Lineární modely

Lineární regrese

1. U výběru 24 jedinců ve věku 21 až 70 let, náhodně vybraných ze stejné etnické skupiny, byl po dobu 2 týdnů denně vždy v 8 hodin ráno měřen systolický krevní tlak. Nalezněte regresní model závislosti krevního tlaku Y [mm] na stáří jedince X [roky] a odpovězte na následující otázky:

   a) Jaký je průměrný věk jedinců celé etnické skupiny?

   b) Jaký je průměrný krevní tlak jedinců celé etnické skupiny?

   c) Jaký bude krevní tlak jedinců starých 65 let? Jak spolehlivostí můžeme tento krevní tlak odhadnout, pracujeme-li s 95% statistickou jistotou?

   d) O kolik mm se zvýší krevní tlak jedince každým rokem? V jakém rozmezí bude tato hodnota, odhadujeme-li tento krevní tlak s 95% statistickou jistotou?

   e) Na základě dat tohoto výběru odhadněte krevní tlak novorozence. V jakém intervalu bude tato hodnota, odhadujeme-li krevní tlak novorozence s 95% statistickou jistotou?

   f) Vypočtěte 95% oboustranný intervalový odhad predikovaného krevního tlaku pro jedince staré 60 a 55 let.

   g) Proveďte diagnostiku navrženého regresního modelu a odhalte vlivné body.

   [Datový soubor: vek_tlak.txt]

2. Cílem studie bylo nalézt závislost mezi tělesným tukem lehkých atletů-běžců Y, kteří týdně trénují asi 12 hodin, a zkonzumovaným tukem v jejich každodenní stravě X. U náhodného vzorku 18 běžců byl měřen jejich tělesný podkožní tuk Y [%] a sledován v závislosti na zkonzumovaném tuku ve stravě X [%]. Ověřte, zda lze uvedenou závislost popsat jednoduchým lineárním regresním modelem $Y = \beta_1 + \beta_2 X$.

   a) Jaký lze očekávat tělesný tuk u běžce, který spotřeboval ve stravě 25 % tuku?

   b) Jaké procento tuku ve stravě očekává běžec, který má tělesný tuk 25 %? Uveďte i rozmezí této hodnoty, a to s 95 % statistickou jistotou.

   c) Proveďte diagnostiku regresní navrženého regresního modelu a odhalte také vlivné body.

   [Datový soubor: tuk_atleti.txt]

3. Pro data představující ukazatele „přírůstku investic“ v USA v miliardách US $ Y v cenové hladině roku 1934 v závislosti na letech X byl v literatuře původně navržen kvadratický model (období 1920–1941). Regresní analýzou vyšetřete, zda by datům lépe vyhovoval polynom vyššího stupně.
   [Datový soubor: prirustek_investic_USA.txt]

4. U náhodného vzorku 20 Američanů byla provedena analýza krve a sledována denní spotřeba tuku ve stravě X v gramech a hodnota celkového cholesterolu Y v mg na 100 ml krve, str. 215 v cit70. Pro tuto závislost byl navržen jednoduchý lineární regresní model $Y = \beta_1 + \beta_2 X$.

   a) Odhadněte parametry daného modelu a ověřte platnost navrženého regresního modelu a existenci vlivných bodů.

   b) Testujte statistickou významnost obou parametrů.

   c) Sestrojte 95%ní oboustranný interval spolehlivosti pro parametr β_1 a dále vysvětlete fakt, že $\beta_1 = 0$.

d) Sestrojte 95% a 99% oboustranný interval spolehlivosti směrnice β_2.

e) Nalezněte 95 % interval spolehlivosti celkového cholesterolu u lidí, kteří denně spotřebují 50 g tuku.

f) Jaký je Pearsonův korelační koeficient mezi celkovým cholesterolem v krvi Y a denní spotřebou tuku X u sledovaných jedinců?

[Datový soubor: tuk_cholesterol.txt]

5. Data reprezentují počet nově diagnostikovaných případů rakoviny v Anglii v tisících (proměnná Y) v letech 1971–2010 (proměnná X).

a) Pro daný datový soubor odhadněte metodou nejmenších čtverců parametry modelu $Y_i = \beta_1 + \beta_2 x_i + \epsilon_i,\ i = 1, 2, \ldots, n$.

b) Odhadněte rozptyl náhodných chyb ϵ_i v tomto modelu.

c) Na hladině významnosti 0,05 ověřte hypotézu, že se jednotlivé regresní koeficienty z modelu statisticky významně liší od nuly.

d) Data proložte parabolou a stanovte intervaly spolehlivosti pro její parametry Rozhodněte, zda je pro daná data lepší model popsaný regresní přímkou nebo model popsaný regresní parabolou.

