propertee: Flexible Covariance Adjustment and Improved Standard Errors in Analyses of Intact Clusters

Joshua Wasserman

Ben Hansen

Department of Statistics, University of Michigan – Ann Arbor

It would be nice if `lm()`'s and `glm()`'s in causal analyses routinely returned estimates of marginal causal effects...

It would be nice if `lm()` and `glm()` always returned (marginal) causal estimates...

Add "experimental" somewhere in here, also because people don't only use lm() or glm() for causal reasons

- Simple to use `lm()` or `glm()` to run a regression
- Challenging to ensure the coefficient on a treatment assignment indicator has the desired interpretation

Expand with another bullet by saying "ATE" is common estimand in experiments and "ATT" is common in observational studies

- Complicated standard error calculations may preclude more satisfying methods for covariance adjustment

Remove SE's bullet point

It would be nice if `lm()` and `glm()` always returned causal estimates…

- Simple to use `lm()` or `glm()` to run a regression
- Challenging to ensure the coefficient on a treatment assignment indicator has the desired interpretation
- Complicated standard error calculations may preclude more satisfying methods for covariance adjustment
- **propertee introduces `lmitt()`, an analog of `lm()` that:**
    1) Reports difference-in-means estimates (optionally, with covariance adjustment) with causal interpretations
    2) Returns a fitted model object equipped with special S3 methods for variance estimation

A call to `lmitt()` in the context of the Work, Family and Health Study (WFHS)[1]

```
# a call to lmitt() for a difference-in-means estimate with covariance adjustment

tmod <- lmitt(
    TOPPRIORITY ~ 1,
    specification = StudySpec,
    data = RespondentData,
    absorb = FALSE,
    offset = cov_adj(logistic_mod),
    weights = “ate”,
    ...
)
```

A call to `lmitt()` in the context of the Work, Family and Health Study (WFHS)[1]

```
# a call to lmitt() for a difference-in-means estimate with covariance adjustment

tmod <- lmitt(
   TOPPRIORITY ~ 1,
   specification = StudySpec,
   data = RespondentData,
   absorb = FALSE,
   offset = cov_adj(logistic_mod),
   weights = "ate",
   ...
)
```

`StudySpec` maps cluster-level treatment assignments and cluster-level inverse propensity weights (computed based on the `weights` argument) to observations in RespondentData

A call to `lmitt()` in the context of the Work, Family and Health Study (WFHS)[1]

```
# a call to lmitt() for a difference-in-means estimate with covariance adjustment

tmod <- lmitt(
   TOPPRIORITY ~ 1,
   specification = StudySpec,
   data = RespondentData,
   absorb = FALSE,
   offset = cov_adj(logistic_mod),
   weights = “ate”,
   ...
)
```

remove "with covariance adjustment"

`StudySpec` maps cluster-level treatment assignments and cluster-level inverse propensity weights (computed based on the `weights` argument) to observations in RespondentData

No need to join cluster-level assignment data with observation-level response data!

Replace studyspec with formula, don't name the arg; this will show we compute correct IPW weights from RespondentData

Don't need the second callout box.

Get rid of absorb and offset args here.

A call to `lmitt()` in the context of the Work, Family and Health Study (WFHS)[1]

```
# a call to lmitt() for a difference-in-means estimate with covariance adjustment

tmod <- lmitt(
   TOPPRIORITY ~ 1,
   specification = StudySpec,
   data = RespondentData,
   absorb = FALSE,
   offset = cov_adj(logistic_mod),
   weights = "ate",
   ...
)
```

`cov_adj()` elicits predictions from a fitted prognostic model and stashes additional information for SE calculations in the output vector.

A call to `lmitt()` in the context of the Work, Family and Health Study (WFHS)[1]

```
# a call to lmitt() for a difference-in-means estimate with covariance adjustment

tmod <- lmitt(
   TOPPRIORITY ~ 1,
   specification = StudySpec,
   data = RespondentData,
   absorb = FALSE,
   offset = cov_adj(logistic_mod),
   weights = "ate",
   ...
)
```

Remove absorb arg

`cov_adj()` elicits predictions from a fitted prognostic model and stashes additional information for SE calculations in the output vector.

