

OBSAH

Charakteristika a význam korelačnej a regresnej analýzy

Jednoduchá lineárna regresná analýza

Viacnásobná lineárna regresná analýza

Vzorové údaje 1.5

Praktické použitie programu SAS (SAS Enterprise Guide) - korelačná analýza

Príklad 1.10 (SAS) - korelačná analýza

Príklad 1.11 (SAS) - jednoduchá lineárna regresná analýza

Príklad 1.12 (SAS) - viacnásobná lineárna regresná analýza (2 nezávisle premenné)

Príklad 1.13 (SAS) - viacnásobná lineárna regresná analýza (viac nezávisle premenných)

Praktické použitie programu EXCEL (Microsoft 365)

Príklad 1.10 (EXCEL) - korelačná analýza

Príklad 1.11 (EXCEL) - jednoduchá lineárna regresná analýza

Praktické použitie programu R (R Studio)

Príklad 1.10 (Program R) - korelačná analýza

Príklad 1.11 (Program R) - jednoduchá lineárna regresná analýza

Príklad 1.12 (Program R) - viacnásobná lineárna regresná analýza (2 nezávisle premenné)

Príklad 1.13 (Program R) - viacnásobná lineárna regresná analýza (viac nezávisle premenných)

Ďalšie príklady použitia korelačnej a regresnej analýzy

Zdroje a zoznam použitej literatúry

Charakteristika a význam korelačnej a regresnej analýzy

Korelačná a regresná analýza umožňujú analyzovať vzťahy medzi jednotlivými spojitými premennými (kvantitatívnymi znakmi) hodnoteného súboru údajov, časti, alebo kompletnej databázy údajov.

Základom korelačnej analýzy je výpočet tzv. korelačnej štatistiky, ktorá by mala byť doplnená preskúmaním bodových grafov jednotlivých vzťahov. Bodové grafy opisujú vzťah medzi dvoma spojitými premennými. Grafické vyjadrenie vzťahov môže napomôcť k lepšiemu pochopeniu a presnejšej interpretácii získaných základných výsledkov analýz. Medzi najčastejšie používanú korelačnú štatistiku patrí Pearsonova korelačná štatistika, ktorá meria stupeň lineárneho vzťahu medzi dvoma premennými pomocou jednoduchého korelačného koeficienta.

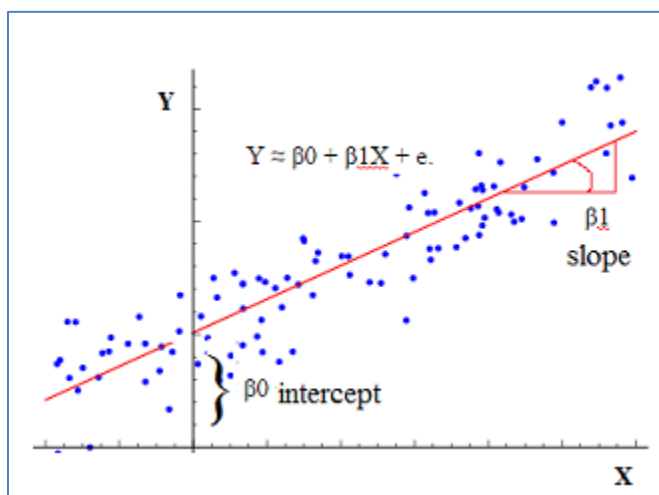
Regresná analýza definuje funkčný vzťah medzi spojitou závislou premennou a spojitou vysvetľujúcou (nezávislou) premennou. Najčastejšie sú používané jednoduchá lineárna závislosť, viacnásobná lineárna závislosť a logistická závislosť (logistická regresia). Predpoklady lineárnej regresie sú: priemer závisle premennej je lineárne závislý na vysvetľujúcej premennej, jednotlivé pozorovania sú nezávislé, chyby vysvetľujúcej premennej majú normálne rozdelenie a chyby ich rozptylu sú rovnaké. V špecifických prípadoch sa ale pri regresnej analýze využívajú aj mnohé ďalšie nelineárne regresné závislosti.

Jednoduchá lineárna regresná analýza

Vzťah medzi závisle premennou (Y) a nezávisle premennou (X) je možné popísať jednoduchou lineárnou regresnou rovnicou:

$$Y = \beta_0 + \beta_1 \cdot X + e$$

Y	závisle premenná
X	nezávisle (vysvetľujúca) premenná
β_0	intercept parameter, ktorý zodpovedá hodnote závisle premennej, keď nezávisle premenná má hodnotu 0 (priesečník osi)
β_1	parameter sklonu priamky, ktorý zodpovedá veľkosti zmeny závisle premennej danej zmene o jednu jednotku nezávisle premennej (sklon priamky)
e	chyba reprezentujúca odchýlku Y o $\beta_0 + \beta_1 X$.



Obr. 5.1

zdroj: https://datacadamia.com/data_mining/simple_regression

Cieľom jednoduchšej lineárnej regresie je charakterizovať vzťah medzi závisle a nezávisle premennými v populácii. Naše údaje, ktoré predstavujú v skutočnosti vzorku údajov, použijeme na odhad neznámych parametrov celej populácie (β_0 , β_1), ktoré definujú predpokladaný vzťah medzi závisle a nezávisle premennými.

Odhady neznámych parametrov populácie možno získať metódou najmenších štvorcov. Metóda poskytuje odhady pomocou priamky, ktorá minimalizuje súčet veľkosti štvorcov medzi pozorovaniami a preloženou priamkou. Vertikálne vzdialenosti pozorovaní od priamky sa často označujú ako chyba.

Pri pomerne veľkej a súčasne reprezentatívnej vzorke údajov by preložená regresná priamka mala byť dobrým odhadom vzťahu medzi závislými a nezávislými premennými v celej populácii. Odhadované parametre získané pomocou metódy najmenších štvorcov by mali byť preto kvalitnými aproximáciami skutočných parametrov populácie.

Ak chceme zistiť, či je jednoduchý lineárny regresný model presnejší ako základný model, musíme porovnať vysvetlenú variabilitu znaku s jej nevysvetlenou variabilitou. Za základný model sa považuje model pri ktorom parameter $\beta_0 = 0$.

Vysvetlená variabilita súvisí s rozdielom medzi regresnou čiarou a priemerom závisle premennej. Suma štvorcov modelu (Sum of Squares Model) je množstvo variability vysvetlenej modelom. Nevysvetlená variabilita súvisí s rozdielom medzi pozorovanými hodnotami a regresnou priamkou. Suma chýb štvorcov (Sum of Squers Error) je množstvo variability, ktorú model nevysvetlil. Celková variabilita súvisí s rozdielom medzi pozorovanými hodnotami a priemerom závisle premennej. Korigovaný celkový súčet štvorcov (Sum of Squers Corrected Total) je súčtom vysvetlenej a nevysvetlenej variability.

Pri jednoduchjej lineárnej regresnej analýzy overujeme v skutočnosti na začiatku nasledovné hypotézy:

Hypotéza H_0 : jednoduchý lineárny regresný model nevyhovuje údajom lepšie ako základný model

$$\beta_1 = 0$$

Odhadovaný jednoduchý lineárny regresný model nezodpovedá údajom lepšie ako základný model, nemáme dostatok dôkazov na to, aby sme mohli konštatovať, že sklon regresnej priamky je výrazne odlišný od 0. Preto nezávislá premenná nevysvetľuje štatisticky významné množstvo variability závislej premennej.

Hypotéza H_1 : jednoduchý lineárny regresný model vyhovuje údajom lepšie ako základný model

$$\beta_1 \neq 0$$

Odhadovaný jednoduchý lineárny regresný model zodpovedá údajom lepšie ako základný model, máme dostatok dôkazov na to, aby sme mohli konštatovať, že sklon regresnej priamky je výrazne odlišný od 0. Preto nezávislá premenná vysvetľuje štatisticky významné množstvo variability závislej premennej.

Viacnásobná lineárna regresná analýza

Viacnásobná lineárna regresná analýza skúma vzťah závisle premennej a viacerých nezávislých premenných súčasne. Pri jednoduchjej lineárnej regresii sme vyjadrili vzťah medzi dvoma premennými pomocou priamky. Ak máme dve nezávislé premenné, môžeme vyjadriť vzťah medzi celkovo tromi premennými pomocou roviny.

Nulová hypotéza pri viacnásobnej lineárnej regresii je, že sklony priamok (odhady regresných koeficientov) pre všetky nezávisle premenné sú rovné nule ($\beta_1 = \beta_2 = \dots = \beta_k = 0$). Alternatívnou hypotézou je, že aspoň jeden sklon priamky (jeden odhadnutý regresný koeficient) sa nerovná nule. Ak odmietnete nulovú hypotézu, musíte určiť, ktoré z nezávislých premenných majú nenulové sklony priamky, a či sú skutočne pri analýze potrebné a štatisticky významné.

Testy odhadov parametrov nám pomáhajú určiť, ktoré konkrétne parametre sú nenulové, ale popri tom treba ich použitie starostlivo zvážiť. Testujú významnosť každej premennej, keď sa pridá k modelu, ktorý už obsahuje všetky ostatné nezávisle premenné.

Existujú viaceré možnosti výberu modelu. Metódy postupného výberu pomáhajú vybrať nezávisle premenné, ktoré sú najužitočnejšie pri vysvetľovaní alebo predpovedaní hodnotenej závisle premennej. Niektoré z metód postupného výberu sú Forward, Backward a Stepwise metódy.

Predpoklady pre použitie viacnásobnej regresie môžeme preskúmať podobným spôsobom ako pri jednoduchej lineárnej regresii vytvorením bodových grafov a grafov rezíduí. Podobne aj pri viacnásobnej lineárnej regresie identifikujeme body, ktoré môžu výrazne ovplyvniť zmenu regresnej rovnice a detekujeme možnú multikolinearitu.

Vzorové údaje 1.5 (zdroj: SAS Enterprise Guide, úprava Candrák 2021)

Športový oddiel chce vyhodnotiť telesnú kondíciu svojich členov, ktorí sa v minulosti intenzívne zaoberali športom. Jedným z meradiel kondície športovcov je spotreba kyslíka. Z uvedeného dôvodu sa preto merala jednotlivým osobám počas pretekov spotreba kyslíka a súčasne sa uskutočnilo aj zisťovanie niekoľkých ďalších meraní a parametrov, ako je vek, pulz, hmotnosť človeka. Jedným z cieľov vyhodnocovania telesnej kondície bývalých športovcov je, či niektorá z týchto ďalších premenných resp. parametrov môže pomôcť spresniť predpoveď spotreby kyslíka ako hlavného ukazovateľa hodnotenia telesnej kondície.

