Advanced Transportation Modeling and Statistics

Homework #2

Hossein Zamani Saghazadeh

29423388

University of Hawaii at Manoa

Department of Civil and Environmental Engineering

Problem 1 - Theory

We have the following:

$$Y = X\beta + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma_\varepsilon^2\right) \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad Y = X\beta + \varepsilon \qquad \hat{Y} = X\hat{\beta} \qquad \hat{\varepsilon} = Y - \hat{Y} \implies Y = \hat{\varepsilon} + \hat{Y}$$

We know that error is distributed normal with a mean of $E[\varepsilon] = 0$ and variance of $E[\varepsilon\varepsilon^T] = \sigma^2 I$.

$$\hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) = (X^T X)^{-1} X^T X \beta + (X^T X)^{-1} X^T \varepsilon = \beta + (X^T X)^{-1} X^T \varepsilon$$

$$\implies E\left[\hat{\beta}\right] = E\left[\beta\right] + E\left[(X^T X)^{-1} X^T\right] E\left[\varepsilon\right]$$

Also we know that $E[\varepsilon] = 0$, and $E\left[\beta\right] = \beta$ since it is non-stochastic, so:

$$E\left[\hat{\beta}\right] = \beta + 0 = \beta$$

So the OLS estimator is unbiased.

For computing the variance we have:

$$Var\left(\hat{\beta}\right) = E\left[\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)^T\right]$$

We know from before that $\hat{\beta} = \beta + (X^T X)^{-1} X^T \varepsilon$ and $E\left[\hat{\beta}\right] = \beta$:

$$\hat{\beta} - E\left[\hat{\beta}\right] = \beta + (X^T X)^{-1} X^T \varepsilon - \beta = (X^T X)^{-1} X^T \varepsilon$$

$$\implies \left(\hat{\beta} - E\left[\hat{\beta}\right]\right)\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)^T = (X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1}$$

So taking the expected value of the above expression we have (notice that X is non-stochastic:

$$E\left[(X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1}\right] = (X^T X)^{-1} X^T X (X^T X)^{-1} E\left[\varepsilon\varepsilon^T\right] = (X^T X)^{-1} \sigma_\varepsilon^2 I$$

$$\implies Var\left(\hat{\beta}\right) = (X^T X)^{-1} \sigma_\varepsilon^2$$

Problem 2 - Theory

We have the following:

$$Y_t = X_t\beta + \varepsilon_t \qquad t = 0, 1, \ldots, 100 \quad \varepsilon_t \sim N\left(0, \sigma_t^2\right) \qquad Cov\left(\varepsilon\right) = 0 \qquad Var\left(\varepsilon\right) = \begin{cases} \sigma_1^2, & \text{if } t \leq 50 \\ \sigma_2^2, & \text{if } x > 50 \end{cases} \qquad R = \frac{\sigma_1^2}{\sigma_2^2}$$

The variance-covariance matrix can be defined as follows:

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_1^2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_1^2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \sigma_2^2 \end{bmatrix}$$

We have $R = \frac{\sigma_1^2}{\sigma_2^2} \implies \sigma_1^2 = R\sigma_2^2$, so the variance-covariance matrix can be written as:

$$\Omega = \sigma_2^2 \begin{bmatrix} R & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & R & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & R & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & R & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Since this case is a Heteroskedasticity case, we can use transformation to make our problem be transformed from GLS to OLS. Then, we solve the OLS, and transform the solution to GLS at the end. Our GLS model is:

$$Y_t = X_t\beta + \varepsilon_t$$

Since Ω is positive definite we have the following:

$$P^T\Omega P = I \implies \Omega = \left(P^T\right)^{-1}\left(P\right)^{-1} \implies \Omega^{-1} = PP^T$$

If we multiply our model with P^T:

$$P^TY_t = P^TX_t\beta + P^T\varepsilon_t$$

Considering the followings:

$$P^TY_t^* = P^TY_t \qquad X_t^* = P^TX_t \qquad \varepsilon_t^* = P^T\varepsilon$$

We have:

$$Y_t^* = X_t^*\beta + \varepsilon_t^*$$

Now we check whether this transformed model is OLS or not:

$$E\left[\varepsilon_t^*\right] = E\left[P^T\varepsilon_t\right] = P^T E\left[\varepsilon_t\right] = P^T.0 = 0 \implies E\left[\varepsilon_t^*\right] = 0$$

$$E\left[\varepsilon_t^*\varepsilon_t^{*T}\right] = E\left[P^T\varepsilon_t\varepsilon_t^T P\right] = P^T E\left[\varepsilon_t\right] P = \sigma_2^2 P^T \Omega P = \sigma_2^T I \implies E\left[\varepsilon_t^*\varepsilon_t^{*T}\right] = \sigma_2^T I$$

So the new transformed model can be solved using OLS. From OLS we have:

$$\hat{\beta} = \left(X_t^* X_t^{*T}\right)^{-1} X_t^{*T} Y_t^*$$

Substituting the start variables with the original ones:

$$\hat{\beta} = \left(X_t^T P P^T X_t\right)^{-1} X_t^T P P^T Y_t$$
$$PP^T = \Omega^{-1} \implies \hat{\beta} = \left(X_t^T \Omega X_t\right)^{-1} X_t^T \Omega^{-1} Y_t$$

All we need to do is to define P so that $P^T\Omega P = I$ is satisfied. We know that if we have the following matrix of Ω:

$$\Omega = \begin{bmatrix} d_1^2 & 0 & 0 & \cdots & 0 \\ 0 & d_2^2 & 0 & \cdots & 0 \\ 0 & 0 & d_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d_n^2 \end{bmatrix}$$

