IV. Multiple Regression.

V. Extensions of Multiple Regression

A. Non-Linear Models (Chapter 9)

B. Dummy (Binary) Variables (Chapter 10)

C. Scaling Variables

VI. Problems and Specification Issues

A. Model Selection/Specification

B. Multicollinearity

C. Heteroskedasticity

D. Autocorrelation

3. Diagnosis (Multicollinearity)

for model $H_0: \beta_1 = \dots = \beta_k = 0$

a. Classic signs: R^2 and F_{calc} tell you your model is good !!

BUT when you simultaneously have low t_{calcs} it suggest a problem

b. Correlation Coefficients

$-1 < r_{X_1 X_2} < +1$ or $0 < r^2_{X_1 X_2} < 1$ $\Rightarrow$ high correlations of 0.8 and above

c. Auxilliary Regressions

$X_2 = a + bX_1 + cX_3$ ← estimate aux. reg. check R^2 → is it high?

d. Variance Inflation Factors

related to c $\Rightarrow$ $VIF = \frac{1}{1-R^2}$ aux. reg. R^2

or - ask SAS or minitab

4. Example – annual per capita demand for chicken. Estimate the following model:

$$chikcons_t = \beta_0 + \beta_1\, pchik_t + \beta_2\, ppork_t + \beta_3\, pbeef_t + \beta_4\, disincom_t + u_t$$

8 α

Model: MODEL1
Dependent Variable: chikcons chikcons

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1127.25901	281.81475	73.87	<.0001
Error	18	68.66969	3.81498		
Corrected Total	22	1195.92870			

Fcalc for model

model great

Root MSE	1.95320	R-Square	0.9426
Dependent Mean	39.66957	Adj R-Sq	0.9298
Coeff Var	4.92367		

great

Parameter Estimates

$\alpha = 0.05$

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	37.23236	3.71770	10.01	<.0001
pchik	pchik	1	-0.61117	0.16285	-3.75*	0.0015
ppork	ppork	1	0.19841	0.06372	3.11*	0.0060
pbeef	pbeef	1	0.06950	0.05099	1.36	0.1896
disincom	disincom	1	0.00501	0.00489	1.02	0.3194

only 2 stat. important

a. Classic Signs – look on your printout for the following ***combination – a contradiction***:

- **Model is good:** Fits well and is significant.

  R^2 is high – suggests a good model.

  F_{calc} is high – suggests variables are important.

- ***BUT*: Individual t_{calcs} suggest variables are not important.** (***Contradicts*** the high R^2 and F_{calc} values)

b. Correlation Coefficients

ρ pop correlation coefficient

```
proc corr data=chicken;
 var pchik ppork pbeef disincom;
run;
```

$H_0: \rho = 0$

Pair-wise correlations – any problems?

$H_a: \rho \neq 0$

The CORR Procedure

4 Variables: pchik ppork pbeef disincom

Pearson Correlation Coefficients, N = 23
Prob > |r| under H0: Rho=0

	pchik	ppork	pbeef	disincom
pchik	1.00000	0.97011	0.92847	0.93168
		<.0001	<.0001	<.0001
ppork	0.97011	1.00000	0.94057	0.95713
	<.0001		<.0001	<.0001
pbeef	0.92847	0.94057	1.00000	0.98588
	<.0001	<.0001		<.0001
disincom	0.93168	0.95713	0.98588	1.00000
	<.0001	<.0001	<.0001	

all very high !! ☹

statistically significant

c. Auxilliary Regressions – are independent variables related?

