Chapter 10. Validation of Regression Models

SKIP!

Chapter 11. Multicollinearity

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}, \text{ if } (X'X) \text{ is non-singular.}$$

Q: When is $X'X$ singular?

Ans: When at least one column of X is *linearly dependent* on the other columns. We say that **collinearity** (or **multicollinearity**) exists among the columns of X.

Def: The data are said to be **ill-conditioned** if there is a "linear dependency" in the columns of X. □

i.e. $\left.\begin{array}{ll} \det(X'X) & \approx 0 \\ Var(\hat{\beta}_i) & \text{large} \end{array}\right\}$ Inadequate data.

Read Section 11.2 for sources of multicollinearity.

Effects of Multicollinearity

Centering and Scaling Data: Consider

$$\mathbf{Y}\sqrt{S_{yy}} = \beta_1\sqrt{S_{11}}\mathbf{X}_1 + \cdots + \beta_k\sqrt{S_{kk}}\mathbf{X}_k + \boldsymbol{\epsilon}',$$

where $Y_i = (y_i - \bar{y})/\sqrt{S_{yy}}$, $X_{ji} = (x_{ji} - \bar{x}_j)/\sqrt{S_{jj}}$,

$j = 1, \ldots, k; i = 1, \ldots, n$. Thus, we have

$$\mathbf{Y} = \alpha_1\mathbf{X}_1 + \cdots + \alpha_k\mathbf{X}_k + \boldsymbol{\epsilon}'', \text{ where } \alpha_j = \beta_j\left(\frac{S_{jj}}{S_{yy}}\right)^{1/2},$$

$j = 1, \ldots, k$.

$\therefore \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ can be estimated through the normal equations

$$\begin{pmatrix} 1 & \cdots & r_{ij} \\ \vdots & 1 & \vdots \\ r_{ij} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_k \end{pmatrix} = \begin{pmatrix} r_{1y} \\ \vdots \\ r_{ky} \end{pmatrix}, \text{ in correlation form.}$$

Here

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} = corr(x_i, x_j), \text{ and } r_{iy} = \frac{S_{iy}}{\sqrt{S_{ii} S_{yy}}} = corr(x_i, y).$$

Then $\hat{\beta}_i = \hat{\alpha}_i (S_{yy}/S_{ii})^{1/2}, i = 1, \ldots, k$ and

$$\hat{\beta}_0 = \bar{y} - \hat{\alpha}_1 \bar{x}_1 - \cdots - \hat{\alpha}_k \bar{x}_k.$$

Note: People in some fields argue that $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ are more meaningful in interpreting the regression.

Ex. $k = 2$, $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$ (centered and scaled).

$$X'X = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

and

$$C = (X'X)^{-1} = \begin{pmatrix} (1 - r_{12}^2)^{-1} & -r_{12}/(1 - r_{12}^2) \\ & (1 - r_{12}^2)^{-1} \end{pmatrix}.$$

$$\therefore \quad \hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}, \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2}.$$

$$Var(\hat{\beta}_1) = Var(\hat{\beta}_2) = \sigma^2/(1 - r_{12}^2),$$

and $Cov(\hat{\beta}_1, \hat{\beta}_2) = -\sigma^2 r_{12}/(1 - r_{12}^2)$.

If $|r_{12}| \to 1$ (x_1, x_2 possess strong multicollinearity), then the LSE of β have large variances and covariance.

For $k > 2$,

$$C_{jj} = \frac{1}{1 - R_j^2}, j = 1, \ldots, k, \qquad \text{(Exercise)}$$

where R_j^2 is the coefficient of multiple determination from the regression of x_j on the remaining $k - 1$ regressors.

$\therefore R_j^2 \longrightarrow 1$ if x_j has strong linear relationship with some subsets of the other regressors.

Thus, $Var(\hat{\beta}_j) \longrightarrow$ large.

Consider

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \text{expected squared distance} = \sum_{j=1}^{k} E(\hat{\beta}_j - \beta_j)^2$$

$$= \sum_{J=1}^{k} Var(\hat{\beta}_j) = \sigma^2 tr(X'X)^{-1} = \sigma^2 \sum_{j=1}^{k} \frac{1}{\lambda_j},$$

where λ_j's are the eigenvalues of $X'X$.

