Econometrics II
Tutorial No. 5

Lennart Hoogerheide & Agnieszka Borowska

15.03.2017

Outline

Summary

Key terms

- **RESET:** A general test for functional form in a multiple regression model; it is an F test of joint significance of the estimated coefficients at the squares, cubes, and perhaps higher powers of the fitted values from the initial OLS estimation.
- **Chow Statistic:** An F statistic for testing the equality of regression parameters across different groups (say, men and women) or time periods (say, before and after a policy change).

Key terms – cont'd

- **Dummy Variable:** A variable that takes on the value zero or one.
- **Dummy Variable Trap:** The mistake of including too many dummy variables among the independent variables; it occurs when an overall intercept is in the model and a dummy variable is included for each group.

Key terms – cont'd

- **Interaction Term:** An independent variable in a regression model that is the product of two explanatory variables.
- **Intercept Shift:** The intercept in a regression model differs by group or time period.

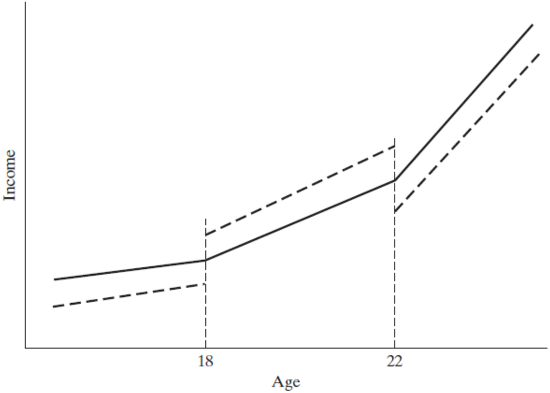
Extra topic: Piecewise linear regression

Piecewise linear regression

Consider modelling income data for individuals of varying ages in a population. Certain patterns with regard to some age *thresholds* will be clearly evident. In general, income will be rising with age, but the slope (i.e. marginal increase) might change at some distinct milestones. For example, the typical individual

1. at age 18 graduates from high school;
2. at age 22 graduates from college/university.

Time profile of income for the typical individual in this population:



How to model such a phenomenon?

- We could fit a linear regression model to each of three subsamples separately.
  This, however, would most likely lead to a discontinuous function (the dashed line in the figure), which is not in line with our assumed pattern of the time profile of income.
- We could use dummy variables, as they can also be used to model **varying slope parameters**.
  Notice that we want to estimate

$$\mathbb{E}(income|age) = \begin{cases} \underline{a_0 + b_0\, age} & \text{if } age < 18, \\ a_1 + b_1\, age & \text{if } age \in [18, 22), \\ a_2 + b_2\, age & \text{if } age \geq 22. \end{cases} \qquad (1)$$

The threshold values (here 18 and 22) are often referred to as **knots** in this context. Define two dummies:

$$D_1 = \mathbb{I}_{\{age \geq 18\}} = \begin{cases} 0 & \text{if } age < 18, \\ 1 & \text{if } age \geq 18, \end{cases}$$

$$D_2 = \mathbb{I}_{\{age \geq 22\}} = \begin{cases} 0 & \text{if } age < 22, \\ 1 & \text{if } age \geq 22. \end{cases}$$

Now, we can combine three parts in (1) as follows:

$$income = \underline{\beta_0 + \beta_1\, age} + \delta_1 D_1 + \gamma_1 D_1 \cdot age + \delta_2 D_2 + \gamma_2 D_2 \cdot age + \varepsilon. \tag{2}$$

where we can see the underlined part as the "baseline" case. Explicitly rewritten, it becomes:

$$income = \begin{cases} \beta_0 + \beta_1\, age + \varepsilon & \text{if } age < 18, \\ \beta_0 + \beta_1\, age + \delta_1 + \gamma_1\, age + \varepsilon & \text{if } age \in [18, 22), \\ \beta_0 + \beta_1\, age + \delta_1 + \gamma_1\, age + \delta_2 + \gamma_2\, age + \varepsilon & \text{if } age \geq 22. \end{cases}$$

$$income = \begin{cases} \beta_0 + \beta_1\, age + \varepsilon & \text{if } age < 18, \\ \beta_0 + \delta_1 + \left(\beta_1 + \gamma_1\right) age + \varepsilon & \text{if } age \in [18, 22), \\ \beta_0 + \delta_1 + \delta_2 + \left(\beta_1 + \gamma_1 + \gamma_2\right) age + \varepsilon & \text{if } age \geq 22, \end{cases}$$

The intercepts in the three segments are: β_0, $\beta_0 + \delta_1$ and $\beta_0 + \delta_1 + \delta_2$, while the slopes are β_1, $\beta_1 + \gamma_1$ and $\beta_1 + \gamma_1 + \gamma_2$.

