

## IIA-3 Econometrics: Supervision 3

Emre Usenmez

2024-11-23

### FACULTY QUESTIONS

#### QUESTION 1

An investigator analysing the relationship between food expenditure, disposable income and prices across a random sample of 25 counties in the UK estimates the relationship

$$\log(FOOD) = 4.7377 + 0.3506 \log(PDI) - 0.5086 \log(PRICE)$$

(0.6805) (0.0899) (0.1010)

where figures in paranthesis are standard errors and where:

FOOD = Average household expenditure on food

PDI = Average personal disposable income

PRICE = Average price of food deflated by a general price index

(i) Give an economic interpretation of the coefficients on  $\log(PDI)$  and  $\log(PRICE)$ .

**Answer:** The coefficient of  $\log(PDI)$  is income elasticity, and the coefficient of  $\log(PRICE)$  is the price elasticity.

---

(ii) Test the hypothesis, using 5% significance level, that the coefficient of  $\log(PRICE)$  is equal to zero against the alternative that it is nonzero.

**Answer:** For this we would use a  $t$ -test for the hypothesis that

$$H_0 : \beta_2 = 0 \text{ vs } H_A : \beta_2 \neq 0$$

The  $t$ -statistic is:

$$\frac{-0.5086 - 0}{0.1010} = -5.035644$$

which is significantly different than  $t_{0.025,24} = -2.064$ , and so we would reject the null hypothesis.

---

(iii) Test the hypothesis using 5% significance level that the coefficient of  $\log(\text{INCOME})$  is equal to 1 against the alternative that it is significantly different from 1.

**Answer:** For this we would use a  $t$ -test for the hypothesis that

$$\mathbb{H}_0 : \beta_1 = 1 \text{ vs } \mathbb{H}_A : \beta_1 \neq 1$$

The  $t$ -statistic is:

$$\frac{0.3506 - 1}{0.0899} = -7.223582$$

which is significantly different than  $t_{0.025,24} = -2.064$ , and so we would reject the null hypothesis.

---

(iv) You are now given the following extra information:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 = 0.52876$$

$$RSS = \sum_{i=1}^n e_i^2 = 0.46276$$

Compute ESS and  $R^2$  for the above regression.

**Answer:**

$$ESS = TSS - RSS = 0.52876 - 0.46276 = 0.066$$

$$R^2 = \frac{ESS}{TSS} = \frac{0.066}{0.52876} = 0.1248203.$$

---

(v) Test the joint hypothesis at the 5% level that the two slope coefficients are all equal to zero against the alternative that at least one slope coefficient is nonzero.

**Answer:** For this we would use an  $F$ -test for the hypothesis that

$$\mathbb{H}_0 : (\beta_1 = 0) \cap (\beta_2 = 0)$$

The  $F$ -statistic is:

$$F = \frac{\frac{ESS}{df_1}}{\frac{RSS}{df_2}} = \frac{\frac{0.066}{2}}{\frac{0.46279}{22}} = 1.568746$$

or

$$F = \frac{\frac{R_u^2 nr - R_r^2 es}{df_1}}{\frac{1 - R_u^2 nr}{df_2}} = \frac{0.1248203}{2} \frac{1 - 0.1248203}{22} = 1.568847$$

giving a marginal difference due to rounding errors. In either case the  $F$ -statistic is significantly different than  $F_{0.05,2,22} = 0.0514$ , and so we would reject the null hypothesis.

-----

## QUESTION 2

# SUPPLEMENTARY QUESTIONS

## QUESTION 1

In a study of the Cobb-Douglass production function, a researcher suspects that the parameters are subject to change over time. Data on output,  $Y$ , labor input,  $X_1$ , and capital stock,  $X_2$ , are available for years 1929 to 1967.  $T$  represents the time trend. The results obtained are as follows (t-values in parantheses):

$$\begin{array}{ll}
 \text{Full Sample: } \widehat{\log Y} = -3.02 + 1.34 \log X_1 + 0.29 \log X_2 + 0.0052 T & \\
 \quad \quad \quad (-6.65) \quad (14.68) \quad \quad (4.89) \quad \quad (2.34) & \\
 \quad \quad \quad R^2 = 0.99535 \quad \quad \quad \hat{\sigma} = 0.03274 & \\
 \\
 \text{1929-48: } \widehat{\log Y} = -3.22 + 1.36 \log X_1 + 0.32 \log X_2 + 0.0051 T & \\
 \quad \quad \quad (-4.63) \quad (4.95) \quad \quad (1.36) \quad \quad (1.40) & \\
 \quad \quad \quad R^2 = 0.97853 \quad \quad \quad \hat{\sigma} = 0.04449 & \\
 \\
 \text{1949-67: } \widehat{\log Y} = -1.56 + 1.02 \log X_1 + 0.33 \log X_2 + 0.0095 T & \\
 \quad \quad \quad (-2.21) \quad (7.58) \quad \quad (2.33) \quad \quad (1.85) & \\
 \quad \quad \quad R^2 = 0.99565 \quad \quad \quad \hat{\sigma} = 0.0135 &
 \end{array}$$

a) Conduct a test of the hypothesis that the four regression coefficients are jointly the same in both sub-periods, against the alternative that they differ.

**Answer:** This question is effectively testing if there is a structural break from the start of 1949 which may be caused by different intercept, different slope coefficient, or both. Suppose the coefficients of the full sample regression are  $\beta$ s, and coefficients of the 1929-48 regression are  $\psi$ s, and coefficients of the 1949-67 are  $\gamma$ s. Our hypothesis is therefore:

$$\mathbb{H}_0 : (\psi_0 = \gamma_0) \cap (\psi_1 = \gamma_1) \cap (\psi_2 = \gamma_2) \cap (\psi_3 = \gamma_3)$$

We can test this in two ways. First is to reparameterize the model and then run an  $F$ -test, and the second is to run a Chow test.

