

Lecture 5: Multivariate Linear Regression (continued)

1 Inference

For the multiple linear regression model in (??), we further assume that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.

Result 7.4 $\hat{\boldsymbol{\beta}}$ is also the maximum likelihood estimate of $\boldsymbol{\beta}$. In addition, $\hat{\boldsymbol{\beta}} \sim N_{r+1}[\boldsymbol{\beta}, \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}]$ and is independent of the residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$. Let $\tilde{\sigma}^2$ be the maximum likelihood estimate of σ^2 . Then,

$$n\tilde{\sigma}^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \sim \sigma^2\chi_{n-r-1}^2.$$

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 $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/n$, 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

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

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

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

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

Proof. Consider the vector $\mathbf{V} = (\mathbf{Z}'\mathbf{Z})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, which is normally distributed with mean zero and covariance matrix $\sigma^2 \mathbf{I}$. Consequently, $\mathbf{V}'\mathbf{V} \sim \sigma^2\chi_{r+1}^2$. In addition, $(n - r - 1)s^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$ is distributed as $\sigma^2\chi_{n-r-1}^2$ and is independent of \mathbf{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

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

Under H_o , the model is

$$\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}. \tag{1}$$

Under H_a , the model is

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

where

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

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

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

where $SS_{res}(\mathbf{Z}_1)$ and $SS_{res}(\mathbf{Z}_2)$ are the sum of squares of the models in (1) and (2), respectively.

Proof: Under the model in (2), the maximized likelihood function is

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

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

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

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

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

which gives rise to the test statistic

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

This completes the proof.

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

2 Inferences from the fitted model

Consider a specific point of interest, say $\mathbf{z}_o = (1, z_{o1}, \dots, z_{or})'$, in the design-matrix space. Then, the model says

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

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

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

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

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

3 Model checking

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

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

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

3.1 Residual plots

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

1. Plot $\widehat{\epsilon}_j$ against the fitted model $\widehat{y}_j = \mathbf{Z}_{j \cdot} \widehat{\boldsymbol{\beta}}$, where $\mathbf{Z}_{j \cdot}$ is the j th row of the design matrix \mathbf{Z} . This plot can be used to check (a) validity of linear model assumption and (b) constant variance of ϵ_j .
2. Plot $\widehat{\epsilon}_j$ against individual explanatory variable, e.g. Z_1 . This is less common when the number of regressors is large.
3. QQ-plot of $\widehat{\epsilon}_j$ to check the normality assumption and possible outliers.
4. Time plot of $\widehat{\epsilon}_j$ to check for serial correlations. This is often accompanied by the Durbin-Watson statistic

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

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.

3.2 High leverage points and influential observations

From $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$, 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, $\sum_{j=1}^n h_{jj} = \text{tr}(\mathbf{H}) = r + 1$, under the assumption of \mathbf{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 j th data point.

The leverage h_{jj} has another interpretation. It measures the distance of \mathbf{Z}_j to the center of the explanatory variables, where \mathbf{Z}_j is the j th data point of the design matrix \mathbf{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 j th 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 $\hat{\boldsymbol{\beta}}$ 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 i th observation is defined as

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'(\mathbf{Z}'\mathbf{Z})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(r + 1)s^2}, \quad (3)$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the LSE of $\boldsymbol{\beta}$ with the i th data point removed, and $s^2 = \hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{\epsilon}} / (n - r - 1)$ is the LSE of σ^2 . See Cook (1977, *Technometrics*). It is the squared distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$

relative to the fixed geometry of $\mathbf{Z}'\mathbf{Z}$. A large D_i indicates the i th 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, $h_{ii}/(1-h_{ii}) = \text{Var}(\hat{Y}_i)/\text{Var}(\hat{\epsilon}_i)$. It can also be shown that $\mathbf{z}'_i(\mathbf{Z}'_{(i)}\mathbf{Z}_{(i)})^{-1}\mathbf{z}_i = h_{ii}/(1-h_{ii})$. Finally,

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

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

4 Variable selection

Problem: Select a “best” submodel of

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \dots + \beta_r Z_{ir} + \epsilon_i.$$

There are 2^r possible submodels, including the one with constant only. Several methods are available.

1. Stepwise (forward selection and backward elimination)
 - Forward selection: Simple to complex
 - Backward elimination: Complex to simple
2. Mallows's C_p

Let $SSR_p(M)$ be the residual sum of squares for the submodel M , where p is the number of explanatory variables in M , and $SSR_r(F)$ be the residual sum of squares of the full model. Then,

$$C_p = \frac{SSR_p(M)}{SSR_r(F)} - (n - 2p).$$

One selects the submodel that is close to the 45° line of the scatter-plot of (p, C_p) .

3. AIC: Akaike (1974). For a submodel M with p explanatory variables,

$$AIC(M) = n \ln \left(\frac{RSS_p(M)}{n} \right) + 2p.$$

One selects the submodel with minimum AIC value.

4. BIC: Scharwtz (1978, Ann. Statistics.). For a submodel M with p explanatory variables,

$$BIC(m) = n \ln \left(\frac{RSS_p(M)}{n} \right) + p \ln(n).$$

5. Stochastic search variable selection: George and McCulloch (1993, JASA).

For each coefficient β_i , introduce an indicator variable γ_i such that

$$\beta_i | \gamma_i \sim (1 - \gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2),$$

where τ_i and c_i are selected in such a way that if $\beta_i \sim N(0, \tau_i^2)$, then it is safe to treat β_i as zero, whereas $c_i > 1$ is chosen such that if $\beta_i \sim N(0, c_i^2 \tau_i^2)$, then β_i has a non-zero estimate.

The augmented variable γ_i is Binomial such that $P(\gamma_i = 0) = 1 - P(\gamma_i = 1) = p_i$. Markov chain Monte Carlo (MCMC) method is used to estimate the model. The posterior distribution of $(\gamma_1, \dots, \gamma_r)$ provides information about the model selection.

What 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 $\mathbf{m} = \mathbf{0}$. Define $\widehat{\mathbf{Y}}^{(0)} = \mathbf{0}$.

- (a) Construct the residuals of the m th iteration as $\widehat{\mathbf{R}}^{(m)} = \mathbf{Y} - \widehat{\mathbf{Y}}^{(m)}$.
 (b) Fit r simple linear regression

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

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 $\widehat{\mathbf{Y}}(j_m) = \widehat{\beta}_{j_m} \mathbf{Z}_{j_m}$.
 (d) Update the fit as $\widehat{\mathbf{Y}}^{(m+1)} = \widehat{\mathbf{Y}}^{(m)} + v \widehat{\mathbf{Y}}(j_m)$, 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 $\hat{\boldsymbol{\beta}}^l$, is obtained by

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

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

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

5 Omitted variables

Suppose that the true model is

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

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

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

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

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

In this case,

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

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

6 Multivariate Multiple Linear Regression