Chapter 7 Multiple Regression Model

남준우·허인 (2018), 제5장

Gujarati/Porter (2018), Chapter 7 & 8

(1) Model and Assumptions

(2) Interpretation of β's

(3) Least Squares Estimator in Three Variable Case

(4) R^2 and Adjusted R^2

(5) Sampling Properties of LS Estimator

(6) Testing Overall Significance

(7) Testing about part of β's

(8) Testing Linear Restrictions

(9) Chow Test: Test of Structural Change(Difference)

- In simple regression model, only one independent variable $\Rightarrow$ dependent variable.

► This is very restrictive in reality.

(Example)

① Household consumption function:

Income, Wealth, # of household family, Region, Age of household head, …

② Income(Earnings):

Education, Gender, Tenure, Square of Tenure, Occupation, …

③ Grade:

Studying hour, Effort, IQ?, Luck?, ….

(1) Model and Assumptions

① Model:

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

► X_{ji}: jth variable, ith observation.

► k= # of independent variables including constant term.

② Assumptions

(a) $E(\varepsilon_i) = 0$ for all i.

(b) $V(\varepsilon_i) = \sigma^2$ for all i.

(c) $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

(d) No exact linear relationship among X variables.

(e) Variation in each variable of X.

(f) X's are non-random.

(2) Interpretation of β's

① Controlling other variables:

$$\beta_j = \frac{\Delta E(y)}{\Delta X_j}\Big|_{\text{all other X's controlled}}$$; partial regression coefficient.

- In regression model,

$$y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \beta_4 W_i + \varepsilon_i$$

$$\beta_2 = \frac{\Delta E(y)}{\Delta X}\Big|_{\Delta Z = \Delta W = 0}, \qquad \beta_3 = \frac{\Delta E(y)}{\Delta Z}\Big|_{\Delta X = \Delta W = 0}$$

► Total derivative vs. Partial derivative

► Simple regression vs. Multiple regression

(Examples)

(a) $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \beta_4 W_i + \varepsilon_i$

y: earnings, X: gender, Z: education, ….

► β_2: (pure) gender discrimination <u>when other things are equal</u>.

(b) $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$

y: output, X: weather, Z: labor input.

► β_2: (pure) weather effect on agricultural output.

(c) To see the schooling(knowledge) effect on heath,

you have to include all the variables (income, occupation …) affecting health status.

Why?

As △schooling(knowledge) --→ △health.

And also, △schooling(knowledge) --→ △ income, △ occupation --→ △health.

② Direct Effect

- Multiple regression coefficients can be understood as direct effect.

► Total effect vs. Direct effect

(a) Interpretation

(b) Simple regression coefficient vs. Multiple regression coefficient:

► $y_i = a_1 + a_2 X_i + e_i$

► $y_i = b_1 + b_2 X_i + b_3 Z_i + e_i$

$$\boxed{a_2 = b_2 + b_3 \cdot d_2}$$

where d_2: slope coefficient of Z on X ($\hat{Z}_i = d_1 + d_2 X_i$)

(Proof)

(c) Empirical Demonstration(Pitcher file)

Games (Game) -----→ Salary (Money)

ⓐ $Money_i = a_1 + a_2 Game_i + e_i$

Model 1: OLS, using observations 1-115
Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	5822.24	1887.45	3.085	0.0026	***
GAME	139.523	55.2550	2.525	0.0130	**

ⓑ $Money_i = b_1 + b_2 Game_i + b_3 Win_i + e_i$

Model 2: OLS, using observations 1-115
Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3756.67	1698.99	2.211	0.0291	**
GAME	55.2160	50.7532	1.088	0.2790	
WIN	1243.64	213.819	5.816	<0.0001	***

ⓒ $Win_i = d_1 + d_2 Game_i + e_i$

Model 3: OLS, using observations 1-115
Dependent variable: WIN

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	1.66090	0.730980	2.272	0.0250	**
GAME	0.0677906	0.0213995	3.168	0.0020	***

► $139.523 \overset{?}{=} 55.21 + 1243.64 \times 0.0677$

◎ <<u>Empirical Demonstration</u>> Multiple Regression Coefficient

- In regression model,

$$y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$$

$$\beta_2 = \frac{\Delta E(y)}{\Delta X}\Big|_{\Delta Z=0}, \qquad \beta_3 = \frac{\Delta E(y)}{\Delta Z}\Big|_{\Delta X=0}$$

(Example) Pitcher File

► $Money_i = \beta_1 + \beta_2 Game_i + \beta_3 Win_i + \varepsilon_i$, $\beta_3 = \frac{\Delta E(Money)}{\Delta Win}\Big|_{\Delta Game=0}$.

