PS5_12265092

12265092

16/02/2022

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

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

```
library(estimatr)
```

```
## Warning: package 'estimatr' was built under R version 4.1.2
```

```
library(Rcpp)
library(readxl)
library(haven)
library(boot)
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.1.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.1.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
data <- read.csv("fatality.csv")
```

```r
#Question 1

##a

reg_1 <- lm(formula = mrall ~ beertax + unrate + pop, data = data)

summary(reg_1)
```

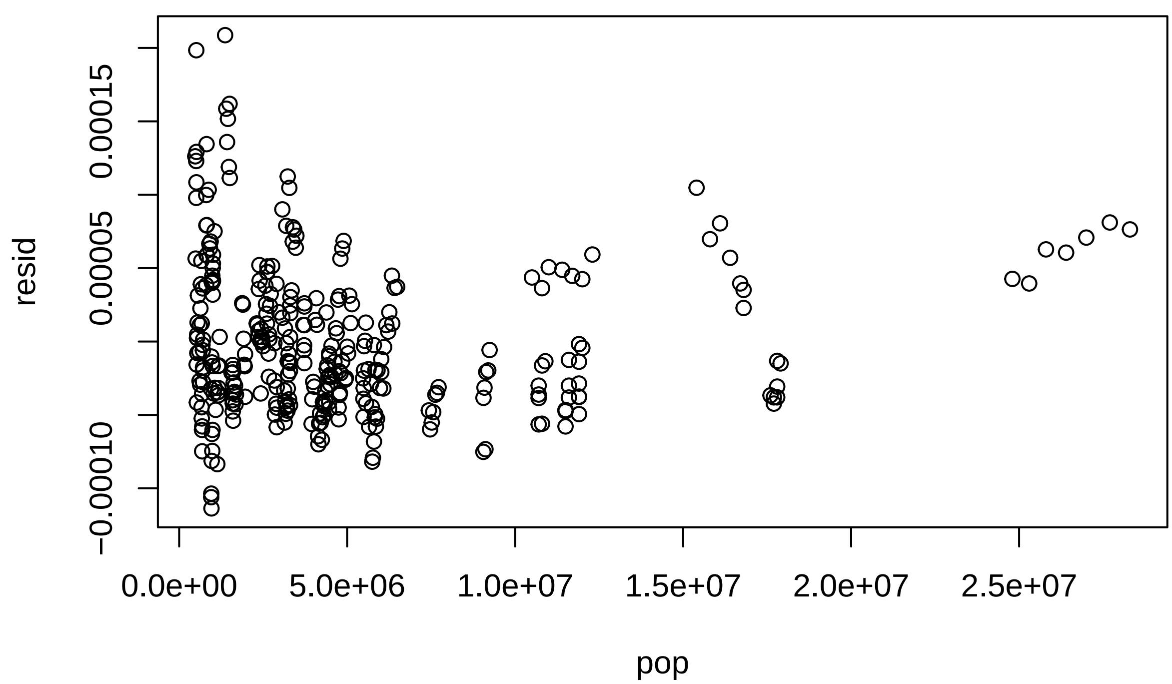
```
##
## Call:
## lm(formula = mrall ~ beertax + unrate + pop, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.137e-04 -3.687e-05 -1.342e-05  3.151e-05  2.087e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.718e-04  9.344e-06  18.389  < 2e-16 ***
## beertax      3.252e-05  5.951e-06   5.464 9.15e-08 ***
## unrate       4.061e-06  1.122e-06   3.621  0.00034 ***
## pop         -2.903e-12  5.612e-13  -5.172 4.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.174e-05 on 332 degrees of freedom
## Multiple R-squared:  0.1841, Adjusted R-squared:  0.1768
## F-statistic: 24.98 on 3 and 332 DF,  p-value: 1.349e-14
```

```r
#Intercept estimate 1.718e-04, se = 9.344e-06
#Coefficient estimate of beertax = 3.252e-05  se = 5.951e-06
#Coefficient estimate of unrate = 4.061e-06  se = 1.122e-06
#s e of beertax = 5.951e-06
#s e of unrate = 1.122e-06
#s e of pop = 5.612e-13
#Coefficient estimate of pop = -2.903e-12 se = 5.612e-13
#F-statistic: 24.98
#p-value: 1.349e-14

#From p-value and t-statistics we can say that the
#variables beertax, unrate, pop are statistically
#significant
```

```r
##b

#Suppose you think that the variance of u depends on population size. What are the
```

```r
#implications of this for the OLS standard errors and test statistics?

