HW10: The Moving to Opportunity Experiment & Multiple Regression with Inference

Millions of low-income Americans live in high-poverty neighborhoods, which also tend to be racially segregated and sometimes have issues with public safety. While social scientists have long believed a lack of investment in these neighborhoods contributes to negative outcomes for the residents living in them, it is often difficult to establish a causal link between neighborhood conditions and individual outcomes. The Moving to Opportunity (MTO) demonstration was designed to test whether offering housing vouchers to families living in public housing in high-poverty neighborhoods could lead to better experiences and outcomes by providing financial assistance to move to higher income neighborhoods.

Between 1994 and 1998 the U.S. Department of Housing and Urban Development enrolled 4,604 low-income households from public housing projects in Baltimore, Boston, Chicago, Los Angeles, and New York in MTO, randomly assigning enrolled families in each site to one of three groups: (1) The low-poverty voucher group received special MTO vouchers, which could only be used in census tracts with 1990 poverty rates below 10% and counseling to assist with relocation, (2) the traditional voucher group received regular section 8 vouchers, which they could use anywhere, and (3) the control group, who received no vouchers but continued to qualify for any project-based housing assistance they were entitled to receive. Today we will use the MTO data to learn if being given the opportunity to move to lower-poverty neighborhoods improved participants' economic and subjective well-being. This exercise is based on the following article:

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, J.R.K., and Sanbonmatsu, L., 2012. "Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults." *Science*, Vol. 337, Issue 6101, pp. 1505-1510.

The file `mto3.csv` includes the following variables for 3,263 adult participants in the voucher and control groups:

Name	Description
`group`	factor with 3 levels: `lpv` (low-poverty voucher), `sec8` (traditional section 8 voucher), and `control`
`econ_ss_zcore`	Standardized measure of economic self-sufficiency, centered around the control group mean and re-scaled such that the control group mean = 0 and its standard deviation = 1. Measure aggregates several measures of economic self-sufficiency or dependency (earnings, government transfers, employment, etc.)
`crime_vic`	Binary variable, `1` if a member of that household was the victim of a crime in the six months prior to being assigned to the MTO program, `0` otherwise
`age`	Age of the head of household

The data we will use are not the original data, this dataset has been modified to protect participants' confidentiality, but the results of our analysis will be consistent with published data on the MTO demonstration. Several of the variables used in this homework are simulated data.

This homework has 75 points.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(tinytex)
```

