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.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

Question 1

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. What linear model might be useful for testing this hypothesis? Write the equation for this linear model. What is the parameter of interest? What is the estimator for this parameter of interest? What are the Null and Alternative hypotheses? Please use a two-sided hypothesis test.

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

$$E_i = \hat{\alpha} + \hat{\beta}_1 * A_i + \hat{\epsilon}_i$$

The linear model that would aptly address or estimate the parameter of interest, the mean change in economic self-sufficiency in the data generating mechanism, is econ_zz_score_i as equal to the Y-intercept (alpha hat) plus the slope (Beta1 hat) multiplied by the covariate for age of household, A_i, plus the estimated epsilon hat (the residual)

A dual sided hypothesis test will show the Null as age having no impact on the level of economic self sufficiency for households in the data generating mechanism. The Alternative shows that age does have an impact on the level of economic self sufficiency for households in the data generating mechanism.

Type 1 error would be rejecting the Null, of no effect, and conclude that the age of a head of household is impactful to self sufficiency when in fact the age of a head of household has no effect on economic self sufficiency. Only younger households would receive section 8 vouchers as a consequence, when all households would have benfited equally from the economic program.

Type 2 error would be failing to reject the Null, that age has no effect on self sufficiency when in fact age is effective for self sufficiency. All households would receive vouchers as a consequence, when the greatest impact would have been porviding section 8 vouchers to younger households.

Seeing as the heart of the issue is to help the most number of households move out of areas with high poverty, a type 1 error would be worse. So, the alpha for this hypothesis test should be set to .01

Question 2

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

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

Run the simple linear regression you describe in Question 1.

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.

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

Answer 2

```
#Examine Distribution of Population Econimc Self Sufficiency
summary(mto3$econ_ss_zscore)
```

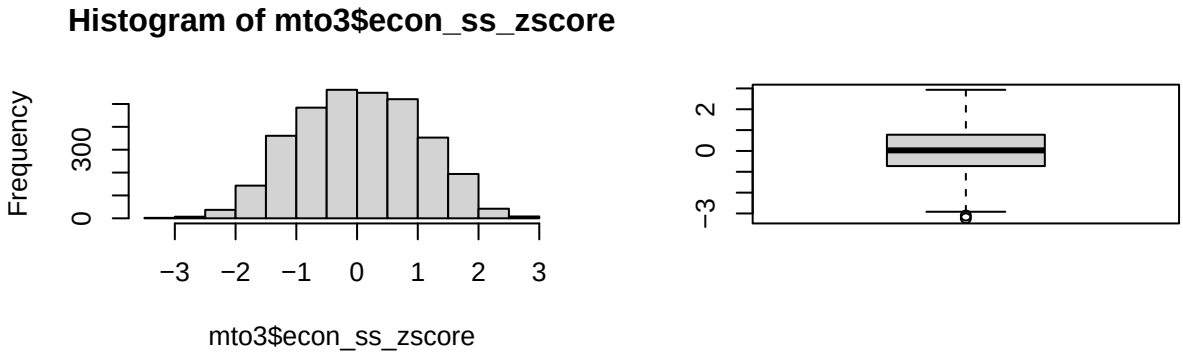
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.23231 -0.72662  0.02777  0.03129  0.77849  2.93332
```

```
par(mfrow = c(2, 2))
hist(mto3$econ_ss_zscore)
boxplot(mto3$econ_ss_zscore)

#Examine Distribution of Population Age
summary(mto3$age)
```

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

```
par(mfrow = c(2, 2))
```




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

#Run Linear Regression
reg1 <- lm(econ_ss_zscore ~ age, data = mto3)
reg1
```

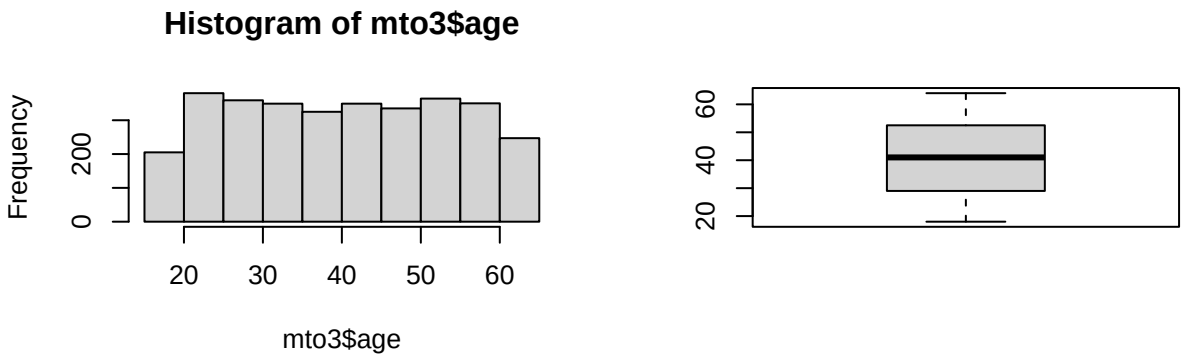
```
##
## Call:
## lm(formula = econ_ss_zscore ~ age, data = mto3)
##
## Coefficients:
## (Intercept)          age
##    -2.45403      0.06119
```

```
summary(reg1)
```