First Approach:

We can create a new coefficient  $\delta = \psi - \gamma$  whereby our model becomes:  $\widehat{\log Y} = \delta_0 + \delta_1 \log X_1 + \delta_2 \log X_2 + \delta_3 T + \varepsilon$  for which the hypothesis becomes:

$$\mathbb{H}_0 : (\delta_0 = 0) \cap (\delta_1 = 0) \cap (\delta_2 = 0) \cap (\delta_3 = 0)$$

We would then use the  $F$ -test for this join hypothesis.

Second Approach:

An alternative appraoch is to use *Chow Test*.<sup>1</sup> This test assumes that:

- The error terms in the subperiod regressions are normally distributed with the same, i.e. homoskedastic, variance  $\sigma^2$ . That is,  $u_{1929-48t} \sim N(0, \sigma^2)$  and  $u_{1949-67t} \sim N(0, \sigma^2)$ .

---

<sup>1</sup>Chow, C Gregory (1960) *Tests of Equality Between Sets of Coefficients in Two Linear Regressions*, Econometrica, 28(3) 591:605

- The two error terms  $u_{1929-48t}$  and  $u_{1949-67t}$  are independently distributed.

The Chow test is an  $F$ -ratio which means we will need the  $RSS$  of both unrestricted and restricted models. Here, *the full sample model is the restricted model* since that is the model we have by imposing the restrictions that all  $\psi_j = \gamma_j$  for  $j = 0, \dots, 3$ . The  $RSS$  of the unrestricted model, on the other hand - and this is a key insight - is the combination of the two sub-sample  $RSS$ s.

#### Why $RSS$ and not $R^2$ form of $F$ -Test

Note that the Chow test uses  $RSS$  and that there is no simple  $R^2$  form of the  $F$ -test if separate regressions have been estimated for each group. This is because the  $TSS$ s are not the same as  $\bar{Y}$  is not the same in both samples.

The steps to carry out a Chow test is as follows:

1. Obtain the restricted model's residual sum of squares,  $RSS_R$ , by estimating the regression for the full sample model with  $(n - k - 1)$  degrees of freedom, where  $n = n_1 + n_2$  with  $n_1$  being the sample size of the first sub-sample, and  $n_2$  being the sample size of the second sub-sample, and where  $k$  is the number of regressors in that model.
2. Estimate the first sub-sample model to obtain its residual sum of squares,  $RSS_1$  with  $n_1 - k - 1$  degrees of freedom.
3. Do the same for the second sub-sample model to obtain  $RSS_2$  with  $n_2 - k - 1$  degrees of freedom.
4. Add the two  $RSS$ s to compute the unrestricted model's residual sum of squares:  $RSS_{UR} = RSS_1 + RSS_2$ .
5. Compute the  $F$ -ratio:

$$F = \frac{\frac{RSS_R - RSS_{UR}}{k + 1}}{\frac{RSS_{UR}}{n - 2(k + 1)}} = \frac{\frac{\text{improvement in fit}}{\text{extra degrees of freedom used up}}}{\frac{\text{residual sum of squares remaining}}{\text{degrees of freedom remaining}}}$$

$\hookrightarrow$  Because we are splitting the sample into two, we are estimating  $k$  regressors for each sample plus their intercepts, so we are using up  $(k + 1)$  degrees of freedom twice.

6. Compare the  $F$ -ratio to the critical  $F$  value with  $((k + 1), n - 2(k + 1))$  degrees of freedom and fail to reject the null hypothesis of *parameter stability*, i.e. no structural change, if  $F$ -ratio does not exceed the critical value at the chosen significance level.

Accordingly, we first need to calculate the  $RSS$ s. For that, recall that  $RSS = \hat{\sigma}^2(n - k - 1)$  where  $n = 39, n_1 = 20, n_2 = 19$  because the dates are inclusive. Therefore:

$$\begin{aligned} RSS_R &= 0.03274^2 \times 35 &= 0.03751677 \\ RSS_{UR} &= RSS_1 + RSS_2 \\ &= 0.04449^2 \times 16 + 0.0135^2 \times 15 \\ &= 0.03440351 \end{aligned}$$

With these we can now calculate our  $F$ -ratio:

$$F = \frac{\frac{0.03751677 - 0.04304995}{4}}{\frac{RSS_{UR}}{n - 2(k + 1)}} = \frac{\frac{0.03751677 - 0.03440351}{4}}{\frac{0.03440351}{31}} = 0.7013157$$

The  $F$ -statistic for  $\alpha = 0.05$  is 2.678667 and for  $\alpha = 0.01$  is 3.992811, and thus we fail to reject the null hypothesis of parameter stability at either of the  $\psi$  values.

```
qf(p=c(0.05, 0.01), df1=4, df2=31, lower.tail = FALSE)
```

```
## [1] 2.678667 3.992811
```

Why  $RSS$  equals  $\hat{\sigma}^2(n - k - 1)$ ?

Consider how we estimate the error variance,  $\sigma^2$ . First notice that  $\sigma^2 = \mathbb{E}(u^2)$ , so an unbiased estimator of  $\sigma^2$  is  $\frac{1}{n} \sum_{i=1}^n u_i^2$ . However, since we do not observe the errors  $u_i$  this is not a true estimator. What we have, though, is the estimates of the errors  $u_i$  which are the OLS residuals  $\hat{u}_i$ . If we replace the errors with the OLS residuals then we have

$$\sigma^2 = \frac{\sum_{i=1}^n u_i^2}{n} = \frac{RSS}{n}$$

which is a true estimator because it gives a computable rule for any sample of data on  $X$ s and  $Y$ . However, this is biased because it does not account for the restrictions that must be satisfied by the OLS residuals. These restrictions are given by the two OLS first order conditions:

$$\sum_{i=1}^n \hat{u}_i = 0, \quad \sum_{i=1}^n X_i \hat{u}_i = 0$$

for a simple regression with one regressor. In a way, if we know  $n - k - 1$  residuals, we can always get the other remaining residuals by using the restrictions implied by the first order conditions. Therefore there are only  $n - k - 1$  degrees of freedom in the OLS residuals, as opposed to  $n$  degrees of freedom in the errors.

