THE UNIVERSITY OF CHICAGO
Booth School of Business
Business 41912, Spring Quarter 2020, Mr. Ruey S. Tsay

Lecture 4: Multivariate Linear Regression

Linear regression analysis is one of the most widely used statistical methods. We shall start with the multiple linear regression (MLR) analysis before studying the multivariate model. Our discussion of the multiple linear regression uses matrix algebra.

1 The classical linear regression model

The multiple linear regression model consists of a single response variable with multiple predictors and assumes the form

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_r Z_{ir} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where the error terms $\{\epsilon_i\}$ satisfy

1. $E(\epsilon_i) = 0$ for all i;
2. $\text{Var}(\epsilon_i) = \sigma^2$, a constant; and
3. $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$.

In matrix notation, Eq. (1) becomes

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Z_{11} & Z_{12} & \cdots & Z_{1r} \\ 1 & Z_{21} & Z_{22} & \cdots & Z_{2r} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & Z_{n1} & Z_{n2} & \cdots & Z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{Z}_{n\times(r+1)} \boldsymbol{\beta}_{(r+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1},$$

and the assumptions become

1. $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$; and
2. $\text{Cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 \boldsymbol{I}_{n\times n}$.

The matrix $\boldsymbol{Z}$ is referred to as the *design matrix*. To be more specific, the assumptions are conditioned on the given $\boldsymbol{Z}$. That is,

1. $E(\boldsymbol{\epsilon}|\boldsymbol{Z}) = \boldsymbol{0}$; and
2. $\text{Cov}(\boldsymbol{\epsilon}|\boldsymbol{Z}) = \sigma^2 \boldsymbol{I}$.

One can also add that $\boldsymbol{\epsilon}$ and $\boldsymbol{Z}$ are independent. If $\boldsymbol{\epsilon}$ and $\boldsymbol{Z}$ are dependent, then we encounter the endogeneity problem. The least squares estimates may not be consistent.

2 Least squares estimation

The least squares estimate (LSE) of $\boldsymbol{\beta}$ is obtained by minimizing the sum of squares of errors,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 Z_{i1} - \cdots - \beta_r Z_{ir})^2 = (\boldsymbol{Y} - \boldsymbol{Z\beta})'(\boldsymbol{Y} - \boldsymbol{Z\beta}).$$

Before deriving the LSE, we first define $\tilde{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$ and $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$, assuming that the inverse exists. Then, it is easy to see that $\boldsymbol{H}$ is symmetric and $\boldsymbol{H}^2 = \boldsymbol{H}$. The latter property says that $\boldsymbol{H}$ is an idempotent matrix.
Define $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}$. Then, using $\boldsymbol{Z}'\boldsymbol{H} = \boldsymbol{Z}'$, we have

$$\boldsymbol{Z}'\tilde{\boldsymbol{\epsilon}} = \boldsymbol{Z}'[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y} = [\boldsymbol{Z}' - \boldsymbol{Z}'\boldsymbol{H}]\boldsymbol{Y} = [\boldsymbol{Z}' - \boldsymbol{Z}']\boldsymbol{Y} = \boldsymbol{0}.$$

In other order, $\tilde{\epsilon}$ is orthogonal to the column space of the design matrix $\boldsymbol{Z}$.

Result 7.1 Assume that $\boldsymbol{Z}$ is of full rank $(r+1)$, where $r + 1 \leq n$. The LSE of $\boldsymbol{\beta}$ is

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

Proof. Since

$$\boldsymbol{Y} - \boldsymbol{Z\beta} = \boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{\beta}} - \boldsymbol{Z\beta} = \boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

we have

$$\begin{array}{rcl} S(\boldsymbol{\beta}) & = & (\boldsymbol{Y} - \boldsymbol{Z\beta})'(\boldsymbol{Y} - \boldsymbol{Z\beta}) \\ & = & (\boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{Z}'\boldsymbol{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ & + & 2(\boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}})'\boldsymbol{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ & = & (\boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{Z}'\boldsymbol{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \end{array}$$

because $(\boldsymbol{Y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}})'\boldsymbol{Z} = \tilde{\boldsymbol{\epsilon}}'\boldsymbol{Z} = \boldsymbol{0}$. The first term of $S(\boldsymbol{\beta})$ does not depend on $\boldsymbol{\beta}$ and the second term is the squared length of $\boldsymbol{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Because $\boldsymbol{Z}$ has full rank, $\boldsymbol{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \neq \boldsymbol{0}$ if $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$, so the minimum of $S(\boldsymbol{\beta})$ is unique and ocurs at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$. Consequently, the LSE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.   Q.E.D.

Note: There are many ways to derive the LSE of the MLR. For instance, one can take the partial derivatives and use the first-order equation to obtain $\hat{\boldsymbol{\beta}}$.

The deviations

$$\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 Z_{i1} - \cdots - \hat{\beta}_r Z_{ir}, \quad i = 1, \ldots, n,$$

are the *residuals* of the MLR model. Let $\widehat{\boldsymbol{Y}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}} = \boldsymbol{HY}$ denote the *fitted values* of $\boldsymbol{Y}$. The $\boldsymbol{H}$ matrix is called the *hat* matrix. The LS residuals then become

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = [\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']\boldsymbol{Y} = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}.$$

Note that $\boldsymbol{I} - \boldsymbol{H}$ is also a symmetric and idempotent matrix.
Our prior discussion shows that the residuals satisfy (a) $\boldsymbol{Z}'\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}$ and (b) $\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{\epsilon}} = 0$. Also, the residual sum of squares (RSS) is

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}.$$

Sum of squares decomposition: Since $\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{\epsilon}} = 0$, we have

$$\boldsymbol{Y}'\boldsymbol{Y} = (\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\epsilon}})'(\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\epsilon}}) = \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}.$$

Since the first column of $\boldsymbol{Z}$ is $\boldsymbol{1}$, the result $\boldsymbol{Z}'\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}$ includes $0 = \boldsymbol{1}'\widehat{\boldsymbol{\epsilon}}$. Consequently, we have $\bar{Y} = \bar{\hat{Y}}$. Subtracting $n\bar{Y}^2 = n(\bar{\hat{Y}})^2$ from the prior decomposition, we have

$$\boldsymbol{Y}'\boldsymbol{Y} - n\bar{Y}^2 = \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} - n(\bar{\hat{Y}})^2 + \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}},$$

or

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2.$$

The quantity

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

is the coefficient of determination. It is the proportion of the total variation in the Y_i's *explained* by the predictors $Z_1, \ldots, Z_p$.

Sampling properties:

1. The LSE $\widehat{\boldsymbol{\beta}}$ satisfies $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

2. The residuals satisfy $E(\widehat{\boldsymbol{\epsilon}}) = \boldsymbol{0}$ and $\text{Cov}(\widehat{\boldsymbol{\epsilon}}) = \sigma^2[\boldsymbol{I} - \boldsymbol{H}]$.

3. $E(\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}) = (n - r - 1)\sigma^2$ so that, defining

$$s^2 = \frac{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}}{n - r - 1} = \frac{\boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}}{n - r - 1},$$

   we have $E(s^2) = \sigma^2$.

