GEOG210B Assignment2: Matrix Linear Regression with R and Diagnostics

Lily Cheng

Part1: Using the following matrices estimate a linear regression model with Y the dependent variable and X the independent.

Create Matrix using the matrix function

1. Create X matrix

$$X$$

```
X <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,
              2,3,1,5,9,11,11,11,12,12,12,
              43,42,43,54,61,35,52,86,45,44,34,
              1,1,1,1,0,0,0,0,0,0,0),11,4)
X
```

```
##       [,1] [,2] [,3] [,4]
##  [1,]    1    2   43    1
##  [2,]    1    3   42    1
##  [3,]    1    1   43    1
##  [4,]    1    5   54    1
##  [5,]    1    9   61    0
##  [6,]    1   11   35    0
##  [7,]    1   11   52    0
##  [8,]    1   11   86    0
##  [9,]    1   12   45    0
## [10,]    1   12   44    0
## [11,]    1   12   34    0
```

2. Transpose X

$$X'$$

```
XT <- t(X)
XT
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## [1,]    1    1    1    1    1    1    1    1    1     1     1
## [2,]    2    3    1    5    9   11   11   11   12    12    12
## [3,]   43   42   43   54   61   35   52   86   45    44    34
## [4,]    1    1    1    1    0    0    0    0    0     0     0
```

3. X transpose X

$$X'X$$

```
XTX <- XT %*% X
XTX
```

```
##      [,1] [,2]  [,3] [,4]
## [1,]   11   89   539    4
## [2,]   89  915  4453   11
## [3,]  539 4453 28541  182
## [4,]    4   11   182    4
```

4. Inverse of (XT*X)

$$(X'X)^{-1}$$

```
XTXI <- solve(XTX)
XTXI
```

```
##              [,1]          [,2]          [,3]         [,4]
## [1,] 10.47963261 -0.7701768238 -0.0344079440 -6.796084892
## [2,] -0.77017682  0.0654004646  0.0008123292  0.553364567
## [3,] -0.03440794  0.0008123292  0.0004971819  0.009552263
## [4,] -6.79608489  0.5533645674  0.0095522633  5.089704353
```

```
det(XTX)
```

```
## [1] 878954
```

X is a square matrix.

5. Check if the inverse satisfies the inverse properties

$$(X'X)^{-1}(X'X) = I$$

```
Check <- XTXI %*% XTX
Check
```

```
##               [,1]          [,2]          [,3]         [,4]
## [1,] 1.000000e+00  8.526513e-14 -4.547474e-13 0.000000e+00
## [2,] 0.000000e+00  1.000000e+00  0.000000e+00 0.000000e+00
## [3,] 5.551115e-17 -4.440892e-16  1.000000e+00 1.387779e-17
## [4,] 0.000000e+00 -2.415845e-13 -2.273737e-13 1.000000e+00
```

The inverse satisfies the inverse properties.

6. Create Y matrix

$$Y$$

```
Y <- matrix(c(4,7,3,9,17,27,13,121,10,11,23),11,1)
Y
```

```
##       [,1]
##  [1,]    4
##  [2,]    7
##  [3,]    3
##  [4,]    9
##  [5,]   17
##  [6,]   27
##  [7,]   13
##  [8,]  121
##  [9,]   10
## [10,]   11
## [11,]   23
```

X transpose Y

$$X'Y$$

```
XTY <- t(X) %*% Y
XTY
```

```
##       [,1]
## [1,]   245
## [2,]  2529
```

```
## [3,] 15861
## [4,]    23
```

Here are the coefficient estimates

$$(X'X)^{-1}(X'Y) = \hat{\beta}$$

```
alphabeta = XTXI %*% XTY
alphabeta
```

```
##             [,1]
## [1,] -82.321550
## [2,]   2.316192
## [3,]   1.729938
## [4,]   2.989840
```

Part2: Using the model in Part 3 of your assignment 1, report every statistical test you learned in Geog 210B and explain what it means and what your findings are. For example, which variables are significantly different than zero? Are the error terms more likely to be heretoskedastic? Are the error terms autocorrelated?

