Multiple Linear Regression I

Emilija Perković

Dept. of Statistics
University of Washington

Notation

Let n be the sample size, y be a dependent variable, and $x_1, \dots, x_p$ be independent variables. The multiple linear regression model is often written as:

$$\underline{y} = \beta_0 + \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \cdots + \beta_p \underline{x}_p + \underline{\epsilon} \qquad (1)$$

where

- $\underline{y} = (y_1, \dots, y_n)'$ is the vector of observations on y,
- $\underline{x}_k = (x_{1k}, \dots, x_{nk})'$ is the vector of observations on covariate x_k, $k = 1, \dots, p$,
- $(\beta_0, \beta_1, \dots, \beta_p)'$ is the vector of regression coefficients, and
- $\underline{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ is the vector of errors such that $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.

Notation

Note that assumptions of linear model in Equation (1) (i.e., errors ϵ_i are independent and identically distributed $\mathcal{N}(0, \sigma^2)$) are stated compactly in matrix form:

$$\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n),$$

where $\underline{0} = (0, \ldots, 0)'$ is $n \times 1$ vector and

$$I_n = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \ldots & 0 & 1 \end{bmatrix} \text{ is the } n \times n \text{ identity matrix.}$$

Notation

Equation (1) stands for the system of n equations:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_1$$
$$\ldots$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \epsilon_n$$

which can also be written in matrix notation as

$$\underline{y} = X\underline{\beta} + \underline{\epsilon} \tag{2}$$

Notation

Writing out the matrices and vectors, we have an equivalent formulation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Dimensions:

$n \times 1$ $\quad$ $n \times (p+1)$ $\quad$ $(p+1) \times 1$ $\quad$ $n \times 1$

Notation

To summarize, the multiple linear regression model can be written in matrix form

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

where

- $\underline{y}$ is a dependent variable,
- X is a design matrix,
- $\underline{\beta}$ is a parameter vector to be estimated, and
- $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.

Mean and variance of response

We can obtain the mean and variance of response vector by using matrix notation and the properties of expectation and variance:

$$\mathrm{E}[\underline{y}] = \mathrm{E}[X\underline{\beta} + \underline{\epsilon}] = X\underline{\beta} + \mathrm{E}[\underline{\epsilon}] = X\underline{\beta},$$
$$\mathrm{Var}[\underline{y}] = \mathrm{Var}[X\underline{\beta} + \underline{\epsilon}] = \mathrm{Var}[\underline{\epsilon}] = \sigma^2 I_n.$$

In fact:

$$\underline{y} \sim \mathcal{N}(X\underline{\beta}, \sigma^2 I_n).$$

Mean and variance of response

Without the use of matrix notation, we can derive the mean and variance functions of y_i. Treating the unknown parameters and observed covariate values as constants, we obtain for the mean:

$$\begin{aligned}\mathrm{E}[y_i] &= \mathrm{E}[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i],\\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \mathrm{E}[\epsilon_i]\\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},\end{aligned}$$

where the last line uses $\mathrm{E}[\epsilon_i] = 0$.
For the variance:

$$\mathrm{Var}[y_i] = \mathrm{Var}[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i] = \mathrm{Var}[\epsilon_i] = \sigma^2.$$

What about covariance $\mathrm{Cov}[y_i, y_j]$?

$$\mathrm{Cov}[y_i, y_j] = \cdots = \mathrm{Cov}[\epsilon_i, \epsilon_j] = 0.$$

OLS

The residual sum of squares as a function of $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is

$$RSS(\underline{\beta}) = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta}).$$

Values $\hat{\underline{\beta}}$ that minimize $RSS(\underline{\beta})$ are called ordinary least squares (OLS) estimates.
To minimize $RSS(\underline{\beta})$ with respect to $\underline{\beta}$ note that

$$\begin{aligned} RSS(\underline{\beta}) &= (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta}) \\ &= \underline{y}'\underline{y} - \underline{\beta}'X'\underline{y} - \underline{y}'X\underline{\beta} + \underline{\beta}'X'X\underline{\beta} \\ &= \underline{y}'\underline{y} - 2(\underline{y}'X)\underline{\beta} + \underline{\beta}'X'X\underline{\beta} \end{aligned}$$

OLS

Next, we find the partial derivative with respect to $\underline{\beta}$:

$$\frac{\partial RSS(\underline{\beta})}{\partial \underline{\beta}} = -2(X'\underline{y}) + 2X'X\underline{\beta}$$

and set the derivative to zero to produce a system of normal equations. The solution of this system of normal equations is $\hat{\underline{\beta}}$.

