ECONOMETRICS 1 / APPLIED PROBLEM SET 1

Topic: Introduction to Multivariate Regression

- This class is an introduction to the **menu-driven features** of EViews 6. The data is available on the website as `cigarette.txt`.
- The problem set covers (i) file import and workfile save, (ii) statistical analysis, (iii) creation of groups, (iv) ordinary least squares, (v) (robust) inference, and (vi) diagnostic testing.
- We use data on (i) the log of cigarette consumption (in packs) per person of smoking age (> 16 years) for 46 U.S. states in 1992: LNC ($\ln(C)$), (ii) the log real price of cigarettes in each state, normalized at 1983\$ per pack: LNP ($\ln(P)$), and (iii) the log of real disposable income per capita in each state, in 1983\$1000: LNY ($\ln(Y)$).
- Perform all of the steps described in Figures **1-85** (answering any corresponding questions, and carefully considering the methods and output, noting any new or confusing tools), and end by responding in full to the following questions:

1. Regress log consumption on log prices **(eq01)**:

$$\widehat{\ln(C_i)} \approx \boxed{[\ 5.09\ ]} + \boxed{[\ -1.20\ ]}\ln(P_i), \qquad y_i = x_i'\beta + u_i,\quad \hat{y}_i = x_i'\hat{\beta},\quad \hat{u}_i = y_i - \hat{y}_i$$

with estimated residuals $\widehat{u}_i^{(1)} \approx \ln(C_i) + \boxed{[\ -5.09\ ]} + \boxed{[\ 1.20\ ]}\ln(P_i)$.

What is the estimated price elasticity of consumption? Answer:

$$\frac{/\Delta \ln C}{/\Delta \ln P} \simeq \frac{100\ \partial \hat{C}/C}{100\ \partial P/P} = \frac{\partial \widehat{\ln(C_i)}}{\partial \ln(P_i)} \approx \boxed{[\ -1.20\ ]} \quad ; \quad \frac{\partial \ln C}{\partial C} = \frac{1}{C}$$

and so a [10%] increase/~~decrease~~ in price results in a [12%] ~~increase~~/decrease in consumption.

2. Regress log consumption on log prices and log income **(eq02)**: no "satiation" ie. Y↑↑ ⇒ C↑↑

$$\widehat{\ln(C_i)} \approx [\,4.30\,] + [\,-1.34\,]\ln(P_i) + [\,0.17\,]\ln(Y_i).$$

3. Regress log income on log prices **(eq03)**: P↑ 10% C↓13% Y↑ 10% C↑2% (insignificant)

$$\widehat{\ln(Y_i)} \approx [\,4.61\,] + [\,0.81\,]\ln(P_i),$$

Y ⊥ P

with estimated residuals $\widehat{u}_i^{(3)} \approx \ln(Y_i) + [\,-4.61\,] + [\,-0.81\,]\ln(P_i)$.

4. Regress log consumption on $\widehat{u}_i^{(3)}$ **(eq04)**:

$$\widehat{\ln(C_i)} \approx [\,4.85\,] + [\,0.17\,]\widehat{u}_i^{(3)}$$

Y ⊥ P

C ⊥ P ↓

5. Regress $\widehat{u}_i^{(1)}$ on $\widehat{u}_i^{(3)}$ **(eq05)**:

$$\widehat{\widehat{u}_i^{(1)}} \approx [\,1.2\times10^{-15}\,] + [\,0.17\,]\widehat{u}_i^{(3)}.$$

6. Regress log consumption on log income, log income squared, and log prices **(eq06)**:

$$\widehat{\ln(C_i)} \approx [\,-34.77\,] + [\,16.43\,]\ln(Y_i) + [\,-1.69\,](\ln(Y_i))^2 + [\,-1.35\,]\ln(P_i),$$

(rounding)

so that a [10%] increase/~~decrease~~ in price results in a [13%] ~~increase~~/decrease in consumption.

~~crease~~ decrease in consumption. The income elasticity of consumption is now:

$$\widehat{\frac{\partial \ln(C_i)}{\partial \ln(Y_i)}} \approx [\ 16.43\ ] + [\ -3.38\ ] \ln(Y_i),$$

see histogram of $\ln Y$. very few observations are < 4.57

$\Rightarrow \frac{\partial \ln C}{\partial \ln Y} \leq 1$

which is greater than (equal to) less than 1 as $\ln(Y_i)$ is *less than (equal to) greater than* [4.57] (as Y_i is *less than (equal to) greater than* \$[96.147]): use the full available accuracy on the estimated coefficients when performing this computation. Interpret your findings carefully.

7. Compare equations eq02, eq04 and eq05. In eq02, [0.17] quantifies the impact of log income on log consumption. In eq04, $\widehat{u}_i^{(3)}$ is the part of **log income not explained by log price**, and so [0.17] quantifies the impact on log consumption of that part of log income not explained by log price. In eq05, $\widehat{u}_i^{(1)}$ is the part of **log consumption not explained by log price**, and so [0.17] quantifies the impact on that part of log consumption not explained by log price, of that part of log income not explained by log price. Carefully explain the intuition behind these results. — each model "controls for" P in a different way $\Rightarrow$ same result for Y coefficient

eq02 $C \sim P, \underline{Y}$

eq04 $C \sim \underline{Y}_{\perp} P$

eq05 $C_{\perp P} \sim \underline{Y}_{\perp P}$

- Special attention should be paid to observations 3 (Arkansas), 15 (Kentucky) and 40 (Utah): Arkansas and Kentucky have particularly high sales, Kentucky is a producer with rather low prices, and Utah has especially low sales because of its high Mormon population (which bans smoking). It is important to build a deep understanding of the structure and peculiarities of your data before you start modelling.

data $i = 1, 2, \dots, n$

$x_1, x_2, \dots, x_n$

ordered data

$x_{(1)} \leq \dots \leq x_{(n)}$

minimum

$x_{(1)}$

maximum

$x_{(n)}$

mean

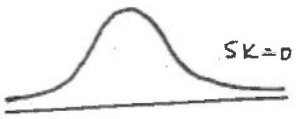
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

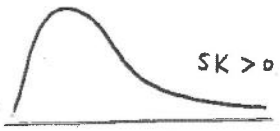
standard dev

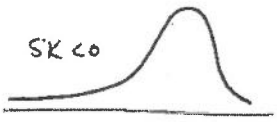
$$SD(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

