Modelos Computacionais em Economia

Marcleiton Morais

Universidade Federal do Tocantins (UFT)

15 de julho de 2021

Multiple Regression [Adams, 2020]

There is clear evidence that other factors are also important, including experience, gender and race. **When we should and should not run a long regression**?

Using Short Regression

y is determined by both x and w (and v) (long regression):

$$y_i = a + bx_i + cw_i + v_i$$

The income is determined by their years of schooling (x_i), by their experience (w_i) and (v_i).
Following short regression:

$$y_i = a + bx_i + v_{wi}$$

Does it matter? Does it matter if we just leave out important explanatory variables? Yes. And No. Maybe. It depends.

Multiple Regression

It makes little difference if we run the short or long regression.

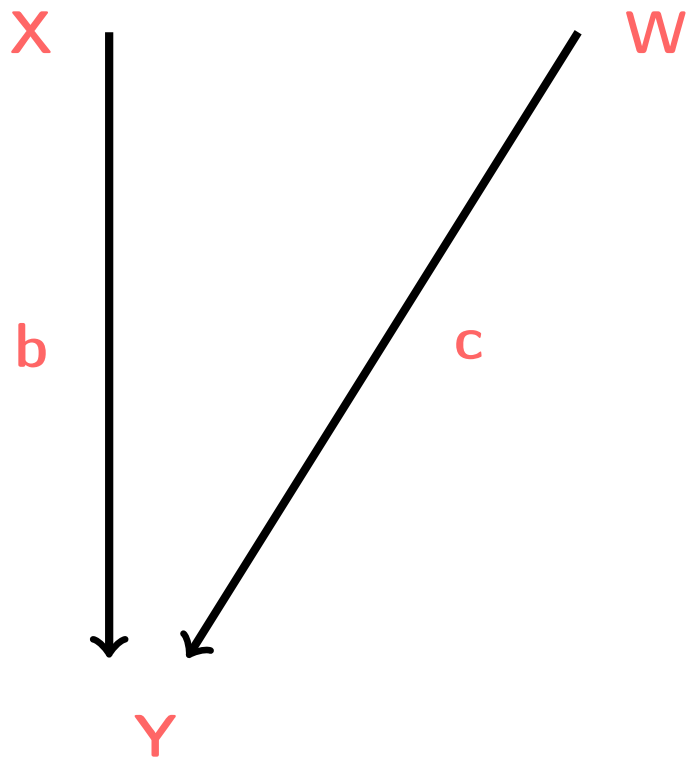

Figura: Two variable causal graph

Multiple Regression

Short regressions are much less trustworthy when there is some sort of dependence between the two variables

