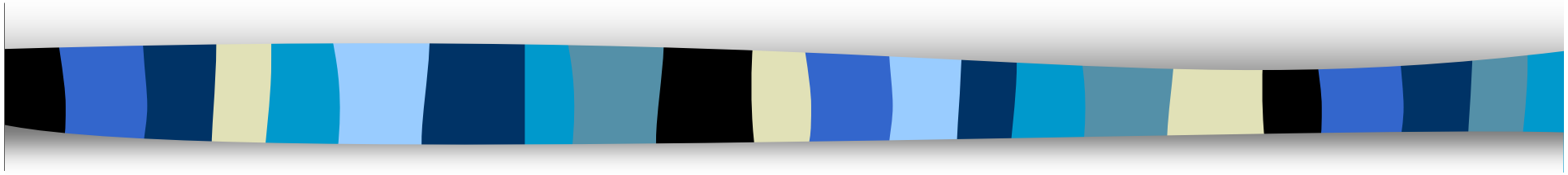
Graduate School

Quantitative Research Methods

Gwilym Pryce



Module II

Lecture 7: Multicollinearity, and Modeling Strategies

Notices:

- **Assignment:**
  - much less guidance than for quants I
  - provided with a data set and expected to construct a regression model from it.
  - The only guidance is regarding the format of the report and a statement saying that you need to follow “good modelling practice”
    - I.e. the strategies to be outlined in this lecture.
- **Important to Attend the Labs & do reading!!**

Plan

- 1. What is multicollinearity?
- 2. Causes
- 3. Consequences
- 4. Detection
- 5. Solutions
- 6. Modeling Strategies

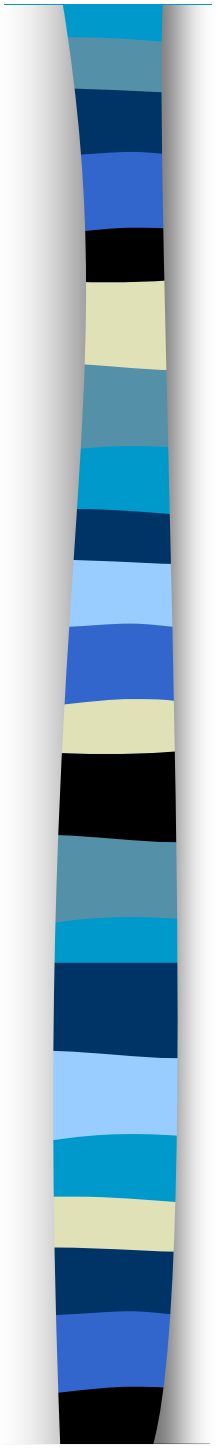
1. What is Multicolinearity?

- multicolinearity occurs when the explanatory variables are highly inter-correlated.
- This may not necessarily be a problem, but it can prevent precise analysis of the individual effects of each variable
- Consider the case of just $k = 2$ explanatory variables and a constant. For either slope coefficient, the square of the standard error is:

$$\mathrm{Var}[b_k] = \frac{\sigma^2}{(1 - r_{12}^2)\sum_i (x_{ik} - \bar{x}_k)^2}$$

- If the two variables are perfectly correlated, $r_{12}^2 = 1$ (where r_{12}^2 is the square of the simple correlation coefficient between x_1 and x_2), then the variance of the estimated slope coefficient will be infinite:

$$\text{Var}[b_k] = \frac{\sigma^2}{(1-r_{12}^2)\sum_i (x_{ik} - \bar{x}_k)^2} = \frac{\sigma^2}{0} = \infty$$

