PRC Methods Presentation - March 12, 2019

- Note: This script is modeled off of an example created by Joshua Rosenberg, Ran Xu, and Ken Frank dated January 20, 2019.

In social research, we rarely have access to complete populations. Instead, we rely on samples to generate estimates and draw inferences regarding the entire population. Because we are typically always working with samples, how confident can we be that the our inferences aren't the result of sampling bias? If the phenomenon we are interested in present within our sample, how confident are we that we didn't just luck out by selecting the cases where this happens to be true? If our sample was slightly diffferent, could this dramatically change our conclusions?

In other words, how much bias must be present in our sample in order to invalidate an inference? This script walks through a new technique developed by Kenneth Frank and colleagues (2013) to provide insights onto this question.

Using Rubin's causal model, this technique calculates the percent of cases that would have to be null cases in order for a significant association to go away. Similar to p-values, there is not set threshold that the results need to cross to demonstrate that bias is not an issue. Instead, it is up to researchers to interpret how much bias they are willing to be okay with – but higher the value the less researchers should be concerned with sampling bias.

```
knitr::opts_knit$set(root.dir = '/Users/krisvelasco/Documents/UT-Austin/Admin/PRC Consultants/March 12 Presentati
on')
```

```
# Clear the workspace and do a gabarbage collection
rm(list=ls())
gc()
```

```
##          used (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
## Ncells 450979 24.1     973217   52         NA   630588 33.7
## Vcells 877053  6.7    8388608   64      16384  1767323 13.5
```

```
# Importing a Stata dataset
library(haven)

lgbti <- read_dta("/Users/krisvelasco/Documents/UT-Austin/Dissertation/Datasets/Master/dissertation_2018only.dta"
)
```

For this example, I will be doing a simple linear regression model (but the tools presented are also available for other types of models like general linear models, mixed models, etc.)

Let's say I am wanting evaluate how ties to transnational advocacy networks predict LGBTI policies, while controlling for % of women in parliament, % internet users, logged population density, trade as % of GDP, measurement of liberal democracy, and logged GDP per capita (note: all variables are lagged one year).

- FYI: lgbti_eigen = eigenvector centrality within transnational LGBTI advocacy networks

```
model1 <- lm(progressive_index ~ lgbti_eigen +
             lag_women_parliament +
             lag_internet_users +
             lag_population_density_ln +
             lag_trade  +
             lag_libdem_vdem +
             lag_gdppercap_ln,
              data = lgbti)

# Now let's look at the output from this model
summary(model1)
```

```
## 
## Call:
## lm(formula = progressive_index ~ lgbti_eigen + lag_women_parliament +
##     lag_internet_users + lag_population_density_ln + lag_trade +
##     lag_libdem_vdem + lag_gdppercap_ln, data = lgbti)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1885 -1.2206 -0.0833  1.2127  6.1424
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.395703   1.864838  -0.748 0.455412
## lgbti_eigen                4.779034   0.974406   4.905 2.48e-06 ***
## lag_women_parliament       0.052624   0.014590   3.607 0.000426 ***
## lag_internet_users         0.027653   0.013317   2.077 0.039610 *
## lag_population_density_ln -0.332825   0.130338  -2.554 0.011696 *
## lag_trade                 -0.006287   0.003689  -1.704 0.090466 .
## lag_libdem_vdem            3.785916   0.833488   4.542 1.16e-05 ***
## lag_gdppercap_ln           0.121869   0.267710   0.455 0.649627
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.944 on 145 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6914, Adjusted R-squared:  0.6765
## F-statistic:  46.4 on 7 and 145 DF,  p-value: < 2.2e-16
```

Question: How many countries would have to have no association between network centrality and progressive policies to invalidate this inference?

