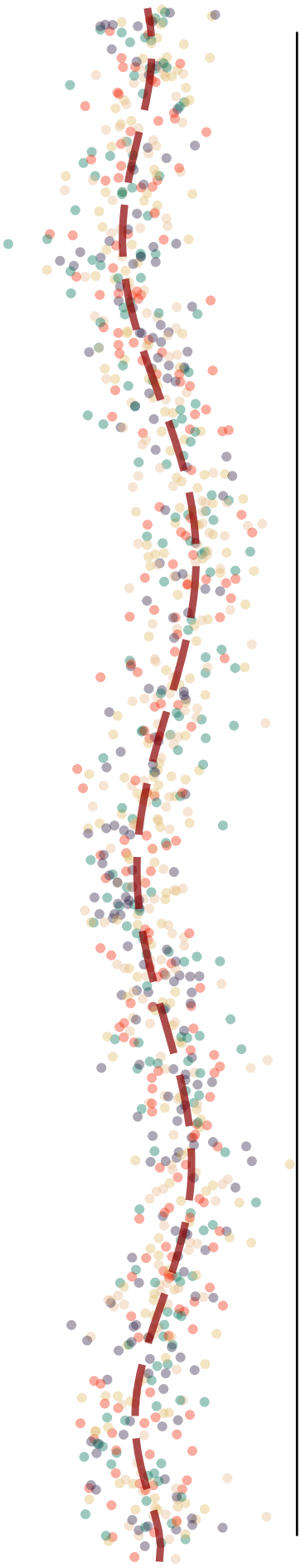


EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Univerzita Palackého v Olomouci
Filozofická fakulta

Lineární statistické modely v psychologii

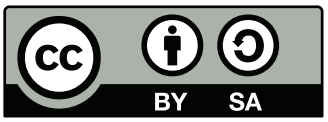
Daniel Dostál

Olomouc
2021

Tato kniha je vydána v rámci projektu Vznik doktorského studijního programu pro inovativní výzkum v psychologii práce a organizace, r.č.: CZ.02.2.69/0.0/0.0/16_018/0002313.

Odborní recenzenti:

Mgr. Vladimír Matlach, Ph.D.
Mgr. Ondřej Vencálek, Ph.D



1. vydání

DOI: 10.5507/ff.21.24458236
ISBN 978-80-244-5823-6 (online: iPDF)

Obsah

Úvod

Úvodní kurzy statistiky vybaví studenta rozmanitými nástroji k popisu dat a k testování platnosti nulových hypotéz. Ukazatele polohy, variability a dalších charakteristik, korelační koeficienty, parametrické i neparametrické bivariátní testy jsou dobrými pomocníky v nespočtu situací. Při navrhování pokročilejších experimentů, rozsáhlých dotazníkových šetření a dalších sofistikovaných designů však často pocítíme nedostatky výše uvedených postupů. Chování proměnných popisují izolovaně a statistické vztahy dokážou zachytit pouze na úrovni dvojic proměnných.

Úkolem těchto skript je představit studentům pokročilejší cestu, jak optikou kvantitativního výzkumu popsat zkoumané fenomény včetně jejich kontextu a spletitých vztahů mezi skupinami proměnných. Studenti budou uvedeni do světa statistického modelování na příkladu lineárních regresních modelů.

Různé druhy regresních modelů jsou ústředním nástrojem výzkumu v psychologii i téměř všech dalších empirických vědách. Bez znalosti lineární regrese lze jen obtížně publikovat výsledek výzkumu a zejména u studentů plánujících postgraduální působení na akademické půdě by měla být dobrá znalost tématu samozřejmostí.

Tato skripta navazují na látku probíranou v základních kurzech statistiky. Pro dobré porozumění by tedy čtenář měl být seznámen s pojmy, jako je rozdělení pravděpodobnosti, střední hodnota, rozptyl, statistické odhady, testy hypotéz a p-hodnoty. Na druhou stranu bylo cílem autora těchto skript učinit poměrně komplikovanou látku přístupnou i těm studentům, kteří zmiňované pojmy někdy slyšeli, ale jejichž hlubokou znalost si nikdy neosvojili nebo ji už dávno ztratili. V důsledku této snahy musel autor často volit mezi matematickou přesností a srozumitelností. V případě, že vás téma statistického modelování zaujme, porovnejte získané znalosti s dalšími, pokročilejšími texty, které mnohá zjednodušení a nepřesnosti doplní a zkorigují.

V následujících kapitolách budeme nejrůznější postupy demonstrovat na několika datových souborech. Většinu z nich si můžete stáhnout v souboru pro MS Excel na adrese

dostal.vyzkum-psychologie.cz/soubory/data_linearni_modely.xlsx

Poděkování patří mému učiteli a kamarádovi Ondřeji Vencálkovi za to, že si učební text přečetl a tam, kde se má tvrzení odchylovala od matematické teorie obzvláště hanebným způsobem, zakročil a přiměl mě dané pasáže přepsat.

autor

1 Statistický model

Všechny modely jsou špatně,
některé jsou však užitečné.[1]

Slavný výrok George Boxe, významného britského statistika a, mimochodem, zetě Ronalda Fishera, dobře ilustruje podstatu toho, co si pod slovem model představit. K čemu vlastně modely jsou, když žádný z nich není správný? A co vlastně slovem model označujeme?

Realita je nekonečně složitá. Je tak složitá, že se nám nikdy doopravdy nepodaří proniknout do nejhlubších zákonitostí jejího fungování. Pravda o tom, jaké zákonitosti vládnou přírodě, včetně lidského chování a prožívání, zůstane navždy lidskému poznání utajena.

Pokud realitu nedokážeme nikdy celou pochopit, jak je možné, že řada věcí, které pochází z dílny člověka, jednoduše *funguje*? Jak to, že můžeme letět na dovolenou letadlem, když se nikomu nepodařilo zjistit, přesně jakými pravidly se řídí proudění vzduchu v turbíně motoru letadla? Jak to, že nás dokáže šikovný terapeut zbavit fobie z výtahů, když nemá ponětí, co přesně se v mozku děje a co fobii způsobilo?

Abychom mohli realitu ovlivňovat a vědecké poznatky nějak převádět do praxe, nepotřebujeme znát celou pravdu se všemi detaily. To, co nám stačí, je šikovné zjednodušení reality, které méně podstatné body vynechá, zároveň však zůstane dostatečně přesné, aby svým chováním realitu připomínalo. Striktně vzato je toto zjednodušení *špatně*, protože neodpovídá skutečnosti. Na druhou stranu toto zjednodušení může být nesmírně *užitečné*. Toto zjednodušení skutečnosti budeme od tohoto okamžiku označovat slovem **model**.

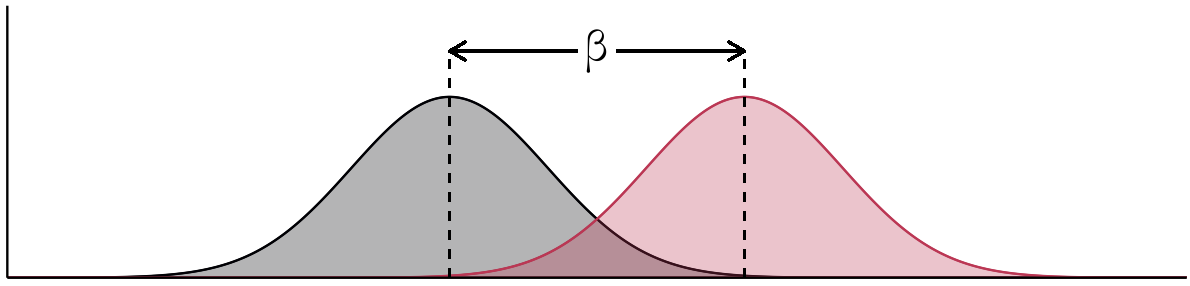
V základním kurzu statistiky jsme se naučili jeden trik, jak realitu zjednodušovat. Přestože předpokládáme, že svět se řídí kauzálními pravidly a je přísně deterministický, často se nám vyplatí řadu vlivů ignorovat a prohlásit, že svou roli zde hraje náhoda. Jevy, které označujeme jako náhodné, jsme se naučili popisovat pomocí náhodných veličin. Pokud se k tomuto přístupu přikloníme při tvorbě našeho modelu, budeme jej označovat jako **statistický model**.

Statistické modely mají řadu výhod. Zejména tu, že je velmi dobře umíme srovnávat se světem kolem nás. Kdykoli můžeme provést pozorování (měření) popisovaného jevu v realitě a to, co vidíme, srovnat s tím, co bychom měli vidět dle našeho modelu. Můžeme tak snadno zjistit, do jaké míry se náš model s realitou shoduje, nebo jí naopak odporuje.

To však není všechno. Statistický model můžeme použít k hledání odpovědí na otázky. Do našeho modelu můžeme zakomponovat nějaké (volné) **parametry**. Jelikož je statistický model popsán matematickými rovnicemi, je tímto parametrem číslo. Není to ovšem

[1] All models are wrong but some are useful.

Obrázek 1: Jednoduchý statistický model



nějaké konkrétní číslo, ale neznámá hodnota, jejíž velikost odhadneme až na základě našich pozorování.

Konkrétní statistický model, který má volné parametry, ilustrujme následujícím příkladem. Chceme prozkoumat, jestli určitý paměťový trénink má vliv na výkon v testu paměti. Oslovili jsme proto skupinu dobrovolníků, náhodně je rozlosovali do dvou skupin, experimentální a kontrolní. Experimentální skupinu jsme poté vystavili paměťovému tréninku, s kontrolní skupinou jsme si místo trénování jen povídali. Nakonec jsme obě skupiny otestovali testem paměti.

Náš model by mohl vypadat takto: Výkon členů experimentální skupiny v paměťovém testu popišme náhodnou veličinou A. Řekněme, že tato náhodná veličina má normální rozdělení. Výkon kontrolní skupiny popišme náhodou veličinou B. Řekněme, že i tato náhodná veličina má normální rozdělení, a navíc předpokládejme, že má stejný rozptyl jako náhodná veličina A. Střední hodnoty náhodných veličin A a B se od sebe liší o nějakou hodnotu, říkejme jí třeba β. Tento rozdíl β je volným parametrem modelu. Pokud je $\beta > 0$, pak lidé vystavení paměťovému tréninku skutečně skórují výš než lidé, kteří mu vystaveni nebyli. Pokud se β rovná nule nebo je menší než nula, pak toto neplatí. Náš model je znázorněn na obrázku 1.

Všimněte si, že jde skutečně jen o model, který se pochopitelně liší od reality. Náhoda přece neexistuje, ale my tu mluvíme o náhodných veličinách. Tvrdíme, že A i B má normální rozdělení, což nejspíš není tak docela pravda, i kdybychom připustili existenci nějakých náhodných veličin. Tvrdíme, že obě mají přesně stejný rozptyl, což není pravda skoro zaručeně. Zjevně tento model je špatně, zkrátka proto, že to je jen model. Na druhou stranu můžeme čekat, že v dané situaci je tento model velmi užitečný.

Ve chvíli, kdy učiníme nějaká pozorování realizací náhodné veličiny A a B, můžeme začít odhadovat velikost parametru β. Dokonce, jelikož jde o statistický model, můžeme testovat platnost hypotéz o velikosti parametru β.

S popisovaným modelem jsme se už mnohokrát sekali, jen jsme si nejspíš neuvědomili, že jde o statistický model. Vytvářeli jsme jej pokaždé, když jsme prováděli t-test pro dva nezávislé výběry.

1.1 Lineární statistické modely

Statistických modelů existuje nespočet a další bychom mohli na počkání vymyslet tak, abychom popsali jakoukoli situaci. Z této pestré plejády si vybereme jen poměrně úzkou skupinu modelů, kterým budeme říkat lineární modely. Ba co víc, z lineárních modelů si vybereme opět jen malou podmnožinu modelů, které si jsou podobné svou strukturou.

Budeme se zabývat jen takovými modely, které popisují chování jedné závislé náhodné veličiny Y, u které předpokládáme, že je ovlivňována jedním až několika faktory X_1, X_2, ..., X_k. Slovo *lineární* značí to, že budeme předpokládat, že hodnota, které náhodná veličina Y dosahuje, odpovídá nějaké **lineární kombinaci** faktorů X_1, X_2, ..., X_k. Když matematik použije termín lineární kombinace, myslí tím prostý součet, kdy však jednotlivým sčítancům dává různé váhy. My budeme tyto váhy značit β_1, β_2, ..., β_k. Každý z modelů, o nichž se budeme na následujících stránkách bavit, bude mít tvar[2]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

V následujícím textu budeme závisle proměnnou značit vždy písmenem Y a budeme o ní referovat jako o závisle proměnné, případně vysvětlované nebo predikované proměnné. Nezávisle proměnné X_1 až X_k budeme označovat slovem **faktor**, **regresor**, případně **prediktor**.

Tím, že požadujeme, aby náš model měl vždy právě tuto strukturu, jsme výrazně omezili své možnosti. Na druhou stranu, většina výzkumných problémů, se kterými se setkáváme v diplomových i disertačních pracích studentů psychologie (a koneckonců i ve většině vědeckých článků), se do této struktury bez potíží vejde. Je obvyklé, že zkoumáme nějakou veličinu Y a klademe si otázku, jak je ovlivňována nějakými faktory. Pokud například píšeme práci s titulem *Poruchy spánku u dětí předškolního věku* a rozhodneme se využít takovýto lineární model, dá se očekávat, že závisle proměnnou bude právě kvalita spánku (operacializovaná například jako skór nějakého dotazníku, který ji měří). Mezi proměnné X bychom pak zařadili třeba počet minut, které dítě večer stráví před televizí nebo hraním počítačových her, jeho pohlaví, věk, stejně tak jako informaci o tom, jestli dítě bylo zařazeno do skupiny, která před usnutím provádí nějakou relaxaci.

[2] Uvedený vztah v mnoha modifikacích uvidíme na stránkách těchto skript nesčetněkrát. Pokročilejší čtenář by proto měl být upozorněn na to, že takovýto zápis není zcela správný, a dali jsme mu přednost, abychom text učinili srozumitelnějším. Ve skutečnosti daný vztah platí pouze „v průměru“, na levé straně rovnice by proto měla být střední hodnota náhodné veličiny Y, tedy $\mathbf{E}(Y)$, nebo bychom měli místo znaménka rovná se použít symbol označující pouze přibližnou rovnost.

To, že je model lineární, nemusí být někdy na první pohled patrné. Například o modelu

$$Y = \beta_0 + \beta_1 \sqrt{Z_1 Z_2} + \beta_2 \left(\frac{4}{Z_3} + 3 \right) + \beta_3 \sin(2\pi Z_4)$$

bychom asi na první pohled neřekli, že patří do rodiny lineárních modelů. Ve skutečnosti si však můžeme představit, že $X_1 = \sqrt{Z_1 Z_2}$, $X_2 = \frac{4}{Z_3} + 3$ a $X_3 = \sin(2\pi Z_4)$, a opět se dostaneme do tvaru $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

Naopak například model

$$Y = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}$$

za lineární považovat nemůžeme, jelikož ani při nejlepší vůli jej do podoby lineární kombinace nějakých regresorů převést nedokážeme.

2 Parametry modelu a jejich odhad

Výše jsme zmínili, že modely obsahují nějaké parametry, jejichž velikost neznáme. Na příkladu modelu v literatuře nazývaném **jednoduchá regrese** si přiblížme, jakou tyto parametry mají roli a jak jejich velikost odhadujeme. Jednoduchá regrese je model, který má vedle jedné závisle proměnné Y jen jediný regresor X. Mohli bychom jej tedy popsat následující rovnicí:

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

Pro lepší představu si vezměme konkrétní příklad. Zkusíme třeba popsat, kolik bodů v zápočtovém testu z kognitivní psychologie student získá (veličina Y) v závislosti na tom, kolik hodin se na něj připravoval (X). Tedy:

$$\text{(počet bodů)} = \beta_0 + \beta_1 \cdot \text{(počet hodin přípravy)}$$

Abychom zjistili, čemu se rovnají váhy β_0 a β_1, musíme se zeptat alespoň dvou studentů, jak dlouho se učili a jak dopadli. Zeptali jsme se Agáty, která dostala z testu 42 bodů a učila se 16 hodin, a Otokara, který dostal 30 bodů a učil se 10 hodin.

$$\text{Agáta:} \qquad 42 = \beta_0 + \beta_1 \cdot 16$$

$$\text{Otokar:} \qquad 30 = \beta_0 + \beta_1 \cdot 10$$

Získaná data odpovídají tomu, co nám říká zdravý rozum – Otokar, který věnoval přípravě o šest hodin méně než Agáta, skutečně získal méně bodů. Váha β_1 tedy zřejmě bude kladné číslo, které říká, kolik bodů nám v průměru přinese každá hodina učení. Problém lze vyřešit jako soustavu rovnic o dvou neznámých (β_0 a β_1). Snadno zjistíme, že nalezená hodnota β_1 je 2 a β_0 10. Tedy

$$\text{Agáta:} \qquad 42 = 10 + 2 \cdot 16$$

$$\text{Otokar:} \qquad 30 = 10 + 2 \cdot 10$$

Náš maličký výzkum můžeme uzavřít tvrzením, že každá hodina přípravy na zápočtový test z kognitivní psychologie vede k zisku dvou bodů. Taky intuitivně chápeme roli koeficientu β_0. Model tvrdí, že pokud bychom se na učivo ani nepodívali ($X = 0$), tak získáme 10 bodů ($\beta_0 = 10$).

Kdybychom prováděli skutečný výzkum, asi bychom se nespolehli na data pouze od dvou respondentů. Zkusme do našeho souboru proto zařadit ještě Uršulu, která testu velkou pozornost nevěnovala, učila se jen dvě hodiny, ale nejspíš se jí podařilo část výsledků opsat od Agáty, nebo je mimořádně talentovaná, těžko říct. Podařilo se jí získat 24 bodů.

$$\text{Uršula:} \qquad 24 = \beta_0 + \beta_1 \cdot 2$$

Zjistíme tak nepříjemnou věc – náš model zde selhává.

$$\text{Uršula:} \qquad 24 \neq 10 + 2 \cdot 2$$

Ba co víc, neexistují takové hodnoty β_0 a β_1, které by vyhovovaly všem třem účastníkům našeho výzkumu. Řešením je rozšíření našeho původního modelu o ještě jeden člen, který nazýváme **reziduum** a značíme ϵ (epsilon)[3]. Můžeme si ho představit jako velikost chyby, které se model u daného jedince dopouští. Rovnice popisující náš model rozšířená o reziduum by tedy měla podobu:

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

Pokud budeme trvat na tom, že $\beta_0 = 10$ a $\beta_1 = 2$, pak by rezidua jednotlivých pozorování nabývala těchto hodnot:

$$\text{Agáta:} \qquad 42 = 10 + 2 \cdot 16 + 0$$

$$\text{Otokar:} \qquad 30 = 10 + 2 \cdot 10 + 0$$

$$\text{Uršula:} \qquad 24 = 10 + 2 \cdot 2 + 10$$

Toto zjevně není moc férové řešení. Náš model sice dokonale odpovídá výsledkům Agáty a Otokara, kde je chyba (reziduum) nulová, ale v případě Uršuly se mýlí o deset bodů. Mohli bychom si pochopitelně zvolit jinou dvojici čísel β_0 a β_1, a získat tak jinou sadu reziduí. Některá řešení budou lepší a jiná horší. Za nejlepší řešení považujeme takové, které produkuje rezidua co nejtěsněji se blížící svými hodnotami k nule.

Abychom mohli jednotlivá řešení srovnávat, musíme vytvořit ukazatel (takzvané minimalizační kritérium), jenž získanou sadu reziduí převede do podoby jediného čísla, které bude odrážet, jak je naše řešení dobré. Ukazuje se, že nejvýhodnějším minimalizačním kritériem není prostý součet reziduí (respektive jejich absolutních hodnot), ale součet jejich druhých mocnin. Tomuto kritériu budeme říkat **reziduální součet čtverců** (RSČ) a předpis pro jeho výpočet má tuto podobu:

$$\text{RSČ} = \sum_{i=1}^{n} \epsilon_i^2$$

Písmeno n označuje rozsah souboru pozorování. V našem případě by n bylo rovné trojce a RSČ je rovno číslu 100 (jelikož $0^2 + 0^2 + 10^2 = 100$).

[3] Upozorněme čtenáře na drobnou terminologickou nepřesnost – v tomto textu nebudeme rozlišovat mezi náhodnou složkou a reziduem, byť jde o dva příbuzné, ne však identické koncepty.

2.1 Metoda nejmenších čtverců

Hledání takových hodnot parametrů, které minimalizují RSČ metodou pokus omyl, by bylo zdlouhavé a pravděpodobně bychom přesný výsledek nikdy nezískali. Výpočet lze nicméně provést s pomocí takzvané **metody nejmenších čtverců**. Metoda nejmenších čtverců je proslulá svou elegancí a všestranným využitím a již v roce 1795 ji používal při svých výpočtech Carl Friedrich Gauss. Pro její výpočet potřebujeme znát základní postupy počítání s maticemi a vektory. Předpis pro odhad parametrů s pomocí metody nejmenších čtverců se skrývá v této stručné formuli:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

kde $\hat{\boldsymbol{\beta}}$ je vektor odhadů parametrů modelu, $\mathbf{X}$ je designová matice, která obsahuje jeden sloupec jedniček a v dalších sloupcích hodnoty regresorů, $\mathbf{Y}$ je vektor hodnot závisle proměnné. Operátory $'$ a $^{-1}$ značí transpozici a inverzi. V našem příkladu by jednotlivé prvky rovnice měly tyto hodnoty:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} 42 \\ 30 \\ 24 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 16 \\ 1 & 10 \\ 1 & 2 \end{pmatrix}$$

Všimněme si, že místo β_0 a β_1 píšeme $\hat{\beta}_0$ a $\hat{\beta}_1$. **Symbol stříšky budeme používat kdykoli, když budeme chtít vyjádřit, že jde o odhad, nikoli o přesnou hodnotu.** Kdybychom náš výpočet opakovali na jiných trojicích studentů, než je Otokar, Agáta a Uršula, dojdeme k různým odhadům $\hat{\boldsymbol{\beta}}$, které se budou pohybovat kolem skutečných nám neznámých hodnot $\boldsymbol{\beta}$. Odhady $\hat{\boldsymbol{\beta}}$ jsou náhodnými veličinami (statistikami), ba co víc, jsou nejlepšími nestrannými odhady parametrů $\boldsymbol{\beta}$, které s rostoucím počtem pozorování získávají čím dál tím vyšší přesnost.

V našem případě po dosazení do rovnice metody nejmenších čtverců získáme hodnoty $\hat{\beta}_0 = 20.27$ a $\hat{\beta}_1 = 1.26$.[4] Po dosazení najdeme tyto hodnoty reziduí:

$$\begin{aligned}
\text{Agáta:} \quad & 42 = 20.27 + 1.26 \cdot 16 + 1.62 \\
\text{Otokar:} \quad & 30 = 20.27 + 1.26 \cdot 10 - 2.84 \\
\text{Uršula:} \quad & 24 = 20.27 + 1.26 \cdot 2 + 1.22
\end{aligned}$$

Reziduální součet čtverců teď vychází 12.16 ($1.62^2 + (-2.84)^2 + 1.22^2$) a dle očekávání oproti původní hodnotě 100 značně poklesl. Jak vyplývá z tvrzení výše, neexistují žádné hodnoty $\hat{\beta}_0$ a $\hat{\beta}_1$, které by v tomto případě vedly k RSČ menšímu než 12.16.

[4] Dodejme, že v tomto textu značíme stejným symbolem $\hat{\beta}$ náhodnou veličinu (estimátor) a hodnotu její realizace (estimát). Jestli jde o číslo, nebo statistiku, čtenář dle kontextu snadno rozezná.

2.2 Nestandardizované regresní koeficienty

Parametry β_0, β_1 až β_k označujeme jako *nestandardizované regresní koeficienty*, případně jako regresní váhy. Pro pochopení výsledků lineárního modelu je klíčové dobře rozumět jejich významu.

Každý z parametrů β_1 až β_k patří jednomu regresoru (X_1 až X_k). Hodnota nestandardizovaného regresního koeficientu říká, **o kolik se v průměru změní velikost závisle proměnné *Y*, když hodnota příslušného regresoru vzroste o jedničku** (zatímco ostatní regresory zůstanou beze změny). V našem příkladu jsme došli k výsledku, že $\hat{\beta}_1 = 1.26$, což můžeme přeložit do srozumitelnější podoby jako tvrzení: *za každou hodinu příprav získá student v průměru přibližně o jeden a čtvrt bodu víc.* Pokud bychom u nějakého regresoru nalezli váhu rovnou nule, znamenalo by to, že bez ohledu na změnu hodnoty daného regresoru se průměrná hodnota závisle proměnné nemění (tzn. daný regresor nemá žádný vliv).

Parametr β_0 na rozdíl od ostatních jmenovaných k žádnému regresoru nepatří (respektive patří k onomu záhadnému sloupečku jedniček v designové matici). Tento parametr označujeme jako **absolutní člen**, **počátek**, či (z angličtiny) **intercept**. **Absolutní člen říká, jaké hodnoty bude v průměru dosahovat závisle proměnná *Y*, pokud všechny regresory budou rovny nule.** Tedy v našem příkladu, kde $\hat{\beta}_0 = 20.27$, by to znamenalo, že zcela nenaučený student, který věnoval přípravě nula hodin, dostane z testu v průměru něco přes 20 bodů.

Pochopení regresních koeficientů nám umožňuje najít odpověď na otázky, jako *kolik bodů pravděpodobně dostaneme, pokud jsme si na přípravu nechali jen noc před zápočtem, řekněme 6 hodin od deseti do čtyř ráno.* Odpověď by byla 28 bodů. Přesněji řečeno $20.27 + 1.26 \cdot 6 = 27.81$ bodů. To zní poměrně optimisticky, pokud je hranice ke splnění zápočtu třeba 25 bodů a my máme v náš model důvěru, což, jak zjistíme dále, v tomto případě není zrovna na místě. Učení proto věnujte raději o něco víc času než poslední noc před testem.

2.3 Standardizované regresní koeficienty

Za určitých okolností nemusí samotné nestandardizované regresní koeficienty čtenáři poskytnout srozumitelnou informaci. Představte si například, že v rámci marketingového výzkumu mapujete, jak jsou uživatelé spokojeni s kvalitou poskytovaného připojení k internetu. Od každého uživatele máte k dispozici údaj o tom, jak je se službou spokojen – číslo mezi jedničkou (extrémně nespokojen) a desítkou (extrémně spokojen), dále jakou částku měsíčně za službu platí, ke kolika výpadkům připojení delším než 5 minut každý měsíc dojde a nakonec jaká je rychlost připojení. Regresor rychlost připojení byl navíc

převeden na škálu dekadického logaritmu (tedy 1 Kbps byl kódován jako 0, 10 Kbps jako 1, 100 Kbps jako 2 atd.). Nalezené nestandardizované regresní koeficienty můžou vypadat třeba takto:

Regresor	$\hat{\beta}$
Absolutní člen	3.950
Cena (Kč/měsíc)	−0.002
Výpadky (počet za měsíc)	−0.140
Rychlost ($\log_{10}$ Kpbs)	0.862

Asi ani při nejlepší vůli bychom z dané tabulky nedokázali vyčíst, co zákazníkovu spokojenost ovlivní výrazně a co téměř vůbec. Dokážeme sice udělat závěry jako *„za každou zaplacenou korunu měsíčně se spokojenost sníží o 0.002 bodu“* a nebo *„s každou poruchou spokojenost zákazníka klesá o 0.14 bodu“*, ale již nemůžeme říct, jestli je to mnoho nebo málo. Nehledě na to, že interpretace váhy 0.862 regresoru *rychlost připojení* se bude do slov převádět poměrně těžko.

A právě v takových situacích používáme takzvané standardizované regresní koeficienty. Standardizované regresní koeficienty získáme přesně stejným postupem jako koeficienty nestandardizované, až na jeden rozdíl: před provedením výpočtu závisle proměnnou i každý regresor převedeme do podoby z-skóru. Převodem na z-skór se rozumí to, že od každé hodnoty daného sloupce datové matice odečteme její aritmetický průměr a každou hodnotu pak vydělíme výběrovou směrodatnou odchylkou daného sloupce. Pokud by tedy průměrná spokojenost zákazníka byla 5.5 a výběrová směrodatná odchylka spokojenosti 2.5, pak, pokud někdo odpověděl, že jeho spokojenost je 4, hodnota z-skóru jeho spokojenosti bude odpovídat číslu -0.6 (jelikož $(4 - 5.5)/2.5 = -0.6$).

Převod na z-skór má pro nás ten benefit, že všechny proměnné získají stejné měřítko. Je tedy jedno, jestli jsme například cenu měřili v korunách, či stokorunách, vždy dojdeme ke stejnému výsledku. Jednotkou všech regresorů se stane jedna směrodatná odchylka. **Standardizovaný regresní koeficient říká, o kolik směrodatných odchylek v průměru vzroste hodnota závisle proměnné Y, pokud hodnota příslušného regresoru vzroste o jednu směrodatnou odchylku.** Standardizované regresní koeficienty budeme v tomto textu označovat symbolem β^*.

Standardizované koeficienty můžeme taky z nestandardizovaných vypočítat tak, že váhu β vynásobíme směrodatnou odchylkou příslušného regresoru a vydělíme směrodatnou odchylkou závisle proměnné. Tedy

$$\hat{\beta}^* = \frac{\hat{\sigma}_X}{\hat{\sigma}_Y}\hat{\beta}$$

Následující tabulka obsahuje výsledky našeho modelu obohacené o standardizované regresní koeficienty:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
Absolutní člen	3.950	
Cena (Kč/měsíc)	−0.002	−0.172
Výpadky (počet za měsíc)	−0.140	−0.463
Rychlost ($\log_{10}$ Kpbs)	0.862	0.351

Standardizované regresní koeficienty lze použít ke srovnání velikosti efektu jednotlivých regresorů, a to dokonce i napříč různými modely. Zde můžeme říct, že největší vliv na spokojenost zákazníka má počet výpadků připojení a nejmenší roli hraje cena služby. Mohli bychom též prohlásit například, že *pokud počet výpadků připojení stoupne o jednu směrodatnou odchylku, spokojenost zákazníka v průměru klesne o bezmála půl směrodatné odchylky.*

Všimněme si, že údaj o standardizovaném regresním koeficientu u absolutního členu v tabulce schází a vynechané místo zřejmě uvidíme i u výstupu z libovolného statistického programu. Je to důsledek faktu, že β_0^* se vždy rovná 0. Tedy, pokud jsme pozorovali průměrné hodnoty všech regresorů (tzn. z-skóry všech regresorů rovné 0), pak očekáváme, že hodnota závisle proměnné převedené na z-skór bude taky průměrná (tedy 0).

Jelikož standardizované regresní koeficienty téměř vždy vychází v rozmezí $[-1; 1]$, lze interpretovat podobným způsobem jako Pearsonův korelační koeficient. Dokonce v případě, že daný regresor je dokonale nekorelovaný se všemi ostatními regresory modelu, pak standardizovaný regresní koeficient přesně odpovídá hodnotě Pearsonova korelačního koeficientu. Pokud bychom mluvili o jednoduché regresi, pak $\hat{\beta}_1^* = r_{X,Y}$ vždy, jelikož jediný regresor modelu pochopitelně s žádným jiným regresorem nekoreluje.

Standardizované regresní koeficienty jsou v psychologii poměrně oblíbené, nicméně existují obory, kde se tento druh ukazatele prakticky nepoužívá. S tím jsou spojená taky různá označení regresních koeficientů – zatímco texty z dílny statistiků značí nestandardizované a standardizované koeficienty (stejně jako tato skripta) β a β^*, v psychologických textech nejčastěji uvidíme označení b a β. Přehledně:

Regresní koeficient	Ve statistice	V psychologii
Nestandardizovaný	β	b
Standardizovaný	β^*	β

Dodejme, že standardizované regresní koeficienty nejsou užitečné vždy. Pokud se budeme rozhodovat, kterou sadu koeficientů prezentovat, pak platí jednoduché pravidlo: prezentujte ty koeficienty, které u daného modelu pomůžou čtenáři porozumět výsledku. Pokud tuto roli v daném případě plní oba typy koeficientů, pak je zřejmě nejlepší rozhodnutí prezentovat obě sady.

2.4 Grafická reprezentace jednoduché regrese

Výhodou jednoduché regrese je to, že ji můžeme znázornit graficky. Zobrazíme-li si naměřené hodnoty formou bodového grafu, kde osy x a y odpovídají náhodným veličinám X a Y, můžeme nalezený model zobrazit v podobě *regresní přímky*. Prohlédněte si obrázek 2. Jde o grafické zobrazení výsledků testu Agáty, Otokara a Uršuly a jejich dalších pěti spolužáků. Odhad parametru β_0 je roven 22.00 a odhad parametru β_1 1.46.

