Assignment 2

Question 1 (*Appendix: Figures 1.0 and onwards*)

Question 2 (*Appendix: Figures 2.0 and onwards*)

Legend	
Tchratio	Students per teacher
Pctel	Share of English second language students in the district
Lnch_pct	Share of students eligible for free lunch
Percap	Per capita income
Percap2	Per capita income squared
Percap3	Per capita income cubed

Legend
tchratio = students-teacher ratio
pctel = share of English second language students in the district
lnch_pct = % eligible for free lunch
percap = per capita income
percap2 = per capita income squared
percap3 = per capita income cubed

Grade 4 & 8 students have some distinct differences along with some similarities. Notably, the regressors explain the testscr values for the grade 8 better. It would explain the significance of the variables and their relation to the age of the student. This can be proven by the higher adj. r sq value for model5. The following relationship will be based on an endogenous relationship between age and independent variables.

Grade 4	**Grade 8**
- With 1 more student per teacher, students' scores increased by *744.025*. Highly sensitive variable. (Significant t value of *-2.387*, depicting that younger students **do not mind** a high tchratio value) - ~~A **negative** relationship with per capita income (score decreases by *3.067* with an increase of 1% of per capita income)~~	- With 1 more student per teacher, students' scores increased by *676.774*. Slightly less sensitive variable. (Students prefer a **smaller** tchratio value, t val. = *-1.254*) - ~~A **positive** relationship with per capita income (score increases by *1.932* with an increase of 1% of per capita income)~~

- Scores decrease by *0.437* scores with a 1 % increase in English learners - Another notable discovery was the impact of the variable '% of free lunches.' (t value of lnch_pct was -*5.976*) - This would imply that the variable lmch_pct is an outlier as it is **significantly different** in models 4&5 as compared to the population dataset. - ~~There seems to be evidence of a **linear** relationship between test scores and per capita income due to percap having the most significant **p-value** (*0.194*) within the model. The exponential variables for per capita income has significant t-values which aims to shed light on the model having a **non-linear** rather than a linear relation. This only applies to grade 4 testscores however.~~	- Scores decreased by *0.353* scores with a 1 % increase in English learners (The fact this result is similar for both grades implies being fluent in English leads to higher test scores.) - **t value** of lnch_pct was -*6.524*. This is a fairly significant number, one that rejects the null hypothesis which states that the difference in the means of the two data sets (MASS_small, MASS) is 0. - ~~There is strong evidence of an **exponential** relationship between test scores and per capita income due to percap2 (*0.843*) and percap3 (*0.586*) having significant **p-values**.~~

Conclusion:

- The same monetary and knowledge-based regressors show a causal effect on the predicted variables in both models differently.
- Determining if external validity exists:
    - The **External validity** question for this exercise is whether our results for fourth graders can be generalized to older students, like 8th graders: we see some evidence against this given that the variables which matter most are different in the two regressions, and especially from the fact that STR doesn't seem to matter for older kids.
    - This assumption is **not coherent** for these two models due to several variables having significant t values greater than *1.96*. (lnch_pct for both models, tchratio for model4, percap2&3 for model4)
    - Moreover, the p-values of percap, percap2, and percap3 for both models show empirical evidence for different types of relationships between the dependent and independent variables for each of the models. This would indicate that the sample variance is very different from the actual population dataset, thus rendering the generalization principle moot.
    - Also, the hypothesis that the true coefficient of tchratio is zero was accepted at the 1% significance level, even after adding variables that control student background and district-wide economic characteristics.

Question 3 (*Appendix: Figures 3.0 and onwards*)

This model scrutinized technology by finding a relationship of log of income per capita by state and the share of English as second language learners on computers per student.

Basic Analysis (*Appendix: Model6*):

- With a 1% increase in the log function of the per capita income, computers per student increases by *0.012* points. The effect of log(income) is insignificant as a determinant of comp_stu in this regression.
- Dummy variable for location is either equal to zero (when the observation is from Mass) or one (when the observation is from Cali) of the per capita income dataset, where computers per student increase marginally by *0.016* points for the California dataset.
- A marginal increase in the share of English second language students in the district decreased the computers per student by *0.001*.

Internal Validity Assumptions:

A study has internal validity if the statistical inferences about causal effects drawn from the study are valid for the population being studied.

