Elements of Econometrics. Lecture 8. Dummy Variables.

FCS, 2022-2023

Dummy Variables

- A dummy (or binary) variable is a variable that takes the value 1 or 0
- Examples: male (= 1 if being male, 0 otherwise), having PhD degree (= 1 if yes, 0 otherwise), vaccinated (=1 if yes, 0 otherwise), etc.

Dummy variables allow to reflect qualitative information.

Consider a model with one standard variable (X) and one dummy (D)

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 D + u$$

This can be interpreted as an intercept shift

If D = 0, then $Y = \beta_1 + \beta_2 X_2 + u$

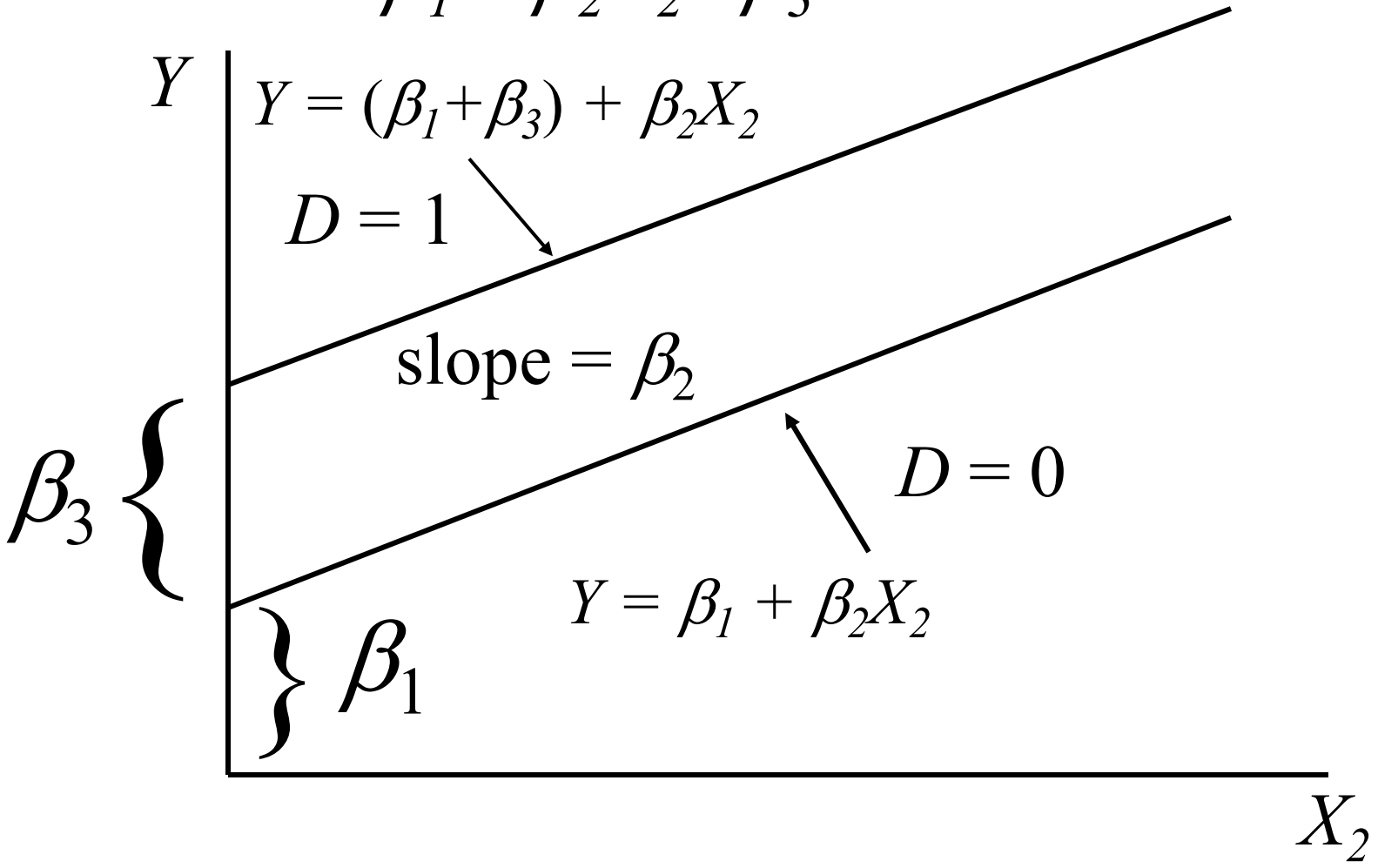
If D = 1, then $Y = (\beta_1 + \beta_3) + \beta_2 X_2 + u$

