# Problem Set 3

Scott Cohn

November 12 2020

## Question 1

The Wald estimates for 1980 are calculated by dividing $\bar{y}_1 - \bar{y}_0$ by $\bar{x}_1 - \bar{x}_0$ when $x_i$ is *More than two children* imply that having more than two children reduced hours worked per week by 5.18 ($-0.311/0.06$). The 1990 result is also negative but smaller in magnitude. This LATE estimate identified by the subpopulation for which the variation in $x$ is induced by the instrument $z$ has a higher variation and therefore a higher standard error than the $x$. The 95% confidence interval for the Wald statistic given does not include 0. Thus, it is likely that the estimated causal effect is statistically significant.

## Question 2

We use the following model:

$$\texttt{score} = \beta_0 + \beta_1 \texttt{choice} + \beta_2 \texttt{faminc} + u_1,$$

where

- `score` is score on statewide test
- `choice` is binary indicating whether student attended a "choice" school
- `faminc` is family income
- `grant` is dollar amount granted to students to use for tuition at choice schools (IV for `choice`)

### Part A

Higher income families are likely to send their children to choice schools. There is a selection bias because wealthier families have the means to be exercise more choice in their childrens' education regardless of aptitude.

### Part B

Yes — if $u_1$ does not contain income, random assignment ensures that grant designation has no correlation with the other covariates or unobservables.

### Part C

The reduced form equation for `choice` is:

$$\texttt{choice} = \delta_0 + \delta_1 \texttt{faminc} + \pi_1 \texttt{grant} + v_2,$$

where we assume $\pi_1 \neq 0$. That is, it is assumed there is a non-zero effect of the grant designation on `choice`.

### Part D

The reduced form equation for `score` is:

$$\texttt{score} = \gamma_0 + \gamma_1 \texttt{faminc} + \gamma_2 \texttt{grant} + \nu_1.$$

This form is useful because we can directly estimate the effect of the changes in the grant amount on test score controlling for family income. Also using `grant` as an IV for `choice`, we can see the effect of `choice` on score indirectly.

## Question 3

Let $\hat{\beta} = (X'P_Z X)^{-1} X' P_Z y$ where $P_Z = Z(Z'Z)^{-1}Z'$. For the case where $\ell = k$, we want to show that $\hat{\beta} = (Z'X)^{-1}Z'y$.

$$\begin{aligned}
\hat{\beta} &= (X'P_Z X)^{-1} X' P_Z y && \texttt{Given.} && (1)\\
&= (X'Z(Z'Z)^{-1}Z'X)^{-1} X'Z(Z'Z)^{-1}Z'y && \texttt{Substitute } P_Z. && (2)\\
&= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y && \texttt{Expand inverse via hint.} && (3)\\
&= (Z'X)^{-1}(Z'Z)(Z'Z)^{-1}Z'y && (X'Z)^{-1}X'Z = I && (4)\\
&= (Z'X)^{-1}Z'y && (Z'Z)(Z'Z)^{-1} = I, && (5)
\end{aligned}$$

as desired. □

## Question 4

Consider the regression model with scalar $x$ given by $y - \beta_1 + \beta_2 x + u$ with $E(u \mid x) = 0$.

### Part A

To choose an instrument for $x$, we are interested in a variable that is correlated with $x$, but not correlated with $u$. In the first case, we have $x$ as an instrument for itself. Surely $x$ correlates with itself — that is, $\text{Cov}(x, x) = 1 \ (\neq 0)$. It is also given that $E(u \mid x) = 0$. Thus we have

$$E\begin{bmatrix} u \\ xu \end{bmatrix} = \begin{bmatrix} y - \beta_1 - \beta_2 x \\ x(y - \beta_1 - \beta_2 x) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Since $x$ is exogenous, $x$ can be an instrument for itself reducing the IV $\beta$ estimators to the OLS estimates.

In the second case, let $z = x^2$. A similar argument applies. We have

$$E\begin{bmatrix} u \\ x^2 u \end{bmatrix} = E\begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

given the initial condition that $E[u \mid x] = 0$.