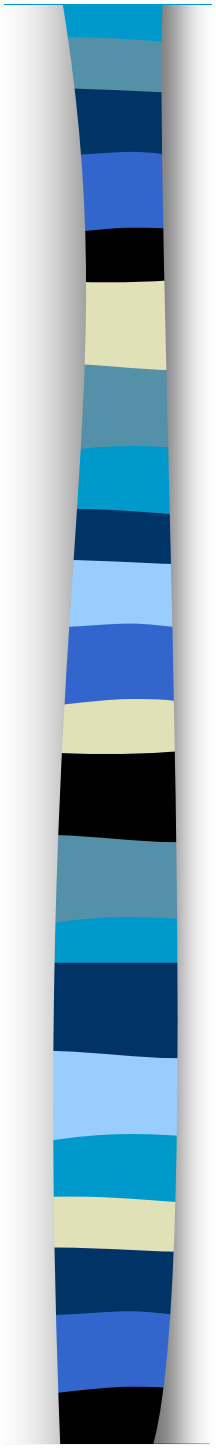


- Perfect multicollinearity usually only occurs because of model misspecification rather than measurement problems
- more common case is where the variables are highly but not perfectly correlated

2. Causes

- Causes of Perfect Multicollinearity:
  - Dummy variable trap
    - Improper use of dummy variables (e.g. failure to exclude one category)
  - Conceptual linear sum
    - Including a variable that can be computed from other variables in the equation
      - e.g. family income = husband's income + wife's income, and the regression includes all 3 income measures

- Two or more measures of the same entity:
  - including the same or almost the same variable twice
    - e.g height in feet and height in inches;
  - more commonly, two different operationalizations of the an identical concept
    - e.g. including two different indices of IQ -- the method of measurement is different but the underlying phenomena is fundamentally the same.



- The above all imply some sort of error on the researcher's part.
- But, it is possible that different causal variables happen to be highly correlated
- or that measurement methods fail to distinguish the underlying concepts we believe to be causes of *y*.

- Causes of Near multicollinearity
  - Measurement failure: unable to distinguish between entities:
    - the variables to be measured were not defined in a way that would allow the separation of different effects when the variables come to be analysed
    - you really need to understand the modelling process *before* you collect your data!

3. Consequences

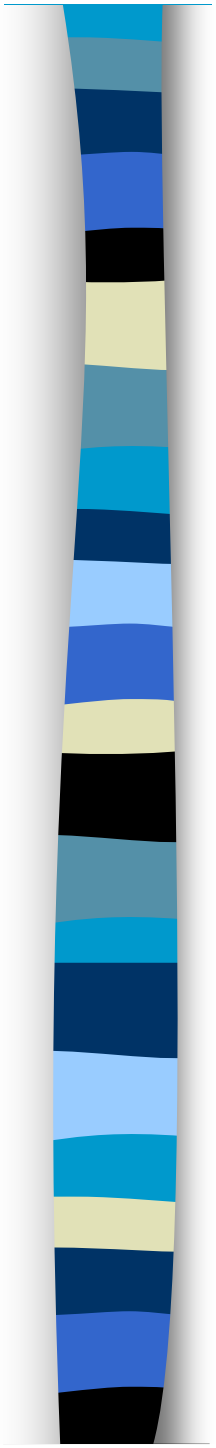
- Perfect Multicollinearity:
  - suppose we attempt to estimate the following regression:

$$Consumption = b_1 + b_2\ nonlabour\ income + b_3\ salary + b_4\ total\ income$$

(Greene p. 267)

  - it will not be possible to separate out individual effects of the components of income ($N + S$) and total income (T)

– In other words, this regression specification allows the same value of C^{hat} for many different values of the slope coefficients.



- This is called the "**identification problem**" and most statistical packages will come up with an error message if you try to run a regression suffering from perfect multicollinearity.
- Note, though, that this is a poorly specified model and the problems of identification have nothing to do with the quality of the data.