To get this idea,

(a) LS of Money on (1, Game), get residual ea.

* gretl 에서 regression output 상단의 [save]/[residual] 을 통해 이름을 [ea]로 설정한다.

► ea: Money 중 Game 의 영향이 제외(△Game=0)된 부분.

(b) LS of Win on (1, Game), get residual eb.

* gretl 에서 regression output 상단의 [save]/[residual] 을 통해 이름을 [eb]로 설정한다.

► eb: Win 중에서 Game 의 영향이 제외(△Game=0)된 부분.

(c) LS of ea on (1,eb)

Dependent variable: ea

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	6.23467e-013	800.963	7.784e-016	1.0000	
eb	1243.64	212.871	5.842	<0.0001	***

기울기 1243.64 는 $Money_i = b_1 + b_2 Game_i + b_3 Win_i + e_i$의 b_3와 정확히 일치함.

► $b_3 = \dfrac{\Delta ea}{\Delta eb} = \dfrac{\Delta Money}{\Delta Win}\Big|_{\Delta Game=0}$ 이다.

(3) Least Squares Estimator in Three Variable Case

► PRF: $y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$

► SRF: $y_i = b_1 + b_2 X_{2i} + \cdots + b_k X_{ki} + e_i$

Find $(b_1, b_2, \cdots, b_k)$ minimizing $\sum_{i=1}^{n} e_i^2$.

• For k=3,

PRF: $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$ SRF: $y_i = b_1 + b_2 X_i + b_3 Z_i + e_i$

F.O.C.: $\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} X_i e_i = 0, \quad \sum_{i=1}^{n} Z_i e_i = 0$.

► $\bar{e} = 0 \Rightarrow$ Regression line passes thru $(\bar{X}, \bar{Z}, \bar{y})$.

► $S_{Xe} = 0, \quad S_{Ze} = 0$

► It is hard (and stupid) to derive the expression of (b_1, b_2, b_3).

(4) R^2 and Adjusted R^2

► $R^2 = 1 - \dfrac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}$.

• However, as $k \uparrow$, $\left(\sum_{i=1}^{n} e_i^2\right) \downarrow$, regardless of the significance of additional variables.

► Use $\overline{R}^2 = 1 - \dfrac{\sum_{i=1}^{n} e_i^{\,2} / (n-k)}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2 / (n-1)} = 1 - (1 - R^2)\dfrac{n-1}{n-k}$.

(Example) Pitcher file

Model 7: OLS, using observations 1-115
Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3756.67	1698.99	2.211	0.0291	**
WIN	1243.64	213.819	5.816	<0.0001	***
GAME	55.2160	50.7532	1.088	0.2790	

Mean dependent var	9992.174	S.D. dependent var	10029.55
Sum squared resid	8.34e+09	S.E. of regression	8627.628
R-squared	0.273003	Adjusted R-squared	0.260021
F(2, 112)	21.02919	P-value(F)	1.76e-08
Log-likelihood	−1203.871	Akaike criterion	2413.743
Schwarz criterion	2421.978	Hannan-Quinn	2417.085

(Notes on R^2)

① In a multiple regression, $\bar{R}^2$ is a better measure.

② If the model does not contain an intercept, $0 \le R^2, \bar{R}^2 \le 1$ can not be guaranteed.

③ If the dependent variables are <u>different</u>, comparing R^2 across two regression models is meaningless.

④ $\bar{R}^2 < R^2$ but $\bar{R}^2 < 0$ is possible.

► The model does not describe the data.

