Effect of 401(k) Eligibility on Net Financial Assets : Reproducing results from Chernozhukov et al. [2018, 2022]

Anuradha Ramachandran

Section 1: Introduction

The data comes from the 1991 Survey of Income and Program Participation (SIPP), and has 9,915 observations. The main "treatment" variable is e401, which indicates whether the individual worked for a company that offers a 401(k) plan. The main outcome of interest is net tfa, standing for the net financial assets of the individual. Key covariates include age, inc (income), educ (years of education), fsize (family size), marr (1 if individual is married), twoearn (1 if a two-earner household), pira (1 if individual participates in IRA plan), and hown (1 if home owner).
Throughout this exercise, let:

- Y_i = net tfa (outcome)
- D_i = e401 (binary indicator of "treatment")
- X_i = a matrix with covariates age, inc, educ, fsize, marr, twoearn, pira, and hown.

Section 2: Reasearch question

A 401(k) plan is an employed sponsored tax-deferred savings option that allows individuals to deduct contributions from their taxable income, and accrue tax-free interest on investments within the plan. Introduced in the early 1980s as an incentive to increase individual savings for retirement, an important question in the savings literature is precisely to quantify the causal impact of 401(k) eligibility (Di) on net financial assets (Yi). Average treatment effect (ATE) of 401(k) eligibility on net financial assets can be expressed mathematically as,
Let the potential outcomes be:
$Y_i(0)$ = Potential net financial asset of an individual if they worked for a company that did not offer 401k plan.
$Y_i(1)$ = Potential net financial asset of an individual if they worked for a company that offered 401k plan.
Under the assumption that ignorability ($Y_i(d) \perp D_i = d$) and consistency holds, ATE can be interpreted causally.
Treatment effect of an individual is $\tau_i = Y_i(1) - Y_i(0)$
Average treatment effect is $E[\tau_i] = E[Y_i(1) - Y_i(0)] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$

Section 3: Naive Estimation

To begin our investigation, a natural point to start is to check whether those individuals with 401(k) eligibility save more than those without it. Since D_i is binary variable, CEF can be approximated by linear regression as $E[Y_i|D_i] = \alpha + \beta D_i + e$.
$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \beta =$ \$19559.34coefficient to treatment D_i in the regression. On average individuals with 401k eligibility have higher net financial assets (\$19559) than those without 401k eligibility.
$E[Y_i|D_i = 0] = \alpha =$ \$10788

$\implies E[Y_i|D_i = 1] = 19559 + 10788 = \$30347$
95% CI using robust standard error is (16790.012, 22328.68).
Conditions of causal intrepretation:
1. Consistency 2. Ignorability which is achieved when D_i is randomly assigned or whenY_i is conditionally independent of D_i given other confounders (ie, $Y_i \perp D_i|X_i$) However in this dataset, D_i was not randomly assigned.
ate.raw can be interpreted causally, if only there are no other observed confounders that can affect both treatment and outcome. However, there might be confounder variables such as age, twoearn, income etc that can affect Y_i and D_i. Without conditioning on those confounders,ate.raw cannot be interpreted causally.

Section 4: Simple Covariate adjustment

Suppose that eligibility for enrolling in a 401(k) plan can be taken "as-if" random after conditioning on X. A causal diagram (Directed Acyclic Graph - DAG) illustrating this assumption is shown below. We consider that the decision to work for a firm that offers a 401(k) plan depends both on the observed covariates X, but also on latent firm characteristics, denoted by F; moreover, X, F, and D are jointly affected by a set of latent factors U. Since in this DAG, it can seen that X is common cause(confounder) for both treatment(D) and outcome(Y). Estimate without conditioning on X could induce bias in the estimate by allowing non-causal association between D and Y.