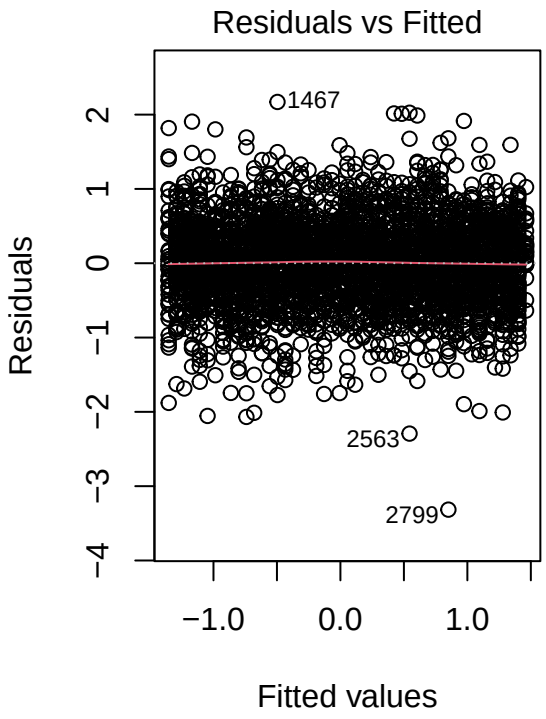
```
##
## Call:
## lm(formula = econ_ss_zscore ~ age, data = mto3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3157 -0.3553 -0.0025  0.3593  2.1702
```

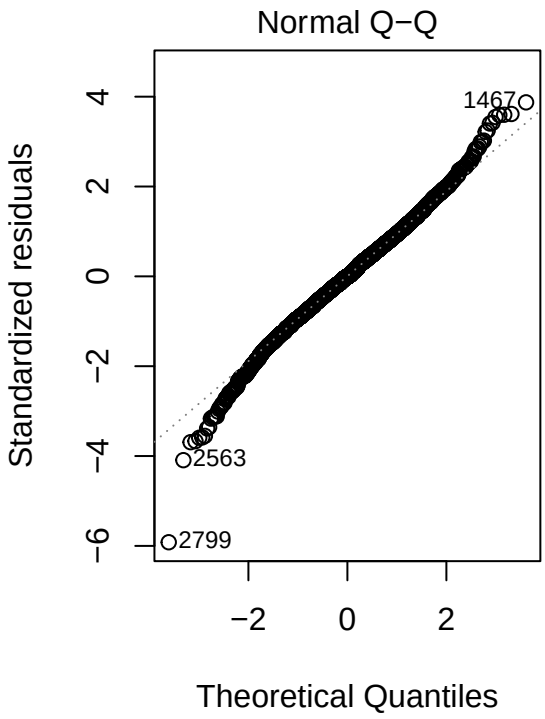
```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.4540300  0.0310497  -79.04   <2e-16 ***
## age          0.0611945  0.0007254   84.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5604 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
```

```
#Examine Linear Regression
par(mfrow = c(1, 2))
```