The matrix P can be defined as:

$$P = \begin{bmatrix} \frac{1}{d_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{d_2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{d_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{d_n} \end{bmatrix}$$

We have the following matrix of Ω:

$$\Omega = \sigma_2^2 \begin{bmatrix} R & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & R & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & R & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & R & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}$$

So the matrix P can be defined as:

$$P = \frac{1}{\sigma_2}\begin{bmatrix} \frac{1}{\sqrt{R}} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{R}} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{R}} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{R}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}$$

and the inverse of Ω:

$$\Omega^{-1} = P^T P = \frac{1}{\sigma_2}\begin{bmatrix} \frac{1}{\sqrt{R}} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{R}} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{R}} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{R}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \frac{1}{\sigma_2}\begin{bmatrix} \frac{1}{\sqrt{R}} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{R}} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{R}} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{R}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\implies \Omega^{-1} = \frac{1}{\sigma_2^2}\begin{bmatrix} \frac{1}{R} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{R} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{R} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{R} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}$$

So our estimation is:

$$\hat{\beta} = \left(X_t^T \Omega X_t\right)^{-1} X_t^T \Omega^{-1} Y_t$$

$$\Omega^{-1} = \frac{1}{\sigma_2^2}\begin{bmatrix} \frac{1}{R} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{R} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{R} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{R} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Problem 3 - Theory

We have the following:

$$Y_t = X_t\beta + \varepsilon_t \qquad t = 1,2 \quad X_1 = X_2 = 1 \qquad E\left[\varepsilon_1\right] = E\left[\varepsilon_2\right] = 0 \qquad Var(\varepsilon_1) = Var(\varepsilon_2) = \sigma^2 \qquad Cov(\varepsilon_1, \varepsilon_2) = \sigma_{12}$$

$$\rho = \tfrac{\sigma_{12}}{\sigma^2} \qquad VC(\hat{\beta}) = (X^TX)^{-1}X^T(\sigma^2\Omega)X(X^TX)^{-1}$$

We know that $\sigma^2\Omega$ is the variance-covariance matrix of ε:

$$VC(\varepsilon) = \begin{bmatrix} \sigma^2 & \sigma_{12} \\ \sigma_{21} & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

$$VC(\hat{\beta}) = (X^TX)^{-1}X^TVC(\varepsilon)X(X^TX)^{-1}$$

$$(X^TX)^{-1} = \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)^{-1} = \tfrac{1}{2} \qquad X^T = \begin{bmatrix} 1 & 1 \end{bmatrix} \qquad VC(\varepsilon) = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \qquad X = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\implies VC(\hat{\beta}) = \frac{\sigma^2\left(\rho+1\right)}{2}$$

$$VC(think) = \sigma^2\left(X^TX\right)^{-1} = \frac{\sigma^2}{2}$$

By comparing these two variance-covariance values, we can see if we have a negative correlation then the variance-covariance value of the OLS estimator is greater than the variance-covariance of the GLS estimator:

$$\frac{\sigma^2}{2} > \frac{\sigma^2\left(\rho+1\right)}{2} \implies 1 > 1+\rho \implies \rho < 0$$

Problem 1 - Computational

We want to estimate a linear regression model of *cnttdhh* as a function of the four variables of *URBAN*, *hhsize*, *numadlt*, and *youngchild*. *URBAN* is a dummy variable:

$$URBAN \begin{cases} 1 & urbrur = 1 \\ 0 & otherwise \end{cases}$$

Before beginning to solve the problem as a review we take a very short look at the different model metrics.

a) **Sum of Squares (SS)**
Measures variance explained by each parameter. A higher value is preferable, which shows that the parameter explains more variance.

$$\sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

b) **Sum of Squared Error(SSE)**
Measures total unexplained variance. A lower value is preferable, which shows better model fit.

$$\sum \left(Y_i - \hat{Y}_i\right)^2$$

c) **Total Sum of Squares(SST)**
Measures how much the observed data deviated from the overall mean.

$$\sum \left(Y_i - \bar{Y}\right)^2$$

$$SST = SS + SSE$$

d) **Standard Error(SE)**
Measures the uncertainty in estimating the regression coefficients. Small values show coefficients are precise.

$$SE_\beta = \sqrt{\frac{SSE}{(N-K)\sum(X_i - \bar{X})^2}}$$

e) **Mean Squared Error(MSE)**
Is the mean of the sum of squared error. The less the better.

$$MSE = \frac{SSE}{N-K}$$

f) **R-squared(R^2)**
Measures the proportion of variance explained by the model. The closer to 1 the better our model is.

$$R^2 = \frac{SS}{SST}$$

g) **P-value of the t-test**
The probability that the coefficient is not significantly different from zero. If its value is less than 0.05 the parameter is statistically significant; otherwise, it is not.

h) **P-value of the F-test**
The probability that the coefficient does not explain variance. If its value is less than 0.05, the parameter significantly improves the model; otherwise, it does not.

The dummy variable can be added to both slope and interception, to only interception, and only slope. In order to decide on how we should add the dummy variable to our model, we perform a F-test between the complete model and the nested models.

