# Vector Autoregressions

- VAR: Vector AutoRegression
  - Nothing to do with VaR: Value at Risk (finance)
- Multivariate autoregression
- Multiple equation model for joint determination of two or more variables
- One of the most commonly used models for applied macroeconometric analysis and forecasting in central banks

# Two-Variable VAR

- Two variables: $y$ and x
- Example: output and interest rate
- Two-equation model for the two variables
- One-Step ahead model
- One equation for each variable
- Each equation is an autoregression plus distributed lag, with $p$ lags of each variable

# VAR(p) in 2 Variables

$$y_t = \mu_1 + \alpha_{11} y_{t-1} + \alpha_{12} y_{t-2} + \cdots + \alpha_{1p} y_{t-p}$$

$$+ \beta_{11} x_{t-1} + \beta_{12} x_{t-1} + \cdots + \beta_{1p} x_{t-p} + e_{1t}$$

$$x_t = \mu_2 + \alpha_{21} y_{t-1} + \alpha_{22} y_{t-2} + \cdots + \alpha_{2p} y_{t-p}$$

$$+ \beta_{21} x_{t-1} + \beta_{22} x_{t-1} + \cdots + \beta_{2p} x_{t-p} + e_{2t}$$

# Multiple Equation System

- In general: k variables

- An equation for each variable

- Each equation includes $p$ lags of $y$ and $p$ lags of $x$

- (In principle, the equations could have different # of lags, and different # of lags of each variable, but this is most common specification.)

- There is one error per equation.
  - The errors are (typically) correlated.

# Unrestricted VAR

- An unrestricted VAR includes all variables in each equation
- A restricted VAR might include some variables in one equation, other variables in another equation
- Old-fashioned macroeconomic models (so-called simultaneous equations models of the 1950s, 1960s, 1970s) were essentially restricted VARs
  - The restrictions and specifications were derived from simplistic macro theory, e.g. Keynesian consumption functions, investment equations, etc.

# VAR Revolution

- Christopher Sims (1942-) of Princeton University
  - 2011 Nobel Prize in Economics
- "Macroeconomics and Reality" (1980)
  - Sims argued that conventional macro models were "incredible" – they were based on ~~non-credible identifying assumptions~~
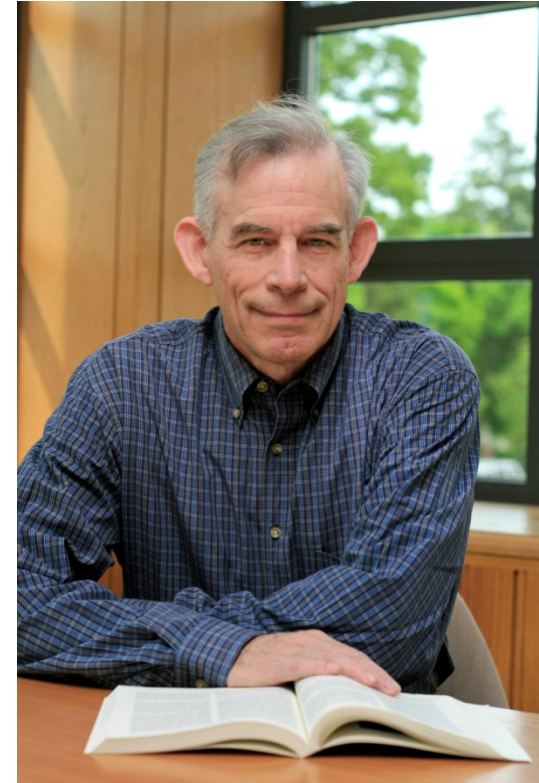
# Sims and VARs

- Sims argued that the conventional models were ~~restricted VARs,~~ and the ~~restrictions had no substantive justification~~
  - ~~Based on incomplete and/or non-rigorous theory, or intuition~~
- Sims argued that economists should instead <mark>use unrestricted models, e.g. VARs</mark>
- He proposed a set of tools for use and evaluation of VARs in practice.

# Estimation

- Each equation estimated by OLS

$$y_t = \mu_1 + \alpha_{11} y_{t-1} + \alpha_{12} y_{t-2} + \cdots + \alpha_{1p} y_{t-p}$$
$$+ \beta_{11} x_{t-1} + \beta_{12} x_{t-1} + \cdots + \beta_{1p} x_{t-p} + e_{1t}$$

$$x_t = \mu_2 + \alpha_{21} y_{t-1} + \alpha_{22} y_{t-2} + \cdots + \alpha_{2p} y_{t-p}$$
$$+ \beta_{21} x_{t-1} + \beta_{22} x_{t-1} + \cdots + \beta_{2p} x_{t-p} + e_{2t}$$

# Estimation in Stata

- To estimate a VAR in the variables *y* & *x* with lags 1 through p included
  - .varbasic y x, lags(1/p)
- For example, using gdp2013.dta and variables gdp and d.t12 with 3 lags
  - .gen rate=d.t12
  - .varbasic rate gdp, lags(1/3)
- Could also use
  - .var rate gdp, lags(1/3)