skewness

$$SK = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

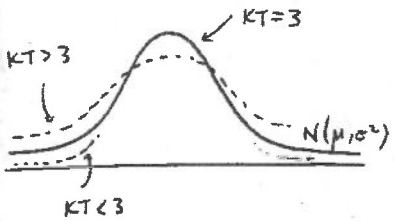


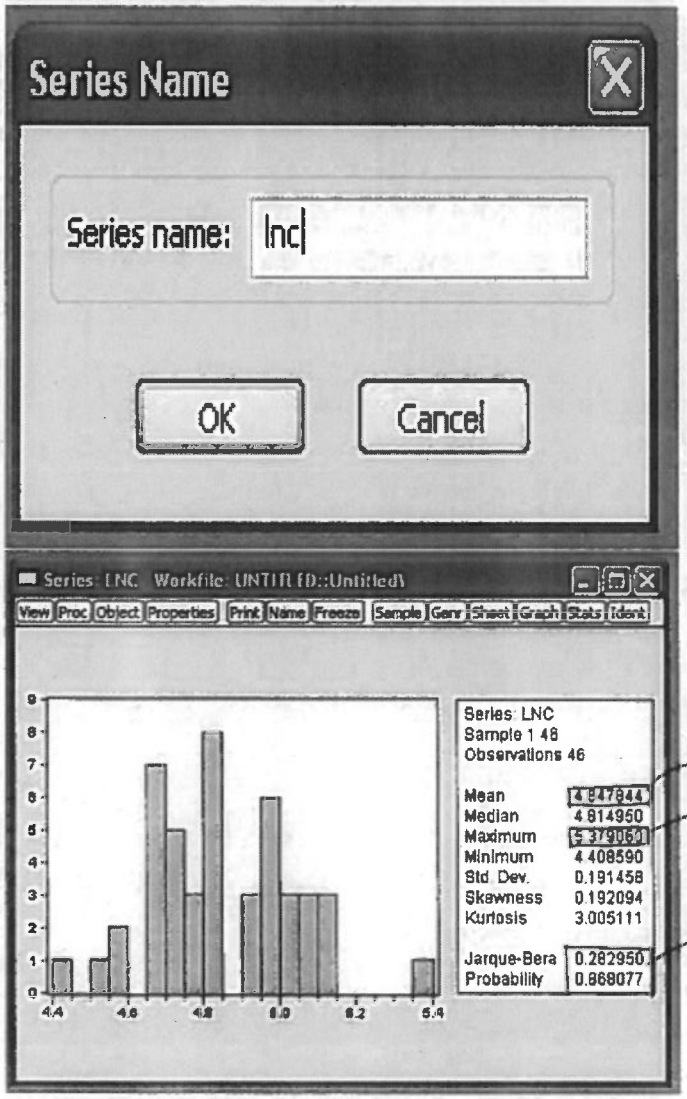




kurtosis

$$KT = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$





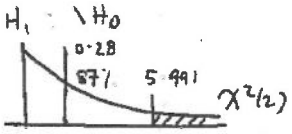
127 packs / person

217 packs / person

$$JB = \frac{n}{6}\left[SK^2 + \frac{(KT-3)^2}{4}\right] \sim \chi^2_{(2)}$$

H_0: $SK = 0$, $KT = 3$

H_1: $\setminus H_0$



d n r H_0

("normality")

Figure 8: Enter the series name: lnc. The histogram of the data also gives some basic descriptive statistics, up to the standardized fourth central moment (kurtosis), and the Jarque-Bera test for normality of the data: this performs the joint test that the skewness = 0 and the kurtosis = 3.

generally, if "Prob" > 0.05,

do not reject (d n r) H_0 at 95% (etc.)

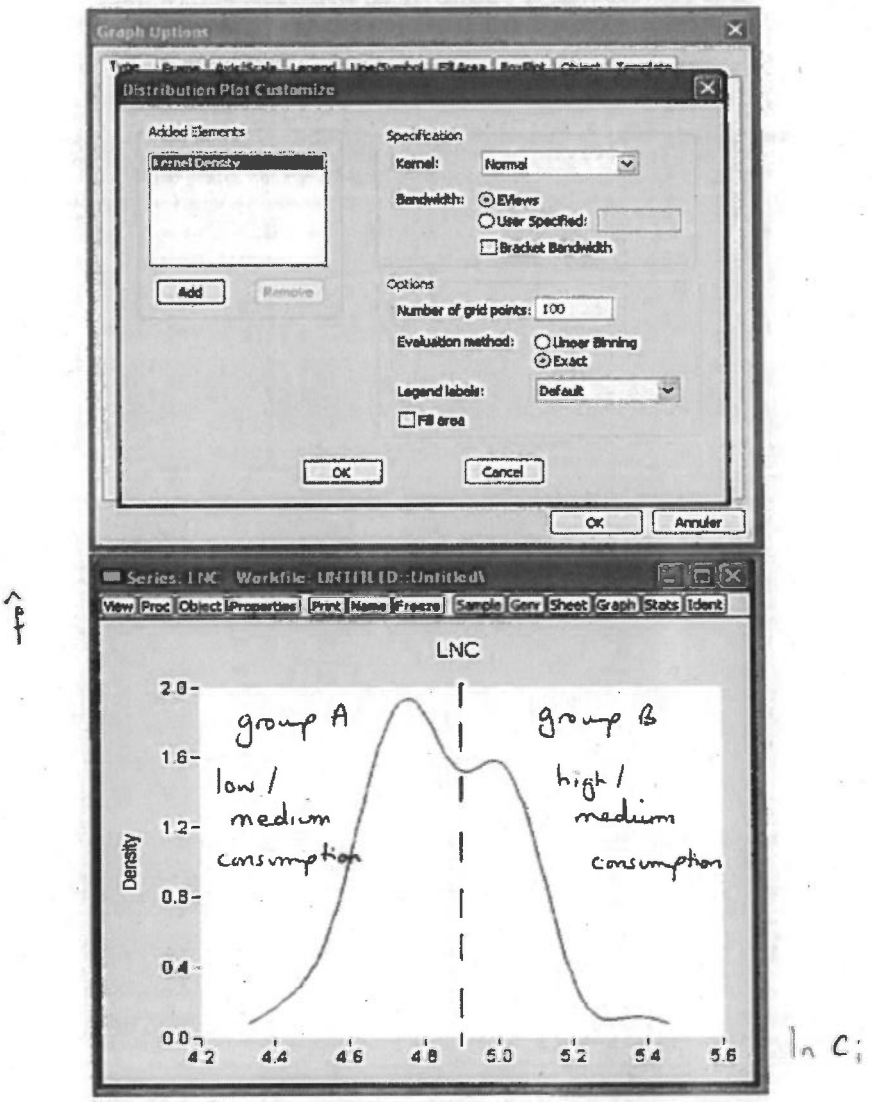
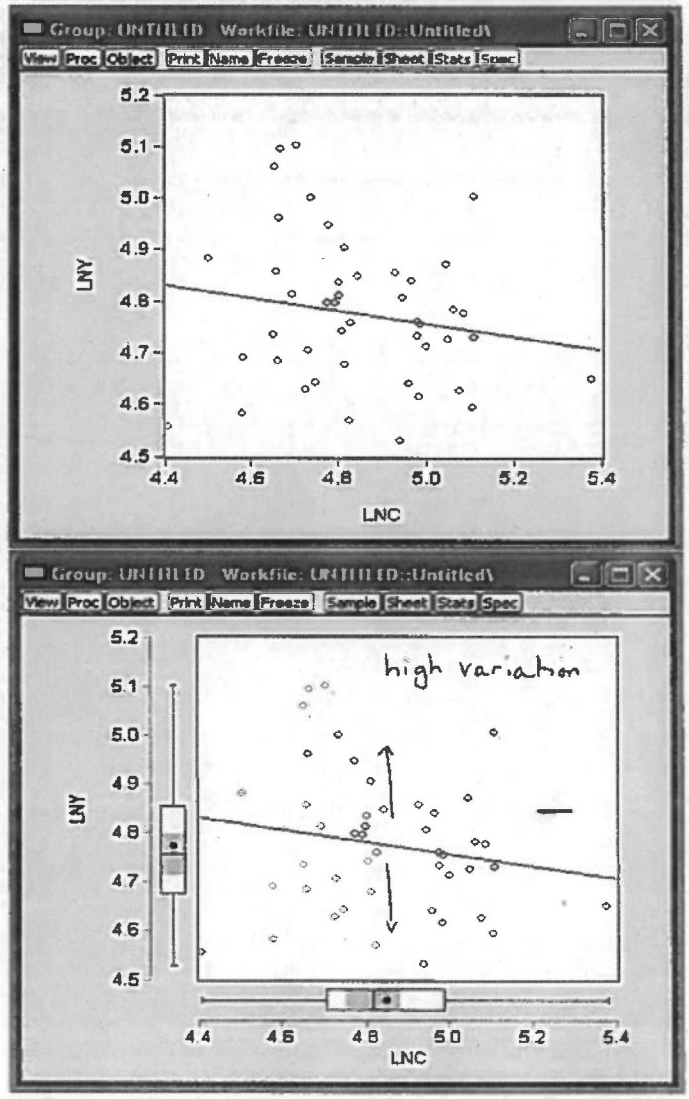


