Multi-collinearity of MEVs in risk modelling: using divergence weighted independence graphs

April 2008

joe.whittaker@lancaster.ac.uk

Provide:

- Sensible math notation for multi-collinearity: vif, R^2, information.
- Graph to highlight dependences, and software.
- Application to assess effect of multi-collinearity macro-economic variables on consumer risk.

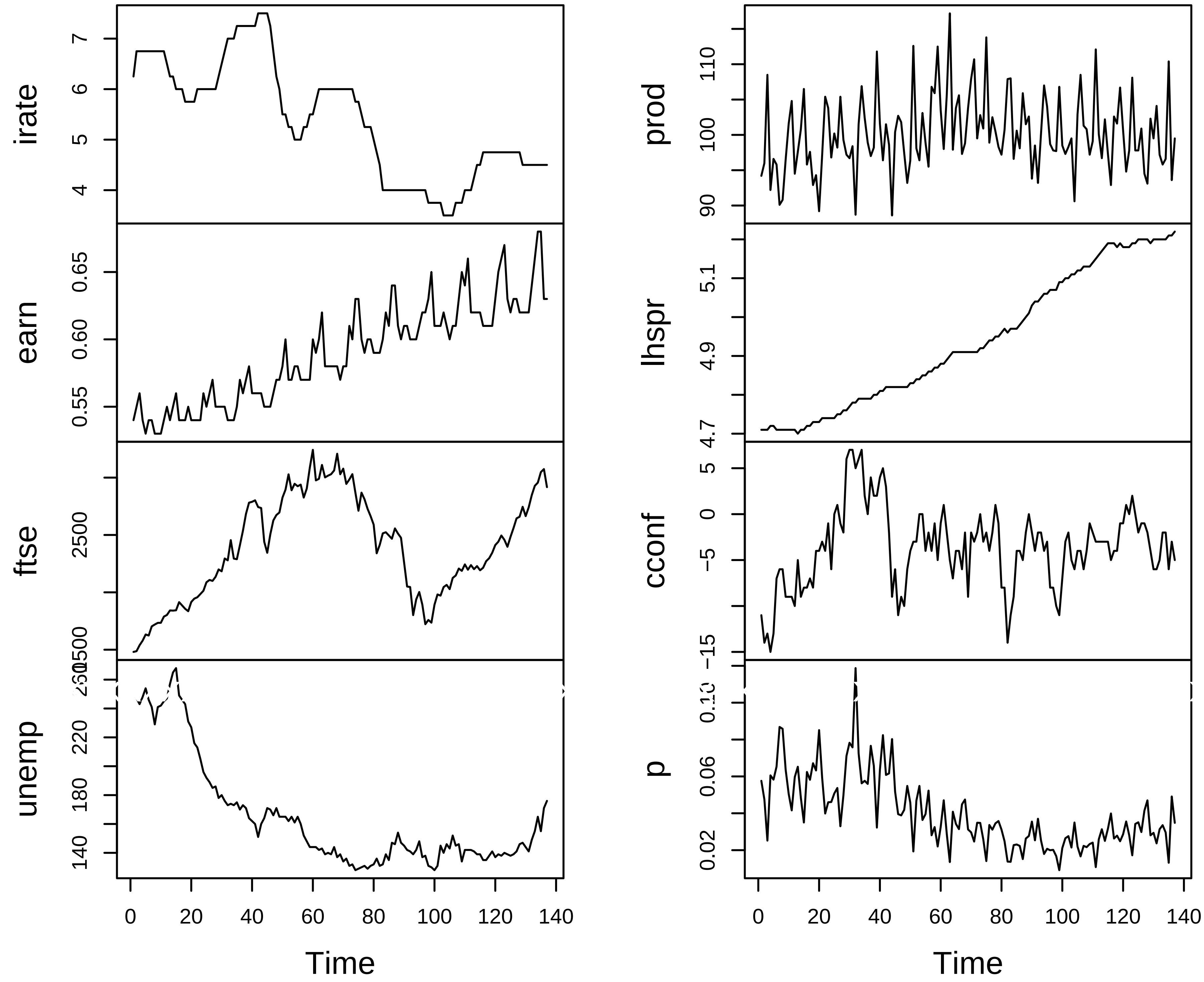
Application:

Bellotti and Crook (2007) 'Credit Scoring with Macroeconomic Variables using Survival Analysis'

Period: 1997 to mid-2005, $n = 137$ months

Response: based on over 200k credit card accts.

- 7 MEVs in X:
  - Interest rate,
  - Earnings ratio of UK earnings to rpi,
  - FTSE all-share index,
  - Unemployment index [males 6-12 mnths, sa],
  - Index of all UK production,
  - House price index (Nationwide), logged,
  - UK consumer confidence index.



irate
earn
ftse
unemp
Time
prod
lhspr
cconf
p
Time

Remarks

Source: ONS, etal.

Not seasonally adjusted, apart from unemp.

- p fitted default probability (BC07)

7 coeffs:

$\widehat{\boldsymbol{\beta}} = (0.133,\ -7.08,\ 0.000216,\ 0.0044,\ -0.0491,\ 0.821,\ 0.0331)'$

linear predictor:

$$\eta = X\widehat{\boldsymbol{\beta}} \quad \text{and} \quad p = \frac{\exp(-.1+\eta)}{1+\exp(-.1+\eta)}$$

Better to have observed defaults as response.

Our aim: indicate how difficult it is to estimate $\widehat{\boldsymbol{\beta}}$ from these MEVs, convention vif > 10?

Multi-collinearity in regression

Causes var($\widehat{\beta}_j$) to be large.

Inference about β_j becomes imprecise:

- lack of statistical significance,
- incorrect estimates of 'importance',
- obscures relation marginal and partial coeffs.

web: 'Multicollinearity in logistic regression models is a result of strong correlations between independent variables.'

Need for precise mathematical description:

$Y \perp\!\!\!\perp X$ statistical independence of rvs Y and X,

Y response; $\boldsymbol{X}$ covariates, explanatory.

LLSP notation

Random variables and vectors

Y,

$\boldsymbol{X} = (X_1, X_2, \ldots, X_7)'$, $\quad X_j$ generic,

$\boldsymbol{X}_{\setminus j}$, the $_{\setminus j}$ indicate the remaining vbles in set,

$\boldsymbol{X}_a$ subvector where $a \subseteq \{1, \ldots, 7\}$.

Expected values, covariance, variance

$\mathsf{E}(Y)$, $\mathsf{E}(\boldsymbol{X})$,

$\mathsf{var}(Y)$, $\mathsf{cov}(Y, \boldsymbol{X})$, $\mathsf{var}(\boldsymbol{X})$.

