Ordinary least squares

In statistics, **ordinary least squares** (**OLS**) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function.

Geometrically, this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression surface – the smaller the differences, the better the model fits the data. The resulting estimator can be expressed by a simple formula, especially in the case of a simple linear regression, in which there is a single regressor on the right side of the regression equation.

The OLS estimator is consistent when the regressors are exogenous, and optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated. Under these conditions, the method of OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances. Under the additional assumption that the errors are normally distributed, OLS is the maximum likelihood estimator.

OLS is used in fields as diverse as economics (econometrics), political science, psychology and engineering (control theory and signal processing).

Contents

Linear model

Suppose the data consists of n observations $\{ y_i, x_i \}_{i=1}^{n}$. Each observation i includes a scalar response y_i and a column vector x_i of values of p predictors (regressors) x_{ij} for $j = 1, \ldots, p$. In a linear regression model, the response variable, $\boldsymbol{y_i}$, is a linear function of the regressors:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

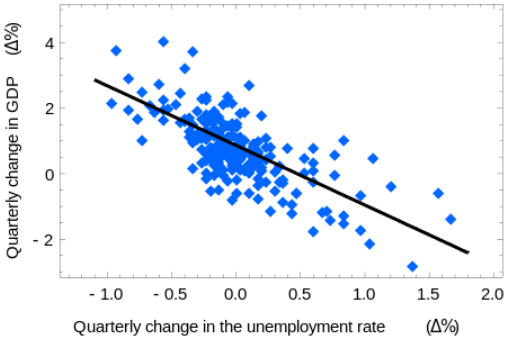
or in vector form,

$$y_i = x_i^T \beta + \varepsilon_i,$$

where β is a $p \times 1$ vector of unknown parameters; the ε_i's are unobserved scalar random variables (errors) which account for influences upon the responses y_i from sources other than the explanators x_i; and $\boldsymbol{x_i}$ is a column vector of the ith observations of all the explanatory variables. This model can also be written in matrix notation as

$$y = X\beta + \varepsilon,$$

where y and ε are $n \times 1$ vectors of the values of the response variable and the errors for the various observations, and X is an $n \times p$ matrix of regressors, also sometimes called the design matrix, whose row i is x_i^T and contains the ith observations on all the explanatory variables.



Okun's law in macroeconomics states that in an economy the GDP growth should depend linearly on the changes in the unemployment rate. Here the ordinary least squares method is used to construct the regression line describing this law.

As a rule, the constant term is always included in the set of regressors X, say, by taking $x_{i1} = 1$ for all $i = 1, \ldots, n$. The coefficient β_1 corresponding to this regressor is called the *intercept*.

There may be some relationship between the regressors. For instance, the third regressor may be the square of the second regressor. In this case (assuming that the first regressor is constant) we have a quadratic model in the second regressor. But this is still considered a linear model because it is linear in the βs.

Matrix/vector formulation

Consider an overdetermined system

$$\sum_{j=1}^{n} X_{ij}\beta_j = y_i, \ (i = 1, 2, \ldots, m),$$

of m linear equations in n unknown coefficients, $\beta_1, \beta_2, \ldots, \beta_n$, with $m > n$. (Note: for a linear model as above, not all of $\boldsymbol{X}$ contains information on the data points. The first column is populated with ones, $\boldsymbol{X_{i1}} = \mathbf{1}$, only the other columns contain actual data, and n = number of regressors + 1.) This can be written in matrix form as

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y},$$

where

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Such a system usually has no solution, so the goal is instead to find the coefficients $\boldsymbol{\beta}$ which fit the equations "best",in the sense of solving the quadratic minimization problem

$$\hat{\boldsymbol{\beta}} = \operatorname{arg\,min}_{\boldsymbol{\beta}} S(\boldsymbol{\beta}),$$

where the objective function S is given by

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} X_{ij}\beta_j \right|^2 = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2.$$

A justification for choosing this criterion is given in properties below. This minimization problem has a unique solution, provided that the n columns of the matrix $\mathbf{X}$ are linearly independent, given by solving the normal equations

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

The matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is known as the Gramian matrix of $\mathbf{X}$, which possesses several nice properties such as being a positive semi-definite matrix, and the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is known as the moment matrix of regressand by regressors.[1] Finally, $\hat{\boldsymbol{\beta}}$ is the coefficient vector of the least-squares hyperplane, expressed as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Estimation

Suppose b is a "candidate" value for the parameter vector β. The quantity $y_i - x_i^{\mathrm{T}} b$, called the **residual** for the i-th observation, measures the vertical distance between the data point $(x_i\, y_i)$ and the hyperplane $y = x^{\mathrm{T}} b$, and thus assesses the degree of fit between the actual data and the model. The **sum of squared residuals** (**SSR**) (also called the **error sum of squares** (**ESS**) or **residual sum of squares** (**RSS**))[2] is a measure of the overall model fit:

$$S(b) = \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}} b)^2 = (y - Xb)^{\mathrm{T}} (y - Xb),$$

where T denotes the matrix transpose, and the rows of X, denoting the values of all the independent variables associated with a particular value of the dependent variable, are $X_i = x_i^{\mathrm{T}}$. The value of b which minimizes this sum is called the **OLS estimator for $\boldsymbol{\beta}$**. The function $S(b)$ is quadratic in b with positive-definite Hessian, and therefore this function possesses a unique global minimum at $\boldsymbol{b} = \hat{\boldsymbol{\beta}}$, which can be given by the explicit formula:[3][proof]

$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} S(b) = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y\,.$$

The product $N{=}X^{\mathrm{T}} X$ is a normal matrix and its inverse, $Q{=}N^{-1}$, is the *cofactor matrix* of β,[4][5][6] closely related to its covariance matrix, C_β. The matrix $(X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}}{=}Q\, X^{\mathrm{T}}$ is called the Moore–Penrose pseudoinverse matrix of X. This formulation highlights the point that estimation can be carried out if, and only if, there is no perfect multicollinearity between the explanatory variables (which would cause the normal matrix to have no inverse).

