

Multicollinearity

Sometimes, the coefficient for an independent variable fails to have a significant t-statistic, even though this result flies in the face of all generally accepted theory. In cases like this it is customary to begin to worry about multicollinearity.

What is Multicollinearity?

The variance of the estimated coefficient for a particular independent variable X_k in a multivariate regression is as follows:

$$V(\hat{\beta}_k) = \frac{\sigma^2}{\sum_i (x_{ki} - \bar{x}_k)^2 (1 - R_k^2)}$$

If the standard error of the estimated coefficient is high, the t-statistic will be low, and the coefficient is likely to be insignificant. From the above formula, three things can cause the standard error to be high:

- 1) High variance of the error term (numerator high).
- 2) Low variation in the independent variable (low value of the summation term in the denominator).
- 3) A high correlation between the independent variable and all the other independent variables (high R_k^2). R_k^2 can be found by regressing the particular independent variable X_k on all the other independent variables in the model. The resulting R^2 tells the percent of the variation in the independent variable X_k which can be jointly explained by the other independent variables.

Multicollinearity refers only to the third item above. It is thus a sufficient condition for a high standard error, but not a necessary condition.

How do we know if we have Multicollinearity?

The most common test for multicollinearity employs the Variance Inflation Factors (VIFs). These are derived directly from the above formula.

$$VIF_k = \frac{1}{(1 - R_k^2)}$$

$$V(\hat{\beta}_k) = \frac{\sigma^2}{\sum_i (x_{ki} - \bar{x}_k)^2} VIF_k$$

One must calculate a VIF for each of the k independent variables. This is easily done in R, by using the package *car*, which contains a function that will calculate the VIF:

```
library(car)
zz<-lm(y~xx)
vif(zz)
```

By convention, a VIF of 10 or greater signals that multicollinearity is a problem. This, however, is a simple rule of thumb, since the VIF (like R^2) is a statistic without a distribution.

If we find that we have a multicollinearity problem, what should we then do?

Principal components analysis is commonly suggested. Here, two or more highly correlated independent variables are replaced with their first principal component. This should only be done when there is an intuitive meaning for the principal component. For example, one can take the first principal component of population and population density, interpreting the new variable as an indicator of size of place. R can carry out principal components with the command *princomp*.

Novices will be tempted to drop highly correlated variables. This is usually a **bad idea**. Omitting relevant variables is a way to introduce bias in the other estimated coefficients. On the other hand, if one has one or more independent variables which in fact are measuring the same underlying construct, then it is a **good idea** to drop all but one. Examples: using both percent white and percent black as measures of racial composition; using both percent high school graduates and percent college graduates as measures of educational attainment.

In most cases, there is little that one can do about multicollinearity. Usually you will talk about multicollinearity in order to explain why your coefficients are not significant. Multicollinearity does not bias your estimated coefficients, it merely explains why the variance of an estimated coefficient may be high.

Homework

The data contain observations for all 138 Tennessee local school districts, and are

described in the following two pages. For your homework, build models explaining two of the 10 test score figures. For each model:

1. Run a test for multicollinearity. Do you think multicollinearity poses a problem?
2. Eliminate irrelevant variables, after conducting an F-test.
3. Present your results in tables, with *ex ante* and *ex post* discussion.

Observations: *S:\TEFF\662\R\school.dbf*

SYSNUM	SN	SYSNUM	SN	SYSNUM	SN
10	AndersonCounK12	275	GibsonSSDK12	621	SweetwaterCiK8
11	ClintonCityK6	280	GilesCountyK12	630	MontgomeryCouK12
12	OakRidgeK12	290	GraingerCounK12	640	MooreCountyK12
20	BedfordCountK12	300	GreeneCountyK12	650	MorganCountyK12
30	BentonCountyK12	301	GreenevilleCK12	660	ObionCountyK12
40	BledsoeCountK12	310	GrundyCountyK12	661	UnionCityK12
50	BlountCountyK12	320	HamblenCountK12	670	OvertonCountK12
51	AlcoaCityK12	330	HamiltonCountK12	680	PerryCountyK12
52	MaryvilleCitK12	331	ChattanoogaCiK12	690	PickettCountK12
60	BradleyCountK12	340	HancockCountK12	700	PolkCountyK12
61	ClevelandCitK12	350	HardemanCounK12	710	PutnamCountyK12
70	CampbellCounK12	360	HardinCountyK12	720	RheaCountyK12
80	CannonCountyK12	370	HawkinsCountK12	721	DaytonCityK8
92	HollowRock-BrucetK12	371	RogersvilleCK8	730	RoaneCountyK12
93	HuntingdonSSK12	380	HaywoodCountK12	731	HarrimanCityK12
94	McKenzieSSDK12	390	HendersonCouK12	740	RobertsonCouK12
95	SouthCarrollK12	391	LexingtonCitK8	750	RutherfordCoK12
97	WestCarrollSK12	400	HenryCountyK12	751	MurfreesboroK8
100	CarterCountyK12	401	ParisSSDK7	760	ScottCountyK12
101	ElizabethtonK12	410	HickmanCountK12	761	OneidaSSDK12
110	CheathamCounK12	420	HoustonCountK12	770	SequatchieCoK12
120	ChesterCountK12	430	HumphreysCouK12	780	SevierCountyK12
130	ClaiborneCouK12	440	JacksonCountK12	790	ShelbyCountyK12
140	ClayCountyK12	450	JeffersonCouK12	791	MemphisCityK12
150	CockeCountyK12	460	JohnsonCountK12	800	SmithCountyK12
151	NewportCityK8	470	KnoxCountyK12	810	StewartCountK12
160	CoffeeCountyK12	480	LakeCountyK12	820	SullivanCounK12
161	ManchesterCiK9	490	LauderdaleCoK12	821	BristolCityK12
162	TallahomaCitK12	500	LawrenceCounK12	822	KingsportCitK12
170	CrockettCounK12	510	LewisCountyK12	830	SumnerCountyK12
171	AlamoCityK6	520	LincolnCountK12	840	TiptonCountyK12
172	BellsCityK5	521	FayettevilleK9	841	CovingtonCitK8
180	CumberlandCoK12	530	LoudonCountyK12	850	TrousdaleCouK12
190	DavidsonCountK12	531	LenoirCityK12	860	UnicoiCountyK12
200	DecaturCountK12	540	McMinnCountyK12	870	UnionCountyK12
210	DeKalbCountyK12	541	AthensCityK9	880	VanBurenCounK12
220	DicksonCountK12	542	EtowahCityK8	890	WarrenCountyK12
230	DyerCountyK12	550	McNairyCountK12	900	WashingtonCoK12
231	DyersburgCitK12	560	MaconCountyK12	901	JohnsonCityK12
240	FayetteCountK12	570	MadisonCountK12	910	WayneCountyK12
250	FentressCounK12	580	MarionCountyK12	920	WeakleyCountK12
260	FranklinCounK12	581	RichardCityK11	930	WhiteCountyK12
271	HumboldtCityK12	590	MarshallCounK12	940	WilliamsonCoK12
272	MilanSSD612	600	MauzyCountyK12	941	FranklinSSDK8
273	TrentonSSDK12	610	MeigsCountyK12	950	WilsonCountyK12
274	BradfordSSDK12	620	MonroeCountyK12	951	LebanonSSDK8