### Part B

For $z = x$, it is exactly the OLS estimator. Observe that in large samples the IV-estimator converges to

$$\text{plim}(\hat{\beta}_2) = \beta_2 + \frac{\text{Corr}(z, u)}{\text{Corr}(z, x)} \cdot \frac{\sigma_u}{\sigma_x}.$$

Then, for $z = x$ we substitute into the above relation:

$$\text{plim}(\hat{\beta}_2) = \beta_2 + \frac{\text{Corr}(x, u)}{\text{Corr}(x, x)} \cdot \frac{\sigma_u}{\sigma_x} = \beta_2 + \frac{0}{1} \cdot \frac{\sigma_u}{\sigma_x} = \beta_2,$$

which is the same as the OLS estimator.

For $z = x^2$ we substitute into the above relation:

$$\text{plim}(\hat{\beta}_2) = \beta_2 + \frac{\text{Corr}(x^2, u)}{\text{Corr}(x^2, x)} \cdot \frac{\sigma_u}{\sigma_x} > \beta_2,$$

which is not the OLS estimator.

### Part C

The estimator in the GMM approach would be the same as the 2SLS approach for both $x$ and $x^2$ because we are in the *exact identified case* where we have the same number of instruments as endogenous variables.

## Question 5

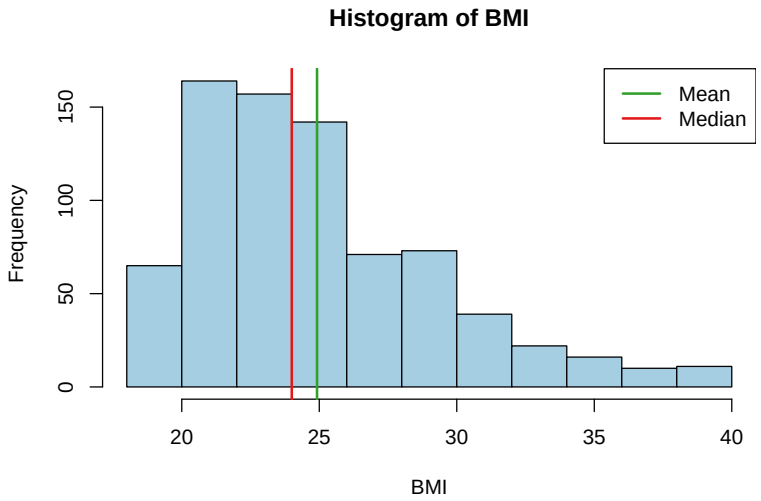
### Part A

```
kbl_q <- children_samp_wm %>%
  summarize(quants = round(quantile(bmi, probs = c(0.1, 0.25, 0.5, 0.75, 0.9)), 2)) %>%
  mutate(tau = c("q10", "q25", "q50", "q75", "q90")) %>%
  rbind(c(round(mean(children_samp$bmi), 2), "mean"))

kbl_q %>%
  select(tau, quants) %>%
  kbl(format = "latex",
              col.names = c("$\\tau$", "Quantile"),
              digits = 2,
              booktabs = TRUE,
              caption = "Quantiles for BMI",
              escape = FALSE
              ) %>%
  kable_styling(latex_options = "hold_position")
```

Table 1: Quantiles for BMI

| $\tau$ | Quantile |
|---|---|
| q10 | 20.2 |
| q25 | 21.7 |
| q50 | 24 |
| q75 | 27.4 |
| q90 | 30.91 |
| mean | 24.45 |

## Part B




Yes, the relative values make sense given that the data has a right skew.

## Part C

```
bmi_x_reg <- lm(bmi ~ educ + age + mombmi + dadbmi, data = children_samp,
                subset = (male == 1 & white == 1))

summary(bmi_x_reg)
```

```
##
## Call:
## lm(formula = bmi ~ educ + age + mombmi + dadbmi, data = children_samp,
##     subset = (male == 1 & white == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2064 -2.8689 -0.9492  2.2170 15.4804
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.045149   1.714538   5.276 1.72e-07 ***
## educ        0.003787   0.097868   0.039    0.969
## age         0.279955   0.060082   4.660 3.74e-06 ***
## mombmi      0.189110   0.028271   6.689 4.33e-11 ***
## dadbmi      0.172827   0.036285   4.763 2.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.154 on 765 degrees of freedom
## Multiple R-squared:  0.1393, Adjusted R-squared:  0.1348
## F-statistic: 30.95 on 4 and 765 DF,  p-value: < 2.2e-16
```

This model indicates that `age`, `mombmi`, and `dadbmi` are statistically significant at the 1% level and all have positive signs. The covariate `educ` is not statistically significant in its association with the dependent variable `bmi`.