In other words, how biased would our sample have to be in order to invalidate this significant association? To answer that question, we are going to use the new Konfound tool developed by Frank and colleagues.

```
library(konfound)
```

```
## Warning: package 'konfound' was built under R version 3.5.2
```

```
## Sensitivity analysis as described in Frank, Maroulis, Duong, and Kelcey (2013) and in Frank (2000).
## For more information visit http://konfound-it.com.
```

We are now going to test the bias in our regression model using the following command: konfound(model1, lgbti_eigen)

- model1 is the object holding the results from our linear regression
- lgbti_eigen is the predictor we are trying to invalidate

```
konfound(model1, lgbti_eigen)
```

```
## Note that this output is calculated based on the correlation-based approach used in mkonfound()
```

```
## Percent Bias Necessary to Invalidate the Inference:
## To invalidate an inference, 59.72% of the estimate would have to be due to bias. This is based on a threshold
of 1.925 for statistical significance (alpha = 0.05).
## To invalidate an inference, 91 observations would have to be replaced with cases for which the effect is 0.
## 
## Impact Threshold for a Confounding Variable:
## An omitted variable would have to be correlated at 0.508 with the outcome and at 0.508 with the predictor of i
nterest (conditioning on observed covariates) to invalidate an inference based on a threshold of 0.163 for statis
tical significance (alpha = 0.05).
## Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be 0.508 X 0.508 = 0.258 to
invalidate an inference.
```

```
## For more detailed output, consider setting `to_return` to table
```

```
## To consider other predictors of interest, consider setting `test_all` to TRUE.
```

59.72% of the estimate would have to be due to bias to invalidate the inference

From the output, we can see that in order to invalidate the inference, 59.72% of the estimate would have to be due to bias – or 91 observations would have to be replaced with cases in which the effect of centrality on progressive policies is 0.

Further, the output tells us the extent to which a confounding variable could nullify our results. In this example, an omitted variable would have to be correlated at .508 with the outcome and predictor variable to invalidate the inference.

