5 Multiple regression analysis with qualitative information

Ezequiel Uriel  
University of Valencia  
Version: 21-09-2012

5.1 Introduction of qualitative information in econometric models.  
5.2 A single dummy independent variable  
5.3 Multiple categories for an attribute  
5.4 Several attributes  
5.5 Interactions involving dummy variables.  
5.5.1 Interactions between two dummy variables  
5.5.2 Interactions between a dummy variable and a quantitative variable  
5.6 Testing structural changes  
5.6.1 Using dummy variables  
5.6.2 Using separate regressions: The Chow test  
Exercises

5.1 Introducing qualitative information in econometric models.

Up until now, the variables that we have used in explaining the endogenous variable have a quantitative nature. However, there are other variables of a qualitative nature that can be important when explaining the behavior of the endogenous variable, such as sex, race, religion, nationality, geographical region etc. For example, holding all other factors constant, female workers are found to earn less than their male counterparts. This pattern may result from gender discrimination, but whatever the reason, qualitative variables such as gender seem to influence the regressand and clearly should be included in many cases among the explanatory variables, or the regressors. Qualitative factors often (although not always) come in the form of binary information, i.e. a person is male or female, is either married or not, etc. When qualitative factors come in the form of dichotomous information, the relevant information can be captured by defining a binary variable or a zero-one variable. In econometrics, binary variables used as regressors are commonly called dummy variables. In defining a dummy variable, we must decide which event is assigned the value one and which is assigned the value zero.

In the case of gender, we can define

\[
female = \begin{cases} 
1 & \text{if the person is a female} \\
0 & \text{if the person is a male} 
\end{cases}
\]

But of course we can also define

\[
male = \begin{cases} 
1 & \text{if the person is a male} \\
0 & \text{if the person is a female} 
\end{cases}
\]

Nevertheless, it is important to remark that both variables, male and female, contain the same information. Using zero-one variables for capturing qualitative information is an arbitrary decision, but with this election the parameters have a natural interpretation.

5.2 A single dummy independent variable

Let us see how we incorporate dichotomous information into regression models. Consider the simple model of hourly wage determination as a function of the years of education (edu):
To measure gender wage discrimination, we introduce a dummy variable for gender as an independent variable in the model defined above,

\[ \text{wage} = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + u \]  

(5-2)

The attribute gender has two categories: male and female. The female category has been included in the model, while the male category, which was omitted, is the reference category. Model 1 is shown in Figure 5.1, taking \( \delta_0 < 0 \). The interpretation of \( \delta_0 \) is the following: \( \delta_0 \) is the difference in hourly wage between females and males, given the same amount of education (and the same error disturbance \( u \)). Thus, the coefficient \( \delta_0 \) determines whether there is discrimination against women or not. If \( \delta_0 < 0 \) then, for the same level of other factors (education, in this case), women earn less than men on average. Assuming that the disturbance mean is zero, if we take expectation for both categories we obtain:

\[
E(\text{wage} | \text{female} = 1, \text{educ}) = \beta_0 + \delta_0 + \beta_1 \text{educ} \\
E(\text{wage} | \text{female} = 0, \text{educ}) = \beta_0 + \beta_1 \text{educ}
\]

(5-3)

As can be seen in (5-3), the intercept is \( \beta_0 \) for males, and \( \beta_0 + \delta_0 \) for females. Graphically, as can be seen in Figure 5.1, there is a shift of the intercept, but the lines for men and women are parallel.

\[ \text{FIGURE 5.1. Same slope, different intercept.} \]

In (5-2) we have included a dummy variable for female but not for male, because if we had included both dummies this would have been redundant. In fact, all we need is two intercepts, one for females and another one for males. As we have seen, if we introduce the female dummy variable, we have an intercept for each gender. Introducing two dummy variables would cause perfect multicollinearity given that \( \text{female} + \text{male} = 1 \), which means that male is an exact linear function of female and of the intercept. Including dummy variables for both genders plus the intercept is the simplest example of the so-called dummy variable trap, as we shall show later on.

If we use male instead of female, the wage equation would be the following:

\[ \text{wage} = \alpha_0 + \gamma_0 \text{male} + \beta_1 \text{educ} + u \]  

(5-4)
Nothing has changed with the new equation, except the interpretation of $\alpha_0$ and $\gamma_0$: $\alpha_0$ is the intercept for women, which is now the reference category, and $\alpha_0 + \gamma_0$ is the intercept for men. This implies the following relationship between the coefficients:

$$\alpha_0 = \beta_0 + \delta_0$$

and

$$\alpha_0 + \gamma_0 = \beta_0 \Rightarrow \gamma_0 = -\delta_0$$

In any application, it does not matter how we choose the reference category, since this only affects the interpretation of the coefficients associated to the dummy variables, but it is important to keep track of which category is the reference category. Choosing a reference category is usually a matter of convenience. It would also be possible to drop the intercept and to include a dummy variable for each category. The equation would then be

$$wage = \mu_0 \text{ male} + \nu_0 \text{ female} + \beta_i \text{ educ} + u$$

where the intercept is $\mu_0$ for men and $\nu_0$ for women.

Hypothesis testing is performed as usual. In model (5-2), the null hypothesis of no difference between men and women is $H_0: \delta_0 = 0$, while the alternative hypothesis that there is discrimination against women is $H_1: \delta_0 < 0$. Therefore, in this case, we must apply a one sided (left) $t$-test.

A common specification in applied work has the dependent variable as the logarithm transformation $\ln(y)$ in models of this type. For example:

$$\ln(wage) = \beta_0 + \delta_0 \text{ female} + \beta_i \text{ educ} + u$$

Let us see the interpretation of the coefficient of the dummy variable in a log model. In model (5-6), taking $u=0$, the wage for a female and for a male is as follows:

$$\ln(wage_F) = \beta_0 + \delta_0 + \beta_i \text{ educ}$$

$$\ln(wage_M) = \beta_0 + \beta_i \text{ educ}$$

Given the same amount of education, if we subtract (5-7) from (5-8), we have

$$\ln(wage_F) - \ln(wage_M) = \delta_0$$

Taking antilogs in (5-9) and subtracting 1 from both sides of (5-9), we get

$$\frac{wage_F}{wage_M} - 1 = e^{\delta_0} - 1$$

That is to say

$$\frac{wage_F - wage_M}{wage_M} = e^{\delta_0} - 1$$

According to (5-11), the proportional change between the female wage and the male wage, for the same amount of education, is equal to $e^{\delta_0} - 1$. Therefore, the exact percentage change in hourly wage between men and women is $100 \times (e^{\delta_0} - 1)$. As an approximation to this change, $100 \times \delta_0$ can be used. However, if the magnitude of the percentage is high, then this approximation is not so accurate.

**EXAMPLE 5.1 Is there wage discrimination against women in Spain?**

Using data from the *wage structure survey* of Spain for 2002 (file *wage02sp*), model (5-6) has been estimated and the following results were obtained:
\[
\ln(\text{wage}) = 1.731 - 0.307 \times \text{female} + 0.0548 \times \text{educ} \\
R^2=0.243 \quad n=2000
\]

where wage is hourly wage in euros, female is a dummy variable that takes the value 1 if it is a woman, and educ are the years of education. (The numbers in parentheses are standard errors of the estimators.)

To answer the question posed above, we need to test \( H_0: \delta_1 = 0 \) against \( H_1: \delta_1 < 0 \). Given that the \( t \) statistic is equal to \(-14.27\), we reject the null hypothesis for \( \alpha=0.01 \). That is to say, there is a negative discrimination in Spain against women in the year 2002. In fact, the percentage difference in hourly wage between men and women is \( 100 \times (e^{0.307} - 1) = 35.9\% \), given the same years of education.

**EXAMPLE 5.2 Analysis of the relation between market capitalization and book value: the role of ibex35**

A researcher wants to study the relationship between market capitalization and book value in shares quoted on the continuous market of the Madrid stock exchange. In this market some stocks quoted are included in the ibex35, a selective index. The researcher also wants to know whether the stocks included in the ibex35 have, on average, a higher capitalization. With this purpose in mind, the researcher formulates the following model:

\[
\ln(\text{marketcap}) = \beta_0 + \delta_1 \times \text{ibex}35 + \beta_1 \times \ln(\text{bookvalue}) + u
\]

where

- marketcap is the capitalization value of a company, which is calculated by multiplying the price of the stock by the number of stocks issued.
- bookvalue is the book value of a company, also referred to as the net worth of the company. The book value is calculated as the difference between a company's assets and its liabilities.
- ibex35 is a dummy variable that takes the value 1 if the corporation is included in the selective Ibex 35.

Using the 92 stocks quoted on 15th November 2011 which supply information on book value (file bolmad11), the following results were obtained:

\[
\ln(\text{marketcap}) = 1.784 + 0.690 \times \text{ibex}35 + 0.675 \times \ln(\text{bookvalue}) \\
R^2=0.893 \quad n=92
\]

The marketcap/bookvalue elasticity is equal to 0.690; that is to say, if the book value increases by 1%, then the market capitalization of the quoted stocks will increase by 0.675%.

To test whether the stocks included in ibex35 have on average a higher capitalization implies testing \( H_0: \delta_1 = 0 \) against \( H_1: \delta_1 > 0 \). Given that the \( t \) statistic is \((0.690/0.179)=3.85\), we reject the null hypothesis for the usual levels of significance. On the other hand, we see that the stocks included in ibex35 are quoted 99.4% higher than the stocks not included. The percentage is obtained as follows: \( 0.690 \times (e^{0.690} - 1) = 99.4\% \).

In the case of \( \beta_1 \), we can test \( H_0: \beta_1 = 0 \) against \( H_1: \beta_1 \neq 0 \). Given that the \( t \) statistic is \((0.675/0.037)=18\), we reject the null hypothesis for the usual levels of significance.

**EXAMPLE 5.3 Do people living in urban areas spend more on fish than people living in rural areas?**

To see whether people living in urban areas spend more on fish than people living in rural areas, the following model is proposed:

\[
\ln(\text{fish}) = \beta_0 + \delta_1 \times \text{urban} + \beta_1 \times \ln(\text{inc}) + u
\]

where fish is expenditure on fish, urban is a dummy variable which takes the value 1 if the person lives in an urban area and inc is disposable income.

