Breast Cancer Data Analysis

Using 3 Different Algorithms

Date

20/12/2019

Submitted by
Rajneesh Gulati

University
of Victoria

1. Brief Introduction to Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1][2]

Concepts of machine learning are used in computer vision and email filtering, where conventional algorithms are difficult to develop.
Machine learning tasks are classified into several broad categories.
In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) designating whether it contained the object. In special cases, the input may be only partially available, or restricted to special feedback. Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels.

Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object.

In unsupervised learning, the algorithm builds a mathematical model from a set of data that contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data.

Active learning algorithms access the desired outputs (training labels) for a limited set of inputs based on a budget and optimize the choice of inputs for which it will acquire training labels. When used interactively, these can be presented to a human user for labeling. Reinforcement learning algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment and are used in autonomous vehicles or in learning to play a game against a human opponent. Other specialized algorithms in machine learning include topic modeling, where the computer program is given a set of natural language documents and finds other documents that cover similar topics. Machine learning algorithms can be used to find the unobservable probability density function in density estimation problems. Meta learning algorithms learn their own inductive bias based on previous experience. In developmental robotics, robot learning algorithms generate their own sequences of learning experiences, also known as a curriculum, to cumulatively acquire new skills through self-guided exploration and social interaction with humans. These robots use guidance mechanisms such as active learning, maturation, motor synergies, and imitation.[1][2]

In this report, three different machine learning algorithms have been used to explore the breast cancer data set.

2. About the data set

D_bc_tr.mat: This matrix will be used for training purposes. It is of size 31 × 480 whose first 30 rows contain 30 medical features for each of 480 patients. The corresponding labels are contained in the 31st row of D_bc_tr. Label "−1" is assigned to the samples associated with those diagnosed as benign, while label "1" is assigned to the samples associated with those diagnosed as malignant.

D_bc_te.mat: This matrix will be used for test purposes. It is of size 31 × 89 whose first 30 rows constitute 30 medical features for each of 89 patients. The corresponding labels are contained in the 31st row of D_bc_tr.
As labels are given all the methods come under supervised learning.

3. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names.

Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

The binary logistic regression model has two levels of the dependent variable: categorical outputs with more than two values are modeled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cut off value and classifying inputs with probability greater than the cut off as one class, below the cut off as the other; this is a common way to make a binary classifier. The coefficients are generally not computed by a closed-form expression, unlike linear least squares.[3][4]

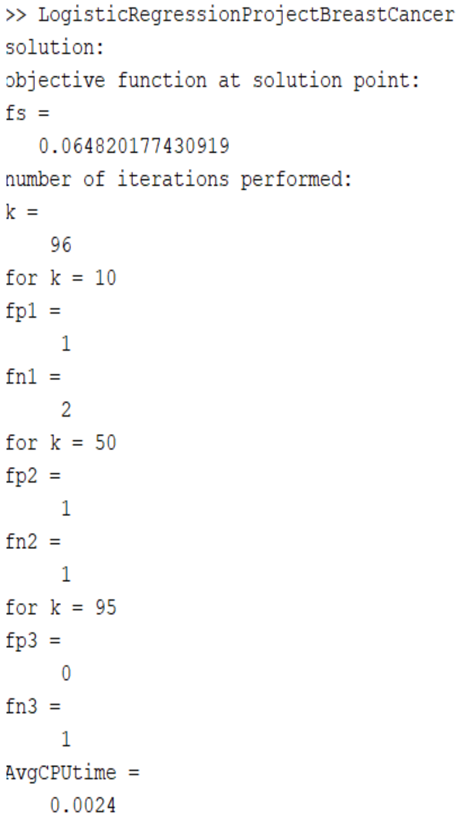
3.1 Objective

The goal is to develop a computer program for automatic diagnosis of breast cancer based on logistic regression where the logistic loss function is defined by a dataset provided by Dr. Wolberg from General Surgery Department, University of Wisconsin, Madison, WI in 1990's. The dataset contains 30 carefully selected features from each of 569 patients. The same dataset has also been made available from the UCI Machine Learning Repository. [5][6][7]

3.2 Results

In the results below:

1. ‘fs’ is the value of objective function at solution point.
2. ‘k’ refers to the number of iterations performed.
3. ‘[fp1 fn1], [fp2 fn2] and [fp3 fn3]’ refers to false positive and false negatives at k=10,50 and 95 respectively.
4. ‘AvgCpuTime’ refers to the average cpu time of the training process.



```
>> LogisticRegressionProjectBreastCancer
solution:
objective function at solution point:
fs =
   0.064820177430919
number of iterations performed:
k =
    96
for k = 10
fp1 =
     1
fn1 =
     2
for k = 50
fp2 =
     1
fn2 =
     1
for k = 95
fp3 =
     0
fn3 =
     1
AvgCPUtime =
    0.0024
```

3.3 Conclusion

Gradient Decent algorithm has been used to minimize the loss function along with back tracking line search, to calculate the step size. It took GD 96 iterations to converge to the solution time with average cpu time of 0.0024secs. With increasing iterations number of false positive and false negatives decreased as can be seen in the results. It can be said that GD gave satisfactory results. GD can be improved by scaling the variables and decreasing the cputime.

4. Multi-Feature Classification

The very simplest case of a single scalar predictor variable x and a single scalar response variable y is known as simple linear regression. The extension to multiple and/or vector-valued predictor variables (denoted with a capital X) is known as multiple linear regression, also known as multivariable linear regression. Nearly all real-world regression models involve multiple predictors, and basic descriptions of linear regression are often phrased in terms of the multiple regression model. [8]

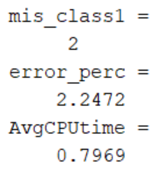
4.1 Objective

In this experiment, we build a multi-output linear model to predict if the patient has breast cancer or not. dataset provided by Dr. Wolberg from General Surgery Department, University of Wisconsin, Madison, WI in 1990's. The dataset contains 30 carefully selected features from each of 569 patients. The same dataset has also been made available from the UCI Machine Learning Repository.

4.2 Results

In the results below:

1. We get two miss classifications.
2. Average cpu time is 0.4531 secs for the process.
3. Error percentage is 2.2472, which is good.

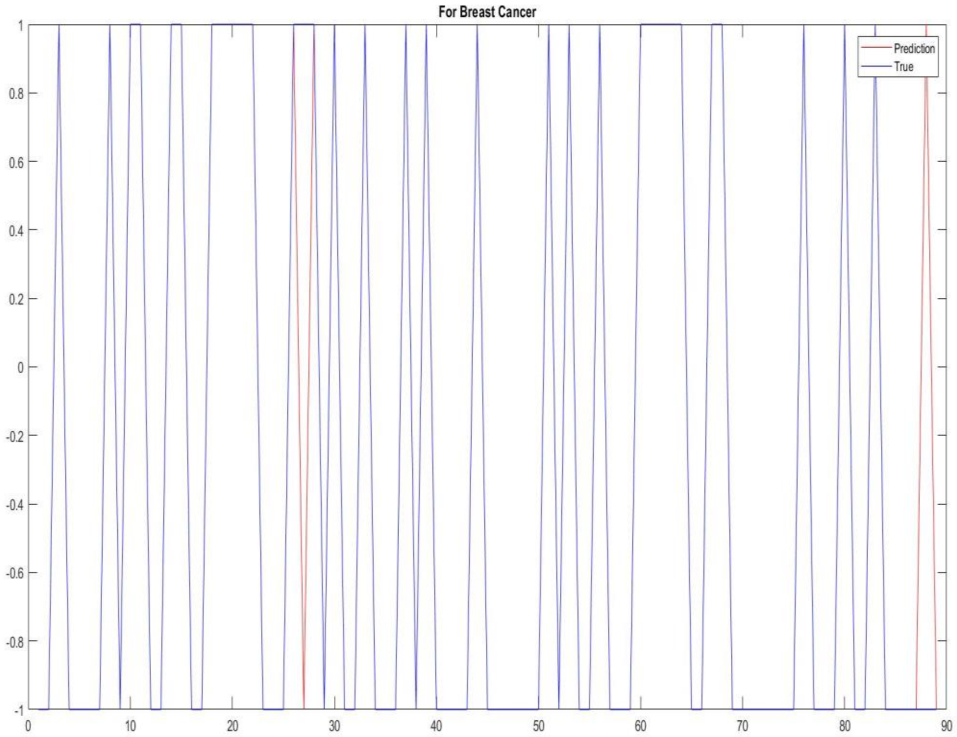


```
mis_class1 =
     2
error_perc =
    2.2472
AvgCPUtime =
    0.7969
```