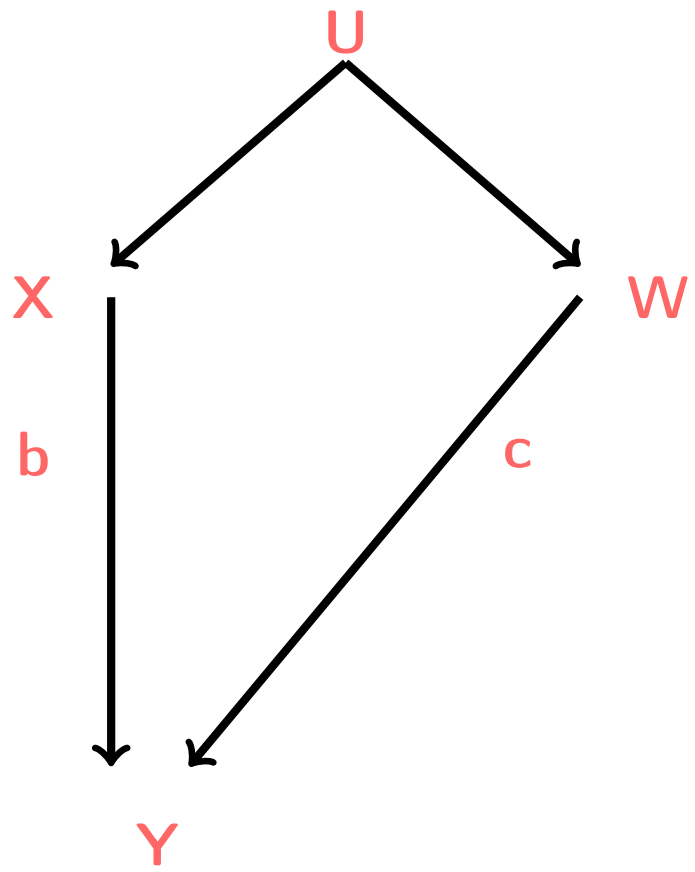

Figura: Two variable causal graph with dependence.

In this case, a short regression will give a biased estimate of b because it will incorporate c. We can trace a relationship from X to Y through the

Multiple Regression

Simulation with Multiple Explanatory Variables

```
set.seed(123456789)
N <- 1000
a <- 2
b <- 3
c <- 4
u_x <- rnorm(N)
alpha <- 0 #x and w affect y independently
x <- x1 <- (1 - alpha)*runif(N) + alpha*u_x
w <- w1 <- (1 - alpha)*runif(N) + alpha*u_x
# this creates two identical variables
u <- rnorm(N)
y <- a + b*x + c*w + u
lm1 <- lm(y ~ x)
lm2 <- lm(y ~ x + w)
```

The common factor is determined by α.

Multiple Regression

Simulation with Multiple Explanatory Variables

```
alpha <- 0.5 #allows for dependence
x <- x2 <- (1 - alpha)*runif(N) + alpha*u_x
w <- w2 <- (1 - alpha)*runif(N) + alpha*u_x
y <- a + b*x + c*w + u
lm3 <- lm(y ~ x)
lm4 <- lm(y ~ x + w)
```

Multiple Regression

Simulation with Multiple Explanatory Variables

If x and w are highly correlated:

```
alpha <- 0.95
x <- x3 <- (1 - alpha)*runif(N) + alpha*u_x
w <- w3 <- (1 - alpha)*runif(N) + alpha*u_x
y <- a + b*x + c*w + u
lm5 <- lm(y ~ x)
lm6 <- lm(y ~ x + w)
# install.packages("stargazer")
require(stargazer) #presenting regression results.
stargazer(list(lm1,lm2,lm3,lm4,lm5,lm6),
          keep.stat = c("n","rsq"),
          float = FALSE, font.size = "small", digits=2,
          keep=c(1:6))
```

To see this in the console set type = "text".

Multiple Regression

Simulation with Multiple Explanatory Variables

	Dependent variable:					
	y					
	(1)	(2)	(3)	(4)	(5)	(6)
x	3.14***	2.81***	6.84***	2.86***	7.01***	0.67
	(0.17)	(0.11)	(0.08)	(0.17)	(0.03)	(1.59)
w		4.05***		4.16***		6.35***
		(0.11)		(0.16)		(1.59)
Constant	3.98***	2.15***	2.14***	2.07***	2.07***	2.08***
	(0.10)	(0.08)	(0.05)	(0.04)	(0.03)	(0.03)
Observations	1,000	1,000	1,000	1,000	1,000	1,000
R^2	0.26	0.68	0.88	0.93	0.98	0.98
Note:						*p<0.1; **p<0.05; ***p<0.01

Multiple Regression

Matrix Algebra of Short Regression

Consider the matrix algebra:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{W}\gamma + v$$

"Dividing" by X for the true and estimated short regression:

$$\hat{\beta} - \beta = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}\gamma + (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'v$$

short regression gives the same answer if $(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'v = 0$ and either $\gamma = 0$ or $(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}$.

- *Xs* are independent of the unobserved characteristic.
- *W*'s have no effect on the outcome (Y).
- There is no correlation between the *X*s and the *W*s.

The short regression gives biased estimates incorporating both the effect of x and w. But the long regression cannot disentangle the two effects.