Using a sample of size 40 (file demand), model (5-13) was estimated:

\[
\ln(\text{fish}) = -6.375 + 0.140 \times \text{urban} + 1.313 \times \ln(\text{inc}) \\
R^2=0.904 \quad n=40
\]

According to these results, people living in urban areas spend 14% more on fish than people living in rural areas. If we test \( H_0: \delta_1 = 0 \) against \( H_1: \delta_1 > 0 \), we find that the \( t \) statistic is \((0.140/0.055)=2.55\). Given that \( t_{0.01} \approx t_{0.05} = 2.44 \), we reject the null hypothesis in favor of the alternative
for the usual levels of significance. That is to say, there is empirical evidence that people living in urban areas spend more on fish than people living in rural areas.

### 5.3 Multiple categories for an attribute

In the previous section we have seen an attribute (gender) that has two categories (male and female). Now we are going to consider attributes with more than two categories. In particular, we will examine an attribute with three categories.

To measure the impact of firm size on wage, we can use a dummy variable. Let us suppose that firms are classified into three groups according to their size: small (up to 49 workers), medium (from 50 to 199 workers) and large (more than 199 workers). With this information, we can construct three dummy variables:

- \( small = \begin{cases} 1 & \text{up to 49 workers} \\ 0 & \text{in other case} \end{cases} \)
- \( medium = \begin{cases} 1 & \text{from 50 to 199 workers} \\ 0 & \text{in other case} \end{cases} \)
- \( large = \begin{cases} 1 & \text{more than 199 workers} \\ 0 & \text{in other case} \end{cases} \)

If we want to explain hourly wages by introducing the firm size in the model, we must omit one of the categories. In the following model, the omitted category is small firms:

\[
\ln(wage) = \beta_0 + \beta_{medium} + \beta_{large} + \beta_{educ} + u
\]  

(5-14)

The interpretation of the \( \theta_j \) coefficients is the following: \( \theta_1 \) (\( \theta_2 \)) is the difference in hourly wage between medium (large) firms and small firms, given the same amount of education (and the same error term \( u \)).

Let us see what happens if we also include the category small in (5-14). We would have the model:

\[
\ln(wage) = \beta_0 + \beta_{small} + \beta_{medium} + \beta_{large} + \beta_{educ} + u
\]  

(5-15)

Now, let us consider that we have a sample of six observations: the observations 1 and 2 correspond to small firms; 3 and 4 to medium ones; and 5 and 6 to large ones. In this case the matrix of regressors \( X \) would have the following configuration:

\[
X = \begin{bmatrix}
1 & 1 & 0 & 0 & educ_1 \\
1 & 1 & 0 & 0 & educ_2 \\
1 & 0 & 1 & 0 & educ_3 \\
1 & 0 & 1 & 0 & educ_4 \\
1 & 0 & 0 & 1 & educ_5 \\
1 & 0 & 0 & 1 & educ_6 \\
\end{bmatrix}
\]

As can be seen in matrix \( X \), column 1 of this matrix is equal to the sum of columns 2, 3 and 4. Therefore, there is perfect multicollinearity due to the so-called dummy variable trap. Generalizing, if an attribute has \( g \) categories, we need to include only \( g-1 \) dummy variables in the model along with the intercept. The intercept for the reference category is the overall intercept in the model, and the dummy variable
coefficient for a particular group represents the estimated difference in intercepts between that category and the reference category. If we include \( g \) dummy variables along with an intercept, we will fall into the dummy variable trap. An alternative is to include \( g \) dummy variables and to exclude an overall intercept. In the case we are examining, the model would be the following:

\[
\ln(\text{wage}) = \theta_{\text{small}} + \theta_{\text{medium}} + \theta_{\text{large}} + \beta_{\text{educ}} + u
\]  

(5-16)

This solution is not advisable for two reasons. With this configuration of the model it is more difficult to test differences with respect to a reference category. Second, this solution only works in the case of a model with only one unique attribute.

**EXAMPLE 5.4** Does firm size influence wage determination?

Using the sample size example 5.1 (file wage02sp), model (5-14) was estimated:

\[
\ln(\text{wage}) = 1.566 + 0.281 \text{medium} + 0.162 \text{large} + 0.0480 \text{educ}
\]

\( \text{RSS} = 406 \quad R^2 = 0.218 \quad n = 2000 \)

To answer the question above, we will not perform an individual test on \( \theta_1 \) or \( \theta_2 \). Instead we must jointly test whether the size of firms has a significant influence on wage. That is to say, we must test whether medium and large firms together have a significant influence on the determination of wage. In this case, the null and the alternative hypothesis, taking (5-14) as the unrestricted model, will be the following:

\[
H_0 : \theta_1 = \theta_2 = 0 \\
H_1 : H_0 \text{ is not true}
\]

The restricted model in this case is the following:

\[
\ln(\text{wage}) = \beta_0 + \beta_{\text{educ}} + u
\]  

(5-17)

The estimation of this model is the following:

\[
\ln(\text{wage}) = 1.657 + 0.0525 \text{educ}
\]

\( \text{RSS} = 433 \quad R^2 = 0.166 \quad n = 2000 \)

Therefore, the \( F \) statistic is

\[
F = \frac{[\text{RSS}_u - \text{RSS}_r]/q}{[\text{RSS}_r/(n-1-k)]} = \frac{[433 - 406]/2}{406/(2000-1-3)} = 66.4
\]

So, according to the value of the \( F \) statistic, we can conclude that the size of the firm has a significant influence on wage determination for the usual levels of significance.

**Example 5.5** In the case of Lydia E. Pinkham, are the time dummy variables introduced significant individually or jointly?

In example 3.4, we considered the case of Lydia E. Pinkham in which sales of a herbal extract from this company (expressed in thousands of dollars) were explained in terms of advertising expenditures in thousands of dollars (\( \text{advexp} \)) and last year’s sales (\( \text{sales}_{t-1} \)). However, in addition to these two variables, the author included three time dummy variables: \( d_1, d_2 \) and \( d_3 \). These dummy variables encompass the various situations which took place in the company. Thus, \( d_1 \) takes 1 in the period 1907-1914 and 0 in the remaining periods, \( d_2 \) takes 1 in the period 1915-1925 and 0 in other periods, and finally, \( d_3 \) takes 1 in the period 1926 - 1940 and 0 in the remaining periods. Thus, the reference category is the period 1941-1960. The final formulation of the model was therefore the following:

\[
\text{sales} = \beta_1 + \beta_2 \text{advexp}_t + \beta_3 \text{sales}_{t-1} + \beta_4 d_1 + \beta_5 d_2 + \beta_6 d_3 + \epsilon
\]  

(5-18)

The results obtained in the regression, using file pinkham, were the following:

\[
\text{sales} = 254.6 + 0.5345 \text{advexp}_t + 0.6073 \text{sales}_{t-1} - 133.35 d_1 + 216.84 d_2 - 202.50 d_3 + \epsilon
\]

\( R^2 = 0.929 \quad n = 53 \)

To test whether the dummy variables individually have a significant effect on sales, the null and alternative hypotheses are:
\[
\begin{align*}
H_0 &: d_i = 0 \\
H_1 &: d_i \neq 0
\end{align*}
\]

The corresponding \( t \) statistics are the following:

\[
\begin{align*}
t_{d_1} &= \frac{-133.35}{89} = -1.50 \\
t_{d_2} &= \frac{216.84}{67} = 3.22 \\
t_{d_3} &= \frac{-202.50}{67} = -3.02
\end{align*}
\]

As can be seen, the regressor \( d_1 \) is not significant for any of the usual levels of significance, whereas on the contrary the regressors \( d_2 \) and \( d_3 \) are significant for any of the usual levels.

The interpretation of the coefficient of the regressor \( d_2 \), for example, is as follows: holding fixed the advertising spending and giving the previous year's sales, sales for one year of the period 1915-1920 are $2.684 higher than for a year of the period 1941-1960.

To test jointly the effect of the time dummy variables, the null and alternative hypotheses are

\[
\begin{align*}
H_0 &: d_1 = d_2 = d_3 = 0 \\
H_1 &: H_0 \text{ is not true}
\end{align*}
\]

and the corresponding test statistic is

\[
F = \frac{(R_{\text{full}}^2 - R_{\text{null}}^2) / q}{(1 - R_{\text{full}}^2) / (n - 1 - k)} = \frac{(0.9290 - 0.8770) / 3}{(1 - 0.9290) / (53 - 1 - 5)} = 11.47
\]

For any of the usual significance levels the null hypothesis is rejected. Therefore, the time dummy variables have a significant effect on sales.

### 5.4 Several attributes

Now we will consider the possibility of taking into account two attributes to explain the determination of wage: gender and length of workday (part-time and full-time). Let \( \text{partime} \) be a dummy variable that takes value 1 when the type of contract is part-time and 0 if it is full-time. In the following model, we introduce two dummy variables: \( \text{female} \) and \( \text{partime} \):

\[
\ln(wage) = \beta_0 + \delta_{\text{female}} + \phi_{\text{partime}} + \beta_{\text{educ}} + u \tag{5-19}
\]

In this model, \( \phi_0 \) is the difference in hourly wage between those who work part-time, given gender and the same amount of education (and also the same disturbance term \( u \)).

Each of these two attributes has a reference category, which is the omitted category. In this case, male is the reference category for gender and full-time for type of contract. If we take expectations for the four categories involved, we obtain:

\[
\begin{align*}
E[\ln(wage) \mid \text{female, partime, educ}] &= \beta_0 + \delta_0 + \phi_0 + \beta_{\text{educ}} \\
E[\ln(wage) \mid \text{female, fulltime, educ}] &= \beta_0 + \delta_0 + \beta_{\text{educ}} \\
E[\ln(wage) \mid \text{male, partime, educ}] &= \beta_0 + \phi_0 + \beta_{\text{educ}} \\
E[\ln(wage) \mid \text{male, fulltime, educ}] &= \beta_0 + \beta_{\text{educ}} 
\end{align*}
\tag{5-20}
\]

The overall intercept in the equation reflects the effect of both reference categories, male and full-time, and so full-time male is the reference category. From (5-20), you can see the intercept for each combination of categories.