Strong multicollinearity causes at least one λ_j to be small, thus, $E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ will be large. Moreover,

$$E(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \; tr(X'X)^{-1} \geq \boldsymbol{\beta}'\boldsymbol{\beta}.$$

Hence, the vector $\hat{\boldsymbol{\beta}}$ is generally longer then $\boldsymbol{\beta}$. □

Detecting Multicollinearity

1. Scatter plots of $(\mathbf{x}_i, \mathbf{x}_j)$. Examine the correlation matrix – only pairs of regressors.

1. By experience or prior knowledge.

   i) $\hat{\beta}_i$ has different sign from anticipated.

   ii) Important explanatory variable yields small t-statistic.

   iii) Sensitive of deletion of a row or a column from X.

2. Check *VIF*'s (Variance Inflation Factors).

   If $\mathbf{x}_j$ is orthogonal to all other columns of X, $\forall j$,

$$X'X = \begin{pmatrix} S_{11} & & 0 \\ & S_{jj} & \\ 0 & & S_{kk} \end{pmatrix} \quad \text{or} \quad R = I.$$

   $\therefore$ $Var(\hat{\beta}_j) = \sigma^2 / S_{jj}$. Consider

$$VIF_j = Var(\hat{\beta}_j) / \left(\frac{\sigma^2}{S_{jj}} \right), j = 1, \ldots, k.$$

Note: 1. VIF_j is a measure of how much σ^2/S_{jj} is inflated by columns of X to $\mathbf{x}_j$. It's the combined effect of the dependencies among the relationship of other regressors on the variance of $\hat{\beta}_j$.

2. For centered and scaled data, $S_{jj} = 1$ (in R),

$$VIF_j = C_{jj} = \frac{1}{1 - R_j^2}, j = 1, \ldots, k.$$

$C_{jj} \to 1 (R_j^2 \to 0)$ orthogonal.

3. A large VIF (> 5 or 10) indicates a strong multi-collinearity.

3. Eigensystem analysis of $X'X$.

   (1) Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of $X'X$. Then one or more of the λ_j's will be ≈ 0, if there is one or more near-linear dependencies in the data. Let

$$\kappa = \frac{\lambda_{\max}}{\lambda_{min}} = \text{ condition number of } X'X$$

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j} = \text{ condition indices of } X'X \text{ for } j = 1, \ldots, k.$$

Note: 1. κ measures the spread in the eigenvalue spectrum of $X'X$.

2. $\kappa \leq 100$, okay;

   $100 < \kappa < 1000$, moderate to strong multicollinearity;

   $\kappa \geq 1000$, severe multicollinearity.

3. If $\lambda_j \approx 0$ then the elements of the associated eigenvector $\mathbf{t}_j = (t_{j1}, \ldots, t_{jk})'$ describe the nature of the linear dependence. i.e. $\sum_{i=1}^{k} t_{ji}\mathbf{x}_i \approx 0$, some of $t_{ji} \approx 0$. Find the linear relationship, e.g. $\mathbf{x}_1 \approx t_2^\star \mathbf{x}_2 + t_4^\star \mathbf{x}_4$.

(2) (Belsley's Method.)

1. Decompose $X'X$ such that $X'X = T\Lambda T'$ where $T'T = TT' = I_k$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_k)$.

Note: 1. $T = (\mathbf{t}_1, \ldots, \mathbf{t}_k)$ and $\mathbf{t}_j$ is the eigenvector associated with λ_j.

2. Singular value decomposition of X:

$$X_{n\times k} = U_{n\times k} D T',$$

where $T'T = I_k$ and U is an $n \times k$ matrix whose columns are the *eigenvectors* associated with the k nonzero eigenvalues of XX' such that $U'U = I_k$, $\quad D = diag(\mu_1, \ldots, \mu_k)$, μ_j's are called the *singular values* of X.

Note: 1. $X'X = TDU'UDT' = TD^2T'$, $\therefore D^2 = \Lambda$, $\mu_j = \sqrt{\lambda_j}$.

