OVB, Sensitivity Analysis, and FWL

Nicholas Ray

2024-11-25

OVB Review

In scalar notation:

$$
\begin{aligned}
Y &= \beta_{0L} + \beta_{1L}X + \gamma Z + e \\
Y &= \beta_{0S} + \beta_{1S}X + u
\end{aligned}
$$

$$
\begin{aligned}
\hat{\beta}_{1S} &= \frac{\text{Cov}[X,Y]}{\text{V}[X]} \\
&= \frac{\text{Cov}[X,(\beta_{0L} + \beta_{1L}X + \gamma Z + e)]}{\text{V}[X]} \\
&= \frac{\text{Cov}[X,\beta_{0L}]}{\text{V}[X]} + \frac{\text{Cov}[X,\beta_{1L}X]}{\text{V}[X]} + \frac{\text{Cov}[X,\gamma Z]}{\text{V}[X]} + \frac{\text{Cov}[X,e]}{\text{V}[X]} \\
&= 0 + \beta_{1L}\frac{\text{Cov}[X,X]}{\text{V}[X]} + \gamma\frac{\text{Cov}[X,Z]}{\text{V}[X]} + 0 \\
&= \beta_{1L} + \gamma\frac{\text{Cov}[X,Z]}{\text{V}[X]} \\
&= \beta_{1L} + \gamma\delta
\end{aligned}
$$

In matrix notation:

$$
\begin{aligned}
Y &= X\beta_L + Z\gamma + e \\
Y &= X\beta_S + u
\end{aligned}
$$

$$
\begin{aligned}
\hat{\beta}_S &= (X'X)^{-1}X'Y \\
&= (X'X)^{-1}X'(X\beta_L + Z\gamma + e) \\
&= (X'X)^{-1}X'X\beta_L + (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'e \\
&= \beta_L + \gamma(X'X)^{-1}X'Z + 0 \\
&= \beta_L + \gamma\delta
\end{aligned}
$$

Sensitivity Analysis

The function `ovb` primarily estimates the average estimated bias in $\hat{\beta}_1$ when the true DGP is $Y = \beta_{0L} + \beta_{1L}X + \gamma Z + e$ but the estimated model is $Y = \beta_{0S} + \beta_{1S}X + u$ (when `return` is set to `"ovb"`). Thus, Z is omitted. We can easily verify that the OVB approximates $\gamma \times \delta$ from our equations above.

The function `ovb` instead returns the data used in the last iteration or simulation conducted when `return` is set to anything else (below, to `"data"`). We can use this toy dataset to get started with sensitivity analysis ala Cinelli and Hazlett (2020) under known conditions.

```
#simulation####################################################################
ovb<-function(obs,sims,b1,gamma,delta,return){
  set.seed(1)
  b0=0
  correlation_matrix=matrix(c(1,delta,
                              delta,1),
                            ncol=2,byrow=T)
  #mvrnorm() is useful for generating random variables that are correlated
  covariates<-data.frame(MASS::mvrnorm(obs,c(0,0),correlation_matrix,
                                       empirical=T))
  data<-data.frame(X=covariates[,1],Z=covariates[,2])
  results<-list()
```

```r
  for (j in 1:sims) {
    for (i in 1:obs) {
      data$Y=b0+b1*data$X+gamma*data$Z+rnorm(sims)
    }
    results[[j]]<-lm(Y~X,data)$coefficients[2]-b1
  }
  if(return=="ovb"){
    return(mean(unlist(results)))
  } else{
    return(data)
  }
}
#theoretical ovb is \gamma times \delta
ovb(100,100,b1=1,gamma=1,delta=0.75,"ovb")
```

```
## [1] 0.7690668
```

```r
data<-ovb(100,100,b1=1,gamma=1,delta=0.75,"data") #data from last simulation
#note that you can make gamma (or delta) negative, just add "reduce=F" to sensemakr() function below

#sensitivity analysis##############################################################
##partial R^2 of Z################################################################
library(asbio)
```

```
## Warning: package 'asbio' was built under R version 4.3.3

## Loading required package: tcltk
```

```r
#partial R^2 of Z on outcome:
partial.R2(lm(Y~X,data), #model without Z
           lm(Y~X+Z,data))  #model with Z
```

```
## [1] 0.340535
```

```r
#partial R^2 of Z on treatment (X):
partial.R2(lm(X~1,data), #model without Z (intercept only)
           lm(X~Z,data)) #model with Z
```

```
## [1] 0.5625
```

```r
#keep these in mind when looking at the plots below

library(sensemakr)
```

```
## Warning: package 'sensemakr' was built under R version 4.3.3

## See details in:

## Carlos Cinelli and Chad Hazlett (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias. Journ
```