##If the variance of u depends on population size, then the homoskedasticity
##assumption does not hold and will result in heteroskedasticity.
##Heteroskedasticity does not cause any bias in the OLS estimators
##As the homoskedasticity assumption does not hold anymore, the variances of
##the estimates are no longer correct. They are different from before and are
##valid. Hence, we cannot use the variances of the estimates to decide on
##testing a hypothesis or in building a confidence interval and t statistic.
##However, the R-squared and the adj-R squared are not affected by heteroskedasticity.
##The standard errors are also not valid and cannot be used for testing a
##hpothesis or building a confidence interval as the case is same with variances
##The tstatistic will no more have a tdistribution even with a large sample.
```

```r
##c

##Estimate the model and capture the residuals in a new variable, resid. Present
##a scatterplot with resid on the y-axis pop on the x-axis.

resid <- resid(reg_1)

plot(data$pop, resid, xlab = "pop")
```



```r
#Do these data ##support your belief that the variance of u
#depends on population size?

##The scatter plot shows that the variance in u is not consistent with pop
##and is changing with pop. This means that there is  heteroskedasticity
##observed in the regression
```

```r
##d

#Carry out a Breusch-Pagan test for heteroskedasticity.
#Report the F-statistic and p-value from this test and state your conclusion.

bptest(reg_1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_1
## BP = 12.49, df = 3, p-value = 0.005879
```

```r
bptest(reg_1, ~ beertax + unrate + pop, data = data)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_1
## BP = 12.49, df = 3, p-value = 0.005879
```

```r
#BP test lm statistic = 12.49
#df = 3
#H0: Heteroskedasticity is not present
#Alternate Hypothesis : Heteroskedasticity is present
#Chi-squared distribution critical value at 5% level is 7.81
# Lm statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis


resid_square <- resid * resid

reg_2 <- lm(formula = resid_square ~ beertax + unrate + pop, data = data)

summary(reg_2)
```

```
##
## Call:
## lm(formula = resid_square ~ beertax + unrate + pop, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.800e-09 -2.305e-09 -1.233e-09  3.230e-10  3.987e-08
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.023e-09  8.526e-10   3.545 0.000448 ***
## beertax     -1.611e-09  5.430e-10  -2.967 0.003225 **
## unrate       1.309e-10  1.023e-10   1.279 0.201799
## pop         -1.040e-16  5.121e-17  -2.032 0.042990 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.721e-09 on 332 degrees of freedom
## Multiple R-squared:  0.03717,    Adjusted R-squared:  0.02847
## F-statistic: 4.273 on 3 and 332 DF,  p-value: 0.005598
```

```r
R_squared_reg2 <- summary(reg_2)$r.squared

F_stat_reg2 <- summary(reg_2)$fstatistic[1]

##Residual standard error: 4.721e-09 on 332 degrees of freedom
##Multiple R-squared:  0.03717
##Adjusted R-squared:  0.02847
##F-statistic: 4.273
##p-value: 0.005598

#df = 3, denominator df = 332
#H0: Heteroskedasticity is not present
#Alternate Hypothesis : Heteroskedasticity is present
#F distribution critical value at 5% level is 2.37
# F statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis
```

```r
##e

##Carry out the White test for heteroskedasticity, according to the
##standard (not alternate) form of the procedure described in class.