# Example: GDP and Interest Rate

```
. varbasic rate gdp

Vector autoregression

Sample:  1954q1 - 2013q4                         No. of obs      =        240
Log likelihood = -896.4798                       AIC             =   7.553999
FPE            =   6.542089                       HQIC            =   7.612434
Det(Sigma_ml)  =   6.018939                       SBIC            =   7.699025
```

| Equation | Parms | RMSE | R-sq | chi2 | P>chi2 |
|----------|-------|------|------|------|--------|
| rate | 5 | .752146 | 0.1282 | 35.29972 | 0.0000 |
| gdp | 5 | 3.35843 | 0.1763 | 51.35579 | 0.0000 |

# Example: GDP and Interest Rate

|  | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. Interval] | |
|---|---|---|---|---|---|---|
| **rate** | | | | | | |
| rate | | | | | | |
| L1. | -.0687148 | .0650433 | -1.06 | 0.291 | -.1961972 | .0587677 |
| L2. | .0106681 | .0624896 | 0.17 | 0.864 | -.1118092 | .1331454 |
| gdp | | | | | | |
| L1. | .0779384 | .0140999 | 5.53 | 0.000 | .0503032 | .1055737 |
| L2. | .0037657 | .0150392 | 0.25 | 0.802 | -.0257106 | .0332421 |
| _cons | -.2639241 | .0718976 | -3.67 | 0.000 | -.4048407 | -.1230075 |
| **gdp** | | | | | | |
| rate | | | | | | |
| L1. | -.6844941 | .2904269 | -2.36 | 0.018 | -1.25372 | -.1152678 |
| L2. | -.7395324 | .2790243 | -2.65 | 0.008 | -1.28641 | -.1926549 |
| gdp | | | | | | |
| L1. | .3160889 | .0629578 | 5.02 | 0.000 | .192694 | .4394839 |
| L2. | .190147 | .0671521 | 2.83 | 0.005 | .0585313 | .3217626 |
| _cons | 1.579774 | .3210322 | 4.92 | 0.000 | .950562 | 2.208985 |

# Order Selection

- A VAR(p) includes $p$ lags of each variable in each equation
- In a two-variable system, the number of coefficients in each equation is $1+2p$
  - The total number is $2(1+2p)=2+4p$
- In a $k$-variable system, the number of coefficients in each equation is $1+kp$
  - The total number is $k(1+2p)=k+2kp$
- How should $p$ be selected?
- Common approach:
  - Information criterion, primarily AIC

# AIC and BIC for VAR Models

$$AIC = -2L + 2\left(k + 2kp\right)$$

$$BIC = -2L + \left(k + 2kp\right)\ln\left(T\right)$$

where *L* is log-likelihood from model
- Select model with smallest AIC (or BIC)

# Stata Implementation

- varsoc command
- To calculate information criterion for a VAR in variables *x* and *y* up to a maximum lag of *pmax*:
  - .varsoc x y, maxlag(pmax)
- Produces a convenient table

# Example: GDP and Interest Rate

```
. varsoc rate gdp, maxlag(8)

   Selection-order criteria
   Sample:  1955q3 - 2013q4                    Number of obs      =       234
```

| lag | LL | LR | df | p | FPE | AIC | HQIC | SBIC |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0 | -908.041 | | | | 8.18559 | 7.77813 | 7.79004 | 7.80766 |
| 1 | -882.051 | 51.979 | 4 | 0.000 | 6.78308 | 7.59018 | 7.62591* | 7.67878* |
| 2 | -876.011 | 12.081 | 4 | 0.017 | 6.66587* | 7.57274* | 7.63228 | 7.7204 |
| 3 | -874.997 | 2.0279 | 4 | 0.731 | 6.83834 | 7.59826 | 7.68162 | 7.80499 |
| 4 | -869.617 | 10.76* | 4 | 0.029 | 6.75844 | 7.58647 | 7.69364 | 7.85226 |
| 5 | -867.435 | 4.3637 | 4 | 0.359 | 6.86471 | 7.60201 | 7.73299 | 7.92687 |
| 6 | -865.116 | 4.6376 | 4 | 0.327 | 6.9647 | 7.61638 | 7.77118 | 8.0003 |
| 7 | -861.706 | 6.8213 | 4 | 0.146 | 7.00073 | 7.62142 | 7.80003 | 8.06441 |
| 8 | -861.084 | 1.2442 | 4 | 0.871 | 7.20695 | 7.65029 | 7.85272 | 8.15234 |

```
   Endogenous:  rate gdp
    Exogenous:  _cons
```

# Result

- For this example
  - AIC selects *p=3*
  - BIC selects *p=2*
- Notice that the AIC value for p=3 in this table (AIC=7.572) is different from that obtained when we estimated the VAR(3) model (AIC=7.553).
  - This is because for the AIC comparison, all estimates are from a common sample, in this case excluding the first 8 observations since the maximum order is set to 8
- The varsoc command is correct

Let's look at the VAR(3) estimates again.