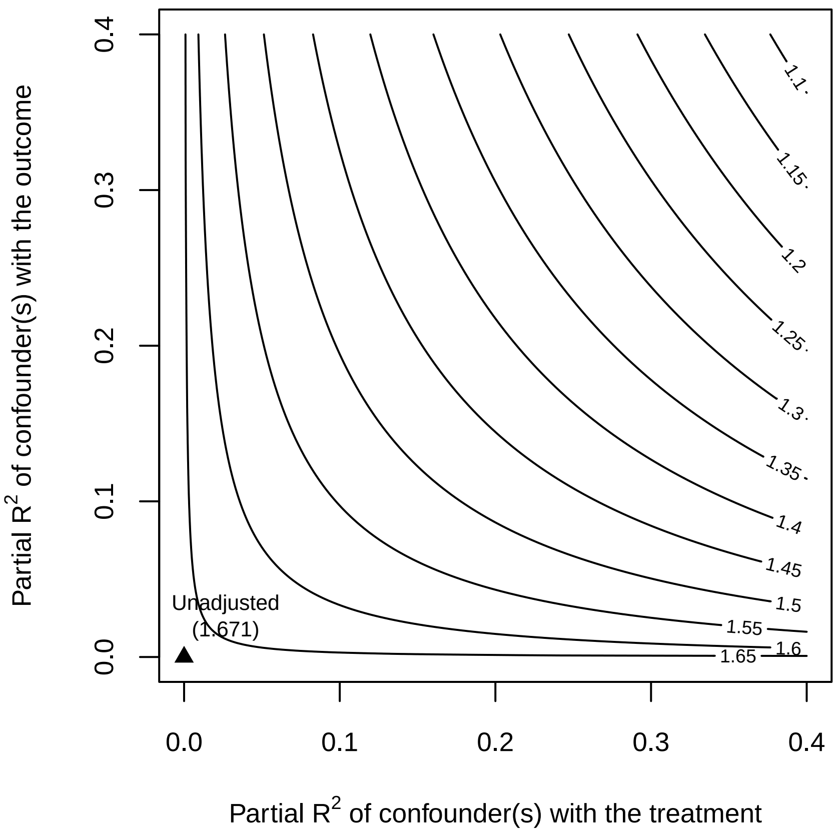
```r
#equivalently (from `sensemakr` package)
sensemakr::partial_r2(lm(Y~X+Z,data)) #partial R^2 of Z on outcome
```

```
## (Intercept)           X           Z
## 0.001351596 0.286198278 0.340535034
```

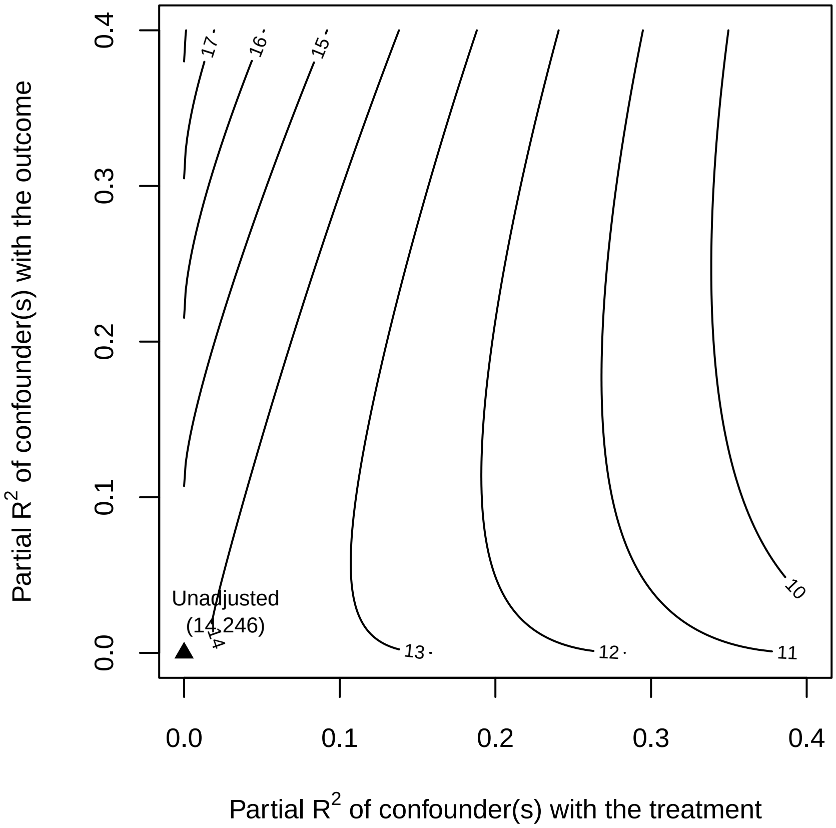
```r
sensemakr::partial_r2(lm(X~Z,data)) #partial R^2 of Z on treatment (X)
```

```
## (Intercept)           Z
##      0.0000      0.5625
```

