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 community violence. 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 actually improved participants' economic and subjective wellbeing. 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 61 points.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

Question 1 [13 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 [5 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 which households to offer a program to that is intended to improve economic self-sufficiency - all households or only households with younger heads of household.

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

The simple linear regression model having age of the head of the household as predictor and $econ\_ss\_zscore$ as outcome variable will be useful to check whether there is a positive association between age of the head of the household and the expected economic well being of the household. The equation for this model is as follows:

$$Econ - ss - zscore_i = \hat{\alpha} + \hat{\beta} * Age_i + \hat{\epsilon}_i$$

Parameter of interest is population parameter or true Beta β (slope) which is the slope of the association between age and economic well being and the estimator for this parameter of interest here is $\hat{\beta}$. ### Answer 1b Null Hypothesis: $\beta = 0$ (where β is the true population parameter for slope) Alternative Hypothesis: $\beta \neq 0$)

Answer 1c

In this study, the Type I error is the probability of falsely rejecting the null hypothesis when it is true; in this case, it is the probability of concluding that there is a an association between age and economic well being when there is not. The real world consequence of this could be that the program manager of the voucher program would only issue vouchers to the households with younger heads of the household and thus the households with older heads would suffer greatly in two ways: one, with poverty and two, the head of the household is too old to earn anything and thus in reality would really need help.
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 is no association between age and economic well being when they do have. The consequence of this could be the program manager would keep on issuing housing vouchers to all of the households and because of it, some number of the households with younger heads would suffer as they really deserve but not all of them are getting voucher program. In my opinion, in this case, Type I error is relatively worse compared to Type II as explained above - so I will suggest Alpha = 0.01 here.

Question 2 [20 pts]

2a [5 points]

Using summary statistics and a figure, describe the distribution of the economic well-being 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 two or more residual plots. For each assumption state whether or not it is violated and how you can tell.

2e [3 points]

State what parts of the regression results are or are not valid based on which assumptions are or are not violated.