##Report the F-statistic and p-value from this test and state your conclusion.
unrate_squared <- data$unrate * data$unrate
pop_squared <- data$pop * data$pop
beertax_squared <- data$beertax * data$beertax

#interrelation variables, multiply each variable with other
unrate_pop <- data$unrate * data$pop
unrate_beertax <- data$unrate * data$beertax

pop_beertax <- data$pop * data$beertax

reg_3 <- lm(formula = resid_square ~ beertax + unrate + pop + beertax_squared +
              unrate_squared + pop_squared + unrate_beertax
            + unrate_pop + pop_beertax, data = data)

summary(reg_3)
```

```
##
## Call:
```

```
## lm(formula = resid_square ~ beertax + unrate + pop + beertax_squared +
##     unrate_squared + pop_squared + unrate_beertax + unrate_pop +
##     pop_beertax, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.698e-09 -2.039e-09 -8.590e-10  7.350e-10  3.787e-08
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.386e-09  2.236e-09   1.514 0.130964
## beertax         -9.203e-09  2.693e-09  -3.417 0.000713 ***
## unrate           1.049e-09  4.558e-10   2.301 0.021999 *
## pop             -1.050e-15  2.804e-16  -3.743 0.000215 ***
## beertax_squared  1.309e-09  7.153e-10   1.831 0.068088 .
## unrate_squared  -4.938e-11  2.662e-11  -1.855 0.064543 .
## pop_squared      3.757e-23  7.106e-24   5.287 2.29e-07 ***
## unrate_beertax   1.157e-10  2.388e-10   0.485 0.628259
## unrate_pop      -1.598e-17  2.494e-17  -0.641 0.522016
## pop_beertax      8.901e-16  2.258e-16   3.942 9.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.482e-09 on 326 degrees of freedom
## Multiple R-squared:  0.1479, Adjusted R-squared:  0.1244
## F-statistic: 6.287 on 9 and 326 DF,  p-value: 3.293e-08
```

```
##F-statistic: 6.287 on 9 and 326 DF
##p-value: 3.293e-08
##Adjusted R-squared:  0.1244
#Multiple R-squared:  0.1479
##
#k = 9, denominator df = 326
#H0: Heteroskedasticity is not present
#Alternate Hypothesis : Heteroskedasticity is present
#F distribution critical value at 5% level fove above is 1.88
# F statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis

R_squared_reg3 <- summary(reg_3)$r.squared

lm_stat_reg_3 <- length(resid_square) * R_squared_reg3
lm_stat_reg_3
```

```
## [1] 49.69352
```

```
##lm stat = 49.69352
#k = 9, denominator df = 326
#H0: Heteroskedasticity is not present
#Alternate Hypothesis : Heteroskedasticity is present
#Chi-square distribution critical value at 5% level for above is 16.92
# lm statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis
```

```r
##f

##Based on the results of the preceding tests, you decide to estimate
##the model using heteroskedasticity-robust standard errors.
##Report the coefficients and standard errors on beertax, unrate, and pop.
##Did the coefficients change from those you calculated in part a?
##What about the standard errors? What do your findings suggest about
##the presence or absence of heteroskedasticity?



summary(reg_1, robust = T)
```

```
##
## Call:
## lm(formula = mrall ~ beertax + unrate + pop, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.137e-04 -3.687e-05 -1.342e-05  3.151e-05  2.087e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.718e-04  9.344e-06  18.389  < 2e-16 ***
## beertax      3.252e-05  5.951e-06   5.464 9.15e-08 ***
## unrate       4.061e-06  1.122e-06   3.621  0.00034 ***
## pop         -2.903e-12  5.612e-13  -5.172 4.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.174e-05 on 332 degrees of freedom
## Multiple R-squared:  0.1841, Adjusted R-squared:  0.1768
## F-statistic: 24.98 on 3 and 332 DF,  p-value: 1.349e-14
```

```r
robust_reg1 <- lm_robust(mrall ~ beertax + unrate + pop, data = data)

summary(robust_reg1)
```

```
##
## Call:
## lm_robust(formula = mrall ~ beertax + unrate + pop, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper  DF
## (Intercept)  1.718e-04  8.503e-06  20.208 8.881e-60  1.551e-04  1.886e-04 332
## beertax      3.252e-05  5.272e-06   6.167 2.021e-09  2.214e-05  4.289e-05 332
## unrate       4.061e-06  1.005e-06   4.042 6.582e-05  2.085e-06  6.037e-06 332
## pop         -2.903e-12  6.509e-13  -4.460 1.126e-05 -4.183e-12 -1.622e-12 332
##
## Multiple R-squared:  0.1841 ,    Adjusted R-squared:  0.1768
## F-statistic: 30.22 on 3 and 332 DF,  p-value: < 2.2e-16
```

```
##Change in F statistic from 24.98 -> 30.22
##Coefficient estimates are same in both the regular and robust regressions
##standard errors are slightly different.