Suppresses mention of the $n = 137$ data points.

Fragment of uncentered X data matrix:

	irate	earn	ftse	unemp	prod	lhspr	cconf
...							
16	6.8	0.54	1579	248	92	4.72	-15
17	6.8	0.53	1633	254	97	4.72	-13
18	6.8	0.54	1624	246	96	4.71	-7
...							
avg	5.4	0.59	2395	165	100	4.93	-3.6

var-cov	irate	earn	ftse	unemp	prod	lhspr	cconf
irate	1.31622	-0.032338	13.52	...			
earn	-0.03234	0.001356	5.19				
ftse	13.52450	5.190200	203462.08				
...				$\ddots$			

Expectation take $\mathsf{E}(\boldsymbol{X}) = 1'X/n$, average.

Center and take $\mathsf{var}(\boldsymbol{X}) = (X'X)/n$.

Definitions

Predicted (fitted) value

- $\widehat{Y} = \mathsf{E}(Y) + \mathsf{cov}(Y, \boldsymbol{X})\,\mathsf{var}(\boldsymbol{X})^{-1}[\boldsymbol{X} - \mathsf{E}(\boldsymbol{X})]$

Multiple correlation coefficient

- $R(Y|\boldsymbol{X}) = \mathsf{corr}(Y, \widehat{Y})$.

Residual variance

- $\mathsf{var}(Y|\boldsymbol{X}) = \mathsf{var}(Y - \widehat{Y})$.

Application to OLS regression

Variance of regression coefficients

$$\begin{aligned} \text{var}(\widehat{\boldsymbol{\beta}}) &= \sigma^2 (X'X)^{-1} \quad \text{from OLS theory} \\ &= \frac{\sigma^2}{n} \text{var}(\boldsymbol{X})^{-1}. \end{aligned}$$

Multi-collinearity is due to nature of $\text{var}(\boldsymbol{X})^{-1}$.

Inverse variance: partition $\boldsymbol{X} = (\boldsymbol{X}_a, \boldsymbol{X}_b)$.

- Invert

$$\text{var}(\boldsymbol{X})^{-1} = \text{var}\left(\begin{bmatrix} \boldsymbol{X}_a \\ \boldsymbol{X}_b \end{bmatrix}\right)^{-1} = \begin{bmatrix} \ddots & \cdots \\ \vdots & \text{var}(\boldsymbol{X}_b | \boldsymbol{X}_a)^{-1} \end{bmatrix}.$$

Multi-collinearity

Applying this inverse variance result

$$\begin{aligned} \text{var}(\widehat{\beta}_j) &= \frac{\sigma^2}{n}\,\text{var}(X_j|\boldsymbol{X}_{\setminus j})^{-1} \\ &= \frac{\sigma^2}{n}\,\text{var}(X_j)^{-1}\frac{1}{1-R(X_j|\boldsymbol{X}_{\setminus j})^2}. \end{aligned}$$

Definition

variance inflation factor $\text{vif} = \dfrac{1}{1-R(X_j|\boldsymbol{X}_{\setminus j})^2}$.

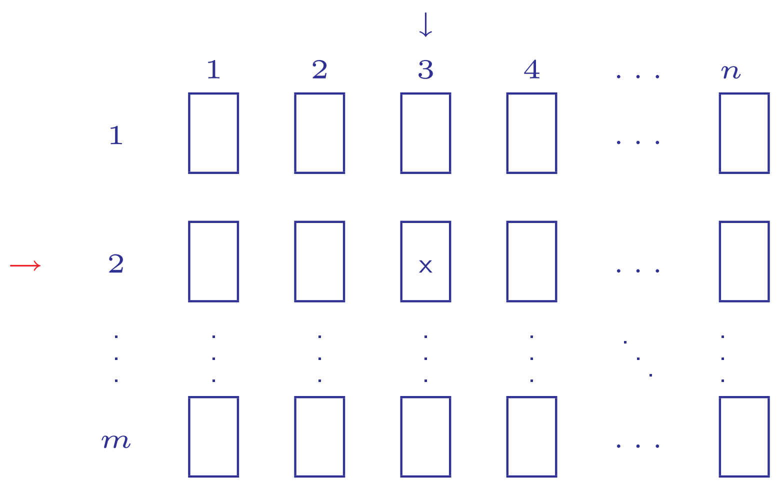
Multi-collinearity bad when

$$\text{var}(X_j|\boldsymbol{X}_{\setminus j})\rightarrow 0 \quad \text{or} \quad \text{vif}\rightarrow\infty.$$

Hartley Information, 1928

Information gained by learning outcome of experiment with 2 equi-probable outcomes □ ☒ is $H(2) = 1$ bit (=1000mbits).

Additivity: $H(n \times m) = H(n) + H(m)$, because if arrange outcomes in rows and columns, to know row and column is sufficient.



$$\Rightarrow H(n) = \log_2 n = -\log_2(\tfrac{1}{n}).$$

Information measures

Hartley, Shannon, Kullback, . . .

If $(Y, \boldsymbol{X})$ jointly normal distributed

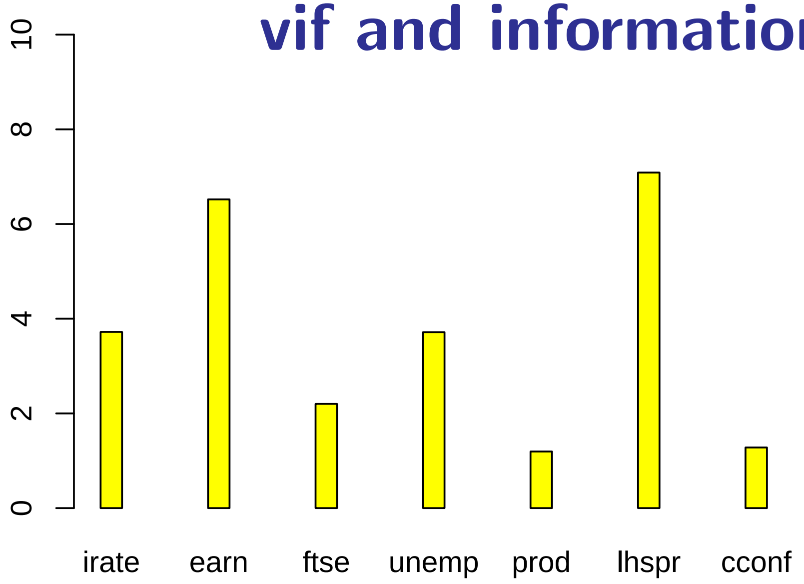
$$\begin{aligned}
\mathsf{Inf}(Y \perp\!\!\!\perp \boldsymbol{X}) &= -\frac{1}{2}\log\left(1 - R(Y|\boldsymbol{X})^2\right), \\
\mathsf{Inf}(X_j \perp\!\!\!\perp \boldsymbol{X}_{\setminus j}) &= -\frac{1}{2}\log\left(1 - R(X_j|\boldsymbol{X}_{\setminus j})^2\right), \\
\mathsf{Inf}(X_i \perp\!\!\!\perp X_j|\boldsymbol{X}_{\setminus ij}) &= -\frac{1}{2}\log\left(1 - \mathsf{corr}(X_i, X_j|\boldsymbol{X}_{\setminus ij})^2\right).
\end{aligned}$$