**EXAMPLE 5.6 The influence of gender and length of the workday on wage determination**

Model (5-19) was estimated by using data from the wage structure survey of Spain for 2006 (file wage06sp):

\[
\begin{align*}
\ln(wage) &= 2.006 - 0.233 \text{ female} - 0.087 \text{ partime} + 0.0531 \text{ educ} \\
\text{ (0.026)} & \quad \text{(0.021)} \quad \text{(0.027)} \quad \text{(0.0023)}
\end{align*}
\]

7
According to the values of the coefficients and corresponding standard errors, it is clear that each one of the two dummy variables, female and partime, are statistically significant for the usual levels of significance.

**EXAMPLE 5.7 Trying to explain the absence from work in the company Buenosaires**

Buenosaires is a firm devoted to the manufacturing of fans, having had relatively acceptable results in recent years. For this purpose, the following model is proposed:

\[
\text{absent} = \beta_0 + \delta_0 \text{bluecoll} + \phi_0 \text{male} + \beta_1 \text{age} + \beta_2 \text{tenure} + \beta_3 \text{wage} + u
\]  

(5-21)

where bluecoll is a dummy indicating that the person is a manual worker (the reference category is white collar) and tenure is a continuous variable reflecting the years worked in the company.

Using a sample of size 48 (file absent), the following equation has been estimated:

\[
\begin{align*}
\text{absent} &= 12.444 + 0.968 \text{bluecoll} + 2.049 \text{male} - 0.037 \text{age} - 0.151 \text{tenure} - 0.044 \text{wage} \\
\text{RSS} &= 161.95 \\
R^2 &= 0.760 \\
n &= 48
\end{align*}
\]

Next, we will look at whether bluecoll is significant. Testing \( H_0 : \delta_0 = 0 \) against \( H_1 : \delta_0 \neq 0 \), the \( t \) statistic is \( \frac{(0.968/0.669)=1.45} {1.68} \). As \( t_{40}^{0.10}/2 = 1.68 \), we fail to reject the null hypothesis for \( \alpha = 0.10 \). And so there is no empirical evidence to state that absenteeism amongst blue collar workers is different from white collar workers. But if we test \( H_0 : \delta_0 = 0 \) against \( H_1 : \delta_0 > 0 \), as \( t_{40}^{0.10} = 1.30 \) for \( \alpha = 0.10 \), then we cannot reject that absenteeism amongst blue collar workers is greater than amongst white collar workers.

On the contrary, in the case of the male dummy, testing \( H_0 : \phi_0 = 0 \) against \( H_1 : \phi_0 \neq 0 \), given that the \( t \) statistic is \( \frac{2.049/0.712}=2.88 \) and \( t_{40}^{0.012} = 2.70 \), we reject that absenteeism is equal in men and women for the usual levels of significance.

**EXAMPLE 5.8 Size of firm and gender in determining wage**

In order to know whether the size of the firm and gender jointly are two relevant factors in determining wage, the following model is formulated:

\[
\ln(\text{wage}) = \beta_0 + \delta_0 \text{female} + \theta_1 \text{medium} + \theta_2 \text{large} + \beta_3 \text{educ} + u
\]

(5-22)

In this case, we must perform a joint test where the null and the alternative hypotheses are

\[
H_0 : \delta_0 = \theta_1 = \theta_2 = 0 \\
H_1 : H_0 \text{ is not true}
\]

In this case, the restricted model is model (5-17) which was estimated in example 5.4 (file wage02sp). The estimation of the unrestricted model is the following:

\[
\begin{align*}
\ln(\text{wage}) &= 1.639 - 0.327 \text{female} + 0.308 \text{medium} + 0.168 \text{large} + 0.0499 \text{educ} \\
\text{RSS} &= 361 \\
R^2 &= 0.305 \\
n &= 2000
\end{align*}
\]

The \( F \) statistic is

\[
F = \frac{[\text{RSS}_{u} - \text{RSS}_{r}]/q}{\text{RSS}_{r} / (n-k)} = \frac{[433 - 361]/3}{361/(2000-1-4)} = 133
\]

Therefore, according to the value of \( F \), we can conclude that the size of the firm and gender jointly have a significant influence in wage determination.

**5.5 Interactions involving dummy variables.**

**5.5.1 Interactions between two dummy variables**

To allow for the possibility of an interaction between gender and length of the workday on wage determination, we can add an interaction term between female and partime in model (5-19), with the model to estimate being the following:

\[
\ln(\text{wage}) = \beta_0 + \delta_0 \text{female} + \phi_0 \text{partime} + \varphi_0 \text{female} \times \text{partime} + \beta_3 \text{educ} + u
\]

(5-23)
This allows working time to depend on gender and vice versa.

**EXAMPLE 5.9 Is the interaction between females and part-time work significant?**

Model (5-23) was estimated by using data from the wage structure survey of Spain for 2006 (file wage06sp):

\[
\text{ln}(\text{wage}) = 2.007 - 0.259 \text{ female} - 0.198 \text{ partime} + 0.167 \text{ female} \times \text{ partime} + 0.0538 \text{ educ} \\
\text{RSS}=363 \quad R^2=0.238 \quad n=2000
\]

To answer the question posed, we have to test \( H_0 : \phi_0 = 0 \) against \( H_0 : \phi_0 \neq 0 \). Given that the \( t \) statistic is \( (0.167/0.058)=2.89 \) and taking into account that \( t_{0.01/2}=2.66 \), we reject the null hypothesis in favor of the alternative hypothesis. Therefore, there is empirical evidence that the interaction between females and part-time work is statistically significant.

**EXAMPLE 5.10 Do small firms discriminate against women more or less than larger firms?**

To answer this question, we formulate the following model:

\[
\text{ln}(\text{wage}) = \beta_0 + \delta_1 \text{ female} + \theta_1 \text{medium} + \theta_2 \text{large} \\
+ \phi_1 \text{ female} \times \text{ medium} + \phi_2 \text{ female} \times \text{ large} + \beta \text{ educ} + \text{ u} \\
\text{(5-24)}
\]

Using the sample of example 5.1 (file wage02sp), model (5-24) was estimated:

\[
\text{ln}(\text{wage}) = 1.624 - 0.262 \text{ female} + 0.361 \text{ medium} + 0.179 \text{ large} \\
- 0.159 \text{ female} \times \text{ medium} - 0.043 \text{ female} \times \text{ large} + 0.0497 \text{ educ} \\
\text{RSS}=359 \quad R^2=0.308 \quad n=2000
\]

If in (5-24) the parameters \( \phi_1 \) and \( \phi_2 \) are equal to 0, this will imply that in the equation for wage determination, there will be non interaction between gender and firm size. Thus to answer the above question, we take (5-24) as the *unrestricted* model. The null and the alternative hypothesis will be the following:

\[
H_0 : \phi_1 = \phi_2 = 0 \\
H_1 : H_0 \text{ is not true}
\]

In this case, the restricted model is therefore model (5-22) estimated in example 5.7. The \( F \) statistic takes the value

\[
F = \frac{[\text{RSS}_R - \text{RSS}_U] / q}{\text{RSS}_U / (n-1-k)} = \frac{[361 - 359] / 2}{359 / (2000-1-6)} = 5.55
\]

For \( \alpha=0.01 \), we find that \( F_{0.01/2}=4.98 \). As \( F=5.61 \), we reject \( H_0 \) in favor of \( H_1 \). As \( H_0 \) has been rejected for \( \alpha=0.01 \), it will also be rejected for levels of 5% and 10%. Therefore, the usual levels of significance, the interaction between gender and firm size is relevant for wage determination.

### 5.5.2 Interactions between a dummy variable and a quantitative variable

So far, in the examples for wage determination a dummy variable has been used to shift the intercept or to study its interaction with another dummy variable, while keeping the slope of \textit{educ} constant. However, one can also use dummy variables to shift the slopes by letting them interact with any continuous explanatory variables. For example, in the following model the \textit{female} dummy variable interacts with the continuous variable \textit{educ}:

\[
wage = \beta_0 + \beta \text{ educ} + \delta_1 \text{ female} \times \text{ educ} + \text{ u} \quad \text{(5-25)}
\]

As can be seen in figure 5.2, the intercept is the same for men and women in this model, but the slope is greater in men than in women because \( \delta_1 \) is negative.

In model (5-25), the returns to an extra year of education depend upon the gender of the individual. In fact,
\[
\frac{\partial wage}{\partial educ} = \begin{cases} 
\beta_i + \delta_i & \text{for women} \\
\beta_i & \text{for men} 
\end{cases}
\] (5-26)

**FIGURE 5.2. Different slope, same intercept.**

**EXAMPLE 5.11 Is the return to education for males greater than for females?**

Using the sample of example 5.1 (file `wage02sp`), model (5-25) was estimated by taking log for \( y \):

\[
\ln(wage) = 1.640 + 0.0632 \times educ - 0.0274 \times educ \times female
\]

\[
\begin{array}{l}
\text{RSS}=400 \\
R^2=0.229 \\
n=2000
\end{array}
\]

In this case, we need to test \( H_0 : \delta_i = 0 \) against \( H_1 : \delta_i < 0 \). Given that the \( t \) statistic is \((-0.0274/0.0021) = -12.81\), we reject the null hypothesis in favor of the alternative hypothesis for any level of significance. That is to say, there is empirical evidence that the return for an additional year of education is greater for men than for women.

### 5.6 Testing structural changes

So far we have tested hypotheses in which one parameter, or a subset of parameters of the model, is different for two groups (women and men, for example). But sometimes we wish to test the null hypothesis that two groups have the same population regression function, against the alternative that it is not the same. In other words, we want to test whether the same equation is valid for the two groups. There are two procedures for this: using dummy variables and running separate regressions through the Chow test.