## Part D

```
# Quantile reg at median (LAD)
bmi_x_reg_quant <- rq(bmi ~ educ + age + mombmi + dadbmi,
                      tau = 0.5,
                      data = children_samp,
                      subset = (male == 1 & white == 1))

# Summarize with 500 replications
summary(bmi_x_reg_quant, se = "boot", R = 500, bsmethod = "xy")
```

```
##
## Call: rq(formula = bmi ~ educ + age + mombmi + dadbmi, tau = 0.5, data = children_samp,
##     subset = (male == 1 & white == 1))
##
## tau: [1] 0.5
##
## Coefficients:
##             Value   Std. Error t value Pr(>|t|)
## (Intercept) 7.68132 1.79215    4.28610 0.00002
## educ        0.05393 0.09801    0.55020 0.58234
## age         0.33584 0.07429    4.52066 0.00001
## mombmi      0.13153 0.03034    4.33564 0.00002
## dadbmi      0.17773 0.04395    4.04432 0.00006
```

### D.i

A 1 year increase in education is associated with a 0.05 unit increase in BMI around the conditional mean holding all else fixed. This effect is not statistically significant at the 5% level.

### D.ii

If `mombmi` and `dadbmi` both increase by 1 unit, then `bmi` increases by $0.13153 + 0.17773 = 0.30926$.

```
x <- as.data.frame(children_samp) %>%
  dplyr::filter(white == 1 & male == 1) %>%
  select(educ, age, mombmi, dadbmi)

# Dependent var
y <- as.data.frame(children_samp) %>%
  dplyr::filter(white == 1 & male == 1) %>%
  select(bmi)

x <- as.matrix(x)
y <- as.matrix(y)

QR.b <- boot.rq(cbind(1,x),y,tau=0.5, R=500, bsmethod = "xy")

z <- t(apply(QR.b$B, 2, quantile, c(0.05,0.95)))[4:5, ]

z <- z[1, ] + z[2, ]

z %>% kbl(col.names = "mombmi + dadbmi", booktabs = T) %>%
  kable_styling(position = "center")
```

| | mombmi + dadbmi |
|---|---|
| 5% | 0.1811684 |
| 95% | 0.4398492 |

For a 90% confidence interval of a one unit increase in both mother's BMI and father's BMI, the lower bound is 0.198 and the upper bound is 0.437.

**D.iii**

```
# Rerun
bmi_x_reg_quant_2 <- rq(bmi ~ educ + age + mombmi + dadbmi,
                        tau = 0.5,
                        data = children_samp,
                        subset = (male == 1 & white == 1))

# Summarize with 500 replications
summary(bmi_x_reg_quant_2, se = "boot", R = 500, bsmethod = "xy")
```

```
##
## Call: rq(formula = bmi ~ educ + age + mombmi + dadbmi, tau = 0.5, data = children_samp,
##     subset = (male == 1 & white == 1))
##
## tau: [1] 0.5
##
## Coefficients:
##             Value   Std. Error t value Pr(>|t|)
## (Intercept) 7.68132 1.69557    4.53022 0.00001
## educ        0.05393 0.09445    0.57096 0.56819
## age         0.33584 0.07224    4.64926 0.00000
## mombmi      0.13153 0.03064    4.29212 0.00002
## dadbmi      0.17773 0.04228    4.20373 0.00003
```

The standard error's and t-statistics are different from before because we are resampling. There is a stochastic component (unless we set a seed) that will result in different samples being estimated and with slightly different underlying distributions for the estimators.