The hypothesis that the true coefficient of log_avginc is zero was accepted at the 1%, 3%, and 5% significance levels, even after adding a variable that controls student background and district-wide economic characteristics. Since the p-value of *0.181*, or $p > 0.05$, we assume the relationship of the regressor on the predicted variable is a linear one.

We can further improve this model by adding the variable lnch_pct. (*Appendix: Model62*) It was concluded that the addition of this variable ended up increasing the coefficient of log_avginc and decreasing the coefficient of the dummy variable. Therefore a causal effect of log_avginc on comp_stu can be assumed to be endogenous. This would mean that the new model would not favor the external validity of our results.

The above-mentioned factors pose a threat to the internal validity of the model as it is clear that omitted variable bias exists.

We could test the dummy variable and if the per capita income is the same in both states (Cali, Mass). This result would help us understand the state-wide impact on the standard of living of a student. However, like the models for question 3, this proposed model also poses a threat to the internal validity of the model due to the errors-in-variables bias. Where the income per capita is a district-wide measure, the student who is a participant in the study might not reflect the measured avginc value. This is a complicated non-classical type of error. Ideally, the income per capita data would be individualized to fit the students better.

Question 4 (*Appendix: Figures 4.0 and onwards*)

Legend
Y = test score
X = log of per capita income
Z = % of free lunch meals

By running the 2SLS by hand model, we split the regressor (X) into two parts (first and second stage). One that might be correlated with $U$and a part that is not. This is done by using an instrument variable, Z, that is considered an exogenenous variable, X variable being the endogenous variable.

1) Determining if an instrument is weak
   a) An instrument is valid if it is
      i) Relevant: $corr(Z, X) \neq 0$
      ii) Exogenous: $corr(Z, U) = 0$

To be able to determine whether an instrument is weak and explain what this means to identify the effects of weak instruments on a model's results, it needs to satisfy these two conditions to have a valid instrument; first, the instrument must be relevant meaning that it must be correlated with the endogenous independent variable, it must also be exogenous meaning it is not correlated with the model's error term (*another way of thinking about the exogeneity condition is that it the instrument must not have a direct impact on the dependent variable*).

- The Relevanance assumption: $corr(Z, X) \neq 0$ condition is checked by running the cor.test(Z,X). This assumption holds as the $corr(Z, X) \neq 0$.
- Exogeneity assumption:
  - First Stage: Isolate the part of X that is uncorrelated with the error term μ by regressing X on Z.

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

  - Because Z_i is uncorrelated with μ_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with μ_i.This **is** the case here, as the mean value for predicted variable and the log_avginc variable is the same.

The instrument variable 'Z' **does not** have have a direct impact on the dependent variable Y, and therefore is considered as a valid instrument.

2) The primary reason why we use 2SLS over OLS is since log(income) is endogenous, therefore breaking it down using the two stage least squares method into an endogenous and exogenous eradicating bias.

Figure 1 and onwards

```r
#Moh Jaiswal, 500916860, Assignment #2, Prof. Turner
#Column 5

rm(list=ls())


library(haven)
MASS <- read_dta("MASS.dta")
View(MASS)
STAR_small <- read_dta("STAR_small.dta")
View(STAR_small)

MASS$log_percap <- log(MASS$percap)
STAR_small$log_avginc <- log(STAR_small$avginc)

MASS_sample = subset(MASS, select = c(totsc4, tchratio, pctel, lnch_pct, log_percap))
STAR_sample = subset(STAR_small, select = c(testscr, str, el_pct, meal_pct, log_avginc))


MASS_sample$meal_pct <- MASS_sample$lnch_pct
MASS_sample$lnch_pct <- NULL

MASS_sample$testscr <- MASS_sample$totsc4
MASS_sample$totsc4 <- NULL

MASS_sample$str <- MASS_sample$tchratio
MASS_sample$tchratio <- NULL

MASS_sample$log_avginc <- MASS_sample$log_percap
MASS_sample$log_percap <- NULL

MASS_sample$el_pct <- MASS_sample$pctel
MASS_sample$pctel <- NULL

#Combine data sets
STAR_sample$state = "Cali"
MASS_sample$state = "Mass"

schooldata <- rbind(MASS_sample, STAR_sample)

rm(STAR,MASS,STAR_small,MASS_small)

# Let's replicate column 5 from the given table
# We will run versions for both Mass and Cali combined, for Mass only, and for Cali only.
```