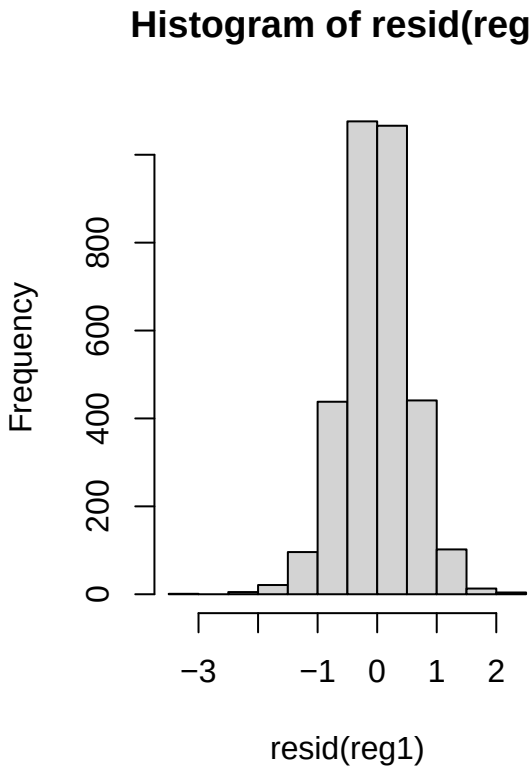


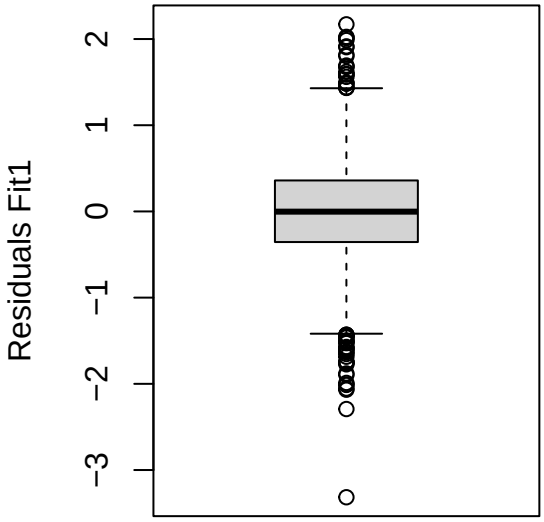

```
plot(reg1, c(1, 2))
```





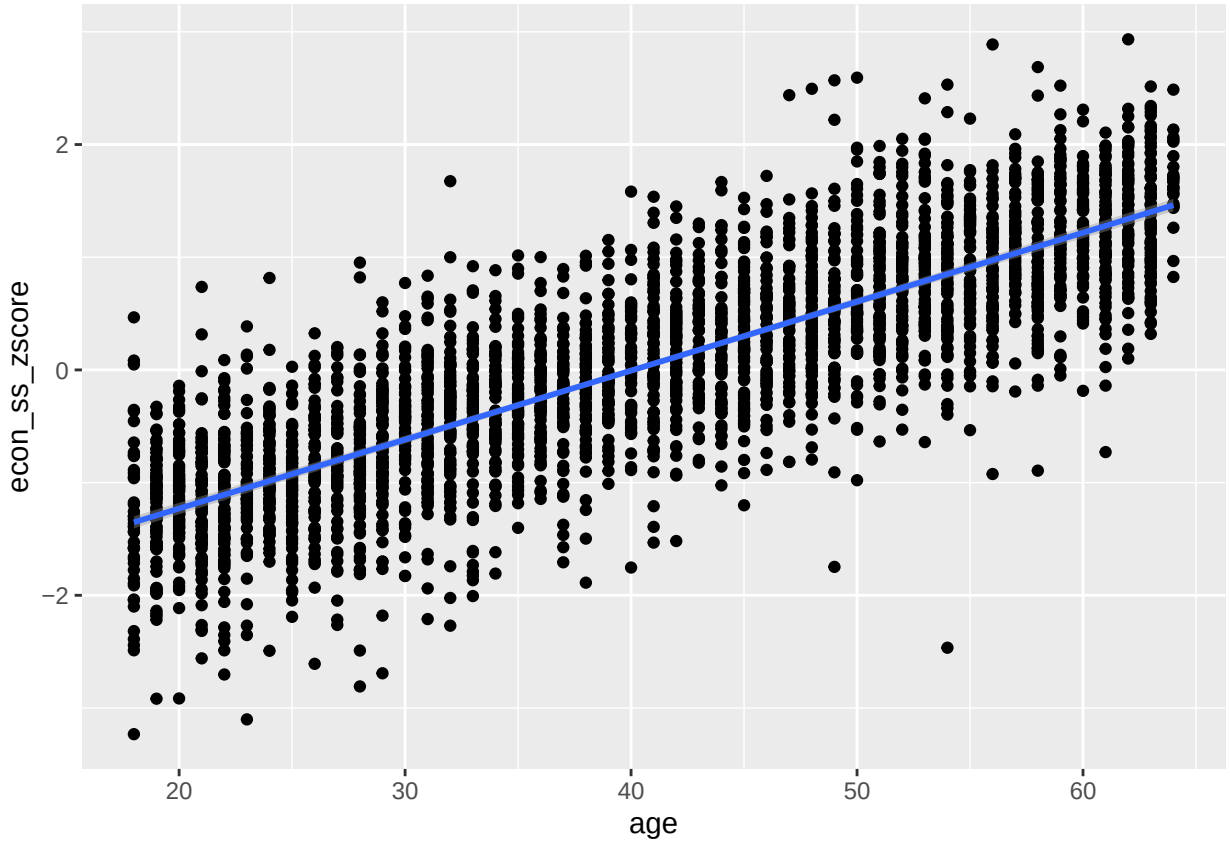

```
hist(resid(reg1))
boxplot(resid(reg1), ylab = "Residuals Fit1")
```






```
mto3 %>%
  ggplot(aes(y = econ_ss_zscore, x = age)) +
  geom_point() +
  geom_smooth(method = 'lm', se = TRUE, level = 0.995)
```

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



#Distribution of Economic Self Sufficiency and Age The distribution of economic self sufficiency in the data genrating mechanism (dgm) visually appears to be normal, unimodal, and near symmetrical in shape, with a few outliers in the lowest edge of the z-score -3. On average, economic self sufficiency in the dgm is .027, with an IQR spanning from -.726 to .778 (~ 1 z-score). The lowest any data point scored is -3.23 and the highest any data point scored is 2.93. The mean (.0312) and the median (.0277) are very close, but not same– meaning the data set skews slightly towards lower economic self sufficiency.

The distribution for age of household in the dgm visually appears to be constant for all age groups and near symmetrical in shape. There appear to be no outliers in this coeficiants data set. On average, age of household in the dgm is 40.6 years old, with an IQR spanning from 29 years old to 52.5 years old (~ 23 years). The lowest age for any data point is 18 and the oldest age of data point is 64. The mean (40.6) and the median (41) are near similar, signaling the data set is distributed normally.

#Running Tests 1) The linear assumption holds for this plot. The mean of residuals appears to be near zero in each segment of the plot as we move from left to right. The regression coefficients and R-squared value are unbiased and interpretable.

2) Constance variance holds. The spread of residuals is similar as we move from left to right across the residual plot. The Residual Standard Error is unbiased and interpretable.

3) Since the residuals have a round shaped distribution, and are symmetric with most values near the mean value, Normality holds. The QQ plot shows that normality is slightly off in the tails, but this small skew is minor and will not cause issues in trusting the inferences garnered. The standard error, T value, and p-value for the slope are unbiased and interpretable.