##se decreased for beertax, se decreased for unrate, se increased for pop
##p value in robust version  = 2.2e-16 , regular p-value = 1.349e-14
##p value is not according to the standard significance level.
##As we can see from the above, we cannot say whether heteroskedasticity is
##present or not on basis of robust test. Other than the fact that the
##standard errors differed and not valid, we cannot conclude concretely on
##whether heteroskedasticity is present or not.
##This is because, we observe that the change in standard
##errors in the robust tests cannot be considered
##significantly enough.We can conclude hetereroskedastic
##if we consider the change in s.e to be significant
```

```
##g
##Suppose you determine that the variance of ui can be consistently estimated by hi_hat . How
##could you transform the equation so that the transformed equation has error variances that
##do not vary with the explanatory variables?  Why might running OLS on this transformed
##equation be preferable to the procedure you identified in (f)



##hi_hat is a function of xi.
##E[ ui /sqrt(hi_hat) | xi ] = 0
##Above means expected value of ui/sqrt(hi_hat) conditional on xi is zero
##Var(ui /sqrt(hi_hat)) = var(hi)
##We get the above from Var (ui | xi) = var(ui)*hi_hat
##We can show that equation in a can be dividied by sqrt(hi_hat)
##After dividing with sqrt(hi_hat), the equation will be
##linear in its parameters.
##From above calculated values, we can say that the modified equation
##satisfies all Gauss-markov assumptions from MLR 1 thrrough MLR6. This means
##that modified ui will have normal distribution. The variance of
#modified ui will be same. So the modified
##equation will be valid for all MLR assumptions.
##We simply do the OLS anlaysis of the modified equation
##The estimates, s e, statistics can now be derived from the modified equation
##Estimators will be differnt in modified equation from the original equation
##The estimators derived from the modified equation are more effieicient in
##estimating than the original equation due to the said reasons.
##We can also derive SSR from the modified equation. SSR/df will now become an
##estimator for the variance of modified residual.
```

```
##h

reg_2_h <- lm(formula = mrall ~ beertax + unrate + pop, weights = pop, data=data)

summary(reg_2)
```

```
##
## Call:
## lm(formula = resid_square ~ beertax + unrate + pop, data = data)
```

```
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.800e-09 -2.305e-09 -1.233e-09  3.230e-10  3.987e-08
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.023e-09  8.526e-10   3.545 0.000448 ***
## beertax     -1.611e-09  5.430e-10  -2.967 0.003225 **
## unrate       1.309e-10  1.023e-10   1.279 0.201799
## pop         -1.040e-16  5.121e-17  -2.032 0.042990 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.721e-09 on 332 degrees of freedom
## Multiple R-squared:  0.03717,    Adjusted R-squared:  0.02847
## F-statistic: 4.273 on 3 and 332 DF,  p-value: 0.005598
```

```
##Residual standard error: 0.09272 on 332 degrees of freedom
##Multiple R-squared:  0.2749,
##Adjusted R-squared:  0.2683
##F-statistic: 41.95 on 3 and 332 DF,
##p-value: < 2.2e-16

#k = 3, denominator df = 332
#H0: Heteroskedasticity is not present
#Alternate Hypothesis : Heteroskedasticity is present
#F distribution critical value at 5% level for above is 2.37
# F statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis


bptest(reg_2_h)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_2_h
## BP = 12.49, df = 3, p-value = 0.005879
```

```
#BP value = 12.49, df = 3, p-value = 0.005879

##calculating bp lm statistic
resid_squared_h <- resid(reg_2_h) * resid(reg_2_h)
reg_3_h <- lm(formula = resid_squared_h ~ beertax + unrate + pop, data = data)
summary(reg_3_h)
```

```
##
## Call:
## lm(formula = resid_squared_h ~ beertax + unrate + pop, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
```

```
## -6.479e-09 -3.212e-09 -1.606e-09  6.010e-10  5.381e-08
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.699e-09  1.188e-09   3.115  0.00200 **
