The standard technique for dealing with issues of endogeneity, in which some of the right hand side or explanatory variables are not independent because their value is determined by some of the other explanatory variables, is to estimate a series of equations simultaneously rather than to estimate just one equation.

For example, if you want to determine the effect of neighborhood watch programs on the crime rate in a neighborhood, you need to recognize that the existence of a neighborhood watch program is endogenous, so you also should simultaneously estimate a model to estimate the probability of a neighborhood having a neighborhood watch program.

For example, the number of chickens you have depends on the number of eggs laid, but the number of eggs laid also depends on the number of chickens you have. You need to simultaneously estimate equations for eggs and chickens, and each will have the other as an explanatory variable.

For example, if you want to estimate a recent graduate’s income as a function of, among other things, whether or not they were an economics major, you should simultaneously estimate a model of the probability that a person majors in economics.

Other tricky examples from the book include:
- population size and food supply
- determination of wages and prices
- exchange rates and international trade and capital flows

It is important to deal properly with endogeneity problems because when you have endogeneity, Classical Assumption III is violated in that the error terms will be correlated with the explanatory variables and, as a result, undue weight will be placed on the estimated coefficients of the explanatory variables. The result is that the estimated coefficients will be biased if the simultaneity is not explicitly recognized in the modeling.

If the standard regression equation is:

$$ Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t $$

then a simultaneous system in one in which the dependent variable has an impact on at least one of the explanatory variables.
If these are written out to express the underlying economic relationships that determine each variable determined within the system, these equations are called the structural equations.

\[
Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \varepsilon_{1t}
\]

\[
Y_{2t} = \beta_0 + \beta_1 Y_{1t} + \beta_2 X_{3t} + \beta_3 X_{2t} + \varepsilon_{2t}
\]

The coefficients estimated here are the structural coefficients.

In the equations shown here, because \(Y_{1t}\) depends on \(Y_{2t}\) and vice versa, there is said to be a feedback loop. This isn’t necessarily the case.

You might consider the example of neighborhood crime rates and the existence of block watch programs in which the crime rate is determined in the first equation and the probability of a block watch program existing is determined in the second equation and there is partial overlap of the other explanatory variables:

\[
Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \varepsilon_{1t}
\]

\[
Y_{2t} = \beta_0 + \beta_1 Y_{1t} + \beta_2 X_{3t} + \beta_3 X_{2t} + \varepsilon_{2t}
\]

To be clear about this, an exogenous or independent variable is one whose value is determined by factors totally outside of what is being estimated. The weather or sunspot activity are examples of exogenous variables.

An endogenous variable is one whose value is determined by the factors that are considered in the system.

Predetermined variables are all exogenous variables and lagged endogenous variables.

If you rewrite the equations so that the \(Y\)’s are all expressed as functions of only the \(X\)’s, these equations are referred to as the reduced form equations.

So, from the above examples:

Structural Equations:

\[
Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \varepsilon_{1t}
\]

\[
Y_{2t} = \beta_0 + \beta_1 Y_{1t} + \beta_2 X_{3t} + \beta_3 X_{2t} + \varepsilon_{2t}
\]

Reduced Form Equations:

\[
Y_{1t} = \pi_0 + \pi_1 X_{1t} + \pi_2 X_{2t} + \pi_3 X_{3t} + \nu_{1t}
\]

\[
Y_{2t} = \pi_4 + \pi_5 X_{1t} + \pi_6 X_{2t} + \pi_7 X_{3t} + \nu_{2t}
\]

Can you solve for the \(\pi\) terms?
Structural Equations:
\[ Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \varepsilon_{1t} \]
\[ Y_{2t} = \beta_0 + \beta_1 X_{2t} + \beta_2 X_{3t} + \varepsilon_{2t} \]

Reduced Form Equations:
\[ Y_{1t} = \pi_0 + \pi_1 X_{1t} + \pi_2 X_{2t} + \pi_3 X_{3t} + \nu_{1t} \]
\[ Y_{2t} = \pi_4 + \pi_5 X_{2t} + \pi_6 X_{3t} + \nu_{2t} \]

(Note that \( X_{1t} \) doesn’t appear in the equation for \( Y_{2t} \).)

But gosh, Allen, I liked the structural equations. They had stories behind them and the coefficients made sense and they weren’t all weird and scary like these reduced form equations are. Why in the heck would we want to use these reduced form equations?

1. Endogeneity problems are avoided with reduced form equations because there’s no inherent simultaneity.
2. Sometimes you can calculate the coefficients of the nice friendly structural equations by using the coefficient estimates from the reduced form equations.
3. The coefficients of the reduced form equations are impact multipliers, meaning that they express the overall change in the value of a dependent variable from a change in the value of an independent variable, allowing for such things as indirect effects and feedback loops.
4. Reduced form equations let you estimate two stage least squares (2SLS) models, which are fun and sexy.