Visualizing Output

```
konfound_output <- konfound(model1, lgbti_eigen, to_return = c("raw_output", "thresh_plot", "corr_plot"))
```

```
## Note that this output is calculated based on the correlation-based approach used in mkonfound()
```

```
## Percent Bias Necessary to Invalidate the Inference:
## To invalidate an inference, 59.72% of the estimate would have to be due to bias. This is based on a threshold
of 1.925 for statistical significance (alpha = 0.05).
## To invalidate an inference, 91 observations would have to be replaced with cases for which the effect is 0.
##
## Impact Threshold for a Confounding Variable:
## An omitted variable would have to be correlated at 0.508 with the outcome and at 0.508 with the predictor of i
nterest (conditioning on observed covariates) to invalidate an inference based on a threshold of 0.163 for statis
tical significance (alpha = 0.05).
## Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be 0.508 X 0.508 = 0.258 to
invalidate an inference.
```

```
## Print output created by default. Created 3 other forms of output. Use list indexing or run summary() on the ou
tput to see how to access.
```

```
konfound_output$thresh_plot
```
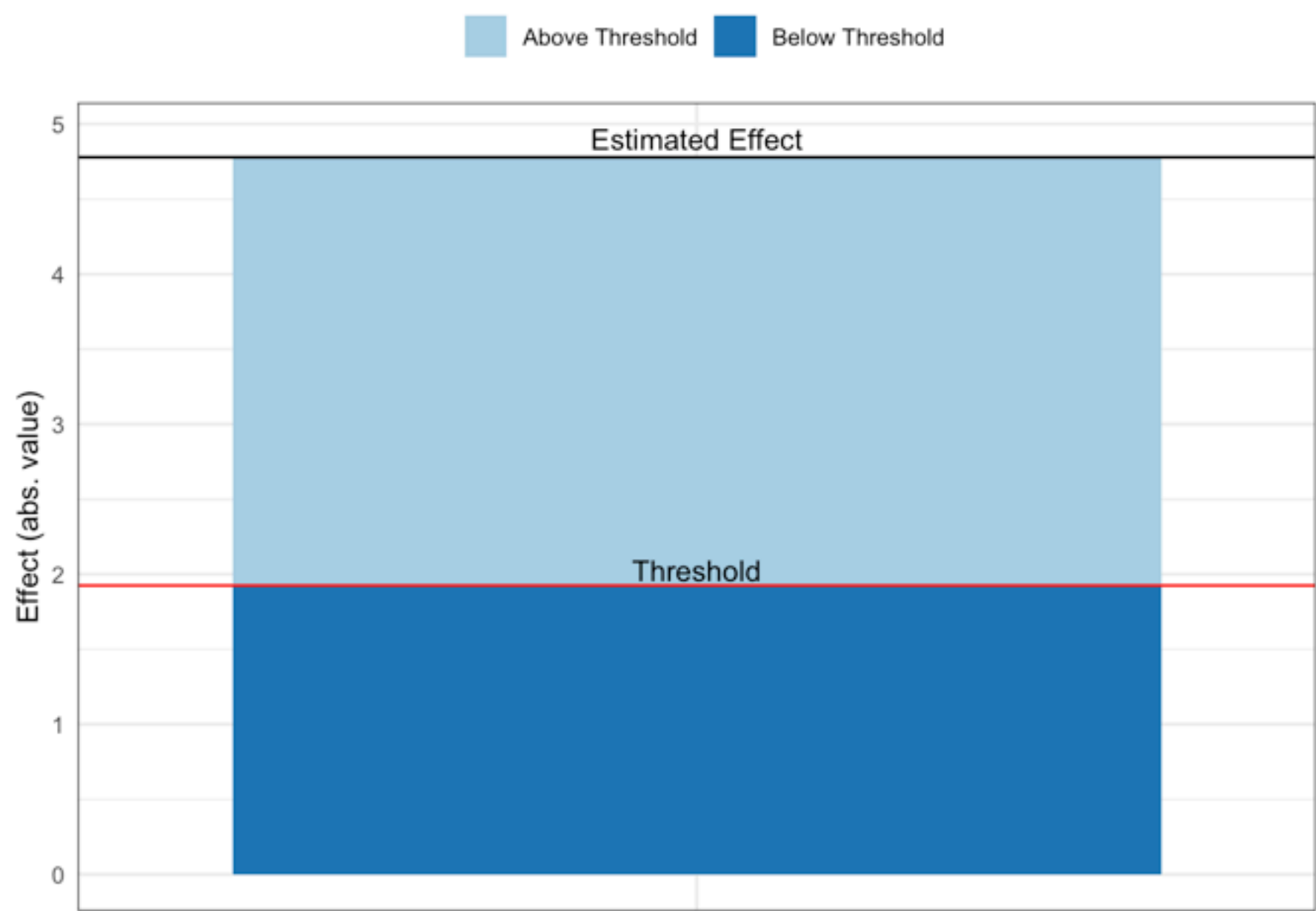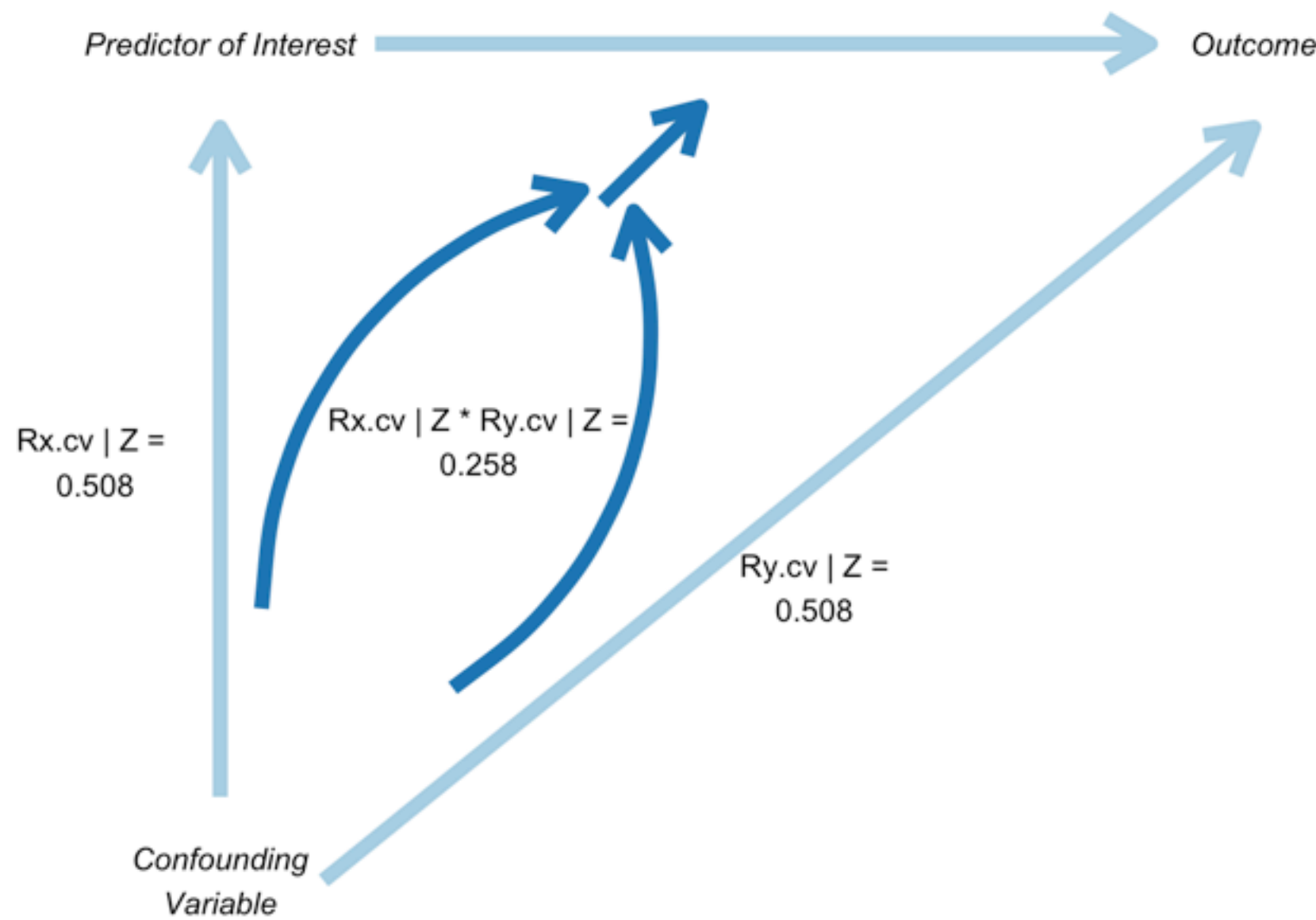