Multiple Regression

The correlation is captured by X′W

```
# calculates the covariance between x1 and w1
cov(x1,w1)

cov(x2,w2)

t(x2)%*%w2
# this corresponds to the linear algebra above
# it measures the correlation between the Xs and Ws.
```

Multiple Regression

Collinearity and Multicollinearity

If the two variables are strongly correlated then the long regression cannot distinguish between the two different effects. **There is a multicollinearity problem.**

The true parameter vector can be written as follows:

$$\beta = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y} - (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'v$$

Standard assumption for OLS:

1) Unobserved characteristic is independent of the observed characteristic.
2) Unobserved characteristic affects the dependent variable additively.
3) The matrix $X'X$ must be invertible (full-column rank). Or will be called "collinearity".

An "almost" **not full-column rank** cause a problem called multicollinearity.

Multiple Regression

Matrix Algebra of Multicollinearity

In the problem we make two standard assumptions:

- The average value of the unobserved characteristic is 0.
- The Xs are independent of the Us. That implies that $X'v$ will be zero when the sample size is large.

When the matrix of observable characteristics is "almost" not full-column rank then this weighted average can diverge quite a lot from 0.

Multiple Regression

Understanding Multicollinearity with R

```
X2 <- cbind(1,x3,w3)
solve(t(X2)%*%X2)%*%t(X2)%*%u
#In a perfect world, this
#would be a vector of 0s

mean(u)

#Covariace between two variables
cov(x3,u)

cov(w3,u)

(1/N)*t(X2)%*%u
```

Again we can look at the main OLS assumptions, that the mean of the unobserved term is zero and the covariance between the unobserved term

Returns to Schooling

Multiple Regression of Returns to Schooling

Standard characteristics that are known to determine income are work experience, race, the region of the country where the individual grew up and the region where the individual currently lives.

$$Income_i = \alpha + \beta Education_i + \gamma Experience_i + \cdots + Unobserved_i$$

A short regression to estimate $\hat{\beta} = 0.052$. What happens if we estimate a long regression?

```
x <- read.csv("nls.csv",as.is=TRUE)
x$wage76 <- as.numeric(x$wage76)
x$lwage76 <- as.numeric(x$lwage76)
# note that this is "el" wage 76.
x1 <- x[is.na(x$lwage76)==0,]
#working years after school
x1$exp <- x1$age76 - x1$ed76 - 6
x1$exp2 <- (x1$exp^2)/100
# experienced squared divided by 100
# this makes the presentation nicer.
```

Returns to Schooling

Multiple Regression of Returns to Schooling

```
lm1 <- lm(lwage76 ~ ed76, data=x1)
lm2 <- lm(lwage76 ~ ed76 + exp +exp2, data=x1)
lm3 <- lm(lwage76 ~ ed76 + exp +exp2 + black + reg76r,
data=x1)
lm4 <- lm(lwage76 ~ ed76 + exp +exp2 + black +reg76r+
smsa76r + smsa66r + reg662 + reg663 + reg664 +
reg665 + reg666 + reg667 + reg668 + reg669, data=x1)
#reg76 refers to living in the south in 1976
#smsa urban or rural in 1976.
#reg region of the US - North, South, West etc.
#66 refers to 1966.
stargazer(list(lm1,lm2,lm3,lm4),
          keep.stat = c("n","rsq"),
          float = FALSE, font.size = "small", digits=2,
          keep=c(1:4))
```

Causal Pathways

Consider the case against Harvard University for discrimination against Asian-Americans in undergraduate admissions.

The questions is then how does this causal relationship work? direct causal relationship between race and admissions or indirect.