[Datový soubor: incidence_Anglie.txt]

6. U automobilu dané značky se měřila spotřeba paliva Y (v litrech na 100 km) v závislosti na jeho rychlosti X [km/hod].

a) Pro daný datový soubor odhadněte metodou nejmenších čtverců parametry modelu: $Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i,\ i = 1, 2, \ldots, n$.

b) Rozhodněte, zda kvadratický člen lze z uvedeného modelu vypustit a tedy, zda stačí data proložit přímkou.

c) Zkonstruujte 99% intervaly spolehlivosti pro parametry dané regresní funkce z bodu a).

d) Odhadněte rozptyl náhodných chyb ϵ_i v modelu a). Stanovte odhad směrodatných chyb odhadů parametrů β_i pro $i = 1, 2, 3$.

[Datový soubor: rychlost_spotreba2.txt]

7. Zemědělský ústav zkoumal závislost hektarového výnosu určité obiloviny Y v [t/ha] na množství hnojiva X, a to ledku amonného v [kg/ha]. Bylo zjištěno, že regresní model je tvořen polynomem m-tého stupně.

a) Stanovte stupeň polynomu m.

b) Věnujte zvláštní pozornost multikolinearitě a pokuste se o její snížení.

c) Jaký hektarový výnos lze očekávat při hnojení 85 kg/ha a 115 kg/ha?

[Datový soubor: hnojivo_vynos.txt.

8. Při vývoji nového benzínového motoru byla sledována závislost mezi objemem válců [dm^3] a maximálním výkonem motoru [kW]. Odhadněte parametry lineární regresní funkce $Y = \beta_1 + \beta_2 X$ vyjadřující závislost výkonu motoru na objemu válců. Jaký výkon by měl motor o objemu 4,2 litru?

[Datový soubor: objem_vykon.txt]

9. V průběhu jedné směny byla sledována kvalita výrobku záznamem řady faktorů, jedním z nich byla i hmotnost Y. Vzorky pěti výrobků byly odebírány v každou celou hodinu a jejich hmotnost Y sledována v závislosti na čase x.

   a) Zjistěte, je-li hmotnost výrobků konstantní, či zda je v ní nějaký trend.

   b) Dá se případný trend vystihnout nějakou závislostí, např. polynomickou? Stanovte stupeň polynomu m této závislosti.

   c) Pokuste se snížit vliv multikolinearity.

   d) Ověřte vhodnost zvoleného regresního modelu.

   [Datový soubor: hmotnost_smena.txt]

10. Při výrobě jistého elektronického výrobku se od pracovníků vyžaduje rychlost a vysoká přesnost. Výstupní kontrola zjistila vysoký podíl vadných výrobků. Byl proveden průzkum, jehož cílem bylo popsat průběh závislosti procenta vadných výrobků Y na výkonu za směnu X. Tabulka obsahuje údaje o výkonu za směnu X a procentu vadných výrobků Y u 20 náhodně vybraných pracovníků. Závislost popište pomocí vhodné polynomické regresní funkce. [Datový soubor: vykon_vadne_vyrobky.txt]

11. Výběr 25 pacientů, nemocných s hyperliprotenemií byl vyšetřován na hladinu lipidů v plasmě totálního cholesterolu Y [mg/100 ml] s přihlédnutím k hmotnosti X_1 [kg] a věku X_2 [roky] pacienta.

   a) Navrhněte regresní model a testujte statistickou významnost jednotlivých parametrů.

   b) Proveďte analýzu vlivných bodů.