## Part E

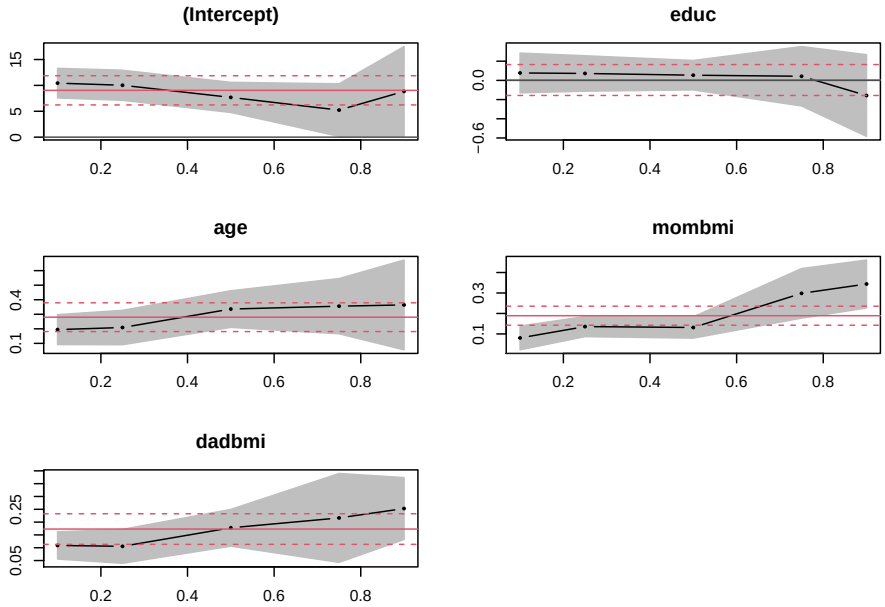
**E.i**

```
# Quantile reg at median (LAD)
bmi_x_reg_quant_5e3 <- rq(bmi ~ educ + age + mombmi + dadbmi,
                        tau = c(0.1, 0.25, 0.5, 0.75, 0.9),
                        data = children_samp,
                        subset = (male == 1 & white == 1))

# Summarize with 500 replications
summ_bmi_x_reg_quant_5e3 <- summary(bmi_x_reg_quant_5e3, se = "boot", R = 500, bsmethod = "xy")

# return summary
summ_bmi_x_reg_quant_5e3
```

```
##
```

```
## Call: rq(formula = bmi ~ educ + age + mombmi + dadbmi, tau = c(0.1,
##     0.25, 0.5, 0.75, 0.9), data = children_samp, subset = (male ==
##     1 & white == 1))
##
## tau: [1] 0.1
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept) 10.43543  1.73957    5.99887  0.00000
## educ         0.07728  0.12598    0.61345  0.53976
## age          0.19465  0.06249    3.11492  0.00191
## mombmi       0.08042  0.03528    2.27951  0.02291
## dadbmi       0.10883  0.03228    3.37152  0.00079
##
## Call: rq(formula = bmi ~ educ + age + mombmi + dadbmi, tau = c(0.1,
##     0.25, 0.5, 0.75, 0.9), data = children_samp, subset = (male ==
##     1 & white == 1))
##
## tau: [1] 0.25
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept) 10.02518  1.76628    5.67588  0.00000
## educ         0.07207  0.11241    0.64116  0.52161
## age          0.20896  0.07186    2.90792  0.00374
## mombmi       0.13588  0.02988    4.54753  0.00001
## dadbmi       0.10553  0.04027    2.62037  0.00896
##
## Call: rq(formula = bmi ~ educ + age + mombmi + dadbmi, tau = c(0.1,
##     0.25, 0.5, 0.75, 0.9), data = children_samp, subset = (male ==
##     1 & white == 1))
##
## tau: [1] 0.5
##
## Coefficients:
##             Value   Std. Error t value Pr(>|t|)
## (Intercept) 7.68132 1.77137    4.33637 0.00002
## educ        0.05393 0.09321    0.57856 0.56306
## age         0.33584 0.07663    4.38276 0.00001
## mombmi      0.13153 0.03168    4.15136 0.00004
## dadbmi      0.17773 0.04362    4.07445 0.00005
##
## Call: rq(formula = bmi ~ educ + age + mombmi + dadbmi, tau = c(0.1,
##     0.25, 0.5, 0.75, 0.9), data = children_samp, subset = (male ==
##     1 & white == 1))
##
## tau: [1] 0.75
##
## Coefficients:
##             Value   Std. Error t value Pr(>|t|)
## (Intercept) 5.22818 3.08721    1.69349 0.09077
## educ        0.04277 0.18819    0.22728 0.82027
## age         0.35572 0.11546    3.08091 0.00214
## mombmi      0.29822 0.07367    4.04823 0.00006
```

```
## dadbmi      0.21623 0.10547    2.05009 0.04070
##
## Call: rq(formula = bmi ~ educ + age + mombmi + dadbmi, tau = c(0.1,
##     0.25, 0.5, 0.75, 0.9), data = children_samp, subset = (male ==
##     1 & white == 1))
##
## tau: [1] 0.9
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  8.84646  5.26132    1.68142  0.09309
## educ        -0.15774  0.25878   -0.60954  0.54235
## age          0.36429  0.18770    1.94082  0.05265
## mombmi       0.34385  0.07120    4.82938  0.00000
## dadbmi       0.25295  0.07343    3.44487  0.00060
```