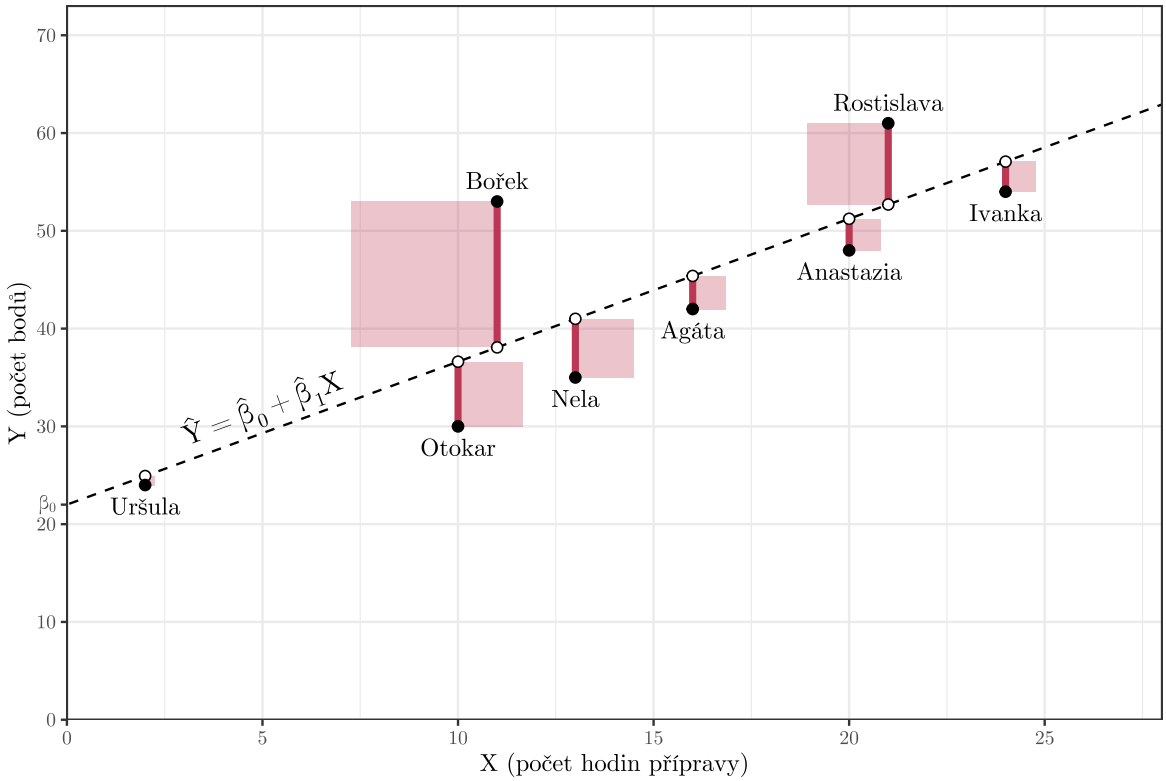
Vyplněné body znázorňují pozorované hodnoty závislé a nezávislé náhodné veličiny. Prázdné body označují očekávané hodnoty veličiny Y dle našeho modelu (značme $\hat{Y}$). Označujeme je taky jako **vyrovnané hodnoty** sledované proměnné. Očekávané hodnoty leží na regresní přímce, která je definována předpisem našeho modelu, tedy $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$. Rozdíly mezi naměřenými a očekávanými hodnotami závisle proměnné jsou nám již známá rezidua (v grafu značeny jako červené úsečky). Červené plochy pak označují druhé mocniny reziduí. Součet červených ploch je tedy RSČ, o jehož minimalizaci usilujeme.

Všimněte si také umístění β_0 – označuje místo, kde regresní přímka protíná osu y. Parametr β_1 udává sklon regresní přímky. V případě, že je kladný, bude přímka růst, pokud je záporný, bude klesat. Kdyby proměnná Y byla nezávislá na X, pak bude regresní přímka vodorovná[5].

Pro rezidua lineárních modelů platí vlastnost, se kterou jsme se již setkali u aritmetického průměru. Součet všech reziduí se rovná nule. Tedy celková velikost reziduí u pozorování nad regresní přímkou je až na znaménko stejná jako u pozorování pod regresní přímkou.

[5] Vztah mezi hodnotou parametru β_1 a úhlu svíraného regresní přímkou a osou x (značme α) lze vyjádřit jako $\beta_1 = \tan(\alpha)$.

Obrázek 2: Grafické znázornění jednoduché regrese



3 Ukazatele kvality modelu

V úvodu jsme hovořili o tom, že statistický model je co nejvěrnějším zjednodušením reality. Odhadneme-li parametry modelu, můžeme se tázat, jak věrně náš model realitu kopíruje. V předchozí kapitole jsme například pracovali s modelem, který říká, že s určitou dávkou zjednodušení je počet bodů, které student získá v testu, určen jen a pouze tím, kolik hodin se učil, a že tento vztah je lineární (tzn. za každou hodinu učení získáme β_1 bodů). Je asi odůvodněné se domnívat, že délka učení je jedním z hlavních faktorů, které ovlivní výsledek testu, nezjednodušili jsme nicméně realitu již příliš? Do jaké míry náš model odpovídá skutečnosti? Při hledání odpovědi na tuto otázku využijeme ukazatele kvality modelu.

Již jsme se seznámili s reziduálním součtem čtverců, který jakýmsi způsobem kvantifikuje, jak velké dělá model chyby, jak moc se liší od reality. Za určitých okolností by proto RSČ mohl posloužit jako míra přesnosti modelu. Obvykle jej tímto způsobem však nevyužíváme, a to hned ze dvou důvodů. RSČ je závislý na počtu pozorování (n), máme-li tedy mnoho pozorování, pak dostaneme obvykle vyšší hodnotu, než když máme pozorování jen několik. Druhou vlastností, která nám může překážet, je to, že RSČ je závislý na měřítku vysvětlované proměnné.

Užitečnějším ukazatelem je **reziduální rozptyl**. Jednoduše se jedná o výběrový rozptyl reziduí našeho modelu. Oproti obvyklému postupu výpočtu rozptylu je zde jediný rozdíl, a to, že počet stupňů volnosti není $n-1$, jak jsme byli zvyklí, ale obecně $n-p$, kde p, je počet odhadovaných parametrů[6]. Tedy

$$S_\epsilon^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{\text{RSČ}}{n-p}$$

Při jednoduché regresi bychom tedy pro výpočet použili $n-2$ stupňů volnosti, jelikož jsme museli odhadnut dva parametry (β_0, β_1). Někdy též narazíme na odmocninu z reziduálního rozptylu S_ϵ, reziduální směrodatnou odchylku (*residual standard error*).

V řadě případů je reziduální rozptyl či směrodatná odchylka užitečným ukazatelem kvality modelu. Není ale vhodný například k porovnání různých modelů z různých studií, jelikož závisí na jednotce měření. Tento problém překonává zdaleka nejpopulárnější ukazatel kvality modelu, kterým je **koeficient determinace** R^2. Koeficient determinace lze vypočítat několika způsoby. Lze jej odvodit například z RSČ jeho standardizací a odečtením od jedničky:

[6] Pokud jsme parametry svázali nějakými podmínkami (viz kapitola 13.2), pak za každou podmínku získáváme jeden stupeň volnosti zpět. Počet stupňů volnosti by tedy byl $n-p+q$, kde q je počet jedinečných podmínek.

$$R^2 = 1 - \frac{\mathrm{RS\check{C}}}{\mathrm{S\check{C}}_Y}$$

kde $\mathrm{S\check{C}}_Y$ je součtem čtverců závislé proměnné, tedy $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Jiný způsob, jak můžeme chápat koeficient determinace, je druhá mocnina Pearsonova korelačního koeficientu mezi závisle proměnnou Y a jejími vyrovnanými (tzn. odhadnutými) hodnotami $\hat{Y}$. Koeficient determinace nabývá libovolných hodnot mezi nulou a jedničkou.

Koeficient determinace je výhodný svým univerzálním použitím – můžeme jej využít ke srovnání modelů bez ohledu na měřítko či velikost souboru. Navíc jej lze velmi intuitivně interpretovat jako **procento vysvětleného rozptylu závisle proměnné**. Tedy pokud by R^2 bylo rovno 1.0, znamená to, že dokážeme bezchybně předpovědět hodnotu proměnné Y z hodnot proměnných X. R^2 rovno 0.5, znamená, že dokážeme vysvětlit 50 % variability Y. Zbývajících 50 % pak zahrnuje faktory, které jsme do našeho modelu nezařadili, spolu s chybou měření proměnné Y.

I přes své vynikající vlastnosti má R^2 jednu slabinu. Pokud do modelu přidáme další regresor, hodnota R^2 vzroste. Ani v případě, že nový regresor je zcela nesmyslný, nemůže dojít k situaci, že by procento vysvětleného rozptylu pokleslo. V nejkrajnějším případě může zůstat stejné, v praxi ale kvůli náhodnému kolísání vždy vzroste – čím máme méně pozorování, tím je náhodné kolísání větší a každý další (byť nesmyslný) regresor o to větší kousek rozptylu vysvětlí. Pokud tedy do modelu zahrneme velké množství regresorů a máme relativně malý soubor pozorování, koeficient determinace začne šplhat k nerealisticky vysokým hodnotám[7].

Tento problém se snaží překonat upravená (adjustovaná) hodnota $R^2_{adj.}$, která zohledňuje počet odhadovaných parametrů:

$$R^2_{adj.} = 1 - \frac{S^2_\epsilon}{S^2_Y} = 1 - \frac{\mathrm{RS\check{C}}}{\mathrm{S\check{C}}_Y} \cdot \frac{n-1}{n-p} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p}$$

Takto získaná hodnota nicméně nemá tak přímočarou interpretaci (v krajním případě může vyjít dokonce záporná), proto je vhodné ji prezentovat spolu s původním R^2. Při prezentaci lineárního modelu obecně vždy používáme ukazatel R^2. V případě, že by čtenář mohl mít podezření, že zde dochází k nadhodnocení R^2 kvůli vysokému počtu odhadovaných parametrů, pak je namístě prezentovat R^2 i $R^2_{adj.}$.

Ukažme si příklad výpočtu ukazatelů kvality modelu na datech od Agáty, Otokara a jejich spolužáků. Tabulka 1 obsahuje hodnoty, které byly použity k sestrojení grafu na obrázku 2.

[7] V krajním případě, kdy je počet odhadovaných parametrů stejný jako počet pozorování ($n = p$), je R^2 vždy rovno 100 % bez ohledu na to, jaké závislé či nezávislé proměnné jsme zvolili.

Tabulka 1: Hodnoty pozorovaných proměnných, predikcí a reziduí

Student	délka přípravy X	počet bodů Y	predikce $\hat{Y}$	reziduum ϵ	čtverec rezidua ϵ^2
Agáta	16	42	45.38	−3.38	11.46
Otokar	10	30	36.62	−6.62	43.76
Uršula	2	24	24.92	−0.92	0.85
Bořek	11	53	38.08	14.92	222.70
Ivanka	24	54	57.08	−3.08	9.47
Anastazia	20	48	51.23	−3.23	10.44
Nela	13	35	41.00	−6.00	36.00
Rostislava	21	61	52.69	8.31	69.02

Pomocí metody nejmenších čtverců zjistíme, že na základě našich osmi pozorování ($n = 8$) odhadneme $\hat{\beta}_0 = 22.00$ a $\hat{\beta}_1 = 1.46$. Pro každého z osmi studentů dosadíme do rovnice modelu $\hat{Y}_i = 22.00 + 1.46 \cdot X_i$ a získáme tak vyrovnané hodnoty, tedy predikce hodnot závisle proměnné. Rezidua vypočítáme jako rozdíly mezi skutečnými a vyrovnanými hodnotami závisle proměnné, tedy $\epsilon_i = Y_i - \hat{Y}_i$. RSČ pak získáme jako součet druhých mocnin reziduí, tedy

$$\text{RSČ} = 11.46 + 43.76 + 0.85 + 222.70 + 9.47 + 10.44 + 36.00 + 69.02 = 403.69$$

Jak je patrné, od oka těžko posoudíme, zdali je hodnota něco přes čtyři sta hodně nebo naopak málo. O něco informativnější je odhad reziduálního rozptylu S^2_ϵ:

$$S^2_\epsilon = \frac{\text{RSČ}}{n - p} = \frac{403.69}{8 - 2} = 67.28$$

a zejména reziduální směrodatné odchylky S_ϵ:

$$S_\epsilon = \sqrt{67.28} = 8.20$$

Směrodatná odchylka reziduí je tedy něco přes osm bodů. Toto nám již pomůže si udělat obrázek – pokud budeme predikovat, jak kdo dopadne dle našeho modelu, můžeme očekávat chyby o velikosti například 5 nebo 10 bodů, ale prakticky můžeme vyloučit, že by se u někoho model zmýlil třeba o 30 nebo 50 bodů.

Abychom mohli přesněji posoudit, jestli je oněch 8.2 bodů hodně, nebo málo, musíme prozkoumat rozmanitost závislé proměnné Y. Vypočítáme průměrný počet bodů $\bar{Y} = 43.38$ a ten využijeme při výpočtu součtu čtverců $\text{SČ}_Y = \sum_{i=1}^{8}(Y_i - 43.38)^2 = 1163.88$. Po vydělení $n - 1$ stupni volnosti získáme výběrový rozptyl proměnné Y, $S^2_Y = 166.27$, respektive po odmocnění její směrodatnou odchylku 12.89 bodů. Toto můžeme chápat

tak, že pokud bychom nevěděli, jak dlouho se který student učil, a u každého bychom očekávali průměrný počet bodů (43.38), budou mít naše predikce směrodatnou chybu 12.89. Pokud použijeme k predikci informaci o době strávené studiem, klesne směrodatná odchylka chyb na 8.20 bodů, což je vcelku uspokojivý výsledek.

Nejvýstižnějším ukazatelem i tak zůstává koeficient determinace R^2:

$$R^2 = 1 - \frac{403.69}{1163.88} = 0.65$$

Můžeme tedy říct, že jsme pomocí informací o době strávené studiem dokázali vysvětlit (popsat, predikovat) 65 % rozptylu počtu bodů, které studenti u zápočtu dostanou. Naopak chybový rozptyl tvoří zbývajících 35 %. Ten zahrnuje všechny vlivy, které jsme do modelu nezařadili (například předešlé znalosti, schopnost učit se, nepřesnost testu, opisování...).

Pokud bychom měli podezření, že je náš výsledek nadhodnocen kvůli malému počtu pozorování ($n = 8$) v poměru k počtu odhadovaných parametrů ($p = 2$), vypočítáme upravenou hodnotu koeficientu determinace $R^2_{adj.}$:

$$R^2_{adj.} = 1 - \frac{67.28}{166.27} = 0.60$$

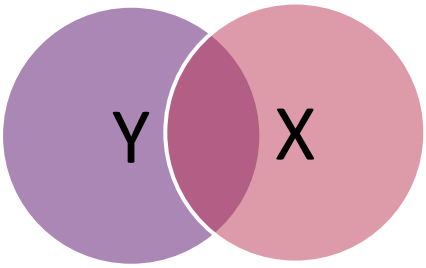
Je vidět, že k určité změně došlo a zřejmě bychom měli čtenáře na možný vliv malého souboru upozornit.

Na místě je otázka, s jakou hodnotou R^2 již můžeme být spokojeni a kterou označíme za nedostatečnou. Odpověď bohužel není jednoznačná a závisí na účelu našeho modelu. Pokud například přijdete se senzačním tvrzením, že v populaci dospělých lidí je hojně rozšířený parazit, který snižuje IQ svého nositele, pak i model, jehož R^2 dosahuje stěží 3 %, bude plnit titulní stránky novin. Naopak pokud se budete snažit doložit, že výsledek státní maturity je dobrým ukazatelem schopností studenta a dokáže předpovědět jeho akademický úspěch na VŠ, pak hodnoty R^2 pod nějakých 30 % za uspokojivé rozhodně neoznačíte. Nakonec, půjde-li o model z oblasti fyziky či některé technické disciplíny, pak modely s R^2 pod 99.9 % nebudou vůbec hodné pozornosti (dodejme, že zde pravděpodobně sáhneme po jiném ukazateli přesnosti, nejspíš po nám známé reziduální směrodatné odchylce).

4 Modely s více regresory

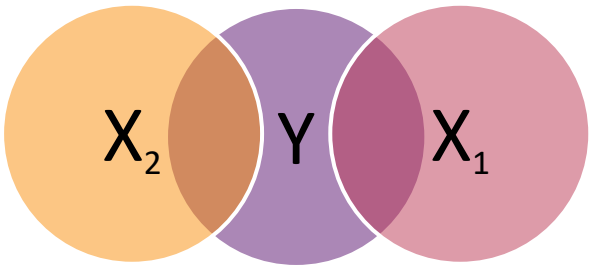
Vše, co jsme se naučili na předešlých stránkách, platí nejen pro jednoduchou regresi, ale i pro modely s více regresory. Na rozdíl od jednoduché regrese nicméně obvykle nelze složitější modely zobrazit formou regresní přímky v bodovém grafu, musíme proto pochopit vztahy mezi proměnnými jen s pomocí hodnot, které nám poskytne statistický program. Je proto užitečné si umět představit, jak vztahy mezi nezávisle proměnnými můžou vypadat a jak se to projeví na jejich regresních vahách.

Představme si rozptyl závisle proměnné Y a nezávisle proměnné X jako dva kruhy.



Obě proměnné spolu část rozptylu sdílí, kruhy se proto překrývají. Pomocí proměnné X bychom tedy mohli vysvětlit část rozptylu proměnné Y. V tomto případě je překryv 25 % plochy kruhu, platilo by tedy, že $R^2 = 0.25$. Jelikož víme, že v případě jednoduché regrese je hodnota R^2 rovna druhé mocnině Pearsonova korelačního koeficientu mezi proměnnými X a Y, respektive druhé mocnině standardizovaného regresního koeficientu β^*, můžeme také říct, že $r_{XY} = \beta^* = \sqrt{R^2} = 0.5$.

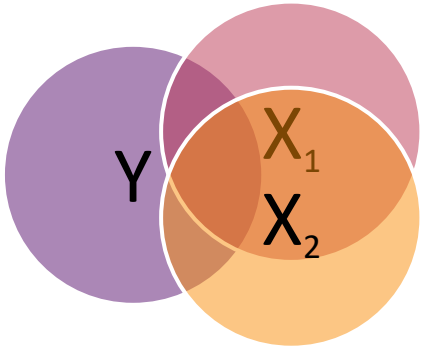
Přidejme do modelu další regresor, X_2.



Z obrázku je patrné, že i tento regresor sdílí s proměnnou Y 25 % rozptylu. Regresory X_1 a X_2 spolu nekorelují (nesdílí žádný rozptyl) a jejich kruhy se tedy nepřekrývají. I v takovémto případě snadno odhadneme hodnoty koeficientu determinace i regresních vah. Pokud regresor X_1 vysvětluje 25 % rozptylu a regresor X_2 jiných 25 % rozptylu, pak $R^2 = 0.5$, jelikož model dokáže vysvětlit celkem polovinu rozptylu závisle proměnné. Dodejme, že i zde platí shoda mezi standardizovanými regresními vahami a Pearsonovými

korelačními koeficienty mezi regresory a závisle proměnnou. V případě nekorelovaných regresorů dále platí také to, že součet druhých mocnin standardizovaných regresních koeficientů je roven koeficientu determinace: $0.5^2 + 0.5^2 = 0.25 + 0.25 = 0.5$.

Ve skutečnosti se však nejčastěji setkáme se třetím případem, kdy regresory sdílí část rozptylu nejen se závisle proměnnou, ale i mezi sebou:



Opět se X_1 překrývá s Y z jedné čtvrtiny a totéž platí pro X_2. Tentokrát jsou však oba regresory korelované a část rozptylu, kterou může vysvětlit X_1, lze stejně dobře vysvětlit s pomocí X_2. Koeficient determinace bude zjevně nižší než původních 50 %. Z obrázku je taky patrné, že přidáním více regresorů nemůže R^2 klesat, zatímco růst ano. Otázkou je, který z regresorů dostane jakou regresní váhu, tedy který regresor bude použit k popisu onoho rozptylu sdíleného všemi třemi proměnnými. Zde naši metaforu s kruhy použít nemůžeme – odpověď najde metoda nejmenších čtverců – a jen těžko bychom hledali racionální pravidlo, které by šlo popsat slovy[8].

Poselství, které bychom si měli z předešlých odstavců vzít, je to, že přidáním dalšího regresoru se obvykle změní hodnoty vah všech ostatních regresorů, a to v libovolném směru. Obvykle je tato změna žádoucí – přidání dalšího regresoru zredukuje část chybového rozptylu a poskytne nám tak čistší pohled na to, jakou mají ve skutečnosti jednotlivé nezávisle proměnné roli. Ilustrujme toto následujícím příkladem.

Ověřme hypotézu o tom, že intenzivní pocity trémy, které doprovází veřejné vystoupení člověka, souvisí s jeho sebehodnocením (*self-esteem*). Mohli bychom za tímto účelem realizovat malý výzkumný projekt, kdy necháme v rámci nějakého předmětu vystoupit 20 studentů s jejich referáty před plnou posluchárnou a pomocí dotazníku budeme měřit, jak silnou trému prožívali. Před experimentem je také necháme vyplnit několik testových

[8] Naše metafora s kruhy taky kulhá v tom, že bychom na jejím základě mohli tvrdit, že R^2 nikdy nepřesáhne součet druhých mocnin koeficientů β^*, tedy součet rozptylů, které vysvětlují jednotlivé regresory zvlášť. To ovšem není pravda. Existuje situace, kdy jednotlivé regresory jsou slabě korelované se závisle proměnnou, nicméně když jsou vloženy do modelu společně, vysvětlí nečekaně velké množství rozptylu. Jde mimochodem o situaci, kdy se můžeme setkat s koeficienty β^* překračujícími hodnotu 1. V literatuře lze tento jev najít pod heslem *supresorová proměnná* (suppressor variable).

metod včetně Rosenbergovy škály sebehodnocení (RSS) a škály Neuroticismus dotazníku NEO-FFI. Získaná data by mohla vypadat jako ta v tabulce 2.

Tabulka 2: Data: Neuroticismus, sebehodnocení a tréma

Student	Neuroticismus	Sebehodnocení	Tréma
1	14	39	2
2	22	31	6
3	16	36	2
4	19	32	8
5	24	36	5
6	24	26	6
7	25	23	5
8	26	36	4
9	20	23	4
10	21	28	5
11	30	24	7
12	18	35	5
13	33	22	6
14	24	31	6
15	23	28	5
16	14	29	1
17	30	25	5
18	28	23	10
19	16	37	2
20	21	31	2

Pokud bychom se soustředili pouze na vztah mezi sebehodnocením a prožívanou trémou, pravděpodobně nás výsledky potěší. Po provedení jednoduché regrese najdeme tyto hodnoty regresních vah:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	10.90	
Sebehodnocení	-0.21	-0.50

Hodnota standardizovaného regresního koeficientu (a tedy i Pearsonova korelačního koeficientu) se rovná pěkným -0.50. Můžeme konstatovat, že jde o středně silný až silný vztah a že výsledek Rosenbergova testu a hodnocení prožívané trémy spolu sdílí 25 % rozptylu.

V případě, že se rozhodneme prozkoumat zvlášť vztah výsledku škály neuroticismu a hodnocení trémy, zjistíme, že i zde je patrná těsná závislost:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	-1.16	
Neuroticismus	0.27	0.64

Korelační koeficient mezi oběma proměnnými je 0.64, a tedy celých 41 % rozptylu trémy lze popsat s pomocí neuroticismu. Mohli bychom se tedy domnívat, že jsme nalezli dva klíčové faktory, které souvisí s prožívanou trémou. V našich úvahách jsme ale vynechali fakt, že konstrukty sebehodnocení a neuroticismus se do velké míry překrývají. Dle našich dat je korelace obou konstruktů -0.65. Nelze tedy vyloučit to, že obě proměnné vysvětlují tentýž rozptyl proměnné tréma. Výsledky mnohonásobné regrese tuto obavu potvrdí:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	1.50	
Neuroticismus	0.23	0.55
Sebehodnocení	-0.06	-0.15

Oba regresory společně dokážou vysvětlit 42 % rozptylu, tedy jen o procento víc, než dokázal samotný neuroticismus. Váha sebehodnocení je blízko nule, zatímco váha neuroticismu zůstává vysoká. Intenzita prožívané trémy je závislá na neuroticismu jedince, zatímco sebehodnocení má jen malý dopad.

Zvolený příklad by mohl ve čtenáři vzbudit pocit, že mnohonásobná regrese je zaručený způsob, jak pokazit slibné výsledky. Opak je však pravdou – mnohonásobná regrese nám pomůže identifikovat relevantní vztahy a posílí naše přesvědčení, že pozorovaná korelace není artefakt vzniklý působením třetího faktoru. Je vysoce žádoucí do modelu zařadit regresory, jako je věk probandů a jejich pohlaví, abychom odstranili jejich nežádoucí vliv. Tyto regresory, jejichž vliv není ve středu našeho zájmu, ale do modelu je zařazujeme, aby nebyly příčinou zkreslení, označujeme jako **kovariáty**.

4.1 Nominální regresory

Doposud jsme se zabývali jen případem, kdy jsou nezávisle proměnné X kvantitativní. Poměrně často ale máme potřebu do našich úvah zahrnout i regresory měřené na nominální škále, jako je například pohlaví probanda, jeho zařazení do některé z experimentálních či kontrolních skupin nebo třeba jeho národnost. S tímto problémem se v rámci lineárních modelů můžeme snadno vyrovnat.

Pro začátek si představme nejjednodušší situaci, kdy náš model obsahuje jedinou nezávisle proměnnou X, která je dichotomická. Byl by to třeba případ, kdy zkoumáme, jaký má vliv intoxikace tetrahydrokanabinolem (THC) na reakční čas člověka. Design by mohl vypadat třeba takto: dobrovolníky náhodně rozdělíme na experimentální a kontrolní sku-

pinu. Experimentální skupině podáme určitou dávku THC a kontrolní skupině placebo. Pak pomocí počítačového testu přeměříme reakční časy obou skupin. Datová matice by mohla vypadat třeba takto:

Tabulka 3: Data: Reakční čas a THC

Proband	THC	RT [ms]
1	0	523
2	0	603
3	0	669
4	0	500
5	0	662
6	0	643
7	1	657
8	1	784
9	1	504
10	1	802
11	1	561
12	1	514

To, do které skupiny jedinec spadá, jsme zakódovali pomocí jedniček a nul, kde jednička značí experimentální (intoxikovanou) skupinu. Chceme-li vytvořit model, který by dopad experimentální podmínky popisoval, překvapivě dojdeme k přesně stejné rovnici, kterou jsme používali při jednoduché regresi:

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

Proměnná X se v tomto případě nazývá **indikátorová proměnná** (*dummy variable*) a jedničkou označuje jedince, kteří patří do experimentální skupiny. Výpočet koeficientů $\hat{\beta}$ i veškerý další postup zůstává přesně stejný jako v případě jednoduché regrese. Metoda nejmenších čtverců nám poskytne tyto odhady:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	600	
THC	37	0.19

Nejen že se nezměnil postup výpočtu, ale stejná zůstává i interpretace regresních vah. Parametr β_0 říká, jaké hodnoty v průměru závisle proměnná nabývá, pokud jsou všechny regresory rovny nule. Pokud se naše indikátorová proměnná rovná nule, znamená to, že se bavíme o probandovi z kontrolní skupiny. Tedy výsledek $\hat{\beta}_0 = 600$ znamená, že reakční čas jedinců z kontrolní skupiny je v průměru roven 600 milisekundám. Parametr β_1 obecně

říká, o kolik bodů v průměru vzroste hodnota závisle proměnné, když hodnota regresoru vzroste o jedničku. V našem případě to, že hodnota regresoru vzroste o jedničku, neznamená nic jiného, než že se jedinec z kontrolní skupiny přesune do skupiny intoxikované THC. Průměry obou skupin se tedy liší o 37 ms, z čehož snadno odvodíme, že probandi z experimentální skupiny měli reakční čas v průměru roven 637 milisekundám.

O něco obtížnější je zde interpretace koeficientu β^*. Opět jde o počet směrodatných odchylek, o které vzroste hodnota Y, když X stoupne o jednu směrodatnou odchylku. Směrodatná odchylka dichotomické proměnné ale není příliš smysluplný koncept. Koeficient β^* proto příliš rozumně interpretovat nejde; můžeme jej nicméně použít jako ukazatel míry účinku, třeba pro srovnání vlivů více proměnných X navzájem.

Situace se o něco zkomplikuje, když bude mít nominální regresor víc než dvě úrovně. Co kdybychom chtěli náš experiment rozšířit na další návykové látky a zařadili třetí skupinu, kterou jsme intoxikovali dávkou etanolu? Jako první by nás mohlo napadnout v druhém sloupečku tabulky označit tuto alkoholovou skupinu číslem 2. Takto získané výsledky by ale byly nesmyslné. Znamenalo by to, že předpokládáme existenci nějakého kontinua nic-THC-etanol, kde má etanol dvakrát vyšší hodnotu než THC. Ve skutečnosti se ale jedná o dvě kvalitativně odlišné věci, které nemůžeme stavět za sebe na jednu osu.

Řešením je zavedení jedné indikátorové proměnné pro THC a další indikátorové proměnné pro etanol. Toto řešení můžeme vidět v tabulce 4.

Tabulka 4: Data: Reakční čas, THC a ethanol

Proband	THC	ethanol	RT
1	0	0	523
2	0	0	603
3	0	0	669
4	0	0	500
5	0	0	662
6	0	0	643
7	1	0	657
8	1	0	784
9	1	0	504
10	1	0	802
11	1	0	561
12	1	0	514
13	0	1	677
14	0	1	790
15	0	1	778
16	0	1	723
17	0	1	646
18	0	1	682

Odhady parametrů opět získáme beze změny postupu pomocí metody nejmenších čtverců. Výše uvedená data nás dovedou k těmto hodnotám:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	600	
THC	37	0.19
ethanol	116	0.56

Přidání další skupiny pochopitelně nijak neovlivnilo průměr v předešlých dvou skupinách – stále platí, že v kontrolní skupině je průměrný reakční čas 600 ms a ve skupině intoxikované THC $600 + 37 = 637$ ms. Navíc však získáváme informaci o tom, že probandi intoxikovaní ethanolem jsou ve srovnání s kontrolní skupinou v průměru o 116 ms pomalejší a jejich průměrný reakční čas je tedy 716 ms.

Použití indikátorových proměnných může být někdy poněkud neintuitivní. Upozorněme proto na několik faktů:

- Ač pracujeme se třemi skupinami, máme pouze dvě indikátorové proměnné. Stejný vztah bychom nalezli i pro vyšší čísla, třeba při práci s nominální proměnnou o deseti úrovních bychom zařadili do modelu 9 indikátorových proměnných. Desátou, vynechanou, proměnnou označujeme jako **referenční skupinu** a můžeme si ji zvolit dle libosti.
- Koeficient $\hat{\beta}_0$ (absolutní člen) je průměrná hodnota proměnné Y v referenční skupině.
- Koeficienty $\hat{\beta}_1, \hat{\beta}_2, ...$ ukazují, o kolik se liší jednotlivé skupiny od referenční skupiny. Pokud chceme srovnat jednotlivé skupiny mezi sebou, můžeme jako referenční skupinu určit jinou úroveň faktoru.
- Pokud zvolíme jinou referenční skupinu, celková přesnost modelu (RSČ, $R^2, ...$) se nezmění, ale pochopitelně se změní regresní koeficienty jednotlivých skupin. Pokud například v našem příkladu zvolíme za referenční skupinu probandy intoxikované THC, získáme výsledky, které nás po bližším zamyšlení dovedou ke stejným skupinovým průměrům:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	637	
ethanol	79	0.38
kontrolní	-37	-0.18

Zařadíme-li do modelu sadu indikátorových proměnných, pak již obvykle nemluvíme o regresi, ale o **všeobecném lineárním modelu** (*general linear model*). Kategoriálních regresorů můžeme do modelu pochopitelně zařadit více než jeden a lze je také volně kombinovat s regresory spojitými.

Je zjevné, že indikátorové kódování je praktické zejména tehdy, když máme jednu kontrolní skupinu, se kterou srovnáváme několik skupin experimentálních, nebo v případě, kdy pracujeme s jedinou dichotomickou proměnnou. Pokud máme srovnat několik skupin, z nichž žádná nemá nějakým způsobem výsadní postavení, aby mohla být považována za referenční skupinu, tato metoda není příliš obratná.

Doplňme také, že existuje celá řada dalších způsobů kódování nominální proměnné, které vedou k jiným interpretacím regresních parametrů. Obecně ale nijak nemění celkovou přesnost modelu. Podrobně se jim věnujeme v kapitole 14.

4.2 Interakce regresorů

Někdy je vztah nezávisle proměnných X k závisle proměnné Y poněkud komplikovanější, než aby šel popsat jako vážený součet. Triviálním příkladem je slazení kávy. Pokud kávu zamícháte, nebude z ničeho nic sladká. Pokud do ní nasypete cukr, její chuť se taky nezmění, protože cukr sedne na dno šálku a nerozpustí se. Řekli bychom tedy, že ani jeden úkon nemá na chuť kávy vliv. Pokud ovšem provedeme oba dva úkony – do kávy nasypeme cukr a zamícháme ji – získáme nesrovnatelně větší účinek. Interakce může fungovat i naopak. Když zůstaneme u prostoduchých kulinářských příkladů, můžeme modelovat, jak je chutná palačinka v závislosti na tom, jaké ingredience na ni dáme. Třeba marmeláda nebo Nutella chutnost palačinky rozhodně zvýší. Stejně tak bude mít pozitivní účinek eidam nebo šunka. Pokud však přidáte všechny tyto přísady současně, efekty se nesečtou, ale naopak bude chutnost výsledného produktu hluboko v záporných číslech.

Do interakce můžou vstupovat jak metrické, tak nominální proměnné. Prozatím odhlédneme od nominálních proměnných s více než dvěma úrovněmi, kde je situace o něco komplikovanější. Ve všech ostatních případech můžeme do modelu zahrnout vliv interakce mezi proměnnou X_1 a X_2 přidáním **interakčního členu** do regresní rovnice. Model s k regresory, kde první dva z nich vstupují do interakce, bychom popsali následující rovnicí:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \cdots + \beta_k X_k$$

Proměnná $X_1 X_2$ je skutečně vynásobením hodnoty regresoru X_1 hodnotou regresoru X_2. Koeficient β_{12} je pak váhou tohoto nově vzniklého regresoru. Samotné zařazení interakce do našeho modelu je tedy po technické stránce velmi snadné. O to větší výzvou je interakční člen správně interpretovat. Demonstrujme práci s interakcí na následující úloze.

Zlé jazyky o nejmenovaném profesorovi tvrdí, že při zkoušení není ani v nejmenším objektivní, ale známky rozdává na základě jediného kritéria, a tím je délka sukně. Studenti se proto rozhodli ověřit tento předpoklad pomocí lineárního modelu. Každému, kdo šel ke zkoušce, jednoduše přeměřili délku sukně případně nohavic a tuto hodnotu v centimetrech zapsali do tabulky. Dále si také poznačili, jestli je dotyčný žena nebo muž, a jakou známku dostal. Tabulka 5 obsahuje záznamy studentů na konci zkouškového období. Pro snazší matematické zpracování byly známky převedeny na čísla od nejlepší známky A označené 1 po nejhorší F nesoucí hodnotu 6.

Úvaha, že hodnocení může být ovlivněno dvěma faktory (pohlavím studenta a délkou sukně či nohavic), se zdá být na první pohled správná. Ověřme proto předpoklad studentů pomocí modelu s těmito dvěma regresory

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot \text{pohlaví} \ + \ \beta_2 \cdot \text{délka}$$

který nás dovede k následujícímu výsledku:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	3.001	
pohlaví	0.045	0.015
délka	0.006	0.092
R^2	1 %	

Nalezené váhy obou regresorů dokážeme snadno interpretovat. Muži, značení jedničkou, dostávají o 0.045 stupně (tedy asi jednu dvaadvacetinu) vyšší (tzn. horší) známku než referenční skupina žen. Za každý centimetr délky sukně se známka zvyšuje (zhoršuje) o 0.006 stupně (tedy asi jednu stosedmasedmdesátinu). Na první pohled je zřejmé, že se bavíme o prakticky nulových efektech. Regresory vysvětlí přibližně 1 % rozptylu sledované proměnné.