The observations with D=0 are the reference (or base) category

MODEL WITH A DUMMY VARIABLE D FOR INTERCEPT

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 D$$

Example of $\beta_3 > 0$



MODEL WITH DUMMY VARIABLE MALE
FEMALE is the REFERENCE CATEGORY

Dependent Variable: EARN
Method: Least Squares
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-26.32962	4.984233	-5.282582	0.0000
S	1.723689	0.249993	6.894949	0.0000
ASVABC	0.165788	0.048232	3.437296	0.0006
MALE	4.105127	0.720810	5.695156	0.0000
PWE	0.412816	0.168736	2.446514	0.0147

R-squared	0.236233	Mean dependent var	13.68988
S.D. dependent var	9.702960	S.E. of regression	8.509745
Sum squared resid	40914.91	F-statistic	43.68865
Durbin-Watson stat	1.881990	Prob(F-statistic)	0.000000

$$\widehat{EARNINGS} = -26.32 + 1.72S + 0.17ASVABC + 4.11MALE + 0.41PWE$$

MORE THAN ONE GROUP OF DUMMIES: EARNINGS FUNCTION WITH DUMMY VARIABLES MALE AND ETHNW (non-white); FEMALE and WHITE are the REFERENCE CATEGORIES

$$\widehat{EARNINGS} = \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 MALE + \beta_5 PWE + \beta_6 ETHNW$$

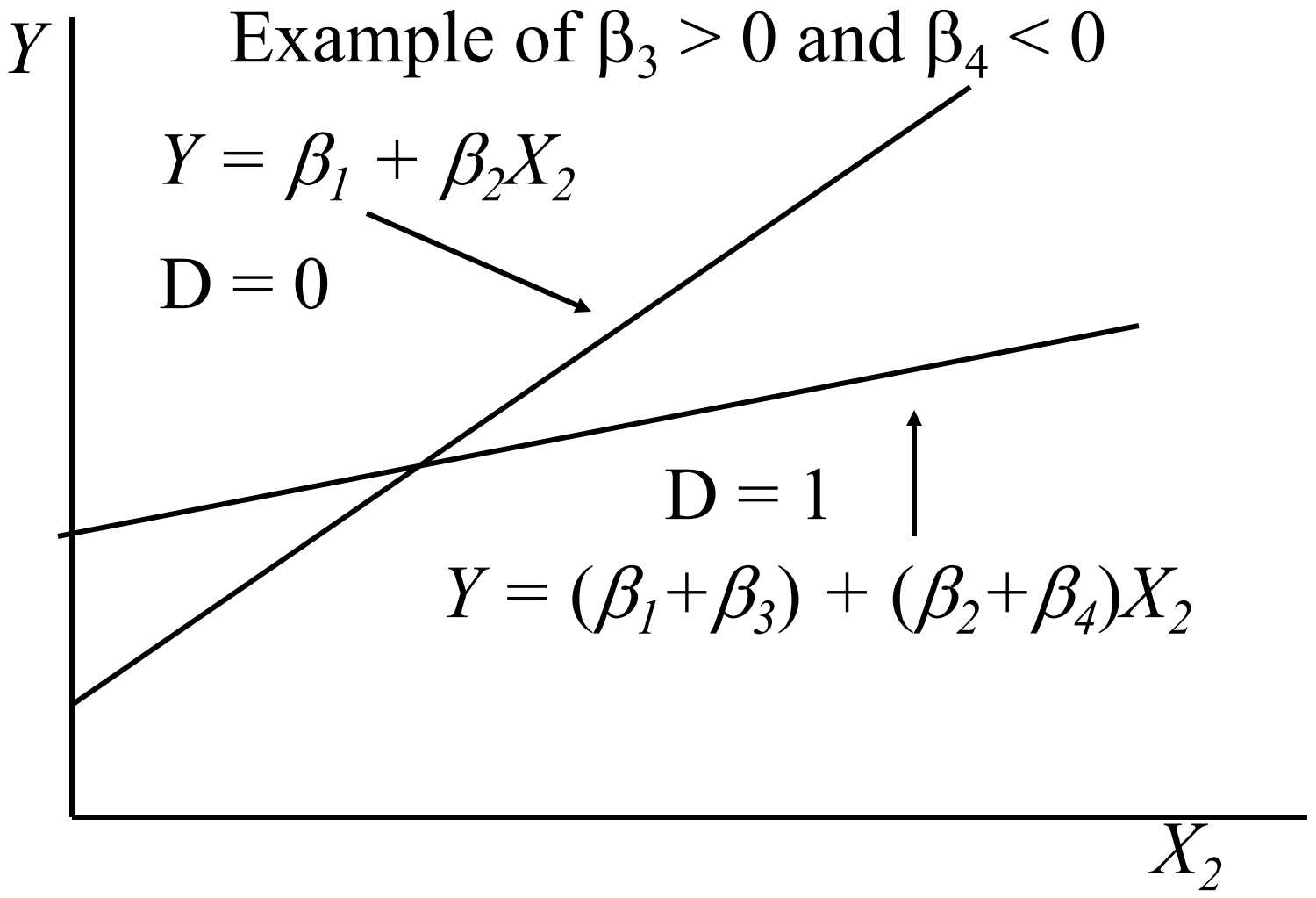
Dependent Variable: EARN Method: Least Squares Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-25.20476	4.995630	-5.045362	0.0000
S	1.797148	0.251515	7.145295	0.0000
ASVABC	0.127068	0.051329	2.475574	0.0136
MALE	4.073660	0.718649	5.668501	0.0000
PWE	0.425351	0.168296	2.527395	0.0118
ETHNW	-2.331664	1.082538	-2.153886	0.0317