Slope and Intercept ($m1$)

```
> # Dummy Variable Interacted to Both Intercept and Slope (m1)
> m1 <- lm(cnttdhh ~ hhsize + numadlt + youngchild + URBAN + URBHHS
+          + URBADL + URBYOCH, data = D_HH_HI)
```

$$\begin{aligned} cnttdhh_i = \beta_0 &+ \beta_1 (hhsize_i) + \beta_2 (numadlt_i) + \beta_3 (youngchild_i) + \beta_4 (URBAN_i) \\ &+ \beta_5 (hhsize_i.URBAN_i) + \beta_6 (numadlt_i.URBAN_i) + \beta_7 (youngchild_i.URBAN_i) + \varepsilon_i \end{aligned}$$

Intercept Only ($m2$)

```
> # Dummy Variable Interacted to Just Intercept (m2)
> m2 <- lm(cnttdhh ~ hhsize + numadlt + youngchild + URBAN, data = D_HH_HI)
```

$$cnttdhh_i = \beta_0 + \beta_1 (hhsize_i) + \beta_2 (numadlt_i) + \beta_3 (youngchild_i) + \beta_4 (URBAN_i) + \varepsilon_i$$

Slope Only ($m3$)

```
> # Dummy Variable Interacted to Just Slope (m3)
> m3 <- lm(cnttdhh ~ hhsize + numadlt + youngchild + URBHHS
+ URBADL + URBYOCH, data = D_HH_HI)
```

$$\begin{aligned} cnttdhh_i = \beta_0 &+ \beta_1 (hhsize_i) + \beta_2 (numadlt_i) + \beta_3 (youngchild_i) \\ &+ \beta_4 (hhsize_i.URBAN_i) + \beta_5 (numadlt_i.URBAN_i) + \beta_6 (youngchild_i.URBAN_i) + \varepsilon_i \end{aligned}$$

Model ($m2$) vs Model ($m1$)
It is observed that the test value of F is greater than the critical value, so our NULL hypothesis of ignoring slope interaction is rejected.

```
> F_test2 <- ((SSE2 - SSE1) / (df2 - df1)) / (SSE1 / df1)
> F_test2
[1] 3.198286
> F_critical2 <- qf(0.95, df2 - df1, df1)
> F_critical2
[1] 2.607718
```

Model ($m3$) vs Model ($m1$)
It is observed that the test value of F is less than the critical value, so our NULL hypothesis of ignoring intercept interaction is accepted.

```
> F_test3<-((SSE3-SSE1)/(df3-df1))/(SSE1/df1)
> F_test3
[1] 0.2181682
> F_critical3 <- qf(0.95,df3-df1,df1)
> F_critical3
[1] 3.844401
```

So we choose ($m3$) as our pooled model. However, looking at the output of this model (t-test) we see that some of the parameters are insignificant.

```
Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)   1.6103     0.1528  10.537 < 0.0000000000000002 ***
hhsize        3.3438     0.6563   5.095          0.000000369 ***
numadlt      -2.9252     0.9455  -3.094              0.00199 **
youngchild   -0.2733     1.6805  -0.163              0.87082
URBHHS       -0.7439     0.6628  -1.122              0.26178
URBADL        1.5435     0.9491   1.626              0.10399
URBYOCH      -2.7597     1.6942  -1.629              0.10344
```

Now we can take one step further and create model $m4$ by ignoring the insignificant parameters of model $m3$ and perform an F-test between these two models to check whether ignoring these parameters is valid or not. It is observed that the test value of F is greater than the critical value, so our NULL hypothesis of ignoring insignificant parameters of model ($m3$) is rejected.

```
> # Removing Insignificant Parameters of m3 (m4)
> m4 <- lm(cnttdhh ~ hhsize + numadlt, data = D_HH_HI)

> F_test4<-((SSE4-SSE3)/(df4-df3))/(SSE3/df3)
> F_test4
[1] 51.24632
> F_critical4 <- qf(0.95,df4-df3,df3)
> F_critical4
[1] 2.374742
```

Our final decision is to choose ($m3$) as our pooled model.

Now we apply market segmentation (Exogenous Approach). We split our dataset into three groups based on income.

a) Low Income: *hhfaminc_imp* $\leq$ 49,999

b) Medium Income: 49,999 $<$ *hhfaminc_imp* $\leq$ 124,999

c) High Income: *hhfaminc_imp* $\geq$ 124,999

We create the 3 sub-groups (models $m5$, $m6$, and $m7$) from the dataset and use the structure of the pooled model ($m3$) on each. In the following table, the results for the pooled model and each of the sub-group models are displayed.

```
> # Low Income Family (m5)
> m5 <- lm(cnttdhh ~ hhsize + numadlt + youngchild + URBHHS
+          + URBADL + URBYOCH, data = D_HH_HI_LOWINC)
> # Medium  Income Family (m6)
> m6 <- lm(cnttdhh ~ hhsize + numadlt + youngchild + URBHHS
+          + URBADL + URBYOCH, data = D_HH_HI_MEDINC)
> # High  Income Family (m7)
> m7 <- lm(cnttdhh ~ hhsize + numadlt + youngchild + URBHHS
+          + URBADL + URBYOCH, data = D_HH_HI_HIGHINC)
```

Model 3					
Variable	**Coefficient**	**Std. Error**	**t-statistic**	$Pr(>\|t\|)$	**Significancy**
Intercept	1.6103	0.1528	10.537	< 0.0000000000000002	***
hhsize	3.3438	0.6563	5.095	0.000000369	***
numadlt	-2.9252	0.9455	-3.094	0.00199	**
youngchild	-0.2733	1.6805	-0.163	0.87082	
URBHHS	-0.7439	0.6628	-1.122	0.26178	
URBADL	1.5435	0.9491	1.626	0.10399	
URBYOCH	-2.7597	1.6942	-1.629	0.10344	
N	3170				
SSE	41725.79				
MSE	13.19184				
R^2	0.3001				