4. $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}}$ are uncorrelated.

Proof. First, $E(\widehat{\boldsymbol{\beta}}) = E[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}] = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'E(\boldsymbol{Y}) = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'[\boldsymbol{Z}\boldsymbol{\beta} + E(\boldsymbol{\epsilon})] = \boldsymbol{\beta}$. Also, $\text{Cov}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\text{Cov}(\boldsymbol{Y})\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\sigma^2\boldsymbol{I})\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. Next, $E(\widehat{\boldsymbol{\epsilon}}) = [\boldsymbol{I} - \boldsymbol{H}]E(\boldsymbol{Y}) = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Z}\boldsymbol{\beta} = [\boldsymbol{Z} - \boldsymbol{Z}]\boldsymbol{\beta} = \boldsymbol{0}$. Also, $\text{Cov}(\widehat{\boldsymbol{\epsilon}}) = [\boldsymbol{I} - \boldsymbol{H}]\text{Cov}(\boldsymbol{Y})[\boldsymbol{I} - \boldsymbol{H}] = \sigma^2[\boldsymbol{I} - \boldsymbol{H}]$, because $[\boldsymbol{I} - \boldsymbol{H}]$ is idempotent.

Next, using $tr(\boldsymbol{H}) = tr(\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}') = tr[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}] = tr(\boldsymbol{I}_{r+1}) = r+1$, we have

$$\begin{aligned} E(\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}) &= E[tr(\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}})] = E[tr(\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}')] \\ &= tr[\text{Cov}(\widehat{\boldsymbol{\epsilon}})] = tr[\sigma^2(\boldsymbol{I}-\boldsymbol{H})] \\ &= \sigma^2[tr(\boldsymbol{I}) - tr(\boldsymbol{H})] = \sigma^2(n-r-1). \end{aligned}$$

Finally, $\text{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\epsilon}}) = \text{Cov}[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}, (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}] = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\text{Cov}(\boldsymbol{Y},\boldsymbol{Y})(\boldsymbol{I}-\boldsymbol{H}) = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\sigma^2\boldsymbol{I})(\boldsymbol{I}-\boldsymbol{H}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{I}-\boldsymbol{H}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'-\boldsymbol{Z}') = \boldsymbol{0}$ so that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}}$ are uncorrelated.

Gauss' least squares theorem. Let $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$, and $\text{Rank}(\boldsymbol{Z}) = r+1$. For any non-zero $\boldsymbol{c}$, the estimator $\boldsymbol{c}'\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{c}'\boldsymbol{\beta}$ has the smallest possible variance among all linear estimators of the form $\boldsymbol{a}'\boldsymbol{Y}$ that are unbiased for $\boldsymbol{c}'\boldsymbol{\beta}$.

Proof. For any fixed $\boldsymbol{c}$, let $\boldsymbol{a}'\boldsymbol{Y}$ be an unibased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$. Then, $E(\boldsymbol{a}'\boldsymbol{Y}) = \boldsymbol{c}'\boldsymbol{\beta}$, whatever the value of $\boldsymbol{\beta}$. But $E(\boldsymbol{a}'\boldsymbol{Y}) = E[\boldsymbol{a}'(\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{\epsilon})] = \boldsymbol{a}'\boldsymbol{Z}\boldsymbol{\beta}$. Consequently, $\boldsymbol{c}'\boldsymbol{\beta} = \boldsymbol{a}'\boldsymbol{Z}\boldsymbol{\beta}$ or equivalently, $(\boldsymbol{c}' - \boldsymbol{a}'\boldsymbol{Z})\boldsymbol{\beta} = 0$ for all $\boldsymbol{\beta}$. In particular, choosing $\boldsymbol{\beta} = (\boldsymbol{c}' - \boldsymbol{a}'\boldsymbol{Z})'$, we obtain $\boldsymbol{c}' = \boldsymbol{a}'\boldsymbol{Z}$ for any unbiased estimator.
Next, $\boldsymbol{c}'\widehat{\boldsymbol{\beta}} = \boldsymbol{c}'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \equiv \boldsymbol{a}^*\boldsymbol{Y}$, where $\boldsymbol{a}^* = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{c}$. Since $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, so $\boldsymbol{c}'\hat{\boldsymbol{\beta}} = \boldsymbol{a}^*\boldsymbol{Y}$ is an unbiased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$. The result of the last paragraph says that $\boldsymbol{c}' = (\boldsymbol{a}^*)'\boldsymbol{Z}$. Finally, for any $\boldsymbol{a}$ satisfying the unbiased requirement $\boldsymbol{c}' = \boldsymbol{a}'\boldsymbol{Z}$, we have

$$\begin{aligned} \text{Var}(\boldsymbol{a}'\boldsymbol{Y}) &= \text{Var}(\boldsymbol{a}'\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{a}'\boldsymbol{\epsilon}) = \text{Var}(\boldsymbol{a}'\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{a}'\boldsymbol{a} \\ &= \sigma^2(\boldsymbol{a} - \boldsymbol{a}^* + \boldsymbol{a}^*)'(\boldsymbol{a} - \boldsymbol{a}^* + \boldsymbol{a}^*) \\ &= \sigma^2[(\boldsymbol{a}-\boldsymbol{a}^*)'(\boldsymbol{a}-\boldsymbol{a}^*) + (\boldsymbol{a}^*)'\boldsymbol{a}^*], \end{aligned}$$

where $(\boldsymbol{a}-\boldsymbol{a}^*)'\boldsymbol{a}^* = (\boldsymbol{a}-\boldsymbol{a}^*)'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{c} = (\boldsymbol{a}'\boldsymbol{Z} - (\boldsymbol{a}^*)'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{c} = (\boldsymbol{c}' - \boldsymbol{c}')(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{c}$ $= 0$. Because $\boldsymbol{a}^*$ is fixed, and $(\boldsymbol{a}-\boldsymbol{a}^*)'(\boldsymbol{a}-\boldsymbol{a}^*)$ is positive unless $\boldsymbol{a} = \boldsymbol{a}^*$, $\text{Var}(\boldsymbol{a}'\boldsymbol{Y})$ is minimized by the choice of $\boldsymbol{a} = \boldsymbol{a}^*$ and we have, $(\boldsymbol{a}^*)'\boldsymbol{Y} = E\boldsymbol{c}'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \boldsymbol{c}'\hat{\boldsymbol{\beta}}$. Q.E.D.

The result says that the LSE $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ is the *best linear unbiased estimator* (BLUE) of $\boldsymbol{c}'\boldsymbol{\beta}$.

3 Inference

For the multiple linear regression model in (1), we further assume that $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.
Result 7.4 $\widehat{\boldsymbol{\beta}}$ is also the maximum likelihood estimate of $\boldsymbol{\beta}$. In addition, $\widehat{\boldsymbol{\beta}} \sim N_{r+1}[\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}]$ and is independent of the residuals $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$. Let $\tilde{\sigma}^2$ be the maximum likelihood estimate of σ^2. Then,

$$n\tilde{\sigma}^2 = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} \sim \sigma^2\chi^2_{n-r-1}.$$

Proof. Follows what we discussed before for the MLE of multivariate model random sample. Q.E.D.

