Robust, easy standard errors with the clubSandwich package

James E. Pustejovsky

April 25, 2018

Conventional regression analysis

A generic regression model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

Conventional regression analysis

A generic regression model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

Statistics 101 regression analysis makes two strong assumptions:

1. Errors are **independent**, so that $\mathrm{corr}(e_i, e_j) = 0$ when $i \neq j$
2. Errors are **homoskedastic**, so $\mathrm{Var}(e_i) = \sigma^2$ for all i

Conventional regression analysis

A generic regression model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

Statistics 101 regression analysis makes two strong assumptions:

1. Errors are **independent**, so that $\mathrm{corr}(e_i, e_j) = 0$ when $i \neq j$
2. Errors are **homoskedastic**, so $\mathrm{Var}(e_i) = \sigma^2$ for all i

Many situations where these assumptions are untenable:

- Multi-stage survey data
- Repeated measurements data
- Longitudinal/panel data
- Cluster-randomized trials

Effect of minimum legal drinking age on motor vehicle fatalities

- Carpenter & Dobkin (2011) examine effects of **changes in minimum legal drinking age** on motor vehicle fatalities among 18-20 year olds.

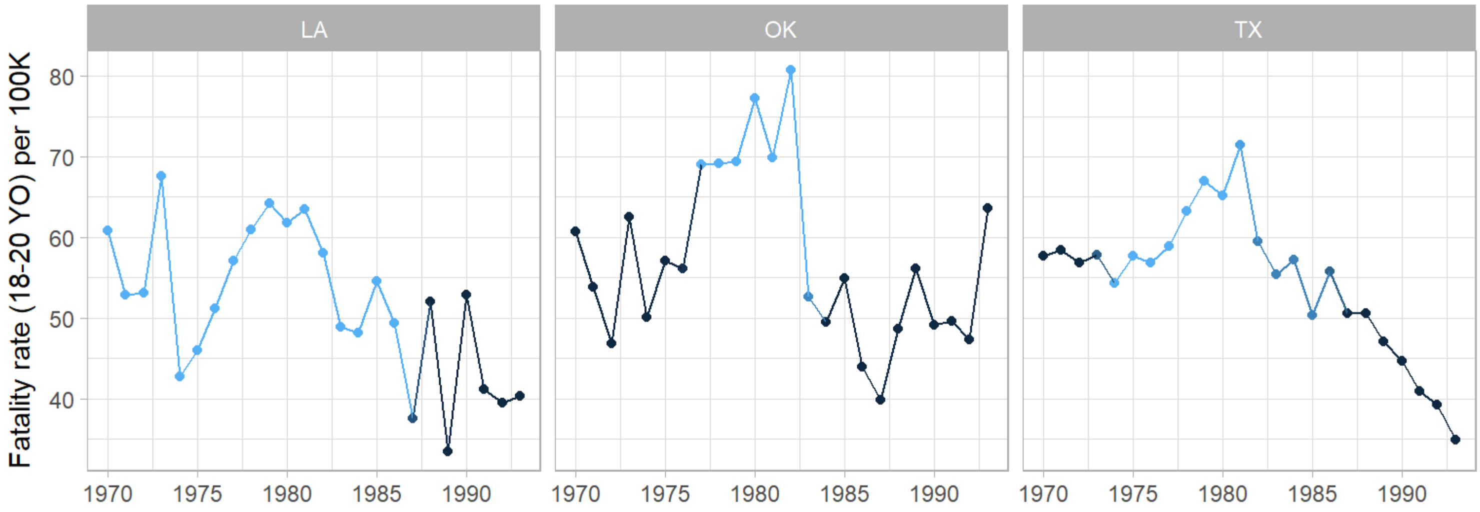
Effect of minimum legal drinking age on motor vehicle fatalities

- Carpenter & Dobkin (2011) examine effects of **changes in minimum legal drinking age** on motor vehicle fatalities among 18-20 year olds.
- *Repeated measures* of annual motor vehicle fatalities for all 50 states + DC, 1970-1983

Effect of minimum legal drinking age on motor vehicle fatalities

- Carpenter & Dobkin (2011) examine effects of **changes in minimum legal drinking age** on motor vehicle fatalities among 18-20 year olds.

