Лекція № 3 Багатовимірна лінійна регресія

Під терміном багатовимірна лінійна регресія ми будемо розуміти метод побудови та дослідження залежності однієї змінної від кількох незалежних змінних. Загалом, багатовимірна регресія регресія дозволяє досліднику визначити, "що є кращим поясненням для...". Наприклад, досліднику в галузі освіти було би цікаво, які фактори є кращими поясненням успішного навчання в середній школі, психолога могло би зацікавити питання, які індивідуальні якості дозволяють краще прогнозувати ступінь соціальної адаптації індивіда, соціологам, імовірно, було б цікаво знайти ті соціальні індикатори, які найкраще пояснюють результат адаптації нової іммігрантської групи і ступінь її злиття із суспільством.

Модель багатовимірної (множинної) лінійної регресії можна записати у вигляді:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n + \xi,$$

де n — кількість факторів (екзогенних змінних) регресії, ξ — випадкова величина, що описує вплив сторонніх факторів. Всі x_i є значеннями спостережуваних факторів.

Застосувавши метод найменших квадратів, можна обчислити оцінки коефіцієнтів множинної регресії. Важливими є припущення про те, що:

- похибки спостережень є нормально розподіленими випадковими величинами з нульовим середнім;
- дисперсія похибок є сталою (гомоскедастичність);
- похибки неавтокорельовані;
- екзогенні змінні незалежні (некорельовані)

Якщо ж деякі з екзогенних змінних корелюють між собою, то говорять про явище мультиколінеарності.

Мультиколінеарність не дозволяє правильно визначити оцінки коефіцієнтів у регресійній моделі при корелюючих змінних, і тому варто переглянути питання специфікації моделі, виходячи з міркувань відповідної (економічної, соціологічної та іншої) теорії.

В пакеті STATISTICA методи багатовимірного регресійного аналізу реалізовані в пункті Multiple Regression меню Statistics. Основне вікно регресійного аналізу (рис. 1) містить крім можливості вибору залежної (Dependent) та незалежних (Independent) змінних ще і можливість вибору опцій покрокового аналізу (Advanced options (stepwise or ridge regression)), перегляду описових статистик та кореляційних матриць (Review descriptive statistics, correlation matrix), застосування високоточних обчислень (Extended precision computations), пакетної обробки (Batch processing), аналізу залишків (Print/report residual analysis). Серед способів видалення пропущених даних маємо додаткову опцію заміни їх середніми значеннями (Mean substitution). Крім того можна вибрати тип файлу даних (Input file:): вихідні дані (Raw Data) чи кореляційна матриця (Correlation Matrix). Кореляційна матриця повинна бути збережена у файлі спеціального формату, який крім самої матриці містить середні значення, стандартні відхилення, об'єм вибірки та індикатор того, що електронна таблиця є матрицею.

Стандартна процедура побудови оцінок параметрів лінійної регресії виводить вікно загальних результатів регресійного аналізу (рис. 2). Кнопка (Summary: Regression results) виводить оцінки параметрів лінійної регресії і статистичну інформацію про значимість кожного параметра та адекватність лінійної моделі (рис. 3). В заголовку таблиці Regression Summary for Dependent Variable наведено статистичні параметри лінійної регресії:

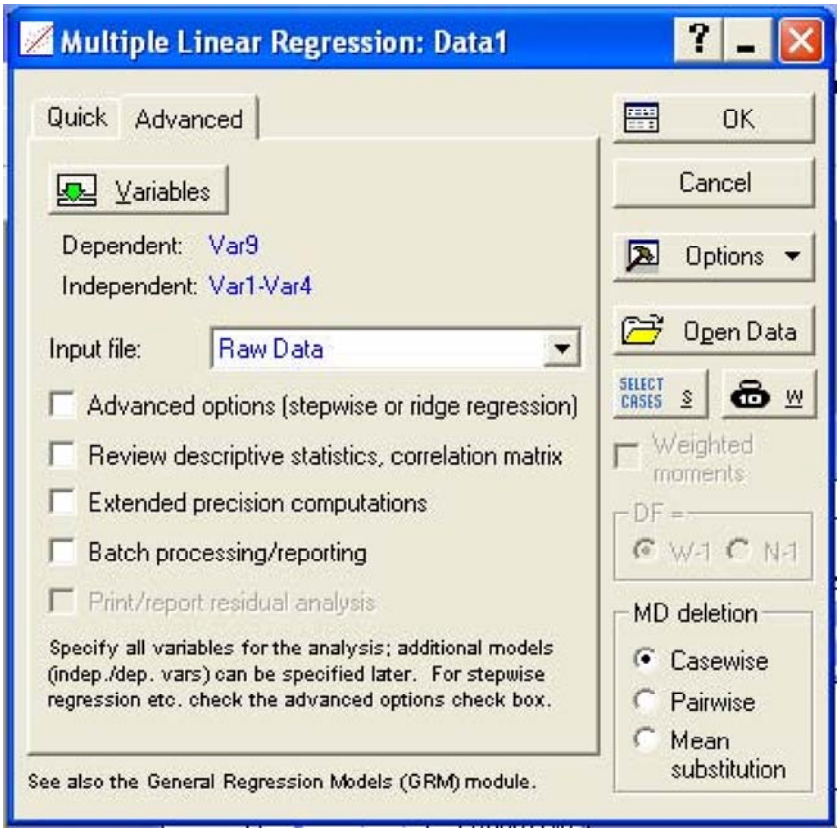

Рис. 1: Множинна регресія

- R — квадратний корінь з коефіцієнта детермінації;
- R? — коефіцієнт детермінації (квадрат коефіцієнта кореляції між прогнозованими та спостережуваними значеннями залежної змінної) — відношення суми квадратів відхилень залежної змінної від свого середнього значення поясненої регресією до загальної суми квадратів відхилень;
- Adjusted R? — скорегований коефіцієнт детермінації — те ж відношення що і для R?, тільки суми квадратів діляться на відповідні кількості ступенів вільності;
- F — значення статистики Фішера перевірки гіпотези про адекватність лінійної моделі;
- p — рівень значущості моделі;
- Std.Error of estimate — міра відхилень спостережуваних значень залежної змінної відносно прогнозованих значень.

Сама ж таблиця містить оцінки параметрів лінійної регресії та їх характеристики:

- Intercept — оцінка вільного члена (b_0);
- B — оцінки коефіцієнтів регресії при незалежних змінних;
- Beta — ті ж оцінки, побудовані за стандартизованими значеннями незалежних змінних (дозволяють порівнювати вклад різних змінних в модель);
- St.Err.of B, St.Err.of Beta — стандартні помилки оцінок параметрів регресії;

