Chapter 3: Multiple Regression

August 14, 2018

1 The multiple linear regression model

The model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon \qquad (1)$$

is called a **multiple linear regression model** with k regressors. The parameters β_j, $j = 0, 1, \cdots, k$, are called the regression coefficients. This model describes a hyperplane in the k-dimensional space of the regressor variables x_j. The parameter β_j

represents the expected change in the response y per unit change in x_j when all of the remaining regressor variables $x_i\ (i \neq j)$ are held constant. For this reason the parameters $\beta_j,\ j = 1, 2, \cdots, k$, are often called partial regression coefficients. To estimate β's in (1), we will use a sample of n observations on y and the associated x's. The model for the ith observation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i, \quad i = 1, 2, \cdots, n.$$

The assumptions for e_i or y_i are analogous to as those for simple linear regression, namely:

1. $E(e_i) = 0$ for $i = 1, 2, \cdots, n$, or, equivalently $E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$.

2. var$(e_i) = \sigma^2$ for $i = 1, 2, \cdots, n$, or, equivalently, var$(y_i)) = \sigma^2$.

3. $cov(e_i, e_j) = 0$ for all $i \neq j$, or, equivalently, $cov(y_i, y_j) = 0$.

2 Terms and Predictions

Regression problems start with a collection of potential predictors, which are either continuous or discrete. From the pool of potential predictors, we create a set of terms that are the X-variable that appear in (1). The terms might include:

- *The intercept* The mean function can be rewritten as

$$E(Y|X) = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where X_0 is a term that is always equal to one.

- *Transformations of predictors* Sometimes the original predictors need to be transformed in some way to make (1) hold to a reasonable approximation.

- *Polynomials* Problems with curbed mean functions can sometimes be accommodated in the multiple linear regression model by including polynomial terms in the predictor variables.

- *Interactions and other combinations of predictors* Products of predictors called interactions are often included in a mean function along with the original predictors to allow for joint effect of two or more variables.

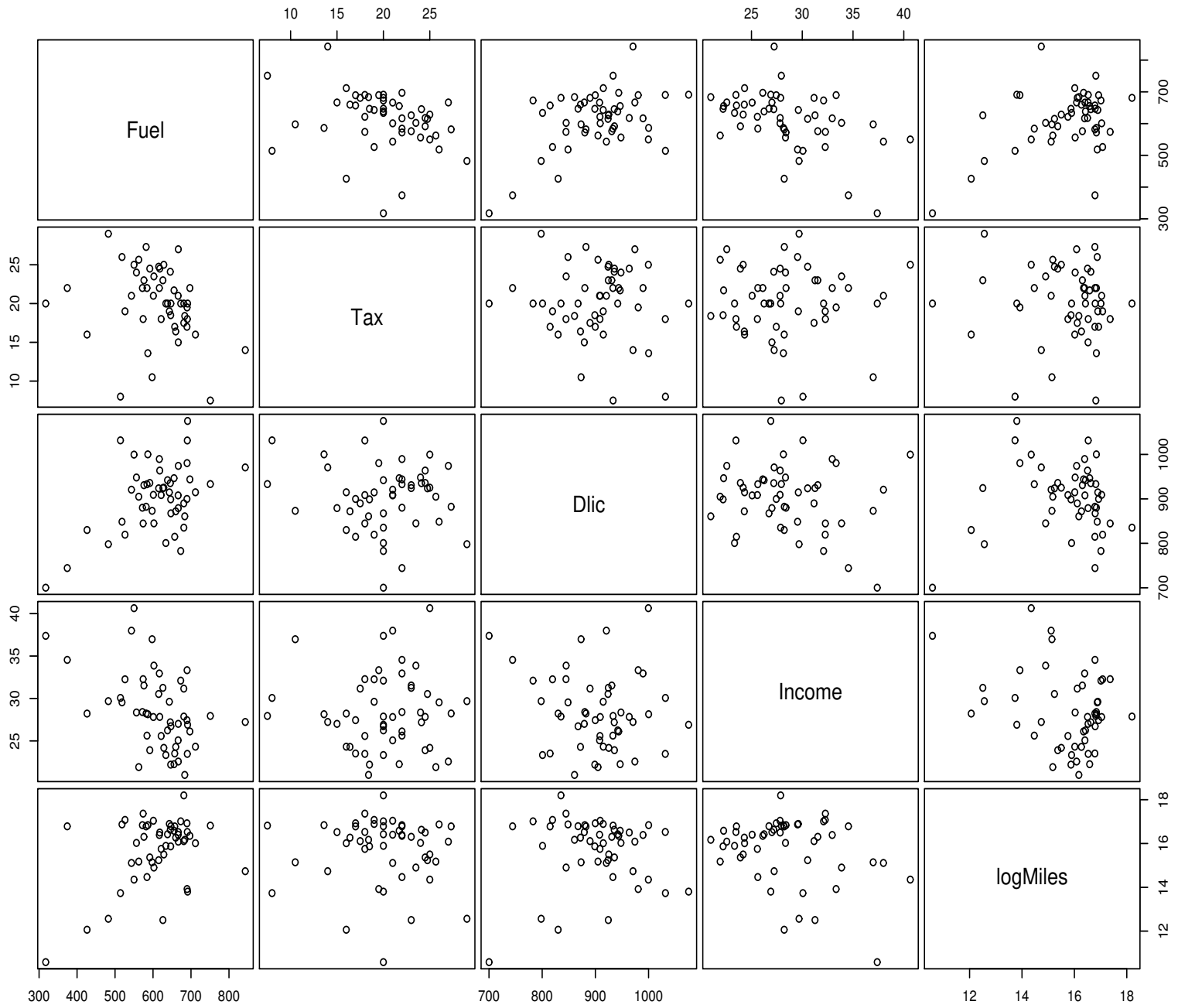
- *Dummy variables and factors* A categorical predictor with two or more levels is called a factor. Factors are included in multiple linear regression using dummy variables, which are typically terms that have only two values, often zero and one, indicating which category is present for a particular observation.