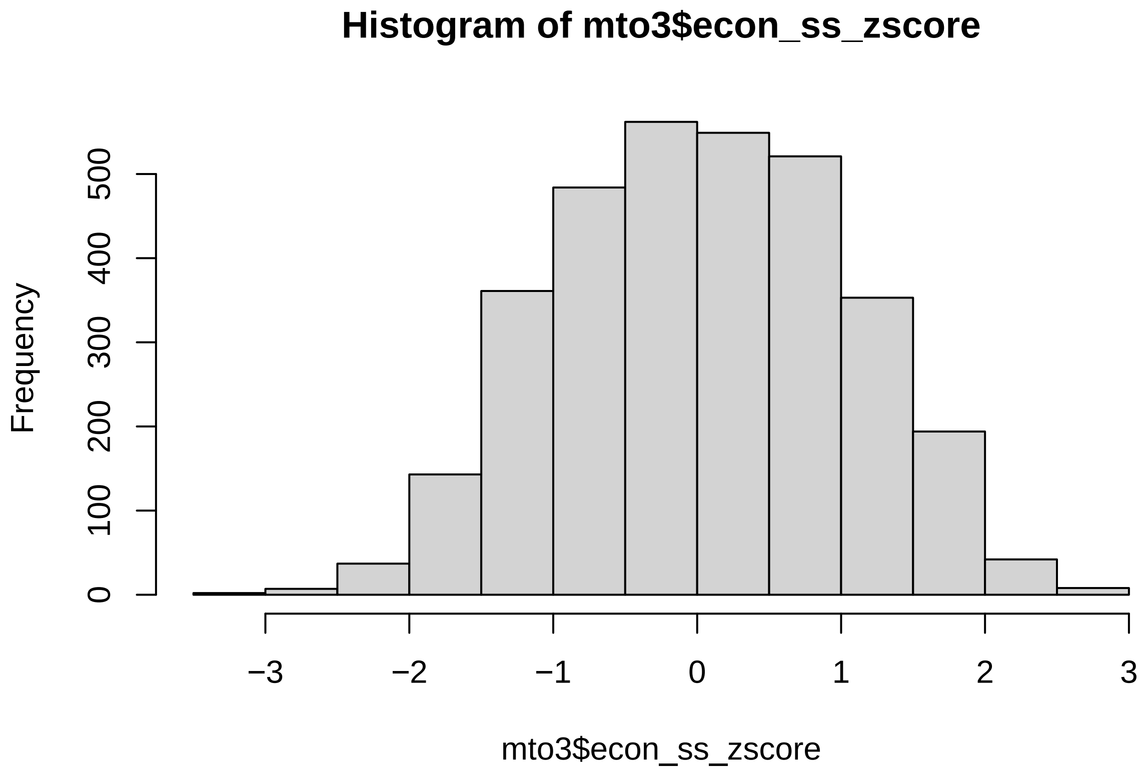
Answer 2

Answer 2a

```
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)
```

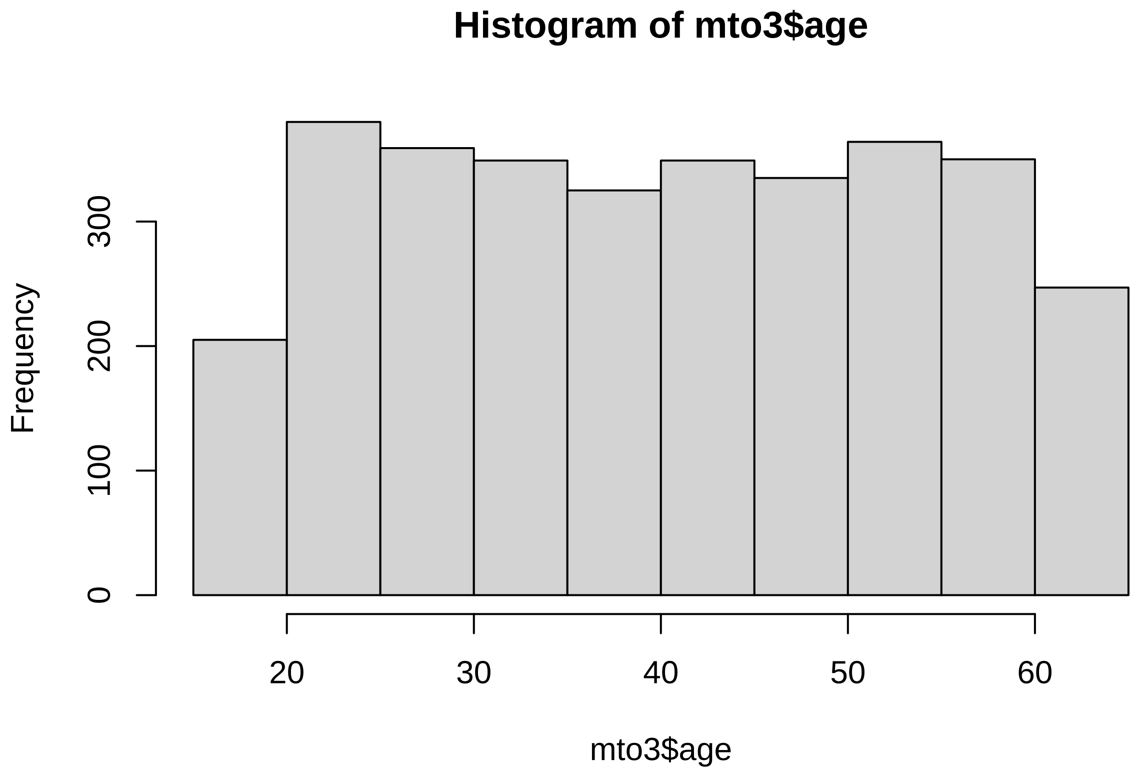


The mean economic well being is 0.03 and the median is 0.02. The z-score of economic well being ranges between -3.23 and 2.93. As it is already converted to z scale, therefore, there is not much variation which is also evident from the histogram of the *econ_ss_zscore* as the distribution is symmetric and uni-modal with no gaps and spikes.
### Answer 2b

```
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)
```