   [Datový soubor: cholesterol_vek_hmotnost.txt]

12. U dvaceti vybraných domácností byly zjištěny údaje o čtvrtletních výdajích na potraviny a nápoje Y [Kč], čtvrtletním příjmu domácnosti X_1 [Kč], počtu dětí X_2, průměrném věku vyděličně činných členů domácnosti X_3 [roky] a počtu členů domácnosti X_4. Rozhodněte, které proměnné významně přispívají k vysvětlení variability hodnot čtvrtletních výdajů a zkonstruujte lineární regresní model s nejlepšími vysvětlujícími proměnnými. Jsou v datech odlehlé hodnoty? [Datový soubor: vydaje_rodiny.txt]

13. Na základě náhodného výběru studentů 2. ročníku technické univerzity je třeba hledat závislost mezi dosaženým studijním průměrem předmětů v 1. ročníku y a studijními výsledky na střední škole, tj. skóre maturitního testu z matematiky X_1, z jazyků X_2 a průměrem z matematických předmětů X_3 a jazyků či verbálních předmětů X_4 středoškolského studia. Postavte regresní model $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4$, popisující závislost mezi Y a X_1, X_2, X_3 a X_4, pokuste se predikovat studijní výsledky Y nových studentů, kteří právě začali studium na univerzitě, z jejich dosavadních výsledků na střední škole. Testujte statistickou významnost jednotlivých parametrů. [Datový soubor: znamky_TU.txt]

Analýza rozptylu

1. V USA jsou dva typy managementu: u prvního A1 šéfové věří, že pracující jsou v podstatě leniví a nedůvěryhodní, u druhého A2 věří svým pracovníkům, považují je za pilné a pracovité, závislé na silných individualitách. Japonci A3 však prosazují třetí směr: širokospektré plánování, konsenzus rozhodování-výroba, oboustrannou loajalitu zaměstnanec-zaměstnavatel. Cílem je porovnat hodinové mzdy [US $]

podobných inženýrských firem se třemi typy managementu (faktor A), když od každého typu bylo náhodně vybráno 6 zaměstnanců. Na hladině významnosti $\alpha = 0{,}05$ vyšetřete, zda lze považovat mzdy u tří managementů za stejné. Ověřte předpoklady použité metody.

[Datový soubor: typ_managementu.txt]

	Hodinová mzda u tří typů managementu [US $]
Management A1	5,20 5,20 6,10 6,00 5,75 5,60
Management A2	6,25 6,80 6,87 7,10 6,30 6,35
Management A3	5,50 5,75 4,60 5,36 5,85 5,90

2. Mezi místy A a B jezdí spoj A1 (jedna tramvajová linka) a spoj A2 (jedna autobusová linka), doprava je též možná spojem A3 (metro s přestupem na tramvaj). V době ranní špičky byl při cestě do práce šestkrát použit A1 ($n_1 = 6$), pětkrát A2 ($n_2 = 5$), a sedmkrát A3 ($n_3 = 7$). Naměřená doba cestování včetně čekání na příslušný spoj je v minutách. Je třeba vyšetřit, zda je doba ranního cestování mezi uvedenými místy u všech spojů (faktor A) stejná, tzn. za daných podmínek tato doba nezávisí na použitém spoji. Liší se významně výsledky spojů A1 a A3? [Datový soubor: doba_cestovani.txt]

3. Výrobce benzinu si potřebuje ověřit, seb který ze 3 typů benzinů A, B, C se dosáhne nejmenší spotřeby paliva. Na nových vozech stejné značky byl zaznamenán počet ujetých kilometrů s plnou nádrží. S každým typem benzinu jeli čtyři vozy.

   a) Popište závislost počtu ujetých kilometrů na typu benzinu pomocí podmíněných průměrů a podmíněných rozptylů.
   b) Proveďte rozklad celkové variability, vypočtěte a interpretujte poměr determinace.
   c) Pomocí F-testu pro $\alpha = 0{,}05$ ověřte, zda je opodstatněný předpoklad o závislosti počtu ujetých kilometrů na typu benzinu a výsledek prakticky interpretujte.

   [Datový soubor: benzin_km.txt]

4. Byla sledována hmotnost [g] strojových součástek vyrobených 4 dělníky, přičemž z produkce každého dělníka bylo náhodně vybráno 5 součástek.

   a) Popište závislost hmotnosti součástek mezi jednotlivými dělníky pomocí podmíněných průměrů a podmíněných rozptylů, načrtněte graf podmíněných rozptylů.
   b) Proveďte rozklad celkové variability, vypočtěte a interpretujte poměr determinace.
   c) Pomocí F-testu pro $\alpha = 0{,}05$ ověřte, zda je opodstatněný předpoklad, že hmotnost strojové součástky závisí na tom, který dělník ji vyrobil. Výsledek prakticky interpretujte.