Note that the MLE $\tilde{\sigma}^2$ of σ^2 is $\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}/n$, which is different from the LSE of $\hat{\sigma}^2$.

Result 7.5. For the Gaussian MLR model, a $100(1-\alpha)$ percent confidence region for $\boldsymbol{\beta}$ is given by

$$(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \le (r+1)s^2F_{r+1,n-r-1}(\alpha),$$

where $s^2 = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}/(n-r-1)$ is the LSE of σ^2. Also, simultaneous $100(1-\alpha)$ percent confidence intervals for the β_j are

$$\hat{\beta}_i \pm \sqrt{\mathrm{Var}(\hat{\beta}_i)}\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}, \quad i = 0, 1, \ldots, r,$$

where $\mathrm{Var}(\hat{\beta}_i)$ is the diagonal element of $s^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ corresponding to $\hat{\beta}_i$.

Proof. Consider the vector $\boldsymbol{V} = (\boldsymbol{Z}'\boldsymbol{Z})^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$, which is normally distributed with mean zero and covariance matrix $\sigma^2\boldsymbol{I}$. Consequently, $\boldsymbol{V}'\boldsymbol{V} \sim \sigma^2\chi^2_{r+1}$. In addition, $(n-r-1)s^2 = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}$ is distributed as $\sigma^2\chi^2_{n-r-1}$ and is independent of $\boldsymbol{V}$. The result then follows. Q.E.D.

Remark: The R command for MLR is `lm`, which stands for linear model.
Likelihood ratio tests for the regression parameters: Consider

$$H_o : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_r = 0, \quad vs \quad H_a : \beta_i \neq 0 \quad \text{for some} \quad q+1 \le i \le r.$$

Under H_o, the model is

$$\boldsymbol{Y} = \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}. \tag{2}$$

Under H_a, the model is

$$\boldsymbol{Y} = \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{3}$$

where

$$\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2], \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}.$$

Result 7.6. Let $\boldsymbol{Z}$ have full rank $r+1$ and $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The likelihood ratio test for the null hypothesis $H_o : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is

$$\frac{[SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z})]/(r-q)}{s^2} \sim F_{r-q,n-r-1},$$

where $SS_{res}(\boldsymbol{Z}_1)$ and $SS_{res}(\boldsymbol{Z}_2)$ are the sum of squares of the models in (2) and (3), respectively.

Proof: Under the model in (3), the maximized likelihood function is

$$L(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \frac{1}{(2\pi)^{n/2}\hat{\sigma}^n}e^{-n/2},$$

where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$ and $\hat{\sigma}^2 = (\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})/n$. On the other hand, under the submodel in (2), the maximized likelihood function is

$$L(\widehat{\boldsymbol{\beta}}_1, \hat{\sigma}_1^2) = \frac{1}{(2\pi)^{n/2}\hat{\sigma}_1^n} e^{-n/2},$$

where $\widehat{\boldsymbol{\beta}}_1 = (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{Y}$ and $\hat{\sigma}_1^2 = (\boldsymbol{Y} - \boldsymbol{Z}_1\widehat{\boldsymbol{\beta}}_1)'(\boldsymbol{Y} - \boldsymbol{Z}_1\widehat{\boldsymbol{\beta}}_1)/n$. Thus, the likelihood ratio is

$$\frac{L(\widehat{\boldsymbol{\beta}}_1, \hat{\sigma}_1^2)}{L(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2)} = \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}^2}\right)^{-n/2} = \left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}^2}{\hat{\sigma}^2}\right)^{-n/2},$$

which gives rise to the test statistic

$$\frac{n(\hat{\sigma}_1^2 - \hat{\sigma}^2)/(r-q)}{n\hat{\sigma}^2/(n-r-1)} = \frac{(SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z}))/(r-q)}{s^2} \sim F_{r-q,n-r-1}.$$

This completes the proof.
Alternatively, one can construct a matrix $\boldsymbol{C}$ such that the null hypothesis becomes H_o : $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$. In this way, $\boldsymbol{C}\widehat{\boldsymbol{\beta}} \sim N_{r-q}(\boldsymbol{C}\boldsymbol{\beta}, \sigma^2\boldsymbol{C}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{C}')$, which can be used to perform the test.

4 Inferences from the fitted model

Consider a specific point of interest, say $\boldsymbol{z}_o = (1, z_{01}, \ldots, z_{0r})'$, in the design-matrix space. Then, the model says

$$E(Y_o|\boldsymbol{z}_o) = \boldsymbol{z}_o'\boldsymbol{\beta},$$

and the LSE of this expectation is $\boldsymbol{z}_o\widehat{\boldsymbol{\beta}}$. In addition, $\text{Var}(\boldsymbol{z}_o'\widehat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{z}_o'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_o$. Consequently, under the normality assumption, a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{z}_o'\boldsymbol{\beta}$ is

$$\boldsymbol{z}_o'\widehat{\boldsymbol{\beta}} \pm t_{n-r-1}(\alpha/2)\sqrt{\boldsymbol{z}_o'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_o s^2}.$$

Forecasting: The point prediction of Y at $\boldsymbol{z}_o$ is $\boldsymbol{z}_o'\widehat{\boldsymbol{\beta}}$, which is an unbiased estimator. Since $Y_o = \boldsymbol{z}_o'\boldsymbol{\beta} + \epsilon_o$, the variance of the forecast is $\sigma^2(1 + \boldsymbol{z}_o'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_o)$, where we use the property $\widehat{\boldsymbol{\beta}}$ and ϵ_o are uncorrelated. Therefore, a $100(1-\alpha)\%$ prediction interval for Y_o is

$$\boldsymbol{z}_o'\widehat{\boldsymbol{\beta}} \pm t_{n-r-1}(\alpha/2)\sqrt{s^2(1 + \boldsymbol{z}_o'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_o)}.$$

5 Model checking

Studentized residuals: From $\hat{\boldsymbol{\epsilon}} = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}$, we have $\text{Cov}(\hat{\boldsymbol{\epsilon}}) = \sigma^2[\boldsymbol{I} - \boldsymbol{H}]$. In particular, $\text{Var}(\hat{\epsilon}_j) = \sigma^2(1 - h_{jj})$, for $j = 1, \ldots, n$. The studentized residuals are

$$\hat{\epsilon}_j^* = \frac{\hat{\epsilon}_j}{\sqrt{s^2(1 - h_{jj})}}, \quad j = 1, \ldots, n.$$

If the fitted regression model is adequate,we expected the studentized residuals to look like independent draws from an $N(0,1)$.

5.1 Residual plots

The residuals $\hat{\epsilon}_j$ or studentized residuals $\hat{\epsilon}_j^*$ are used to obtain various residual plots for model checking:

1. Plot $\hat{\epsilon}_j$ against the fitted model $\hat{y}_j = \boldsymbol{Z}_{j.}\widehat{\boldsymbol{\beta}}$, where $\boldsymbol{Z}_{j.}$ is the jth row of the design matrix $\boldsymbol{Z}$. This plot can be used to check (a) validity of linear model assumption and (b) constant variance of ϵ_j.