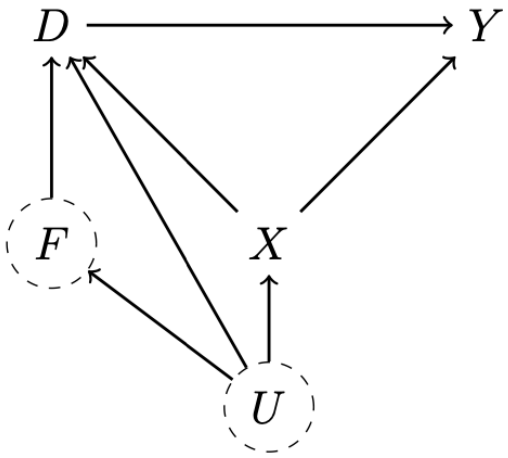


Figure 1: A possible causal DAG for the 401(K) example.

From the DAG, it can be seen that X is an observed confounder whereas U is a latent or unobserved confounder. ie $U_i \to X_i \to D_i$ and $U_i \to X_i \to Y_i$. Moreover X_i is a descendant of U_i. Hence controlling for a descendant of a variable is equivalent to "partially" controlling for that variable (Source: Cinelli 2020). Moreover U and F affects only D and not Y. Hence X is sufficient for control of confounding in this DAG.
Conditional Ignorability condition becomes:
$Y_i(d) \perp D_i = d|X_i$
Under this DAG, ATE can be identified non-paramterically as:
Assuming consistency ($D_i = d \implies Y_i(d) = Y_i$)and conditional ignorability condition specified above.
$ATE = E[Y_i(1) - Y_i(0)] = E[E[Y_i(1)|X_i]] - E[E[Y_i(0)|X_i]]$ (by tower property)
$ATE = E[E[Y_i(1)|D_i = d, X_i]] - E[E[Y_i(0)|D_i = d, X_i]]$ (by conditional ignorability)
$ATE = E[E[Y_i|D_i = 1, X_i]] - E[E[Y_i|D_i = 1, X_i]]$ (By consistency)

Now suppose the CEF is linear:
Model 1: $E[Y_i|D_i, X_i] = \alpha + \tau D_i + X_i^T \beta$

ATE can be estimated from simple linear regression of Y_i on $D_i + X_i$
$ATE = E[\alpha + \tau + X_i^T\beta - (\alpha + X_i^T\beta)] = E[\tau] = \tau =$ coefficient of D_i in the regression $Y_i \sim D_i, X_i$
$ATE = \tau =$ \$4949.023
The 95% CI using robust standard error for ATE is (1868.458, 8029.587).
Controlling for X, the estimate for ATE is lower than the naive estimate (section 3) as it may have reduced the bias (due to confoundedness caused by X) in the estimation of τ.
On average, there is an increase of \$4949.023 in net financial assets of among individuals who were 401k eligible than those who were not while conditioning on all other covariates such as age, income etc.
Note: Code for all these estimates can be found in the appendix.

Section 5: Flexible Covariate adjustment

The assumption of linearity may be too strong. Thus, here we will relax this assumption, and assume instead that the CEF is partially linear on D_i
CEF follows Model 2: $E[Y_i|D_i, X_i] = \alpha + \tau D_i + g(X_i)$
where $g(X_i)$ is an extended covariate set X_i^{est} consisting of polynomials of order 3 for age, inc, educ, and fsize; polynomial of order 1 for the remaining variables; and first order interactions of all these. There are 124 covariates in X_i^{est}. In total there are 126 features in the regression model including the constant, D_i and X_i^{est}. Here orthogonal polynomials are used for better numerical stability as specified in poly() function with default argument RAW = FALSE.
$ATE = E[Y_i|D_i = 1, X_i] - E[Y_i|D_i = 0, X_i] = \alpha + \tau + g(X_i) - (\alpha + g(X_i)) = \tau$
Estimated $ATE = \tau =$ \$9077.974
95% CI using robust standard error is (6704.489, 11451.459)
On average, individuals who were eligible for 401(k) have higher net financial asset ($\sim$ \$9078) than those who were not eligible for 401(k) while conditioning on age, income, and other demographic data.
Under linearity assumption of X, ATE was found to be \$4949.023 (section 4), which is lower than the ATE when the linearity assumption of X is relaxed.
Note: Code for all these estimates can be found in the appendix.

