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

VI. Problems and Specification Issues

A. Model Selection/Specification

1. Two Possible Mistakes: omit important independent variables; include variables that don't belong
2. Omitted Variables: estimators are ***biased***
3. Irrelevant Variables: estimators are ***unbiased***, but inefficient – OLS estimators are no longer BLUE
4. Tools and Tests

4. Tools

a. Start with theory and literature review:
   - ✓ Do you have all the right variables? What did other researchers find?

b. Check the Adjusted R^2:
   - ✓ Do additional variables explain variation in Y?

c. Joint F-tests for additional variables

d. Specification tests:
   - ✓ Regression Error Specification Test (RESET)
   - ✓ SAS – use the "SPEC" (specification test) option

RESET Test

i. Estimate: $Y_i = \beta_0 + \beta_0 X_{1i} + \ldots + \beta_K X_{Ki} + u_i$

ii. Save predicted values:

iii. Create new variables:

iv. Estimate again:

v. Hypothesis test:

White's Specification Test (SPEC option)

i. Estimate: $Y_i = \beta_0 + \beta_0 X_{1i} + \ldots + \beta_K X_{Ki} + u_i$

ii. Save the errors and square:

iii. Create new variables:

iv. Estimate the auxiliary regression:

v. Hypothesis test:

B. Multicollinearity

1. **Definition**: The presence of ***linear association*** among independent variables – i.e. linear association or correlation between X_1 and X_2.

- ***Sample Problem*** – the problem lies in your sample data.
- There is ***no causal relationship*** between X_2 and the other independent variables.
- Ie., X_2 does not "cause" X_1

2. Consequences:

- OLS estimators – remain unbiased.
- **Standard errors are inflated.**
- **Calculated t-statistics are deflated.**
- **What is the ultimate problem?**

2. Consequences of Multicollinearity

- Multiple reg. standard error – 2 indep. variables:

- If X_1 and X_2 are strongly and linearly associated:

- Multiple regression variances are inflated – the *variance inflation factor*

1. Multiple regression variance:

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2}$$

2. Multiply by "1":

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2} \cdot \frac{1/\left(\sum x_{1i}^2 \sum x_{2i}^2\right)}{1/\left(\sum x_{1i}^2 \sum x_{2i}^2\right)}$$

3. Rearrange terms and note that some stuff cancels:

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 \sum x_{2i}^2 \Big/ \left(\sum x_{1i}^2 \sum x_{2i}^2\right)}{\frac{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}} = \frac{\hat{\sigma}^2 \Big/ \sum x_{1i}^2}{\frac{\sum x_{1i}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2} - \frac{(\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}}$$

4. Simplify the denominator:

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 \Big/ \sum x_{1i}^2}{\frac{\sum x_{1i}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2} - \frac{(\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}} = \frac{\hat{\sigma}^2 \Big/ \sum x_{1i}^2}{1 - \frac{(\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}}$$

5. Two familiar terms:

Simple regression variance:

$$s_{b_1}^2 = \frac{\hat{\sigma}^2}{\sum x_{1i}^2}$$

Squared correlation coefficient:

$$r_{x1x2}^2 = \frac{(\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}$$

6. Viola! The variance of the multiple regression estimator is the variance of the simple regression estimator multiplied by the ***VIF:***

$$s_{b_1}^2 = \frac{\hat{\sigma}^2}{\sum x_{1i}^2} \cdot \left(\frac{1}{1 - r_{x_1x_2}^2}\right)$$

3. Diagnosis (Multicollinearity)

a. Classic signs:

b. Correlation Coefficients

c. Auxilliary Regressions

d. Variance Inflation Factors

4. Example – annual per capita demand for chicken.

Estimate the following model:

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

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			

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

Parameter Estimates

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

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

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

Pair-wise correlations – any problems?

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	

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		

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

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		

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)

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		

Parameter Estimates

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

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

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

Correlations for relative prices – these look pretty good.

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:

Model: MODEL5
Dependent Variable: lnqchik

Note: Restrictions have been applied to parameter estimates.

$$b_1 + b_2 + b_3 + b_4 = 0$$

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		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	2.37252	0.11899	19.94	<.0001	0
lnpchik		1	-0.61227	0.09844	-6.22	<.0001	12.51797
lnppork		1	0.11561	0.08991	1.29	0.2132	30.63080
lnpbeef		1	0.11561	0.08991	1.29	0.2132	30.63194
lndinc		1	0.38104	0.08430	4.52	0.0002	60.64222
RESTRICT		-1	0.03483	0.01898	1.84	0.0647*	.
RESTRICT		-1	0.00614	0.00778	0.79	0.4448*	.

*** Probability computed using beta distribution.**

F-tests of the two restrictions:

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		

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		