2)

**Model 1**

Explanatory Variables:

ß1 = AGE\_REF

ß2 = FAM\_SIZE

ß3 = VEHQ

ß4 = VEHQL

ß5 = ELCTRCPQ

ß6 = ALLFULPQ

ß7 = GASMOPQ

ß8 = HEALTHPQ

ß9 = HLTHINPQ

ß10 = HLTHINPQ

ß11 = PERSCACQ

Respond Variable:

Fraction of transport spent of the total expenditure.

rm(list = ls(all = TRUE))  
load("cex\_2012.RData")  
  
attach(data\_cex)  
  
fraction\_transport <- TRANSPQ/TOTEXPPQ # fraction of spent on transport of totalexpenditure  
fraction\_transport[is.infinite(fraction\_transport)] <- NA  
  
fraction\_transport[fraction\_transport<0] <- NA   
summary(fraction\_transport)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00000 0.05473 0.09977 0.12840 0.16310 0.99060 1

summary(fraction\_transport[as.logical(FAM\_SIZE)])

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00000 0.05473 0.09977 0.12840 0.16310 0.99060 1

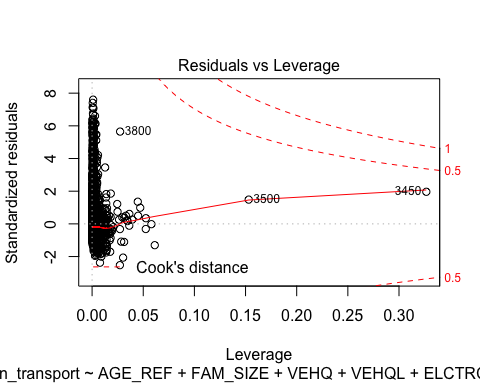
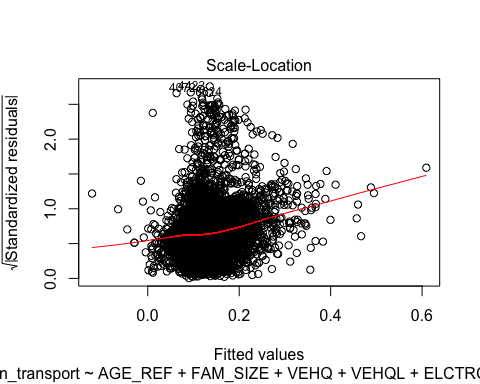
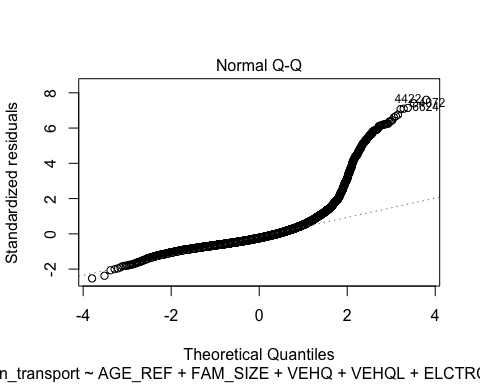
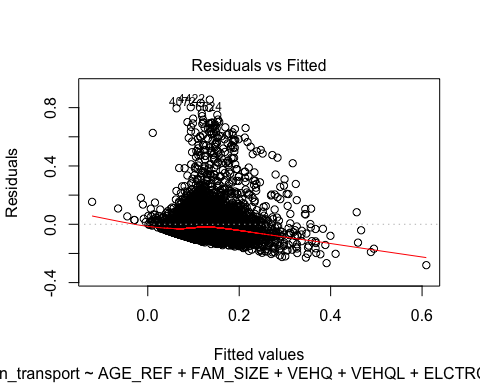
FAM\_SIZE[is.infinite(FAM\_SIZE)] <- NA  
summary(fraction\_transport[as.logical(VEHQ)])

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00000 0.06631 0.10930 0.13980 0.17100 0.99060 1

#make sure we do not have negative transport spent  
  
model1 <- lm(fraction\_transport~AGE\_REF + FAM\_SIZE + VEHQ + VEHQL + ELCTRCPQ + ALLFULPQ + GASMOPQ + HEALTHPQ + HLTHINPQ + HLTHINPQ + PERSCACQ)  
summary(model1)