After we have estimated β, the **fitted values** (or **predicted values**) from the regression will be

$$\hat{y} = X\hat{\beta} = Py,$$

where $P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the projection matrix onto the space V spanned by the columns of X. This matrix P is also sometimes called the hat matrix because it "puts a hat" onto the variable y. Another matrix, closely related to P is the *annihilator* matrix $M = I_n - P$; this is a projection matrix onto the space orthogonal to V. Both matrices P and M are symmetric and idempotent (meaning that $P^2 = P$), and relate to the data matrix X via identities $\boldsymbol{PX = X}$ and $MX = 0$.[7] Matrix M creates the **residuals** from the regression:

$$\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta} = My = M(X\beta + \varepsilon) = (MX)\beta + M\varepsilon = M\varepsilon.$$

Using these residuals we can estimate the value of σ^2, called the **reduced chi-squared**:

$$s^2 = \frac{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}}{n-p} = \frac{(My)^{\mathrm{T}} My}{n-p} = \frac{y^{\mathrm{T}} M^{\mathrm{T}} My}{n-p} = \frac{y^{\mathrm{T}} My}{n-p} = \frac{S(\hat{\beta})}{n-p}, \qquad \hat{\sigma}^2 = \frac{n-p}{n}\, s^2$$

The numerator, $n{-}p$, is the statistical degrees of freedom. The first quantity, s^2, is the OLS estimate for σ^2, whereas the second, $\hat{\sigma}^2$, is the MLE estimate for σ^2. The two estimators are quite similar in large samples; the first one is always unbiased, while the second is biased but minimizes the mean squared error of the estimator. In practice s^2 is used more often, since it is more convenient for the hypothesis testing. The square root of s^2 is called the **regression standard error**,[8] **standard error of the regression**,[9][10] or **standard error of the equation**.[7]

It is common to assess the goodness-of-fit of the OLS regression by comparing how much the initial variation in the sample can be reduced by regressing onto X. The **coefficient of determination** R^2 is defined as a ratio of "explained" variance to the "total" variance of the dependent variable y:[11]

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{y^{\mathrm{T}} P^{\mathrm{T}} L P y}{y^{\mathrm{T}} L y} = 1 - \frac{y^{\mathrm{T}} M y}{y^{\mathrm{T}} L y} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

where TSS is the **total sum of squares** for the dependent variable, $L = I_n - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n$, and $\mathbf{1}$ is an $n\times 1$ vector of ones. (L is a "centering matrix" which is equivalent to regression on a constant; it simply subtracts the mean from a variable.) In order for R^2 to be meaningful, the matrix X of data on regressors must contain a column vector of ones to represent the constant whose coefficient is the regression intercept. In that case, R^2 will always be a number between 0 and 1, with values close to 1 indicating a good degree of fit.

The variance in the prediction of the independent variable as a function of the dependent variable is given in the article Polynomial least squares.

Simple linear regression model

If the data matrix X contains only two variables, a constant and a scalar regressor x_i, then this is called the "simple regression model".[12] This case is often considered in the beginner statistics classes, as it provides much simpler formulas even suitable for manual calculation. The parameters are commonly denoted as (α, β):

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

The least squares estimates in this case are given by simple formulas

$$\hat{\beta} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\mathrm{Cov}[x,y]}{\mathrm{Var}[x]}$$
$$\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}\,,$$

where Var(.) and Cov(.) are sample parameters.

Alternative derivations

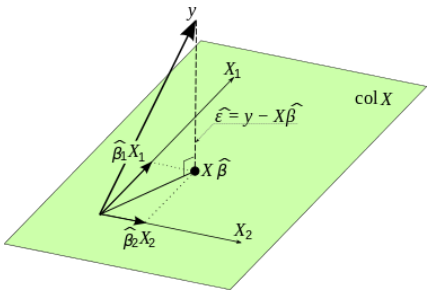
In the previous section the least squares estimator $\hat{\beta}$ was obtained as a value that minimizes the sum of squared residuals of the model. However it is also possible to derive the same estimator from other approaches. In all cases the formula for OLS estimator remains the same: $\hat{\beta} = (X^T X)^{-1} X^T y$; the only difference is in how we interpret this result.

Projection

For mathematicians, OLS is an approximate solution to an overdetermined system of linear equations $X\beta \approx y$, where β is the unknown. Assuming the system cannot be solved exactly (the number of equations n is much larger than the number of unknowns p), we are looking for a solution that could provide the smallest discrepancy between the right- and left- hand sides. In other words, we are looking for the solution that satisfies

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|,$$

where $||\cdot||$ is the standard L^2 norm in the n-dimensional Euclidean space $\mathbf{R}^n$. The predicted quantity $X\beta$ is just a certain linear combination of the vectors of regressors. Thus, the residual vector $y - X\beta$ will have the smallest length when y is projected orthogonally onto the linear subspace spanned by the columns of X. The OLS estimator $\hat{\beta}$ in this case can be interpreted as the coefficients of vector decomposition of $\hat{y} = Py$ along the basis of X.



OLS estimation can be viewed as a projection onto the linear space spanned by the regressors. (Here each of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ refers to a column of the data matrix.)