Skutečně šlo o křivé nařčení a dlužíme obviněnému profesorovi omluvu? Dříve než přijmeme tento závěr, zamysleme se trošku podrobněji nad tím, co náš model říká. Pomůže nám v tom zobrazení modelu pomocí regresní přímky. Jelikož víme, že regresor *pohlaví* u všech žen nabývá hodnoty 0 a u všech mužů hodnoty 1, můžeme se podívat, co model říká o mužích a o ženách zvlášť, jednoduše tak, že nulu nebo jedničku místo pohlaví do modelu dosadíme.

Tabulka 5: Záznamy o známkách studentů a studentek a délce sukní či nohavic

Student	Známka	Pohlaví	Délka	Student	Známka	Pohlaví	Délka
1	B (2)	1	104	25	E (5)	1	59
2	D (4)	0	84	26	D (4)	0	80
3	B (2)	1	86	27	B (2)	0	21
4	B (2)	0	48	28	D (4)	0	70
5	E (5)	1	92	29	C (3)	1	99
6	A (1)	1	90	30	E (5)	1	52
7	C (3)	1	95	31	C (3)	0	48
8	D (4)	0	88	32	E (5)	0	89
9	E (5)	0	59	33	E (5)	0	83
10	A (1)	1	101	34	C (3)	1	79
11	C (3)	0	38	35	F (6)	1	30
12	E (5)	0	76	36	B (2)	0	39
13	E (5)	0	82	37	B (2)	1	102
14	A (1)	0	28	38	B (2)	1	100
15	B (2)	0	29	39	C (3)	0	46
16	D (4)	1	59	40	E (5)	0	92
17	C (3)	1	100	41	E (5)	1	91
18	A (1)	0	41	42	C (3)	0	88
19	C (3)	0	48	43	C (3)	1	95
20	B (2)	0	48	44	C (3)	1	80
21	E (5)	0	79	45	F (6)	1	41
22	A (1)	0	33	46	B (2)	0	41
23	F (6)	0	70	47	F (6)	1	52
24	C (3)	0	40				

Muži jsou značeni číslem 1, ženy 0. Délka sukně či nohavic je uvedena v centimetrech.

V případě žen (pohlaví = 0):

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot \text{pohlaví} \ + \ \beta_2 \cdot \text{délka}$$

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot 0 \ + \ \beta_2 \cdot \text{délka}$$

$$\text{známka} = \beta_0 \ + \ \beta_2 \cdot \text{délka}$$

$$\text{známka} = 3.001 \ + \ 0.006 \cdot \text{délka}$$

V případě mužů (pohlaví = 1):

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot \text{pohlaví} \ + \ \beta_2 \cdot \text{délka}$$

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot 1 \ + \ \beta_2 \cdot \text{délka}$$

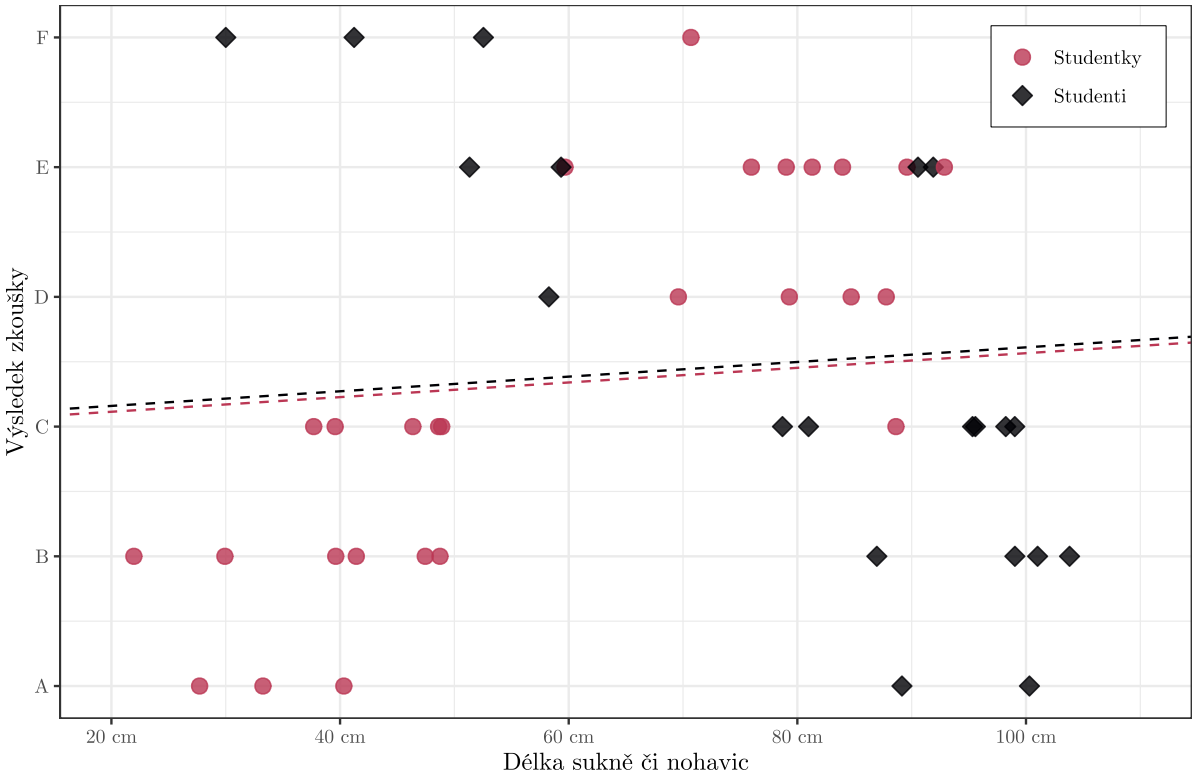
$$\text{známka} = (\beta_0 + \beta_1) \ + \ \beta_2 \cdot \text{délka}$$

$$\text{známka} = (3.001 + 0.045) \ + \ 0.006 \cdot \text{délka}$$

$$\text{známka} = 3.046 \ + \ 0.006 \cdot \text{délka}$$

Obě regresní přímky tedy mají stejný sklon (0.006) a liší se jen ve vertikálním posunutí (počátky jsou rovny 3.001 a 3.046). Na obrázku 3 je podobnost obou přímek patrná.

Obrázek 3: Nesprávně zvolený model bez interakčního členu



Řada čtenářů si zřejmě již teď uvědomuje chybu, které jsme se při tvorbě modelu dopustili. Délka sukně či nohavic má poněkud odlišný dopad v závislosti na tom, jestli se bavíme o studentce, či studentovi. Aby náš model dokázal přesněji popsat realitu, musíme jej navrhnout tak, aby váha regresoru *délka* mohla nabývat různých hodnot v závislosti na hodnotě regresoru *pohlaví*. A právě tuto možnost mu získáme přidáním interakčního členu *pohlaví* × *délka*.

Jak jsme uvedli výše, pro přidání interakce rozšíříme datovou tabulku o ještě jeden sloupec, ve kterém budou hodnoty *pohlaví* vynásobené hodnotami *délky* (prvních několik řádků obsahuje tabulka 6). Jelikož regresor *pohlaví* nabývá hodnot 0 a 1, interakční člen se rovná u žen nule a u mužů hodnotě regresoru *délka*. Postup výpočtu by byl ale stejný i u dvou kvantitativních regresorů.

Tabulka 6: Ukázka dat pro model s interakčním členem

Student	Známka	Pohlaví	Délka	Interakce
1	B (2)	1	104	104
2	D (4)	0	84	0
3	B (2)	1	86	86
4	B (2)	0	48	0
5	E (5)	1	92	92
6	A (1)	1	90	90
7	C (3)	1	95	95
8	D (4)	0	88	0
9	E (5)	0	59	0
10	A (1)	1	101	101
⋮	⋮	⋮	⋮	⋮

Model s interakcí má v tomto případě podobu

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot \text{pohlaví} \ + \ \beta_2 \cdot \text{délka} \ + \ \beta_3 \cdot \text{pohlaví} \cdot \text{délka}$$

a po přepočítání nás dovede k těmto hodnotám odhadů parametrů:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	0.260	
pohlaví	7.759	2.539
délka	0.052	0.853
interakce	−0.109	−3.038
R^2	63 %	

Než okomentujeme drastický nárůst vysvětleného rozptylu závisle proměnné, zamysleme se nad tím, co jednotlivé nestandardizované koeficienty modelu s interakcí znamenají. Pomůže nám opět prozkoumání modelu zvlášť u mužů a žen.

V případě žen (pohlaví = 0):

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot \text{pohlaví} \ + \ \beta_2 \cdot \text{délka} \ + \ \beta_3 \cdot \text{pohlaví} \cdot \text{délka}$$

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot 0 \ + \ \beta_2 \cdot \text{délka} \ + \ \beta_3 \cdot 0 \cdot \text{délka}$$

$$\text{známka} = \beta_0 \ + \ \beta_2 \cdot \text{délka}$$

$$\text{známka} = 0.260 \ + \ 0.052 \cdot \text{délka}$$

V případě mužů (pohlaví = 1):

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot \text{pohlaví} \ + \ \beta_2 \cdot \text{délka} \ + \ \beta_3 \cdot \text{pohlaví} \cdot \text{délka}$$

$$\text{známka} = \beta_0 \ + \ \beta_1 \cdot 1 \ + \ \beta_2 \cdot \text{délka} \ + \ \beta_3 \cdot 1 \cdot \text{délka}$$

$$\text{známka} = (\beta_0 + \beta_1) \ + \ (\beta_2 + \beta_3) \cdot \text{délka}$$

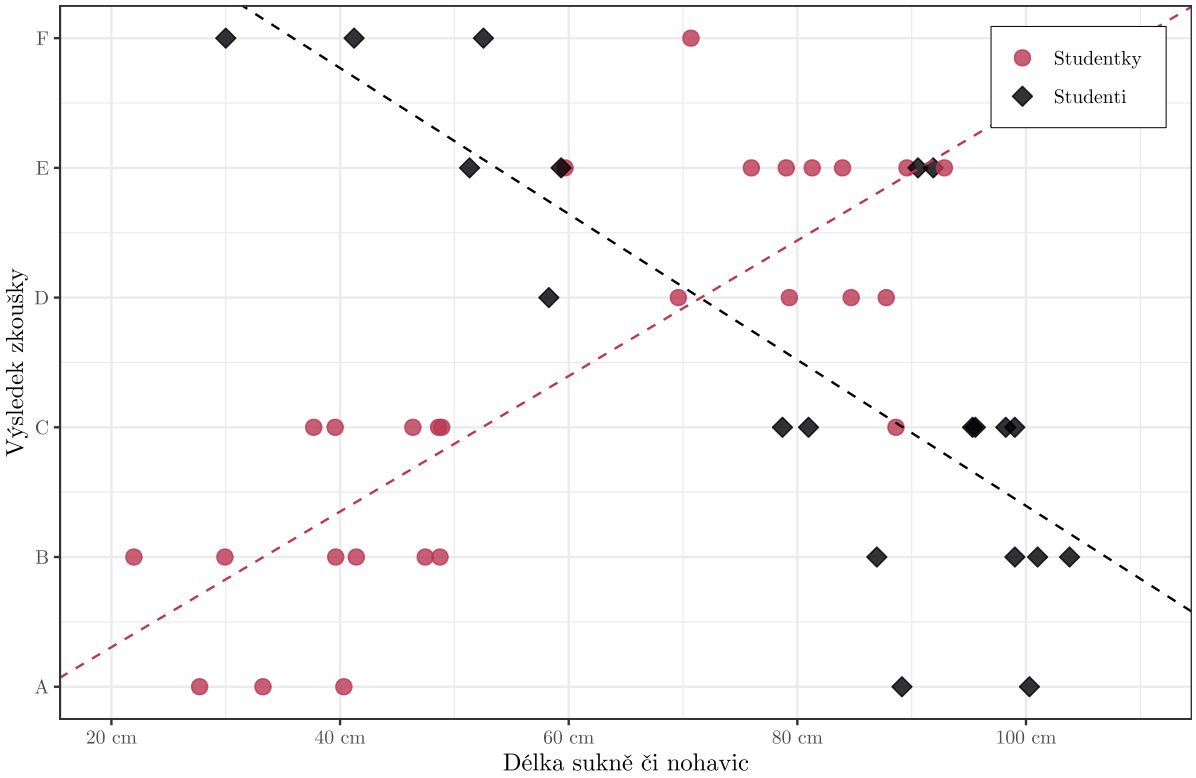
$$\text{známka} = (0.260 + 7.759) \ + \ (0.052 - 0.109) \cdot \text{délka}$$

$$\text{známka} = 8.020 \ - \ 0.056 \cdot \text{délka}$$

Parametry β_0 a β_2 lze interpretovat jako počátek a sklon regresní přímky v referenční skupině žen. Parametry β_1 a β_3 říkají, o kolik se počátek a sklon liší ve skupině mužů ve srovnání s referenční skupinou žen. Pozor, β_1 a β_3 nejsou hodnoty počátku a sklonu regresní přímky u mužů! Počátek pro muže je $\beta_0 + \beta_1$ a sklon regresní přímky $\beta_2 + \beta_3$. Toto je zřejmě nejčastější chyba, které se studenti při práci s interakcemi dopouští.

Srovnejme regresní přímky na obrázku 3 s přímkami na obrázku 4. Jak je patrné, za každých 20 centimetrů délky sukně se známky studentek v průměru zhorší přibližně o jeden stupeň (přesněji o $0.052 \cdot 20 = 1.045$ stupně), zatímco u studentů každých 20 centimetrů délky nohavic známku o přibližně jeden stupeň zlepšuje ($-0.056 \cdot 20 = -1.125$ stupně).

Obrázek 4: Model s interakčním členem



Bylo tedy podezření studentů oprávněné? V tomto případě by $R^2 = 63$ % bylo extrémně vysoké. Výsledek v podstatě říká, že téměř dvě třetiny rozptylu hodnocení závisí na tom, jestli je zkoušený mužem, nebo ženou a co má na sobě. Ve zbývající třetině rozptylu se pak tísní všechny ostatní vlivy, například znalosti, které by měly být hlavním faktorem.

Jak jsme výše několikrát zmínili, do interakce může vstupovat libovolný typ proměnné. Pokud bychom chtěli zkoumat interakci nominální proměnné s více než dvěma úrovněmi, museli bychom vytvořit interakční člen pro každou indikátorovou proměnnou, kterou jsme z původní nominální proměnné vytvořili. Tedy pro výpočet interakce nominální proměnné se třemi úrovněmi s nominální proměnnou se čtyřmi úrovněmi bychom museli vytvořit 6 ($= (3-1) \cdot (4-1)$) interakčních členů.

Analogicky lze vytvářet i interakce vyššího řádu – můžeme tedy vynásobit i více než dvě proměnné mezi sebou. Nicméně již v případě dvou proměnných je interpretace často nesnadná a v případě tří či více proměnných obvykle téměř nemožná.

Při zkoumání interakcí nemá smysl do našich úvah zahrnovat standardizované regresní koeficienty β^*, jelikož nepřináší prakticky žádnou relevantní informaci.

4.2.1 Centrování regresorů

Při popisu výsledků předešlého příkladu jsme se vyhýbali interpretaci absolutního členu β_0 i váhy regresoru *pohlaví*. Obě zmiňované váhy přitom své interpretace mají, ač, jak vzápětí zjistíme, nejsou v našem případě příliš smysluplné. Absolutní člen β_0 říká, jaké hodnoty nabývá závislá proměnná, pokud jsou všechny regresory rovny nule. Jde tedy o průměrnou známku ženy (*pohlaví* $= 0$), jejíž sukně má délku 0 cm. Tento scénář je s trochou fantazie ještě představitelný, nicméně samotná hodnota 0.26 již smysl nemá (je to známka o tři čtvrtě stupně lepší než A).

Váha β_1 popisuje ještě kurióznější situaci: jde o rozdíl mezi průměrnou známkou ženy bez sukně (*délka* $= 0$ cm) a muže bez kalhot. Hodnota mezi sedmičkou a osmičkou ($\beta_1 = 7.759$) je opět nesmyslná, když vezmeme v potaz, že známky mají jen 6 stupňů.

Na podobně nesmyslné váhy můžeme při práci s lineárními modely narazit poměrně často. A nemusí jít jen o modely s interakčním členem, ač pro ně je tato situace typická. Představme si, že budeme predikovat výši platu pomocí IQ jedince. Počátkem pak bude průměrný plat neexistujícího člověka, který má IQ rovno nule. Když do modelu zahrneme věk, bude počátek hovořit o platu novorozence a tak podobně.

Všechny popisované situace jsou z matematického pohledu v pořádku, těžko však popřeme to, že se příčí zdravému rozumu. Abychom se podobným absurditám vyhnuli, vyplatí se nám naučit se regresory *centrovat*.

Centrováním regresoru se rozumí to, že vypočítáme jeho aritmetický průměr a ten odečteme od hodnoty regresoru u každého pozorování. Pokud bychom chtěli centrovat délku sukně či nohavic v předešlém příkladu, spočítali bychom průměrnou délku, která se rovná 67.979 cm, a tuto hodnotu bychom od všech délek odečetli. Proměnná *délka* tedy už neudává absolutní délku, ale to, o kolik se délka liší od průměrné délky sukně či nohavic v rámci našeho souboru. Novou podobu datové matice zobrazuje tabulka 7. Po provedení centrování se odhady regresních vah modelu změní následovně:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	3.812	
pohlaví	0.384	0.126
délka	0.052	0.853
interakce	−0.109	−1.135
R^2	63 %	

Tabulka 7: Ukázka centrovaného regresoru

Student	Známka	Pohlaví	Délka	Interakce
1	B (2)	1	36.021	36.021
2	D (4)	0	16.021	0
3	B (2)	1	18.021	18.021
4	B (2)	0	−19.979	0
5	E (5)	1	24.021	24.021
6	A (1)	1	22.021	22.021
7	C (3)	1	27.021	27.021
8	D (4)	0	20.021	0
9	E (5)	0	−8.979	0
10	A (1)	1	33.021	33.021
⋮	⋮	⋮	⋮	⋮

Vycentrováním regresoru se přesnost nalezeného modelu ani hodnoty předpovědí, které nám model poskytne, nemění. Stejné zůstaly i sklony regresních přímek. Změnila se pouze hodnota a interpretace váhy β_0 i β_1. Počátek stále říká, v průměru jakou hodnotu závisle proměnné získáme, pokud jsou všechny regresory rovny nule. Jelikož je ale regresor *délka* centrovaný, nulová hodnota znamená, že se bavíme o průměrně dlouhé sukni. Váha β_1 pak kvantifikuje rozdíl v očekávané známce muže a ženy, kteří mají průměrně dlouhé nohavice, respektive sukni.

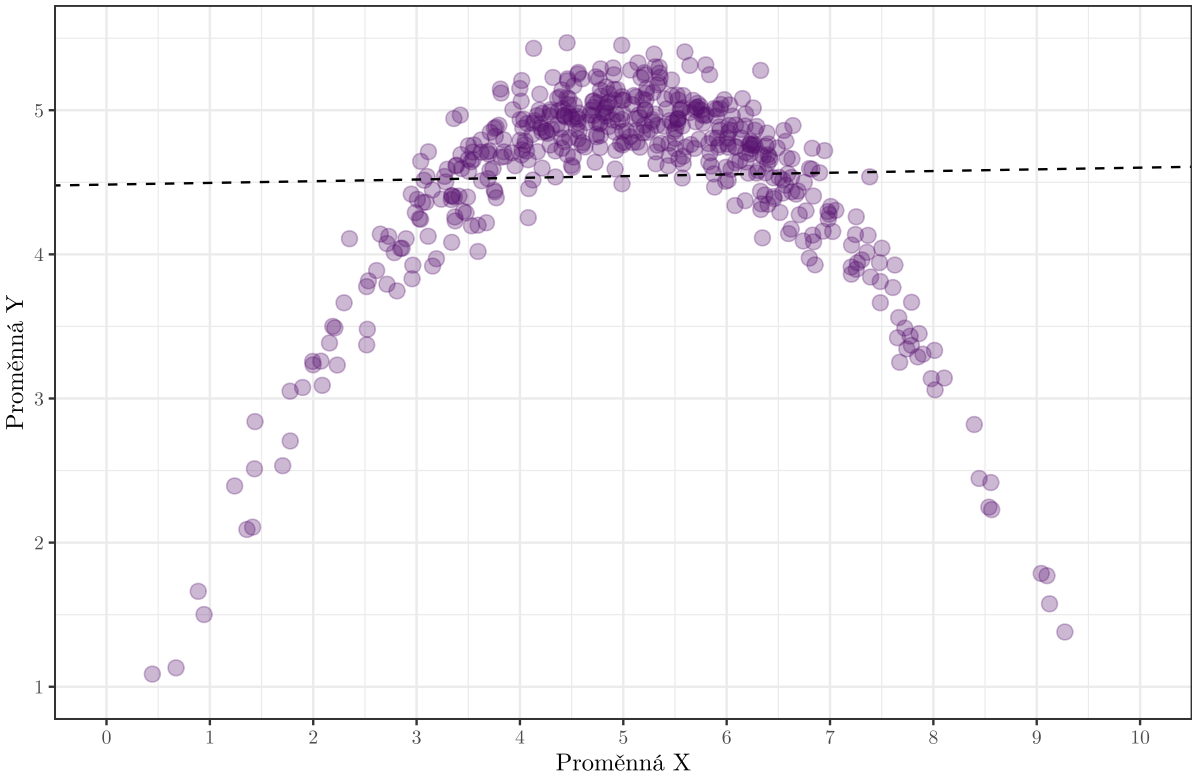
Pokud bychom centrovali regresor *IQ*, absolutní člen se bude týkat průměrně chytrého člověka, pokud regresor *věk*, pak průměrně starého atd. K centrování nemusíme použít aritmetický průměr, ale i jakoukoli jinou konstantu (ač pak nejde o centrování v pravém smyslu slova). Například místo průměru 67.979 cm jsme mohli použít číslo 70, nebo dokonce jsme mohli centrovat muže a ženy zvlášť jejich skupinovými průměry. Žádný z těchto kroků nezmění výsledky, jaké nám model poskytne. Změní se však hodnoty a interpretace dotčených regresních koeficientů.

Dichotomické proměnné obvykle nemá žádný smysl centrovat. Pokud náš model obsahuje interakci dvou kvantitativních proměnných, je jejich centrování téměř vždy nutností, pokud stojíme o srozumitelné výsledky.

4.3 Zkoumání nelineárních vztahů

Klasická poučka, kterou známe ze základních kurzů statistiky, říká, že Pearsonův korelační koeficient popisuje pouze *lineární* vztah mezi sledovanými proměnnými. Tuto vlastnost mají i regresní přímky, se kterými jsme pracovali v předešlých kapitolách. V praxi se však často setkáváme se situacemi, kdy mezi sledovanými proměnnými existuje jiný vztah než přímá úměra.

Obrázek 5: Data nesprávně proložená přímkou



Představme si, že pracujeme se dvěma proměnnými z bodového grafu na obrázku 5. Mezi těmito proměnnými zjevně existuje úzký vztah. Pokud se však tuto závislost pokusíme popsat pomocí modelu jednoduché regrese, nedokážeme ji odhalit.

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	4.484	
proměnná X	0.012	0.026
R^2	0 %	

Lineární model ve skutečnosti dokáže popsat široké spektrum nelineárních vztahů, musíme tomu však přizpůsobit použité regresory[9]. Pokud v psychologii mluvíme o nelineárním vztahu, v drtivé většině případů máme na mysli vztah kvadratický. Ten předpokládá, že závislost má tvar paraboly – připomíná tedy písmeno U (nebo Ω), případně nějakou jeho část. K popisu kvadratické závislosti musíme do modelu přidat daný regresor v druhé mocnině:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

[9] Poměrně rozšířená je domněnka, že slovo *lineární* v názvu modelu nějak souvisí s druhem vztahu mezi X a Y. Jde však o omyl – slovo lineární znamená, že závisle proměnná je modelována jako lineární kombinace (tzn. vážený součet) regresorů, kde jako váhy figurují parametry modelu.

V praxi bychom tuto změnu provedli tak, že do datové tabulky přidáme nový sloupec, který obsahuje hodnoty původního regresoru umocněné na druhou. Několik řádků takto upravené datové matice obsahuje tabulka 8. Další kroky výpočtu jsou identické s nám již známým postupem. Data z grafu 5 nás dovedou k následujícím odhadům parametrů.

Regresor	$\hat{\beta}$	$\hat{\beta}^*$
(počátek)	−0.012	
proměnná X	2.004	4.371
proměnná X^2	−0.200	−4.449
R^2	92 %	

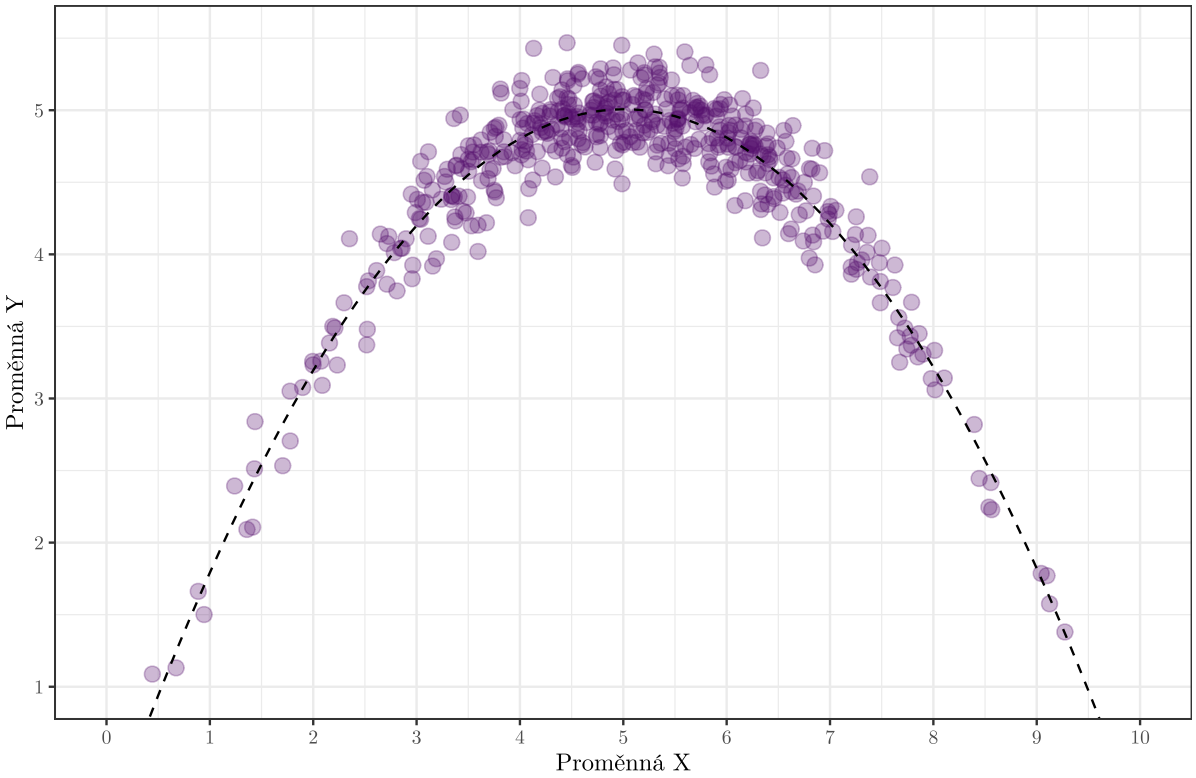
Tabulka 8: Ukázka dat pro modelování kvadratické závislosti

Y	X	X^2
13.2	1.5	2.25
21.2	3.0	9.00
21.4	3.2	10.24
23.6	4.0	16.00
24.4	4.2	17.64
25.3	4.8	23.04
24.8	5.5	30.25
24.3	5.7	32.49
23.7	6.0	36.00
16.9	7.8	60.84
10.4	8.8	77.44
⋮	⋮	⋮

Proložení dat parabolou přináší v našem případě velmi uspokojivý výsledek. Za pravdu nám dává nejen vysoká hodnota koeficientu determinace, ale i graf na obrázku 6, který naši odhadnutou parabolu znázorňuje.

Pokud modelujeme vliv některého regresoru jako kvadratický, nalezené regresní váhy ztratí své elegantní interpretace. Standardizované regresní koeficienty nám interpretaci neusnadní vůbec (všimněte si, že v našem příkladu vyšly v absurdně vysokých hodnotách) a i ty nestandardizované poskytují jen omezenou informaci. Koeficient β u regresoru X^2 říká, jestli je křivka prohnuta ve tvaru U (kladné hodnoty) nebo ve tvaru $\cap$ (záporné hodnoty). **Pokud je váha kvadratického regresoru blízko nule, znamená to, že závislost má lineární průběh a daný regresor bylo zbytečné do modelu zařazovat.** Koeficient regresoru X poskytuje ještě méně relevantních informací. Pokud je jeho hodnota blízko nule, znamená to, že parabola má svůj vrchol (respektive dno údolí)

Obrázek 6: Data proložená kvadratickou funkcí



horizontálně umístěný poblíž hodnoty 0. Hodnoty různé od nuly značí posunutí doleva či doprava (pokud oba koeficienty mají stejná znaménka, tak doleva, v opačném případě doprava). Toto jsou ale informace, které nám obvykle k žádnému užitku neslouží.

4.3.1 Další nelineární vztahy

V psychologii si obvykle vystačíme s lineárními vztahy, vzácně se uchylujeme ke kvadratickým. Lineární modely ve skutečnosti umožňují modelovat jakýkoli další vztah, který jsme schopni zapsat jako vážený součet nějakých regresorů nebo jejich kombinací. Do rovnice bychom mohli vedle regresoru X^2 přidat třeba i regresor X^3 nebo X^4, ba co víc, můžeme modelovat exponenciální závislost, funkce s absolutní hodnotou, goniometrické funkce atd.

Jako příklad zajímavého využití této plejády možností zmiňme modelování křivky funkce sinus. S tímto druhem závislosti se setkáme typicky tehdy, když regresor X značí čas a závisle proměnná Y vypovídá o hodnotách události, která se odehrávala v různých časech. Sinusoida říká, že hodnoty proměnné Y opakovaně rostou a klesají, a že tento průběh má vždy stejnou délku odpovídající určité *periodě*. Budeme-li předpokládat, že tuto periodu známe, potřebujeme určit hodnoty dvou dalších parametrů: amplitudu (výši kolísání) a horizontální posunutí (tedy čas, ve kterém začíná první pozorovaný cyklus).

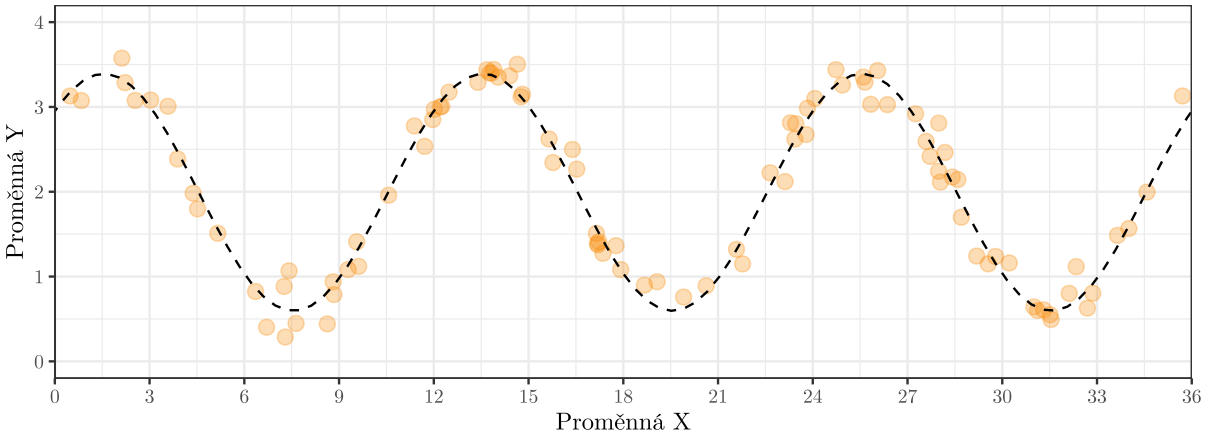
Lze odvodit, že pro popis tohoto druhu závislosti můžeme využít vztah:

$$Y = \beta_0 + \beta_1 \cdot \sin\left(\frac{2\pi}{s}X\right) + \beta_2 \cdot \cos\left(\frac{2\pi}{s}X\right)$$

kde s je délka periody – pokud by tedy hodnoty regresoru X byly v měsících a cyklus měl roční periodu, pak $s = 12$. Vhodné využití tohoto modelu znázorňuje obrázek 7.

Všimněte si, že odhadnuté regresní váhy nelze přímo interpretovat jako amplitudu a posunutí. Tyto hodnoty nicméně dokážeme vypočítat: amplituda $\rho = \sqrt{\beta_1^2 + \beta_2^2}$ a posunutí $\theta = \arcsin\frac{\beta_1}{\rho} = \arccos\frac{\beta_2}{\rho}$. Model pak můžeme přepsat do srozumitelnější podoby $Y = \beta_0 + \rho\cos(\frac{2\pi}{s}X - \theta)$.

Obrázek 7: Data proložená funkcí sinus



5 Testy statistické významnosti

Ze základních kurzů víme, že jakákoli popisná statistika je různě přesným či nepřesným odhadem skutečné hodnoty určitého parametru. Když jsme například v jedné z předešlých kapitol vypočítali, že průměrný reakční čas u šesti náhodně zvolených probandů byl 600 ms, neznamená to, že libovolní lidé skutečně mají v průměru reakční čas přesně 600 ms. Pokud bychom experiment zopakovali na dalších osobách, dostali bychom třeba hodnotu 558 nebo 613. Víme totiž, že aritmetický průměr je statistika – estimátor, jehož úkolem je co nejpřesněji vypovídat o skutečné hodnotě zkoumaného parametru.

V rámci lineárních modelů hrají roli estimátorů všechny regresní váhy $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$. Ač se jejich hodnoty u různých skupin probandů budou měnit, víme, že kolísají kolem nějakých nám neznámých, ale neproměnlivých hodnot $\beta_0, \beta_1, ..., \beta_k$. Pokud bychom měli k dispozici nekonečně velký soubor pozorování, naše odhady by byly dokonale přesné[10].

