Assignment 2: Simple linear regression for modeling household trip generation
Submitted by : 0424042406 , Arnob Protim Roy
R Code _ Github Link : https://github.com/trewto/CE-6511-Assignments/blob/main/Assignment_2/assignment2code.R

Ans No 1

HHFAMINC : Household income impacts trip generation because higher-income families may own more vehicles, which is a main factor of trip generation.
HHSIZE: Family Size, Larger family size generate more trip.

Ans No 2

Here is the info of developed two model by adding HHFAMINC and HHSIZE

Model 1

```
summary(model1)
lm(formula = CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHFAMINC, data = hh_ga)

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  3.08206    0.14241  21.642 < 2e-16 ***
WRKCOUNT     1.61140    0.07331  21.980 < 2e-16 ***
HHVEHCNT     0.62181    0.05525  11.255 < 2e-16 ***
HHFAMINC     0.19607    0.02474   7.924 2.6e-15 ***

Residual standard error: 5.137 on 8321 degrees of freedom
Multiple R-squared:  0.1421,     Adjusted R-squared:  0.1418
F-statistic: 459.4 on 3 and 8321 DF,  p-value: < 2.2e-16
```

Model 2

```
summary(model2)
lm(formula = CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHFAMINC + HHFAMINC + HHSIZE, data = hh_ga)

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.54266    0.14327   3.788 0.000153 ***
WRKCOUNT     0.42206    0.07259   5.814 6.31e-09 ***
HHVEHCNT     0.16388    0.05147   3.184 0.001458 **
HHFAMINC     0.19835    0.02252   8.808 < 2e-16 ***
HHSIZE       2.12586    0.05114  41.569 < 2e-16 ***

Residual standard error: 4.675 on 8320 degrees of freedom
Multiple R-squared:  0.2896,     Adjusted R-squared:  0.2893
F-statistic: 848.1 on 4 and 8320 DF,  p-value: < 2.2e-16
```

Ans No 3

From Model 1 ,
lm(formula = CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHFAMINC, data = hh_ga)

Coefficients:
Intercept: 3.08206 , **WRKCOUNT**: 1.61140 , **HHVEHCNT**: 0.62181 , **HHFAMINC**: 0.19607

All signs are positive, so positive correlation.

- WRKCOUNT (1.61140) suggests that households with more workers generate more trips
- HHVEHCNT (0.62181) suggests higher vehicle availability leads to more trips
- HHFAMINC (0.19607) suggests Higher income likely increases the ability and tendency to travel.

For Model 2

```
lm(formula = CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHFAMINC + HHFAMINC + HHSIZE, data =
hh_ga)
```

Coefficients:
Intercept: 0.54266 ,**WRKCOUNT**: 0.42206, **HHVEHCNT**: 0.16388 , **HHFAMINC**: 0.19835, **HHSIZE**: 2.12586

- WRKCOUNT (0.42206) remains positive, though smaller than in Model 1. This still aligns with the expectation that more workers contribute to higher trip generation.
- HHVEHCNT: The positive estimate (0.16388) remains consistent with Model 1, supporting the notion that households with more vehicles are likely to make more trips.
- HHFAMINC: The estimate for HHFAMINC (0.19835) is again positive and similar to Model 1, showing that higher income positively correlates with trip generation.
- Family size (HHSIZE), has a large positive estimate (2.12586). This matches our theoretical expectation, as larger households would have more people needing to travel for various reasons.

The signs of the estimates in both models support our theoretical basis. WRKCOUNT, HHVEHCNT, HHFAMINC and HHSIZE are all positively corelated with household trip generation.

Ans No 4

If a parameter's p-value is below a standard significance level (usually 0.05), we can conclude that it is statistically significant.

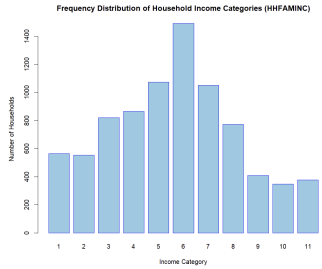
Model 1 Analysis

From (Ans 2) Data, It is seen The t-statistics and p-values for **Model 1** are as follows:
All t-statistics are large, and all p-values are far below 0.05, indicating that each coefficient is statistically significantly different from zero.
WRKCOUNT, HHVEHCNT, and HHFAMINC all significantly contribute to explaining household trip generation in Model 1

Model 2 Analysis

The t-statistics and p-values for Model 2 are as follows-
All p-values are below 0.05, showing that each coefficient is statistically significantly different from zero.
WRKCOUNT, HHVEHCNT, HHFAMINC, and HHSIZE are all significant predictors of household trip generation in Model 2.