(5) Sampling Properties of LS Estimator

① Gauss-Markov theorem applies

② $E(b_j) = \beta_j$ for all $j = 1, 2, \cdots, k$.

③ $V(b_j) = \left[\sigma^2 (X'X)^{-1}\right]_{jj}$.

④ $s^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-k}$ for estimator of σ^2, then $E(s^2) = \sigma^2$.

⑤ $\hat{V}(b_j) = s_{b_j}{}^2 = s^2\left[(X'X)^{-1}\right]_{jj}$, then $E(s_{b_j}{}^2) = \sigma_{b_j}{}^2$.

► $s_{b_j} = \sqrt{s_{b_j}{}^2}$: standard error of b_j.

⑥ With $\varepsilon_i \sim N(0,\sigma^2)$,

► $\dfrac{b_j - \beta_j}{\sigma_{b_j}} \sim N(0, 1)$.

► $\dfrac{b_j - \beta_j}{s_{b_j}} \sim t(n-k)$.

⑦ 100(1-α)% C.I. for $\beta_j = \left(b_j \pm t_{(n-k;\alpha/2)}\, s_{b_j} \right)$.

⑧ Testing for single coefficient:

$H_0 : \beta_j = \beta_j *, \quad H_1 : \beta_j \neq \beta_j *,$

Compare $t = \dfrac{b_j - \beta_j *}{s_{b_j}}$ with $t_{(n-k;\alpha/2)}$.

► The same logic (2-t rule of thumb, P-value) applies for one-sided hypothesis test.

Model 7: OLS, using observations 1-115
Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3756.67	1698.99	2.211	0.0291	**
WIN	1243.64	213.819	5.816	<0.0001	***
GAME	55.2160	50.7532	1.088	0.2790	

Mean dependent var	9992.174	S.D. dependent var	10029.55
Sum squared resid	8.34e+09	S.E. of regression	8627.628
R-squared	0.273003	Adjusted R-squared	0.260021
F(2, 112)	21.02919	P-value(F)	1.76e-08
Log-likelihood	−1203.871	Akaike criterion	2413.743
Schwarz criterion	2421.978	Hannan-Quinn	2417.085

(6) Testing Overall Significance(모형의 적합도 검정)

► Individual test vs. Joint test

◎ <u>Overall Significance Test</u>(모형의 적합도 검정)

► $y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$

• $H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0,$ $H_1 :$ at least one of the β_j's$(\beta_2, \beta_3, \cdots, \beta_k)$ are not zero.

$\Leftrightarrow$ $H_0 : R^2 = 0$

① <u>Using sum squares of residual</u>

(a) (Restricted model) Under H_0, $y_i = \beta_1 + \varepsilon_i$. SRF: $y_i = b_1{*} + e_i{*}$

► LS of y on 1, get $SSR^R \equiv \sum_{i=1}^{n} e_i{*}^2$.

(b) (General(Unrestricted) model) $y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$

SRF: $y_i = b_1 + b_2 X_{2i} + \cdots + b_k X_{ki} + e_i$

► LS of y on $(1, X_2, \cdots, X_k)$, get $SSR^U = \sum_{i=1}^{n} e_i^2$.

$$F = \frac{\left(SSR^{R} - SSR^{U}\right) \big/ (k-1)}{SSR^{U} \big/ (n-k)}$$

- $$= \frac{\left(\sum_{i=1}^{n} e_i{*}^2 - \sum_{i=1}^{n} e_i^{\,2}\right) \Big/ (k-1)}{\sum_{i=1}^{n} e_i^{\,2} \Big/ (n-k)}$$

► Since $F \sim F(k-1, n-k)$ under H_0,

if $F \geq F_{(k-1, n-k;\alpha)}$, reject H_0.

if $F < F_{(k-1, n-k;\alpha)}$, do not reject H_0.

② <u>Using R^2</u>

From restricted model, SRF: $y_i = b_1 * + e_i *$

► LS of y on 1, $\sum_{i=1}^{n} e_i *^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2$.

$$F = \frac{\left(\sum_{i=1}^{n} (y_i - \overline{y})^2 - \sum_{i=1}^{n} e_i^2 \right) \Big/ (k-1)}{\sum_{i=1}^{n} e_i^2 \Big/ (n-k)} = \frac{R^2 \big/ (k-1)}{1 - R^2 \big/ (n-k)}$$

◎ <u>Procedure for overall significance test</u>:

• To test $H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$,

► LS of y on $(1, X_2, \cdots, X_k)$, get R^2.

