Multiple Linear Regression

The formulas are primarily derived from *Linear Regression Analysis, Fifth Edition* by Montgomery et al. and/or the materials from the STAT 6021 course taught by Dr. Woo at the University of Virginia. Except where specially cited, it is believed that the materials are well-known equations and concepts in the public domain. If you believe otherwise, please reach out to me through my Github account so that I can correct the material.

The model

Observations and features

Suppose we wish to model a variable of interest y as a function of k-many features $(x_1, x_2, ..., x_k)$.

When we want to speak of a single observation and its outcome, we will use

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

where $\mathbf{x}$ represents a single observation of the k features and y is the measured outcome. We call the elements of $\mathbf{x}_i$ the regressor variables. We call the y the response.

In general, we assume that there are n-many observations and n associated responses, resulting in the following matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Note that we add a constant column to the observation matrix. This simplifies our later formulas.

The linear model

In multiple linear regression, we model the response in terms of the regressor variables:

$$y = \mathbf{x}'\boldsymbol{\beta} + e$$

where

$$\boldsymbol{\beta}^T = \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_k \end{bmatrix}$$

are the regression coefficients that define the model and e is an error term. The error term is not always explicitly written. Note that error can result from various causes, such as model inadequacy, inaccurate measurements, and random variation in the y values.

In practice, we shall have to use estimators for the $\boldsymbol{\beta}$ model parameter, calculated from a set of observations. In this case, we will resort to the usual $\hat{\boldsymbol{\beta}}$ notation for the vector of estimators or $\hat{\beta}_i$ for individual estimators. These estimators are called the least-squares estimators.

Model assumptions

As in simple linear regression, we make the following assumptions:

- We assume the e_i to be normally distributed
- We assume that $\text{Var}(e) = \sigma^2$ for some fixed value of σ
- We assume $E(e) = 0$.

- We assume that observations on the regressors $\mathbf{x}_i = \begin{bmatrix} x_1 & x_2 & \dots & x_k \end{bmatrix}$ are independent and not dependent on the regression coefficients $\boldsymbol{\beta}$ or σ.
- We assume, of course, that the data can be modeled by a linear relationship.

Basic formulas

The model parameters

The $\hat{\beta}_i$ can be calculated as follows:

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

where $\mathbf{A}'$ is our notation for the matrix transpose of $\mathbf{A}$.

Fitted values and residuals

We also define a useful "hat matrix" $\mathbf{H}$:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}'$$

The fitted values $\hat{\mathbf{y}}$ can be calculated from the observed responses $\mathbf{y}$ using

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

We define the residuals as the difference between the observed values and our fitted values:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

Note that the $e_i, ..., e_n$ give the residual for each individual observation. You can also use the hat matrix:

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

Utility Matrices

We define the adjusted model matrix (the book calls it the `"centered" model matrix`):

$$\mathbf{X}_c = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1k} - \bar{x}_k \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{nk} - \bar{x}_k \end{bmatrix}$$

As a notational convenience (borrowed from the book) let $\mathbf{C}$ be defined as follows:

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$$

This is the scaled covariance matrix. In fact, per page 80 of the book,

$$\sigma^2\mathbf{C} = \text{Var}(\hat{\boldsymbol{\beta}})$$

Sums of squares, variances, and expected values

Sums of squares

The sums of squares have matric forms.

The residual sum of squares:

$$SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

The regression sum of squares:

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

The total sum of squares:

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Note that the SS_T can also be calculated with the formula used in simple linear regression (see p. 26 of the textbook).

Mean squared error

The residual mean squared error is

$$MS_{Res} = \frac{SS_{Res}}{n-p}$$

where $p = k + 1$. Recall that k is the number of features. This is an unbiased estimator, according to the book.

Also:

$$MS_R = SS_R/k$$

Error variance and expected value

We have the following:

$$E(e) = 0$$

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

where $p = k + 1$ is one more than the number of features. Note that this is not the same estimate as the maximum-likelihood estimate. See p. 83 of the textbook for details.

The regression coefficients

The least squares estimators (for the regression coefficients) are unbiased estimators:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

We also have (p. 80):

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$$

where C is the utility matrix previously discussed. Hence, $\mathrm{se}(\hat{\beta}_j) = \sigma\sqrt{C_{jj}}$.

R squared

The R^2 is

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

The adjusted R^2 is

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

where $MS_T = SS_T/(n-1)$.

Note: Recall that $p = k + 1$.

Test statistics and distributions

The overall F test

Given the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ and alternative hypothesis H_a : one of the β_j is nonzero, the F test statistic is

$$F_0 = \frac{MS_R}{MS_{Res}}$$

When the null hypothesis is true, F_0 follows an $F_{k,n-p}$ distribution. When it is not true, another distribution holds. In both cases, if $F_0 > F_{\alpha,k,n-k-1}$, we can reject the null hypothesis (for our level of significant α).

For details, including an ANOVA table, see p. 85 from our textbook.

The partial F test

Read the book starting at p. 89. Seriously, there's a lot of detail that we skip over here. This is the basic method for checking significance when attempting to reduce a model (that is, reduce the number of regressors).

Let

$$\boldsymbol{\beta}_1, \boldsymbol{\beta}_1$$

be subsets of the regression coefficients such that

$$\boldsymbol{\beta} = \left[\frac{\boldsymbol{\beta}_1}{\boldsymbol{\beta}_2}\right]$$

Consider the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ and the alternative hypothesis $H_a : \boldsymbol{\beta}_2 \neq \mathbf{0}$, where $textbf0$ is a vector of all zeroes of an appropriate size.