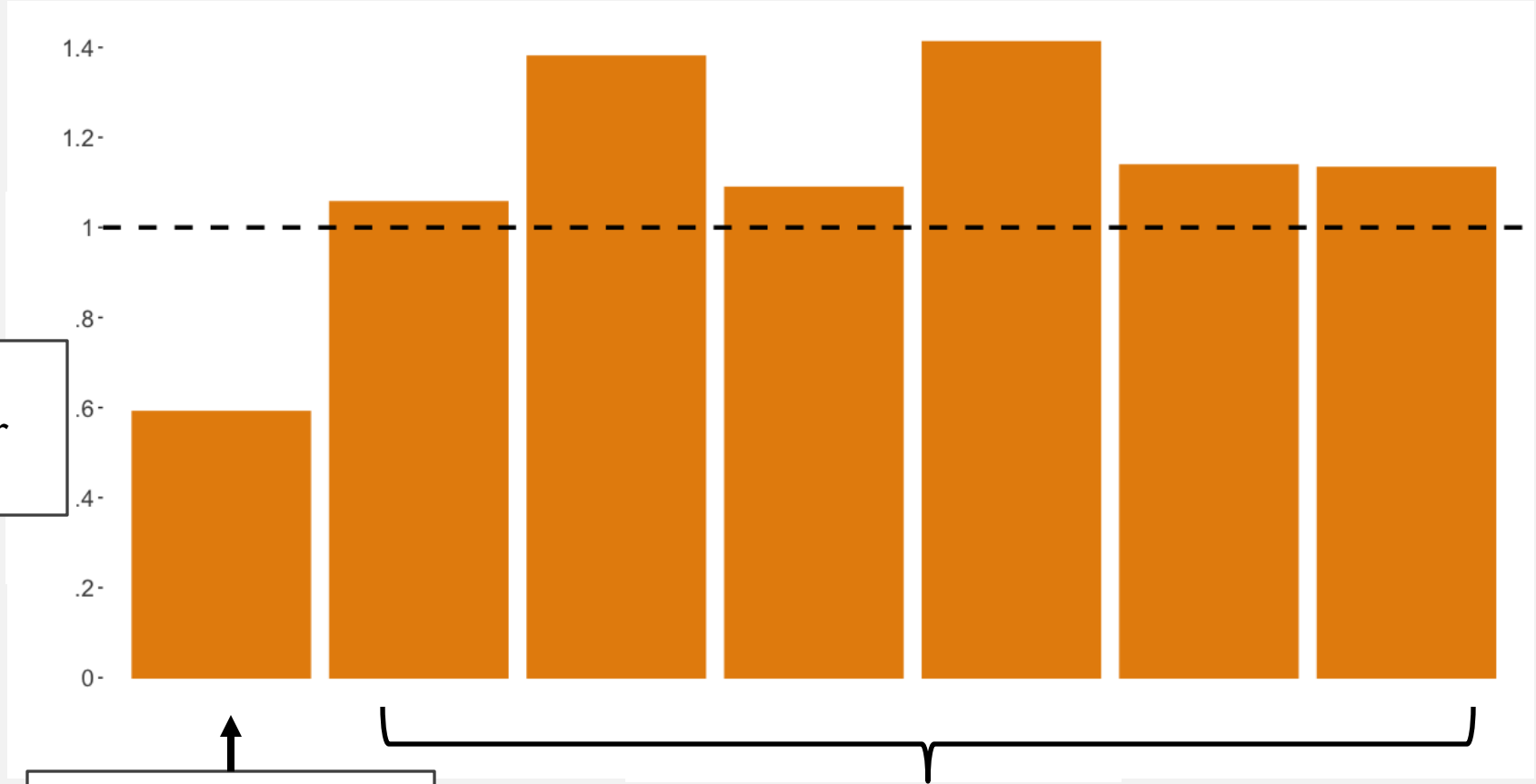
Passing prognostic scores to `offset` yields a (likely more precise) difference-in–differences type estimate, with differences of the form $Y(A) - \widehat{Y}(0)$

"model predictions" in place of "prognostic scores"

cov_adj() is an analog of predict(fitted_model, type = "response")

Why should SE's incorporate information from the prognostic score model?

"prognostic model"



Estimated SE / True SE of linear approximation

Two-sample t-test SE

propertee SE's with different combinations of CR0, CR1, and CR2 corrections to residuals

Example output of `summary()` called on a model fit with `lmitt()`

```
> summary(tmod, vcov.type = "CR2", cov_adj_rcorrect = "HC1", loco_residuals = TRUE)
Call:
lmitt(TOPPRIORITY ~ 1, StudySpec, RespondentData, offset = cov_adj(logistic_mod), weights = "ate")

 Treatment Effects :
                              Estimate Std. Error t value Pr(>|t|)
CONDITION.                    -0.05699    0.05499  -1.036     0.31
TOPPRIORITY:(Intercept)  0.63360    0.04209  15.052 2.24e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Std. Error calculated via type "CR2"
```

Title: "Better SE's via lmitt(), summary() (and vcov_tee())"

Example output of `summary()` called on a model fit with `lmitt()`

```
> summary(tmod, vcov.type = "CR2", cov_adj_rcorrect = "HC1", loco_residuals = TRUE)
Call:
lmitt(TOPPRIORITY ~ 1, StudySpec, RespondentData, offset = cov_adj(logistic_mod), weights = "ate")

 Treatment Effects :
                            Estimate Std. Error t value Pr(>|t|)
CONDITION.                  -0.05699    0.05499  -1.036     0.31
TOPPRIORITY:(Intercept)  0.63360    0.04209  15.052 2.24e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Std. Error calculated via type "CR2"
```

Degrees of freedom for reference t-distribution reflect clustering and known improvements for CR2 variance estimates

Update the output here to reflect covariance adjustment

These are the arguments that provide the best model-based SE based on our work (Wasserman, 2025+). Design-based SE's are available in some circumstances as well.

Just get rid of this slide

Find propertee on CRAN soon and, until then, Github!

REFERENCES

1. Work, Family and Health Network. Work, Family, and Health Study (WFHS) . Inter-university Consortium for Political and Social Research [distributor], 2018-10-03. https://doi.org/10.3886/ICPSR36158.v2

Add citation to my (in progress) dissertation, Github URL, Wang and Hansen (arXiv)

Wrap up with some other things propertee can do that users can look into

Add note about github here, pointing to QR