# Least Squares (LS)

## Contents



## 1 Introduction

There can be difficulties working with linear regression models in GAMS. An explicit minimization problem will be non-linear as it needs to express a sum of squares: this model may be difficult to solve. Alternatively, it is well known that a linear formulation using the normal equations (X'X)b=X'y will introduce numerical instability.

Therefore we have introduced a compact notation where we replace the objective by a dummy equation: the solver will implicitly understand that we need to minimize the sum of squared residuals. The LS solver will understand this notation and can apply a stable QR decomposition to solve the model quickly and accurately.

## 2 Basic Usage

A least squares model contains a dummy objective and a set of linear equations:

```
sumsq..   sse =n= 0;
fit(i)..  data(i,'y') =e= b0 + b1*data(i,'x');

option lp = ls;
model leastsq /fit,sumsq/;
solve leastsq using lp minimizing sse;
```

Here `sse` is a free variable that will hold the sum of squared residuals after solving the model. The variables `b0` and `b1` are the statistical coefficients to be estimated. On return the levels are the estimates and the marginals are the standard errors. The `fit` equations describe the equation to be fitted.

The constant term or intercept is included in the above example. If you don't specify it explicitly, and the solver detects the absence of a column of ones in the data matrix X, then a constant term will be added automatically. When you need to do a regression without intercept you will need to use an option `add_constant_term 0`.

It is not needed or beneficial to specify initial values (levels) or an advanced basis (marginals) as they are ignored by the solver.

The estimates are returned as the levels of the variables. The marginals will contain the standard errors. The row levels reported are the residuals errors. In addition a GDX file is written which will contain all regression statistics.

Several complete examples of LS solver usage are available in testlib starting with GAMS Distribution 22.8. For example, model ls01 takes the data from the Norris dataset found in the NIST collection of statistical reference datasets and reproduces the results and regression statistics found there.

Erwin Kalvelagen is the original author and further information can be found at Amsterdam Optimization Modeling Group's web site.

# 3 Options

The following options are recognized:

| Option | Description | Default |
|---|---|---|
| maxn | Maximum number of cases or observations. This is the number of rows (not counting the dummy objective). When the number of rows is very large, this is probably not a regression problem but a generic LP model. To protect against those, we don't accept models with an enormous number of rows. | 1000 |
| maxp | Maximum number of coefficients to estimate. This is the number of columns or variables (not counting the dummy objective variable). When the number of variables is very large, this is probably not a regression problem but a generic LP model. To protect against those, we don't accept models with an enormous number of columns. | 25 |
| add_constant_term | Must be 0, 1, or 2. If this number is zero, no constant term or intercept will be added to the problem. If this option is one, then always a constant term will be added. If this option is two, the algorithm will add a constant term only if there is no data column with all ones in the matrix. In this automatic mode, if the user already specified an explicit intercept in the problem, no additional constant term will be added. As the default is two, you will need to provide an option add_constant_term 0 in case you want to solve a regression problem without an intercept. | 2 |
| gdx_file_name | Name of the GDX file where results are saved. | ls.gdx |