Chapter 11

Analysis of Collinear Data

Let us recall the multicollinearity problem in Chapter 3. Multicollinearity implies near-linear dependence among the predictors. Multicollinearity can seriously disturb the least-squares fit and in some situations render the regression model almost useless. For example, regression coefficients can have wrong sign or many of the predictors are not statistically significant when the overall F-test is highly significant. Thus, when the model includes more than one predictor it is important to assess whether strong correlations exist among the predictors. Several techniques has been proposed for detecting multicollinearity. Read chapter 11 of textbook for the details. In this course, we will use the variation inflation factor (VIF) among these diagnostic measures. We will see how to overcome multicollinearity through this chapter.

Example

Recall the house price data in Table B.4 of Homework 2. The least-squares fit was $\hat{Y} = 14.92765 + 1.92472X_1 + 7.00053X_2 + .14918X_3 + 2.72281X_4 + 2.00668X_5 - .41012X_6 - 1.40324X_7 - .03715X_8 + 1.55945X_9$ where Y=sale price of the house, X_1=taxes, X_2=number of baths, X_3=lot size, X_4=living space, X_5=number of garage stalls, X_6=number of rooms, X_7=number of bedrooms, X_8=age of the home, X_9=number of fireplaces. The least-squares regression results showed that the overall F-test is highly significant, but all the predictors are not statistically significant. We saw this phenomenon resulted from multicollinearity.

We may ask what causes the multicollinearity problem in these data?

11.1 Sources of Multicollinearity

- The data collection method can lead to multicollinearity.

  Often regression models are fit to data collected in an observational study. Since the modeler has no control over the design points, the predictors can be very dependent upon each other.

If the modeler can control the design, the predictors can be chosen orthogonal to each other or nearly so.

Note: If there is no linear relationship between the predictors, they are said to be orthogonal.

- Constraints on the model or in the population can cause multicollinearity.

  Example. Suppose that an electric utility is investigating the effect of family income (X_1) and house size (X_2) on residential electricity consumption. A physical constraint in the population has caused that families with higher incomes generally have larger homes than families with lower incomes. When physical constraints such as this are present, multicollinearity will exist regardless of the data collection method employed. In this situation, it may be helpful to provide new predictor or artificial predictor.

11.2 Methods for Dealing with Multicollinearity

Several methods have been suggested for dealing with multicollinearity including

- variable selection
- ridge regression
- principal components regression.

When the method of least squares is applied to collinear data, very poor estimates of the regression coefficients can be obtained. The variance of the least-squares estimates of the regression coefficients may be considerably inflate in the presence of near-linear dependencies among the predictors. This implies that the least-squares estimates of regression coefficients are very unstable, that is, their magnitudes and signs may change considerably given a different sample.

The problem with the least-squares estimation method is the requirement that $\hat{\boldsymbol{\beta}}$ be an unbiased estimate of $\boldsymbol{\beta}$. Recall that the least-squares estimate $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate of $\boldsymbol{\beta}$ (Gauss-Markov theorem in Chapter 3). Though ordinary least squares gives unbiased estimates and indeed enjoy the minimum variance of all linear unbiased estimates, there is no upper bound on the variance of the estimates and the presence of multicollinearity may produce large variances. As a result, one can visualize that, under the condition of multicollinearity, a huge price is paid for the unbiasedness property that one achieves by using ordinary least squares. One way to alleviate this problem is to drop the requirement that the estimate of β be unbiased. Biased estimation is used to attain a substantial reduction in variance with an accompanied increase in stability of the regression coefficients. The coefficients become biased and, simply put, if one is successful, the reduction in variance is of greater magnitude than the bias induced in the estimates.

Regression model using standardized variables

Revisit Section 3.3. For convenience, we will consider the following model

$$y_i = \beta_0 + \beta_1 x_{i1}^s + \beta_2 x_{i2}^s + \cdots + \beta_k x_{ik}^s + \epsilon_i, \quad i = 1, \ldots, n,$$

where

$$x_{ij}^s = \frac{x_{ij} - \bar{x}_j}{S_{jj}}, \; j = 1, \ldots, k,$$

with $S_{jj} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = (n-1)s_{x_j}$. This model can be written as

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}_s \boldsymbol{\beta}_* + \boldsymbol{\epsilon},$$

where

$$\mathbf{X}_s = \begin{pmatrix} x_{11}^s & x_{12}^s & \cdots & x_{1k}^s \\ x_{21}^s & x_{22}^s & \cdots & x_{2k}^s \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^s & x_{n2}^s & \cdots & x_{nk}^s \end{pmatrix}, \quad \boldsymbol{\beta}_* = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Under this setting, notice that

$$\mathbf{X}_s^T \mathbf{X}_s = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} \end{pmatrix}$$

which is the correlation matrix of the predictors $X_1, \ldots, X_k$.