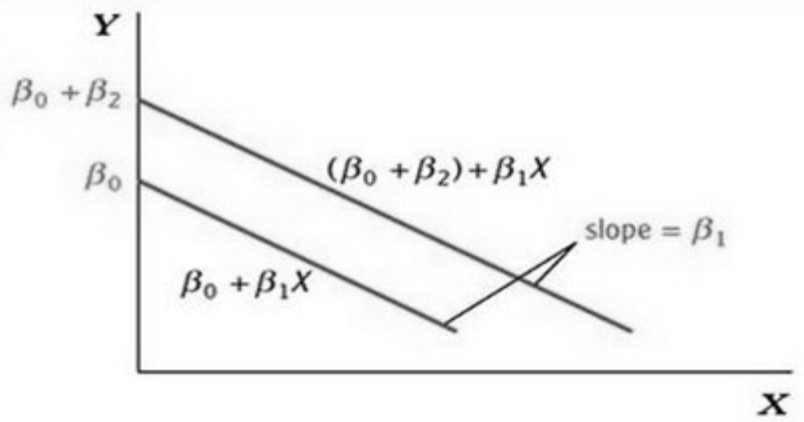
R-squared	0.242464	Mean dependent var	13.68988
S.D. dependent var	9.702960	S.E. of regression	8.482471
Sum squared resid	40581.10	F-statistic	36.10389
Durbin-Watson stat	1.880030	Prob(F-statistic)	0.000000

Model with Intercept and Slope Dummy Variables

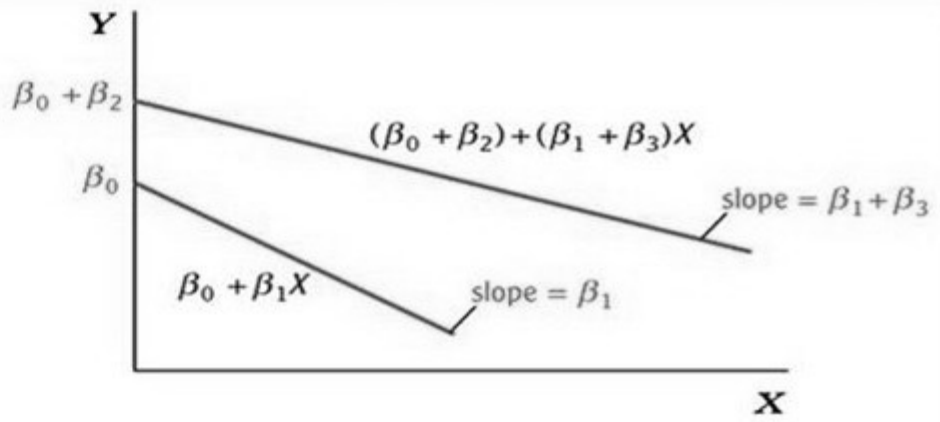
$$Y = \beta_1 + \beta_3 D + \beta_2 X_2 + \beta_4 DX_2$$



