5-2 Multiple Linear Regression and Least Squares Estimates MAT 374

It is common for more than one factor to influence an outcome. Fitting regression models to data involving two or more predictors is one of the most widely used statistical procedures. In this chapter we consider multiple linear regression problems involving modeling the relationship between a dependent variable, Y and two or more predictor variables X_1, X_2, X_3 , etc.

Multiple linear regression model

A multiple linear regression (MLR) model that relates a response variable y to $p-1$ predictor variables is written as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where

$$\varepsilon_i \sim \text{Normal}(0, \sigma^2), \quad \text{independent.}$$

- The subscript i refers to the i^{th} individual (or unit) in the population.
- In x_{ij}, the first subscript i denotes the individual and the second subscript j denotes which predictor variable it is.
- The word "linear" in multiple linear regression refers to linearity in the *parameters* β_j.
- Predictors x_{ij} can be raw variables or transformations (e.g., x^2, interaction terms).
- The model includes $p-1$ predictors but p parameters (including the intercept β_0).

Interpretation of the parameters

– Each β_j represents the change in the mean response $E(Y)$ per unit increase in x_j, when all other predictors are held constant.

– β_0 represents the mean response when all predictors are equal to zero (may or may not be meaningful).

– For example: Consider the multiple regression function with two predictors:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

a. The parameter β_0 is the Y-intercept of the regression plane. If the scope of the model includes $X_1 = 0$ and $X_2 = 0$, then β_0 represents the mean response $E(Y)$ at that point. Otherwise, β_0 does not generally have any practical meaning as a separate term.

b. The parameter β_1 indicates the change in the mean response $E(Y)$ per unit increase in X_1 when X_2 is held constant. Say, $\beta_1 = 2$, then the mean response $E(Y)$ increases by 2 units for every unit increase in X_1, regardless of the level of X_2.

c. The parameter β_2 indicates the change in the mean response $E(Y)$ per unit increase in X_2 when X_1 is held constant. Similarly, say, $\beta_2 = 5$ indicates that the mean response increases by 5 units for every unit increase in X_2, when X_1 is held constant.

Matrix formulation

We now present the principal results for the multiple linear regression model in matrix terms. Define the following matrices:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}_{n \times p},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{p\times 1}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}.$$

– The multiple linear regression model can then be written concisely as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where:

- $\mathbf{Y}$ is the $n \times 1$ vector of responses,
- $\mathbf{X}$ is the $n \times p$ design matrix (including a column of ones for the intercept and one column for each of the $p-1$ predictors),
- $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, and
- $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of errors.

– We assume:

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

– Consequently, the random vector $\mathbf{Y}$ has

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \text{ and } \text{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n.$$

Estimation

– The least squares method is again used to estimate the regression coefficients. The estimator $\hat{\boldsymbol{\beta}}$ minimizes the sum of squared errors:

$$SSE = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

$$\frac{\partial SSE}{\partial \beta} \overset{\text{set}}{=} 0$$

$$-2\, X^T (Y - X\beta) = 0$$

$$X^T Y - X^T X \beta = 0$$

$$X^T X \beta = X^T Y$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

– Setting derivatives to zero yields the **normal equations**:

$$\mathbf{X}^\top \mathbf{X}\, \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y}.$$

– Provided $\mathbf{X}^\top \mathbf{X}$ is nonsingular, the least squares solution is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Properties of least squares estimates

– The expected value of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} E(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) &= E\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \;\middle|\; \mathbf{X}\right] \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(\mathbf{Y} \mid \mathbf{X}) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}) \\ &= \boldsymbol{\beta}. \end{aligned}$$

– The variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) &= \mathrm{Var}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \;\middle|\; \mathbf{X}\right] \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \mathrm{Var}(\mathbf{Y} \mid \mathbf{X})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}. \end{aligned}$$

– Since $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ and $\hat{\boldsymbol{\beta}}$ is a linear function of $\mathbf{Y}$, it follows that:

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right).$$

This distributional result provides the basis for statistical inference on regression coefficients, including the construction of confidence intervals and hypothesis testing (t-tests and F-tests).

Model evaluation

– Recall the following sum of squares from the simple linear regression

a. SST: the total corrected sum of squares of the Y's is given by $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

b. SSE: the residual sum of squares (error sum of squares) is given by $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

c. SSR: the regression sum of squares (i.e., sum of squares explained by the regression model) is given by $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

$I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$

– Define:

- $\bar{Y} = \frac{1}{n}JY$ where J is the $n \times n$ matrix of ones. $J = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$
- $\hat{Y} = HY$ where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix.
- Residuals: $e = Y - \hat{Y} = (I - H)Y$.