Calculate $F = \dfrac{R^2 / (k-1)}{\left(1 - R^2\right) / (n-k)}$

► Since $F \sim F(k-1, n-k)$ under H_0,

if $F \geq F_{(k-1, n-k; \alpha)}$, reject H_0.

if $F < F_{(k-1, n-k; \alpha)}$, do not reject H_0.

<Review> F-distribution

- Suppose $W_1 \sim \chi^2(k_1)$, $W_2 \sim \chi^2(k_2)$. W_1 and W_2 are independent.

Then, $F = \dfrac{W_1/k_1}{W_2/k_2} \sim F(k_1, k_2)$

(Example) Pitcher file

Model 7: OLS, using observations 1-115
Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3756.67	1698.99	2.211	0.0291	**
WIN	1243.64	213.819	5.816	<0.0001	***
GAME	55.2160	50.7532	1.088	0.2790	

Mean dependent var	9992.174	S.D. dependent var	10029.55
Sum squared resid	8.34e+09	S.E. of regression	8627.628
R-squared	0.273003	Adjusted R-squared	0.260021
F(2, 112)	21.02919	P-value(F)	1.76e-08
Log-likelihood	−1203.871	Akaike criterion	2413.743
Schwarz criterion	2421.978	Hannan-Quinn	2417.085

TABLE D.3 Upper Percentage Points of the F Distribution

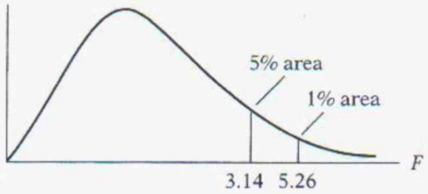
Example

$\Pr(F > 1.59) = 0.25$

$\Pr(F > 2.42) = 0.10$ for df $N_1 = 10$

$\Pr(F > 3.14) = 0.05$ and $N_2 = 9$

$\Pr(F > 5.26) = 0.01$



df for denominator N_2	df for numerator N_1												
	Pr	1	2	3	4	5	6	7	8	9	10	11	12
1	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.36	9.41
	.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.5	60.7
	.05	161	200	216	225	230	234	237	239	241	242	243	244
2	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.39
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41
	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4
3	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.45
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22
	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1
4	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
	.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4
5	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89
6	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.77
	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	.01	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	.25	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.69	1.68
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
	.01	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47
8	.25	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.63	1.62
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
	.01	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67
9	.25	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38

TABLE D.3 Upper Percentage Points of the *F* Distribution *(Continued)*

df for denominator N_2	df for numerator N_1												
	Pr	1	2	3	4	5	6	7	8	9	10	11	12
22	.25	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.38	1.37
	.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12
24	.25	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.37	1.36
	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85	1.83
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.21	2.18
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03
26	.25	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.36	1.35
	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.84	1.81
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96
28	.25	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.81	1.79
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90
30	.25	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.35	1.34
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.73	1.71
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.29
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.27	1.26
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.62	1.60
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.87	1.83
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34
200	.25	1.33	1.39	1.38	1.36	1.34	1.32	1.31	1.29	1.28	1.27	1.26	1.25
	.10	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.60	1.57
	.05	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80
	.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27
∞	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.24
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.57	1.55
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18

③ (Special case) k=2

- Overall significance test: $H_0 : \beta_2 = 0$ in $y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$

► Test statistics: $F = \dfrac{\left(\sum_{i=1}^{n} (y_i - \overline{y})^2 - \sum_{i=1}^{n} e_i^2 \right) \Big/ (k-1)}{\sum_{i=1}^{n} e_i^2 \Big/ (n-k)} \Rightarrow$

$$F = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} e_i^2 \Big/ (n-2)}$$

$$= \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{s^2} = \frac{b_2^2 \sum_{i=1}^{n} (X_i - \overline{X})^2}{s^2} = \left(\frac{b_2}{s_{b_2}} \right)^2$$

► Critical value: $F_{(1,n-2;\alpha)} = {t_{(n-2;\alpha/2)}}^2$.