Model: MODEL2
Dependent Variable: ppork ppork

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	26356	8785.33502	177.66	<.0001
Error	19	939.57493	49.45131		
Corrected Total	22	27296			

Root MSE	7.03216	**R-Square**	0.9656
Dependent Mean	90.40000	**Adj R-Sq**	0.9601
Coeff Var	7.77894		

$\Rightarrow VIF = \frac{1}{1-0.9656}$

$= 29.07$

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	**1**	-17.30398	12.78269	-1.35	0.1917
pchik	pchik	**1**	1.95988	0.37629	5.21	<.0001
pbeef	pbeef	**1**	-0.22436	0.17621	-1.27	0.2183
disincom	disincom	**1**	0.04015	0.01502	2.67	0.0150

d. Regression results with VIFs:

Model: MODEL3
Dependent Variable: chikcons chikcons

Clearly we have a problem!! ☹

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1127.25901	281.81475	73.87	<.0001
Error	18	68.66969	3.81498		
Corrected Total	22	1195.92870			

Root MSE	1.95320	**R-Square**	0.9426
Dependent Mean	39.66957	**Adj R-Sq**	0.9298
Coeff Var	4.92367		

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

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	37.23236	3.71770	10.01	<.0001	0
pchik	pchik	1	-0.61117	0.16285	-3.75	0.0015	18.90128
ppork	ppork	1	0.19841	0.06372	3.11	0.0060	29.05099
pbeef	pbeef	1	0.06950	0.05099	1.36	0.1896	39.76141
disincom	disincom	1	0.00501	0.00489	1.02	0.3194	52.70104

5. Solutions – fixing the problem

a. ~~***Sample data problem – get new sample data***~~ (Not a good suggestion – the new sample will probably have the same problem ☹)

b. ***Eliminate the offensive variable*** (But your results will be biased if that variable was important ☹)

* c. ***It's linear Association – use non-linear forms*** (Ok – this might work. Eg., try a log-log model (important need logs on right-hand side)

* d. ***Data transformations – try ratios of variables*** (This is often great, but the ratios must make sense!)

!! ☺ e. ***Use "non-sample" information - Restrictions*** (Great possible solution – but you need to have some theoretical result to use as a restriction) Theory !! ☺

c. Use non-linear form:

Model: MODEL4
Dependent Variable: lnqchik

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.76105	0.19026	249.93	<.0001
Error	18	0.01370	0.00076127		
Corrected Total	22	0.77475			

Root MSE	0.02759	**R-Square**	0.9823
Dependent Mean	3.66389	**Adj R-Sq**	0.9784
Coeff Var	0.75306		

non linear model did NOT solve the problem

Parameter Estimates

elasticities

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	2.18979	0.15571	14.06	<.0001	0
lnpchik		1	-0.50459	0.11089	-4.55	0.0002	17.48577
lnppork		1	0.14855	0.09967	1.49	0.1535	41.43312
lnpbeef		1	0.09110	0.10072	0.90	0.3776	42.30710
lndinc		1	0.34256	0.08327	4.11	0.0007	65.11460

!!

Correlations for the log variables – no improvement

The CORR Procedure
5 Variables: lnpchik lnppork lnpbeef lndinc

Pearson Correlation Coefficients, N = 23
Prob > |r| under H0: Rho=0

	lnpchik	lnppork	lnpbeef	lndinc
lnpchik	1.00000	0.94675 <.0001	0.93306 <.0001	0.90717 <.0001
lnppork	0.94675 <.0001	1.00000	0.95428 <.0001	0.97246 <.0001
lnpbeef	0.93306 <.0001	0.95428 <.0001	1.00000	0.97900 <.0001
lndinc	0.90717 <.0001	0.97246 <.0001	0.97900 <.0001	1.00000

yikes!!

d. Data transformation – ratios of variables – relative prices:

$lnrpchik = \ln\left(\frac{pchik}{pbeef}\right)$

$lnrdin = \ln\left(\frac{dinc}{pbeef}\right)$

Model: MODEL6
Dependent Variable: lnqchik

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.75851	0.25284	295.85	<.0001
Error	19	0.01624	0.00085463		
Corrected Total	22	0.77475			