2. There will be one small singular value for *each* near-linear dependence.

Let

$$\eta_j = \frac{\mu_{max}}{\mu_j}, j = 1, \ldots, k, \text{ condition indices of } X;$$
$$\eta = \frac{\mu_{max}}{\mu_{min}} = \text{ condition number of } X.$$

Note: 1. This approach deals directly with the data matrix X.

2. The algorithms for singular-value decomposition are more stable numerically than those for eigensystem analysis.

Recall $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1} = \sigma^2 T \Lambda^{-1} T'$ or

$$Var(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{k} \frac{t_{ji}^2}{\mu_i^2} = \sigma^2 \sum_{i=1}^{k} \frac{t_{ji}^2}{\lambda_i} = \sigma^2 C_{jj}.$$

$$\therefore \quad VIF_j = \sum_{i=1}^{k} \frac{t_{ji}^2}{\mu_i^2} = \sum_{i=1}^{k} \frac{t_{ji}^2}{\lambda_i}.$$

Thus, one or more small μ_i^2 (or λ_i) can inflate $Var(\hat{\beta}_j)$ dramatically.

4. Variance-decomposition proportions.

$$\pi_{ij} = \frac{t_{ji}^2/\mu_i^2}{VIF_j}, \quad i = 1, \ldots, k, \text{ for each } j = 1, \ldots, k.$$

Note: 1. π_{ij} measures the multicollinearity = proportion of $Var(\hat{\beta}_j)$ contributed by the i^{th} singular value.

2. $\eta_j > 30$ and $\pi_{ij} > 0.5$ are recommended guidelines for detecting multicollinearity. It indicates that the corresponding regressors are of possible multicollinearity.

		x_1	x_2	$\dots$	$\dots$	x_k
λ_1	η_1	π_{11}	π_{12}	$\dots$	$\dots$	π_{1k}
$\vdots$	$\vdots$	π_{21}	π_{22}	$\dots$	$\dots$	
$\geq$	$\leq$	$\vdots$			$\vdots$	
$\vdots$	$\vdots$	$\vdots$			$\vdots$	
$\vdots$	30	$\vdots$				
λ_k	η_k	π_{k1}	π_{k2}	$\dots$	$\dots$	π_{kk}
$\downarrow$	$\downarrow$		$\dots$			
0	∞	$\sum = 1$	$\dots$			

6. $|X'X| \to 0$. $(0 \leq |X'X| \leq 1)$ or $100[|X'X|^{-1/2} - 1]$ large.

Note: $|X'X|^{-1/2}$ is the size of the confidence region enlarged due to multicollinearity. $(|X'X| = 1$ if X is orthogonal.)

When multicollinearity occurs, one can

1) Add prior information.

e.g. In $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. 'Known' (x_1, x_2) are highly correlated, i.e. $\beta_1 = c\beta_2$.

$\Longrightarrow$ Let $z = x_1 + cx_2$ and consider $y = \beta_0 + \beta_1 z + \epsilon$.

2) Combine models. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$,

(x_{i1}, x_{i2}) collinear.

i) Use *previous* information of $(y_i^\star, x_{i2}^\star)$ to fit

$y_i^\star = \beta_0^\star + \beta_2^\star x_{i2}^\star + \epsilon_i^\star \Longrightarrow$ get $\hat{\beta}_0^\star, \hat{\beta}_2^\star$.

ii) Let $\tilde{y}_i = y_i - \hat{\beta}_0^\star - \hat{\beta}_2^\star x_{i2}$ and consider $\tilde{y}_i = \beta_0' + \beta_1 x_{i1} + \epsilon_i'$.

3) Delete collinear variables. Must check for model adequacy, etc.

4) Principle component analysis.

5) Make transformation on x's to eliminate correlations.

6) Take more data.

7) Bayesian approach.

Homework 10: (Page 365) 11.2, 11.3, 11.4, 11.5, 11.12.

Due: Jan. 16, 2009.

Ridge Regression

Intended to overcome "ill-conditioned" solutions.