# Example: GDP and Interest Rate

|  | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. Interval] | |
|---|---|---|---|---|---|---|
| **rate** | | | | | | |
| rate | | | | | | |
| L1. | -.0687148 | .0650433 | -1.06 | 0.291 | -.1961972 | .0587677 |
| L2. | .0106681 | .0624896 | 0.17 | 0.864 | -.1118092 | .1331454 |
| gdp | | | | | | |
| L1. | .0779384 | .0140999 | 5.53 | 0.000 | .0503032 | .1055737 |
| L2. | .0037657 | .0150392 | 0.25 | 0.802 | -.0257106 | .0332421 |
| _cons | -.2639241 | .0718976 | -3.67 | 0.000 | -.4048407 | -.1230075 |
| **gdp** | | | | | | |
| rate | | | | | | |
| L1. | -.6844941 | .2904269 | -2.36 | 0.018 | -1.25372 | -.1152678 |
| L2. | -.7395324 | .2790243 | -2.65 | 0.008 | -1.28641 | -.1926549 |
| gdp | | | | | | |
| L1. | .3160889 | .0629578 | 5.02 | 0.000 | .192694 | .4394839 |
| L2. | .190147 | .0671521 | 2.83 | 0.005 | .0585313 | .3217626 |
| _cons | 1.579774 | .3210322 | 4.92 | 0.000 | .950562 | 2.208985 |

# Interpretation

- It is difficult to interpret the large number of coefficients in the VAR model
- Main tools for interpretation
  - Impulse responses

# Impulse Response Analysis

- VAR(1) with no intercept

$$y_t = \alpha_{11} y_{t-1} + \beta_{11} x_{t-1} + e_{1t}$$

$$x_t = \alpha_{21} y_{t-1} + \beta_{21} x_{t-1} + e_{2t}$$

- The impulse responses are the time-paths of to y and x in response to shocks

# Impulse Response Analysis

- The errors may be correlated.
- <mark>We "orthogonalize" them</mark>

$$e_{1t} = u_{1t}$$

$$e_{2t} = \rho e_{1t} + u_{2t}$$

$$= \rho u_{1t} + u_{2t}$$

# Orthogonalized Model

$$y_t = \alpha_{11} y_{t-1} + \beta_{11} x_{t-1} + u_{1t}$$

$$x_t = \alpha_{21} y_{t-1} + \beta_{21} x_{t-1} + \rho u_{1t} + u_{2t}$$

- The <mark>shocks $u_1$ and $u_2$ are uncorrelated</mark>
- The ordering matters
  - The shock to y affects both y and x in period t
  - The shock to x affects only x in period t
- The impulse responses are the time-paths of to y and x in response to the shocks $u_1$ and $u_2$
- Imagine y=0 and x=0. Set $u_1$=1. Trace the history of y and x

# Impulse Responses by Recursion

$$y_1 = \alpha_{11}0 + \beta_{11}0 + 1 = 1$$

$$x_1 = \alpha_{21}0 + \beta_{21}0 + \rho 1 = \rho$$

$$y_2 = \alpha_{11}y_1 + \beta_{11}x_1 = \alpha_{11} + \beta_{11}$$

$$x_2 = \alpha_{21}y_1 + \beta_{21}x_1 = \alpha_{21} + \beta_{21}\rho$$

$$y_3 = \alpha_{11}y_2 + \beta_{11}x_2 = \alpha_{11}(\alpha_{11} + \beta_{11}) + \beta_{11}(\alpha_{21} + \beta_{21}\rho)$$

$$x_3 = \alpha_{21}y_2 + \beta_{21}x_2 = \alpha_{21}(\alpha_{11} + \beta_{11}) + \beta_{21}(\alpha_{21} + \beta_{21}\rho)$$

# Impulse Responses

- The impulse responses are these time-paths of y and x due to the shocks $u_1$ and $u_2$

- They are found by this recursion formula

- They are functions of the estimated VAR coefficients

# Impact of Shocks on Variables

- In a 2-variable system, <mark>there are 4 impulse</mark> response functions
  - The effect on y of a shock to y ($u_1$)
  - The effect on y of a shock to x ($u_2$)
  - The effect on x of a shock to y ($u_1$)
  - The effect on x of a shock to x ($u_2$)
- In a k-variable system, there are $k^2$ impulse response functions!

# Stata Calculation

- Impulse response automatically calculated with **varbasic** command
- A kxk matrix of impulse response is created