MASS_sample	220 obs. of 6 variables
schooldata	640 obs. of 6 variables
STAR_sample	420 obs. of 6 variables

```r
#run regression
model <- lm(testscr ~ str + el_pct + meal_pct + log_avginc, data = schooldata)
model2 <- lm(testscr ~ str + el_pct + meal_pct + log_avginc, data = subset(schooldata,
                                                                  state=="Mass"))
model3 <- lm(testscr ~ str + el_pct + meal_pct + log_avginc, data = subset(schooldata,
                                                                  state=="Cali"))

library(sandwich)
library(lmtest)
library(jtools)
summ(model, digits=3, robust = "HC1")
summ(model2, digits=3, robust = "HC1")
summ(model3, digits=3, robust = "HC1")
```

```
> summ(model, digits=3, robust = "HC1")
MODEL INFO:
Observations: 640
Dependent Variable: testscr
Type: OLS linear regression

MODEL FIT:
F(4,635) = 462.994, p = 0.000
R² = 0.745
Adj. R² = 0.743

Standard errors: Robust, type = HC1
-----------------------------------------------------
                       Est.      S.E.     t val.        p
----------------- ---------- --------- ---------- -------
(Intercept)         755.325    11.397     66.274    0.000
str                  -3.337     0.312    -10.699    0.000
el_pct               -0.182     0.053     -3.459    0.001
meal_pct             -0.752     0.050    -15.116    0.000
log_avginc            3.250     2.874      1.131    0.259
-----------------------------------------------------
```

```
> summ(model2, digits=3, robust = "HC1")
MODEL INFO:
Observations: 220
Dependent Variable: testscr
Type: OLS linear regression

MODEL FIT:
F(4,215) = 112.284, p = 0.000
R² = 0.676
Adj. R² = 0.670

Standard errors: Robust, type = HC1
----------------------------------------------------
                       Est.     S.E.    t val.       p
----------------- --------- -------- -------- -------
(Intercept)         682.432   11.497   59.356   0.000
str                  -0.689    0.270   -2.553   0.011
el_pct               -0.411    0.306   -1.341   0.181
meal_pct             -0.521    0.078   -6.715   0.000
log_avginc           16.529    3.146    5.255   0.000
----------------------------------------------------
```

```
> summ(model3, digits=3, robust = "HC1")
MODEL INFO:
Observations: 420
Dependent Variable: testscr
Type: OLS linear regression

MODEL FIT:
F(4,415) = 405.359, p = 0.000
R² = 0.796
Adj. R² = 0.794

Standard errors: Robust, type = HC1
----------------------------------------------------
                       Est.    S.E.     t val.       p
----------------- --------- ------- ---------- -------
(Intercept)         658.552   8.642     76.208   0.000
str                  -0.734   0.257     -2.860   0.004
el_pct               -0.176   0.034     -5.215   0.000
meal_pct             -0.398   0.033    -12.004   0.000
log_avginc           11.569   1.819      6.361   0.000
```