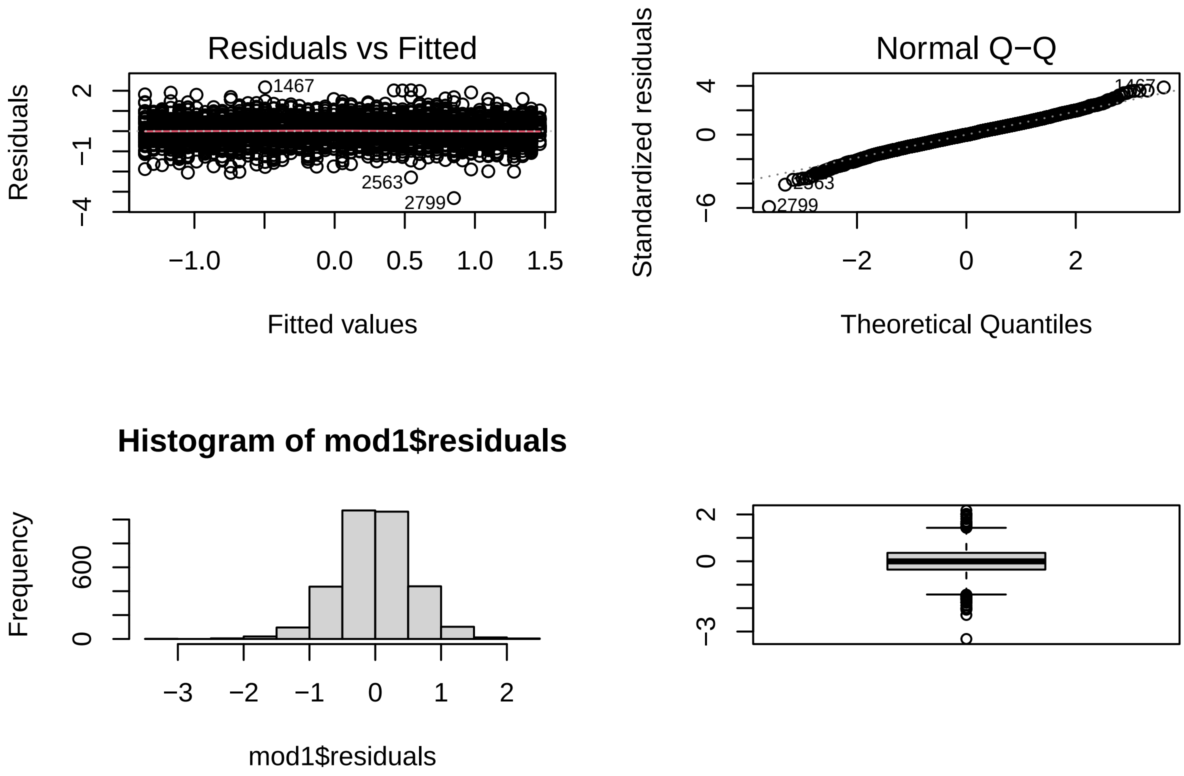
The mean age in the data frame is 40.61 and the median is 41 years. The distribution ranges between 18 to 64 years. The distribution is bi-modal, almost flat in shape and have no gaps and spikes ### Answer 2c

```
mod1 <- lm(data = mto3, econ_ss_zscore ~ age)
summary(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