#### 5.6.1 Using dummy variables

In this procedure, testing for differences across groups consists in performing a joint significance test of the dummy variable, which distinguishes between the two groups and its interactions with all other independent variables. We therefore estimate the model with (unrestricted model) and without (restricted model) the dummy variable and all the interactions.
From the estimation of both equations we form the $F$ statistic, either through the RSS or from the $R^2$. In the following model for the determination of wages, the intercept and the slope are different for males and females:

$$\text{wage} = \beta_0 + \delta_0 \text{female} + \beta_{\text{educ}} + \delta_1 \text{female} \times \text{educ} + u$$ (5-27)

This model is represented in figure 5.3. As can be seen, if female=1, we obtain

$$\text{wage} = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) \text{educ} + u$$ (5-28)

For women the intercept is $\beta_0 + \delta_0$, and the slope $\beta_1 + \delta_1$. For female=0, we obtain equation (5-1). In this case, for men the intercept is $\beta_0$, and the slope $\beta_1$. Therefore, $\delta_0$ measures the difference in intercepts between men and women and, $\delta_1$ measures the difference in the return to education between males and females. Figure 5.3 shows a lower intercept and a lower slope for women than for men. This means that women earn less than men at all levels of education, and the gap increases as educ gets larger; that is to say, an additional year of education shows a lower return for women than for men.

Estimating (5-27) is equivalent to estimating two wage equations separately, one for men and another for women. The only difference is that (5-27) imposes the same variance across the two groups, whereas separate regressions do not. This set-up is ideal, as we will see later on, for testing the equality of slopes, equality of intercepts, and equality of both intercepts and slopes across groups.

**EXAMPLE 5.12 Is the wage equation valid for both men and women?**

If parameters $\delta_0$ and $\delta_1$ are equal to 0 in model (5-27), this will imply that the equation for wage determination is the same for men and women. In order to answer the question posed, we take (5-27), as the *unrestricted* model but express wage in logs. The null and the alternative hypothesis will be the following:

$$H_0 : \delta_0 = \delta_1 = 0$$
$$H_1 : H_0 \text{ is not true}$$

Therefore, the restricted model is model (5-17). Using the same sample as in example 5.1 (file *wage02sp*), we have obtained the following estimation of models (5-27) and (5-17):

$$\ln(\text{wage}) = 1.739 - 0.3319 \text{female} + 0.0539 \text{educ} - 0.0027 \text{educ} \times \text{female}$$

$$\text{RSS}=393 \quad R^2=0.243 \quad n=2000$$
\[
\ln(\text{wage}) = 1.657 + 0.0525 \text{ educ}
\]
\[
\text{RSS}=433 \quad R^2=0.166 \quad n=2000
\]

The \( F \) statistic takes the value
\[
F = \frac{[\text{RSS}_a - \text{RSS}_b]/q}{\text{RSS}_b / (n-1-k)} = \frac{[433-393]/2}{393/(2000-1-3)} = 102
\]

It is clear that for any level of significance, the equations for men and women are different.

When we tested in example 5.1 whether there was discrimination in Spain against women (\( H_0 : \delta_0 = 0 \) against \( H_1 : \delta_0 < 0 \)), it was assumed that the slope of \( \text{educ} \) (model (5-6)) is the same for men and women. Now it is also possible to use model (5-27) to test the same null hypothesis, but assuming a different slope. Given that the \( t \) statistic is \((-0.3319/0.0546)=-6.06\), we reject the null hypothesis by using this more general model than the one in example 5.1.

In example 5.11 it was tested whether the coefficient \( \delta_1 \) in model (5-25) was 0, assuming that the intercept is the same for males and females. Now, if we take (5-27) as the unrestricted model, we can test the same null hypothesis, but assuming that the intercept is different for males and females. Given that the \( t \) statistic is \((0.0027/0.0054)=0.49\), we cannot reject the null hypothesis which states that there is no interaction between gender and education.

**EXAMPLE 5.13 Would urban consumers have the same pattern of behavior as rural consumers regarding expenditure on fish?**

To answer this question, we formulate the following model which is taken as the unrestricted model:

\[
\ln(\text{fish}) = \beta_0 + \delta_0 \text{urban} + \beta_1 \ln(\text{inc}) + \delta_1 \ln(\text{inc}) \times \text{urban} + u
\]  

(5-29)

The null and the alternative hypothesis will be the following:

\( H_0 : \delta_0 = \delta_1 = 0 \)

\( H_1 : H_0 \) is not true

The restricted model corresponding to this \( H_0 \) is

\[
\ln(\text{fish}) = \beta_0 + \beta_1 \ln(\text{inc}) + u
\]

(5-30)

Using the sample of example 5.3 (file demand), models (5-29) and (5-30) were estimated:

\[
\ln(\text{fish}) = -6.551 + 0.678 \text{urban} + 1.337 \ln(\text{inc}) - 0.075 \ln(\text{inc}) \times \text{urban}
\]

\[
\text{RSS}=1.123 \quad R^2=0.904 \quad n=40
\]

\[
\ln(\text{fish}) = -6.224 + 1.302 \ln(\text{inc})
\]

\[
\text{RSS}=1.325 \quad R^2=0.887 \quad n=40
\]

The \( F \) statistic takes the value
\[
F = \frac{[\text{RSS}_a - \text{RSS}_b]/q}{\text{RSS}_b / (n-1-k)} = \frac{[1.325-1.123]/2}{1.123/(40-4)} = 3.24
\]

If we look up in the \( F \) table for 2 \( df \) in the numerator and 35 \( df \) in the denominator for \( \alpha=0.10 \), we find \( F_{2.35}^{0.10} \approx 2.46 \). As \( F > 2.46 \) we reject \( H_0 \). However, as \( F_{2.35}^{0.05} \approx 3.27 \), we fail to reject \( H_0 \) in favour of \( H_1 \) for \( \alpha=0.05 \) and, therefore, for \( \alpha=0.01 \). Conclusion: there is no strong evidence that families living in rural areas have a different pattern of fish consumption than families living in rural areas.

**Example 5.14 Has the productive structure of Spanish regions changed?**

The question to be answered is specifically the following: Did the productive structure of Spanish regions change between 1995 and 2008? The problem posed is a problem of structural stability. To specify the model to be taken as a reference in the test, let us define the dummy \( y_{2008} \), which takes the value 1 if the year is 2008 and 0 if the year is 1995.

The reference model is a Cobb-Douglas model, which introduces additional parameters to collect the structural changes that may have occurred. Its expression is:
\[ \ln(q) = \gamma_0 + \alpha_0 \ln(k) + \beta_0 \ln(l) + \gamma_1 y_{2008} + \alpha_1 y_{2008} \times \ln(k) + \beta_1 y_{2008} \times \ln(l) + u \]  

(5-31)

It is easily seen, according to the definition of the dummy \( y_{2008} \), that the elasticities production/capital are different in the periods 1995 and 2008. Specifically, they take the following values:

\[ e_{Q/K}^{(1995)} = \frac{\partial \ln(Q)}{\partial \ln(K)} = \alpha_0 \quad e_{Q/K}^{(2008)} = \frac{\partial \ln(Q)}{\partial \ln(K)} = \alpha_0 + \alpha_1 \]

In the case that \( \alpha_1 \) is equal to 0, then the elasticity of production/capital is the same in both periods.

Similarly, the production/labor elasticities for the two periods are given by

\[ e_{L/K}^{(1995)} = \frac{\partial \ln(L)}{\partial \ln(K)} = \beta_0 \quad e_{L/K}^{(2008)} = \frac{\partial \ln(L)}{\partial \ln(K)} = \beta_0 + \beta_1 \]

The intercept in the Cobb-Douglas is a parameter that measures efficiency. In model (5-31), the possibility that the efficiency parameter (PEF) is different in the two periods is considered. Thus

\[ PEF^{(1995)} = \gamma_0 \quad PEF^{(2008)} = \gamma_0 + \gamma_1 \]

If the parameters \( \alpha_1, \beta_1 \), and \( \gamma_1 \) are zero in model (5-31), the production function is the same in both periods. Therefore, in testing structural stability of the production function, the null and alternative hypotheses are:

\[ H_0 : \gamma_1 = \alpha_1 = \beta_1 \]
\[ H_1 : H_0 \text{ is not true} \]

(5-32)

Under the null hypothesis, the restrictions given in (5-32) lead to the following restricted model:

\[ \ln(q) = \gamma_0 + \alpha_0 \ln(k) + \beta_0 \ln(l) + u \]

(5-33)

The file prodsp contains information for each of the Spanish regions in 1995 and 2008 on gross value added in millions of euros (gdp), occupation in thousands of jobs (labor), and productive capital in millions of euros (captot). You can also find the dummy \( y_{2008} \) in that file.

The results of the unrestricted regression model (5-31) are shown below. It is evident that we cannot reject the null hypothesis that each of the coefficients \( \alpha_1, \beta_1 \) and \( \gamma_1 \), taken individually, are 0, since none of the \( t \) statistics reaches 0.1 in absolute value.

\[ \ln(gva) = 0.0559 + 0.6743 \ln(captot) + 0.3291 \ln(labour) \\
- 0.1088 y_{20108} + 0.0154 y_{2008} \times \ln(captot) - 0.0094 y_{2008} \times \ln(labour) \]

\[ R^2=0.99394 \quad n=34 \]

The results of the restricted model (5-33) are the following:

\[ \ln(gva) = -0.0690 + 0.6959 \ln(captot) + 0.3111 \ln(labour) \]

\[ R^2=0.99392 \quad n=34 \]

As can be seen, the \( R^2 \) of the two models are virtually identical because they differ only from the fifth decimal. It is not surprising, therefore, that the \( F \) statistic for testing the null hypothesis (5-32) takes a value close to 0:

\[ F = \frac{(R_{10}^2 - R_{00}^2) / q}{(1-R_{10}^2) / (n-1-k)} = \frac{(0.99394 - 0.99392) / 3}{(1-0.99394) / (34-1-5)} = 0.0308 \]

Thus, the alternative hypothesis that there is structural change in the productive economy of the Spanish regions between 1995 and 2008 is rejected for any significance level.

