Graduate School Class Reminders

- Maintain six feet of distancing
- Please sit in the same chair each class time
- Observe entry/exit doors as marked
- Use hand sanitizer when you enter/exit the classroom
- Use a disinfectant wipe/spray to wipe down your learning space before and after class
- Media Services: 414 955-4357 option 2

Documentation on the web

- CRAN: `http://cran.r-project.org`
- R manuals: `https://cran.r-project.org/manuals.html`
- SAS: `http://support.sas.com/documentation`
- Step-by-Step Programming with Base SAS 9.4 (SbS): `https://documentation.sas.com/api/docsets/basess/9.4/content/basess.pdf`
- SAS 9.4 Programmer s Guide: Essentials (PGE): `https://documentation.sas.com/api/docsets/lepg/9.4/content/lepg.pdf`
- Wiki: `https://wiki.biostat.mcw.edu` (MCW/VPN)

Eigen C++ class library and RcppEigen

- Eigen is C++ header-only class library that provides linear algebra calculations with objects like vectors and matrices
- RcppEigen is the R package that provides Eigen within Rcpp
- R is quite fast for linear algebra, but there are occasions where something faster like Eigen is needed
- The Eigen documentation can be found at `https://eigen.tuxfamily.org/dox/GettingStarted.html`
- And RcppEigen is documented in BateEdde13

Two-stage least squares

- Suppose that we a continuous treatment, T, (like the dosage of a drug infusion); and a continuous outcome, Y, (like the size of a myocardial infarct)
- In randomized experiments, the patient's treatment is assigned at random, i.e., ignoring the patient's characteristics
- In non-randomized experiments, we assume that the patient's treatment is NOT assigned at random, i.e., the patient's characteristics likely influence the decision
- Such characteristics are known as *confounders*
- If these confounders are observed, then they can be adjusted for by linear regression to estimate the treatment effect
- However, in many cases, these confounders are unobserved, therefore, you can NOT adjust for them by linear regression
- There is a causal inference method known as two-stage least squares or 2SLS which can still estimate the treatment effect in the presence of unobserved confounding provided *instruments* are observed

Causal diagrams and 2SLS

$$
\begin{array}{rcl}
\text{Confounders} & & \text{Instruments} \\
X & & Z \\
\downarrow & \searrow & \downarrow \\
Y & \leftarrow & T \\
\text{Outcome} & & \text{Treatment}
\end{array}
$$

$$
\begin{aligned}
t_i &= \alpha_0 + z_i'\alpha_z + x_i'\alpha_x + \epsilon_{1_i} && \text{Stage 1} \\
y_i &= \beta_0 + \hat{t}_i\beta_t + x_i'\beta_x + \epsilon_{2_i} && \text{Stage 2} \\
\beta_t & \text{ is the treatment effect} \\
& \textbf{where } \hat{t}_i = \hat{\alpha}_0 + z_i'\hat{\alpha}_z + x_i'\hat{\alpha}_x
\end{aligned}
$$

But, what is the variance estimate of $\hat{\beta}_t$: $\mathrm{V}\left[\hat{\beta}_t\right]$?

It is based on $\mathrm{V}\left[\hat{\gamma}_t\right]$ as in $y_i = \gamma_0 + t_i\gamma_t + x_i'\gamma_x + \epsilon_{0_i}$: Stage 0.

H. Theil. Repeated least squares applied to complete equation systems. The Hague: Central Planning Bureau, 1953.

The US National Longitudinal Survey of Young Men (NLSYM)

- This data set is contained in the `nlsym` data frame: `/data/shared/04224/nlsym.rds`
- This data set contains 3613 observations for men in 1976
- NLSYM began in 1966 with 5525 men aged 14:24 and continued with follow-up surveys through 1981
- The question here is:
  Are there monetary returns of post-secondary education?
- See the R program `/data/shared/04224/nlsym.R` which organizes the data for analysis

The US National Longitudinal Survey of Young Men (NLSYM)

- Treatment, T: `ed76` years of education in 1976
- Confounders, X: `exp76` years of experience in 1976,
  `exp762` centered years of experience squared in 1976,
  `black` African-American,
  `smsa76r` residing in an SMSA 1976 and
  `reg76r` residing in the south 1976
- Outcome, Y: `lwage76` log wages in 1976 (outliers trimmed)
- Instruments, Z: `nearc2` grew up near 2-year college
  and `nearc4` grew up near 4-year college

$w_i = \exp y_i$	Wages
$\mathbf{E}[y_i] = \beta_0 + \hat{t}_i\beta_t + x_i'\beta_x$	
$u_i = \exp[\hat{t}_i\beta_t]\exp[\beta_0 + x_i'\beta_x]$	$\hat{t}_i$ years of education
$w_i = \exp[(\hat{t}_i + 1)\beta_t]\exp[\beta_0 + x_i'\beta_x]$	$\hat{t}_i + 1$ years
$w_i = \exp\beta_t u_i$	Multiple for **+1** years

Matrix inversion of real-valued square matrices

- $A_{n \times n} A^{-1}_{n \times n} = I_{n \times n}$
- The R function to compute A^{-1} is `> solve(A)`
- But, singular matrices have no unique matrix inverse
- The *condition number* is an indicator of how numerically unstable the matrix inversion is likely to be
  or, how close to a singular matrix do we have here?
  very large condition numbers suggest singularity
- The R function to compute the condition number is
  `> kappa(A)`
- For example, the condition number of a singular matrix in the `kappa` function documentation is about 10^{17}
- By way of comparison, the current Big Bang model suggests that the universe is 13.8 billion years old or 4.4×10^{17} seconds

Cholesky Decomposition of real-valued square matrices

- For symmetric, positive definite matrices: a matrix *square root*
- Cholesky decomposition: $A_{n\times n} = LL'$
  where $L_{n\times n}$ is a lower triangular matrix
  (all of the elements above the diagonal are zero)
- Provided by the `chol` function which produces the alternative representation: $A = R'R$ where $R' = L$
- And it is useful for calculating the matrix inverse in a numerically stable way: $A^{-1} = (L^{-1})'L^{-1} = R^{-1}(R^{-1})'$
- The formula for calculating L is as follows

$$L_{ij} = L_{jj}^{-1}\left[A_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk}\right] \text{ where } i > j$$

$$L_{jj} = \sqrt{A_{jj} - \sum_{k=1}^{j-1} L_{jk}^2}$$

See `cholesky.R`

Linear regression with linear algebra

$$\begin{aligned}
y_i &= \beta_0 + x_i'\beta_x + \epsilon_i \\
\epsilon_i &\sim \mathrm{N}\left(0,\ \sigma^2\right) \text{ where } i = 1, \ldots, n \\
X &= \begin{bmatrix} 1 & x_1' \\ \vdots & \vdots \\ 1 & x_n' \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_x \end{bmatrix} \\
\hat{\beta} &= (X'X)^{-1}X'y \\
\widehat{\mathrm{V}\left[\hat{\beta}\right]} &= \hat{\sigma}^2(X'X)^{-1}
\end{aligned}$$

HW: 2SLS with RcppEigen

- In BateEdde13, there is a nice example of linear regression with matrix inversion via Cholesky decomposition
- see the RcppEigen source code `lmEigen.h` and `lmEigen.cpp` and its call at the bottom of `nlsym.R` (commented out)
- We will adapt this code to perform 2SLS by creating a new function: `TSLS`
- We can compare the results that we get from the `tsls` function from the `sem` package (also at the bottom of `nlsym.R`)
- What is the income multiplier for an additional year of education?