In other words, the gradient equations at the minimum can be written as:

$$(\mathbf{y} - X\hat{\boldsymbol{\beta}})^{\mathrm{T}} X = 0.$$

A geometrical interpretation of these equations is that the vector of residuals, $\mathbf{y} - X\hat{\boldsymbol{\beta}}$ is orthogonal to the column space of X, since the dot product $(\mathbf{y} - X\hat{\boldsymbol{\beta}}) \cdot X\mathbf{v}$ is equal to zero for *any* conformal vector, $\mathbf{v}$. This means that $\mathbf{y} - X\hat{\boldsymbol{\beta}}$ is the shortest of all possible vectors $\mathbf{y} - X\boldsymbol{\beta}$, that is, the variance of the residuals is the minimum possible. This is illustrated at the right.

Introducing $\hat{\boldsymbol{\gamma}}$ and a matrix K with the assumption that a matrix $[X\ K]$ is non-singular and $K^{\mathrm{T}} X = 0$ (cf. Orthogonal projections), the residual vector should satisfy the following equation:

$$\hat{\mathbf{r}} \triangleq \mathbf{y} - X\hat{\boldsymbol{\beta}} = K\hat{\boldsymbol{\gamma}}.$$

The equation and solution of linear least squares are thus described as follows:

$$\mathbf{y} = \begin{bmatrix} X & K \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix},$$
$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{bmatrix} X & K \end{bmatrix}^{-1} \mathbf{y} = \begin{bmatrix} (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \\ (K^{\mathrm{T}} K)^{-1} K^{\mathrm{T}} \end{bmatrix} \mathbf{y}.$$

Another way of looking at it is to consider the regression line to be a weighted average of the lines passing through the combination of any two points in the dataset.[13] Although this way of calculation is more computationally expensive, it provides a better intuition on OLS.

Maximum likelihood

The OLS estimator is identical to the maximum likelihood estimator (MLE) under the normality assumption for the error terms.[14][proof] This normality assumption has historical importance, as it provided the basis for the early work in linear regression analysis by Yule and Pearson. From the properties of MLE, we can infer that the OLS estimator is asymptotically efficient (in the sense of attaining the Cramér–Rao bound for variance) if the normality assumption is satisfied.[15]

Generalized method of moments

In iid case the OLS estimator can also be viewed as a GMM estimator arising from the moment conditions

$$\mathrm{E}\big[\, x_i (y_i - x_i^T \beta) \,\big] = 0.$$

These moment conditions state that the regressors should be uncorrelated with the errors. Since x_i is a p-vector, the number of moment conditions is equal to the dimension of the parameter vector β, and thus the system is exactly identified. This is the so-called classical GMM case, when the estimator does not depend on the choice of the weighting matrix.

Note that the original strict exogeneity assumption $\mathrm{E}[\varepsilon_i \mid x_i] = 0$ implies a far richer set of moment conditions than stated above. In particular, this assumption implies that for any vector-function f, the moment condition $\mathrm{E}[f(x_i)\cdot\varepsilon_i] = 0$ will hold. However it can be shown using the Gauss–Markov theorem that the optimal choice of function f is to take $f(x) = x$, which results in the moment equation posted above.

Properties

Assumptions

There are several different frameworks in which the linear regression model can be cast in order to make the OLS technique applicable. Each of these settings produces the same formulas and same results. The only difference is the interpretation and the assumptions which have to be imposed in order for the method to give meaningful results. The choice of the applicable framework depends mostly on the nature of data in hand, and on the inference task which has to be performed.

One of the lines of difference in interpretation is whether to treat the regressors as random variables, or as predefined constants. In the first case (**random design**) the regressors x_i are random and sampled together with the y_i's from some population, as in an observational study. This approach allows for more natural study of the asymptotic properties of the estimators. In the other interpretation (**fixed design**), the regressors X are treated as known constants set by a design, and y is sampled conditionally on the values of X as in an experiment. For practical purposes, this distinction is often unimportant, since estimation and inference is carried out while conditioning on X. All results stated in this article are within the random design framework.

Classical linear regression model

The classical model focuses on the "finite sample" estimation and inference, meaning that the number of observations n is fixed. This contrasts with the other approaches, which study the asymptotic behavior of OLS, and in which the number of observations is allowed to grow to infinity.

- **Correct specification**. The linear functional form must coincide with the form of the actual data-generating process.
- **Strict exogeneity**. The errors in the regression should have conditional mean zero:[16]

$$\mathbf{E}[\varepsilon \mid X] = 0.$$

  The immediate consequence of the exogeneity assumption is that the errors have mean zero: $\mathrm{E}[\varepsilon] = 0$, and that the regressors are uncorrelated with the errors: $\mathrm{E}[X^{\mathrm{T}}\varepsilon] = 0$.

  The exogeneity assumption is critical for the OLS theory. If it holds then the regressor variables are called *exogenous*. If it doesn't, then those regressors that are correlated with the error term are called *endogenous*,[17] and then the OLS estimates become invalid. In such case the method of instrumental variables may be used to carry out inference.

- **No linear dependence**. The regressors in X must all be linearly independent. Mathematically, this means that the matrix X must have full column rank almost surely:[18]

$$\mathbf{Pr}\big[\,\mathbf{rank}(X) = p\big] = 1.$$

  Usually, it is also assumed that the regressors have finite moments up to at least the second moment. Then the matrix $Q_{xx} = \mathrm{E}[X^{\mathrm{T}}X/n]$ is finite and positive semi-definite.
  When this assumption is violated the regressors are called linearly dependent or perfectly multicollinear. In such case the value of the regression coefficient β cannot be learned, although prediction of y values is still possible for new values of the regressors that lie in the same linearly dependent subspace.