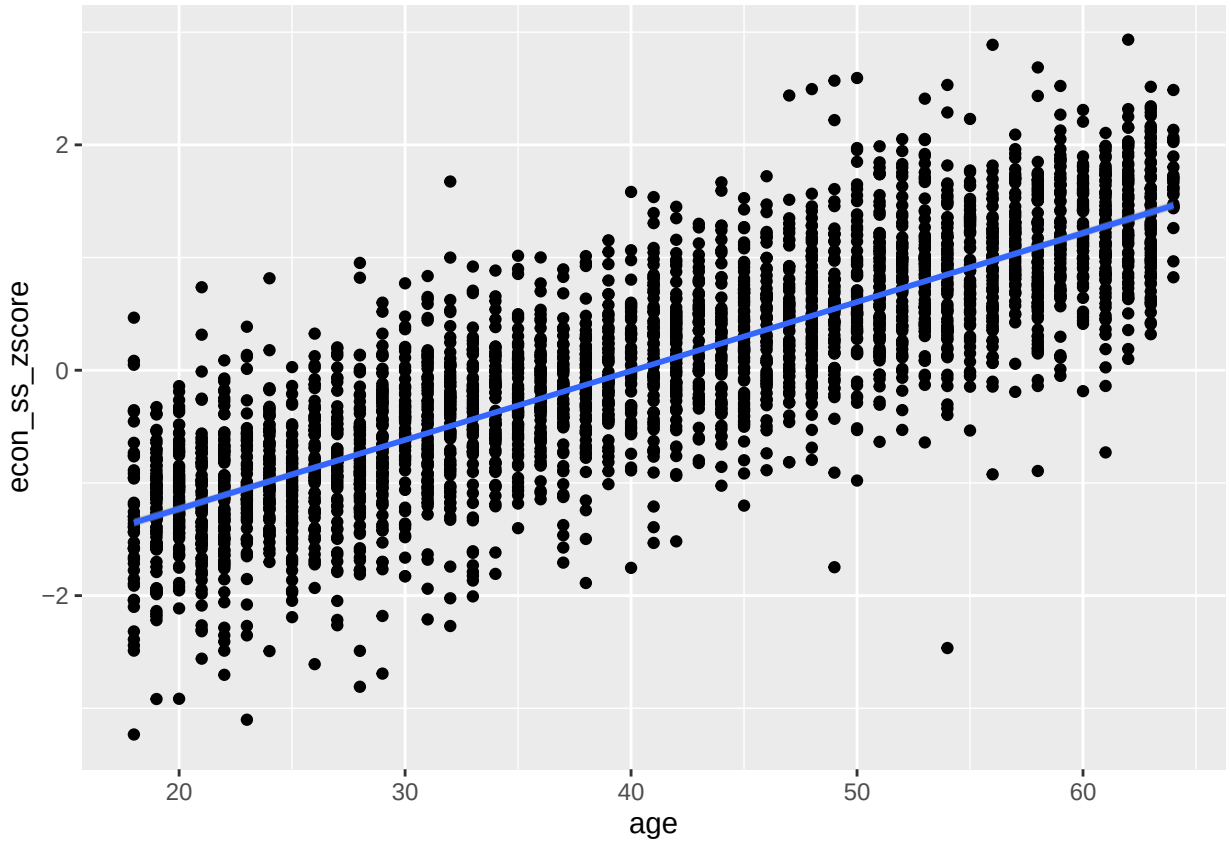
Question 3

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

Answer 3

```
mto3 %>%
ggplot(aes(y = econ_ss_zscore, x = age)) +
geom_point() +
geom_smooth(method = 'lm', se = TRUE)
```

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



#Interpretations Our model results suggests that a one unit increase in age of household is associated with an approximate .0611 point increase in the mean z_score of economic self sufficiency. For a household with a head at the age of 0, the economic efficiency shows to be -2.454. This y-intercept value is not meaningful to interpret, since it falls outside the scope of the data (which has our lowest age at 18).

R-squared: 69% of the variation in economic self sufficiency (it varied from -3.23 to 2.93 and is now -3.32 to 2.17) is accounted for by its linear relationship with the age of household.

RMSE: On average, the distance the points are from the estimated regression line for economic self sufficiency by age of household is 0.5604. This is the average distance each data point's actual economic z-score is away from the mean of 'econ_ss_zscore' for those observations with the same age of household value.

The p-value for the slope coefficient on age is far smaller than the stated alpha value, so we will reject the Null Hypothesis that this coefficient equals zero in favor of the Alternative Hypothesis that it does not equal zero.

#Learning from Interpretations These interpretations lend themselves to proving the alternative: that age does have an impact on economic self sufficiency.

#Sampling Description Sampling distribution under the Null Hypothesis: The shape of the sampling distribution for Beta1_Hat over repeated sampling is a moundshaped, t-distribution with 3261 degrees of freedom. The mean of this sampling distribution is 0, given that Beta1 in the Null Hypothesis is set equal to 0. The estimated standard error of this sampling distribution is 0.0007254. The points on the horizontal axis should be labeled .0007, 0.0014, 0.0021,0 0.0007, 0.0014, 0.00215.

Question 4

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

Answer 4

```
#CreateVariable
lpv <- if_else(mto3$group == 'lpv', 1, 0)
mean(lpv)
```

```
## [1] 0.4459087
```

```
#DataFrame
mto3$lpv <- mto3 %>%
  mutate(lpv = if_else(mto3$group == 'lpv', 1, 0))
mean(mto3$lpv$lpv)
```

```
## [1] 0.4459087
```

```
#2 Sided Test
t.test(econ_ss_zscore ~ lpv, data = mto3$lpv, var.equal = TRUE,
       alternative = 'two.sided', conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  econ_ss_zscore by lpv
## t = -4.8027, df = 3261, p-value = 1.636e-06
```

```
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.23729148 -0.09971096
## sample estimates:
## mean in group 0 mean in group 1
##     -0.04384251      0.12465871
```

#Specific Casual Question What is the effect of receiving a low poverty voucher to move to a low-poverty neighborhood relative to receiving usual housing support in high-poverty neighborhoods on economic self-sufficiency for low-income Americans, holding age of household constant, between 1994 and 1998?

The potential outcomes for a single household are what it's economic self-sufficiency would be if it were to receive a low poverty voucher and what it's economic self-sufficiency would be if it received usual housing support.

$$E_i = \hat{\alpha} + \hat{\beta}_1 * A_i + \hat{\beta}_2 * L_i + \hat{\epsilon}_i$$

The linear model that would aptly address or estimate the parameter of interest, the mean change in economic self-sufficiency in the data generating mechanism, is econ_zz_score as equal to the Y-intercept (alpha_hat) plus the slope (Beta1_hat) multiplied by the covariate for age of household, (A_i), plus the slope (Beta2_hat) multiplied by the covariate for whether or not a household received the treatment (a low poverty voucher) (L_i) plus the estimated epsilon (the residual hat).

The parameter of interest is Beta2. The Null Hypothesis holds Beta2 = 0, and the alternative holds Beta2 as > 0 OR < 0.

A 2 sided t.test shows that the mean for the Null is -0.044, and that the mean for the Alternative is 0.125.

