

# Semestrální práce

## STP202

### 1 Jednoduchá regrese

Jako zkoumaný soubor jsem si vybral statistiku 50 nejaktivnějších členů projektu Find-a-Drug. Tento projekt se zabývá hledáním aktivních inhibitorů některých proteinů, u nichž je známá jejich role v metabolismu rakovinových buněk, viru HIV a dalších onemocnění. Jednotlivci do tohoto projektu zapojují své počítače, na které jsou pak centrálními servery rozesílány úkoly, které jsou po zpracování vráceny zpět serveru. Každý člen může k zpracování věnovat i více počítačů (někteří až stovky). V každém úkolu je zpracováno až 10100 molekul, u kterých se očekává možnost aktivní vazby na cílový protein.

Jako vysvětlovanou proměnnou jsem zvolil počet bodů, který je vypočítán z času, který počítače vynaložili na zpracování a na hodnocení rychlosti počítačů. Jako vysvětlující proměnnou jsem zvolil počet zásahů – tedy množství nalezených aktivních vazeb mezi zkoumanými molekulami a cílovým proteinem. Volím jej z toho důvodu, že v dřívější době byla tato proměnná používána k hodnocení členů namísto bodů a zajímá mě tedy její vztah k bodovému hodnocení.

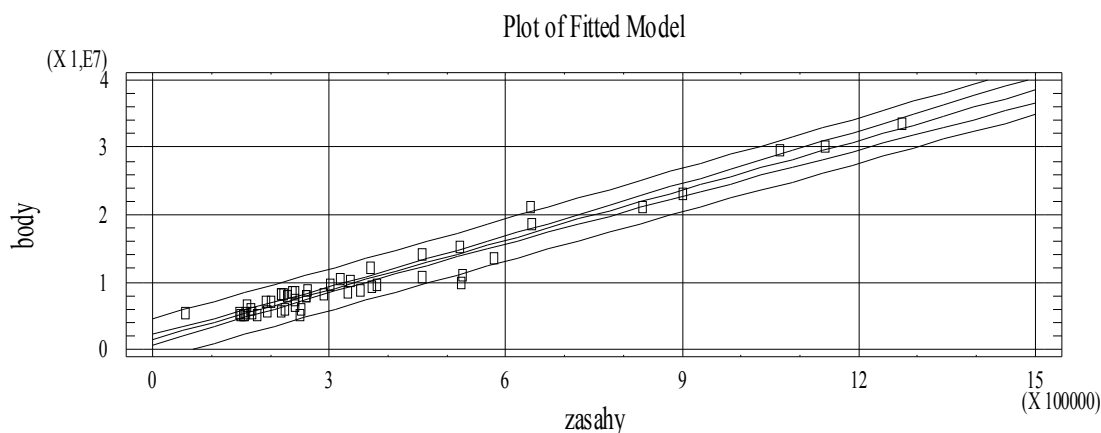
### Výběr vhodné regresní funkce

Při apriorním výběru vhodné regresní funkce bych se přiklonil k volbě přímky, jelikož při tvorbě úloh administrátory projektu je přihlíženo k tomu, aby déle trávající úlohy měly větší množství zásahů, tedy byly ohodnoceny více body. Ze zkušenosti ale vím, že tomu tak není vždy.

Při empirickém posouzení vycházím z tabulky vygenerované programem Statgraphics pro porovnání alternativních modelů.

Srovnání alternativních modelů		
Model	Correlation	R-Squared
Linear	0,9749	95,04%
Square root-Y	0,9669	93,49%
Square root-X	0,9485	89,97%
Exponential	0,9402	88,39%
Multiplicative	0,9255	85,65%
Logarithmic-X	0,8808	77,58%
Double reciprocal	0,7215	52,06%
S-curve	-0,6771	45,85%
Reciprocal-X	-0,5861	34,36%
Reciprocal-Y		<no fit>
Logistic		<no fit>
Log probit		<no fit>

Z této tabulky je zřejmé, že nejvhodnějším modelem je lineární model, jehož  $R^2$  charakteristika je přes 95% procent, znamená to tedy, že tento model se nám regresní přímkou podařilo vysvětlit na 95%. Vybraný model znázorňuje následující graf:



Rovnice výsledného modelu je:

$$\text{body} = 1,44294E6 + 24,6689 * \text{zasahy}$$

### **Komentář**

Vzhledem k faktu, že 95% variability lze vysvětlit zvoleným regresním modelem, lze konstatovat, že původní hodnocení členů nebylo příliš odlišné od dnešního. V tabulce předpokládaných hodnot si lze všimnout, že při srovnání náhodně vybraných naměřených hodnot vysvětlované proměnné a modelem předpovězených hodnot u známých hodnot vysvětlující proměnné se naměřené hodnoty nacházejí v intervalu spolehlivosti předpokládaných hodnot, tedy že regresní přímkou velmi dobře vysvětluje zvolený model (tabulka je v příloze 2). I dle grafu je většina hodnot vysvětlované proměnné v intervalu spolehlivosti pro predikovanou individuální proměnnou.