```
mto3 <- read.csv("data3/mto3.csv")
```

Question 1 [12 points]

One of the outcomes of interest in this dataset is economic self-sufficiency. The researchers hypothesized that older heads of household would have greater economic self-sufficiency.

1a [4 points]

What linear regression model might be useful for testing this hypothesis? Write the equation for this linear regression model. What is the parameter of interest? What is the estimator (sample summary statistic) for this parameter of interest?

1b [3 points]

What are the Null and Alternative hypotheses? Please use a two-sided hypothesis test.

1c [5 points]

Consider a situation where the manager of the voucher program is planning on using the results of this hypothesis test to determine whether to change a program that is intended to improve economic self-sufficiency. The program currently is offered to all low income households. If we conclude that age is positively associated with economic self-sufficiency then they will offer the program only to households with younger heads of household and stop making it available to households with older heads of households (all across the

country). If we do not find this association they will continue to offer it to all households and also will continue to collect data on the success of the program for households with different demographics.

For this study what is a Type I error and what are its consequences? What is a Type II error and what are its consequences? What alpha level do you suggest for this hypothesis test?

Answer 1

Answer 1a

we can use a simple linear regression with one predictor. the parameter of interest coefficient on predictor age. The estimator would be the sample mean of the age.

Answer 1b

the null hypothesis : h0= 0 , B^1=0 ha=0, B^1!=0

the null would be h0 : it would be that there is no improvement in the self sufficiency scores as the age increases and therefore B^1 is zero.

the alternate would mean that there is an improvement in the self sufficiency score is non zero as age increases B^ !=0.

null and alternate hypothesis: if the conclusion of this research study is that the older heads of the household improves economic well being relative to not being an older head in the household then then they will continue the Low Poverty Voucher Program at its current size and continue to collect more data to determine what the effects are in a larger study

Answer 1c type - 1 error: Type I error is the probability of rejecting the null hypothesis when it is true; in this case, it is the probability of concluding that there is a positive association between age and the econ sufficiency outcome which means that we falsely reject the null which is that there is no change in econ self sufficiency based on age, which means now that they will offer the program to young heads at households and completely get rid of the program for older folks. This is a type - 1 error. A potential consequence is that they would not offer the program to folks that are older and disadvantaged due to the result of the study. even if there isn't a change observed in the study between both the groups, rejecting the null would mean scrapping the program for the older folks, this could unfairly benefit only one group while leaving the other group disadvantaged.

type -2 error: A Type II error is the probability of failing to reject the null hypothesis when the null hypothesis is false; in this case it is the probability of concluding that there was infact no change on the economic well being of households that were under the younger household heads relative to the older household heads even when there is change. a potential issue would be when the research study will continue the study Program across different demographics and offer it to all households until they reach a conclusive diagnosis, they will continue to measure the success of the program amongst all

households. . this is a failure to reject the null even when it is false, leading to money and time spent collecting more data.

the alpha level I recommend is 0.01 % .

Question 2 [17 pts]

2a [5 points]

Using summary statistics and a figure, describe the distribution of the economic self-sufficiency variable among everyone in the data set.

2b [5 points]

Using summary statistics and a figure, describe the distribution of the age variable among everyone in the data set.

2c [1 point]

Run the simple linear regression you describe in Question 1.

2d [6 points]

Check the three assumptions of linear regression for this model by making and then assessing the 2 or 3 standard residual plots. For each assumption state whether or not it is violated and how you can tell.

Answer 2

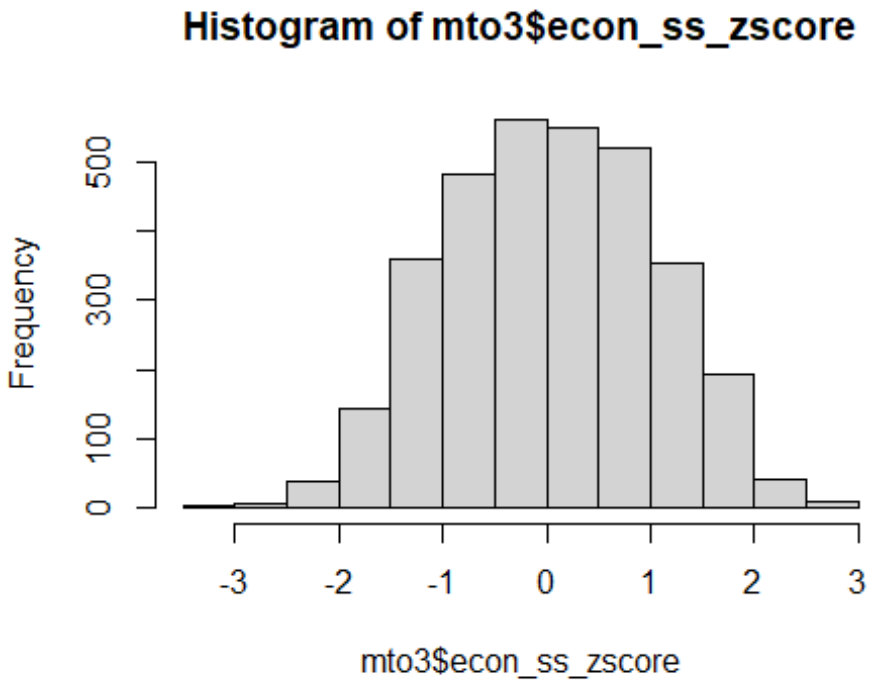
Answer 2a

It ranges from -3.23 - 2.93. the iqr is -0.726 - 0.77. It has a median of 0.02 and a mean of 0.03. the shape is bell shaped and is unimodal, there is a slight skew to the left. it has no visible outliers.

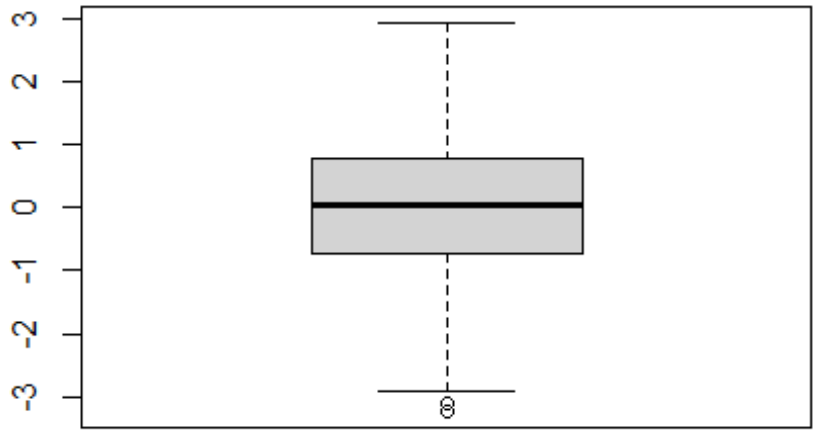
```
summary(mto3$econ_ss_zscore)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.23231 -0.72662  0.02777  0.03129  0.77849  2.93332
```