Yes, there are differences. The covariates `mombmi` and `dadbmi` are both increasing as the quantiles increase. And, on the outer quantiles they are out the range of the OLS confidence intervals for their estimates (given by the red dashed lines). The covariate `educ` seems fairly consistent with no statistically significant effect across the quantiles. The covariate `age` has an increasing effect in magnitude and is statistically significant across quantiles.

### E.ii

Yes, given the change in the coefficients and the changing intervals around the coefficient estimates I would presume that their is heteroskedasticity in the OLS model in part (c). We can test this directly.

```
# Breusch-Pagan test on Part (c) model
bptest(bmi_x_reg)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bmi_x_reg
## BP = 29.505, df = 4, p-value = 6.171e-06
```

We do find evidence to reject the null hypothesis of homoskedasticity at the 1% level of significance.

**E.iii**

```
children_samp_wm <- children_samp %>% subset(male == 1 & white == 1)

model_e10 <- rq(bmi ~ educ + age + mombmi + dadbmi, tau = 0.10,
                data = children_samp_wm)
model_e25 <- rq(bmi ~ educ + age + mombmi + dadbmi, tau = 0.25,
                data = children_samp_wm)
model_e50 <- rq(bmi ~ educ + age + mombmi + dadbmi, tau = 0.50,
                data = children_samp_wm)
model_e75 <- rq(bmi ~ educ + age + mombmi + dadbmi, tau = 0.75,
                data = children_samp_wm)
model_e90 <- rq(bmi ~ educ + age + mombmi + dadbmi, tau = 0.90,
                data = children_samp_wm)

anova.rq(model_e10, model_e25, model_e50, model_e75, model_e90, joint = FALSE, R = 500)
```

```
## Warning in summary.rq(x, se = se, R = R, covariance = TRUE): 2 non-positive fis
```

```
## Quantile Regression Analysis of Deviance Table
##
## Model: bmi ~ educ + age + mombmi + dadbmi
## Tests of Equality of Distinct Slopes: tau in {  0.1 0.25 0.5 0.75 0.9  }
##
##        Df Resid Df F value    Pr(>F)
## educ    4     3846  0.3627 0.8353201
## age     4     3846  1.1960 0.3103874
## mombmi  4     3846  5.2697 0.0003128 ***
## dadbmi  4     3846  1.6682 0.1544565
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We fail to reject the null hypothesis that age is the same value for all five slopes with a p-value of 0.31.

**E.iv**

```
anova.rq(model_e50, model_e90, joint = FALSE, R = 500)
```

```
## Quantile Regression Analysis of Deviance Table
##
## Model: bmi ~ educ + age + mombmi + dadbmi
## Tests of Equality of Distinct Slopes: tau in {  0.5 0.9  }
##
##        Df Resid Df F value    Pr(>F)
## educ    1     1539  1.0617 0.3030001
## age     1     1539  0.0459 0.8303651
## mombmi  1     1539 14.9350 0.0001159 ***
## dadbmi  1     1539  1.1807 0.2773867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have evidence to reject the null hypothsis that the mombmi coefficient is the same for the 50% and 90% quantiles at the 1% level, with a p value of 0.00012. However, we fail to reject the null hypothesis that the dadbmi coefficient is the same for the 50% and 90% quantiles.