- *Repeated measures* of annual motor vehicle fatalities for all 50 states + DC, 1970-1983



An easy fix with sandwich estimators

- Calculate regression coefficient estimates $\hat{\beta}$ per usual (ordinary least squares)

An easy fix with sandwich estimators

- Calculate regression coefficient estimates $\hat{\beta}$ per usual (ordinary least squares)
- Use *sandwich* estimators for standard errors of $\hat{\beta}$.

An easy fix with sandwich estimators

- Calculate regression coefficient estimates $\hat{\beta}$ per usual (ordinary least squares)

- Use ***sandwich*** estimators for standard errors of $\hat{\beta}$.

- Sandwich estimators are based on ***weaker*** assumption that observations can be grouped into J ***clusters*** of independent observations:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + e_{ij}$$

  - $\text{cor}(e_{hj}, e_{ik}) = 0$ if observations are in different clusters $(j \neq k)$
  - $\text{cor}(e_{hj}, e_{ij}) = \rho_{hij}$ for observations in the same cluster
  - $\text{Var}(e_{ij}) = \phi_{ij}$, allowing for heteroskedasticity

Plain sandwich estimators

Actual variance of coefficient estimate $\hat{\beta}$:

$$\mathrm{Var}(\hat{\beta}) = \frac{1}{J} \mathrm{B} \left(\frac{1}{J} \sum_{j=1}^{J} X_j' \Phi_j X_j \right) \mathrm{B}$$

where $\Phi_j = \mathrm{Var}(e_j)$ and $\mathrm{B} = \left(\frac{1}{J} \sum_{j=1}^{J} X_j' X_j \right)^{-1}$.

Plain sandwich estimators

Actual variance of coefficient estimate $\hat{\beta}$:

$$\mathrm{Var}(\hat{\beta}) = \frac{1}{J} B \left(\frac{1}{J} \sum_{j=1}^{J} X_j' \Phi_j X_j \right) B$$

where $\Phi_j = \mathrm{Var}(e_j)$ and $B = \left(\frac{1}{J} \sum_{j=1}^{J} X_j' X_j \right)^{-1}$.

The plain sandwich estimator:

$$V^{plain} = \frac{1}{J} B \left(\frac{1}{J} \sum_{j=1}^{J} X_j' e_j e_j' X_j \right) B$$

for residuals $e_j = Y_j - X_j \hat{\beta}$

A plain sandwich

```
# fit regression from Carpenter & Dobkin (2011)
MLDA_fit <- lm(mrate ~ 0 + legal + beertaxa + totpercap
               + factor(State) + factor(year),
               weights = pop,
               data = MV_deaths)
```

A plain sandwich

```
# fit regression from Carpenter & Dobkin (2011)
MLDA_fit <- lm(mrate ~ 0 + legal + beertaxa + totpercap
               + factor(State) + factor(year),
               weights = pop,
               data = MV_deaths)
```

```
library(clubSandwich)

# type = "CR0" is the plain sandwich variance estimator
MLDA_plain <- vcovCR(MLDA_fit, cluster = MV_deaths$State,
                     type = "CR0")
```

A plain sandwich

```
# fit regression from Carpenter & Dobkin (2011)
MLDA_fit <- lm(mrate ~ 0 + legal + beertaxa + totpercap
               + factor(State) + factor(year),
               weights = pop,
               data = MV_deaths)
```

```
library(clubSandwich)

# type = "CR0" is the plain sandwich variance estimator
MLDA_plain <- vcovCR(MLDA_fit, cluster = MV_deaths$State,
                     type = "CR0")
```

```
coef_test(MLDA_fit, vcov = MLDA_plain, test = "z", coefs = 1:3)
```

```
##        Coef Estimate   SE p-val (z) Sig.
## 1     legal     3.17 1.75    0.0690    .
## 2  beertaxa     3.25 4.81    0.4989
## 3 totpercap     7.71 3.61    0.0327    *
```