# GDP/Interest Rate Example



Graphs by irfname, impulse variable, and response variable
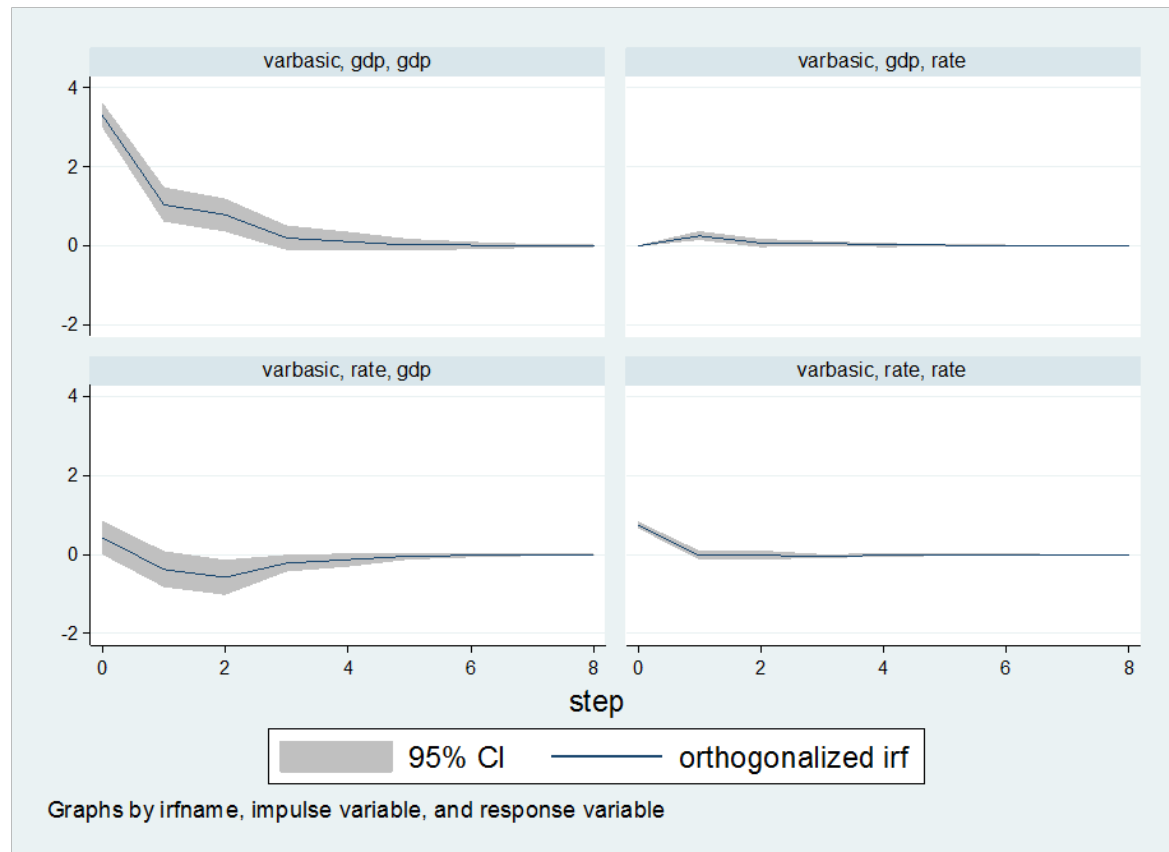
# Interpretation

- Labeled "Graphs by irfname, impulse variable, and response variable"
  - "Impulse variable" means the source of the shock
  - "Response variable" means the variable being affected
- Upper left: "varbasic, gdp, gdp"
  - Impact of a gdp shock on the time-path of gdp
- Upper right: "varbasic, gdp, rate"
  - Impact of a gdp shock on the time-path of interest rates
- Lower left: "varbasic, rate, gdp,"
  - Impact of an interest rate shock on the time-path of gdp
- Lower fight: "varbasic, rate, rate"
  - Impact of an interest rate shock on the time-path of interest rates
- The impulse response is graphed as a function of forward time periods

# Scale

- The graphs are all created on the same scale, so difficult to read

- It may be better to create graphs separate for each impulse response

```
. irf graph oirf, impulse(gdp) response(rate)
```