Dual path model

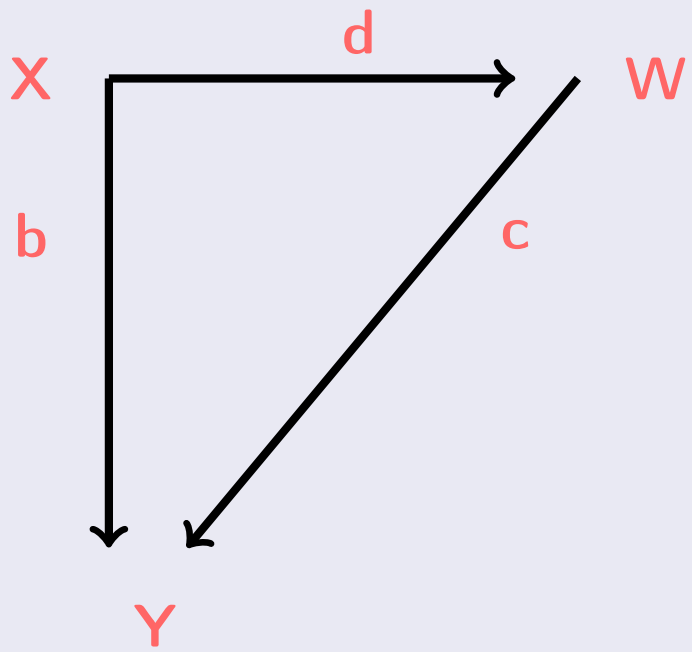

Figura: Dual path causal graph.

- Direct causal effect of X on Y which has value b.
- The indirect effect of X on Y is c times d.

Causal Pathways

In algebra

$y_i = a + bx_i + cw_i + v_i$ and $w_i = dx_i + v_{wi}$. Substituting

$$y_i = a + (b + cd)x_i + v_i + cv_{wi}$$

- The full relationship of X on Y $(b + cd)$.
- Is straightforward to estimate $b + cd$ and it is straightforward to estimate d. It is not straightforward to estimate c.
- Running the regression of Y on W gives $c + \frac{b}{d}$. Where $\frac{b}{d}$ is the **backdoor relationship**. OLS estimates c using the observed variation between W and Y but to some extent that variation is being determined by b and d.
  We will make the problem go away by assuming that $b = 0$.

Causal Pathways

Simulation of Dual Path Model

```
set.seed(123456789)
N <- 50
a <- 1
b <- 0
c <- 3
d <- 4
x <- round(runif(N)) # creates a vector of 0s and 1s
u_w <- runif(N)
w <- d*x + u_w
u <- rnorm(N)
y <- a + b*x + c*w + u
shortreg <- lm(y ~ x)
summary(shortreg)
longreg <- lm(y ~ x + w)
summary(longreg)
```

Causal Pathways

The issue with the standard long regression is **multicollinearity**. It hard for the algorithm to separate the effect of x on y from the effect of w on y.

Simulation of Dual Path Model

There is a better way to do the estimation:

```
e_hat <- lm(y ~ x)$coef[2]
# element 2 is the slope coefficient of interest.
c_hat <- lm(y ~ w)$coef[2]
d_hat <- lm(w ~ x)$coef[2]
# Estimate of b
e_hat - c_hat*d_hat
```

$\hat{b}$ will tend to be much closer to the true value of zero than the standard estimate from the long regression.

Causal Pathways

Dual Path Estimator Versus Long Regression

```
set.seed(123456789)
b_mat <- matrix(NA,100,3)
for (i in 1:100) {
        x <- round(runif(N))
        u_w <- runif(N)
        w <- d*x + u_w
        u <- rnorm(N)
        y <- a + b*x + c*w + u
        lm2_temp <- summary(lm(y ~ x + w))
        # coefficients object (item 4)
        b_mat[i,1] <- lm2_temp[[4]][2]
        # The 4th item is the results vector.
        # The second item is the coefficient on x.
        b_mat[i,2] <- lm2_temp[[4]][8]
        # the 8th item is the T-stat on the coefficient on x.
        e_hat <- lm(y ~ x)$coef[2]
        c_hat <- lm(y ~ w)$coef[2]
        d_hat <- lm(w ~ x)$coef[2]
        b_mat[i,3] <- e_hat - c_hat*d_hat }
```

Causal Pathways

Dual Path Estimator Versus Long Regression

```
colnames(b_mat) <-
c("Standard Est","T-Stat of Standard","Proposed Est")
summ_tab <- summary(b_mat)
rownames(summ_tab) <- NULL
print(xtable(summ_tab), floating=FALSE)
```

The standard estimates vary from over -5 to 5, while the proposed estimates have a much much smaller variance.

