Problem 8.7 from Wooldridge

October 23, 2020

Econometrics II, Bachelor degree in Economics

Universitat Autònoma de Barcelona

Student, Roylan Martinez Vargas

Professor, Michael Creel

NIU: 1539069

EXAMPLE 8.7 **Demand for Cigarettes**

We use the data in SMOKE to estimate a demand function for daily cigarette consumption. Since most people do not smoke, the dependent variable, *cigs*, is zero for most observations. A linear model is not ideal because it can result in negative predicted values. Nevertheless, we can still learn something about the determinants of cigarette smoking by using a linear model.

The equation estimated by ordinary least squares, with the usual OLS standard errors in parentheses, is

$$\begin{aligned} \widehat{cigs} = {} & \underset{(24.08)}{-3.64} + \underset{(.728)}{.880} \log(income) - \underset{(5.773)}{.751} \log(cigpric) \\ & - \underset{(.167)}{.501}\, educ + \underset{(.160)}{.771}\, age - \underset{(.0017)}{.0090}\, age^2 - \underset{(1.11)}{2.83}\, restaurn \\ & n = 807,\ R^2 = .0526, \end{aligned} \tag{8.35}$$

where

cigs = number of cigarettes smoked per day.
income = annual income.
cigpric = the per-pack price of cigarettes (in cents).
educ = years of schooling.
age = age measured in years.
restaurn = a binary indicator equal to unity if the person resides in a state with restaurant smoking restrictions.

Since we are also going to do weighted least squares, we do not report the heteroskedasticity-robust standard errors for OLS. (Incidentally, 13 out of the 807 fitted values are less than zero; this is less than 2% of the sample and is not a major cause for concern.)

Neither income nor cigarette price is statistically significant in (8.35), and their effects are not practically large. For example, if income increases by 10%, *cigs* is predicted to increase by $(.880/100)(10) = .088$, or less than one-tenth of a cigarette per day. The magnitude of the price effect is similar.

Each year of education reduces the average cigarettes smoked per day by one-half of a cigarette, and the effect is statistically significant. Cigarette smoking is also related to age, in a quadratic fashion. Smoking increases with age up until $age = .771/[2(.009)] \approx 42.83$, and then smoking decreases with age. Both terms in the quadratic are statistically significant. The presence of a restriction on smoking in restaurants decreases cigarette smoking by almost three cigarettes per day, on average.

Do the errors underlying equation (8.35) contain heteroskedasticity? The Breusch-Pagan regression of the squared OLS residuals on the independent variables in (8.35) [see equation (8.14)] produces $R^2_{\hat{u}^2} = .040$. This small R-squared may seem to indicate no heteroskedasticity, but we must remember to compute either the F or LM statistic. If the sample size is large, a seemingly small $R^2_{\hat{u}^2}$ can result in a very strong rejection of homoskedasticity. The LM statistic is $LM = 807(.040) = 32.28$, and this is the outcome of a χ^2_6 random variable. The p-value is less than .000015, which is very strong evidence of heteroskedasticity.

Therefore, we estimate the equation using the feasible GLS procedure based on equation (8.32). The weighted least squares estimates are

$$\begin{aligned}\widehat{cigs} &= \underset{(17.80)}{5.64} + \underset{(.44)}{1.30}\log(income) - \underset{(4.46)}{2.94}\log(cigpric) \\ &\quad - \underset{(.120)}{.463}\,educ + \underset{(.097)}{.482}\,age - \underset{(.0009)}{.0056}\,age^2 - \underset{(.80)}{3.46}\,restaurn \\ n &= 807,\ R^2 = .1134.\end{aligned} \quad [8.36]$$

The income effect is now statistically significant and larger in magnitude. The price effect is also notably bigger, but it is still statistically insignificant. [One reason for this is that *cigpric* varies only across states in the sample, and so there is much less variation in log(*cigpric*) than in log(*income*), *educ*, and *age*.]

The estimates on the other variables have, naturally, changed somewhat, but the basic story is still the same. Cigarette smoking is negatively related to schooling, has a quadratic relationship with *age*, and is negatively affected by restaurant smoking restrictions.

[39]:
```
# TOOLS
import numpy as np
import pandas as pd
from statsmodels.formula.api import ols, wls
from statsmodels.compat import lzip
import statsmodels.stats.api as sms
```

[40]:
```
# Data
dataExcel = pd.read_csv("smoke.csv")
dataExcel
```

```
[40]:      educ  cigpric  white  age  income  cigs  restaurn    lincome  agesq  \
     0    16.0   60.506      1   46   20000     0         0   9.903487   2116
     1    16.0   57.883      1   40   30000     0         0  10.308950   1600
     2    12.0   57.664      1   58   30000     3         0  10.308950   3364
     3    13.5   57.883      1   30   20000     0         0   9.903487    900
     4    10.0   58.320      1   17   20000     0         0   9.903487    289
     ..    …       …      …   …     …     …         …          …
     802  18.0   61.818      0   52   30000    20         0  10.308950   2704
     803  18.0   61.676      1   31   12500     0         0   9.433484    961
     804  16.0   60.707      1   30   20000     0         0   9.903487    900
     805  10.0   59.988      1   18   20000     0         0   9.903487    324
     806  10.0   59.652      1   47   30000    20         0  10.308950   2209

          lcigpric
     0    4.102743
     1    4.058424
     2    4.054633
     3    4.058424
     4    4.065945
     ..        …
     802  4.124195
     803  4.121895
     804  4.106059
     805  4.094144
     806  4.088528

     [807 rows x 10 columns]
```