A regression with k predictors may combine to give fewer than k terms or expand to require more

than k terms.

Figure 1 shows the scatterplot matrix for the fuel consumption data. In this plot, the relationships between all pairs of terms appear to be very weak, suggesting that for this problem the marginal plots including Fuel are quite information about the multiple linear regression problem.

A more traditional and less informative, summary of the two-variable relationships is the matrix


Fuel
Tax
Dlic
Income
logMiles
10 15 20 25
25 30 35 40
300 500 700
10 15 20 25
700 800 900 1000
25 30 35 40
12 14 16 18
300 400 500 600 700 800
700 800 900 1000
12 14 16 18

of sample correlations, shown in Table 3.2. In this instance, the correlation matrix helps to reinforce the relationships we see in the scatterplot matrix, with fairly small correlations between the predictors and Fuel, and essentially no correlation between the predictors themselves.

Table 1: Sample correlation for the fuel data.

	Tax	Dlic	Income	logMiles	Fuel
Tax	1.0000	-0.0858	-0.0107	-0.0437	-0.2594
Dlic	-0.0858	1.0000	-0.1760	0.0306	0.4685
Income	-0.0107	-0.1760	1.0000	-0.2959	-0.4644
logMiles	-0.0437	0.0306	-0.2959	1.0000	0.4220
Fuel	-0.2594	0.4685	-0.4644	0.4220	1.0000

3 Ordinary Least Squares

3.1 Parameter estimation

In matrix notation, the model given by eq. (1) is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

We wish to find the vector of least squares estimators, $\hat{\boldsymbol{\beta}}$, that minimizes

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$

The least squares estimators must satisfy

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

which simplifies to

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}.$$

The above equations are the least squares normal equations. The least squares estimator of $\boldsymbol{\beta}$ is

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

provided that the inverse matrix $(\boldsymbol{X}'\boldsymbol{X})^1$ exists. The matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ will always exist if the regressors are linearly independent, that is, if no column of the $\boldsymbol{X}$ matrix is a linear combination of the other columns.