The unbiased estimator of the error variance is therefore:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n - k - 1} = \frac{RSS}{n - k - 1}$$

where in simple regression with one regressor,  $k = 1$ .

(b) It is believed that irrespective of whether the form of the relationship has changed over the two periods (i.e. whether the coefficients of the equation have changed), there has still been a structural break. Test this hypothesis and use this result to comment on the assumptions made in part (a). What limitations are there to these methods for testing for stability over the whole period.

**Answer:** What this question is effectively asking is that whether the error variances of the two subperiod regressions are the same. Recall from part (a) one of the two key assumptions of the Chow test is that the variances of the errors are homoskedastic. Therefore the hypothesis is:

$$\begin{aligned} \mathbb{H}_0 : \sigma_\psi &= \sigma_\gamma & \text{or} & & \mathbb{H}_0 : \frac{\sigma_\psi}{\sigma_\gamma} &= 1 \\ \mathbb{H}_1 : \sigma_\psi &\neq \sigma_\gamma & \text{or} & & \mathbb{H}_0 : \frac{\sigma_\psi}{\sigma_\gamma} &\neq 1 \end{aligned}$$

If the variances of the two subpopulations are the same, i.e.  $\sigma_\psi = \sigma_\gamma$ , as assumed by the Chow test, then the ratios of the ratios of estimated error variance to population error variance has an  $F$  distribution with  $(n_1 - k - 1)$  and  $(n_2 - k - 1)$  degrees of freedom in the numerator and denominator, respectively. That is,

$$\frac{\frac{\hat{\sigma}_\psi^2}{\sigma_\psi^2}}{\frac{\hat{\sigma}_\gamma^2}{\sigma_\gamma^2}} \sim F_{(n_1-k-1), (n_2-k-1)}.$$

Notice that if  $\sigma_\psi = \sigma_\gamma$  then this ratio and thus the  $F$ -test becomes:

$$F = \frac{\sigma_\psi^2}{\sigma_\gamma^2}$$

where by convention the larger estimated variance is in the numerator.

Accordingly, our  $F$ -statistic is:

$$F = \frac{0.04449^2}{0.0135^2} = 10.86069.$$

Since this is a two-tailed test, but we are putting the higher variance in the numerator, then we can treat it as one-sided test with alternative hypothesis is greater than 1, The critical values for  $\alpha = 0.05$  with (16, 15) degrees of freedom is 2.384875, and for  $\alpha = 0.01$  it is 3.485246.

Thus at both  $\alpha$  levels we can reject the null hypothesis and conclude that the subperiod variances are not the same at  $\alpha = 0.01$ . This means, the assumption of Chow test does not hold and we shouldn't use the Chow test, at least not in this form. There are modifications to Chow test that can be utilized but that is beyond this class.

Another point regarding the Chow test to bear in mind is that it is sensitive to the choice of the time at which we divide the subperiods. The  $F$  values would be different if the cut-off point was 1947 or 1949.

Finally, the Chow test will tell us only if the two regressions are different but not whether the difference is due to the intercepts, the slopes, or both. We can use dummy variables for that, though.

```
qf(c(0.05, 0.01), df1 = 16, df2 = 15, lower.tail = FALSE)
```

```
## [1] 2.384875 3.485246
```

-----

## QUESTION 2

The following demand for money function was estimated from 60 observations, for which  $\sum(M - \bar{M})^2 = 45600$ :

$$\hat{M}_t = 284 + 0.56Y_t - 0.43M_{t-1} \quad R^2 = 0.841$$

When a further 8 observations became available, the equation was re-estimated. The pooled data had  $\sum(M - \bar{M})^2 = 50100$ , and the re-estimated equation  $R^2 = 0.818$ . Carry out a Chow test for predictive failure. What do you conclude from your results? Explain carefully the role of the dummy variable in this test.

**Answer:** The approach to the Chow test is similar to Question 1 whereby the  $RSS$  for restricted model is the pooled data. However, notice we are not running a two separate regressions on the subsamples. So we follow the first approach from Question 1(a) by taking the differences of the model:

$$\mathbb{H}_0 : (\delta_0 = 0) \cap (\delta_1 = 0) \cap (\delta_2 = 0)$$

where  $\delta$ s are the coefficients of model  $D_t$  which is the difference between the initial model and the model after the new observations.

Accordingly, we first need to derive the respective  $RSS$ s using the identity  $R^2 = 1 - \frac{RSS}{TSS}$  or  $RSS = (1 - R^2)TSS$  where  $TSS = \sum(M - \bar{M})^2$ :

$$RSS_R = (1 - 0.841) \times 45,600 = 7,250.4$$

$$RSS_{UR} = (1 - 0.818) \times 50,100 = 9,118.2$$

We can then calculate the Chow test whereby

$$F = \frac{9118.2 - 7250.4}{8} \frac{7250.4}{57} = 1.835495$$

The  $F$  critical values for  $\alpha = 0.05$  and for  $\alpha = 0.01$  with  $(8, 57)$  degrees of freedom are 2.105599 and 2.840694 respectively. Accordingly we cannot reject the null hypothesis and conclude that there is no predictive failure.

```
qf(c(0.05, 0.01), df1 = 8, df2 = 57, lower.tail = FALSE)
```

```
## [1] 2.105599 2.840694
```