Consequences of Near Multicollinearity:

- When the correlation between explanatory variables is high but not perfect, then the difficulty in estimation is not one of identification but of precision.
- The higher the correlation between the regressors, the less precise our estimates will be (I.e. the greater the standard errors on the slope parameters):

$$\mathrm{Var}[b_k] = \frac{\sigma^2}{(1 - r_{12}^2)\sum_i (x_{ik} - \bar{x}_k)^2}$$

- But even where there is extreme multicollinearity, so long as it is not perfect OLS assumptions will not be violated.
  - OLS estimates of *that particular model* are still BLUE (Best Linear Unbiased Estimators)
  - Alterations to the model, however, may increase efficiency
    - I.e. reduce the variance of the estimated slopes

- When high multicollinearity is present, confidence intervals for coefficients tend to be very wide and t-statistics tend to be very small.
- Note, however, that large standard errors can be caused by things other than multicollinearity
  - e.g. if σ^2, the standard error of the residuals, is large

$$\mathrm{Var}[b_k] = \frac{\sigma^2}{(1-r_{12}^2)\sum_i (x_{ik} - \bar{x}_k)^2}$$

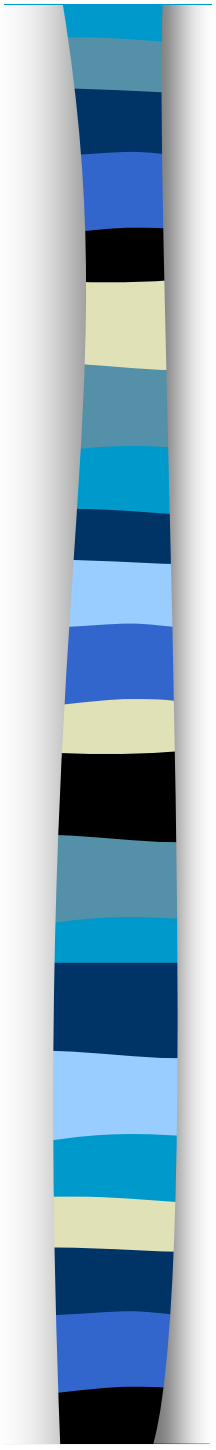
- When two explanatory variables are highly and *positively* correlated, their slope coefficient estimators will tend to be highly and *negatively* correlated.
- But a different sample could easily produce the opposite result if there is multicollinearity because coefficient estimates tend to be very unstable from one sample to the next.
- Coefficients can have implausible magnitude

4. Detection

- *Check for unstable parameter values across subsamples:*
  - *Step 1:* create an arbitrary random variable, *Q* and order your sample by *Q* (alternatively you can use the random subsample facility in SPSS)
  - *Step 2:* run the same regression on different sub-samples (e.g. first 100 observations vs rest)
  - *Step 3:* do F-tests to see if the slopes change

- *Check for unstable Parameters Across Specification:*
  - try a slightly different specification of a model using the same data. See if seemingly “innocuous” changes (adding a variable, dropping a variable, using a different operationalization of a variable) produce big shifts.
  - As variables are added, look for changes in the signs of effects (e.g. switches from positive to negative) that seem theoretically questionable.

- *Check the t ratios:*
  - If none of the t-ratios for the individual coefficients are statistically significant, yet the overall F statistic is, then you may have multicolinearity.
  - Note, however, the word of caution from Greene:

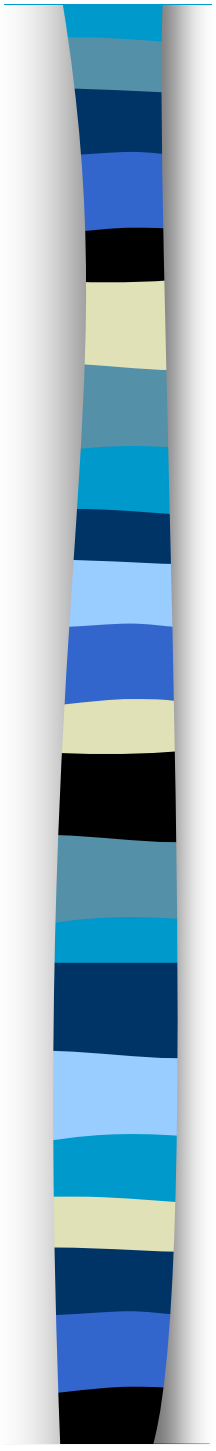


- “It is tempting to conclude that a variable has a low *t* ratio, or is significant, because of multicolinearity. One might (some authors have) then conclude that if the data were not collinear, the coefficient would be significantly different from zero.
- Of course, this is not necessarily true.
  - Sometimes a coefficient turns out to be insignificant because the variable does not have any explanatory power in the model”

- *Check the Simple Correlation Matrix:*
  - The simple correlation coefficient, *r*(*x*,*z*), has the same sign as the covariance but only varies between -1 and 1 and is unaffected by any scaling of the variables.

$$r_{xz} = \frac{\sigma_{xz}}{\sigma_x \sigma_z}$$

  - This measure is useful if we have only two explanatory variables.
  - If the number of explanatory variables is greater than 2, the method is useless since near multicolinearity can occur when any one explanatory variable is a near *linear combination* of any collection of the others.



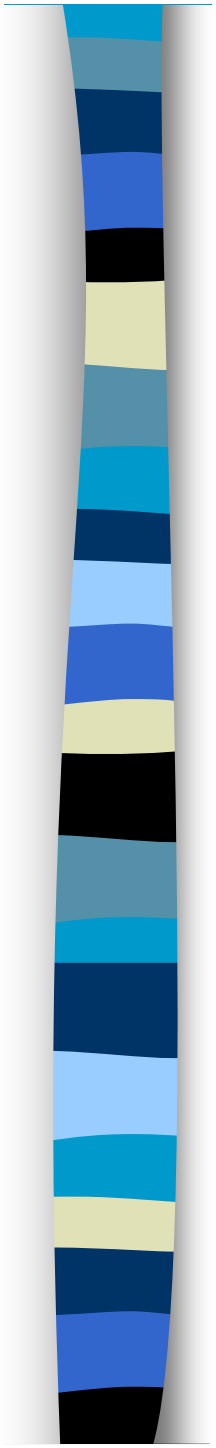
- Thus, it is quite possible for one *x* to be a linear combination of several *x*'s, and yet not be highly correlated with any one of them :
  - the correlation coefficient (which only measures bivariate correlation) to be small,
  - but for the squared multiple correlation coefficient (I.e. the R^2, which measures multivariate correlation) between the explanatory variables to be high.
- It is also hard to decide on a cut-off point. The smaller the sample, the lower the cut-off point should probably be.

- *Check* R_k^2
  - when you have more than one explanatory variable, you should run regressions of each on the others to see if there is multicolinearity
    - this is probably the best way of investigating multicolinearity since examining coefficients will also help you find the source of the multicolinearity.
  - If you have lots of regressors, however, this can be a daunting task, so you may want to start by looking at the *Tolerance* and *VIF*...

- *Check the Tolerance and VIF*
  - the general formula (as opposed to the one where you have just 2 regressors) for the variance of the slope coefficient estimate is:

$$\mathrm{Var}[b_k] = \frac{\sigma^2}{(1-R_k^2)\sum_i (x_{ik} - \bar{x}_k)^2}$$

  - where R_k^2 is the squared multiple correlations coefficient between x_k and the other explanatory variables
    - e.g. R^2 from the regression: $x_1 = a_1 + a_2x_2 + a_3x_3$