Section 6: Difference in ATE estimation among income quartiles

In this section assume that CEF follows Model 2. Some scholars may argue that the causal effect of 401(k) eligibility may be different depending on your level of income. For example, the effect of 401(k) eligibility may be higher for an individual in the top quartile of income, as compared to an individual in the lower quartile. With this in mind, let G_{ji}, for $j \in \{1, 2, 3, 4\}$, denote binary indicators of the income quartile the individual belongs to. That is, $G_{1i} = 1$ means individual i has an income level in the first quartile of the distribution (i.e., the bottom 25% of the distribution), whereas $G_{4i} = 1$ means individual i has an income level in the last quartile of the distribution (i.e., the top 25% of the distribution).
Model 3: $E[Y_i|D_i, X_i] = \sum_{j=1}^{4} \tau_j D_i G_{ji} + g(X_i)$
where G_{ji} is a binary variable that represent the income quartile the individual belongs to and τ_j is the coefficient of treatment D_i in the j^{th} income quartile.
$ATE = \tau_j$
ATE for q1 income quartile is: \$4087.97
95% CI using robust std error is(2161.27 , 6014.669)
ATE for q2 income quartile is: \$3192.739
95% CI using robust std error is(1083.622 , 5301.856)
ATE for q3 income quartile is: \$7710.935
95% CI using robust std error is(4513.04 , 10908.83)
ATE for q4 income quartile is: \$18843.81
95% CI using robust std error is(11597.62 , 26090)
ATE estimate (effect of 401(k) eligibility) varies across different income groups. For each group, the net

financial asset for individuals with 401(k) eligibility is higher than those without 401(k) eligibility.
Note: Code for all these estimates can be found in the appendix.

Section 6.1: Difference of the effect between the last and first quartiles

Difference between causal effect of 401(k) eligibility on net financial assets for those individuals on the q1 as compared to those in the q4:
$\triangle_{q4-q1} = 18843.81 - 4087.97 = \$14755.84$
95% CI using robust SE for $\triangle_{q4-q1}$ is:
2.5%: 11597.62 - 2161.27 = 9436.35
97.5%: 26090 - 6014.669 = 20075.331
95% CI using robust SE for $\triangle_{q4-q1}$ is: (9436.35, 20075.331)
Another method:
From 1(d) it was shown that $\frac{\hat{\beta}_{OLS}-\beta}{\hat{SE}_{rob}(\hat{\beta})} \xrightarrow{d} N(0,1) \implies \hat{\beta}_{OLS} \sim N(\beta, \hat{SE}_{rob}(\hat{\beta}))$
By property of normal distribution if $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2), \implies X_1 - X_2 \sim N(\mu_1 - \mu 2, \sigma_1^2 + \sigma_2^2)$
Similarly, $\hat{\beta}_{q4} - \hat{\beta}_{q1} \sim N(\beta_{q4} - \beta_{q1}, \hat{SE}^2_{rob.q4} + \hat{SE}^2_{rob.q1})$
From part(b), $\hat{SE}_{rob.q4}$ = (18843.81 -11597.62)/1.96 = 3697.03
From part(b), $\hat{SE}_{rob.q1}$ = (4087.97 -2161.27)/1.96 = 983.01
$\hat{SE}_{rob.q4-q1} = \sqrt{3697.03^2 + 983.01^2} = 3825.4$
95% CI using robust SE for $\triangle_{q4-q1}$ is:
2.5% quartile = 14755.84 - 1.96 X 3825.4 = 7258.06
97.5% quartile = 14755.84 + 1.96 X 3825.4 = 22253.6.
Hence the 95% quartile using property of normal distribution is : (7258.06,22253.6)
Note: Code for all these estimates can be found in the appendix.