## **2Mnohonásobná regrese**

K vysvětlujícím proměnným jsem přidal další tři znaky. Jsou to počet zpracovaných úkolů, počet zpracovaných molekul a celkový čas věnovaný výpočtům v hodinách. Dá se předpokládat určitá závislost počtu molekul na počtu úkolů, vzhledem k faktu, že maximální počet molekul v jednom úkolu je 10100, zároveň ale platí, že ne všechny molekuly jsou v každém úkolu zpracované, u časově náročných úkolů je počet nezpracovaných molekul poměrně vysoký.

Souvislost mezi počtem úkolů a počtem zásahů neočekávám příliš vysokou, počet zásahů ve vztahu k různým úkolům je velmi variabilní, záleží tedy na tom, jaké je rozdělení podobných úkolů mezi jednotlivými členy.

Stejně tak je nejasný vztah mezi počtem úkolů a počtem bodů, vzhledem k jejich různé náročnosti. Vliv počtu hodin věnovaných výpočtům je těžké posoudit, pouze věnovaný čas totiž nevypovídá o rychlosti a vytíženosti zapojených počítačů, takže není jasné, nakolik ovlivní celkové hodnocení.

### **Výběr vhodné regresní funkce**

#### **Celkový F test**

Výsledek vícenásobné regrese vygenerované programem Statgraphics.

Multiple Regression Analysis					
-----					
Dependent variable: body					
-----					
Parameter	Estimate	Standard Error	T Statistic	P-Value	
-----					
CONSTANT	1,68278E6	331903,0	5,07011	0,0000	
hodiny	-207,745	131,245	-1,58288	0,1205	
molekuly	-0,209873	0,0429456	-4,88696	0,0000	
ukoly	2091,96	425,352	4,91819	0,0000	
zasahy	26,4508	1,52331	17,3641	0,0000	
-----					
Analysis of Variance					
-----					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
-----					
Model	2,19636E15	4	5,4909E14	340,56	0,0000
Residual	7,25539E13	45	1,61231E12		
-----					
Total (Corr.)	2,26892E15	49			
-----					
R-squared = 96,8023 percent					
R-squared (adjusted for d.f.) = 96,518 percent					
Standard Error of Est. = 1,26977E6					
Mean absolute error = 1,01761E6					
Durbin-Watson statistic = 2,11658					

Testová statistika F má hodnotu 340,56. Minimální hladina významnosti (p-value) má hodnotu nižší než 0,05 i než 0,001, což potvrzuje hypotézu, že alespoň jedna proměnná v modelu má opodstatnění (což ostatně vyplynulo už z jednoduché regrese s vysvětlující proměnnou počet zásahů).

### **Individuální t-testy**

Při pohledu do tabulky vícenásobné regrese vygenerované Statgraphicsem vidíme, že její p-hodnota je větší než 0,05, tedy v modelu bude pravděpodobně zbytečná. Na druhou stranu, pokud ji použijeme jako jedinou vysvětlující proměnnou, vysvětluje tento model na 70,65%.

Vzájemnou multikolaritu zjistím pomocí korelační matice.

Ve vygenerované matici je vysoký párový korelační koeficient mezi vysvětlujícími proměnnými zásahy a hodiny, dále je vysoký koeficient mezi vysvětlovanou proměnnou a zásahy. To potvrzuje vhodnost vypuštění proměnné hodiny z modelu. Vysoký párový korelační koeficient je také mezi proměnnými molekuly a úkoly, proto je vhodné prozkoumat, zda není možné jednu z nich vypustit.

