Final Project

Group 4

Question

1. **How does an individual's race affect the amount of cigarettes smoked per day.**
2. The effect of income on amount of cigarettes smoked in a day.
3. The effect of education on the amount of cigarettes smoked in a day.

Why is this important?

1. Results will help programs against smoking target certain demographics that are need the most help or that are more likely to smoke than others.
    a. Government programs
    b. School programs
2. Marketing teams can more efficiently spend their budgeted dollars targeting at risk individuals.
3. Cigarette companies will know who they should target as their customers.

Data

Summary statistics of the variables (Mean, std, median, and etc):

```
> summary(df)
      educ          cigpric          white             age             income      
 Min.   : 6.00   Min.   :44.00   Min.   :0.0000   Min.   :17.00   Min.   :  500  
 1st Qu.:10.00   1st Qu.:58.14   1st Qu.:1.0000   1st Qu.:28.00   1st Qu.:12500  
 Median :12.00   Median :61.05   Median :1.0000   Median :38.00   Median :20000  
 Mean   :12.47   Mean   :60.30   Mean   :0.8786   Mean   :41.24   Mean   :19305  
 3rd Qu.:13.50   3rd Qu.:63.18   3rd Qu.:1.0000   3rd Qu.:54.00   3rd Qu.:30000  
 Max.   :18.00   Max.   :70.13   Max.   :1.0000   Max.   :88.00   Max.   :30000  
      cigs           restaurn         lincome           agesq          lcigpric    
 Min.   : 0.000   Min.   :0.0000   Min.   : 6.215   Min.   : 289   Min.   :3.784  
 1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 9.433   1st Qu.: 784   1st Qu.:4.063  
 Median : 0.000   Median :0.0000   Median : 9.903   Median :1444   Median :4.112  
 Mean   : 8.686   Mean   :0.2466   Mean   : 9.687   Mean   :1990   Mean   :4.096  
 3rd Qu.:20.000   3rd Qu.:0.0000   3rd Qu.:10.309   3rd Qu.:2916   3rd Qu.:4.146  
 Max.   :80.000   Max.   :1.0000   Max.   :10.309   Max.   :7744   Max.   :4.250  
```

Data

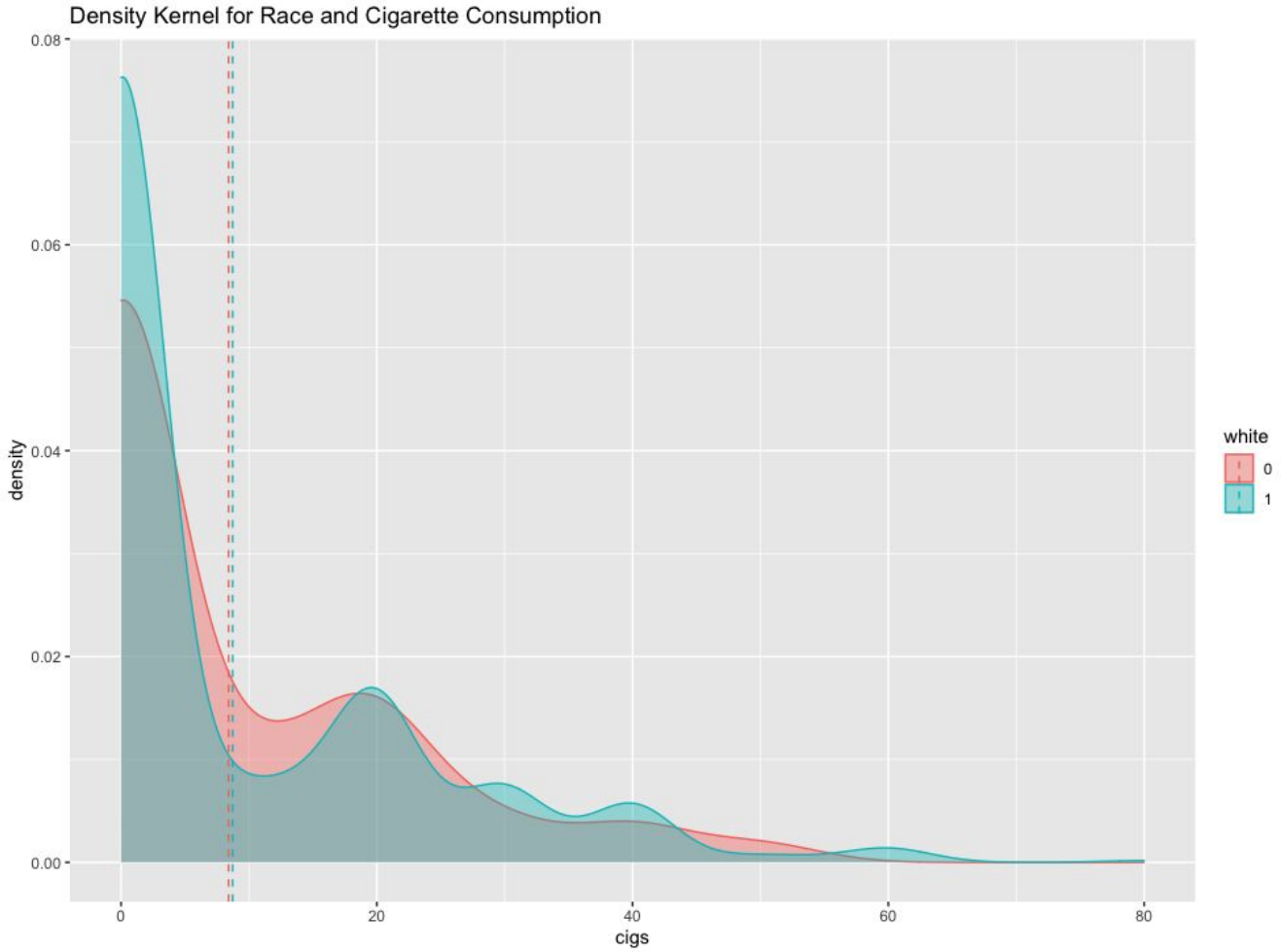
In order to specifically analyze the impact of each element on the number of cigarettes smoking per day, choose “cigs smoked per day” as explained variable (“Y”), “years of schooling”, “annual income,$” and “if white” as explanatory variable.