$\Longrightarrow$ Find out how the ill-conditioning occurs, and add specific additional information to the problem to remove the ill-conditioning.

Motivation: Find a "biased" estimator of $\boldsymbol{\beta}$ that has a much smaller variance than the LSE.

Let $\theta > 0$ (usually $\theta \in (0, 1)$). Define the ridge regression estimate $\hat{\boldsymbol{\beta}}_R$ of $\boldsymbol{\beta}$ as the solution to

$$(X'X + \theta I)\hat{\boldsymbol{\beta}}_R = X'\mathbf{y} \quad \text{or} \quad \hat{\boldsymbol{\beta}}_R = (X'X + \theta I)^{-1}X'\mathbf{y}.$$

Note: 1. Here the data are centered and scaled.

2. $\hat{\boldsymbol{\beta}}_R = (X'X + \theta I)^{-1}(X'X)(X'X)^{-1}X'\mathbf{y} = Z_\theta \hat{\boldsymbol{\beta}}$. ($\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}$ when $\theta = 0$.)

3. θ is called the **biasing parameter** since $E(\hat{\boldsymbol{\beta}}_R) = Z_\theta \boldsymbol{\beta}$ is biased.

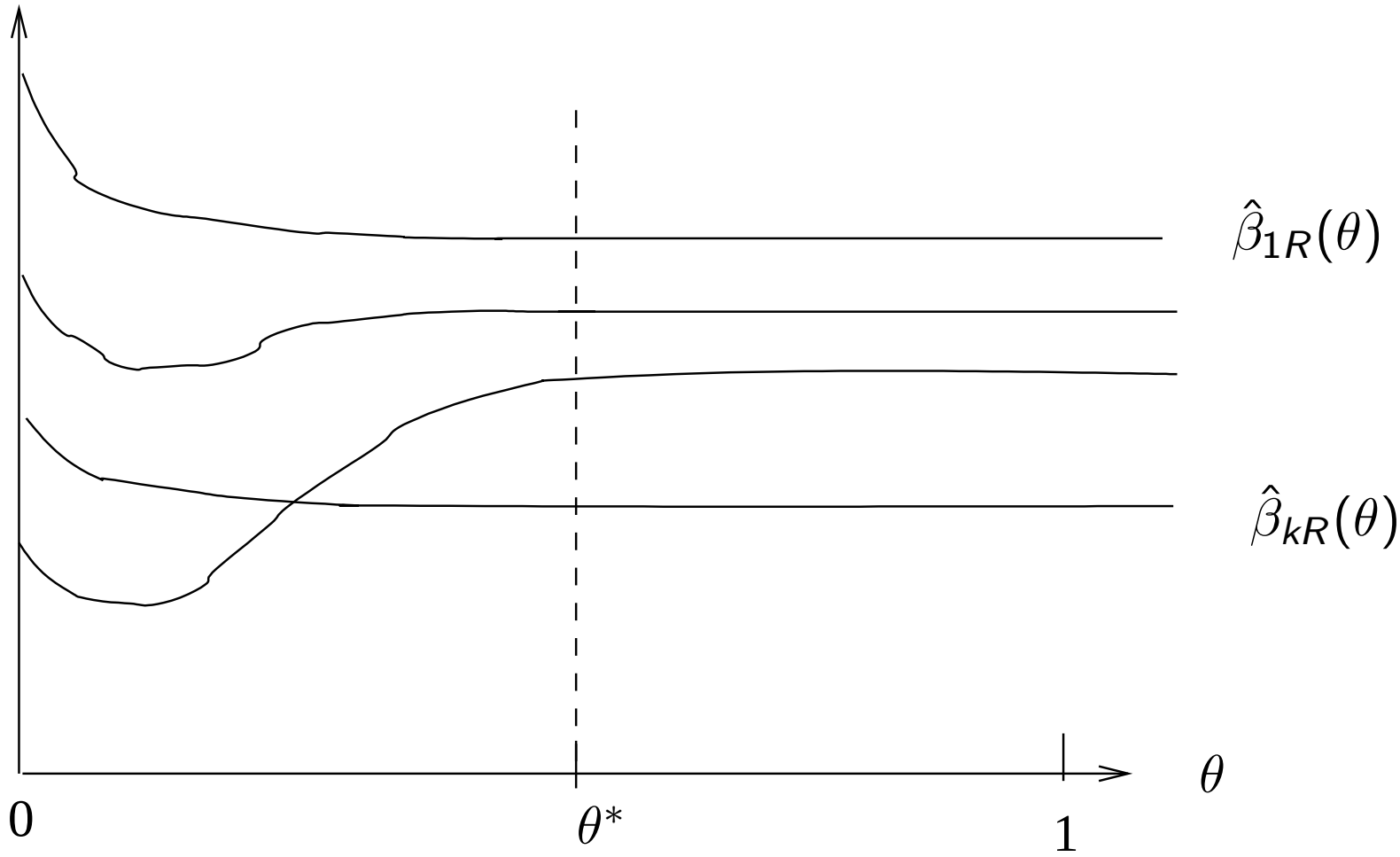
4. $Var(\hat{\boldsymbol{\beta}}_R) = \sigma^2(X'X + \theta I)^{-1}X'X(X'X + \theta I)^{-1}$.