So most likely we will still end up with the dashed line! Hence, simply employing the dummies would not help in solving the problem of discontinuity from the previous point!

To make the function *continuous* we need to impose that that its value in two adjacent segment is equal in the separating knot (so, simply speaking, that "segments join at the knots").

Hence:

$$\begin{cases} [\beta_0 + \beta_1\, age]\big|_{age=18} & = [\beta_0 + \delta_1 + \left(\beta_1 + \gamma_1\right) age]\big|_{age=18}\,, \\ [\beta_0 + \delta_1 + \left(\beta_1 + \gamma_1\right) age]\big|_{age=22} & = [\beta_0 + \delta_1 + \delta_2 + \left(\beta_1 + \gamma_1 + \gamma_2\right) age]\big|_{22}\,, \end{cases}$$

$$\begin{cases} \beta_0 + \beta_1 \cdot 18 & = \beta_0 + \delta_1 + \left(\beta_1 + \gamma_1\right) \cdot 18, \\ \beta_0 + \delta_1 + \left(\beta_1 + \gamma_1\right) \cdot 22 & = \beta_0 + \delta_1 + \delta_2 + \left(\beta_1 + \gamma_1 + \gamma_2\right) \cdot 22 \end{cases}$$

$$\begin{cases} 0 & = \delta_1 + \gamma_1 \cdot 18, \\ 0 & = \delta_2 + \gamma_2 \cdot 22, \end{cases}$$

This means that we need for the dummy coefficients:

$$\delta_1 = -\gamma_1 \cdot 18,$$
$$\delta_2 = -\gamma_2 \cdot 22.$$

When we plug this in the original regression (2), we obtain:

$$\begin{aligned} income =& \beta_0 + \beta_1\, age + (-\gamma_1 \cdot 18) D_1 + \gamma_1 D_1 \cdot age \\ & + (-\gamma_2 \cdot 22) D_2 + \gamma_2 D_2 \cdot age + \varepsilon \\ =& \beta_0 + \beta_1 \underbrace{age}_{=:x_1} + \gamma_1 D_1 \cdot \underbrace{(age - 18)}_{=:x_2} \\ & + \gamma_2 D_2 \cdot \underbrace{(age - 22)}_{=:x_3} + \varepsilon. \end{aligned}$$

Constrained least squares estimates are obtained by multiple regression, using a constant and the variables x_1, x_2 and x_3.

Notice that the latter two need to be obviously multiplied by the corresponding dummies, so that they only affect the relevant segments.

We can test the hypothesis that the slope of the function is constant with the joint test of the two restrictions $\gamma_1 = 0$ and $\gamma_2 = 0$.

Warm-up Exercises

RESET 1

Can you include $\hat{y}_i$ as an explanatory variable in the test regression of the RESET? What would happen then?

That would not make sense! Recall that

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik},$$

and for the RESET we consider

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3 + u_i,$$

where we test the null $H_0 : \delta_1 = \delta_2 = 0$.

Then if we additionally include $\hat{y}_i$ as an explanatory variable in the test regression we obtain

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \underline{\delta_0 \hat{y}_i} + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3 + u_i \\ =& \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \underline{\delta_0 \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik} \right)} \\ & + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3 + u_i, \end{aligned}$$

so that testing the insignificance of δ_0 is equivalent to testing the insignificance of the original regression model (all its variables at the same time).

Also, we **could not** perform OLS in the test regression, because $\hat{y}_i$ is a linear combination of the explanatory variables, which leads to **perfect multicollinearity**:

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \delta_0 \big(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}\big) \\ & + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3 + u_i. \end{aligned}$$

RESET 2

Consider a regression with a constant term and a single variable x_i. *What does the RESET specification look like in this case?*

In this simple case we have

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i,$$

with the fitted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}.$$

Hence, the RESET becomes

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3 + u_i \\
&= \beta_0 + \beta_1 x_{i1} + \delta_1 \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}\right)^2 + \delta_2 \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}\right)^3 + u_i \\
&= \beta_0 + \beta_1 x_{i1} + \delta_1 \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}\right)^2 + \delta_2 \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}\right)^3 + u_i \\
&= \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i1}^2 + \tilde{\beta}_3 x_{i1}^3 + u_i
\end{aligned}$$

RESET 3

How to check for both non-linearity and heteroskedasticity?

In case of heteroskedasticity the tests for other features (like RESET and Chow test) need to take this into account, so that White standard errors need to be used in the test regression then.