```
par(mfrow = c(2, 2))
plot(mod1, c(1, 2))
hist(mod1$residuals)
boxplot(mod1$residuals)
```



We do not see any obvious non-linearity 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. There is a very small (almost negligible) left skew in the residuals with several high and low outlying values (as the economic well being variable is transformed to z-scale, therefore no vivid variation) and can be considered almost uniform which makes us conclude that there is a constant variation of residuals along the fitted line and constant variation assumption holds. As the data set contains 3263 observations,and the residuals follow a uniform distribution therefore normality assumption also holds. So, none of the assumptions seems to be violated.

Answer 2e

Here, all three assumptions hold. To start with, as linearity assumption hold, therefore, the estimated regression coefficients and R^2 values are unbiased and interpretable. As constant variation assumption holds therefore, the model RMSE is also likely to be a useful summary statistic. An lastly, as the normality assumption holds, therefore, the standard errors and p-values associated with our estimated coefficients are likely to be correct, 95% confidence intervals for the true coefficient values are likely to be valid.

Question 3 [12 pts]

Question 3a [5 points]

Interpret the model results as indicated by your answer to Question 2. Interpret each of the following aspects of the model if they are valid to interpret: estimated y-intercept, estimated slope, r-squared value, RMSE, and p-value for the slope.

Question 3b [2 points]

What do these interpretations tell you about the hypothesis test stated in Question 1?

Question 3c [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 standard error of the sampling distribution.

Question 3d [2 points]

Make a scatterplot of the age and economic self-sufficiency data and add the estimated regression line to the figure.

Answer 3

Answer 3a

Y Intercept: Expected value of the outcome when the predictor is 0. Here, on average, the economic well being of the household where the age of the head of the household is 0 years is -2.45. It is beyond the support of data and has no meaning in real world. Estimated slope: With a one year increase in the age of the head of the household, on average, the economic well being of the household increases by 0.61 points. On a meaningful scale, with a 10 years increase in age of the head of the household, on average, the economic well being increases by 6 points. Rˆ2: 68% of the uncertainty in the economic well being of the household is explained by the age variable and the linear relationship between age and economic well being. RMSE:0.56, On average, the economic well being of the household varies around 0.56 points away from the estimated value than we would expect based on the age of the head of the household. (P value:<2e-16) How many SEs away the estimated slope $\hat{\beta}$ is from 0. Over repeated sampling, the p-value is the probability that if the Null Hypothesis were true, that we would observe an estimated slope $\hat{\beta}$ of the data this far from the Null Value (or any farther), just by chance. ### Answer 3b As the p-value or the probability of observing an estimated slope $\hat{\beta}$ of 0.061 being 0.0007 standard errors away from the true $\beta = 0$ under the Null Hypothesis, just by chance is lesser than 2e-16 which is very very small so we reject the Null Hypothesis which stated that there is no association between age of the head and the economic well being of the household and thus accept the alternative hypothesis and conclude that there is an association between age of the head and economic well being of the household.

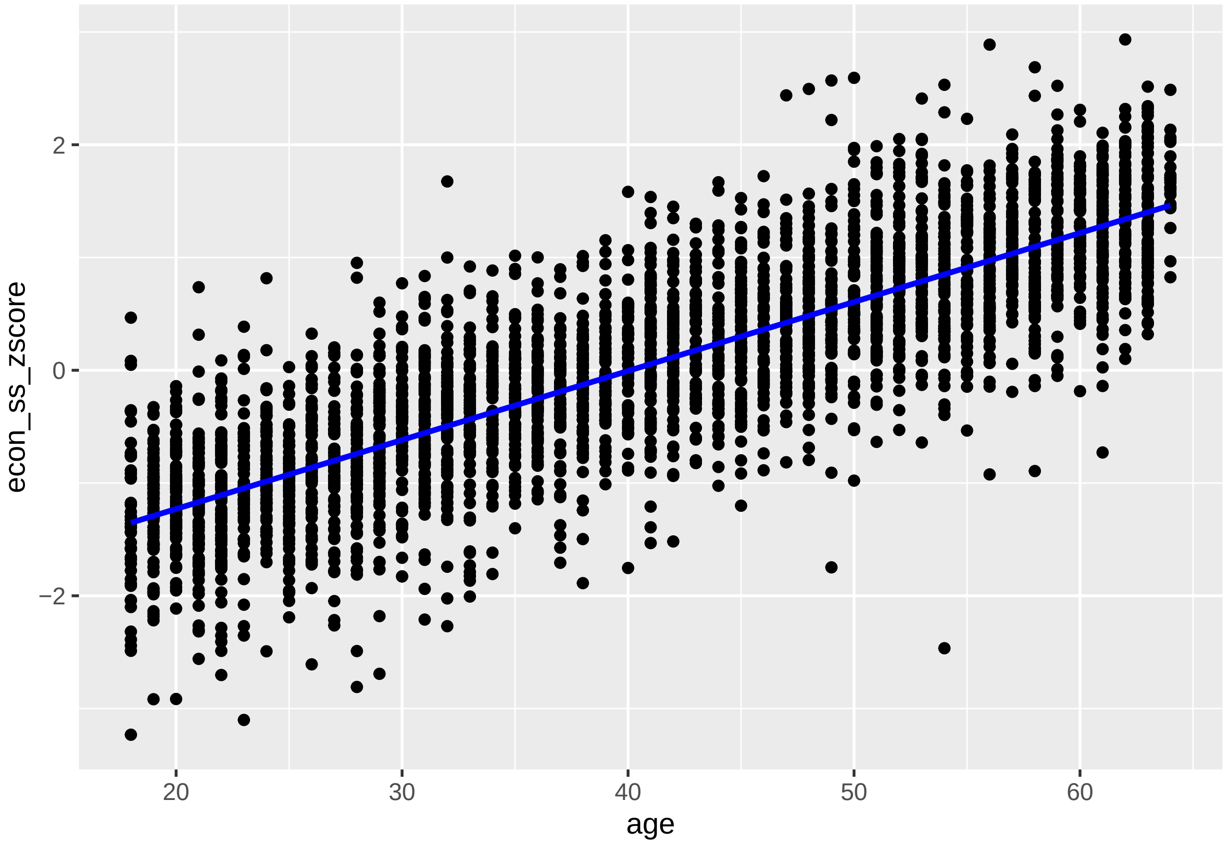
Answer 3c

If the assumptions of linear regression are not violated, then CLT holds, so over repeated sampling under the null hypothesis, the sampling distribution of the estimated slope coefficient for age $\hat{\beta}$ would be a normal distribution (0r a t distribution with 3261 degrees of freedom). The mean of this distribution would be equal to the true value of the slope coefficient β in the population which under the null is 0, and the standard error would be equal to the standard error of the estimated slope coefficient which here is 0.0007.

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’
```