Section 7: Flexible adjustment with Lasso and FWL

For this problem, we will consider a more flexible covariate set X^{ext2}, including: (i) polynomials of order 8 for age, inc, educ, and fsize; (ii) a polynomial of order 1 for the remaining variables; (iii)first order interactions of all these.
Let $E[Y_i|D_i, X_i] = \alpha + \tau D_i + f(X_i)$ where $f(X_i)$ is a polynomial covariate matrix. Let $E[Y_i|D_i, X_i]$ can be approximated by OLS estimator.
Similarly $E[Y_i|X_i]$ can be approximated by running linear regression of Y_i on X_i. and $E[D_i|X_i]$ can be approximated by running linear regression of D_i on X_i.
Let$Y \sim E[Y_i|D_i, X_i] + \epsilon = \alpha + \tau D_i + f(X_i) + e$
Hence residuals can also be written as:
$\tilde{Y_i} = Y_i^{\perp X_i} = Y_i - E[Y_i|X_i] = Y_i - OLS(Y_i|X_i)$
$\tilde{D_i} = D_i^{\perp X_i} = D_i - E[D_i|X_i] = D_i - OLS(D_i|X_i)$
By property of FWL theorem, $f(X_i)^{\perp X_i} = 0$ and $e^{\perp X_i} = 0$
$\frac{cov(Y_i^{\perp X_i}, D_i^{\perp X_i})}{var(D_i^{\perp X_i})} = \frac{cov([\alpha+\tau D_i+f(X_i)+e]^{\perp X_i}, D_i^{\perp X_i})}{var(D_i^{\perp X_i})} = \frac{cov(\alpha^{\perp X_i}, D_i^{\perp X_i}) + cov(\tau D_i^{\perp X_i}, D_i^{\perp X_i})}{var(D_i^{\perp X_i})}$
$cov(\alpha^{\perp X_i}, D_i^{\perp X_i}) = 0$
$\frac{cov(Y_i^{\perp X_i}, D_i^{\perp X_i})}{var(D_i^{\perp X_i})}$ is the coefficient of regressing $Y_i^{\perp X_i}$ on $D_i^{\perp X_i}$
Hence, $\frac{cov(Y_i^{\perp X_i}, D_i^{\perp X_i})}{var(D_i^{\perp X_i})} = \tau \frac{var(D_i^{\perp X_i})}{var(D_i^{\perp X_i})} = \tau = ATE$
Hence ATE is recovered from simple regression of $Y_i^{\perp X_i}$ on $D_i^{\perp X_i}$.
Using the extended covariate set X^{ext2}, with Lasso and 10-fold cross-validation to estimate $E[Y_i|X_i]$ and $E[D_i|X_i]$.
Number of non-zero coefficients that produces the best fit(ie, min MSE) using set.seed(123) is:
For estimation of $E[Y|X^{est2}] = 87$

For estimation of $E[D|X^{est2}] = 35$
Note: these numbers change as the randomization changes. I noticed that without set.seed, cv.glmnet produces different results each time since K fold cross validation samples a subset from data randomly each time. Similarly the plot of CV-MSE vs penalty parameter also changes each time with randomization.
Estimated τ from residuals $\tilde{Y}_i$ and $\tilde{D}_i$ is \$8857.542.
95% CI for τ using robust standard error is (6513.360 ,11201.725)
Estimated τ from residuals $\tilde{Y}_i$ and $\tilde{D}_i$ is \$8857.542.
95% CI for τ using robust standard error is (6513.360 ,11201.725)
Note: Code for all these estimates can be found in the appendix.