```
hist(mto3$econ_ss_zscore)
```




```r
boxplot(mto3$econ_ss_zscore)
```



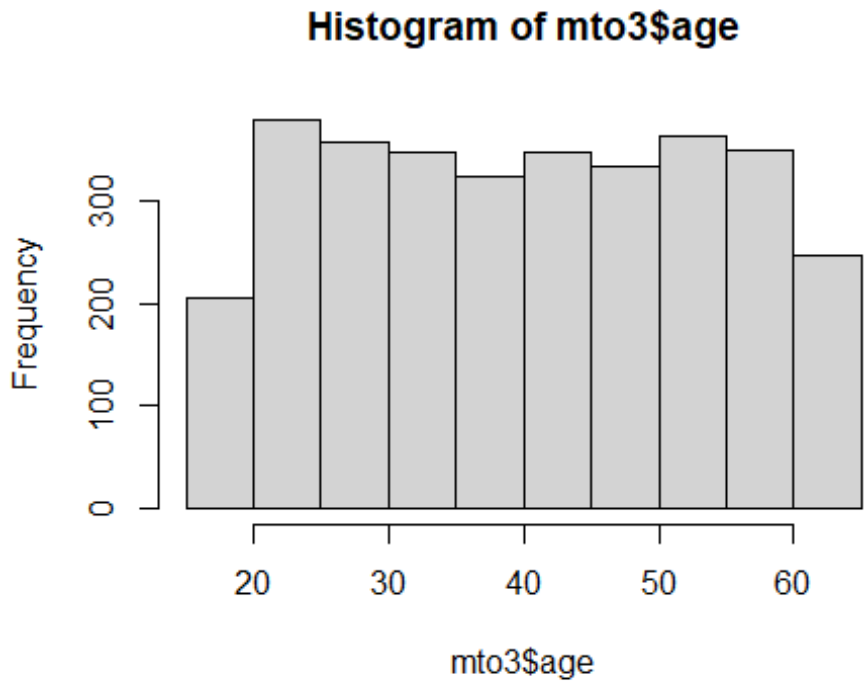
Answer 2b

the range is 18 - 64 years. the mean age is 40.61 and the median is 41. the iqr is 29 - 52.50. the distribution seems balanced and there are no visible outliers. the distribution is symmetric and is non modal uniform distribution.

```
summary(mto3$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   29.00   41.00   40.61   52.50   64.00
```