Figure 28: Select 'Exact' instead of 'Linear Binning', to give the Gaussian kernel density plot of the data: $\widehat{f}(x) = (nh)^{-1} \sum_i K((x - X_i)h^{-1})$, where n is the sample size, h is the kernel bandwidth (chosen automatically by EViews), and $K()$ is the kernel (here, the N(0,1) pdf). **Compare this to the histogram of the data that was plotted earlier.**



ln Y

suggests that
Y ↑ ⇒ C ↓
(strange?!)

ln C

Figure 40: Scatter plot of lny against lnc, with ordinary least squares (OLS) fit from a regression of lny on a constant and lnc. Plot the same figure with boxplots of lny and lnc on the axes!

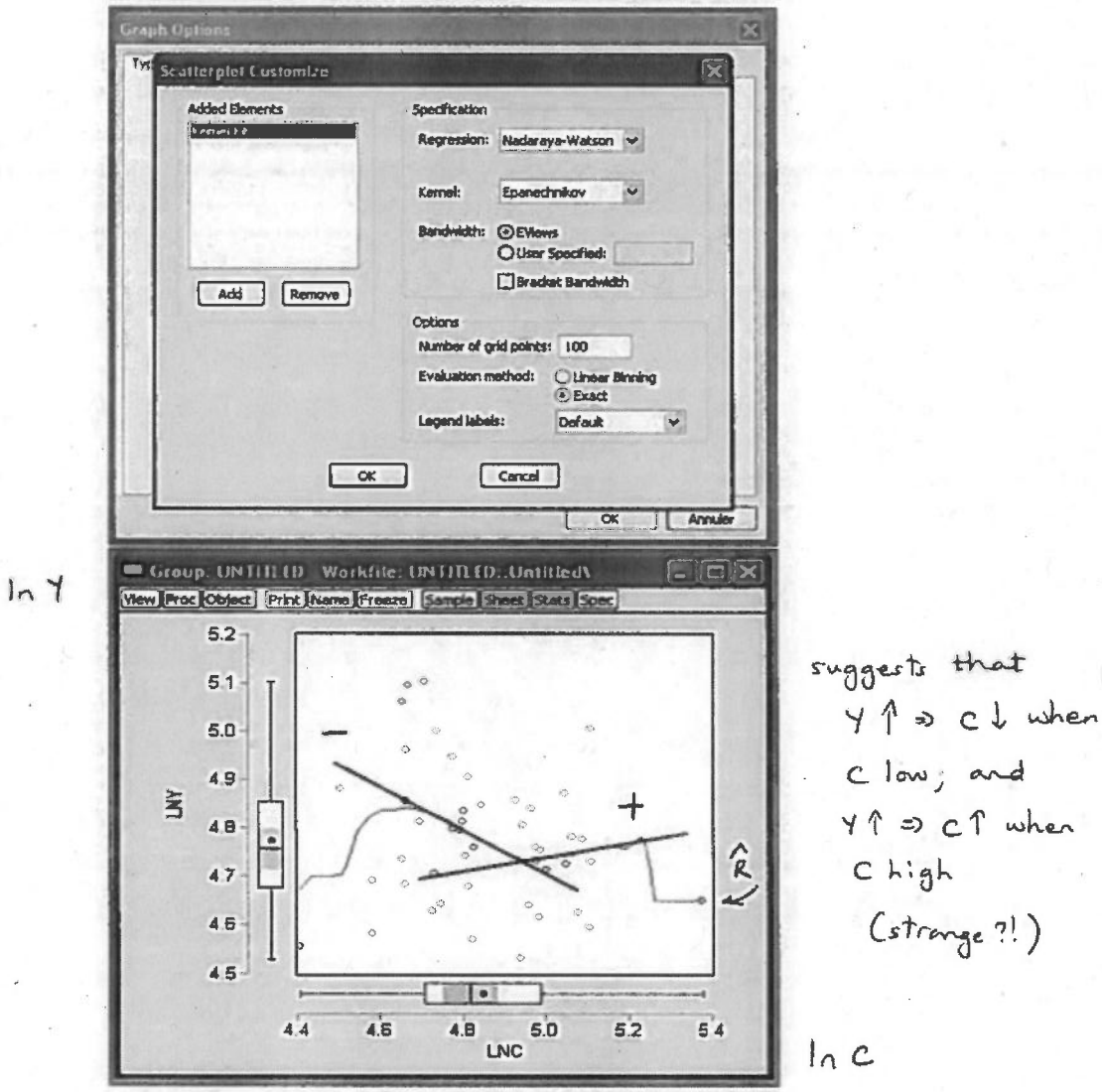
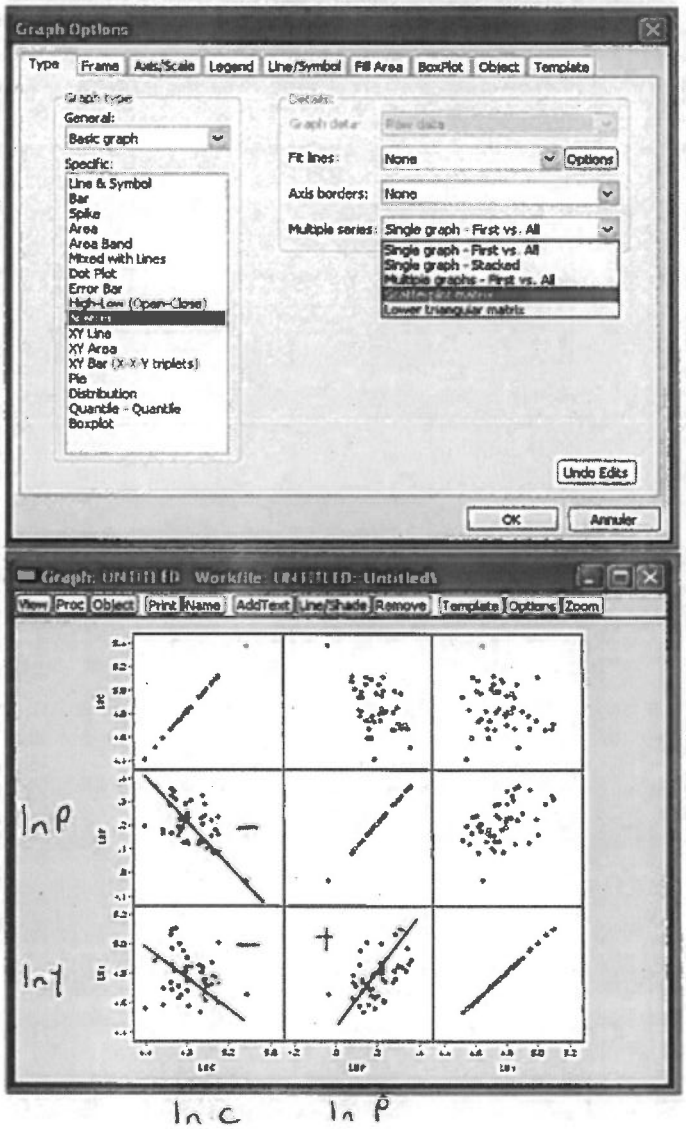