11.2.1 Ridge regression

A number of procedures have been developed for obtaining biased estimates of regression coefficients. One of these procedures is ridge regression, originally proposed by Hoerl and Kennard (1970). Note that there are several ways of motivating ridge regression. The ridge estimate is found by solving a slightly modified version of the normal equations. Specifically we define the ridge estimates $\hat{\boldsymbol{\beta}}_{*,R}$ as the solution to

$$(\mathbf{X}_s^T \mathbf{X}_s + \kappa \mathbf{I}) \hat{\boldsymbol{\beta}}_{*,R} = \mathbf{X}_s^T \mathbf{y}$$

or

$$\hat{\boldsymbol{\beta}}_{*,R} = (\mathbf{X}_s^T \mathbf{X}_s + \kappa \mathbf{I})^{-1} \mathbf{X}_s^T \mathbf{y}$$

where $\kappa \geq 0$ is a constant selected by the analyst.

Properties of the ridge estimates $\hat{\boldsymbol\beta}_{*,R}$

- $\mathrm{E}[\hat{\boldsymbol\beta}_{*,R}] = (\mathbf{X}_s^T\mathbf{X}_s+\kappa\mathbf{I})^{-1}\mathbf{X}_s^T(\beta_0\mathbf{1}+\mathbf{X}_s\boldsymbol\beta_*) = (\mathbf{X}_s^T\mathbf{X}_s+\kappa\mathbf{I})^{-1}\mathbf{X}_s^T\mathbf{X}_s\boldsymbol\beta_* = \boldsymbol\beta_* - \kappa(\mathbf{X}_s^T\mathbf{X}_s+\kappa\mathbf{I})^{-1}\boldsymbol\beta_*$. That is, $\hat{\boldsymbol\beta}_{*,R}$ is a biased estimate of $\boldsymbol\beta_*$.

- $\mathrm{Var}(\hat{\boldsymbol\beta}_{*,R}) = \sigma^2(\mathbf{X}_s^T\mathbf{X}_s+\kappa\mathbf{I})^{-1}\mathbf{X}_s^T\mathbf{X}_s(\mathbf{X}_s^T\mathbf{X}_s+\kappa\mathbf{I})^{-1}$

- $\mathrm{MSE}(\hat{\boldsymbol\beta}_{*,R}) = \mathrm{tr}\left\{\mathrm{Var}(\hat{\boldsymbol\beta}_{*,R})\right\} + \left\{\mathrm{Bias}(\hat{\boldsymbol\beta}_{*,R})\right\}^2 = \sigma^2\sum_{j=1}^{k}\frac{\lambda_j}{(\lambda_j+\kappa)^2} + \kappa^2\boldsymbol\beta_*^T(\mathbf{X}_s^T\mathbf{X}_s+\kappa\mathbf{I})^{-2}\boldsymbol\beta_*$ where $\lambda_1,\ldots,\lambda_k$ are the eigenvalues of $\mathbf{X}_s^T\mathbf{X}_s$.

  If $\kappa>0$, note that the bias in $\hat{\boldsymbol\beta}_{*,R}$ increases with κ. However, the variance decreases as κ increases.

Note: The ridge estimate $\hat{\boldsymbol\beta}_{*,R}$ shrinks the ordinary least-squares estimate $\hat{\boldsymbol\beta}_* = (\mathbf{X}_s^T\mathbf{X}_s)^{-1}\mathbf{X}_s^T\mathbf{y}$ toward the origin. Consequently, ridge estimates are sometimes called shrinkage estimates and κ is often referred to as a shrinkage parameter. See figure 11.7 and read pp.351-352 in textbook.