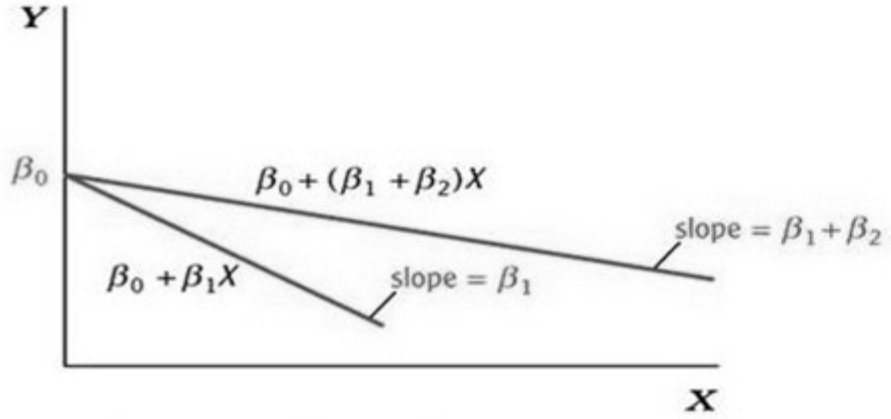
Interactions Between Binary and Continuous Variables



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1 X + \beta_2 D$ allows for different intercepts but has the same slope; (b) $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$ allows for different intercepts and different slopes; and (c) $\beta_0 + \beta_1 X + \beta_2 (X \times D)$ has the same intercept but allows for different slopes.

MORE THAN TWO CATEGORIES AND SEVERAL GROUPS OF DUMMIES

Dependent Variable: EARN
Method: Least Squares
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-25.12985	5.004870	-5.021080	0.0000
S	1.791735	0.252261	7.102694	0.0000
ASVABC	0.128165	0.051480	2.489636	0.0131
MALE	4.083965	0.719913	5.672861	0.0000
PWE	0.420852	0.168994	2.490332	0.0130
ETHBLACK	-2.077643	1.334459	-1.556917	0.1201
ETHHISP	-2.707987	1.583041	-1.710623	0.0877

R-squared	0.242607	Mean dependent var	13.68988
S.D. dependent var	9.702960	S.E. of regression	8.489199
Sum squared resid	40573.44	F-statistic	30.05662
Durbin-Watson stat	1.879758	Prob(F-statistic)	0.000000

$$\widehat{EARNINGS} = -25.13 + 1.79S + 0.13ASVABC + 4.08MALE + 0.42PWE - 2.08ETHBLACK - 2.71ETHHISP$$

CHANGING REFERENCE CATEGORIES

Dependent Variable: EARN
Method: Least Squares
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-23.75388	5.049820	-4.703906	0.0000
S	1.791735	0.252261	7.102694	0.0000
ASVABC	0.128165	0.051480	2.489636	0.0131
PWE	0.420852	0.168994	2.490332	0.0130
FEMALE	-4.083965	0.719913	-5.672861	0.0000
ETHBLACK	0.630345	1.933364	0.326035	0.7445
ETHWHITE	2.707987	1.583041	1.710623	0.0877

R-squared	0.242607	Mean dependent var	13.68988
S.D. dependent var	9.702960	S.E. of regression	8.489199
Sum squared resid	40573.44	F-statistic	30.05662
Durbin-Watson stat	1.879758	Prob(F-statistic)	0.000000

$$\widehat{EARNINGS} = -23.75 + \mathbf{1.79}\,S + \mathbf{0.13}\,ASVABC - \mathbf{4.08}\,FEMALE + \mathbf{0.42}\,PWE$$
$$+ 0.63\,ETHBLACK + \mathbf{2.71}\,ETHWHITE$$

INTERCEPT DUMMIES and SLOPE DUMMIES.
INTERACTION DUMMIES.