Súbor údajov **Fitness.xls** obsahuje nasledovné premenné:

Meno	meno člena športového oddielu
Pohlavie	muž - M, žena -Z
Beh	čas odbehnutia preteku v minútach
Vek	vek v rokoch
Hmot	hmotnosť športovca v kilogramoch
Kyslík	spotreba kyslíka (miera schopnosti využiť kyslík)
Pulz_beh	pulz meraný tesne po preteku
Pulz_kľud	pulz meraný v kľude
Pulz_max	maximálny pulz počas preteku
Výkon	celková hodnota telesnej kondície

Vzor súboru údajov:

Meno	Pohlavie	Beh	Vek	Hmot	Kyslík	Pulz_beh	Pulz_kľud	Pulz_max	Výkon
Donna	Z	8,17	42	68	59,57	166	40	172	14
Gracie	Z	8,63	38	82	60,06	170	48	186	13
Luanne	Z	8,65	43	86	54,30	156	45	168	13
Mimi	Z	8,92	50	71	54,63	146	48	155	11
Chris	M	8,95	49	81	49,16	180	44	185	11
Allen	M	9,22	38	89	49,87	178	55	180	12
Nancy	Z	9,40	49	76	48,67	186	56	188	10
Patty	Z	9,63	52	76	45,44	164	48	166	10
Suzanne	Z	9,93	57	59	50,55	148	49	155	9

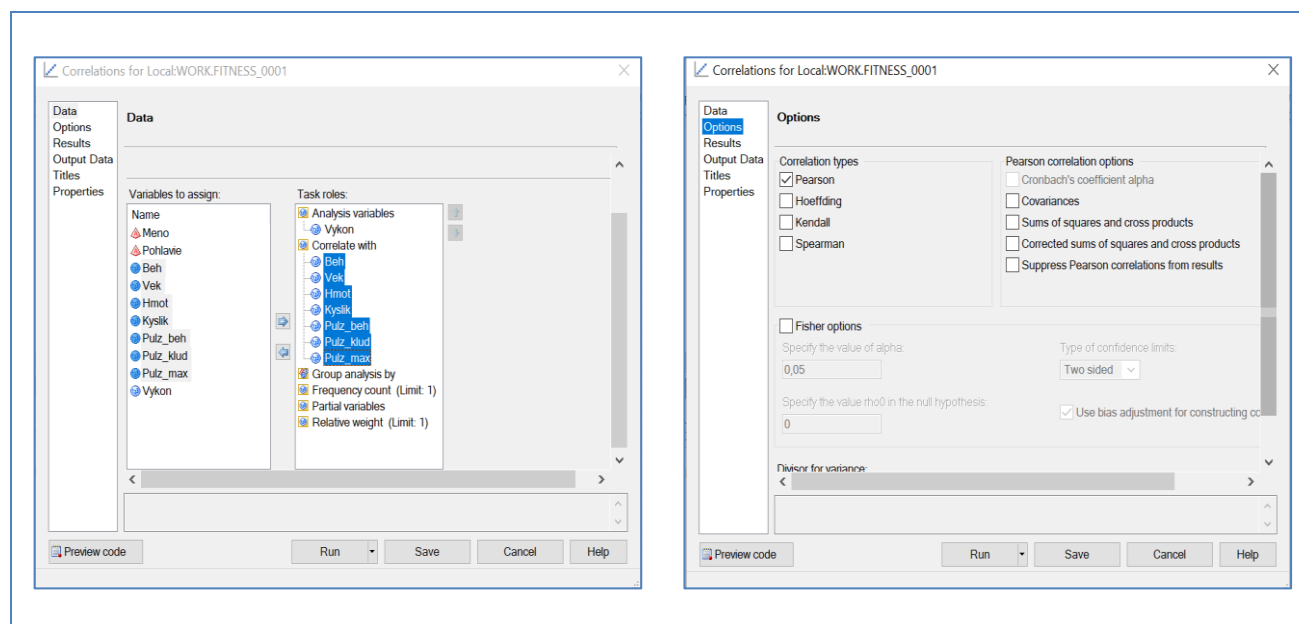
Praktické použitie programu SAS (SAS Enterprise Guide) - korelačná analýza

Úlohy (Tasks): Analyze - Multivariate - Correlation

Príklad 1.10 (SAS) - korelačná analýza

Na základe korelačnej analýzy zistíte aký je vzťah medzi výkonom bývalých športovcov a ostatnými sledovanými ukazovateľmi. Zadaná úloha má nasledovné nastavenia a parametre, ktoré umožnia vypočítať požadované korelačné koeficienty a zostaviť jednoduché bodové grafy analyzovaných vzťahov:

Tab. 5.1a Nastavenie korelačnej analýzy - príklad 1.10

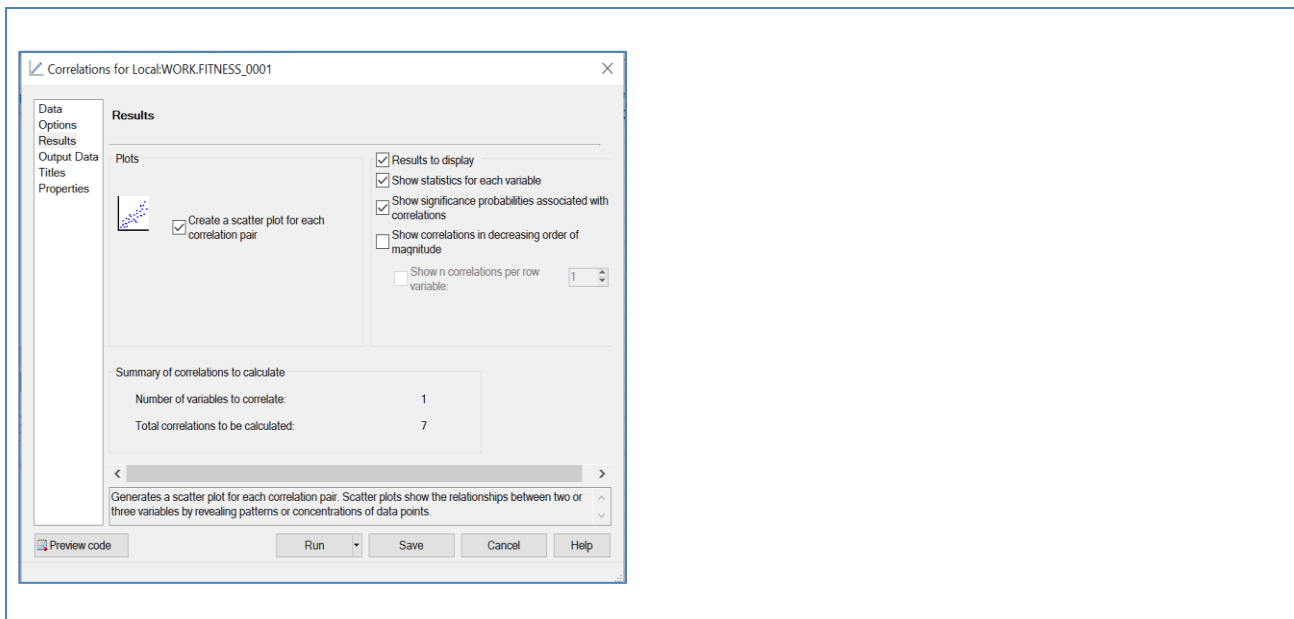


Vybratým typom korelačného koeficienta je jednoduchý (Pearsonov) korelačný koeficient, ktorý vyjadruje mieru intenzity (tesnosť) lineárneho vzťahu medzi dvoma náhodnými premennými.

Uvedené nastavenia nám poskytnú základnú analýzu vťahu výkonu športovca k ostatným sledovaným spojitým premenným.

V rámci jednotlivých nastavení je možnosť voľby viacerých typov korelačných koeficientov a ďalších doplnkových výpočtov (štandardne je nastavený Pearsonov korelačný koeficient). Samostatne je možné nastaviť zostavenie jednoduchých bodových grafov:

Tab. 5.1b Nastavenie grafov korelačnej analýzy - príklad 1.10



V prípade zostavovania jednoduchých bodových grafov vyjadrenia závislosti hodnotených premenných je ale možné použiť aj iné programovacie prostriedky. Ako príklad uvádzame možnosti zostavovania grafov v programe Microsoft Excel, ktorý umožňuje zostavovať veľmi kvalitné grafické výstupy s rozmanitou možnosťou tvorby trendových regresných grafov spolu s odhadnutými parametrami regresných funkcií.

Výsledky korelačnej analýzy a ich interpretácia

Tab. 5.2 Výsledky korelačnej analýzy - príklad 1.10

Correlation Analysis						
The CORR Procedure						
7 With Variables:		Beh	Vek	Hmot	Pulz_beh	Pulz_klud
1 Variables:		Kyslik				
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Beh	31	10.58613	1.38741	328.17000	8.17000	14.03000
Vek	31	47.67742	5.26236	1478	38.00000	57.00000
Hmot	31	77.44452	8.32857	2401	59.08000	91.63000
Pulz_beh	31	169.64516	10.25199	5259	146.00000	186.00000
Pulz_klud	31	53.45161	7.61944	1657	40.00000	70.00000
Pulz_max	31	173.77419	9.16410	5387	155.00000	192.00000
Vykon	31	8.00000	3.11983	248.00000	0	14.00000
Kyslik	31	47.37581	5.32777	1469	37.39000	60.06000
Pearson Correlation Coefficients, N = 31						
Prob > r under H0: Rho=0						
						Kyslik
						-0.86219
						<.0001
Beh						-0.31162
						0.0879
Vek						-0.16289
						0.3813
Hmot						-0.39808
						0.0266
Pulz_beh						-0.39935
						0.0260
Pulz_klud						-0.23677
						0.1997
Pulz_max						0.86377
						<.0001
Vykon						

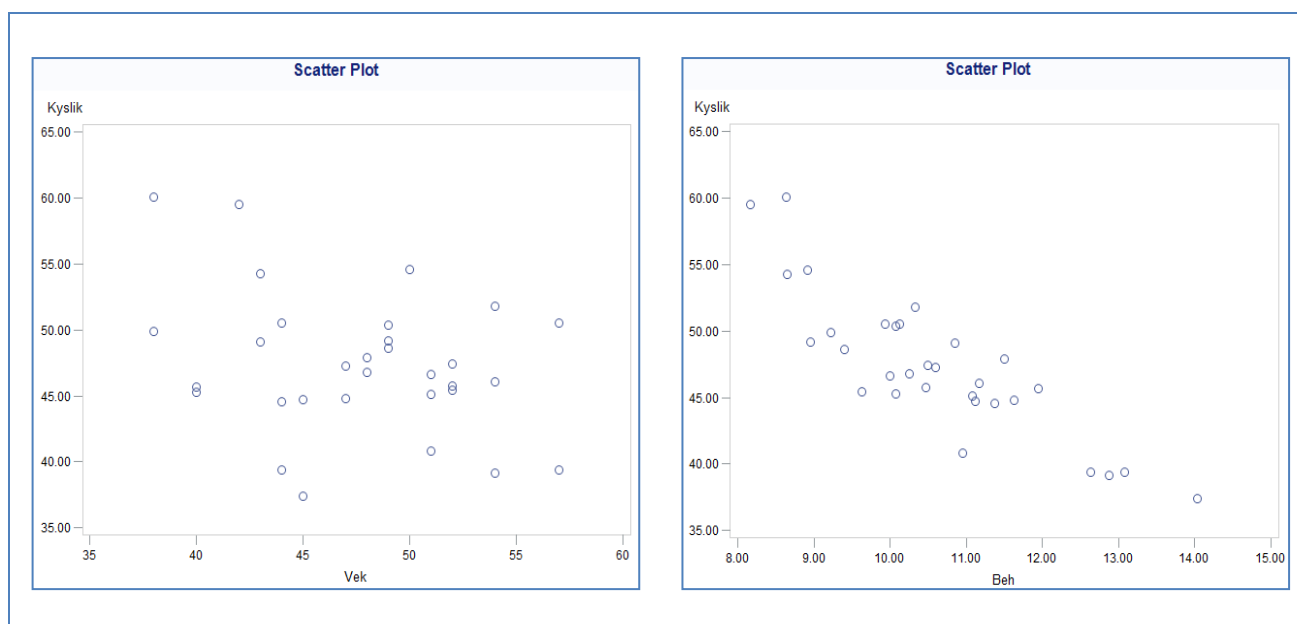
Korelačný koeficient vyjadrujúci vzťah medzi spotrebou kyslíka a samotným výkonom športovca je $+0,86377$. Hodnota p je veľmi nízka (<0.001), čo znamená, že tento koeficient sa výrazne líši od hodnoty nula. Druhý najväčší korelačný koeficient v zápornej hodnote $-0,86219$ bol zistený pri analýze vzťahu spotreba kyslíka a čas odbehnutia preteku.