4. The graph proves the point in 1.

4.3 Conclusion

Multifeatured classification works pretty good as there are two miss classifications. Downside is the when the data is large the cpu time will increase significantly, as cputime here directly depends on the size of the data set.



5. Linear Regression

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors,

rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.
Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.[9][10][11][12]

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

1. If the goal is prediction, or forecasting, or error reduction, [clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

2. If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

5.1 Objective

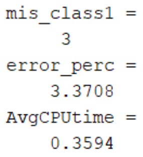
In this experiment, we investigate a technique for multi-category classification based on binary classifications. The technique is then applied to dataset provided by Dr. Wolberg from General Surgery Department, University of Wisconsin, Madison, WI in 1990’s. The dataset contains 30 carefully selected features from each of 569 patients. In this experiment, the above data set was divided into two sets, one for training and the other for testing.

5.2 Results

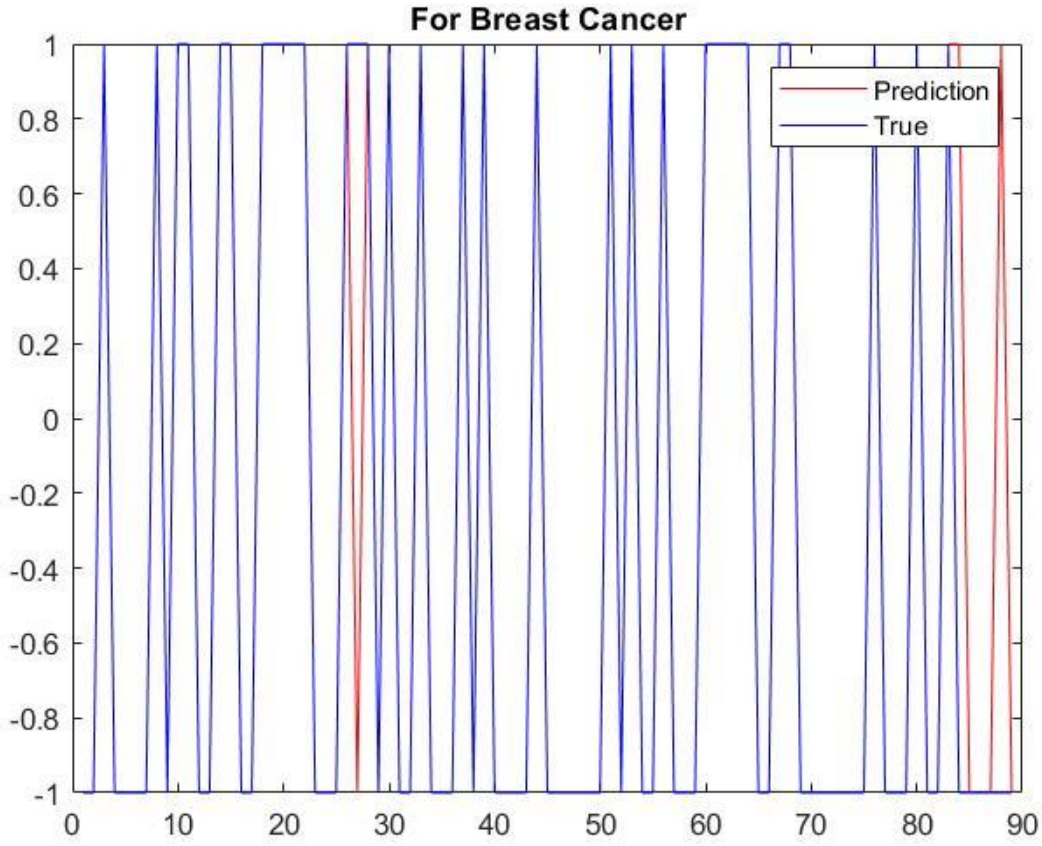
In the results below:

1. We get three miss classifications.
2. Average cpu time is 0.3594 secs for the process.
3. Error percentage is 3.3708, which is good.