**E.v**

```
# 10 and 90 quantile estimates
model_e5 <- rq(bmi ~ educ + age + mombmi + dadbmi, tau = c(0.10, 0.90),
               data = children_samp_wm)

# define prediction data of means
x_means <- data.frame(
  educ = mean(children_samp_wm$educ),
  age = mean(children_samp_wm$age),
  mombmi = mean(children_samp_wm$mombmi),
  dadbmi = mean(children_samp_wm$dadbmi)
)

# create prediction interval
predict.rqs(model_e5, newdata = x_means, level = 0.9)
```

```
##   tau= 0.1 tau= 0.9
## 1 20.73985 30.82972
```

When comparing to Table 1, in question 5A, the prediction interval using the means is slightly tighter than the unconditional interval. However, the change is only to a few decimal places so they are largely the same.

# Question 6

## Part A

```
# How many never recieved a voucher?
no_voucher <- voucher %>%
  dplyr::filter(selectyrs == 0) %>%
  nrow()

print(paste(no_voucher, "students never recieved a voucher."))
```

```
## [1] "468 students never recieved a voucher."
```

```
# How many had a voucher for all four years?
four_voucher <- voucher %>%
  dplyr::filter(selectyrs == 4) %>%
  nrow()

print(paste(four_voucher, "students recieved a voucher for all four years."))
```

```
## [1] "108 students recieved a voucher for all four years."
```

```
# How many students actually attended a choice school for four years?
four_choice <- voucher %>%
  dplyr::filter(choiceyrs == 4) %>%
  nrow()

print(paste(four_choice, "students attended a choice school for all four years."))
```

```
## [1] "56 students attended a choice school for all four years."
```

## Part B

```
reg_6b <- lm(choiceyrs ~ selectyrs, data = voucher)
summary(reg_6b)
```

```
##
## Call:
## lm(formula = choiceyrs ~ selectyrs, data = voucher)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08725 -0.01992 -0.01992  0.21325  1.21325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01992    0.02461   0.809    0.419
## selectyrs    0.76683    0.01259  60.931   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.576 on 988 degrees of freedom
## Multiple R-squared:  0.7898, Adjusted R-squared:  0.7896
## F-statistic:  3713 on 1 and 988 DF,  p-value: < 2.2e-16
```

Yes, these variables are related in the direction that I would expect. That is, an increase in years receiving a voucher are associated with an increase in years attending a choice school. This relationship is significant at the 1% level. Since `selectyrs` is random and moves in the same direction, it is a sensible IV candidate for `choiceyears`.

## Part C

```
reg_6c <- lm(mnce ~ choiceyrs, data = voucher)
summary(reg_6c)
```

```
##
## Call:
## lm(formula = mnce ~ choiceyrs, data = voucher)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.234 -13.234   0.603  12.766  60.114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.2344     0.8507  54.348  < 2e-16 ***
## choiceyrs    -1.8370     0.5255  -3.495 0.000494 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.75 on 988 degrees of freedom
## Multiple R-squared:  0.01222,    Adjusted R-squared:  0.01122
## F-statistic: 12.22 on 1 and 988 DF,  p-value: 0.0004943
```

The result is not what I expected. I would have assumed that the effect of going to choice schools would have a positive effect on exam scores. Here, the effect is negative. An increase in years at a choice school

are associated with a lower exam score. However, adding `black`, `hispanic`, and `female`, we see:

```
reg_6c2 <- lm(mnce ~ choiceyrs + black + hispanic + female, data = voucher)
summary(reg_6c2)
```

```
##
## Call:
## lm(formula = mnce ~ choiceyrs + black + hispanic + female, data = voucher)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.122 -12.507   0.108  12.156  60.156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.1219     1.6567  34.479  < 2e-16 ***
## choiceyrs    -0.5652     0.5307  -1.065    0.287
## black       -16.0174     1.7944  -8.926  < 2e-16 ***
## hispanic    -13.4029     2.3168  -5.785 9.73e-09 ***
## female        1.3527     1.2758   1.060    0.289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.99 on 985 degrees of freedom
## Multiple R-squared:  0.08677,    Adjusted R-squared:  0.08307
## F-statistic:  23.4 on 4 and 985 DF,  p-value: < 2.2e-16
```

For females (`female == 1`), their average exam score is higher than male counterparts. No surprise there. For Black and Hispanic students (`black == 1` and `hispanic == 1`), the partial effect is negative and large in magnitude. The partial effect of `choiceyrs` is much smaller when controlling for race and sex.