Podobně si můžeme představit, že koeficient determinace R^2 je odhadem parametru, který bychom mohli značit třeba ρ^2 a jehož hodnotu bychom opět znali jen tehdy, pokud bychom pracovali s hypotetickým nekonečně rozsáhlým souborem. A stejně tak dříve zmiňovaný reziduální rozptyl S^2_ϵ je ne zcela přesným odhadem skutečné, ač nám neznámé hodnoty σ^2_ϵ.

Jak jsme zvyklí ze základních kurzů statistiky, o parametrech můžeme formulovat hypotézy, jejichž platnost ověřujeme pomocí statistických testů. Dříve nabyté znalosti o testových statistikách, p-hodnotách, nulových či alternativních hypotézách uplatníme beze zbytku i v kontextu lineárních modelů. Jediný znatelný rozdíl bude v tom, že se tentokrát nemusíme učit desítky rozmanitých statistických testů, ale pro ověření libovolné hypotézy si vystačíme s jediným statistickým testem: s testem podmodelu.

5.1 Test podmodelu

Podmodel je statistickým modelem, který jsme vytvořili z nějakého původního modelu vypuštěním jednoho či více regresorů, případně tím, že jsme regresory svázali nějakými podmínkami. Uvažujme teď pouze první z těchto možností, o modelech s podmínkami se dočteme v kapitole 13.2. Představme si například lineární model se třemi regresory definovaný rovnicí

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

[10] Nestandardizované regresní váhy mají jako estimátory řadu dobrých vlastností. Jsou to nejlepší nestranné odhady parametrů β. S rostoucím rozsahem výběru snižují svůj rozptyl, a stávají se tak přesnějšími. Můžeme je proto označit za slabě konzistentní.

K tomuto modelu můžeme vytvořit například tyto podmodely:

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3$$
$$Y = \beta_0 + \beta_1 X_1$$
$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

nebo třeba podmodel bez regresorů

$$Y = \beta_0$$

U libovolného modelu či podmodelu dokážeme zjistit, jak věrně odpovídá naměřeným datům. K tomuto účelu můžeme použít koeficient determinace R^2 odvozený z RSČ. Lze očekávat, že když z modelu odstraníme jeden či více regresorů, pak se hodnota R^2 sníží (RSČ vzroste), jelikož model ztratí svou přesnost. Čím důležitější roli vyřazené regresory v modelu hrály, tím bude pokles R^2 citelnější. Naopak, kdyby se po vyřazení regresorů hodnota R^2 nezměnila, znamená to, že vyřazené regresory jsou v modelu zbytečné – jejich regresní váhy se rovnají nule. A právě toto je podstata testu podmodelu. **Test podmodelu testuje platnost nulové hypotézy $\rho^2_{model} = \rho^2_{podmodel}$, což je ekvivalentní tvrzení s tím, že váhy odstraněných regresorů byly nulové.**

Testovou statistiku, která nám pomůže o platnosti této hypotézy rozhodnout, lze odvodit ze tří vlastností, které má RSČ. Dodejme, že tyto vlastnosti platí jen za několika předpokladů, kterým budeme věnovat kapitolu 7, prozatím je považujme za splněné.

- Odhadujeme-li p regresních vah pomocí n pozorování, pak statistika $\frac{\text{RSČ}}{\sigma^2_\epsilon}$ má rozdělení chí kvadrát s $n - p$ stupni volnosti. Zapisujeme $\frac{\text{RSČ}}{\sigma^2_\epsilon} \sim \chi^2_{n-p}$.
- Je-li skutečná hodnota regresních vah regresorů, které jsme z modelu odstranili, rovna nule (tedy platí-li nulová hypotéza), pak platí $\frac{\text{RSČ}_{podmodel} - \text{RSČ}_{model}}{\sigma^2_\epsilon} \sim \chi^2_h$, kde h je počet odstraněných parametrů, tedy rozdíl stupňů volnosti modelu a podmodelu.
- Statistiky RSČ_{model} a $\text{RSČ}_{podmodel} - \text{RSČ}_{model}$ jsou vzájemně nezávislé.

Fisherovo rozdělení pravděpodobnosti, které známe ze základních kurzů statistiky, definujeme jako podíl dvou nezávislých náhodných veličin s rozdělením χ^2 dělených jejich stupni volnosti. Můžeme tedy říct, že náhodná veličina

$$F = \frac{\frac{\text{RSČ}_{podmodel} - \text{RSČ}_{model}}{sv_{podmodel} - sv_{model}}}{\frac{\text{RSČ}_{model}}{sv_{model}}}$$

kde sv označuje počet stupňů volnosti (tedy $sv = n - p$), má za platnosti nulové hypotézy Fisherovo rozdělení pravděpodobnosti s počty stupňů volnosti odpovídajícími jmenovatelům obou zlomků podílu, tedy $sv_{podmodel} - sv_{model}$ a sv_{model}. Řada čtenářů dává při výpočtu přednost R^2 před RSČ, uveďme proto ekvivalentní podobu statistiky F vycházející z koeficientu determinace:

$$F = \frac{\frac{R^2_{model} - R^2_{podmodel}}{sv_{podmodel} - sv_{model}}}{\frac{1 - R^2_{model}}{sv_{model}}}$$

Oba zápisy vedou pochopitelně k identickému výsledku. Statistiku můžeme zjednodušit ještě víc, když si uvědomíme, že počet stupňů volnosti modelu a podmodelu se liší právě o počet vynechaných parametrů v podmodelu. Pokud jsme tedy podmodel vytvořili vynecháním jediného parametru, pak $sv_{podmodel} - sv_{model}$ se rovná jedné, pokud jsme vynechali dva parametry, pak dvěma atd. Označíme-li počet vynechaných parametrů písmenem h a rozdíl koeficientů determinace modelu a podmodelu ΔR^2, pak si statistiku F pro test podmodelu můžeme zapamatovat ve tvaru:

$$F = \frac{\frac{\Delta R^2}{h}}{\frac{1 - R^2_{model}}{n - p}} \sim F_{h,\, n-p}$$

Obecně tedy můžeme říct, že **test podmodelu ověřuje platnost nulové hypotézy o tom, že vyřazením jednoho či více regresorů nedošlo k poklesu procenta vysvětleného rozptylu závisle proměnné**.

Jak test podmodelu využíváme v praxi? Typicky testujeme tyto podmodely:

- Podmodel vzniklý vyřazením jediného regresoru s vahou β_j. Nulová hypotéza říká, že model vysvětluje stejné množství rozptylu bez ohledu na to, jestli obsahuje daný regresor, nebo ne. Tato hypotéza je ekvivalentní s tvrzením, že daný regresní koeficient se rovná nule, tedy $H_0 : \beta_j = 0$.
- Podmodel vzniklý vyřazením všech regresorů (tedy $Y = \beta_0$). Jelikož podmodel bez regresorů nevysvětluje žádný rozptyl, tak tento test ověřuje to, zda náš model vůbec vysvětluje nenulové množství rozptylu. Nulová hypotéza by tedy šla zapsat jako: $H_0 : \rho^2 = 0$.
- Podmodel vzniklý vyřazením všech indikátorů patřících k jednomu nominálnímu regresoru. Tento test nám umožní ověřit nulovou hypotézu o tom, že daný nominální regresor model nijak nezpřesňuje, a nemá tedy hledaný vztah k závisle proměnné Y.

5.2 Waldova statistika a intervaly spolehlivosti regresních vah

Pokud si klademe otázku, které regresory našeho modelu mají statisticky významný dopad (tedy jejich regresní koeficienty se liší od nuly), můžeme si představit, že počítačový program všechny regresory modelu projde, vždy daný regresor z modelu vyřadí, vypočítá R^2 daného podmodelu, pak regresor do modelu vrátí, vyřadí jiný atd. Ve skutečnosti tímto způsobem testování jednotlivých regresorů *neprobíhá*. Pro výpočet jejich významnosti se používá Waldova statistika. Ta poskytuje identické p-hodnoty jako test podmodelu, je však výpočetně méně náročná a má několik dalších užitečných vlastností.

Pro výpočet Waldovy statistiky musíme nejprve odhadnout varianční matici (*covariance matrix*) regresních vah. Připomeňme, že jednotlivé odhady $\hat{\beta}_j$ jsou náhodné veličiny, a mají tedy své rozptyly. Navíc obecně nejsou nezávislé, takže pro libovolné dva odhady koeficientů můžeme odhadnout jejich kovarianci. Varianční matice právě tyto informace obsahuje – má tvar čtverce s odhady rozptylů jednotlivých regresních koeficientů na diagonále (značme S_j^2) a s odhady kovariancí mimo diagonálu (značme S_{ij}):

$$\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} S_0^2 & S_{0,1} & \cdots & S_{0,k} \\ S_{1,0} & S_1^2 & \cdots & S_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k,0} & S_{k,1} & \cdots & S_k^2 \end{pmatrix}$$

Odhad varianční matice je poměrně snadný, musíme však opět sáhnout po maticových počtech:

$$\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}}) = S_\epsilon^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Při výpočtu Waldovy statistiky si vystačíme s diagonálními prvky této matice, tedy s rozptyly odhadů regresních koeficientů. Waldova statistika T je podíl odhadu regresní váhy β_j její směrodatnou odchylkou S_j. Za platnosti nulové hypotézy $H_0 : \beta_j = 0$ má Waldova statistika Studentovo rozdělení s $n - p$ stupni volnosti:

$$T = \frac{\hat{\beta}_j}{\sqrt{S_j^2}} \sim t_{n-p}$$

Při testování významnosti skupin regresorů Waldovu statistiku v této podobě nelze využít, při testování jednotlivých regresorů je nicméně testem první volby. Není proto překvapením, že statistické programy k odhadům regresních vah obvykle doplňují údaj o jejich směrodatné odchylce (S_j) a o hodnotě Waldovy statistiky t.

Opět však zopakujme, že Waldova statistika vede ke stejným výsledkům jako test příslušného podmodelu. Dokonce si můžeme všimnout vztahu obou statistik: $t^2 = F$. To odpovídá vztahu mezi Studentovým a Fisherovým rozdělením:

$$t^2_{n-p} = F_{1,n-p}$$

Práce se směrodatnou odchylkou odhadu regresní váhy (takzvanou směrodatnou chybou) má také tu výhodu, že nám umožňuje vytvářet **intervaly spolehlivosti** (též označované jako konfidenční intervaly) pro jednotlivé regresní váhy. Pro odhad regresních koeficientů platí $\hat{\beta}_j \sim N(\beta_j, \sigma^2_j)$, pro odhad rozptylu koeficientu $\hat{\beta}_j$ pak $S^2_j \sim \frac{\sigma^2_j}{n-p}\chi^2_{n-p}$, s tím, že obě statistiky jsou nezávislé. Interval spolehlivosti regresní váhy proto můžeme vytvořit analogicky k intervalu spolehlivosti pro střední hodnotu, který známe ze základních kurzů:

$$I_{1-\alpha} = \hat{\beta}_j \pm S_j \cdot t_{n-p;1-\frac{\alpha}{2}}$$

kde $t_{n-p;1-\frac{\alpha}{2}}$ je příslušný kvantil Studentova t rozdělení (hodnotu α nejčastěji nastavujeme na 5 %). Dodejme, že využití intervalových odhadů patří k dobré praxi při reportování výsledků, jelikož čtenáři umožní si udělat obrázek o věrohodnosti nalezených výsledků lépe než p-hodnota[11].

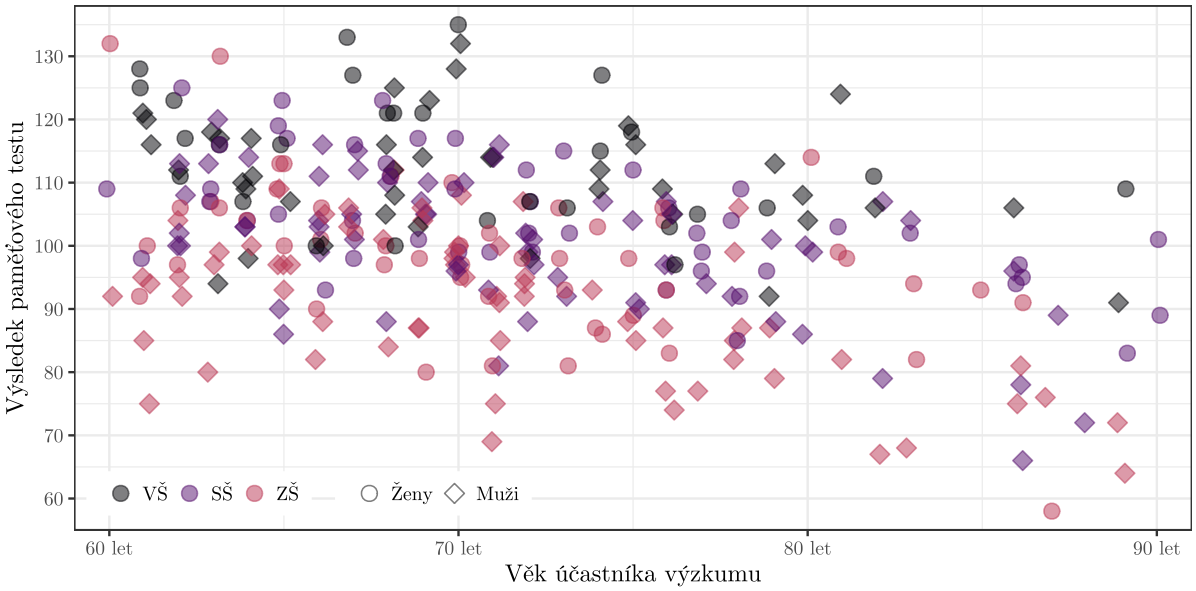
5.3 Ekvivalence s bivariátními testy

Pro lepší představu logiky testů nulových hypotéz demonstrujme jejich využití na simulovaných datech. Představme si situaci, kdy na datech získaných z rozsáhlého souboru seniorů modelujeme úbytek paměťových kompetencí v průběhu času. U každého z účastníků výzkumu jsme zaznamenali jeho pohlaví, věk, nejvyšší dosažené vzdělání a výsledek paměťového testu. Naměřené hodnoty znázorňuje obrázek 8. Abychom mohli interpretovat absolutní člen našich modelů, odečteme od proměnné *věk* číslo 60, takže nula bude odpovídat šedesáti rokům věku.

Dříve než přikročíme k testování nulových hypotéz v rámci modelu s mnoha regresory, demonstrujme pro řadu čtenářů možná překvapivý fakt: mnohé z bivariátních testů (tzn. testů dvojice proměnných), které známe ze základních kurzů statistiky, jsou ve skutečnosti zvláštními případy testu podmodelu. Zjistíme, že t-testy, analýza rozptylu či test Pearsonova korelačního koeficientu jsou jeden a tentýž test, pokud se na ně podíváme optikou lineárních modelů.

[11] Intervaly spolehlivosti můžeme vytvářet nejen pro jednotlivé regresní váhy, ale například i pro dvojice či větší skupiny regresorů dohromady. Pak by šlo o konfidenční elipsy či vícerozměrné elipsoidy. Jelikož tento postup obvykle v psychologii nevyužíváme, odkažme čtenáře na pokročilejší statistickou literaturu.

Obrázek 8: Vizualizace ukázkových dat



T-test pro dva nezávislé výběry

Řekněme, že se tážeme, jestli muži i ženy v daném populačním segmentu získávají v průměru stejně vysoká skóre v našem paměťovém testu, přičemž nezohledňujeme vzdělání ani věk. Standardně bychom sáhli po t-testu pro dva nezávislé výběry. V našem konkrétním případě by přinesl výsledek $t(298) = -3.708, p < 0.001$. Stejnou hypotézu však můžeme testovat pomocí modelu

$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví}$$

Waldova statistika koeficientu β_1 přesně odpovídá výše uvedené statistice t a nezmění se ani stupně volnosti či p-hodnota:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$	Statistika	p
(počátek)	104.383		$t(298) = 90.561$	< 0.001
pohlaví	-5.644	-0.210	$t(298) = -3.708$	< 0.001
R^2	4.4 %		$F(1; 298) = 13.749$	< 0.001

Standardizovaný koeficient β_1^* odpovídá hodnotě bodově-biseriálního korelačního koeficientu mezi proměnnou pohlaví a výkonem v testu. Míru účinku, Cohenovo d, bychom získali jako podíl rozdílu obou skupin (β_1) reziduální směrodatnou odchylkou $\hat{\sigma}_\epsilon$.

Test Pearsonova korelačního koeficientu

Podobně, jako jsme izolovaně testovali významnost regresoru *pohlaví*, můžeme statistickým testem ověřit, jestli mezi věkem probandů a výkonem v paměťovém testu existuje

lineární souvislost. Pokud budeme ignorovat vliv dalších faktorů – pohlaví a vzdělání – zřejmě bychom sáhli po Pearsonovu korelačním koeficientu a statistickém testu, který ověřuje, jestli je, či není korelace nulová. Nalezli bychom hodnoty $r = -0.427, t(298) = -8.148, p < 0.001$. K přesně stejnému výsledku nás dovede test regresoru *věk* v modelu

$$Y = \beta_0 + \beta_1 \cdot \text{věk}$$

Regresor	$\hat{\beta}$	$\hat{\beta}^*$	Statistika	p
(počátek)	110.050		t(298) = 84.929	< 0.001
Věk	−0.782	−0.427	t(298) = −8.148	< 0.001
R^2	18.2 %		F(1; 298) = 66.385	< 0.001

Ba co víc, hodnotě korelačního koeficientu přesně odpovídá standardizovaná váha $\hat{\beta}_1^*$ a koeficient determinace R^2 odpovídá jeho druhé mocnině.

Analýza rozptylu

Pokud bychom se tázali, zdali probandi dle různé úrovně vzdělání dostávají v průměru různé bodové ohodnocení, a nezohledňovali bychom pohlaví či věk, byla by adekvátní volbou analýza rozptylu při jednoduchém třídění. Test by nás dovedl k výsledku $F(2; 297) = 59.741$, $p < 0.001$. I tento test lze realizovat v rámci lineárního modelu

$$Y = \beta_0 + \beta_1 \cdot \text{SŠ} + \beta_2 \cdot \text{ZŠ}$$

kde *SŠ* a *ZŠ* jsou indikátorové proměnné regresoru vzdělání. Analýze rozptylu odpovídá test celé skupiny indikátorových proměnných vážících se k proměnné *vzdělání*, respektive test významnosti celého modelu. Pochopitelně nezáleží na tom, kterou skupinu označíme za referenční.

Regresor	$\hat{\beta}$	$\hat{\beta}^*$	Statistika	p
(počátek)	112.536		t(297) = 82.856	< 0.001
SŠ	−10.766	−0.392	t(297) = −6.246	< 0.001
ZŠ	−18.646	−0.685	t(297) = −10.905	< 0.001
R^2	28.7 %		F(2; 297) = 59.741	< 0.001

V kontextu analýzy rozptylu bývá mezi post-hoc testy zmiňován LSD test (*Least Significant Difference*). Jde o test, který srovná všechny dvojice skupin mezi sebou, na rozdíl od například Scheffého a Tukeyho testu ale není součástí LSD testu korekce na mnohonásobné testování. Výsledky LSD testu odpovídají výsledkům testů statistické významnosti jednotlivých indikátorových proměnných.

T-test pro jeden výběr

Hypotézu o tom, že se průměrná hodnota, které probandi dosahují v paměťovém testu, liší od 100, by bylo možné ověřit pomocí t-testu pro jeden výběr. V našem případě by vedl k výsledku $t(299) = 1.492, p = 0.137$. I zde lze test provést v rámci lineárního modelu, a to tak, že budeme srovnávat hodnotu absolutního členu se zadanou hodnotou 100, k čemuž bychom ovšem museli využít model s podmínkou. Případně bychom použili test pomocí Waldovy statistiky v obecnější podobě

$$T = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{S_j^2}} = \frac{\hat{\beta}_j - 100}{\sqrt{S_j^2}} \sim t_{n-p}$$

Zřejmě nejpohodlnější cesta však bude transformovat závisle proměnnou odečtením čísla 100 a ověřit hypotézu, že $\beta_0 = 0$, v rámci modelu

$$(Y - 100) = \beta_0$$

I v tomto případě získáme identický výsledek.

Regresor	$\hat{\beta}$	$\hat{\beta}^*$	Statistika	p
(počátek)	1.147		t(299) = 1.492	0.137

5.4 Příklad použití testů nulových hypotéz

V praxi bychom samozřejmě netestovali každou hypotézu pomocí zvláštního modelu, jelikož tím bychom přišli o výhody, které nám statistické modelování nabízí. Správnou cestou je použití jediného komplexního modelu, s jehož pomocí realizujeme jednotlivé testy. Mohl by vypadat například následovně:

$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví} + \beta_2 \cdot \text{věk} + \beta_3 \cdot \text{SŠ} + \beta_4 \cdot \text{ZŠ}$$

Regresor	$\hat{\beta}$	$\hat{\beta}^*$	Statistika	p
(počátek)	123.901		t(295) = 81.034	< 0.001
Věk	−0.755	−0.412	t(295) = −9.994	< 0.001
Pohlaví	−5.995	−0.223	t(295) = −5.436	< 0.001
Vzdělání			F(2; 295) = 81.977	< 0.001
SŠ	−9.488	−0.346	t(295) = −6.541	< 0.001
ZŠ	−18167	−0.668	t(295) = −12.677	< 0.001
R^2	50.4 %		F(4; 295) = 74.821	< 0.001

Nalezené p-hodnoty jednotlivých regresorů se vztahují k nulovým hypotézám ve tvaru $H_0 : \beta_j = 0$. V případě věku a pohlaví je jejich význam zjevný, v případě indikátorových proměnných *ZŠ* a *SŠ* se test týká nulové hypotézy o tom, že se daná skupina neliší od referenční skupiny, kterou je v našem případě *VŠ*. Pokud bychom chtěli ověřit statistickou významnost rozdílu mezi *ZŠ* a *SŠ*, nejjednodušší cestou by bylo vybrat některou jinou referenční skupinu a provést test znovu. F statistika na řádku *Vzdělání* je výsledek testu trsu indikátorových proměnných (tedy *ZŠ* a *SŠ* dohromady).

Statistický test na posledním řádku se týká modelu jako celku a ověřuje, jestli se množství vysvětleného rozptylu (ρ^2) liší od nuly.

Testy interakcí

Výše uvedený model předpokládá, že efekt věku bude stejný u mužů i žen, i u lidí libovolného vzdělání. Ať už se tedy bavíme o kterékoli skupině, každý rok nás v průměru stojí přibližně tři čtvrtiny bodu. Zároveň však platí to, že muž libovolného věku má v průměru o 6 bodů méně než stejně stará žena a že mezi vysokoškolákem a člověkem se základním vzdělání je rozdíl přibližně 18 bodů (za předpokladu, že jsou oba stejného věku a pohlaví). Odpovídá toto skutečnosti? Bylo by poměrně odůvodněné spekulovat nad tím, že rychlost úpadku paměťových kompetencí (tedy efekt věku) je v různých skupinách rozdílná.

Významnost interakce regresorů *pohlaví* a *věk* je v modelu reprezentována jediným regresorem, který vznikl jako součin dvou jmenovaných regresorů. K ověření významnosti interakce bychom mohli využít příslušnou Waldovu statistiku nebo test podmodelu, kdy z modelu vypouštíme interakční člen (označený červeně):

$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví} + \beta_2 \cdot \text{věk} + \beta_3 \cdot \text{SŠ} + \beta_4 \cdot \text{ZŠ} + \beta_5 \cdot \text{pohlaví} \cdot \text{věk}$$

Nalezená regresní váha $\hat{\beta}_5$ se téměř přesně rovná nule a efekt není statisticky významný, $t(294) = 0.045, p = 0.964$. Nenalezli jsme tedy důkazy o tom, že by úbytek paměťových kompetencí byl u mužů a žen různě strmý.

Vstupuje-li do interakce nominální regresor s více úrovněmi, otestujeme jeho významnost pomocí testu podmodelu, kdy vypustíme všechny interakční členy. Pokud se budeme ptát, jestli je rychlost úpadku paměti různá u různých skupin dle vzdělání, reprezentovali bychom to tímto modelem a podmodelem:

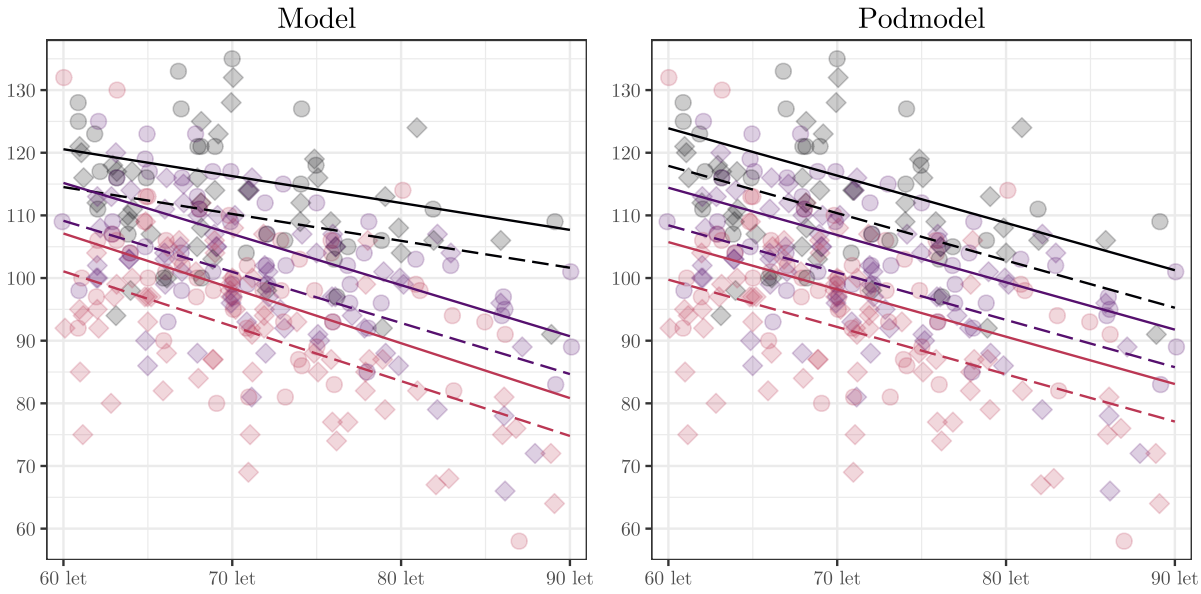
$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví} + \beta_2 \cdot \text{věk} + \beta_3 \cdot \text{SŠ} + \beta_4 \cdot \text{ZŠ} + \beta_5 \cdot \text{SŠ} \cdot \text{věk} + \beta_6 \cdot \text{ZŠ} \cdot \text{věk}$$

Rozdíl modelu a příslušného podmodelu (tedy bez červeně označených členů) znázorňuje obrázek 9. Statistický test přinese následující výsledky:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$	Statistika	p
(počátek)	120.560	0.000	$t(293) = 57.139$	< 0.001
Věk	−0.429	−0.234	$t(293) = -2.657$	0.008
Pohlaví	−6.040	−0.225	$t(293) = -5.505$	< 0.001
Vzdělání			$F(2; 293) = 14.515$	< 0.001
SŠ	−5.380	−0.196	$t(293) = -2.039$	0.042
ZŠ	−13.452	−0.494	$t(293) = -5.218$	< 0.001
Vzdělání × věk			$F(2; 293) = 2.661$	0.072
SŠ × věk	−0.386	−0.218	$t(293) = -1.926$	0.055
ZŠ × věk	−0.447	−0.237	$t(293) = -2.212$	0.028
R^2	51.2 %		$F(6; 293) = 51.329$	< 0.001

Ač na obrázku 9 zřetelně vidíme, že přinejmenším v našem souboru úpadek paměťových funkcí probíhá u vysokoškoláků o něco pomaleji než ve zbývajících dvou skupinách, rozdíl není signifikantní ($p = 0.072$). Nenašli jsme tedy podporu pro naši hypotézu, že rychlost úpadku paměťových kompetencí souvisí se vzděláním jedince. Poněkud paradoxně můžou působit dílčí výsledky – rozdíl efektu věku mezi vysokoškoláky a lidmi se základním vzděláním signifikantní je ($p = 0.028$) a mezi vysokoškoláky a středoškoláky je na hranicích statistické významnosti ($p = 0.055$). Příčinou je to, že párová srovnání nejsou nijak korigována, zatímco souhrnný test nominálního regresoru s faktem, že zkoumáme více proměnných, kalkuluje. Při interpretaci významnosti interakčního členu *ZŠ* × *věk* bychom tedy měli být opatrní, jelikož nejde o příliš přesvědčivý důkaz.

Obrázek 9: Model s interakcí pohlaví × věk a bez ní



Stejně jako v obrázku 8 jsou barvami odlišeny skupiny dle vzdělání a tvarem značky muži a ženy. Vynesené přímky pak odpovídají regresním přímkám jednotlivých skupin dle vzdělání a pohlaví (ženy jsou značeny plnou čarou, muži přerušovanou). Totéž platí pro přímky i křivky v následujících obrázcích.

Testování kvadratického vztahu

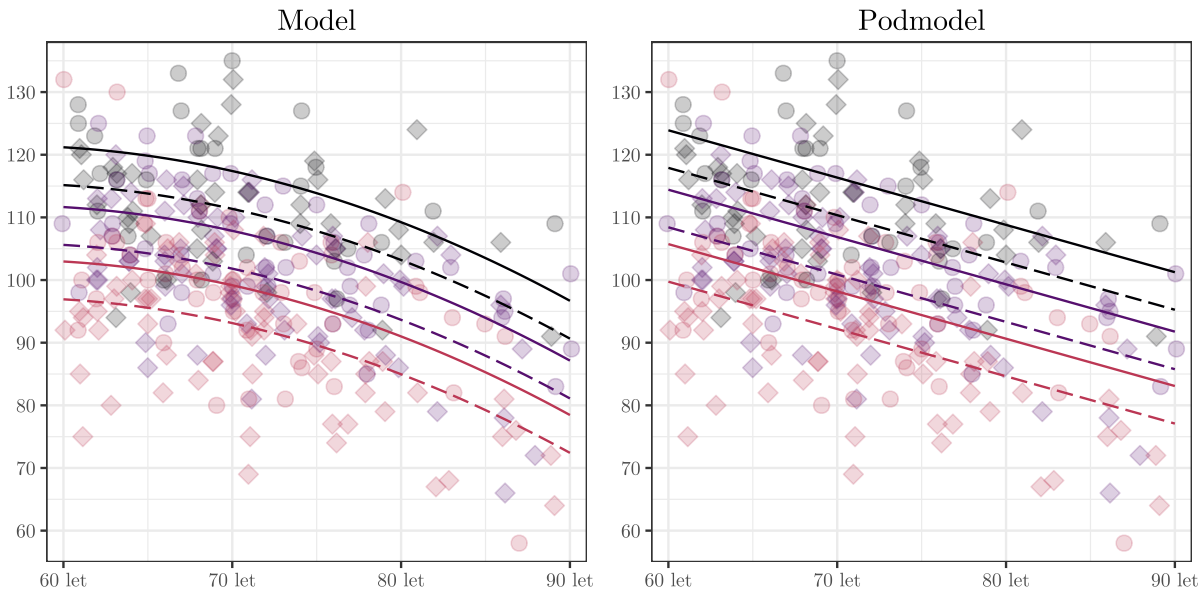
Rozumné by bylo také předpokládat, že úpadek neprobíhá stabilní rychlostí, ale že se v čase zrychluje. Můžeme jej modelovat jako kvadratickou závislost. Hypotézu bychom ověřili pomocí následujícího modelu a jeho podmodelu:

$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví} + \beta_2 \cdot \text{věk} + \beta_3 \cdot \text{SŠ} + \beta_4 \cdot \text{ZŠ} + \beta_5 \cdot \text{věk}^2$$

Rozdíl mezi modelem a podmodelem znázorňuje obrázek 10 a statistickou významnost rozdílu následující tabulka:

Regresor	$\hat{\beta}$	$\hat{\beta}^*$	Statistika	p
(počátek)	121.192		$t(294) = 64.434$	< 0.001
Věk	−0.161	−0.088	$t(294) = -0.628$	0.530
Věk^2	−0.022	−0.339	$t(294) = -2.434$	0.016
Pohlaví	−6.039	−0.225	$t(294) = -5.520$	< 0.001
Vzdělání			$F(2; 294) = 83.879$	< 0.001
SŠ	−9.545	−0.348	$t(294) = -6.634$	< 0.001
ZŠ	−18.232	−0.670	$t(294) = -12.826$	< 0.001
R^2	51.3 %		$F(5; 294) = 61.040$	< 0.001

Obrázek 10: Model s kvadratickým a lineárním vlivem věku



Jelikož jde o test jediného regresoru, můžeme použít Waldovu statistiku, $t(294) = -2.434, p = 0.016$, a konstatovat, že naši hypotézu o kvadratickém vlivu věku můžeme přijmout.

Velmi neintuitivní je interpretace statistického testu regresoru *věk*. Nejedná se o test hypotézy o tom, že *věk* má vliv na paměťové kompetence, ale o poměrně nezajímavou hypotézu o tom, že vrchol oblouku paraboly vlivu věku je umístěný nad nulou (tedy v 60 letech). Tímto problémem jsme se již zabývali v kapitole 4.3. Přestože je regresor *věk* v modelu nezbytný, nemá smysl jeho hodnotu a statistickou významnost interpretovat. Pokud bychom chtěli otestovat statistickou významnost věku s tím, že předpokládáme, že *věk* může mít kvadratický vliv, museli bychom vytvořit podmodel bez obou regresorů *věk* i *věk*^{2}:

$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví} + \beta_2 \cdot \text{věk} + \beta_3 \cdot \text{SŠ} + \beta_4 \cdot \text{ZŠ} + \beta_5 \cdot \text{věk}^2$$

Příslušný test potvrdí statisticky významný vliv věku, $F(2; 294) = 53.739, p < 0.001$.