## beertax     -2.211e-09  7.563e-10  -2.923  0.00371 **
## unrate       2.738e-10  1.426e-10   1.921  0.05563 .
## pop         -2.964e-16  7.133e-17  -4.156 4.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.575e-09 on 332 degrees of freedom
## Multiple R-squared:  0.07211,    Adjusted R-squared:  0.06372
## F-statistic:   8.6 on 3 and 332 DF,  p-value: 1.633e-05
```

```
##Residual standard error: 6.575e-09 on 332 degrees of freedom
##Multiple R-squared:  0.07211,
##Adjusted R-squared:  0.06372
##F-statistic:   8.6 on 3 and 332 DF,
##p-value: 1.633e-05

#k = 3, denominator df = 332
#H0: Heteroskedasticity is not present
#Alternate Hypothesis : Heteroskedasticity is present
#F distribution critical value at 5% level for above is 2.37
# F statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis

R_squared_reg3_h <- summary(reg_3_h)$r.squared

bp_lm_statistic <- length(resid_squared_h) * R_squared_reg3_h

bp_lm_statistic
```

```
## [1] 24.22743
```

```
##bp lm statistic  = 24.22743
#H0: Heteroskedasticity is not present
#Alternate Hypothesis : Heteroskedasticity is present
#Chi-square distribution critical value at 5% level for above is 7.8
# lm statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis

##The weighting process followed in the above resulted in p-values that validate
##presence of heteroskedasticity. This is true for both
##the cases.
```

```
##i

reg_i <- lm(mrall ~ beertax + unrate + pop , data = data)

resid_i <- residuals(reg_i)
```

```r
resid_i2 <- (resid_i)^2

data$resid_i2 <- resid_i2

inverse_pop = 1/data$pop

reg_i2 <- lm(resid_i2 ~ inverse_pop, data = data)

summary(reg_i2)
```

```
##
## Call:
## lm(formula = resid_i2 ~ inverse_pop, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.116e-09 -1.930e-09 -1.083e-09  4.350e-10  4.035e-08
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.384e-09  3.584e-10   3.862 0.000135 ***
## inverse_pop 2.511e-03  5.064e-04   4.958 1.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.629e-09 on 334 degrees of freedom
## Multiple R-squared:  0.06856,    Adjusted R-squared:  0.06577
## F-statistic: 24.58 on 1 and 334 DF,  p-value: 1.134e-06
```

```r
##Residual standard error: 4.629e-09 on 334 degrees of freedom
##Multiple R-squared:  0.06856,
##Adjusted R-squared:  0.06577
##F-statistic: 24.58 on 1 and 334 DF,
##p-value: 1.134e-06


## The estimated coefficient of inverse pop (2.511e-03) is significantly
##different from zero.
##The above can be said from using t statistic and p-values
##This means, hypothese hetereskedasticity is present cannot be rejected

##in h, our assumption that the states with larger populations
##will have more precisely measured vehicle fatality rates is not correct.

##The above will conclude that the original regression of mrall on beertax, pop
##and unrate did not result in homoskedasticity.