Podobne aj tento vzťah je vysoko štatisticky preukazný. Štatisticky preukazné boli zistené aj vzťahy medzi spotrebou kyslíka a pulzom krátko po behu ($-0,39808$, $p=0,0266$) a pulzom v kľude ($-0,39935$, $p=0,0260$).

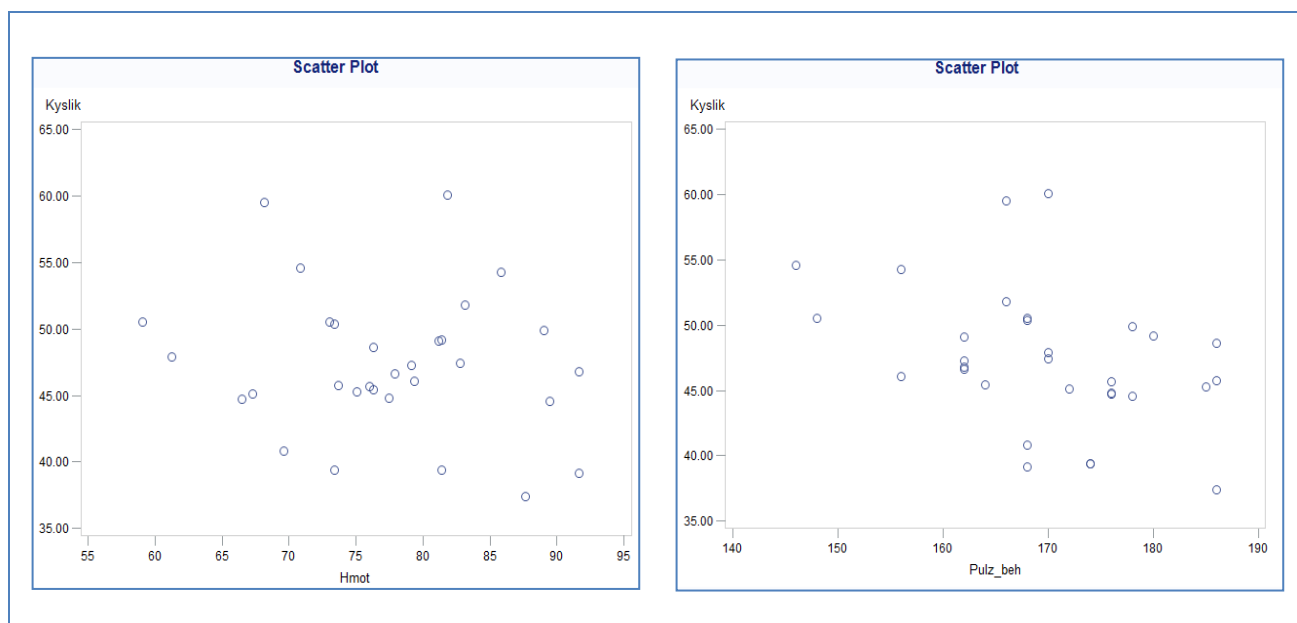
Pri týchto vzťahoch ale hodnoty korelačných koeficientov nedosiahli vysoké absolútne hodnoty. Uvedené výsledky korelačnej analýzy naznačujú, že viacero premenných môže byť dobrými predpoveďami skutočnej spotreby kyslíka v rámci športovej činnosti človeka. Následne vykonaná detailnejšia regresná analýza môže využiť práve tieto uvedené výsledky a závery.

Grafické znázornenie (bodové grafy) vzťahov medzi analyzovanými ukazovateľmi

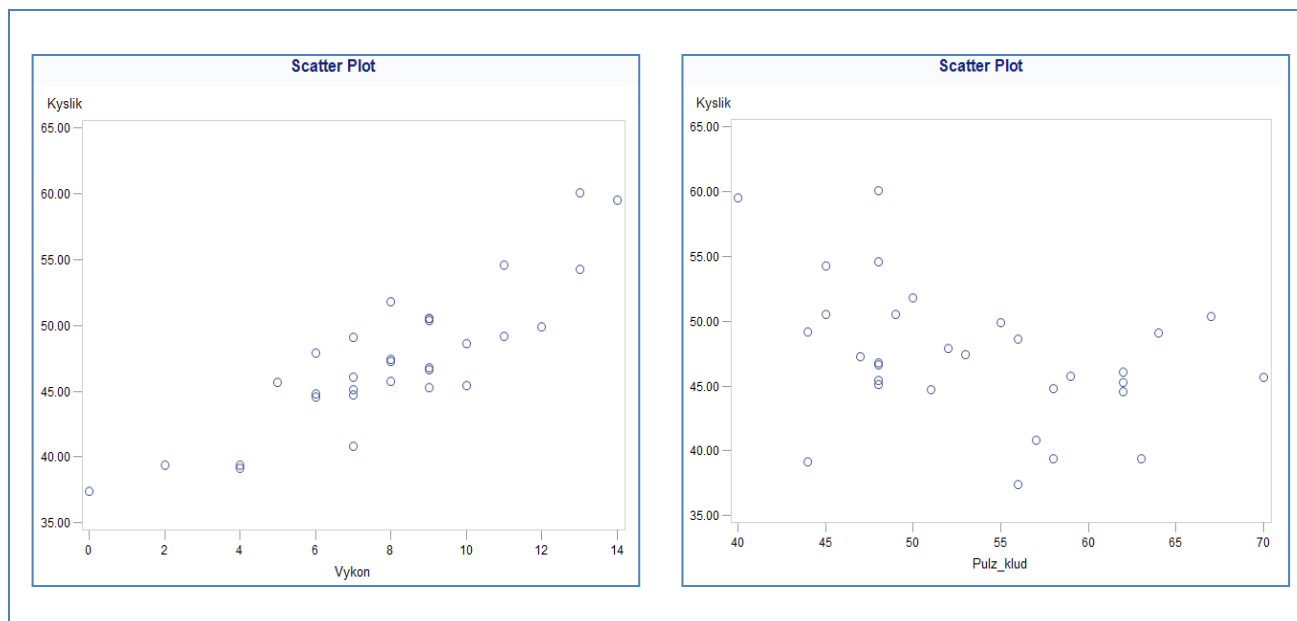
Tab. 5.3a Grafické zobrazenie korelačnej analýzy - príklad 1.10



Tab. 5.3b Grafické zobrazenie korelačnej analýzy - príklad 1.10



Tab. 5.3c Grafické zobrazenie korelačnej analýzy - príklad 1.10



Graf, ktorý vyjadruje vzťah spotreby kyslíka a behu naznačuje jednoznačne, že čím dlhšie jednotlivcovi trvá beh na stanovenú vzdialenosť, tým je nižšia spotreba kyslíka. Medzi spotrebou kyslíka a vekom je pravdepodobne veľmi slabý lineárny vzťah. Podobne to je aj medzi spotrebou kyslíka a hmotnosťou športovca. V prípade vzťahu pulzu na konci behu a pulzu v kľude k spotrebe kyslíka je dokázaný a štatisticky preukazný ale iba mierny negatívny vzťah. Naopak vzťah medzi spotrebou kyslíka a celkovým výkonom športovca je v absolútnej hodnote podobný vzťahu

spotreby kyslíka a behu čo znamená, že so zvyšujúcim sa výkonom stúpa jednoznačne spotreba kyslíka.

Jednou z dôležitých vlastností analýz využívajúcich lineárne regresné modelí je lineárna nezávislosť v rámci vysvetľujúcich premenných. Pokiaľ dochádza k porušeniu tohto predpokladu a vysvetľujúce premenné sú lineárne závislé, potom pri niektorých regresných koeficientoch môžeme mať problém s ich odhadom.

Tento problém nazývame multikolinearita. V prípade existencie výraznej lineárnej závislosti, môžu mať odhady koeficientov pomocou metódy najmenších štvorcov veľký rozptyl, čo znamená, že pri malej zmene hodnôt vysvetľujúcich premenných sa zmení výrazne odhad samotných koeficientov. Uvedené môže spôsobiť veľký problém ich praktického použitia. Práve z uvedeného dôvodu existencie multikolinearity a získania presnejších výsledkov regresnej analýzy je vždy dobrou praxou analyzovať korelácie medzi možnými vysvetľujúcimi (nezávislými) premennými. Je to jedna z metód ako odhaliť multikolinearitu v korelačno regresných analýzach.

V tomto konkrétnom prípade riešeného príkladu musíme uskutočniť samostatnú korelačnú analýzu, ktorou vygenerujeme maticu korelačných koeficientov medzi nezávislými premennými. Preukazné hodnoty korelačných koeficientov nad 0,8 resp. nad 0,9 upozorňujú na možný problém s multikolinearitou.

V tabuľke 5.4 uvádzame výsledky korelačnej analýzy medzi všetkými hodnotenými ukazovateľmi, ktoré nám môžu odhaliť niektoré zaujímavé vzťahy.

Tab. 5.4 Výsledky korelačnej analýzy - príklad 1.10

Correlation Analysis							
The CORR Procedure							
7 Variables: Beh Vek Hmot Pulz_beh Pulz_klud Pulz_max Vykon							
Pearson Correlation Coefficients, N = 31							
Prob > r under H0: Rho=0							
	Beh	Vek	Hmot	Pulz_beh	Pulz_klud	Pulz_max	Vykon
Beh	1.00000	0.19523	0.14351	0.31365	0.45038	0.22610	-0.98841
Vek	0.19523	1.00000	-0.24050	-0.31607	-0.15087	-0.41490	-0.22943
Hmot	0.14351	-0.24050	1.00000	0.18152	0.04397	0.24938	-0.10544
Pulz_beh	0.31365	-0.31607	0.18152	1.00000	0.35246	0.92975	-0.31369
Pulz_klud	0.45038	-0.15087	0.04397	0.35246	1.00000	0.30512	-0.47957
Pulz_max	0.22610	-0.41490	0.24938	0.92975	0.30512	1.00000	-0.22035
Vykon	-0.98841	-0.22943	-0.10544	-0.31369	-0.47957	-0.22035	1.00000
	<.0001	0.2144	0.5724	0.0857	0.0063	0.2336	

V rámci tejto korelačnej analýzy boli zistené dve štatisticky významné kombinácie veľmi silnej korelácie medzi behom a výkonom (-0,98841) a medzi pulzom tesne po behu a maximálnym pulzom (+0,92975). Uvedené je preto potrebné zohľadniť pokiaľ predpokladáme ďalšie využitie analýz na základe viacnásobnej lineárnej regresnej závislosti.

