GEOG210B Assignment1:Linear Regression with R

Lily Cheng

```r
library(readr)

SmallHHfile <- read.csv("~/Desktop/Winter_2018/210B/Week1_Basic_Concepts/Smal
lHHfile.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 ...
```

Part1: Report a table of descriptive statistics using package *psych* of the variables in the dataset called SmallHHfile.

```
library(psych)
describe(SmallHHfile)
```

```
##           vars     n       mean         sd     median    trimmed       mad
## X...SAMPN    1 42431 2588378.63 1641345.14 1971814.00 2195483.36 847148.74
## INCOM        2 42431      13.18      26.29       5.00       5.51      2.97
## HHSIZ        3 42431       2.57       1.37       2.00       2.40      1.48
## HHEMP        4 42431       1.22       0.88       1.00       1.18      1.48
## HHSTU        5 42431       0.64       1.02       0.00       0.44      0.00
## HHLIC        6 42431       1.86       0.85       2.00       1.81      0.00
## DOW          7 42431       4.02       1.99       4.00       4.02      2.97
## HTRIPS       8 42431       8.29       7.78       6.00       7.14      5.93
## Mon          9 42431       0.14       0.34       0.00       0.05      0.00
## Tue         10 42431       0.14       0.35       0.00       0.06      0.00
## Wed         11 42431       0.14       0.35       0.00       0.06      0.00
## Thu         12 42431       0.15       0.35       0.00       0.06      0.00
## Fri         13 42431       0.14       0.35       0.00       0.05      0.00
## Sat         14 42431       0.14       0.35       0.00       0.05      0.00
## Sun         15 42431       0.15       0.35       0.00       0.06      0.00
## TotDist     16 42431      68.09     118.52      33.89      45.44     45.13
## center      17 42431       0.28       0.45       0.00       0.23      0.00
## suburb      18 42431       0.29       0.45       0.00       0.23      0.00
## exurb       19 42431       0.23       0.42       0.00       0.16      0.00
## rural       20 42431       0.20       0.40       0.00       0.13      0.00
## other       21 42431       0.00       0.00       0.00       0.00      0.00
## highinc     22 42431       0.41       0.49       0.00       0.39      0.00
## HHVEH       23 42431       1.86       1.00       2.00       1.81      1.48
## HHBIC       24 42431       1.58       3.79       1.00       1.20      1.48
## VEHNEW      25 42431       2.15       2.02       2.00       1.57      1.48
## OWN         26 42431       1.24       0.56       1.00       1.16      0.00
## CarBuy      27 42431       0.45       0.50       0.00       0.44      0.00
## snglhm      28 42431       0.82       0.39       1.00       0.90      0.00
## ownhm       29 42431       0.77       0.42       1.00       0.84      0.00
## MilesPr     30 42431      27.12      43.46      14.50      18.40     18.19
## TrpPrs      31 42431       3.28       2.58       3.00       3.02      2.22
##               min        max      range skew kurtosis      se
## X...SAMPN 1031985 7212388.00 6180403.00 2.04     3.09 7968.16
## INCOM           1      99.00      98.00 2.92     6.62    0.13
## HHSIZ           1       8.00       7.00 1.03     0.90    0.01
## HHEMP           0       6.00       6.00 0.47     0.33    0.00
## HHSTU           0       8.00       8.00 1.66     2.52    0.00
## HHLIC           0       8.00       8.00 0.60     1.70    0.00
## DOW             1       7.00       6.00 0.00    -1.24    0.01
## HTRIPS          0      99.00      99.00 1.72     4.88    0.04
## Mon              0       1.00        1.00  2.12      2.49    0.00
## Tue              0       1.00        1.00  2.03      2.10    0.00
## Wed              0       1.00        1.00  2.02      2.08    0.00
## Thu              0       1.00        1.00  1.99      1.96    0.00
## Fri              0       1.00        1.00  2.08      2.33    0.00
## Sat              0       1.00        1.00  2.06      2.26    0.00
## Sun              0       1.00        1.00  1.99      1.98    0.00
## TotDist          0    5838.26     5838.26  8.38    196.69    0.58
## center           0       1.00        1.00  0.97     -1.05    0.00
## suburb           0       1.00        1.00  0.94     -1.12    0.00
## exurb            0       1.00        1.00  1.29     -0.34    0.00
## rural            0       1.00        1.00  1.49      0.21    0.00
## other            0       0.00        0.00   NaN       NaN    0.00
## highinc          0       1.00        1.00  0.35     -1.88    0.00
## HHVEH            0       8.00        8.00  0.80      2.26    0.00
## HHBIC            0      99.00       99.00 20.40    513.75    0.02
## VEHNEW           1       9.00        8.00  2.38      4.20    0.01
## OWN              1       9.00        8.00  5.96     67.49    0.00
## CarBuy           0       1.00        1.00  0.19     -1.96    0.00
## snglhm           0       1.00        1.00 -1.65      0.71    0.00
## ownhm            0       1.00        1.00 -1.31     -0.29    0.00
## MilesPr          0    1167.65     1167.65  5.15     47.24    0.21
## TrpPrs           0      32.00       32.00  1.27      3.68    0.01
```

Part2: Estimate the following model (called Model 1 herein): Dependent variable (y): MilesPr Independent variables (x): Mon + Tue + Wed + Thu + Fri+ Sat + HHVEH + HHSIZ + suburb + exurb+ rural

2.1 Report in a table the regression coefficients, their standard errors, t-stats, and R-square (it is ok to just use the standard reporting of R for object lm). Note: I'm supposed to be the only in the class to add interaction in the model and I added the interaction of household lives in rural environment (variable: rural) and daily number of household trips (variable: TRIPS).