So, the ordering to test for both non-linearity and heteroskedasticity would then:

1. run the RESET with White standard errors (and correct a potential functional misspecification);
2. run e.g. the Breusch-Pagan test for heteroskedasticity (for the model that has the correct specification for the conditional mean of y given x).

Dummy variables 1

Explain the dummy variable trap.

Including in the regression model a constant term and dummy variables for all categories. This introduces perfect collinearity because one of the categories can be expressed as a perfect linear function of the remaining categories and the constant term.

Or in other words: the sum of the dummy variables for all categories is equal to the constant term 1, because every observation belongs to exactly 1 group.

Dummy variables 2

Let d be a dummy variable and let z be a quantitative variable. Consider the model

$$y = \beta_0 + \delta_0 d + \beta_1 z + \delta_1 d \cdot z + u,$$

which is a general version of a model with an interaction between a dummy variable and a quantitative variable.

Dummy variables 2(a)

Give the relationship between y and z as a function of d.

When $d = 0$ we have

$$\begin{aligned} y &= \beta_0 + \delta_0 \cdot 0 + \beta_1 z + \delta_1 \cdot 0 \cdot z + u \\ &= \beta_0 + \beta_1 z + u, \end{aligned}$$

while when $d = 1$ we have

$$\begin{aligned} y &= \beta_0 + \delta_0 \cdot 1 + \beta_1 z + \delta_1 \cdot 1 \cdot z + u \\ &= \big(\beta_0 + \delta_0\big) + \big(\beta_1 + \delta_1\big) z + u. \end{aligned}$$

Dummy variables 2(b)

Give the relationship between the expected value of y *and* z *as a function of* d*. Give a geometric interpretation of the results.*

We simply set the error term to zero as $\mathbb{E}(u|d,x) = 0$. Then using the results from the previous point we obtain for $d = 0$

$$\mathbb{E}(y|d,x) = \beta_0 + \beta_1 z,$$

while for $d = 1$

$$\mathbb{E}(y|d,x) = \big(\beta_0 + \delta_0\big) + \big(\beta_1 + \delta_1\big)z.$$

We can see that these are simply two linear functions in z.

Dummy variables 2(c)

Assume that $\delta_1 \neq 0$*. What does this assumption mean? Find* z^**, a value of* z *such that the conditional expectation of* y *given* z *and given* $d = 0$ *is equal to the conditional expectation of* y *given* z *and given* $d = 1$*. When is* z^* *positive?*

When $\delta_1 \neq 0$ then two lines from the previous point are not parallel. Then, at z^* they intersect:

$$\begin{aligned}
\mathbb{E}(y|d=0,x) &= \mathbb{E}(y|d=1,x), \\
\beta_0 + \beta_1 z &= \big(\beta_0 + \delta_0\big) + \big(\beta_1 + \delta_1\big) z, \\
\big(\beta_1 - \beta_1 - \delta_1\big) z &= -\beta_0 + \beta_0 + \delta_0, \\
\delta_1 z &= -\delta_0, \\
z &= -\frac{\delta_0}{\delta_1}.
\end{aligned}$$

Obviously, z^* is positive if and only if δ_0 and δ_1 have opposite signs.

Dummy variables 2(d)

Suppose that we have estimated the following model

$$\hat{y} = 2.289 - 0.357 female + 0.50 educ + 0.030 female \cdot educ$$

where y is the log wage using, $female$ is a gender dummy and $educ$ is the number of total years of education. Use the above equation to find the value of educ such that the predicted values of log wage are the same for men and women.

Using the result from the previous point we obtain:

$$educ^* = -\frac{-0.357}{0.030} = 11.9.$$

Dummy variables 2(e)

Based on the equation in part (d), can women realistically get enough years of college so that their earnings catch up to those of men? Explain.

The estimated years of college where women catch up to men of almost 12 years is much too high to be practically relevant.

While the estimated coefficient on $female \cdot educ$ shows that the gap is reduced at higher levels of eduction, it is never closed – not even close. In fact, at four years of college, the difference in predicted log wage is still

$$-0.357 + 0.030 \cdot 4 = -0.237$$

less for women.