```
hist(mto3$age)
```



Answer 2c

```
reg_mod1 <- lm(econ_ss_zscore ~ age, data = mto3)
summary(reg_mod1)
```

```
##
## Call:
## lm(formula = econ_ss_zscore ~ age, data = mto3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3157 -0.3553 -0.0025  0.3593  2.1702
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -2.4540300  0.0310497  -79.04   <2e-16 ***
## age          0.0611945  0.0007254   84.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5604 on 3261 degrees of freedom
## Multiple R-squared:  0.6858, Adjusted R-squared:  0.6857
## F-statistic:  7117 on 1 and 3261 DF,  p-value: < 2.2e-16
```

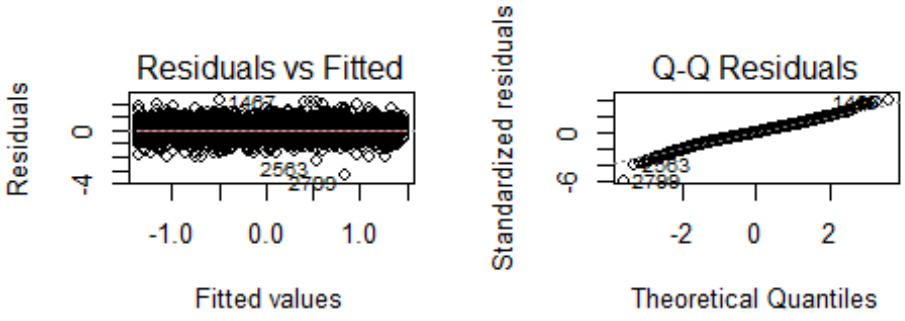
Answer 2d

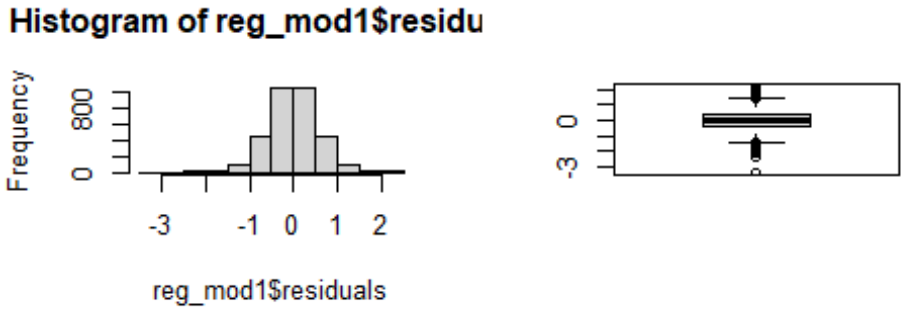
We do not see any obvious non-linearities in the residuals; the mean of the residuals appears to be near zero in all segments of the residual plot as we move from left to right so the linearity assumption likely holds.

The constant variance assumption is also likely to not be violated; the spread across the residuals seems to maintain a uniform average distribution.

The normality is valid there is a small skew and variations in the tails are minor and will not cause the inferences to be incorrect. the plot for the histogram appears to be balanced and symmetric, it does not have any gaps and reasonably problematic outliers except one.