Here is the model I used in Part3 of my assignment1

```
SmallHHfile <- read.csv("~/Desktop/Winter_2018/210B/Week1_Basic_Concepts/SmallHHfile.csv")
#
# inspect the data we imported
#
View(SmallHHfile)
#
# display the data.frame
str(SmallHHfile)
```

```
## 'data.frame':    42431 obs. of  31 variables:
##  $ X...SAMPN: int  1031985 1032036 1032053 1032425 1032558 1033586 1033660 1033944 1034462 1034878 .
##  $ INCOM    : int  3 7 2 7 1 3 2 6 1 3 ...
##  $ HHSIZ    : int  2 5 6 2 1 3 1 1 2 1 ...
##  $ HHEMP    : int  0 1 1 2 0 1 0 1 0 0 ...
##  $ HHSTU    : int  0 3 3 1 0 0 0 0 0 0 ...
##  $ HHLIC    : int  2 2 1 2 1 3 1 0 0 1 ...
##  $ DOW      : int  2 6 4 1 5 5 1 2 2 5 ...
##  $ HTRIPS   : int  4 31 46 0 6 10 0 15 0 5 ...
##  $ Mon      : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ Tue      : int  1 0 0 0 0 0 0 1 1 0 ...
##  $ Wed      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Thu      : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ Fri      : int  0 0 0 0 1 1 0 0 0 1 ...
##  $ Sat      : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Sun      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ TotDist  : num  36.28 164.9 42.44 0 2.98 ...
##  $ center   : int  0 0 0 0 0 0 0 1 1 1 ...
##  $ suburb   : int  0 1 0 0 1 0 0 0 0 0 ...
##  $ exurb    : int  1 0 0 1 0 1 1 0 0 0 ...
##  $ rural    : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ other    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ highinc  : int  0 1 0 1 0 0 0 1 0 0 ...
##  $ HHVEH    : int  2 1 2 2 0 2 1 0 0 1 ...
##  $ HHBIC    : int  2 4 2 3 0 1 1 1 0 2 ...
##  $ VEHNEW   : int  1 1 2 2 2 2 2 2 2 2 ...
```

```
##  $ OWN      : int  1 1 2 1 2 2 1 1 2 2 ...
##  $ CarBuy   : int  1 1 0 0 0 0 0 0 0 0 ...
##  $ snglhm   : int  1 1 1 1 1 1 1 1 0 0 ...
##  $ ownhm    : int  1 1 0 1 0 0 1 1 0 0 ...
##  $ MilesPr  : num  18.14 32.98 7.07 0 2.98 ...
##  $ TrpPrs   : num  2 6.2 7.67 0 6 ...
```

```
Model2= lm(TrpPrs ~ Mon + Tue + Wed + Thu + Fri + Sat + HHVEH +
    HHSIZ + suburb + exurb + rural + HTRIPS + rural * HTRIPS, data=SmallHHfile)
summary(Model2)
```

```
##
## Call:
## lm(formula = TrpPrs ~ Mon + Tue + Wed + Thu + Fri + Sat + HHVEH +
##     HHSIZ + suburb + exurb + rural + HTRIPS + rural * HTRIPS,
##     data = SmallHHfile)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2774  -0.6571  -0.0674   0.4306  18.8316
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   3.422758   0.023477  145.789  < 2e-16 ***
## Mon           0.097503   0.023698    4.114 3.89e-05 ***
## Tue           0.127968   0.023464    5.454 4.96e-08 ***
## Wed           0.114077   0.023433    4.868 1.13e-06 ***
## Thu           0.135934   0.023325    5.828 5.66e-09 ***
## Fri           0.114288   0.023625    4.838 1.32e-06 ***
## Sat           0.097590   0.023485    4.155 3.25e-05 ***
## HHVEH        -0.016650   0.006970   -2.389   0.0169 *
## HHSIZ        -1.102960   0.005853 -188.441  < 2e-16 ***
## suburb       -0.153380   0.016822   -9.118  < 2e-16 ***
## exurb        -0.192149   0.017937  -10.712  < 2e-16 ***
## rural        -0.411024   0.024354  -16.877  < 2e-16 ***
## HTRIPS        0.334771   0.001046  319.951  < 2e-16 ***
## rural:HTRIPS  0.015596   0.002143    7.279 3.42e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.296 on 42417 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7474
## F-statistic:  9657 on 13 and 42417 DF,  p-value: < 2.2e-16
```

$\hat{TrpPrs} = 145.789 + 4.114Mon + 5.454Tue + 4.868Wed + 5.828Thu + 4.838Fri + 4.155Sat - 2.389HHVEH - 188.441HHSIZ - 9.$

1. Check if the mean of the residual is zero.

```
Model2.res = resid(Model2)
summary(Model2.res)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -15.27738  -0.65715  -0.06741   0.00000   0.43063  18.83159
```

The mean of the residual is zero.