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}.$$

OLS

The system of normal equations contains $(p+1)$ equations and $(p+1)$ unknowns.
If matrix $X'X$ is non-singular (i.e., $rank(X'X) = p+1$), we can obtain the least squares estimates as

$$\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y} \tag{3}$$

for a given design matrix X and observed response vector $\underline{y}$.
Note: $rank(X) = rank(X'X)$ and $rank(X) = p+1$ if and only if

- there are more distinct data points than parameters in the model and
- the $p+1$ columns of design matrix X are linearly independent.

Example of issues: Temperature is recorded in both degrees of Fahrenheit and Celsius, and both variables are in the model.

OLS properties

Properties of the least squares estimate $\underline{\hat{\beta}}$ include:

1. $\underline{\hat{\beta}}$ is a linear function of $\underline{y}$.
2. $\underline{\hat{\beta}}$ is unbiased, $\mathrm{E}[\underline{\hat{\beta}}] = \underline{\beta}$.
3. $\mathrm{Var}[\underline{\hat{\beta}}] = \sigma^2(X'X)^{-1}$.
4. Gauss-Markov Theorem.

OLS properties

OLS property I: $\underline{\hat{\beta}}$ is a linear function of $\underline{y}$,
follows from:

$$\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}.$$

OLS property II: $\underline{\hat{\beta}}$ is unbiased, $\mathrm{E}[\underline{\hat{\beta}}] = \underline{\beta}$.
Recall, the rule for expectation of a random vector $\underline{v}$: $\mathrm{E}[A\underline{v}] = A\mathrm{E}[\underline{v}]$,
where A is a constant matrix.
Proof:

$$\begin{aligned}\mathrm{E}[\underline{\hat{\beta}}] &= \mathrm{E}[(X'X)^{-1}X'\underline{y}] \\ &= (X'X)^{-1}X'\mathrm{E}[\underline{y}] \\ &= (X'X)^{-1}X'X\underline{\beta} = \underline{\beta}.\end{aligned}$$

Notation Distribution of response OLS OLS properties Interpretation

OLS properties

OLS property III: $\mathrm{Var}[\underline{\hat{\beta}}] = \sigma^2(X'X)^{-1}.$,
Recall, the rule for variance of a random vector $\underline{v}$:
$\mathrm{Var}[A\underline{v}] = A\mathrm{Var}[\underline{v}]A'$, where A is a constant matrix.
Proof: (Assuming that σ^2 is known)

$$\begin{aligned}\mathrm{Var}[\underline{\hat{\beta}}] &= (X'X)^{-1}X'\mathrm{Var}[\underline{y}][(X'X)^{-1}X']' \\ &= (X'X)^{-1}X'\sigma^2 I_n[(X'X)^{-1}X']' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}.\end{aligned}$$

In practice, we usually estimate σ^2 with:

$$\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}.$$

OLS properties

OLS property IV: **Gauss-Markov Theorem** The least squares estimator $\underline{\hat{\beta}}$ has the smallest sampling variance among the class of linear unbiased estimators.

If we assume $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$, then, using property I, we obtain that $\underline{\hat{\beta}}$ is multivariate normal:

$$\underline{\hat{\beta}} \sim \mathcal{N}(\underline{\beta}, \sigma^2 (X'X)^{-1}).$$

Note: to derive the mean and variance of $\underline{\hat{\beta}}$ we did not require the assumption of normality but only assumptions of linearity, constant variance, and independence.