Question 4 [9 pts]

Question 4a [2 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). For this hypothesis, state the specific causal question.

Question 4b [4 points]

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

```
mto3.new <- mto3 %>%
  mutate(lpv = if_else(group == 'lpv', 1, 0))
num.1 = nrow(subset(mto3.new, lpv == 1))
num.0 = nrow(subset(mto3.new, lpv == 0))
num.1
```

```
## [1] 1455
```

```
num.0
```

```
## [1] 1808
```

SCQ: Controlling for age, what is the impact of Low Poverty Voucher (LPV)(intervention group) relative to the usual housing support (control and section 8 groups combined (combined control group)) on the economic self sufficiency of the low income households located in the high-poverty neighborhoods in the US. ### Answer 4b Potential outcome for a single household: What the economic self sufficiency of the household would be if that household is given LPV intervention. What the economic self sufficiency of the household would be if that household is given usual housing support.

Since we are trying to study the relationship between economic well being of the households which are randomly assigned to any of the two groups of LPV and usual housing support (dichotomous variable) and age as predictor variables therefore, multi-linear regression model will be useful to test this hypothesis. Parameter of interest here is true beta_2 β_2 (slope).

$$Econ - ss - zscore_i = \hat{\alpha} + \hat{\beta}_1(Age_i) + \hat{\beta}_2(lpv_i) + \hat{\epsilon}_i$$

### Answer 4c The Null Hypothesis is: $\beta_2 = 0$
The Alternative Hypothesis is: $\beta_2 \neq 0)$

Question 5 [19 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 [5 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?

Question 5c [3 points]

Briefly describe what the two regression lines would look like on a graph where the horizontal axis is age and the vertical axis is the economic self-sufficiency measure.