- 1 - R_k^2 is referred to as the *Tolerance* of x_k.
  - A *tolerance* close to 1 means there is little multicolinearity, whereas a value close to 0 suggests that multicolinearity may be a threat.
- The reciprocal of the tolerance is known as the *Variance Inflation Factor (VIF)*.
  - The *VIF* shows us how much the variance of the coefficient estimate is being inflated by multicolinearity.
  - A *VIF* near to one suggests there is no multicolinearity, whereas a VIF near 5 might cause concern.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-5863.031	674.178		-8.697	.000		
	Bedrooms	12885.593	280.324	.344	45.967	.000	.730	1.371
	PublicRooms	26431.578	439.243	.434	60.175	.000	.785	1.273
	HasGarden	347.620	516.156	.005	.673	.501	.843	1.186
	Time on the Market (number of days)	-15.213	1.289	-.076	-11.798	.000	.997	1.003

a. Dependent Variable: SellingPrice

- All the *VIF* levels in this regression are near to one so there is no real problem.
- If *VIF* where high for a particular regressor, say *z*, then we might want to run a regression of *z* on the other explanatory variables to see variables are closely related.
- We could then consider whether to omit one or more of the variables
  - e.g. if on deliberation we decide that they are in fact measuring the same thing

- *Check the Eigenvalues and Condition Index:*
  - eigenvalues indicate how many distinct dimensions there are among the regressors
  - when several eigenvalues are close to zero, there may be a high level of multicolinearity.
  - Condition Indices are the square roots of the ratio of the largest eigenvalue to each successive eigenvalue.
    - **Condition Index Values above 30 suggest a problem**

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Bedrooms	PublicRooms	HasGarden	Time on the Market (number of days)
1	1	4.028	1.000	.01	.01	.01	.01	.02
	2	.597	2.597	.00	.01	.01	.04	.88
	3	.211	4.370	.04	.03	.09	.92	.05
	4	8.810E-02	6.761	.12	.39	.86	.03	.01
	5	7.643E-02	7.259	.83	.57	.03	.01	.05

a. Dependent Variable: SellingPrice

- Two of the eigenvalues are pretty small, but:
- the Condition Indices are all below 10 so there is unlikely to be a problem with multicolinearity here.

- Problems with the Condition Index Approach:
  - the condition number can change by a reparametrization of the variables: “it can be made equal to one with suitable transformations of the variables” (Maddala, p. 275)
  - such transformations can be meaningless
  - does not tell you whether the multicolinearity is actually causing problems or how to go about resolving the problems if they exist.

Regression Collinearity Diagnostics

Index Next

This table displays statistics that help you determine if there are any problems with collinearity.

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Infant Mortality	%Urban
1	1	1.751	1.000	.12	.12	
	2	.249	2.653	.88	.88	
2	1	2.461	1.000	.01	.02	.01
	2	.514	2.189	.00	.26	.05
	3	.025	9.913	.99	.72	.94
3	1	3.349	1.000	.00	.01	.00
	2	.581	2.400	.00	.07	.06
	3	.045	8.585	.01	.79	.05
	4	.024	11.786	.99	.14	.89
4	1	4.130	1.000	.00	.00	.00
	2	.589	2.647	.00	.06	.05
	3	.229	4.243	.02	.03	.03
	4	.028	12.173	.11	.91	.22
	5	.023	13.305	.87	.00	.69

This table displays statistics that help you determine if there are any problems with collinearity.

Collinearity (or multicollinearity) is the undesirable situation where the correlations among the independent variables are strong.

Eigenvalues provide an indication of how many distinct dimensions there are among the independent variables.

When several eigenvalues are close to zero, the variables are highly intercorrelated and small changes in the data values may lead to large changes in the estimates of the coefficients.

Model	Dimension	Eigenvalue	Condition Index
1	1	1.751	1.000
	2	.249	2.653
2	1	2.461	1.000
	2	.514	2.189
	3	.025	9.913
3	1	3.349	1.000
	2	.581	2.400
	3	.045	8.585
	4	.024	11.786
4	1	4.130	1.000
	2	.589	2.647
	3	.229	4.243
	4	.028	12.173
	5	.023	13.305
5	1	5.014	1.000
	2	.672	2.732
	3	.249	4.485
	4	.035	11.888
	5	.026	14.006
	6	.004	36.783

Condition indices are the square roots of the ratios of the largest eigenvalue to each successive eigenvalue.