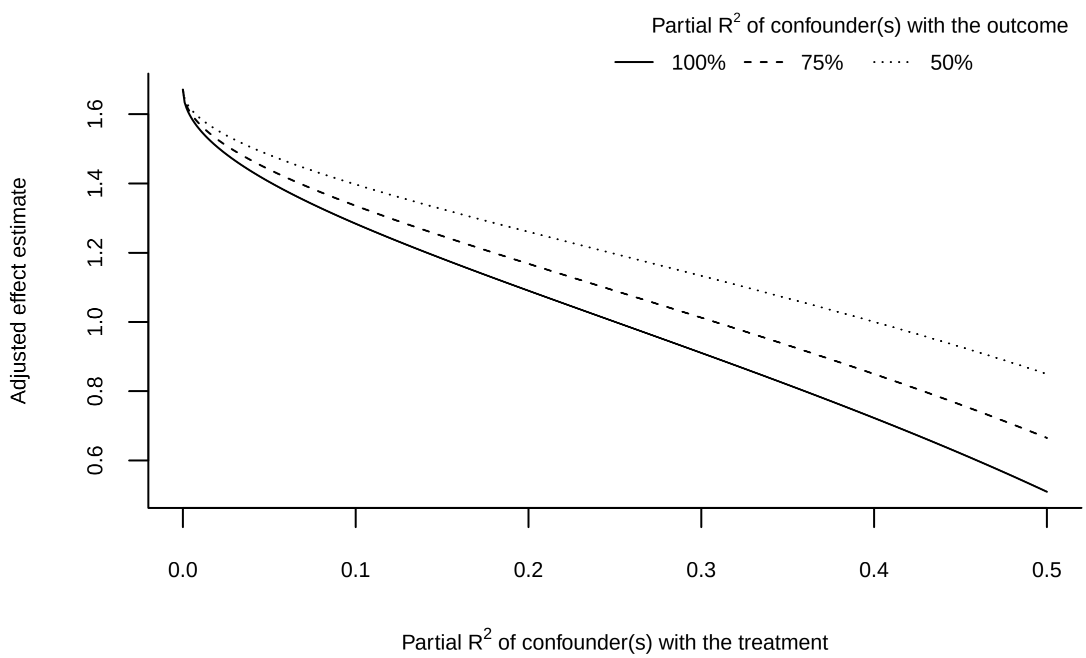
```r
##using sensemakr()###############################################################]
sensitivity<-sensemakr(lm(Y~X,data),treatment="X") #model with known OVB

#given what we know about partial R^2 of omitted variable (Z), below plot implies X would go to about one (tru
plot(sensitivity) #partial R^2 rescaling of \gamma on Y axis and partial R^2 rescaling of \delta on X axis (an
```




```r
plot(sensitivity,sensitivity.of="t-value") #X would still remain highly statistically different from 0
```



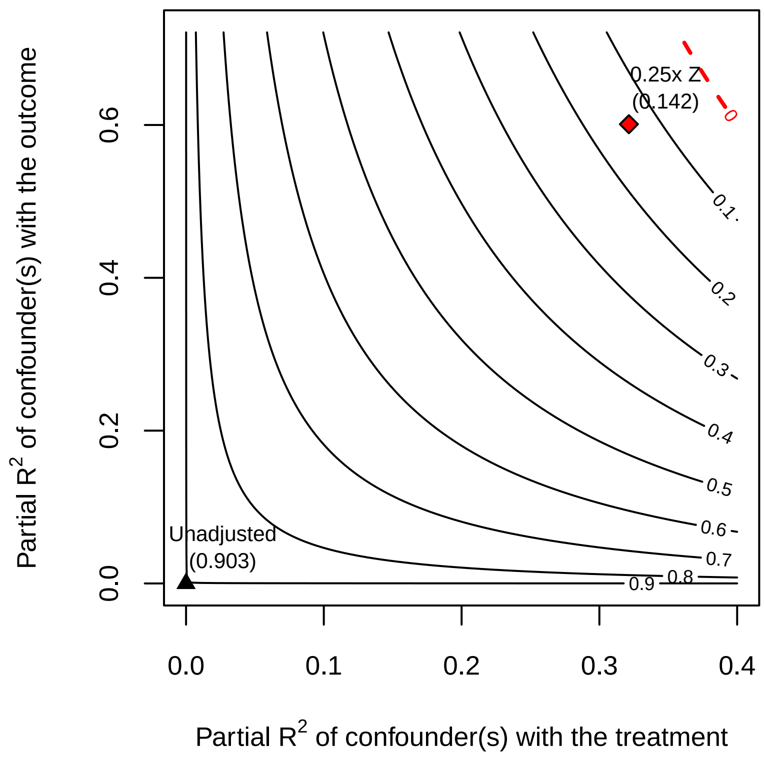

```
#again keep in mind our known confounder has a partial R^2 of 0.340535 on the outcome and 0.5625 on the treatm
plot(sensitivity,type="extreme")
```