Section 7.1: Difference of the effect between the last and first quartiles using X^{ext2}
Let $Y_i = E[Y_i|D_i, X_i] = \sum_{j=1}^{4} \tau_j D_i G_{ji} + g(X_i) + e$
$Y_i^{\perp X_i} = [\sum_{j=1}^{4} \tau_j D_i G_{ji}]^{\perp X_i} + g(X_i)^{\perp X_i} + e^{\perp X_i}$
Since $e^{\perp X_i} = 0$ and $g(X_i)^{\perp X_i} = 0$
$Y_i^{\perp X_i} = \sum_{j=1}^{4} \tau_j G_{ji} D_i^{\perp X_i}$
Let $Z_i = G_{ji} D_i^{\perp X_i}$
$\frac{cov(Y_i^{\perp X_i}, Z_i)}{var(Z_i)}$ = coefficient of Z_i on regressing Y_i on $\sum_{i=1}^{4} Z_i$
$\frac{cov(Y_i^{\perp X_i}, Z_i)}{var(Z_i)} = \frac{cov(\sum_{i=1}^{4} \tau_i Z_i, Z_i)}{var(Z_i)}$
Since $cov(Z_i, Z_k) = 0 \forall i \neq k$
Hence $\frac{cov(Y_i^{\perp X_i}, Z_i)}{var(Z_i)} = \tau_i$ = ATE of income quartile "i".
tau1: ATE for individual in q1 is \$4114.792 with 95% CI in (2010.445, 6219.140) tau2: ATE for individual in q2 is \$3199.0302 with 95% CI in (850.1893, 5547.8711) tau3: ATE for individual in q3 is \$6755.068 with 95% CI in (3198.345, 10311.791) tau4: ATE for individual in q4 is \$18503.19 with 95% CI in (11741.56, 25264.83)
Note: Code for all these estimates can be found in the appendix.

Section 8: Appendix:

Section 8.1 Code for Section 3

```
# load data
pension.data = read.csv("/Users/anuram/Library/Mobile Documents/com~apple~CloudDocs/MS Stats/Winter 202
```

```
# part(a) Linear regression coefficient
ols2 = lm(net_tfa ~ e401, data = pension.data)
ate.raw = coef(ols2)["e401"]
ate.raw
```

```
##     e401
## 19559.34
```

```
# part(a) 95% CI using robust error
CI = Confint(ols2, vcov. = vcovHC(ols2, type = "HC0"))
```

```
## Standard errors computed by vcovHC(ols2, type = "HC0")
```

```
CI["e401", ]
```

```
## Estimate    2.5 %   97.5 %
## 19559.34 16790.01 22328.68
```

Section 8.1 Code for Section 4

```
# Part(d) ATE estimate using linear regression
ols3 = lm(net_tfa ~ e401 + age + inc + educ + fsize + marr +
    twoearn + pira + hown, data = pension.data)
ate.est1 = coef(ols3)["e401"]
ate.est1
```

```
##     e401
## 4949.023
```

```
# part(d) 95% CI using robust se
CI = Confint(ols3, vcov. = vcovHC(ols3, type = "HC0"))
```

```
## Standard errors computed by vcovHC(ols3, type = "HC0")
```

```
CI["e401", ]
```

```
## Estimate    2.5 %   97.5 %
## 4949.023 1868.458 8029.587
```