Root MSE	0.02923	R-Square	0.9790
Dependent Mean	3.66389	Adj R-Sq	0.9757
Coeff Var	0.79790		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	2.38310	0.12093	19.71	<.0001	0
lnrpchik		1	-0.61246	0.09942	-6.16*	<.0001	9.22794
lnrppork		1	0.15750	0.10548	1.49°	0.1518	3.77977
lnrdinc		1	0.38228	0.08516	4.49*	0.0003	8.44858

!! :)

clearly helped

Correlations for relative prices – these look pretty good.

mitigated multicollinearity !! :)

The CORR Procedure
3 Variables: rpchik rppork rdincom

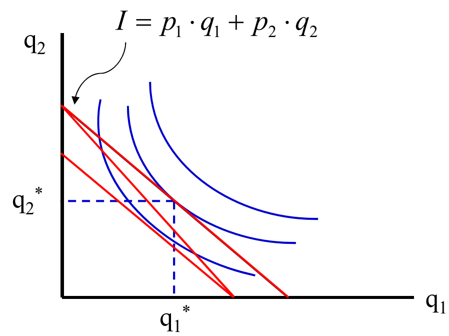
Pearson Correlation Coefficients, N = 23
Prob > |r| under H0: Rho=0

	rpchik	rppork	rdincom
rpchik	1.00000	0.27577 0.2028	-0.80185 <.0001
rppork	0.27577 0.2028	1.00000	0.20395 0.3506
rdincom	-0.80185 <.0001	0.20395 0.3506	1.00000

e. Use non-sample information - restricted least squares:

Non-sample information – theoretical restriction for population parameters.
Example: Consumer demands are homogeneous of degree zero – if all prices and income change by the same proportion, demand does not change.

- p_1, p_2, and I increase by same proportion (λ).
- What happens to q_1^* and q_2^*?

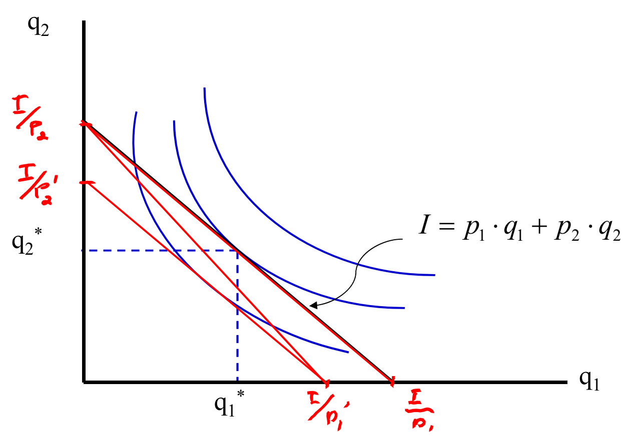


e. Use non-sample information - restricted least squares:

- p_1, p_2, and I increase by same proportion (λ).
- What happens to q_1^* and q_2^*?

Nothing!

consumer demands are homogeneous of degree zero



What does that mean for the demand function for q_1:

$$q_1 = q_1(p_1, p_2, I)$$

Totally differentiate the demand function:

$$d\,q_1 = \left(\frac{\partial q_1}{\partial p_1}\right) d\,p_1 + \left(\frac{\partial q_1}{\partial p_2}\right) d\,p_2 + \left(\frac{\partial q_1}{\partial I}\right) d\,I$$

Divide by q_1 and multiply by "1":

$$\frac{d\,q_1}{q_1} = \frac{\partial q_1}{\partial p_1}\frac{1}{q_1}\left(\frac{p_1}{p_1}\right) d\,p_1 + \frac{\partial q_1}{\partial p_2}\frac{1}{q_1}\left(\frac{p_2}{p_2}\right) d\,p_2 + \frac{\partial q_1}{\partial I}\frac{1}{q_1}\left(\frac{I}{I}\right) d\,I$$