Figure 42: Choose 'Nadaraya-Watson' regression, with 'Epanechnikov' kernel, 'Exact' rather than 'Linear Binning', and bandwidth chosen by 'Eviews', to give the Nadaraya-Watson kernel regression fit, superimposed on the scatter plot of lny against lnc. The Nadaraya-Watson kernel estimator of Y_i on X_i is given by $\widehat{R}(x) = \arg\min_\psi \sum_i (Y_i - \psi)^2 K((x - X_i)/h)$, where ψ is a locally fit constant, and $K(u) = (3/4)(1 - u^2)$ on $[-1, 1]$ is the Epanechnikov kernel.



mechanism :

$Y \uparrow \Rightarrow P \uparrow \Rightarrow C \downarrow$

price discrimination — demand curve

if P not included, suggests that $Y \uparrow \Rightarrow C \downarrow$; we need to include P as a control variable

Figure 49: Plot a matrix scatterplot of lnc, lnp and lny, and name the graph graph01.

(even if we are only interested in Y)

$y = X\beta + u$ (n×k)

$y_i = x_i'\beta + u_i$

$i = 1, 2, \ldots, n$

$n = 46$

$k = 2$

$y_i = \ln c_i$

$x_i = \begin{pmatrix} 1 \\ \ln p_i \end{pmatrix}$

$\hat{\beta}_{k\times 1} = (X'X)^{-1}X'y$

$\hat{Var}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$

$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n-k}$

$\hat{Var}(\hat{\beta}) = \begin{pmatrix} \hat{Cov}(\hat{\beta}_1,\hat{\beta}_1) & \hat{Cov}(\hat{\beta}_1,\hat{\beta}_2) \\ \hat{Cov}(\hat{\beta}_1,\hat{\beta}_2) & \hat{Cov}(\hat{\beta}_2,\hat{\beta}_2) \end{pmatrix} = \begin{pmatrix} \hat{Var}(\hat{\beta}_1) & \hat{Cov}(\hat{\beta}_1,\hat{\beta}_2) \\ \hat{Cov}(\hat{\beta}_1,\hat{\beta}_2) & \hat{Var}(\hat{\beta}_2) \end{pmatrix}$

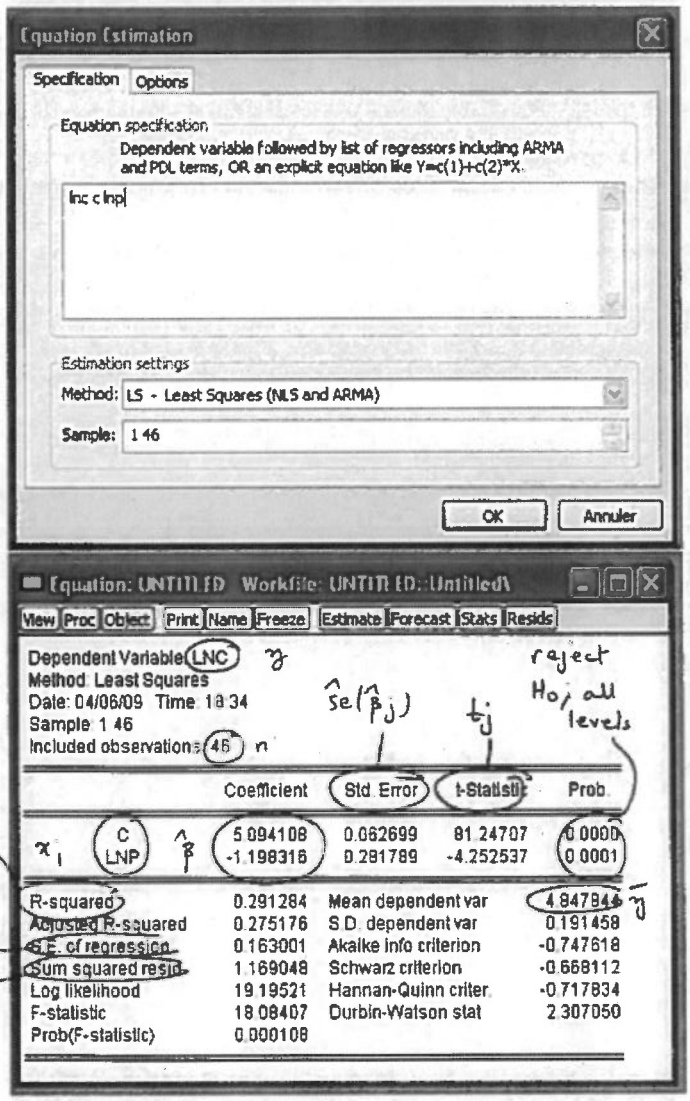
$\hat{se}(\hat{\beta}_j) = \sqrt{\hat{Var}(\hat{\beta}_j)}$

$t_j = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} \sim t(n-k)$

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

(test of individual significance)



$\hat{se}(\hat{\beta}_j)$, t_j, reject H_0 all levels

y, $\bar{y}$

R^2 measures quality of linear fit (0% - 100%)

but need to include constant in model for this to work

here, $R^2 \approx 29\%$ ("good" R^2 depends on context)

$\hat{\sigma}$

$\hat{u}'\hat{u}$

Figure 54: **(eq01)** Enter 'lnc c lnp' for a regression of log consumption on a constant and log price. Observe that the regression output gives estimated coefficients (by ordinary least squares), standard errors, probabilities for individual tests of significance, and various statistics, including coefficients of determination, the sum of squared residuals, the Durbin-Watson statistic for autocorrelation, the Akaike and Schwarz Information Criteria, and the F statistic for significance of the entire regression (*we will return to these statistics later in the course*).

