of deviance by using formula given in section 3.3. In summary, these results are given in the Table 8. The residual deviance (7.9077) in Table 8 is smaller than residual sum of squares (19.387) in Table 4. Similarly, both coefficient of determination and adjusted coefficient of determination given in the Table 8 are bigger than those of the Table 4. It can be said that PLRM provides a better fit than LRM. However, the difference between the adjusted coefficients of determination for PLRM 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 8, it can be said that all of parametric coefficients of PLRM are also significant to $F-statistic$ (parametric) that obtain by means of the Eq. (18). Furthermore, according to the Npar-F in the Table 2, the nonparametric component is also able to test that significant or not. In addition to, it can perform an approximate $F-test$ whether the nonparametric component of model is linear or whether PLRM provides a significantly better fit. For this goal, $F-statistic$ (nonparametric) computed by using Eq.(19) is given Table 8. An equivalent computation using ***gam package*** in **S-plus** and **R** is given in Table 9. $F-statistic$ (nonparametric) derived by Eq.(19) is equivalent to F in Table 9.

TABLE VIII ANALYSIS OF DEVIANCE TABLE FOR PLRM

Source of variation	DF	Deviance	Mean Deviance
Regression	5	49.8419	9.96982
Residual	94	7.9077	0.08412
Null	99	57.7496	0.58333
R^2	0.8631	**F-stat (Parametric)** = 118.519 **F-stat (Nonparametric)** = 46.485	
R^2_{Adj}	0.8558		

TABLE IX ANALYSIS OF VARIANCE TABLE

Model	Res. Df	Res.Sum Sq	Df	Sum Sq	F	Pr(>F)
LRM	97	19.3871				
PLRM	94	7.9077	3	11.4794	45.485	2.2e-16

According to Table 9, it is said that the nonparametric function or nonparametric component of model is significant curve and provide a better fit.

Similarly to the expressions mentioned above, to indicate the generalization of the sum of squares in LRM to various deviances in PLRM, we performed an analysis of deviance for real data set. In summary, these results are given in the Table 10. The residual deviance (0.0094) in Table 10 is smaller than residual sum of squares (0.54061) in Table 7. As previous example, both coefficient of determination and adjusted coefficient of determination given in the Table 10 are bigger than those of the Table 7. Accordingly, PLRM provides a better fit than LRM. However, the difference between the adjusted coefficients of determination for PLRM and LRM are smaller than the difference between non-adjusted coefficients of determination. Therefore, adjusted coefficients of determination are more realistic in assessing the overall model performance. As shown Table 10, parametric coefficients of PLRM are significant according to $F-statistic$ (parametric). Furthermore, according to the Npar-F in the Table 5, the nonparametric component of PLRM is also significant. In addition to, according to $F-statistic$ (nonparametric) in Table 10 and its equivalent to F in Table 11, it is said that nonparametric component of PLRM is significant curve and provide a better fit than LRM.

TABLE X ANALYSIS OF DEVIANCE TABLE FOR PLRM

Source of variation	**DF**	**Deviance**	**Mean Deviance**
Regression	26	3.1036	0.11937
Residual	189	0.0094	0.0000496
Null	215	3.113	0.01448
R^2	0.9970	**F-stat (Parametric)** = 2406.65 **F-stat (Nonparametric)** = 765.12	
R^2_{Adj}	0.9966		

TABLE XI ANALYSIS OF VARIANCE TABLE

Model	**Res. Df**	**Res.Sum Sq**	**Df**	**Sum Sq**	**F**	**Pr(>F)**
LRM	203	0.54061				
PLRM	189	0.00936	14	0.53125	766.07	2.2e-16

V. CONCLUSIONS AND DISCUSSION

In the Gaussian family of distributions, we have demonstrated that the residual deviance can be easily reduces to the residual sum of squares. Besides, it is shown that the null deviance can be also reduces to the total sum of squares. Furthermore, coefficient of determination and adjusted coefficient of determination play quite important role in assessing of the goodness of fit of the regression models. We have indicated that these coefficients obtained by using linear regression models (LRM) can be easily generalized to partial linear regression models (PLRM). Especially, adjusted coefficient of determination in PLRM is very proper for assessment of the model goodness of fit because it detects the degrees of complexity of the PLRM.

It is shown that both of examples, the estimation performances of PLRM are better than LRM.

REFERENCES

[1] Mayers, Raymond. H., Classical and Modern Regression with Applications, Duxbury Classical Series, United States, 1990.

[2] Montgomarey, C. Douglas., Peck, A. Elizabeth., Vining, G. Geoffrey., Introduction to Linear Regression Analysis, John Wiley&Sons,Inc., Toronto, 2001.

[3] Hardle, Wolfang., Müller, Marlene., Sperlich, Stefan., Weratz, Axel., Nonparametric and Semiparametric Models, Springer, Berlin, 2004.

[4] Eubank, R. L., Nonparametric Regression and Smoothing Spline, Marcel Dekker Inc., 1999.

[5] Wahba, G., Spline Model for Observational Data, Siam, Philadelphia Pa., 1990.

[6] Green, P.J. and Silverman, B.W., Nonparametric Regression and Generalized Linear Models, Chapman & Hall, 1994.

[7] Schimek, G. Michael, Estimation and Inference in Partially Linear Models with Smoothing Splines, Journal of Statistical Planning and Inference, 91, 525-540, 2000.

[8] Hastie, T.J. and Tibshirani, R.J., Generalized Additive Models, Chapman & Hall /CRC, 1999.

[9] Wood, N. Simon., Generalized Additive Models An Introduction With R, Chapman & Hall/CRC, New York, 2006.

[10] Hastie, T., The gam Package, Generalized Additive Models, R topic documented, http://cran.r.project.org/packages/gam.pdf, February 16, 2008.

Dursun Aydın: I was born in 1969 at Ardahan, the city of Turkey. I graduated from Anadolu University Science Faculty Statistics department in 1992 at Eskişehir, the city of Turkey. Then I graduated from Marmara University Graduate School of Social in 2001 at İstanbul, the city of Turkey. Finally I graduated Anadolu University Graduate School of the Science, department of statistics, Doctorate Degree (Ph.D) in 2005.

He has been working for 13 years at Anadolu University as a teaching assistant, but now I am an instructor doctor at the same university. He has got many papers in national and international journals.

Dr. Aydın's interest fields: Multivariate Statistics, Basics Statistics, Parametric Regression, Nonparametric Regression, Semi-parametric Regression, Generalized Additive Models, Time Series