Slope Dummies:

$MALES = MALE * S;\ MALEASVABC = MALE * ASVABC;\ MALEPWE = MALE * PWE;$

Interaction Dummy: $MALENW = MALE * NW.$

$Sample\ Sorted\ by\ MALE: MALE = 1\ for\ i = 1,\ldots,325;\quad MALE = 0\ for\ i = 326,\ldots,570$

$$\widehat{EARNINGS} = \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 ETHNW + \beta_5 PWE \quad (1)$$

$$\widehat{EARNINGS} = \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 ETHNW + \beta_5 PWE + \beta_6 MALE + \\ + \beta_7 MALES + \beta_8 MALEASVABC + \beta_9 MALEPWE + \beta_{10} MALENW \ (2)$$

$H_0: \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0 \Leftrightarrow$

Relationship is the same for Male and Female subsamples

$F - test\ for\ a\ set\ of\ Dummy\ Variables$:

$$F(5{,}570 - 10) = \frac{(SSR_1 - SSR_2)/5}{SSR_2/(570 - 10)} = \frac{(42893.07 - 39629.74)/5}{39629.74/560} = 9.223$$

$F_{\text{crit, } 0{,}1\%}(5; 560) = 4.17$ $\qquad$ $H_0\quad rejected$

INTERCEPT DUMMIES and SLOPE DUMMIES

$$\widehat{EARNINGS} = \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 ETHNW + \beta_5 PWE + \beta_6 MALE + \\ + \beta_7 MALES + \beta_8 MALEASVABC + \beta_9 MALEPWE + \beta_{10} MALENW \quad (2)$$

Dependent Variable: EARN Method: Least Squares Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-11.66633	7.715561	-1.512053	0.1311
S	1.273872	0.402525	3.164699	0.0016
ASVABC	0.157333	0.088742	1.772925	0.0768
ETHNW	0.286580	1.679629	0.170621	0.8646
PWE	-0.160567	0.257096	-0.624539	0.5325
MALE	-19.56550	10.03029	-1.950642	0.0516
MALES	0.866114	0.513473	1.686775	0.0922
MALEASVABC	-0.033272	0.108446	-0.306808	0.7591
MALENW	-4.422123	2.195731	-2.013964	0.0445
MALEPWE	1.002124	0.338438	2.961031	0.0032

R-squared 0.260224 Sum squared resid 39629.74 F-statistic 21.88727

The variables MALENW and MALEPWE are significant at 5% level, while the variables ETHNW and PWE are not. For these variables slope dummies for MALE are significant. For S and ASVABC – the slope does not differ significantly for the subsamples.

THE DUMMY VARIABLES TRAP

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \delta_2 D_2 + \ldots \delta_s D_s + u \quad (1)$$

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \delta_1 D_1 + \delta_2 D_2 + \ldots \delta_s D_s + u \quad (2)$$

$$Y = \beta_2 X_2 + \ldots + \beta_k X_k + \delta_1 D_1 + \delta_2 D_2 + \ldots \delta_s D_s + u \quad (3)$$

Observation	Category	X_1	D_1	D_2	D_3	D_4
1	4	1	0	0	0	1
2	3	1	0	0	1	0
3	1	1	1	0	0	0
4	2	1	0	1	0	0
5	2	1	0	1	0	0
6	3	1	0	0	1	0
7	1	1	1	0	0	0
8	4	1	0	0	0	1

$$\sum_{i=1}^{4} D_i = X_1$$

Dummy Variable Trap – perfect multicollinearity in the specification (2).

You may use either specification (1) with the intercept and dummies for all but one categories, or specification (3) with the dummies for all categories without intercept.

(1) And (3) are equivalent, but in the model without intercept the R^2 formula is invalid, and it is more difficult to test for differences between the parameters.