An Example from the Book: Who doesn’t like a nice, refreshing carbonated beverage?
Imagine the following structural equations relating the quantity demanded ($Q_{Dt}$) and quantity supplied ($Q_{St}$) as functions of the price ($P_t$), dollars spent on advertising ($X_{1t}$), other demand side variables ($X_{2t}$) and some supply side variables ($X_{3t}$).

$$Q_{Dt} = \alpha_0 + \alpha_1 P_t + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \epsilon_{Dt}$$

$$Q_{St} = \beta_0 + \beta_1 P_t + \beta_2 X_{3t} + \epsilon_{St}$$

$$Q_{Dt} = Q_{St}$$

(There must be as many endogenous variables as there are equations.)

The endogenous variables here are $Q_{Dt}$, $Q_{St}$ and $P_t$.

The reduced form equations would be:

$$Q_t = \pi_0 + \pi_1 X_{1t} + \pi_2 X_{2t} + \pi_3 X_{3t} + v_{Qt}$$

$$P_t = \pi_4 + \pi_5 X_{1t} + \pi_6 X_{2t} + \pi_7 X_{3t} + v_{Pt}$$

**Two Stage Least Squares (2SLS)**

2SLS is a technique that is widely used to deal with issues of endogeneity.

The underlying idea is that we would like to replace the endogenous variable (the $Y_{2t}$ term) with a new variable that is a good proxy for $Y_{2t}$ and is also uncorrelated with the error term, which would be called an *instrumental variable*. An instrumental variable (casually known as an instrument) replaces an endogenous explanatory variable. The trick, of course, lies in finding a good instrument. This can be hard.

2SLS is a method of systematically creating an instrument.

2SLS has two stages:

**Stage 1**
Use OLS to do regressions of each of the reduced form equations for each of the endogenous explanatory variables and save the predicted values of each endogenous explanatory variable.

**Stage 2**
Use the predicted values from Stage 1 in the structural equations and then estimate these modified versions of the structural equations.

So, you start out with:
Structural Equations:

\[ Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_t + \varepsilon_{1t} \]
\[ Y_{2t} = \beta_0 + \beta_1 Y_{1t} + \beta_2 Z_t + \varepsilon_{2t} \]

Reduced Form Equations:

\[ Y_{1t} = \pi_0 + \pi_1 X_t + \pi_2 Z_t + \nu_{1t} \]
\[ Y_{2t} = \pi_3 + \pi_4 X_t + \pi_5 Z_t + \nu_{2t} \]

Estimate the reduced form equations to get

\[ \hat{Y}_{1t} = \hat{\pi}_0 + \hat{\pi}_1 X_t + \hat{\pi}_2 Z_t \]
\[ \hat{Y}_{2t} = \hat{\pi}_3 + \hat{\pi}_4 X_t + \hat{\pi}_5 Z_t \]

You then substitute these predicted values into the original structural equations and then use OLS to estimate them:

\[ Y_{1t} = \alpha_0 + \alpha_1 \hat{Y}_{2t} + \alpha_2 X_t + u_{1t} \]
\[ Y_{2t} = \beta_0 + \beta_1 \hat{Y}_{1t} + \beta_2 Z_t + u_{2t} \]

However, estimating these revised structural equations using OLS will result in the standard errors of the estimated coefficients being wrong, so you should use SPSS’s 2SLS procedure.

Properties of 2SLS:

1. Coefficient estimates are still biased in small samples, but as the sample size increases the expected value of \( \hat{\beta} \) approaches the actual value of \( \beta \).
2. The bias is opposite the bias in OLS.
3. If the reduced form equations have small R² values then 2SLS won’t work very well.
4. If the predetermined (or exogenous) variables are highly correlated then 2SLS won’t work very well.
5. t-tests are more accurate under 2SLS than under OLS.

A 2SLS Example From Studenmund

Consider the following model of the U.S. economy, which you may or may not be satisfied with. The structural equations are:

\[ Y_t = \text{CO}_t + \text{I}_t + \text{G}_t + \text{NX}_t \]
\[ \text{CO}_t = \beta_0 + \beta_1 \text{YD}_t + \beta_2 \text{CO}_{t-1} + \varepsilon_{1t} \]
\[ \text{YD}_t = Y_t - T_t \]

\[ I_t = \beta_3 + \beta_4 Y_t + \beta_5 r_{t-1} + \varepsilon_{2t} \]

\[ r_t = \beta_6 + \beta_7 Y_t + \beta_8 M_t + \varepsilon_{3t} \]

where

\[ Y = \text{GDP} \]
\[ \text{CO} = \text{consumption} \]
\[ I = \text{investment} \]
\[ G = \text{government spending} \]
\[ \text{NX} = \text{net exports} \]
\[ T = \text{taxes} \]
\[ r = \text{interest rate} \]
\[ M = \text{money supply} \]
\[ \text{YD} = \text{disposable income} \]

It is worth noting that only the second, fourth and fifth equations are stochastic equations (equations with an error term) that would need to be estimated, so only the explanatory variables from these equations might be endogenous. The other equations, the first and third, are accounting identities and don’t need to be estimated. Only the equations for \( \text{CO}_t, I_t \) and \( r_t \) can be estimated here.

Step 1: Figure out which explanatory variables are endogenous

O.K., for a variable to be endogenous it has to be not predetermined and it has to be one the left hand side of one of the equations and on the right hand side of a stochastic equation.

