Galton: Řekni mi výšku rodičů a já ti povím, kolik bude měřit dítě

V tomto cvičení navážu na slavné výzkumy Francise Galtona, který tvrdil, že z výšky rodičů lze predikovat výšku dítěte.

Teoretické zakotvení

Francis Galton, významný britský vědec 19. století, se intenzivně zabýval studiem dědičnosti a statistiky. V roce 1886 publikoval průlomovou studii, ve které zkoumal vztah mezi výškou rodičů a jejich dětí. Jeho pozorování vedla k formulaci konceptu známého jako **regrese k průměru**. Galton zjistil, že extrémní výšky rodičů (tedy velmi vysokých nebo velmi malých) mají tendenci se v následující generaci přibližovat k průměru populace. To znamená, že děti velmi vysokých rodičů bývají v průměru nižší než jejich rodiče, zatímco děti velmi malých rodičů bývají vyšší. Tento jev nazval "regression towards mediocrity" neboli regresí k průměru. Ve svém článku "Regression Towards Mediocrity in Hereditary Stature" z roku 1886 uvedl:

> *"It appears from these experiments that the offspring did not tend to resemble their parent seeds in size, but to be always more mediocre than they; to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small."*

Galtonův výzkum zahrnoval analýzu výšek 930 dospělých dětí a jejich 205 rodičovských párů. Pro zjednodušení srovnání převedl ženské výšky na ekvivalentní mužské hodnoty pomocí koeficientu 1,08. Jeho analýza ukázala, že výšky dětí mají tendenci regresovat směrem k průměru populace, což kvantifikoval jako dvě třetiny odchylky rodičů od průměru. To znamená, že pokud jsou rodiče vyšší než průměr o určitou hodnotu, jejich děti budou v průměru vyšší o dvě třetiny této hodnoty. Tento vztah vyjádřil následovně: *"The height-deviate of the offspring is, on the average, two-thirds of the height-deviate of its mid-parentage." (*Galton, 1886*).*

Galton také zdůraznil, že děti dědí své vlastnosti nejen od svých rodičů, ale i od předchozích generací. Jak se genealogie rozšiřuje do minulosti, předkové se stávají různorodějšími a jejich průměrné vlastnosti se blíží průměru populace. Tento koncept ilustroval analogií s ředěním vína vodou:

> *"The combination of the zero of the ancestry with the deviate of the mid-parentage is the combination of nothing with something, and the result resembles that of pouring a uniform proportion of pure water into a vessel of wine."* (Galton, 1892).

Galtonovy objevy položily základy moderní statistiky a genetiky. Jeho koncept regrese k průměru vedl k rozvoji regresní analýzy, klíčového nástroje v mnoha vědeckých disciplínách. Jeho práce také ovlivnila studium dědičnosti komplexních znaků, jako je výška, a ukázala, jak genetické a environmentální faktory společně ovlivňují lidské vlastnosti (Stigler, 2010).

Data a další informace o této zprávě jsou dostupné na adrese: *https://dostal.vyzkum-psychologie.cz/stat4?i=402*.

Dataset a analýza dat

Vycházím z volně dostupného datasetu, který obsahuje následující proměnné u 934 respondentů.

- Výška dítěte
- Výška matky
- Výška otce
- Pohlaví

Data jsem z původních hodnot, což byly palce, převedla na centimetry, aby to pro českého čtenáře bylo lépe představitelné. Podíváme se na základní charakteristiky souboru v následující tabulce

	Počet	Průměr dítěte	Průměr otec	Sm.odch otec	Průměr matka	Sm. odch matka
Dívky	453	162,82	175,92	6,72	162,94	5,74
Chlapci	481	175,85	175,61	5,86	162,64	5,90

Tabulka 1: Základní charakteristiky souboru

Na analýzu použiji lineární regresi, v softwaru TIBCO Statistica.

Model č. 1

Pro první model začneme jednoduchou regresí, bez jakýchkoli vyšších mocnin a interakcí. Níže uvádím tabulku regresních koeficientů.

Model vysvětluje 63,5 % variability výšky dítěte ($R^2 = 0,635$). Všechny proměnné jsou statisticky významné. Chlapci jsou v průměru o 13,25 cm vyšší než ženy. Každý 1 cm navíc u otce zvyšuje predikovanou výšku dítěte o 0,39 cm. Každý 1 cm navíc u matky přidává 0,32 cm k výšce dítěte

Proměnná	Regresní koeficient	p-hodnota
Konstanta	41,964	< 0,001
Pohlaví	13,246	< 0,001
Výška otce	0,393	< 0,001
Výška matky	0,318	< 0,001

Tabulka 2: Model č.1

Model č. 2

Do modelu nyní zahrnu výšku rodičů v druhé mocnině. Podíváme se na to, jak se kvalita modelu změní.