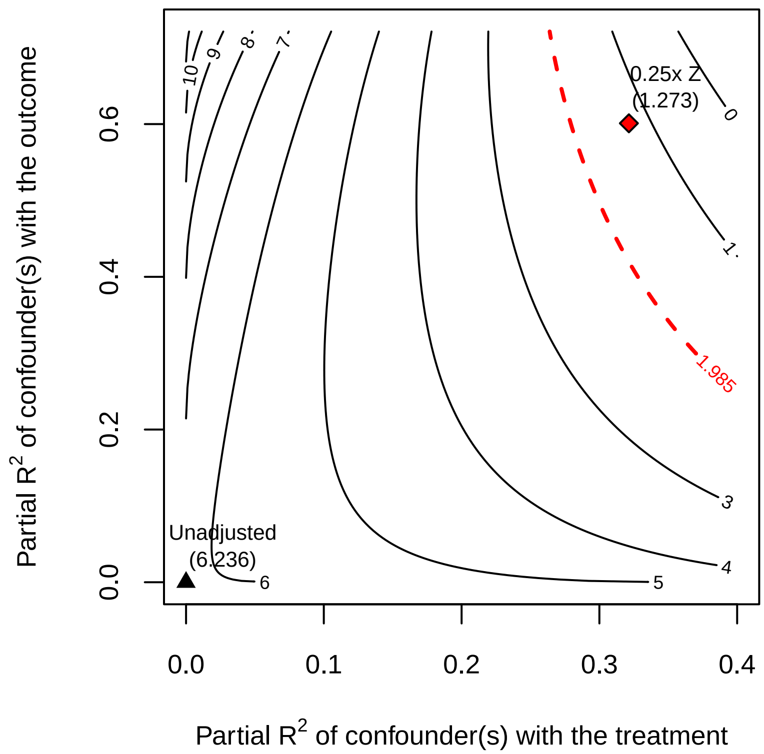
```r
summary(sensitivity) #very useful descriptions
```

```
## Sensitivity Analysis to Unobserved Confounding
##
## Model Formula: Y ~ X
##
## Null hypothesis: q = 1 and reduce = TRUE
## -- This means we are considering biases that reduce the absolute value of the current estimate.
## -- The null hypothesis deemed problematic is H0:tau = 0
##
## Unadjusted Estimates of 'X':
##   Coef. estimate: 1.6707
##   Standard Error: 0.1173
##   t-value (H0:tau = 0): 14.2464
##
## Sensitivity Statistics:
##   Partial R2 of treatment with outcome: 0.6744
##   Robustness Value, q = 1: 0.7374
##   Robustness Value, q = 1, alpha = 0.05: 0.6896
##
## Verbal interpretation of sensitivity statistics:
##
## -- Partial R2 of the treatment with the outcome: an extreme confounder (orthogonal to the covariates) that 
##
## -- Robustness Value, q = 1: unobserved confounders (orthogonal to the covariates) that explain more than 73
##
## -- Robustness Value, q = 1, alpha = 0.05: unobserved confounders (orthogonal to the covariates) that explai
```

```r
##senemakr() with benchmark#####################################################
#what if there was a third confounder, benchmarked to Z?
sensitivity2<-sensemakr(lm(Y~X+Z,data),treatment="X",benchmark_covariates="Z",kd=0.25) #had to set kd to be lo
#feel free to vary the variance of Y

#a third OV that was a quarter the "strength" of Z would have a partial R^2 of about 0.35 with the treatment a
#thus, a third OV that was about a third of the strength of Z would drive X to 0
plot(sensitivity2)
```




```r
#even easier to make X statistically indistinguishable from 0
plot(sensitivity2,sensitivity.of="t-value")
```



```r
summary(lm(Y~X+Z,data))
```

```
##
## Call:
## lm(formula = Y ~ X + Z, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2835 -0.7095  0.0185  0.8655  2.2552
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03451    0.09524  -0.362    0.718
## X            0.90252    0.14472   6.236 1.17e-08 ***
## Z            1.02422    0.14472   7.077 2.31e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9524 on 97 degrees of freedom
## Multiple R-squared:  0.7853, Adjusted R-squared:  0.7808
## F-statistic: 177.4 on 2 and 97 DF,  p-value: < 2.2e-16
```

Applied Example Using Martin et al. (2024)