Figure 2 and onwards

MASS	220 obs. of 19 variables
MASS_small	220 obs. of 8 variables
model4	List of 12
model5	List of 13

```r
#Question 2

rm(list=ls())

library(haven)
MASS <- read_dta("MASS.dta")
View(MASS)

MASS$percap2 = MASS$percap^2
MASS$percap3 = MASS$percap^3

MASS_small = subset(MASS, select = c(totsc4, totsc8, tchratio, pctel, lnch_pct, percap,
                     percap2, percap3))
View(MASS_small)

#run regressions

model4 <- lm(totsc4 ~ tchratio + pctel + lnch_pct + percap + percap2 + percap3, data =
               MASS_small)
model5 <- lm(totsc8 ~ tchratio + pctel + lnch_pct + percap + percap2 + percap3, data =
               MASS_small)
summ(model4, digits=3, robust = "HC1")
summ(model5, digits=3, robust = "HC1")
```

```
> summ(model4, digits=3, robust = "HC1")
MODEL INFO:
Observations: 220
Dependent Variable: totsc4
Type: OLS linear regression

MODEL FIT:
F(6,213) = 77.232, p = 0.000
R² = 0.685
Adj. R² = 0.676

Standard errors: Robust, type = HC1
---------------------------------------------------
                       Est.      S.E.    t val.       p
----------------- --------- --------- --------- -------
(Intercept)         744.025    21.318    34.902   0.000
tchratio             -0.641     0.268    -2.387   0.018
pctel                -0.437     0.303    -1.441   0.151
lnch_pct             -0.582     0.097    -5.976   0.000
percap               -3.067     2.353    -1.304   0.194
percap2               0.164     0.085     1.918   0.056
percap3              -0.002     0.001    -2.246   0.026
```

```
> summ(model5, digits=3, robust = "HC1")
MODEL INFO:
Observations: 180 (40 missing obs. deleted)
Dependent Variable: totsc8
Type: OLS linear regression

MODEL FIT:
F(6,173) = 144.485, p = 0.000
R² = 0.834
Adj. R² = 0.828

Standard errors: Robust, type = HC1
---------------------------------------------------
                       Est.      S.E.    t val.       p
----------------- --------- --------- --------- -------
(Intercept)         676.774    23.403    28.919   0.000
tchratio             -0.342     0.272    -1.254   0.212
pctel                -0.353     0.233    -1.518   0.131
lnch_pct             -0.649     0.099    -6.524   0.000
percap                1.932     2.606     0.742   0.459
percap2               0.019     0.096     0.198   0.843
percap3              -0.001     0.001    -0.545   0.586
```

Figure 3 and onwards

```r
#Question 3

rm(list=ls())

library(haven)
MASS <- read_dta("MASS.dta")
View(MASS)
STAR_small <- read_dta("STAR_small.dta")
View(STAR_small)

# Computers per student is 1/(students per computer)

MASS$comp_stu = 1.0/MASS$s_p_c
MASS$comp_stu[!is.finite(MASS$comp_stu)] <- 0
MASS$s_p_c <- NULL

MASS$log_avginc <- log(MASS$percap)
STAR_small$log_avginc <- log(STAR_small$avginc)


MASS_sample = subset(MASS, select = c(comp_stu, log_avginc, pctel))
STAR_sample = subset(STAR_small, select = c(comp_stu, log_avginc, el_pct))

MASS_sample$el_pct <- MASS_sample$pctel
MASS_sample$pctel <- NULL

#Combine data sets
STAR_sample$state = "Cali"
MASS_sample$state = "Mass"

schooldata <- rbind(MASS_sample, STAR_sample)

#rm(STAR,MASS,STAR_small,MASS_small, model, model2, model3, model4, model5)

schooldata$dummy = ifelse(schooldata$state=="Cali", 1, 0)

model6 <- lm(comp_stu ~ log_avginc + el_pct + dummy, data = schooldata)
summ(model6, digits=3, robust = "HC1")

#Improving omitted variable bias by testing dummy variable on state wide per capita data from our model.

MASS_sample2 = subset(MASS, select = c(comp_stu, log_avginc, pctel,lnch_pct))
STAR_sample2 = subset(STAR_small, select = c(comp_stu, log_avginc, el_pct, meal_pct))

MASS_sample2$el_pct <- MASS_sample2$pctel
MASS_sample2$pctel <- NULL

STAR_sample2$lnch_pct <- STAR_sample2$meal_pct
STAR_sample2$meal_pct <- NULL

STAR_sample2$state = "Cali"
MASS_sample2$state = "Mass"

schooldata2 <- rbind(MASS_sample2, STAR_sample2)

schooldata2$dummy = ifelse(schooldata2$state=="Cali", 1, 0)
schooldata2$inter = schooldata2$dummy*schooldata2$log_avginc

model62 <- lm(comp_stu ~ log_avginc + el_pct + dummy + inter + lnch_pct, data = schooldata2)
summ(model62, digits=3, robust = "HC1")
```

```
> summ(model6, digits=3, robust = "HC1")
MODEL INFO:
Observations: 640
Dependent Variable: comp_stu
Type: OLS linear regression

MODEL FIT:
F(3,636) = 10.283, p = 0.000
R² = 0.046
Adj. R² = 0.042

Standard errors: Robust, type = HC1
-----------------------------------------------------
                      Est.    S.E.   t val.        p
----------------- -------- ------- -------- -------
(Intercept)          0.101   0.025    4.023   0.000
log_avginc           0.012   0.009    1.338   0.181
el_pct              -0.001   0.000   -4.528   0.000
dummy                0.016   0.006    2.731   0.006
```

```
> summ(model62, digits=3, robust = "HC1")
MODEL INFO:
Observations: 640
Dependent Variable: comp_stu
Type: OLS linear regression

MODEL FIT:
F(5,634) = 6.188, p = 0.000
R² = 0.047
Adj. R² = 0.039

Standard errors: Robust, type = HC1
-----------------------------------------------------
                      Est.    S.E.   t val.        p
----------------- -------- ------- -------- -------
(Intercept)          0.104   0.058    1.797   0.073
log_avginc           0.010   0.020    0.525   0.600
el_pct              -0.001   0.000   -3.702   0.000
dummy                0.002   0.059    0.026   0.979
inter                0.005   0.021    0.231   0.817
lnch_pct             0.000   0.000    0.354   0.723
```