R^2 cannot ↓ as k ↑, even if irrelevant variables added

adjusted R^2 (denoted $\bar{R}^2$) controls for k: $\bar{R}^2 = 1 - \left(\frac{n-1}{n-k}\right)(1-R^2)$

⇒ $\bar{R}^2$ must be used to compare models with different k

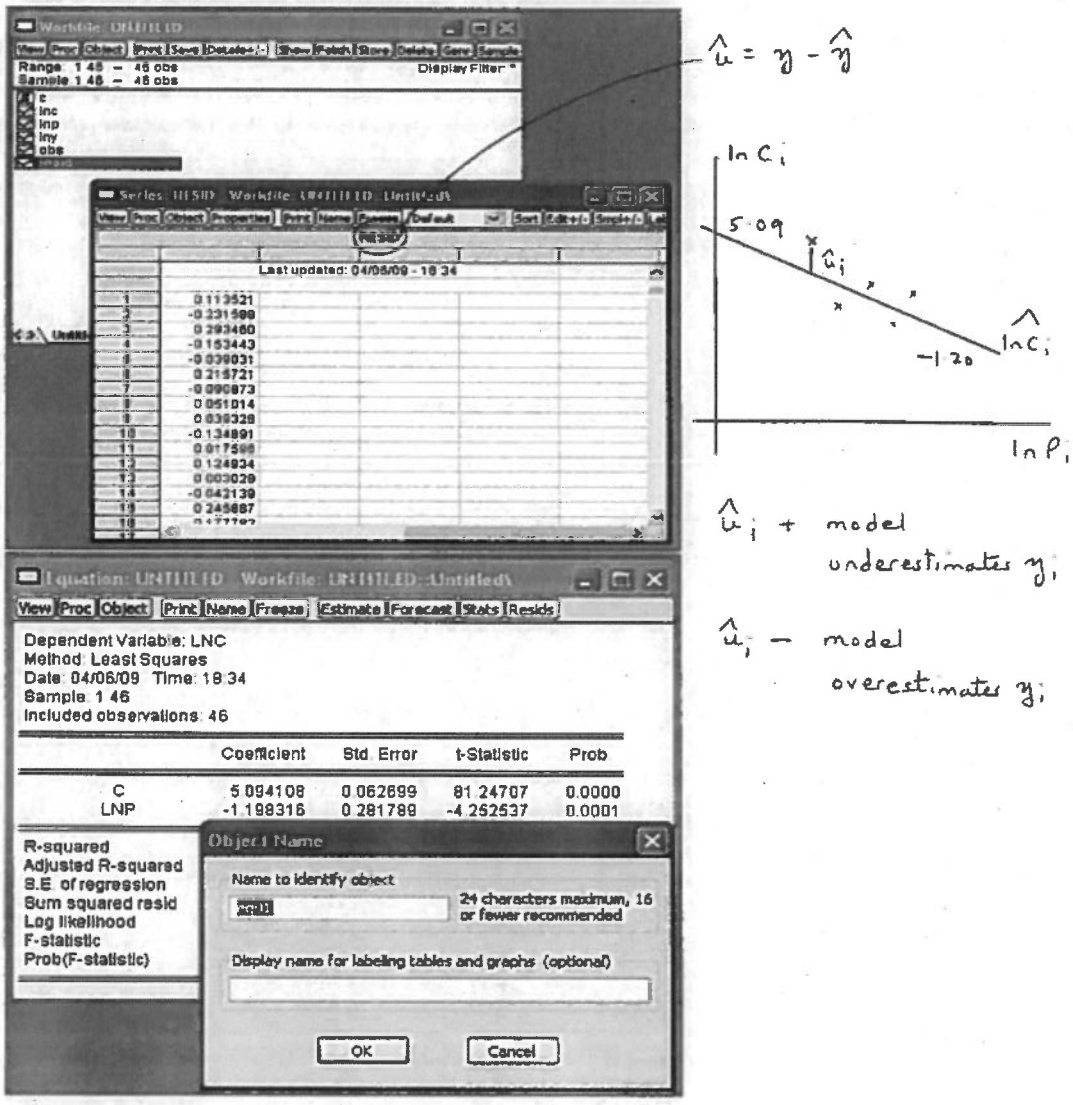


Figure 56: Check that the estimated residuals have been recorded in the workfile object resid: **note that the estimated residuals are overwritten each time that a new regression is performed.** Name the equation 'eq01'.

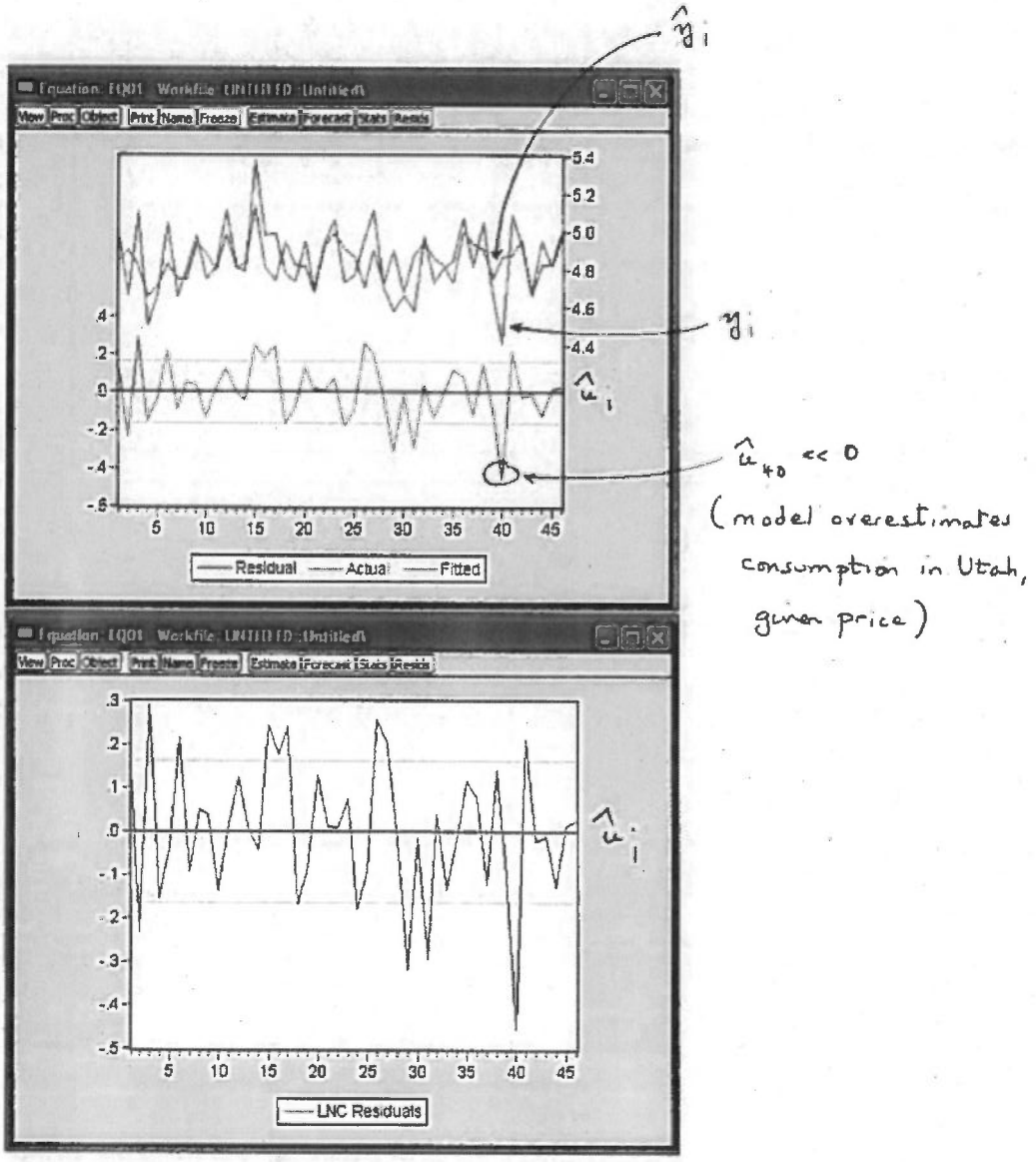


Figure 59: Several representations of the fitted residuals $\widehat{u}$.