Model se téměř nezlepšil oproti prvnímu (R^2 = 0,635). Přidané kvadratické termíny nejsou statisticky významné. Původní lineární koeficienty pro výšku otce a matky ztratily významnost. Pohlaví zůstává jediným významným prediktorem. Výšky rodičů i jejich kvadratické členy nejsou statisticky významné. Tento model není lepší než původní lineární.

Proměnná	Regresní koeficient	p-hodnota
Konstanta	176,217	0,170
Pohlaví	13,267	< 0,001
Výška otce	-0,418	0,680
Výška matky	-0,459	0,706
Výška otce^2	0,002	0,424
Výška matky^2	0,002	0,523

Tabulka 3: Model č.2

Model č. 3

Vrátíme se tedy k modelu č.1, do kterého nyní zahrneme interakci proměnné *Pohlaví* s výškou obou rodičů.

Model se prakticky nezlepšil oproti původnímu (R^2 = 0,635). Interakce nejsou statisticky významné. Samotná proměnná *Pohlaví* ztratila významnost, což naznačuje možnou kolinearitu. Přidání interakčních členů tedy nepomohlo. Významnost pohlaví v původním modelu pravděpodobně souvisela spíš s absolutním rozdílem mezi muži a ženami než s interakcí s výškou rodičů. Výška otce a výška matky jsou stále významné prediktory výšky dítěte.

Proměnná	Regresní koeficient	p-hodnota
Konstanta	47,837	< 0,001
Pohlaví	1,217	0,930
Výška otce	0,373	< 0,001
Výška matky	0,304	< 0,001
Pohlaví × Výška otce	0,045	0,436
Pohlaví × Výška matky	0,025	0,684

Tabulka 4: Model č.3

Model č. 4

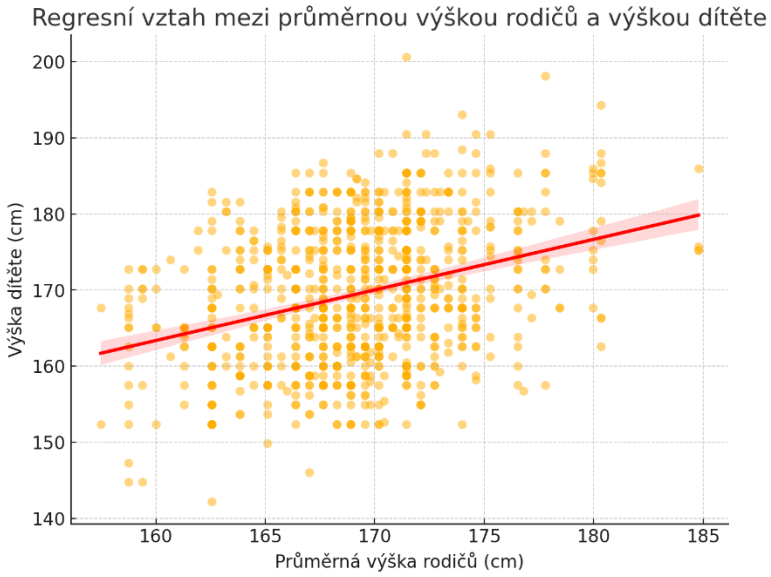
Jak bychom mohli model ještě změnit? Můžeme uvažovat dědičnost jinak. Místo výšek otce a matky zvlášť bychom mohli použít průměrnou výšku rodičů jako hlavní prediktor.

Model má prakticky stejnou přesnost jako původní model (R^2 = 0,682). Pohlaví i průměrná výška jsou statisticky významné. Muži jsou v průměru o 13,25 cm vyšší než ženy. Každý 1 cm navíc v průměrné výšce rodičů přidává 0,716 cm k výšce dítěte (silnější efekt než jednotlivé výšky rodičů v původním modelu). Použití průměrné výšky rodičů je jednodušší a má stejnou nebo lepší predikční sílu než oddělené výšky matky a otce. Menší počet proměnných má ale lepší interpretovatelnost. Umožňuje nám data lehce graficky zpracovat.