Table 1: Estimation Results for Model 3

Estimation Results for Models 5, 6, and 7															
Variable	**Model 5**					**Model 6**					**Model 7**				
	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.
Intercept	1.3669	0.2387	5.727	0.0000000146	***	2.05602	0.23352	8.804	< 0.0000000000000002	***	2.3755	0.3365	7.060	0.00000000000319	***
hhsize	8.8460	2.3126	3.825	0.000141	***	2.57722	0.90280	2.855	0.00437	**	6.8675	2.1406	3.208	0.00138	**
numadlt	-8.2985	2.4762	-3.351	0.000843	***	-2.44789	1.48393	-1.650	0.09925		-6.4228	2.5516	-2.517	0.01199	*
youngchild	-11.0000	4.3263	-2.543	0.011196	*	4.95366	3.14462	1.575	0.11542		-6.1325	4.5146	-1.358	0.17467	
URBHHS	-6.7489	2.3188	-2.911	0.003711	**	-0.04171	0.91512	-0.046	0.96365		-3.9994	2.1464	-1.863	0.06272	
URBADL	7.0943	2.4839	2.856	0.004402	**	0.90379	1.48996	0.607	0.54422		4.6899	2.5529	1.837	0.06651	
URBYOCH	8.8084	4.3450	2.027	0.042978	*	-8.20230	3.16208	-2.594	0.00959	**	3.2224	4.5303	0.711	0.47708	
N	788					1415					967				
SSE	7308.95					17720.68					15167.73				
MSE	9.358451					12.58571					15.79972				
R^2	0.2534					0.2505					0.335				

Table 2: Comparison of Estimation Results for Models 5, 6, and 7

We can have a quick review of pooled model and segmented models. We use MSE instead of SSE since each model has its own unique number of data points.

By comparing these models, we observe that **Number of Significant Variables** in $m5$ is more than all other models. Model $m5$ and $m6$ has better **MSE** than $m3$. Model $m7$ has better $\boldsymbol{R}^2$ than model $m3$.

To see the effect of market segmentation we perform an F-test between the pooled model ($m3$) and segmented models ($m5$, $m6$, and $m7$). It is observed that the segmented models are more significant than pooled model, so our NULL hypothesis of ignoring market segmentation is rejected.

```
Model 1: cnttdhh ~ hhsize + numadlt + youngchild + URBHHS + URBADL + URBYOCH
Model 2: cnttdhh ~ hhsize + numadlt + youngchild + URBHHS + URBADL + URBYOCH +
    as.factor(hhfaminc_imp)
  Res.Df   RSS Df Sum of Sq      F                Pr(>F)
1   3163 41726
2   3153 40351 10    1374.4 10.739 < 0.00000000000000022 ***
```

When we don't use the segmentation we treat all families, regardless of their income, in the same way, and using the same model parameters for all of them; however, by checking the value and the degree of significance of the model parameters between different groups, we see these values are vary from one income group to another. For one thing, low-income households show a stronger bond between household size and their location (urban or non-urban), while other groups do not offer such a relation. The pooled model fails to capture these nuances.

Problem 2 - Computational

For this part, we use the Endogenous segmentation approach. We have three categories, and in order to do so we define the following dummy variables:

$$\text{LOWINC} = \begin{cases} 1 & \text{if } hhfaminc\_imp \leq 49{,}999 \\ 0 & \text{otherwise} \end{cases} \qquad \text{MEDINC} = \begin{cases} 1 & \text{if } 49{,}999 < hhfaminc\_imp \leq 124{,}999 \\ 0 & \text{otherwise} \end{cases}$$

Model ($m8$)

```
> #### Market Segmentation Based on Income (m8-Endogenous) ####
> m8 <- lm(cnttdhh ~ hhsize + numadlt + youngchild + URBHHS
+          + URBADL + URBYOCH + LOWINC + MEDINC + LOWINCHHS + LOWINCADL
+          + LOWINCYOCH + LOWINCURBHHS + LOWINCURBADL + LOWINCURBYOCH
+          + MEDINCHHS + MEDINCAD + MEDINCYOCH + MEDINCURBHHS
+          + MEDINCURBADL + MEDINCURBYOCH, data = D_HH_HI)
```

$$\begin{aligned} cnttdhh_i = {} & \beta_0 + \beta_1(hhsize_i) + \beta_2(numadlt_i) + \beta_3(youngchild_i) \\ & + \beta_4(hhsize_i \cdot URBAN_i) + \beta_5(numadlt_i \cdot URBAN_i) + \beta_6(youngchild_i \cdot URBAN_i) \\ & + \beta_7(LOWINC_i) + \beta_8(MEDINC_i) \\ & + \beta_9(hhsize_i \cdot LOWINC_i) + \beta_{10}(numadlt_i \cdot LOWINC_i) + \beta_{11}(youngchild_i \cdot LOWINC_i) \\ & + \beta_{12}(hhsize_i \cdot URBAN_i \cdot LOWINC_i) + \beta_{13}(numadlt_i \cdot URBAN_i \cdot LOWINC_i) \\ & + \beta_{14}(youngchild_i \cdot URBAN_i \cdot LOWINC_i) \\ & + \beta_{15}(hhsize_i \cdot MEDINC_i) + \beta_{16}(numadlt_i \cdot MEDINC_i) + \beta_{17}(youngchild_i \cdot MEDINC_i) \\ & + \beta_{18}(hhsize_i \cdot URBAN_i \cdot MEDINC_i) + \beta_{19}(numadlt_i \cdot URBAN_i \cdot MEDINC_i) \\ & + \beta_{20}(youngchild_i \cdot URBAN_i \cdot MEDINC_i) + \varepsilon_i \end{aligned}$$