Figure 4 and onwards

```
#Question 4

#2SLS BY HAND MODEL
rm(list=ls())


library(haven)
STAR_small <- read_dta("STAR_small.dta")
View(STAR_small)

# Let's implement 2sls

#Suppose we let Y = test score, X = log of per capita income, Z = % of free lunch meals

# Here, the idea is that lnch_pct on its own has no causal effect on Y but with the instrument for X (avginc) which is measured with error
# Since the 8th number is column 6, we only work with the Cali data/STAR_small

STAR_sample = subset(STAR_small, select = c(dist_cod,str,meal_pct,el_pct,avginc,testscr,expn_stu))

# Now let's run 2SLS "by hand".
# First stage: get the x-hats and write them to MASS_small
# To do this, run the regression as usual (with lm) then use the predict command to get the predicted values of x.

STAR_sample$log_avginc <- log(STAR_small$avginc)

firststage <- lm(log_avginc ~ meal_pct,data=STAR_sample)
STAR_sample$predicted <- predict(firststage)

summary(STAR_sample$log_avginc)
summary(STAR_sample$predicted)

# Second stage

secondstage <- lm(testscr ~ predicted, data=STAR_sample)

library(jtools)
summ(secondstage, digits=3, robust="HC1")

# Simple OLS Model

olsmodelq4 <- lm(testscr ~ log_avginc, data=STAR_sample)
summ(olsmodelq4, digits=3, robust = "HC1")
coefci(olsmodelq4, vcov = vcovHC, type="HC1", level=0.95)
coefci(olsmodelq4, vcov = vcovHC, type="HC1", level=0.90)
linearHypothesis(olsmodelq4, c("log_avginc =0"), white.adjust = "hc1")
question4b <- plot(STAR_sample$testscr,STAR_sample$log_avginc, main="Relationship between testscore and average income",
                   xlab="Income by District (Moh Jaiswal)",
                   ylab="Testscore",
                   las=1)

#Weak instrument Validity

cor.test(STAR_sample$meal_pct, STAR_sample$log_avginc)
```