► Rejection region: Reject if

(Joint test) $F = \left(\dfrac{b_2}{s_{b_2}}\right)^2 \geq F_{(1,n-2;\alpha)}$

$\Leftrightarrow \ \left|\text{t-ratio}\right| = \left|\dfrac{b_2}{s_{b_2}}\right| \geq t_{(n-2;\alpha/2)}$ (Single individual test)

(Example) Pitcher file

Model 8: OLS, using observations 1-115
Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	5161.97	1104.55	4.673	<0.0001	***
WIN	1310.08	205.080	6.388	<0.0001	***

Mean dependent var	9992.174	S.D. dependent var	10029.55
Sum squared resid	8.42e+09	S.E. of regression	8634.634
R-squared	0.265320	Adjusted R-squared	0.258818
F(1, 113)	40.80847	P-value(F)	3.87e-09
Log-likelihood	−1204.476	Akaike criterion	2412.952
Schwarz criterion	2418.442	Hannan-Quinn	2415.180

(7) Testing about part of β's

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

► $H_0 : \beta_2 = \beta_3 = \cdots = \beta_{J+1} = 0$ (# of restrictions=J)

(a) (Restricted model)

Under H_0, $y_i = \beta_1 + \beta_{J+2} X_{J+2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$.

SRF: $y_i = b_1 * + b_{J+2} * X_{J+2i} + \cdots + b_k * X_{ki} + e_i *$

► LS of y on $(1, X_{J+2}, \cdots, X_k)$,

$$SSR^R \equiv \sum_{i=1}^{n} e_i *^2 \text{ and } R*^2.$$

(b) (General model)

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

SRF: $y_i = b_1 + b_2 X_{2i} + \cdots + b_k X_{ki} + e_i$

► LS of y on $(1, X_2, \cdots, X_k)$,

$$SSR^U \equiv \sum_{i=1}^{n} e_i^2 \text{ and } R^2.$$

① Using sum squares of residual approach

$$F = \frac{\left(SSR^R - SSR^U\right)/J}{SSR^U/(n-k)}$$

- $$= \frac{\left(\sum_{i=1}^{n} e_i^{*2} - \sum_{i=1}^{n} e_i^2\right)/J}{\sum_{i=1}^{n} e_i^2/(n-k)}$$

► Since $F \sim F(J, n-k)$ under H_0,

if $F \geq F_{(J,n-k;\alpha)}$, reject H_0.

if $F < F_{(J,n-k;\alpha)}$, do not reject H_0.

② Using R^2

Use $F = \dfrac{\left(R^2 - R*^2\right)/J}{\left(1 - R^2\right)/(n-k)}$

(Example) Pitcher file

Model: $Money_i = \beta_1 + \beta_2 Win_i + \beta_3 Save_i + \beta_4 Four_i + \beta_5 Lose_i + \beta_6 Year_i + \varepsilon_i$

$$H_0 : \beta_4 = \beta_5 = 0$$

(a)

Model 11: OLS, using observations 1-115

Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	52.3272	1503.49	0.03480	0.9723	
WIN	1119.04	268.607	4.166	<0.0001	***
SAVE	407.865	134.312	3.037	0.0030	***
YEAR	818.325	146.250	5.595	<0.0001	***
FOUR	−%s	62.7995	−%#.4g	0.9801	
LOSE	79.7402	310.720	0.2566	0.7979	

Mean dependent var	9992.174	S.D. dependent var	10029.55
Sum squared resid	6.08e+09	S.E. of regression	7470.273
R-squared	0.469566	Adjusted R-squared	0.445234
F(5, 109)	19.29843	P-value(F)	1.02e-13
Log-likelihood	−1185.746	Akaike criterion	2383.492
Schwarz criterion	2399.961	Hannan-Quinn	2390.177

(b)