- This creates the impulse response for the impact of a gdp shock on the time-path of interest rates
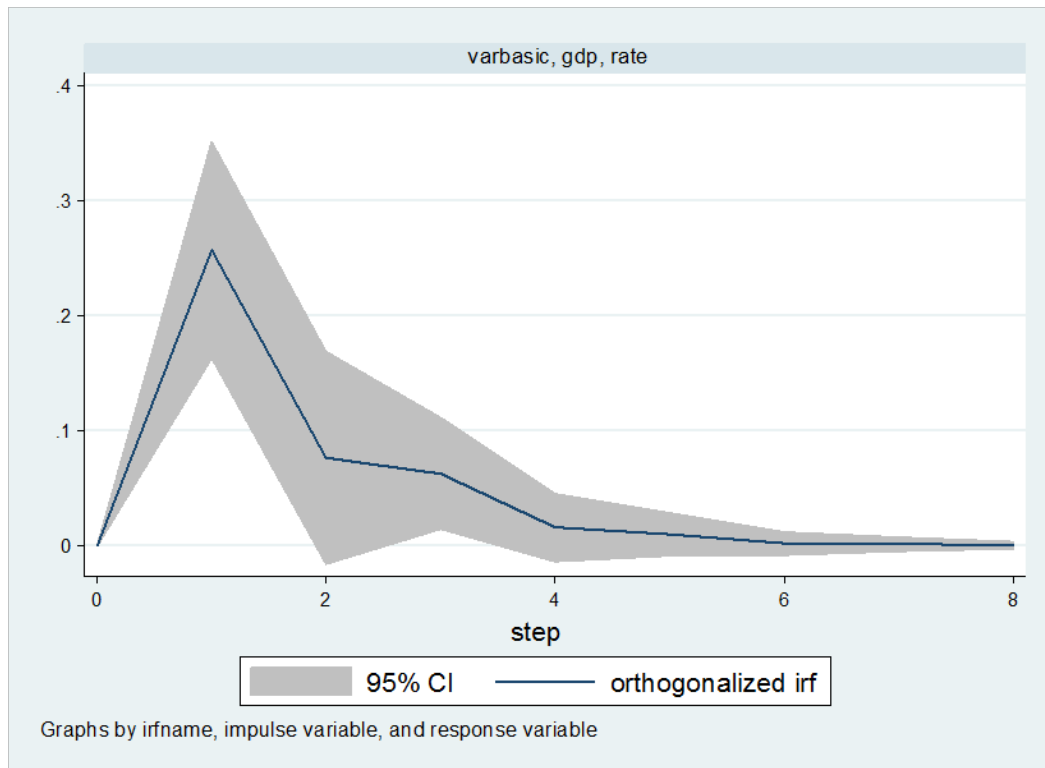
# GDP on GDP
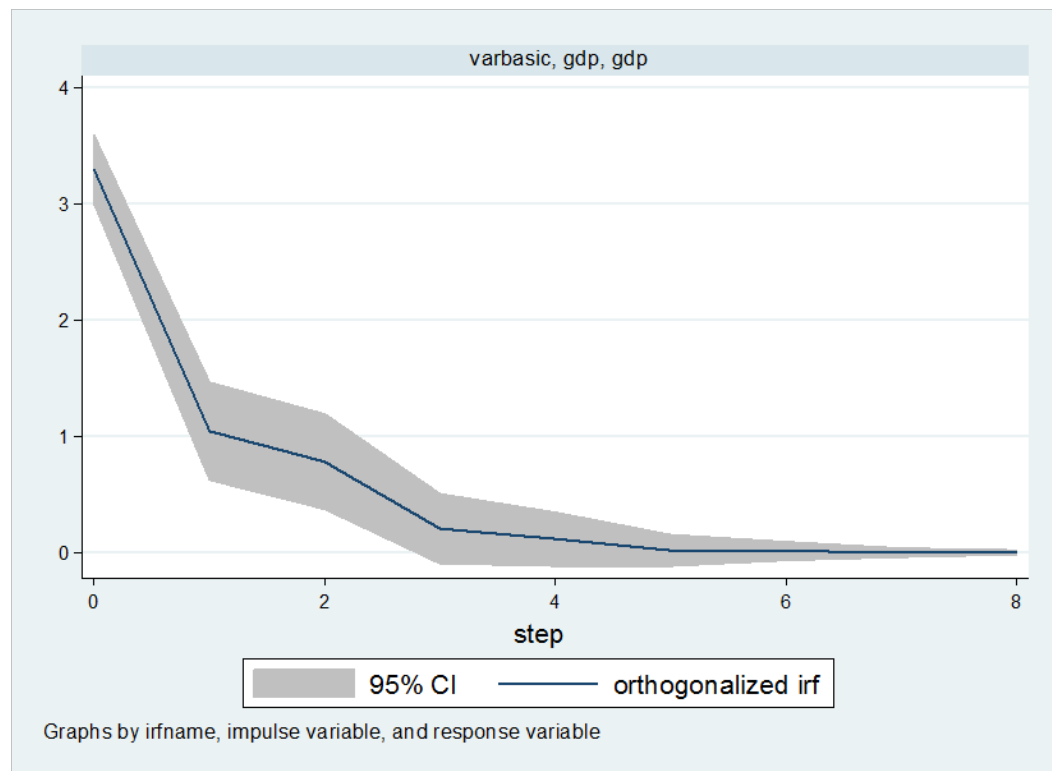
# GDP on Interest Rates