-----

### QUESTION 3

(a) Use the dataset sup3.xls to import the following variables for the period 1955-1990:

W = Wages and Salaries (£ million), CSO code: CFAJ\_AU

P = Implied Deflator for Consumers Expenditure, CSO code: GIEF\_AU

E = Employees in Employment, CSO code: BCAD\_AU

WF = Workforce, CSO code: DYDB\_AU

U = Unemployed, CSO code: BCAB\_AU

Load the libraries:



```
libraries <- c("haven",      # to import/export SPSS, STATA, SAS files
              "readxl",     # to import/export Excel files
              "tidyverse",   # for tidy data
              "rstatix")     # converts stats functions to a tidyverse-friendly format

# lapply(libraries, library, character.only=TRUE) will load the libraries
```

Load the data:

```
salaries_df <- read_excel("../Data/sup3.xls", sheet = 1)
```

Briefly examine the data frame

```
# You can use any of the following to examine data frame (df):
# `dim()`: for its dimensions, by row and column
# `str()`: for its structure
# `summary()`: for summary statistics on its columns
# `colnames()`: for the name of each column
# `head()`: for the first 6 rows of the data frame
# `tail()`: for the last 6 rows of the data frame
# `View()`: for a spreadsheet-like display of the entire data frame

summary(salaries_df)
```

```
##      Year      CFAJ_AU      GIEF_AU      BCAD_AU
## Min.   :1955   Min.    : 10210   Min.    : 12.20   Min.    :21067
## 1st Qu.:1964   1st Qu.: 17430   1st Qu.: 15.55   1st Qu.:21890
## Median :1972   Median : 35890   Median : 24.75   Median :22466
## Mean   :1972   Mean    : 76269   Mean    : 47.06   Mean    :22285
## 3rd Qu.:1981   3rd Qu.:127400   3rd Qu.: 81.03   3rd Qu.:22740
## Max.   :1990   Max.    :271394   Max.    :127.90   Max.    :23257
##      DYDB_AU      BCAB_AU
## Min.   :24180   Min.    : 210.0
## 1st Qu.:25211   1st Qu.: 406.0
## Median :25634   Median : 626.5
## Mean   :26005   Mean    :1175.9
## 3rd Qu.:26692   3rd Qu.:1602.8
## Max.   :28437   Max.    :3229.0
```

We can change the column names to something more memorable:

```
colnames(salaries_df) <- c("Year", "W", "P", "E", "WF", "U")
```

(b) Use the data set to estimate the following wage equation by the OLS method:

$$\Delta \ln W_t = \beta_0 + \beta_1 \ln P + \beta_2 \ln P_{t-1} + \beta_3 \ln E_t + \beta_4 \ln E_{t-1} + \beta_5 \ln UR_t + \beta_6 \ln UR_{t-1} + \varepsilon_t$$

where  $\beta_i$  are constants,  $\Delta$  denotes the first difference operator and  $UR_t = \frac{U_t}{W_t}$ .

Verify that your estimate of  $\beta_1$  is .91170. Also show that the above is a generalized version of the following hypothesis:

$$\Delta \ln \left( \frac{W}{EP} \right)_t = \beta_0 + \varepsilon.$$

How would you explain the inclusion of the  $UR$  terms?

Answer:

```
#Transform the data to get it ready for the regression
salaries_df <- salaries_df %>%
  mutate(lnW = log(W),
         DlnW = lnW - lag(lnW,1),
         lnP = log(P),
         lag_lnP = lag(lnP,1),
         lnE = log(E),
         lag_lnE = lag(lnE,1),
         lnUR = log(U/WF),
         lag_lnUR = lag(lnUR,1))

#Let's put these in a new dataframe to keep them neat:
salaries_new_df <- data.frame(Year = salaries_df$Year,
                              lnW = salaries_df$lnW,
                              DlnW = salaries_df$DlnW,
                              lnP = salaries_df$lnP,
                              lag_lnP = salaries_df$lag_lnP,
                              lnE = salaries_df$lnE,
                              lag_lnE = salaries_df$lag_lnE,
                              lnUR = salaries_df$lnUR,
                              lag_lnUR = salaries_df$lag_lnUR,
                              stringsAsFactors = FALSE)
```

Since we are taking the differences those columns should have any value for 1955:

```
#Look at total number of `NA` values per column:
colSums(is.na(salaries_new_df))
```

```
##      Year      lnW      DlnW      lnP lag_lnP      lnE lag_lnE      lnUR
##      0         0         1         0         1         0         1         0
## lag_lnUR
##      1
```

Thus we can adjust our sample size to take account of this:

```
n_unr <- nrow(salaries_new_df) - sum(apply(is.na(salaries_df), 1, any))
# Note, if we replace 1 with 2 within apply function it will give us 4, the total number of NAs.
n_unr
```

```
## [1] 35
```

Thus our sample size is 35.

```
# Run the regression
lm_Q3b <- lm(DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR, data = salaries_new_df)
summary(lm_Q3b)

##
## Call:
## lm(formula = DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR,
##     data = salaries_new_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.052235 -0.010767  0.001441  0.007637  0.046109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.87957    1.96811   0.447  0.6584
## lnP           0.91170    0.10094   9.032 8.67e-10 ***
## lag_lnP       -0.89866    0.10072  -8.923 1.12e-09 ***
## lnE           0.62385    0.38833   1.607  0.1194
## lag_lnE       -0.71830    0.35308  -2.034  0.0515 .
## lnUR          -0.03547    0.02731  -1.299  0.2046
## lag_lnUR       0.01994    0.02391   0.834  0.4114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02058 on 28 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.8324, Adjusted R-squared:  0.7965
## F-statistic: 23.18 on 6 and 28 DF,  p-value: 1.175e-09
```

We can therefore see that only  $\ln P$  and  $\text{lag\_}\ln P$  coefficients are significant. However,  $F$ -stat is pretty high which suggests presence of multicollinearity. We can look at the variance inflation factor,  $VIF$ :

```
car::vif(lm_Q3b)

##      lnP      lag_lnP      lnE      lag_lnE      lnUR      lag_lnUR
## 560.948084 539.661549  8.784345  7.269653 40.557237 32.997711
```

where we see each of the  $VIF$  value is greater than 5 suggesting serious collinearity.