5.6.2 Using separate regressions: The Chow test

This test was introduced by the econometrician Chow (1960). He considered the problem of testing the equality of two sets of regression coefficients. In the Chow test, the restricted model is the same as in the case of using dummy variables to distinguish between groups. The unrestricted model, instead of distinguishing the behaviour of the
two groups by using dummy variables, consists simply of separate regressions. Thus, in
the wage determination example, the *unrestricted* model consists of two equations:

\[
\begin{align*}
\text{female:} & \quad \ln(\text{wage}) = \beta_{01} + \beta_{11}\text{educ} + u \\
\text{male:} & \quad \ln(\text{wage}) = \beta_{02} + \beta_{12}\text{educ} + u
\end{align*}
\] (5-34)

If we estimate both equations by OLS, we can show that the $RSS$ of the
unrestricted model, $RSS_{UR}$, is equal to the sum of the $RSS$ obtained from the estimates
for women, $RSS_1$, and for men, $RSS_2$. That is to say,

$$RSS_{UR} = RSS_1 + RSS_2$$

The null hypothesis states that the parameters of the two equations in (5-34) are
equal. Therefore

$$H_0 : \begin{cases} \beta_{01} = \beta_{02} \\ \beta_{11} = \beta_{12} \end{cases}$$
$$H_1 : \text{No } H_0$$

By applying the null hypothesis to model (5-34), you get model (5-17), which is
the restricted model. The estimation of this model for the whole sample is usually called
the pooled ($P$) regression. Thus, we will consider that the $RSS_R$ and $RSS_P$ are equivalent
expressions.

Therefore, the $F$ statistic will be the following:

$$F = \frac{RSS_P - (RSS_1 + RSS_2)}{(k + 1) / [n - 2(k + 1)]} / \frac{RSS_1 + RSS_2}{n - 2(k + 1)}$$

(5-35)

It is important to remark that, under the null hypothesis, the error variances for
the groups must be equal. Note that we have $k+1$ restrictions: the slope coefficients
(interactions) plus the intercept. Note also that in the unrestricted model we estimate
two different intercepts and two different slope coefficients, and so the df of the model
are $n-2(k+1)$.

One important limitation of the Chow test is that under the null hypothesis there
are no differences at all between the groups. In most cases, it is more interesting to
allow partial differences between both groups as we have done using dummy variables.

The Chow test can be generalized to more than two groups in a natural way.
From a practical point of view, to run separate regressions for each group to perform the
test is probably easier than using dummy variables.

In the case of three groups, the $F$ statistic in the Chow test will be the following:

$$F = \frac{RSS_P - (RSS_1 + RSS_2 + RSS_3)}{2 \times (1 + k) / \left(\frac{RSS_1 + RSS_2 + RSS_3}{(n-3(1+k))}\right)}$$

(5-36)

Note that, as a general rule, the number of the df of the numerator is equal to the
(number of groups-1)×(1+k), while the number of the df of the denominator is equal to
$n$ minus (number of groups)×(1+k).

**EXAMPLE 5.15** Another way to approach the question of wage determination by gender

Using the same sample as in example 5.1 (file wage02sp), we have obtained the estimation of
the equations in (5-34) for men and women, which taken together gives the estimation of the *unrestricted*
model:
Female equation

\[
\ln(wage) = 1.407 + 0.0566 \times \text{educ}
\]

\[
\text{RSS} = 104 \quad R^2 = 0.236 \quad n = 617
\]

Male equation

\[
\ln(wage) = 1.739 + 0.0539 \times \text{educ}
\]

\[
\text{RSS} = 289 \quad R^2 = 0.175 \quad n = 1383
\]

The restricted model, estimated in example 5.4, has the same configuration as the equations in (5-34) but in this case refers to the whole sample. Therefore, it is the pooled regression corresponding to the restricted model. The \(F\) statistic takes the value

\[
F = \frac{[\text{RSS}_p - (\text{RSS}_u + \text{RSS}_p)] / (1 + k)}{\text{RSS}_u} = \frac{433 - (104 + 289)}{104 + 289} / (2000 - 2 \times 2) = 102
\]

The \(F\) statistic must be, and is, the same as in example 5.12. The conclusions are therefore the same.

**EXAMPLE 5.16 Is the model of wage determination the same for different firm sizes?**

In other examples the intercept, or the slope on education, was different for three different firm sizes (small, medium and large). Now we shall consider a completely different equation for each firm size. Therefore, the unrestricted model will be composed by three equations:

- small: \(\text{wage} = \beta_0 + \delta_{01} \times \text{female} + \beta_1 \times \text{educ} + u\)
- medium: \(\text{wage} = \beta_0 + \delta_{02} \times \text{female} + \beta_2 \times \text{educ} + u\)
- large: \(\text{wage} = \beta_0 + \delta_{03} \times \text{female} + \beta_3 \times \text{educ} + u\)

(5-37)

The null and the alternative hypothesis will be the following:

- \(H_0: \beta_1 = \beta_2 = \beta_3\)
- \(H_1: \delta_{01} = \delta_{02} = \delta_{03}\)
- \(H_1: \beta_1 = \beta_2 = \beta_3\)

Given this null hypothesis, the restricted model is model (5-2).

The estimations of the three equations of (5-37), by using file wage02sp, are the following:

- small
  \[
  \ln(wage) = 1.706 - 0.249 \times \text{female} + 0.0396 \times \text{educ}
  \]
  \[
  \text{RSS} = 121 \quad R^2 = 0.160 \quad n = 801
  \]

- medium
  \[
  \ln(wage) = 1.934 - 0.422 \times \text{female} + 0.0548 \times \text{educ}
  \]
  \[
  \text{RSS} = 123 \quad R^2 = 0.302 \quad n = 590
  \]

- large
  \[
  \ln(wage) = 1.749 - 0.303 \times \text{female} + 0.0554 \times \text{educ}
  \]
  \[
  \text{RSS} = 114 \quad R^2 = 0.273 \quad n = 609
  \]

The pooled regression has been estimated in example 5.1. The \(F\) statistic takes the value

\[
F = \frac{\left[\text{RSS}_p - (\text{RSS}_u + \text{RSS}_p + \text{RSS}_u)\right] / 2 \times (1 + k)}{\text{RSS}_u} = \frac{[393 - (121 + 123 + 114)] / 6}{(121 + 123 + 114) / (2000 - 3 \times 3)} = 32.5
\]

For any level of significance, we reject that the equations for wage determination are the same for different firm sizes.

**EXAMPLE 5.17 Is the Pinkham model valid for the four periods?**

In example 5.5, we introduced time dummy variables and we tested whether the intercept was different for each period. Now, we are going to test whether the whole model is valid for the four periods considered. Therefore, the unrestricted model will be composed by four equations:

\[
\ln(wage) = \beta_0 + \delta_{01} \times \text{female} + \beta_1 \times \text{educ} + u
\]
1907-1914 \( sales_i = \beta_{01} + \beta_{11} advexp_i + \beta_{21} sales_{i-1} + u_i \) 
1915-1925 \( sales_i = \beta_{02} + \beta_{12} advexp_i + \beta_{22} sales_{i-1} + u_i \) 
1926-1940 \( sales_i = \beta_{03} + \beta_{13} advexp_i + \beta_{23} sales_{i-1} + u_i \) 
1941-1960 \( sales_i = \beta_{04} + \beta_{14} advexp_i + \beta_{24} sales_{i-1} + u_i \) 
(5-38)

The null and the alternative hypothesis will be the following:

\[
H_0 : \begin{cases} 
\beta_{01} = \beta_{02} = \beta_{03} = \beta_{04} \\
\beta_{12} = \beta_{13} = \beta_{14} \\
\beta_{23} = \beta_{24} = \beta_{2a}
\end{cases}
\]

\[
H_1 : \text{No } H_0
\]

Given this null hypothesis, the restricted model is the following model:

\[
(5-39)
\]

The estimations of the four equations of (5-38) are the following:

\[
\begin{align*}
\text{1907-1914} & \quad \widehat{sales}_i = 64.84 + 0.9149 \text{advexp}_i + 0.4630 \text{sales}_{i-1} \quad SSR = 36017 \quad n = 7 \\
\text{1915-1925} & \quad \widehat{sales}_i = 221.5 + 0.1279 \text{advexp}_i + 0.9319 \text{sales}_{i-1} \quad SSR = 400605 \quad n = 11 \\
\text{1926-1940} & \quad \widehat{sales}_i = 446.8 + 0.4638 \text{advexp}_i + 0.4445 \text{sales}_{i-1} \quad SSR = 201614 \quad n = 15 \\
\text{1941-1960} & \quad \widehat{sales}_i = -182.4 + 1.6753 \text{advexp}_i + 0.3042 \text{sales}_{i-1} \quad SSR = 187332 \quad n = 20
\end{align*}
\]

The pooled regression, estimated in example 3.4, is the following:

\[
\begin{align*}
\text{1907-1914} & \quad \widehat{sales}_i = 138.7 + 0.3288 \text{advexp}_i + 0.7593 \text{sales}_{i-1} \quad SSR = 2527215 \quad n = 53 \\
\end{align*}
\]

The \( F \) statistic takes the value

\[
F = \frac{\left[ SSR_p - (SSR_1 + SSR_2 + SSR_3 + SSR_4) \right] / 3 \times (1+k)}{(SSR_1 + SSR_2 + SSR_3 + SSR_4) / (n-4(1+k))} \\
= \frac{\left[ 2527215 - (36017 + 400605 + 201614 + 187332) \right] / 9}{(36017 + 400605 + 201614 + 187332) / (53 - 4 \times 3)} = 9.16
\]

For any level of significance, we reject that the model (5-39) is the same for the four periods considered.

**Exercises**

**Exercise 5.1** Answer the following questions for a model with explanatory dummy variables:

a) What is the interpretation of the dummy coefficients?

b) Why are not included in the model so many dummy variables as categories there are?