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

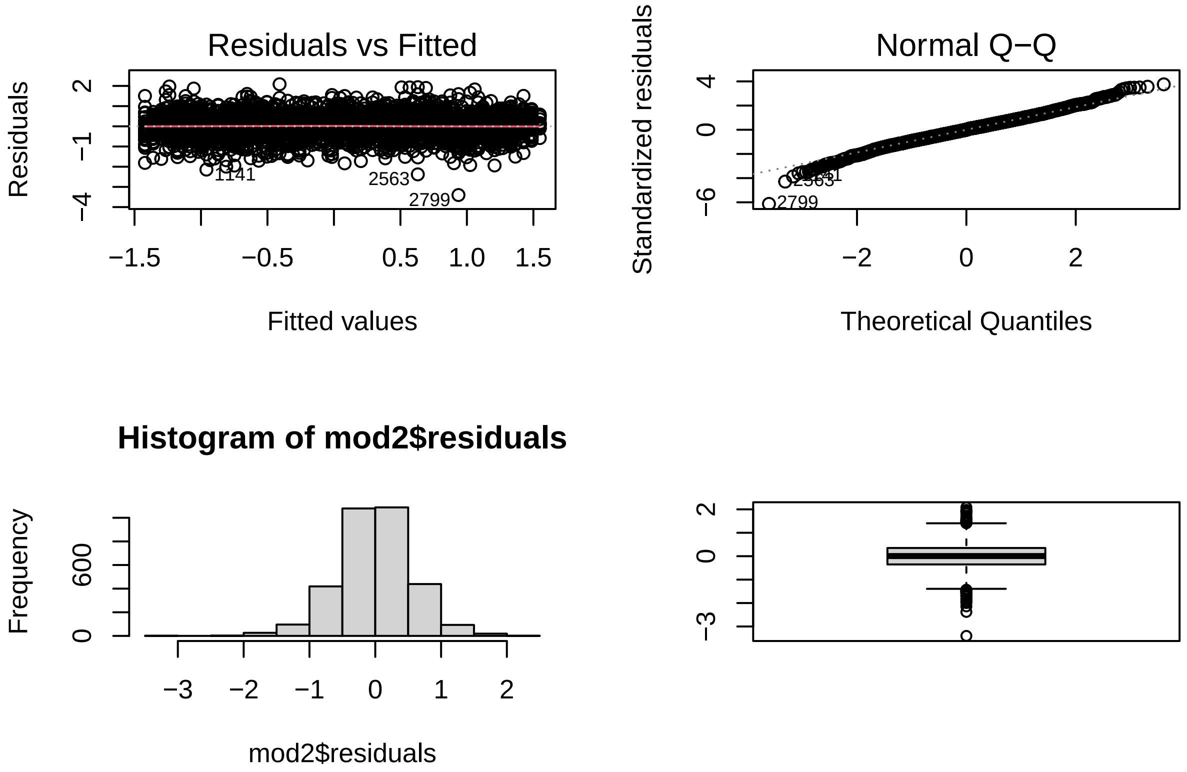
Answer 5a

```
mod2 <- lm(data = mto3.new, econ_ss_zscore ~ age + lpv)
summary(mod2)
```

```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv, data = mto3.new)
##
## 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(mod2, c(1, 2))
hist(mod2$residuals)
boxplot(mod2$residuals)
```



We do not see any obvious non-linearity 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. There is a very small (almost negligible) left skew in the residuals with several high and low outlying values (as the economic well being variable is transformed to z-scale, therefore no vivid variation) and can be considered almost uniform which makes us conclude that there is a constant variation of residuals along the fitted line and constant variation assumption holds. As the data set contains 3263 observations which is a reasonably large sample,and the residuals follow a uniform distribution therefore normality assumption also holds. So, none of the assumptions seems to be violated. ### Answer 5b Y Intercept: -2.522: Y-intercept is the expected value of the outcome when the predictor is 0. Here, the expected economic well being of the households where the age of the head of the household is 0 years and the household is in control group is -2.522. It is beyond the support of data and has no meaning in real world. Estimated slope: With a one year increase in the age of the head of the household, on average, the economic well being of the household increases by 0.61 points in each treatment arm. On a meaningful scale, with a 10 years increase in age of the head of the household, on average, the economic well being increases by 6.1 points within each treatment arm. Rˆ2: 69.18% of the uncertainty in the economic well being of the household is explained by the two predictors of age and treatment arm (lpv and control group) and the linear relationship of these variables with the economic well being. RMSE:0.56, On average, the economic well being of the household varies around 0.56 from what we would expect based on the age of the head of the household and treatment status. Comparison of RMSE and Rˆ2 between Simple and Multiple Linear Regression: In simple linear regression, RMSE was 0.5604 which has reduced to 0.555 in multiple linear regression. Rˆ2 in simple linear regression was 0.6858 or 68.58% while in multiple linear regression Rˆ2 has increased to 69.18, thus the treatment effect seem to have little association with the outcome of economic well being and thereby have caused little improvement in explaining the uncertainty*

Answer 5c

The two lines for the LPV group and the control group will be parallel to each other because both have same positive slope,and also these are upward sloping on a graph having age on the horizontal axis and economic self-sufficiency on the y-axis. One of these lines will be for the control group and the other will be for the LPV group. These lines will have different Y-intercepts: for control group it will be -2.52 while for the treatment group it will be -2.37.
### Answer 5d (P value: 1.64e-15) How many SEs away the estimated slope is from 0. Over repeated sampling, the p-value is the probability that if the Null Hypothesis were true, that we would observe an estimated slope of the data this far from the Null Value (or any farther), just by chance. Since the p-value is less than the alpha level of 5%, so we reject the Null Hypothesis in favor of Alternate Hypothesis and conclude that on average, there is an impact of LPV relative to combined alternative group on the economic well-being of the households (can be positive or negative - two sided alternative) ### Answer 5e

```r
confint(mod2, level = 0.95)
```