Small Computer Exercise

Generate a sample of size 100 *from the model* $y_i = 2 + \sqrt{x_i} + \varepsilon_i$, *where* x_i *are independent and uniformly distributed on the interval* $[0, 20]$ *and the* ε_i *are independent and distributed as* $\mathcal{N}(0, 0.01)$. *Regress* y *on a constant and* x. *Perform a RESET.*

Dependent Variable: Y
Method: Least Squares
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.228553	0.045917	26.75607	0.0000
X	0.175351	0.003998	43.86014	0.0000

R-squared	0.951526	Mean dependent var	2.972152
Adjusted R-squared	0.951032	S.D. dependent var	1.038389
S.E. of regression	0.229783	Akaike info criterion	-0.083564
Sum squared resid	5.174430	Schwarz criterion	-0.031461
Log likelihood	6.178196	Hannan-Quinn criter.	-0.062477
F-statistic	1923.712	Durbin-Watson stat	1.991067
Prob(F-statistic)	0.000000		

Ramsey RESET Test
Equation: EQ
Specification: Y C X
Omitted Variables: Powers of fitted values from 2 to 3

	Value	df	Probability
F-statistic	178.6239	(2, 96)	0.0000
Likelihood ratio	155.2091	2	0.0000

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	4.078462	2	2.039231
Restricted SSR	5.174430	98	0.052800
Unrestricted SSR	1.095968	96	0.011416

LR test summary:

	Value	df
Restricted LogL	6.178196	98
Unrestricted LogL	83.78274	96

Unrestricted Test Equation:
Dependent Variable: Y
Method: Least Squares
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.828686	0.121918	14.99930	0.0000
X	0.750358	0.058806	12.75981	0.0000
FITTED^2	-0.922465	0.116923	-7.889521	0.0000
FITTED^3	0.078278	0.012864	6.085098	0.0000

R-squared	0.989733	Mean dependent var	2.972152
Adjusted R-squared	0.989412	S.D. dependent var	1.038389
S.E. of regression	0.106847	Akaike info criterion	-1.595655
Sum squared resid	1.095968	Schwarz criterion	-1.491448
Log likelihood	83.78274	Hannan-Quinn criter.	-1.553480
F-statistic	3084.791	Durbin-Watson stat	1.732123

Recall that the null for the RESET is that the functional specification is correct. The obtained value of the F test statistic is 178.62 and under the null it follows the $F(2, n-k-3)$ distribution. The corresponding p-value is 0, so at any significance level we can reject the null.

This shows that the RESET is flexible enough to detect various forms of non-linearity, including roots of variables (and not only their powers).

Note that this is the built-in version of the RESET in EViews, which assumes homoskedasticity.

Computer Exercises

Exercise 1

For the data **`bankwages.wf1`** *consider the model*

$$y_i = \alpha + \gamma D_{gi} + \mu D_{mi} + \beta x_i + \varepsilon_i, \qquad (3)$$

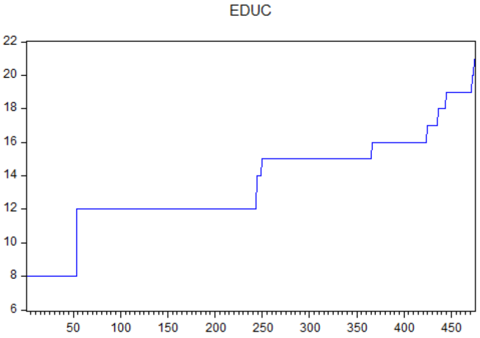
where y_i *is the logarithm of yearly wage,* D_g *is a gender dummy (*1 *for males,* 0 *for females),* D_m *is a minority dummy (*1 *for minorities,* 0 *otherwise) and* x_i *is the number of completed years of education. The education ranges from* 8 *to* 21 *years. The* $n = 474$ *employees in the sample are ordered according to the values of* x*, starting with the lowest education:*

- *those with ranking number* 365 *or lower have at most* 15 *years of education (*$x \leq 15$*);*
- *those with ranking number* 366–424 *have exactly* 16 *years of education (*$x = 16$*);*
- *those with ranking number* 425 *or higher have over* 16 *years of education (*$x \geq 17$*).*

Exercise 1(i)

Test whether an additional year of education gives the same relative increase in wages for lower and higher levels of education (i.e. investigate the marginal effect of β of education on salary).

To this end, perform the Chow tests on parameter variations in (3)*, where the break point is at observation* 425 *(with education at least* 17 *years). Check the outcomes on a break.*



We run the OLS on the full sample of $n = 474$ employees and on two subsamples, of $n_1 = 424$ employees with $x \leq 16$ and of $n_2 = 50$ employees with $x > 16$. We run the OLS on the full sample of $n = 474$ employees and on two subsamples, of $n_1 = 424$ employees with $x \leq 16$ and of $n_2 = 50$ employees with $x > 16$.