Result: There exists a $\theta^\star > 0$ such that $MSE_{\theta^\star}(\hat{\beta}_R) < MSE(\hat{\beta})$ if $\beta'\beta$ is bounded.

Q: $\theta^\star$ depends on σ^2 and β. $\theta^\star =$???

Solutions:

1. Ridge trace: plot of $\hat{\beta}_{jR}(\theta)$ versus θ, for each $j = 1, \ldots, k$.



At a certain value of θ, the system will stabilize. □

Note that

$$\begin{aligned} SS_{Res}(\hat{\beta}_R) &= (\mathbf{y} - X\hat{\beta}_R)'(\mathbf{y} - X\hat{\beta}_R) \\ &= (\mathbf{y} - X\hat{\beta})'(\mathbf{y} - \hat{\beta}) + (\hat{\beta}_R - \hat{\beta})'X'X(\hat{\beta}_R - \hat{\beta}) \\ & \qquad SS_{Res} \qquad\qquad\qquad \uparrow \text{ in } \theta. \end{aligned}$$

Large $\theta \Longrightarrow$ ridge trace gets stable, but with large SS_{Res}.

So θ can't be too large.

$\therefore$ Select a reasonable *small* θ at which $\hat{\beta}_R$ are stable.

Note: Usually, the ridge estimates do a better job of predicting future observations than LSE.

2. $\theta^\star = p\hat{\sigma^2}/(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}})$, where $\hat{\sigma^2} = MS_{Res}$ based on LSE.

Note that

$$\theta^\star = \underbrace{\frac{\hat{\sigma^2}}{(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}})/p}}_{\text{sample variance}} \longrightarrow \underbrace{\frac{\sigma^2}{\sigma^2_\beta}}_{\text{prior variance}},$$

sensible from a Bayesian point of view (as $\boldsymbol{\beta} \sim N(0, \sigma^2_\beta I)$).

3. Iterative procedure. Let $\theta_0^\star = p\hat{\sigma^2}/(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}) \Longrightarrow \hat{\boldsymbol{\beta}}_R(\theta_0^\star)$.

   Compute $\theta_{j+1}^\star = p\hat{\sigma^2}/(\hat{\boldsymbol{\beta}}_R'(\theta_j^\star)\hat{\boldsymbol{\beta}}_R(\theta_j^\star)), \quad j = 0, 1, 2, \ldots$

   until 'convergence'!

   i.e. Stop until $|\theta_{j+1}^\star - \theta_j^\star|/\theta_j^\star \leq \delta$.

   e.g. $\delta = 20(tr(X'X)^{-1}/k)^{-1.3}$ is suggested.

Justification:

1. Bayesian regression analysis: $\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta})$ prior distribution, $\hat{\boldsymbol{\beta}} = ?$
2. LSE with restriction: Known $\boldsymbol{\beta}'\boldsymbol{\beta} \leq c^2$, $\hat{\boldsymbol{\beta}} = ?$