STA3010 Regression Analysis

Feng YIN

The Chinese University of Hong Kong (Shenzhen)

yinfeng@cuhk.edu.cn

February 10, 2020

Overview

1. Multiple Linear Regression
   - Multiple Linear Regression Model
   - Least-Squares (LS) Parameter Estimation
   - Maximum-likelihood (ML) Parameter Estimation
   - Hypothesis Testing on Parameters
   - Coefficient of Determination

Multiple Linear Regression Model

A multiple linear regression model is given by

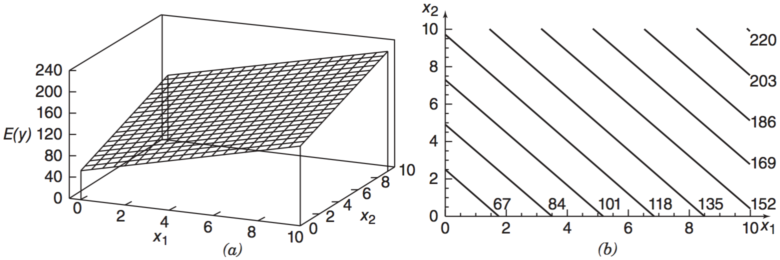
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{1}$$

where

- unknown model parameters $\beta_j, j = 1, 2, ..., k$ are often called regression coefficients
- $x_j, j = 1, 2, ..., k$ are the inputs/regressors and y is the output/response
- input $x_j, j = 1, 2, ..., k$ are deterministic and precisely known
- ε is random error term with zero mean and unknown variance σ^2 (Note that, no specific error distribution is assumed herein).

An Illustrating Example

In the following example, we have two inputs/regressors, x_1 and x_2.



(a) The regression plane for the model $E(y) = 50 + 10x_1 + 7x_2$. (b) The contour plot. Source: textbook.

More Complicated Representation

Examples:

1. Polynomial model (in one input variable):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon \tag{2}$$

2. Interaction model (in two input variables):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \tag{3}$$

In the first example, if we set $x_j = x^j$; and in the second example, if we set $\beta_{12} = \beta_3, x_3 = x_1 x_2$, they can be rewritten as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{4}$$

Key message: Any regression model that is linear in the parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_k]$ is a linear regression model, regardless of the functional shape it demonstrates.

Multiple Linear Regression Model

Assume that we have in total n observations, the above regression model can be written in a compact matrix form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5}$$

where

$$\mathbf{y} = [y_1, y_2, ..., y_n]^T \tag{6}$$

$$\mathbf{x}_j = [x_{1,j}, x_{2,j}, ..., x_{n,j}]^T, j \in \{1, ..., k\} \tag{7}$$

$$\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k] \tag{8}$$

$$\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_k]^T \tag{9}$$

$$\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T \tag{10}$$

$\mathbf{y}$ is an $n \times 1$ vector of the observations, $\mathbf{X}$ is an $n \times p$ $(\text{note:} p = k + 1)$ matrix of the regressor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the model parameters, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of uncorrelated random error terms.

LS Parameter Estimation: β

Derive (in compact matrix form) the least-squares (LS) estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \triangleq (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{11}$$

The LS estimator (in matrix form) is given by

$$\frac{\partial S}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0} \tag{12}$$

which simplifies to

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \tag{13}$$

If $(\mathbf{X}^T\mathbf{X})^{-1}$ exists, we finally have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{14}$$

LS Parameter Estimation: σ^2

As a result, the vector of fitted values $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_n]^T$ is computed by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y} \tag{15}$$

where the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the hat matrix, which plays an important role in regression analysis.

Consequently, the vector of residuals, $\mathbf{e} = [e_1, e_2, ..., e_n]^T$, is computed as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{16}$$

It can be easily verified that both $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are idempotent matrices.

LS Parameter Estimation: σ^2

Moreover, **residual/error sum of squares** is defined to be

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^T \mathbf{e} \tag{17}$$

An estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-p} = MS_{Res} \tag{18}$$

Properties of the LS Estimator: $\hat{\sigma}^2$

When we perform the following analyses, it is assumed that the data is truly from the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

It can be proven that :

1. An unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-p} = MS_{Res}. \tag{19}$$

2. under the assumption that the error $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, SS_{Res}/σ^2 follows χ^2_{n-p} distribution, where $p = k + 1$.

Properties of the LS Estimator: $\hat{\sigma}^2$

1. To prove $\frac{SS_{Res}}{n-p} = MS_{Res}$ is an unbiased estimator of σ^2, we need:

Theorem 1

Let $\mathbf{A}$ be a $k \times k$ matrix of constants and $\mathbf{y}$ be a $k \times 1$ multivariate random vector with mean $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$. Let U be the quadratic form defined by $U = \mathbf{y}^T\mathbf{A}\mathbf{y}$, then $E(U) = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$.