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.199980	0.058687	156.7634	0.0000
GENDER	0.261131	0.025511	10.23594	0.0000
MINORITY	-0.132673	0.028946	-4.583411	0.0000
EDUC	0.077366	0.004436	17.44229	0.0000

R-squared	0.586851	Mean dependent var	10.35679
Adjusted R-squared	0.584214	S.D. dependent var	0.397334
S.E. of regression	0.256207	Akaike info criterion	0.122741
Sum squared resid	30.85177	Schwarz criterion	0.157857
Log likelihood	-25.08970	Hannan-Quinn criter.	0.136552
F-statistic	222.5344	Durbin-Watson stat	1.347522
Prob(F-statistic)	0.000000		

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474 IF EDUC<=16
Included observations: 424

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.463702	0.063095	149.9906	0.0000
GENDER	0.229931	0.023801	9.660543	0.0000
MINORITY	-0.111687	0.027462	-4.066947	0.0001
EDUC	0.055783	0.004875	11.44277	0.0000

R-squared	0.426202	Mean dependent var	10.27088
Adjusted R-squared	0.422103	S.D. dependent var	0.310519
S.E. of regression	0.236055	Akaike info criterion	-0.040113
Sum squared resid	23.40327	Schwarz criterion	-0.001908
Log likelihood	12.50392	Hannan-Quinn criter.	-0.025018
F-statistic	103.9882	Durbin-Watson stat	1.408086
Prob(F-statistic)	0.000000		

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474 IF EDUC>16
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.953242	0.743176	13.39284	0.0000
GENDER	0.830174	0.263948	3.145213	0.0029
MINORITY	-0.346533	0.126096	-2.748175	0.0085
EDUC	0.019132	0.041108	0.465418	0.6438

R-squared	0.302888	Mean dependent var	11.08534
Adjusted R-squared	0.257424	S.D. dependent var	0.293434
S.E. of regression	0.252861	Akaike info criterion	0.164663
Sum squared resid	2.941173	Schwarz criterion	0.317624
Log likelihood	-0.116564	Hannan-Quinn criter.	0.222911
F-statistic	6.662173	Durbin-Watson stat	2.007423
Prob(F-statistic)	0.000788		

The Chow F-test statistic is given by

$$F = \frac{\frac{SSR_0 - SSR_1 - SSR_2}{k}}{\frac{SSR_1 + SSR_2}{n_1 + n_2 - 2k}} \overset{H_0}{\sim} F(k, n_1 + n_2 - 2k),$$

where the null is no break at the chosen point.

The statistic from the lecture is a special case of that general statistic!

The Chow test from the lecture was an F-test with the null that the coefficients at the dummy and its product with x are 0. Under homoskedasticity it can be shown that that F statistic is equal to the one here.

The one here makes sense: it is large if SSR_0 is much larger than $SSR_1 + SSR_2$ – which happens if the quality of the model becomes much worse if we force the coefficients to be the same among the two groups.

Plugging in the regression results we obtain

$$F = \frac{\frac{30.852-23.403-2.941}{4}}{\frac{23.403+2.941}{424+50-8}} = 19.932 \overset{H_0}{\sim} F(4, 466),$$

with the corresponding p-value of 0. At any significance level the null hypothesis (that all four coefficients are equal among the two groups) is clearly rejected.

Hence, the Chow test confirms that there is a break at observation 425 in the marginal effect β of education (and the constant term, gender and minority) on salaries.

Alternatively, we can use the EViews built-in test `chow 425`, which obviously gives the same result.

Chow Breakpoint Test: 425
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1 474

F-statistic	19.93222	Prob. F(4,466)	0.0000
Log likelihood ratio	74.86199	Prob. Chi-Square(4)	0.0000
Wald Statistic	79.72890	Prob. Chi-Square(4)	0.0000

Exercise 1(ii)

Check the effect of changing of the break point: now set it at observation 366 *(with education at least* 16 *years). Perform the Chow tests on parameter variations in* (3) *and check the outcomes on a break. Compare the results with these from (i).*

The original regression stays at it was, but the two subsample regressions, for the $n_1 = 365$ employees with $x < 16$ and of $n_2 = 109$ employees with $x \geq 16$, are now given by:

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474 IF EDUC<16
Included observations: 365

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.761619	0.055637	175.4532	0.0000
GENDER	0.228921	0.020448	11.19511	0.0000
MINORITY	-0.069008	0.022853	-3.019604	0.0027
EDUC	0.027696	0.004495	6.161612	0.0000

R-squared	0.371390	Mean dependent var	10.19693
Adjusted R-squared	0.366166	S.D. dependent var	0.235566
S.E. of regression	0.187543	Akaike info criterion	-0.498720
Sum squared resid	12.69721	Schwarz criterion	-0.455982
Log likelihood	95.01645	Hannan-Quinn criter.	-0.481735
F-statistic	71.09422	Durbin-Watson stat	1.932875
Prob(F-statistic)	0.000000		