Causal Pathways

Histogram and density plot of standard estimator compared to the minimum and maximum values from the proposed estimator (dashed lines) from simulated data.

```
hist(b_mat[,1],freq=FALSE,xlab="Estimate of b",main="")
# plots the histogram, xlab labels the x-axis,
# freq type of histogram.
# main provide a title for the chart, here it is empty.
lines(density(b_mat[,1]), type="l",lwd=3)
# lines is used to add another plot to an existing plot.
# calculates an approximation of the density function.
# type l means that the plot is a line
# lwd determines the line width.
abline(v=c(min(b_mat[,3]),max(b_mat[,3])), lty=2, lwd=3)
# v determines vertical lines.
```

Causal Pathways

Matrix Algebra of the Dual Path Estimator

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{W}\gamma + \epsilon$$

$$\mathrm{W} = \mathrm{X}\Delta + \Upsilon$$

- y is a $(Nx1)$, X is a (NxJ) and W is a (NxK).
- Parameters of interest ($\beta(Jx1)$ is the direct effect and ($\gamma(Kx1)$ and $\Delta(JxK)$) the indirect effect.
- $\epsilon(Nx1)$.
- $\Upsilon(NxK)$.

Causal Pathways

Matrix Algebra of the Dual Path Estimator

Then remember that we can derive β by estimating Δ and y, as $\beta = \tilde{\beta} - \Delta\gamma$.

$$\hat{\tilde{\beta}} = (X'X)^{-1}X'y$$

$$\hat{\Delta} = (X'X)^{-1}X'W$$

$$\hat{\gamma} = (W'W)^{-1}W'y$$

Substituting the results of the last two regressions into the appropriate places we get our proposed estimator for the direct effect of X on Y.

$$\hat{\beta} = (X'X)^{-1}X'y - (X'X)^{-1}X'W(W'W)^{-1}W'y$$

Causal Pathways

Dual Path Estimator in R

```
X <- cbind(1,x)
W <- cbind(1,w)
beta_tilde_hat <- solve(t(X)%*%X)%*%t(X)%*%y
Delta_hat <- solve(t(X)%*%X)%*%t(X)%*%W
gamma_hat <- solve(t(W)%*%W)%*%t(W)%*%y
beta_tilde_hat - Delta_hat%*%gamma_hat
```

Causal Pathways

Are Bankers Racist or Greedy?

Blacks are substantially more likely to be denied mortgages than Whites. Determining the causal pathway has implications for policy.

```
x <- read.csv("hmda_aer.csv", as.is = TRUE)
x$deny <- ifelse(x$s7==3,1,NA)
x$deny <- ifelse(x$s7==1 | x$s7==2,0,x$deny)
#S7 == 3 = Application denied
#See codebook for variable names at
#https://sites.google.com/view/microeconometricswithr
#S13 = Applicant race
x$black <- x$s13==3 # creating  a dummy
```

Causal Pathways

Are Bankers Racist or Greedy?

[Munnell et al., 1996] being black reduces 20% the likelihood of getting a mortgage.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1131	0.0069	16.30	0.0000
blackTRUE	0.1945	0.0174	11.21	0.0000