2. Plot $\hat{\epsilon}_j$ against individual explanatory variable, e.g. Z_1. This is less common when the number of regressors is large.

3. QQ-plot of $\hat{\epsilon}_j$ to check the normality assumption and possible outliers.

4. Time plot of $\hat{\epsilon}_j$ to check for serial correlations. This is often accompanied by the Durbin-Watson statistic

$$DW = \frac{\sum_{j=2}^n (\hat{\epsilon}_j - \hat{\epsilon}_{j-1})^2}{\sum_{j=1}^n \hat{\epsilon}_j^2} \approx 2(1-\hat{\rho}_1),$$

   where $\hat{\rho}_1$ is the lag-1 autocorrelation function of the residuals defined as

$$\hat{\rho}_1 = \frac{\sum_{j=2}^n \hat{\epsilon}_j \hat{\epsilon}_{j-1}}{\sum_{j=1}^n \hat{\epsilon}_j^2}.$$

   The range of DW-statistic is [0,4] with 2 as the ideal value. A DW statistic greater than 2 indicates negative correlation between the residuals. In practice, when the data have time or spatial characteristics, one should also check higher lags of autocorrelations of the residuals.

5.2 High leverage points and influential observations

From $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, we have

$$\hat{y}_j = \sum_{i=1}^n h_{ji} y_i = h_{jj} y_j + \sum_{i \neq j} h_{ji} y_i.$$

In addition, it can be shown that $0 < h_{jj} < 1$ for all j. In fact, $\sum_{j=1}^n h_{jj} = tr(\boldsymbol{H}) = r+1$, under the assumption of $\boldsymbol{Z}$ is full rank $r+1$. Thus, if h_{jj} is large relatively to other h_{ji} (in magnitude), then y_j will be a major contributor to the fitted value $\hat{y}_j$. Consequently, h_{jj} is called the *leverage* of the linear regression. A large h_{jj} tends to pull the regression line toward the jth data point.

The leverage h_{jj} has another interpretation. It measures the distance of $\boldsymbol{Z}_{j.}$ to the center of the explanatory variables, where $\boldsymbol{Z}_{j.}$ is the jth data point of the design matrix $\boldsymbol{Z}$. For instance, consider the simple linear regression $y_j = \beta_0 + \beta_1 z_j + \epsilon_j$. It can be shown that

$$h_{jj} = \frac{1}{n} + \frac{(z_j - \bar{z})^2}{\sum_{i=1}^n (z_i - \bar{z})^2}.$$

Consequently, if the jth data point of the explanatory variables is far away from the center, then it has a high leverage and pulls the model fit toward itself. It is, therefore, useful in linear regression analysis to check the high leverage points.

Influential observations of a linear regression model are defined as those points that significantly affect the inferences drawn from the data. Methods for assessing the influence are often derived from the change in the LSE $\widehat{\boldsymbol{\beta}}$ if the observations are removed from the data. The well-known statistics for assessing influential observations is the Cook's distance. The Cook's distance for the ith observation is defined as

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{Z}'\boldsymbol{Z})(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{(r+1)s^2}, \tag{4}$$

where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the LSE of $\boldsymbol{\beta}$ with the ith data point removed, and $s^2 = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}/(n-r-1)$ is the LSE of σ^2. See Cook (1977,*Technometrics*). It is the squared distance between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(i)}$ relative to the fixed geometry of $\boldsymbol{Z}'\boldsymbol{Z}$. A large D_i indicates the ith data point is influential, because removing it from the data leads to a substantial change in the parameter estimates. It can be shown that

$$D_i = \frac{1}{r+1}\frac{\hat{\epsilon}_i h_{ii}}{s^2(1-h_{ii})^2} = \frac{1}{r+1}\frac{h_{ii}}{1-h_{ii}}(\hat{\epsilon}_i^*)^2,$$

where $\hat{\epsilon}_i^*$ is the studentized residual.
This expression leads to several interpretations for the Cook's distance. For instance, $h_{ii}/(1-h_{ii})$ is a monotonic function of the leverage h_{ii} and $\hat{\epsilon}_i^*$ is large for an outlying observation. Thus, D_i is the product of a random deviation and a leverage measure. In addition, $h_{ii}/(1-h_{ii}) = \mathrm{Var}(\hat{Y}_i)/\mathrm{Var}(\hat{\epsilon}_i)$. It can also be shown that $\boldsymbol{z}_i'(\boldsymbol{Z}_{(i)}'\boldsymbol{Z}_{(i)})^{-1}\boldsymbol{z}_i = h_{ii}/(1-h_{ii})$. Finally,

$$\frac{h_{ii}}{1-h_{ii}} = \frac{\sum_{j=1}^n \mathrm{Var}(\boldsymbol{z}_j'\widehat{\boldsymbol{\beta}}_{(i)}) - \sum_{j=1}^n \mathrm{Var}(\boldsymbol{z}_j'\widehat{\boldsymbol{\beta}})}{\sigma^2}.$$

Thus, $h_{ii}/(1-h_{ii})$ is proportional to the total change in the variance of prediction at $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ when $\boldsymbol{z}_i$ is deleted.

6 Variable selection

For the MLR model considered, we have r predictors. In theory, there are $2^r - 1$ possible linear regression models available. Because the r predictors may be highly correlated, this

can lead to the problem of *multi-collinearity*, which, in turn, results in a unstable or over-fitted model. An important question then is to select the *best* MLR model among those candidate models. This is the variable selection problem in MLR analysis.
The problem of variable selection has been widely studied in the literature. It is an on-going project for many statisticians and researchers. The current research is to investigate variable selection when both r and n (the sample size) go to infinity. See LASSO, boosting and some Bayesian methods in the literature. In this lecture, we focus on the traditional approach of variable selection. That is, r is fixed, but n may increase.
The methods we discuss include (a) Stepwise regression, (b) Mallow's C_p criterion, (c) Information criteria such as AIC and BIC, and (d) Stochastic search variable selection by George and McCulloch (1993, JASA, *Variable selection via Gibbs sampling*).

6.1 Stepwise regression

It is an iterative procedure alternating between *forward selection* and *backward elimination*. The procedure requires two prespecified thresholds, which I shall denote by F_{in} and F_{out}. Typically, F_{in} and F_{out} are squares of some critical values of a Student-t distribution with $n-1$ degrees of freedom. In some cases, one may use $F_{in} = F_{out} = 4$.
To describe the procedure, we divide the r regressors into two subsets, say S_{in} and S_{out}, where S_{in} contains all the regressors already selected and S_{out} denotes the remaining regressors. S_{in} and S_{out} may be the empty set. Suppose that S_{in} has p variables and denote the corresponding MLR model as

$$Y_i = \beta_0 + \beta_1 Z^*_{i1} + \cdots + \beta_p Z^*_{pi} + e_i, \quad i = 1, \ldots, n. \tag{5}$$

Forward-selection step: The goal of forward selection to expand the regression by adding a regressor to the model. Let $\hat{e}_i$ be the residuals of the MLR in (5).