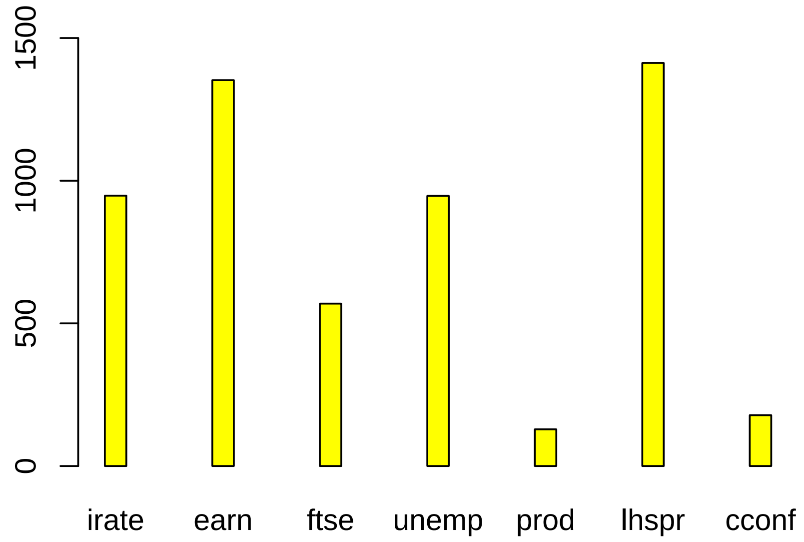
$\cdot \perp\!\!\!\perp \cdot | \cdot$ denotes conditional independence.

Using Inf puts correlation onto an additive scale.

- Define $\mathsf{Inf}_{\mathsf{vif}} = \mathsf{Inf}(X_j \perp\!\!\!\perp \boldsymbol{X}_{\setminus j})$

vif and information





vif $\qquad\qquad$ $\text{Inf}_{\text{vif}} = \text{Inf}(X_j \perp\!\!\!\perp \boldsymbol{X}_{\setminus j})$

Variance inflation: $6 \times$ marginal,

theoretical limit vif > 1, Inf > 0

substantial dependence, except prod, ccon.

Wider variation on the Inf scale,

Inf_{vif} scale meaningful: millibits.

Divergence weighted independence graph

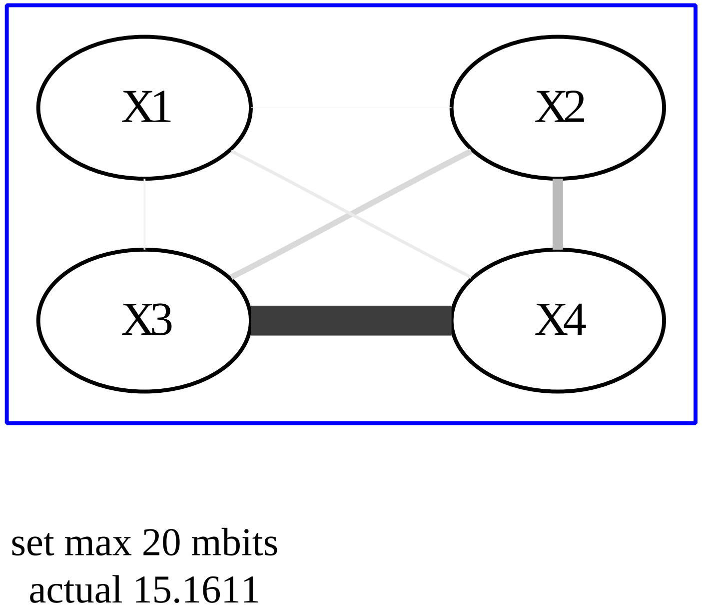
Consider a graph with vertices = variables, edges between all vertices, and weights on each edge.

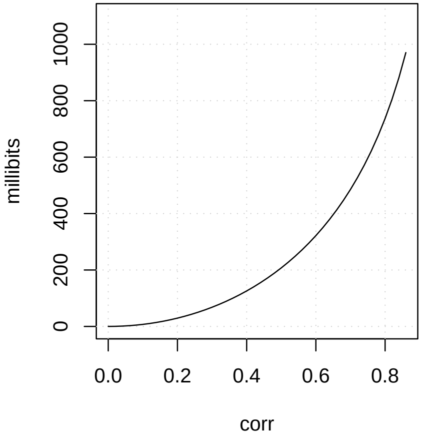
The DWIG is the graph with edge weights

- $$\begin{aligned} w_{ij} &= \mathsf{Inf}(X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{\setminus ij}) \\ &= -\frac{1}{2}\log\left(1 - \mathsf{corr}(X_i, X_j | \boldsymbol{X}_{\setminus ij})^2\right). \end{aligned}$$

w_{ij} is the extra information for predicting X_i provided by X_j when the rest have been taken into account.