(c) Append the variable *IP* (strength of Income Policy Index) in Table 1 to your data set (use the generate command in Stata to generate a new variable, or copy and paste the data from the question sheet) and run the above regression including *IP* as an additional regressor and give both a statistical and an economic interpretation of your results.

**Answer:** Once we append *IP*, the model we will be testing is:

$$\Delta \ln W_t = \beta_0 + \beta_1 \ln P + \beta_2 \ln P_{t-1} + \beta_3 \ln E_t + \beta_4 \ln E_{t-1} + \beta_5 \ln UR_t + \beta_6 \ln UR_{t-1} + \beta_7 IP_t + \varepsilon_t$$

*#Append IP with data from the question*

```
salaries_new_df <- salaries_new_df %>%
  mutate(IP = 0)
salaries_new_df$IP[salaries_new_df$Year == 1962] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1965] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1966] <- 1.5
salaries_new_df$IP[salaries_new_df$Year == 1967] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1968] <- 1.5
salaries_new_df$IP[salaries_new_df$Year == 1969] <- 1.75
salaries_new_df$IP[salaries_new_df$Year == 1975] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1976] <- 3.0
salaries_new_df$IP[salaries_new_df$Year == 1977] <- 4.5
salaries_new_df$IP[salaries_new_df$Year == 1978] <- 1.0
salaries_new_df$IP[salaries_new_df$Year == 1979] <- 1.0
```

*# Now the number of regressors increased to 7:*

```
k_3c <- 7
```

*#Run the regression*

```
lm_Q3c <- lm(DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR + IP, data = salaries_new_df)
summary(lm_Q3c)
```

```
##
## Call:
## lm(formula = DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR +
##      IP, data = salaries_new_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.024935 -0.009426 -0.001330  0.006936  0.037838
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.008500   1.700503  -1.181  0.24785
## lnP          0.951886   0.080394  11.840 3.37e-12 ***
## lag_lnP      -0.961580   0.081034 -11.866 3.20e-12 ***
## lnE          1.098018   0.327043   3.357  0.00235 **
## lag_lnE      -0.887196   0.282076  -3.145  0.00401 **
## lnUR         -0.003416   0.022894  -0.149  0.88250
## lag_lnUR      0.012673   0.018987   0.667  0.51016
## IP           -0.014722   0.003492  -4.216  0.00025 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.01627 on 27 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared: 0.899, Adjusted R-squared: 0.8728
## F-statistic: 34.31 on 7 and 27 DF, p-value: 7.754e-12
```

With the introduction of  $IP$ , we now see that all are statistically significant except for  $UR$  and its lag, as well as the intercept. Thus we can see that by adding a required variable,  $IP$ , we reduce the standard error for the whole equation, which in turn, increases the  $t$ -values of the other coefficients to rise; except for  $UR$  which, as we will see, is because  $UR$  has no effect.

$R^2$  is not helpful here since it will always increase when a new variable is added. Adjusted  $R^2$  also increases from 0.7965 to 0.8728. However, a more useful indicated here is the  $F$ -statistic which tests that all slope coefficients are zero. If this is higher, then the equation is an improvement. With the introduction of  $IP$ , we see an increase from 23.18 to 34.31.

We can also check the significance of  $IP$  by looking at its  $t$ -value which is  $-4.216$ . Since the  $p$ -value is 0.00025 we know it is statistically significant and we don't need to calculate the  $t$ -statistic. As long as  $p$ -value is less than 0.05 it is significant at the 5 level.

---

(d) Test the following set of restrictions jointly by re-specifying your preferred equation:

$$\beta_1 + \beta_2 = 0; \quad \beta_3 + \beta_4 = 0; \quad \beta_5 + \beta_6 = 0$$

Verify your results using the Wald Test (the `test` command in Stata). Interpret your results.

**Answer:**

We can do this in two different ways. First is to use the `test` command in Stata, or `linearHypothesis` command from the `car` package in R for the Wald Test. The second way is to impose the constraint on the equation and then compare the constrained and unconstrained equations, testing the restriction on the constrained equation.

*# Approach 1:*

```
car::linearHypothesis(lm_Q3c, c("lnP=-lag_lnP", "lnE=-lag_lnE", "lnUR=-lag_lnUR"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## lnP + lag_lnP = 0
## lnE + lag_lnE = 0
## lnUR + lag_lnUR = 0
##
## Model 1: restricted model
## Model 2: DlnW ~ lnP + lag_lnP + lnE + lag_lnE + lnUR + lag_lnUR + IP
##
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1      30 0.0077553
## 2      27 0.0071515  3 0.00060376 0.7598 0.5265
```

which gives us the  $F$ -statistic as 0.7598.

In the second approach, the restricted model we will test is:

$$\Delta \ln W_t = \beta_0 + \beta_1 \Delta \ln P_t + \beta_3 \Delta \ln E_t + \beta_5 \Delta \ln UR_t + \beta_7 IP_t + \Delta \varepsilon_t$$

Why restricted model is a *dynamic* regression model?