**Exercise 5.2** Using a sample of 560 families, the following estimations of demand for rental are obtained:

\[
\hat{q}_i = 4.17 - 0.247 \ p_i + 0.960 \ y_i \\
R^2 = 0.371 \quad n=560
\]

\[
\hat{q}_i = 5.27 - 0.221 \ p_i + 0.920 \ y_i + 0.341 \ d_i \ y_i \\
R^2 = 0.380
\]
where $q_i$ is the log of expenditure on rental housing of the $i^{th}$ family, $p_i$ is the logarithm of rent per m$^2$ in the living area of the $i^{th}$ family, $y_i$ is the log of household disposable income of the $i^{th}$ family and $d_i$ is a dummy variable that takes value one if the family lives in an urban area and zero in a rural area.

(The numbers in parentheses are standard errors of the estimators.)

a) Test the hypothesis that the elasticity of expenditure on rental housing with respect to income is 1, in the first adjusted mode.

b) Test whether the interaction between the dummy variable and income is significant. Is there a significant difference in the housing expenditure elasticity between urban and rural areas? Justify your answer.

Exercise 5.3 In a linear regression model with dummy variables, answer the following questions:

a) The meaning and interpretation of the coefficients of dummy variables in models with endogenous variable in logs.

b) Express how a model is affected when a dummy variable is introduced in a multiplicative way with respect to a quantitative variable.

Exercise 5.4 In the context of a multiple linear regression model,

a) What is a dummy variable? Give an example of an econometric model with dummy variables. Interpret the coefficients.

b) When is there perfect multicollinearity in a model with dummy variables?

Exercise 5.5 The following estimation is obtained using data for workers of a company:

$$\text{wage}_i = 500 + 50\text{tenure}_i + 200\text{college}_i + 100\text{male}_i$$

where wage is the wage in euros per month, tenure is the number of years in the company, college is a dummy variable that takes value 1 if the worker is graduated from college and 0 otherwise and male is a dummy variable which takes value 1 if the worker is male and 0 otherwise.

a) What is the predicted wage for a male worker with six years of tenure and college education?

b) Assuming that all working women have college education and none of the male workers do, write a hypothetical matrix of regressors (X) for six observations. In this case, would you have any problem in the estimation of this model? Explain it.

c) Formulate a new model that allows to establish whether there are wage differentials between workers with primary, secondary and college education.

Exercise 5.6 Consider the following linear regression model:

$$y_i = \alpha + \beta x_i + \gamma_1 d_{i1} + \gamma_2 d_{i2} + u_i$$

where $y$ is the monthly salary of a teacher, $x$ is the number of years of teaching experience $y$, $d_{i1}$ and $d_{i2}$ are two dummy variables taking the following values:

$$d_{i1} = \begin{cases} 1 & \text{if the teacher is male} \\ 0 & \text{otherwise} \end{cases}, \quad d_{i2} = \begin{cases} 1 & \text{if the teacher is white} \\ 0 & \text{otherwise} \end{cases}$$

a) What is the reference category in the model?
b) Interpret $\gamma_1$ and $\gamma_2$. What is the expected salary for each of the possible categories?

c) To improve the explanatory power of the model, the following alternative specification was considered:

$$y_i = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \gamma_3 (d_{1i} d_{2i}) + u_i$$  \hspace{1cm} (2)

d) What is the meaning of the term $(d_{1i} d_{2i})$? Interpret $\gamma_3$.

e) What is the expected salary for each of the possible categories in model (2)?

**Exercise 5.7** Using a sample of 36 observations, the following results are obtained:

$$\hat{y}_i = 1.10 - 0.96 x_{i1} - 4.56 x_{i2} + 0.34 x_{i3}$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 109.24 \quad \sum_{i=1}^{n} \hat{u}_i^2 = 20.22$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the individual significance of the coefficient associated with $x_2$.

b) Calculate the coefficient of determination, $R^2$, and explain its meaning.

c) Test the joint significance of the model.

d) Two additional regressions, with the same specification, were made for the two categories $A$ and $B$ included in the sample ($n_1=21$ y $n_2=15$). In these estimates the following RSS were obtained: 11.09 y 2.17, respectively. Test if the behavior of the endogenous variable is the same in the two categories.

**Exercise 5.8** To explain the time devoted to sport ($sport$), the following model was formulated

$$sport = \beta_0 + \delta_0 female + \varphi_0 smoker + \beta_1 age + u$$  \hspace{1cm} (1)

where sport is the minutes spent on sports a day, on average; female and smoker are dummy variables taking the value 1 if the person is a woman or smoker of at least five cigarettes per day, respectively.

a) Interpret the meaning of $\delta_0$, $\varphi_0$ and $\beta_1$.

b) What is the expected time spent on sports activities for all possible categories?

c) To improve the explanatory time spent on sports activities for all possible categories?

d) What are the possible marginal effects of sport with respect to age in the model (2)? Describe them.

**Exercise 5.9** Using information for Spanish regions in 1995 and 2000, several production functions were estimated.

For the whole of the two periods, the following results were obtained:

$$\ln(q) = 5.72 + 0.26 \ln(k) + 0.75 \ln(l) - 1.14 f + 0.11 f \times \ln(k) - 0.05 f \times \ln(l)$$  \hspace{1cm} (1)
Moreover, the following models were estimated separately for each of the years:

1995
\[ \ln(q) = 5.72 + 0.26 \ln(k) + 0.75l \]  
\[ R^2 = 0.9527 \quad \bar{R}^2 = 0.9459 \quad RSS = 0.6052 \]  
(3)

2000
\[ \ln(q) = 4.58 + 0.37 \ln(k) + 0.70l \]  
\[ R^2 = 0.9629 \quad \bar{R}^2 = 0.9555 \quad RSS = 0.3331 \]  
(4)

where \( q \) is output, \( k \) is capital, \( l \) is labor and \( f \) is a dummy variable that takes the value 1 for 1995 data and 0 for 2000.

\( a) \) Test whether there is structural change between 1995 and 2000.
\( b) \) Compare the results of estimations (3) and (4) with estimation (1).
\( c) \) Test the overall significance of model (1).

**Exercise 5.10** With a sample of 300 service sector firms, the following cost function was estimated:

\[ \hat{\text{cost}}_i = 0.847 + 0.899 qty_i \quad RSS = 31.074 \quad n = 300 \]

where \( qty_i \) is the quantity produced.

The 300 firms are distributed in three big areas (100 in each one). The following results were obtained:

Area 1:
\[ \hat{\text{cost}}_i = 1.053 + 0.876 qty_i \quad \hat{\sigma}^2 = 0.457 \]

Area 2:
\[ \hat{\text{cost}}_i = 3.279 + 0.835 qty_i \quad \hat{\sigma}^2 = 3.154 \]

Area 3:
\[ \hat{\text{cost}}_i = 5.279 + 0.984 qty_i \quad \hat{\sigma}^2 = 4.255 \]

\( a) \) Calculate an unbiased estimation of \( \sigma^2 \) in the cost function for the sample of 300 firms.
\( b) \) Is the same cost function valid for the three areas?

**Exercise 5.11** To study spending on magazines (\( mag \)), the following models have been formulated:

\[ \ln(mag) = \beta_0 + \beta_1 \ln(inc) + \beta_2 age + \beta_3 male + u \]  
(1)

\[ \ln(mag) = \beta_0 + \beta_1 \ln(inc) + \beta_2 age + \beta_3 male + \beta_4 prim + \beta_5 sec + u \]  
(2)

where \( inc \) is disposable income, \( age \) is age in years, \( male \) is a dummy variable that takes the value 1 if he is male, \( prim \) and \( sec \) are dummy variables that take the value 1 when the individual has reached, at most, primary and secondary level respectively.

With a sample of 100 observations, the following results have been obtained

\[ \ln(mag_i) = 1.27 + 0.756 \ln(inc_i) + 0.031 age_i - 0.017 male_i \]

\[ RSS = 1.1575 \quad R^2 = 0.9286 \]
\[
\ln(mag_i) = 1.26 + 0.811\ln(inc_i) + 0.030\ age_i + 0.003\ male_i - 0.250\ prim_i + 0.108\ sec_i
\]

\[
RSS=0.0306 \quad R^2=0.9981
\]

a) Is education a relevant factor to explain spending on magazines? What is the reference category for education? Is spending on magazines higher for men than for women? Justify your answer.

b) Interpret the coefficient on the male variable. Is spending on magazines higher for men than for women? Justify your answer.

Exercise 5.12 Let \(fruit\) be the expenditure on fruit expressed in euros over a year carried out by a household and let \(r_1, r_2, r_3,\) and \(r_4\) be dichotomous variables which reflect the four regions of a country.

a) If you regress \(fruit\) only on \(r_1, r_2, r_3,\) and \(r_4\) without an intercept, what is the interpretation of the coefficients?

b) If you regress \(fruit\) only on \(r_1, r_2, r_3,\) and \(r_4\) with an intercept, what would happen? Why?

c) If you regress \(fruit\) only on \(r_2, r_3,\) and \(r_4\) without an intercept, what is the interpretation of the coefficients?

d) If you regress \(fruit\) only on \(r_1-r_2, r_2, r_4-r_3,\) and \(r_4\) without an intercept, what is the interpretation of the coefficients?

Exercise 5.13 Consider the following model

\[
wage = \beta_0 + \delta_{female} + \beta_{educ} + u
\]

Now, we are going to consider three possibilities of defining the \(female\) dummy variable.

1) \(female = \begin{cases} 1 & \text{for female} \\ 0 & \text{for male} \end{cases}\)

2) \(female = \begin{cases} 2 & \text{for female} \\ 1 & \text{for male} \end{cases}\)

3) \(female = \begin{cases} 2 & \text{for female} \\ 0 & \text{for male} \end{cases}\)

a) Interpret the dummy variable coefficient for each definition.

b) Is one dummy variable definition preferable to another? Justify the answer.

Exercise 5.14 In the following regression model:

\[
wage = \beta_0 + \delta_{female} + u
\]

where \(female\) is a dummy variable, taking value 1 for female and value 0 for a male.