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474 IF EDUC>=16
Included observations: 109

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.322807	0.344479	27.06350	0.0000
GENDER	0.339568	0.069099	4.914234	0.0000
MINORITY	-0.317896	0.084394	-3.766811	0.0003
EDUC	0.078384	0.021137	3.708366	0.0003

R-squared	0.441430	Mean dependent var	10.89210
Adjusted R-squared	0.425471	S.D. dependent var	0.358938
S.E. of regression	0.272066	Akaike info criterion	0.270466
Sum squared resid	7.772112	Schwarz criterion	0.369231
Log likelihood	-10.74038	Hannan-Quinn criter.	0.310519
F-statistic	27.66002	Durbin-Watson stat	1.958952
Prob(F-statistic)	0.000000		

The new value of the Chow statistic is given by

$$F = \frac{\frac{30.852-12.697-7.772}{4}}{\frac{12.697+7.772}{365+109-8}} = 58.524 \overset{H_0}{\sim} F(4, 466),$$

with the corresponding p-value of 0. So again, we reject the null that there is no change in the marginal effect β of education on salaries, but this time at observation 366.

Chow Breakpoint Test: 365
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1 474

F-statistic	58.52394	Prob. F(4,466)	0.0000
Log likelihood ratio	192.9329	Prob. Chi-Square(4)	0.0000
Wald Statistic	234.0958	Prob. Chi-Square(4)	0.0000

Exercise 1(iii)

Formulate a model with two different values of β in (3): one for education levels less than 16 years (observations $i \leq 365$) and another for education levels of 16 years or more (observations $i > 366$). Estimate this model, and give an interpretation of the outcomes.

[Hint: think how to make the expected log wage a continuous function of education.]

We know that dummy variables are a helpful tool to remove parameter variation. So we need to work with dummies. But how? Possibly, we could think of the following three cases.

Case 1

Including dummies for low and high levels of education:

$$y_i = \alpha + \gamma D_{gi} + \mu D_{mi} + \beta^* x_i + \beta^*_{low} \mathbb{I}_{\{x_i < 16\}} + \beta^*_{high} \mathbb{I}_{\{x_i \geq 16\}} + \varepsilon_i.$$

Case 1

Obviously, this is a dummy variable trap!

Case 2

Considering the low and high levels of education separately:

$$y_i = \alpha + \gamma D_{gi} + \mu D_{mi} + \beta_{low}\mathbb{I}_{\{x_i<16\}} \cdot x_i + \beta_{high}\mathbb{I}_{\{x_i\geq 16\}} \cdot x_i + \varepsilon_i.$$

Case 2

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.738164	0.061737	157.7357	0.0000
GENDER	0.252722	0.021307	11.86109	0.0000
MINORITY	-0.105051	0.024243	-4.333192	0.0000
EDUC*EDU_LOW	0.029301	0.004997	5.864159	0.0000
EDUC*EDU_HIGH	0.056925	0.003968	14.34436	0.0000

R-squared	0.712639	Mean dependent var	10.35679
Adjusted R-squared	0.710188	S.D. dependent var	0.397334
S.E. of regression	0.213901	Akaike info criterion	-0.236110
Sum squared resid	21.45856	Schwarz criterion	-0.192215
Log likelihood	60.95802	Hannan-Quinn criter.	-0.218847
F-statistic	290.7737	Durbin-Watson stat	1.917430
Prob(F-statistic)	0.000000		

This has a disadvantage that the expected log wage is not a continuous function of education.

Case 3

Considering the additional effect of the high level of education:

$$y_i = \alpha + \gamma D_{gi} + \mu D_{mi} + \beta x_i + \beta_{high} \mathbb{I}_{\{x_i \geq 16\}} \cdot (x_i - 16) + \varepsilon_i.$$

Case 3

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.438310	0.064279	146.8329	0.0000
GENDER	0.241536	0.024315	9.933567	0.0000
MINORITY	-0.119494	0.027483	-4.348004	0.0000
EDUC	0.057855	0.004963	11.65735	0.0000
(EDUC-16)*EDU_HIGH	0.125909	0.017039	7.389416	0.0000

R-squared	0.629936	Mean dependent var	10.35679
Adjusted R-squared	0.626779	S.D. dependent var	0.397334
S.E. of regression	0.242739	Akaike info criterion	0.016829
Sum squared resid	27.63442	Schwarz criterion	0.060723
Log likelihood	1.011558	Hannan-Quinn criter.	0.034092
F-statistic	199.5867	Durbin-Watson stat	1.489866
Prob(F-statistic)	0.000000		

Here we make the log wage a continuous, **piecewise-linear**, function of education.
We can see that, as expected, there is a "bonus" from having a high level of education, here defined as at least 16 years of education. The increase in the slope in the hight education segment is 0.126 and it is statistically significant.