2. To prove under another assumption that the error $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, SS_{Res}/σ^2 follows χ^2_{n-p} distribution, we need:

Theorem 2

Let $\mathbf{A}$ be a $k \times k$ idempotent matrix of constants with rank p' and $\mathbf{y}$ be a $k \times 1$ multivariate Gaussian random vector with mean $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. Let U be the quadratic form defined by $U = \mathbf{y}^T\mathbf{A}\mathbf{y}$, then $\frac{U}{\sigma^2} \sim \chi^{2'}_{p',\lambda'}$, where $\lambda' = \frac{\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}}{\sigma^2}$.

Properties of the LS Estimator: $\hat{\boldsymbol{\beta}}$

When we perform the following analyses, it is assumed that the data is truly from the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Prove (on the white board and in the manuscript) that:

1. $E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$
2. $\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\mathbf{C} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, where the variance of $\hat{\beta}_j$ is $\sigma^2 C_{jj}$ and the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$ is $\sigma^2 C_{ij}$
3. the LS estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ (Gauss-Markov Theorem)

Gauss-Markov Theorem: Special Case

Gauss-Markov Theorem, $E(\varepsilon) = \mathbf{0}$, $Cov(\varepsilon) = \sigma^2\mathbf{I}$

If the observations are of the general linear model form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{20}$$

where $\mathbf{X}$ is a known $n \times p$ matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of model parameters/regression coefficients to be fitted, and $\boldsymbol{\varepsilon}$ (the probability density function can be arbitrary) is a $n \times 1$ vector with zero mean and covariance matrix $\sigma^2\mathbf{I}$, then the BLUE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{21}$$

and the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$Cov\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \tag{22}$$

Maximum-Likelihood (ML) Estimation

LS estimation

The LS estimation does not assume any specific distribution of the random error ε

ML estimation

The ML estimation does assume a known statistical distribution of the random error ε

The idea of the ML estimation is to find the point estimate of $\boldsymbol{\beta}$ that maximizes the likelihood function, defined to be $p(\mathbf{y}; \boldsymbol{\beta})$ herein, namely the value that makes the observed data (i.e., the output) the most probable!

Maximum-Likelihood (ML) Estimation

For the multiple linear regression model with i.i.d. Gaussian errors, i.e., $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, the likelihood function is

$$p(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left[\frac{-(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)}{2\sigma^2}\right] \tag{23}$$

Very often, we take the logarithm of the likelihood function, short for log-likelihood, namely

$$\ln p(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) \tag{24}$$

Taking the derivative of $\ln p(\mathbf{y}; \boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ and σ^2 respectively and setting them equal to zero, yields:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \qquad \hat{\sigma}^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})}{n} \tag{25}$$

Maximum-Likelihood (ML) Estimation

ML estimator is asymptotically optimal!

Asymptotic Properties of the ML Estimator

If the probability density function $p(\mathbf{y}; \boldsymbol{\theta})$ of the output $\mathbf{y}$ satisfies some regularity conditions, then the ML estimator of the unknown parameters $\boldsymbol{\theta}$ is asymptotically distributed (i.e., for very large data records) according to

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}\left(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta})\right) \tag{26}$$

where $I(\boldsymbol{\theta})$ is the Fisher information evaluated at the true value of the unknown parameter and $I^{-1}(\boldsymbol{\theta})$ is well known as Cramer-Rao lower bound for benchmarking unbiased parameter estimators.

Reading recommendation: Steven Kay, Fundamentals of Statistical Signal Processing: Estimation Theory

Hypothesis Testing on Model Parameters

Additional Assumption

The error terms $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are Gaussian/normally and independently distributed with zero mean and variance σ^2.

Task-I: Test for significance of regression determines if there is a linear relationship between the output y and any of the inputs, $x_1, x_2, ..., x_k$.
Hypotheses:

$$H_0 : \beta_1 = ... = \beta_k = 0, \quad H_1 : \beta_j \neq 0 \text{ for at least one } j \tag{27}$$

We use analysis of variance (ANOVA) for this purpose. Recall that,

$$SS_T = SS_R + SS_{Res}. \tag{28}$$

- $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is called the corrected sum of squares
- $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is called regression/model sum of squares
- (Recall) $SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is called residual sum of squares

It can be proven (cf. the manuscript) that:

1. SS_{Res}/σ^2 follows χ^2_{n-p} distribution.
2. SS_R/σ^2 follows χ^2_k distribution if the null hypothesis H_0 is true.
3. SS_R and SS_{Res} are independent.