Then

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \\ SS_R(\boldsymbol{\beta}_1) &= \hat{\boldsymbol{\beta}}_1'\mathbf{X}_1'\mathbf{y} \\ SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) &= SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1) \end{aligned}$$

where $\mathbf{X}_1$ is the subset of the $\mathbf{X}$ matrix with only the features included in $\boldsymbol{\beta}_1$.

Then the F statistic (of doom!) is

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/r}{MS_{Res}}$$

where r is the number of regression coefficients in $\boldsymbol{\beta}_2$.

We won't go into which distributions hold (see p. 90), but we can test the F statistic as if the statistic follows an $F_{r,n-p}$ distribution. If your level of significance is α, check for $F_0 > F_{\alpha,r,n-p}$ and if so, reject the null hypothesis.

Please read the book for some serious caveats to this method. The method can be expanded, and we leave the discussion of this to the book.

T tests for individual coefficient

Given the null hypothesis $H_0 : \beta_j = 0$ and the alternative hypothesis $H_a : \beta_j \neq 0$, we can use the following t statistic:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$$

where C_{jj} is the element in the scaled correlation matrix. If the null hypothesis is true, then the statistic follows the t_{n-k-1} distribution.

Additional tests

There are several more tests. We leave the discussion of these to the book.

Confidence intervals

Basic formula

For β_j:

$$\hat{\beta}_j - t_{A/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{A/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

or, if you want to estimate the standard error differently, the more general formula:

$$\hat{\beta}_j - t_{A/2,n-p}\text{se}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{A/2,n-p}\text{se}(\hat{\beta}_j)$$

For the mean response ($E(y|\mathbf{x}_0)$):

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \leq E(y|\mathbf{x}_0) \leq \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

Joint confidence regions for regression coefficients

Confidence region

See p. 101 of the book. We can use an approximate F statistic to build a region of confidence.

Scaled intervals

We can set up a series of confidence intervals over all $\hat{\beta}_j$ simultaneously, so that we're $(1-a)$ confident that the full set of intervals contains the values of their respective parameters One way of doing this is to pick a coefficient Δ to the standard error that sizes them with this goal:

$$\hat{\beta}_j \pm \Delta \text{ se}(\hat{\beta}_j), \quad j = 0...k$$

The book discusses three ways to do this. See p. 102.

Method	Δ
Bonferroni confidence intervals	Let $\Delta = t_{\alpha/(2p),n-p}$
the Scheffé method	see the book
the maximum modulus t procedure	see the book

About the alpha values: Note that we use the book's notation, rather than Dr. Woo's.

Standardized regression coefficients

See p. 113-115 of the textbook. The main purpose of scaling is to produce "dimensionless regression coefficients" (as the textbook states it). The units of the β_j are normally based on the units of the data, but by scaling (and centering) the data, we can normalize the values of the β_j. Note that we do not need the coefficient term anymore. See the book for details, if I haven't mentioned that you should look at the book yet.

Unit normal scaling

The following is effectively quoted from p.113-114 of the textbook. In this approach, we perform the following:

Let

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad j = 1, 2, \ldots, k$$

Let

$$y_i^* = \frac{y_i - \bar{y}}{s_y} \quad i = 1, 2, ..., n$$

where

$$s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x})^2)}{n-1}$$

and

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2)}{n-1}$$

are the sample variances of the x_j, y.

We write the model as $y_i^* = b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \epsilon_i$ for $i = 1, 2, \ldots, n$. The book writes the formula for the regression coefficients as

$$\hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}^*$$

Unit length scaling

This approach is effectively the same, except we use the sum of squares of differences from the means. The following is quoted from p. 114 of the textbook:

Let

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{jj}^2}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, k$$

and

$$y_i^0 = \frac{y_i - \bar{y}}{SS_T^{1/2}}, \qquad i = 1, 2, \ldots, n$$

where each

$$S_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

are what the book calls the "corrected sum of squares for regressor x_j".

We then fit the model

$$y_i^0 = b_q w_{i1} + b_2 w_{i2} + \cdots + b_k w_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

The coefficients can be calculated as

$$\hat{\mathbf{b}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}^0$$

where $\mathbf{W}'\mathbf{W}$ is the correlation matrix. Effectively, $(\mathbf{W}'\mathbf{W})_{ij} = r_{ij}$ is the correlation computed using the Pearson correlation method (the book doesn't identify the method).

The book doesn't say this in this section, but recall that $SS_T = \sum(y - \bar{y})$ (p. 26), so SS_T is the y analogue to the s_{jj}.

Relationship

Both forms of scaling produce the same $\hat{\mathbf{b}}$ coefficient estimates.

In fact, in terms of the matrices, there is only a scaling factor difference:

$$\mathbf{Z}\mathbf{Z}' = (n-1)\mathbf{W}'\mathbf{W}$$

The coefficients in the original model (before scaling) and the coefficients in the scaled model are also related by a scaling factor:

$$\hat{\beta}_j = \hat{b}_j \left(\frac{SS_T}{S_{jj}}\right)^{1/2} \quad j = 1, 2, \ldots, k$$

So in theory you could calcaulate the original model and then calculate the coefficients for the scaled model.