The fitted regression model corresponding to

the levels of the regressor variables $\boldsymbol{x}' = [1, x_1, x_2, \cdots, x_k$

is

$$\hat{y} = \boldsymbol{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j.$$

The vector of fitted values $\hat{y}_i$ is

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

where $n \times n$ matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is usually called the **hat matrix.** The n residuals may

be conveniently written as

$$\widehat{e} = y - \hat{y} = y - X\hat{\beta} = y - Hy = (\mathrm{I} - H)y.$$

3.2 Properties of the least-squares estimators

Theorem 3.1 *If* $E(y) = X\beta$, *then* $\hat{\beta}$ *is an unbiased estimator for* $\hat{\beta}$.

Theorem 3.2 *If cov(*y*)=*$\sigma^2\mathrm{I}$, *the covariance matrix for* $\hat{\beta}$ *is given by* $\sigma^2(X'X)^{-1}$.

Theorem 3.3 *(Gauss-Markov theorem) If* $E(y) =$

$\boldsymbol{X}\boldsymbol{\beta}$ *and* $\text{cov}(\boldsymbol{y}) = \sigma^2\mathbf{I}$*, the least squares estimators* $\hat{\beta}_j,\ j = 0, 1, \cdots, k$*, have minimum variance among all linear unbiased estimators.*

3.3 Estimation of σ^2

The residual sum of squares

$$\begin{aligned} SS_{Res} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \boldsymbol{e}'\boldsymbol{e} \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'(I - H)\boldsymbol{y}. \end{aligned}$$

The residual mean square is

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

and it is an unbiased estimate of σ^2.

3.4 Fuel Consumption Data

Fit the fuel data by a multiple linear regression model with mean function $E(Fuel|X) = \beta_0 + \beta_1 Tax + \beta_2 Dlic + \beta_3 Income + \beta_4 \log(Miles)$.

The 5×5 matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is given by

	Intercept	Tax	Dlic	Income	logMiles
Intercept	9.0215	-2.85e-02	-4.08e-03	-5.98e-02	-1.93e-01
Tax	-0.0285	9.79e-04	5.60e-06	4.26e-05	1.60e-04
Dlic	-0.0041	5.60e-06	3.92e-06	1.19e-05	5.40e-06
Income	-0.0598	4.26e-05	1.19e-05	1.14e-03	1.00e-03
logMiles	-0.1932	1.60e-04	5.40e-06	1.00e-03	9.95e-03

The coefficients can then be calculated as

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$= (154.193, -4.228, 0.472, -6.135, 18.545)'.$$