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

The 95% CI for the slope between LPV and the mean economic well being of the household is (0.12, 0.19). Over repeated sampling, 95% of CI constructed in this manner will contain the population slope, and 5% will not. Based on this data, under the null hypothesis, we do not think true slope of economic well being $\beta_2 = 0$ lies within this CI therefore, we reject the Null Hypothesis. ### Answer 5f These results tell us that controlling for age, on average, the estimated treatment effect is not 0 but the estimated treatment effect of LPV intervention relative to usual housing support is marginal and here it is 0.2.

Question 6 [12 pts]

6a [6 points]

The researchers also hypothesize that the low poverty voucher will have different effects on economic self-sufficiency (relative to usual housing support) for households with older versus younger heads of household.

What linear regression model might be useful to test this hypothesis? What is the parameter of interest? What are the null and (two-sided) alternative hypotheses?

6b [6 points]

Run this multiple linear regression. In practice when running multiple linear regression we would check the model assumptions by making and assessing the residual plots, as you did in previous exercises. However, for the purposes of this homework, you can skip making these plots for this subsection - we've checked them and found that all 3 assumptions of linear regression hold.
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. Using an alpha level of 0.10, carry out the hypothesis test you stated in 6a (interpret the p-value and give the conclusion of the hypothesis test) and interpret its implications for the researcher's hypothesis.

## Answer 6

Answer 6a

The multi-linear regression model with interaction between age and treatment arm will be useful to test this hypothesis. Parameter of interest is population parameter or true β_3 (slope). The Null Hypothesis is: $\beta_3 = 0$
The Alternative Hypothesis is: $\beta_3 \neq 0$)

$$Econ - ss - zscore_i = \hat{\alpha} + \hat{\beta}_1(Age_i) + \hat{\beta}_2(lpv_i) + \hat{\beta}_3 lpv_i * Age_i + \hat{\epsilon}_i$$

### Answer 6b

```
mto3.new.mod3 <- lm(data = mto3.new, econ_ss_zscore ~ age + lpv +lpv*age)
summary(mto3.new.mod3)
```

```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv + lpv * age, data = mto3.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4126 -0.3499  0.0018  0.3528  2.0902
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.4974800  0.0407065 -61.353   <2e-16 ***
## age          0.0605448  0.0009514  63.637   <2e-16 ***
## lpv          0.0989838  0.0621341   1.593    0.111
## age:lpv      0.0014152  0.0014513   0.975    0.330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.555 on 3259 degrees of freedom
## Multiple R-squared:  0.6919, Adjusted R-squared:  0.6917
## F-statistic:  2440 on 3 and 3259 DF,  p-value: < 2.2e-16
```

```
mto3.new.mod3$coef
```