Correlations					
	body	hodiny	molekuly	ukoly	zasahy
body		0,8406 ( 50)	0,7205 ( 50)	0,7218 ( 50)	0,9749 ( 50)
hodiny	0,8406 ( 50)		0,6644 ( 50)	0,6671 ( 50)	0,8607 ( 50)
molekuly	0,7205 ( 50)	0,6644 ( 50)		0,9996 ( 50)	0,7250 ( 50)
ukoly	0,7218 ( 50)	0,6671 ( 50)	0,9996 ( 50)		0,7228 ( 50)
zasahy	0,9749 ( 50)	0,8607 ( 50)	0,7250 ( 50)	0,7228 ( 50)	

### Stepwise metoda

O správnosti výběru proměnných se přesvědčím ještě stepwise metodami. Při zpětné metodě, jejíž postup je na výpisu níže, byla vypuštěna pouze proměnná hodiny, tak, jak vyplývalo i z předchozích úvah, avšak při dopředné metodě dojde pouze k výběru proměnné zásahy. To je pravděpodobně způsobeno faktem, že testová statistika F sekvenčního F-testu, má u proměnné molekuly významnou hodnotu až tehdy, je-li přidávána jako třetí proměnná, avšak testová statistika F pro druhou proměnnou není dost vysoká.

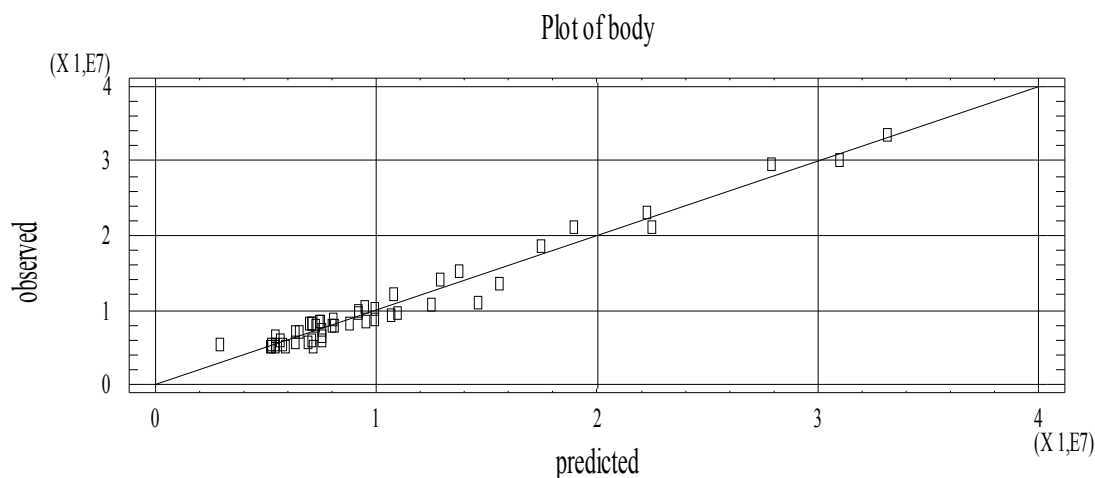
```

Stepwise regression
-----
Method: backward selection
F-to-enter: 4,0
F-to-remove: 4,0
  Step 0:
  -----
  4 variables in the model.  45 d.f. for error.
  R-squared =  96,80%      Adjusted R-squared =  96,52%      MSE = 1,61231E12
  Step 1:
  -----
  Removing variable hodiny with F-to-remove = 2,50551
  3 variables in the model.  46 d.f. for error.
  R-squared =  96,62%      Adjusted R-squared =  96,40%      MSE = 1,66508E12
  Final model selected.

```

Na výpisu zpětné metody je vidět, že proměnná hodiny byla z modelu vypuštěna v prvním kroku, přičemž testovací hodnota F sekvenčního F-testu byla 2,50551, vypuštění této proměnné se na modelu projevilo snížením  $R^2$  charakteristiky z 96,8% na 96,62%.

Graf výsledného modelu vícenásobné regrese je zobrazen níže. Tento model vysvětluje variabilitu vysvětlované proměnné na 96,62%.



Rovnice výsledného modelu je:

$$\text{body} = 1,49351\text{E}6 - 0,18798 * \text{molekuly} + 1872,22 * \text{ukoly} + 24,6109 * \text{zasahy}$$

## Závěr

Ve srovnání jsou výsledků s původní úvahou se ukázal nejasný vliv počtu hodin na celkový model, ač počet hodin sám o sobě vysvětluje variabilitu bodů na 70%, v celkovém modelu je jeho vliv spíše mizivý. Zvláštní je také vztah mezi úkoly a molekulami, které mají silný vzájemný vztah, ale přesnost celkového modelu zvyšují jen při přítomnosti obou.

Při pohledu do grafu je vidět, že pozorované hodnoty vysvětlované proměnné těsně kopírují přímku modelu, takže pozorované hodnoty odpovídají předvídaným. Z tabulky neobvyklých hodnot vygenerované Stagraphicsem vyplývá, že pouze proměnné na 5. a 12. řádku mají hodnoty, které se modelem nepodařilo vysvětlit.