## Part D

There may be unobservables contained in the error term, such as family income or "education enthusiasm", that lead to higher exam scores that appear via the covariate `choiceyrs` and are not independent. There is not sufficient randomness (selection bias) in who attends choice schools that endogeneity may be a problem in the current model.

## Part E

```
voucher_iv1 <- ivreg(
    mnce ~ choiceyrs + black + hispanic + female |
      black + hispanic + female + selectyrs,
    data = voucher
    )

summary(voucher_iv1)
```

```
##
## Call:
## ivreg(formula = mnce ~ choiceyrs + black + hispanic + female |
##     black + hispanic + female + selectyrs, data = voucher)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -56.0680 -12.5098  -0.0476  12.0769  59.2142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.0680     1.6577  34.426  < 2e-16 ***
## choiceyrs    -0.2413     0.6053  -0.399    0.690
## black       -16.3169     1.8148  -8.991  < 2e-16 ***
## hispanic    -13.7754     2.3412  -5.884 5.49e-09 ***
## female        1.3197     1.2763   1.034    0.301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.99 on 985 degrees of freedom
## Multiple R-Squared: 0.08643, Adjusted R-squared: 0.08272
## Wald test: 23.15 on 4 and 985 DF,  p-value: < 2.2e-16
```

We estimate `mnce` using `selectyrs` as an instrumental variable for `choiceyrs`. The sign of `choiceyrs` is still negative, however the magnitude of the point estimate of the coefficient of the covariate is smaller and not statistically significant.

The coefficients on the other explanatory variables are almost identical.

## Part F

```
# add mnce90 iv
voucher_iv2 <- ivreg(
    mnce ~ choiceyrs + black + hispanic + female + mnce90 |
      black + hispanic + female + mnce90 + selectyrs,
    data = voucher)

summary(voucher_iv2)
```

```
##
## Call:
## ivreg(formula = mnce ~ choiceyrs + black + hispanic + female +
##     mnce90 | black + hispanic + female + mnce90 + selectyrs,
##     data = voucher)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.171 -11.160   1.067  10.883  50.802
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.53886    3.64548   5.908 8.79e-09 ***
## choiceyrs    1.79938    0.86019   2.092 0.037236 *
## black       -9.06711    2.57142  -3.526 0.000483 ***
## hispanic    -5.00373    3.39279  -1.475 0.141240
## female      -1.02048    1.78630  -0.571 0.568205
## mnce90       0.62881    0.04874  12.901  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.12 on 322 degrees of freedom
## Multiple R-Squared: 0.4173,  Adjusted R-squared: 0.4082
```

```
## Wald test: 47.64 on 5 and 322 DF,  p-value: < 2.2e-16
```

```
# add mnce90 ols
voucher_ols <- lm(mnce ~ choiceyrs + black + hispanic + female + mnce90,
                  data = voucher)

summary(voucher_ols)
```

```
##
## Call:
## lm(formula = mnce ~ choiceyrs + black + hispanic + female + mnce90,
##     data = voucher)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.921 -11.669   0.773  10.686  50.838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.1529     3.6204   6.119 2.73e-09 ***
## choiceyrs     0.4106     0.7359   0.558  0.57726
## black        -8.3052     2.5461  -3.262  0.00123 **
## hispanic     -4.1050     3.3624  -1.221  0.22303
## female       -0.8829     1.7760  -0.497  0.61945
## mnce90        0.6204     0.0484  12.817  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.03 on 322 degrees of freedom
##   (662 observations deleted due to missingness)
## Multiple R-squared:  0.4237, Adjusted R-squared:  0.4147
## F-statistic: 47.34 on 5 and 322 DF,  p-value: < 2.2e-16
```

The $\beta_1$ coefficient differs greatly between the IV regression ($= 1.799$) and the OLS regression ($= 0.411$). For the IV estimate, each year in a choice school is worth 1.799 percentage points on the math percentile score. This result is significant at the 5% level. This is not practically a large effect. Two percentage points is not a lot.

## Part G

The analysis from part (f) was not terribly convincing. Controlling for `mnce` dropped our number of observations by about 70%. This drop occurs because we are controlling for 1990 scores and the IV was not in effect at this time. This decreased variation will increase our standard errors and make our interval estimates wider.