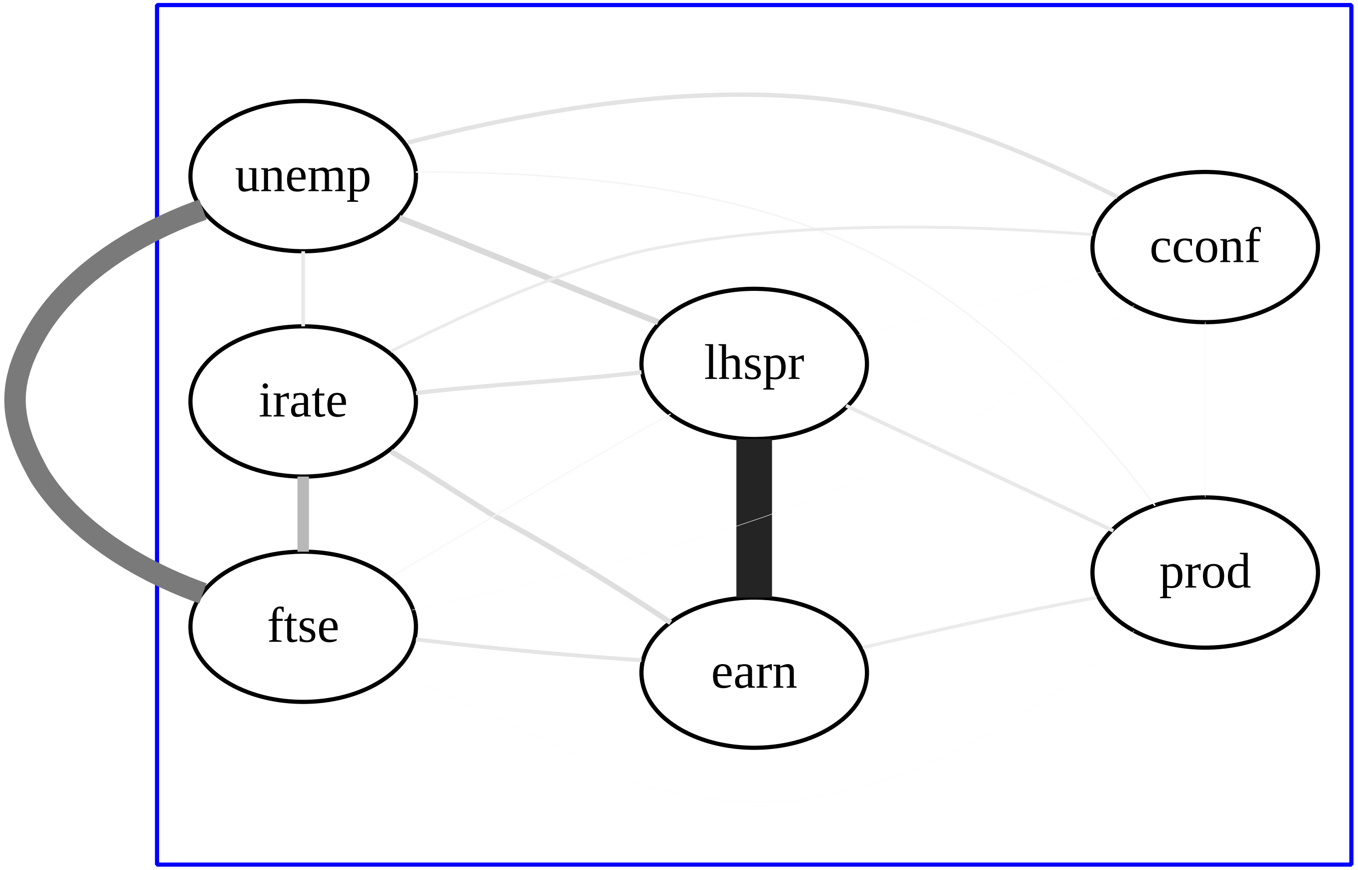
Toy example





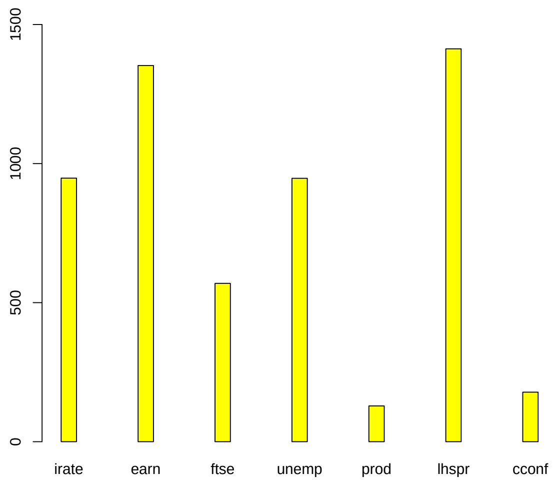
Graph is complete, all edges appear, natural display sets edge width (and tone) $\propto w_{ij}$.

DWIG of macro-economic covariates



set max 500 mbits
actual 426.752

Interesting connections.



Inf_{vif} again

Remarks on the DWIG

DWIG make sense of $\mathrm{Inf}_{\mathrm{vif}}$

- largest related to vertices with thickest edges
- smallest to vertices with invisible edges,
- new information: inter-vertex relationships.

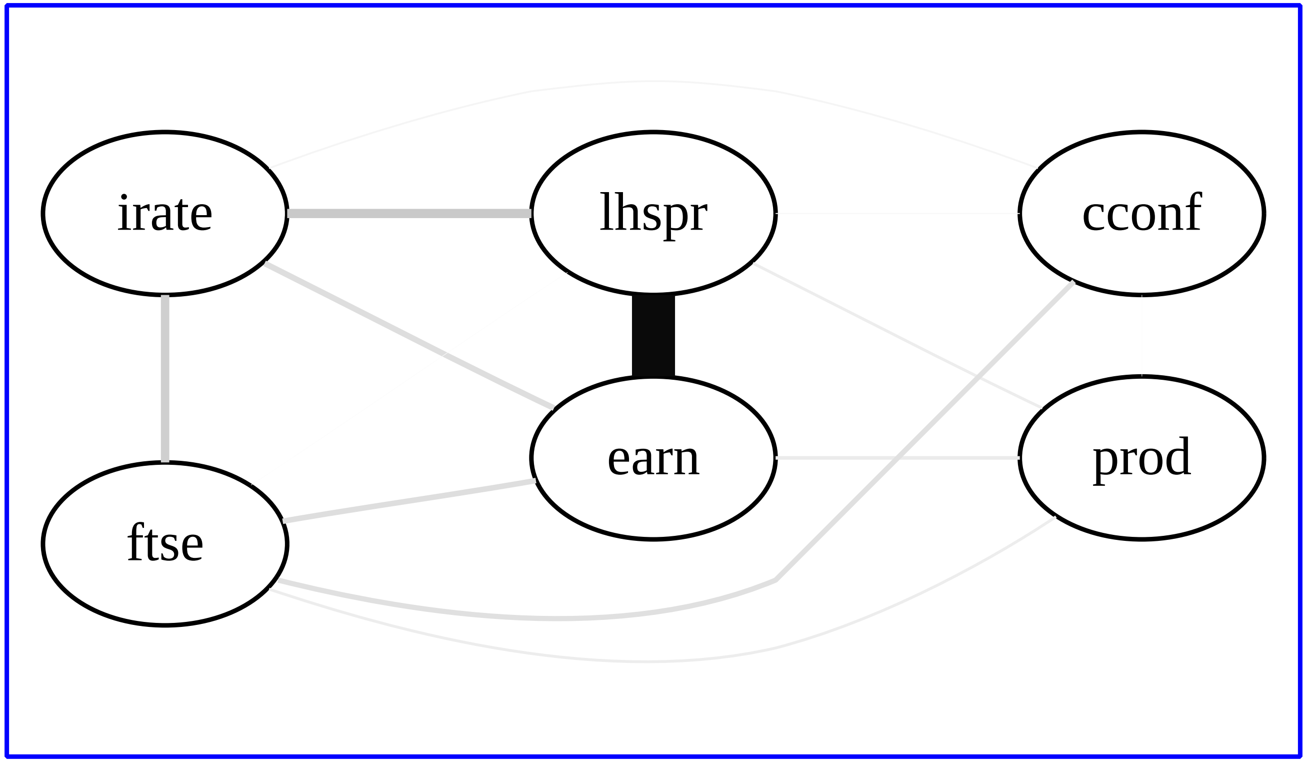
Strategy:

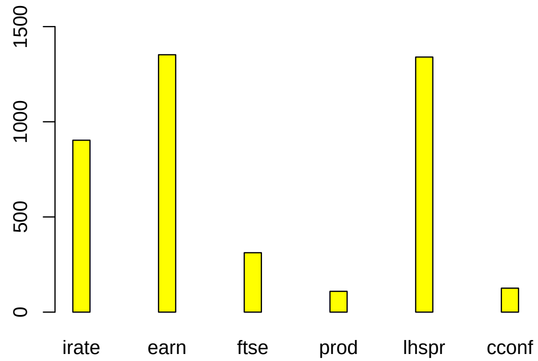
use DWIG to reduce dependency in MEVs by eliminating highly dependent variables.

Use DWIG to test out new explanatory variables.

Reduce dependence by marginalising

Take out unemp [least sig $\widehat{\beta}$ in BC07 paper]

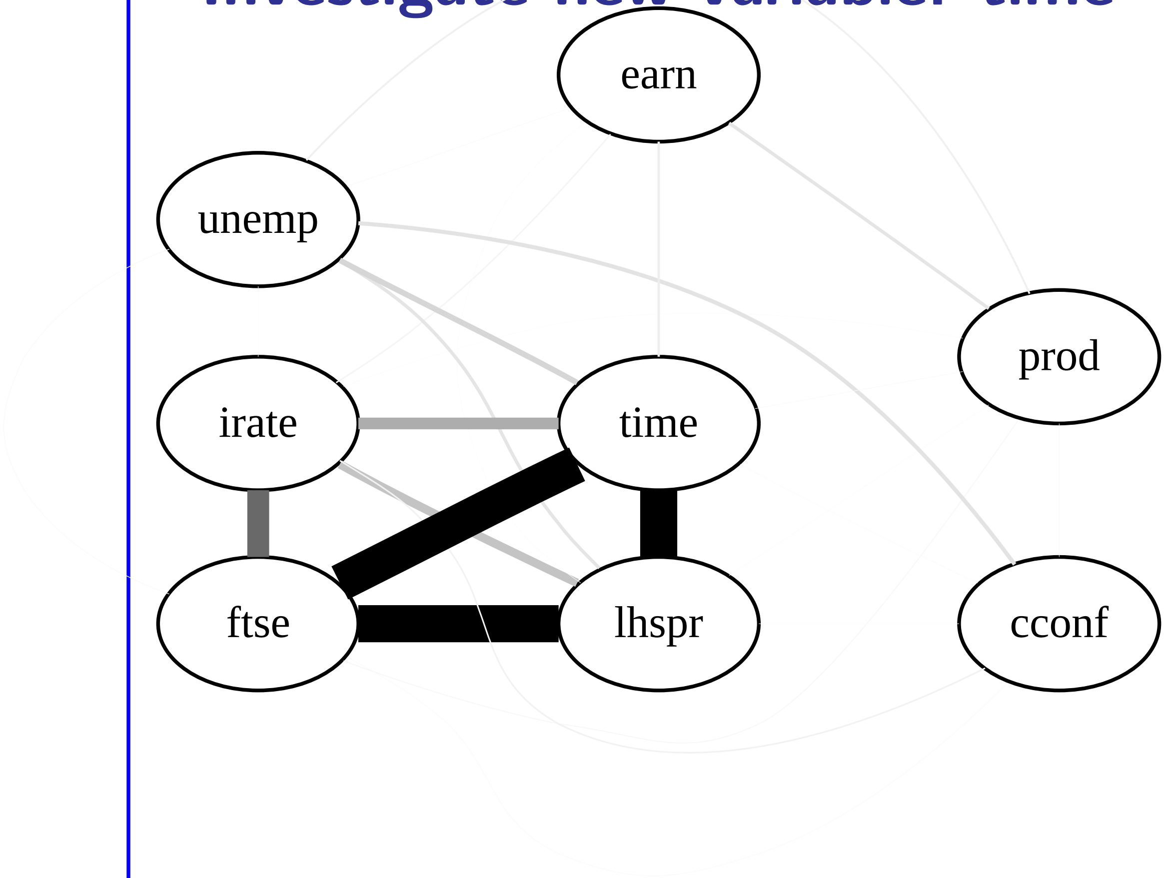




set max 500 mbits
actual 476.0549

And ftse $\mathrm{Inf}_{\mathrm{vif}}$ halves.