Prove that applying the OLS formulas for simple regression you obtain that

\[
\hat{\beta}_0 = \overline{wage}_M
\]
\[
\delta_0 = \overline{wage}_F - \overline{wage}_M
\]

where \(F\) indicates female and \(M\) male.

In order to obtain a solution, consider that in the sample there are \(n_1\) females and \(n_2\) males: the total sample is \(n = n_1 + n_2\).

Exercise 5.15 The data of this exercise were obtained from a controlled marketing experiment in stores in Paris on coffee expenditure, as reported in A. C. Bemmaor and D. Mouchoux, “Measuring the Short-Term Effect of In-Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment”, Journal of Marketing Research.
In this experiment, the following model has been formulated to explain the quantity sold of coffee per week:

\[ \ln(\text{coffqty}) = \beta_0 + \delta_1 \text{advert} + \beta_2 \ln(\text{coffee price}) + \delta_3 \text{advert} \times \ln(\text{coffpric}) + u \]

where \( \text{coffpric} \) takes three values: 1, for the usual price, 0.95 and 0.85; \( \text{advert} \) is a dummy variable that takes value 1 if there is advertising in this week and 0 if there is not. The experiment lasted for 18 weeks. The original model and three other models were estimated, using file \( \text{coffee2} \):

1) \( \ln(\text{coffqty}) = 5.85 + 0.2565 \text{advert} - 3.9760 \ln(\text{coffpric}) - 1.069 \text{advert} \times \ln(\text{coffpric}) \)
   \[ R^2 = 0.9468 \quad n = 18 \]

2) \( \ln(\text{coffqty}) = 5.83 + 0.3559 \text{advert} - 4.2539 \ln(\text{coffpric}) \)
   \[ R^2 = 0.9412 \quad n = 18 \]

3) \( \ln(\text{coffqty}) = 5.88 - 3.6939 \ln(\text{coffpric}) - 2.9575 \text{advert} \times \ln(\text{coffpric}) \)
   \[ R^2 = 0.9214 \quad n = 18 \]

a) In model (2), what is the interpretation of the coefficient on \( \text{advert} \)?

b) In model (3), what is the interpretation of the coefficient on \( \text{advert} \times \ln(\text{coffpric}) \)?

c) In model (2), does the coefficient on \( \text{advert} \) have a significant positive effect at 5% and at 1%?

d) Is the same model valid for weeks with advertising and for weeks without advertising?

e) Is the intercept the same for weeks with advertising and for weeks without advertising?

f) In model (1), is the coffee demand/price elasticity different for weeks with advertising and for weeks without advertising?

g) In model (4), is the coffee demand/price elasticity smaller than -4?

Exercise 5.16 (Continuation of exercise 4.39). Using file \( \text{timuse03} \), the following models have been estimated:

1) \( \text{houswork}_i = 132 + 2.787 \text{educ}_i + 1.847 \text{age}_i - 0.2337 \text{paidwork}_i \)
   \[ R^2 = 0.142 \quad n = 1000 \]

2) \( \text{houswork}_i = -3.02 + 3.641 \text{educ}_i + 1.775 \text{age}_i - 0.1568 \text{paidwork}_i + 32.11 \text{female}_i \)
   \[ R^2 = 0.298 \quad n = 1000 \]

3) \( \text{houswork}_i = -8.04 + 4.847 \text{educ}_i + 1.333 \text{age}_i - 0.0871 \text{paidwork}_i + 32.75 \text{female}_i - 0.1650 \text{ educ}_i \times \text{female}_i + 0.1019 \text{age}_i \times \text{female}_i - 0.02625 \text{ paidwork}_i \times \text{female}_i \)
   \[ R^2 = 0.306 \quad n = 1000 \]

a) In model (1), is there a statistically significant tradeoff between time devoted to paid work and time devoted to housework?
b) All other factors being equal and taking as a reference model (2), is there evidence that women devote more time to housework than men?
c) Compare the $R^2$ of models (1) and (2). What is your conclusion?
d) In model (3), what is the marginal effect of time devoted to housework with respect to time devoted to paid work?
e) Is interaction between paidwork and gender significant?
f) Are the interactions between gender and the quantitative variables of the model jointly significant?

Exercise 5.17 Using data from Bolsa de Madrid (Madrid Stock Exchange) on November 19, 2011 (file bolmad11), the following models have been estimated:

\begin{align*}
\ln(mktval_i) &= 1.784 + 0.6998 ibex35_i + 0.6749 \ln(bookval_i) \\
&\quad \text{RSS} = 35.69 \quad R^2 = 0.8931 \quad n=92 \\
\ln(mktval_i) &= 1.828 + 0.4236 ibex35_i + 0.6678 \ln(bookval_i) \\
&\quad + 0.0310 ibex35_i \times \ln(bookval_i) \\
&\quad \text{RSS} = 35.622 \quad R^2 = 0.8933 \quad n=92 \\
\ln(mktval_i) &= 2.323 + 0.1987 ibex35_i + 0.6688 \ln(bookval_i) \\
&\quad + 0.0369 ibex35_i \times \ln(bookval_i) - 0.6613 services_i - 0.6698 consump_i \\
&\quad - 0.1931 energy_i - 0.3895 industry_i - 0.7020 itc_i \\
&\quad \text{RSS} = 30.781 \quad R^2 = 0.9078 \quad n=92 \\
\ln(mktval_i) &= 1.366 + 0.7658 \ln(bookval_i) \\
&\quad \text{RSS} = 41.625 \quad R^2 = 0.8753 \quad n=92
\end{align*}

For finance=1

\begin{align*}
\ln(mktval_i) &= 0.558 + 0.9346 \ln(bookval_i) \\
&\quad \text{RSS} = 2.7241 \quad R^2 = 0.9415 \quad n=13
\end{align*}

where

- $mktval$ is the capitalization value of a company.
- $bookval$ is the book value of a company.
- $ibex35$ is a dummy variable that takes the value 1 if the corporation is included in the selective Ibex 35.
- $services$, $consumption$, $energy$, $industry$ and $itc$ (information technology and communication) are dummy variables. Each of them takes the value 1 if the corporation is classified in this sector in Bolsa de Madrid. The category of reference is $finance$.

a) In model (1), what is interpretation of the coefficient on $ibex35$?
b) In model (1), is the $mktval/bookval$ elasticity equal to 1?
c) In model (2), is the elasticity $mktval/bookval$ the same for all corporations included in the sample?
d) Is model (4) valid both for corporations included in ibex 35 and for corporations excluded?
e) In model (3), what is interpretation of the coefficient on $consump$?
f) Is the coefficient on $\text{consump}$ significatively negative?
g) Is the introduction of dummy variables for different sectors statistically justifiable?
h) Is the $\text{marktval/bookval}$ elasticity for the financial sector equal to 1?

**Exercise 5.18** (Continuation of exercise 4.37). Using file `rdspain`, the equations which appear in the attached table have been estimated.

The following variables appear in the table:
- $\text{rintens}$ is expenditure on research and development (R&D) as a percentage of sales,
- $\text{sales}$ are measured in millions of euros,
- $\text{exponsal}$ is exports as a percentage of sales;
- $\text{medtech}$ and $\text{hightech}$ are two dummy variables which reflects if the firm belongs to a medium or a high technology sector. The reference category corresponds to the firms with low technology,
- $\text{workers}$ is the number of workers.

<table>
<thead>
<tr>
<th></th>
<th>(1) $\text{rintens}$</th>
<th>(2) $\text{rintens}$</th>
<th>(3) $\text{rintens}$</th>
<th>(4) $\text{rintens}$ for $\text{hightech}=1$</th>
<th>(5) $\text{rintens}$ for $\text{medtech}=1$</th>
<th>(6) $\text{rintens}$ for $\text{lowtech}=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{exponsal}$</td>
<td>0.0136 (0.00195)</td>
<td>0.0101 (0.00193)</td>
<td>0.00968 (0.00189)</td>
<td>0.00584 (0.00792)</td>
<td>0.0116 (0.00300)</td>
<td>0.00977 (0.00169)</td>
</tr>
<tr>
<td>$\text{workers}$</td>
<td>0.000433 (0.0000740)</td>
<td>0.000392 (0.0000725)</td>
<td>0.000394 (0.000208)</td>
<td>0.00196 (0.000338)</td>
<td>0.0000563 (0.0000815)</td>
<td>0.000393 (0.000121)</td>
</tr>
<tr>
<td>$\text{hightech}$</td>
<td>1.448 (0.141)</td>
<td>0.976 (0.151)</td>
<td>0.472 (0.112)</td>
<td>0.00153 (0.000271)</td>
<td>0.0000326 (0.000222)</td>
<td></td>
</tr>
<tr>
<td>$\text{medtech}$</td>
<td>0.361 (0.109)</td>
<td>0.472 (0.112)</td>
<td>0.00153 (0.000271)</td>
<td>0.0000326 (0.000222)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{hightech} \times \text{workers}$</td>
<td>-0.000326 (0.000222)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{medtech} \times \text{workers}$</td>
<td>-0.000326 (0.000222)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{intercept}$</td>
<td>0.394 (0.0598)</td>
<td>0.137 (0.0691)</td>
<td>0.143 (0.0722)</td>
<td>1.211 (0.313)</td>
<td>0.577 (0.103)</td>
<td>0.142 (0.0443)</td>
</tr>
</tbody>
</table>

$\text{R}^2$: 0.0507, 0.0986, 0.138, 0.113, 0.0278, 0.0459
$\text{RSS}$: 9282.7, 8815.0, 8425.3, 4409.0, 2483.6, 1527.5
$\text{F}$: 52.90, 54.06, 52.90, 18.71, 8.776, 25.72
$\text{df}_n$: 2, 4, 6, 2, 2, 2

Standard errors in parentheses

a) In model (2), all other factors being equal, is there evidence that expenditure on research and development (expressed as a percentage of sales) in high technology firms is greater than in low technology firms? How strong is the evidence?
b) In model (2), all other factors being equal, is there evidence that \textit{rdintens} in medium technology firms is equal to low technology firms? How strong is the evidence?

c) Taking as reference model (2), if you had to test the hypothesis that \textit{rdintens} in high technology firms is equal to medium technology firms, formulate a model that allows you to test this hypothesis without using information on covariance matrix of the estimators.

d) Is the influence of workers on \textit{rdintens} associated with the level of technology in the firms?

e) Is the model (1) valid for all firms regardless of their technological level?