Choosing κ

The ridge trace is a very pragmatic procedure of choosing the shrinkage parameter. The ridge trace is a plot of the elements $\hat{\boldsymbol\beta}_{*,R}$ versus κ for values of κ usually in the interval [0,1]. For values of κ close to zero, multicollinearity will cause rapid changes in the coefficients. These quick changes occur in an interval of κ in which one expects coefficients variance to be inflated. As κ grows, variances reduce, and coefficients are no longer changing rapidly. The analyst simply allows κ to increase until stability is indicated in all coefficients.

Example

See the Acetylene data example in Example 11.2 of textbook.

SAS Program

```
title 'Acetylene Data';
proc means data=acetylene; var Temperature H2ratio Time; run;
proc standard data=acetylene mean=0 std=1 out=s_acetylene;
var Temperature H2ratio Time; run;
data acetylene2;
set s_acetylene;
TempH2 = Temperature*H2ratio;
TempTime = Temperature*Time;
H2Time = H2ratio*Time;
Temp2 = Temperature**2;
H2ratio2 = H2ratio**2;
```

```
Time2 = Time**2;
run;
proc reg data=acetylene2;
model Conversion = Temperature H2ratio Time
                   TempH2 TempTime H2Time Temp2 H2ratio2 Time2 / vif;
run;
/* Ridge regression */
proc reg data=acetylene2 outest=b ridge=.006 to .04 by .002;
model Conversion = Temperature H2ratio Time
                   TempH2 TempTime H2Time Temp2 H2ratio2 Time2 / noprint;
plot / ridgeplot nomodel;  /* Ridge trace to choose the shrinkage parameter */
run;
proc reg data=acetylene2 outest=b2 ridge=.035;
model Conversion = Temperature H2ratio Time TempH2 TempTime H2Time Temp2 H2ratio2 Time2;
run; quit;
proc print data=b2; run;
```

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	2118.83379	235.42598	289.72	<.0001
Error	6	4.87558	0.81260		
Corrected Total	15	2123.70937			

Root MSE	0.90144	R-Square	0.9977
Dependent Mean	36.10625	Adj R-Sq	0.9943
Coeff Var	2.49664		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	35.89579	1.09158	32.88	<.0001	0
Temperature	1	4.00378	4.50870	0.89	0.4087	375.24776
H2ratio	1	2.77831	0.30708	9.05	0.0001	1.74063
Time	1	-8.04233	6.07066	-1.32	0.2335	680.28004
TempH2	1	-6.45678	1.46603	-4.40	0.0045	31.03706
TempTime	1	-26.98038	21.02129	-1.28	0.2467	6563.34519
H2Time	1	-3.76814	1.65535	-2.28	0.0631	35.61129
Temp2	1	-12.52359	12.32380	-1.02	0.3487	1762.57537
H2ratio2	1	-0.97272	0.37460	-2.60	0.0408	3.16432
Time2	1	-11.59322	7.70628	-1.50	0.1832	1156.76628

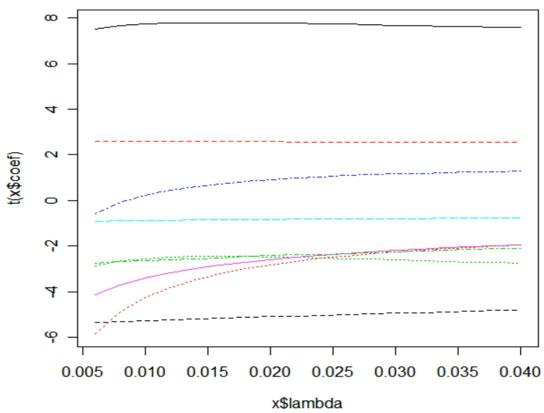

Figure 11.1: Ridge trace for acetylene data.

```
Obs _MODEL_ _TYPE_  _DEPVAR_   _RIDGE_ _PCOMIT_  _RMSE_ Intercept Temperature H2ratio
 1  MODEL1  PARMS  Conversion    .          .    0.90144  35.8958    4.00378   2.77831
 2  MODEL1  RIDGE  Conversion  0.035        .    1.43358  35.0193    6.37296   2.51238

Obs   Time      TempH2   TempTime    H2Time     Temp2   H2ratio2     Time2  Conversion
 1  -8.04233  -6.45678  -26.9804  -3.76814  -12.5236  -0.97272  -11.5932      -1
 2  -4.46413  -3.04433   -0.8775   0.26579    1.8738  -0.51450   -0.2960      -1
```

11.3 Principal-Components Regression

Principal components regression represents another biased estimation technique for combating multicollinearity. With this method, we perform least squares estimation on a set of artificial predictors called the principal components of the correlation matrix $\mathbf{X}_s^T\mathbf{X}_s$. Based on the nature of the analysis, we eliminate a certain number of the principal components to effect a substantial reduction in variance. The method varies somewhat in philosophy from ridge regression but, like ridge, gives biased estimates; when used successfully, this method results in estimation and prediction that is superior to ordinary least squares.