##   
## Call:  
## lm(formula = fraction\_transport ~ AGE\_REF + FAM\_SIZE + VEHQ +   
## VEHQL + ELCTRCPQ + ALLFULPQ + GASMOPQ + HEALTHPQ + HLTHINPQ +   
## HLTHINPQ + PERSCACQ)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.28037 -0.05994 -0.02643 0.02302 0.85486   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.189e-01 5.532e-03 21.490 < 2e-16 \*\*\*  
## AGE\_REF -5.016e-04 8.432e-05 -5.948 2.84e-09 \*\*\*  
## FAM\_SIZE -5.835e-03 1.015e-03 -5.748 9.40e-09 \*\*\*  
## VEHQ 1.667e-02 1.086e-03 15.350 < 2e-16 \*\*\*  
## VEHQL 6.906e-02 6.601e-03 10.462 < 2e-16 \*\*\*  
## ELCTRCPQ -5.125e-05 7.121e-06 -7.197 6.82e-13 \*\*\*  
## ALLFULPQ -1.763e-05 8.945e-06 -1.971 0.0488 \*   
## GASMOPQ 8.496e-05 3.620e-06 23.473 < 2e-16 \*\*\*  
## HEALTHPQ -1.327e-05 2.425e-06 -5.474 4.55e-08 \*\*\*  
## HLTHINPQ 2.318e-06 3.495e-06 0.663 0.5072   
## PERSCACQ -5.401e-05 2.512e-05 -2.150 0.0316 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1125 on 6826 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.1658, Adjusted R-squared: 0.1646   
## F-statistic: 135.7 on 10 and 6826 DF, p-value: < 2.2e-16

plot(model1)



vcov(model1)# make sure that each explanatory variables are not strongly relateive.

## (Intercept) AGE\_REF FAM\_SIZE VEHQ  
## (Intercept) 3.060075e-05 -3.906128e-07 -2.951579e-06 -4.793648e-07  
## AGE\_REF -3.906128e-07 7.110117e-09 2.454770e-08 -7.133136e-09  
## FAM\_SIZE -2.951579e-06 2.454770e-08 1.030214e-06 -1.980058e-07  
## VEHQ -4.793648e-07 -7.133136e-09 -1.980058e-07 1.179018e-06  
## VEHQL -9.357687e-07 -2.304570e-10 -1.585383e-07 9.126504e-07  
## ELCTRCPQ -4.935543e-10 -7.192728e-11 -1.165085e-09 -4.061106e-10  
## ALLFULPQ 2.414568e-09 -4.954670e-11 -1.224377e-11 -1.859742e-10  
## GASMOPQ -2.750756e-09 4.530921e-11 -4.366865e-10 -1.224637e-09  
## HEALTHPQ 2.093912e-10 -9.727807e-12 5.027021e-11 -6.196866e-11  
## HLTHINPQ 1.059637e-09 -2.615571e-11 -5.701194e-11 -1.295076e-10  
## PERSCACQ -6.395250e-09 -8.477191e-11 -1.957706e-09 -4.092780e-09  
## VEHQL ELCTRCPQ ALLFULPQ GASMOPQ  
## (Intercept) -9.357687e-07 -4.935543e-10 2.414568e-09 -2.750756e-09  
## AGE\_REF -2.304570e-10 -7.192728e-11 -4.954670e-11 4.530921e-11  
## FAM\_SIZE -1.585383e-07 -1.165085e-09 -1.224377e-11 -4.366865e-10  
## VEHQ 9.126504e-07 -4.061106e-10 -1.859742e-10 -1.224637e-09  
## VEHQL 4.356988e-05 -8.565519e-10 -9.704407e-10 -1.943148e-09  
## ELCTRCPQ -8.565519e-10 5.070959e-11 -5.289913e-12 -7.386659e-12  
## ALLFULPQ -9.704407e-10 -5.289913e-12 8.000748e-11 -7.554025e-13  
## GASMOPQ -1.943148e-09 -7.386659e-12 -7.554025e-13 1.310139e-11  
## HEALTHPQ -1.758725e-10 -1.473246e-12 -6.404387e-13 -6.194293e-13  
## HLTHINPQ -1.072078e-09 -2.203213e-13 5.770182e-14 -4.717968e-16  
## PERSCACQ -9.951979e-09 1.978195e-11 -7.456583e-12 7.151723e-12  
## HEALTHPQ HLTHINPQ PERSCACQ  
## (Intercept) 2.093912e-10 1.059637e-09 -6.395250e-09  
## AGE\_REF -9.727807e-12 -2.615571e-11 -8.477191e-11  
## FAM\_SIZE 5.027021e-11 -5.701194e-11 -1.957706e-09  
## VEHQ -6.196866e-11 -1.295076e-10 -4.092780e-09  
## VEHQL -1.758725e-10 -1.072078e-09 -9.951979e-09  
## ELCTRCPQ -1.473246e-12 -2.203213e-13 1.978195e-11  
## ALLFULPQ -6.404387e-13 5.770182e-14 -7.456583e-12  
## GASMOPQ -6.194293e-13 -4.717968e-16 7.151723e-12  
## HEALTHPQ 5.878944e-12 -6.503634e-12 -1.198854e-12  
## HLTHINPQ -6.503634e-12 1.221193e-11 2.584072e-12  
## PERSCACQ -1.198854e-12 2.584072e-12 6.312047e-10