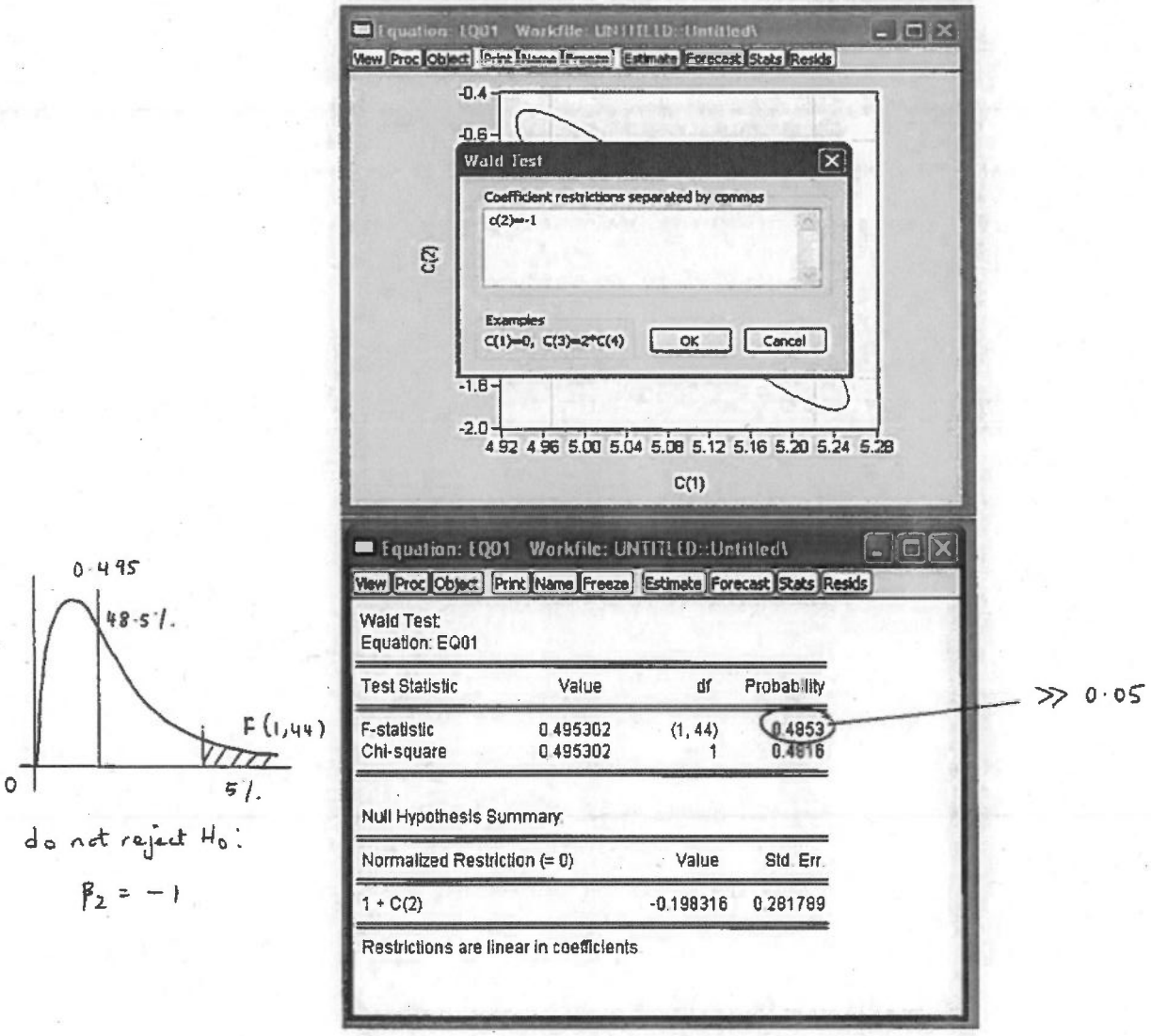


Figure 63: Enter 'c(2)=-1' to perform a Wald test of $H_0 : \beta_2 = -1$. Observe that both exact (F) and asymptotic (χ^2) statistics are reported: the result of the test is that we do not reject the null at the 90% level (say). **Why is an F test reported, not a t test?**