Ans No 5



Every group has more than 30 or 50 number. So each group is applicable.
Here is the dummy variable equation

```
CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHSIZE + HHFAMINC_1 + HHFAMINC_2 +
    HHFAMINC_3 + HHFAMINC_4 + HHFAMINC_5 + HHFAMINC_6 + HHFAMINC_7 +
    HHFAMINC_8 + HHFAMINC_9 + HHFAMINC_10 + HHFAMINC_11
After Modeling:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.73730    0.28738   9.525  < 2e-16 ***
WRKCOUNT     0.42458    0.07270   5.840 5.41e-09 ***
HHVEHCNT     0.16510    0.05187   3.183 0.001464 **
HHSIZE       2.12766    0.05128  41.494  < 2e-16 ***
HHFAMINC_1  -1.97784    0.32611  -6.065 1.38e-09 ***
HHFAMINC_2  -1.53642    0.32435  -4.737 2.21e-06 ***
HHFAMINC_3  -1.78119    0.30040  -5.929 3.16e-09 ***
HHFAMINC_4  -1.54876    0.29381  -5.271 1.39e-07 ***
HHFAMINC_5  -1.10593    0.28357  -3.900 9.69e-05 ***
HHFAMINC_6  -0.96571    0.27194  -3.551 0.000386 ***
HHFAMINC_7  -0.83258    0.28145  -2.958 0.003104 **
HHFAMINC_8  -0.91071    0.29393  -3.098 0.001952 **
HHFAMINC_9  -0.21164    0.33376  -0.634 0.526027
HHFAMINC_10 -0.03181    0.34781  -0.091 0.927125
HHFAMINC_11       NA         NA      NA       NA
---
Residual standard error: 4.675 on 8311 degrees of freedom
Multiple R-squared:  0.2904, Adjusted R-squared:  0.2893
F-statistic: 261.6 on 13 and 8311 DF,  p-value: < 2.2e-16
```

- The negative coefficients for the income categories (HHFAMINC_1 to HHFAMINC_8) indicate that households in these income categories have a lower expected count of CNTTDHH compared to the reference category.
- The magnitude of the coefficients decreases as the income categories increase, indicating that households in lower income categories are expected to have significantly fewer trips compared to those in higher income categories.

Maybe, we can group based on income level ,
1-3 GROUP A , 4-7 GROUP B , 9-11 GROUP C

For group

CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHSIZE + HHFAMINC_A + HHFAMINC_B + HHFAMINC_C

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.75003    0.22151   7.900 3.14e-15 ***
WRKCOUNT     0.44059    0.07253   6.074 1.30e-09 ***
HHVEHCNT     0.18905    0.05129   3.686 0.000229 ***
HHSIZE       2.12639    0.05124  41.497  < 2e-16 ***
HHFAMINC_A  -0.81559    0.21137  -3.859 0.000115 ***
HHFAMINC_B  -0.15601    0.18402  -0.848 0.396576
HHFAMINC_C   0.81754    0.21850   3.742 0.000184 ***
Residual standard error: 4.677 on 8318 degrees of freedom
Multiple R-squared:  0.2891,     Adjusted R-squared:  0.2886
F-statistic: 563.8 on 6 and 8318 DF,  p-value: < 2.2e-16
```

The grouping of household income into three categories is a practical approach to simplify the analysis and interpret the coefficients more easily. Though the R^2 value of group data is slightly less than non group 0.2891<0.2904 but it is memory efficient when the data size is very big.

Observation: Group B's coefficient is not statistically significant. it may be worth merge it with either Group A or Group C

Ans No 6

Comparation of different model Adjusted R^2 Value

Model 1	CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHFAMINC Adjusted R^2 - 0.1418
Model 2	CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHFAMINC + HHSIZE Adjusted R^2 - 0.2893
Model with Dummies	CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHSIZE + HHFAMINC_1 + HHFAMINC_2 + HHFAMINC_3 + HHFAMINC_4 + HHFAMINC_5 + HHFAMINC_6 + HHFAMINC_7 + HHFAMINC_8 + HHFAMINC_9 + HHFAMINC_10 + HHFAMINC_11 Adjusted R^2 - 0.2893
Model with Grouped Dummies	CNTTDHH ~ WRKCOUNT + HHVEHCNT + HHSIZE + HHFAMINC_A + HHFAMINC_B + HHFAMINC_C Adjusted R^2 : 0.2886

The best model is chosen can be chosen as **Model 2** 0.2893 (Model with Dummies has the same adjusted R^2 , but it has a lot of non-efficient data so we can reject it). Model 2 is best based on adjusted R^2. But i will suggest the **model with Grouped Dummies**, because though it is slightly less R^2 value, it is very good memory efficient for operation than model 2 when big data is used.