Now we can show the equivalency of the Endogenous and Exogenous approaches. From Table 3, we can observe:

Model ($m5$) **and Model** ($m8$)

$$Intercept_5 = Intercept_8 + LOWINC_8 = 2.3755 + -1.0086 = 1.3669$$

$$hhsize_5 = hhsize_8 + LOWINCHHS_8 = 6.8675 + 1.9785 = 8.8460$$

$$numadlt_5 = numadlt_8 + LOWINCADL_8 = -6.4228 + -1.8757 = -8.2985$$

$$youngchild_5 = youngchild_8 + LOWINCYOCH_8 = -6.1325 + -4.8675 = -11.0000$$

$$URBHHS_5 = URBHHS_8 + LOWINCURBHHS_8 = -3.9994 + -2.7495 = -6.7489$$

$$URBADL_5 = URBADL_8 + LOWINCURBADL_8 = 4.6899 + 2.4044 = 7.0943$$

$$URBYOCH_5 = URBYOCH_8 + URBYOCH_8 = 3.2224 + 5.5860 = 8.8084$$

Model ($m6$) **and Model** ($m8$)

$$Intercept_6 = Intercept_8 + MEDINC_8 = 2.3755 + -0.3195 = 2.05602$$

$$hhsize_6 = hhsize_8 + MEDINCHHS_8 = 6.8675 + -4.2903 = 2.57722$$

$$numadlt_6 = numadlt_8 + MEDINCADL_8 = -6.4228 + 3.9749 = -2.44789$$

$$youngchild_6 = youngchild_8 + MEDINCYOCH_8 = -6.1325 + 11.0861 = 4.95366$$

$$URBHHS_6 = URBHHS_8 + MEDINCURBHHS_8 = -3.9994 + 3.9577 = -0.04171$$

$$URBADL_6 = URBADL_8 + MEDINCURBADL_8 = 4.6899 + -3.7862 = 0.90379$$

$$URBYOCH_6 = URBYOCH_8 + URBYOCH_8 = 3.2224 + -11.4247 = -8.20230$$

Model ($m7$) and Model ($m8$)

$$Intercept_7 = Intercept_8 = 2.3755$$

$$hhsize_7 = hhsize_8 = 6.8675$$

$$numadlt_7 = numadlt_8 = -6.4228$$

$$youngchild_7 = youngchild_8 = -6.1325$$

$$URBHHS_7 = URBHHS_8 = -3.9994$$

$$URBADL_7 = URBADL_8 = 4.6899$$

$$URBYOCH_7 = URBYOCH_8 = 3.2224$$

Model 8					
Variable	**Coefficient**	**Std. Error**	**t-statistic**	$Pr(>\|t\|)$	**Significance**
Intercept	2.3755	0.3024	7.855	0.00000000000000545	***
hhsize	6.8675	1.9240	3.569	0.000363	***
numadlt	-6.4228	2.2935	-2.800	0.005136	**
youngchild	-6.1325	4.0580	-1.511	0.130832	
URBHHS	-3.9994	1.9293	-2.073	0.038250	*
URBADL	4.6899	2.2947	2.044	0.041055	*
URBYOCH	3.2224	4.0721	0.791	0.428808	
LOWINC	-1.0086	0.4113	-2.452	0.014256	*
MEDINC	-0.3195	0.3831	-0.834	0.404416	
LOWINCHHS	1.9785	3.3162	0.597	0.550803	
LOWINCADL	-1.8757	3.6910	-0.508	0.611363	
LOWINCYOCH	-4.8675	6.4805	-0.751	0.452650	
LOWINCURBHHS	-2.7495	3.3251	-0.827	0.408354	
LOWINCURBADL	2.4044	3.6988	0.650	0.515708	
LOWINCURBYOCH	5.5860	6.5064	0.859	0.390662	
MEDINCHHS	-4.2903	2.1280	-2.016	0.043878	*
MEDINCAD	3.9749	2.7375	1.452	0.146597	
MEDINCYOCH	11.0861	5.1475	2.154	0.031339	*
MEDINCURBHHS	3.9577	2.1381	1.851	0.064255	
MEDINCURBADL	-3.7862	2.7418	-1.381	0.167401	
MEDINCURBYOCH	-11.4247	5.1694	-2.210	0.027174	
N	3170				
SSE	40197.36				
MSE	12.76512				
R^2	0.3257				

Estimation Results for Models 5, 6, and 7

Variable	**Model 5**					**Model 6**					**Model 7**				
	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.
Intercept	1.3669	0.2387	5.727	0.0000000146	***	2.05602	0.23352	8.804	<0.0000000000000002	***	2.3755	0.3365	7.060	0.00000000000319	***
hhsize	8.8460	2.3126	3.825	0.000141	***	2.57722	0.90280	2.855	0.00437	**	6.8675	2.1406	3.208	0.00138	**
numadlt	-8.2985	2.4762	-3.351	0.000843	***	-2.44789	1.48393	-1.650	0.09925		-6.4228	2.5516	-2.517	0.01199	*
youngchild	-11.0000	4.3263	-2.543	0.011196	*	4.95366	3.14462	1.575	0.11542		-6.1325	4.5146	-1.358	0.17467	
URBHHS	-6.7489	2.3188	-2.911	0.003711	**	-0.04171	0.91512	-0.046	0.96365		-3.9994	2.1464	-1.863	0.06272	
URBADL	7.0943	2.4839	2.856	0.004402	**	0.90379	1.48996	0.607	0.54422		4.6899	2.5529	1.837	0.06651	
URBYOCH	8.8084	4.3450	2.027	0.042978	*	-8.20230	3.16208	-2.594	0.00959	**	3.2224	4.5303	0.711	0.47708	
N	788					1415					967				
SSE	7308.95					17720.68					15167.73				
MSE	9.358451					12.58571					15.79972				
R^2	0.2534					0.2505					0.335				