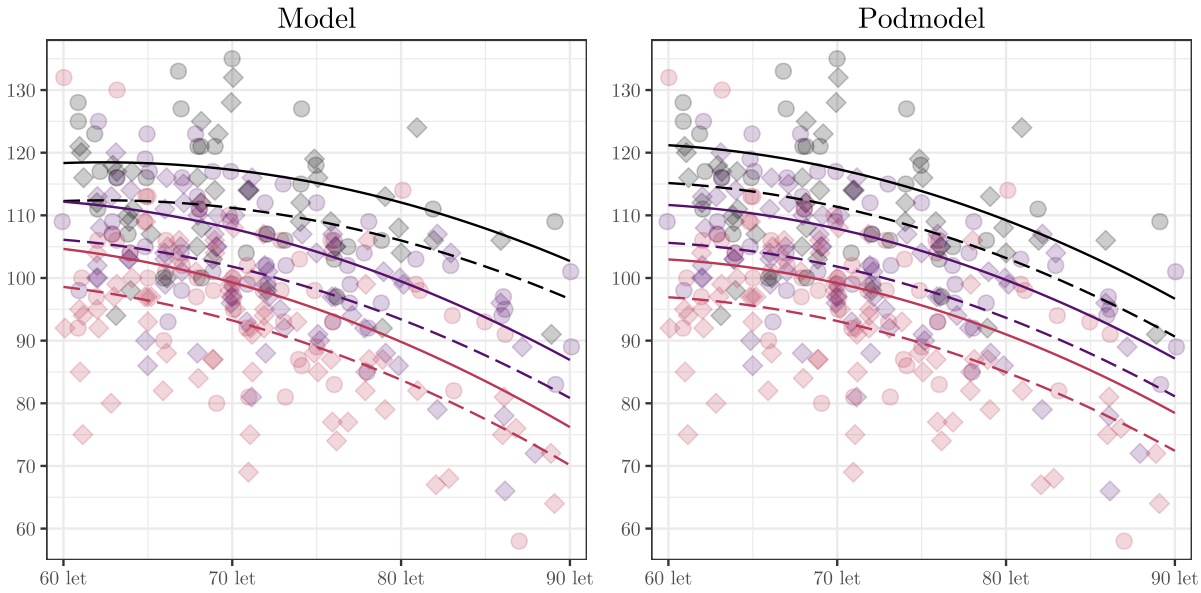
Situace bude ještě o poznání komplikovanější, když do modelu zahrneme kvadratický vliv věku a zároveň budeme předpokládat, že tento vliv je různý pro různé skupiny dle vzdělání. Pokud bychom testovali tento model a podmodel

$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví} + \beta_2 \cdot \text{věk} + \beta_3 \cdot \text{SŠ} + \beta_4 \cdot \text{ZŠ} + \beta_5 \cdot \text{SŠ} \cdot \text{věk} + \beta_6 \cdot \text{ZŠ} \cdot \text{věk} + \beta_7 \cdot \text{věk}^2$$

předpokládáme, že křivka věku je pro všechny skupiny stejným způsobem prohnutá, ale kromě posunutí nahoru a dolů, může být též pro jednotlivé skupiny posunutá vlevo a vpravo. Opět tedy testujeme ne příliš zajímavou hypotézu o tom, zda vrcholy jednotlivých parabol jsou lokalizované nad stejnou hodnotou osy x nebo ne.

Výsledek není signifikantní, $F(2; 292) = 2.303, p = 0.102$, ač grafická reprezentace na obrázku 11 je poměrně pěkně interpretovatelná.

Obrázek 11: Model s kvadratickým vlivem věku s interakčním členem a bez něj



Smysluplnější by zřejmě bylo modelovat závisle proměnnou s pomocí modelu s interakcí *vzdělání* × *věk*, ale i *vzdělání* × *věk*^{2}:

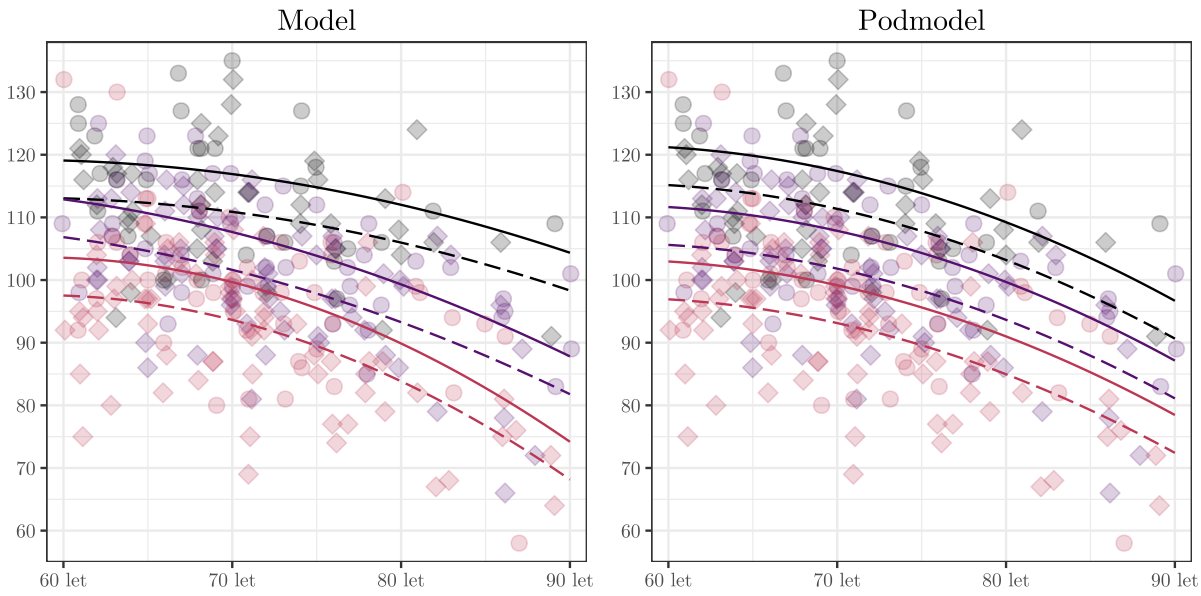
$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví} + \beta_2 \cdot \text{věk} + \beta_3 \cdot \text{SŠ} + \beta_4 \cdot \text{ZŠ} + \beta_5 \cdot \text{SŠ} \cdot \text{věk} + \beta_6 \cdot \text{ZŠ} \cdot \text{věk} + \beta_7 \cdot \text{věk}^2 + \beta_8 \cdot \text{SŠ} \cdot \text{věk}^2 + \beta_9 \cdot \text{ZŠ} \cdot \text{věk}^2$$

Test podmodelu bez červeně vyznačených členů ověřuje hypotézu o tom, že křivka úbytku paměťových kompetencí má různý tvar v závislosti na vzdělání (viz obrázek 12). Ani v tomto případě jsme nenašli důkazy k zamítnutí nulové hypotézy, $F(4; 290) = 1.300$, $p = 0.270$, což není překvapivé, jelikož (jak můžeme vidět při srovnání obrázků 11 a 12) náš model přináší skoro stejné řešení jako předchozí uvedený, přitom k popisu role věku využívá hned čtyři parametry místo původních dvou.

Pokud bychom i přesto, že jsme neprokázali různý vliv věku u různých skupin, tuto možnost připouštěli, změnil by se způsob, jakým bychom testovali obecnou hypotézu o tom, že regresor *věk* ovlivňuje výsledek paměťového testu. Srovnali bychom výše uvedený model s podmodelem, který postrádá všechny regresory (ať už interakční členy nebo hlavní efekty) obsahující regresor *věk* (v libovolné mocnině):

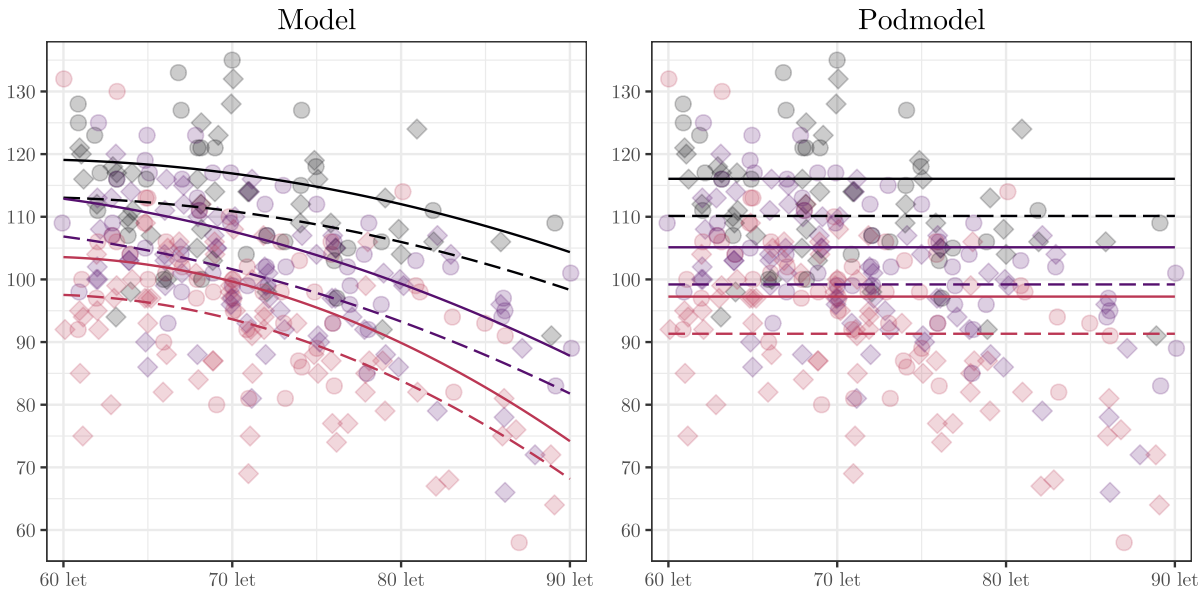
$$Y = \beta_0 + \beta_1 \cdot \text{pohlaví} + \beta_2 \cdot \text{věk} + \beta_3 \cdot \text{SŠ} + \beta_4 \cdot \text{ZŠ} + \beta_5 \cdot \text{SŠ} \cdot \text{věk} + \beta_6 \cdot \text{ZŠ} \cdot \text{věk} + \beta_7 \cdot \text{věk}^2 + \beta_8 \cdot \text{SŠ} \cdot \text{věk}^2 + \beta_9 \cdot \text{ZŠ} \cdot \text{věk}^2$$

Obrázek 12: Model s kvadratickým vlivem věku s interakcemi a bez nich



Výsledek bude opět statisticky významný, $F(6; 290) = 18.853, p < 0.001$. Rozdíl mezi modelem a podmodelem je patrný z obrázku 13.

Obrázek 13: Model s kvadratickým vlivem věku včetně interakcí a bez vlivu věku



6 Predikce a intervalové odhady

Ještě jednou se vraťme k úvaze o tom, že všechny číselné údaje, které získáme během výpočtů v rámci lineárního modelu, jsou realizace náhodných veličin. Nejde tedy o neměnné „pravdivé“ hodnoty – pokud bychom naši datovou matici ztratili a výzkum realizovali ještě jednou, tak přestože dodržíme všechny kroky a skutečný vztah zkoumaných fenoménů se nezmění, získáme mírně odlišné výsledky. Ze základních kurzů statistiky víme, že často je užitečné místo bodového odhadu (který je sice často velmi přesný, ale zákonitě se vždy od pravdivé hodnoty alespoň nepatrně liší) uvést interval spolehlivosti (tzv. konfidenční interval), který nám říká, v jakém rozmezí je rozumné skutečnou hodnotu odhadovaného parametru očekávat. V kapitole 5.2 jsme již popisovali intervalové odhady jednotlivých koeficientů β. V této kapitole prozkoumáme možnost konstruovat konfidenční intervaly pro celou regresní přímku a také predikční intervaly pro nové datové body.

6.1 Interval pro regresní přímku a kolem regresní přímky

V kapitole 2.2 jsme se tázali, kolik by dostal Otokar, Agáta, Uršula nebo nějaký jiný spolužák bodů ze zápočtového testu, kdyby se učili 6 hodin. Řekněme, že k odpovědi tentokrát použijeme data od osmi studentů z tabulky 1 na straně 20. Již dříve jsme odhadli regresní koeficienty, takže náš model můžeme zapsat ve tvaru

$$Y = 22.00 + 1.462 \cdot X$$

kde Y je počet získaných bodů a X počet hodin přípravy. Bodový odhad $\hat{Y}$ za předpokladu $X = 6$ je $22.00 + 1.462 \cdot 6 = 30.769$. Bez ohledu na počet regresorů bychom mohli předpis pro $\hat{Y}$ vyjádřit jakou součin dvou vektorů $\hat{Y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$, kde $\hat{\boldsymbol{\beta}}$ je vektor odhadů regresních vah $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ a $\mathbf{x}$ vektor hodnot regresorů, pro které chceme spočítat predikci. Tento vektor začíná vždy číslem 1, které patří k absolutnímu členu, v našem případě platí $\mathbf{x} = (1, 6)'$.

Náš model tedy říká, že lidé, kteří se na zápočet připravují 6 hodin, získají v průměru necelých 31 bodů. Ve skutečnosti se ale toto tvrzení opírá o bodové odhady $\hat{\boldsymbol{\beta}}$, které samy o sobě nemusí odpovídat skutečnosti, jelikož jsou zatíženy náhodným kolísáním. Pokud bychom chtěli být precizní, mohli bychom stanovit konfidenční interval, který s pravděpodobností $1 - \alpha$ (typicky 95 %) pokrývá střední hodnotu veličiny $\hat{Y}$ pro dané $\mathbf{x}$.

Tento úkol je poměrně jednoduchý, jelikož víme, že náhodná veličina $\hat{Y}$ vzniká jako součet náhodných veličin $\hat{\boldsymbol{\beta}}$ vynásobených zadanými konstantami $\mathbf{x}$. Také víme to, že odhady jednotlivých koeficientů $\hat{\beta}_j$ mají při dodržení předpokladů normální rozdělení

$N(\beta_j, \sigma_j^2)$ a především to, že rozptyly a kovariance jednotlivých regresních vah dokážeme odhadnout pomocí vztahu $\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}}) = S_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}$. V našem případě

$$\widehat{\mathbf{VAR}}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = 67.282 \cdot \begin{pmatrix}8 & 117\\ 117 & 2067\end{pmatrix}^{-1} = 67.282 \cdot \begin{pmatrix}0.726 & -0.041\\ -0.041 & 0.003\end{pmatrix} = \begin{pmatrix}48.849 & -2.765\\ -2.765 & 0.189\end{pmatrix}$$

Víme tedy, že $\widehat{\mathbf{VAR}}(\hat{\beta}_0) = 48.849$, $\widehat{\mathbf{VAR}}(\hat{\beta}_1) = 0.189$ a $\widehat{\mathbf{COV}}(\hat{\beta}_0, \hat{\beta}_0) = -2.765$. Budeme-li se tedy ptát na rozptyl $\hat{Y}$, pak se znalostí ze základních kurzů statistiky odvodíme vztah:

$$\begin{aligned}\widehat{\mathbf{VAR}}(\hat{Y}) = \widehat{\mathbf{VAR}}(1 \cdot \hat{\beta}_0) + \widehat{\mathbf{VAR}}(6 \cdot \hat{\beta}_1) + 2 \cdot \widehat{\mathbf{COV}}(1 \cdot \hat{\beta}_0, 6 \cdot \hat{\beta}_1) = \\ 1^2 \cdot \widehat{\mathbf{VAR}}(\hat{\beta}_0) + 6^2 \cdot \widehat{\mathbf{VAR}}(\hat{\beta}_1) + 1 \cdot 6 \cdot 2 \cdot \widehat{\mathbf{COV}}(\hat{\beta}_0, \hat{\beta}_1) = \\ 1 \cdot 48.849 + 36 \cdot 0.189 + 12 \cdot (-2.765) = 22.475\end{aligned}$$

Přesně stejný postup bychom mnohem snáze zapsali v maticové formě jako $\widehat{\mathbf{VAR}}(\hat{Y}) = S_\epsilon^2 \cdot \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$.

Jelikož náhodná veličina $\hat{Y}$ vzniká váženým součtem náhodných veličin s normálním rozdělením, má i ta normální rozdělení. Podobně jako při konstrukci konfidenčního intervalu pro střední hodnotu v základních kurzech jde o situaci, kdy rozptyl máme ve formě bodového odhadu a skutečnou hodnotu parametru neznáme. Budeme proto místo normálního rozdělení využívat t-rozdělení s $n - p$ stupni volnosti, kde p je počet odhadovaných parametrů (tedy zde 2).

Konfidenční interval pro střední hodnotu náhodné veličiny $\hat{Y}$ pro zadané hodnoty regresorů $\mathbf{x}$ můžeme zapsat ve tvaru

$$I_{1-\alpha} = \mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{n-p,(1-\frac{\alpha}{2})} \cdot S_\epsilon \sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

kde výraz $t_{n-p,(1-\frac{\alpha}{2})}$ značí příslušný kvantil náhodné veličiny se Studentovým t-rozdělením. Pokud dosadíme hodnoty z našeho příkladu a hladinu α nastavíme na obvyklých 5%, získáme tento výsledek:

$$I_{95\%} = 30.769 \pm 2.447 \cdot 4.741 = 30.769 \pm 11.600 = (19.169; 42.369)$$

Můžeme tedy říct, že kdybychom měli nekonečně velký soubor studentů, kteří se na test připravovali přesně 6 hodin, pak jejich průměrný počet bodů bude roven nějakému číslu, které můžeme očekávat v intervalu $(19.169; 42.369)$.

V praxi by mohlo být užitečné vypočítat konfidenční interval pro všechny časy přípravy, a vytvořit tak konfidenční interval pro každý bod regresní přímky. Toto řešení se

skutečně používá a v literatuře je najdeme pod názvem **pás spolehlivosti kolem regresní funkce**. Určité omezení tohoto postupu však plyne z faktu, že každý z vytvořených konfidenčních intervalů sice má zadanou spolehlivost, nicméně nemůžeme tvrdit, že pás, který tímto postupem vytvoříme, v $1 - \alpha$ případech regresní přímku pokryje. Opět zde narážíme na problém mnohonásobného testování – pokud má každý z vytvořených intervalů 5% pravděpodobnost, že nebude pokrývat hledanou hodnotu, pak pravděpodobnost, že selže alespoň jeden, je pochopitelně citelně vyšší.

Řešením může být využití některé z korekcí na mnohonásobné testování, a jelikož na vytvoření pásu spolehlivosti potřebujeme *nekonečno* intervalů, nabízí se využíti Scheffého přístupu, který je například oproti Bonferronimu šetrnější, když vytváříme velký počet intervalů. Princip Scheffého věty zde nebudeme odvozovat, uvedeme samotné řešení, které zajišťuje, že vytvořený pás konfidenčních intervalů s požadovanou mírou jistoty obsáhne celou regresní přímku, respektive křivku:

$$I_{1-\alpha} = \mathbf{x}'\hat{\boldsymbol{\beta}} \pm S_\epsilon \sqrt{p \cdot F_{p,n-p,(1-\alpha)} \cdot \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

kde $F_{p,n-p,(1-\alpha)}$ značí příslušný kvantil náhodné veličiny s Fisherovým rozdělením. Takto uvedenou oblast označujeme jako **pás spolehlivosti pro regresní funkci**.

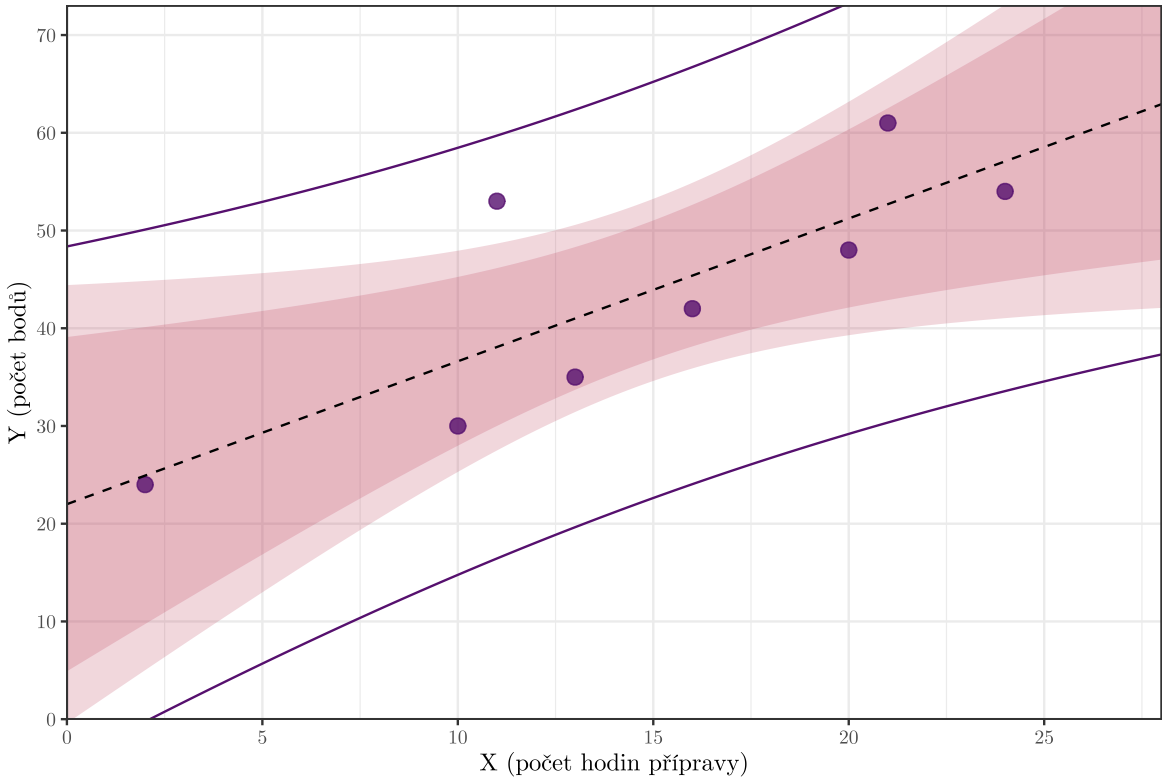
6.2 Predikční interval

Kromě hledání intervalů pro střední hodnotu veličiny $\hat{Y}$ můžeme vytvářet i takzvané predikční intervaly. Predikční interval popisuje chování budoucích měření – jde o interval, do kterého libovolné budoucí pozorování s hodnotami regresorů $\mathbf{x}$ padne s pravděpodobností $1 - \alpha$. Predikční interval konstruujeme podobným způsobem jako interval konfidenční, rozdíl je však v tom, že při predikci máme místo jednoho hned dva zdroje nepřesnosti. Budeme-li odhadovat rozptyl zkoumané náhodné veličiny, pak započteme stejně jako v předešlém případě rozptyl plynoucí z nepřesnosti odhadu regresních vah β, k němu ale přičteme ještě rozmanitost jednotlivých hodnot kolem regresní přímky, kterou charakterizuje reziduální rozptyl S_ϵ^2:

$$P_{1-\alpha} = \mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{n-p,(1-\frac{\alpha}{2})} \cdot S_\epsilon \sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

Pokud víme, že očekávaná pozorování budou měřena s jinou přesností než dosavadní, pak můžeme nahradit jedničku ve vzorci číslem, které bude vyjadřovat, kolikrát přesnější či méně přesná nová měření jsou (přesněji řečeno, kolikrát větší či menší rozptyl mají). Opět můžeme vypočítat interval pro libovolný vektor $\mathbf{x}$, čímž vytvoříme pás, do kterého jakékoli příští pozorování padne se stanovenou pravděpodobností.

Obrázek 14: Pásy pro regresní funkci, kolem regresní funkce a predikční intervaly



Tmavá plocha znázorňuje pás kolem regresní funkce a světlá pás pro regresní funkci. Fialové křivky vymezují predikční intervaly. Všechny pásy jsou nejužší nad průměrnou hodnotou regresoru.

Vrátíme-li se k otázce, kolik bodů dostaneme, pokud se budeme připravovat na zápočtový test přesně 6 hodin, pak nejlepší odpovědí bude interval $(7.59, 53.95)$, jelikož takové skóre můžeme s 95% pravděpodobností očekávat. Jak je vidět, s odhady provedenými na pouhých osmi pozorováních je náš model poměrně nespolehlivý. Grafické znázornění predikčních intervalů a pásů pro regresní přímku i kolem ní obsahuje obrázek 14.

7 Předpoklady užití lineárních modelů

Ze základních kurzů statistiky již víme, že každý statistický test je svázán určitými předpoklady, které na popisované náhodné veličiny klademe. Výjimkou nejsou ani testy v rámci lineárních modelů a určité podmínky použití má i samotná metoda nejmenších čtverců. Pokud dobře porozumíme podmínkám lineárních modelů, budeme do detailu znát i podmínky pro výpočet t-testů, analýzy rozptylu nebo testu korelačního koeficientu, jelikož, jak již víme, jde o speciální případy lineárních modelů.

Tradičně hovoříme o pěti podmínkách lineárních modelů. K těm přidáme ještě krátkou úvahu o vhodném rozsahu souboru a taky jakousi nultou podmínku, která říká, že závisle proměnná musí být kvantitativní (metrickou) proměnnou a regresory musí být kvantitativní nebo alternativní proměnné (respektive nominální, převedené na alternativní indikátory). Situace, kdy pracujeme s jinými proměnnými, lze modelovat také, nicméně tehdy bychom museli sáhnout po o poznání složitějších *nelineárních* statistických modelech.

7.1 Správnost modelu

Každý model může popsat jen takové vztahy, pro které je vytvořen. Tedy například jednoduchá regrese popisuje jen lineární vztah, ale třeba závislost ve tvaru nějaké křivky by již popsat nedokázala (toto ilustrují obrázky 5 a 6 v kapitole 4.3). Použijeme-li nevhodný model, snadno přehlédneme existující vztah, a budeme ponecháni v mylném přesvědčení, že dané regresory nijak nesouvisí se závisle proměnnou.

7.2 Nezávislost náhodné složky

Ze základních kurzů statistiky víme, že statistické testy pracují s pozorováními, která se mezi sebou neovlivňují, jsou tedy nezávislá. V kontextu lineárních modelů tuto podmínku o něco zpřesníme – budeme hovořit o nezávislosti náhodné složky. Představme si, že známe přesné hodnoty parametrů β a můžeme tedy bezchybně nakreslit regresní přímku (nebo křivku) daného modelu. Provedeme-li pozorování, naměřené hodnoty Y budou pochopitelně kolem této regresní přímky náhodně kolísat. Nezávislostí náhodné složky se rozumí to, že toto kolísání bude pro jednotlivá pozorování nezávislé. Tedy například pokud u někoho pozorujeme hodnotu vysoko nad regresní přímkou, neprozrazuje nám to nic o tom, jaký výsledek naměříme u kteréhokoli jiného pozorování[12].

[12] V některých textech tato podmínka bývá označená jako *nezávislost reziduí*. Toto označení však není zcela přesné, jelikož rezidua modelu vypočítaná jako $Y_i - \hat{Y}_i$ vždy alespoň slabě závislá jsou. Příčinou je to, že výpočet libovolného rezidua se opírá o odhady $\hat{\beta}$, které jsme však učinili s využitím všech ostatních pozorování. Takže přesná hodnota libovolného rezidua je do nějaké míry ovlivněna každým prvkem datového souboru.

Aby si čtenář lépe představil, co se podmínkou nezávislosti rozumí, uveďme několik příkladů toho, za jakých okolností dochází k jejímu porušení. Představte si, že měříte spokojenost zaměstnanců z několika týmů jednoho podniku. Lze očekávat, že pokud v rámci některého týmu panuje špatná nálada, jeho členové se jí nakazí a budou na škále spokojenosti skórovat níže. Pozorování tedy nejsou nezávislá. Abychom nezávislost obnovili, musíme informaci o tom, ve kterém je kdo týmu, do modelu zařadit – mezi regresory proto přidáme nominální proměnnou *tým*. De facto tím vyjadřujeme předpoklad, že každý tým má svou náladu, která přičítá či odečítá nějaký pevný počet bodů od výsledku každého svého člena[13].

Jiný příklad porušení podmínky nezávislosti může být situace, kdy Uršula z prvního příkladu v těchto skriptech skutečně od Agáty opisovala. Tehdy jsou výsledky obou studentek opět nějakým způsobem svázané. Ještě jeden příklad: zkoumáme výskyt určitých komunikačních vzorců u mužů a žen v partnerském vztahu. Datový soubor je tvořen jednotlivci, ale my přehlédneme fakt, že někteří účastníci výzkumu spolu tvoří partnerský pár. Pozorování v rámci těchto dvojic se opět vzájemně ovlivňují.

Dodejme, že nezávislost náhodné složky je poměrně zásadní a přehlédnutí vzájemně provázaných skupin pozorování obvykle vyústí v řadu falešně pozitivních zjištění.

7.3 Nepřítomnost kolinearity

Pojmem kolinearita se rozumí vzájemná korelace regresorů. Obecně mohou být regresory korelované, ovšem pouze do nějaké rozumné míry. Váhy vysoce korelovaných regresorů (řekněme $|r| > 0.9$) se obtížně odhadují a přináší velkou chybovost, jelikož v rámci statistického modelu se jen těžko rozlišuje, který ze dvou téměř identických regresorů hraje roli. Pokud by se stalo, že některé dva regresory jsou dokonale korelované ($|r| = 1.0$), pak odhady s pomocí metody nejmenších čtverců vůbec nelze provést (při výpočtu by vyšla matice $\mathbf{X}'\mathbf{X}$ singulární, a nelze k ní tedy stanovit inverzi). Statistický program by nás nejspíš upozornil nějakým chybovým hlášením nebo by automaticky některý regresor z výpočtu vyřadil.

Samotná kolinearita je poměrně dobře odhalitelný problém – stačí si zobrazit korelační matici jednotlivých regresorů a snadno zjistíme, které regresory v modelu korelují. O poznání větší oříšek je ale takzvaná **multikolinearita**. O multikolinearitě mluvíme tehdy, když v modelu existuje skupina regresorů, z nichž lze vytvořit taková lineární kombinace, která vysoce koreluje s jiným regresorem. Opět při korelaci 1.0 je úloha neřešitelná. Ty-

[13] Toto elegantní řešení bychom dokonce mohli využít i tehdy, když budeme provádět u každého respondenta opakovaná měření. Řádky datové tabulky patřící témuž člověku jsou pochopitelně závislé, nezávislost je však obnovena, pokud do modelu vložíme regresor *respondent*. Toto řešení nás však obvykle dovede k využití takzvaných náhodných faktorů (viz kapitola 15 o modelech se smíšenými efekty), případně o poznání vzácněji k zobecněné metodě nejmenších čtverců (viz kapitola 13.1).

picky na multikolinearitu naráží studenti, kteří při přípravě indikátorových proměnných zapomenou na to, že jich má být o jednu méně, než má daná nominální proměnná úrovní, a zakódují všechny úrovně včetně té referenční. Pokud žádný indikátor nevynecháme, pak platí, že korelace mezi libovolným indikátorem a součtem zbývajících indikátorů je -1.0, a statistický program vyprodukuje upozornění místo výsledku.

Multikolinearitu nelze odhalit pohledem na korelační matici regresorů, ale k jejímu zjištění používáme statistiky jménem *variance inflation factor* (VIF) a *tolerance*. Toleranci lze vypočítat pro libovolný regresor. Je to číslo mezi 0 a 1, které říká, kolik unikátního rozptylu (nesdíleného s jinými regresory) daný regresor obsahuje. Vypočítá se opět pomocí lineárního modelu, tentokrát za závisle proměnnou však zvolíme zkoumaný regresor a jako nezávisle proměnné zbývající regresory. Tolerance pak odpovídá nevysvětlenému rozptylu, tedy $1 - R^2$. VIF je jednoduše převrácená hodnota tolerance, tedy $\frac{1}{\text{tolerance}}$.

VIF říká, kolikrát se zvětšil rozptyl odhadu váhy daného regresoru v porovnání se situací, kdy by daný regresor s žádným jiným regresorem nekoreloval. Je-li tedy VIF třeba 4, je rozptyl odhadu čtyřnásobný, tedy směrodatná odchylka, respektive šířka konfidenčního intervalu pro parametr β, dvojnásobná. Za problém je obvykle považována hodnota VIF vyšší než 5, případně 10 (tedy nejméně 10–20 % unikátního rozptylu). Pokud najdeme v našem modelu regresor, který tuto podmínku porušuje, měli bychom se zamyslet nad tím, jestli některé závisle proměnné neříkají téměř totéž, a nejsou tedy zbytečné.

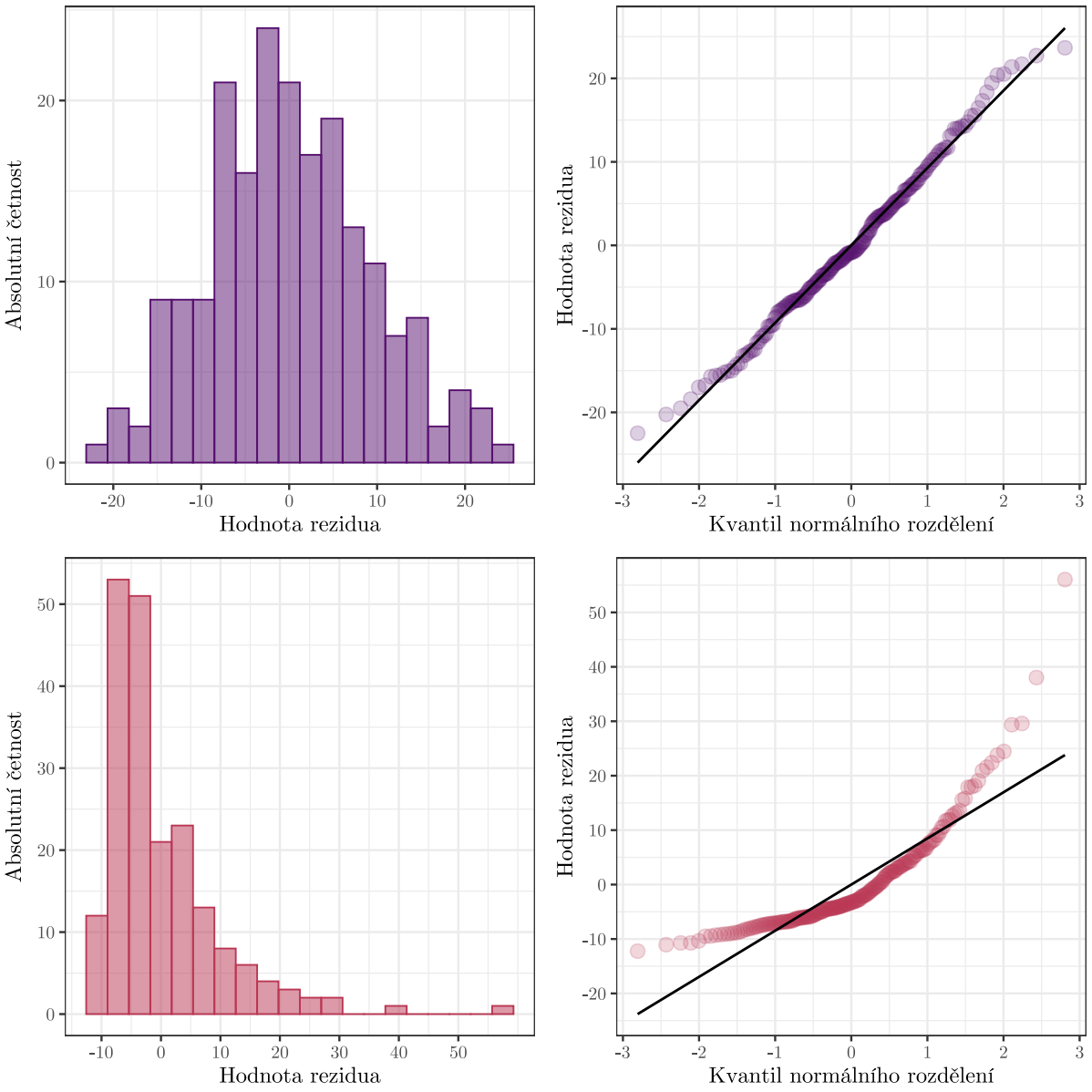
7.4 Normální rozdělení reziduí

Podmínka normálního rozdělení provázela všechny parametrické testy ze základních kurzů statistiky. Není proto překvapením, že tato podmínka je přítomna i u lineárních modelů. Překvapivé je však možná to, že i kdyby náhodná veličina Y ani žádný z regresorů normální rozdělení neměly, nemusí nutně být tato podmínka porušena. Normální rozdělení vyžadujeme jen a pouze u reziduí modelu.