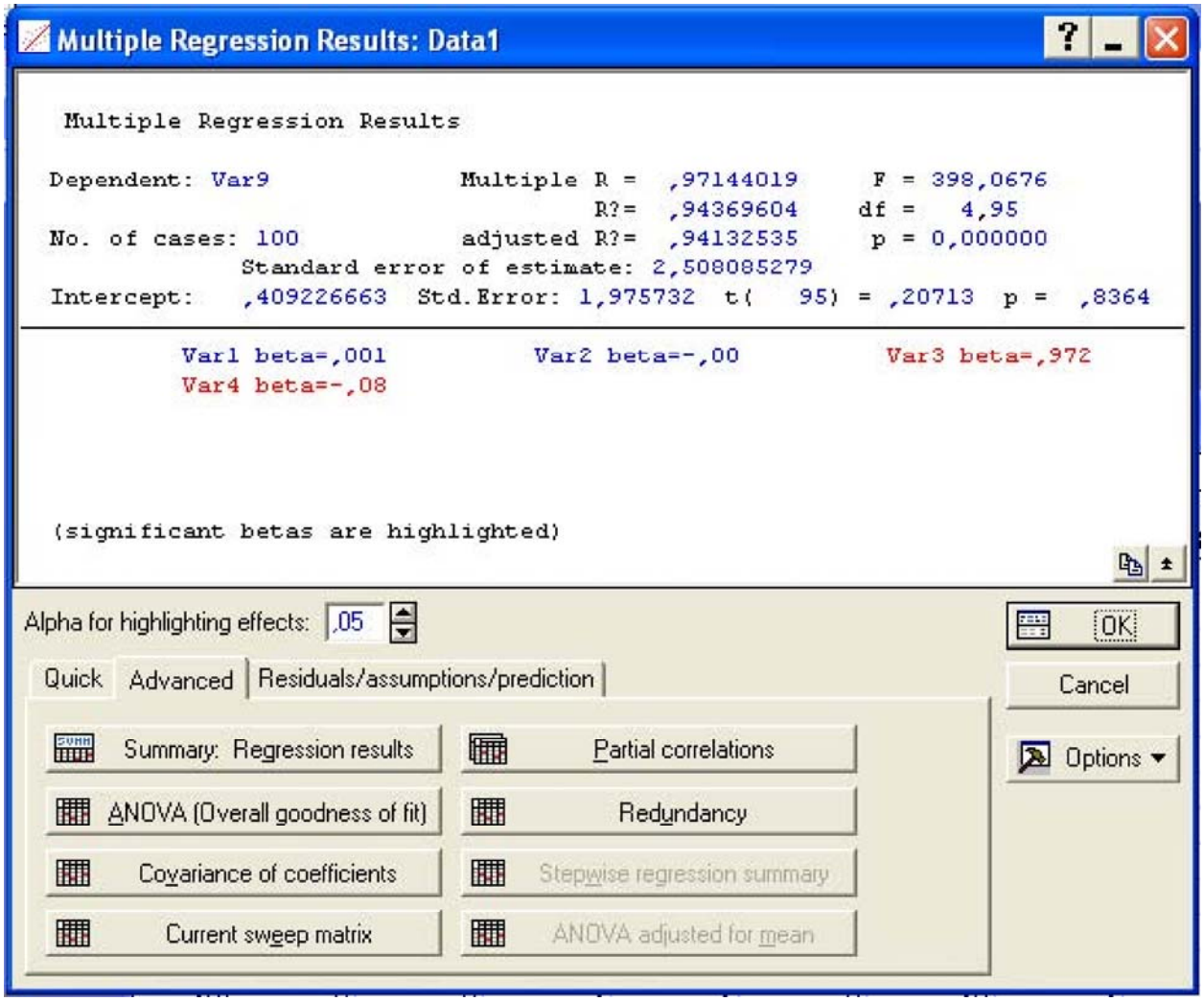

Рис. 2: Загальні результати регресійного аналізу

Data: Regression Summary for Dependent Variable: Var9 (Data1)*

Regression Summary for Dependent Variable: Var9 (Data1)
R= ,97144019 R?= ,94369604 Adjusted R?= ,94132535
F(4,95)=398,07 p<0,0000 Std.Error of estimate: 2,5081

N=100	Beta	Std.Err. of Beta	B	Std.Err. of B	t(95)	p-level
Intercept			0,409227	1,975732	0,20713	0,836354
Var1	0,000611	0,024786	0,002047	0,083077	0,02464	0,980397
Var2	-0,002995	0,024576	-0,022748	0,186677	-0,12186	0,903269
Var3	0,971832	0,024650	6,944827	0,176152	39,42513	0,000000
Var4	-0,076889	0,024462	-0,940353	0,299167	-3,14324	0,002229

Data: Summary Statistics; DV: Var9 (Data1)*

Summary Statistics; DV: Var9 (Data1)

Statistic	Value
Multiple R	0,9714
Multiple R?	0,9437
Adjusted R?	0,9413
F(4,95)	398,0676
p	0,0000
Std.Err. of Estimate	2,5081

Рис. 3: Підсумки регресійного аналізу

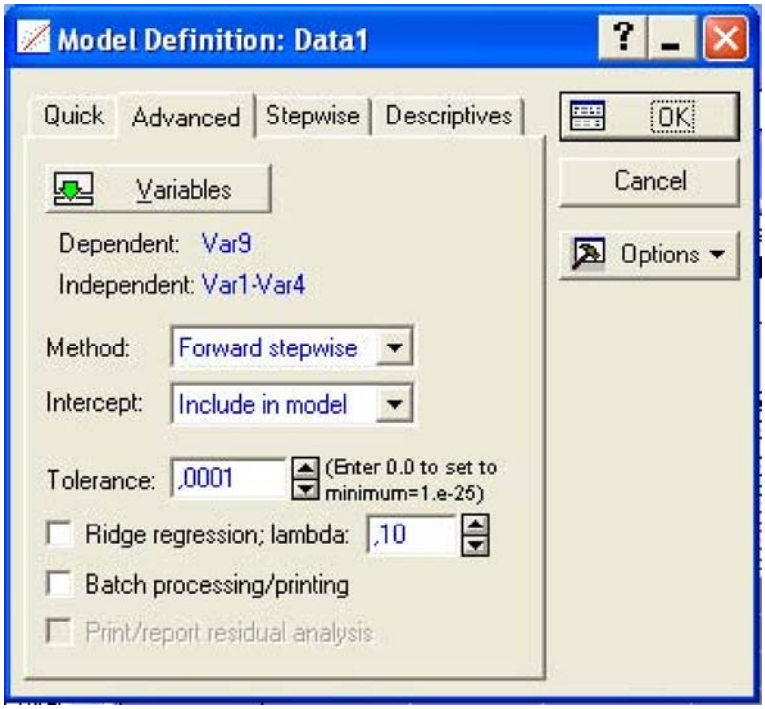

Рис. 4: Параметри моделі покрокового регресійного аналізу

- t( ) — значення статистики критерію перевірки гіпотези про значимість (відмінність від 0) параметрів регресії;
- p-level — максимальне значення рівеня значущості, на якому ця гіпотеза підтверджується.

Червоним кольором відмічені ті параметри, в яких p-level$<$0,05 (цю межу можна змінити в Alpha for highlighting effects).