Proměnná	Koeficient	p-hodnota
Konstanta	41,520	<0,001
Pohlaví	13,250	<0,001
Průměrná výška rodičů	0,716	<0,001

Tabulka 5: Model č. 4

Níže uvádím scatterplot s regresní přímkou. Vidíme jasný pozitivní trend – čím vyšší jsou rodiče, tím vyšší je i dítě. Data se rozptylují kolem regresní přímky, což ukazuje variabilitu výšky dítěte i při stejné průměrné výšce rodičů. Některé body leží dál od přímky, což může naznačovat vliv dalších faktorů (genetika mimo rodiče, výživa, zdravotní faktory).



Graf 1: Regresní přímka

Dále vykreslíme boxplot rozdělení výšky podle pohlaví. Chlapci (1) jsou v průměru vyšší než dívky (0) – to odpovídá očekávání. Rozptyl výšky je podobný u obou pohlaví, což naznačuje, že variabilita růstu není zásadně odlišná. Některé extrémní hodnoty (outliery) – mohou představovat neobvykle vysoké nebo nízké jedince.

Graf 2: Krabicové grafy dle pohlaví

Model č. 5

A pro úplnost si ještě zkusme spočítat model, který vychází z modelu č. 4, ale navazuje na pediatrickou praxi. Vycházím z pediatrické praxe, kde se s tímto odhadem na základě průměrné výšky můžeme setkat. Tato metoda, známá jako mid-parentální výška, se vypočítává následovně:

Cílová výška chlapců = výška otce + (výška matky +13) / 2 ± 8,5 cm
Cílová výška dívek = výška matky + (výška otce –13) / 2 ± 8,5 cm

Tento výpočet poskytuje odhad střední výšky dítěte v dospělosti s 95% pravděpodobností v rozmezí ±8,5 cm od vypočtené hodnoty (*Růstový hormon*, n.d.). Model má prakticky stejnou přesnost jako původní model (R^2 = 0,634). Pohlaví i průměrná výška rodičů jsou statisticky významné. Muži jsou v průměru o 3,94 cm vyšší než ženy. Každý 1 cm navíc v průměrné výšce rodičů přidává 0,716 cm k výšce dítěte.

Proměnná	Koeficient	p-hodnota
Konstanta	46,170	<0,001
Pohlaví	3,939	<0,001
Průměrná výška rodičů	0,716	<0,001

Tabulka 6: Model č. 5

Srovnání modelu č. 4 a č. 5

Podíváme se na krátké srovnání modelu č. 4 a 5, kde byl použit podobný, ale lehce upravený výpočet.

Proměnná	Model 4	Model 5
Konstanta	41,520	46,170
Pohlaví	13,250	3,939
Průměrná výška rodičů	0,716	0,716
R^2	0,682	0,634

Tabulka 7: Srovnání modelů

Model č. 4 má vyšší R^2 (0,682) než model č. 5 (0,634), to znamená, že model č. 4 lépe vysvětluje výšku dítěte na základě zadaných proměnných. V model č. 5 má pohlaví menší efekt než v modelu č. 4. To znamená, že v modelu č. 5 pohlaví nehraje tak zásadní roli, zatímco v modelu č. 4 je rozdíl mezi chlapci a dívkami mnohem výraznější. Tento rozdíl naznačuje, že model č. 5 možná nepostihuje některé interakce nebo vlivy pohlaví dostatečně přesně. V obou modelech je koeficient průměrné výšky rodičů téměř identický. To znamená, že bez ohledu na model má výška rodičů stabilní vliv na výšku dítěte

Závěr

Dataset je uživatelsky přívětivý pro lineární regresi, nicméně čekala jsem, že najdu nějakou zajímavější závislost či problematiku. Zůstali jsme nakonec u jednoduché regrese, bez nelineárních vztahů či interakcí, což je škoda.

Data jako taková jsou označena jako cvičná, velmi by mě zajímala reálná data z populace – možná je to skvělý nápad pro další úkol.

Zdroje

- Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263. https://www.york.ac.uk/depts/maths/histstat/galton_reg.htm
- Galton, F. (1892). *Hereditary Genius: An Inquiry into Its Laws and Consequences* (2nd ed.). London: Macmillan. https://galton.org/books/hereditary-genius/text/v5/galton-1869-hereditary-genius-v5.htm
- Stigler, S. M. (2010). Darwin, Galton and the Statistical Enlightenment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3), 469–482. https://doi.org/10.1111/j.1467-985X.2010.00646.x
- Růstový hormon. (n.d.). *Co ovlivňuje růst dítěte*. Růstový hormon. https://www.rustovyhormon.cz/co-ovlivnuje-rust