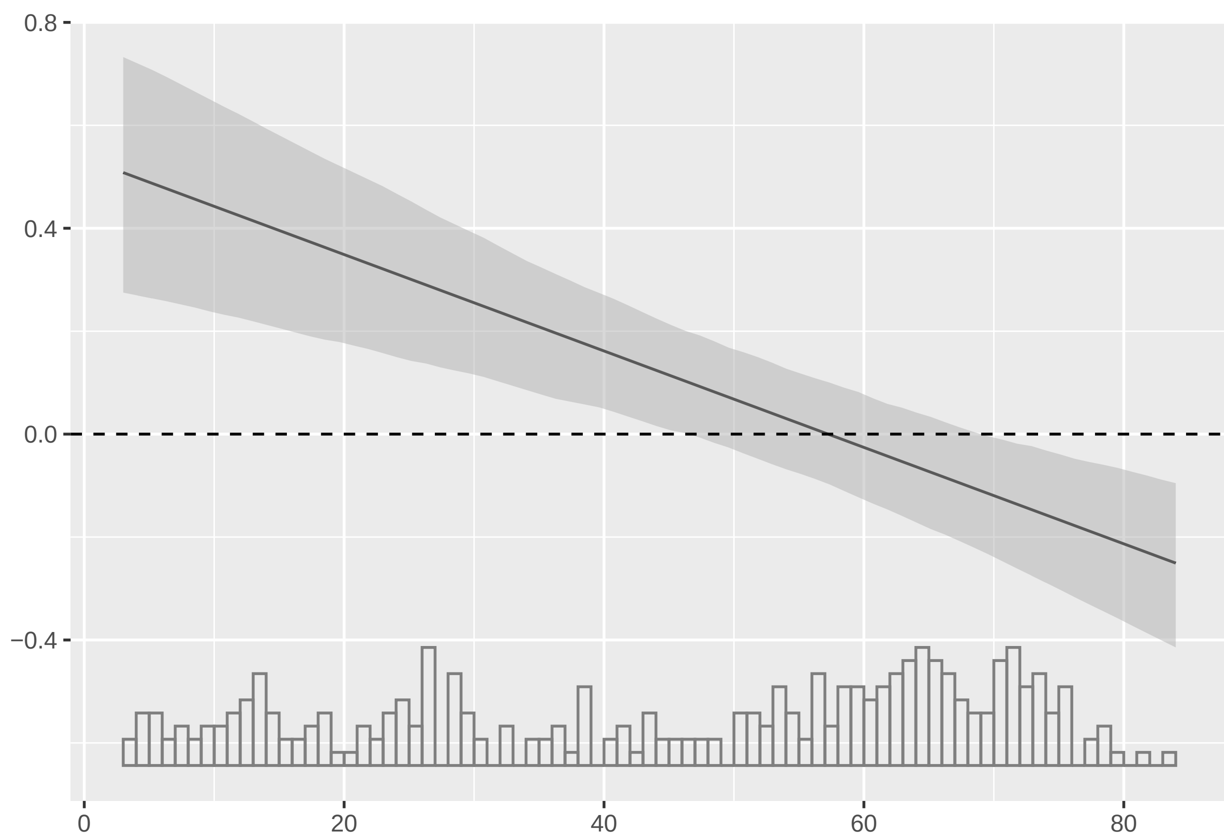
The authors are interested in explaining variation in "incumbency rates" across the top 70 or so most populous democracies (see Martin, McClean, and Strom 2024).

```r
#data##########################################################################
#url from our public DropBox (note, have to set dl=1 instead of dl=0 at the end)
url<-"https://www.dropbox.com/scl/fi/tofgmxv4neuu0qweewkuv/martin_etal_data.csv?rlkey=roy6wsju5jqxdqzpyh34agzs
martin<-read.csv("https://www.dropbox.com/scl/fi/tofgmxv4neuu0qweewkuv/martin_etal_data.csv?rlkey=roy6wsju5jqx

#main analysis and figure######################################################
#their basic OLS model
model<-lm(IncumbencyRate~LegOrg+Corruption+LegOrg*Corruption,martin)

#plotting the interaction term
library(interplot); library(tidyverse)
```

```
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.3

## Loading required package: abind

## Loading required package: arm

## Loading required package: MASS

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.3.2

## Loading required package: lme4

## Warning: package 'lme4' was built under R version 4.3.2
```

```
##
## arm (Version 1.13-1, built: 2022-8-25)

## Working directory is C:/Users/nicho/OneDrive/Documents/Teaching/602_F_24/11_25_OVB_Sensitivity_Analysis_and

## Warning: package 'lubridate' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x lubridate::pm() masks asbio::pm()
## x dplyr::select() masks MASS::select()
## x tidyr::unpack() masks Matrix::unpack()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
interplot(model,"LegOrg","Corruption",hist=T)+
  geom_hline(yintercept=0,linetype="dashed")
```