Model1

```
Model1= lm(MilesPr ~    Mon +   Tue +   Wed +   Thu +   Fri+    Sat +   HHVEH
+   HHSIZ   +   suburb  +   exurb+  rural + rural*HTRIPS, data=SmallHHfile)
summary(Model1)

## 
## Call:
## lm(formula = MilesPr ~ Mon + Tue + Wed + Thu + Fri + Sat + HHVEH + 
##     HHSIZ + suburb + exurb + rural + rural * HTRIPS, data = SmallHHfile)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -107.09  -18.99  -10.47    3.77 1156.78 
## 
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.59550    0.75623  27.234  < 2e-16 ***
## Mon          -2.87185    0.76333  -3.762 0.000169 ***
## Tue          -3.33177    0.75580  -4.408 1.04e-05 ***
## Wed          -2.88092    0.75481  -3.817 0.000135 ***
## Thu          -3.11690    0.75133  -4.148 3.35e-05 ***
## Fri           0.20520    0.76098   0.270 0.787426
## Sat           2.23267    0.75647   2.951 0.003165 **
## HHVEH         5.12377    0.22450  22.823  < 2e-16 ***
## HHSIZ        -7.38386    0.18853 -39.164  < 2e-16 ***
## suburb        4.17253    0.54186   7.700 1.39e-14 ***
## exurb         7.92725    0.57778  13.720  < 2e-16 ***
## rural         5.33671    0.78446   6.803 1.04e-11 ***
## HTRIPS        1.49025    0.03370  44.217  < 2e-16 ***
## rural:HTRIPS  0.66116    0.06901   9.580  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.76 on 42417 degrees of freedom
## Multiple R-squared:  0.07714,    Adjusted R-squared:  0.07686
## F-statistic: 272.7 on 13 and 42417 DF,  p-value: < 2.2e-16
```

2.1 Write the equation that corresponds to this model.

$$\widehat{MilesPr} = 27.234 - 3.762Mon - 4.408Tue - 3.817Wed - 4.148Thu + 2.951Sat + 22.823HHVEH - 39.164HHSIZ + 7.7suburb + 13.720exurb + 6.803rural + 44.217HTRIPS + 9.580rural * HTRIPS$$

2.3 Write a short summary of the model in a similar fashion as our discussion in class highlighting which coefficients are significantly different than zero and what they tell us.

From model1, when all the variables are 0, a person's travel distance will be 27.234miles (p<0.001). If all the variables are not 0, then a person's travel distance on dairy is significantly related to which day of the week the person chooses to travel, the number of cars the household owns, the number of persons in household, the region the household lives, and daily number of household trips, as well as the interaction between daily number of household trips and the household lives in rural areas. For each more Monday a person travels, the distance will decrease by 3.762miles (p<0.001). For each more Tuesday a person travels, the distance will decrease by 4.408miles (p<0.001). For each more Wednesday a person travels, the distance will decrease by 3.817miles (p<0.001). However, for each more Saturday a person travels, the distance will increase by 2.951miles (p<0.05). For each more car the household owns, the distance will increase by 22.823miles (p<0.001). For each more person lives in household, the distance will decrease by 39.164miles (p<0.001). For each more household lives in suburb area, the distance will increase by 7.7miles (p<0.001). For each more household lives in exurb area, the distance will increase by 13.72miles (p<0.001). For each more household lives in rural area, the distance will increase by 6.803miles (p<0.001). For each more daily trips I household makes, the distance will increase by 44.217miles (p<0.001). There is also significant difference between household trips and rural areas: for each trip that a household make, if the household is from rural area, the distance will increase by 9.58miles (p<0.001).

Part3: Estimate a model using just one of the following as the dependent variable (called Model 2 herein). Possible y: TrpPrs (this is the number of trip per person) or HTRIPS (this is the number of trips for each household).3.1 Report in a table the regression coefficients, their standard errors, t-stats, and R-square (it is ok to just use the standard reporting of R for the object lm.

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

```
##
## Call:
## lm(formula = TrpPrs ~ Mon + Tue + Wed + Thu + Fri + Sat + HHVEH +
##     HHSIZ + suburb + exurb + rural + 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
```

Model2

$$\widehat{TrpPrs} = 145.789 + 4.114Mon + 5.454Tue + 4.868Wed + 5.828Thu + 4.838Fri + 4.155Sat - 2.389HHVEH - 188.441HHSIZ - 9.118suburb - 10.712exurb - 16.877rural + 319.951HTRIPS + 7.279rural * HTRIPS$$

3.2 Write a comparison summary between Model 1 and Model 2.

Comparing Model1 and Model2, we could find that individual trip numbers increase with the number of days (either weekday or weekend, Model2) people spend to complete their dairy while how far a person goes on dairy decreases with more weekdays but increases with more weekends (Model1). This means that the more often people go out on diary, more trips they are going to make(Model2). Those trips on dairy are shorter if they go out on weekdays and longer if they go out on weekend (Model1). A household with more cars will tend to make less trips (Model2) but longer distance for each trip (Model1). A household with more people will tend to make shorter (Model1) and much less trips (Model2) on dairy. Among households live in 3 different regions (suburb, exurb, rural), they all tend to make less trips (Model2) on dairy with longer distance (Model1) per trip. More specifically, household lives in rural area will tend to make least trips (Model2) and household lives in exurb area will tend to make longest distance per trip (Model1). Each person's trip amount is closely related to the household trip amount (Model2) and trip distance (Model1). For household lives in rural areas, a person's trip amount (Model2) and distance (Model1) on dairy is will both increase with household trip amount.