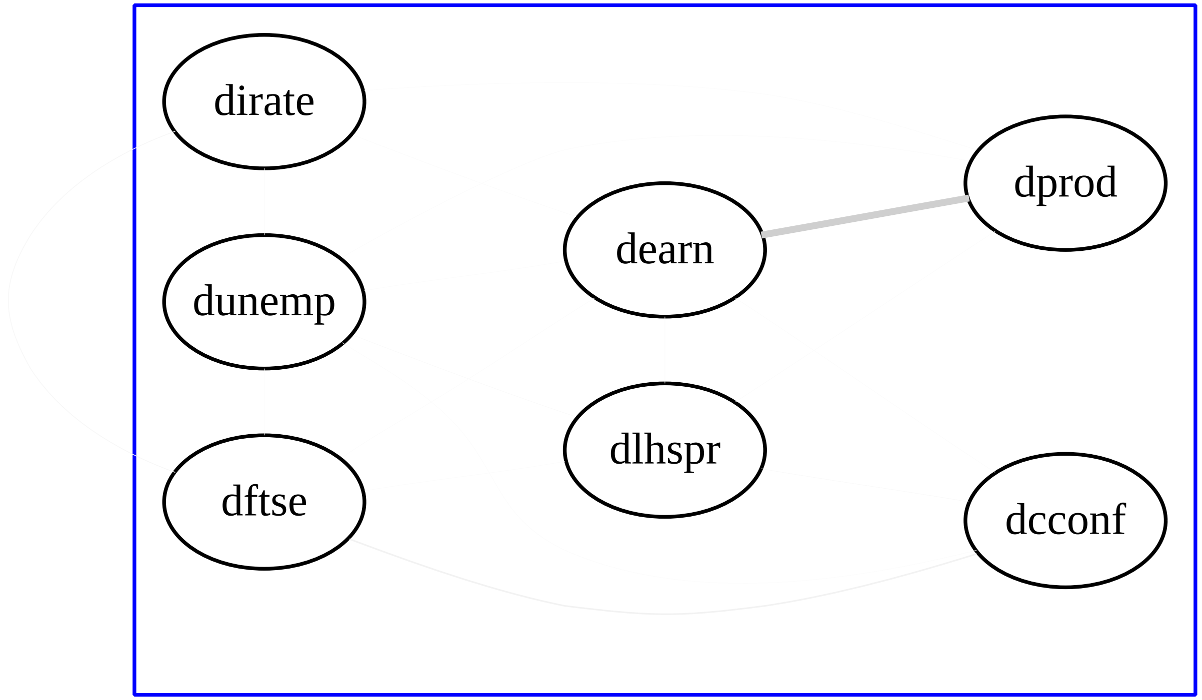
Investigate new variable: time



set max 500 mbits
actual 2240.2976

Not a good idea.

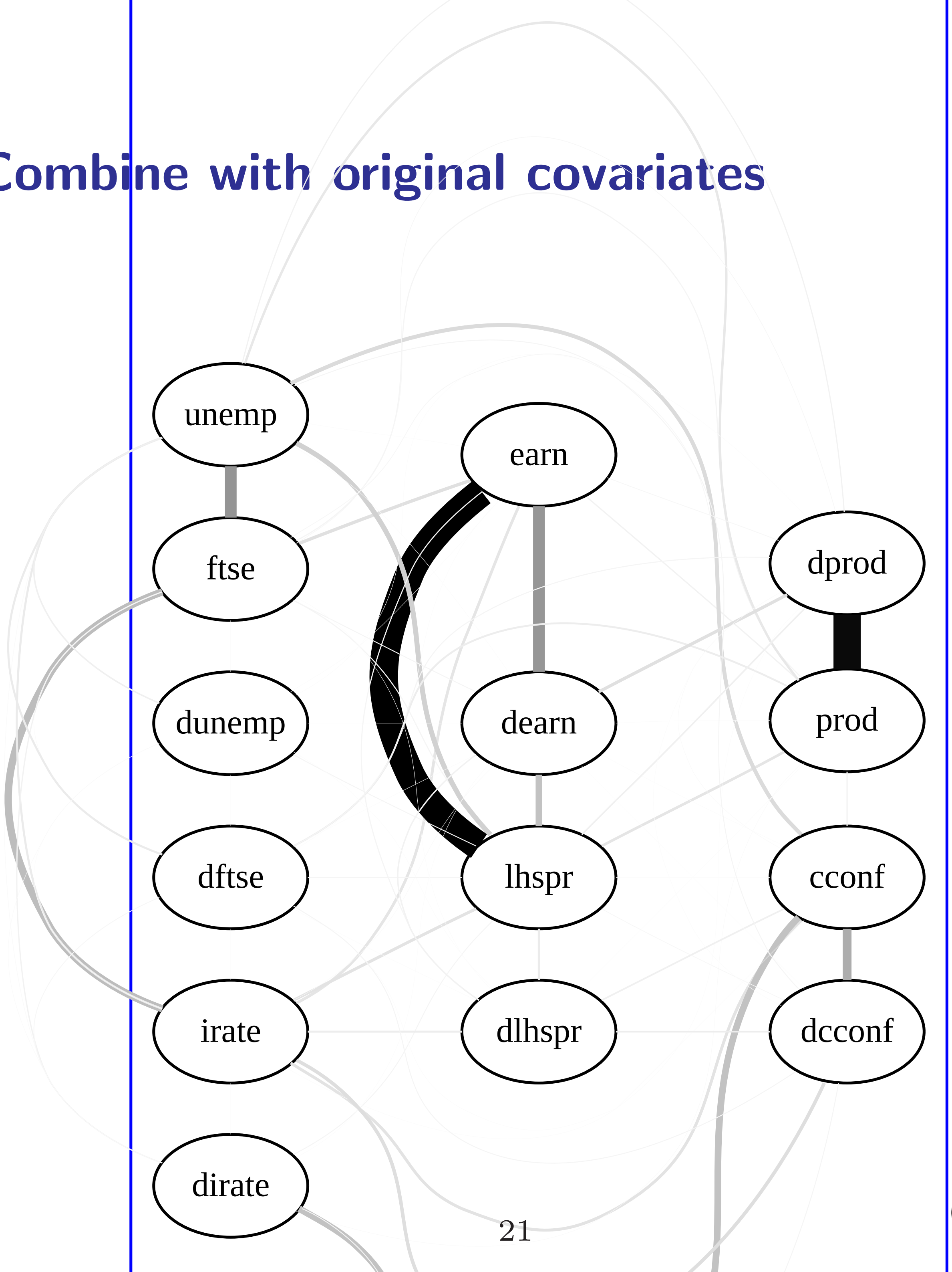
Investigate differenced covariates



set max 500 mbits
actual 93.0542

Nearly complete independence!

Combine with original covariates



Complex but feasible.

Example R code, requires graphviz

Eventual package, easy syntax:

```
xdf = read.table(file='charts.csv',header = T, sep = ",")
xlabs = c('irate','earn','ftse','unemp','prod','lhspr','cconf')
names(xdf) = xlabs

infmat = getGGinf(xdf)$offdiag

xmfo = ~irate+earn+ftse+unemp+prod+lhspr+cconf
bloc = list(xmfo,  ~irate+unemp+ftse, ~prod+cconf, ~earn+lhspr)

mkchaindotfile('MEVdwig',xlabs,list(bloc),infmat,maxdiv=500)
```

gives graph.

Default risk

Default on credit card is binary, with prob

$p_{it} = \mathsf{P}(\mathsf{def}|z_i, x_t)$

z_i are individual score card variables

x_t ME variables.

First, assume homogeneous over individuals,

number of portfolio defaults $\sim \mathsf{Bino}(n_t, p_t)$,

estimate by logistic regression.

Then, divide into groups for interactions.

[Guess work, as data confidential.]