Unusual Residuals				
Row	Y	Predicted		Studentized
		Y	Residual	Residual
5	2,10952E7	1,89158E7	2,17943E6	2,20
12	1,08752E7	1,45788E7	-3,70356E6	-3,20

## Příloha 1: Soubor dat

Členské číslo	Počet úkolů	Počet zásahů	Počet molekul	Počet bodů	Počet hodin
4003019	24270	1274276	240205542	33539070	13682
3003019	97402	1143258	962909495	30062290	12535
3002089	23473	1065488	232631344	29678728	7401
23683312	27926	901423	285775431	23182711	6601
4000614	68516	641774	673736463	21095233	7426
2000003	32604	832420	321901174	21013878	12563
1005702	16953	644848	168196616	18652769	4112
25483479	16799	522411	170556520	15245651	5404
1000220	11669	457593	115402370	14075332	6851
26367528	24588	580612	246079621	13571535	6661
5000650	9914	369270	97791913	12252430	4345
2000499	13896	525479	137586819	10875204	8946
27513391	13890	457367	139811692	10644203	4054
5001303	6669	319567	65776383	10522268	5848
4000740	9103	336609	89872175	10279779	3712
23702338	27012	524976	297013244	9774088	3514
4000241	9532	381227	94415926	9715915	6688
3001709	11388	302339	112019422	9677894	2081
4002407	4251	371104	42039135	9245459	3380
4003112	4510	262510	44552517	8839251	2218
5000703	15289	352119	153561257	8720761	3006
5002405	4319	241084	42685660	8574524	2452
2000376	7389	237102	73044971	8478254	6146
22586610	8469	330569	84853596	8441803	4523
5001012	8335	217404	81815911	8314568	2488
5000498	7768	290665	76656009	8186434	3084
5000995	4701	217735	46390231	8122061	3623
5001430	7325	221666	72267727	8105624	3890
3001045	5855	229134	57684220	8036009	2279
2000545	12235	261179	120823626	7924208	3585
5000367	7064	260705	69890144	7771041	3410
2000395	5877	242284	58124144	7457153	5862
4003179	4252	193261	42072921	7092311	1757
4002653	3686	201371	36507694	7070957	2331
3002773	3613	159079	35860741	6456548	2298
5000029	6258	242274	61977260	6416967	3237
2002399	3426	166445	33828644	6069068	2346
3001048	22918	253131	229288400	6009996	2402
4001639	4179	224235	41235954	5859561	2560
3002099	3641	193910	35936574	5619637	1922
4002586	3135	217229	30891072	5596935	2539
28249087	6669	148242	66017274	5490399	2591
6000414	3905	169166	38520300	5375195	2231
3002026	6651	157191	65890704	5296580	1486
2001002	3162	54483	31041451	5264762	2301
2002582	3486	156770	34418962	5230949	1945
3002049	3289	177920	32679271	5192860	1168
20168871	3809	149696	37883869	5048874	1507
21170640	4481	153559	44707114	5029010	1629
24983243	8173	249957	84088718	4998534	3337

## Příloha 2: předpovědi pro jednoduchou regresi

Předpovězené hodnoty

Predicted		95,00% Prediction Limits		95,00% Confidence Limits		Naměřené hodnoty
X	Y	Lower	Upper	Lower	Upper	
1,27428E6	3,2878E7	2,94332E7	3,63228E7	3,13301E7	3,44258E7	33539070
901423,0	2,368E7	2,04509E7	2,69092E7	2,2702E7	2,46581E7	23182711
832420,0	2,19778E7	1,87774E7	2,51782E7	2,10994E7	2,28562E7	21013878
457593,0	1,27312E7	9,61949E6	1,5843E7	1,22708E7	1,31917E7	14075332
369270,0	1,05524E7	7,44428E6	1,36605E7	1,01171E7	1,09877E7	12252430
262510,0	7,91876E6	4,80607E6	1,10315E7	7,452E6	8,38553E6	8839251
193261,0	6,21047E6	3,08959E6	9,33134E6	5,69193E6	6,729E6	7092311
159079,0	5,36723E6	2,24081E6	8,49366E6	4,81632E6	5,91815E6	6456548
193910,0	6,22648E6	3,10569E6	9,34726E6	5,70852E6	6,74443E6	5619637
156770,0	5,31027E6	2,18344E6	8,43711E6	4,75704E6	5,86351E6	5230949