1. Find the variable, say Z^*_{p+1}, in S_{out} that has the maximum correlation with $\hat{e}_i$.
2. Perform the MLR

$$Y_i = \beta_0 + \beta_1 Z^*_{i1} + \cdots + \beta_p Z^*_{pi} + \beta_{p+1} Z^*_{(p+1)i} + e_i, \quad i = 1, \ldots, n.$$

3. Let t_{p+1} be the t-ratio of the LSE $\hat{\beta}_{p+1}$ of the prior regression. Compare t^2_{p+1} with F_{in}. If $t^2_{p+1} \geq F_{in}$, then move Z^*_{p+1} from S_{out} to S_{in} and go to the backward-elimination step. If $t^2_{p+1} < F_{in}$, stop the procedure.

Backward-elimination step. The goal of backward elimination is to simplify the existing model by removing one variable out of S_{in}. Again, consider the existing model in (5).

1. Find the regressor in S_{in} that has the smallest t-ratio, in absolute value, of the MLR in (5). Denote the regressor by Z^*_j and the corresponding t-ratio by t_j. Compare t^2_j with F_{out}. If $t^2_j < F_{out}$, move Z^*_j from S_{in} to S_{out} and go to the forward-selection step. On the other hand, if $t^2_j \geq F_{out}$, do not move Z^*_j and go to the forward selection.

In practice, one starts with S_{in} being the empty set and S_{out} containing all regressors. One can also start the backward-elimination step with S_{in} containing all regressors. The stepwise procedure will stop after certain iterations. Obviously, the result would depend on the choices of F_{in} and F_{out}.

6.2 Mallow's C_p

Mallow (1973, *Technometrics*, Vol. 15) proposes the following criterion for model selection:

$$C_p = \frac{SSE_p}{s_F^2} - n + 2(p+1),$$

where p denotes the number of variables in the regression

$$Y_i = \beta_0 + \beta_1 Z_{i1}^* + \cdots + \beta_p Z_{ip}^* + \epsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where $\{Z_1^*, \ldots, Z_p^*\}$ is a subset of the available regressors $\{Z_1, \ldots, Z_r\}$, $SSE_p = \sum_{i=1}^n (Y_i - \hat{Y}_{pi})^2$ with $\hat{Y}_{pi}$ denoting the fitted value of the MLR in (6), and s_F^2 is the residual mean squared errors of the MLR model in (1), i.e., the residual mean squared error when *ALL* predictors are used. Here we use the subscript F to denotes the FULL model. Suppose that the subset model in (6) is the *true* model. Then, s_F^2 converges to σ^2 (variance of ϵ_i) as n increases and

$$E(C_p) = \frac{E(SSE_p)}{\sigma^2} - n + 2(p+1) = \frac{(n-p-1)\sigma^2}{\sigma^2} - n + 2(p+1) = p + 1.$$

Therefore, in practice, one selects a model such that C_p is less than, but close to $(p+1)$.

Remark: In the discussion of C_p, I use a constant term in the MLR. If no constant term is used or the constant term is treated as a predictor, then C_p is defined as $C_p = \frac{SSE_p}{s_F^2} - n + 2p$.

6.3 Information criteria

Assume that ϵ_i is normally distributed. For a given subset of regressors, say $S_{in} = \{Z_1^*, \ldots, Z_p^*\}$, one estimates the model parameters of MLR (6) by the maximum likelihood method. As shown before, the estimate of the coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the same as the LSE. However, the MLE of the variance of ϵ_t is

$$\tilde{\sigma}^2 = \frac{SSE}{n},$$

where SSE denotes the sum of squared errors. The AIC criterion for the fitted MLR model is then

$$\text{AIC}(p) = \ln(\tilde{\sigma}^2) + \frac{2p}{n}. \tag{7}$$

The first term of the AIC is a goodness-of-fit statistic measuring the fidelity of the model to the data and the second term is a penalty function. From the definition, AIC uses a penalty of 2 for any parameter used. Note that the constant term is assumed to be included for all model so that it is not counted. In practice, one entertains a set of possible MLR models for Y_i and selects the model that has the smallest AIC values.
If one changes the penalty term to $p\ln(n)/n$, then the criterion becomes the BIC.

Remark: In small sample, Hurvich and Tsai (1989, *Biometrika*) recommend a correction to AIC to improve its selection (via second-order approximation, instead of the first -order approximation). The resulting criterion is referred to as AIC_c. The criterion becomes