Table 3: Estimation Results for Models 8, 5, 6, and 7

Problem 3 - Computational

We want to estimate a linear regression model of *cnttdhh* as a function of the three variables of *URBAN*, *hhsize*, and *ADLT0TO4*. *URBAN* is a dummy variable:

$$URBAN \begin{cases} 1 & urbrur = 1 \\ 0 & otherwise \end{cases}$$

The dummy variable can be added to both slope and interception, to only interception, and only slope. In order to decide on how we should add the dummy variable to our model, we perform a F-test between the complete model and the nested models.

Slope and Intercept ($m9$)

```
# Dummy Variable Interacted to Both Intercept and Slope (m9)
m9 <- lm(cnttdhh ~ hhsize + ADLT0T04 + URBAN + URBHHS
         + URBADLT0T04, data = D_HH_HI)
```

$$\begin{aligned} cnttdhh_i = \beta_0 + \beta_1\left(hhsize_i\right) + \beta_2\left(ADLT0TO4_i\right) + \beta_3\left(URBAN_i\right) \\ + \beta_4\left(hhsize_i.URBAN_i\right) + \beta_5\left(ADLT0TO4_i.URBAN_i\right) + \varepsilon_i \end{aligned}$$

Intercept Only ($m10$)

```
> # Dummy Variable Interacted to Intercept Only (m10)
> m10 <- lm(cnttdhh ~ hhsize + ADLT0T04 + URBAN, data = D_HH_HI)
```

$$cnttdhh_i = \beta_0 + \beta_1\left(hhsize_i\right) + \beta_2\left(ADLT0TO4_i\right) + \beta_3\left(URBAN_i\right) + \varepsilon_i$$

Slope Only ($m11$)

Intercept Only ($m10$)

```
> # Dummy Variable Interacted to Just Slope (m11)
> m11 <- lm(cnttdhh ~ hhsize + ADLT0T04  + URBHHS
+           + URBADLT0T04, data = D_HH_HI)
```

$$\begin{aligned} cnttdhh_i = \beta_0 + \beta_1\left(hhsize_i\right) + \beta_2\left(ADLT0TO4_i\right) + \\ + \beta_3\left(hhsize_i.URBAN_i\right) + \beta_4\left(ADLT0TO4_i.URBAN_i\right) + \varepsilon_i \end{aligned}$$

Model ($m10$) **vs Model** ($m9$)
It is observed that the test value of F is less than the critical value, so our NULL hypothesis of ignoring slope interactions is accepted.

```
> F_test10<-((SSE10-SSE9)/(df10-df9))/(SSE9/df9)
> F_test10
[1] 0.8535379
> F_critical10 <- qf(0.95,df10-df9,df9)
> F_critical10
[1] 2.99857
```

Model ($m11$) vs Model ($m9$)
It is observed that the test value of F is less than the critical value, so our NULL hypothesis of ignoring intercept interaction is accepted.

```
> F_test11<-((SSE11-SSE9)/(df11-df9))/(SSE9/df9)
> F_test11
[1] 0.1235122
> F_critical11 <- qf(0.95,df11-df9,df9)
> F_critical11
[1] 3.8444
```

Now we can see both $m10$ and $m11$ models are better than $m9$ model. The question is what model to pick. By comparing the MSE, SSE, and R^2 of these models we finally pick model $m11$.

$MSE_{10} = 13.48 \quad MSE_{11} = 13.48 \quad SSE_{10} = 42678.7 \quad SSE_{11} = 42657.35 \quad R^2_{10} = 0.2841 \quad R^2_{11} = 0.2844$

```
Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)   2.0040     0.1464  13.693 < 0.0000000000000002 ***
hhsize        3.4322     0.6624   5.182         0.000000234 ***
ADLTOTO4     -2.8920     0.9470  -3.054             0.00228 **
URBHHS       -0.8834     0.6689  -1.321             0.18668
URBADLTOTO4   1.2944     0.9511   1.361             0.17364
```

Now we can take one step further and create model $m12$ by ignoring the insignificant parameters of model $m11$ and perform an F-test between these two models to check whether ignoring these parameters is valid or not. It is observed that the test value of F is less than the critical value, so our NULL hypothesis of ignoring insignificant parameters of the model ($m11$) is accepted.

```
> # Removing Insignificant Parameters of m11 (m12)
> m12 <- lm(cnttdhh ~ hhsize + ADLTOTO4, data = D_HH_HI)

> F_test12<-((SSE12-SSE11)/(df12-df11))/(SSE11/df11)
> F_test12
[1] 0.9282733
> F_critical12 <- qf(0.95,df12-df11,df11)
> F_critical12
[1] 2.99857
```

Because models ($m10$) and ($m11$) are so close, we can perform another F-test between $m10$ and $m12$. We see again that the test value of F is less than the critical value, so our NULL hypothesis of ignoring insignificant parameters of the model ($m10$) is accepted.

```
> F_test12prime<-((SSE12-SSE10)/(df12-df10))/(SSE10/df10)
> F_test12prime
[1] 0.272494
> F_critical12prime <- qf(0.95,df12-df10,df10)
> F_critical12prime
[1] 3.844398
```

Our final decision is to choose ($m12$) as our pooled model.

