Compulsory assignment in econometrics with applications

Question 1

For this question use only the sample of smokers (observations with cigs > 0). We are interest in the effect of cigarettes price on cigarettes consumption controlling for all the other observable variables (do not include in the models the same variable in log and level.). Estimate two models, one that gives you an estimate of the marginal effect of cigarettes price on cigarette consumption, and a second one which gives you an estimate of a semi-elasticity. Report and comment the results.

The first model we will be estimating is the following one, which gives an estimate of the marginal effect of cigarettes price on consumption while controlling for all the other variables:

$$cigs_i = \beta + \beta_1 educ_i + \beta_2 cigpric_i + \beta_3 white_i + \beta_4 white_i + \beta_5 age_i + \beta_6 income_i + \beta_7 restaurn_i + \beta_8 agesq_i + \epsilon_i$$

Here is the code we set up in R for the first task:

```
#opgave 1
(smokers <- subset(data, cigs > 0))

m1 <- lm(
  cigs ~ educ + cigpric + white + age + income + restaurn + agesq ,
  data = smokers
);

summary(m1)

m2 <- lm(
  cigs ~ educ + lcigpric + white + age + lincome + restaurn + agesq ,
  data = smokers
);

summary(m2)
```

The first line in the code makes a new dataset with only smokers. Then we run a linear regression of *cigs* on all the other observable variables. Then we make a summary of the regression through R Studio's summary function.

$$\sum\left(\chi_i - \overline{\chi}\right) = \sum \chi_i - n\overline{\chi} = \sum \chi_i - \sum \chi_i = 0$$

While for the second term:

$$\sum \chi_i \left(\chi_i - \overline{\chi}\right) = \sum \chi_i^2 - \overline{\chi}\sum \chi_i = \sum \chi_i^2 - n\left(\overline{\chi}\right)^2 = \sum\left(\chi_i - \overline{\chi}\right)^2$$

Now the final term:

$$\sum \varepsilon_i \chi_i - \overline{\chi}\sum \varepsilon_i$$

Since $E\left(\varepsilon_i\right) = 0$, and this means that the second term is zero. Next, we have that $E\left(\varepsilon_i \chi_i\right) = 0$. Then the expected value of the above term is zero, and our estimator is unbiased. In fact, the OLS estimator is unbiased because the variance of the error term doesn't enter its calculations.
Overall, the consequences of Heteroskedasticity in the OLS estimators are still unbiased and consistent. Because none of the explanatory variables is correlated with the error term. So if the equation is correctly it will give us values of estimated coefficient that is close to the parameters.
Why is the OLS estimators inefficient, that is because the affects the distribution of the estimated coefficients increasing the variances of the distributions. Finally, we have that the underestimates the variance of the estimators, leads to higher values of F and t statistics.

Question 4

Test for heteroskedasticity using a Breusch-Pagan test. Which is the main drawback of this test? Is there any other test that can be used instead?

Here we make a new variable with the residuals of the regressions, and then we run a new regression with the ressiduals as the dependent variable. Here is the code:

summary(data_sub1)	Educ	Cigpric	White	Age
Min	6.00	45.84	1.00	19.00
1st qu	10.00	57.49	1.00	28.00
Median	12.00	60.29	1.00	36.00
Mean	11.97	59.85	1.00	39.26
3rd qu	13.5	63.31	1.00	52.00
Max	16.00	69.90	1.00	71.00
	Income	Cigs	Restaurn	Lincome
Min	6500.00	2.00	0.00	8.78
1st Qu	12500.00	15.00	0.00	9.43
Median	20000.00	20.00	0.00	9.90
Mean	20651.00	23.96	0.00	9.84
3rd Qu	30000.00	30.00	0.00	10.31
Max	30000.00	60.00	0.00	10.31
	Agesq	Lcigpric		
Min	361.00	3.825		
1st	784.00	4.05		
Median	1296.00	4.09		
Mean	1738.00	4.08		
3rd qu	2704.00	4.15		
Max	5041.00	4.25		

Now we run the 3 models again to get the new results, which should be more useable due to no outliers.

M1 with outliers

Residuals:

Min 1Q Median 3Q Max

-27.192 -7.718 -2.355 7.498 48.914

M1 without outliers

Residuals:

Min 1Q Median 3Q Max

-24.027 -7.452 -2.570 6.199 36.837