```
par(mfrow = c(2, 2))
plot(reg_mod1, c(1, 2))
hist(reg_mod1$residuals)
boxplot(reg_mod1$residuals)
```





Answer 2e

Question 3 [12 pts]

Question 3a [5 points]

Make a scatterplot of the age and economic self-sufficiency data and add the estimated regression line to the figure. Interpret each of the following aspects of the model if they are valid to interpret: estimated y-intercept (and whether it has a real-world interpretation), estimated slope.

Question 3b [3 points]

Under the null hypothesis, describe what the sampling distribution of the estimated slope coefficient for age would be over repeated sampling, if all three assumptions of linear regression held (were not violated) - include information about the shape, mean, and (estimated) standard error of the sampling distribution.

Question 3c [4 points]

How many estimated standard errors away from the Null Value is the estimated slope coefficient? Interpret the p-value for the hypothesis test you stated in Q1b (assuming all 3 assumptions of linear regression hold). What is the conclusion of this hypothesis test at the alpha level that you recommended in Q1c (assuming all 3 assumptions of linear regression hold)?

Answer 3

Answer 3a

y intercept(econ score) : -2.45 + 0.061 ageB^1 + E^i

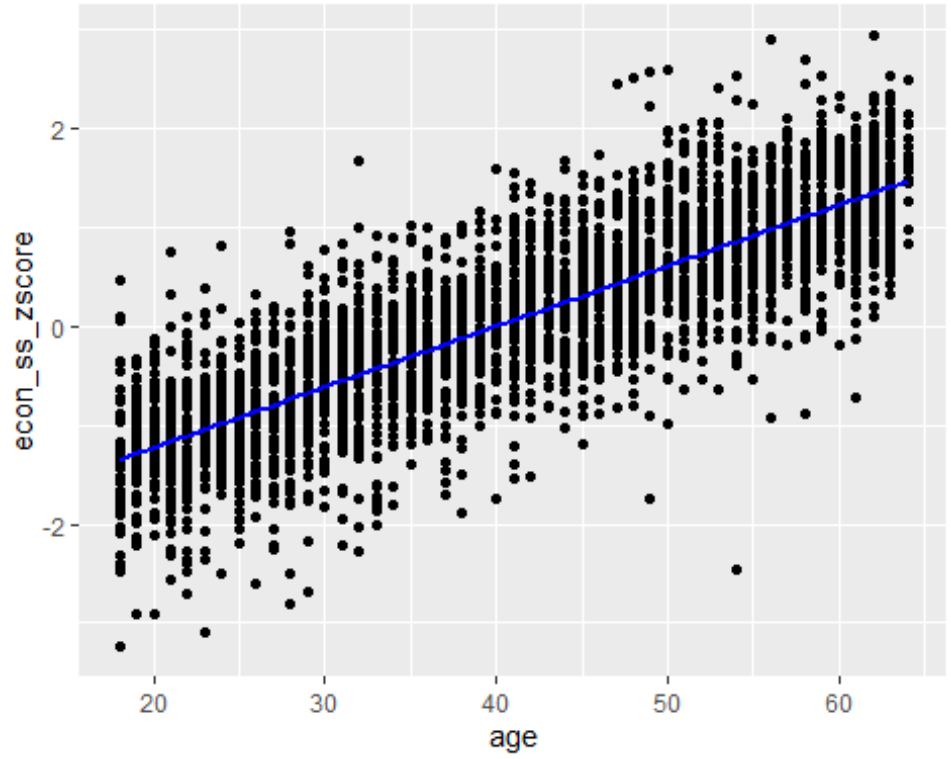
The y-intercept, -2.45, is the average economic well being score of those households aged 0 years old. This is not meaningful as this point is outside the data for age range and meaningless. It is not interpretable.

estimated slope -> 0.061 on the age variable: the slope tells us how much more change the economic well being score changes for each additional year old the household value is. therefore, for a one year increase in the age of the household head there is a 0.06 increase in average economic well being score for the households. In this case, for a 10 year increase in the age of household head on average the household economic score increases by 0.6 points. ### Answer 3b Sampling distribution of the β^1(slope on the age variable) under the Null Hypothesis: The shape of the sampling distribution for β^1 over repeated sampling is a t-distribution with 3261 degrees of freedom which approximates to a normal distribution. The mean of this sampling distribution under the null hypothesis is zero. The estimated standard error of this sampling distribution is 0.0007. over repeated sampling, this would be a normal distribution, considering the size is >500 and also there are no outliers and the parameter of interest is a mean. ### Answer 3c The observed value of β^1 is 0.06 which is difficult to put on the sketch of the sampling distribution under the null hypothesis as it is 84.36 standard errors above the assumed null value. The parts of the sampling distribution that would be shaded and correspond to the p-value, are the areas

under the curve from 0.06 to the right and from -0.06 to the left. p-value for the slope coefficient on x1: The p-value is very small (much smaller than any standard alpha level=0.01) so we reject the Null Hypothesis that this difference in older ages and young equals zero in favor of the Alternative Hypothesis that it does not equal zero. The p-value is the probability of obtaining data that result in a β^1 estimate this far (or any farther) from the null value just by chance, if the null hypothesis were true.therefore we may seek to provide the treatment to a specific group as the null hypothesis was rejected. ### Answer 3d

```
mto3 %>%
ggplot(aes(y = econ_ss_zscore, x = age)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, color = "blue")