A condition index greater than 15 indicates a possible problem and an index greater than 30 suggests a serious problem with collinearity.

Model	Dimension	Eigenvalue	Condition Index
1	1	1.751	1.000
	2	.249	2.653
2	1	2.461	1.000
	2	.514	2.189
	3	.025	9.913
3	1	3.349	1.000
	2	.581	2.400
	3	.045	8.585
	4	.024	11.786
4	1	4.130	1.000
	2	.589	2.647
	3	.229	4.243
	4	.028	12.173
	5	.023	13.305
5	1	5.014	1.000
	2	.672	2.732
	3	.249	4.485
	4	.035	11.888
	5	.026	14.006
	6	.004	**36.783**

Model	Dimension	Variance Proportions (Constant)	Infant Mortality	%Urban	Fertility	Birth Death Ratio
1	1	.12	.12			
	2	.88	.88			
		.01	.02	.01		
		.00	.26	.05		
		.99	.72	.94		
		.00	.01	.00	.01	
		.00	.07	.06	.01	
		.01	.79	.05	.91	
	4	.99	.14	.89	.08	
4	1	.00	.00	.00	.00	.01
	2	.00	.06	.05	.01	.00
	3	.02	.03	.03	.00	.53
	4	.11	.91	.22	.64	.35
	5	.87	.00	.69	.35	.11

Collinearity is a problem when a component associated with a high condition index contributes substantially to the variance of two or more variables.

5. Solutions

- Solving Perfect Multicolinearity
  - check whether you have made any obvious errors
    - e.g. improper use of computed or dummy variables (particularly for perfect multicol^{y}).

- Solutions to Near Multicolinearity:
  - Do nothing!
    - NB: only needs “solving” if it is having an adverse effect on your model
    - e.g. large SEs, unstable signs on coefficients.
  - Factor analysis, Principle components or some other means to create a scale from the X’s.
    - This solution is not recommended in most instances since the meaning of coefficients on your created factors are difficult to interpret:

– e.g. 3 problems of Princ. Comp. (Greene p. 273):
  - "First, the results are quite sensitive to the scale of measurement in the variables. The obvious remedy is to standardize the variables, but, unfortunately, this has substantial effects on the computed results.
  - Second, the principle components are not chosen on the basis of any relationship of the regressors to *y* , the variable we are attempting to explain.

- Lastly, the calculation makes ambiguous the interpretation of results. The principle components estimator is a mixture of all of the original coefficients. It is unlikely that we shall be able to interpret these combinations in any meaningful way.”

- Use joint hypothesis tests:
  - I.e. as well as doing t-tests for individual coefficients, do an F test for a group of coefficients So, if x_1, x_2, and x_3 are highly correlated, do an F test of the hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$.
- Omitted Variables Estimation:
  - I.e. "drop" the offending variable. But, if the variable really belongs in the model, this can lead to specification error, which can have far worse consequences (I.e. bias) than multicolinear model (which is BLUE).

– Ridge Regression:

- Deliberately adds bias to the estimates to reduce the standard errors
- “it is difficult to attach much meaning to hypothesis tests about an estimator that is biased in an unknown direction” (Greene)

6. Modelling Strategies

- Whether or not you present the results of the diagnostics to your audience, you MUST construct your model using them otherwise:
  - how do you know that you have specified it correctly?
  - How do you know that it can be generalised beyond your little sample!?
- <u>E.g. A Salutary Tale…</u>
  - You construct a model of mortality rate:
  - mortality rate = b1 + b2 smoking rate + b3 ave age
  - you did not include in your model a whole range of variables because when you entered them in individually, there were not significant (I.e. t < 2)