Question 5

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. 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? 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. 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 for this hypothesis test. What is the conclusion of this hypothesis test? 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). 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

```
#Run Linear Regression
reg2 <- lm(econ_ss_zscore ~ age + lpv, mto3$lpv)
reg2
```

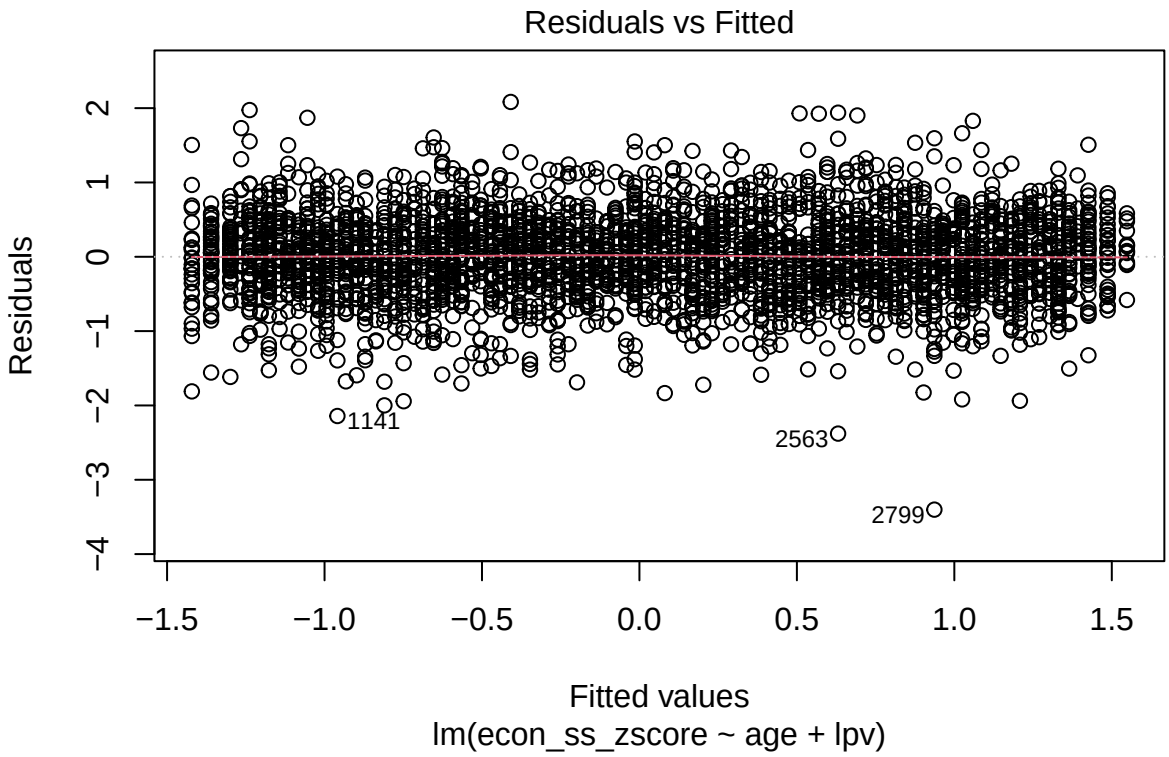
```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv, data = mto3$lpv)
##
```

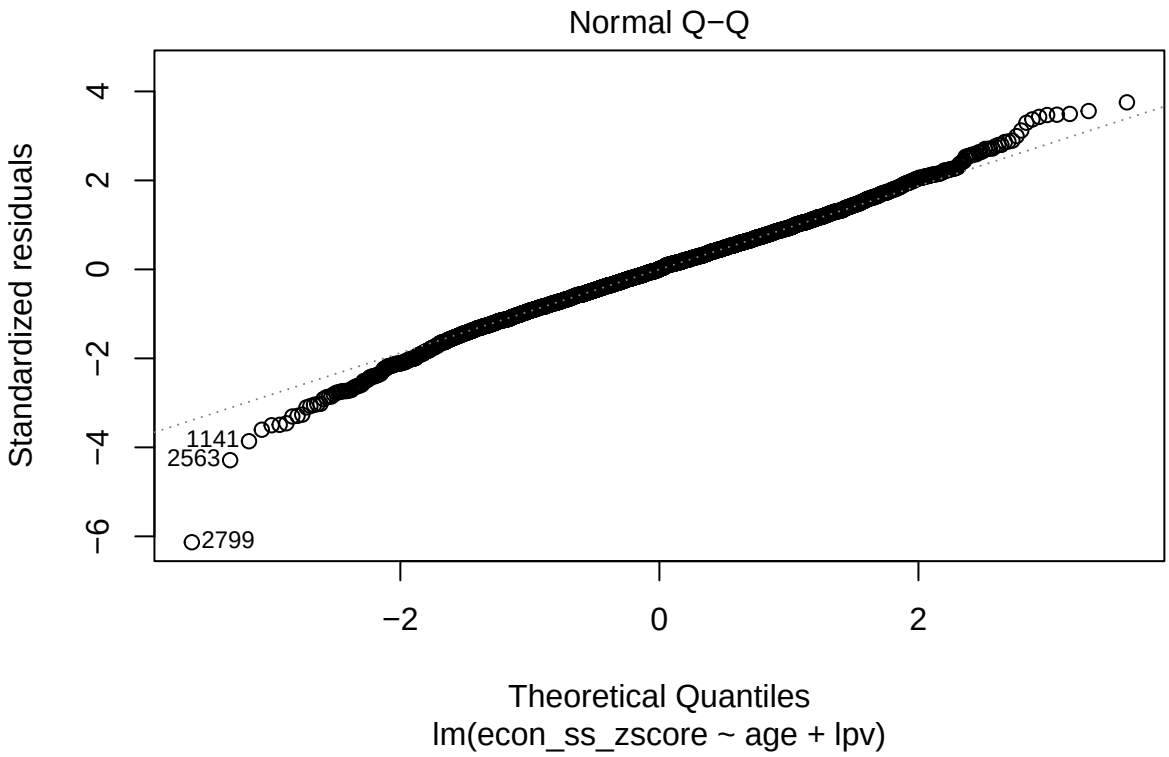
```
## Coefficients:
## (Intercept)          age          lpv
##    -2.52213      0.06115      0.15649
```

```
summary(reg2)
```