Geometric illustration

Consider the linear regression model with two predictors:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}, \text{ where } X = (\underline{1}, \underline{x}_1, \underline{x}_2).$$

Assume the observed data are:

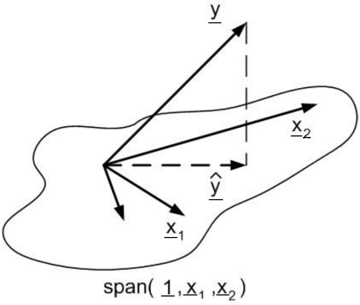
$$\underline{y} = (y_1, \dots, y_n)', \quad \underline{x}_1 = (x_{11}, \dots, x_{n1})' \quad \underline{x}_2 = (x_{12}, \dots, x_{n2})'.$$

The least squares estimate of $\underline{\beta}$ minimizes the squared distance:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = ||\underline{y} - \underline{\hat{y}}||^2,$$

that is, the Euclidean distance between $\underline{y}$ and $\underline{\hat{y}}$.

Geometric illustration



Geometric illustration

The fitted values are given by:

$$\underline{\hat{y}} = X\underline{\hat{\beta}} = X(X'X)^{-1}X'\underline{y}.$$

Since the vector of fitted values $\underline{\hat{y}}$ is a linear combinations of vectors in the design matrix X, $\underline{\hat{y}}$ belongs to span$(\underline{1}, \underline{x}_1, \underline{x}_2)$.
See Linear Models Handout for more details.
One can use geometry to understand what some notions in regression mean, for example:
Small eigenvalues of $X'X$ correspond to collinearity.
Collinearity may happen when the angle between $\underline{x}_1$ and $\underline{x}_2$ is very small. If the angle between $\underline{x}_1$ and $\underline{x}_2$ is small, this means that the hyperplane span$(\underline{1}, \underline{x}_1, \underline{x}_2)$ will be very sensitive to small changes and hence not reliable.

Interpretation of regression parameters

Example: Fuel consumption

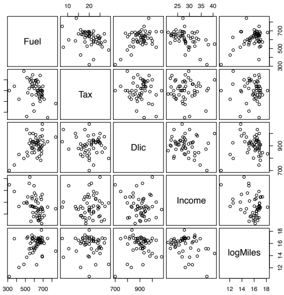
What is the effect of the state gasoline tax on fuel consumption?

Variables:

- Dlic - 1000×[number of licensed drivers in the state]/[population of the state older than 16 in 2001].
- Income - yearly personal income in the year 2000.
- Fuel - 1000×[gasoline sold in thousands of gallons]/[population of the state older than 16 in 2001].
- logMiles - log(Miles), where Miles denotes the miles of Federal-aid highway in the state.
- Tax - Gasoline state tax rate in cents per gallon.

Data: `fuel2001` from R package `alr4`.

Example: Fuel consumption



Example: Fuel consumption

```
lm(formula = Fuel ~ Tax + Dlic + Income + logMiles,
data = new.fuel)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.192845 194.906161   0.791 0.432938
Tax          -4.227983   2.030121  -2.083 0.042873
Dlic          0.471871   0.128513   3.672 0.000626
Income       -0.006135   0.002194  -2.797 0.007508
logMiles     26.755176   9.337374   2.865 0.006259
---

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-squared:  0.5105,Adjusted R-squared:  0.4679
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```


Example: Fuel consumption

Interpret coefficient $\hat{\beta}_1$.

For every unit increase of x_1, **while** $x_2, \dots, x_p$ **are held constant**, we estimate that y increases by $\hat{\beta}_1$ on average.
Analogous interpretation for coefficients $\hat{\beta}_2, \dots, \hat{\beta}_p$.

Interpret coefficient $\hat{\beta}_0$.

The estimated average value of y for $(x_1, \dots, x_p) = 0$.
Be careful with extrapolation!

Interpretation of regression coefficients

Note that performing a multiple linear regression **will not in general** produce the same coefficient estimates as performing many simple linear regressions.

```
lm(formula = Fuel ~ Tax, data = new.fuel)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  715.485     55.770  12.829   <2e-16
Tax           -5.078      2.701  -1.881    0.066 .
---
Residual standard error: 86.79 on 49 degrees of freedom
Multiple R-squared:  0.06731,Adjusted R-squared:  0.04828
F-statistic: 3.536 on 1 and 49 DF,  p-value: 0.06599
```

Exception: If the sample correlation between the predictor vectors is 0. Usually only true in some designed experiments.