Model 10: OLS, using observations 1-115
Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	128.912	1230.24	0.1048	0.9167	
WIN	1150.34	178.744	6.436	<0.0001	***
SAVE	409.819	132.346	3.097	0.0025	***
YEAR	822.359	143.346	5.737	<0.0001	***

Mean dependent var	9992.174	S.D. dependent var	10029.55
Sum squared resid	6.09e+09	S.E. of regression	7405.396
R-squared	0.469175	Adjusted R-squared	0.454828
F(3, 111)	32.70281	P-value(F)	3.18e-15
Log-likelihood	−1185.788	Akaike criterion	2379.576
Schwarz criterion	2390.556	Hannan-Quinn	2384.033

► Since $F=\dfrac{(6.09\times10^{9}-6.08\times10^{9})/2}{6.08\times10^{9}/(115-6)}=0.09<F_{(2,109;0.05)}=3.07$, do not reject H_0.

(8) Testing Linear Restrictions

(Examples)

① $H_0 : \beta_2 + \beta_3 = \beta *$

② $H_0 : \beta_2 + \beta_3 + \beta_4 = 1$ and $\beta_2 = \beta_3$

► Testing above hypothesis by using t-test but it is trivial.

◎ <u>Using sum squares of residual approach</u>

(a) (Restricted model) Under H_0, get $SSR^R \equiv \sum_{i=1}^{n} e_i{*}^2$.

(b) (General model) $y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$

► LS of y on $(1, X_2, \cdots, X_k)$, get $SSR \equiv \sum_{i=1}^{n} e_i^2$.

Calculate $F = \dfrac{\left(SSR^R - SSR\right)/J}{SSR/(n-k)} = \dfrac{\left(\sum_{i=1}^{n} e_i{*}^2 - \sum_{i=1}^{n} e_i^2\right)/J}{\sum_{i=1}^{n} e_i^2 /(n-k)}$,

where J=# of restrictions(NOT # of parameters)

Restricted model:

① $H_0 : \beta_2 + \beta_3 = 1$

- Under restriction, $\beta_2 = 1 - \beta_3$.

$$\begin{aligned} & y_i = \beta_1 + (1-\beta_3)X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \\ \Rightarrow\ & y_i - X_{2i} = \beta_1 + \beta_3(-X_{2i} + X_{3i}) + \cdots + \beta_k X_{ki} + \varepsilon_i \end{aligned}.$$

SRF: LS of $y - X_2$ on $\left(1, (X_3 - X_2), \cdots, X_k\right)$,

get $SSR^R \equiv \sum_{i=1}^{n} e_i{*}^2$.

► Can we use R^2 approach?

② $H_0: \beta_2 + \beta_3 + \beta_4 = 1 \quad \text{and} \quad \beta_2 = \beta_3$

- Under restriction, $\beta_4 = 1 - 2\beta_3 \quad \text{and} \quad \beta_2 = \beta_3$.

$$
\begin{aligned}
& y_i = \beta_1 + \beta_3 X_{2i} + \beta_3 X_{3i} + (1-2\beta_3) X_{4i} + \cdots + \beta_k X_{ki} + \varepsilon_i \\
& \Rightarrow \; y_i - X_{4i} = \beta_1 + \beta_3 (X_{2i} + X_{3i} - 2X_{4i}) + \cdots + \beta_k X_{ki} + \varepsilon_i
\end{aligned}
.
$$

SRF: LS of $y - X_4$ on $\left(1, (X_2 + X_3 - 2X_4), \cdots, X_k\right)$,

get $SSR^R \equiv \sum_{i=1}^{n} e_i *^2$.