$$AIC_c(p) = AIC(p) + \frac{1}{n} \times \frac{2(p+1)(p+2)}{n-p-2}.$$

Example. Consider silver-zinc battery data set of the textbook. The dependent variable is the logarithm of the cycles to failure of the battery. The five predictors are (1) charge rate (Z_1), (2) discharge rate (Z_2), (3) depth of discharge (Z_3), (4) temperature (Z_4) and end of charge voltage (Z_5). We shall use the `leaps` package and the command `step` in R to carry out the computation.
Since $r = 5$ is small, we have $2^5 - 1 = 31$ possible regressions and may entertain all of them in model selection. When r is large, some procedure will become useful in reducing the number of MLR models entertained. The `leaps` package can be used to consider all possible regressions.

```
> setwd("C:/Users/rst/teaching/ama")
> da=read.table("T7-5.DAT",header=T)
> da
      z1   z2    z3 z4   z5   y
1  0.375 3.13  60.0 40 2.00 101
2  1.000 3.13  76.8 30 1.99 141
3  1.000 3.13  60.0 20 2.00  96
4  1.000 3.13  60.0 20 1.98 125
5  1.625 3.13  43.2 10 2.01  43
6  1.625 3.13  60.0 20 2.00  16
7  1.625 3.13  60.0 20 2.02 188
8  0.375 5.00  76.8 10 2.01  10
9  1.000 5.00  43.2 10 1.99   3
10 1.000 5.00  43.2 30 2.01 386
11 1.000 5.00 100.0 20 2.00  45
12 1.625 5.00  76.8 10 1.99   2
13 0.375 1.25  76.8 10 2.01  76
14 1.000 1.25  43.2 10 1.99  78
15 1.000 1.25  76.8 30 2.00 160
```

```
16 1.000 1.25  60.0  0 2.00   3
17 1.625 1.25  43.2 30 1.99 216
18 1.625 1.25  60.0 20 2.00  73
19 0.375 3.13  76.8 30 1.99 314
20 0.375 3.13  60.0 20 2.00 170
> y=log(da$y)
> x=as.matrix(da[,1:5])
> x
         z1   z2    z3 z4   z5
 [1,] 0.375 3.13  60.0 40 2.00
 [2,] 1.000 3.13  76.8 30 1.99
 [3,] 1.000 3.13  60.0 20 2.00
 [4,] 1.000 3.13  60.0 20 1.98
 [5,] 1.625 3.13  43.2 10 2.01
 [6,] 1.625 3.13  60.0 20 2.00
 [7,] 1.625 3.13  60.0 20 2.02
 [8,] 0.375 5.00  76.8 10 2.01
 [9,] 1.000 5.00  43.2 10 1.99
[10,] 1.000 5.00  43.2 30 2.01
[11,] 1.000 5.00 100.0 20 2.00
[12,] 1.625 5.00  76.8 10 1.99
[13,] 0.375 1.25  76.8 10 2.01
[14,] 1.000 1.25  43.2 10 1.99
[15,] 1.000 1.25  76.8 30 2.00
[16,] 1.000 1.25  60.0  0 2.00
[17,] 1.625 1.25  43.2 30 1.99
[18,] 1.625 1.25  60.0 20 2.00
[19,] 0.375 3.13  76.8 30 1.99
[20,] 0.375 3.13  60.0 20 2.00
> library(leaps)
> help(leaps)
>
> leaps(x,y,nbest=3)
$which
      1     2     3     4     5
1 FALSE FALSE FALSE  TRUE FALSE
1 FALSE  TRUE FALSE FALSE FALSE
1  TRUE FALSE FALSE FALSE FALSE
2 FALSE  TRUE FALSE  TRUE FALSE
2 FALSE FALSE FALSE  TRUE  TRUE
2 FALSE FALSE  TRUE  TRUE FALSE
3 FALSE  TRUE FALSE  TRUE  TRUE
3  TRUE  TRUE FALSE  TRUE FALSE
3 FALSE  TRUE  TRUE  TRUE FALSE
4  TRUE  TRUE FALSE  TRUE  TRUE
```

```
4 FALSE  TRUE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE FALSE
5  TRUE  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"
[6] "5"

$size
 [1] 2 2 2 3 3 3 4 4 4 5 5 5 6

$Cp
 [1]  4.086863 22.044063 24.296971  2.604347  4.670782  5.514947  2.879936
 [8]  4.148842  4.440754  4.429345  4.699115  5.746078  6.000000

> > leaps(x,y,nbest=1)
$which
      1     2     3    4     5
1 FALSE FALSE FALSE TRUE FALSE
2 FALSE  TRUE FALSE TRUE FALSE
3 FALSE  TRUE FALSE TRUE  TRUE
4  TRUE  TRUE FALSE TRUE  TRUE
5  TRUE  TRUE  TRUE TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"
[6] "5"

$size
[1] 2 3 4 5 6

$Cp
[1] 4.086863 2.604347 2.879936 4.429345 6.000000

> x1=data.frame(x)
> nn=lm(y~.,data=x1)
> step(nn)
Start:  AIC=7.58
y ~ z1 + z2 + z3 + z4 + z5

       Df Sum of Sq    RSS     AIC
- z3    1    0.4917 16.523  6.1810
- z1    1    0.8006 16.832  6.5514
<none>              16.032  7.5768
- z5    1    1.9995 18.031  7.9275
- z2    1    3.9387 19.971  9.9705
- z4    1   24.5815 40.613 24.1673

Step:  AIC=6.18
y ˜ z1 + z2 + z4 + z5

       Df Sum of Sq    RSS     AIC
- z1    1    0.5160 17.039  4.7960
<none>              16.523  6.1810
- z5    1    1.9690 18.492  6.4327
- z2    1    4.5731 21.097  9.0675
- z4    1   24.3922 40.916 22.3156

Step:  AIC=4.8
y ˜ z2 + z4 + z5

       Df Sum of Sq    RSS     AIC
<none>              17.039  4.7960
- z5    1    1.9747 19.014  4.9890
- z2    1    4.3410 21.380  7.3349
- z4    1   25.8384 42.878 21.2524

Call:
lm(formula = y ˜ z2 + z4 + z5, data = x1)

Coefficients:
(Intercept)           z2           z4           z5
   -64.4322      -0.3365       0.1175      33.5971

>
> m1=step(nn,trace=F)
> m1

Call:
lm(formula = y ˜ z2 + z4 + z5, data = x1)

Coefficients:
(Intercept)           z2           z4           z5
   -64.4322      -0.3365       0.1175      33.5971

> m2=lm(y˜z2+z4+z5,data=x1)
> summary(m2)

Call:
lm(formula = y ˜ z2 + z4 + z5, data = x1)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.7954 -0.7995  0.2129  0.6183  1.4406 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -64.43215   49.34720  -1.306 0.210121    
z2           -0.33647    0.16666  -2.019 0.060573 .  
z4            0.11754    0.02386   4.926 0.000152 ***
z5           33.59708   24.67298   1.362 0.192161    
---
Residual standard error: 1.032 on 16 degrees of freedom
Multiple R-squared: 0.6421,     Adjusted R-squared: 0.575 
F-statistic: 9.568 on 3 and 16 DF,  p-value: 0.0007419
```

6.4 Stochastic search variable selection

This approach uses a hierarchical model and Gibbs sampling to perform variable selection. Readers are referred to George and McCulloch (1993, JASA) for details. Given the observed response $\boldsymbol{Y}$ and the design matrix $\boldsymbol{Z}$, the approach assumes that

$$\boldsymbol{Y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{Z}\boldsymbol{\beta},\sigma^2\boldsymbol{I}),$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix. The key concept of this hierarchical approach is that each component of $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_r)'$ follows a mixture of two normal distributions with mean zero, but different variances. Specifically,

$$\beta_i \sim (1-\gamma_i)N(0,\tau_i^2) + \gamma_i N(0, c_i\tau_i^2), \quad i = 0, \ldots, r \tag{8}$$

where c_i is a sufficiently large positive real number, and γ_i is a Bernoulli variable such that

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = p_i. \tag{9}$$

The idea is as follows: First, one sets τ_i close to zero so that if $\gamma_i = 0$, then β_i would probably be so small that it could be *safely* estimated by 0. Second, one sets c_i large, (in practice $c_i > 1$ always) so that if $\gamma_i = 1$, then a non-zero estimate of β_i should probably be included in the final model. See George and McCulloch (1993) for discussion of selecting c_i and τ_i. Based on the setup, p_i can be regarded as the prior probability that β_i will require a non-zero estimate, i.e., the predictor Z_i should be selected. From (8),

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{D}_\gamma \boldsymbol{R} \boldsymbol{D}_\gamma),,$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_r)'$, $\boldsymbol{R}$ is a prior correlation matrix, and

$$\boldsymbol{D}_\gamma = \text{diag}\{a_0\tau_0, a_1\tau_1, \ldots, a_r\tau_r\},$$

where $a_i = 1$ if $\gamma_i = 0$ and $a_i = c_i$ if $\gamma_i = 1$.
Finally, the prior distribution of σ^2 given $\boldsymbol{\gamma}$ is inverse Gaussian as $\mathrm{IG}(v_\gamma/2, v_\gamma\lambda_\gamma/2)$. This says that $v_\gamma\lambda_\gamma/\sigma^2$ is distributed as $\chi^2_{v_\gamma}$.
Under the framework described earlier, the variable selection is governed by the posterior distribution of $\boldsymbol{\gamma}$ given the data. We denote it as $f(\boldsymbol{\gamma}|\boldsymbol{Y})$. The Markov chain Monte Carlo (MCMC) methods with Gibbs sampling can be used to obtain this posterior distribution.