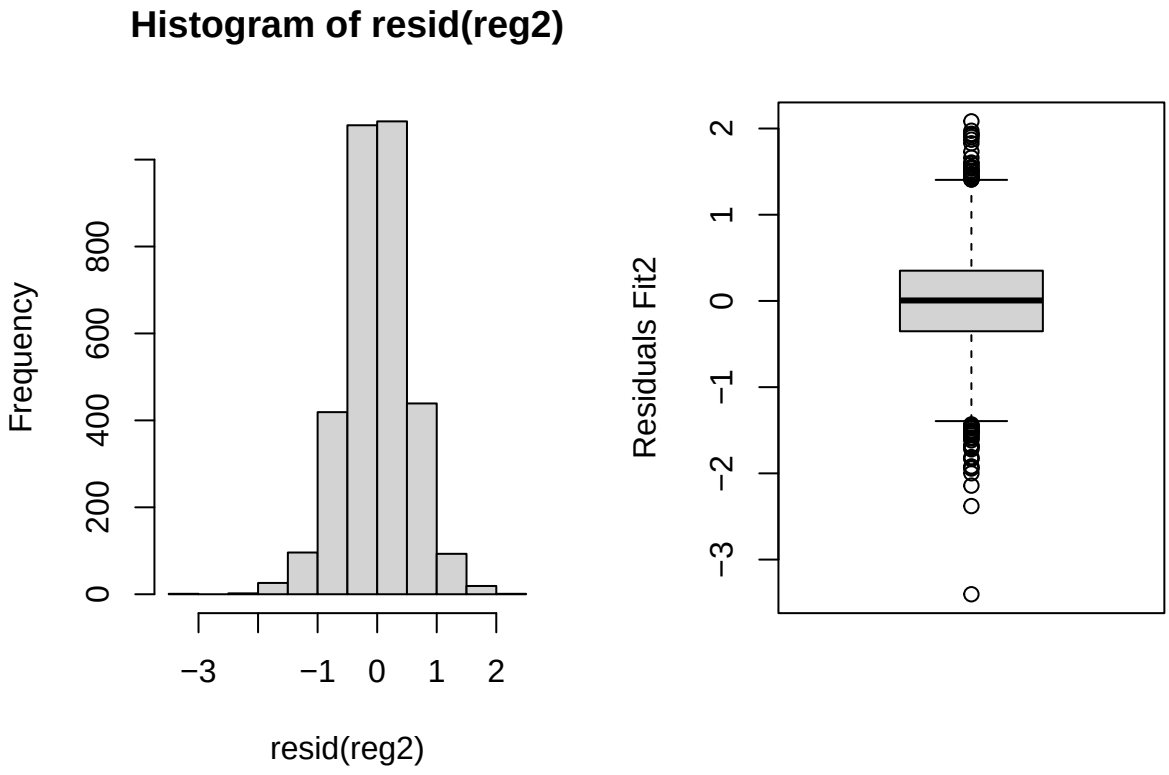
```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv, data = mto3$lpv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4019 -0.3515  0.0058  0.3513  2.0831
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.5221266  0.0319083 -79.043  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

```
#Examine Linear Regression
plot(reg2, c(1, 2))
```


Residuals vs Fitted
Residuals
Fitted values
lm(econ_ss_zscore ~ age + lpv)
1141
2563
2799
−4
−3
−2
−1
0
1
2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5




```
par(mfrow = c(1, 2))
hist(resid(reg2))
boxplot(resid(reg2), ylab = "Residuals Fit2")
```




```
confint(reg1, level = 0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) -2.51490885 -2.39315120
## age          0.05977226  0.06261665
```

```
confint(reg2, level = 0.95)
```

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