Suppose we have the following model:

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

This model is known as *level form*. Since this level form holds true for every time period, this can also be written as their lagged values:

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + u_{t-1}$$

If we subtract the latter from the former, we will get what is called *first difference form*:

$$\Delta Y_t = \beta_1 \Delta X_t + \Delta u_t$$

where the *first difference operator*,  $\Delta$ , tells us to take successive differences of the variables in the equation.

So, for example, if  $X$  and  $Y$  in the level form represent the logarithms of income and consumption expenditure, then  $\Delta X$  and  $\Delta Y$  represent changes in the logs of income and consumption expenditure, respectively. So, if we are interested in the relationships between the variables in their growth form, then the difference form may be more appropriate. This is because a change in the log of a variable is a relative change, or a percentage change if multiplied by 100.

If the error term in the level form satisfies the OLS assumptions, especially the assumption of no autocorrelation, then the error term  $v_t = \Delta u_t = u_t - u_{t-1}$  is autocorrelated. To see this we need to take the expectation and variance of  $v_t$ :

$$\mathbb{E}(v_t) = \mathbb{E}(u_t - u_{t-1}) = \mathbb{E}(u) - \mathbb{E}(u_{t-1}) = 0$$

since  $\mathbb{E}(u) = 0$  for each  $t$ .

$$\begin{aligned} \text{Var}(v_t) &= \text{Var}(u_t - u_{t-1}) \\ &= \text{Var}(u_t) + \text{Var}(-u_{t-1}) \\ &= \sigma^2 + \sigma^2 \\ &= 2\sigma^2 \end{aligned}$$

thus  $v_t$  is homoskedastic. However we also need to look at the covariance of  $v_t$  and  $v_{t-1}$ :

$$\begin{aligned} \text{Cov}(v_t, v_{t-1}) &= \mathbb{E}(v_t v_{t-1}) - \text{mathbbE}(v_t) \text{mathbbE}(v_{t-1}) \\ &= \mathbb{E}((u_t - u_{t-1})(u_{t-1} - u_{t-2})) - 0 \\ &= \mathbb{E}(u_t u_{t-1} - u_t u_{t-2} - u_{t-1}^2 + u_{t-1} u_{t-2}) \\ &= \mathbb{E}(-u_{t-1}^2) \\ &= -\sigma^2 \end{aligned}$$

which is nonzero. Therefore, although the  $u$ 's are not autocorrelated the  $v$ 's are.

Now to see why the restrictions in the question are expressed in the first difference form, consider a general distributed-lag model with a finite lag of  $k$  time periods:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \cdots + \beta_{k+1} X_{t-k} + u_t$$

The coefficient  $\beta_1$  is known as the *short-run multiplier*, or *impact multiplier*, because it gives the change in the mean value of  $Y$  following a unit change in  $X$  in the same time period. If the change in  $X$  is maintained at the same level thereafter, then  $(\beta_1 + \beta_2)$  gives the change in the mean value of  $Y$  in the next period,  $(\beta_1 + \beta_2 + \beta_3)$  in the following period, so on. These partial sums are called *interim multipliers* or *intermediate multipliers*. After  $k$  time periods we obtain:

$$\sum_{i=1}^k +1\beta_i = \beta_1 + \beta_1 + \dots + \beta_k + 1 = \beta$$

which is known as the *long-run distributed-lag multiplier* or *total distributed-lag multiplier*.

In this question, what we are testing is if that total distributed-lag multiplier is different from 0 for first differences form.

### #Approach 2

*#Modify the data frame to add the constrained versions:*

```
salaries_new_df <- salaries_new_df %>%
  mutate(DlnP = lnP - lag_lnP,
         DlnE = lnE - lag_lnE,
         DlnUR = lnUR - lag_lnUR)
```

*#Estimate the constrained model:*

```
lm_Q3d <- lm(DlnW ~ DlnP + DlnE + DlnUR + IP, data = salaries_new_df)
summary(lm_Q3d)
```

```
##
## Call:
## lm(formula = DlnW ~ DlnP + DlnE + DlnUR + IP, data = salaries_new_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.029306 -0.009883  0.000174  0.007155  0.036716
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.033841   0.004885   6.928 1.08e-07 ***
## DlnP         0.975417   0.064695  15.077 1.53e-15 ***
## DlnE         0.961982   0.259391   3.709 0.000845 ***
## DlnUR        -0.008418   0.017647  -0.477 0.636796
## IP          -0.012631   0.003071  -4.114 0.000279 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01608 on 30 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.8904, Adjusted R-squared:  0.8758
## F-statistic: 60.94 on 4 and 30 DF, p-value: 5.662e-14
```

We will then need to calculate the  $F$ -statistic:

$$F = \frac{\frac{RSS_R - RSS_{UR}}{df_1}}{\frac{RSS_{UR}}{df_2}}$$

which means we need to obtain the  $RSS$ s from both regressions which can be done either by `deviance` function in R or by calculating the  $RSS$  manually:

```
#RSS unrestricted
# Manual calculation: sum(resid(lm_Q3c)^2)
RSS_unr_Q3d <- deviance(lm_Q3c)

#RSS restricted
#Manual calculation: sum(resid(lm_Q3d)^2)
RSS_res_Q3d <- deviance(lm_Q3d)

# Difference between the two models (ie number of restrictions):
q_3d <- 3

#F-statistic with (3,27) df:
Fstat_Q3d <- ((RSS_res_Q3d - RSS_unr_Q3d)/q_3d)/(RSS_unr_Q3d/(n_unr-k_3c-1))

#F Critical value:
Crit_Value_Q3d <- qf(0.05, df1=q_3d, df2=n_unr-k_3c-1,lower.tail = FALSE)

as.table(c("RSS_unr"=RSS_unr_Q3d,
           "RSS_res"=RSS_res_Q3d,
           "F-Stat"=Fstat_Q3d,
           "Critical_Value"=Crit_Value_Q3d,
           "df_1"=q_3d,
           "df_2"=n_unr-k_3c-1))
```

```
##          RSS_unr          RSS_res          F-Stat Critical_Value          df_1
##    0.007151510    0.007755266    0.759812374    2.960351318    3.000000000
##          df_2
##    27.000000000
```

We again obtain the same  $F$ -statistic of 0.7598 as in Approach 1, which is much lower than the critical value of 2.96, thus we cannot reject the null hypothesis. That is, the constraints seem to be correct.