(Example) Pitcher file

Model: $Money_i = \beta_1 + \beta_2 Win_i + \beta_3 Save_i + \beta_4 Four_i + \beta_5 Lose_i + \beta_6 Year_i + \varepsilon_i$

$$H_0 : \beta_2 = 2\beta_3$$

Model 12: OLS, using observations 1-115
Dependent variable: MONEY

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	−%s	1449.09	−%#.4g	0.8589	
YEAR	831.232	145.094	5.729	<0.0001	***
FOUR	22.1565	55.1132	0.4020	0.6885	
LOSE	83.6026	310.156	0.2696	0.7880	
TWINSAVE	483.699	94.1012	5.140	<0.0001	***

Mean dependent var	9992.174	S.D. dependent var	10029.55
Sum squared resid	6.12e+09	S.E. of regression	7457.636
R-squared	0.466509	Adjusted R-squared	0.447109
F(4, 110)	24.04728	P-value(F)	2.62e-14
Log-likelihood	−1186.076	Akaike criterion	2382.152
Schwarz criterion	2395.877	Hannan-Quinn	2387.723

► (gretl Tips) TWINSAVE=2*WIN+SAVE 변수를 나타내는 것으로 gretl에서 [Add]/[Define a new variable]에서 식을 입력하여 TWINSAVE 변수를 생성한 것임.

► Since $F = \dfrac{(6.12\times 10^9 - 6.08\times 10^9)/1}{6.08\times 10^9/(115-6)} = 0.72 < F_{(1,109;0.05)} = 3.92$, do not reject H_0.

(8) Chow Test: Test of Structural Change(Difference)

$$y_i = \alpha_1 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, 2, \cdots, n_1 \qquad (n_1 \text{ observations})$$

$$y_i = \alpha_2 + \beta_2 X_i + \varepsilon_i, \qquad i = n_1 + 1, \cdots, n \qquad (n_2 \text{ observations})$$

$$n_1 + n_2 = n$$

$H_0 : \alpha_1 = \alpha_2 (\equiv \alpha),\ \beta_1 = \beta_2 (\equiv \beta)$ (Structural no change)

► # of independent variables =k

(Examples)

① consumption function

② time series model

③ cross-sectional model

(a) (Restricted model) Under H_0, $y_i = \alpha + \beta X_i + \varepsilon_i$, $i = 1, 2, \cdots, n$.

LS of y on $(1, X)$ with n observations and get $SSR^* = \sum_{i=1}^{n} e_i^{*2}$. (pooled regression)

(b) (General model) (two separate regressions)

For $i = 1, 2, \cdots, n_1$ (n_1 observations), LS of y on $(1, X)$ and get $SSR_1 = \sum_{i=1}^{n_1} e_i^2$.

For $i = n_1 + 1, \cdots, n$ (n_2 observations), LS of y on $(1, X)$ and get $SSR_2 = \sum_{i=n_1+1}^{n} e_i^2$.

► $$F = \frac{\left(SSR^{R} - (SSR_1 + SSR_2)/k\right)}{(SSR_1 + SSR_2)/(n-2k)} \sim F(k, n-2k)$$

If $F \geq F_{(k,n-2k;\alpha)}$, reject H_0.

If $F < F_{(k,n-2k;\alpha)}$, do not reject H_0.

(Examples) Artprice file

◎ 그림 가격을 화가의 연령(AGE), 연령의 제곱(AGE^2), 그림의 크기(SIZE), 그림크기의 제곱(SIZE^2), 수상회수(ARD), 전시회수(EXB), 생몰여부(1 만일 생존화가이면, 0=사망화가이면)의 변수에 대해 회귀분석하는 모형에서 한국화(TYPE=0)와 서양화(TYPE=1) 간의 그림 구조에 차이가 있을 것이라는 가설을 검정하기 위하여 다음 세 회귀식의 결과를 구하였다.

$$\log \Pr ice_i = \alpha_1 + \alpha_2 Age_i + \alpha_3 Age^2{}_i + \alpha_4 Size_i + \alpha_5 Size^2{}_i + \alpha_6 Ard_i + \alpha_7 Exb_i + \alpha_8 Die_i + \varepsilon_i, \qquad \text{if Type}_\text{i}\text{=0}$$

$$\log \Pr ice_i = \beta_1 + \beta_2 Age_i + \beta_3 Age^2{}_i + \beta_4 Size_i + \beta_5 Size^2{}_i + \beta_6 Ard_i + \beta_7 Exb_i + \beta_8 Die_i + \varepsilon_i, \qquad \text{if Type}_\text{i}\text{=1}$$

(1) 전체 관찰치(n=250)