Variables: *S:\TEFF\662\R\school.dbf*

<u>NAME_</u>	<u>LABEL_</u>
MEDFAMIN	MEDIAN FAMILY INCOME 1989
AVGFAMI0	AVG FAMILIY INCOME: FAMS W 0 WRKRS 1989
AVGFAMI1	AVG FAMILIY INCOME: FAMS W 1 WRKR 1989
AVGFAMI2	AVG FAMILIY INCOME: FAMS W 2 WRKRS 1989
PHS1	% over 25 w/ 9-12 years education, no diploma
PEED	% over 25 w/ 0-8 years education, no diploma
PCL1	% over 25 w/ some college, no diploma
PCL2	% over 25 w/ associates degree
PCL3	% over 25 w/ bachelors degree
PCL4	% over 25 w/ degree past bachelors
PMGR	% of labor force managerial professional
PPRIV	% county's elementary and secondary teachers working in private schools
PQ	housing price index for county
PUPILS	number of pupils in system
TXPS	total expenditures per pupil
PWHITE	% pupils white
PBLACK	% pupils black
PHISPA	% pupils hispanic
PASIAN	% pupils asian
PNATIV	% pupils native american
TEAWAG	average teacher's wage
OVERSIZE	% classes oversize
LIMENG	number of students in limited English program
PUPMILE	pupils transported/miles transported
EXPEL	% pupils expelled
SUSPEND	%pupils suspended
LOCALREV	% revenue local
FEDREV	% revenue federal
STATEREV	% revenue state
PEMPTEA	% employment teachers
PEMPASS	% employment teaching assistants
PEMPSUP	% employment support
PEMPADM	% employment administrator
PEMPOTH	% employment other
VARE	value added test score reading
VAMA	value added test score math
VALA	value added test score language
VASO	value added test score social studies
VASC	value added test score science
NPRE	national percentile test score reading
NPMA	national percentile test score math
NPLA	national percentile test score language
NPSO	national percentile test score social studies
NPSC	national percentile test score science
ATK6	attendance rate k-6
AT7U	attendance rate 7-12
PRMO	promotion rate
CHRT	cohort drop-out rate

Monte-Carlo Program for Multicollinearity and Omitted Variable Bias---S:\TEFF\662\R\r06mc.R

```

#--Monte Carlo: Multicollinearity--
rm(list=ls(all=TRUE))
#--Set path to your own directory--
setwd("S:/teff/662/R/")
options(echo=TRUE)
library(foreign)
library(AER)

#---We make up our data, 500 times---
#---Though our data are randomly generated, we KNOW our coefficients--
#---We estimate the coefficients and compare the estimates with true values--

estcoef<-NULL
#--empty object to store our results from the 500 simulations--
x<-matrix(rnorm(100),50,2) #make 2 independent variables
x<-princomp(x)$scores #make 2 orthogonal indep. varbs
#x[,1]<-x[,2]^3 #make 2 highly correlated indep. varbs.
oC<-cor(x)[1,2] #correlation between indep. varbs.
for (i in 1:500){
y<-x%*%c(.5,7)+rnorm(50) #same as y=.5*x[,1]+7*x[,2]+error
o1<-coef(lm(y~x[,1])) #include only first indep. varb.
o2<-coef(lm(y~x)) #include both indep. varbs.
estcoef<-rbind(estcoef,cbind(o1[2],o2[2],.5,oC))
}
estcoef<-data.frame(estcoef)
names(estcoef)<-c("drop1","useBoth","trueVal","corr")

#--take mean and standard deviation of results--
apply(estcoef,2,mean)
apply(estcoef,2,sd)

xrange<-range(estcoef[,c("drop1","useBoth","trueVal")])

#--plot histograms of results--
layout(matrix(1:2,1,2))
hist(estcoef$drop1,breaks=30,xlim=xrange)
lines(density(estcoef$drop1),col="blue")
abline(v=.5,col="red",lty=2,lwd=2)
hist(estcoef$useBoth,breaks=30,xlim=xrange)
lines(density(estcoef$useBoth),col="blue")
abline(v=.5,col="red",lty=2,lwd=2)
layout(1)

```