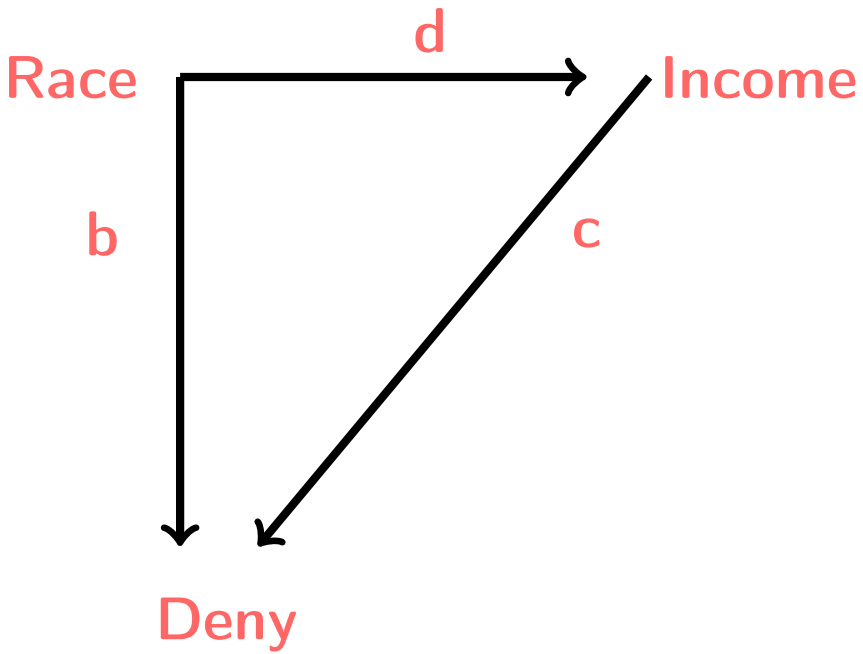

Figura: Dual path causal graph of mortgage denials.

Causal Pathways

Are Bankers Racist or Greedy?

Estimating the Direct Effect: We can create a number of variables from the data set that may mediate race:

```
x$lwage <- NA
#(S31A) Total monthly income of applicant($)
x[x$s31a > 0 & x$s31a < 999999,]$lwage <-
log(x[x$s31a > 0 & x$s31a < 999999,]$s31a)
# Make sure that NAs are not misclassified.
#Credit history - mortgage payments
x$mhist <- x$s42
#(S43) Credit history - consumer payments Codes
x$chist <- x$s43
#(S44) Credit history - public records
x$phist <- x$s44
#(S25A) Years employed in applicable line of work
x$emp <- x$s25a
x$emp <- ifelse(x$emp > 1000, NA, x$emp)
```

Causal Pathways

Are Bankers Racist or Greedy?

Estimating the Direct Effect:

```
Y1 <- x$deny
X1 <- cbind(1,x$black)
W1 <- cbind(1,x$lwage,x$chist,x$mhist,x$phist,x$emp)
index <- is.na(rowSums(cbind(Y1,X1,W1)))==0
# this removes missing values.
X2 <- X1[index,]
Y2 <- Y1[index]
W2 <- W1[index,]
beta_tilde_hat <- solve(t(X2)%*%X2)%*%t(X2)%*%Y2
Delta_hat <- solve(t(X2)%*%X2)%*%t(X2)%*%W2
gamma_hat <- solve(t(W2)%*%W2)%*%t(W2)%*%Y2
beta_tilde_hat - Delta_hat%*%gamma_hat
```

Adding these variables reduces the possible direct effect of race on mortgage denials by almost half.

Causal Pathways

Are Bankers Racist or Greedy? Adding in More Variables:

```
x$married <- x$s23a=="M"
#(S45) Debt-to-income ratio
x$dr <- ifelse(x$s45>999999,NA,x$s45)
#(S41) Consumer credit lines on credit reports
x$clines <- ifelse(x$s41>999999,NA,x$s41)
x$male <- x$s15==1
#(S11) County
x$suff <- ifelse(x$s11 > 999999, NA, x$s11)
#(S35) Liquid assets (in thousands)
x$assets <- ifelse(x$s35 > 999999, NA, x$s35)
#(S6) Loan amount (in thousands)
x$s6 <- ifelse(x$s6 > 999999, NA, x$s6)
#(S50) Appraised value (in thousands)
x$s50 <- ifelse(x$s50 > 999999, NA, x$s50)
#(S33) Purchase price (in thousands)
x$s33 <- ifelse(x$s33 > 999999, NA, x$s33)
#(S6) Loan amount (in thousands)
x$lr <- x$s6/x$s50
x$pr <- x$s33/x$s50
```