First step: Compute the OLS model (model A) and perform a Bresuch-Pagan test to check if there is homoskedasticity

```
[41]: # Initial model
modelA = ols("cigs ~ np.log(income) + np.log(cigpric) + educ + age + agesq +␣
 ↪restaurn", data = dataExcel).fit()
print(modelA.summary())

#perform Bresuch-Pagan test
names = ['Lagrange multiplier statistic', 'p-value', 'f-value', 'f p-value']
test = sms.het_breuschpagan(modelC.resid, modelC.model.exog)
print('Bresuch - Pagan test')
for x in list(zip(names, test)):
    print(f'{x[0]} : {x[1]}')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   cigs   R-squared:                       0.053
Model:                            OLS   Adj. R-squared:                  0.046
```

```
Method:                 Least Squares   F-statistic:                     7.423
Date:                Fri, 23 Oct 2020   Prob (F-statistic):           9.50e-08
Time:                        20:01:31   Log-Likelihood:                -3236.2
No. Observations:                 807   AIC:                             6486.
Df Residuals:                     800   BIC:                             6519.
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
===
                      coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
---
Intercept          -3.6398     24.079     -0.151      0.880     -50.905
43.625
np.log(income)      0.8803      0.728      1.210      0.227      -0.548
2.309
np.log(cigpric)    -0.7509      5.773     -0.130      0.897     -12.084
10.582
educ               -0.5015      0.167     -3.002      0.003      -0.829
-0.174
age                 0.7707      0.160      4.813      0.000       0.456
1.085
agesq              -0.0090      0.002     -5.176      0.000      -0.012
-0.006
restaurn           -2.8251      1.112     -2.541      0.011      -5.007
-0.643
==============================================================================
Omnibus:                      225.317   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              494.255
Skew:                           1.536   Prob(JB):                    4.72e-108
Kurtosis:                       5.294   Cond. No.                     1.33e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.33e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
Bresuch - Pagan test
Lagrange multiplier statistic : 30.353321951457545
p-value : 3.367162727883872e-05
f-value : 5.2110048338798425
f p-value : 2.833124416308335e-05
```

Besides of the not significant p-values of the model on some of the regressors, I performed a Bresuch Pagan test for heterokedasticity **(null hypothesis: homoskedastic variance, alternative hypothesis: heteroskedastic variance)** and as we see the p value is **3.367162727883872e-05** therefore we shall reject the homoskedastic variance hypothesis with a **99.99 % confidence level.**

Second step: We must compute, log of the squared errors of the model A, create a new model (model B) with the transform errors of the model A as the dependent variable, save the predicted values of the model B and finally compute H.

[42]:
```
# Get residuals of the model A, square them and log them
dataExcel['logu2'] = np.log(modelA.resid ** 2)

# New model with log of u squared as the dependent variable
modelB = ols("logu2 ~ lincome + lcigpric + educ + age + agesq + restaurn", data␣
 ↪= dataExcel).fit()

# New variable with the fitted values of the model B
dataExcel['h'] = np.e ** modelB.predict(dataExcel.loc[:, ['lincome',␣
 ↪'lcigpric', 'educ', 'age', 'agesq', 'restaurn']])
```

Third step: Run a Weighted Least Squares model with 1 / h as the weight

[43]:
```
# WLS model using 1 / h as the weight
modelC = wls("cigs ~ lincome + lcigpric + educ + age + agesq + restaurn", data␣
 ↪= dataExcel, weights = (1 / dataExcel['h'])).fit(cov_type='HC1')
print(modelC.summary())
```

```
                            WLS Regression Results
==============================================================================
Dep. Variable:                   cigs   R-squared:                       0.113
Model:                            WLS   Adj. R-squared:                  0.107
Method:                 Least Squares   F-statistic:                     23.56
Date:                Fri, 23 Oct 2020   Prob (F-statistic):           1.01e-25
Time:                        20:01:32   Log-Likelihood:                -3207.8
No. Observations:                 807   AIC:                             6430.
Df Residuals:                     800   BIC:                             6462.
Df Model:                           6
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.6353     37.323      0.151      0.880     -67.517      78.788
lincome        1.2952      0.535      2.421      0.015       0.246       2.344
lcigpric      -2.9403      8.970     -0.328      0.743     -20.522      14.641
educ          -0.4634      0.149     -3.109      0.002      -0.756      -0.171
age            0.4819      0.115      4.191      0.000       0.257       0.707
agesq         -0.0056      0.001     -4.781      0.000      -0.008      -0.003
restaurn      -3.4611      0.716     -4.835      0.000      -4.864      -2.058
==============================================================================
Omnibus:                      325.055   Durbin-Watson:                   2.050
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1258.137
Skew:                           1.908   Prob(JB):                    6.29e-274
Kurtosis:                       7.780   Cond. No.                     2.30e+05
==============================================================================

Warnings:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 2.3e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

As we see, we obtain the same regressor coeficcients and the same adjusted R-squared as in the example given by the book. To build this model I used the same **weight vector** and the same covariance type **HC1**