Section 8.1 Code for Section 5

```
# Problem 3.4 (a)
ols_trial = lm(net_tfa ~ e401 + (poly(age, 3) + poly(inc, 3) +
    poly(fsize, 3) + poly(educ, 3) + marr + twoearn + pira +
    hown)^2, data = pension.data)

CI = Confint(ols_trial, vcov. = vcovHC(ols_trial, type = "HC0"))
```

```
## Standard errors computed by vcovHC(ols_trial, type = "HC0")
```

```
tot.covariate = nrow(CI)
tot.covariate
```

```
## [1] 126
```

```
CI["e401", ]
```

```
##  Estimate     2.5 %    97.5 %
##  9077.974  6704.489 11451.459
```

```
# part(b) ATE estimate for different groups creating income
# quartiles based on pendion.data$inc
pension.data$groups = cut(pension.data$inc, quantile(pension.data$inc,
    c(0, 0.25, 0.5, 0.75, 1), na.rm = TRUE), labels = c("q1",
    "q2", "q3", "q4"), include.lowest = T)

for (i in unique(pension.data$groups)) {
    data.new = pension.data[pension.data$groups == i, ]
    ols_new = lm(net_tfa ~ e401 + (poly(age, 3) + poly(inc, 3) +
        poly(fsize, 3) + poly(educ, 3) + marr + twoearn + pira +
        hown)^2, data = data.new)
    ate = coef(ols_new)["e401"]
    cat("ATE for", i, "income quartile is:", ate)
    cat("\n")
    CI = Confint(ols_new, vcov. = vcovHC(ols_new, type = "HC0"))
```

```
    cat("95% CI using robust std error is(", CI["e401", 2], ",",
        CI["e401", 3], ")")
    cat("\n")
}
```

```
## ATE for q2 income quartile is: 3192.739
## Standard errors computed by vcovHC(ols_new, type = "HC0")
## 95% CI using robust std error is( 1083.622 , 5301.856 )
## ATE for q3 income quartile is: 7710.935
## Standard errors computed by vcovHC(ols_new, type = "HC0")
## 95% CI using robust std error is( 4513.04 , 10908.83 )
## ATE for q4 income quartile is: 18843.81
## Standard errors computed by vcovHC(ols_new, type = "HC0")
## 95% CI using robust std error is( 11597.62 , 26090 )
## ATE for q1 income quartile is: 4087.97
## Standard errors computed by vcovHC(ols_new, type = "HC0")
## 95% CI using robust std error is( 2161.27 , 6014.669 )
```

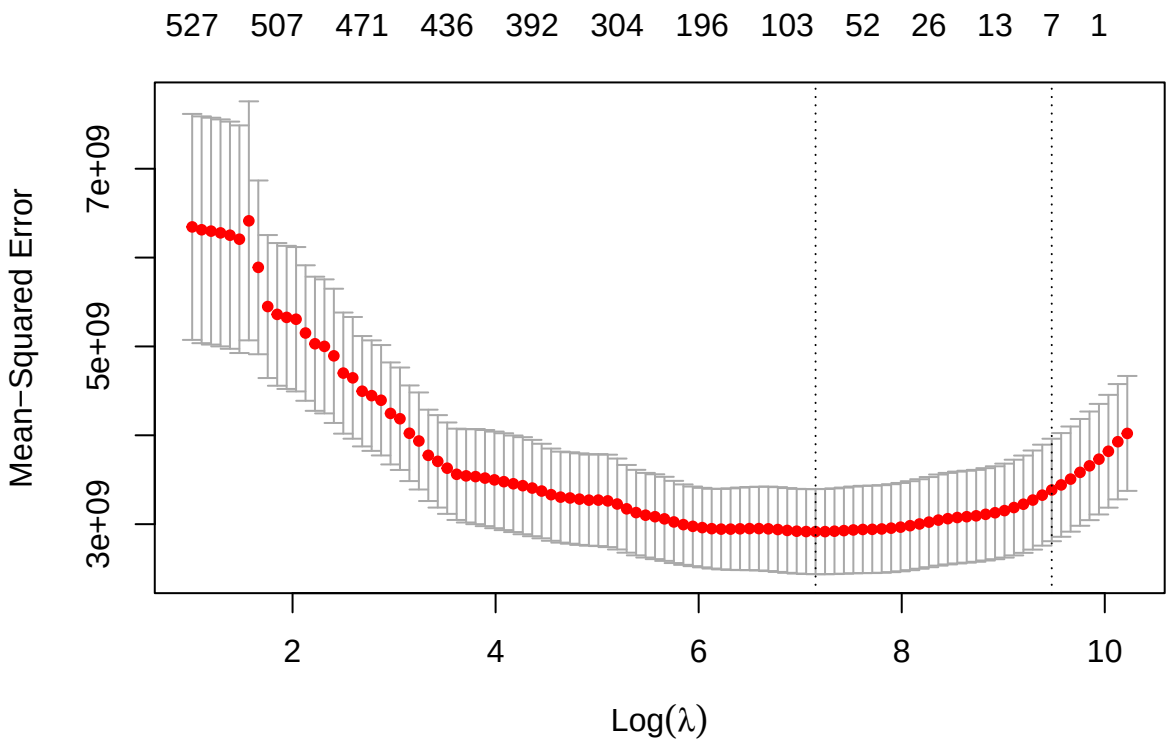
Section 8.1 Code for Section 6

```
# part(b)Estimating $E[Y|X^{est2}]$ and $E[D|X^{est2}]$
# using lasso and 10 fold CV Extended covariate matrix
# X_est2
X_est2 = model.matrix(~(poly(age, 8) + poly(inc, 8) + poly(fsize,
    8) + poly(educ, 8) + marr + twoearn + pira + hown)^2, data = pension.data)
Y = pension.data$net_tfa
D = pension.data$e401
```

```
# Regressing Y on X_est2 to estimate E[Y|X] using Lasso and
# 10fold CV
set.seed(123)
E.yx = cv.glmnet(X_est2, Y, type = "mse", alpha = 1)
E.yx
```

```
##
## Call:  cv.glmnet(x = X_est2, y = Y, type.measure = "mse", alpha = 1)
##
## Measure: Mean-Squared Error
##
##     Lambda Index   Measure        SE Nonzero
## min   1276    34 2.916e+09 478810475      87
## 1se  13062     9 3.385e+09 576584693       7
```