Now we apply market segmentation (Exogenous Approach). We split our dataset into three groups based on vehicle ownership.

a) 0 Vehicle Family: $hhvehcnt = 0$

b) 1 Vehicle Family: $hhvehcnt = 1$

c) 2+ Vehicle Family: $hhvehcnt \geq 2$

We create the 3 sub-groups (models $m13$, $m14$, and $m15$) from the dataset and use the structure of the pooled model ($m12$) on each. In the following table, the results for the pooled model and each of the sub-group models are displayed.

```
> # 0 Vehicle Family (m13)
> m13 <- lm(cnttdhh ~ hhsize + ADLTOTO4, data = D_HH_HI_0VEH)
> # 1 Vehicle Family (m14)
> m14 <- lm(cnttdhh ~ hhsize + ADLTOTO4, data = D_HH_HI_1VEH)
> # Plus 2 Vehicles Family (m15)
> m15 <- lm(cnttdhh ~ hhsize + ADLTOTO4, data = D_HH_HI_2PLUSVEH)
```

Model 12					
Variable	**Coefficient**	**Std. Error**	**t-statistic**	$Pr(>\|t\|)$	**Significance**
Intercept	2.00285	0.14635	13.69	< 0.0000000000000002	***
hhsize	2.56383	0.09254	27.70	< 0.0000000000000002	***
ADLT0TO4	-1.61548	0.12688	-12.73	< 0.0000000000000002	***
N	3170				
SSE	42682.37				
MSE	13.47722				
R^2	0.284				

Table 4: Estimation Results for Model 12

Estimation Results for Models 13, 14, and 15															
Variable	**Model 13**					**Model 14**					**Model 15**				
	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.
Intercept	1.8494	0.3112	5.944	0.00000000757	***	1.9470	0.2022	9.630	< 0.0000000000000002	***	2.7268	0.2718	10.034	< 0.0000000000000002	***
hhsize	1.4752	0.3881	3.801	0.000174	***	2.6515	0.1748	2.855	< 0.0000000000000002	***	2.5140	0.1196	21.028	< 0.0000000000000002	***
ADLT0TO4	-1.0300	0.4852	-2.123	0.034555	*	-1.7896	0.2265	-7.901	0.00000000000000614	***	-1.7167	0.1718	-9.992	< 0.0000000000000002	***
N	310					1224					1636				
SSE	2458.348					11669.2					27774.15				
MSE	8.007648					9.557086					17.00805				
R^2	0.09032					0.2171					0.2617				

Table 5: Comparison of Estimation Results for Models 13, 14, and 15

We can have a quick review of pooled model and segmented models. We use MSE instead of SSE since each model has its own unique number of data points.

By comparing these models, we observe that model $m13$ and $m14$ has better **MSE** than $m12$.

To see the effect of market segmentation we perform an F-test between the pooled model ($m12$) and segmented models ($m13$, $m14$, and $m15$). It is observed that the segmented models are more significant than the pooled model, so our NULL hypothesis of ignoring market segmentation is rejected.

```
Model 1: cnttdhh ~ hhsize + ADLTOTO4
Model 2: cnttdhh ~ hhsize + ADLTOTO4 + as.factor(hhvehcnt)
  Res.Df   RSS Df Sum of Sq      F          Pr(>F)
1   3167 42682
2   3160 41862  7    820.21 8.8449 0.0000000000806 ***
```

Problem 4 - Computational

For this part, we use the Endogenous segmentation approach. We have three categories, and in order to do so we define the following dummy variables:

$$\text{NCAR1} = \begin{cases} 1 & \text{if } hhvehcnt = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{NCAR2} = \begin{cases} 1 & \text{if } hhvehcnt \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

Model ($m16$)
As it was asked in the problem, we only add our dummy variables for Endogenous approach to the intercept of our pooled model. In other words, we do not consider the slope interactions.

```
> #### Market Segmentation Based on Income (Endogenous) ####
> m16 <- lm(cnttdhh ~ hhsize + ADLT0TO4 + NCAR1 + NCAR2, data = D_HH_HI)
```

$$cnttdhh_i = \beta_0 + \beta_1(hhsize_i) + \beta_2(ADLT0TO4_i) + \beta_3(NCAR1_i) + \beta_4(NCAR2_i) + \varepsilon_i$$

Now we can check the equivalency of the Endogenous and Exogenous approaches. From Table 6, we can observe:

Model ($m13$), ($m14$), and ($m15$) and Model ($m16$)
We can see neither the interceptions nor other parameters are the same for Endogenous and Exogenous models. Unlike the **Problem 2**, where we saw the equivalency between Endogenous and Exogenous models. This is rooted in not considering the slope interactions.

		Model 16			
Variable	**Coefficient**	**Std. Error**	**t-statistic**	$Pr(>\|t\|)$	**Significance**
Intercept	1.25753	0.23306	5.396	0.0000000732914	***
hhsize	2.50330	0.09237	27.100	0.0000000000000002	***
ADLT0TO4	-1.73493	0.12754	-13.603	0.0000000000000002	***
NCAR1	0.86150	0.23186	3.716	0.000206	***
NCAR2	1.54542	0.23660	6.532	0.0000000000755	***
N	3170				
SSE	42041.35				
MSE	13.28321				
R^2	0.2948				