Po vytvoření modelu je proto užitečné si zobrazit histogram reziduí nebo jejich Q-Q graf (viz níže) a zvážit, jestli se nalezený tvar od normálního rozdělení neodlišuje příliš zásadním způsobem. Někdy může být výhodné použít i testy normality, jejichž použití je nicméně zatíženo logickou chybou, kterou známe ze základních kurzů. Pokud pracujeme s malým rozsahem souboru, test normality má malou sílu a pravděpodobně signifikantní odchýlení nenajde. Je-li rozsah souboru v řádu stovek pozorování, je síla statistického testu vysoká a i nepatrná odchylka od normálního rozdělení povede k nalezení velmi nízké p-hodnoty. Toto je v ostrém kontrastu s faktem, že statistické testy, které provádíme v rámci lineárních modelů, se stávají s rostoucím rozsahem souboru vysoce robustní proti porušení podmínky normálního rozdělení. Pracujeme-li tedy s mnoha desítkami či stovkami pozorování, můžou být výsledky modelu relevantní, přestože testy normality hlásí vážné porušení podmínky.

Většina statistických programů kromě možnosti zobrazit histogram reziduí nabízí i takzvaný Q-Q graf (*Q-Q plot*, *quantile-quantile plot*). Q-Q graf obecně slouží ke srovnání tvaru rozdělení pozorovaného u dvou souborů nebo mezi souborem a zadanou distribucí. V našem případě srovnáváme rozdělení reziduí s normálním rozdělením. Při konstrukci grafu vypočítáme, jakým výběrovým kvantilům odpovídají jednotlivá rezidua (získáme tedy sadu hodnot α_1 až α_n ležících mezi nulou a jedničkou). Pro každé z těchto čísel najdeme hodnotu kvantilu normovaného normálního rozdělení ϕ_α. Nakonec vytvoříme bodový graf, kdy na jednu z os vynášíme hodnoty reziduí a na druhou hodnoty příslušných kvantilů normovaného normálního rozdělení ϕ_α. Pokud rozdělení reziduí připomíná normální rozdělení, bude mít vzniklý obrazec přibližně tvar rovné čáry táhnoucí se od levého spodního rohu grafu po pravý horní. Pokud rezidua normální rozdělení nevykazují, bude vzniklý obrazec tvořen nějak prohnutou křivkou. Dodejme, že čtení Q-Q grafů není tak přímočaré jako čtení histogramu a vyžaduje víc zkušeností. Pro srovnání vizte histogramy reziduí a příslušné Q-Q grafy na obrázku 15.

Obrázek 15: Histogramy reziduí a Q-Q grafy



Horní dva grafy vychází z dat, jejichž rozdělení se blíží normálnímu rozdělení. Spodní dva pochází z kladně zešikmeného rozdělení a podmínka normality je tedy porušena.

7.5 Homoskedasticita

Homoskedasticita je označení pro homogenitu rozptylu reziduí. O homoskedasticitě mluvíme, když pro libovolný regresor platí to, že rezidua mají stejný rozptyl pro všechny úrovně (respektive hodnoty) tohoto regresoru. Pokud by tedy regresorem bylo pohlaví, pak by zde byl požadavek, aby byl reziduální rozptyl u mužů i u žen stejný. Kdyby regresorem byla inteligence, požadavek by se týkal neměnného reziduálního rozptylu napříč všemi hodnotami IQ od nízkého až po vysoké. Opakem homoskedasticity je heteroskedasticita.

Nejsnazší kontrolu přítomnosti heteroskedasticity můžeme provést pomocí bodového grafu. Na osu x vyneseme hodnoty příslušného regresoru jednotlivých měření, na osu y jejich rezidua. Tento proces bychom museli opakovat pro každý regresor, což u větších modelů může být zdlouhavé, proto často volíme jednodušší cestu, kdy na osu x vynášíme hodnoty predikcí $\hat{Y}$ a na osu y rezidua (případně standardizovaná rezidua). Ať už použijeme jakýkoli přístup, v grafu by neměly být vidět žádné smysluplné útvary a měl by připomínat obrázek 16a. Typickým příkladem heteroskedasticity je obrázek 16b. Obrázek 16c vypovídá taky o heteroskedasticitě, na první pohled zde navíc dokážeme říct, že její příčinou je opomenutí nějakého zásadního vlivu, který na proměnnou Y působí, ale my jsme jej nezahrnuli do modelu.

Na rozdíl od podmínky normálního rozdělení škodlivý vliv heteroskedasticity neslábne s rostoucím rozsahem výběru. Pokud na tento problém narazíme, měli bychom se zamyslet nad tím, zda jsme použili vhodný model, který nepotřebuje nějaká vylepšení. Kromě grafické kontroly existují i statistické testy homoskedasticity. Jejich použití je však zatíženo stejným problémem, na který jsme naráželi u použití F testu před výběrem vhodného t-testu.

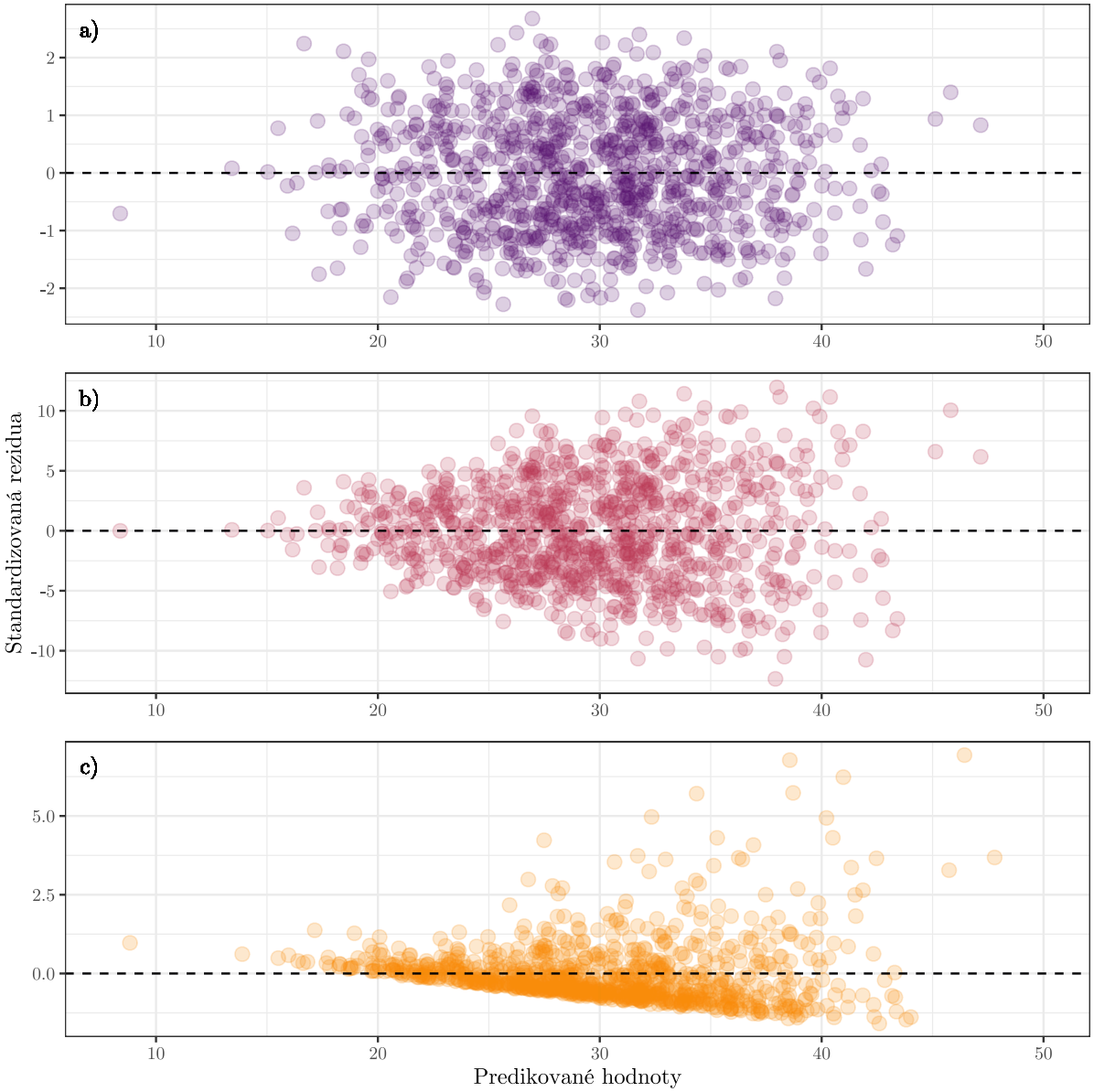
7.6 Poznámka k rozsahu souboru

Při práci s lineárními modely často narážíme na otázku, jaký je nejmenší počet pozorování, abychom vůbec mohli naučené postupy uplatnit. Na tuto otázku neexistuje jednoznačná odpověď, jelikož ani samotná otázka není položena jednoznačně.

Co znamená dostatečný rozsah souboru? Postupy, které jsme se naučili v rámci lineárních modelů, nejsou asymptotické, a fungují tedy pro téměř libovolně malý rozsah souboru. Metodu nejmenších čtverců lze použít, jakmile máme alespoň tolik pozorování, kolik odhadujeme parametrů. Pokud máme víc pozorování než parametrů modelu, můžeme odhadovat rozptyl odhadů regresních vah, a provádět tedy libovolné testy statistické významnosti.

Existují nicméně dva důvody, proč bychom uvítali poněkud větší rozsah souboru než $p + 1$. První je ten, že soubor malého rozsahu není robustní vůči porušení normality reziduí a malý počet datových bodů nám ani neumožní si udělat obrázek o tom, jestli tato

Obrázek 16: Homoskedasticita a heteroskedasticita



podmínka platí, nebo ne. Druhým důvodem je statistická síla. Při malém rozsahu souboru mají testy v rámci lineárních modelů jen malou šanci zamítnout nulovou hypotézu, respektive, pokud bychom spočítali konfidenční intervaly odhadů regresních vah, zjistili bychom, že jsou tak široké, že v podstatě neříkají vůbec nic.

Z tohoto důvodu v literatuře narážíme na jakási hrubá pravidla, která nám mají pomoci stanovit počet pozorování potřebný k tomu, aby nalezené odhady měly nějaký smysl. Tato pravidla obvykle odvozují rozsah souboru podle počtu regresorů, které model obsahuje (značíme k). V psychologii se autoři nejčastěji odvolávají na následující doporučení[14]:

- $n \geq 104 + k$ pro test hypotézy o rozdílu R^2 od nuly,
- $n \geq 50 + 8k$ pro testy jednotlivých regresorů.

Dodejme, že toto doporučení je jen velmi hrubým odhadem. Kromě počtu regresorů taky záleží na tom, do jaké míry jsou tyto regresory korelované (vysoká korelovanost výrazně oslabuje testy), a zejména pak na tom, jak velké efekty se pokoušíme odhalit. Nejvhodnější by bylo využít postupy analýzy síly testu, ty jsou však v případě lineárních modelů obtížně proveditelné, jelikož jen těžko dopředu odhadneme korelační strukturu všech zařazených proměnných. Navržený postup selhává u velmi jednoduchých modelů (například pro t-test pro dva nezávislé výběry požaduje 105 nebo 58 pozorování, což je nesmyslně vysoká hranice), ale i u modelů velmi složitých, kde potřebný počet pozorování spíše podceňuje (například pro práci s modelem s 20 regresory si vystačí s nějakými 124 nebo 210 pozorováními, což bychom v praxi za uspokojivý rozsah nejspíš nepovažovali).

Odpověď je tedy taková, že u jednoduchých modelů můžeme pracovat se soubory o několika málo desítkách pozorování, ale musíme mít důvod se domnívat, že zde jsou splněny předpoklady. U velmi složitých modelů bychom pak měli požadovat několikasethlavý výzkumný soubor. Pro ostatní případy nám můžou pomoci výše uvedená pravidla, berme je však spíše jako orientační doporučení.

[14] Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate behavioral research*, *26*(3), 499–510.

8 Detekce problematických pozorování

Chceme-li mít důvěru ve výsledky, které jsme získali s pomocí statistického modelu, obvykle po odhadnutí parametrů učiníme několik kroků, kterým se souhrnně říká *diagnostika modelu*. Mezi tyto kroky by patřilo jednak prozkoumání rozdělení reziduí, ověření nepřítomnosti heteroskedasticity a vypočítání VIF pro jednotlivé regresory. Tyto kroky již známe z předešlé kapitoly.

Vedle toho je často užitečné věnovat pozornost přítomnosti pozorování, která jsou v rámci našeho modelu něčím netypická a vybočují od ostatních. Obvykle jde o zkoumání přítomnosti různých druhů odlehlých či vlivných pozorování. Existuje nespočet statistik, které používáme k popisu chování jednotlivých pozorování. Při diagnostice modelu zpravidla nebudeme pracovat se všemi, ale spokojíme se s jedním či dvěma ukazateli, kterým věnujeme pozornost. S několika z nich, které se používají poměrně často, se seznámíme.

8.1 Rezidua (*raw residuals*)

Jedna ze základních pouček parametrické statistiky říká, že odlehlá pozorování můžou mít citelný vliv na věrohodnost výsledku. Identifikovat odlehlé pozorování (tzv. outlier) při práci s jednou či dvěma proměnnými je poměrně snadný úkol. V případě komplexnějších designů však můžeme narazit na situaci, kdy pozorování není outlierem v žádném sledovaném znaku, ale přesto je v nápadném rozporu s modelem. Naopak můžou existovat pozorování, která nápadně vybočují třeba hned v několika znacích, a přesto jsou s modelem v dobrém souladu. Při hledání pozorování, která se neshodují s tím, co model očekává, může být užitečné ověřit přítomnost outlierů v reziduích. Extrémně vysoké nebo extrémně nízké hodnoty nás můžou upozornit na pozorování, která jsou v nějakém smyslu problematická (například jsme při přepisování výsledků do datové matice udělali chybu).

Pro grafickou prezentaci se někdy používá místo původních reziduí jejich absolutní hodnota nebo odmocnina z absolutní hodnoty. Výhoda této úpravy je, že pak stačí sledovat nejvyšší hodnoty, informační přínos se ale nezmění. Někdy se také prezentuje hodnota reziduí vydělených odhadem směrodatné odchylky σ_ϵ.

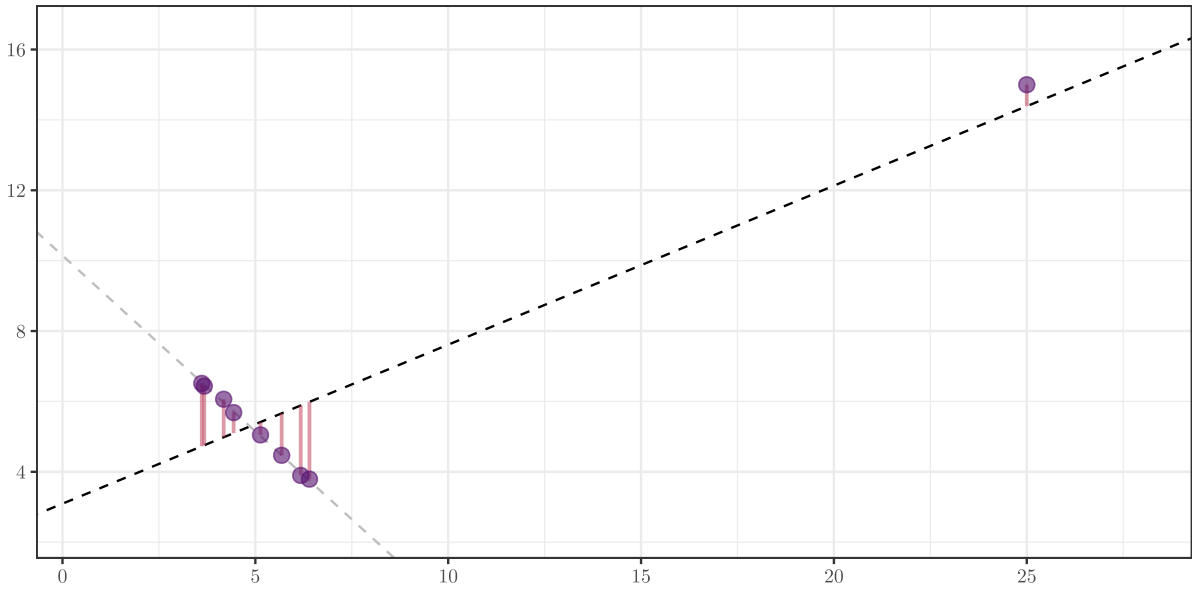
8.2 Rezidua s odstraněným pozorováním (*deleted residuals*)

I přes výše uvedené výhody nejsou ani rezidua schopna nás upozornit na každé problematické pozorování. Obrázek 17 demonstruje situaci, kdy je v malém datovém setu přítomno jedno odlehlé pozorování, které natolik změnilo výsledky, že jeho reziduum je jedno z nejbližších nule. Prozkoumat samotná rezidua by nám v tomto případě nepomohlo. Jedním z přístupů, jak tuto situaci vyřešit, je spočítat takzvaná *deleted residuals*. Opět jde o kla-

sická rezidua, jak je známe, tentokrát však při výpočtu každého z nich použijeme odhad parametrů udělaný na všech datových bodech vyjma toho, jehož reziduum počítáme. Pokud tedy počítáme tuto statistiku pro i-té pozorování, odhadneme hodnoty β pomocí celého datového souboru bez i-tého pozorování, ze získaných vah uděláme predikci pro i-té pozorování a spočítáme jeho reziduum.

Pro srovnání reziduí a reziduí po odstranění daného pozorování spočítáme jednoduše jejich rozdíl. V literatuře tuto statistiku najdeme někdy pod názvem DFFIT.

Obrázek 17: Vlivné pozorování



8.3 Vliv pozorování (*leverage*)

Problém vlivných pozorování řeší i statistika s názvem **leverage** (v doslovném překladu páka, pákový bod, volněji vliv pozorování). Jedná se již o něco sofistikovanější statistiku, pro jejíž pochopení se musíme nejprve seznámit s takzvanou **projekční maticí**.

Projekční matice $\mathbf{H}$ (*hat matrix*) je pozoruhodným konceptem s řadou překvapivých vlastností[15]. Její podobu lze odvodit ze dvou vztahů, které již známe:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad\qquad \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

[15] Jedna z nich je například idempotence. Pokud idempotentní matici $\mathbf{H}$ vynásobíme samu sebou, pak zase získáme matici $\mathbf{H}$, tedy $\mathbf{HH} = \mathbf{H}$.

Když dosadíme $\hat{\beta}$ z první rovnice do druhé, získáme vztah

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

který přepíšeme jako

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \qquad \text{kde} \qquad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Projekční matice $\mathbf{H}$ má rozměry $n \times n$. Jak říká vztah výše, pokud sloupec (vektor) pozorování $\mathbf{Y}$ vynásobíme maticí $\mathbf{H}$, získáme predikce $\hat{\mathbf{Y}}$. Čtenář seznámený s principem maticového násobení si může všimnout mechanismu, jak jsou jednotlivé predikce pomocí matice $\mathbf{H}$ počítány. Pokud počítáme predikci pro i-té pozorování, vezmeme si i-tý řádek matice $\mathbf{H}$ a jeho první prvek vynásobíme hodnotou Y_1, druhý prvek hodnotou Y_2 a tak dál včetně i-tého prvku, který násobíme hodnotou Y_i. Všechny získané výsledky pak sečteme. Všimněte si, že i-tý prvek z i-tého řádku (tedy jakýkoli diagonální prvek) vždy říká, do jaké míry hodnota i-tého pozorování (Y_i) ovlivňuje svou vlastní predikci ($\hat{Y}_i$). Jinými slovy, jak moc je schopno dané pozorování ovlivnit výsledek ve svůj prospěch.

Tyto hodnoty diagonály matice $\mathbf{H}$ (značíme je h_i) jsou právě označovány pojmem *leverage*. Tato statistika vždy leží mezi 0 a 1 (obvykle mnohem blíže nule než jedničce) a vysoké hodnoty nám pomůžou identifikovat vlivná pozorování.

8.4 Standardizovaná rezidua (*standardized residuals*)

Statistika h_i má kromě identifikace vlivu pozorování využití při výpočtu standardizovaných reziduí. Pokud bychom chtěli standardizovat rezidua, asi by nás napadlo hodnotu každého z nich vydělit odhadnutou směrodatnou chybou modelu S_ϵ. Tento přístup, ač má svou logiku, přehlíží jeden fakt: přestože náhodná složka modelu má napříč pozorováními stejný rozptyl, rozptyl jednotlivých reziduí se pro jednotlivá pozorování může značně lišit. Rezidua vlivných pozorování (typicky pozorování s okrajovými hodnotami některého z regresorů) mají menší rozptyl než méně vlivná rezidua. Rozptyl rezidua ϵ_i lze odhadnout jako

$$\widehat{\mathbf{VAR}}(\epsilon_i) = (1 - h_i)S_\epsilon^2$$

kde h_i je vliv (*leverage*) daného pozorování popsaný v předešlých odstavcích. Vydělíme-li hodnoty jednotlivých reziduí odmocninou rozptylu jejich odhadů, získáme standardizovaná rezidua. Ač se standardizovaná rezidua obvykle příliš neliší od reziduí převedených na z-skór (tzn. vydělených hodnotou S_ϵ), jde jistě o přesnější odpověď na otázku, které pozorování vybočuje více než jiné. Standardizovaná rezidua se někdy označují jako vnitřně studentizovaná (*internally studentized*).

8.5 Studentizovaná rezidua (*studentized residuals*)

Předešlé úvahy o standardizaci reziduí lze propojit s úvahou o postupném vynechávání jednotlivých pozorování při výpočtu jednotlivých reziduí. V případě, že uplatníme postup popsaný v kapitole o standardizaci reziduí, ale při výpočtu i-tého rezidua vynecháme i-tý prvek, pak hovoříme o takzvaných studentizovaných reziduích (přesněji navenek studentizovaných, *externally studentized*).

Studentizovaná rezidua mají při dodržení podmínek Studentovo t-rozdělení s $n-p-1$ stupni volnosti, hodí se proto pro provádění nejrůznějších statistických testů.

8.6 Cookova vzdálenost (*Cook's distance*)

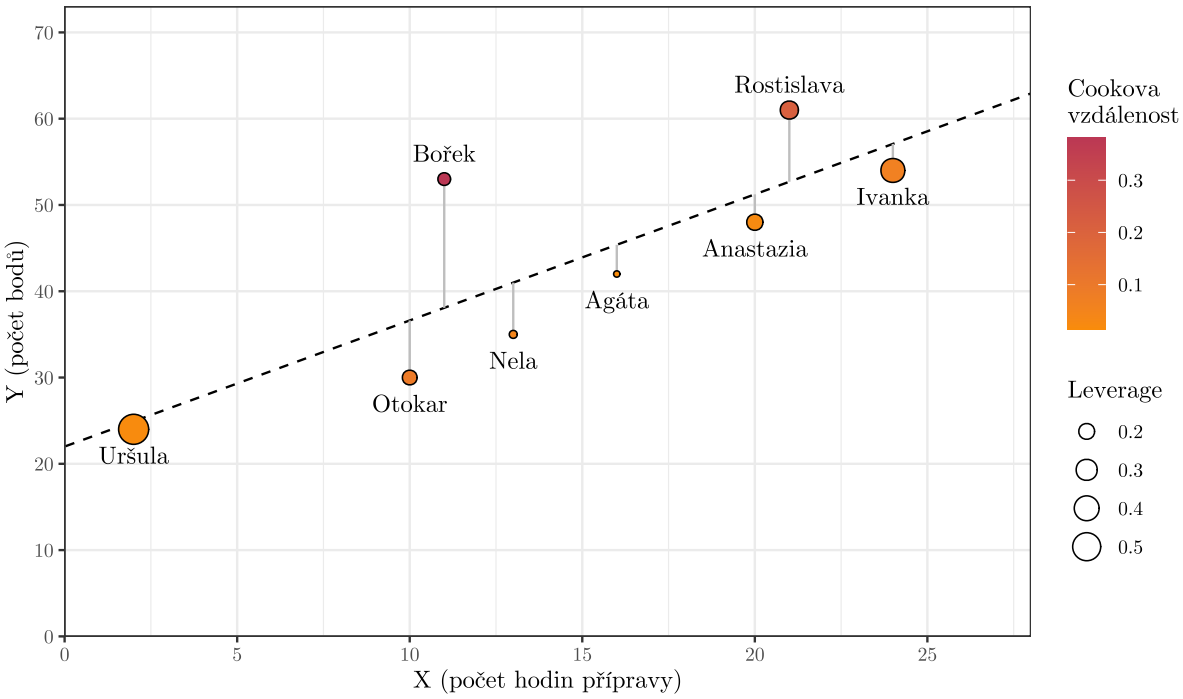
Podobný problém jako *leverage* řeší Cookova vzdálenost. Tentokrát však nehodnotíme pouze to, do jaké míry dané pozorování ovlivňuje predikci své vlastní hodnoty, ale do jaké míry ovlivňuje predikce všech n pozorování. Pro výpočet tedy kromě predikcí $\hat{Y}_i$ potřebujeme i predikce získané, pokud bychom k výpočtu regresních vah použili všechny pozorování kromě prvního (značme $\hat{Y}_{i(1)}$), kromě druhého ($\hat{Y}_{i(2)}$) a tak dál až kromě posledního ($\hat{Y}_{i(n)}$). Cookovu vzdálenost pro j-té pozorování pak spočítáme jako

$$D_j = \frac{1}{pS_\epsilon^2} \sum_{i=1}^{n} \left(\hat{Y}_i - \hat{Y}_{i(j)}\right)^2$$

Cookova vzdálenost má při dodržení předpokladů Fisherovo F rozdělení s p a $n-p$ stupni volnosti.

Diagnostiku vlivných pozorování můžeme demonstrovat na souboru Otokara, Agáty a dalších šesti spolužáků z příkladu v kapitole 2. Graf 18 znázorňuje leverage a Cookovu vzdálenost každého pozorování. Tytéž a další údaje jsou shrnuty v tabulce 3. Jak je patrné, nejúčinněji si regresní přímku k sobě přitáhne Uršula, která nám poskytuje vzácnou informaci o tom, co se stane, když jdeme k zápočtu nenaučení. Na druhé straně to jsou pak tři studentky, které se učily 20 a více hodin. Pokud bychom s pomocí Cookových vzdáleností hodnotili, kdo model ovlivňuje nejvíce, zjistili bychom, že zde vyčnívá Bořek s Rostislavou, kteří sice nemají zásadní vliv na sklon přímky, ale citelně ji posouvají nahoru.

Obrázek 18: Vlivná pozorování



Tabulka 9: Diagnostika modelu

Student	Y	X	Reziduum	Stand. reziduum	Stud. reziduum	Leverage	Cookova vzdálenost
Agáta	42	16	−3.38	−0.44	−0.41	0.13	0.01
Otokar	30	10	−6.62	−0.89	−0.88	0.19	0.09
Uršula	24	2	−0.92	−0.17	−0.16	0.57	0.02
Bořek	53	11	14.92	1.99	3.10	0.16	0.38
Ivanka	54	24	−3.08	−0.47	−0.44	0.37	0.07
Anastazia	48	20	−3.23	−0.44	−0.41	0.21	0.03
Nela	35	13	−6.00	−0.79	−0.76	0.13	0.05
Rostislava	61	21	8.31	1.16	1.20	0.24	0.21

9 Transformace závisle proměnné

V kapitole 7.4 jsme konstatovali, že ani velké zešikmení závisle proměnné nemusí nutně znamenat, že rezidua modelu budou porušovat předpoklad normálního rozdělení. Toto tvrzení je pravda v teorii, v praxi ale zjistíme, že ve většině případů, kdy pracujeme s vysoce zešikmenou závisle proměnnou, toto zešikmení zůstává přítomno i v reziduích, ba co víc, často bude doprovázené heteroskedasticitou.

Pokud budeme tento problém konzultovat se statistikem, nejspíš navrhne opustit svět lineárních modelů a pokusí se připravit model nelineární, který bude podivnému rozdělení našich proměnných ušitý na míru. Na naší úrovni poznání nicméně takovými nástroji vybaveni nejsme, budeme se proto muset smířit s nějakým trikem, který nám umožní si vystačit se znalostmi, jež máme. Tímto trikem může být právě vhodně zvolená transformace závisle proměnné Y.

Transformací se rozumí to, že na proměnnou Y aplikujeme nějakou funkci (která je monotonní, byť nelineární). V literatuře můžeme narazit na funkce, jako je například $\sqrt{Y}$, $1/Y$ nebo $\log(Y)$. Tuto upravenou závisle proměnnou pak obvyklým způsobem vložíme do lineárního modelu, který s trochou štěstí již bude splňovat podmínky normálního rozdělení reziduí a jejich homoskedasticity. Pokud bychom pak chtěli pomocí tohoto modelu provádět predikce, na výsledek (ať už bodový nebo na meze intervalového odhadu) uplatníme naši transformaci naopak (inverzně), abychom se vrátili k původním jednotkám, ve kterých bylo měřeno Y. Pro tři výše uvedené transformace by to bylo $\hat{Y}^2$, $1/\hat{Y}$ a $\exp(\hat{Y})$.

Jedním z hlavních důvodů, proč transformaci závisle proměnné nelze paušálně doporučit a proč by ji mnoho statistiků označilo za neelegantní, případně až barbarskou, je ten, že většina transformací znemožní interpretovat nalezené regresní koeficienty. Ty totiž popisují chování transformované veličiny, která se často vzdaluje představitelné realitě (co například znamená tvrzení, že se *odmocnina počtu symptomů zvýší o půl bodu?*). Mnohdy nám tak nezbude než rezignovat na přesný popis vztahů mezi proměnnými, ale jen konstatovat statistickou významnost, kterou případně doplníme standardizovanými regresními koeficienty k dokreslení míry účinku.

9.1 Log-normální regrese

Existují transformace závisle proměnné, které výše uvedeným neduhem netrpí. Nějakým způsobem sice změní způsob, jak o modelu budeme přemýšlet, ale možnost interpretace získaných koeficientů neztratíme. Asi nejoblíbenější z těchto modelů s transformovanou závislou veličinou je tak zvaná log-normální regrese.

Log-normální regresí se rozumí lineární model, který pracuje s logaritmem závisle proměnné[16]:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Výše uvedený předpis nám sice umožňuje pomocí metody nejmenších čtverců získat odhady parametrů β, ale neprozrazuje, jak o těchto číslech přemýšlet. Dle známých pouček bychom mohli uvažovat, že β_1 říká, o kolik bodů vzroste logaritmus veličiny Y, když regresor X_1 vzroste o jeden bod. To je pravdivé, avšak krajně neuspokojivé tvrzení. Abychom odhalili skutečný smysl výsledků modelu, vzpomeňme si na pravidla počítání s logaritmy a s exponenciální funkcí a přepišme rovnici modelu do poněkud odlišné podoby.

Začněme prvním krokem, kdy obě strany rovnice exponenciálně transformujeme. Pamatujme, že exponenciální a logaritmická transformace jsou protiklady (takzvané inverzní funkce), a tedy $\exp(\log(Y)) = Y$. Tedy:

$$Y = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon)$$

Dále víme, že platí vztah $\exp(A + B) = \exp(A) \cdot \exp(B)$. Všimněte si, že znaménko plus se mění na znaménko krát. Náš model proto přepíšeme do podoby

$$Y = \exp(\beta_0) \cdot \exp(\beta_1 X_1) \cdot \exp(\beta_2 X_2) \cdot \ldots \cdot \exp(\beta_k X_k) \cdot \exp(\epsilon)$$

Nakonec uplatníme poznatek $\exp(AB) = \exp(A)^B$:

$$Y = \exp(\beta_0) \cdot \exp(\beta_1)^{X_1} \cdot \exp(\beta_2)^{X_2} \cdot \ldots \cdot \exp(\beta_k)^{X_k} \cdot \exp(\epsilon)$$

Než se zamyslíme nad tím, co nám tato rovnice prozrazuje, uvědomme si, že náš model již nemá aditivní, ale multiplikativní podobu. Tedy všude, kde jsme se byli zvyklí tázat *o kolik*, budeme teď pokládat otázku *kolikrát*. Taky v našich úvahách již nebudeme pracovat se samotnými odhady $\hat{\beta}$, ale s jejich exponenciálně transformovanou podobou $\exp(\hat{\beta})$. To však nepředstavuje žádný problém, jelikož hodnoty $\hat{\beta}$ jsme odhadli obvyklým způsobem již dříve a exponenciální transformaci dokážeme snadno provést s pomocí kalkulačky nebo tabulkového editoru. Pod $\exp(\hat{\beta})$ si tedy můžeme představit nějaká konkrétní nám známá čísla.