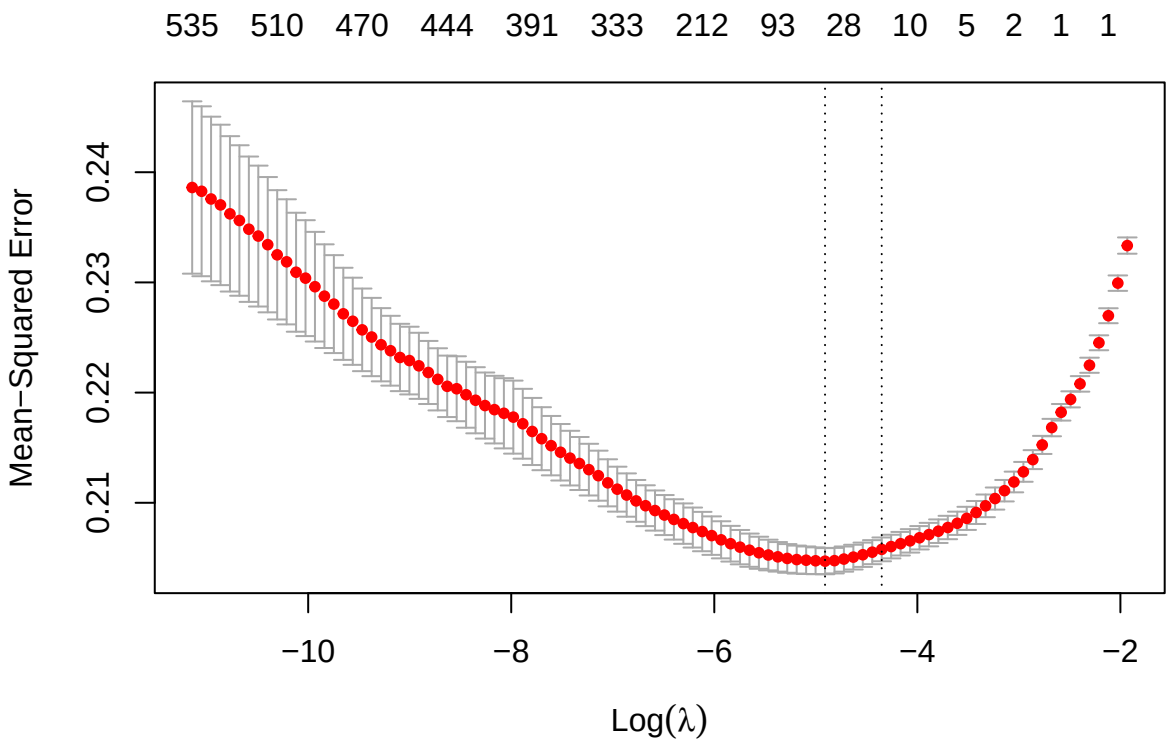
```
# plotting CV MSE as a function of penalty parameter
plot(E.yx)
```




```
# Regressing D on X_est2 to estimate E[D|X] using Lasso and
# 10fold CV
set.seed(123)
E.dx = cv.glmnet(X_est2, D, type = "mse", alpha = 1)
E.dx
```

```
##
## Call:  cv.glmnet(x = X_est2, y = D, type.measure = "mse", alpha = 1)
##
## Measure: Mean-Squared Error
##
##       Lambda Index Measure       SE Nonzero
## min 0.007377    33  0.2047 0.001186      35
## 1se 0.012892    27  0.2058 0.001080      14
```

```
# plotting CV MSE as a function of penalty parameter
plot(E.dx)
```




```
# (c) Residuals of $\tilde{Y_i}$
Y_pred = predict(E.yx, newx = X_est2, s = "lambda.min")
y.tilde = Y - Y_pred
# (c) Residuals of $\tilde{\D_i}$
D_pred = predict(E.dx, newx = X_est2, s = "lambda.min")
d.tilde = D - D_pred

# Regression of $\tilde{Y_i}$ on $\tilde{\D_i}$
ols.tildeyd = lm(y.tilde ~ d.tilde)
tau.est = coef(ols.tildeyd)["d.tilde"]
tau.est
```

```
##  d.tilde
## 8857.542
```

```
# 95%
CI = Confint(ols.tildeyd, vcov. = vcovHC(ols.tildeyd, type = "HC0"))
```

```
## Standard errors computed by vcovHC(ols.tildeyd, type = "HC0")
```

```
CI["d.tilde", ]
```

```
##  Estimate     2.5 %    97.5 %
##  8857.542  6513.360 11201.725
```

```
pension.data$q1 = ifelse(pension.data$groups == "q1", 1, 0)