Praktické použitie programu SAS (SAS Enterprise Guide) - regresná analýza

Úlohy (Tasks): Analyze - Regression - Linear Regression

Príklad 1.11 (SAS) - jednoduchá lineárna regresná analýza

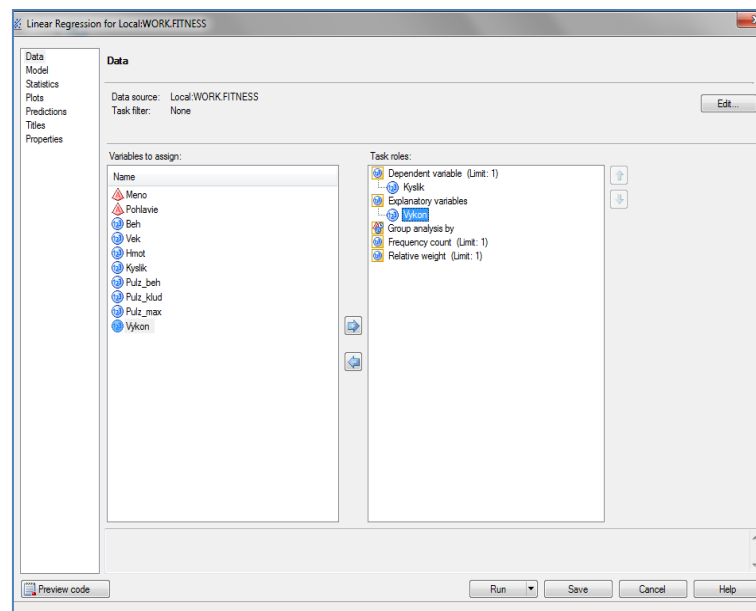
Na základe uskutočnenej korelačnej analýzy sme zistili, že spotreba kyslíka má najvýznamnejší vzťah s celkovou výkonnosťou hodnotených športovcov (+0,86377). Preto ďalšie detailnejšie skúmanie uskutočníme medzi týmito premennými na základe jednoduchej lineárnej regresie spotreby kyslíka a výkonu športovca (Príklad 1.11).

Vzťah medzi závisle premennou (spotrebou kyslíka) a nezávisle premennou (výkonom) je možné popísať jednoduchou lineárnou regresnou rovnicou:

$$\text{kyslík} = \beta_0 + \beta_1 \cdot \text{výkon} + e$$

Základné nastavenie parametrov regresnej analýzy pre jednoduchú lineárnu regresnú rovnicu:

Tab. 5.5 Nastavenie regresnej analýzy - príklad 1.11



V rámci nastavenia grafických výstupov je automaticky označené nastavenie všetkých vhodných a potrebných výstupov s možnosťou ich manuálneho výberu, alebo zrušenia.

Výsledky jednoduchšej lineárnej regresnej analýzy a ich interpretácia

Tab. 5.6 Výsledky lineárnej regresnej analýzy - príklad 1.11

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Kyslik

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	635.34150	635.34150	85.22	<.0001
Error	29	216.21305	7.45562		
Corrected Total	30	851.55455			

Root MSE	2.73050	R-Square	0.7461
Dependent Mean	47.37581	Adj R-Sq	0.7373
Coeff Var	5.76349		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	35.57526	1.36917	25.98	<.0001
Vykon	1	1.47507	0.15979	9.23	<.0001

Základné výsledky lineárnej regresnej analýzy obsahujú aj výsledky analýza rozptylu, ktorá poskytuje analýzu variability vysvetlenú pomocou regresnej priamky (Model), variability ktorá nie je vysvetlená modelom (Error) a celkovú variabilitu hodnoteného znaku (Corrected Total).

Pre posúdenie štatistickej významnosti regresného modelu je dôležitá vypočítaná hodnota F-testu (85,22) a p-hodnota spojená s hodnotou F-testu. F-hodnota v skutočnosti testuje, či je sklon regresnej priamky rovný 0. p-hodnota je v prípade nášho príkladu menšia ako 0,0001, takže môžeme zamietnuť nulovú hypotézu a dospieť k záveru, že jednoduchý lineárny regresný model zodpovedá údajom lepšie ako základný model. Inými slovami výkon športovca vysvetľuje vysoko štatisticky preukazne variabilitu kyslíka.

Dalším dôležitým parametrom analýzy je koeficient determinácie R^2 , ktorý môže nadobúdať hodnoty od 0 do 1. V našej analýze koeficient dosiahol hodnotu 0,7461.

Uvedené predstavuje podiel variability, ktorá je vysvetlená regresnou priamkou, čo znamená, že regresná priamka vysvetľuje približne 75 % celkovej variability spotreby kyslíka. V skutočnosti sa jedná o druhú mocninu už vypočítaného Pearsonovho korelačného koeficienta. Vo výsledkoch je ešte uvedený korigovaný koeficient determinácie ($\text{Adj } R^2$), ktorý je upravený pre počet parametrov v modeli a dosiahol hodnotu 0,7373. Tento typ koeficienta determinácie je užitočný pri viacnásobnej regresnej analýze. Najdôležitejšou časťou výsledkov jednoduchej regresnej analýzy je vlastný odhad regresných parametrov (koeficientov), odhad chýb parametrov a testovacích charakteristík. Pre posúdenie štatistickej významnosti je publikovaná p-hodnota spojená s t-štatistikou. V skutočnosti sa testuje, či sa odhad parametra štatisticky významne odlišuje od 0. V našom príklade sa parameter β_0 a parameter β_1 (sklon priamky a aj výkon športovca) vysoko štatisticky preukazne odlišujú od 0.

Odhadnutú jednoduchú lineárnu regresnú rovnicu môžeme preto na základe uvedených výsledkov zapísať nasledovne:

$$\text{kyslík} = 35,57526 + 1,47507 \cdot \text{výkon}$$

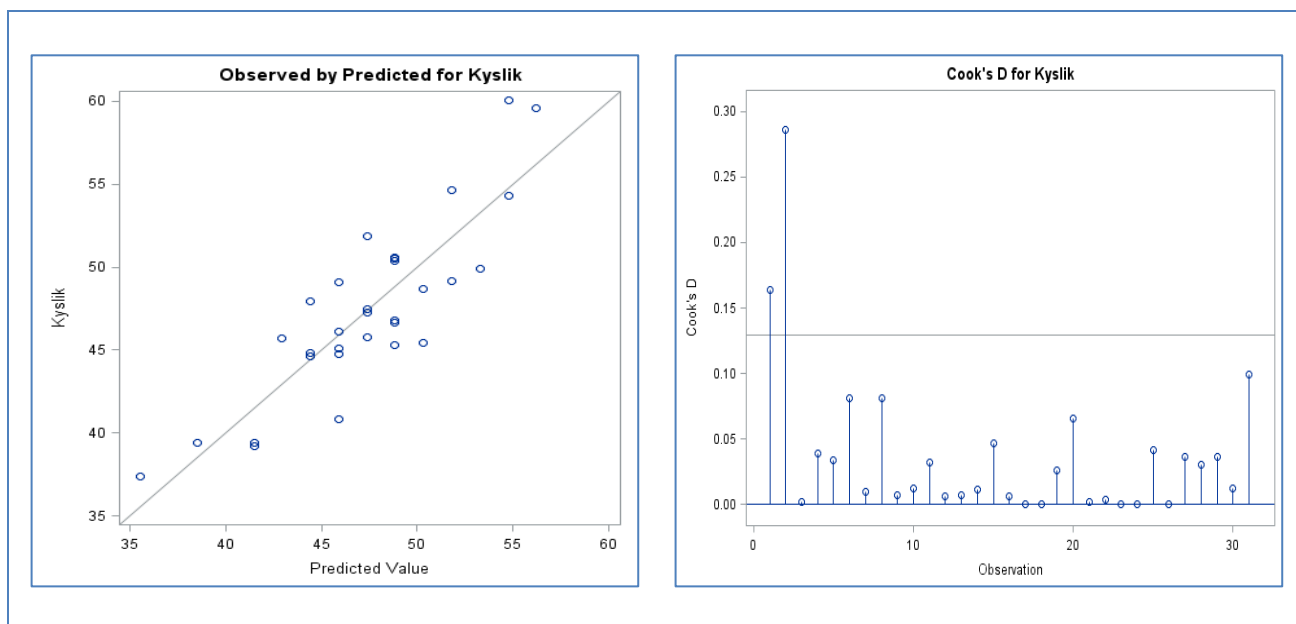
Odhadnutá lineárna regresná rovnica umožňuje tvrdiť, že zvýšenie výkonu športovca o jednu jednotku predstavuje zvýšenie spotreby kyslíka o 1,48 jednotky. Uvedená rovnica je však vhodná len v rozsahu hodnôt, ktoré sme pozorovali pri premennej výkon športovca (rozsah od 0 do 14).

Odhadov parametru β_0 tiež ukazuje, že parameter sa nerovná 0. Test pre tento parameter má však praktický význam len vtedy, keď rozsah hodnôt pre premennú výkon športovca obsahuje hodnotu 0. V tomto príklade má test praktický význam, pretože výkon, ktorý sa rovná hodnote 0 sa nachádza v rozsahu hodnôt, ktoré sme analyzovali (rozsah od 0 do 14).

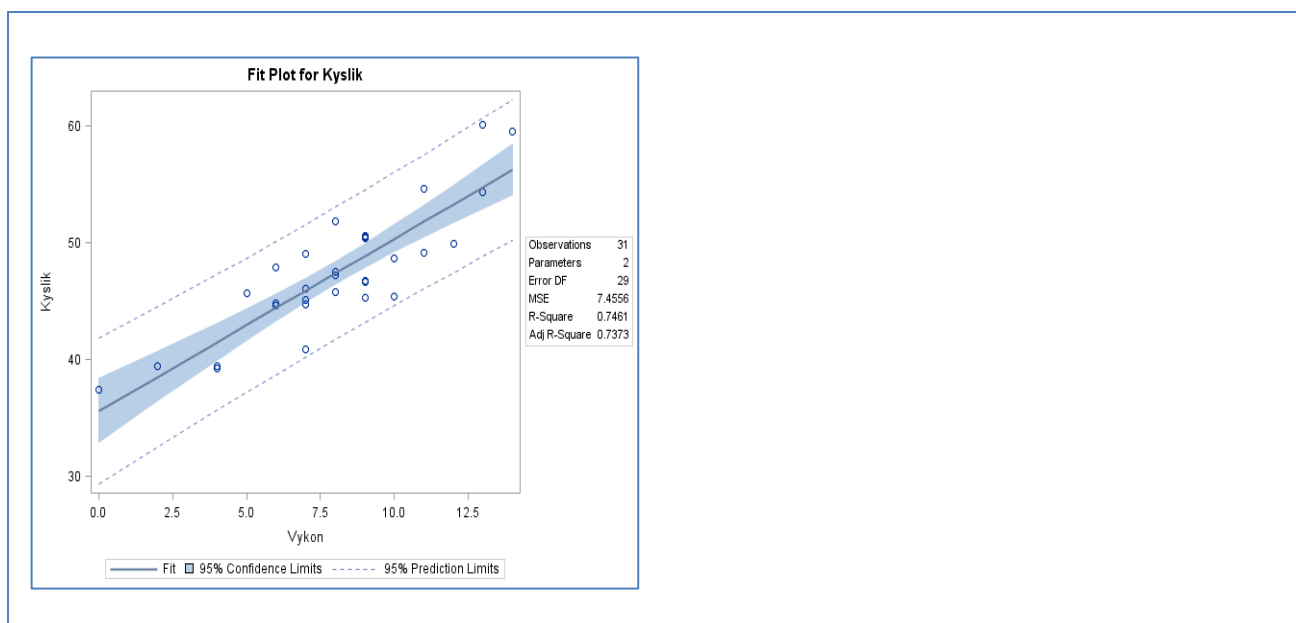
Na základe odhadnutej lineárnej regresnej rovnice je možné veľmi jednoducho predikovať hodnoty spotreby kyslíka pri stanovenej hodnote výkonu športovca.