<table>
<thead>
<tr>
<th>LHS</th>
<th>RHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_t )</td>
<td>( \text{YD}_t )</td>
</tr>
<tr>
<td>( \text{CO}_t )</td>
<td>( \text{CO}_{t-1} ) predetermined</td>
</tr>
<tr>
<td>( \text{YD}_t )</td>
<td>( Y_t )</td>
</tr>
<tr>
<td>( I_t )</td>
<td>( r_{t-1} ) predetermined</td>
</tr>
<tr>
<td>( r_t )</td>
<td>( Y_t )</td>
</tr>
<tr>
<td></td>
<td>( M_t ) exogenous</td>
</tr>
</tbody>
</table>

So, the two endogenous explanatory variables are \( \text{YD}_t \) and \( Y_t \).

So, we need to estimate reduced form equations for \( \text{YD}_t \) and \( Y_t \).

\[ \text{YD}_t = \pi_0 + \pi_1 G_t + \pi_2 \text{NX}_t + \pi_3 \text{CO}_{t-1} + \pi_4 T_t + \pi_5 r_{t-1} + \pi_6 M_t + v_{\text{YD}t} \]

\[ Y_t = \pi_7 + \pi_8 G_t + \pi_9 \text{NX}_t + \pi_{10} \text{CO}_{t-1} + \pi_{11} T_t + \pi_{12} r_{t-1} + \pi_{13} M_t + v_{Yt} \]

Estimation of these reduced form equations will generate estimates of \( \text{YD}_t \) and \( Y_t \) which can then be used in the structural equations:
\[ CO_t = \beta_0 + \beta_1 Y_D t + \beta_2 CO_{t-1} + \varepsilon_{1t} \]
\[ I_t = \beta_3 + \beta_4 Y_t + \beta_5 r_{t-1} + \varepsilon_{2t} \]
\[ r_t = \beta_6 + \beta_7 Y_t + \beta_8 M_t + \varepsilon_{3t} \]

There are a few details about standard errors that need to be differently because this is 2SLS rather than OLS, but a software package will take care of these for you.

It is worth mentioning that the 2SLS results might not be very different from the OLS results, but it is still important to do them in order to conduct a proper analysis.

**The Identification Problem**

The identification problem is related to the different structural equations that you have and whether or not they can be properly estimated (or identified) based on the variables that are include and, more importantly, the variables that are excluded from that equation.

In fact, whether or not you are actually able to estimate one of your structural equations will depend on some of the exogenous or predetermined explanatory variables being absent from that structural equation.

If you’ve got several equations in a system, it might be the case that you can estimate equations for both of them or one of them or none of them. Put differently, it might be that both of them are identified, or that one of them is identified or that none of them is identified.

Here are examples in which X, V and Z are exogenous or predetermined:

**Both Identified**

\[ Y_{1t} = a_0 + a_1 V_t + \varepsilon_{1t} \]
\[ Y_{2t} = b_0 + b_1 Z_t + \varepsilon_{2t} \]

\[ Y_{1t} = a_0 + a_1 X_t + a_2 V_t + \varepsilon_{1t} \]
\[ Y_{2t} = b_0 + b_1 X_t + b_2 Z_t + \varepsilon_{2t} \]

Both are identified because Z isn’t in the first equation and V isn’t in the second.

**One Identified**

\[ Y_{1t} = a_0 + a_1 X_t + a_2 V_t + \varepsilon_{1t} \quad (identified \ because \ Z \ is \ absent) \]
\[ Y_{2t} = b_0 + b_1 X_t + b_2 V_t + b_3 Z_t + \varepsilon_{2t} \quad (not \ identified) \]

\[ Y_{1t} = a_0 + a_1 X_t + a_2 V_t + \alpha_3 Z_t + \varepsilon_{1t} \quad (not \ identified) \]
\[ Y_{2t} = b_0 + b_1 X_t + b_2 Z_t + \varepsilon_{2t} \quad (identified \ because \ V \ is \ absent) \]

**Neither Identified**
\[ Y_{1t} = \alpha_0 + \alpha_1 V_t + \varepsilon_{1t} \]
\[ Y_{2t} = \beta_0 + \beta_1 V_t + \varepsilon_{2t} \]

\[ Y_{1t} = \alpha_0 + \alpha_1 X_t + \alpha_2 V_t + \alpha_3 Z_t + \varepsilon_{1t} \]
\[ Y_{2t} = \beta_0 + \beta_1 X_t + \beta_2 V_t + \beta_3 Z_t + \varepsilon_{2t} \]

**The Order Condition for Identification**

The Order Condition is an easy way to determine whether a structural equation will be identified.

A necessary condition for a structural equation to be identified is:

The number of predetermined variables must be greater than the number of slope coefficients in the structural equation.

Put differently, this means that a predetermined variable must be missing from a structural equation in order for that structural equation to be identified.

Further…
1. If exactly one predetermined variable is missing, a structural equation is said to be *exactly identified*.
2. If more than one predetermined variable is missing, a structural equation is said to be *overidentified*. 