$F(1, n-k) = [t(n-k)]^2$

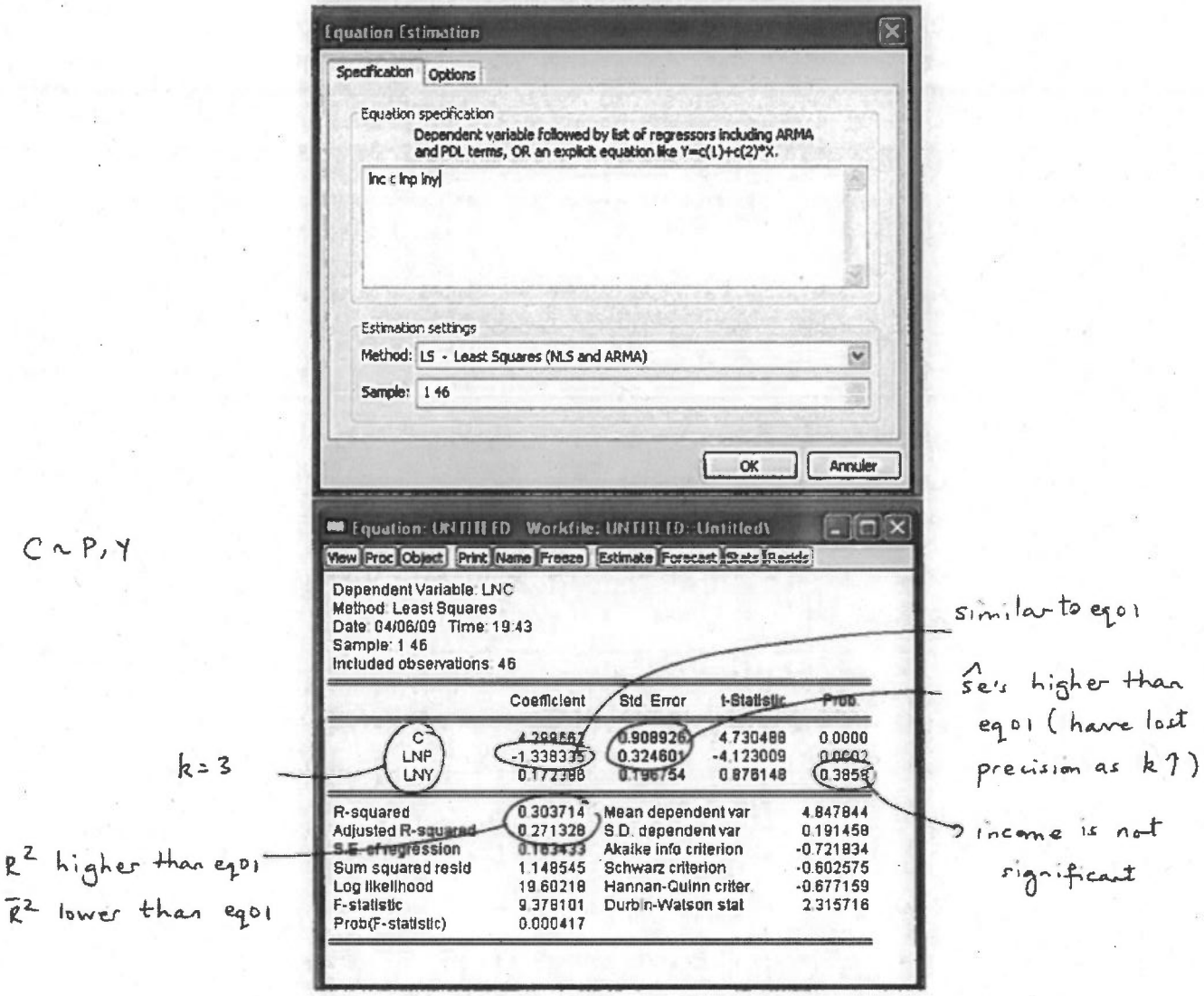
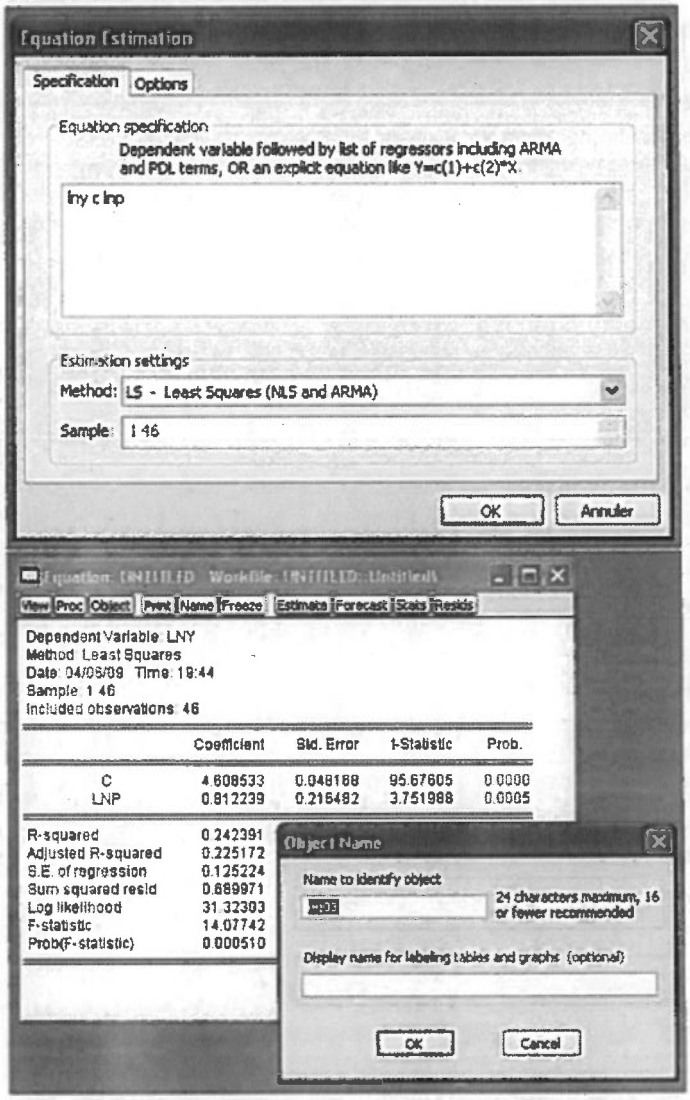


Figure 72: **(eq02)** Run the regression of log consumption on a constant, log price, and log income, using ordinary least squares, and consider the regression output.



Y ~ P

Correlation

≠ causality

Figure 74: (eq03) Regress log income on a constant and log price, using ordinary least squares, and consider the regression output. Name this equation 'eq03'.

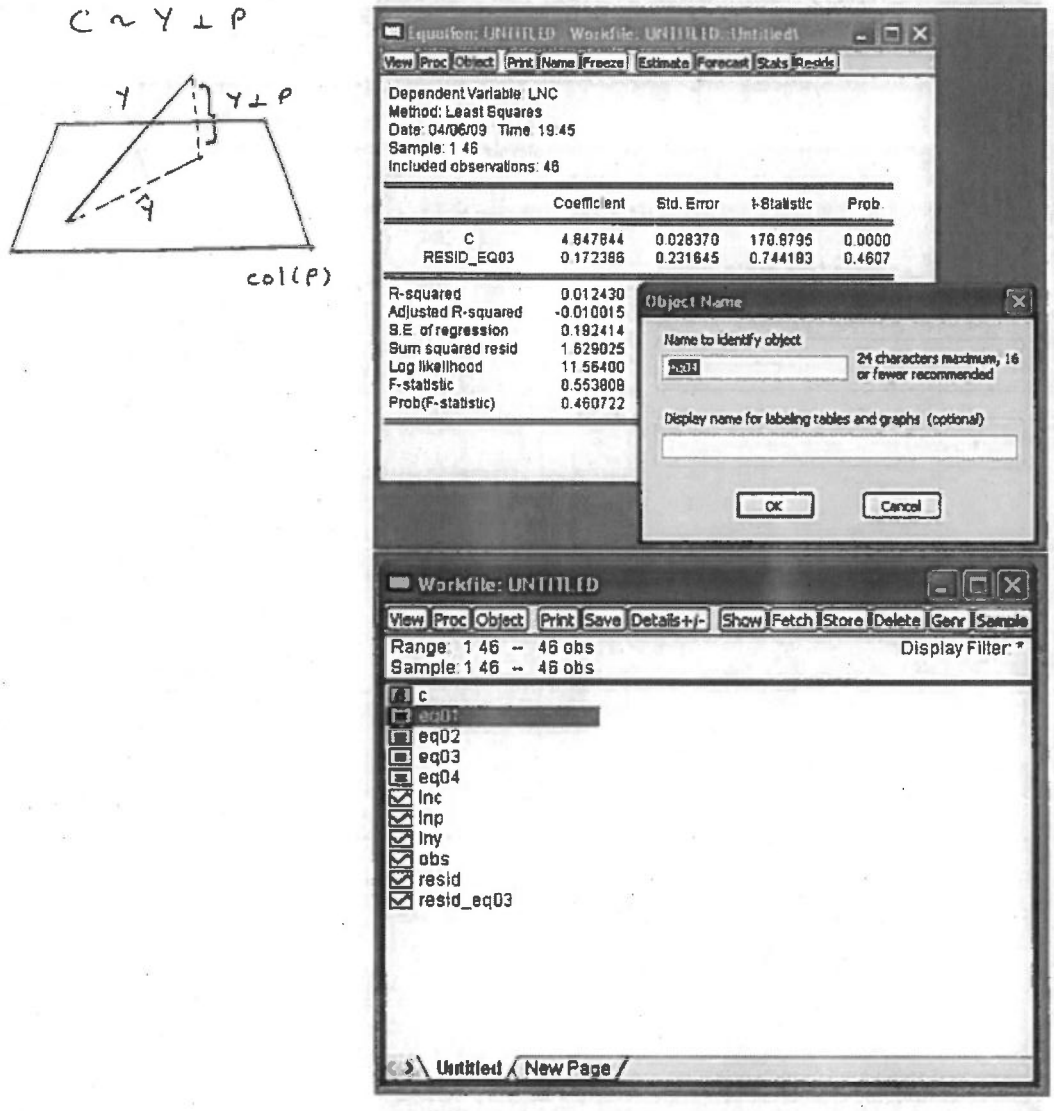
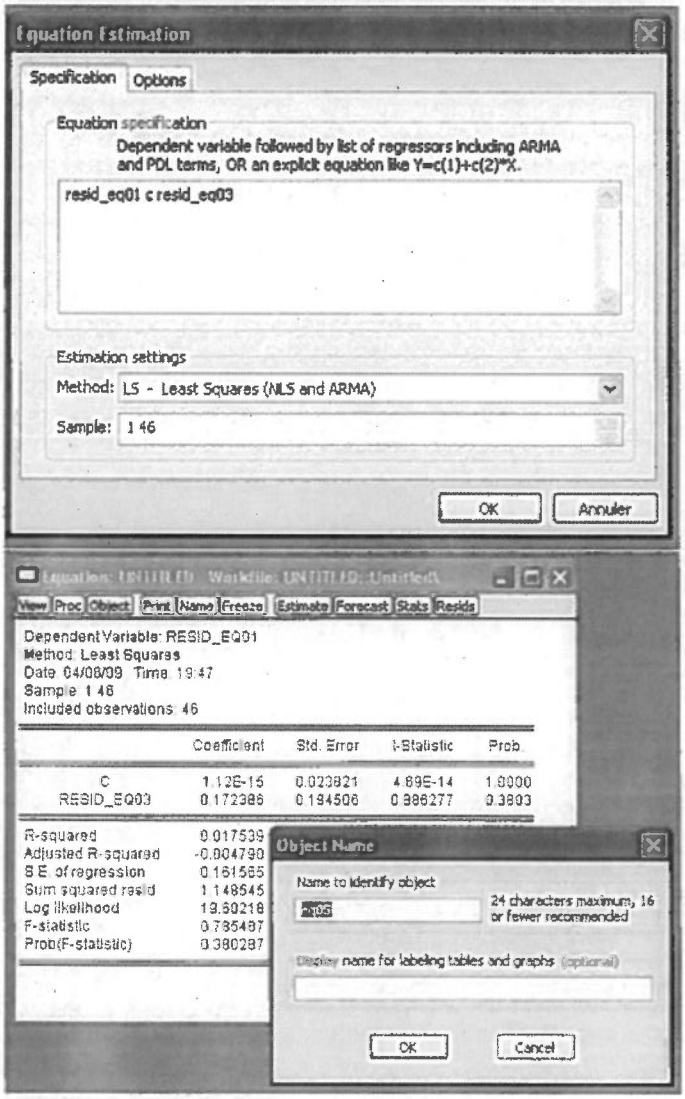