- **Spherical errors**:[18]

$$\mathbf{Var}[\varepsilon \mid X] = \sigma^2 I_n,$$

  where I_n is the identity matrix in dimension n, and σ^2 is a parameter which determines the variance of each observation. This σ^2 is considered a nuisance parameter in the model, although usually it is also estimated. If this assumption is violated then the OLS estimates are still valid, but no longer efficient.
  It is customary to split this assumption into two parts:

  - **Homoscedasticity**: $\mathrm{E}[\,\varepsilon_i^2 \mid X] = \sigma^2$, which means that the error term has the same variance σ^2 in each observation. When this requirement is violated this is called heteroscedasticity, in such case a more efficient estimator would be weighted least squares. If the errors have infinite variance then the OLS estimates will also have infinite variance (although by the law of large numbers they will nonetheless tend toward the true values so long as the errors have zero mean). In this case, robust estimation techniques are recommended.
  - **No autocorrelation**: the errors are uncorrelated between observations: $\mathrm{E}[\,\varepsilon_i\varepsilon_j \mid X] = 0$ for $i \neq j$. This assumption may be violated in the context of time series data, panel data, cluster samples, hierarchical data, repeated measures data, longitudinal data, and other data with dependencies. In such cases generalized least squares provides a better alternative than the OLS. Another expression for autocorrelation is *serial correlation*.

- **Normality**. It is sometimes additionally assumed that the errors have normal distribution conditional on the regressors:[19]

$$\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 I_n).$$

  This assumption is not needed for the validity of the OLS method, although certain additional finite-sample properties can be established in case when it does (especially in the area of hypotheses testing). Also when the errors are normal, the OLS estimator is equivalent to the maximum likelihood estimator (MLE), and therefore it is asymptotically efficient in the class of all regular estimators. Importantly, the normality assumption applies only to the error terms; contrary to a popular misconception, the response (dependent) variable is not required to be normally distributed.[20]

Independent and identically distributed (iid)

In some applications, especially with cross-sectional data, an additional assumption is imposed — that all observations are independent and identically distributed. This means that all observations are taken from a random sample which makes all the assumptions listed earlier simpler and easier to interpret. Also this framework allows one to state asymptotic results (as the sample size $n \to \infty$), which are understood as a theoretical possibility of fetching new independent observations from the data generating process. The list of assumptions in this case is:

- **iid observations**: (x_i, y_i) is independent from, and has the same distribution as, (x_j, y_j) for all $i \neq j$;
- **no perfect multicollinearity**: $Q_{xx} = \mathrm{E}[\, x_i x_i^{\mathrm{T}} \,]$ is a positive-definite matrix;
- **exogeneity**: $\mathrm{E}[\, \varepsilon_i \,|\, x_i \,] = 0$;
- **homoscedasticity**: $\mathrm{Var}[\, \varepsilon_i \,|\, x_i \,] = \sigma^2$.

Time series model

- The stochastic process $\{x_i, y_i\}$ is stationary and ergodic; if $\{x_i, y_i\}$ is nonstantionary, OLS results are often spurious unless $\{x_i, y_i\}$ is co-integrating.
- The regressors are *predetermined*: $\mathrm{E}[x_i \varepsilon_i] = 0$ for all $i = 1, \ldots, n$;
- The $p \times p$ matrix $Q_{xx} = \mathrm{E}[\, x_i x_i^{\mathrm{T}} \,]$ is of full rank, and hence positive-definite;
- $\{x_i \varepsilon_i\}$ is a martingale difference sequence, with a finite matrix of second moments $Q_{xx\varepsilon^2} = \mathrm{E}[\, \varepsilon_i^2 x_i x_i^{\mathrm{T}} \,]$.

Finite sample properties

First of all, under the *strict exogeneity* assumption the OLS estimators $\hat{\beta}$ and s^2 are unbiased, meaning that their expected values coincide with the true values of the parameters:[21][proof]

$$\mathbf{E}[\, \hat{\beta} \mid X \,] = \beta, \quad \mathbf{E}[\, s^2 \mid X \,] = \sigma^2.$$

If the strict exogeneity does not hold (as is the case with many time series models, where exogeneity is assumed only with respect to the past shocks but not the future ones), then these estimators will be biased in finite samples.

The *variance-covariance matrix* (or simply *covariance matrix*) of $\hat{\beta}$ is equal to [22]

$$\mathbf{Var}[\, \hat{\beta} \mid X \,] = \sigma^2 (X^T X)^{-1} = \sigma^2 Q.$$

In particular, the standard error of each coefficient $\hat{\beta}_j$ is equal to square root of the j-th diagonal element of this matrix. The estimate of this standard error is obtained by replacing the unknown quantity σ^2 with its estimate s^2. Thus,

$$\widehat{\mathrm{s.e.}}(\hat{\beta}_j) = \sqrt{s^2 (X^T X)^{-1}_{jj}}$$

It can also be easily shown that the estimator $\hat{\beta}$ is uncorrelated with the residuals from the model:[22]

$$\mathbf{Cov}[\, \hat{\beta}, \hat{\varepsilon} \mid X \,] = 0.$$

The **Gauss–Markov theorem** states that under the *spherical errors* assumption (that is, the errors should be uncorrelated and homoscedastic) the estimator $\hat{\beta}$ is efficient in the class of linear unbiased estimators. This is called the **best linear unbiased estimator (BLUE)**. Efficiency should be understood as if we were to find some other estimator $\tilde{\beta}$ which would be linear in y and unbiased, then [22]

$$\mathbf{Var}[\, \tilde{\beta} \mid X \,] - \mathbf{Var}[\, \hat{\beta} \mid X \,] \geq 0$$

in the sense that this is a nonnegative-definite matrix. This theorem establishes optimality only in the class of linear unbiased estimators, which is quite restrictive. Depending on the distribution of the error terms ε, other, non-linear estimators may provide better results than OLS.

Assuming normality

The properties listed so far are all valid regardless of the underlying distribution of the error terms. However, if you are willing to assume that the *normality assumption* holds (that is, that $\varepsilon \sim N(0, \sigma^2 I_n)$), then additional properties of the OLS estimators can be stated.

