

Collinearity

Applied Regression and Other Multivariable Methods
Sections 12-5

15

Overview

- Consider the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i$$

and suppose $X_{i2} = 5 + 2X_{i1}$

- This is perfect collinearity $\rightarrow r_{X_1, X_2}^2 = 1$
- What happens to estimates?
- Can rewrite model to be

$$Y_i = (\beta_0 + 5\beta_2) + (\beta_1 + 2\beta_2)X_{i1} + E_i$$

and get unique estimates for the slope $(\beta_1 + 2\beta_2)$ and intercept $(\beta_0 + 5\beta_2)$.

- Estimates for $\beta_0, \beta_1, \beta_2$, however are not unique.
 $\beta_0 = 2, \beta_1 = 0, \beta_2 = 4 \rightarrow \text{slope}=8, \text{int}=22$
 $\beta_0 = 7, \beta_1 = 2, \beta_2 = 3 \rightarrow \text{slope}=8, \text{int}=22$
- Without unique estimates, cannot test parameters. Both the estimate and std dev are indeterminate.

15-1

Collinearity

- When fitting the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i$$

It can be shown that the estimate β_1, β_2 , and $\bar{Y} - \beta_0$ are all proportional to

$$\left(\frac{1}{1 - r_{X_1, X_2}^2} \right)$$

and thus so are the variances of each estimate

- We can extend this idea to more than two variables using the squared multiple correlation coefficient

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

where VIF is the variance inflation factor and R_j^2 is based on the regression X_j vs all other predictors

15-2

Collinearity

- If X_j perfectly predicted, the $\text{VIF} = 1/0 = \infty$.
- If X_j completely uncorrelated, the $\text{VIF} = 1/1 = 1$.
- Rule of Thumb : X_j troublesome if $\text{VIF} > 10$
- This is equivalent to $R_j^2 > .90$ and $R_j > .95$
- Some prefer to discuss tolerance

$$\text{tolerance}_j = 1/\text{VIF}_j = 1 - R_j^2$$

$$\text{tolerance}_j \rightarrow 0 \leftrightarrow \text{VIF}_j \rightarrow \infty$$

- Methods to avoid collinearity impasse
 - Remove the troublesome variables from analysis
 - Use computational methods that detect collinearity problems
 - Doing a principal components analysis to eliminate collinearity and possibly reduce the number of predictors
 - Centering/scaling

15-3

Centering/Scaling

- Centering involves transforming the predictors
- Instead of X_j use $X_j - \bar{X}_j$
- The mean on the predictor is now zero
- Standardized predictors are obtained by

$$\frac{X_j - \bar{X}_j}{S_j}$$

where $S_j = \sqrt{\sum (X_{ij} - \bar{X}_j)^2 / (n - 1)}$.

- This predictor has mean 0 and variance 1
- The parameters from this analysis are known as the standardized regression coefficients
- Associated with standard estimates in following way

$$\hat{\beta}_j^* = \hat{\beta}_j \left(\frac{S_j}{S_Y} \right)$$

15-4

Example Problem

Consider the hypothetical sample of 32 white males over the age of 40 from the town of Angina that was described in Chapter 5 (Prob #2). Earlier in Topic 9 we commented on the strong correlations among several independent variables. We demonstrate here how centering can help.

```
options nocenter; goptions reset=global colors=(none);

data problem81;
infile 'i:\www\datasets502\EX0502.DAT' firstobs=2 dlm='09'x;
input person sbp quet age smk;
quet_age = quet*age;
quet_smk = quet*smk;
age_smk = age*smk;
proc reg corr;
model sbp = age quet smk quet_smk quet_age age_smk / collin;
proc means;
var quet age smk;
data centered;
set problem81;
quetc = quet - 3.4410938 ;
agec = age - 53.25 ;
smkc = smk - 0.53125 ;
quet_smk = quetc*smkc;
quet_age = quetc*agec;
age_smk = agec*smkc;
proc reg corr;
model sbp = agec quetc smkc quet_smk quet_age age_smk / collin;
run;
quit;
```

15-5

Variable	Correlation			
	age	quet	smk	quet_smk
age	1.0000	0.8028	-0.1395	-0.0128
quet	0.8028	1.0000	-0.0714	0.1191
smk	-0.1395	-0.0714	1.0000	0.9732
quet_smk	-0.0128	0.1191	0.9732	1.0000
quet_age	0.9429	0.9512	-0.1053	0.0624
age_smk	-0.0245	0.0430	0.9853	0.9907
sbp	0.7752	0.7420	0.2473	0.3717

Variable	Correlation		
	quet_age	age_smk	sbp
age	0.9429	-0.0245	0.7752
quet	0.9512	0.0430	0.7420
smk	-0.1053	0.9853	0.2473
quet_smk	0.0624	0.9907	0.3717
quet_age	1.0000	0.0155	0.8095
age_smk	0.0155	1.0000	0.3501
sbp	0.8095	0.3501	1.0000

It appears that interaction variables created are highly correlated with the main effect variables. There are several pairs of variables that have correlations larger than .94. These only look at pairwise relationships. It could be that a combination of variables even better predicts some of these interactions.