Interpretace hodnoty $\exp(\hat{\beta}_0)$ je analogická k interpretaci absolutního členu v klasickém lineárním modelu. Jde o **očekávanou hodnotu proměnné Y za předpokladu, že se všechny regresory X_1 až X_k rovnají nule**. Vyplývá to z faktu, že jakékoli číslo umocněné na nultou se rovná jedné a $\exp(\hat{\beta}_0) \cdot 1$ se opět rovná $\exp(\hat{\beta}_0)$. Odhady $\exp(\hat{\beta}_1)$ až

[16] Označením log v těchto skriptech myslíme přirozený logaritmus, tedy logaritmus o základu $e \approx 2.718$. Pokud bychom se rozhodli zvolit jiný základ, například 10, musíme jím Eulerovo číslo e nahradit ve všech výpočtech.

$\exp(\hat{\beta}_k)$ pak říkají, **kolikrát v průměru vzroste hodnota $\hat{Y}$, když hodnota daného regresoru vzroste o jedničku**.

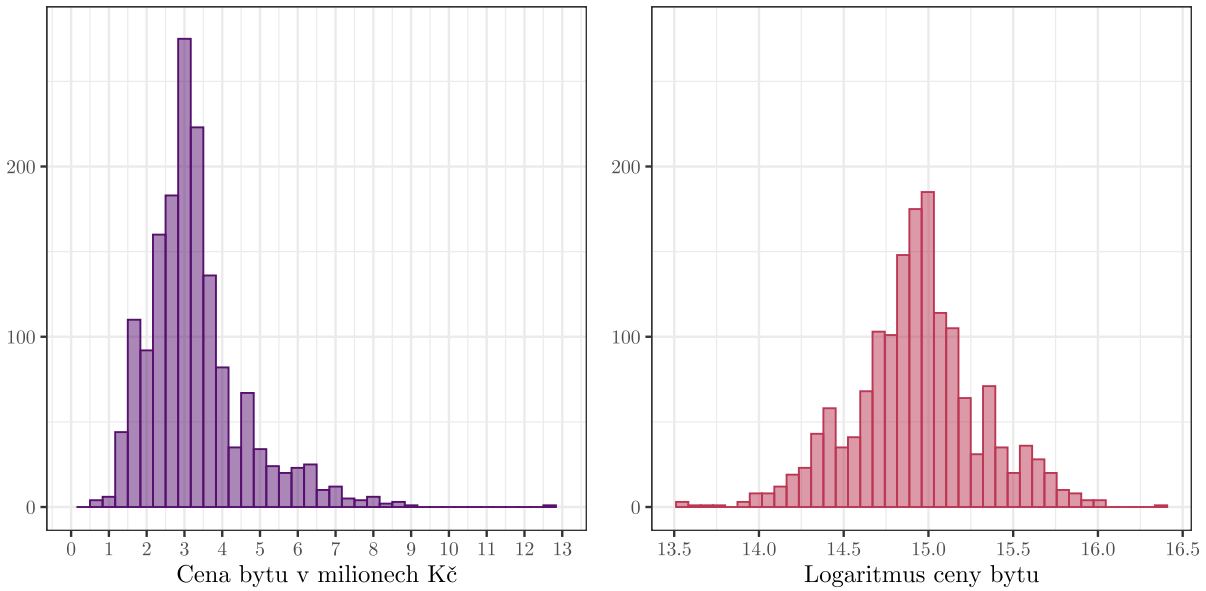
Záměna *o kolik* za *kolikrát* může být ze začátku matoucí. Demonstrujme si proto popsaný postup na příkladu. Zvolme tentokrát téma mimo obor psychologie.

Představme si, že jsme v létě roku 2019 koupili byt v Olomouci. Můžeme se tázat, jestli to byla ve srovnání s jinými byty v tomtéž městě dobrá koupě, nebo naopak jsme prodejci zaplatili více, než co odpovídá běžné ceně bytu. Za tímto účelem jsme z webových stránek největšího prodejce nemovitostí zkopírovali všechny nabídky prodeje bytů v Olomouci, které byly k dispozici od července do listopadu. Celkem jsme tak získali bezmála 1600 záznamů.

Cena bytu je dána mnoha faktory. Pro jednoduchost příkladu věnujme pozornost jen těmto: podlahová plocha, počet pokojů, samostatná kuchyň (+1 nebo +kk), třída energetické náročnosti, stav bytu a přítomnost balkonu nebo terasy.

Daný příklad bychom mohli řešit jako klasickou regresi nebo jako regresi log-normální. Histogram proměnné cena (viz obrázek 19) naznačuje, že závisle proměnná sice vykazuje kladné zešikmení, ale ne tak zásadní, aby vzhledem k rozsáhlému souboru zneplatnilo výsledky statistických testů. Obě varianty toho, jak situaci modelovat, jsou odůvodnitelné. Při rozhodnutí zvažme, která z těchto vět nám dává větší smysl: *Za balkon si v průměru připlatíte 140 000 Kč* nebo *Balkon zvyšuje cenu bytu v průměru o 5 %*. Přikloňme se k druhé variantě (ač i ta první může dávat smysl), a zvolme tedy log-normální regresi.

Obrázek 19: Histogram ceny bytu a jejího logaritmu



Uvedené hodnoty jsou skutečné záznamy z webu s realitami, nikoli simulovaná data.

Tabulka 10: Odhady koeficientů log-normálního modelu

Regresor	Efekt	$exp(\hat{\beta})$	$\hat{\beta}$	$t(1575)$	p-hodnota
Počátek		2 212 135 Kč	14.61	574.36	< 0.001
Podlahová plocha	7 %	1.07	0.07	25.46	< 0.001
Samostatná kuchyň	−12 %	0.88	−0.13	−7.02	< 0.001
Má balkon / terasu	5 %	1.05	0.05	4.80	< 0.001
1 pokoj	−22 %	0.78	−0.24	−13.92	< 0.001
2 pokoje	0 %				(ref)
3 pokoje	17 %	1.17	0.16	12.78	< 0.001
4 a více pokojů	18 %	1.18	0.17	6.81	< 0.001
Energeticky nenáročný (A–B)	6 %	1.06	0.05	2.19	0.029
Energeticky středně náročný (C–F)	0 %				(ref)
Energeticky náročný (G)	−2 %	0.98	−0.02	−0.98	0.326
Novostavba / projekt	27 %	1.27	0.24	7.68	< 0.001
Velmi dobrý / rekonstruovaný	18 %	1.18	0.16	9.30	< 0.001
Dobrý	0 %				(ref)
Před rekonstrukcí	−5 %	0.95	−0.05	−1.38	0.168

Hodnoty ve sloupci *efekt* byly vypočítány jako $(exp(\beta) - 1) \cdot 100$ % a říkají, o kolik procent se zvýší cena, vzroste-li hodnota daného regresoru o jedničku. Ve výsledcích jsou uvedeny pro přehlednost i referenční úrovně nominálních regresorů. Jako paradox se může jevit to, že byty se samostatnou kuchyní jsou levnější než ty s kuchyňským koutem. Skutečně nejde o chybu, ale o trend známý z mnoha měst v ČR.

Závisle proměnnou tedy bude $log(cena)$. Regresory použijeme tak, jak je běžné, jen pro přehlednost regresor *rozloha bytu* zmenšíme o 60 a vydělíme 10. Absolutní člen tak bude vypovídat o 60metrovém bytu a nově vzniklý regresor bude říkat, o kolik desítek metrů se rozloha daného bytu liší od 60 metrů. Za referenční skupinu zvolíme byt $2 + kk$, energeticky středně náročný, v dobrém stavu, bez balkonu. Odhady regresních koeficientů obsahuje tabulka 10.

Výsledky říkají, že pokud bychom se bavili o bytu $2 + kk$ o rozloze 60 m^2 energeticky středně náročném, v dobrém stavu, bez balkonu, pak můžeme očekávat cenu přibližně 2 212 135 Kč.

Pokud jsme však koupili novostavbu $3 + kk$ o rozloze 80 m^2, která je energeticky nenáročná a má balkon, pak bychom postupovali následovně. Vezměme výchozí cenu $exp(\hat{\beta}_0)$. Za to, že jde o třípokojový byt, tuto částku vynásobme číslem 1.17. Za nízkou energetickou náročnost číslem 1.06 a za to, že jde o novostavbu, číslem 1.27. Dále nám cenu o 5 % zvedne fakt, že byt má balkon, budeme tedy násobit ještě číslem 1.05. Poslední úprava je zahrnutí plochy. Náš byt je o 2 desítky metrů větší, než bylo na začátku stanovených 60 m^2. Cenu musíme vynásobit číslem 1.07 za každých 10 metrů navíc. Budeme tedy násobit

číslem $1.07^2 = 1.14$. Kdyby měl popisovaný byt rozlohu třeba jen 30 m^2, pak bychom násobili číslem $1.07^{-3} = 1/1.07^3 = 0.81$.

Postup je tedy podobný tomu, co již známe, jen s tím rozdílem, že kde jsme dřív používali znaménko plus, tam použijeme krát, a kde jsme použili znaménko krát, tam umocňujeme. Pokud vám tento postup připadá neintuitivní, můžete spočítat predikci $log(\hat{Y})$ přímo z parametrů $\hat{\beta}$ tak, jak jsme zvyklí, a až samotný výsledek exponenciálně transformovat. Obě cesty nás dovedou ke stejnému řešení.

Očekávaná hodnota našeho bytu $3 + kk$ o 80 m^2 nám vyjde 4 159 964 Kč. Pokud nejsme spokojeni s bodovým odhadem, můžeme vypočítat predikční interval. Jednoduše to můžeme udělat tak, že vypočítáme predikční interval pro $\log(\hat{Y})$ a obě meze exponenciálně transformujeme. Když budeme hledat například 80% predikční interval, nalezneme po úpravě hodnoty (3 304 955 Kč, 5 236 167 Kč).

Než se pustíme do přemýšlení nad tím, zda jsme prodělali, nebo vydělali, zamysleme se nad ještě jednou vlastností log-normálních modelů. Naše predikce je vytvořená tak, aby co nejpřesněji vystihla střední hodnotu proměnné $\log(\hat{Y})$. Tu když transformujeme, nebude náš odhad již patřit střední hodnotě (tedy průměrné ceně), ale hodnotě mediánové. S tímto faktem někdy můžeme být spokojení, ale pokud bychom chtěli říct, kolik *v průměru* byt s danými parametry stojí, pak budeme muset výsledek ještě upravit. Lze dokázat, že úprava, která z našeho odhadu udělá odhad střední hodnoty, má podobu $\exp(\log(\hat{Y}) + \frac{S_\epsilon^2}{2})$, kde S_ϵ^2 je odhad reziduálního rozptylu log-normálního modelu.

Pokud jsme v našem případě odhadli S_ϵ^2 rovné hodnotě 0.032, pak je poctivá predikce střední hodnoty rovna číslu 4 227 278 Kč. Meze 80% predikčního intervalu bychom upravili analogicky na (3 358 434 Kč, 5 320 896 Kč). Pokud nás zakoupený byt stál, dejme tomu, 3.3 miliony korun, pak nás může hřát u srdce, že jsme koupili jeden z 10 % nejpodhodnocenějších bytů.

10 Kroková a hierarchická regrese

V předchozích kapitolách jsme předpokládali, že statistický model vzniká najednou – tedy že přesně víme, které regresory do něj chceme zařadit, a pak se teprve ptáme po jejich vahách. V této kapitole se seznámíme s dvěma přístupy, které, ač jsou vzájemně značně odlišné, mají společné to, že vytváří statistický model po krocích.

10.1 Kroková regrese

Představme si situaci, kdy máme velké množství regresorů, z nichž některé v modelu hrají nezastupitelnou roli a jiné naše predikce nijak nezpřesňují. Zejména tehdy, pokud chceme náš model použít k dělání předpovědí, nikoli zkoumání vztahů mezi proměnnými, oceníme jednoduchý, avšak přesný model, který obsahuje pouze takové regresory, které nám přináší užitek. K tomuto cíli nás může dovést kroková regrese (*stepwise regression*). Krokovou regresi můžeme realizovat dvěma postupy: zpětnou metodou a dopřednou metodou.

Při využití **zpětné metody** (*backward elimination*) začneme s původním navrženým modelem, který obsahuje všechny regresory, které máme k dispozici. Následně hledáme takový regresor, který má v modelu nejmenší váhu, a tážeme se, jestli jeho zařazení nebylo zbytečné. Pokud usoudíme, že tento regresor skutečně nepřináší žádné zpřesnění, z modelu jej vyřadíme. Poté model zkontrolujeme znovu a opět hledáme takový regresor, který by bylo možné vyřadit. Postup opakujeme, dokud v modelu není žádný regresor, který jsme vyhodnotili jako zbytečný.

Naopak **metoda dopředná** (*forward selection*) začíná s modelem bez regresorů a hledá, která z nezávisle proměnných, jež máme k dispozici, by predikční schopnosti modelu nejvíce zlepšila. V případě, že existuje proměnná, která poskytuje dostatečné zpřesnění, do modelu ji přidáme. Postup opět opakujeme do té doby, než zjistíme, že žádná další proměnná, kterou bychom mohli do modelu přidat, již zlepšení nepřináší.

Při posuzování, která proměnná přináší dostatečné zpřesnění, můžeme použít několik kritérií. Ve statistických programech to je nejčastěji hodnota statistiky F testu srovnávajícího přesnost modelu, který daný regresor obsahuje, s podmodelem, do kterého tento regresor zařazen není. Pokud je nastavena nízká hodnota, budeme v modelu ponechávat i ty regresory, které nejsou statisticky významné, naopak vysoké hodnoty zachovávají jen regresory s velmi malými p-hodnotami.

Na první pohled by se mohlo zdát, že zpětná i dopředná metoda vede ke stejnému výsledku. Ve skutečnosti tomu tak vždy být nemusí. Existují situace, kdy máme více regresorů, z nichž každý sám o sobě přináší jen malé zpřesnění. Pokud se však octnou v modelu společně, tak přesto, že jsme nepřidali jejich interakci, přináší citelné zpřesnění. Takováto skupina by nebyla do modelu zahrnuta pomocí dopředné metody, jelikož ta

přidává regresory po jednom a žádný z kandidátů by vstupní kritérium nesplnil. Naopak metoda zpětného vyřazování by celou skupinu v modelu zachovala, jelikož odstranění libovolného regresoru by vedlo k propadu přesnosti.

Kroková regrese je v některých situacích vítaným pomocníkem pro tvorbu modelu, řada statistiků ji však oprávněně kritizuje. Problém tkví v tom, že při hledání nejúčinnějších regresorů prosséváme všechny možné kandidáty a snadno zaměníme dobrý regresor s falešně pozitivním nálezem – tedy regresorem, který ve skutečnosti citelné zpřesnění nepřináší, ale dílem náhody se tak v našem souboru jeví. Při obvyklé regresi bychom pravděpodobně zpozorněli v situaci, kdy třeba z padesátky regresorů identifikujeme tři jako statisticky významné, jelikož při pětiprocentní hladině významnosti na každých dvacet statistických testů za platnosti nulové hypotézy připadá v průměru jeden falešně pozitivní výsledek. Také by nás nejspíš varovala hodnota $R^2_{adj.}$, která by se zřejmě povážlivě lišila od klasického R^2. Pokud však použijeme krokovou regresi, získáme elegantní model se třemi regresory, jejichž významnost nás zřejmě zpochybnit nenapadne.

Pokud i přes tuto zrádnou vlastnost krokové regrese chcete tento postup použít, přečtěte si kapitolu 12 o rizicích přeučení modelu a cestách, jak jim čelit. Křížová validace je totiž dobrým lékem na všechny zmiňované problémy.

10.2 Hierarchická regrese

Podobně jako při dopředné krokové regresi, i při hierarchické regresi přidáváme do modelu regresory postupně. Na rozdíl od krokové regrese, která je takzvaně *řízená daty* (*data driven*), je hierarchická regrese *řízená teorií* (*theory driven*). Při rozhodování o tom, v jakém pořadí budeme při hierarchické regresi přidávat regresory, nehledíme na jejich statistickou významnost, ale naše rozhodnutí je dáno racionálně s oporou v teorii, ze které vycházíme. Typicky regresory rozdělíme do několika skupin, které do modelu v předem určeném pořadí přidáváme. Při každém kroku pak zkoumáme statistickou významnost množství vysvětleného rozptylu (značíme ΔR^2).

V jedné studii jsme se například tázali, jestli je kreativní úspěšnost (tzn. množství a kvalita kreativních počinů, které jedinec za svůj život vykonal) ovlivněna dvěma méně známými rysy s názvem *systemizace* a *empatizace* (pro podrobnosti viz teorie Simona Barona-Cohena). Do modelu se závisle proměnnou *kreativní úspěšnost* jsme kromě skórů systemizace a empatizace zařadili i pohlaví a zejména věk respondentů, jelikož obě tyto proměnné můžou hrát významnou roli. Také bychom se ale mohli tázat, jestli případný efekt systemizace a empatizace není způsoben jen tím, že tyto rysy korelují již s dobře známými osobnostními dimenzemi velké pětky, u kterých již byla doložena korelace s kreativitou, a není-li tedy úplně zbytečné je do našich úvah zahrnovat. Hierarchická regrese by mohla mít následující podobu.

V prvním kroku porovnáme model bez regresorů s modelem, který obsahuje jen regresory pohlaví a věk. Pokud bude rozdíl v přesnosti modelů významný, pak můžeme říct, že tyto základní biologické charakteristiky souvisí s kreativitou. V druhém kroku porovnáme model s regresory pohlaví a věk s modelem, do kterého přidáme obecné osobnostní dimenze velké pětky. Opět se tážeme, jestli došlo ke statisticky významnému zlepšení predikčních schopností modelu. Pokud ano, můžeme konstatovat, že obecné dimenze osobnosti souvisí s kreativní úspěšností jedince. Navíc bychom uvedli, kolik procent rozptylu nad rámec základních biologických charakteristik osobnostní profil vysvětluje (tzn. ΔR^2). Poslední krok je pro nás nejzajímavější. Porovnáme model obsahující regresory věk, pohlaví a dimenze velké pětky s modelem, do kterého navíc přidáme proměnné našeho zájmu – systematizaci a empatizaci. Ověříme statistickou významnost a spočítáme, o kolik vzrostlo procento vysvětleného rozptylu tentokrát. Hypotéza, jejíž platnost testujeme, říká, že empatizace a systemizace přináší nějaké relevantní informace o kreativitě jedince nad rámec toho, co již víme díky jeho pohlaví, věku a obecným dimenzím osobnosti. Tímto postupem ověřujeme inkrementální validitu konstruktu systemizace a empatizace. Pokud by v tomto testu neuspěly a do modelu by již novou informaci nepřinesly, znamenalo by to, že tyto dvě proměnné nejsou ve výzkumu kreativity relevantní vzhledem k tomu, co jsme schopni popsat s využitím již zažitých konstruktů.

Ve většině případů si vystačíme jen s dvěma kroky, kdy v první skupině regresorů jsou ty, které již známe a nejsou předmětem našeho zájmu, zatímco v druhé jsou ty, jejichž vliv chceme prozkoumat. Pokud by v druhé skupině byl jediný regresor, nemusíme hierarchickou regresi používat, jelikož test podmodelu by splýval s testem pomocí Waldovy statistiky. Hierarchická regrese je považována za velmi poctivý přístup při práci s daty.

11 Odstranění vlivu vybraných proměnných

Představte si, že si kladete odvěkou otázku, do jaké míry je inteligence člověka ovlivněna dědičností. Naivní přístup by nás dovedl k jednoduchému výzkumnému designu, kdy bychom otestovali IQ testem děti určitého věku i jejich rodiče. Mezi oběma veličinami bychom pochopitelně nalezli těsný vztah. Je to ovšem odpověď na naši otázku? Není, jelikož rodiče kromě toho, že dětem předávají nějaké biologické dědictví (zejména prostřednictvím genů), je zároveň vystavují podnětům, které se můžou napříč rodinami výrazně lišit a nejspíš inteligenci dítěte ovlivňují. Nasnadě je řešení, které by bylo metodologicky (i eticky) problematické – přimět rodiče, aby kupovali dětem stejné množství hraček, knih, přihlásili je do stejného množství kroužků, dali do stejných škol, a poskytli jim stejné množství místa v pokojíčku atp. Mnohem snáze bychom tento úkol uchopili matematicky.

Stačilo by u každého dítěte kromě jeho IQ a IQ jeho rodičů zmapovat rozmanité faktory, které se můžou na utváření inteligence podílet, a proměnnou IQ dítěte následně od vlivu pocházejících z prostředí *očistit*.

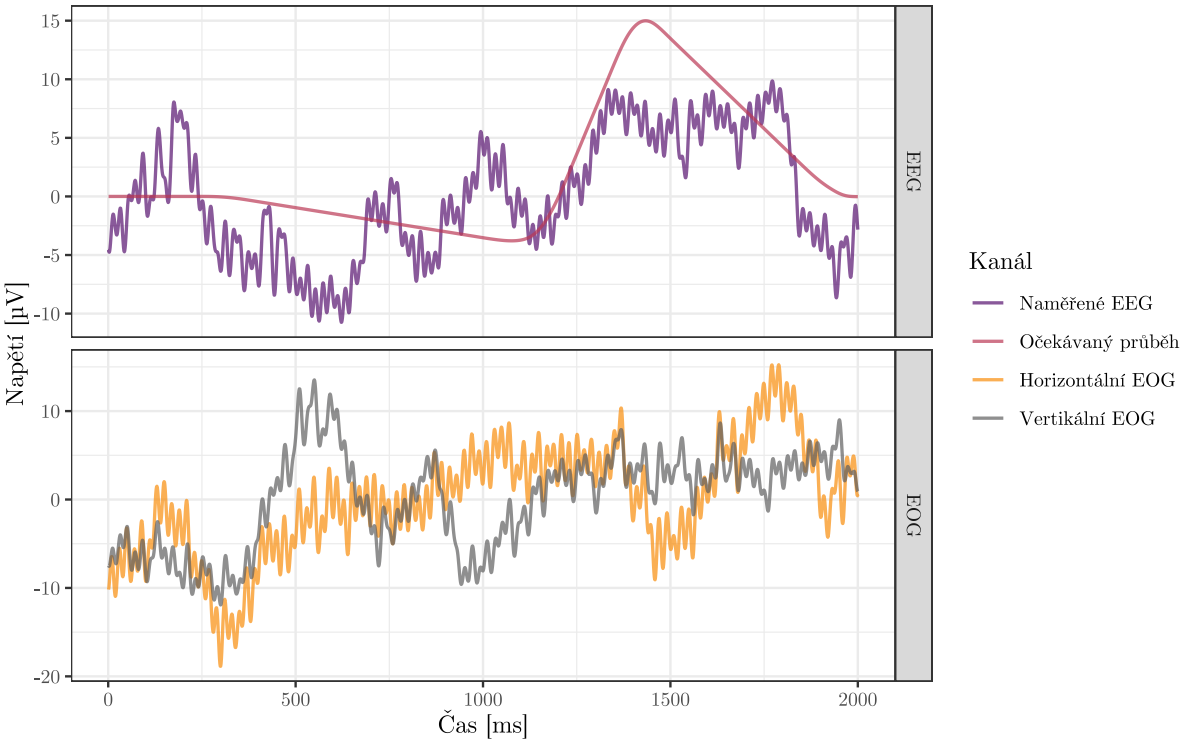
Toto očištění provedeme ne překvapivě pomocí lineárního modelu, kdy jako závisle proměnnou Y použijeme tu vlastnost, kterou budeme očišťovat, a za regresory X všechny faktory, jejichž vliv chceme zohlednit. Možná překvapivě jsou nové, očištěné hodnoty IQ dětí rezidua tohoto modelu, jelikož rezidua říkají, o kolik si dítě vede v testu lépe nebo hůře, než bychom čekali s přihlédnutím k prostředí, ve kterém vyrůstalo.

Kromě toho, že si pro jakékoli další výpočty uložíme takto pořízená rezidua, můžeme provést jednu kosmetickou úpravu. Rezidua mají vždy průměr 0, což často (třeba právě v případě IQ) není moc smysluplná hodnota. Můžeme proto ke každému uloženému reziduu přičíst průměrnou hodnotu Y (nebo jinou smysluplnou hodnotu, třeba 100), čímž vrátíme nově získanou proměnnou na původní úroveň.

Jiným příkladem odstranění vlivu rušivých proměnných může být práce s fyziologickými daty. Představte si situaci, kdy s pomocí elektroencefalografu (EEG) měříte mozkovou aktivitu probanda, kterého vystavujeme určitému podnětu. EEG reaguje velmi nespecificky a místo sledovaného potenciálu můžeme snadno naměřit změny napětí způsobené například motorickou aktivitou. Typickým zdrojem artefaktů je činnost okohybných svalů, jejichž stopy jsou patrné zejména tehdy, když úloha neumožňuje použít fixační kříž (třeba při řízení auta). Jednoduché řešení opět poskytuje lineární model[17].

[17] Pokud bychom se tímto problémem zabývali hlouběji, zjistíme, že v praxi existuje řada mnohem pokročilejších a účinnějších postupů, které tento problém řeší.

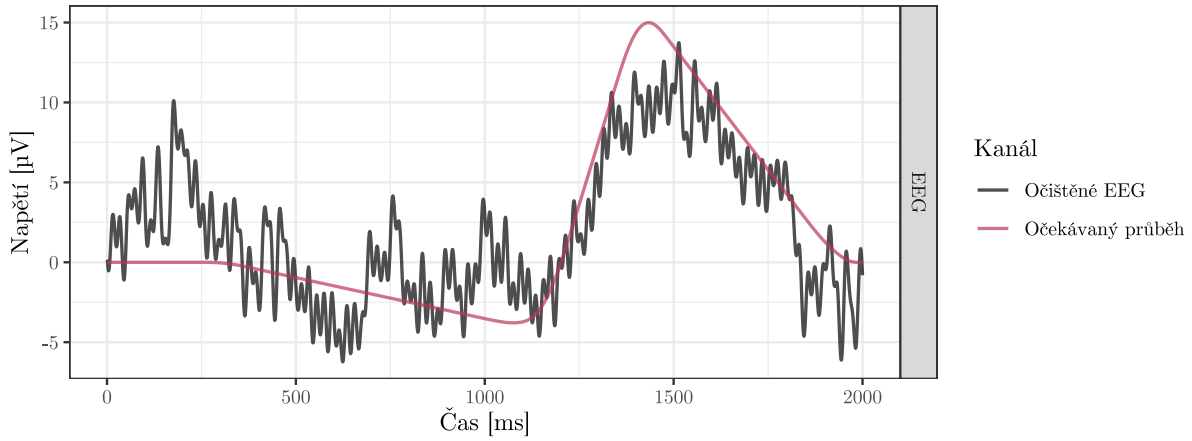
Obrázek 20: Signál EEG a EOG



Do experimentu bychom kromě elektrod EEG zapojili i elektrookulograf (EOG). Pokud použijeme dvě dvojice elektrod EOG, můžeme sledovat zvlášť vertikální a zvlášť horizontální pohyby oka. Všechny naměřené údaje ilustruje obrázek 20. Kromě údajů z jednoho kanálu EEG je zde vykreslen i průběh, který očekáváme, že budeme pozorovat. Tážeme se, do jaké míry kopíruje signál EEG naše očekávání. Již teď pozorujeme poměrně slušnou shodu – korelační koeficient EEG a očekávaných hodnot je 0.69. Na první pohled však vidíme, že v naměřeném průběhu je přítomna řada dalších artefaktů, které můžou být způsobeny právě očními pohyby.

Oba údaje o pohybech očí můžeme z původního EEG odstranit pomocí lineárního modelu. Ten bude mít tvar $\text{EEG} = \beta_0 + \beta_1 \text{EOG}_h + \beta_2 \text{EOG}_v$ a v našem případě bude schopen vysvětlit (odstranit) 24 % rozptylu EEG. Rezidua z tohoto modelu jsou signálem z EEG očištěným od hodnot EOG. Pro snazší grafickou prezentaci můžeme k těmto reziduím (která mají z definice průměr 0) přičíst nějakou konstantu, třeba původní průměr nebo průměr našeho očekávání (tuto situaci znázorňuje obrázek 21). Korelační koeficient vzrostl na hodnotu 0.82, což dává naší hypotéze ještě větší podporu.

Obrázek 21: Očištěný signál EEG



11.1 Parciální a semiparciální korelační koeficient

Očištěné hodnoty obvykle používáme v dalších výpočtech, s čímž se vážou dva pojmy. Pojmem *parciální korelační koeficient* označujeme Pearsonův korelační koeficient spočítaný mezi dvěma proměnnými, kdy jsme z obou ještě před výpočtem odstranili vliv sady jedné nebo více proměnných. Pokud bychom toto očištění provedli jen na jedné z těchto dvou proměnných, pak hovoříme o *semiparciálním korelačním koeficientu* (ten jsme počítali i v příkladu s EEG výše).

Statistické programy obvykle nabízí výpočet parciálního korelačního koeficientu jako samostatnou funkci a není potřeba vytvářet lineární model a ručně dělat všechny mezikroky. Pokud nicméně celou proceduru z nějakého důvodu potřebujeme provést ručně, měli bychom vzít v potaz malou změnu týkající se ověřování statistické významnosti tohoto koeficientu. Vzorec pro výpočet testové statistiky T se změní do podoby

$$T = \frac{R_{YZ.\mathbf{X}}}{\sqrt{1 - R_{YZ.\mathbf{X}}}}\sqrt{n-k-2} \sim t_{n-k-2}$$

kde k označuje počet faktorů, jejichž vliv jsme odstraňovali (v příkladu s EEG tedy $k = 2$), a $R_{YZ.\mathbf{X}}$ je parciální korelační koeficient mezi proměnnými Y a Z s odstraněním vlivu proměnných $\mathbf{X} = (X_1, X_2, ..., X_k)$[18].

[18] Řada statistických programů tento malý rozdíl nicméně zanedbává a používá obvyklý vzorec pro test významnosti Pearsonova korelačního koeficientu. Taky dodejme, že v konkrétním příkladě s EEG by uvedený statistický test nebyl přesný, jelikož se nejedná o sadu nezávislých pozorování, ale o takzvanou *časovou řadu*. Těžko v tomto kontextu můžeme předpokládat nezávislost náhodných složek jednotlivých pozorování.

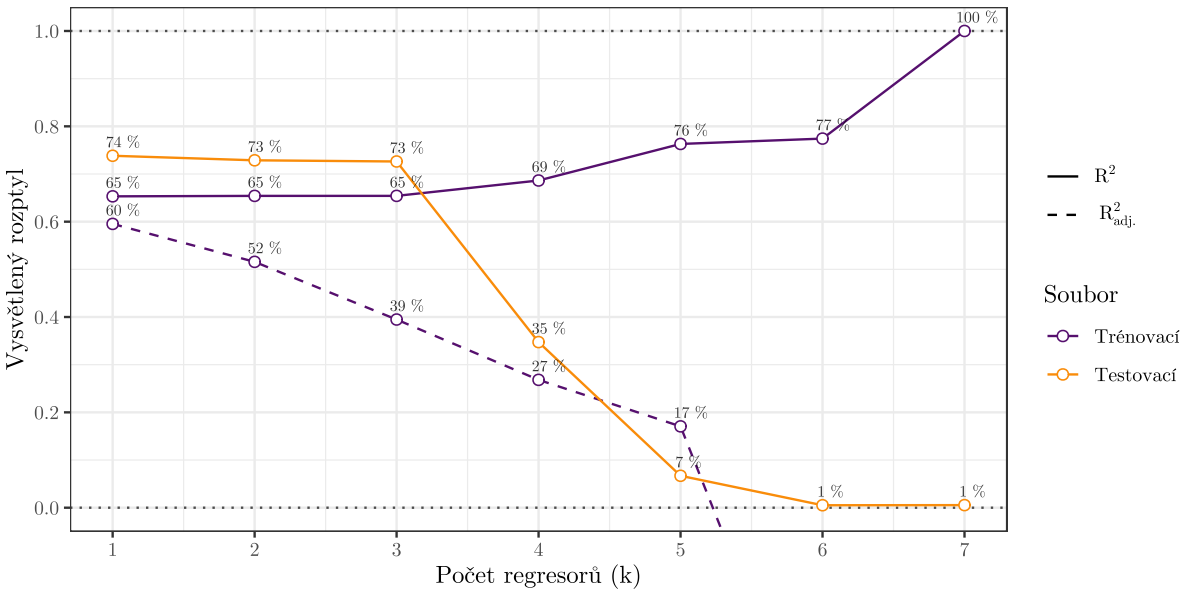
12 Problém přeučení a křížová validace

Při tvorbě modelu máme k dispozici široké spektrum možností. Můžeme se rozhodovat, které regresory do něj zařadíme, jaké interakční členy prvního či vyššího řádu přidáme, zda budeme spojité proměnné modelovat jako lineární, kvadratické či v mocninách ještě vyšších řádů. Poměrně snadno můžeme sklouznout k tomu, že vytvoříme tak košatý model, že nám ve skutečnosti již vůbec žádnou informaci neposkytne.

Ilustrujme to příkladem. Opět se vrátíme k datům z tabulky 1 o Agátě, Otokarovi a dalších šesti spolužácích. Zjistili jsme, že předpovídat výsledek zápočtového testu pomocí počtu hodin, které student věnoval přípravě, je vcelku racionální postup. Co když ale opustíme předpoklad, že pozorovaný vztah je lineární, ale budeme ho modelovat jako kvadratický pomocí přidání regresoru *počet hodin na druhou*? A co když nám toto nebude stačit a přidáme dále regresor *počet hodin na třetí*, abychom modelovali vztah jako kubický? Tak bychom mohli pokračovat a přidávat regresor *počet hodin* ve vyšších a vyšších mocninách, až po hodnotu k. Výsledky budou poměrně přesvědčivé – s přidáním každého dalšího regresoru R^2 vzroste. Jakmile dosáhneme $k = 7$, bude R^2 rovno 100 %. Model tedy dosáhne dokonalé shody s daty.

Nejspíš namítnete, že na pouhých 8 pozorování jsme do modelu zařadili příliš moc parametrů (při $k = 7$ jich odhadujeme 8). Na toto nás upozorní i hodnota $R^2_{adj.}$, která začne povážlivě klesat, pokud je regresorů v modelu příliš mnoho. Vývoj hodnot R^2 i $R^2_{adj.}$ při postupném přidávání regresorů znázorňují fialové čáry v obrázku 22.