##IV

fitted = fitted(reg_i2)

inverse_fitted = 1/fitted
```

```r
fitted_reg <- lm(formula = mrall ~ beertax + pop + unrate,
                 weights = inverse_fitted, data = data)

summary(fitted_reg)
```

```
##
## Call:
## lm(formula = mrall ~ beertax + pop + unrate, data = data, weights = inverse_fitted)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7735 -0.7219 -0.1818  0.7319  3.9009
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.693e-04  8.947e-06  18.928  < 2e-16 ***
## beertax      4.033e-05  5.142e-06   7.843 6.06e-14 ***
## pop         -1.940e-12  4.662e-13  -4.161 4.03e-05 ***
## unrate       2.612e-06  1.024e-06   2.550   0.0112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9946 on 332 degrees of freedom
## Multiple R-squared:  0.2311, Adjusted R-squared:  0.2242
## F-statistic: 33.27 on 3 and 332 DF,  p-value: < 2.2e-16
```

```r
##Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
##(Intercept)  1.693e-04  8.947e-06  18.928  < 2e-16 ***
##beertax      4.033e-05  5.142e-06   7.843 6.06e-14 ***
##pop         -1.940e-12  4.662e-13  -4.161 4.03e-05 ***
##unrate       2.612e-06  1.024e-06   2.550   0.0112 *


##Residual standard error: 0.9946 on 332 degrees of freedom
##Multiple R-squared:  0.2311,
##Adjusted R-squared:  0.2242
##F-statistic: 33.27 on 3 and 332 DF,
##p-value: < 2.2e-16


##bp test

resid_squared_h_IV <- resid(fitted_reg) * resid(fitted_reg)
bp_reg_hIV <- lm(resid_squared_h_IV ~ beertax + unrate + pop, data = data)
summary(bp_reg_hIV)
```

```
##
## Call:
## lm(formula = resid_squared_h_IV ~ beertax + unrate + pop, data = data)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -4.771e-09 -2.614e-09 -1.324e-09  3.760e-10  4.490e-08
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.313e-09  9.574e-10   3.461 0.000609 ***
## beertax     -1.764e-09  6.097e-10  -2.893 0.004067 **
## unrate       1.563e-10  1.149e-10   1.360 0.174608
## pop         -1.723e-16  5.750e-17  -2.996 0.002944 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.301e-09 on 332 degrees of freedom
## Multiple R-squared:  0.04814,    Adjusted R-squared:  0.03954
## F-statistic: 5.597 on 3 and 332 DF,  p-value: 0.000937
```

```
##Coefficients:

#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)  3.313e-09  9.574e-10   3.461 0.000609 ***
#beertax     -1.764e-09  6.097e-10  -2.893 0.004067 **
#unrate       1.563e-10  1.149e-10   1.360 0.174608
#pop         -1.723e-16  5.750e-17  -2.996 0.002944 **


#Residual standard error: 5.301e-09 on 332 degrees of freedom
#Multiple R-squared:  0.04814,
#Adjusted R-squared:  0.03954
#F-statistic: 5.597 on 3 and 332 DF,
#p-value: 0.000937

#k = 3, denominator df = 332
#H0: Heteroskedasticity is not present,  coefficients = 0
#Alternate Hypothesis : Heteroskedasticity is present
#F distribution critical value at 5% level for above is 2.37
# F statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis


R_square_resid_sq <- summary(bp_reg_hIV)$r.squared

R_square_resid_sq
```

```
## [1] 0.04813836
```

```
# R_square_resid_sq = 0.04813836

bp_lm_statistic_hIV <- length(resid_squared_h_IV) * R_square_resid_sq
bp_lm_statistic_hIV
```

```
## [1] 16.17449
```

```r
##bp lm statistic = 16.17449
#H0: Heteroskedasticity is not present, coefficients = 0
#Alternate Hypothesis : Heteroskedasticity is present
#Chi-square distribution critical value at 5% level for above is 7.8
# lm statistic is greater than the critical value
#We can reject the null hypothesis and cannot reject the alternate hypothesis

##The weighting process followed in the above resulted in p-values that validate
##presence of heteroskedasticity and do not elimate the alternate hypothesis.