(e) Test the further restrictions that  $\beta_5 = \beta_6 = 0$ . (N.B. your unrestricted equation is always your currently preferred equation, so in this case you only need to test  $\beta_5 = 0$  explicitly.)

**Answer:** We can respecify the constraint and run the regression on this model. Our constrained is  $\beta_5 = 0$ . This is because in part (d) we have  $\beta_5 + \beta_6 = 0$  thus if  $\beta_5 = 0$  then  $\beta_6 = 0$ . a t-test or check the t-statistic of the regression summary:



```
summary(lm_Q3d)$coefficients["DlnUR", "t value"]
```

```
## [1] -0.4770329
```

We see that the  $t$ -stat for  $DlnUR$  is  $-0.4770329$ .

But the question asks this to be done by respecifying the model. Also note that this time our unrestricted model is the restricted model in (d) since we are imposing further restrictions. Accordingly the model we test is:

$$\Delta \ln W_t = \beta_0 + \beta_1 \Delta \ln P_t + \beta_3 \Delta \ln E_t + \beta_7 IP_t + \Delta \varepsilon_t$$

```
#Estimate the constrained model:
```

```
lm_Q3e <- lm(DlnW ~ DlnP + DlnE + IP, data = salaries_new_df)
summary(lm_Q3e)
```

```
##
## Call:
## lm(formula = DlnW ~ DlnP + DlnE + IP, data = salaries_new_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.028085 -0.010693  0.000627  0.007825  0.034508
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.033677   0.004812   6.999 7.43e-08 ***
## DlnP         0.971549   0.063380  15.329 5.12e-16 ***
## DlnE         1.051632   0.176547   5.957 1.39e-06 ***
## IP          -0.012928   0.002969  -4.354 0.000135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01588 on 31 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.8896, Adjusted R-squared:  0.8789
## F-statistic: 83.26 on 3 and 31 DF, p-value: 6.321e-15
```

As we did in part (d) we can obtain the  $F$ -statistic:

```
# New RSS unrestricted:
```

```
RSS_unr_Q3e <- RSS_res_Q3d
```

```
# Number of regressors in the new unrestricted model:
```

```
k_3e <- 4
```

```
#RSS restricted
```

```
#Manual calculation: sum(resid(lm_Q3e)^2)
```

```
RSS_res_Q3e <- deviance(lm_Q3e)
```

```
# Difference between the two models (ie number of restrictions):
```

```
q_3e <- 1
```

```

#F-statistic with (1,30) df:
Fstat_Q3e <- ((RSS_res_Q3e - RSS_unr_Q3e)/q_3e)/(RSS_unr_Q3e/(n_unr-k_3e-1))

#F Critical value:
Crit_Value_Q3e <- qf(0.05, df1=q_3e, df2=n_unr-k_3e-1,lower.tail = FALSE)

as.table(c("RSS_unr"=RSS_unr_Q3e,
           "RSS_res"=RSS_res_Q3e,
           "F-Stat"=Fstat_Q3e,
           "Critical_Value"=Crit_Value_Q3e,
           "df_1"=q_3e,
           "df_2"=n_unr-k_3e-1))

```

```

##          RSS_unr          RSS_res          F-Stat Critical_Value          df_1
##    0.007755266    0.007814092    0.227560346    4.170876786    1.000000000
##          df_2
##    30.000000000

```

If we take the square of the t-statistic from the summary table, we should also get this F-statistic:

```
summary(lm_Q3d)$coefficients["DlnUR", "t value"]^2
```

```
## [1] 0.2275603
```

which gives us the F-Stat as expected.

---

(f) Interpret your preferred equation in light of your results to (c), (d), and (e).

**Answer:** The equation in (e),  $\Delta \ln W_t = \beta_0 + \beta_1 \Delta \ln P_t + \beta_3 \Delta \ln E_t + \beta_7 IP_t + \varepsilon_t$ , should be our new preferred equation since on all measures it is the best equation so far. This is good for the first hypothesis, i.e. the real wage resistance hypothesis, as this is exactly what we would expect if the hypothesis is correct.

---

(g) Test the hypothesis that  $\beta_3 = 1$  by respecifying your preferred equation in terms of average wages and salaries. (Your F-statistic for this test should be about 0.086)

**Answer:** One way to answer this question is to run a *t*-test on the coefficient of  $\Delta \ln E$ :

```
(summary(lm_Q3e)$coefficients["DlnE","Estimate"]-1)/summary(lm_Q3e)$coefficients["DlnE","Std. Error"]
```

```
## [1] 0.2924517
```

We check against the critical  $t$  value

```
qt(p=0.025, df=34, lower.tail = FALSE)
```

```
## [1] 2.032245
```

Thus we cannot reject the null hypothesis.