The F test statistic is constructed by

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-p)} \sim F_{k,n-p}. \tag{29}$$

Test procedure: If the observed value of F_0 is large, then it is likely that at least one slope $\beta_j \neq 0$. More precisely, if $F_0 > F_{c,k,n-p}$, we reject the null hypothesis H_0. Here, $F_{c,k,n-p}$ is the one-sided c percentage point.

To prove SS_R and SS_{Res} are independent, we need the following theorem:

Theorem 3

Let $\mathbf{A}$ and $\mathbf{B}$ be two $k \times k$ matrices of constants and $\mathbf{y}$ be a $k \times 1$ multivariate Gaussian random vector with mean $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$. Let U and V be the quadratic form defined by $U = \mathbf{y}^T\mathbf{A}\mathbf{y}$ and $V = \mathbf{y}^T\mathbf{B}\mathbf{y}$, respectively. The two quadratic forms, U and V, are independent if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B} = \mathbf{0}_{k\times k}$.

Hypothesis Testing on Model Parameters

Additional Assumption

The error terms $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are Gaussian/normally and independently distributed with zero mean and variance σ^2.

Task II: Tests on individual parameter estimator: Without loss of generality, the hypotheses for testing the significance of β_j are:

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0 \tag{30}$$

We use the following test statistic:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{MS_{Res} C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-p} \tag{31}$$

for which we need to prove:

1. $\frac{(n-p)MS_{Res}}{\sigma^2} = \frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p}$ (Proven already)
2. MS_{Res} and $\hat{\beta}_j$ are independent (similar to what we have proven before)

To prove MS_{Res} (or equivalently SS_{Res}) and $\hat{\beta}$ are independent, we need the following theorem:

Theorem 4

Let $\mathbf{B}$ be a $q \times k$ matrix of constants and let $\mathbf{W}$ be the linear form $\mathbf{W} = \mathbf{B}\mathbf{y}$, where $\mathbf{y}$ is a $k \times 1$ multivariate Gaussian random vector with mean $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$. Let U be the quadratic form defined by $U = \mathbf{y}^T\mathbf{A}\mathbf{y}$. U and $\mathbf{W}$ are independent if $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}_{q\times k}$.

Hypothesis Testing on Model Parameters

Test procedure:

1. compute a realization of t_0, given the observed data $\mathcal{S}$
2. compare the value of t_0 with the upper $c/2$ percentage point of the t_{n-p} distribution $t_{c/2,n-p}$
3. reject the null hypothesis $H_0 : \beta_j = 0$, if $|t_0| > t_{c/2,n-p}$

Note that: This is partial test because the model parameter estimator $\hat{\beta}_j$ depends on all of the other regressor variables that are in the model. Hence, this is a test of the contribution of x_j given other regressors in the model.

Hypothesis Testing on Model Parameters

Revisit the key findings on the above hypothesis test related to the significance of regression is:

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0 \tag{32}$$

Key conclusions:

- Accepting H_0 implies:
  - a there is NO linear relationship between x_j and y
  - b there may exist nonlinear relationship between x_j and y
- Rejecting H_0 implies:
  - a there is linear relationship between x_j and y
  - b there may exist nonlinear relationship between x_j and y

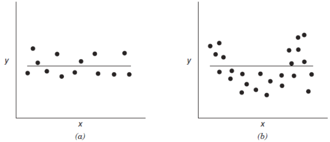


Figure 2.2 Situations where the hypothesis H_0: $\beta_1 = 0$ is not rejected.

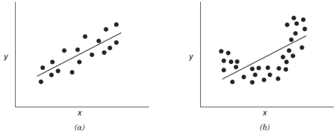


Figure 2.3 Situations where the hypothesis H_0: $\beta_1 = 0$ is rejected.

Coefficient of Determination

Two other metrics that in some sense reflects the model adequacy are

1. coefficient of determination:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \tag{33}$$

   Note that $0 \leq R^2 \leq 1$.

2. adjusted coefficient of determination:

$$R^2_{adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)} \tag{34}$$

   where degrees of freedom are taken into account.

Summary

To summarize with some keywords:

1. Multiple linear regression model
2. Matrix formulation of the model
3. LS and ML parameter estimation
4. Gauss-Markov Theorem
5. Best-Linear-Unbiased-Estimator (BLUE)
6. ANalysis-Of-VAriance (ANOVA)
7. F-test versus T-test
8. Coefficient of determination