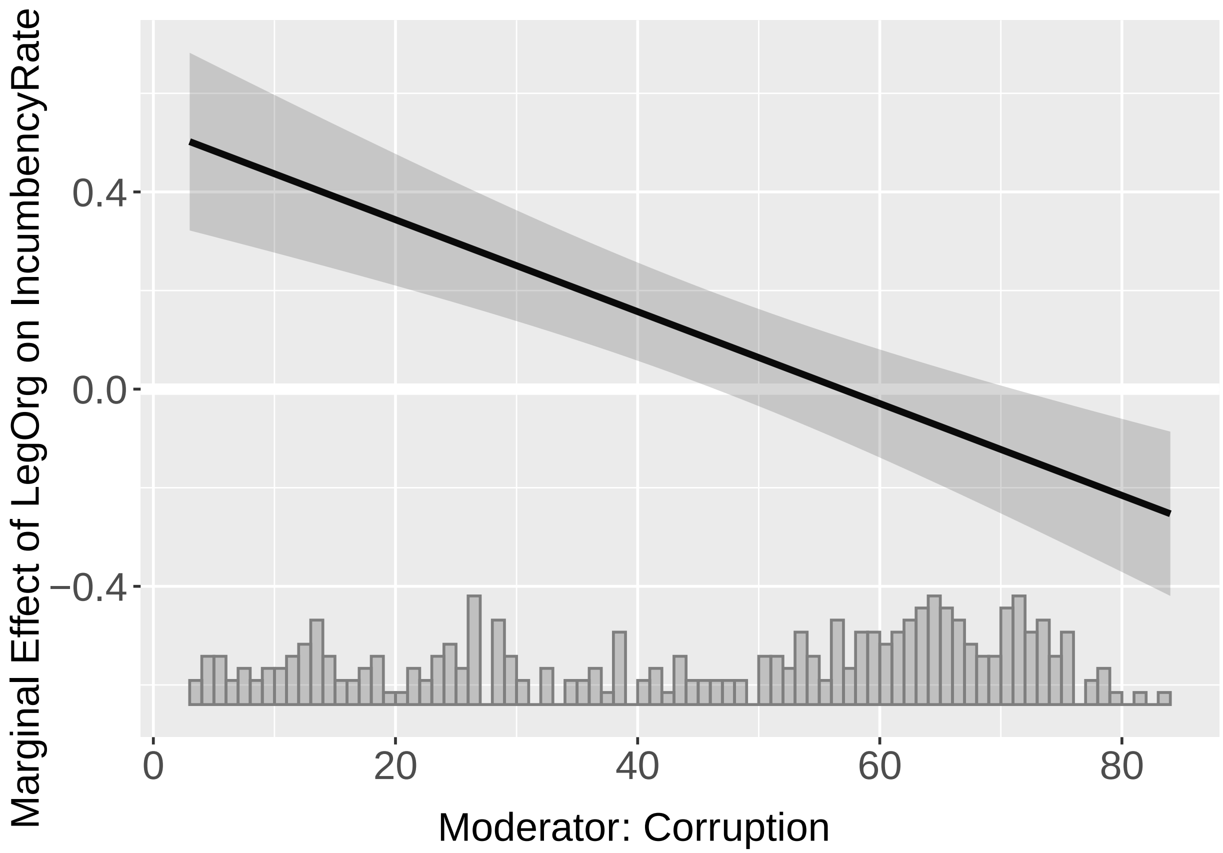



```r
#or
library(interflex)
```

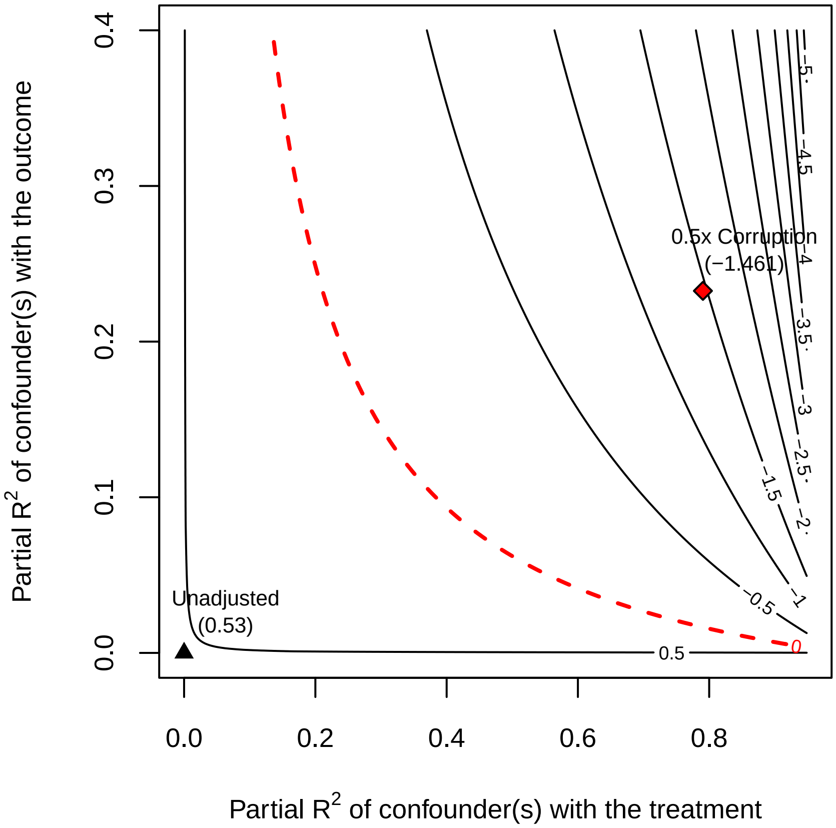
```
## Warning: package 'interflex' was built under R version 4.3.3
```

```
## ## Syntax has changed since v.1.2.1.
##
## ## See http://bit.ly/interflex for more info.
## ## Comments and suggestions -> zyliu2020@uchicago.edu.
```