4. The graph proves the point in 1.



```
mis_class1 =
     3
error_perc =
    3.3708
AvgCPUtime =
    0.3594
```



5.3 Conclusions

Linear regression works pretty good as there are three miss classifications. Downside is the when the data is large the cpu time will increase significantly, as cputime here directly depends on the size of the data set.

6. Comparison

Logistic regression worked the best with minimum cputime out of all the algorithms implemented. It took GD 96 iterations to converge to solution point. GD has just one miss classification which is a false negative. Multi output classification has two miss classifications and linear model has three miss classifications, but the cputime for is high when compared to logistic regression. Depending upon the size of the data set cputime for these two models can increase even more.

7. Code for logistic regression

```
load D_bc_te.mat
load D_bc_tr.mat

train = D_bc_tr;
test = D_bc_te;

X_train = train(1:30, :);
Y_train = train(31, :);

X_test = test(1:30, :);
Y_test = test(31, :);
for i = 1:30
     xi = X_train(i,:);
     mi = mean(xi);
     vi = sqrt(var(xi));
X_train(i,:) = (xi - mi)/vi;
end
X_train = [X_train; ones(1, 480); Y_train];
for i = 1:30
 xi = X_test(i,:);
 mi = mean(xi);
 vi = sqrt(var(xi));
X_test(i,:) = (xi - mi)/vi;
end

initial_cpu_time = cputime;
X_test = [X_test; ones(1, 89)];
w = zeros(31,1);
[xs,fs,k, xs1, xs2, xs3] = grad_desc('f_logistic',
'g_logistic', w, 1e-2, X_train);
[fp1, fn1] = classify(X_test, Y_test, xs1);
[fp2, fn2] = classify(X_test, Y_test, xs2);
[fp3, fn3] = classify(X_test, Y_test, xs3);

disp('for k = 10');
fp1
fn1

disp('for k = 50');
fp2
fn2
```

```
  disp('for k = 95');
  fp3
  fn3
  final_cpu_time = cputime;

  AvgCPUtime = (final_cpu_time - initial_cpu_time)/k

  function [fp, fn] = classify(X_test,Y_test, xs)
    result = zeros(89, 1);
        fp = 0;
        fn = 0;
        for i = 1:89
         y = xs' * X_test(:, i);
             if y > 0
                 result(i) = 1;
             else
                 result(i) = -1;
             end
             if result(i) == 1 && result(i) ~=
Y_test(i)
                  fn = fn + 1;
        end
        if result(i) == -1 && result(i) ~= Y_test(i)
            fp = fp + 1;
        end
    end
end
```

1. Code for GD and back tracking line search taken from the course website.

8. Code for multi feature classification

```
clc
clear
close all


load D_bc_te.mat;
load D_bc_tr.mat;
initial_cpu_time = cputime;
Xtr = D_bc_tr(1:30,:);
Ytr = D_bc_tr(31,:);
```

```
Xte = D_bc_te(1:30,:);
Yte = D_bc_te(31,:);

one=ones(1,480);
one=one.';
Xtr=Xtr.';

Xcap=[Xtr,one];

Xtemp=Xcap.'*Xcap;

I=eye(31);

Xtemp2=Xtemp+0.01*I;

Xinverse=inv(Xtemp2);

Xtemp3=Xinverse*Xcap.';

WandB=Xtemp3*Ytr.';

Wstar1= WandB(1:30,1);

Bstar1=WandB(31,1);




%Testing
Ypre1=Wstar1.'*Xte+Bstar1;


Ypre=sign(Ypre1);
mis_class1=0;
for i=1:89
    if Ypre(:,i)~=Yte(:,i)
        mis_class1=mis_class1+1;
    end
end

mis_class1
```

```
error_perc=(mis_class1/length(Yte))*100



error=norm(Yte-Ypre,'fro');
e=norm(Yte,'fro');
error_relative=error/e;

figure, plot(Ypre,'r-');
hold on;
plot(Yte(1,:),'b-');
legend('Prediction','True');
title('For Breast Cancer');
final_cpu_time = cputime;
AvgCPUtime = (final_cpu_time - initial_cpu_time)
```