6.5 Penalized likelihood approaches

What would happen when r is really large and n is not large? For instance, $r = 3000$ and $n = 100$.

1. Boosting: Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics*, **34**, 559-583.

   Bühlmann and Yu (2003). Boosting with the L2-loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324-339.

   L2 Boosting: For simplicity, assume that the data are mean-corrected so that there is no constant term in the linear regression. Boosting is an iterated procedure as follows. Let m = 0. Define $\widehat{\boldsymbol{Y}}^{(0)} = \boldsymbol{0}$.

   (a) Construct the residuals of the mth iteration as $\widehat{\boldsymbol{R}}^{(m)} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}^{(m)}$.

   (b) Fit r simple linear regression

   $$\hat{R}_i^{(m)} = \beta_j Z_{ij} + \epsilon_{ij}$$

   and compute the resulting sum of squares of residuals.

   (c) Let j_m be the explanatory variable that has the smallest sum of squares of residuals, i.e. the maximum R^2 among the r simple linear regression. Compute the fitted value of this simple linear regression $\widehat{\boldsymbol{Y}}(j_m) = \hat{\beta}_{j_m}\boldsymbol{Z}_{j_m}$.

   (d) Update the fit as $\widehat{\boldsymbol{Y}}^{(m+1)} = \widehat{\boldsymbol{Y}}^{(m)} + v\widehat{\boldsymbol{Y}}(j_m)$, where $v \in (0, 1]$ is a tuning parameter.

   (e) Advance m by 1 and go to Step 1.

   The iteration is stopped by some information criteria such as AIC. A smaller v requires longer iteration, but limited experience shows that it works better. For example, $v = 0.05$ has been used.

2. LASSO: Tibshirani (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

   The LASSO estimate of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}^l$, is obtained by

   $$\widehat{\boldsymbol{\beta}}^l = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{r} Z_{ij}\beta_j \right)^2,$$

$$\text{subject to} \quad \sum_{j=1}^{r} |\beta_j| \leq s,$$

where s is a tuning parameter. In practice, s can be chosen by cross-validation. Typically, a 10-fold cross-validation is used.

3. Variants of LASSO: Group LASSO of Yuan and Lin (2006, JRSSB), Gamma LASSO of Taddy (2013), etc.

4. An important idea to simplify the computation of LASSO-type regressions is the method of *least angle regression* of Efron et al. (2004, Annals of Statistics, Vol. 32, pp. 407-499. You should study this paper if you are interested in high-dimensional statistical analysis. (The paper is easily available from the Internet or Library).

7 Omitted variables

In linear regression analysis, if some relevant predictors are omitted, then the coefficient estimates may contain bias and the residuals may have some serial dependence.
Suppose that the true model is

$$\boldsymbol{Y} = \boldsymbol{Z}_1\boldsymbol{\beta}^{(1)} + \boldsymbol{Z}_2\boldsymbol{\beta}^{(2)} + \boldsymbol{\epsilon},$$

where the dimension of $\boldsymbol{Z}_1$ is $q+1 < r+1$. Suppose that the investigator unknowingly fits the model

$$\boldsymbol{Y} = \boldsymbol{Z}_1\boldsymbol{\beta}^{(1)} + \boldsymbol{\epsilon}^{(1)}.$$

The LSE of $\boldsymbol{\beta}^{(1)}$ is

$$\widehat{\boldsymbol{\beta}}^{(1)} = (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{Y}.$$

In this case,

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}^{(1)}) &= (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'E(\boldsymbol{Y}) = (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'(\boldsymbol{Z}_1\boldsymbol{\beta}^{(1)} + \boldsymbol{Z}_2\boldsymbol{\beta}^{(2)} + E(\boldsymbol{\epsilon})) \\ &= \boldsymbol{\beta}^{(1)} + (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{Z}_2\boldsymbol{\beta}^{(2)}. \end{aligned}$$

Consequently, $\widehat{\boldsymbol{\beta}}^{(1)}$ is a biased estimate of $\boldsymbol{\beta}^{(1)}$ unless $\boldsymbol{Z}_1'\boldsymbol{Z}_2 = \boldsymbol{0}$.

Appendix

In this appendix, we consider some useful matrix properties for linear regression model.
Property 1. Assume that the inverse matrices exist, then

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}' & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{F}\boldsymbol{E}^{-1}\boldsymbol{F}' & -\boldsymbol{F}\boldsymbol{E}^{-1} \\ -\boldsymbol{E}^{1}\boldsymbol{F}' & \boldsymbol{E}^{-1} \end{bmatrix},$$

where $\boldsymbol{E} = \boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B}$ and $\boldsymbol{F} = \boldsymbol{A}^{-1}\boldsymbol{B}$.

Property 2. Assume that the design matrix $\boldsymbol{Z}_{n\times(r+1)}$ is of full rank $(r+1)$. The matrix $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ and $\boldsymbol{I} - \boldsymbol{H}$ are idempotent matrices.

Property 3. Partition the design matrix as $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$, where $\boldsymbol{Z}_1$ is an $n \times q$ matrix of full rank q. Define the hat matrix of the design matrix $\boldsymbol{Z}_1$ as $\boldsymbol{H}_1 = \boldsymbol{Z}_1(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1$. Regress $\boldsymbol{Z}_2$ on $\boldsymbol{Z}_1$. The residuals of such a linear regression are $\boldsymbol{W}_2 = [\boldsymbol{I} - \boldsymbol{H}_1]\boldsymbol{Z}_2$. Now, consider $\boldsymbol{W}_2$ as a design matrix and construct the associated hat matrix $\boldsymbol{T}$ as

$$\boldsymbol{T} = \boldsymbol{W}_2'(\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2.$$

Then, we have

$$\boldsymbol{H} = \boldsymbol{H}_1 + \boldsymbol{T}.$$

Proof: From the partition, we have

$$\boldsymbol{Z}'\boldsymbol{Z} = \left[\begin{array}{cc} \boldsymbol{Z}_1'\boldsymbol{Z}_1 & \boldsymbol{Z}_1'\boldsymbol{Z}_2 \\ \boldsymbol{Z}_2'\boldsymbol{Z}_1 & \boldsymbol{Z}_2'\boldsymbol{Z}_2 \end{array}\right].$$