Data Head:

	educ	white	income	cigs
1	16.0	1	20000	0
2	16.0	1	30000	0
3	12.0	1	30000	3
4	13.5	1	20000	0
5	10.0	1	20000	0
6	6.0	1	6500	0

Source: Wooldridge R Package

DATA


Density Kernel for Race and Cigarette Consumption
density
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
cigs
white
0
1

Empirical Framework

We are going to use OLS technique to simulate the model.

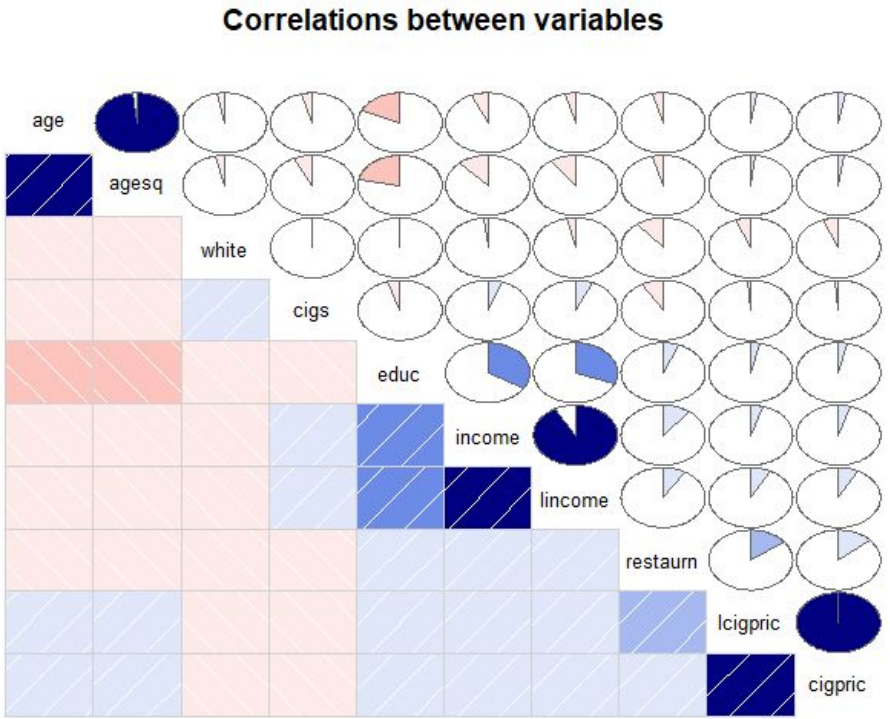
Now we are going to take a look at if those MRM assumptions compatible with our model:

- Zero condition mean
- No Perfect collinearity
- homoskedasticity
- Normality of errors

Empirical Framework

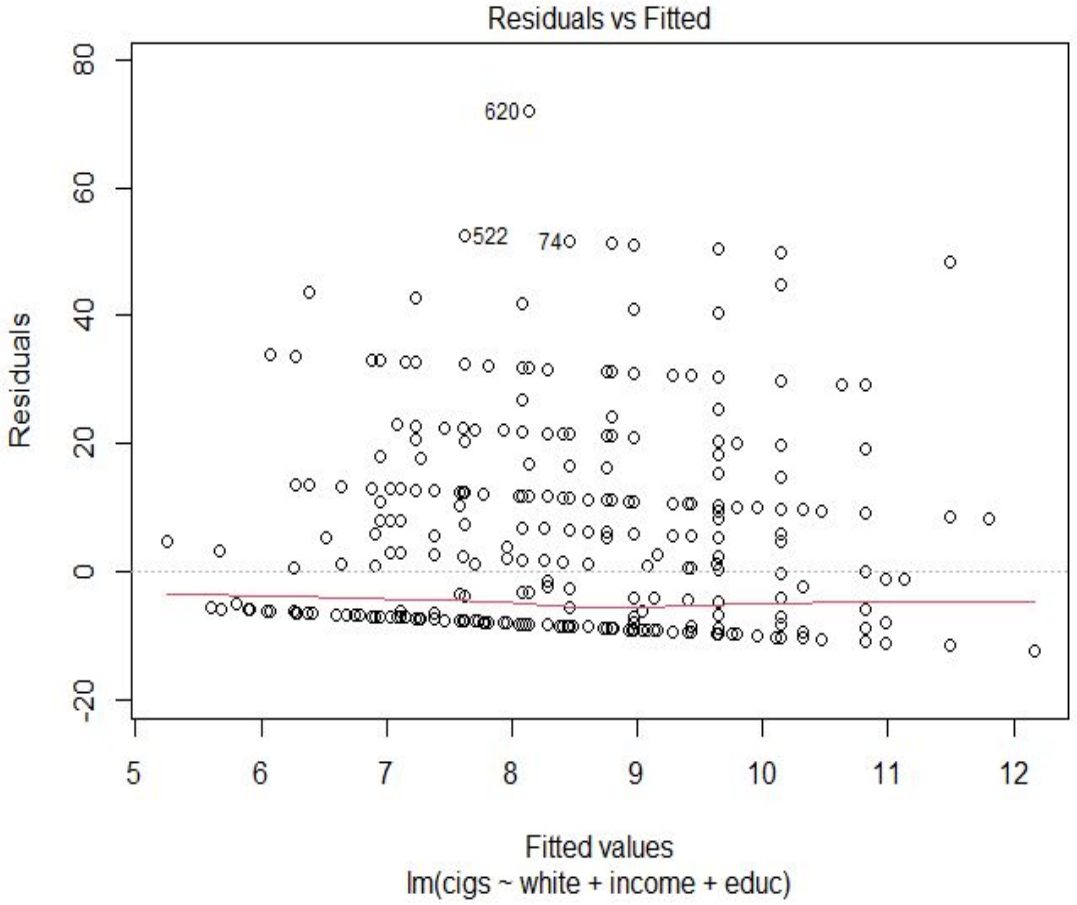
By using “Corrgram” function in R. We acquired the correlations among variables.

There is no such strong correlation among the variables we are currently studying.



Empirical Framework

we see that there are no systematic deviations of the observation from the predicted relation from using this model, Which means the zero condition mean will not be violated.



Empirical Framework

True Model:

$$cigs = \beta_0 + \beta_1 white + \beta_2 income + \beta_3 educ + u$$

Estimated Model:

$$cigs = 10.29 + 0.356 white + 0.001 income - 0.33 educ + u$$

Empirical Framework

Run BP test for the model, acquiring p-value about 0.147 which is not significant enough to reject the null: homoskedasticity, Within even the level of 10%.