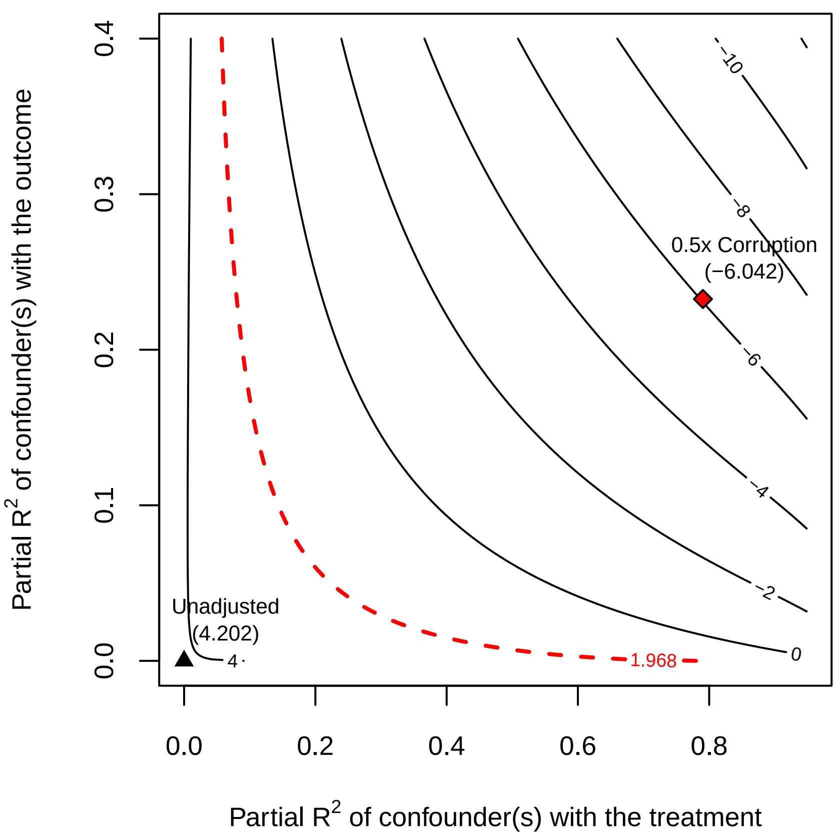
```
plot(interflex("linear","IncumbencyRate","LegOrg","Corruption",data=martin))
```




```
#basic sensitivity analysis#####################################################
sensitivity<-sensemakr(model,"LegOrg","Corruption",kd=0.5)
plot(sensitivity)
```




```r
plot(sensitivity,sensitivity.of="t-value")
```



```r
summary(sensitivity)
```

```
## Sensitivity Analysis to Unobserved Confounding
##
## Model Formula: IncumbencyRate ~ LegOrg + Corruption + LegOrg * Corruption
##
## Null hypothesis: q = 1 and reduce = TRUE
## -- This means we are considering biases that reduce the absolute value of the current estimate.
## -- The null hypothesis deemed problematic is H0:tau = 0
##
## Unadjusted Estimates of 'LegOrg':
##   Coef. estimate: 0.53
##   Standard Error: 0.1261
##   t-value (H0:tau = 0): 4.2024
##
## Sensitivity Statistics:
##   Partial R2 of treatment with outcome: 0.0585
##   Robustness Value, q = 1: 0.2202
##   Robustness Value, q = 1, alpha = 0.05: 0.1239
##
## Verbal interpretation of sensitivity statistics:
##
## -- Partial R2 of the treatment with the outcome: an extreme confounder (orthogonal to the covariates) that 
##
## -- Robustness Value, q = 1: unobserved confounders (orthogonal to the covariates) that explain more than 22
##
## -- Robustness Value, q = 1, alpha = 0.05: unobserved confounders (orthogonal to the covariates) that explai
##
## Bounds on omitted variable bias:
##
## --The table below shows the maximum strength of unobserved confounders with association with the treatment 
##
##       Bound Label R2dz.x R2yz.dx Treatment Adjusted Estimate Adjusted Se
##  0.5x Corruption 0.7905  0.2326    LegOrg           -1.4611      0.2418
##  Adjusted T Adjusted Lower CI Adjusted Upper CI
##     -6.0424           -1.9371           -0.9852
```