Model 1: OLS, using observations 1-250

Dependent variable: logprice

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	−%s	2.02697	−%#.4g	0.1092	
AGE	0.172248	0.0615860	2.797	0.0056	***
ARD	0.0885329	0.0385425	2.297	0.0225	**
EXB	−%s	0.00256821	−%#.4g	0.7290	
LIFE	0.363957	0.170114	2.139	0.0334	**
SIZE	0.0288943	0.00481476	6.001	<0.0001	***
sq_AGE	−%s	0.000458325	−%#.4g	0.0153	**
sq_SIZE	−%s	1.43586e-05	−%#.4g	<0.0001	***

Mean dependent var	3.909161	S.D. dependent var	1.304825
Sum squared resid	328.2114	S.E. of regression	1.164580
R-squared	0.225806	Adjusted R-squared	0.203412
F(7, 242)	10.08328	P-value(F)	4.64e-11
Log-likelihood	−388.7593	Akaike criterion	793.5185
Schwarz criterion	821.6902	Hannan-Quinn	804.8568

(2) 한국화($n_1 = 47$)

Model 3: OLS, using observations 203-249 (n = 47)

Dependent variable: logprice

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3.37757	3.99428	0.8456	0.4029	
AGE	−%s	0.119113	−%#.4g	0.5953	
ARD	0.127880	0.0794365	1.610	0.1155	
EXB	−%s	0.00799204	−%#.4g	0.6921	
LIFE	0.127768	0.340998	0.3747	0.7099	
SIZE	0.100573	0.0387068	2.598	0.0132	**
sq_AGE	0.000626785	0.000895866	0.6996	0.4883	
sq_SIZE	−%s	0.000662179	−%#.4g	0.0366	**

Mean dependent var	3.583671	S.D. dependent var	1.151653
Sum squared resid	35.50723	S.E. of regression	0.954171
R-squared	0.418010	Adjusted R-squared	0.313551
F(7, 39)	4.001643	P-value(F)	0.002185
Log-likelihood	−60.10044	Akaike criterion	136.2009
Schwarz criterion	151.0021	Hannan-Quinn	141.7707

► (gretl Tips) 전체 표본 중 일부인 TYPE=0인 표본설정을 위해서는 [Sample]/[Restrict based on criterion]에서 TYPE=0 을 입력하여 표본을 재 설정하여야 한다.

(3) 서양화($n_2 = 203$)

Model 4: OLS, using observations 1-203

Dependent variable: logprice

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	−%s	2.29839	−%#.4g	0.1278	
AGE	0.173325	0.0713544	2.429	0.0160	**
ARD	0.120000	0.0437528	2.743	0.0067	***
EXB	−%s	0.00275563	−%#.4g	0.6286	
LIFE	0.635370	0.197006	3.225	0.0015	***
SIZE	0.0310363	0.00509961	6.086	<0.0001	***
sq_AGE	−%s	0.000535786	−%#.4g	0.0434	**
sq_SIZE	−%s	1.48552e-05	−%#.4g	<0.0001	***

Mean dependent var	3.984521	S.D. dependent var	1.329030
Sum squared resid	257.8539	S.E. of regression	1.149925
R-squared	0.277309	Adjusted R-squared	0.251366
F(7, 195)	10.68928	P-value(F)	2.26e-11
Log-likelihood	−312.3220	Akaike criterion	640.6440
Schwarz criterion	667.1497	Hannan-Quinn	651.3672

► (gretl Tips) 전체 표본 중 일부인 TYPE=1인 표본설정을 위해서는 [Sample]/[Restrict based on criterion]에서 TYPE=1 을 입력하여 표본을 재 설정하여야 한다.

► 전체 회귀식의 잔차항 제곱=328.2

한국화 회귀식의 잔차항 제곱=35.5

서양화 회귀식의 잔차항 제곱=257.9

이상의 결과에서

$$\mathrm{F} = \frac{\{328.2 - (257.9 + 35.5)\} / 8}{(257.9 + 35.5) / (250 - 2 \times 8)} = 3.48 > F(8, 234; 0.05) = 1.98$$

이므로 두 장르(한국화와 서양화)의 가격구조가 같다는 가설을 기각한다.