The region above the threshold represents the percent of the estimated coefficient that would have to be replaced with null cases in order to invalidate the significant effect.

The next plot showcases how much an omitted variable would have to correlate with both our predictor of interest and the outcome variable in order to invalidate the inference.

```
konfound_output$corr_plot
```

To invalidate an inference

Predictor of Interest ──────────────────► Outcome

Rx.cv | Z = 0.508

Rx.cv | Z * Ry.cv | Z = 0.258

Ry.cv | Z = 0.508

Confounding Variable

Testing bias when you don't have access to the data.

Let's say you are reading an article and are curious how much bias is factoring into the significant associations. This package also includes a tool to test the bias in coefficients using key pieces of information like standard errors, sample sizes, etc.

```
pkonfound(est_eff = -2.2,
          std_err = .65,
          n_obs = 200,
          n_covariates = 3)
```

```
## Percent Bias Necessary to Invalidate the Inference:
## To invalidate an inference, 41.732% of the estimate would have to be due to bias. This is based on a threshold
of -1.282 for statistical significance (alpha = 0.05).
## To invalidate an inference, 83 observations would have to be replaced with cases for which the effect is 0.
##
## Impact Threshold for a Confounding Variable:
## An omitted variable would have to be correlated at 0.334 with the outcome and at 0.334 with the predictor of i
nterest (conditioning on observed covariates) to invalidate an inference based on a threshold of -0.14 for statis
tical significance (alpha = 0.05).
## Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be 0.334 X 0.334 = 0.112 to
invalidate an inference.
```

```
## For other forms of output, change `to_return` to table, raw_output, thres_plot, or corr_plot.
```

```
## For models fit in R, consider use of konfound().
```