$$\frac{d\,q_1}{q_1}^{0} = \underbrace{\boxed{\left(\frac{\partial q_1}{\partial p_1}\frac{p_1}{q_1}\right)}}_{\varepsilon_{p_1}} \overbrace{\left[\frac{d\,p_1}{p_1}\right]}^{\lambda} + \underbrace{\left(\frac{\partial q_1}{\partial p_2}\frac{p_2}{q_1}\right)}_{\varepsilon_{p_2}} \overbrace{\left[\frac{d\,p_2}{p_2}\right]}^{\lambda} + \underbrace{\left(\frac{\partial q_1}{\partial I}\frac{I}{q_1}\right)}_{\eta} \overbrace{\left[\frac{d\,I}{I}\right]}^{\lambda}$$

$$\frac{d\,q_1}{q_1} = \left(\frac{\partial q_1}{\partial p_1}\frac{p_1}{q_1}\right)\frac{d\,p_1}{p_1} + \left(\frac{\partial q_1}{\partial p_2}\frac{p_2}{q_1}\right)\frac{d\,p_2}{p_2} + \left(\frac{\partial q_1}{\partial I}\frac{I}{q_1}\right)\frac{d\,I}{I}$$

The terms in parentheses are ***elasticities***:

$$\frac{d\,q_1}{q_1} = \varepsilon_{p_1}\left(\frac{d\,p_1}{p_1}\right) + \varepsilon_{p_2}\left(\frac{d\,p_2}{p_2}\right) + \varepsilon_I\left(\frac{d\,I}{I}\right)$$

If all prices and income *increase by the same proportion* (λ), there is *no change in quantity demanded* ($dq_1 = 0$):

$$\frac{d\,q_1}{q_1} = \varepsilon_{p_1}\lambda + \varepsilon_{p_2}\lambda + \varepsilon_I\lambda = \boxed{\left(\varepsilon_{p_1} + \varepsilon_{p_2} + \varepsilon_I\right)}\lambda = 0$$

homogeneity tells us the sum of all price and income elasticities must be zero

e. Use non-sample information - restricted least squares:

homog.

$b_1 + b_2 + b_3 + b_4 = 0$

$b_2 - b_3 = 0$

3.

Model: MODEL5
Dependent Variable: lnqchik

Note: Restrictions have been applied to parameter estimates.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.75799	0.37900	452.31	<.0001
Error	20	0.01676	0.00083792		
Corrected Total	22	0.77475			

Root MSE	0.02895	**R-Square**	0.9784
Dependent Mean	3.66389	**Adj R-Sq**	0.9762
Coeff Var	0.79006		

restrictions appear true (if they were not true, the results would be biased)

$a = 0.05$

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	b_0 2.37252	0.11899	19.94	<.0001	0
lnpchik		1	b_1 -0.61227	0.09844	-6.22	<.0001	12.57797
lnppork		1	b_2 0.11561	0.08991	1.29	0.2132	30.63080
lnpbeef		1	b_3 0.11561	0.08991	1.29	0.2132	30.63194
lndinc		1	b_4 0.38104	0.08430	4.52	0.0002	60.64222
RESTRICT	Homog.	-1	0.03483	0.01898	1.84	0.0647*	
RESTRICT		-1	0.00614	0.00778	0.79	0.4448*	

*** Probability computed using beta distribution.**

← used to test if restr. is true

F-tests of the two restrictions:

1. Homogeneity: $H_0:\ \beta_1 + \beta_2 + \beta_3 + \beta_4 = 0;$

$H_0:\ \beta_1 + \beta_2 + \beta_3 + \beta_4 \neq 0$

Same conclusions as the "Beta" tests above.

The REG Procedure
Model: MODEL4
Test 1 Results for Dependent Variable lnqchik

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.00254	3.33	0.0847
Denominator	18	0.00076127		

2. Equal substitution effects: $H_0:\ \beta_2 - \beta_3 = 0;$

$H_0:\ \beta_2 - \beta_3 \neq 0$

Model: MODEL4
Test 2 Results for Dependent Variable lnqchik

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.00023365	0.31	0.5864
Denominator	18	0.00076127		