`. irf graph oirf, impulse(gdp) response(rate)`

# Interest Rates on Interest Rates

# Interest Rates on GDP

```
. irf graph oirf, impulse(rate) response(gdp)
```



Graphs by irfname, impulse variable, and response variable
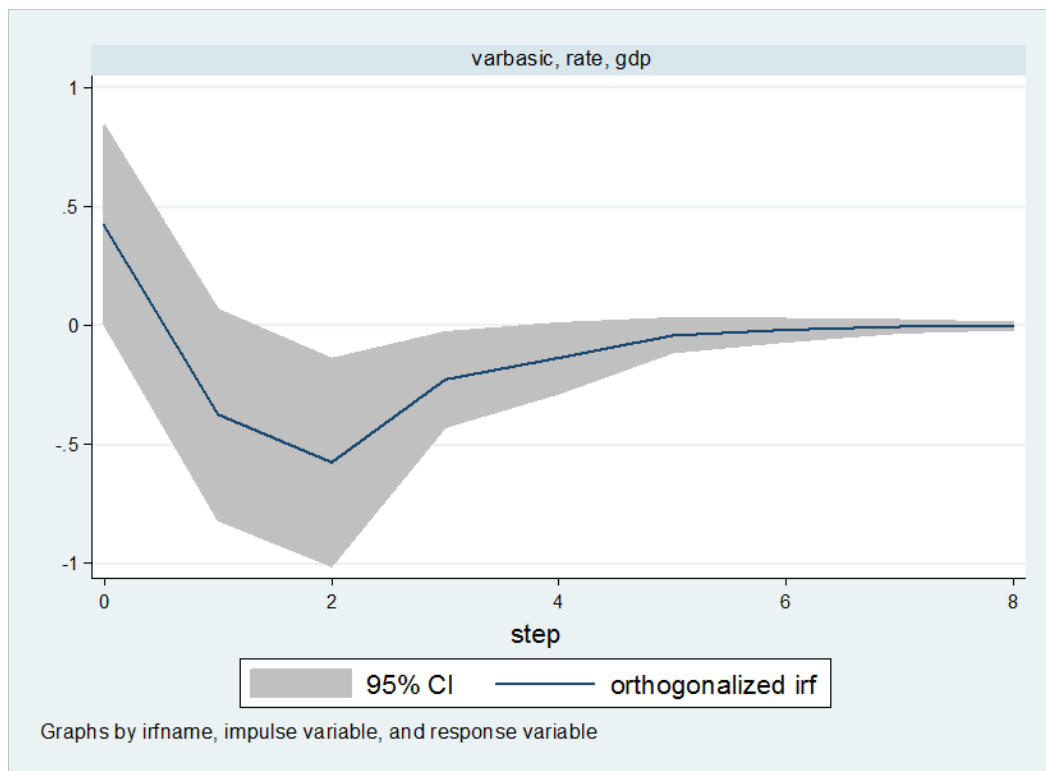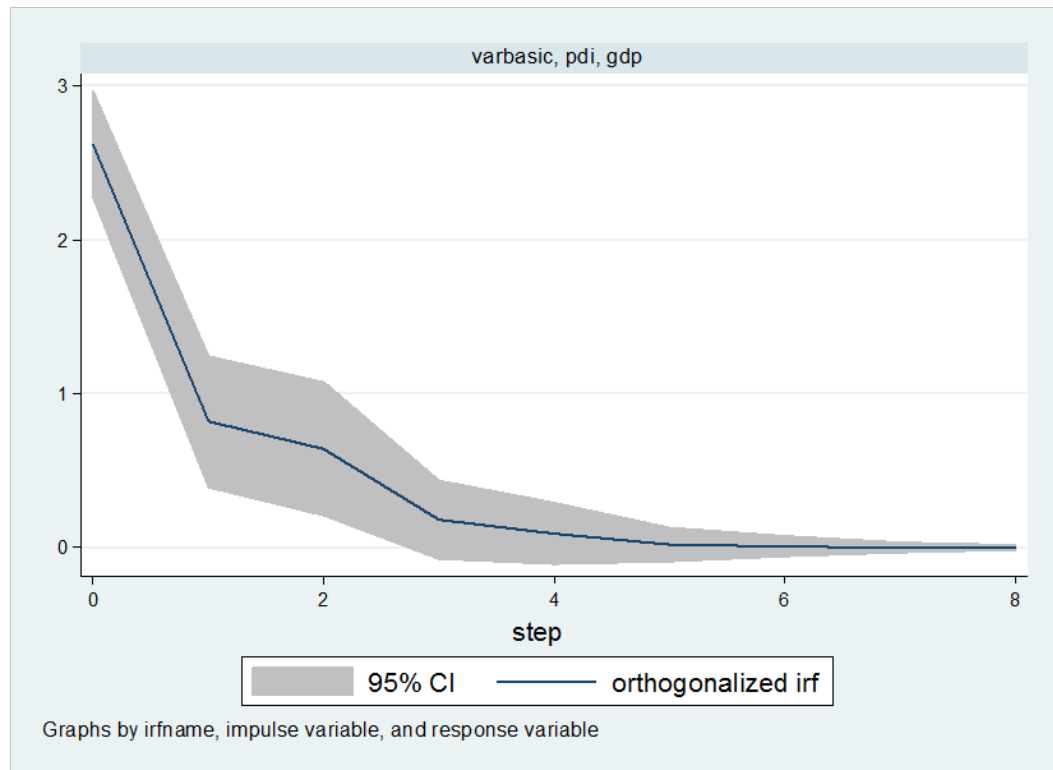
# 3-variable system

- Interest Rate Change (12-month T-Bill)
- Investment Growth Rate
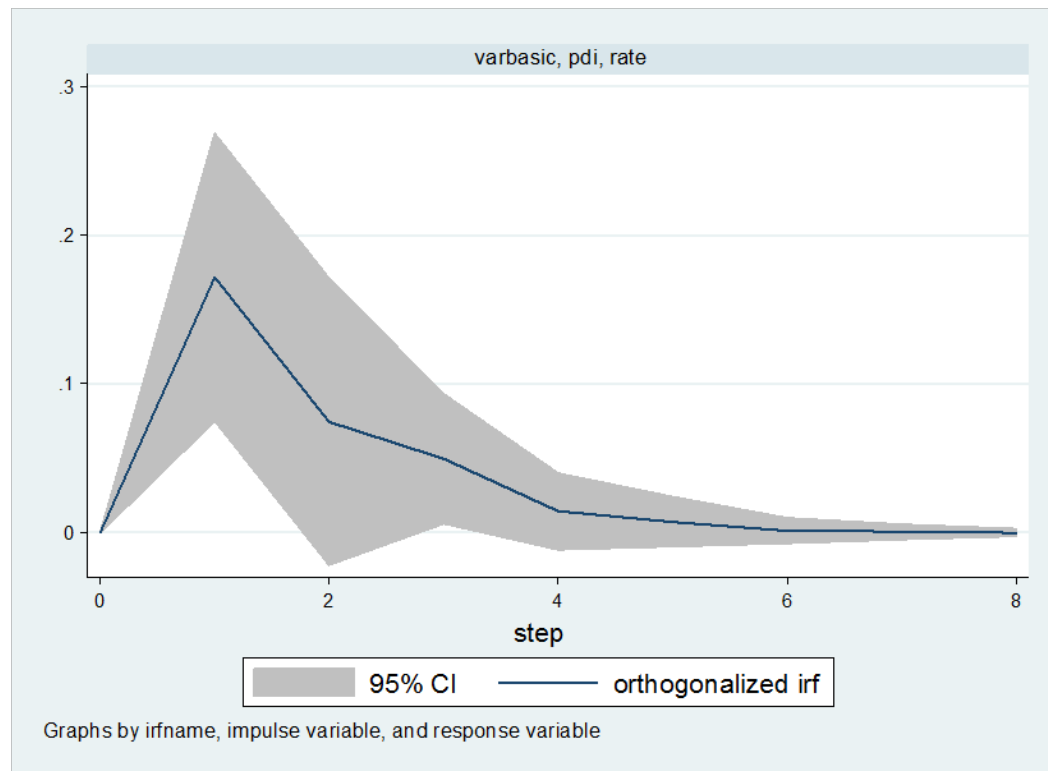- GDP Growth Rate