   [Datový soubor: hmotnost_delnici.txt]

5. Byla sledována hodnota skoku do výšky [cm] u 4 různých skupin sportovců a to mezi fotbalisty, tenisty, ragbisty a basketbalisty. Předpokládejme, že chceme otestovat rozdíly ve schopnosti skákat do výšky u těchto skupin sportovců. Na hladině významnosti 0,05 vyšetřete, zda lze považovat výšku výskoku u těchto skupin sportovců za stejnou. Ověřte předpoklady použité metody.

   [Datový soubor: vyska_sport.txt]

6. Pokuste se prokázat, že směnový výkon dělníků závisí na osvětlení pracoviště A1 až A3 (faktor A), za předpokladu normality a shody rozptylů (ověřte). Výsledky jsou zaznamenány u 16 náhodně vybraných osob. Analýzu rozptylu proveďte na hladině významnosti 0,05. [Datový soubor: vykon_osvetleni.txt]

7. Test znalosti žáků z aritmetiky je do značné míry ovlivněn úrovní znalostí z předešlé školy. Na hladině významnosti 0,05 je třeba testovat, je-li vliv dosažených znalostí předešlé školy A1 až A3 (faktor A) opravdu významný. Datový soubor obsahuje skóre u zkoušky z matematiky Liší se významně vysoká A1 a nízká A3 kvalita předešlé školy? Je splněn předpoklad normality rozdělení každého výběru?
[Datový soubor: skola_aritmetika.txt]

8. Dosažené vzdělání značně ovlivňuje finanční příjem v zaměstnání. Byl proveden náhodný výběr zaměstnanců a data přinášejí souvislost mezi dosažených vzděláním či počtem let strávených na studiích A1 (8 let a méně), A2 (9–11 let), A3 (12 let), A4 (13–15 let) a A5 (16 let a více) a celoživotním příjmem v tisících US $. Na hladině významnosti 0,05 vyšetřete, zda vzdělání opravdu významně ovlivňuje příjem. Pro jaký počet let studií bylo dosaženo silně odlišného finančního příjmu?
[Datový soubor: studium_prijem.txt]

9. Byla sledována chybovost počítačového programátora. V náhodně vybraných dnech byl počítán počet chyb při sestavování programu u čtyř testovaných programátorů A1 až A4 (faktor A). Na hladině významnosti 0,05 vyšetřete, zda chybovost testovaných programátorů je shodná, či zda se liší. Lze přijmout předpoklad stejné přesnosti programátorů, vyjádřené rozptylem? Který programátor dosáhl silně odlišných výsledků od ostatních programátorů. Jsou splněny výběrové předpoklady?
[Datový soubor: programator_chyby.txt]

10. Studenti byli vyučováni předmětu za využití pěti pedagogických metod (faktor A): A1 tradiční způsob, A2 programová výuka, A3 audio technika, A4 audiovizuální technika a A5 vizuální technika. Z každé skupiny byl vybrán náhodný vzorek studentů a všichni byli podrobeni stejnému písemnému testu. Jsou znalosti všech studentů stejné, nezávislé na užité pedagogické metodě? Testujte na hladině významnosti 0,05. Je variabilita v datech dokonale popsána jediným faktorem? Lze přijmout předpoklad stejné variability metod, vyjádřené rozptylem? Vykazují výběry normální rozdělení?
[Datový soubor: test_ped_metody.txt]

11. Vyšetřete vliv věku řidiče A1 (20–39 let), A2 (40–59 let) A3 (60 a více let) a množství vypitého alkoholu B1 (žádný alkohol), B2 (1 sklenička), B3 (2 skleničky) na reakční čas řidiče v sekundách, když každé měření bylo 3x opakováno. Na hladině významnosti 0,05 vyšetřete, zda oba faktory mají významný vliv na reakční čas řidiče a zda existuje významná interakce mezi věkem řidiče a vlivem alkoholu. Vezměte první sloupec reakčního času bez alkoholu B1 za kontrolní a porovnejte s ním zbývající dva sloupce B2 a B3. Jsou zde statisticky významné rozdíly? [Datový soubor: reakcni_cas_vek_alkohol.txt]