**Explanation(model1)**:

In this linear model equation, we set the alpha=0.05. In order for the linear regression model work, we need to determine whether the linear regression model is statistical significance. To do so, we need to use the null and alternative hypothesis for ANOVA and individual coefficient hypothesis test.

**Hypothesis test for ANOVA:**

Null: ß1 = ß2 = …… = B11

Alternative: At least one coefficient does not equal to 0

If the all of coefficient equal to 0, it is telling that they are not significant linear relationship between the variables, and should not be used for linear regression.

F-statistic: 135.7 on 10 and 6823 DF, p-value: < 2.2e-16

The result indicate that we should reject the null hypothesis that ß1 through ß11 have no effect on fraction of transport spent. Also, since the P-value < 0.05, this linear model is statistical significance at 5% level of significance.

**Hypothesis for individual coefficient:**

Null: ßk= 0

Alternative: ßk ≠ 0

Where k = 1, 2, 3,……, 11

Bases on my result from R, ALLFULPQ(ß6) and HLTHINPQ(ß10) have T statistic of -1.971 & 0.663, and P-value of 0.0488 & 0.5072. Since the P-value is greater than 0.05, we fail to reject the null hypothesis. Then, these two variables are not statistical significance and not useful to predicting the fraction of transport spent of the total expenditure.

The coefficient of determination(R^2) is 0.1658, and it indicates that there is weak linear relationship between the explanatory variables (ß1 to ß11) and the respond variable (fraction of transport spent). The explanatory variable variables (ß1 to ß11) explain approximate 16.58% of variation in fraction of transport spent of total expenditure, but a much larger 83.42% remain unexplained.

**Model 2**

fraction\_babysitting<- BBYDAYPQ/FINCATAX  
fraction\_babysitting1<- is.finite(fraction\_babysitting)  
  
model2 <- lm(fraction\_babysitting[fraction\_babysitting1] ~ AGE\_REF[fraction\_babysitting1])  
summary(model2)

##   
## Call:  
## lm(formula = fraction\_babysitting[fraction\_babysitting1] ~ AGE\_REF[fraction\_babysitting1])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.125 -0.032 -0.022 -0.012 101.373   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.0583520 0.0502176 1.162 0.245  
## AGE\_REF[fraction\_babysitting1] -0.0007181 0.0009407 -0.763 0.445  
##   
## Residual standard error: 1.307 on 6158 degrees of freedom  
## Multiple R-squared: 9.462e-05, Adjusted R-squared: -6.776e-05   
## F-statistic: 0.5827 on 1 and 6158 DF, p-value: 0.4453

anova(model2)

## Analysis of Variance Table  
##   
## Response: fraction\_babysitting[fraction\_babysitting1]  
## Df Sum Sq Mean Sq F value Pr(>F)  
## AGE\_REF[fraction\_babysitting1] 1 1 0.99502 0.5827 0.4453  
## Residuals 6158 10515 1.70758