However, the question actually ask us to respecify the equation. For this, we set the coefficient on  $\Delta \ln E$  to 1, i.e.  $\beta_3 = 1$  and then take it over to the other side of the equation so that the model becomes:

$$\Delta \ln W_t - \Delta \ln E_t = \beta_0 + \beta_1 \Delta \ln P_t + \beta_7 IP_t + \Delta \varepsilon_t$$

or

$$\Delta \ln \left( \frac{W_t}{E_t} \right) = \beta_0 + \beta_1 \Delta \ln P_t + \beta_7 IP_t + \Delta \varepsilon_t$$

We can run the regression on this:

```
# First modify the data frame to reflect the new dependent variable:
salaries_new_df <- salaries_new_df %>%
  mutate("DlnWE" = DlnW-DlnE)

# Then run the regression:
lm_Q3g <- lm(DlnWE ~ DlnP + IP, salaries_new_df)
summary(lm_Q3g)
```

```
##
## Call:
## lm(formula = DlnWE ~ DlnP + IP, data = salaries_new_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.029237 -0.010758  0.000967  0.007744  0.034665
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.033934   0.004663   7.278 2.85e-08 ***
## DlnP         0.968616   0.061681  15.704 < 2e-16 ***
## IP          -0.012897   0.002925  -4.410 0.000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01565 on 32 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.8862, Adjusted R-squared:  0.8791
## F-statistic: 124.6 on 2 and 32 DF,  p-value: 7.943e-16
```

We can now derive the  $F$ -statistic as before:

```

# New RSS unrestricted:
RSS_unr_Q3g <- RSS_res_Q3e

# Number of regressors in the new unrestricted model:
k_3g <- 3

#RSS restricted
#Manual calculation: sum(resid(lm_Q3g)^2)
RSS_res_Q3g <- deviance(lm_Q3g)

# Difference between the two models (ie number of restrictions):
q_3g <- 1

#F-statistic with (1,30) df:
Fstat_Q3g <- (RSS_res_Q3g - RSS_unr_Q3g/q_3g)/(RSS_unr_Q3g/(n_unr-k_3g-1))

#F Critical value:
Crit_Value_Q3g <- qf(0.05, df1=q_3g, df2=n_unr-k_3g-1,lower.tail = FALSE)

as.table(c("RSS_unr"=RSS_unr_Q3g,
           "RSS_res"=RSS_res_Q3g,
           "F-Stat"=Fstat_Q3g,
           "Critical_Value"=Crit_Value_Q3g,
           "df_1"=q_3g,
           "df_2"=n_unr-k_3g-1))

```

```

##      RSS_unr      RSS_res      F-Stat Critical_Value      df_1
## 0.007814092 0.007835651 0.085527985 4.159615098 1.000000000
##      df_2
## 31.000000000

```

we can check that this is the square of the  $t$ -stat:

```
sqrt(Fstat_Q3g)
```

```
## [1] 0.2924517
```

---

(h) In the light of your preferred specification for the total period does it appear that price inflation is fully passed on in wage claims?

**Answer:** This question is basically asking if the coefficient of  $\Delta \ln P$  is equal to 1 or not. We can use the  $t$ -test for this:

```
(summary(lm_Q3g)$coefficients["DlnP","Estimate"]-1)/summary(lm_Q3g)$coefficients["DlnP","Std. Error"]
```

```
## [1] -0.5088118
```

We check against the critical  $t$  value

```
qt(p=0.025, df=34, lower.tail = FALSE)
```

```
## [1] 2.032245
```

Thus we cannot reject the null hypothesis.

---

(i) Interpret and comment upon your findings (especially in relation to the hypothesis suggested in (b)) and explain why your preferred equation is now given by:

$$\Delta \ln \left( \frac{W_t}{E_t} \right) = \beta_0 + \beta_1 \Delta \ln P_t + \beta_7 IP_t + \Delta \varepsilon$$

**Answer:** We have tested back as far as we can go and at each stage the real wage resistance hypothesis gives us the best equation. It is not clear what an hypothesis would have to do in order to be better.

---

(j) Why might it be expected that a structural break took place around 1979/1980? Test the hypothesis that the structure of wage determination changed significantly from 1980 onwards using the method of dummy variables (NB this is a different test to those you used in Q1 and Q2). Discuss your results.

**Answer:** The short answer is Thatcher. End to corporatism and wage bargaining etc.

To test this, rather than using the Chow Test, use the dummy variables method. For this, we define a variable,  $M$  which is 0 for the period 1956-1979, and 1 for 1980 onwards. We then estimate the following:

$$\Delta \ln W_t = \beta_0 + \beta_1 \Delta \ln P_t + \beta_8 M_t + \beta_9 M \Delta \ln P_t + \Delta \varepsilon$$

```

# First, add Z column to the data frame:

salaries_new_df <- salaries_new_df %>%
  mutate(
    M = case_when(
      Year < 1980 ~ 0,
      Year >= 1980 ~ 1
    ),
    MDlnP = M * DlnP
  )

# Then run the regression:

lm_Q3j <- lm(DlnW ~ DlnP + M + MDlnP, salaries_new_df)
summary(lm_Q3j)

##
## Call:
## lm(formula = DlnW ~ DlnP + M + MDlnP, data = salaries_new_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.053529 -0.016011 -0.000561  0.018882  0.037861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.036104   0.008534   4.230 0.000192 ***
## DlnP         0.866873   0.100856   8.595 1.04e-09 ***
## M           0.020680   0.019471   1.062 0.296387
## MDlnP       -0.337055   0.256127  -1.316 0.197834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02537 on 31 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.7182, Adjusted R-squared:  0.6909
## F-statistic: 26.33 on 3 and 31 DF,  p-value: 1.164e-08

```

Notice that there is not a  $MlnP$  term because if we include it we would have perfect multicollinearity. From the regression output we can see that neither  $M$  nor  $MDlnP$  are significant - see their t-stats, so we can conclude that there is not any evidence of structural break for these years.

---

(k) Discuss the advantages of using the method of dummy variables over the Chow Test you used in part 1(a).

**Answer:** With the dummy variable we find out which variable or variables have changed. Whereas Chow test tells us only that something has changed. That is, it will tell us if the two sub-regressions are different but not whether the difference is due to the intercepts, the slope, or both.