pension.data$q2 = ifelse(pension.data$groups == "q2", 1, 0)
pension.data$q3 = ifelse(pension.data$groups == "q3", 1, 0)
pension.data$q4 = ifelse(pension.data$groups == "q4", 1, 0)
z1 = d.tilde * pension.data$q1
z2 = d.tilde * pension.data$q2
z3 = d.tilde * pension.data$q3
z4 = d.tilde * pension.data$q4
```

```r
ols.model3 = lm(y.tilde ~ z1 + z2 + z3 + z4)
CI = Confint(ols.model3, vcov. = vcovHC(ols.model3, type = "HC0"))
```

```
## Standard errors computed by vcovHC(ols.model3, type = "HC0")
```

```r
cat("tau1: ATE for individual in q1")
```

```
## tau1: ATE for individual in q1
```

```r
cat("\n")
```

```r
tau1 = CI["z1", ]
tau1
```

```
## Estimate    2.5 %   97.5 %
## 4114.792 2010.445 6219.140
```

```r
cat("tau2: ATE for individual in q2")
```

```
## tau2: ATE for individual in q2
```

```r
cat("\n")
```

```r
tau2 = CI["z2", ]
tau2
```

```
##  Estimate     2.5 %    97.5 %
## 3199.0302  850.1893 5547.8711
```

```r
cat("tau3: ATE for individual in q3")
```

```
## tau3: ATE for individual in q3
```

```r
cat("\n")
```

```r
tau3 = CI["z3", ]
tau3
```

```
##  Estimate     2.5 %    97.5 %
##  6755.068  3198.345 10311.791
```

```r
cat("tau4: ATE for individual in q4")
```

```
## tau4: ATE for individual in q4
```

```r
cat("\n")
```

```r
tau4 = CI["z4", ]
tau4
```

```
## Estimate    2.5 %   97.5 %
## 18503.19 11741.56 25264.83
```