Exercise 1(iv)

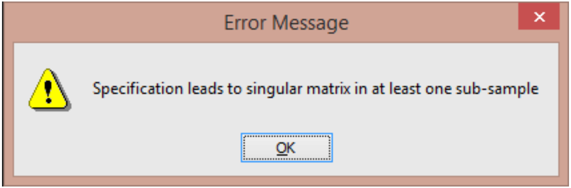
Perform a sequence of Chow break tests for all segments where the variable 'education' changes. Notice that this variable takes on ten different values, so that there are nine possible break points. Comment on the outcomes.

We have ten segments split by the level of education:

$$x_i = \begin{cases} 8 & i = 1, \dots, 53, \\ 12 & i = 54, \dots, 243, \\ 14 & i = 244, \dots, 249, \\ 15 & i = 250, \dots, 365, \\ 16 & i = 366, \dots, 424, \\ 17 & i = 425, \dots, 435, \\ 18 & i = 436, \dots, 444, \\ 19 & i = 445, \dots, 471, \\ 20 & i = 472, 473, \\ 21 & i = 474, \end{cases}$$

which indeed indicates 9 possible break point.

Notice, however, that if we cannot to perform the Chow test for any observation with $i \leq 53$ as then the education variable for one subsample is constant $x_i = 8$ – which obviously results in the following error:



We also cannot perform the Chow test for observations with $i \geq 427$, as then the gender variable is constant, which leads to the same problem.

Hence, we can only effectively consider two more break points (in addition to the two previously analysed): $i = 244$ (with $x_i \geq 14$ for $i \geq 244$) and $i = 250$ (with $x_i \geq 15$ for $i \geq 250$).

In both case the p-value for the test F statistic is zero, so we have at any significance level we can reject the null about no break at the given point.

Chow Breakpoint Test: 244
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1 474

F-statistic	36.96346	Prob. F(4,466)	0.0000
Log likelihood ratio	130.6208	Prob. Chi-Square(4)	0.0000
Wald Statistic	147.8538	Prob. Chi-Square(4)	0.0000

Chow Breakpoint Test: 250
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1 474

F-statistic	37.05569	Prob. F(4,466)	0.0000
Log likelihood ratio	130.9056	Prob. Chi-Square(4)	0.0000
Wald Statistic	148.2228	Prob. Chi-Square(4)	0.0000

Interestingly, it turns that for any "admissible" point (i.e. $i = 55, \ldots, 426$) the conclusion from the test is the same!

This shows that the Chow test is rather robust in this case, in the sense that its rejection does not depend much on where to put the "border" between the groups.

This finding points at particular features of the current dataset, because:

1. the null hypothesis is so "heavily" violated;

   [If the null was only be "slightly" violated, then it might matter where the "border" is chosen. Then it may be important to choose the "border" somewhere close to the median, to have two groups of a reasonable (similar) size.]

2. because the number of observations is not small.

   [The test results also depend on the total number of observations. If the number of observations is small, then it may be more important to choose the "border" somewhere close to the median, to have two groups of a reasonable (approximately equal) size.]

Exercise 2

Consider data in `coffee.wf1` *on weekly coffee sales for one brand. There are* $n = 12$ *weekly observations for the weeks when marketing actions were taken. In particular, there were six weeks with price reductions without advertisement, and six weeks with joint price reductions and advertisement. As there are no advertisements without simultaneous price reductions, we formulate the model*

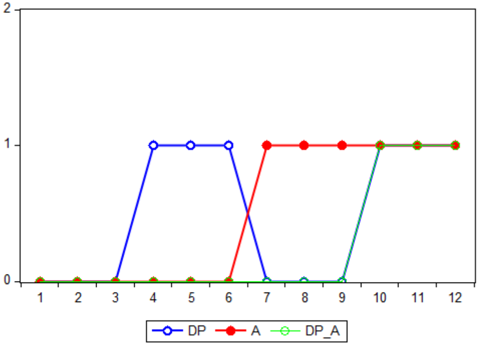
$$y = \beta_1 + \beta_2 D_p + \beta_3 D_a + \beta_4 D_p D_a + \varepsilon,$$

where y *denotes the logarithm of weekly sales,* D_p *is a dummy variable with the value* 0 *if the price reduction is* 5% *and the value* 1 *if this reduction is* 15%*, and* D_a *is a dummy variable that is* 0 *if there is no advertisement and* 1 *if there is advertisement.*

Exercise 2(i)

Give an economic motivation for the above model. Estimate this model and test the null hypothesis that $\beta_2 = 0$. What is the p-value of this test?