– We have:

a. $\left(I - \frac{1}{n}J\right)^\top\left(I - \frac{1}{n}J\right) = \left(I - \frac{1}{n}J\right)$

b. $H^\top H = H$ $\quad (AB)^T = B^T A^T$, $(X^TX)^T = X^TX$, $\left((X^TX)^{-1}\right)^T = (X^TX)^{-1}$

$H^TH = \left(X(X^TX)^{-1}X^T\right)^T\left(X(X^TX)^{-1}X^T\right) = X(X^TX)^{-1}X^T \cdot X(X^TX)^{-1}X^T = X\underbrace{(X^TX)^{-1}X^TX}_{I}(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T$

c. $(I - H)^\top(I - H) = (I - H)$

$$\begin{aligned} (I-H)^T(I-H) &= I^TI - I^TH - H^TI + H^TH \\ &= I - H - H^T + H \\ &= I - H - H + H \\ &= I - H \end{aligned}$$

– Using the matrix formulation, the three sums of squares for the multiple regression model are

a. $SST = (Y-\bar{Y})^T(Y-\bar{Y}) = (Y-\frac{1}{n}JY)^T(Y-\frac{1}{n}JY) = \left((I-\frac{1}{n}J)Y\right)^T\left(I-\frac{1}{n}J\right)Y = Y^T(I-\frac{1}{n}J)^T(I-\frac{1}{n}J)Y = Y^T(I-\frac{1}{n}J)Y$

b. $SSR = (HY-\bar{Y})^T(HY-\bar{Y}) = (HY-\frac{1}{n}JY)^T(HY-\frac{1}{n}JY) = \left((H-\frac{1}{n}J)Y\right)^T\left(H-\frac{1}{n}J\right)Y = Y^T(H-\frac{1}{n}J)^T(H-\frac{1}{n}J)Y = Y^T(H-\frac{1}{n}J)Y$

c. $SSE = (Y-HY)^T(Y-HY) = \left((I-H)Y\right)^T\left(I-H\right)Y = Y^T(I-H)^T(I-H)Y = Y^T(I-H)Y$

– The ANOVA table for the multiple regression model is

	SS	DF	MS
Regression	$SSR = Y^\top(H-\frac{1}{n}J)Y$	$p-1$	$MSR = \dfrac{SSR}{p-1}$
Error	$SSE = Y^\top(I-H)Y$	$n-p$	$MSE = \dfrac{SSE}{n-p}$
Total	$SST = Y^\top(I-\frac{1}{n}J)Y$	$n-1$	

– For the simple linear regression model, there is only one slope parameter about which one can perform hypothesis tests. And the F test from the ANOVA table is testing the slope parameter. However, for the multiple linear regression model, there are three different hypothesis tests for slopes that one could conduct. They are:

a. a hypothesis test for testing that one slope parameter is 0 $H_0: \beta_1 = 0$ v.s. $H_a: \beta_1 \neq 0$ (t-test)

b. a hypothesis test for testing that all of the slope parameters are 0 $H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ v.s. H_a: at least one Full rsn't. (F-test)

c. a hypothesis test for testing that a subset (more than one, but not all) of the slope parameters are 0
$H_0: \beta_2 = \beta_3 = 0$ v.s. H_a: at least one of β_2, β_3 isn't zero

– The above hypotheses can be tested using the analysis of variance F-test. The **general linear F-test** involves three basic steps:

a. Define a larger full model: By "larger," we mean one with more parameters (i.e., containing additional predictors).

b. Define a smaller reduced model: By "smaller," we mean one with fewer parameters (a restricted version of the full model).

c. Construct the F-statistic: to decide whether to reject the smaller (reduced) model in favor of the larger (full) model.

– In this framework, the null hypothesis H_0 always pertains to the *reduced model*, while the alternative hypothesis H_a always pertains to the *full model*.