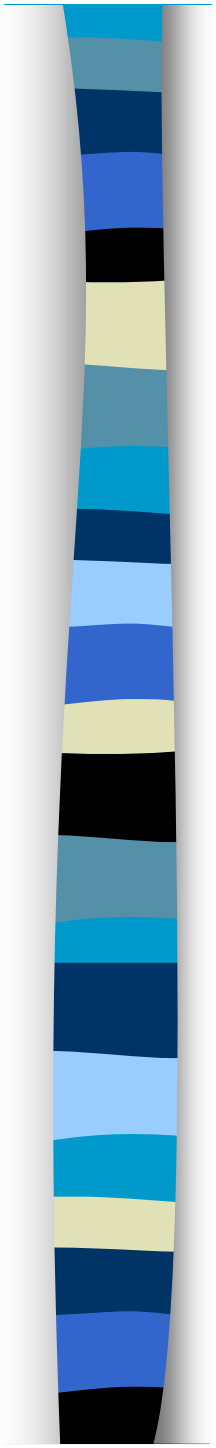
- however, it turns out that your model suffered from heteroscedasticity and so the t-tests were incorrect:
  - if used White's SEs , Unemployment and School Achievement both signif.
- You used simple correlation coefficients between variables to identify multicolinearity
  - => kept Smoking Rate and Age but dropped Unemployment etc
  - but your method was spurious: actually should drop age and keep Unt and School Achieve

- You did not test for parameter stability across subsamples:
  - Your model was not stable across different parts of the country or over time
    - » in some areas, unemployment was actually the most important driver
    - » estimates based on a subsample of the most recent 4 years showed unemployment to have a much larger coefficient than in your model
  - your model was actually totally inapplicable to certain areas (Highlands) and subsample Chow tests would have revealed this.

- You did not check for non-linearities or interactive effects
  - turns out that there is a highly significant quadratic relationship with unemployment and a strong interaction with whether or not the area is urban

CONCLUSION:

- your model is USELESS!!!
  - Worse than that, it is misleading and could distort policy outcomes
  - A few years later, other models are developed (with equal disregard to diagnostics) which produce radically different results,
  - As a result, policy makers become disillusioned with statistical models and resort to their own “good judgement”!
  - The world comes to an end an it was all YOUR fault!!!



To avoid this nightmare scenario you need…

… a sound modelling strategy:

- General to Specific
  - start with all variables & all sample
  - reduce & refine as necessary
- Specific to General
  - start with few variables & specific sample
  - expand & refine incrementally
- One balance, I would recommend the first of these approaches, but both are defensible if used in conjunction with thorough diagnostic testing...

General to Specific model steps:

- (i) Theory
- (ii) Anticipated Regression Model
- (iii) Data Collection
- (iv) General Model
- (v) Diagnostic Checks and Refinement
- (vi) Specific Model
- (vii) Revise Theory?
- (iix) Present Final Model

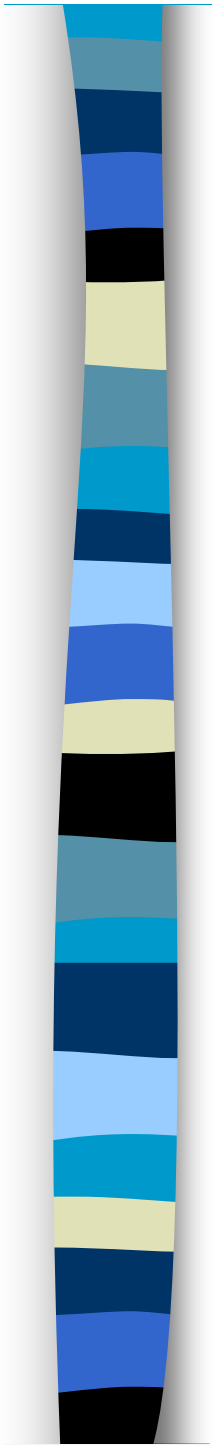
- (i) Theory
  - Always start with theory (qualitative research may help here).
  - Try to cater for all possible determinants
  - Try to identify specific hypotheses you want to test
- (ii) Anticipated Regression Model
  - identify the regression model that follows from your theory and that will allow you to test the hypotheses you are most interested in.
- (iii) Data Collection & Coding
  - make sure the data collect, the way you collect it (I.e. unbiased sampling, large *n*, precise measurement) & the coding will allow you to build your general model and test specific hypotheses