The authors' basic model is $\text{IncumbencyRate}_{it} = \beta_0 + \beta_1\text{LegOrg}_{it} + \beta_2\text{Corruption}_{it} + \beta_3\text{LegOrg}_{it} \times \text{Corruption}_{it} + e_{it}$. What would be their conditional independence assumption (CIA) here, if they were estimating $\mathbb{E}[\tau]$ under the potential outcomes model?

Note that decide to use an instrument because they believe CIA to be violated. What is the instrument? What are the assumptions necessary for it to be a "valid" instrument? How does this instrument differ from how it was introduced in class?

You should read Lal et al. (2024) and keep in mind that people recommend using a "robust" F-stat to evaluate relevance or the "First-stage" (which implies a F-stat of at least 23.1) (see Montiel Olea and Pflueger 2013).

FWL Review

Let's return to our long regression in matrix form: $Y = X\beta_L + Z\gamma + e$.

FWL says that β_L can be recovered from the bivariate regression $\tilde{Y} = \tilde{X}\beta_S + u$, where $\tilde{X} = X - Z\gamma$ and $\tilde{Y} = Y - Z\gamma$. In other words, $\tilde{X}$ are the residuals from the bivariate regression $X = Z\gamma + \tilde{X}$ and $\tilde{Y}$ are the residuals from the bivariate regression $Y = Z\gamma + \tilde{Y}$.

```r
#full or "long" regressions
summary(lm(Y~X[,2]+Z))
```

```r
#\tilde{Y} regressed on \tilde{X}
summary(lm(lm(Y~Z)$residuals ~ lm(X[,2]~Z)$residuals))
```

Note that running the bivariate regression $Y = \tilde{X}\beta_S + w$ recovers β_L but gives the incorrect standard errors. You need "partial" out Z from both the covariate X and the outcome Y.

```r
#Y regressed on \tilde{X}
summary(lm(Y~lm(X[,2]~Z)$residuals))
```

We could also do this using our projection and annihilator matrices, P and M.

```r
#P and M in terms of only Z
P_Z<-Z%*%solve(t(Z)%*%Z)%*%t(Z)
M_Z<-diag(nrow(data))-P_Z

#\beta_L using P_Z and M_Z (see 5.1 slides on FWL/Hansen)
solve(t(X[,2])%*%M_Z%*%X[,2])%*%t(X[,2])%*%M_Z%*%Y

#can also regress \tilde{Y} on \tilde{X} like we did before
Y_tilde=M_Z%*%Y
X_tilde=M_Z%*%X[,2]
summary(lm(Y_tilde~X_tilde))

#equivalent to full or "long" models (absent some rounding error, I think)
summary(lm(Y~X[,2]+Z))
```

Two examples of table output:

```r
library(stargazer)
stargazer(model)
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Nov 25, 2024 - 12:17:58 PM

```r
library(huxtable)
```

```
##
## Attaching package: 'huxtable'

## The following object is masked from 'package:dplyr':
##
##     add_rownames

## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

```r
huxreg(model)
```

Bibliography

Cinelli, Carlos, and Chad Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Jounral of the Royal Statistical Society B.*

Lal, Apoorva, Mackenzie Lockhart, Yiqing Xu, and Ziwen Zu. 2024. "How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies." *Political Analysis.*

Martin, Shane, Charles T. McClean, and Kaare W. Strom. 2024. "Legislative Resources, Corruption, and Incumbency." *British Journal of Political Science.*

Montiel Olea, José Luis, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business and Economic Statistics.*

Table 1

	Dependent variable:
	IncumbencyRate
LegOrg	0.530***
	(0.126)
Corruption	−0.290***
	(0.100)
LegOrg:Corruption	−0.009***
	(0.002)
Constant	55.184***
	(5.379)
Observations	288
R^2	0.478
Adjusted R^2	0.473
Residual Std. Error	16.719 (df = 284)
F Statistic	86.696*** (df = 3; 284)
Note:	*p<0.1; **p<0.05; ***p<0.01

	(1)
(Intercept)	55.184 ***
	(5.379)
LegOrg	0.530 ***
	(0.126)
Corruption	-0.290 **
	(0.100)
LegOrg:Corruption	-0.009 ***
	(0.002)
N	288
R2	0.478
logLik	-1217.812
AIC	2445.625

*** p < 0.001; ** p < 0.01; * p < 0.05.