The Proc Reg procedure has collinearity diagnostics. Let's look at these for the uncentered variables.

15-6

Number	Eigenvalue	Collinearity Diagnostics			
		Condition Index	Intercept	age	quet
1	5.99656	1.00000	0.00000737	0.00000697	0.00000635
2	0.95284	2.50865	0.00002745	0.00003064	0.00002611
3	0.03952	12.31816	0.00288	0.00003187	0.00002562
4	0.00713	29.00845	0.00263	0.00707	0.00679
5	0.00318	43.41632	0.00561	0.00287	0.00259
6	0.00069333	92.99993	0.00033051	0.11889	0.09403
7	0.00007430	284.08581	0.98851	0.87110	0.89653

Number	Collinearity Diagnostics			
	smk	quet_smk	quet_age	age_smk
1	0.00014086	0.00005502	0.00000707	0.00005385
2	0.00134	0.00050483	0.00003377	0.00049932
3	0.02229	0.00605	0.00263	0.00053801
4	0.06948	0.11218	0.00264	0.03179
5	0.88657	0.00227	0.00542	0.27560
6	0.01573	0.85674	0.00198	0.67946
7	0.00445	0.02219	0.98729	0.01205

This is doing a principal components analysis. The conditional number greater than 30 suggests high to severe collinearity. Here it is 284.09. In fact, four of the conditional indices are greater than or equal to 29 suggesting severe collinearity.

When we look at the eigenvalues, we see that there are as many as 4 collinearities (four eigenvalues near zero). This model has a lot of problems.

15-7

Based on the means of QUET, AGE, and SMK, we can center these variables.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
quet	32	3.4410938	0.4970781	2.3680000	4.6370000
age	32	53.2500000	6.9560834	41.0000000	65.0000000
smk	32	0.5312500	0.5070073	0	1.0000000

Correlation				
Variable	agec	quetc	smkc	quet_smk
agec	1.0000	0.8028	-0.1395	0.0353
quetc	0.8028	1.0000	-0.0714	0.2962
smkc	-0.1395	-0.0714	1.0000	0.0091
quet_smk	0.0353	0.2962	0.0091	1.0000
quet_age	0.1658	0.1379	0.0347	-0.0038
age_smk	-0.1699	0.0345	0.0175	0.8046
sbp	0.7752	0.7420	0.2473	0.1338

Variable	quet_age	age_smk	sbp
agec	0.1658	-0.1699	0.7752
quetc	0.1379	0.0345	0.7420
smkc	0.0347	0.0175	0.2473
quet_smk	-0.0038	0.8046	0.1338
quet_age	1.0000	0.0169	0.3183
age_smk	0.0169	1.0000	-0.0321
sbp	0.3183	-0.0321	1.0000

Notice now that the highest pairwise correlations are below 0.81.

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	-----Proportion of Variation-----		
			Intercept	agec	quetc
1	1.95288	1.00000	0.00715	0.02111	0.03664
2	1.90450	1.01262	0.04314	0.04231	0.02597
3	1.48494	1.14679	0.13074	0.01757	0.00606
4	0.97304	1.41668	0.00575	0.00282	0.00662
5	0.37014	2.29697	0.72307	0.00149	0.01018
6	0.18282	3.26835	0.07812	0.56501	0.28985
7	0.13169	3.85088	0.01202	0.34969	0.62467

Collinearity Diagnostics

Number	-----Proportion of Variation-----			
	smkc	quet_smk	quet_age	age_smk
1	0.00318	0.04760	0.00120	0.03414
2	0.00623	0.00860	0.05576	0.02414
3	0.01682	0.01643	0.12165	0.02369
4	0.94905	0.00014793	0.00015540	0.00197
5	0.00325	0.01502	0.75772	0.01444
6	0.01867	0.21131	0.05604	0.44668
7	0.00281	0.70089	0.00748	0.45493

The conditional number is now only 3.85. There also appear to be no eigenvalues close to zero. This is an example where centering was very beneficial.

This improvement in collinearity can also be seen in the standard errors of each main effect parameter and the intercept.

Uncentered model

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	182.44353	75.20462	2.43	0.0228
age	1	-1.44503	1.43978	-1.00	0.3252
quet	1	-29.06756	23.37732	-1.24	0.2253
smk	1	0.04327	22.25931	0.00	0.9985
quet_smk	1	-2.00905	10.35857	-0.19	0.8478
quet_age	1	0.72081	0.39542	1.82	0.0803
age_smk	1	0.31123	0.68643	0.45	0.6542

Centered Model

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	142.70634	1.69596	84.14	<.0001
agec	1	1.20069	0.34393	3.49	0.0018
quetc	1	8.24833	5.02711	1.64	0.1134
smkc	1	9.70293	2.62453	3.70	0.0011
quet_smk	1	-2.00905	10.35857	-0.19	0.8478
quet_age	1	0.72081	0.39542	1.82	0.0803
age_smk	1	0.31123	0.68643	0.45	0.6542