The estimator $\hat{\beta}$ is normally distributed, with mean and variance as given before:[23]

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^{\mathrm{T}} X)^{-1})$$

where Q is the cofactor matrix. This estimator reaches the Cramér–Rao bound for the model, and thus is optimal in the class of all unbiased estimators.[15] Note that unlike the Gauss–Markov theorem, this result establishes optimality among both linear and non-linear estimators, but only in the case of normally distributed error terms.

The estimator s^2 will be proportional to the chi-squared distribution:[24]

$$s^2 \sim \frac{\sigma^2}{n-p} \cdot \chi^2_{n-p}$$

The variance of this estimator is equal to $2\sigma^4/(n-p)$, which does not attain the Cramér–Rao bound of $2\sigma^4/n$. However it was shown that there are no unbiased estimators of σ^2 with variance smaller than that of the estimator s^2.[25] If we are willing to allow biased estimators, and consider the class of estimators that are proportional to the sum of squared residuals (SSR) of the model, then the best (in the sense of the mean squared error) estimator in this class will be $\tilde{\sigma}^2 = \mathrm{SSR}/(n-p+2)$, which even beats the Cramér–Rao bound in case when there is only one regressor ($p = 1$).[26]

Moreover, the estimators $\hat{\beta}$ and s^2 are independent,[27] the fact which comes in useful when constructing the t- and F-tests for the regression.

Influential observations

As was mentioned before, the estimator $\hat{\beta}$ is linear in y, meaning that it represents a linear combination of the dependent variables y_i. The weights in this linear combination are functions of the regressors X, and generally are unequal. The observations with high weights are called **influential** because they have a more pronounced effect on the value of the estimator.

To analyze which observations are influential we remove a specific j-th observation and consider how much the estimated quantities are going to change (similarly to the jackknife method). It can be shown that the change in the OLS estimator for β will be equal to [28]

$$\hat{\beta}^{(j)} - \hat{\beta} = -\frac{1}{1-h_j} (X^{\mathrm{T}} X)^{-1} x_j^{\mathrm{T}} \hat{\varepsilon}_j \,,$$

where $h_j = x_j^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} x_j$ is the j-th diagonal element of the hat matrix P, and x_j is the vector of regressors corresponding to the j-th observation. Similarly, the change in the predicted value for j-th observation resulting from omitting that observation from the dataset will be equal to [28]

$$\hat{y}_j^{(j)} - \hat{y}_j = x_j^{\mathrm{T}} \hat{\beta}^{(j)} - x_j^{\mathrm{T}} \hat{\beta} = -\frac{h_j}{1-h_j} \hat{\varepsilon}_j$$

From the properties of the hat matrix, $0 \le h_j \le 1$, and they sum up to p, so that on average $h_j \approx p/n$. These quantities h_j are called the **leverages**, and observations with high h_j are called **leverage points**.[29] Usually the observations with high leverage ought to be scrutinized more carefully, in case they are erroneous, or outliers, or in some other way atypical of the rest of the dataset.

Partitioned regression

Sometimes the variables and corresponding parameters in the regression can be logically split into two groups, so that the regression takes form

$$y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon,$$

where X_1 and X_2 have dimensions $n \times p_1$, $n \times p_2$, and β_1, β_2 are $p_1 \times 1$ and $p_2 \times 1$ vectors, with $p_1 + p_2 = p$.

The **Frisch–Waugh–Lovell theorem** states that in this regression the residuals $\hat{\varepsilon}$ and the OLS estimate $\hat{\beta}_2$ will be numerically identical to the residuals and the OLS estimate for β_2 in the following regression:[30]

$$M_1 y = M_1 X_2 \beta_2 + \eta \,,$$

where M_1 is the annihilator matrix for regressors X_1.

The theorem can be used to establish a number of theoretical results. For example, having a regression with a constant and another regressor is equivalent to subtracting the means from the dependent variable and the regressor and then running the regression for the de-meaned variables but without the constant term.

Constrained estimation

Suppose it is known that the coefficients in the regression satisfy a system of linear equations

$$A\colon \quad Q^T\beta = c,$$

where Q is a $p\times q$ matrix of full rank, and c is a $q\times 1$ vector of known constants, where $q < p$. In this case least squares estimation is equivalent to minimizing the sum of squared residuals of the model subject to the constraint A. The **constrained least squares (CLS)** estimator can be given by an explicit formula:[31]

$$\hat\beta^c = \hat\beta - (X^TX)^{-1}Q\Big(Q^T(X^TX)^{-1}Q\Big)^{-1}(Q^T\hat\beta - c).$$

This expression for the constrained estimator is valid as long as the matrix X^TX is invertible. It was assumed from the beginning of this article that this matrix is of full rank, and it was noted that when the rank condition fails, β will not be identifiable. However it may happen that adding the restriction A makes β identifiable, in which case one would like to find the formula for the estimator. The estimator is equal to [32]

$$\hat\beta^c = R(R^TX^TXR)^{-1}R^TX^Ty + \Big(I_p - R(R^TX^TXR)^{-1}R^TX^TX\Big)Q(Q^TQ)^{-1}c,$$

where R is a $p\times(p-q)$ matrix such that the matrix $[Q\ R]$ is non-singular, and $R^TQ = 0$. Such a matrix can always be found, although generally it is not unique. The second formula coincides with the first in case when X^TX is invertible.[32]

Large sample properties

The least squares estimators are point estimates of the linear regression model parameters β. However, generally we also want to know how close those estimates might be to the true values of parameters. In other words, we want to construct the interval estimates.

Since we haven't made any assumption about the distribution of error term ε_i, it is impossible to infer the distribution of the estimators $\hat\beta$ and $\hat\sigma^2$. Nevertheless, we can apply the central limit theorem to derive their *asymptotic* properties as sample size n goes to infinity. While the sample size is necessarily finite, it is customary to assume that n is "large enough" so that the true distribution of the OLS estimator is close to its asymptotic limit.