```
#t.test(mto3$lpv$lpv, alternative = 'two.sided', conf.level = 0.90)
t.test(econ_ss_zscore ~ lpv, data = mto3$lpv, var.equal = TRUE,
       alternative = 'two.sided', conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  econ_ss_zscore by lpv
## t = -4.8027, df = 3261, p-value = 1.636e-06
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.23729148 -0.09971096
## sample estimates:
## mean in group 0 mean in group 1
##     -0.04384251      0.12465871
```

#Running Tests 1) The linear assumption holds for this plot. The mean of residuals appears to be near zero in each segment of the plot as we move from left to right. The regression coefficients and R-squared value are unbiased and interpretable.

2) Constance variance holds. The spread of residuals is similar as we move from left to right across the residual plot. The Residual Standard Error is unbiased and interpretable.

3) Since the residuals have a round shaped distribution, and are symmetric with most values near the mean value, Normality holds. the standard error, T value, and p-value for the slope are unbiased and interpretable.

#Interpretations Our model results suggests that a one unit increase in age of household is associated with an approximate .0611 point increase in the mean z_score of economic self sufficiency. For the age of a household is equal to 0, the economic efficiency shows to be -2.454. This y-intercept value is not meaningful to interpret, since it falls outside the scope of the data (which has our lowest age at 18).

R-squared: 69% of the variation in economic self sufficiency (it varied from -3.23 to 2.93 and is now varying -3.4019 to 2.0831) is accounted for by its linear relationship with the age of household and the lpv treatment.

RMSE: On average, the distance the points are from the estimated regression line for economic self sufficiency by age of household is 0.5604. This is the average distance each data point's actualeconomic z-score is away from the mean of 'econ_ss_zscore' for those observations with the same age of household value.

The p-value for the slope coefficient on age is far smaller than the stated alpha value, so we will reject the Null Hypothesis that this coefficient equals zero in favor of the Alternative Hypothesis that it does not equal zero.

Sampling distribution under the Null Hypothesis: The shape of the sampling distribution for Beta1_Hat over repeated sampling is a moundshaped, t-distribution with 3261 degrees of freedom. The mean of this sampling distribution is 0, given that Beta1 in the Null Hypothesis is set equal to 0. The estimated standard error of this sampling distribution is 0.0007254. The points on the horizontal axis should be labeled .0007, 0.0014, 0.0021,0 0.0007, 0.0014, 0.00215.

For the change in the mean of econ_ss_zscore associated with a one unit change in age (the slope of the line when a equals 0) we get a ci of (0.98, 1.25), for the change in the mean of econ_ss_zscore when lpv goes from zero to one (holding age constant) we get (0.118, 0.195). Over repeated sampling with the same regression model run on each data set, 95% of the confidence intervals constructed in this way will contain the true slope between lpv and economic_ss_zscore and 5% will not. The change in the mean of economic self sufficiency with a one unit change in lpv is unlikely to be less than 0.118 or larger than 0.194.

The confidence intervals are positive decimals that do not contain the value of zero. Stated again, the confidence intervals do not contain the Null hypothesis value,(and together with the incredibly tiny p-value) signals that the Alternative is true. This 2 sided test communicates that while holding age constant, lpv has an impact on low poverty neighborhoods of 0.156 increase per 1 unit shift in econ_ss_zscore. This is on a meaningful scale.

Question 6

The researchers also hypothesize that the low poverty voucher will have different effects on economic self-sufficiency (relative to control) 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 alternative hypotheses? Use a two-sided hypothesis test.

Answer 6

$$E_i = \hat{\alpha} + \hat{\beta}_1 * A_i + \hat{\beta}_2 * L_i + \hat{\beta}_3 * A_i * L_i + \hat{\epsilon}_i$$

The linear model that would aptly address or estimate the parameter of interest, the mean change in economic self-sufficiency in the data generating mechanism, is econ_zz_score as equal to the Y-intercept (alpha_hat) plus the slope (Beta1_hat) multiplied by the covariate for age of household, (A_i), plus the slope (Beta2_hat) multiplied by the covariate for whether or not a household received the treatment (a low poverty voucher) (L_i) plus an interaction between covariates age and lpv (A_i * L_i) times the slope (Beta3_hat) plus the estimated epsilon (the residual hat).

The parameter of interest is Beta3. The Null Hypothesis is that Beta3 equals 0. The Alternative Hypothesis is that Beta3 is not equal to 0.

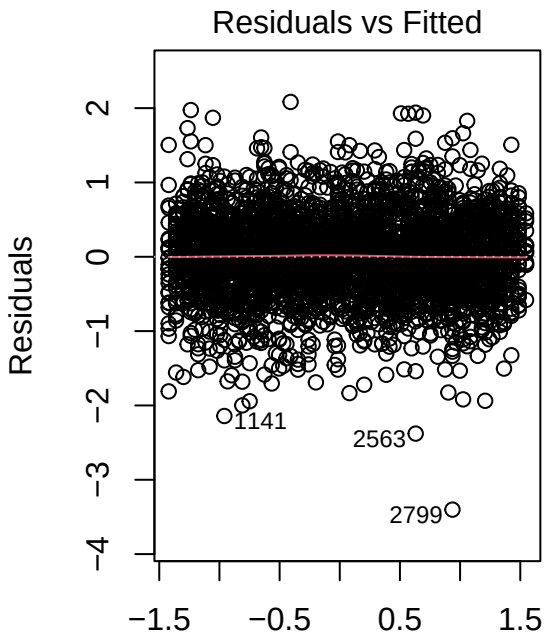
```
#Run Linear Regression
reg3 <- lm(econ_ss_zscore ~ age + lpv + (age:lpv), mto3$lpv)
reg3
```

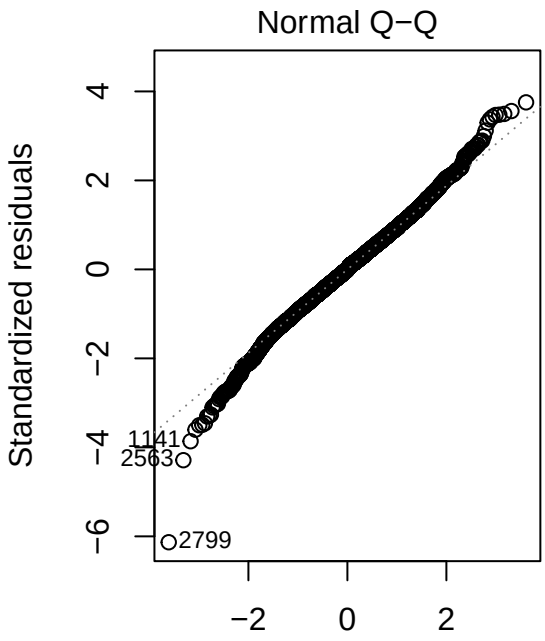
```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv + (age:lpv), data = mto3$lpv)
##
## Coefficients:
## (Intercept)          age          lpv      age:lpv
##   -2.497480     0.060545     0.098984     0.001415
```