2. Look at the ANOVA table of this regression model

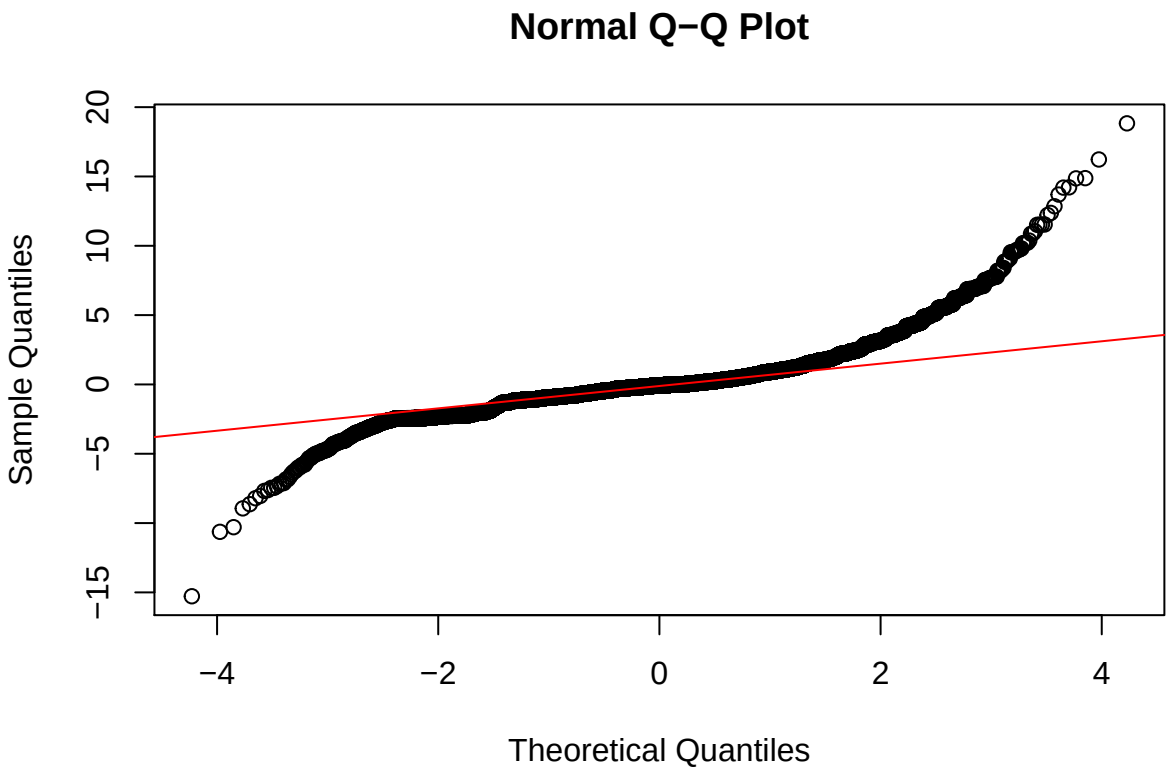
```
anova(Model2)
```

```
## Analysis of Variance Table
##
## Response: TrpPrs
##                  Df Sum Sq Mean Sq    F value    Pr(>F)
## Mon               1    135     135     80.504 < 2.2e-16 ***
## Tue               1    737     737    438.755 < 2.2e-16 ***
## Wed               1    799     799    475.596 < 2.2e-16 ***
## Thu               1   1497    1497    890.903 < 2.2e-16 ***
## Fri               1   1764    1764   1049.626 < 2.2e-16 ***
## Sat               1    439     439    261.530 < 2.2e-16 ***
## HHVEH             1    859     859    511.182 < 2.2e-16 ***
## HHSIZ             1    107     107     63.718 1.471e-15 ***
## suburb            1     60      60     35.815 2.187e-09 ***
## exurb             1    121     121     72.292 < 2.2e-16 ***
## rural             1   4672    4672   2780.307 < 2.2e-16 ***
## HTRIPS            1 199686  199686 118827.754 < 2.2e-16 ***
## rural:HTRIPS      1     89      89     52.986 3.417e-13 ***
## Residuals     42417  71280       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By reading the sum