A plain sandwich

```r
# fit regression from Carpenter & Dobkin (2011)
MLDA_fit <- lm(mrate ~ 0 + legal + beertaxa + totpercap
               + factor(state) + factor(year),
               weights = pop,
               data = MV_deaths)
```

```r
library(clubSandwich)

# type = "CR0" is the plain sandwich variance estimator
MLDA_plain <- vcovCR(MLDA_fit, cluster = MV_deaths$state,
                     type = "CR0")
```

```r
coef_test(MLDA_fit, vcov = MLDA_plain, test = "z", coefs = 1:3)
```

```
##        Coef Estimate   SE p-val (z) Sig.
## 1     legal     3.17 1.75    0.0690    .
## 2  beertaxa     3.25 4.81    0.4989
## 3 totpercap     7.71 3.61    0.0327    *
```

- Similar methods implemented in the `sandwich` package (Zeileis, 2004).

Problems with plain sandwiches

Plain sandwich estimators ***require a large number of clusters*** to work well.

- Downward bias if the number of clusters is not big enough
- Hypothesis tests have inflated type-I error
- Confidence intervals have less-than-advertised coverage

Problems with plain sandwiches

Plain sandwich estimators ***require a large number of clusters*** to work well.

- Downward bias if the number of clusters is not big enough
- Hypothesis tests have inflated type-I error
- Confidence intervals have less-than-advertised coverage

What counts as "large enough" depends on:

- ***number of clusters***, not number of observations
- distribution of predictors X within and across clusters

Problems with plain sandwiches

Plain sandwich estimators ***require a large number of clusters*** to work well.

- Downward bias if the number of clusters is not big enough
- Hypothesis tests have inflated type-I error
- Confidence intervals have less-than-advertised coverage

What counts as "large enough" depends on:

- ***number of clusters***, not number of observations
- distribution of predictors X within and across clusters

How can you tell whether your plain sandwich estimators are edible?

Fancy sandwiches

- Adjust the residuals so that they are unbiased under a working model (Bell & McCaffrey, 2002, 2006; Pustejovsky & Tipton, 2016):

$$V^{club} = \frac{1}{J} B \left(\frac{1}{J} \sum_{j=1}^{J} X_j' \color{red}{A_j} e_j e_j' \color{red}{A_j} X_j \right) B$$

Fancy sandwiches

- Adjust the residuals so that they are unbiased under a working model (Bell & McCaffrey, 2002, 2006; Pustejovsky & Tipton, 2016):

$$V^{club} = \frac{1}{J} B \left(\frac{1}{J} \sum_{j=1}^{J} X_j' A_j e_j e_j' A_j X_j \right) B$$

- Use degrees-of-freedom adjustments for hypothesis tests and confidence intervals.

Fancy sandwiches

- Adjust the residuals so that they are unbiased under a working model (Bell & McCaffrey, 2002, 2006; Pustejovsky & Tipton, 2016):

$$V^{club} = \frac{1}{J} B \left(\frac{1}{J} \sum_{j=1}^{J} X_j' A_j e_j e_j' A_j X_j \right) B$$

- Use degrees-of-freedom adjustments for hypothesis tests and confidence intervals.

- These methods work well ***even when*** J ***is small*** and even when the working model isn't correct.

Fancy sandwiches

- Adjust the residuals so that they are unbiased under a working model (Bell & McCaffrey, 2002, 2006; Pustejovsky & Tipton, 2016):

$$V^{club} = \frac{1}{J} B \left(\frac{1}{J} \sum_{j=1}^{J} X_j' A_j e_j e_j' A_j X_j \right) B$$

- Use degrees-of-freedom adjustments for hypothesis tests and confidence intervals.