# GDP/Investment/Interest Rate
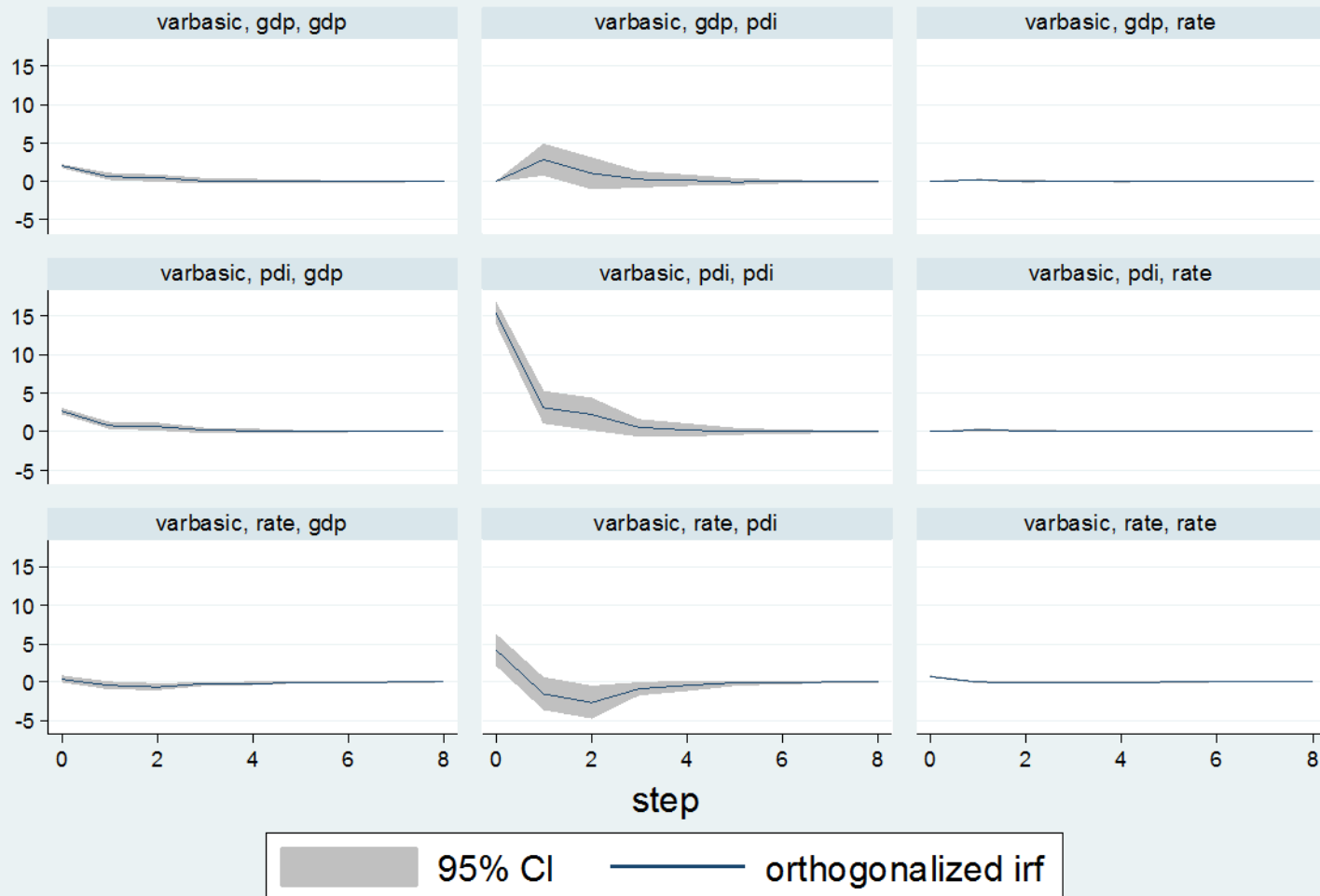


Graphs by irfname, impulse variable, and response variable
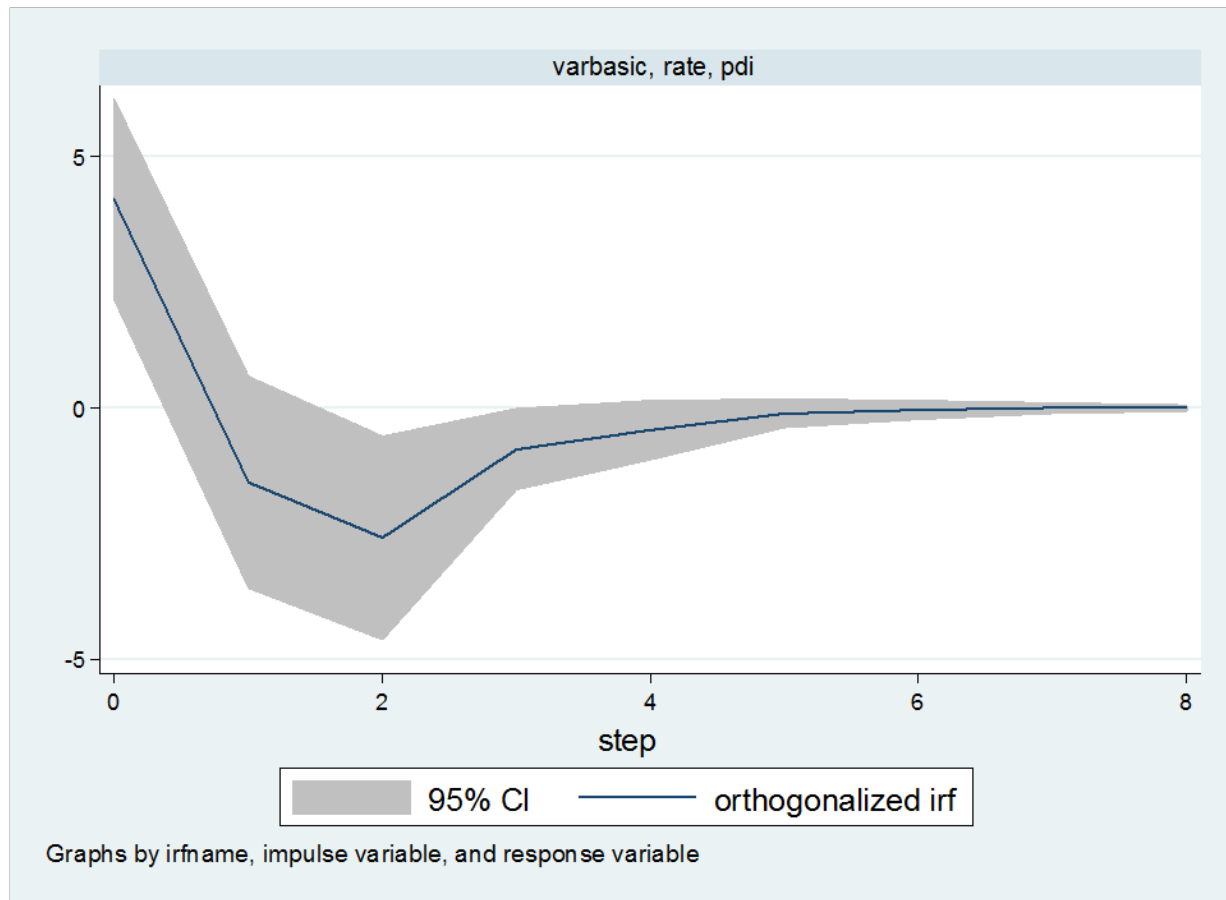
# Investment Shock on GDP



varbasic, pdi, gdp

Graphs by irfname, impulse variable, and response variable

# Investment Shock on Interest Rate



varbasic, pdi, rate

Graphs by irfname, impulse variable, and response variable

# Interest Rate Shock on Investment



Graphs by irfname, impulse variable, and response variable

# High Dimensional Estimation

- What if you have a situation where the number of regressions $p$ exceeds the number of observations $n$ ?

- Classic example: gene array data
  - Goal: Determine which gene causes cancer
  - Number of regressors $p$ = number of genes (5000)
  - Number of observations $p$ = 50 (or similar)

# LASSO

- One solution is LASSO estimation
- Similar idea: LARS, SCAD, Elastic Net
- Idea: Minimize the sum-of-squared errors subject to a penalty based on the sum of the absolute value of the coefficients

$$x_t = \mu_2 + \alpha_{21} y_{t-1} + \alpha_{22} y_{t-2} + \cdots + \alpha_{2p} y_{t-p}$$

$$+ \beta_{21} x_{t-1} + \beta_{22} x_{t-1} + \cdots + \beta_{2p} x_{t-p} + e_{2t}$$

# LASSO

Model

$$y_t = \mu + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} + e_t$$

Minimize sum-of-squared errors plus penalty

$$\sum_{t=1}^{T} \left( y_t - \mu + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} \right)^2$$

$$+ \lambda \sum_{j=1}^{p} \left| \beta_j \right|$$

The penalty changes the problem.

Most coefficient estimates are zero.

# LASSO and Forecasting

- Lasso very popular in high-dimensional statistics
- I haven't yet seen Lasso being discussed in economic forecasting
- It is just a matter of time
- Not programmed in Stata
- If interested, I recommend the R package

# Software after UW??

- You are unlikely to have access to Stata outside a university environment
  - Some corporations may have a few licenses
  - Non-academic price is expensive
- Excel widely available
  - Often used for regression analysis in corporations
  - Highly limited & clumsy
- R is a viable option
  - Free, open-source
  - Continuously updated
  - Popular among statisticians
  - http://www.r-project.org/
  - A different style; may need to do more programming
  - Documentation sometimes limited