## `geom_smooth()` using formula = 'y ~ x'
```

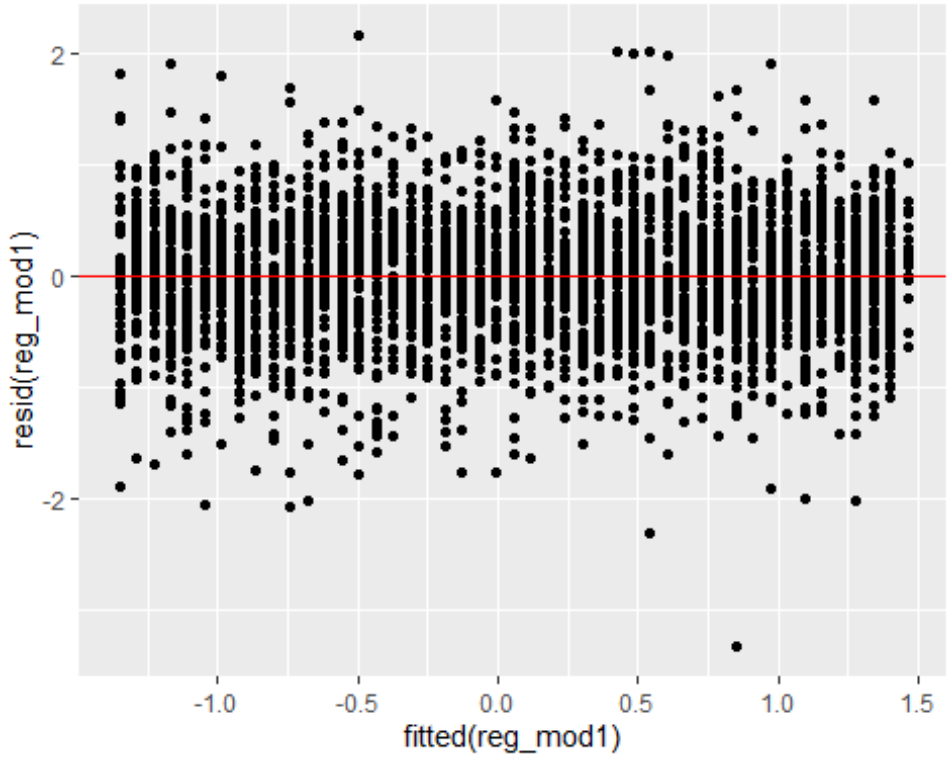



```
summary(reg_mod1)

##
## Call:
## lm(formula = econ_ss_zscore ~ age, data = mto3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3157 -0.3553 -0.0025  0.3593  2.1702
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.4540300  0.0310497  -79.04   <2e-16 ***
## age          0.0611945  0.0007254   84.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5604 on 3261 degrees of freedom
## Multiple R-squared:  0.6858, Adjusted R-squared:  0.6857
## F-statistic:  7117 on 1 and 3261 DF,  p-value: < 2.2e-16
```

```
mto3 %>%
   ggplot(aes(x = fitted(reg_mod1), y = resid(reg_mod1))) +
  geom_point()+
  geom_hline(yintercept=0 ,color= "red")
```



Question 4 [14 pts]

Question 4a [5 points]

Create a dichotomous `lpv` variable that takes the value 1 for all households with a *low poverty voucher* and takes the value 0 for all other households. The researchers also hypothesize that the *low poverty voucher* will improve economic self-sufficiency (relative to control and Section 8 groups combined - these can be referred to as *usual housing support*). State the specific causal question that corresponds with this hypothesis.

Question 4b [6 points]

State the potential outcomes for a single low income household. What linear regression model (include age as a covariate in the model and the dichotomous `lpv` variable) might be useful to test this hypothesis? Write out the model. What is the parameter of interest?

Question 4c [3 points]

What are the null and alternative hypotheses? Use a two-sided hypothesis test.

Answer 4

Answer 4a

SCQ: what is the impact on the economic well being score for households that experience the Low poverty vouchers relative to receiving usual housing support for low income high proverty neighbourhoods ? we must use the multiple regression model to accomadate one continuous predictor (age) and one dichotomous predictor(LPV (treatment=0,1)).

```
mto3$lpv <- ifelse(mto3$group=="lpv",1,0)
table(mto3$lpv)
```

```
##
##    0    1
## 1808 1455
```

```
reg_mod2 <- lm(econ_ss_zscore ~ age + lpv, data = mto3)
summary(reg_mod2)
```

```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv, data = mto3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4019 -0.3515  0.0058  0.3513  2.0831
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.5221266  0.0319083 -79.043  < 2e-16 ***
## age          0.0611529  0.0007185  85.117  < 2e-16 ***
```