CHOW TEST for COEFFICIENTS STABILITY
(CHOW BREAKPOINT TEST)

$$\widehat{EARNINGS} = \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 PWE + \beta_5 ETHNW$$
$$for\ the\ whole\ sample,\ and\ for\ MALE\ and\ FEMALE\ subsamples$$

H_o: *the coefficients are the same for subsamples*
H_1: *at least one coefficient differ*

$$SSR_{MALE} = 32214.01\ (325obs);\ SSR_{FEMALE} = 7415.77\ (245obs)$$
$$SSR_{MALE} + SSR_{FEMALE} = 39629.74\ (570obs)$$

$$F(5{,}570-10) = \frac{(SSR_1 - (SSR_{MALE} + SSR_{FEMALE})/5}{(SSR_{MALE} + SSR_{FEMALE})/(570-10)} =$$
$$= \frac{(42893.07 - (32214.01 + 7415.73))/5}{(32214.01 + 7415.73)/560} = 9.223$$

$F_{\text{crit, }0{,}1\%}(5; 560) = 4.17$ $\qquad$ H_0 *rejected*

Chow test is equivalent to the F – test for the group of all dummies

CHOW BREAKPOINT TEST FOR SEVERAL SUBSAMPLES

$$m\ subsamples:\ SSR_i\ for\ i = 1,\ldots,m;\ SSR_0 - the\ whole\ sample$$

$$F(k(m-1), n-mk) = \frac{(SSR_0 - (SSR_1 + \ldots + SSR_m))/(k(m-1))}{(SSR_1 + \ldots + SSR_m)/(n-mk)}$$

$$H_0: coefficients\ are\ the\ same\ for\ all\ samples$$

Chow test is easy to implement, but it does not show where is the difference

In EViews : estimate regression for the whole sample, then

View - Stability tests - Chow breakpoint test - list the breakpoints

CHOW TEST for COEFFICIENTS STABILITY
(CHOW BREAKPOINT TEST)

$$\widehat{EARNINGS} = \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 PWE + \beta_5 MALE$$
$$for\ the\ whole\ sample,\ and\ for\ WHITE,\ BLACK\ and\ HISP\ subsamples$$

$$H_o: the\ coefficients\ are\ the\ same\ for\ all\ subsamples$$
$$H_1: at\ least\ one\ coefficient\ differ$$

$$SSR_{WHITE} = 38127.78\ (488obs);\ SSR_{HISP} = 676.95\ (32obs);\ SSR_{BLACK} = 724.66\ (50obs);$$
$$SSR_{WHITE} + SSR_{HISP} + SSR_{BLACK} = 39529.39\ (570obs)$$

$$F(10{,}570 - 15) = \frac{(SSR_0 - (SSR_{WHITE} + SSR_{HISP} + SSR_{BLACK}))/10}{(SSR_{WHITE} + SSR_{HISP} + SSR_{BLACK})/(570 - 15)} =$$
$$= \frac{(40914.91 - 39529.39)/10}{39529.39/555} = 1.945$$

$$F_{\text{crit, } 1\%}(10{,}555) = 2.22$$
$$F_{\text{crit, } 5\%}(10{,}555) = 1.85$$

$$H_0\ rejected\ at\ 5\%,\ but\ not\ at\ 1\%$$

DUMMY VARIABLES in PREDICTION

Sample: 1981 2011

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.298460	0.194922	-1.531173	0.1340
LGDPI	1.036576	0.006497	159.5366	0.0000
LGPRHOUS	-0.423765	0.045451	-9.323628	0.0000

$$\hat{Y}_{2012} = -0.2985 + 1.0366 LGDPI - 0.4238 LGPRHOUS$$

Sample: 1981 2012

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.298460	0.194922	-1.531173	0.1340
LGDPI	1.036576	0.006497	159.5366	0.0000
LGPRHOUS	-0.423765	0.045451	-9.323628	0.0000
D2012	-0.041629	0.017210	-2.418867	0.0205

$$\hat{Y}_{2012} = -0.2985 + 1.0366 LGDPI - 0.4238 LGPRHOUS - 0.0416$$

The observation-specific dummy variable in 2012 guarantees perfect fit for the prediction year. The coefficient of *D2012* equals to minus the prediction error for 2012.The standard error of the coefficient of *D2012* is the standard error of the prediction error for that year.