V nasledovnej tabuľke 5.7 uvádzame graf pozorovaných hodnôt spotreby kyslíka versus predikované hodnoty na základe výkonu športovca. Hodnoty, ktoré sú umiestnené tesne v okolí priamky môžeme považovať za veľmi presné predikcie uskutočnené na základe použitej jednoduchej lineárnej regresnej rovnice.

Tab. 5.7a Grafické zobrazenie regresnej analýzy - príklad 1.11



Tab. 5.7b Grafické zobrazenie regresnej analýzy - príklad 1.11



Z dôvodu identifikácie bodov, ktoré negatívne ovplyvňujú použitý regresný model sa používa stanovenie a znázornenie tzv. Cookovej vzdialenosti (Cook's D "distance").

Body, ktoré sa nachádzajú nad vodorovnou priamkou predstavujú práve tie body, ktoré môžu ovplyvňovať odhadnuté parametre jednoduchšej lineárnej regresnej rovnici. V riešenom príklade nepredstavujú ale významnú početnosť. Z hľadiska grafu rezíduí bola splnená požadovaná podmienka zachovania normality chýb.

Z hľadiska interpretácie a prezentovania výsledkov regresnej analýzy je veľmi užitočný grafický výstup znázornenia nami analyzovaného lineárneho vzťahu so zohľadnením konfidenčných a predpovedných (predikčných) intervalov na stanovenej úrovni 95 %.

Ak chceme posúdiť úroveň presnosti na úrovni priemerných odhadov spotreby kyslíka, môžete vytvoriť aj iné intervaly spoľahlivosti. Interval spoľahlivosti na úrovni 95 % pre priemeré hodnoty znamená, že sme si na 95 % istí, že náš interval obsahuje priemernú hodnotu kyslíka (Y) pre konkrétnu hodnotu výkonu športovca (X).

Naopak intervaly predikcie majú význam hlavne vtedy, ak máme záujem o odvodenie budúcich pozorovaní. Interval s hodnotou 95 % predstavuje, že si môžeme byť na 95 % istý, že bude obsahovať aj nové pozorovanie. Intervaly predikcie sú širšie ako intervaly spoľahlivosti, pretože jednotlivé pozorovania majú väčšiu variabilitu ako priemery údajov.

Príklad 1.12 (SAS) - viacnásobná lineárna regresná analýza (2 nezávisle premenné)

Pri riešení príkladu 1.12 sme rozšírili regresnú analýzu o ďalšiu nezávisle premennú (pulz v kľude). Vzťah medzi závisle premennou (spotrebou kyslíka) a dvomi nezávisle premennými (výkon športovca a pulz v kľude) je možné popísať nasledovnou viacnásobnou lineárnou regresnou rovnicou:

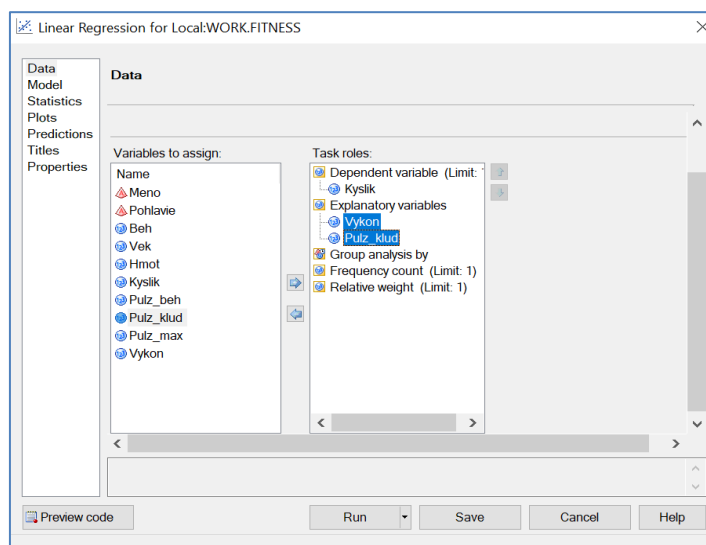
$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + e$$

Y	závisle premenná (kyslík)
X_1, X_2	nezávislé (vysvetľujúce) premenné (výkon a pulz v kľude)
β_0	intercept
β_1	parameter, ktorý zodpovedá veľkosti zmeny závisle premennej danej zmene o jednu jednotku prvej nezávisle premennej
β_2	parameter, ktorý zodpovedá veľkosti zmeny závisle premennej danej zmene o jednu jednotku druhej nezávisle premennej
e	chyba reprezentujúca odchýlku Y o $\beta_0 + \beta_1 X_1 + \beta_2 X_2$.

Rozšírenie viacnásobnej lineárnej regresnej analýzy o tretiu nezávisle premennú (beh športovca) by nebolo správne, pretože aj keď bol zistený veľmi silný vzťah medzi spotrebou kyslíka a behom (-0,86219), musí sa zohľadniť existencia multikolinearity medzi behom a výkonom športovca. V skutočnosti by nastalo porušeniu predpokladu nezávislosti medzi nezávislými premennými v modeli, čo je podmienkou správneho použitia lineárneho regresného modelu.

Základné nastavenie parametrov regresnej analýzy pre viacnásobnú lineárnu regresnú rovnicu (2 nezávisle premenné): **kyslík = $\beta_0 + \beta_1 \cdot \text{výkon} + \beta_2 \cdot \text{pulz v kľude} + e$**

Tab. 5.8 Nastavenie viacnásobnej regresnej analýzy - príklad 1.12



Nastavenia grafických výstupov, ktoré sú automatické, je pri viacnásobnej lineárnej regresii veľmi rovnaké ako pri jednoduchej lineárnej regresnej analýze. Dôležitým prvkom je optimálne využitie narastajúceho počtu grafických výstupov a logických interpretácií výsledkov analýzy. Grafy rezíduí nám potvrdzujú, alebo zamietajú z hľadiska ich normality správne použitie modelu.

Výsledky a interpretácia viacnásobnej lineárnej regresnej analýzy (2 nezávisle premenné)

Tab. 5.9 Výsledky lineárnej regresnej analýzy - príklad 1.12

Linear Regression Results					
The REG Procedure					
Model: Linear_Regression_Model					
Dependent Variable: Kyslik					
Number of Observations Read				31	
Number of Observations Used				31	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	635.58665	317.79333	41.20	<.0001
Error	28	215.96790	7.71314		
Corrected Total	30	851.55455			
Root MSE		2.77725	R-Square	0.7464	
Dependent Mean		47.37581	Adj R-Sq	0.7283	
Coeff Var		5.86218			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	34.72590	4.96357	7.00	<.0001
Vykon	1	1.49090	0.18521	8.05	<.0001
Pulz_klud	1	0.01352	0.07584	0.18	0.8598

Hodnota F-testu je dosiahla 41,20 a štatistická preukaznosť bola menej ako 0,0001. Prvým výsledkom analýzy je preto zamietnutie nulovej hypotézy, čo znamená, že aspoň jeden parameter sklonu priamky (regresný koeficient) nemá hodnotu 0. Koeficient determinácie R^2 dosiahol hodnotu 0,7464 čo je iba mierne vyššia hodnota akú sme dosiahli pri jednoduchnej lineárnej regresnej analýze.

Korigovaný koeficient determinácie ale dosiahol hodnotu 0,7283, čo je nižšia hodnota ako rovnaký koeficient determinácie pri jednoduchnej lineárnej regresnej analýze (0,7373). Uvedené jednoznačne naznačuje, že premenná pulz v kľude nevysvetľuje lepšie spotrebu kyslíka, ak poznáme výkon športovca.

Pri viacnásobnej lineárnej regresnej analýze je presnejšie a odporúčané použitie práve tohto koeficienta determinácie, ktorý zohľadňuje väčší počet zohľadňovaných premenných ako pri jednoduchnej lineárnej regresnej analýze. Jednoduchú regresnú analýzu môžeme považovať z tohto hľadiska dokonca za presnejšiu ako dvojnásobnú regresnú analýzu. Odhadnutý regresný parameter pre pulz v kľude nedosiahol štatistickú preukaznosť (hodnota $p=0,8598$). Na základe uvedeného môžeme konštatovať, že odhad parametra pre pulz v kľude sa výrazne nelíši od hodnoty 0 a tým nemá prakticky žiadny významný vplyv na spotrebu kyslíka.

Odhad parametra pre výkon (1,49090) je veľmi podobný tomu, čo bol v predchádzajúcom jednoduchom modeli (1,47507), ale jeho interpretácia je predsa len o niečo odlišná. Zvýšenie o jednu jednotku výkonu predstavuje zvýšenie spotreby kyslíka o približne 1,49 so zohľadnením pulzu v kľude. Inými slovami, ak sa udrží pulz v kľude konštantný, bude sklon priamky medzi kyslíkom a výkonom športovca mierne výraznejší.

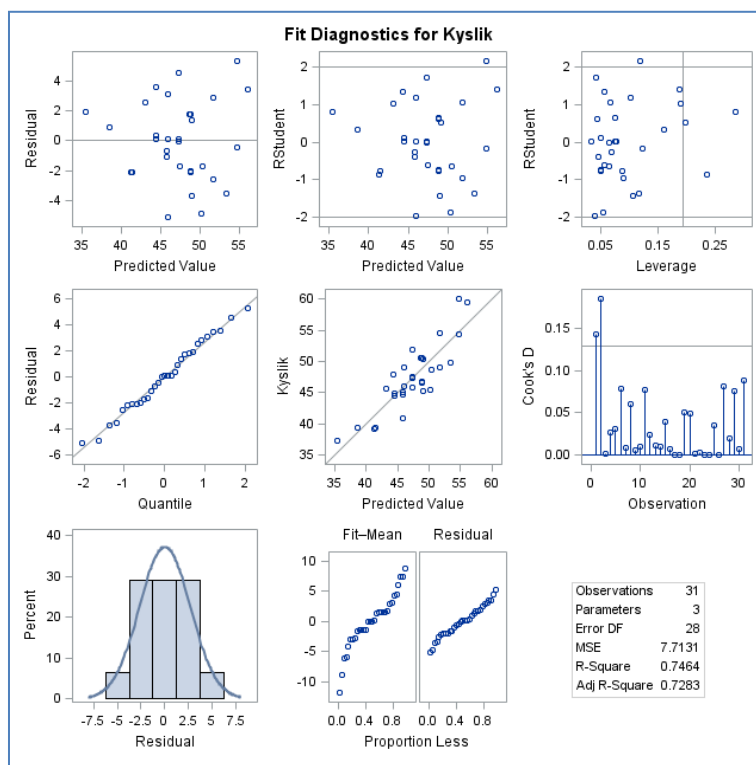
Hodnota p je pre tento parameter menšia ako 0,0001, takže sklon je samozrejme tiež výrazne odlišný od hodnoty 0.