Causal Pathways

Are Bankers Racist or Greedy? Adding in More Vvriables

```
#(S16) Co-applicant sex
x$coap <- x$s16==4
x$school <- ifelse(x$school > 999999, NA, x$school)
#Times application was reviewed by underwriter
x$s57 <- ifelse(x$s57 > 999999, NA, x$s57)
#(S48) Term of loan (months)
x$s48 <- ifelse(x$s48 > 999999, NA, x$s48)
#(S39) Number of commercial credit reports
x$s39 <- ifelse(x$s39 > 999999, NA, x$s39)
#(chvalc) Change in median value of
#property in a given tract, 1980-1990
x$chval <- ifelse(x$chval > 999999, NA, x$chval)
#(S20) Number of units in property purchased
x$s20 <- ifelse(x$s20 > 999999, NA, x$s20)
```

Causal Pathways

Are Bankers Racist or Greedy? Adding in More Vvriables

```
x$lwage_coap <- NA
#(S31C) Total monthly income of coapplicant ($)
x[x$s31c > 0 & x$s31c < 999999,]$lwage_coap <-
log(x[x$s31c > 0 & x$s31c < 999999,]$s31c)
x$lwage_coap2 <- ifelse(x$coap==1,x$lwage_coap,0)
x$male_coap <- x$s16==1
```

Causal Pathways

Are Bankers Racist or Greedy? Adding in More Variables

```
W1 <- cbind(1,x$lwage,x$chist,x$mhist,x$phist,x$emp,
        x$emp^2,x$married,x$dr,x$clines,x$male,
        x$suff,x$assets,x$lr,x$pr,x$coap,x$s20,
        x$s24a,x$s27a,x$s39,x$s48,x$s53,x$s55,x$s56,
        x$s57,x$chval,x$school,x$bd,x$mi,x$old,
        x$vr,x$uria,x$netw,x$dnotown,x$dprop,
        x$lwage_coap2,x$lr^2,x$pr^2,x$clines^2,x$rtdum
#x$rtdum measures the racial make up of the neighborho
index <- is.na(rowSums(cbind(Y1,X1,W1)))==0
X2 <- X1[index,]
Y2 <- Y1[index]
W2 <- W1[index,]
```

Causal Pathways

Are Bankers Racist or Greedy? Bootstrap Dual Path Estimator in R

```
set.seed(123456789)
K <- 1000
bs_mat <- matrix(NA,K,2)
for (k in 1:K) {
        index_k <- round(runif(length(Y2),min=1,max=length(Y2)))
        Y3 <- Y2[index_k]
        X3 <- X2[index_k,]
        W3 <- W2[index_k,]
        beta_tilde_hat <- solve(t(X3)%*%X3)%*%t(X3)%*%Y3
        Delta_hat <- solve(t(X3)%*%X3)%*%t(X3)%*%W3
        gamma_hat <- solve(t(W3)%*%W3)%*%t(W3)%*%Y3
        bs_mat[k,] <- beta_tilde_hat - Delta_hat%*%gamma_hat
        # print(k)
}
tab_res <- matrix(NA,2,4)
tab_res[,1] <- colMeans(bs_mat)
tab_res[,2] <- apply(bs_mat, 2, sd)
tab_res[1,3:4] <- quantile(bs_mat[,1], c(0.025,0.975))
```

Causal Pathways

Are Bankers Racist or Greedy? Bootstrap Dual Path Estimator in R

```
tab_res <- matrix(NA,2,4)
tab_res[,1] <- colMeans(bs_mat)
tab_res[,2] <- apply(bs_mat, 2, sd)
tab_res[1,3:4] <- quantile(bs_mat[,1], c(0.025,0.975))
#first row, third and fourth column.
tab_res[2,3:4] <- quantile(bs_mat[,2], c(0.025,0.975))
colnames(tab_res) <- c("Estimate", "SD", "2.5%", "97.5%")
rownames(tab_res) <- c("intercept","direct effect")
```

Bootstrapped estimates of the proposed approach to estimating the direct effect of race on mortgage denials.

References

Adams, C. P. (2020).
Learning Microeconometrics with R.
Chapman and Hall/CRC.

Munnell, A. H., Tootell, G. M., Browne, L. E., and McEneaney, J. (1996).
Mortgage lending in boston: Interpreting hmda data.
The American Economic Review, pages 25–53.