```
## lpv          0.1564944  0.0195483   8.006 1.64e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.555 on 3260 degrees of freedom
## Multiple R-squared:  0.6918, Adjusted R-squared:  0.6917
## F-statistic:  3660 on 2 and 3260 DF,  p-value: < 2.2e-16
```

Answer 4b

potential outocome #1 : what the economic self sufficinecy for a low income household would be if they underwent treatment condition and were provided low poverty vouchers

potential outcome #2: what the economic self sufficiency score for a low income household would be if they were given the usual housing support.

the parameter of interest is the coefficient on the treatment variable B^2. ### Answer 4c null and alternate hypothesis:

if the conclusion of this research study is that the Low Poverty Voucher program improves economic well being relative to the control and Section 8 programs (combined) then policy makers will *eliminate* the Section 8 program and replace it with an expanded Low Poverty Voucher Program. If the conclusion of this research study is that the Low Poverty Voucher Program seems to have a similar effect on economic well being as the Control and Section 8 programs (combined) then they will continue the Low Poverty Voucher Program at its current size and continue to collect more data to determine what the effects are in a larger study.

the null hypothesis: the null hypothesis would be that there is no effect of the treatment on the economic self- sufficiency score holding age constant. B^2=0

THe alternate hypothesis: this means that under the alternate the change is non zero in the economic self sufficiency score effected by treatment variable. B^2!=0.

Question 5 [20 pts]

Question 5a [4 points]

Run the multiple linear regression you describe in Question 4. Check the three assumptions of linear regression for this model by making and then assessing appropriate residual plots. For each assumption state whether or not it is violated.

Question 5b [6 points]

Interpret each of the following aspects of the regression model if they are valid to interpret: r-squared value, RMSE, slope coefficient for age. How do the r-squared value and RMSE differ from what they were in the simple linear regression you ran in Q3?

Question 5c [3 points]

Create a scatter plot of age and economic self-sufficiency with the two regression lines added. Color the data points to indicate for each household if they were in the *low poverty voucher group* or group receiving usual housing support.

Question 5d [3 points]

If valid to do so, interpret the p-value that is relevant for the hypothesis test you state in Question 4. If valid, use an alpha level of 0.05 and give the conclusion of this hypothesis test.

Question 5e [2 points]

Create a 95% confidence interval for the parameter of interest (if valid to do so). Give a statistical interpretation of this confidence interval (if valid).

Question 5f [2 points]

What do these results tell you about the specific causal question you stated in Question 4? State and interpret the estimated treatment effect.

Answer 5

Answer 5a

We do not see any obvious non-linearities in the residuals; the mean of the residuals appears to be near zero in all segments of the residual plot as we move from left to right so the linearity assumption likely holds.

The constant variance assumption is also likely to not be violated; the spread across the residuals seems to maintain a uniform average distribution and is not having varying across the spread.

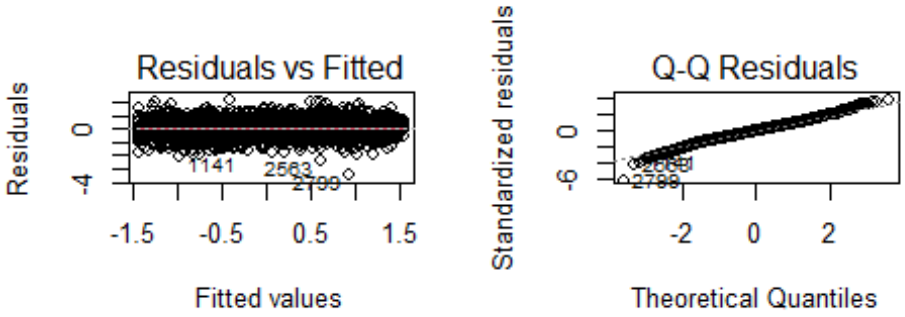
The normality is valid there is a small skew towards the left and variations in the tails are minor and will not cause the inferences to be incorrect. the plot for the histogram appears

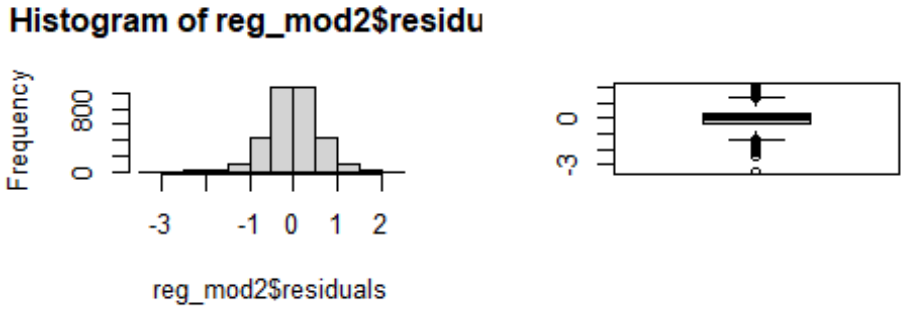
to be balanced and symmetric, it does not have any gaps and reasonably problematic outliers except one outlier which will not significantly alter our results.