Viacnásobnú lineárnu regresnú rovnicu môžeme preto na základe získaných výsledkov zapísať nasledovne:

$$\text{kyslík} = 34,72590 + 1,49090 \cdot \text{výkon} + 0,01352 \cdot \text{pulz v kľude}$$

Súhrnný grafický výstup dvojnásobnej lineárnej regresnej analýzy je uvedený v nasledujúcej tabuľke 5.10.

Tab. 5.10 Grafické zobrazenie viacnásobnej regresnej analýzy - príklad 1.12

**Príklad 1.13 (SAS) - viacnásobná lineárna regresná analýza (viac nezávisle premenných)**

Výhodou viacnásobnej lineárnej regresnej analýzy je súčasné zohľadnenie viacerých premenných na hodnotený znak, alebo vlastnosť. Pri jej použití ale narážame aj na niekoľko problémov. Jedným z nich je problém s jednoznačným určením, ktorý model, alebo ktorý postup je lepší a správnejší. Problém môže nastať aj so správnou interpretáciou a použitím výsledkov analýzy. V prípade analýzy biologických údajov resp. procesov ale výhody viacnásobnej lineárnej regresie prevyšujú jednoduchšie prístupy. Väčšina prírodných a biologických javov je totiž ovplyvňovaná viacerými faktormi a veličinami súčasne s možnou existenciou vzájomných interakcií rôznej úrovne a sily.

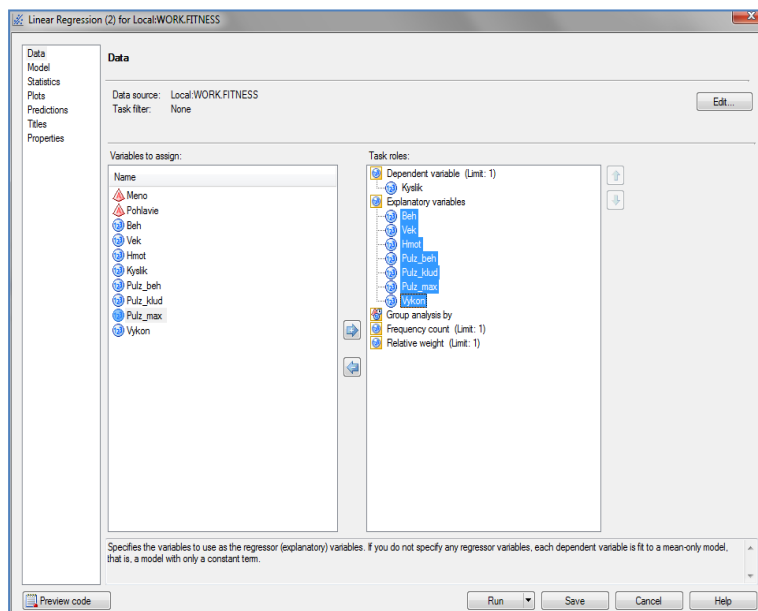
Viacnásobná lineárnej regresná analýza je veľmi silným nástrojom pri navrhovaní jednotlivých experimentov a pokusov. S jej pomocou vieme testovať možný vplyv činiteľov a efektov pri praktickom zakladaní pokusov. Na základe výsledkov tejto analýzy vieme modelovať stav a vývoj sledovaných znakov na ktoré vplyva rôznorodé množstvo nezávislých javov a hodnôt.

Základná rovnica parametrov regresnej analýzy pre viacnásobnú lineárnu regresnú rovnicu (použité sú všetky nezávisle premenné zo súboru údajov):

$$\text{kyslík} = \beta_0 + \beta_1.\text{beh} + \beta_2.\text{vek} + \beta_3.\text{hmot} + \beta_4.\text{pulz_beh} + \beta_5.\text{pulz_kl'dud} + \beta_6.\text{pulz_max} + \beta_7.\text{výkon} + e$$

Základné nastavenie parametrov regresnej analýzy pre viacnásobnú lineárnu regresnú rovnicu:

Tab. 5.11 Nastavenie viacnásobnej regresnej analýzy - príklad 1.13



Všetky ostatné nastavenia zostávajú zachované v pôvodnom stave ako to bolo pri predchádzajúcich analýzách.

Výsledky viacnásobnej lineárnej regresnej analýzy a ich interpretácia

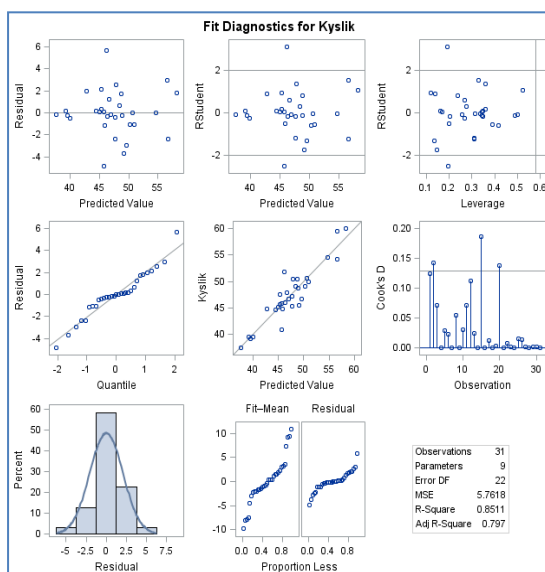
Tab. 5.12 Výsledky viacnásobnej lineárnej regresnej analýzy - príklad 1.13

Linear Regression Results					
The REG Procedure					
Model: Linear_Regression_Model					
Dependent Variable: Kyslik					
Number of Observations Read				31	
Number of Observations Used				31	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.03251	103.14750	18.32	<.0001
Error	23	129.52204	5.63139		
Corrected Total	30	851.55455			
Root MSE		2.37306	R-Square	0.8479	
Dependent Mean		47.37581	Adj R-Sq	0.8016	
Coeff Var		5.00900			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	93.33753	36.49782	2.56	0.0176
Beh	1	-2.08804	2.22856	-0.94	0.3585
Vek	1	-0.21066	0.10519	-2.00	0.0571
Hmot	1	-0.07741	0.05681	-1.36	0.1862
Pulz_beh	1	-0.36618	0.12299	-2.98	0.0067
Pulz_klud	1	-0.01389	0.07114	-0.20	0.8469
Pulz_max	1	0.30490	0.13990	2.18	0.0398
Vykon	1	0.25756	1.02373	0.25	0.8036

Analýza rozptylu ukazuje, že viacnásobný lineárny regresný model zodpovedá údajom lepšie ako predchádzajúce model. Korigovaný koeficient determinácie dosiahol hodnotu 0,8016, čo je vyššie ako rovnaký koeficient dosiahol pri predchádzajúcom modeli (0,7283), čo jednoznačne naznačuje, že tento model vysvetľuje variabilitu spotreby kyslíka lepšie ako predchádzajúci model (približne o viac ako 7 %).

Mnohé z odhadnutých parametrov pre nezávisle premenné majú príliš veľké p-hodnoty (väčšie ako 0,05), a preto sa ich odhady významne nelíšia od 0. Tieto nezávisle premenné vysvetľujú iba malú časť variability závisle premennej a preto sa môžu z analýzy odstrániť. Ich odstránením nám zostane model, ktorý sa jednoduchšie vysvetľuje a interpretuje. Takýto zjednodušený model stále ale dobre vysvetľuje a popisuje variabilitu hodnotenej závisle premennej. Grafické výstupy viacnásobnej lineárnej regresnej analýzy sú uvedené v nasledujúcej tabuľke 5.13

Tab. 5.13 Grafické zobrazenie viacnásobnej lineárnej regresnej analýzy - príklad 1.13



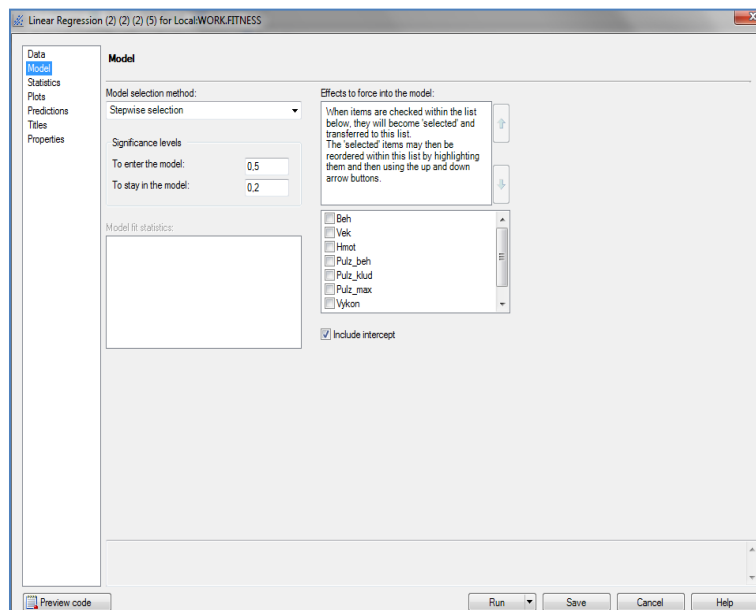
Na základe uskutočnených viacerých regresných analýz je potrebné zvoliť prístup k samotnému výberu vhodného modelu. Jeden z možných prístupov je nasledovný:

- zostaviť komplexný model so všetkými nezávisle premennými,
- najskôr odstrániť nezávisle premennú s najmenej významným odhadom regresného parametra (premenná s najväčšou p-hodnotou za predpokladu, že má p-hodnotu väčšiu ako 0,05),
- uskutočniť novú analýzu podľa upraveného modelu,
- znovu odstrániť premennú s najmenej významným odhadom regresného parametra (premenná s najväčšou p-hodnotou za predpokladu, že má p-hodnotu väčšiu ako 0,05)
- postup neustále opakujeme pokiaľ je ďalej možný

Takýto postup je príkladom metódy manuálneho spätného výberu a je vhodný hlavne pre malé súbory údajov. Pri rozsiahlych súboroch údajov môže byť ale časovo náročný. Riešenie regresnej úlohy má ale aj niekoľko metód automatického výberu optimálneho modelu, ktoré ponúka program SAS (SAS Enterprise Guide).

Ukážka jednej z metód automatického výberu optimálneho modelu pri viacnásobnej lineárnej regresnej analýze (Stepwise selection, automatická metóda postupného výberu):

Tab. 5.14 Nastavenie výberu optimálneho modelu regresnej analýzy - príklad 1.13



Automatická metóda postupného výberu bola, pri nastavení východiskových parametrov, ukončená v siedmom kroku analýzy (minimálna preukaznosť premennej v modeli 0,2).