```
summary(reg3)
```

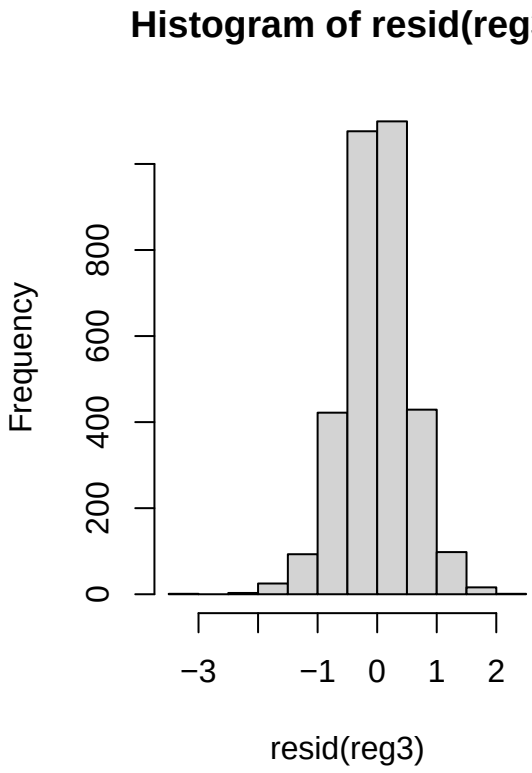
```
##
## Call:
## lm(formula = econ_ss_zscore ~ age + lpv + (age:lpv), data = mto3$lpv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4126 -0.3499  0.0018  0.3528  2.0902
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.4974800  0.0407065 -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
```

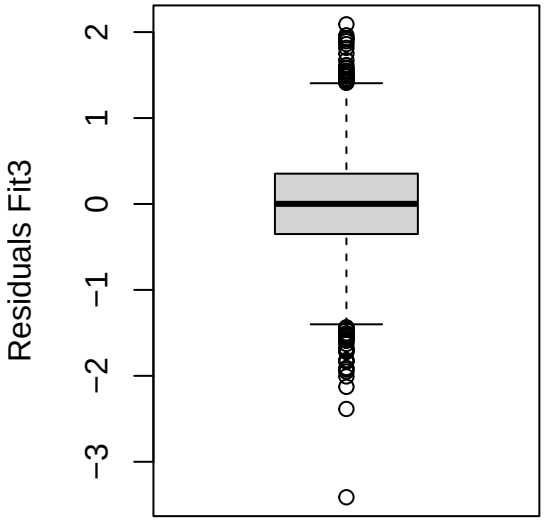
```
#Examine Linear Regression
par(mfrow = c(1, 2))
plot(reg2, c(1, 2))
```






```
hist(resid(reg3))
boxplot(resid(reg3), ylab = "Residuals Fit3")
```






```
mto3$lpv %>%
  ggplot(aes(y = econ_ss_zscore, x = age, color = factor(lpv))) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
```

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


```
confint(reg3, level = 0.90)
```

```
##                         5 %         95 %
## (Intercept) -2.5644553312 -2.430504767
## age          0.0589794065  0.062110156
## lpv         -0.0032466999  0.201214356
## age:lpv     -0.0009727013  0.003803087
```

```
t.test(mto3$lpv$lpv, alternative = 'two.sided', conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  mto3$lpv$lpv
## t = 51.236, df = 3262, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  0.4315894 0.4602280
## sample estimates:
## mean of x
## 0.4459087
```

#Running Tests 1) The linear assumption holds for this plot. The mean of residuals appears to be near zero in each segment of the plot as we move from left to right. The regression coefficients and R-squared value are unbiased and interpretable.

2) Constance variance holds. The spread of residuals is similar as we move from left to right across the residual plot. The Residual Standard Error is unbiased and interpretable.

3) Since the residuals have a round shaped distribution, and are symmetric with most values near the mean value, Normality holds. the standard error, T value, and p-value for the slope are unbiased and interpretable.

We estimate that the interaction is 0.0014, meaning that for units where a = 1 our slope decreases (flattens) by 0.0014 relative to the group where a = 0. However, this term is not statistically significant significant ($p > 0.10$) so we would not reject the null hypothesis that the slopes of the two lines are the same. The p-values on the other terms are small, indicating that the terms are statistically significantly different from zero - so the slopes are non-zero and the distance between the lines is non-zero.

The p-value is 0.330, meaning that 33% of the time we will expect the Null Hypothesis to fall within our expected data set. Given the alpha level is set to .10 for this test, this p-value does not support the hypothesis.

We find that the R2 is 0.69, meaning that 69% of the variation in the econ_ss_zscore is explained by a linear relationship with age, lpv, and their interaction. The residual standard error is 0.56, meaning that on average our observations lie about half a unit away from the regression line. In other words, each observation's outcome value is on average only 0.56 units away from the mean outcome for observations with the same age and lpv values.

Sampling distribution of the beta3_hat under the Null Hypothesis: The shape of the sampling distribution for beta3_hat over repeated sampling is a t-distribution with 3259 degrees of freedom. The mean of this sampling distribution under the null hypothesis is zero. The estimated standard error of this sampling distribution is 0.0014513. The points on the horizontal axis should be labeled -0.0045, -0.0030, -0.0015, 0, 0.0015, 0.0030, 0.0045.

For the change in the mean of econ_ss_zscore associated with a one unit change in age (the slope of the line when a equals 0) we get a ci of (0.98, 1.25), for the change in the mean of econ_ss_zscore when lpv goes from zero to one (holding age constant) we get (0.118, 0.195) and for the interaction we get (-0.0009, 0.0038). Over repeated sampling and estimation of this regression equation, 90% of the confidence intervals constructed in this way will contain the Beta3 parameter lpv*age and 10% will not. As the last of these confidence intervals contains zero, we cannot reject that the difference in slopes between the two lines is zero (The Null Hypothesis).

We fail to reject the Null that Beta3 = 0. It means repeated testing or scaled testing may be necessary to gain more concrete conclusions oer who should received low poverty vouchers by age of household or not.