Multicollinearity

- When we have more than one explanatory variable there is a possible problem.
- There may be two or more variables that explain the dependent variable well, but they may be closely correlated.
- This could mean that is difficult to distinguish the individual effects of both variables.
- The practical questions are when does it become a problem and what can we do?

Problem and identification

Consider perfect multicollinearity with

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

but where $X_2 = 2X_1$meaning that the correlation between the two variables is 1, so $r_{12}^2 = 1$

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 2X_1 + u \\ &= \beta_0 + (\beta_1 + 2\beta_2)X_1 + u \end{aligned}$$

Thus only $(\beta_1 + 2\beta_2)$ can be estimated. It is not possible to get separate estimates of $\beta_1$and β_2

In practise unlikely to find this extreme , but may get close to it. There has to be some multicollinearity, so the question is identifying when it is important and so when it is a problem.

- What will tend to get is high $R^2$and F test statistics, but low individual significance of the individual coefficients.
- Parameter estimates become very sensitive to the addition or deletion of observations. So can test by dropping observations and seeing what happens
- Predictions will be worse than those of a model with only a subset of variables
- Standard errors of the regression coefficients will be high. Though in fact this is not necessarily the result ofβ_1 multicollinearity alone. The variance of $\hat{\beta}_i$is

  $$\frac{\sigma^2}{\sum x_i^2(1 - r_{12}^2)}$$

  will be high if:

  $\sigma^2$is high
  $\sum x_i^2$is low
  $r_{12}^2$is high

  so just knowing r_{12}^2doesn’t really tell us if there is multicollinearity or not.
- Rules of thumb: Klein argued that it is not necessarily a problem unless the intercorrelation is high relative to the overall degree of multiple correlation i.e. $R_y^2 < R_i^2$, where R_i^2 is from a regression of explanatory variable i on all of the others.
- Are other measures but all of limited value in practise

Cures

- Get more data. Most obvious solution as could get rid of any spurious relation if increase information. Could also get over a problem that is of the limited range of data, but not going to help if the new observations still have multicollinearity. But what is important is not the number of observations but the informational content.

- Drop variables: this seems the most obvious solution and may work in some cases where not interested in individual parameter values. But there is a problem of omitted variable bias and have to consider the trade off between bias an variance -use relative mean squared error
- Simply present the OLS estimates and the variance-covariance matrix to allow an assessment of multicollinearity.
- Rather than the null restriction try some others across the variables.
- Use extraneous estimates: e.g. from cross section estimates
- Transform data: Using ratios or first differences will get rid of any multicollinearity caused by common trend. But see discussion of dynamic models.
- Principal components: a statistical devise that creates a linear combination of a number of variables and uses them. Also called latent variables.
- Ridge regression: available on some packages, but don't recommend as a general solution, especially in its simplest form.

So no easy solution to multicollinearity, but have to be aware of the problem when dropping seemingly insignificant variables from the regression. Consider:

	Xs important	Xs not important
Include Xs	✓	inefficiency
Exclude Xs	specification bias	✓