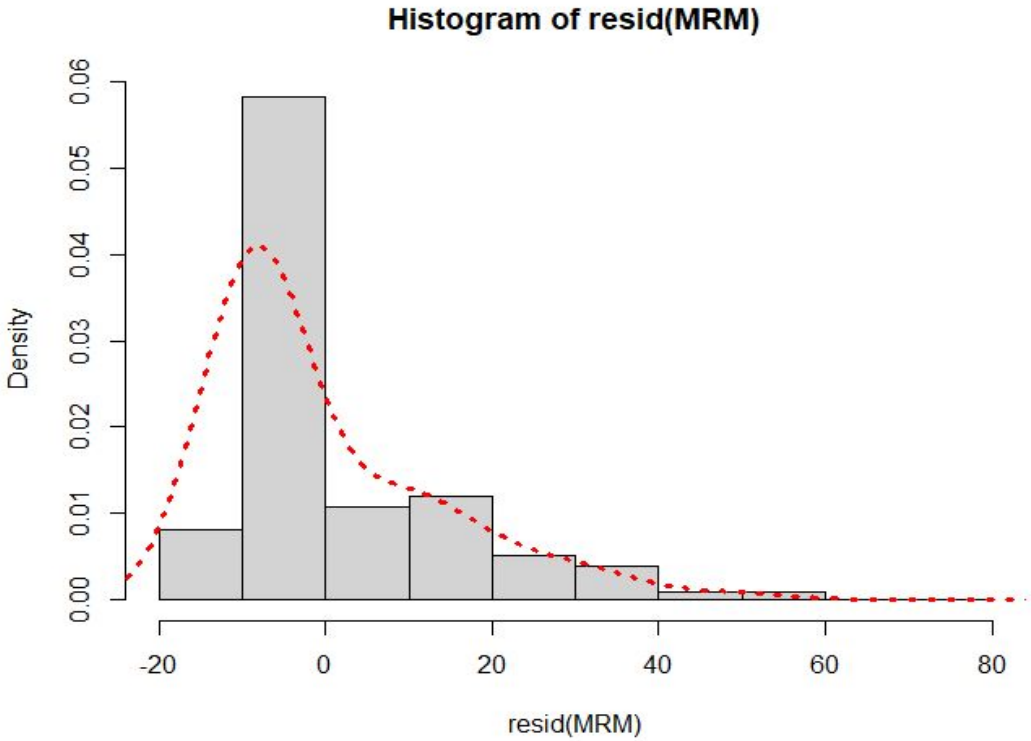
```
> bptest(MRM)

        studentized Breusch-Pagan test

data:  MRM
BP = 5.364, df = 3, p-value = 0.147
```

Empirical framework

The distribution of residual is somehow right-skewed.



	Dependent variable:
	cigs
white	0.356
	(1.476)
income	0.0001**
	(0.0001)
educ	-0.336**
	(0.167)
Constant	10.294***
	(2.441)

Observations	807
R2	0.008
Adjusted R2	0.004
Residual Std. Error	13.69
(df = 803)	
F Statistic	2.126*
(df = 3; 803)	

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

Additional Insights

By adding additional variables, we could increase the R-squared value. Some of these may include:

- Age
- Parent Smoking (Y/N)
- Exercise Frequency
- Log of Income (Quadratic Effect?)

RESULTS

The model can initially pass the test of economic significance, both of the coefficients and symbols have economic significance.

Under the assumption that other variables remain unchanged, the white generally smoke 0.356 more cigarettes per day than non-white people do on average.

Under the assumption that other variables remain unchanged, every $10000 increases in annual income results in 1 more cigarette smoke per day of an individual.

Under the assumption that other variables remain unchanged, 1 more year of schooling decreases cigarettes smoke per day by 0.336.

RESULTS

F-test:

H0:
“white=0,income=0,educ=0”

We get F statistic of 2.12 with p-value of 0.09546

We can reject the null in the level of 10% significance.

```
Hypothesis:
white = 0
income = 0
educ = 0

Model 1: restricted model
Model 2: cigs ~ white + income + educ

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    806 151754
2    803 150558  3    1195.9 2.1262 0.09546 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

RESULTS

T test:

```
> t <- (0.356-1)/1.476
> t
[1] -0.4363144
```

H0: “White=0”

The absolute value of t statistic is quite small which means the variable is insignificant

H0: “income=0”
H0: “educ=0”

From stargazer function we acquiring the significant level of variable income as well as educ in the model.

We can reject the nulls in the level of 5%.

income	0.0001**
	(0.0001)
educ	-0.336**
	(0.167)

Conclusion

From the estimated model there is an increasing effect of the amount of cigarettes smoked per day by white individuals compared to non-whites. While this was the conclusion from the model, we found that the effect is not significant.

Individuals who have higher levels of education, on average, are less likely to smoke more cigarettes per day compared to individuals with less educations- all else held constant.