Estimation Results for Models 13, 14, and 15

Variable	**Model 13**					**Model 14**					**Model 15**				
	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.
Intercept	1.8494	0.3112	5.944	0.00000000757	***	1.9470	0.2022	9.630	0.0000000000000002	***	2.7268	0.2718	10.034	0.00000000000319	***
hhsize	1.4752	0.3881	3.801	0.000174	***	2.6515	0.1748	15.171	0.0000000000000002	***	2.5140	0.1196	21.028	0.0000000000000002	***
ADLT0TO4	-1.0300	0.4852	-2.123	0.034555	*	-1.7896	0.2265	-7.901	0.00000000000000614	***	-1.7167	0.1718	-9.992	0.0000000000000002	***
N	310					1224					1636				
SSE	2458.348					11669.2					27774.15				
MSE	8.007648					9.557086					17.00805				
R^2	0.09032					0.2171					0.2617				

Table 6: Estimation Results for Models 16, 13, 14, and 15

Conclusion
We can see that if we only consider $NCAR1$ and $NCAR2$ without interactions in the slope (interactions with other variables in the pooled model), Endogenous and Exogenous models have different intercepts and slopes for each group. This means that these two models no longer are the same. This happens because $m16$ does not include interactions, but still the effect of *hhsize* and *ADLT0TO4* may be different across vehicle ownership groups. The pooled model averages out these effects and this causes the difference between these two approaches. Without slope interaction terms, the intercept absorbs some of the slope interactions variations. In a more simple explanation, without slope interactions, the slope of *hhsize* and *ADLT0TO4* are forced to be the same across all groups. This means the intercept shifts (β_3 and β_4) cannot fully capture differences.

Ignoring the F-test and Considering *URBAN*
Because it was asked in the problem to include $URBAN$ as one of the variables, we have repeated the **Problem 3 and 4** ignoring the F-test between $m12$ and $m11$ (that F-test showed us removing the insignificant variables improves the model), so our poolede model is $m11$. You can follow this approach through models $m13prime, m14prime, m15prime$, and $m16prime$. This approach makes no difference in the final result, as it can be see in Table 7. Also, one may wonder why we have both $ADLT0TO4$ and $URBADLT0TO4$ in $m15prime$, but only we have $ADLT0TO4$ in $m13prime$ and $m14prime$. This happened because we saw collinearity between $ADLT0TO4$ and $URBADLT0TO4$ in those two models. This happened because in our data almost all families with 0 vehicles and 1 vehicle live in urban areas.

```
> #### Market Segmentation Based on Car Ownership (Exogenous) - Considering Modell 11 ####
> # 0 Vehicle Family (m13prime)
> m13prime <- lm(cnttdhh ~ hhsize + ADLT0TO4 + URBHHS, data = D_HH_HI_0VEH)
> # 1 Vehicle Family (m14prime)
> m14prime <- lm(cnttdhh ~ hhsize + ADLT0TO4 + URBHHS, data = D_HH_HI_1VEH)
> # Plus 2 Vehicles Family (m15)
> m15prime <- lm(cnttdhh ~ hhsize + ADLT0TO4 + URBHHS
+                + URBADLT0TO4, data = D_HH_HI_2PLUSVEH)
> #### Market Segmentation Based on Income (Endogenous) - Considering Modell 11 ####
> m16prime <- lm(cnttdhh ~ hhsize + ADLT0TO4 + URBHHS
+                + URBADLT0TO4 + NCAR1 + NCAR2, data = D_HH_HI)
```

Model 16prime					
Variable	**Coefficient**	**Std. Error**	**t-statistic**	$Pr(>\|t\|)$	**Significance**
Intercept	1.2580	0.2331	5.398	0.0000000725496	***
hhsize	3.3828	0.6576	5.144	0.0000002851541	***
ADLT0TO4	-3.0767	0.9406	-3.271	0.001083	**
URBHHS	-0.8937	0.6640	-1.346	0.178423	
URBADLT0TO4	1.3587	0.9443	1.439	0.150270	
NCAR1	0.8620	0.2319	3.718	0.000205	***
NCAR2	1.5485	0.2366	6.544	0.0000000000698	***
N	3170				
SSE	42013.71				
MSE	13.28287				
R^2	0.2952				

Estimation Results for Models 13prime, 14prime, and 15prime

Variable	Model 13prime					Model 14prime					Model 15prime				
	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.	Coef.	Std. Err.	t-stat	$Pr(>\|t\|)$	Sig.
Intercept	1.8493	0.3115	5.937	0.00000000787	***	1.9465	0.2022	9.628	0.0000000000000002	***	2.7314	0.2719	10.046	0.0000000000000002	***
hhsize	0.5949	1.4815	0.402	0.6883		1.8710	0.7980	2.345	0.0192	*	3.2826	0.7833	4.191	0.0000293	***
ADLT0TO4	-1.0196	0.4859	-2.098	0.0367	*	1.7818	0.2266	-7.862	0.00000000000000829	***	-2.8662	1.1582	-2.475	0.0134	*
URBHHS	0.8741	1.4197	0.616	0.5385		0.7765	0.7747	1.002	0.3164		-0.7851	0.7927	-0.990	0.3221	
URBADLT0TO4	-	-	-	-	-	-	-	-	-	-	1.1681	1.1624	1.005	0.3151	
N	310					1224					1636				
SSE	2455.306					11659.6					27756.81				
MSE	8.023876					9.55705					17.01827				
R^2	0.09145					0.2177					0.2622				

Table 7: Estimation Results for Models 16prime, 13prime, 14prime, and 15prime

At the end, if we consider the slope interactions in Endogenous model, the Exogeneous and Endogenous models become identical and give us the same results. Moreover, if we use *numadlt* and *youngchild* attributes instead of $NUMADL0TO4$ (which is a sum of these two attributes), our F-test between $m11$ and $m12$ will tell us to keep the insignificant parameters of $m11$, and we will use $m11$ as our pooled model (just like we did in **Problems 1 and 2**.