Application to logistic regression

Variance of regression coefficients

$$
\begin{aligned}
\text{var}(\widehat{\boldsymbol{\beta}}) &= (X'VX)^{-1} \quad \text{with} \\
v_t &= p_t(1-p_t), \quad V \text{ diagonal.}
\end{aligned}
$$

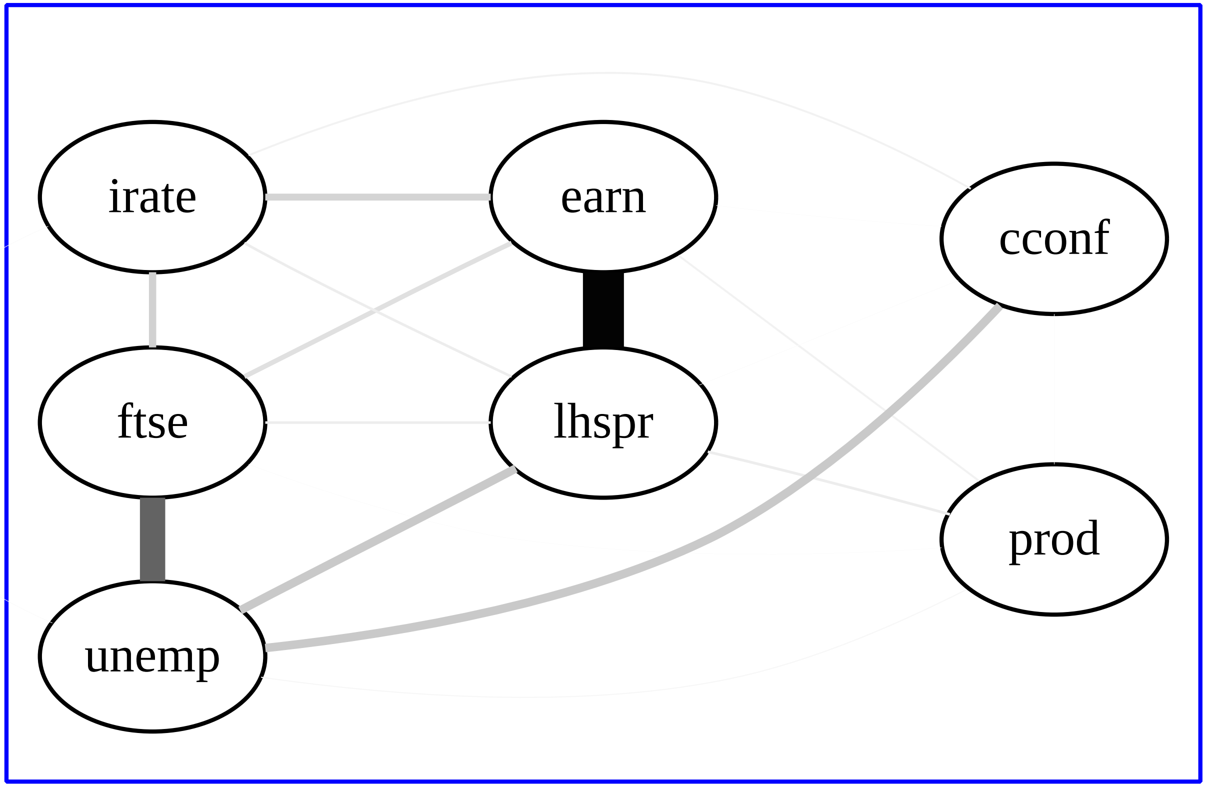
Use BC07 main effects $p_t = \dfrac{\exp(-.1+\eta_t)}{1+\exp(-.1+\eta_t)}$

with $\eta_t = x_t\widehat{\boldsymbol{\beta}}$.

Take $\text{var}(\boldsymbol{X}) = (X'VX)/n$,

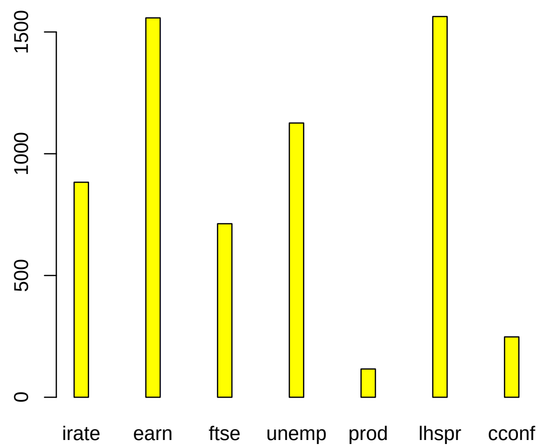
same vif and information theory goes through.