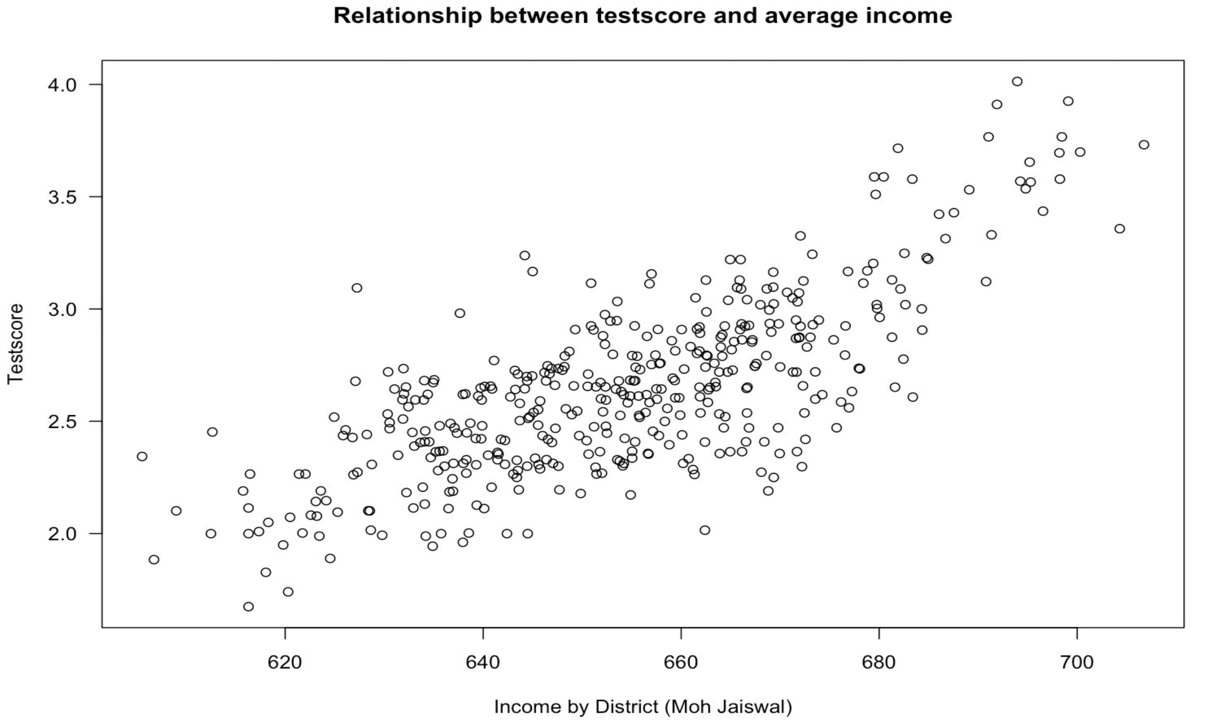
```
summary(STAR_sample$log_avginc)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.674   2.365   2.619   2.645   2.870   4.013
summary(STAR_sample$predicted)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.034   2.400   2.677   2.645   2.882   3.139
```

```
> summ(secondstage, digits=3, robust="HC1")
MODEL INFO:
Observations: 420
Dependent Variable: testscr
Type: OLS linear regression

MODEL FIT:
F(1,418) = 1286.486, p = 0.000
R² = 0.755
Adj. R² = 0.754

Standard errors: Robust, type = HC1
-----------------------------------------------------
                       Est.    S.E.    t val.       p
----------------- --------- ------- --------- -------
(Intercept)         508.090   4.264   119.171   0.000
predicted            55.227   1.629    33.905   0.000
```

```
> summ(olsmodelq4, digits=3, robust = "HC1")
MODEL INFO:
Observations: 420
Dependent Variable: testscr
Type: OLS linear regression

MODEL FIT:
F(1,418) = 537.444, p = 0.000
R² = 0.563
Adj. R² = 0.561

Standard errors: Robust, type = HC1
-----------------------------------------------------
                       Est.    S.E.    t val.       p
----------------- --------- ------- --------- -------
(Intercept)         557.832   3.840   145.271   0.000
log_avginc           36.420   1.397    26.071   0.000
```

```
> linearHypothesis(olsmodelq4, c("log_avginc =0"), white.adjust = "hc1")
Linear hypothesis test

Hypothesis:
log_avginc = 0

Model 1: restricted model
Model 2: testscr ~ log_avginc

Note: Coefficient covariance matrix supplied.

  Res.Df Df     F                Pr(>F)
1    419
2    418  1 679.7 < 0.00000000000000022 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> coefci(olsmodelq4, vcov = vcovHC, type="HC1", level=0.95)
                2.5 %    97.5 %
(Intercept) 550.28427 565.38027
log_avginc   33.67378  39.16559
> coefci(olsmodelq4, vcov = vcovHC, type="HC1", level=0.90)
                  5 %      95 %
(Intercept) 551.50210 564.16244
log_avginc   34.11681  38.72255
```




```
> cor.test(STAR_sample$meal_pct, STAR_sample$log_avginc)

	Pearson's product-moment correlation

data:  STAR_sample$meal_pct and STAR_sample$log_avginc
t = -24.2, df = 418, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8010203 -0.7208889
sample estimates:
       cor
-0.7638831
```