(iv) General Model

- attempt your first regression model
  - start with all available variables and all available observations
- make obvious modifications before starting the diagnostic/refinement process

(v) Diagnostic Checks and Refinement

- Examine Residual plots
  - scatter plots of residuals on y & xs
  - should be “spherical”
  - normal probability plots
  - outliers (use Cook’s distances etc.)
- Heteroscedasticity
  - Test using B-P etc.
  - If heterosk. exists, use White’s SEs & Chow’s 2nd Test
- Wrong signs
  - t-tests & multicolinearity tests
  - RAMSEY reset test.
  - Non-linear Transformations
  - interactions



- Low Adjusted R^2
  - Transform variables
  - drop irrelevant variables
  - get data on new variables
- F-Tests
  - structural stability (Chow)
  - linear restrictions
- Multicolinearity
  - check VIF, eigenvalues, Condition indices etc.
  - present joint hypothesis tets.

(vi) Specific Model

- should be “well behaved”
  - stable
  - passes general misspecification tests if possible
    - » e.g. RESET test
- coefficients should be meaningful
  - do the coefficients make sense?
  - How do they relate to your theory/intuition?
  - Alternative explanations/interpretations

(vii) Revise Theory?

- Do your empirical mean that you need to modify your initial theory, hypotheses and anticipated empirical model?
- Often, it is only when you start the empirical process that you really grasp the key aspects or limitations of your theory

(iix) Present the Final model (to an academic audience)

- you should present your (revised) theory first
- then the (revised) anticipated regression model
- then discuss the data and measurement of (revised) anticipated variables
- then present a selection of regression models
  - present a series of “preferred” regressions which might vary by:
    - selection of regressors
    - measurement of dependent variable
    - and/or sample selection

- present the selection of regressions in columns all in a single table rather than as separate tables -- this will assist comparison
- only present statistics that you explain/discuss in your text
  - always present sample size, Adjusted R^2, t values on individual coefficients or SEs or Sig.

- then offer a full discussion
  - I.e. of the different regressions and statistics that you have presented and discuss any relevant elements of the refinement process
  - this discussion should lead you to select a final "preferred" model(s) (if there is one) on the basis of the diagnostics, intuition and relevance to the theory
    - it is a good idea to present this in a separate table in more detail -- e.g. with confidence intervals for the coefficients
  - you should comment on the limitations of you model given the data and the anticipated effect of measurement problems, omitted variables, bias in sample, insufficient sample size etc.

- Then present the results of your specific hypothesis tests
  - these should be run on your final preferred model(s) and include a full discussion of their meaning and the limitations implied by the inadequacies of your model.
- If you are presenting to a non-academic audience, you will have to select which of the above are likely to be most meaningful/important to them.
  - Whether or not you present the results of the diagnostics, you MUST construct your model using them otherwise:
    - how do you know that you have specified it correctly?
    - How do you know that it can be generalised beyond your little sample!?

Reading

On multicolinearity:

- Kennedy chapter 11.
- Field, A. (2001) “Discovering Statistics” p. 131 onwards
- Maddala, G.S. (1992) Introduction to Econometrics, 2nd ed, Maxwell, chapter 7.
- Greene, W. H. (1993) “Econometric Analysis” p.273
  - Excellent but technical.
- Montgomery, D.C., Peck, E.A. and Vining, G. (2001) “Introduction to Linear Regression Analysis”, Wiley: New York
  - not in library, but good technical analysis of VIFs & Eigenvalue analysis and other regression topics if you want to purchase a good book for reference.