Věk	B1 (žádný alkohol)	B2 (1 sklenička)	B3 (2 skleničky)
A1 (20–39 let)	0,42 0,43 0,41	0,47 0,46 0,46	0,65 0,66 0,68
A2 (40–59 let)	0,51 0,53 0,52	0,62 0,63 0,62	0,66 0,68 0,66
A3 (60 a více let)	0,57 0,58 0,57	0,73 0,73 0,72	0,79 0,80 0,80

12. Psycholog vyšetřuje, zda zapomínání u člověka souvisí s jeho inteligenčním kvocientem IQ a zda ovlivní výsledek psychologického testu. Na hladině významnosti 0,05 testujte, zda hladina inteligenčního kvocientu A1 (nízké IQ), A2 (střední IQ), A3 (vysoké IQ) a velikost zapomínání B1 (zapomíná zřídka), B2

(zapomíná občas), B3 (zapomíná často) mají významný vliv na výsledek psychologického testu.
[Datový soubor: psych_test_IQ_zapominani.txt]

13. Taxikářská firma se rozhoduje o větším nákupu osobních aut, vhodných pro taxi službu. Vybírá mezi pěti značkami aut, jež jsou srovnatelné co do pořizovací ceny a co do měsíční údržby. Bylo proto testováno několik vozů, dva A1 a A2 (faktor A) od každé značky. Rozhodnutí o nákupu značky B1 až B5 (faktor B) padne až podle spotřeby benzinu, tzn. počtu mil ujetých na 1 galon benzinu. Každé auto bylo testováno 3x. Vyšetřete na hladině významnosti $\alpha = 0{,}05$, zda jsou oba vozy od dané značky stejné a zda všechny typy aut jsou stejné co do počtu ujetých mil na 1 galon benzinu. Stanovte, zda je významný rozdíl mezi auty různých značek za předpokladu, že v rámci téže značky považujeme auta za stejná.
[Datový soubor: spotreba_aut.txt]

14. Zajímá nás vliv hlučnosti (faktor A) na úrovních absolutní ticho A1, hluk z ulice A2, hlasitá reprodukce hudby A3 a dále vliv osvětlení (faktor B) na úrovních přímé denní světlo B1, osvětlení stolní lampou B2 a stropní osvětlení B3 na čas, potřebný k provedení určitého příkladu elektronickou kalkulačkou. Bylo vybráno 18 pracovnic výpočetního centra a každá z nich nezávisle na ostatních řešila stejnou výpočetní úlohu. Pracovnice byly náhodně rozděleny mezi kombinace úrovní sledovaných faktorů tak, že každá kombinace byla přidělena vždy dvěma z nich. Doba v minutách k vyřešení úlohy je v datech. Na hladině významnosti $\alpha = 0{,}05$ vyšetřete, zda uvedené faktory mají významný vliv na sledovanou dobu výpočtu.
[Datový soubor: doba_vypoctu_hlucnost_osvetleni.txt]

15. Produkce určitého výrobky je ovlivněna tiskařským strojem A1 a A2 a těsnícím materiálem B1 (korkové těsnění), B2 (gumové těsnění), B3 (plastické těsnění). Na hladině významnosti $\alpha = 0{,}05$ vyšetřete, zda počet výrobků [tisíce ks] je ovlivněn druhem těsnění nebo tiskařským strojem. Existuje statisticky významná interakce obou faktorů a má logický smysl?
[Datový soubor: pocet_vyrobku_stroj_tesneni.txt]

16. Sušenky ztrácejí svou charakteristickou křupavost, když špatným skladováním zvlhnou. Vedle starého balení v tvrdém kartonu (obal B) byly testovány i nově navržené obaly, a to krabice z voskovaného papíru (obal B2), s kovovou fólií (obal B3), plastická krabice (obal B4) a kombinovaná plastická krabice s kovovou fólií (obal B5). Zboží bylo vystaveno 24hodinovému působení 80 % vlhkosti a pak byly z každé krabice náhodně odebrány 4 sušenky a stanoven obsah vody v mg, který za 24 hodin absorbovaly. Měření bylo opakováno 3x. Na hladině významnosti 0,05 zjistěte, zda záleží na druhu sušenky A1 až A4 (faktor A) a zda se funkce obalu krabic B1 až B5 (faktor B) liší v izolaci vůči vlhkosti. Existuje obal, který dosahuje silně odlišných výsledků od ostatních obalů? [Datový soubor: vlhkost_susenky_obal.txt]