Let $\mathbf{P}$ be the matrix of normalized eigenvectors associated with the eigenvalues $\lambda_1, \ldots, \lambda_k$ of $\mathbf{X}_s^T\mathbf{X}_s$. That is, $\mathbf{P}^T\mathbf{X}_s^T\mathbf{X}_s\mathbf{P} = \text{diag}(\lambda_1, \ldots, \lambda_k) := \mathbf{\Lambda}$. Consider the canonical form of the model

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

where $\mathbf{Z} = \mathbf{X}_s\mathbf{P}$, $\boldsymbol{\alpha} = \mathbf{P}^T\boldsymbol{\beta}_*$. The columns of $\mathbf{Z}$, which define a new set of orthogonal predictors, such as

$$
\begin{aligned}
\mathbf{Z} &= [\mathbf{x}_1^s, \mathbf{x}_2^s, \cdots, \mathbf{x}_k^s] \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix} \\
&= [p_{11}\mathbf{x}_1^s + \cdots + p_{k1}\mathbf{x}_k^s, p_{12}\mathbf{x}_1^s + \cdots + p_{k2}\mathbf{x}_k^s \cdots, p_{1k}\mathbf{x}_1^s + \cdots + p_{kk}\mathbf{x}_k^s] \\
&= [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k]
\end{aligned}
$$

are referred to as principal components. Then, the least-squares estimate of $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} = \boldsymbol{\Lambda}^{-1}\mathbf{Z}^T\mathbf{y}.$$

Notice that

$$\mathrm{E}[\hat{\boldsymbol{\alpha}}] = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathrm{E}[\mathbf{y}] = \boldsymbol{\alpha}$$

and the variance covariance matrix of $\hat{\boldsymbol{\alpha}}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\alpha}}) = \sigma^2\boldsymbol{\Lambda}^{-1}.$$

Thus, a small eigenvalue of $\mathbf{X}_s^T\mathbf{X}_s$ means that the variance of the corresponding orthogonal regression coefficient will be large.

Transformation back to original variables

Objection to principal components regression are quite often the result of the artificiality of the principal components themselves. Without a doubt, if principal components regression is used successfully, the analyst can expect the resulting model in the original variables to improve. Suppose, for example, with k predictors and hence k principal components, $r < k$ components are eliminated. With the retention of all components, we can write $\boldsymbol{\alpha} = \mathbf{P}^T\boldsymbol{\beta}_*$, and hence

$$\boldsymbol{\beta}_* = \mathbf{P}\boldsymbol{\alpha}.$$

Clearly then, if we eliminate the last r components, the least squares estimates of the regression coefficients for all k parameters are given by

$$\hat{\boldsymbol{\beta}}_{*,PC} = \begin{pmatrix} \hat{\beta}_{1,PC} \\ \hat{\beta}_{2,PC} \\ \vdots \\ \hat{\beta}_{k,PC} \end{pmatrix} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{k-r}] \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{k-r} \end{pmatrix} := \mathbf{P}_{k-r}\hat{\boldsymbol{\alpha}}_{k-r}.$$

Properties of the principal components estimate $\hat{\boldsymbol{\beta}}_{*,PC}$

Let $\mathbf{P} = [\mathbf{P}_{k-r}, \mathbf{P}_r]$ and $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_{k-r}^T, \boldsymbol{\alpha}_r^T]^T$. Then,

- $\mathrm{E}[\hat{\boldsymbol{\beta}}_{*,PC}] = \mathbf{P}_{k-r}\boldsymbol{\alpha}_{k-r} = \mathbf{P}_{k-r}\mathbf{P}_{k-r}^T\boldsymbol{\beta}_* = \boldsymbol{\beta}_* - \mathbf{P}_r\mathbf{P}_r^T\boldsymbol{\beta}_* = \boldsymbol{\beta}_* - \mathbf{P}_r\boldsymbol{\alpha}_r$.
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{*,PC}) = \sigma^2\mathbf{P}_{k-r}\boldsymbol{\Lambda}_{k-r}^{-1}\mathbf{P}_{k-r}^T$.