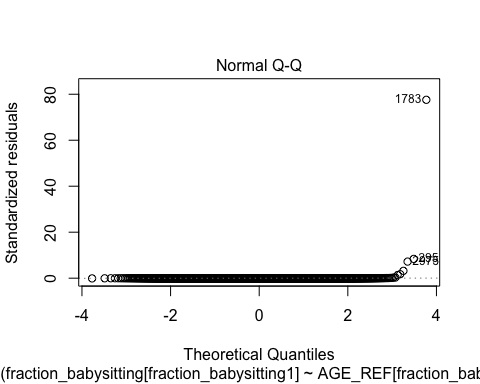
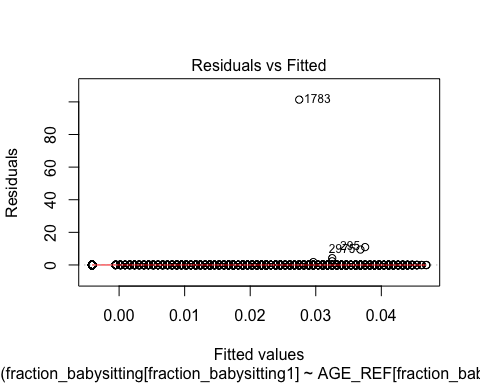
coefficients(model2)

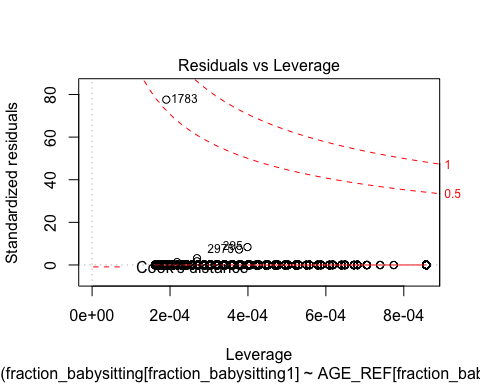
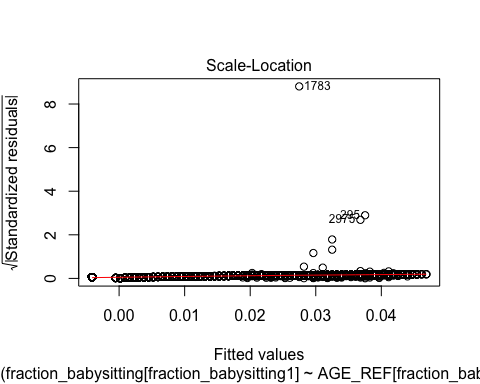
## (Intercept) AGE\_REF[fraction\_babysitting1]   
## 0.0583519688 -0.0007180643

vcov(model2)

## (Intercept)  
## (Intercept) 2.521807e-03  
## AGE\_REF[fraction\_babysitting1] -4.456634e-05  
## AGE\_REF[fraction\_babysitting1]  
## (Intercept) -4.456634e-05  
## AGE\_REF[fraction\_babysitting1] 8.848599e-07

plot(model2)





**Explanation(model2):**

we failed to reject the null hypothesis, because the P value of intercept & Age are 0.245 & 0.445, and they are greater than 0.05.

The R^2 is 9.462e-05, it indicates that they have very weak linear relationship between them.

**Importation of Coefficient of determination:**

* If R^2 is closer to 1, it indicated stronger linear relationship among the variables.
* If R^2 is zero, it indicated that there is no linear relationship among the variables.
* Measure proportion of the variance in the respond variable that is predictable from the explanatory variables.

**Limitation of Coefficient of determination:**

* The coefficient of determination cannot explain whether the linear relationship is positive or negative.

3)

* 1. Fraction of reference person having a college degree spend more than 20% on health insurance is 0.020968357. 2.09%
  2. Fraction of reference person having less than college degree spend more than 20% is 0.041734861. 4.14%
  3. By preforming the hypothesis test for the 2 population statistical test, the difference is 2.05%, the standard error is 0.429 and the t statistic is 4.773, and the P-value is <.0001 for 2 tail test. We reject the null hypothesis at 5% level of significance, Since the P-value is less than 0.05. Therefore, the different in proportion is statistically significant.
  4. The overall share people with any degree is 41%. 26% of people spending more than 20% is made up with any college degree.

|  |
| --- |
| Min. 1st Qu. Median Mean 3rd Qu. Max. |
| -0.17430 0.00000 0.02373 0.04640 0.06937 0.78730 |

The average fraction of spending going to insurance is 4.6%. I think the +/- 20% is not reasonable for the break point, because most people spend approximately 5% of spending on insurance. I think +/- 10% should be better to measure high or low medical spending. Therefore, +10% consider as high medical spending and -10% consider as low medical spending.

f. In the data\_cex, we have sample size of 6838, but in the sample subset you provided (6289) with missing those individual did not state their educational background (548). Those 548 people who did not state their educational background could have spent money on insurances. Therefore, we should use create another column for those 548 people who did not state their educational background.