To obtain $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$, we apply Property 1 with $\boldsymbol{A} = \boldsymbol{Z}_1'\boldsymbol{Z}_1$ and use the property that $\boldsymbol{I} - \boldsymbol{H}_1$ is an idempotent matrix. In particular,

$$\begin{aligned}
\boldsymbol{E} &= \boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B} \\
&= \boldsymbol{Z}_2'\boldsymbol{Z}_2 - (\boldsymbol{Z}_2'\boldsymbol{Z}_1)(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}(\boldsymbol{Z}_1'\boldsymbol{Z}_2) \\
&= \boldsymbol{Z}_2'[\boldsymbol{I} - \boldsymbol{Z}_1(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1']\boldsymbol{Z}_2 \\
&= \boldsymbol{Z}_2'(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Z}_2 \\
&= \boldsymbol{W}_2'\boldsymbol{W}_2, \\
\boldsymbol{F} &= (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{Z}_2,
\end{aligned}$$

$$\boldsymbol{F}\boldsymbol{E}^{-1} = (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{Z}_2[\boldsymbol{Z}_2'(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Z}_2]^{-1},$$

and

$$\boldsymbol{Z}_1(\boldsymbol{A}^{-1} + \boldsymbol{F}\boldsymbol{E}^{-1}\boldsymbol{F}')\boldsymbol{Z}_1' = \boldsymbol{H}_1 + \boldsymbol{H}_1\boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{Z}_2'\boldsymbol{H}_1.$$

Finally,

$$\begin{aligned}
\boldsymbol{H} &= \boldsymbol{Z}_1(\boldsymbol{A}^{-1} + \boldsymbol{F}\boldsymbol{E}^{-1}\boldsymbol{F}')\boldsymbol{Z}_1' - \boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{F}'\boldsymbol{Z}_1' - \boldsymbol{Z}_1\boldsymbol{F}\boldsymbol{E}^{-1}\boldsymbol{Z}_2' + \boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{Z}_2' \\
&= \boldsymbol{H}_1 + \boldsymbol{H}_1\boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{Z}_2'\boldsymbol{H}_1 - \boldsymbol{Z}_2\boldsymbol{E}^{1}\boldsymbol{Z}_2'\boldsymbol{H}_1 - \boldsymbol{H}_1\boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{Z}_2' + \boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{Z}_2' \\
&= \boldsymbol{H}_1 - (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{Z}_2'\boldsymbol{H}_1 + (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{Z}_2' \\
&= \boldsymbol{H}_1 + (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Z}_2\boldsymbol{E}^{-1}\boldsymbol{Z}_2'(\boldsymbol{I} - \boldsymbol{H}_1) \\
&= \boldsymbol{H}_1 + \boldsymbol{W}_2(\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'.
\end{aligned}$$

This property says that we can project on the subspace of $\boldsymbol{Z}_1$, then project on the subspace of $\boldsymbol{Z}_2$ that is orthogonal to $\boldsymbol{Z}_1$.

Property 4. Let $\boldsymbol{Z}_1 = \mathbf{1}$ be the n-dimensional vector of ones. Applying Property 3, we have

$$\boldsymbol{H} = \mathbf{1}\mathbf{1}'/n + \boldsymbol{H}^*,$$

where $\boldsymbol{H}^*$ is the hat matrix of the centered design matrix $\boldsymbol{Z}^*_{n\times r}$ given by

$$\boldsymbol{Z}^* = \begin{bmatrix} Z_{11} - \bar{Z}_1 & Z_{12} - \bar{Z}_2 & \cdots & Z_{1r} - \bar{Z}_r \\ Z_{21} - \bar{Z}_1 & Z_{22} - \bar{Z}_2 & \cdots & Z_{2r} - \bar{Z}_r \\ \vdots & \vdots & & \vdots \\ Z_{n1} - \bar{Z}_1 & Z_{n2} - \bar{Z}_2 & \cdots & Z_{nr} - \bar{Z}_r \end{bmatrix},$$

where $\bar{Z}_i = \sum_{j=1}^n Z_{ji}/n$ is the mean of the ith column of $\boldsymbol{Z}$.

Property 5. Let $\boldsymbol{z}_i'$ be the ith row of the design matrix $\boldsymbol{Z}$ and $\boldsymbol{Z}_{(i)}$ be the matrix $\boldsymbol{Z}$ with the $\boldsymbol{z}_i'$ removed. Then,

$$(\boldsymbol{Z}_{(i)}'\boldsymbol{Z}_{(i)})^{-1} = (\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{z}_i\boldsymbol{z}_i')^{-1} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1} + \frac{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i\boldsymbol{z}_i'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}}{1 - \boldsymbol{z}_i'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i}.$$

Proof: Let $\boldsymbol{A} = \boldsymbol{Z}'\boldsymbol{Z}$, $\boldsymbol{a} = -\boldsymbol{z}_i'$ and $\boldsymbol{b} = \boldsymbol{z}_i$. The result then follows the identity

$$(\boldsymbol{A} + \boldsymbol{a}'\boldsymbol{b})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{a}'(\boldsymbol{I} + \boldsymbol{b}\boldsymbol{A}^{-1}\boldsymbol{a}')^{-1}\boldsymbol{b}\boldsymbol{A}^{-1},$$

where $\boldsymbol{A}$ is a $k \times k$ symmetric and invertible matrix and $\boldsymbol{a}$ and $\boldsymbol{b}$ are $q \times k$ matrices of rank q.

Note that $\boldsymbol{z}_i'\widehat{\boldsymbol{\beta}} = \hat{Y}_i$ is the fitted value of the ith observation so that the ith residual is $\hat{\epsilon}_i = Y_i - \boldsymbol{z}_i'\widehat{\boldsymbol{\beta}}$. Also, from the definition $h_{ii} = \boldsymbol{z}_i'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i$.

Property 6. $\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}} = \frac{-(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i\hat{\epsilon}_i}{1-h_{ii}}$.

Proof. Note that $\boldsymbol{Z}'\boldsymbol{Y} = \boldsymbol{Z}_{(i)}'\boldsymbol{Y}_{(i)} + \boldsymbol{z}_iY_i$, where Y_i is the ith element of $\boldsymbol{Y}$ and $\boldsymbol{Y}_{(i)}$ is the vector $\boldsymbol{Y}$ with Y_i removed. Using Property 5, we have

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{(i)} &= (\boldsymbol{Z}_{(i)}'\boldsymbol{Z}_{(i)})^{-1}\boldsymbol{Z}_{(i)}'\boldsymbol{Y}_{(i)} \\ &= \left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1} + \frac{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i\boldsymbol{z}_i'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}}{1 - h_{ii}}\right](\boldsymbol{Z}'\boldsymbol{Y} - \boldsymbol{z}_iY_i) \\ &= \widehat{\boldsymbol{\beta}} + \frac{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i\boldsymbol{z}_i'\widehat{\boldsymbol{\beta}}}{1 - h_{ii}} - (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_iY_i - \frac{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_ih_{ii}Y_i}{1 - h_{ii}} \\ &= \widehat{\boldsymbol{\beta}} + \frac{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i\hat{Y}_i}{1 - h_{ii}} - \frac{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_iY_i}{1 - h_{ii}} \\ &= \widehat{\boldsymbol{\beta}} + \frac{-(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i\hat{\epsilon}_i}{1 - h_{ii}}. \end{aligned}$$

Consequently,

$$\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}} = \frac{-(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_i\hat{\epsilon}_i}{1-h_{ii}}.$$

Property 7. The Cook's distance can be written as

$$D_i = \frac{1}{r+1}\frac{\hat{\epsilon}_i^2 h_{ii}}{s^2(1-h_{ii})^2}.$$