Tab. 5.15 Výsledky výberu optimálneho modelu regresnej analýzy - príklad 1.13

Stepwise Selection: Step 7

Variable **Vykona** Removed: R-Square = 0.8469 and C(p) = 4.1469

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.20532	144.24106	27.66	<.0001
Error	25	130.34923	5.21397		
Corrected Total	30	851.55455			

Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	101.33835	11.86474	380.36418	72.95	<.0001
Beh	-2.68846	0.34202	322.17052	61.79	<.0001
Vek	-0.21217	0.09437	26.35286	5.05	0.0336
Hmot	-0.07332	0.05360	9.75445	1.87	0.1836
Pulz_beh	-0.37071	0.11770	51.71988	9.92	0.0042
Pulz_max	0.30603	0.13452	26.98596	5.18	0.0317

Bounds on condition number: 8.7438, 104.92

All variables left in the model are significant at the 0.2000 level.

No other variable met the 0.5000 significance level for entry into the model.

Finálne vybraný model viacnásobnej lineárnej regresnej analýzy spotreby kyslíka obsahuje 5 nezávislých (vysvetľujúcich) premenných: beh, vek, hmotnosť, pulz počas behu a pulz maximálny. Koeficient determinácie dosiahol hodnotu 0,8469, ktorá je podobná ako pri úplnom modeli.

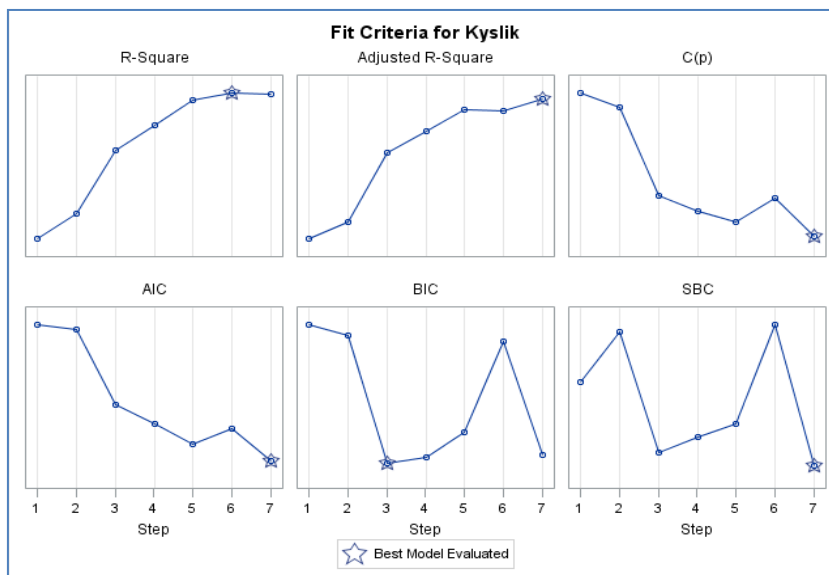
Došlo k zjednodušeniu modelu viacnásobnej lineárnej regresnej analýzy o dve nezávislé premenné. V súhrnej tabuľke 5.16 sú premenné uvedené v poradí, v akom boli vyberané, resp. odstránené, zo samotného modelu s parciálnymi a celkovými hodnotami koeficientov determinácie.

Tab. 5.16 Súhrnná tabuľka finálneho modelu regresnej analýzy - príklad 1.13

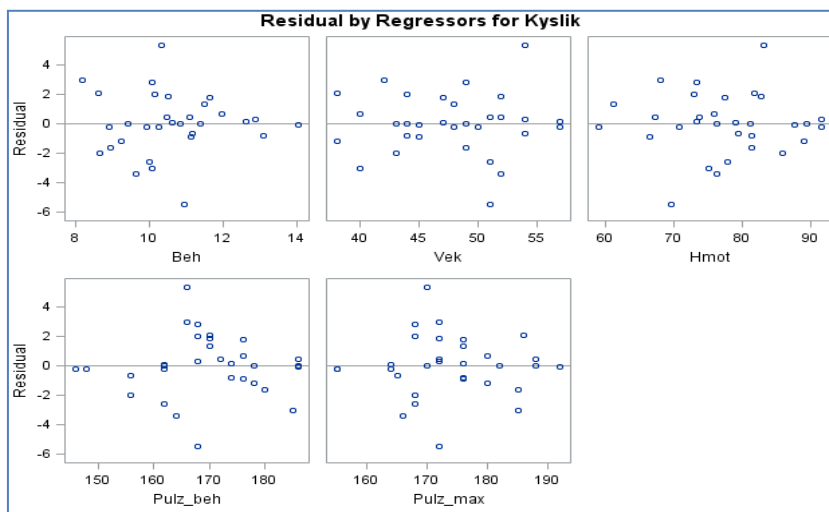
Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Vykon		1	0.7461	0.7461	11.3942	85.22	<.0001
2	Pulz_beh		2	0.0179	0.7640	10.6839	2.13	0.1559
3	Pulz_max		3	0.0432	0.8072	6.1523	6.05	0.0206
4	Vek		4	0.0181	0.8253	5.4191	2.69	0.1130
5	Hmot		5	0.0168	0.8421	4.8787	2.66	0.1155
6	Beh		6	0.0056	0.8476	6.0381	0.88	0.3587
7		Vykon	5	0.0007	0.8469	4.1469	0.11	0.7393

Na záver analýzy je ešte potrebné overiť či boli dodržané hlavné predpoklady použitia finálneho modelu (grafické znázornenia vyhovujúcich kritérií podľa jednotlivých krokov postupného výberu a zobrazenie rezíduí podľa jednotlivých nezávisle premenných).

Tab. 5.17 Grafické zobrazenie výsledkov finálneho modelu regresnej analýzy - príklad 1.13



Tab. 5.18 Grafické zobrazenie reziduí finálneho modelu regresnej analýzy - príklad 1.13



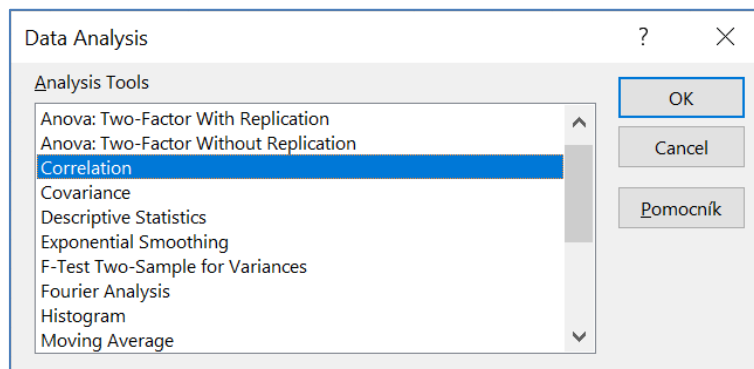
Praktické použitie programu EXCEL (Microsoft 365)

V rámci doplnkov programu Microsoft Excel je možné využiť doplnok analytických nástrojov **Data Analysis** (Analýza údajov), v ktorom nájdeme možnosť uskutočnenia korelačnej a regresnej analýzy.

Príklad 1.10 (EXCEL) - korelačná analýza

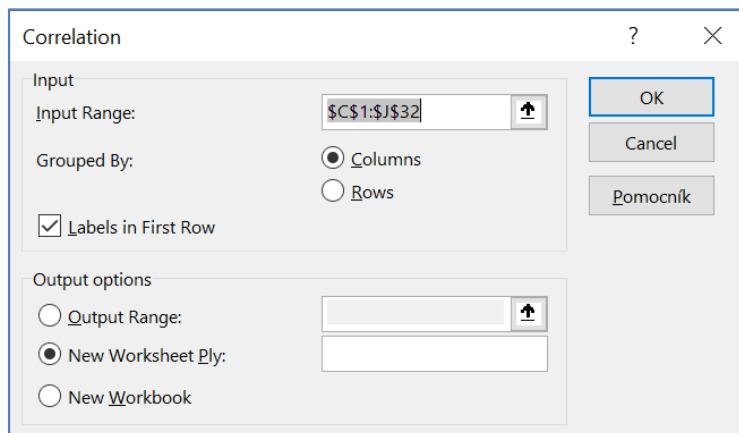
Riešime rovnaký príklad ako pri použití programu SAS Korelačnú analýzu vzťahov medzi sledovanými ukazovateľmi testovania výkonnosti členov športového klubu. Nastavenie analýzy uvádzame v tabuľke 5.19.

Tab. 5.19 Nastavenie korelačnej analýzy - príklad 1.10



Pre nastavenie korelačnej analýzy je užitočné a aj praktické pre následnú interpretáciu výsledkov nastaviť možnosť popisiek v prvom riadku (Label in First Row).

Tab. 5.19 Nastavenie korelačnej analýzy - príklad 1.10



Výsledky korelačnej analýzy (EXCEL)

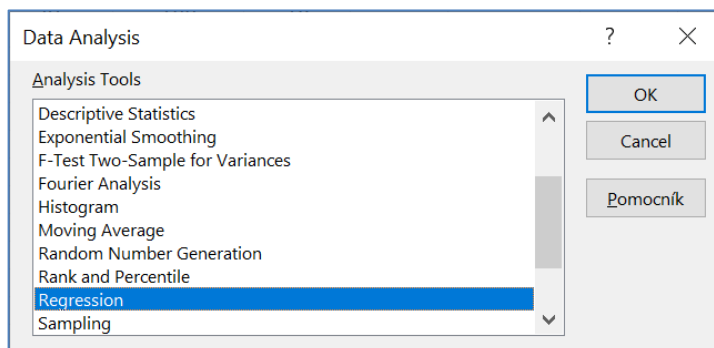
Tab. 5.20 Výsledky korelačnej analýzy - príklad 1.10

	<i>Beh</i>	<i>Vek</i>	<i>Hmot</i>	<i>Kyslik</i>	<i>Pulz_beh</i>	<i>Pulz_klud</i>	<i>Pulz_max</i>	<i>Vykon</i>
Beh	1							
Vek	0.195228	1						
Hmot	0.143508	-0.2405	1					
Kyslik	-0.86219	-0.31162	-0.16289	1				
Pulz_beh	0.313648	-0.31607	0.181516	-0.39808	1			
Pulz_klud	0.450383	-0.15087	0.043974	-0.39935	0.352461	1		
Pulz_max	0.226103	-0.4149	0.249381	-0.23677	0.929754	0.305124	1	
Vykon	-0.98841	-0.22943	-0.10544	0.863769	-0.31369	-0.47957	-0.22035	1

Hlavným výsledkom korelačnej analýzy v programe Excel sú odhadnuté Spearmanove korelačné koeficienty. Spoľahlivosť vypočítaných korelačných koeficientov sa priamo nezobrazuje, ale sa musí riešiť pomocou vloženie príslušných štatistických funkcií, ktoré vypočítajú samotnú hodnotu pravdepodobnosti p (p vyjadruje štatistickú preukaznosť rozdielu od hodnoty 0).

Príklad 1.11 (EXCEL) – jednoduchá regresná analýza

Tab. 5.21 Nastavenie korelačnej analýzy - príklad 1.11



Druhým príkladom úlohou riešenou v programe Excel je identická úloha ako v prípade použitia programu SAS Pomocou jednoduchšej regresnej analýzy odhadneme vzťah medzi závisle premennou (spotrebou kyslíka) a nezávisle premennou (výkonom).