```

```r
##j

pop_inv_2 = 1/data$pop

reg_j1 <- lm_robust(mrall ~ beertax + unrate + pop, data = data)
summary(reg_j1)
```

```
##
## Call:
## lm_robust(formula = mrall ~ beertax + unrate + pop, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper  DF
## (Intercept)  1.718e-04  8.503e-06  20.208 8.881e-60  1.551e-04  1.886e-04 332
## beertax      3.252e-05  5.272e-06   6.167 2.021e-09  2.214e-05  4.289e-05 332
## unrate       4.061e-06  1.005e-06   4.042 6.582e-05  2.085e-06  6.037e-06 332
## pop         -2.903e-12  6.509e-13  -4.460 1.126e-05 -4.183e-12 -1.622e-12 332
##
## Multiple R-squared:  0.1841 ,    Adjusted R-squared:  0.1768
## F-statistic: 30.22 on 3 and 332 DF,  p-value: < 2.2e-16
```

```r
#Coefficients:
#               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper  DF
#(Intercept)  1.718e-04  8.503e-06  20.208 8.881e-60  1.551e-04  1.886e-04 332
#beertax      3.252e-05  5.272e-06   6.167 2.021e-09  2.214e-05  4.289e-05 332
#unrate       4.061e-06  1.005e-06   4.042 6.582e-05  2.085e-06  6.037e-06 332
#pop         -2.903e-12  6.509e-13  -4.460 1.126e-05 -4.183e-12 -1.622e-12 332

#Multiple R-squared:  0.1841 ,  Adjusted R-squared:  0.1768
#F-statistic: 30.22 on 3 and 332 DF,
#p-value: < 2.2e-16

#using inv fitted
reg_j2 <- lm_robust(mrall ~ beertax + unrate + pop,
                    weights = inverse_fitted, data = data)
summary(reg_j2)
```

```
##
## Call:
## lm_robust(formula = mrall ~ beertax + unrate + pop, data = data,
```

```
##     weights = inverse_fitted)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper  DF
## (Intercept)  1.693e-04  8.172e-06  20.722 8.440e-62  1.533e-04  1.854e-04 332
## beertax      4.033e-05  5.048e-06   7.990 2.253e-14  3.040e-05  5.026e-05 332
## unrate       2.612e-06  9.306e-07   2.807 5.302e-03  7.812e-07  4.443e-06 332
## pop         -1.940e-12  5.739e-13  -3.381 8.097e-04 -3.069e-12 -8.111e-13 332
##
## Multiple R-squared:  0.2311 ,    Adjusted R-squared:  0.2242
## F-statistic: 31.92 on 3 and 332 DF,  p-value: < 2.2e-16
```

```
#Coefficients:
#               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper  DF
#(Intercept)  1.693e-04  8.172e-06  20.722 8.440e-62  1.533e-04  1.854e-04 332
#beertax      4.033e-05  5.048e-06   7.990 2.253e-14  3.040e-05  5.026e-05 332
#unrate       2.612e-06  9.306e-07   2.807 5.302e-03  7.812e-07  4.443e-06 332
#pop         -1.940e-12  5.739e-13  -3.381 8.097e-04 -3.069e-12 -8.111e-13 332

#Multiple R-squared:  0.2311 ,  Adjusted R-squared:  0.2242
#F-statistic: 31.92 on 3 and 332 DF,  p-value: < 2.2e-16

reg_j3 <- lm_robust(mrall ~ beertax + unrate + pop,
                    weights = inverse_pop, data = data)
summary(reg_j3)
```

```
##
## Call:
## lm_robust(formula = mrall ~ beertax + unrate + pop, data = data,
##     weights = inverse_pop)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper  DF
## (Intercept)  1.710e-04  1.401e-05 12.2038 1.499e-28  1.434e-04  1.985e-04 332
## beertax      1.358e-06  1.129e-05  0.1203 9.043e-01 -2.085e-05  2.356e-05 332
## unrate       9.122e-06  1.719e-06  5.3054 2.058e-07  5.740e-06  1.250e-05 332
## pop         -7.025e-12  1.110e-12 -6.3291 7.987e-10 -9.209e-12 -4.842e-12 332
##
## Multiple R-squared:  0.1555 ,    Adjusted R-squared:  0.1479
## F-statistic: 20.59 on 3 and 332 DF,  p-value: 2.911e-12
```

```
#Coefficients:
#               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper  DF
#(Intercept)  1.710e-04  1.401e-05 12.2038 1.499e-28  1.434e-04  1.985e-04 332
#beertax      1.358e-06  1.129e-05  0.1203 9.043e-01 -2.085e-05  2.356e-05 332
#unrate       9.122e-06  1.719e-06  5.3054 2.058e-07  5.740e-06  1.250e-05 332
#pop         -7.025e-12  1.110e-12 -6.3291 7.987e-10 -9.209e-12 -4.842e-12 332

#Multiple R-squared:  0.1555 ,  Adjusted R-squared:  0.1479
```

```
#F-statistic: 20.59 on 3 and 332 DF,  p-value: 2.911e-12



##robust results obtained in part f