Зауважимо. що в стандартному методі використовуються всі незалежні змінні для пояснення значень залежної змінної. На відміну від нього в покроковому регресійному аналізі спочатку найбільш впливова змінна вводиться до моделі. потім — наступна за впливом змінна, і так до тих пір, поки внесок чергової змінної не виявиться меншим деякого порогового значення. Для виконання покрокового аналізу потрібно вибрати Advanced options (stepwise or ridge regression) (рис. 1) та вибрати параметри моделі (рис. 4). Результати покрокового регресійного аналізу (рис. 5) показують, що до моделі слід включати тільки ті незалежні змінні, які найбільш повно задають залежну змінну.

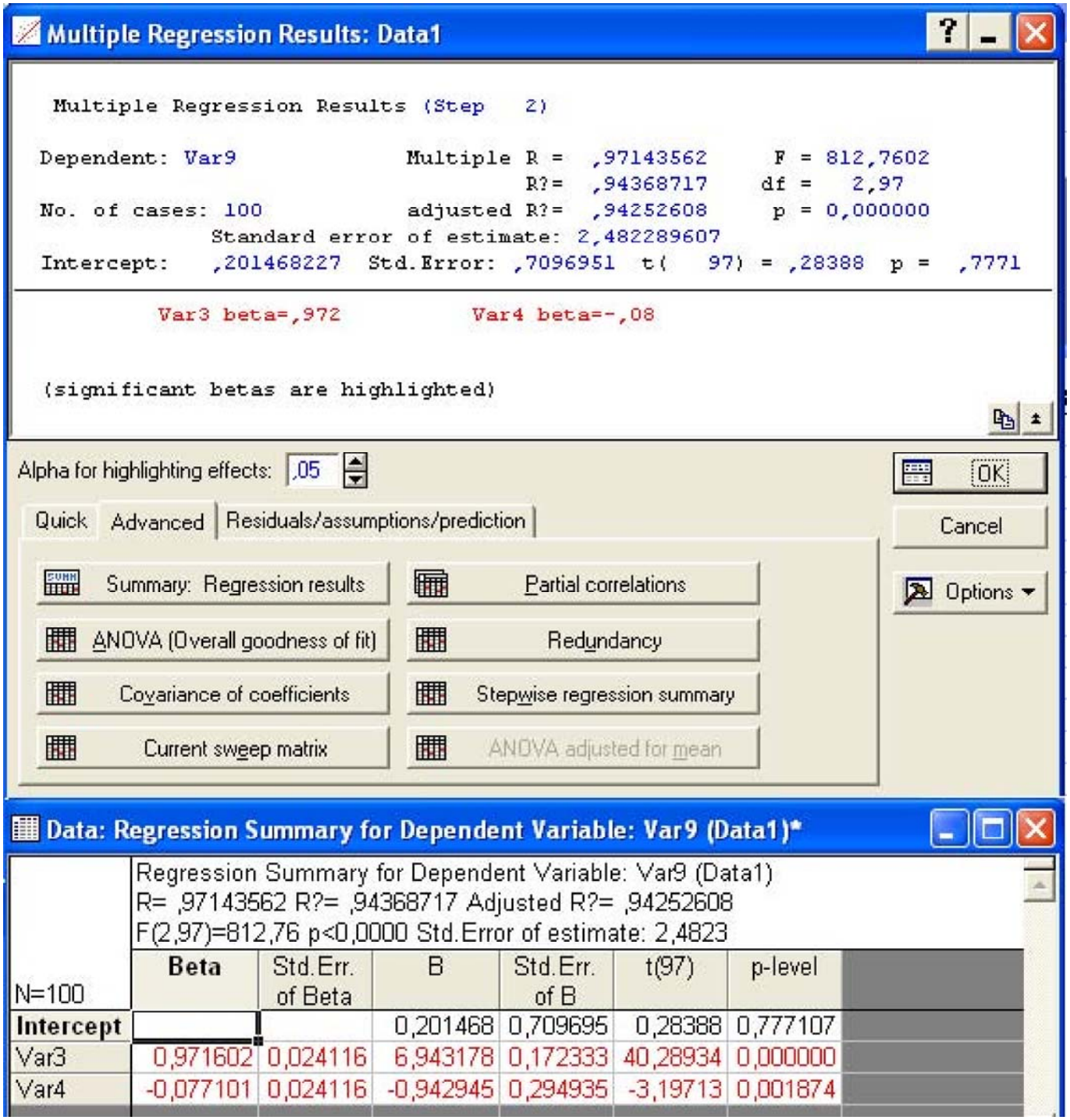

Рис. 5: Результати покрокового регресійного аналізу