- These methods work well ***even when*** J ***is small*** and even when the working model isn't correct.

- Degrees-of-freedom are *diagnostic*, so low d.f. implies:

    - little information available for variance estimation
    - asymptotic approximations haven't "kicked in"

Plain vs. club sandwich estimators

```
coef_test(MLDA_fit, vcov = MLDA_plain, test = "z", coefs = 1:3)
```

```
##          Coef Estimate   SE p-val (z) Sig.
## 1       legal     3.17 1.75    0.0690    .
## 2   beertaxa     3.25 4.81    0.4989     
## 3 totpercap     7.71 3.61    0.0327    *
```

Plain vs. club sandwich estimators

```r
coef_test(MLDA_fit, vcov = MLDA_plain, test = "z", coefs = 1:3)
```

```
##        Coef Estimate   SE p-val (z) Sig.
## 1     legal     3.17 1.75    0.0690    .
## 2  beertaxa     3.25 4.81    0.4989     
## 3 totpercap     7.71 3.61    0.0327    *
```

```r
# type = "CR2" for small-sample adjustments
MLDA_club <- vcovCR(MLDA_fit,
                    cluster = MV_deaths$State,
                    type = "CR2")
coef_test(MLDA_fit, vcov = MLDA_club, coefs = 1:3)
```

```
##        Coef Estimate   SE d.f. p-val (Satt) Sig.
## 1     legal     3.17 1.93 6.52        0.148     
## 2  beertaxa     3.25 5.20 8.23        0.548     
## 3 totpercap     7.71 3.42 5.73        0.067    .
```

R package `clubSandwich`

Methods work with many sorts of regression models:

- logistic/generalized linear models with `glm()`
- multivariate regression with `mlm` objects
- instrumental variables with `AER::ivreg()`
- panel data models with `plm::plm()`
- generalized least squares with `nlme::gls()`
- hierarchical linear models with `nlme::lme()`
- meta-analysis with `metafor::rma()` and `metafor::rma.mv()`

R package `clubSandwich`

Methods work with many sorts of regression models:

- logistic/generalized linear models with `glm()`
- multivariate regression with `mlm` objects
- instrumental variables with `AER::ivreg()`
- panel data models with `plm::plm()`
- generalized least squares with `nlme::gls()`
- hierarchical linear models with `nlme::lme()`
- meta-analysis with `metafor::rma()` and `metafor::rma.mv()`

Object-oriented design for extensibility.

R package `clubSandwich`

Methods work with many sorts of regression models:

- logistic/generalized linear models with `glm()`
- multivariate regression with `mlm` objects
- instrumental variables with `AER::ivreg()`
- panel data models with `plm::plm()`
- generalized least squares with `nlme::gls()`
- hierarchical linear models with `nlme::lme()`
- meta-analysis with `metafor::rma()` and `metafor::rma.mv()`

Object-oriented design for extensibility.

Under active development

- Available on CRAN
- Development repo: https://github.com/jepusto/clubSandwich

Thanks!

pusto@austin.utexas.edu

http://jepusto.github.io

References

- Bell, R. M., & McCaffrey, D. F. (2002). *Survey Methodology*. http://www.statcan.gc.ca/pub/12-001-x/2002002/article/9058-eng.pdf
- McCaffrey, D. F., & Bell, R. M. (2006). *Statistics in Medicine*. http://doi.org/10.1002/sim.2502
- Carpenter, C., & Dobkin, C. (2011). *Journal of Economic Perspectives*. http://doi.org/10.1257/jep.25.2.133
- Pustejovsky, J. E. & Tipton, E. (2016). *Journal of Business and Economic Statistics*. https://doi.org/10.1080/07350015.2016.1247004
- Zeileis, A. (2004). *Journal of Statistical Software*. http://www.jstatsoft.org/v11/i10/.