Example

Recall the acetylene data.

SAS Program

```
/* Principal components regression */
proc princomp data=acetylene2 out=pc_acetylene std;
var Temperature H2ratio Time TempH2 TempTime H2Time Temp2 H2ratio2 Time2;
run;
proc reg data=pc_acetylene;
model Conversion = prin1 prin2 prin3 prin4 prin5 prin6 prin7 prin8 prin9 / vif;
run;
proc reg data=pc_acetylene;
model Conversion= prin1 prin2 prin3 prin4;
run; quit;
```

Output

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.20523034	2.04323121	0.4672	0.4672
2	2.16199913	1.02332224	0.2402	0.7075
3	1.13867689	0.09820180	0.1265	0.8340
4	1.04047509	0.65524461	0.1156	0.9496
5	0.38523048	0.33569241	0.0428	0.9924
6	0.04953807	0.03591281	0.0055	0.9979
7	0.01362526	0.00849746	0.0015	0.9994
8	0.00512780	0.00503087	0.0006	1.0000
9	0.00009694		0.0000	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
Temperature	-.338691	0.105762	0.649400	-.012103	0.142553
H2ratio	-.132359	0.339016	-.001450	0.724723	-.583891
Time	0.413622	-.097999	-.469273	0.075249	-.018469
TempH2	0.219512	0.540056	0.087054	-.361627	-.166331
TempTime	-.449170	0.086411	-.287945	-.189378	-.094467
H2Time	-.252758	-.517197	-.054713	0.344885	0.201039

Temp2	0.405575	-.074762	0.441687	0.219505	0.144309
H2ratio2	-.025556	0.531633	-.221360	0.342909	0.734417
Time2	0.466601	-.097142	0.143135	0.132722	-.034777

	Prin6	Prin7	Prin8	Prin9
Temperature	-.249490	0.221473	0.538744	0.174280
H2ratio	0.020702	0.011304	0.028780	-.003324
Time	0.016032	0.168848	0.712893	0.236874
TempH2	0.366240	0.589696	-.110707	0.002548
TempTime	0.029220	-.062038	-.150412	0.797341
H2Time	0.315494	0.620188	-.148413	0.008333
Temp2	0.540689	-.318931	-.047691	0.411690
H2ratio2	-.070755	0.002471	-.075636	0.005021
Time2	-.636550	0.290459	-.368523	0.328882

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	2118.83379	235.42598	289.72	<.0001
Error	6	4.87558	0.81260		
Corrected Total	15	2123.70937			

Root MSE	0.90144	R-Square	0.9977
Dependent Mean	36.10625	Adj R-Sq	0.9943
Coeff Var	2.49664		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	36.10625	0.22536	160.22	<.0001	0
Prin1	1	-8.59379	0.23275	-36.92	<.0001	1.00000
Prin2	1	-0.08690	0.23275	-0.37	0.7217	1.00000
Prin3	1	7.64254	0.23275	32.84	<.0001	1.00000
Prin4	1	2.84485	0.23275	12.22	<.0001	1.00000
Prin5	1	-0.06880	0.23275	-0.30	0.7775	1.00000
Prin6	1	-0.57399	0.23275	-2.47	0.0487	1.00000
Prin7	1	-0.53247	0.23275	-2.29	0.0621	1.00000
Prin8	1	0.44318	0.23275	1.90	0.1056	1.00000
Prin9	1	-0.28123	0.23275	-1.21	0.2724	1.00000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2105.43545	526.35886	316.84	<.0001

Error	11	18.27393	1.66127
Corrected Total	15	2123.70937	

Root MSE	1.28890	R-Square	0.9914
Dependent Mean	36.10625	Adj R-Sq	0.9883
Coeff Var	3.56975		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	36.10625	0.32223	112.05	<.0001
Prin1	1	-8.59379	0.33279	-25.82	<.0001
Prin2	1	-0.08690	0.33279	-0.26	0.7988
Prin3	1	7.64254	0.33279	22.96	<.0001
Prin4	1	2.84485	0.33279	8.55	<.0001