Zadanie a nastavenie riešenej úlohy so všetkými možnými grafickými výstupmi je nasledovné:

Tab. 5.22 Podrobné nastavenie korelačnej analýzy - príklad 1.11

Výsledky regresnej analýzy (EXCEL)

Tab. 5.23 Výsledky regresnej analýzy - príklad 1.11

Regresní statistika						
Násobné R	0.8637685					
Hodnota spolehlivosti R	0.7460961					
Nastavená hodnota spolehlivosti R	0.7373408					
Chyba stř. hodnoty	2.7304986					
Pozorování	31					
ANOVA						
	Rozdíl	SS	MS	F	Významnost F	
Regrese	1	635.3415014	635.3415014	85.2164254	3.92334E-10	
Rezidua	29	216.2130535	7.455622533			
Celkem	30	851.5545548				
	Koeficienty	Chyba stř. hodnoty	t Stat	Hodnota P	Dolní 95%	Horní 95%
Hranice	35.575259	1.369165074	25.98317703	1.2285E-21	32.77500155	38.37551546
Výkon	1.4750685	0.159790344	9.23127431	3.9233E-10	1.148260551	1.801876435

Interpretácia výsledkov jednoduchšej regresnej analýzy v číselnej aj v grafickej podobe je rovnaká ako v prípade jej uskutočnenia v programe SAS.

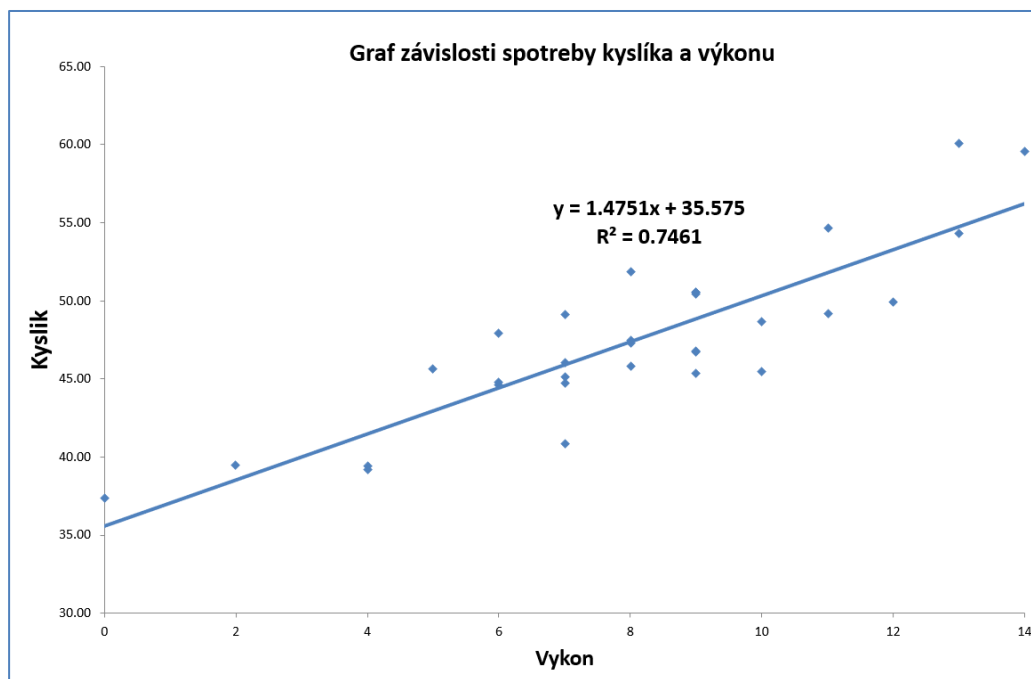
Zobrazovanými grafickými výstupmi sú: graf odhadnutej jednoduchkej regresnej priamky, graf reziduí a percentilový graf pravdepodobnosti.

Program Excel umožňuje vytvoriť ešte jeden veľmi zaujímavý výstup výsledkov regresnej analýzy pomocou zostavenia samostatného grafu závislosti. Odporúčané je nastavenie grafu ako bodového grafu so zobrazením trendovej priamky spolu s odhadnutou lineárnou regresnou rovnicou a koeficientom determinácie.

V prípade analýz nelineárnych vzťahov je možné využiť zobrazenia rôznych typov nelineárnych trendových kriviek resp. zvoliť univerzálny nelineárny prístup na základe polynómov.

Príklad zostavenia bodového grafu s nastavenou trendovou priamkou riešeného príkladu 1.11 uvádzame v tabuľke 5.24.

Tab. 5.24 Grafické výsledky regresnej analýzy - príklad 1.11



Program Excel umožňuje zostaviť a zrealizovať aj viacnásobnú lineárnu regresnú analýzu spolu s kombináciami všetkých grafických výstupov. Použitie viacnásobnej lineárnej regresnej analýzy ale predpokladá správne umiestnenie všetkých nezávisle premenných vedľa seba v jednej súvisle definovanej oblasti údajov.

Praktické použitie programu R (R Studio)

```
# Import a zobrazenie údajov (databáza projektu KEGA)
library(readxl)
Fitness <- read_excel("C:/Kega 2019-2021/Modul1/Fitness.xlsx", sheet = "b_fitness")
View(Fitness)
```

Pokiaľ používame R Studio môžeme použiť možnosť komfortného importu údajov rôzneho typu v rámci možnosti: **Import Dataset**

Príklad 1.10 (Program R) - korelačná analýza

Pomocou korelačnej analýzy vypočítajte vzťah iba pre jednu dvojicu analyzovaných ukazovateľov.

Zadanie analýzy (program R)

```
# Korelačná analýza
# Výpočet korelačných koeficientov
cor.test (Fitness$Kyslik, Fitness$Vykon, method = "pearson")
```

Výsledky korelačnej analýzy (program R)

```
# Korelačná analýza
# Výpočet korelačných koeficientov
Pearson's product-moment correlation
data: Fitness$Kyslik and Fitness$Vykon
t = 9.2313, df = 29, p-value = 3.923e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.7341185 0.9326534
sample estimates:
cor
0.8637685
```

Príklad 1.11 (Program R) - jednoduchá lineárna regresná analýza

Pomocou jednoduchej lineárnej regresie odhadnite vzťah medzi závisle premennou (spotrebou kyslíka) a nezávisle premennou (výkonom).

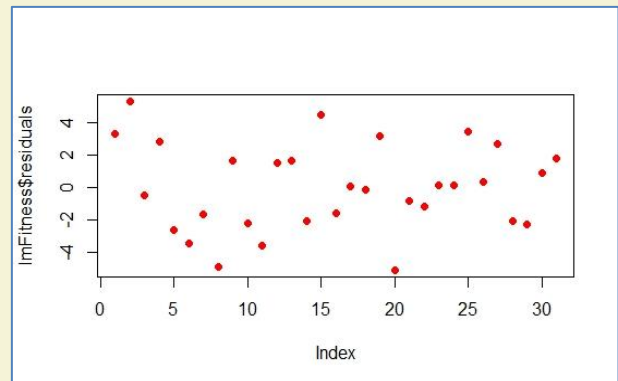
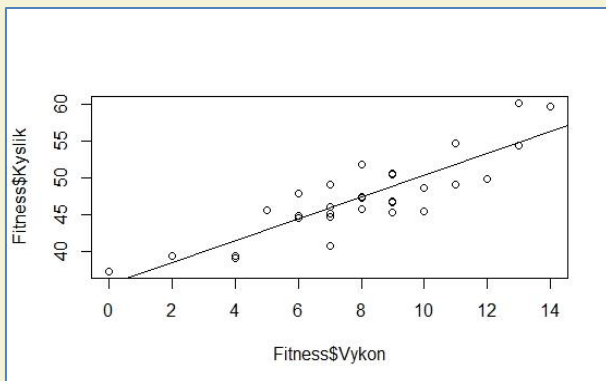
Zadanie analýzy (program R)

```
# Jednoduchá regresná analýza
# Výpočet regresných koeficientov
lmFitness = lm (Kyslik ~ Vykon, data = Fitness)
summary(lmFitness)
# Zobrazenie jednoduchého grafu závislosti analyzovaných premenných a grafu rezíduí
plot(Fitness$Vykon, Fitness$Kyslik)
abline(lmFitness)
plot(lmFitness$residuals, pch = 16, col = "red")
```

Výsledky jednoduchej regresnej analýzy (program R)

```
# Jednoduchá regresná analýza
Residuals:
  Min    1Q  Median    3Q   Max
-5.0607 -2.0732  0.0942  1.7569  5.3089

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    35.5753    1.3692  25.983 < 2e-16 ***
Vykon          1.4751    0.1598   9.231 3.92e-10 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.73 on 29 degrees of freedom
Multiple R-squared:  0.7461,    Adjusted R-squared:  0.7373
F-statistic: 85.22 on 1 and 29 DF, p-value: 3.923e-10
```

Graf závislosti a graf rezíduí

Príklad 1.12 (Program R) - viacnásobná lineárna regresná analýza (2 premenné)

Pomocou viacnásobnej lineárnej regresie odhadnite vzťah medzi závisle premennou (spotreba kyslíka) a dvoma nezávisle premennými (výkon, pulz v kľude).

Zadanie analýzy (program R)

```
# Viacnásobná regresná analýza
# Výpočet regresných koeficientov
lmFitness = lm (Kyslik ~ Vykon + Pulz_klud, data = Fitness)
summary(lmFitness)
```

Výsledky viacnásobnej regresnej analýzy (program R)

```
# Viacnásobná regresná analýza
Residuals:
    Min     1Q   Median     3Q    Max
-5.0929 -2.0433  0.0795  1.8473  5.3034
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.72590   4.96357   6.996 1.31e-07 ***
Vykon          1.49090   0.18521   8.050 9.15e-09 ***
Pulz_klud      0.01352   0.07584   0.178   0.86
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.777 on 28 degrees of freedom
Multiple R-squared:  0.7464,    Adjusted R-squared:  0.7283
F-statistic: 41.2 on 2 and 28 DF, p-value: 4.555e-09
```

Príklad 1.13 (Program R) - viacnásobná lineárna regresná analýza (viacej premenných)

Pomocou viacnásobnej lineárnej regresie odhadnite vzťah medzi závisle premennou (spotreba kyslíka) a viacerými nezávisle premennými (beh, vek, hmot, pulz_beh, pulz_kľud, pulz_max a výkon).

Zadanie analýzy (program R)

```
# Viacnásobná regresná analýza
# Výpočet regresných koeficientov
lmFitness = lm (Kyslik ~ Beh + Vek + Hmot + Pulz_beh + Pulz_klud + Pulz_max + Vykon, data = Fitness)
summary(lmFitness)
```

Výsledky viacnásobnej regresnej analýzy (program R)

```
# Viacnásobná regresná analýza
Residuals:
  Min    1Q  Median    3Q   Max
-5.4357 -0.8278 -0.0270  1.0324  5.4787

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   93.33753   36.49782   2.557  0.01761 *
Beh            -2.08804    2.22856  -0.937  0.35852
Vek            -0.21066    0.10519  -2.003  0.05713 .
Hmot           -0.07741    0.05681  -1.363  0.18623
Pulz_beh       -0.36618    0.12299  -2.977  0.00674 **
Pulz_klud      -0.01389    0.07114  -0.195  0.84694
Pulz_max       0.30490    0.13990   2.179  0.03979 *
Vykon          0.25756    1.02373   0.252  0.80359

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.373 on 23 degrees of freedom
Multiple R-squared:  0.8479,    Adjusted R-squared:  0.8016
F-statistic: 18.32 on 7 and 23 DF, p-value: 5.171e-08
```

Ďalšie príklady použitia korelačnej a regresnej analýzy**Zdroje a zoznam použitej literatúry**