Figure 77: Name this equation 'eq04', and then select the eq01 object.



$C \perp P \sim$

$Y \perp P$

Figure 80: **(eq05)** Run the regression of resid_eq01 on a constant and resid_eq03, and name this equation 'eq05'.

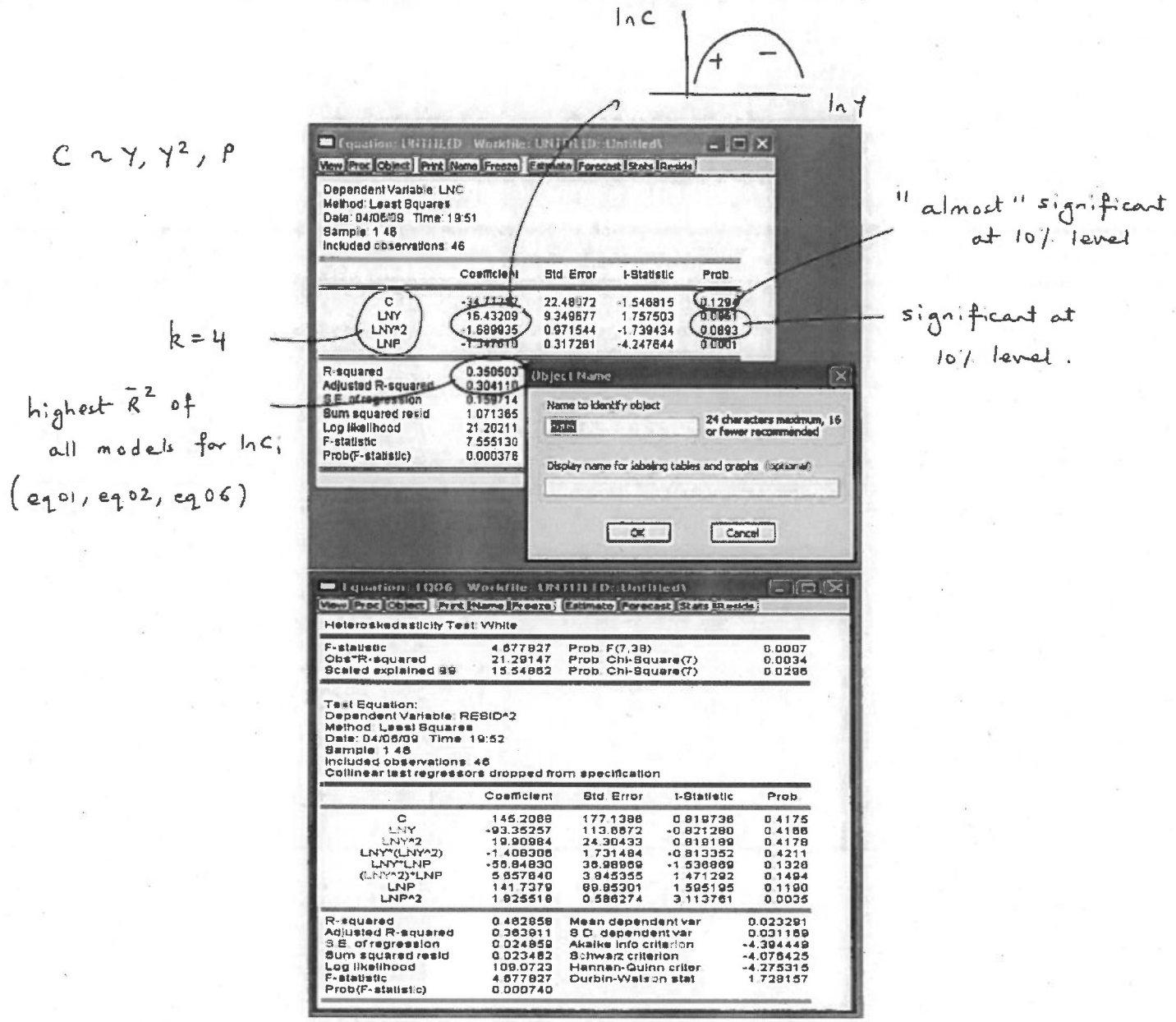


Figure 84: **(eq06)** Perform the regression of log consumption on a constant, log income, log income squared, and log price, and name the equation 'eq06'. Perform White's nR^2 test for heteroscedasticity: note that homoscedasticity is rejected at all usual levels of significance. (This is evidence against the classical assumption $Var(u) = \sigma^2 I_n$).