#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)  1.718e-04  9.344e-06  18.389  < 2e-16 ***
#beertax      3.252e-05  5.951e-06   5.464 9.15e-08 ***
#unrate       4.061e-06  1.122e-06   3.621  0.00034 ***
#pop         -2.903e-12  5.612e-13  -5.172 4.01e-07 ***

#Residual standard error: 5.174e-05 on 332 degrees of freedom
#Multiple R-squared:  0.1841,   Adjusted R-squared:  0.1768
#F-statistic: 24.98 on 3 and 332 DF,  p-value: 1.349e-14




##The standard errors of the coefficients has decreased in the model as compared
##to the model from part f. This means that the model is more precise
```

```
##k

##The model where we regressed square of residuals obtained in standard OLS
##on inverse of pop, and we used robust model with weights equal to the inverse
##of the fitted values is more precise. Due to the reasons that we observed that
##the standard errors we obtained are less compared to other models.
##And the model also has better R suared and the Adjusted r squared models.
##Better goodness of fit values
##The statistics (F and lm) derived in the Breaush-Pagan test seem to be
##effective in rejecting the null hypothesis and validate presence of
##heteroskedasaticty which is lower.
```

```
#Question 2

#In the given question, the variable test score performance has to be
#estimated against the per pupil eucational expenditures at a district level
#and average student teacher ratios in the district. It does not make sense
#to include additional variables that cannot be held constant when we change
#the variable of interest. This is because of the ceteres paribus
#interpretation of multuvariate regression. If there is a change in
#independent variable "spending", we expect a similar change proportionally
#in the other independent variable "ratio" i.e we expect a reasonable change
#in average student teacher ratio when we change per pupil expenditure .
#However we donot want to keep ratio constant while there is a change in
#"spending". This is why we should not include any kind of control for the
#student teacher ratios along with per pupil educational expenditure variables
#in the multiple regression model.Adding more and more variables to explain a
#model leads to over controlling.
```

```
#Question 3

#model saving = B0 + B1educ + B2inc + B3hhsize + B4age + u
#E(u | inc, hhsize, educ, age) = 0

#a)

#Having just age data for age >=60 does not create bias in the estimates.
#However it will increase the error with which the model is estimated.
#Because the error will still be uncorrelated with the sample.
#This is because, having only age >= 60 will lead to a non-random sampling
#scenario where the missing data is unrelated to the variables,
#thus causing an non-random/exogenous sampling scenario.
#Moreover, the missing data does not infact disturb the
#conditional mean assumption. Hence, we get unbiased estimates
#and the standard error will increase.

#b)
#savings >= $10000
#Only married couples without children, hhsize = 2
#This case is similar to above scenario of non-random/exogenous sampling.
#Data is available for hhsize=2. Because the error will be still be
#uncorrelated with the hhsize variable.
#Hence missing data for values other than hhsize=2 will not create bias.
#Similar to b, the conditional mean is not disturbed, hence we can still
#estimate the parameters of the model and they will still remain unbiased.


#c)

#This case is different from above. The conditional mean will no longer be same.
#This is endogeneous selection of sample. We have selected a non-random sample
#on the basis of dependent variable. Thus this will lead to bias in the
#estimated parameters. Regarding the error, by selecting specific values
#of dependent variable(savings) we are again selecting only a few values of
#error(u) which depends on hhsize. This will lead to correlation between
#hhsize and u which will result in bias in the estimated coefficieints.
```