Both dummies, D_p and D_a, as well as their product, D_pD_a over time:



- A price reduction of some type (low or high) was always on, so we only consider the additional effect of a big price cut (i.e. by 15%).
- We know that there are no advertisements without simultaneous price reductions (of any type).
- The sales are expected to increase when there is a big price reduction or when there is advertisement.
- When there are both, a big price reduction and advertisement, the sales are likely to increase even more, as more people will consider to buy the product due to the advertisement, and the more will be likely to actually purchase it due to the lower price.

Then:

- the dummy D_p measures an additional effect of the big price cut when there is no advertisement;
- the dummy D_a captures the effect of advertisement given there is the small price reduction;
- the product of dummies D_pD_a shows the joint effect of advertisement and the big price reduction, so the extra effect of advertisement when there is the big price cut.

Notice that the sum of the coefficients for D_p and D_pD_a measures the impact of the big price reduction when advertisement is launched.

Dependent Variable: LOGQ
Method: Least Squares
Sample: 1 12
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.175711	0.042136	146.5667	0.0000
DP	0.280838	0.059589	4.712916	0.0015
A	0.188664	0.059589	3.166085	0.0133
DP*A	0.280319	0.084272	3.326368	0.0104

R-squared	0.955505	Mean dependent var	6.480542
Adjusted R-squared	0.938819	S.D. dependent var	0.295056
S.E. of regression	0.072981	Akaike info criterion	-2.136023
Sum squared resid	0.042610	Schwarz criterion	-1.974387
Log likelihood	16.81614	Hannan-Quinn criter.	-2.195866
F-statistic	57.26474	Durbin-Watson stat	1.410736
Prob(F-statistic)	0.000009		

The t-statistic for this test is

$$\frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{0.2808 - 0}{0.0596} = 4.7114 \overset{H_0}{\sim} t_{n-k} = t_{12-4} = t_8,$$

with the corresponding p-value of 0.0015. So at any conventional significance level we can reject the null and conclude that β_2 is significantly different from zero.

Exercise 2(ii)

Estimate the above model, replacing D_a by the alternative dummy variable D_a^*, which has the value 0 if there is advertisement and 1 if there is not. The model then becomes

$$y = \beta_1^* + \beta_2^* D_p + \beta_3^* D_a^* + \beta_4^* D_p D_a^* + \varepsilon,$$

Compare the estimated price coefficient and its t-value and p-value with the results obtained in (i).

Dependent Variable: LOGQ
Method: Least Squares
Sample: 1 12
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.364375	0.042136	151.0442	0.0000
DP	0.561157	0.059589	9.417111	0.0000
A_STAR	-0.188664	0.059589	-3.166085	0.0133
DP*A_STAR	-0.280319	0.084272	-3.326368	0.0104

R-squared	0.955505	Mean dependent var	6.480542
Adjusted R-squared	0.938819	S.D. dependent var	0.295056
S.E. of regression	0.072981	Akaike info criterion	-2.136023
Sum squared resid	0.042610	Schwarz criterion	-1.974387
Log likelihood	16.81614	Hannan-Quinn criter.	-2.195866
F-statistic	57.26474	Durbin-Watson stat	1.410736
Prob(F-statistic)	0.000009		

The t-statistic for this test is

$$\frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{0.5811 - 0}{0.0596} = 9.4171 \overset{H_0}{\sim} t_{n-k} = t_{12-4} = t_8,$$

with the corresponding p-value of 0. So again at any conventional significance level we can reject the null and conclude that β_2 is significantly different from zero.

Exercise 2(ii)

Explain why the two results for the price dummy differ in (i) and (ii). Discuss the relevance of this fact for the interpretation of coefficients of dummy variables in regression models.

With dummy variables you always choose one category as the **reference category**.

- The estimate for the constant term refers to the expectation of the dependent variable when the dummy is "switched-off", so for the non-reference category.
- The estimate for the coefficient for the dummy itself shows the average additional effect from "switching-on" the dummy.
- So the sum of the estimate for the constant and the estimate for the coefficient for the dummy describe the expected effect for the reference category.

Hence, if you change the reference category as above:

- the estimates for the constant term and for the remaining variables (which do not include the dummy) will change accordingly (here, for D_p);
- the signs the variables 'related' to the dummy will change (here, for D_a and $D_p D_a$);
- however, the measures for the whole model (like e.g. R^2 and the fitted values $\hat{y}_i$ for all observations) will not be affected: this is still "the same model", only with a different interpretation of the coefficients.