The output gives the estimates $\hat{\boldsymbol{\beta}}$ and their stan-

dard errors computed based on $\hat{\sigma}^2$ and the diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.1928   194.9062   0.791 0.432938
Tax          -4.2280     2.0301  -2.083 0.042873
Dlic          0.4719     0.1285   3.672 0.000626
Income       -6.1353     2.1936  -2.797 0.007508
logMiles     18.5453     6.4722   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: 9.33e-07
```

4 The analysis of variance

The test for significance of regression is a test to determine if there is a linear relationship between the response y and any of the regressor variables $x_1, x_2, \cdots, x_k$. This procedure is often thought of as an overall or global test of model adequacy. The appropriate hypothesis are

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \quad \text{for at least one } j$$

Rejection of this null hypothesis implies that at least one of the regressors $x_1, \cdots, x_k$ contributes significantly to the model.

The test is based on the identity:

$$SS_T = SS_R + SS_{Res},$$

where $SS_R = \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y} - n\bar{\boldsymbol{y}}^2$, $SS_{Res} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y}$, and $SS_T = \boldsymbol{y}'\boldsymbol{y} - n\bar{\boldsymbol{y}}^2$; and the following ANOVA table

Therefore, to test the null hypothesis, compute

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regress	SS_R	k	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

Table 2: Analysis of Variance (ANOVA) for testing significance of regression

the test statistic F_0 and reject H_0 if $F_0 > F_{\alpha,k,n-1}$.

4.1 R^2 and Adjusted R^2

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

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

The adjusted R^2 penalizes us for adding terms that are not helpful, so it is very useful in evaluating and comparing candidate regression models.

The overall ANOVA table for the fuel data is given by

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regress	201994	4	50499	11.992
Residual	193700	46	4211	
Total	395694	50		

To get a significance level for the test, we would compare $F_0 = 11.992$ with the $F(4, 46)$ distribution. Since the probability $Pr(> F_0) = 9.33e - 07$, a very small number, leading to a very strong

evidence against the null hypothesis that the mean function does not depend on any of the terms. The value of $R^2 = 201994/395694 = 0.5105$ indicates that about half the variation in Fuel is explained by the terms.

4.2 Tests on individual regression coefficients

The hypotheses for testing the significance of any individual regression coefficient, such as β_j, are

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_j \neq 0$$

If H_0 is not rejected, then this indicates that the regressor x_j can be deleted from the model. The

test statistic for this hypothesis is

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

where C_{jj} is the diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ corresponding to $\hat{\beta}_j$. The null hypothesis is rejected if $|t_0| > t_{\alpha/2,n-k-1}$. Note that this is really a partial or marginal test because the regression coefficient $\hat{\beta}_j$ depends on all of the other regressor variables $x_i (i \neq j)$ that are in the model.

Thus, this is a test of the contribution of x_j given the other regressors in the model.

Consider the regression model with k regressors

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e},$$

where $p = k + 1$, $\boldsymbol{\beta}_1$ is a $(p - r)$-vector of coefficients, and $\boldsymbol{\beta}_2$ is a r-vector of coefficients. We wish to test the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \quad H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

To find the contribution of the terms in $\boldsymbol{\beta}_2$ to the regression, fit the model assuming that the null hypothesis H_0 is true. The reduced model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{e}.$$

The LS estimator of $\boldsymbol{\beta}_1$ in the reduced model is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}'\boldsymbol{y}$. The regression sum of squares is

$$SS_R(\boldsymbol{\beta}_1) = \hat{\boldsymbol{\beta}}_1\boldsymbol{X}_1'\boldsymbol{y} - (\sum_{i=1}^{n} y_i)^2/n$$

The regression sum of squares due to $\boldsymbol{\beta}_2$ given that $\boldsymbol{\beta}_1$ is

$$SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1)$$

with $p-(p-r) = r$ degrees of freedom. This sum of squares is called the **extra sum of squares due to** $\boldsymbol{\beta}_2$ because it measures the increase in the regression sum of squares that results from adding the regressors $\boldsymbol{X}_2$ to a model that already contains $\boldsymbol{X}_1$. Now $SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ is independent of

MS_{Res}, and the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ may be tested by the statistic

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

If $F_0 > F_{\alpha,r,n-p}$, we reject H_0, concluding that at least one of the parameters in $\boldsymbol{\beta}_2$ is not zero, and consequently at least one of the regressors $x_{k-r+1}, \cdots, x_k$ in $\boldsymbol{X}_2$ contribute significantly to the regression model. This test is also a partial F test because it measures the contribution of the

regressors in $\boldsymbol{X}_2$ given that the other regressors in $\boldsymbol{X}_1$ are in the model.

```
Analysis of Variance Table

Model 1: Fuel ~ Dlic + Income + logMiles
Model 2: Fuel ~ Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     47 211964
2     46 193700  1     18264 4.3373 0.04287
```

Note that the t-statistic for Tax is $t = -2.083$, and $t^2 = (-2.083)^2 = 4.34$, the same as the F-statistic we just computed.

5 Confidence interval: estimation of the mean response

We may construct a confidence interval on the mean response at a particular point $\boldsymbol{x}_0 = (1, x_{01}, \cdots, x_{0k})'$. The fitted value at this point is

$$\hat{y}_0 = \boldsymbol{x}_0' \hat{\boldsymbol{\beta}}$$

This is an unbiased estimate of $E(y|\boldsymbol{x}_0)$, and the variance of $\hat{y}_0$ is

$$Var(\hat{y}_0) = \sigma^2 \boldsymbol{x}_0' (\boldsymbol{X}'\boldsymbol{x})^{-1} \boldsymbol{x}_0$$

Therefore, a $100(1-\alpha)$ percent confidence interval on the mean response at the point $\boldsymbol{x}_0$ is

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

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

6 Prediction of new observations

A point estimate of the future observation y_0 at the point $\boldsymbol{x}_0$ is

$$\hat{y}_0 = \boldsymbol{x}_0' \hat{\boldsymbol{\beta}}.$$

A $100(1-\alpha)$ percent prediction interval for this future observation is

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

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