DWIG of MEVs via LogReg



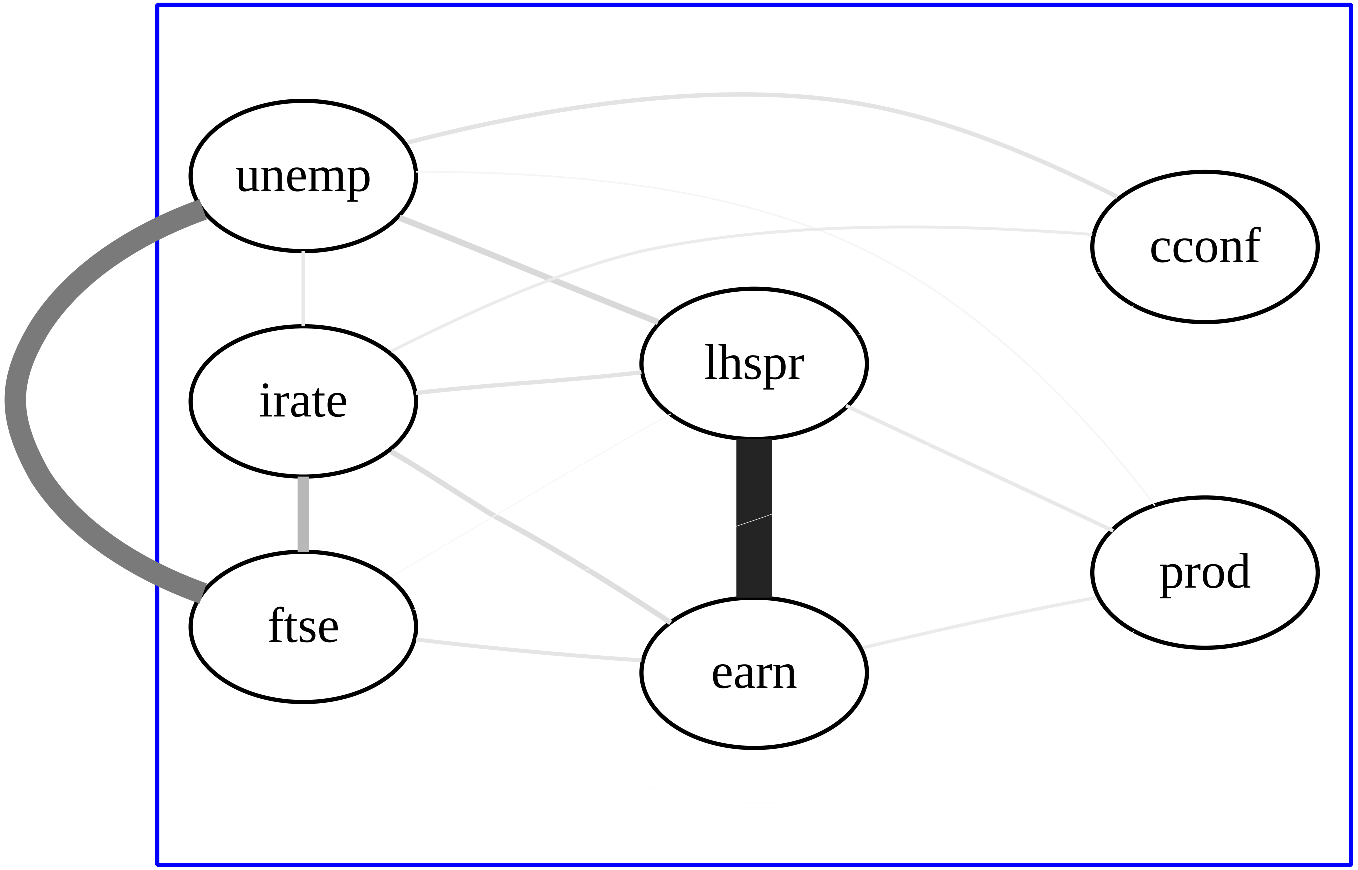
set max 500 mbits
actual 491.333

Very similar.

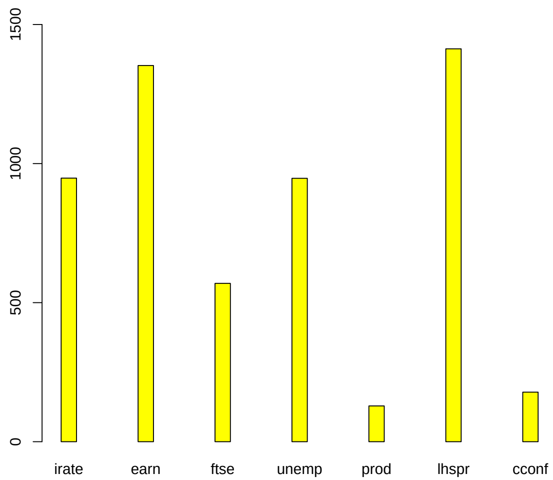


Inf $_{vif}$

DWIG of MEVs via OLS



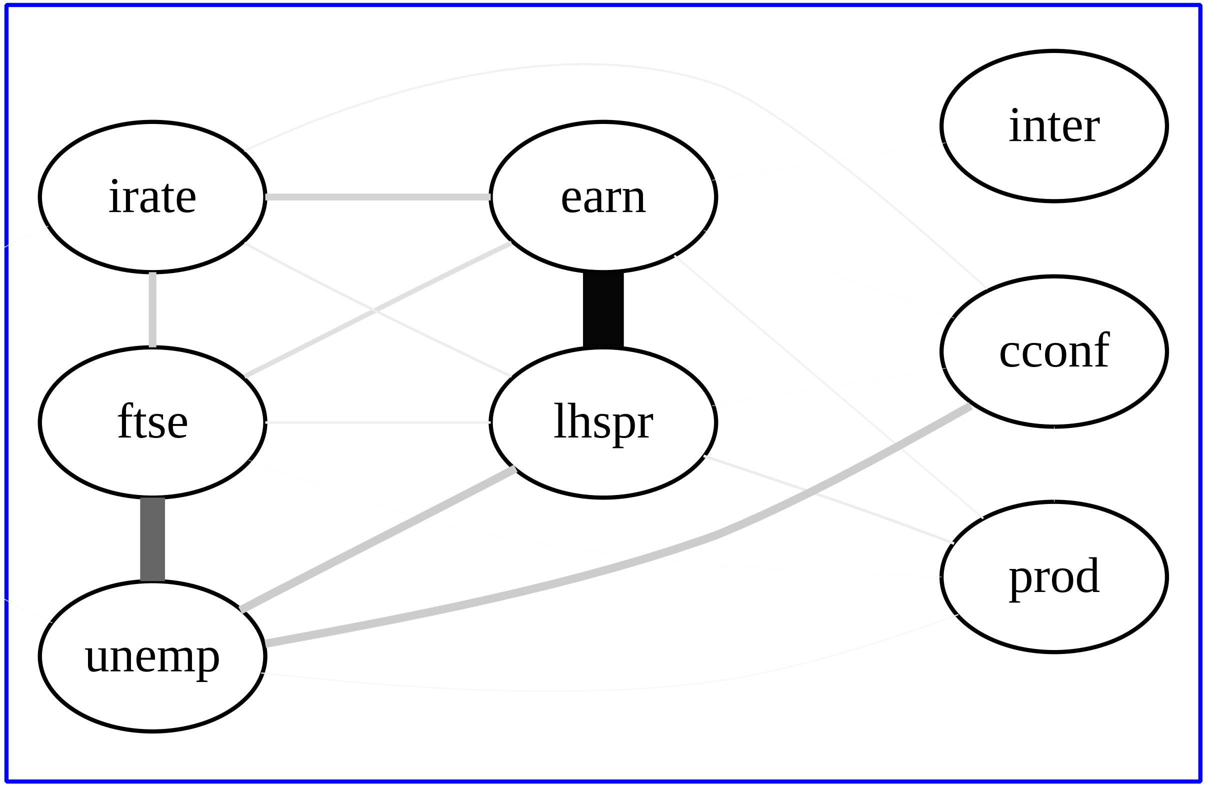
set max 500 mbits
actual 426.752



Inf $_{vif}$

DWIG of MEVs + interaction via LogReg

Add Earnings * private tenant (y/n), $\widehat{\beta}_j = 8.59$ to linear predictor, assuming 50% =yes



set max 500 mbits
actual 489.3911

Little difference.

Concluding remarks

- DWIG gives bird's-eye view of relationships
  - additive information scale, and
  - conditional measures of dependence.
  - No directions, not causal.
- DWIG strength, Inf_{vif} measured on same scale.
- DWIG examines dependences in explanatory variables, with view to clarifying statements about effects on response variable.
- DWIG methodology extends to logistic reg, w/wo interaction, similar results for survival analysis (conjecture).