– Let SSE_R and SSE_F denote the error sums of squares for the reduced and full models, respectively. If the reduced model has df_R error degrees of freedom and the full model has df_F error degrees of freedom, then the test statistic is

$$F^* = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{MSE_F}, \qquad MSE_F = \frac{SSE_F}{df_F}.$$

(SSE_R: 161198, SSE_F: 13322, $SSE_R - SSE_F$: 2875.7, $df_R - df_F$: 36, df_F: 34, SSE_F: 13322, df_F: 34)

– Under H_0, this statistic follows an F distribution with $(df_R - df_F)$ and df_F degrees of freedom. The decision rule is:

$$\text{Reject } H_0 \quad \text{if} \quad F^* > F_{1-\alpha;\, df_R - df_F,\, df_F}.$$

Example 1

Are a person's brain size and body size predictive of his or her intelligence? To investigate this question, researchers (Willerman, et al., 1991) collected the following data (`IQsize.csv`) on a sample of $n = 38$ college students: (1) performance IQ scores (PIQ) from the revised Wechsler Adult Intelligence Scale, which served as the investigator's measure of the individual's intelligence, (2) brain size based on the count obtained from MRI scans (measured in units of count/10,000), (3) height in inches, and (4) weight in pounds.

a. Fit a multiple regression model above using the dataset. Report the estimated regression equation.

$$\widehat{PIQ} = 111 + 2.06 \times Brain - 2.73 \times Ht + .0006\ wt$$

b. Interpret the estimated coefficient for brain size.

Holding height and weight constant, each additional unit of brain size (10,000 MRI counts) is associated with an estimated increase of about 2.06 points in PIQ.

c. Perform an overall F-test to determine whether at least one of the predictors (brain size, height, weight) is significantly associated with PIQ. State the null and alternative hypotheses and report your conclusion.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ v.s. H_a: at least one of $\beta_1, \beta_2, \beta_3$ is not zero.

F-test statistic $= 4.741 \sim F_{(3,34)}$ p-value $= .007215 < \alpha = .05$

3 predictors; $n - p = 38 - 4 = 34$

Yes, at least of one the is significant!

d. Determine whether brain size is a significant predictor of PIQ, controlling for height and weight. Clearly state the hypotheses, p-value, and conclusion at the $\alpha = 0.05$ level.

Yes!

$H_0: \beta_1 = 0$ v.s. $H_a: \beta_1 \neq 0$.

t-stat $= 3.657$ p-value $= .000856 < \alpha = .05$

e. Determine whether height is a significant predictor of PIQ, controlling for brain size and weight. Clearly state the hypotheses, p-value, and conclusion at the $\alpha = 0.05$ level.

yes

$H_0: \beta_2 = 0$ v.s. $H_a: \beta_2 \neq 0$

t-stat $= -2.22$ p-value $= .033 < \alpha = .05$

f. Determine whether height and weight jointly contribute to explaining PIQ after accounting for brain size. Clearly state the hypotheses, p-value, and conclusion at the $\alpha = 0.05$ level.

$H_0: \beta_2 = \beta_3 = 0$ v.s. H_a: at least one β_2, β_3 isn't zero.

F-stat $= 3.6696$, p-value $= .03606$

g. Report the coefficient of determination and interpret it in the context of this study.

This means about 29.49 % of the variation in PIQ scores is explained by brain size, height, and weight.

– Final notes:

1. Brain size is a strong and reliable predictor of performance IQ in this sample. Height adds a weaker, negative effect once brain size is accounted for. Weight provides no predictive value.
2. The model explains about 30% of the variation in PIQ scores, leaving considerable unexplained variation, so intelligence is clearly influenced by many factors beyond body and brain size.

Example 2

A young chef is opening a new Italian restaurant in Manhattan, to be featured on an international reality TV show. The restaurant's goals are to provide the highest quality Italian food, incorporate state-of-the-art décor, and set a new standard for service. The location will be no further south than the Flatiron District and will be situated either east or west of Fifth Avenue. Your role is to determine a competitive dinner menu price by analyzing survey data from 168 Italian restaurants in the area. The dataset (`NYC.csv`) includes the following variables:

- Price: the cost of dinner in US dollars, including one drink and tip
- Food: customer rating of the food out of 30
- Décor: customer rating of the décor out of 30
- Service: customer rating of the service out of 30
- East: dummy variable equal to 1 if east of Fifth Avenue and 0 if west

a Fit a multiple regression model using all of the predictors. Report the estimated regression equation.

b Among Food, Décor, and Service, which has the largest estimated *effect* on Price? Is the largest effect also the most statistically significant?

c If the aim is to choose the location to maximize the achievable dinner price, should the restaurant be on the east or west side of Fifth Avenue? Justify using the sign, magnitude, and significance of the coefficient.

d Does it seem possible to achieve a price premium for "setting a new standard for high-quality service in Manhattan" for Italian restaurants?

e Test whether Food, Décor, and Service jointly contribute to Price after controlling for East. Clearly state the hypotheses, p-value, and conclusion at the $\alpha = 0.05$ level.

f Test whether East and Service jointly add explanatory power beyond Food and Décor. Clearly state the hypotheses, p-value, and conclusion at the $\alpha = 0.05$ level.