## Part H

```
voucher_iv3 <- ivreg(
    mnce ~ choiceyrs1 + choiceyrs2 + choiceyrs3 + black + hispanic + female |
      black + hispanic + female + selectyrs1 + selectyrs2 + selectyrs3,
    data = voucher
    )

summary(voucher_iv3)
```

```
##
## Call:
## ivreg(formula = mnce ~ choiceyrs1 + choiceyrs2 + choiceyrs3 +
##     black + hispanic + female | black + hispanic + female + selectyrs1 +
##     selectyrs2 + selectyrs3, data = voucher)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -56.02503 -12.91203   0.03358  12.09604  58.08797
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.96642    1.66194  34.277  < 2e-16 ***
## choiceyrs1    0.05861    2.38696   0.025    0.980
## choiceyrs2    0.47336    4.12922   0.115    0.909
## choiceyrs3   -4.60395    3.86407  -1.191    0.234
## black       -16.05439    1.88756  -8.505  < 2e-16 ***
## hispanic    -13.20877    2.48859  -5.308 1.37e-07 ***
## female        1.38691    1.27890   1.084    0.278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.01 on 983 degrees of freedom
## Multiple R-Squared: 0.08599, Adjusted R-squared: 0.08041
## Wald test: 15.65 on 6 and 983 DF,  p-value: < 2.2e-16
```

## Part I

```
voucher_iv4 <- ivreg(
    mnce ~ choiceyrs + black + hispanic + female |
      black + hispanic + female + selectyrs1 + selectyrs2 + selectyrs3,
    data = voucher
    )

summary(voucher_iv4)
```

```
##
## Call:
## ivreg(formula = mnce ~ choiceyrs + black + hispanic + female |
##     black + hispanic + female + selectyrs1 + selectyrs2 + selectyrs3,
##     data = voucher)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -56.217731 -12.591599  -0.003092  12.996908  61.832299
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   57.218      1.666  34.344  < 2e-16 ***
## choiceyrs     -1.141      1.133  -1.007    0.314
## black        -15.485      2.020  -7.666 4.25e-14 ***
## hispanic     -12.740      2.588  -4.922 1.00e-06 ***
## female         1.411      1.281   1.102    0.271
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20 on 985 degrees of freedom
## Multiple R-Squared: 0.08568, Adjusted R-squared: 0.08197
## Wald test: 23.34 on 4 and 985 DF,  p-value: < 2.2e-16
```

The $\beta_1$ coefficient changes from $-0.24$ to $-1.14$. The standard error increases by a similar magnitude as well.

```
formula1 <- mnce ~ choiceyrs + black + hispanic + female
formula2 <- ~ black + hispanic + female + selectyrs1 + selectyrs2 + selectyrs3

summary(gmm(formula1, formula2, data = voucher))
```

```
##
## Call:
## gmm(g = formula1, x = formula2, data = voucher)
##
##
## Method:  twoStep
##
## Kernel:  Quadratic Spectral(with bw =  0.77297 )
##
## Coefficients:
##              Estimate      Std. Error    t value       Pr(>|t|)
## (Intercept)    5.7316e+01    1.8881e+00    3.0356e+01   2.0960e-202
## choiceyrs     -1.0888e+00    1.1560e+00   -9.4186e-01    3.4627e-01
## black         -1.5639e+01    2.1995e+00   -7.1104e+00    1.1573e-12
## hispanic      -1.2803e+01    2.4708e+00   -5.1818e+00    2.1977e-07
## female         1.3518e+00    1.2019e+00    1.1247e+00    2.6071e-01
##
## J-Test: degrees of freedom is 2
##                 J-test   P-value
## Test E(g)=0:    0.76775  0.68122
##
## Initial values of the coefficients
## (Intercept)   choiceyrs       black    hispanic      female
##   57.217731   -1.141299  -15.484833  -12.740396    1.411493
```

The p-value of the Sargan's J test is 0.6812. Thus we fail to reject the null hypothesis that the over-identifying restrictions are valid.

## Part J

If I wanted to predict `mnce` based on the explanatory variables in part (d) and also the `selectyrs` variable, but was not interested in the causal effect of `choiceyrs`, I would run a regression like the following:

$$\texttt{mnce} = \beta_1 + \beta_2 \texttt{black} + \beta_3 \texttt{hispanic} + \beta_4 \texttt{female} + \beta_5 \texttt{selectyrs} + u.$$

This is the reduced form.