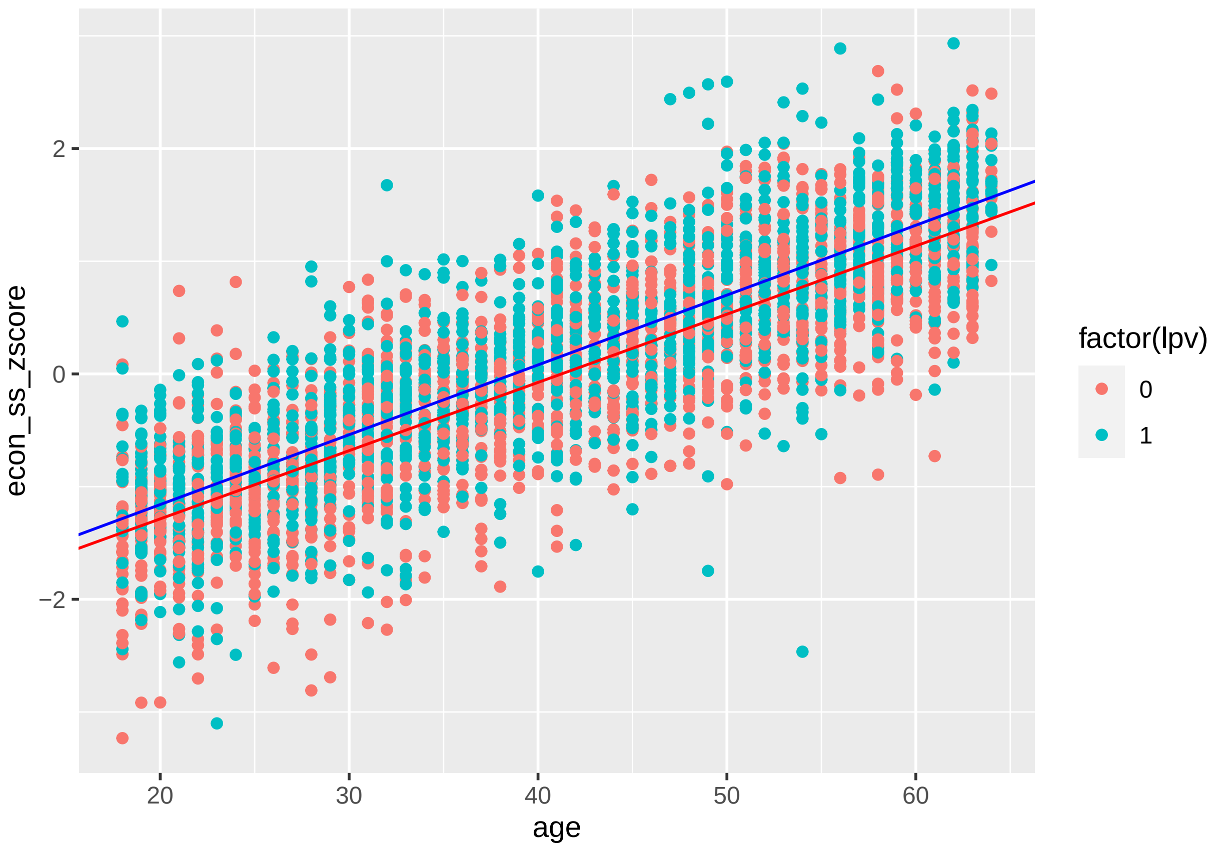
```
##  (Intercept)          age          lpv      age:lpv
## -2.497480049  0.060544781  0.098983828  0.001415193
```

```
#since the combined control group (two groups of households other than LPV program) is coded as 0,
#the regression model intercept is the y-intercept for the combined control group
int_control = mto3.new.mod3$coef['(Intercept)']
#the slope for the age variable is just the regression model #coefficient on the continuous variable
slope_control = mto3.new.mod3$coef['age']

#since LPV households are coded as 1, the intercept for the LPV
# household line is the regression intercept PLUS the coefficient on the
# dichotomous variable
int_lpv = mto3.new.mod3$coef['(Intercept)'] + mto3.new.mod3$coef['lpv']

#the slope for the LPV line is the coefficient on the #continuous variable of age PLUS the coefficient
slope_interact = mto3.new.mod3$coef['age'] + mto3.new.mod3$coef['age:lpv']

mto3.new %>%
  ggplot(aes(y = econ_ss_zscore, x = age, color = factor(lpv))) +
  geom_point() +
  geom_abline(aes(intercept = int_control, slope = slope_control),
              color= 'red') +  #regression line for households in control group
  geom_abline(aes(intercept = int_lpv, slope = slope_interact ),
              color = 'blue') #regression line for LPV households
```



Since, the model holds for all 3 conditions, so the CLT holds, over repeated sampling the sample summary statistics $\hat{\beta}_3$ will follow a t-distribution with 3259 degrees of freedom. Since, the sample size is large, we could assume that the sampling distribution would be a normal distribution. We know that the normal distribution would cover around 68% of area within 1 standard errors and our observed summary statistics $\hat{\beta}_3$ is <1 standard error away from true population parameter, we could not reject the null hypothesis at alpha value of 0.10. Also, the p-value (0.3) is > alpha (0.1) so we fail to reject the null hypothesis. Its implication for the researchers hypothesis is that we do not observe enough evidence to state that low poverty voucher will have different effects on economic self sufficiency (relative to usual housing support) for households with older versus younger heads of the household.