We can show that under the model assumptions, the least squares estimator for β is consistent (that is $\hat\beta$ converges in probability to β) and asymptotically normal:[proof]

$$(\hat\beta - \beta) \xrightarrow{d} \mathcal{N}\big(0,\ \sigma^2Q_{xx}^{-1}\big),$$

where $Q_{xx} = X^TX$.

Intervals

Using this asymptotic distribution, approximate two-sided confidence intervals for the j-th component of the vector $\hat\beta$ can be constructed as

$$\beta_j \in \left[\hat\beta_j \pm q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}}\sqrt{\hat\sigma^2\left[Q_{xx}^{-1}\right]_{jj}}\right] \quad \text{at the } 1-\alpha \text{ confidence level},$$

where q denotes the quantile function of standard normal distribution, and $[\cdot]_{jj}$ is the j-th diagonal element of a matrix.

Similarly, the least squares estimator for σ^2 is also consistent and asymptotically normal (provided that the fourth moment of ε_i exists) with limiting distribution

$$(\hat\sigma^2 - \sigma^2) \xrightarrow{d} \mathcal{N}\left(0,\ \mathrm{E}\left[\varepsilon_i^4\right] - \sigma^4\right).$$

These asymptotic distributions can be used for prediction, testing hypotheses, constructing other estimators, etc.. As an example consider the problem of prediction. Suppose x_0 is some point within the domain of distribution of the regressors, and one wants to know what the response variable would have been at that point. The mean response is the quantity $y_0 = x_0^T\beta$, whereas the predicted response is $\hat y_0 = x_0^T\hat\beta$. Clearly the predicted response is a random variable, its distribution can be derived from that of $\hat\beta$:

$$(\hat{y}_0 - y_0) \xrightarrow{a} \mathcal{N}\left(0,\ \sigma^2 x_0^{\mathrm{T}} Q_{xx}^{-1} x_0\right),$$

which allows construct confidence intervals for mean response y_0 to be constructed:

$$y_0 \in \left[x_0^{\mathrm{T}}\hat{\beta} \pm q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sqrt{\hat{\sigma}^2 x_0^{\mathrm{T}} Q_{xx}^{-1} x_0} \right] \quad \text{at the } 1-\alpha \text{ confidence level.}$$

Hypothesis testing

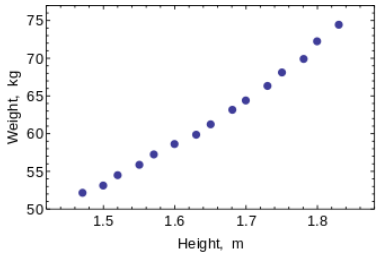
Two hypothesis tests are particularly widely used. First, one wants to know if the estimated regression equation is any better than simply predicting that all values of the response variable equal its sample mean (if not, it is said to have no explanatory power). The null hypothesis of no explanatory value of the estimated regression is tested using an F-test. If the calculated F-value is found to be large enough to exceed its critical value for the pre-chosen level of significance, the null hypothesis is rejected and the alternative hypothesis, that the regression has explanatory power, is accepted. Otherwise, the null hypothesis of no explanatory power is accepted.

Second, for each explanatory variable of interest, one wants to know whether its estimated coefficient differs significantly from zero—that is, whether this particular explanatory variable in fact has explanatory power in predicting the response variable. Here the null hypothesis is that the true coefficient is zero. This hypothesis is tested by computing the coefficient's t-statistic, as the ratio of the coefficient estimate to its standard error. If the t-statistic is larger than a predetermined value, the null hypothesis is rejected and the variable is found to have explanatory power, with its coefficient significantly different from zero. Otherwise, the null hypothesis of a zero value of the true coefficient is accepted.

In addition, the Chow test is used to test whether two subsamples both have the same underlying true coefficient values. The sum of squared residuals of regressions on each of the subsets and on the combined data set are compared by computing an F-statistic; if this exceeds a critical value, the null hypothesis of no difference between the two subsets is rejected; otherwise, it is accepted.

Example with real data

The following data set gives average heights and weights for American women aged 30–39 (source: *The World Almanac and Book of Facts, 1975*).



Scatterplot of the data, the relationship is slightly curved but close to linear

Height (m)	1.47	1.50	1.52	1.55	1.57	1.60	1.63	1.65	1.68	1.70	1.73	1.75	1.78	1.80	1.83
Weight (kg)	52.21	53.12	54.48	55.84	57.20	58.57	59.93	61.29	63.11	64.47	66.28	68.10	69.92	72.19	74.46

When only one dependent variable is being modeled, a scatterplot will suggest the form and strength of the relationship between the dependent variable and regressors. It might also reveal outliers, heteroscedasticity, and other aspects of the data that may complicate the interpretation of a fitted regression model. The scatterplot suggests that the relationship is strong and can be approximated as a quadratic function. OLS can handle non-linear relationships by introducing the regressor `HEIGHT`^{2}. The regression model then becomes a multiple linear model:

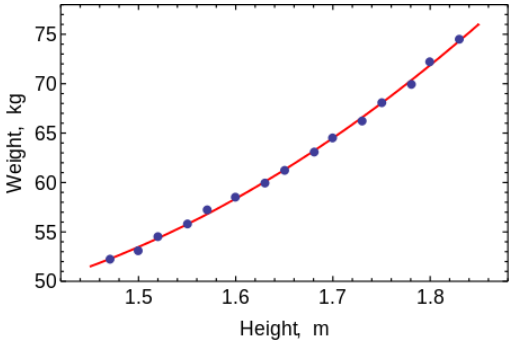
$$w_i = \beta_1 + \beta_2 h_i + \beta_3 h_i^2 + \varepsilon_i.$$

The output from most popular statistical packages will look similar to this:

Method	Least squares
Dependent variable	WEIGHT
Observations	15

Parameter	Value	Std error	t-statistic	p-value
β_1	128.8128	16.3083	7.8986	0.0000
β_2	−143.1620	19.8332	−7.2183	0.0000
β_3	61.9603	6.0084	10.3122	0.0000

R^2	0.9989	S.E. of regression	0.2516
Adjusted R^2	0.9987	Model sum-of-sq.	692.61
Log-likelihood	1.0890	Residual sum-of-sq.	0.7595
Durbin–Watson stat.	2.1013	Total sum-of-sq.	693.37
Akaike criterion	0.2548	F-statistic	5471.2
Schwarz criterion	0.3964	p-value (F-stat)	0.0000



Fitted regression

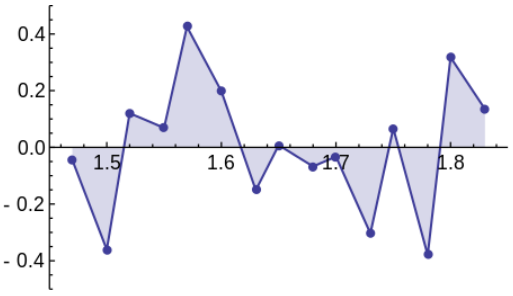
In this table:

- The *Value* column gives the least squares estimates of parameters β_j
- The *Std error* column shows standard errors of each coefficient estimate: $\hat{\sigma}_j = \left(\hat{\sigma}^2 \left[Q_{xx}^{-1}\right]_{jj}\right)^{\frac{1}{2}}$
- The *t-statistic* and *p-value* columns are testing whether any of the coefficients might be equal to zero. The *t*-statistic is calculated simply as $t = \hat{\beta}_j / \hat{\sigma}_j$. If the errors ε follow a normal distribution, *t* follows a Student-t distribution. Under weaker conditions, *t* is asymptotically normal. Large values of *t* indicate that the null hypothesis can be rejected and that the corresponding coefficient is not zero. The second column, *p*-value, expresses the results of the hypothesis test as a significance level. Conventionally, *p*-values smaller than 0.05 are taken as evidence that the population coefficient is nonzero.
- *R-squared* is the coefficient of determination indicating goodness-of-fit of the regression. This statistic will be equal to one if fit is perfect, and to zero when regressors *X* have no explanatory power whatsoever. This is a biased estimate of the population *R-squared*, and will never decrease if additional regressors are added, even if they are irrelevant.
- *Adjusted R-squared* is a slightly modified version of R^2, designed to penalize for the excess number of regressors which do not add to the explanatory power of the regression. This statistic is always smaller than R^2, can decrease as new regressors are added, and even be negative for poorly fitting models:

$$\overline{R}^2 = 1 - \frac{n-1}{n-p}(1-R^2)$$

- *Log-likelihood* is calculated under the assumption that errors follow normal distribution. Even though the assumption is not very reasonable, this statistic may still find its use in conducting LR tests.
- *Durbin–Watson statistic* tests whether there is any evidence of serial correlation between the residuals. As a rule of thumb, the value smaller than 2 will be an evidence of positive correlation.
- *Akaike information criterion* and *Schwarz criterion* are both used for model selection. Generally when comparing two alternative models, smaller values of one of these criteria will indicate a better model.[33]
- *Standard error of regression* is an estimate of *σ*, standard error of the error term.
- *Total sum of squares*, *model sum of squared*, and *residual sum of squares* tell us how much of the initial variation in the sample were explained by the regression.
- *F-statistic* tries to test the hypothesis that all coefficients (except the intercept) are equal to zero. This statistic has *F*(*p*–1,*n*–*p*) distribution under the null hypothesis and normality assumption, and its *p-value* indicates probability that the hypothesis is indeed true. Note that when errors are not normal this statistic becomes invalid, and other tests such as Wald test or LR test should be used.

Ordinary least squares analysis often includes the use of diagnostic plots designed to detect departures of the data from the assumed form of the model. These are some of the common diagnostic plots:



Residuals plot

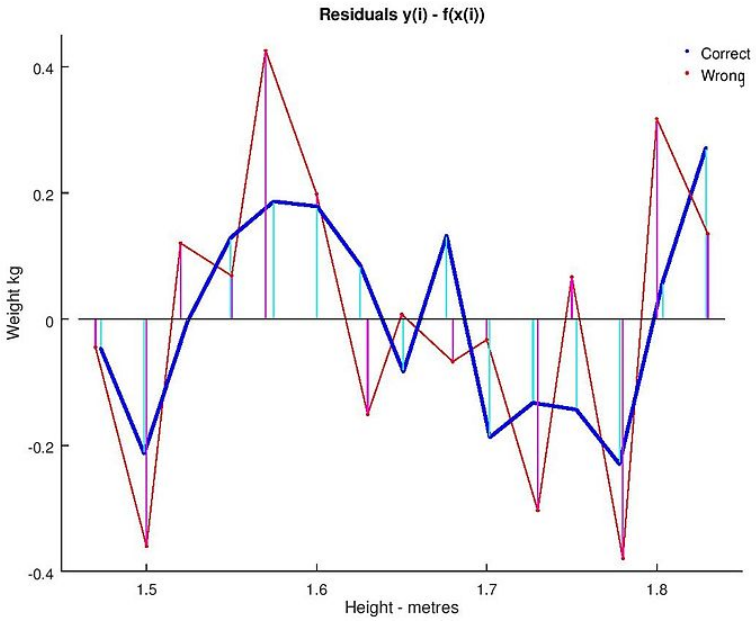
- Residuals against the explanatory variables in the model. A non-linear relation between these variables suggests that the linearity of the conditional mean function may not hold. Different levels of variability in the residuals for different levels of the explanatory variables suggests possible heteroscedasticity.
- Residuals against explanatory variables not in the model. Any relation of the residuals to these variables would suggest considering these variables for inclusion in the model.
- Residuals against the fitted values, $\hat{y}$.
- Residuals against the preceding residual. This plot may identify serial correlations in the residuals.