9. Code for linear regression

```
clc
clear
close all


load D_bc_te.mat;
load D_bc_tr.mat;
initial_cpu_time = cputime;
Xtr = D_bc_tr(1:30,:);
Ytr = D_bc_tr(31,:);
```

```
Xte = D_bc_te(1:30,:);
Yte = D_bc_te(31,:);


X_hat_new=[Xtr;ones(1,480)];
X_hat_new=X_hat_new.';
W_hat= pinv(X_hat_new.'*X_hat_new)*X_hat_new.'*Ytr';
for i=1:30
    Wstar(i)=W_hat(i,:);
end

Bstar=W_hat(31,:);
Yfinal1= Wstar*Xte+Bstar;

Ypre=sign(Yfinal1);

mis_class1=0;
for i=1:89
    if Ypre(:,i)~=Yte(:,i)
        mis_class1=mis_class1+1;
    end
end

mis_class1

error_perc=(mis_class1/length(Yte))*100
error=norm(Yte-Ypre,'fro');
e=norm(Yte,'fro');

error_relative=error/e;

figure, plot(Ypre,'r-');
hold on;
plot(Yte(1,:),'b-');
legend('Prediction','True');
title('For Breast Cancer');
final_cpu_time = cputime;
AvgCPUtime = (final_cpu_time - initial_cpu_time)
```

References

[1]] The definition "without being explicitly programmed" is often attributed to Arthur Samuel, who coined the term "machine learning" in 1959, but the phrase is not found verbatim in this publication, and may be a paraphrase that appeared later. Confer "Paraphrasing Arthur Samuel (1959), the question is: How can computers learn to solve problems without being explicitly programmed?" in Koza, John R.; Bennett, Forrest H.; Andre, David; Keane, Martin A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151–170. doi:10.1007/978-94-009-0279-4_9.

[2] Jump up to:a b c d Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2

[3] Tolles, Juliana; Meurer, William J (2016). "Logistic Regression Relating Patient Characteristics to Outcomes". JAMA Jama. 316 (5): 533–4. doi:10.1001/jama.2016.7653. ISSN 0098-7484. OCLC 6823603312. PMID 27483067.

[4] Jump up to:a b Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54 (1/2): 167–178. doi:10.2307/2333860. JSTOR 2333860.

[5] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear feature extraction for breast tumor diagnosis,” in IS?T/SPIE Int. Symp. Electronic Imaging: Science and Technology, vol. 1905, pp. 861-870, San Jose, CA., 1993.

[6] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, “Breast cancer diagnosis and prognosis via linear programming,” AAAI Tech. Report SS-94-01, 1994.

[7] UCI Machine Learning, http://archive.ics.uci.edu/ml, University of California Irvine, School of Information and Computer Science.

[8]https://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_linear_regression

[9] David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 26. A simple regression equation has on the right-hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression equation has two or more explanatory variables on the right-hand side, each with its own slope coefficient

[10] Rencher, Alvin C.; Christensen, William F. (2012), "Chapter 10, Multivariate regression – Section 10.1, Introduction", Methods of Multivariate Analysis, Wiley

Series in Probability and Statistics, 709 (3rd ed.), John Wiley & Sons, p. 19, ISBN 9781118391679.

[11] Hilary L. Seal (1967). "The historical development of the Gauss linear model". Biometrika. 54 (1/2): 1–24. doi:10.1093/biomet/54.1-2.1. JSTOR 2333849.

[12] Yan, Xin (2009), Linear Regression Analysis: Theory and Computing, World Scientific, pp. 1–2, ISBN 9789812834119, Regression analysis ... is probably one of the oldest topics in mathematical statistics dating back to about two hundred years ago. The earliest form of the linear regression was the least squares method, which was published by Legendre in 1805, and by Gauss in 1809 ... Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the sun.