of squares of ANOVA table, I know that trips a household made (HTRIPS) contributes most to the number of trips a person made(TrpPrs). The number of household living in a rural area (rural) contributes the second to the number of trips a person made. Shopping on diary on Friday (Fri) and Thursday (Thu) contribute at the third level to the number of trips a person made. Then number of cars a household owns (HHVEH) contribute the fourth to the number of trips a person made. Shopping on Tuesday (Tue) or Wednesday (Wed) contribute at the fifth level to the number of trips a person made. Next level of contributions come from shopping on Saturday (Sat), and then the contribution orders goes by shopping on Monday (Mon), how many household lives in exurb (exurb), and the size of household (HHSIZ). The number of household trips made by household that lives in rural areas (rural:HTRIPS) and number of household lives in suburb areas (suburb) contribute least out of the model to the number of trips a person made.

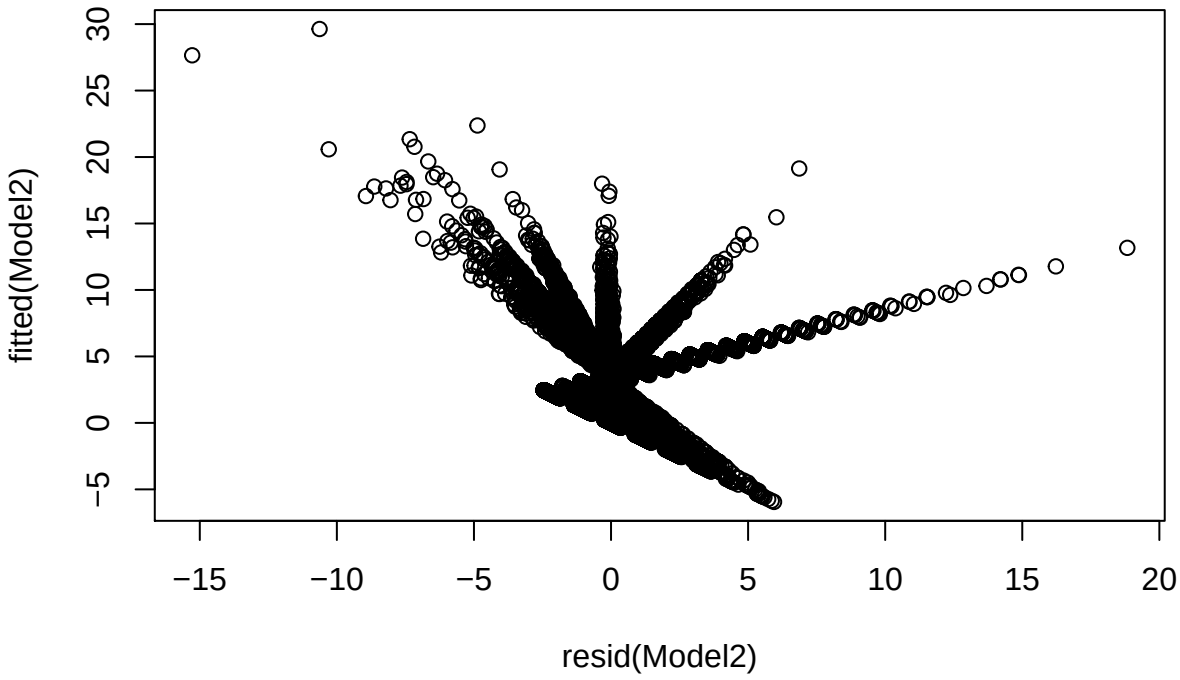
3. In order to whether the distribution of residual fits the theoretical distribution, I plot Q-Q Plot.