```
summary(reg_mod2)
```

```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv, data = mto3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4019 -0.3515  0.0058  0.3513  2.0831
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.5221266  0.0319083 -79.043  < 2e-16 ***
## age          0.0611529  0.0007185  85.117  < 2e-16 ***
## lpv          0.1564944  0.0195483   8.006 1.64e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.555 on 3260 degrees of freedom
## Multiple R-squared:  0.6918, Adjusted R-squared:  0.6917
## F-statistic:  3660 on 2 and 3260 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(reg_mod2, c(1, 2))
hist(reg_mod2$residuals)
boxplot(reg_mod2$residuals)
```





Answer 5b

Interpretation of RMSE: The average distance the points are from the estimated regression line (the distance between observed and fitted values of the outcome (y(economic well being z score))) is 0.55. This is the average distance each observation's actual y-value(econ-zscore) is away from the mean of y for those observations with the same x1 age and x2 treatment values.

Interpretation of R-squared: 69.18% of the variability in economic well being zscore is accounted for by its linear relationship with age and the dichotomous predictor lpv(poverty voucher treatment) and x2. is explained by the age and treatment status of the households, adn its lineaer relationship with the econ self -sifficiency.

slope coefficient for age: 0.06: for the control group(lpv=0), across all ages, for every 1 year increase in the age there is a 0.06 point increase in the econ self sufficiency.

0.56 and 68.58% is the R squared value. the RMSE has slightly decreased and the R-squared has increased indicating a more accurate prediction by the two predictors.

Answer 5c

Answer 5d

p-value: The p-value is very small (much smaller than any standard alpha level=0.01) so we reject the Null Hypothesis. The p-value is the probability of obtaining data that result in a

B^2 variable lpv estimate this far (or any farther) from the null value just by chance, if the null hypothesis were true. ### Answer 5e

CI for the slope coefficient on age : A 95% CI for Lpv is (0.11, 0.19). Over repeated sampling and estimation of this regression equation, 95% of the confidence intervals constructed in this way will contain the population mean of the treatment effect parameter and 5% will not. Based on this data, the population mean of econ_ss_zscore treatment change with a one year increase in B^1(age) (holding x2 constant) is unlikely to be 0.11 or less or larger than 0.19.

```
reg_mod2 = lm(formula = econ_ss_zscore ~ age + lpv, data = mto3)
summary(reg_mod2)
```

```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv, data = mto3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4019 -0.3515  0.0058  0.3513  2.0831
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.5221266  0.0319083 -79.043  < 2e-16 ***
## age          0.0611529  0.0007185  85.117  < 2e-16 ***
## lpv          0.1564944  0.0195483   8.006 1.64e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.555 on 3260 degrees of freedom
## Multiple R-squared:  0.6918, Adjusted R-squared:  0.6917
## F-statistic:  3660 on 2 and 3260 DF,  p-value: < 2.2e-16
```

```
reg_mod2$coef
```

```
## (Intercept)         age         lpv
## -2.52212658  0.06115295  0.15649438
```

```
#since no treatment for households with no lpv is coded as 0, the regression
model intercept
# is the intercept for precincts that recieved late cct progresa alternative
int_nolpv = reg_mod2$coef['(Intercept)']

#since treatment for precincts with the early cct progresa(treatment)is coded
as 1, the intercept for precincts with (treat=lpv) is the intercept of this
regression model
#PLUS the coefficient on the dichotomous variable
int_yeslpv = reg_mod2$coef['(Intercept)'] + reg_mod2$coef['lpv']

#since there is no interaction effect, both lines have the same slope
```

```
slope = reg_mod2$coef['age']  ##coefficient on the continuous variable

mto3 %>%
  ggplot(aes(y = econ_ss_zscore, x = age, color = factor(lpv))) +
  geom_point() +
  geom_abline(aes(intercept = int_nolpv, slope = slope, color = '0')) +
  #regression line for no treatment(late progresa)
  geom_abline(aes(intercept = int_yeslpv, slope = slope, color = '1'))
```


```
  #regression line for countries that had the treatment lpv
confint(reg_mod2, level = 0.95)

##                    2.5 %      97.5 %
## (Intercept) -2.58468892 -2.45956425
## age          0.05974428  0.06256162
## lpv          0.11816628  0.19482248
```

Answer 5f

SCQ: what is the impact on the economic well being score for households that experience the Low poverty vouchers relative to receiving usual housing support ?

the impact on the economic well being score for households that experience Lpv is greater than those that experience the control condition in low income neighbourhoods. We reject the null hypothesis which is that there is no change in the economic well being score between both the treatment(LPV)and the control group in favour of the alternative that there infact is a change between the two groups. the estimated treatment effect is 0.15.