An important consideration when carrying out statistical inference using regression models is how the data were sampled. In this example, the data are averages rather than measurements on individual women. The fit of the model is very good, but this does not imply that the weight of an individual woman can be predicted with high accuracy based only on her height.

Sensitivity to rounding

This example also demonstrates that coefficients determined by these calculations are sensitive to how the data is prepared. The heights were originally given rounded to the nearest inch and have been converted and rounded to the nearest centimetre. Since the conversion factor is one inch to 2.54 cm this is *not* an exact conversion. The original inches can be recovered by Round(x/0.0254) and then re-converted to metric without rounding. If this is done the results become:

	Const	Height	Height^2
Converted to metric with rounding.	128.8128	−143.162	61.96033
Converted to metric without rounding.	119.0205	−131.5076	58.5046



Residuals to a quadratic fit for correctly and incorrectly converted data.

Using either of these equations to predict the weight of a 5' 6" (1.6764m) woman gives similar values: 62.94 kg with rounding vs. 62.98 kg without rounding. Thus a seemingly small variation in the data has a real effect on the coefficients but a small effect on the results of the equation.

While this may look innocuous in the middle of the data range it could become significant at the extremes or in the case where the fitted model is used to project outside the data range (extrapolation).

This highlights a common error: this example is an abuse of OLS which inherently requires that the errors in the independent variable (in this case height) are zero or at least negligible. The initial rounding to nearest inch plus any actual measurement errors constitute a finite and non-negligible error. As a result, the fitted parameters are not the best estimates they are presumed to be. Though not totally spurious the error in the estimation will depend upon relative size of the x and y errors.

See also

- Bayesian least squares
- Fama–MacBeth regression
- Non-linear least squares
- Numerical methods for linear least squares
- Nonlinear system identification

References

1. Goldberger, Arthur S. (1964). "Classical Linear Regression" (https://books.google.com/books?id=KZq5AAAAIAAJ&pg=PA156). *Econometric*

Theory. New York: John Wiley & Sons. pp. 156–212 [p. 158]. ISBN 0-471-31101-4.

2. Hayashi (2000, page 15)
3. Hayashi (2000, page 18)
4. [1] (https://books.google.com.br/books?id=hZ4mAOXVowoC&lpg=PA538&dq=cofactor&pg=PA160#v=onepage&q&f=false)
5. [2] (https://books.google.com.br/books?id=Np7y43HU_m8C&lpg=PA227&dq=cofactor%20matrix%20least%20squares&pg=PA263#v=onepage&q&f=false)
6. [3] (https://books.google.com.br/books?id=peYFZ69HqEsC&lpg=PA151&dq=cofactor%20matrix%20least%20squares&pg=PA134#v=onepage&q&f=false)
7. Hayashi (2000, page 19)
8. Julian Faraway (2000), *Practical Regression and Anova using R*, [4] (https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf)
9. Kenney, J.; Keeping, E. S. (1963). *Mathematics of Statistics*. van Nostrand. p. 187.
10. Zwillinger, D. (1995). *Standard Mathematical Tables and Formulae*. Chapman&Hall/CRC. p. 626. ISBN 0-8493-2479-3.
11. Hayashi (2000, page 20)
12. Hayashi (2000, page 5)
13. Akbarzadeh, Vahab. "Line Estimation" (http://mlmadesimple.com/2014/05/07/line-estimation/).
14. Hayashi (2000, page 49)
15. Hayashi (2000, page 52)
16. Hayashi (2000, page 7)
17. Hayashi (2000, page 187)
18. Hayashi (2000, page 10)
19. Hayashi (2000, page 34)
20. Williams, M. N; Grajales, C. A. G; Kurkiewicz, D (2013). "Assumptions of multiple regression: Correcting two misconceptions" (http://www.pareonline.net/getvn.asp?v=18&n=11). *Practical Assessment, Research & Evaluation*. **18** (11).
21. Hayashi (2000, pages 27, 30)
22. Hayashi (2000, page 27)
23. Amemiya (1985, page 13)
24. Amemiya (1985, page 14)
25. Rao (1973, page 319)
26. Amemiya (1985, page 20)
27. Amemiya (1985, page 27)
28. Davidson & Mackinnon (1993, page 33)
29. Davidson & Mackinnon (1993, page 36)
30. Davidson & Mackinnon (1993, page 20)
31. Amemiya (1985, page 21)
32. Amemiya (1985, page 22)
33. Burnham, Kenneth P.; David Anderson (2002). *Model Selection and Multi-Model Inference* (2nd ed.). Springer. ISBN 0-387-95364-7.

Further reading

- Amemiya, Takeshi (1985). *Advanced econometrics*. Harvard University Press. ISBN 0-674-00560-0.
- Davidson, Russell; Mackinnon, James G. (1993). *Estimation and inference in econometrics*. Oxford University Press. ISBN 978-0-19-506011-9.
- Greene, William H. (2002). *Econometric analysis* (http://stat.smmu.edu.cn/DOWNLOAD/ebook/econometric.pdf) (PDF) (5th ed.). New Jersey: Prentice Hall. ISBN 0-13-066189-9. Retrieved 2016-01-13.
- Hayashi, Fumio (2000). *Econometrics*. Princeton University Press. ISBN 0-691-01018-8.
- Rao, C.R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley & Sons.
- Wooldridge, Jeffrey M. (2013). *Introductory Econometrics: A Modern Approach* (5th international ed.). Australia: South Western, Cengage Learning. ISBN 9781111534394.