```
qqnorm(Model2.res)
qqline(Model2.res,col="red")
```



From the plot above, we can see that the quantile data of residual is not aligned with the predicted quantile distribution. Thus, there is variability in residual distribution and then I plot the residual with fitted y to look more into the variability.

```
plot(resid(Model2),fitted(Model2))
```



From the above plot, I can see that residual variability decreases with fitted y when fitted y is between -5 and around 2.5. However, the residual variability increases with fitted y as fitted y goes over around 2.5.

Put side by side a model with White's correction for the standard errors of the coefficient

estimates and without and decide if the correction changes your findings about the significance of coefficients.

4.

```
library(sandwich)
library(lmtest)
```

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

```
## Warning: package 'zoo' was built under R version 3.4.3
```

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

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

```
coeftest(Model2,vcov = vcovHC(Model2, type = "const"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)   3.4227581  0.0234774  145.7893 < 2.2e-16 ***
## Mon           0.0975031  0.0236978    4.1144 3.889e-05 ***
## Tue           0.1279681  0.0234639    5.4538 4.957e-08 ***
## Wed           0.1140773  0.0234331    4.8682 1.130e-06 ***
## Thu           0.1359343  0.0233253    5.8278 5.658e-09 ***
## Fri           0.1142883  0.0236248    4.8376 1.318e-06 ***
## Sat           0.0975898  0.0234847    4.1555 3.253e-05 ***
## HHVEH        -0.0166499  0.0069696   -2.3889    0.0169 *
## HHSIZ        -1.1029596  0.0058531 -188.4406 < 2.2e-16 ***
## suburb       -0.1533800  0.0168222   -9.1177 < 2.2e-16 ***
## exurb        -0.1921486  0.0179373  -10.7122 < 2.2e-16 ***
## rural        -0.4110237  0.0243538  -16.8772 < 2.2e-16 ***
## HTRIPS        0.3347711  0.0010463  319.9511 < 2.2e-16 ***
## rural:HTRIPS  0.0155956  0.0021425    7.2792 3.417e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(Model2,vcov = vcovHC(Model2, type = "HC0")) # white's adjustment
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)   3.4227581  0.0290795  117.7034 < 2.2e-16 ***
## Mon           0.0975031  0.0220802    4.4159 1.009e-05 ***
## Tue           0.1279681  0.0225847    5.6662 1.470e-08 ***
## Wed           0.1140773  0.0224531    5.0807 3.776e-07 ***
## Thu           0.1359343  0.0226691    5.9964 2.033e-09 ***
## Fri           0.1142883  0.0229096    4.9887 6.104e-07 ***
## Sat           0.0975898  0.0228753    4.2662 1.993e-05 ***
## HHVEH        -0.0166499  0.0074449   -2.2364 0.0253301 *
## HHSIZ        -1.1029596  0.0093931 -117.4220 < 2.2e-16 ***
## suburb       -0.1533800  0.0177408   -8.6456 < 2.2e-16 ***
```

```
## exurb        -0.1921486  0.0180905  -10.6215 < 2.2e-16 ***
## rural        -0.4110237  0.0330015  -12.4547 < 2.2e-16 ***
## HTRIPS        0.3347711  0.0022053  151.8050 < 2.2e-16 ***
## rural:HTRIPS  0.0155956  0.0044772    3.4834 0.0004956 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By comparing the model with and without White's correction for the standard error of the coefficient estimates and without, the correction doesn't change the significance of coefficients.

5. The Breusch-Pagan test for heteroskedasticity

```
bptest(Model2, studentize = TRUE)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Model2
## BP = 2922.7, df = 13, p-value < 2.2e-16
```

With p value of BP test being significant, the Model2 is heteroskedastic.

6. Calculate autocorrelation.

```
dwtest(Model2)
```

```
##
##  Durbin-Watson test
##
## data:  Model2
## DW = 1.9853, p-value = 0.06481
## alternative hypothesis: true autocorrelation is greater than 0
```

As the value of DW is between 0 and 2, there is positive autocorrelation of the data.

7. Using stargazer to make a nicer formatting table with lm objects

```
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 3.4.3
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```
Model2 = lm(MilesPr ~   Mon +   Tue +   Wed +   Thu +   Fri+    Sat +   HHVEH   +   HHSIZ   +   suburb
            +   exurb+  rural + HTRIPS + rural*HTRIPS, data=SmallHHfile)
stargazer(Model2, type="text", title="Regression Results",
          dep.var.labels=c("Number of Miles per Person"),
          covariate.labels=c( "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday",
                              "Number of vehicles a household owns", "Number of people a
                              household has","Residence in Suburb Env", "Residence in Exurb Env",
                              "Residence in Rural Env","the Number of Trips a household made",
                              "the interaction between rural and the number of trips a household made"
                              ), out="Februray, 20.txt")
```

```
##
## Regression Results
```

	Dependent variable:
	Number of Miles per Person
Monday	-2.872***
	(0.763)
Tuesday	-3.332***
	(0.756)
Wednesday	-2.881***
	(0.755)
Thursday	-3.117***
	(0.751)
Friday	0.205
	(0.761)
Saturday	2.233***
	(0.756)
Number of vehicles a household owns	5.124***
	(0.224)
household has	-7.384***
	(0.189)
Residence in Suburb Env	4.173***
	(0.542)
Residence in Exurb Env	7.927***
	(0.578)
Residence in Rural Env	5.337***
	(0.784)
the Number of Trips a household made	1.490***
	(0.034)
the interaction between rural and the number of trips a household made	0.661***
	(0.069)
Constant	20.595***
	(0.756)
Observations	42,431
R2	0.077
Adjusted R2	0.077
Residual Std. Error	41.756 (df = 42417)
F Statistic	272.747*** (df = 13; 42417)

```
## Note:                                                    *p<0.1; **p<0.05; ***p<0.01
```