\textbf{Exercise 5.19} To explain the overall satisfaction of people (\textit{stsfglo}), the following model were estimated using data from the file \textit{hdr2010}:

\begin{equation}
\text{stsfglo}_i = -0.375 + 0.0000207 \text{gnipc}_i + 0.0858 \text{lifexpec}_i \\
R^2 = 0.642 \quad n = 144
\end{equation}

\begin{equation}
\text{stsfglo}_i = 2.911 + 0.0000381 \text{gnipc}_i + 1.215 \text{lifexpec}_i \\
+ 1.215 \text{dlatam}_i - 0.7901 \text{dafrica}_i \\
R^2 = 0.748 \quad n = 144
\end{equation}

\begin{equation}
\text{stsfglo}_i = 1.701 + 0.0000327 \text{gnipc}_i + 0.0527 \text{lifexpec}_i + 1.166 \text{dlatam}_i \\
- 3.096 \text{dafrica}_i + 0.0000673 \text{gnipc}_i \times \text{dafrica}_i - 0.0699 \text{lifexpec}_i \times \text{dafrica}_i \\
R^2 = 0.760 \quad n = 144
\end{equation}

where

- \textit{gnipc} is gross national income per capita expressed in PPP 2008 US dollar terms,
- \textit{lifexpec} is life expectancy at birth, i.e., number of years a newborn infant could be expected to live,
- \textit{dafrica} is a dummy variable that takes value 1 if the country is in Africa,
- \textit{dlatam} is a dummy variable that takes value 1 if the country is in Latin America.

a) In model (2), what is the interpretation of the coefficients on \textit{dlatam} and \textit{dafrica}?

b) In model (2), do \textit{dlatam} and \textit{dafrica} individually have a significant positive influence on global satisfaction?

c) In model (2), do \textit{dlatam} and \textit{dafrica} have a joint influence on global satisfaction?

d) Is the influence of life expectancy on global satisfaction smaller in Africa than in other regions of the world?

e) Is the influence of the variable \textit{gnipc} greater in Africa than in other regions of the world at 10%?

f) Are the interactions of people living in Africa and the variables \textit{gnipc} and \textit{lifexpec} jointly significant?

\textbf{Exercise 5.20} The equations which appear in the attached table have been estimated using data from the file \textit{timuse03}. This file contains 1000 observations corresponding to
a random subsample extracted from the time use survey for Spain carried out in 2002-2003.

The following variables appear in the table:
- *educ* is years of education attained,
- *sleep, paidwork* and *unpaidwrk* are measured in minutes per day,
- *female, workday* (Monday to Friday), *spaniard* and *houswife* are dummy variables.

  a) In model (1), is there a statistically significant tradeoff between time devoted to paid work and time devoted to sleep?
  b) In model (1), is the coefficient on *unpaidwrk* statistically significant?
  c) Is there evidence that women sleep more than men?
  d) In model (3), are *workday* and *spaniard* individually significant? Are they jointly significant?
  e) Is the coefficient on *housewife* statistically significant?
  f) Is each of the interactions between *female* and *educ, paidwork* and *unpaidwrk* statistically significant?
Exercise 5.21 To study infant mortality in the world, the following models have been estimated using data from the file `hdr2010`:

\[
\text{deathinf}_i = 93.02 - 0.00037 \text{gnipc}_i - 0.6046 \text{physicn}_i - 0.003 \text{contrcep}_i \\
\]

(1)

\[
\text{RSS}=40285 \quad R^2=0.6598 \quad n=108
\]

\[
\text{deathinf}_i = 78.55 - 0.00042 \text{gnipc}_i - 0.3809 \text{physicn}_i - 0.6989 \text{contrcep}_i + 17.92 \text{dafrcia}
\]

(2)
\[ \text{RSS} = 35893 \quad R^2 = 0.6851 \quad n = 108 \]

\[ \text{deathinf}_i = 72.58 - 0.00044 \text{gnipc} - 0.3994 \text{physicn}_i - 0.5857 \text{contrcep}_i + 17.92 \text{dafrica} - 0.0000914 \text{gnipc} \times \text{dafrica} - 2.0013 \text{physicn} \times \text{dafrica} - 0.2172 \text{contrcep} \times \text{dafrica} \]

\[ \frac{(6.76)}{(0.002)} \frac{(0.1879)}{(0.1234)} \frac{(5.05)}{(0.000826)} \frac{(2.2351)}{(0.2716)} \]

\[ \text{RSS} = 34309 \quad R^2 = 0.7109 \quad n = 108 \]

where

- `deathinf` is number of infant deaths (one year or younger) per 1000 live births in 2008,
- `gnipc` is gross national income per capita expressed in PPP 2008 US dollar terms,
- `physicn` are physicians per 10,000 people in the period 2000-2009,
- `contrcep` is the contraceptive prevalence rate using any method, expressed as % of married women aged 15–49 for the period 1990-2008,
- `dafrica` is a dummy variable that takes value 1 if the country is in Africa.

\( (a) \) In model (1), what is interpretation of the coefficients on `gnipc`, `physicn` and `contrcep`?

\( (b) \) In model (2), what is the interpretation of the coefficient on `dafrica`?

\( (c) \) In model (2), all other factors being equal, do the countries of Africa have a greater infant mortality than the countries of other regions of the world?

\( (d) \) What is the marginal effect of variable `gnipc` on infant mortality in model (3)?

\( (e) \) Is the slope corresponding to the regressor `contrcep` significantly greater for the countries of Africa?

\( (f) \) Are the slopes corresponding to the regressors `gnipc`, `physicn` and `contrcep` jointly different for the countries of Africa?

\( (g) \) Is the model (1) valid for all countries of the world?

**Exercise 5.22** Using a random subsample of 2000 observations extracted from the time use surveys for Spain carried out in the periods 2002-2003 and 2009-2010 (file `timus309`), the following models have been estimated to explain time spent watching television:

\[ \text{watchtv} = 114 - 3.523 \text{educ} + 1.330 \text{age} - 0.1111 \text{paidwork} \]

\[ R^2 = 0.169 \quad n = 2000 \]

\[ (1) \]

\[ \text{watchtv} = 127 - 3.653 \text{educ} + 1.291 \text{age} - 0.120 \text{paidwork} - 25.146 \text{female} + 17.137 \text{y2009} \]

\[ R^2 = 0.184 \quad n = 2000 \]

\[ (2) \]

\[ \text{watchtv} = 123 - 3.583 \text{educ} + 1.302 \text{age} - 0.105 \text{paidwork} - 24.869 \text{female} + 24.536 \text{y2009} - 0.050 \text{y2009} \times \text{paidwork} \]

\[ R^2 = 0.186 \quad n = 2000 \]

\[ (3) \]

where

- `educ` is years of education attained,
- `watchtv` and `paidwork` are measured in minutes per day.
- female is a dummy variable that takes value 1 if the interviewee is a female
- y2009 is a dummy variable that takes value 1 if the survey was carried out in 2008-2009

a) In model (1), what is interpretation of the coefficient on educ?
b) In model (1), is there a statistically significant tradeoff between time devoted to work and time devoted to watching television?
c) All other factors being equal and taking as reference model (2), is there evidence that men watch television more than women? How strong is the evidence?
d) In model (2), what is the estimated difference in watching television between females surveyed in 2008-2009 and males surveyed in 2002-2003? Is this difference statistically significant?
e) In model (3), what is the marginal effect of time devoted to paid work on time devoted to watching television?
f) Is there a significant interaction between the year of the survey and time devoted to paid work?

Exercise 5.23 Using the file consumsp, the following models were estimated to analyze if the entry of Spain into the European community in 1986 had any impact on the behavior of Spanish consumers:

\[
\text{c}_{it} = -7.156 + 0.3965 \text{inc}_{it} + 0.5771 \text{c}_{i,t-1} \\
R^2=0.9967 \quad \text{RSS}=1891320 \quad n=56
\]  
(1)

\[
\text{c}_{it} = -102.4 + 0.3573 \text{inc}_{it} + 0.5992 \text{c}_{i,t-1} + 148.60 \text{y}_{1986} \\
R^2=0.9968 \quad \text{RSS}=1802007 \quad n=56
\]  
(2)

\[
\text{c}_{it} = 79.17 + 0.5181 \text{inc}_{it} + 0.4186 \text{c}_{i,t-1} + 819.82 \text{y}_{1986} \\
-0.5403 \text{inc}_{it} \times \text{y}_{1986} + 0.5424 \text{c}_{i,t-1} \times \text{y}_{1986} \\
R^2=0.9972 \quad \text{RSS}=1600714 \quad n=56
\]  
(3)

\[
\text{c}_{it} = 117.03 + 0.3697 \text{inc}_{it} + 0.5823 \text{c}_{i,t-1} + 41.62 \text{y}_{1986} \\
+0.0104 \text{inc}_{it} \times \text{y}_{1986} \\
R^2=0.9968 \quad \text{RSS}=1798423 \quad n=56
\]  
(4)

\[
\text{c}_{it} = 120.1 + 0.3750 \text{inc}_{it} + 0.5758 \text{c}_{i,t-1} + 0.0141 \text{inc}_{it} \times \text{y}_{1986} \\
R^2=0.9968 \quad \text{RSS}=1798927 \quad n=56
\]  
(5)

(The numbers in parentheses are standard errors of the estimators.)

a) Test whether the marginal propensity to consume in the short term was reduced in 1986 and beyond.
b) Are the interactions between y1986 and the quantitative variables of the model jointly significant?
c) Test whether there was a structural change in the consumption function in 1986.
d) Test whether the coefficient on c_{i,t-1} changed in 1986 and beyond.
e) Was there a gap between consumption before 1986, with respect to 1986 and beyond?