This can also be re-written as the following:

```
pkonfound(-2.2, .65, 200, 3)
```

```
## Percent Bias Necessary to Invalidate the Inference:
## To invalidate an inference, 41.732% of the estimate would have to be due to bias. This is based on a threshold
of -1.282 for statistical significance (alpha = 0.05).
## To invalidate an inference, 83 observations would have to be replaced with cases for which the effect is 0.
##
## Impact Threshold for a Confounding Variable:
## An omitted variable would have to be correlated at 0.334 with the outcome and at 0.334 with the predictor of i
nterest (conditioning on observed covariates) to invalidate an inference based on a threshold of -0.14 for statis
tical significance (alpha = 0.05).
## Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be 0.334 X 0.334 = 0.112 to
invalidate an inference.
```

```
## For other forms of output, change `to_return` to table, raw_output, thres_plot, or corr_plot.
```

```
## For models fit in R, consider use of konfound().
```

How to learn more about sensitivity analysis

To learn more about sensitivity analysis, please visit:

- The Introduction to konfound vignette (https://jrosen48.github.io/konfound/articles/Introduction_to_konfound.html), with detailed information about each of the functions (`pkonfound()`, `konfound()`, and `mkounfound()`)
- The causal inference section of Ken Frank's website here (https://msu.edu/~kenfrank/research.htm#causal)
- The konfound interactive web application (https://jmichaelrosenberg.shinyapps.io/shinykonfound/), with links to PowerPoints and key publications

References

- Frank, K.A., Maroulis, S., Duong, M., and Kelcey, B. 2013. What would it take to change an inference?: Using Rubin's causal model to interpret the robustness of causal inferences. *Education, Evaluation and Policy Analysis*. Vol 35: 437-460. https://msu.edu/~kenfrank/What%20would%20it%20take%20to%20Change%20an%20Inference%20published.docx (https://msu.edu/~kenfrank/What%20would%20it%20take%20to%20Change%20an%20Inference%20published.docx)

- Frank, K.A., Gary Sykes, Dorothea Anagnostopoulos, Marisa Cannata, Linda Chard, Ann Krause, Raven McCrory. 2008. Extended influence: National Board Certified Teachers as help providers. *Education, Evaluation, and Policy Analysis*. Vol 30(1): 3-30. https://msu.edu/~kenfrank/papers/Does%20NBPTS%20Certification%20Affect%20the%20Number%20of%20Colleagues%20a%20Teacher%20Helps%20wi (https://msu.edu/~kenfrank/papers/Does%20NBPTS%20Certification%20Affect%20the%20Number%20of%20Colleagues%20a%20Teacher%20Helps%20w

- Frank, K. A. and Min, K. 2007. Indices of Robustness for Sample Representation. *Sociological Methodology*. Vol 37, 349-392. https://msu.edu/~kenfrank/papers/INDICES%20OF%20ROBUSTNESS%20TO%20CONCERNS%20REGARDING%20THE%20REPRESENTATIVENESS%20O (https://msu.edu/~kenfrank/papers/INDICES%20OF%20ROBUSTNESS%20TO%20CONCERNS%20REGARDING%20THE%20REPRESENTATIVENESS%20O (co first authors)

- Frank, K. 2000. "Impact of a Confounding Variable on the Inference of a Regression Coefficient." *Sociological Methods and Research*, 29(2), 147-194 https://msu.edu/~kenfrank/papers/impact%20of%20a%20confounding%20variable.pdf (https://msu.edu/~kenfrank/papers/impact%20of%20a%20confounding%20variable.pdf)