Obrázek 22: Vývoj hodnoty koeficientu determinace při různých počtech regresorů

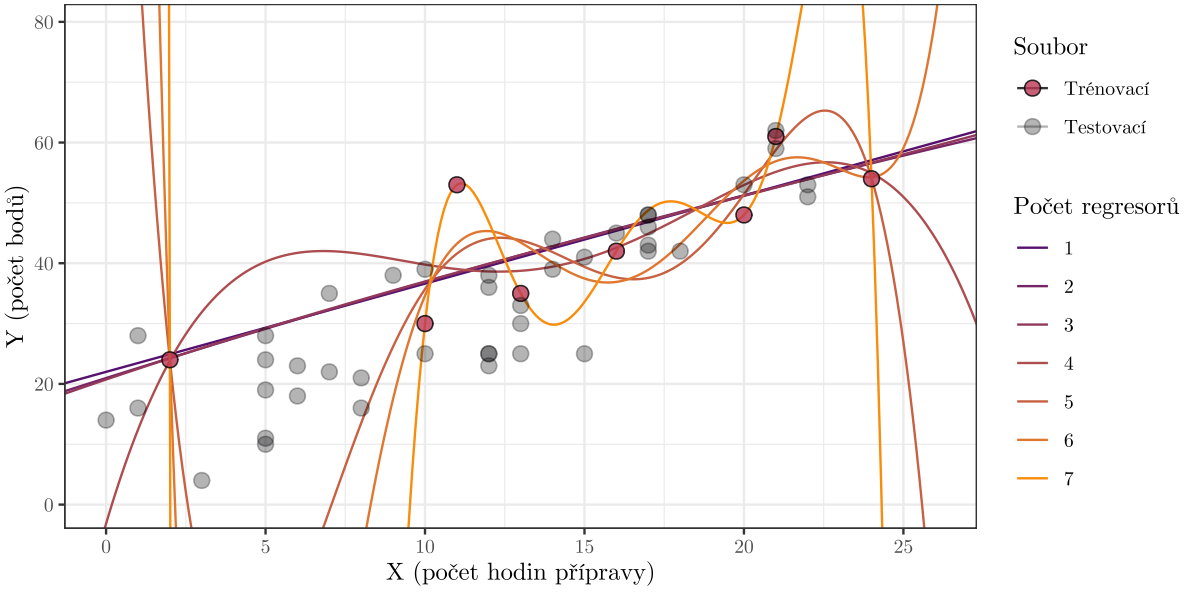


V praxi je ale situace často mnohem méně přehledná. Pokud při tvorbě modelu procházíme velkým množstvím potenciálních kandidátů (třeba tak, jak jsme zvyklí z kapitoly o krokové regresi), snadno vyrobíme model, který obsahuje jen několik málo regresorů (a tedy $R^2_{adj.}$ zůstane vysoké) a přitom se s daty shoduje velmi těsně. Pro čtenáře je velmi těžké zpochybnit kvalitu daného modelu, jelikož všechny kvantitativní ukazatele mluví v jeho prospěch, a informace o tom, z kolika potenciálních kandidátů byl finální model vybrán, nám zůstává obvykle utajena.

Problém, kterému zde čelíme, se označuje jako **přeučení**, nebo též přefitování (*overfitting*) modelu. Přeučením se obecně rozumí to, že naše odhady parametrů jsou dokonale přizpůsobené našim datům, ale už nebudou odpovídat jakýmkoli novým datům. Pokud tedy někdo náš výzkum zopakuje, nebude jeho výsledek ten náš připomínat ani vzdáleně.

Co se děje při přeučení, znázorňuje obrázek 23. Jako červené body jsou zde znázorněni studenti našeho souboru a oranžovofialové křivky představují očekávané výsledky dle jednotlivých modelů. Jak je patrné, model lineární, kvadratický i kubický se v daném případě chová skoro stejně, a všechny tedy poskytují skoro stejné hodnoty R^2. U $k = 4$ se však začne dít zvláštní věc – regresní křivka začne kolísat v širokém rozpětí nízkých i vysokých hodnot a při $k > 4$ opouští jakékoli rozumné meze. Křivka z modelu s $k = 7$ prochází bezchybně všemi osmi body, nicméně za cenu extrémního kolísání.

Obrázek 23: Pozorované hodnoty a regresní křivky jednotlivých modelů



V našem příkladu odhalíme přítomnost přeučení jediným pohledem na obrázek 23, u složitějších modelů se tento úkol ale stává prakticky neřešitelným. Oblíbeným a v podstatě jediným účinným řešením problému přeučení je takzvaná **křížová validace** (*cross-validace*). Křížová validace vychází z úvahy, že není dobrý nápad ověřovat kvalitu modelu

na stejných datech, na kterých jsme odhadovali jeho parametry. Před jakýmkoli výpočtem proto soubor rozdělíme na dva podsoubory: **trénovací** a **testovací**. Pro odhad parametrů použijeme pouze pozorování z trénovacího souboru, ale kvalitu shody s daty budeme ověřovat jen na prvcích souboru testovacího.

Testovací soubor bývá obvykle o něco menší než trénovací. V našem případě nebudeme dělit osm pozorování na dvě skupiny, ale do souboru přidáme dalších 42 spolužáků Agáty a Otokara. V obrázku 23 jsou značeni jako šedá kolečka. Můžeme vidět, že čím se náš model může víc přizpůsobit trénovacímu souboru, tím se víc vzdaluje od souboru testovacího. V obrázku 22 oranžová křivka znázorňuje vývoj R^2 počítaného na testovacím souboru. Ze začátku vychází velmi obstojně (shodou okolností dokonce o něco lépe něž na trénovacím souboru). Od $k = 4$ ale dochází k rapidnímu propadu a pro $k = 6$ a 7 už prakticky nedokáže vysvětlit vůbec žádný rozptyl.

Provedení křížové validace výrazně zvyšuje věrohodnost našich výsledků. V psychologii sice tento postup není příliš oblíbený, ale v řadě jiných oborů (zejména v rámci strojového učení[19]) je křížová validace nezbytný krok a jeho vynecháním autor své výsledky diskvalifikuje z jakékoli diskuze[20].

V případě potřeby si vystačíme s křížovou validací tak, jak je popsaná v těchto skriptech. V odborné literatuře nicméně můžeme narazit na sofistikovanější metody křížové validace. Například oblíbená metoda označovaná jako K-násobná (*K-fold*) křížová validace rozděluje soubor na K stejně velkých podsouborů. Postupně se všech K podsouborů vystřídá v roli testovacího souboru, zatímco zbývající pozorování vždy hrají roli trénovací sady. Získaných K odhadů přesnosti modelu pak zprůměrujeme.

[19] Strojové učení se zabývá tvorbou vysoce komplexních modelů, které mají za úkol nacházet správné řešení (predikci) v rámci náročných problémů. Oblíbenou třídou modelů jsou pak umělé neuronové sítě, které běžně obsahují stovky až tisíce parametrů – přeučení je zde proto ústředním tématem.

[20] Důvodem malého užití křížové validace v psychologii je to, že zde obvykle vytváříme modely za účelem testování statistické významnosti vybraných regresorů, nikoli pro účely predikce. V případě, že je statistický model přeučený, odhady směrodatných odchylek jednotlivých parametrů neúměrně narůstají a je jen malá šance, že bude nalezen nějaký signifikantní vztah. Potenciální riziko falešně pozitivního nálezu tedy klesá.

13 Modifikace metody nejmenších čtverců

Metoda nejmenších čtverců je univerzálním nástrojem pro práci s daty. V praxi však můžeme narazit na situace, kde nám tento postup ve své základní podobě nestačí. V těchto situacích nám někdy můžou pomoci některé z modifikací metody nejmenších čtverců. Zmiňme alespoň dva takovéto postupy.

13.1 Zobecněná metoda nejmenších čtverců

Jedním ze základních předpokladů pro použití metody nejmenších čtverců bylo to, že všechna měření jsou realizována se stejnou přesností a že po zohlednění vlivu regresorů jsou libovolná dvě pozorování na sobě nezávislá. Pokud bychom tedy znali varianční matici náhodné složky modelu, označme ji $\mathbf{\Sigma}$, opakovala by se na její diagonále stále tatáž hodnota σ_ϵ^2 (reziduální rozptyl) a mimo diagonálu by byly nuly (tedy nulové kovariance libovolných dvojic pozorování). Matici sigma si můžeme představit jako součin σ_ϵ^2 a nějaké matice $\mathbf{V}$. Jsou-li dodrženy zmiňované podmínky užití metody nejmenších čtverců, pak by matice $\mathbf{\Sigma}$ a $\mathbf{V}$ pro maličký soubor o čtyřech pozorováních ($n = 4$) vypadaly takto:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\epsilon^2 & 0 & 0 & 0 \\ 0 & \sigma_\epsilon^2 & 0 & 0 \\ 0 & 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & 0 & \sigma_\epsilon^2 \end{pmatrix} = \sigma_\epsilon^2 \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \sigma_\epsilon^2 \mathbf{V}$$

Matice $\mathbf{V}$ má tedy podobu takzvané jednotkové matice. Co když popisovaná podmínka ale není splněna? Matice $\mathbf{V}$ potom může různě měnit svou podobu. Nejčastěji se setkáme se situací, kdy mimo diagonálu zůstanou nuly, ale diagonála bude obsahovat rozmanité hodnoty. Takováto situace by vyjadřovala stav, kdy sice jsou jednotlivá měření nezávislá, ale mají různou přesnost. Například matice

$$\mathbf{V} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

by popisovala situaci, kdy jsou první dvě měření učiněna s dvojnásobnou přesností (tedy polovičním rozptylem) než druhá dvě měření. Mohlo by jít třeba o situaci, kdy jednotlivá pozorování jsou průměrné hodnoty naměřené na různě velkých skupinách. Diagonální prvky matice $\mathbf{V}$ by pak odpovídaly převráceným hodnotám rozsahů jednotlivých skupin. Ve statistických programech bývá tato možnost úpravy matice $\mathbf{V}$ dostupná prostřednictvím vah jednotlivých pozorování. Těmito vahami se rozumí právě převrácené hodnoty diagonálních prvků matice $\mathbf{V}$.

V psychologii méně častá možnost by se týkala matice $\mathbf{V}$, která obsahuje nenulové hodnoty mimo diagonálu. Šlo by o situaci, kdy je porušen předpoklad nezávislosti a jednotlivá pozorování jsou vzájemně korelovaná. Podmínkou však je, že víme, v jakém poměru jsou vůči sobě jednotlivé rozptyly a kovariance. Například matice

$$\mathbf{V} = \begin{pmatrix} 1 & -0.1 & -0.1 & -0.1 \\ -0.1 & 1 & -0.1 & -0.1 \\ -0.1 & -0.1 & 1 & -0.1 \\ -0.1 & -0.1 & -0.1 & 1 \end{pmatrix}$$

by popisovala čtyři pozorování měřená se stejnou přesností a mezi libovolnou dvojicí pozorování existuje negativní kovariance odpovídající jedné desetině reziduálního rozptylu.

Ať už použijeme jakoukoli[21] matici $\mathbf{V}$, odhad regresních koeficientů získáme pomocí takzvané zobecněné metody nejmenších čtverců, která je popsána rovnicí

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

Odhad varianční matice by se analogicky změnil do podoby

$$\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}}) = S_\epsilon^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

Všimněte si, že (klasická) metoda nejmenších čtverců je jen speciálním případem zobecněné metody nejmenších čtverců, kdy $\mathbf{V}$ je jednotkovou maticí. Jednotková matice má totiž tu vlastnost, že její inverze je opět jednotkovou maticí, a pokud jednotkovou maticí vynásobíme jinou matici, nijak tím hodnoty jejích prvků nezměníme.

13.2 Modely s podmínkou

V kapitole 5.1 o testu podmodelu jsme předpokládali, že jediná restrikce, kterou můžeme omezit náš model, a vytvořit z něj tak podmodel, je to, že z něj odstraníme jeden či více regresorů. Ve skutečnosti je toto jen jeden speciální případ mnohem univerzálnějšího a účinnějšího postupu. Tímto postupem je použití podmínek na regresní koeficienty.

Podmínkou se rozumí určitá restrikce, kterou vyřkneme před tím, než provedeme odhad regresních vah. Podmínkou by třeba bylo to, kdybychom předpokládali, že dva regresory mají přesně stejnou (ač neznámou) váhu: $\beta_1 = \beta_2$, nebo to, že některý z regresních koeficientů má určitou pevně danou hodnotu: $\beta_3 = 18$. Do modelu můžeme přidat jakoukoli podmínku nebo skupinu podmínek, které lze vyjádřit ve tvaru *lineární kombinace*

[21] Matice $\mathbf{V}$ úplně *jakákoli* být nemůže. Jelikož je z ní odvozena varianční matice, musí jít takzvaně o pozitivně definitní matici. Z matematického pohledu to znamená, že tato matice má kladný determinant, respektive kladná všechna vlastní čísla. My si to můžeme představit jako varianční matici, ve které nedochází k žádnému paradoxu, jako je třeba proměnná s nulovým nebo záporným rozptylem nebo kovariance vyšší než umožňují rozptyly dvou příslušných proměnných.

regresních vah se rovná určité konstantě. Zapsáno pomocí matic

$$\mathbf{B}\boldsymbol{\beta} + \mathbf{b} = \mathbf{0}$$

kde $\mathbf{B}$ je matice definující lineární kombinaci vah, $\mathbf{b}$ je vektor konstant a $\mathbf{0}$ je vektor nul. Použijeme-li jako příklad dvě výše uvedené podmínky, pak by u modelu se čtyřmi regresory $(\beta_0, \beta_1, \beta_2, \beta_3)$ šlo o tyto hodnoty:

$$\begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} 0 \\ -18 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Zapsáno bez matic naše podmínky říkají $\beta_1 - \beta_2 = 0$ a $\beta_3 - 18 = 0$, tedy $\beta_1 = \beta_2$ a $\beta_3 = 18$, jak jsme chtěli.

Unikátních podmínek, které do modelu vložíme, nemůže být pochopitelně neomezené množství. Ve chvíli, kdy by jich bylo stejně jako počet parametrů modelu, nezbyly by žádné volné parametry k odhadu, a kdyby jich bylo ještě víc, model by žádné řešení neměl. Záměrně zde používáme slovo unikátní, jelikož ne každá podmínka do modelu další restrikci skutečně přidává. Představme si, že bychom do modelu přidali další dvě podmínky: $\beta_1 = 0$ a $\beta_2 = 0$. Poslední uvedená podmínka zjevně již žádné omezení nepřináší, jelikož jsme již dříve vymezili, že $\beta_1 = \beta_2$ a $\beta_1 = 0$, z čehož fakt, že $\beta_2 = 0$, vyplývá[22].

Možnost přidání podmínek nám poskytuje kromě ještě větší svobody při vytváření lineárních modelů i testování pokročilejších statistických hypotéz. Opět bychom využili test podmodelu, tentokrát bychom však místo odebírání regresorů z původního modelu přidávali podmínky. De facto je totiž odebrání regresoru přesně totéž co omezení jeho váhy na hodnotu 0. Mezi hypotézy, které jsme dříve nemohli testovat, by třeba patřila hypotéza o tom, jestli mají dva regresory stejnou váhu nebo se jejich váhy liší, nebo o tom, jestli se některý z regresních koeficientů rovná nějaké pevně dané hodnotě.

Za každou unikátní podmínku, kterou jsme přidali do modelu, získáváme zpět jeden stupeň volnosti, čehož si všimneme při výpočtu odhadu chybového rozptylu, který se změní do podoby

$$S_\epsilon^2 = \frac{1}{n - p + q} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

kde q je počet unikátních podmínek. To se promítne i do stupňů volnosti testové statistiky F testu podmodelu:

$$F = \frac{\frac{\Delta R^2}{h}}{\frac{1 - R^2_{model}}{n-p+q}} \sim F_{h,\, n-p+q}$$

[22] Odpověď na otázku, kolik má model unikátních podmínek, nemusí být vždycky tak zjevná jako zde. Pro přesné určení jejich počtu můžeme vypočítat *hodnost* matice $\mathbf{B}$, která počtu unikátních podmínek odpovídá.

Číslo h tentokrát udává, kolik jsme přidali unikátních podmínek, když jsme z modelu odvozovali podmodel, q je pak počet podmínek původního modelu (tzn. většinou nula).

Největší změna po přidání podmínek se týká podoby rovnic pro odhad parametrů β pomocí metody nejmenších čtverců a odhad jejich varianční matice, které se stanou o poznání složitější:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{I} - \mathbf{C}^{-1}\mathbf{B}'\left(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}'\right)^{-1}\mathbf{B}\right)\mathbf{C}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - \mathbf{C}^{-1}\mathbf{B}'(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}')^{-1}\mathbf{b}$$

$$\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}}) = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}'(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{C}^{-1}$$

kde

$$\mathbf{C}^{-\mathbf{1}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} \qquad \text{a} \qquad \boldsymbol{\Sigma} = S_\epsilon^2\mathbf{V}$$

Odhady $\hat{\beta}$ jsou pak opět nejlepšími nestrannými odhady skutečných hodnot parametrů β za platnosti stanovených podmínek.

I přes velkou užitečnost tohoto přístupu většina standardních statistických programů práci s lineárními modely s podmínkami neumožňuje.

14 Testy kontrastů a kódování nominálních regresorů

V kapitole 4.1 jsme se seznámili se způsobem, jak do statistického modelu zahrnout nějaký nominální regresor. Popisovaný způsob s využitím indikátorových proměnných (*dummy variables*) je ve většině případů uspokojivá cesta. Ve skutečnosti existuje celá řada jiných kódování, po kterých můžeme sáhnout. Některá nám můžou posloužit jako elegantní nástroj k testování různých komplikovanějších hypotéz týkajících se srovnávání několika skupin.

Každé řešení, které budeme v této kapitole probírat, aplikujeme na ukázkovou datovou matici a interpretujeme nalezené koeficienty. Datová matice je tvořena 18 pozorováními, na každém z nich jsme kromě závisle proměnné Y zaznamenali příslušnost k jedné ze skupin a, b, c, d. Naměřené hodnoty, rozsahy skupin a skupinové průměry jsou následující:

$a:$	$9, 53, 16$	$n_a = 3$	$\bar{y}_a = 26$
$b:$	$77, 10, 143, 42$	$n_b = 4$	$\bar{y}_b = 68$
$c:$	$3, 43, 138, 246, 180$	$n_c = 5$	$\bar{y}_c = 122$
$d:$	$284, 133, 169, 150, 141, 251$	$n_d = 6$	$\bar{y}_d = 188$

Celkový průměr ze všech 18 hodnot bez ohledu na skupinu je $\bar{y} = 116$ a průměr průměrů, tedy $(26 + 68 + 122 + 188)/4$, je $\bar{\bar{y}} = 101$.

Indikátorové kódování

Použijeme-li nám již známé indikátorové kódování, určíme si jednu referenční skupinu. Řekněme, že jsme si zvolili skupinu a. Prvky patřící k jednotlivým skupinám bychom pak kódovali pomocí indikátorů X_1, X_2 a X_3 následovně:

skupina	X_0	X_1	X_2	X_3
a	1	0	0	0
b	1	1	0	0
c	1	0	1	0
d	1	0	0	1
$\hat{\beta}_j$	26	42	96	162

V tabulce jsme pro názornost vyznačili i sloupeček počátku, který by v datech znázorněn nebyl, a také odhady koeficientů $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ a $\hat{\beta}_3$. Koeficient $\hat{\beta}_0$ odpovídá průměrné hodnotě v referenční skupině a ($\bar{y}_a = 26$). Koeficienty $\hat{\beta}_1$, $\hat{\beta}_2$ a $\hat{\beta}_3$ pak říkají po řadě, o kolik je vyšší průměr ve skupině b, c, respektive d ve srovnání se skupinou a.

Kontrastové kódování

Kontrastové kódování (*contrast coding*) je velmi blízké indikátorovému kódování. Opět vyžaduje volbu referenční úrovně (použijeme zase úroveň a). Tabulka kontrastů bude vypadat následovně:

skupina	X_0	X_1	X_2	X_3
a	1	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
b	1	$\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
c	1	$-\frac{1}{4}$	$\frac{3}{4}$	$-\frac{1}{4}$
d	1	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{3}{4}$
$\hat{\beta}_j$	101	42	96	162

Z výsledku je patrné, že ke změně došlo jen u koeficientu β_0, který se teď rovná průměru průměrů všech skupin. Koeficienty β_1, β_2 a β_3 mají stejnou interpretaci jako v minulém případě – srovnávají danou úroveň s referenční skupinou.

Součtové kódování

Součtové kódování (*sum coding*) vyžaduje podobně jako předchozí dvě kódování volbu jedné referenční úrovně. Zvolme tentokrát třeba úroveň d. Tabulka kontrastů by pak vypadala následovně:

skupina	X_0	X_1	X_2	X_3
a	1	1	0	0
b	1	0	1	0
c	1	0	0	1
d	1	-1	-1	-1
$\hat{\beta}_j$	101	-75	-33	21

Koeficient β_0 se opět rovná průměru průměrů všech skupin. Koeficienty β_1, β_2 a β_3 pak říkají, o kolik se odlišuje průměr dané skupiny od β_0. Na rozdíl od indikátorového kódování tedy nesrovnáváme jednotlivé úrovně s referenční úrovní, ale s průměrnou hodnotou skupinových průměrů. Ve výsledku chybí ta skupina, která byla zvolena za referenční. Pokud bychom chtěli i ji srovnat s β_0, pak bychom museli určit jinou referenční skupinu a provést výpočet znovu.

Helmertovo kódování

V případě, že jsou kategorie regresoru ordinálně uspořádané, může být na místě využití Helmertova kódování:

skupina	X_0	X_1	X_2	X_3
a	1	-1	-1	-1
b	1	1	-1	-1
c	1	0	2	-1
d	1	0	0	3
$\hat{\beta}_j$	101	21	25	29

Regresní koeficient absolutního členu β_0 je opět roven průměru všech skupinových průměrů. Koeficient β_1 srovnává průměr skupiny b proti skupině a. Koeficient β_2 pak průměr skupiny c proti průměru průměrů skupin a a b. Analogicky β_3 poskytuje srovnání skupiny d proti průměru průměrů všech předchozích skupin, tedy a, b a c. Drobnou komplikací může být to, že pokud bychom zmiňované regresní koeficienty chtěli použít ke kvantifikaci rozdílu mezi uvedenými průměry, museli bychom jejich hodnoty vynásobit číslem 2 u koeficientu β_1, číslem 3 u β_2 a 4 u β_3. Tedy například $\bar{y}_b - \bar{y}_a = 68 - 26 = 42 = 2\beta_1$. Podobně pak $\bar{y}_c - \frac{\bar{y}_a + \bar{y}_b}{2} = 122 - 47 = 75 = 3\beta_2$.

Abychom se vyhnuli této interpretační nepříjemnosti, můžeme Helmertovo kódování přepsat do následující podoby:

skupina	X_0	X_1	X_2	X_3
a	1	$-\frac{1}{2}$	$-\frac{1}{3}$	$-\frac{1}{4}$
b	1	$\frac{1}{2}$	$-\frac{1}{3}$	$-\frac{1}{4}$
c	1	0	$\frac{2}{3}$	$-\frac{1}{4}$
d	1	0	0	$\frac{3}{4}$
$\hat{\beta}_j$	101	42	75	116

V obou případech bude výsledek testů statistické významnosti jednotlivých regresorů identický. Změnou bude pouze to, že v druhém případě nemusíme odhady koeficientů násobit.

Polynomiální kódování

Polynomiální kódování vychází z předpokladu, že úrovně a, b, c, d jsou nejen ordinálně seřazené, ale mají mezi sebou i shodné vzdálenosti. Polynomiální kontrasty se pak netáží po rozdílu mezi skupinovými průměry, ale zkoumají, zda jsou průměry jednotlivých skupin nějak smysluplně uspořádané – například zda netvoří přímku, parabolu atd. Nejvyšší stupeň polynomu, který tento druh kódování zkoumá, je určen počtem srovnávaných skupin. V našem případě půjde o lineární, kvadratický a kubický vztah:

skupina	X_0	X_1	X_2	X_3
a	1	-3	1	-1
b	1	-1	-1	3
c	1	1	-1	-3
d	1	3	1	1
$\hat{\beta}_j$	101	27	6	0

Nalezené koeficienty β_1, β_2 a β_3 nemají na první pohled zřejmou interpretaci. Využijeme je zejména k testování statistických hypotéz. V našem případě by signifikantně odlišný od nuly vyšel pouze koeficient β_1, který indikuje přítomnost lineárního vztahu. Pro existenci kvadratického a kubického vztahu jsme žádné důkazy nenašli.

Pokud by náš regresor měl jiný počet úrovní než 4, hodnoty kontrastů by se změnily. V každém případě by však platilo to, že hodnoty ve sloupci X_1 pochází z přímky, hodnoty sloupce X_2 z paraboly atp. Jednotlivé kódy by nám nejspíš poskytl statistický software. Dodejme, že hodnoty v jednotlivých sloupcích také závisí na použitém programu – v tomto textu jsme dali přednost celým číslům, jinde může být však upřednostněno jiné kritérium. Tato volba sice ovlivní hodnoty regresních koeficientů, ale nijak se nedotkne nalezených p-hodnot.

Další ad hoc kódování

Ve skutečnosti existuje způsobů, jak zakódovat nominální proměnnou, nekonečně mnoho – pro každou hypotézu bychom si mohli nějaké vymyslet. Například pokud bychom chtěli z nějakého důvodu srovnat, zda existuje rozdíl mezi průměry dvojic skupin a a b proti c a d, dále a a d proti b a c a nakonec a a c proti b a d, mohli bychom navrhnout například tyto kontrasty:

skupina	X_0	X_1	X_2	X_3
a	1	1	1	1
b	1	1	-1	-1
c	1	-1	-1	1
d	1	-1	1	-1
$\hat{\beta}_j$	101	-54	6	-27

Při tvorbě smysluplného kódování nominálního regresoru se musíme držet těchto pravidel: a) součty hodnot v jednotlivých sloupcích tabulky použité ke kódování se musí rovnat 0, b) kontrastů bude o jeden méně než je počet úrovní nominálního regresoru a c) každá dvojice sloupců musí být *ortogonální*. Tuto vlastnost můžeme ověřit tak, že půjdeme postupně po řádcích kódovací tabulky a vždy hodnotu z jednoho sloupce vynásobíme hodnotou z druhého sloupce. Součet výsledků se pak musí rovnat 0.

Další nápady při práci s nominální proměnnou

Pro zvídavého čtenáře ještě zmiňme další možnosti, jak pracovat s nominální proměnnou. Zamysleme se nad následujícím kódováním:

skupina	X_1	X_2	X_3	X_4
a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1
$\hat{\beta}_j$	26	68	122	188

Všimněte si, že v tabulce schází absolutní člen, který jsme do modelu nezařadili, a naopak zde jsou indikátorové proměnné pro všechny čtyři srovnávané skupiny. Hodnoty parametrů lze obvyklým způsobem odhadnout, ba co víc, budou velmi dobře interpretovatelné: rovnají se průměrům jednotlivých skupin. Náš model si můžeme představit tak, že obsahuje vlastně čtyři počátky, pro každou skupinu jeden. Určitá nevýhoda je, že i v případě, kdy náš model obsahuje více než jeden nominální regresor, může být takto zakódován pouze jeden.

Náš nápad můžeme ještě rozvíjet. Představme si, že bychom chtěli vedle našeho nominálního regresoru prozkoumat i vliv nějaké spojité proměnné a předpokládali bychom, že tento vliv může být v každé skupině jiný. Zkoumali bychom tedy interakci. Když se budeme držet našeho poněkud méně konvenčního modelu se čtyřmi počátky, můžeme vložit novou spojitou proměnnou také originálním způsobem – místo toho, abychom tuto proměnnou přidali do modelu tak, jak je (tedy jako hlavní efekt), přidáme pouze interakční členy. Ty ovšem budou čtyři, nikoli tři, jak jsme zvyklí – pro každou ze čtyřech skupin jeden.

Zejména tehdy, když prezentujeme jednodušší model posluchačům, kteří nejsou seznámeni s indikátorovým kódováním, může být tento způsob nejpřehlednějším řešením. Kromě toho, že koeficienty β_1 až β_4 představují počátky v jednotlivých skupinách, mají koeficienty interakčních členů (β_5 až β_8) taky velmi přímočarou interpretaci: kvantifikují sklon regresní přímky v jednotlivých skupinách.

15 Formáty *long*, *wide* a modely se smíšenými efekty

Základní kurzy statistiky nás naučily, že obvyklým způsobem jak formátovat datovou tabulku je vyhradit řádky jednotlivým probandům a do sloupců vepisovat sledované statistické znaky. Tento formát bývá označován jako *wide* (široký). Toto však není jediná možnost reprezentace dat. Řadu problémů uchopíme mnohem obratněji, když data převedeme do podoby označované jako *long* (dlouhý). Rozdíl mezi formáty a jejich užitek demonstrujme na následujícím příkladu.

V rámci psychofyziologického výzkumu si klademe otázku, jakou roli hraje neuroticismus v zátěžové situaci. Máme hypotézu, že jedinci vysoko skórující na škále neuroticismu budou vykazovat větší vzrušení při řešení frustrujícího úkolu ve srovnání s méně neurotickými jedinci. Míru vzrušení operacionalizujeme jako srdeční frekvenci měřenou EKG.

Každého probanda připojíme k EKG a zadáme mu postupně tři stresující úlohy: Stroopův test, aritmetickou úlohu (hlasité odříkávání čísel v klesající posloupnosti 300, 293, 286, 279...) a test verbální fluence (vyjmenovávání co nejvíc slov na zadané písmeno v časovém limitu). Mezi jednotlivé úkoly jsou zařazeny fáze relaxace, kdy jsou probandovi prezentovány uklidňující stimuly. Fází relaxace se začíná i končí, každý proband tedy prochází sedmi experimentálními podmínkami. V každé ze sedmi fází je měřena srdeční frekvence probanda a před začátkem testování je administrován inventář měřící neuroticismus. Výzkumu se zúčastní 62 dobrovolníků. Tabulka 11 zobrazuje ukázku naměřených hodnot ve formátu *wide*.

Tabulka 11: Data ve formátu *wide*

		Srdeční frekvence [bpm]						
Proband	Neurot.	Relax. 1	Stroopův test	Relax. 2	Aritmet. úloha	Relax. 3	Test verb. fluence	Relax. 4
1	10	61	106	75	98	75	92	69
2	2	72	84	79	85	71	75	73
3	12	116	160	149	161	138	148	131
4	14	70	90	67	94	68	97	67
5	16	91	107	83	115	78	118	77
6	17	66	90	65	82	61	88	60
7	21	82	112	104	118	97	120	92
8	4	60	78	65	86	64	79	61
9	23	66	81	58	77	61	78	62
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Hledáme-li odpověď na otázku, jakým způsobem ovlivňuje neuroticismus srdeční frekvenci za rozmanitých okolností, narazíme na obtíže při tvorbě statistického modelu. De facto zde máme 7 závisle proměnných a jediný faktor. Zřejmě bychom si museli pomoct

nějakým průměrováním – třeba tak, že pro každého probanda vypočítáme průměrnou srdeční frekvenci při frustrujících úlohách a průměrnou srdeční frekvenci při relaxaci. Obě proměnné vzájemně odečteme a budeme hledat korelační koeficient mezi tímto rozdílem a neuroticismem. Tento postup však zdaleka nevyužívá všechny informace, které máme, a dosahuje proto jen omezené síly testu.

Vhodnější cestou by mohlo být převedení datové tabulky do formátu *long*. V tomto formátu budou jednotlivé řádky reprezentovat jednotlivá měření tepové frekvence. Tedy každý proband bude v tabulce zastoupen hned sedmi řádky odpovídajícími sedmi experimentálním podmínkám. Data ve formátu *long* znázorňuje tabulka 12. Jak je patrné, takováto tabulka bude mít $7 \cdot 62 = 434$ řádků.

Tabulka 12: Data ve formátu *long*

Proband	Neuroticismus	Fáze	Srdeční frekvence
1	10	Relaxace 1	61
1	10	Stroopův test	106
1	10	Relaxace 2	75
1	10	Aritmet. úloha	98
1	10	Relaxace 3	75
1	10	Test verb. fluence	92
1	10	Relaxace 4	69
2	2	Relaxace 1	72
2	2	Stroopův test	84
2	2	Relaxace 2	79
2	2	Aritmet. úloha	85
2	2	Relaxace 3	71
2	2	Test verb. fluence	75
2	2	Relaxace 4	73
3	12	Relaxace 1	116
3	12	Stroopův test	160
⋮	⋮	⋮	⋮

Podoba lineárního modelu, který by dokázal situaci popsat, je teď již zjevnější. Závisle proměnnou bude sloupec srdeční frekvence a jako regresory použijme neuroticismus, fázi a jejich vzájemnou interakci.

Pokud máte pocit, že jsme se museli dopustit nějakého pochybení, jelikož jsme jen přeformátováním našich dat zvětšili rozsah souboru z 62 na 434 řádků, tak se nemýlíte. Podmínkou každého statistického testu je nezávislost jednotlivých pozorování. Tedy hodnoty na prvním řádku nesmí být nijak svázány s hodnotami na druhém, třetím a na dalších řádcích. V tomto případě tuto podmínku nicméně hrubě porušujeme – každých sedm řádků představuje skupinu a dá se čekat, že pokud jste měli šestkrát vysokou srdeční

frekvenci, budete ji mít vysokou i po sedmé, takže o nějaké nezávislosti nemůže být řeč.

Abychom závislost odstranili, musíme přidat do modelu kategoriální proměnnou proband (celkem 62 úrovní). Tvrdíme tím, že řádky v rámci jednotlivých sedmic mají společné to, že některá sedmice je o něco posunutá k vyšším číslům a jiná k nižším, což pomocí nového regresoru kompenzujeme.

Vyřešením problému nezávislosti bohužel vzniká jiný problém: proměnná neuroticismus a proměnná proband jsou dokonale lineárně závislé (tzn. je zde přítomna úplná multikolinearita). Za těchto okolností neposkytuje metoda nejmenších čtverců žádné řešení. Naštěstí i na tento problém známe lék, a tím jsou takzvané *modely se smíšenými efekty*.

Modely se smíšenými efekty (*mixed-effect models*) nejsou lineární modely v pravém slova smyslu. Jsou výpočetně náročné a opírají se o teorii citelně komplikovanější než klasické lineární modely. Pro použití ve výzkumu stačí znát teoretická východiska jen povrchově – nalezené regresní váhy budeme interpretovat stejně, jak jsme byli doposud zvyklí, přestože jsme k nim došli jinou cestou.

Model se smíšenými efekty obsahuje dva druhy regresorů: pevné (*fixed*) a náhodné (*random*). První zmiňované odpovídají tomu, co jsme se učili o regresorech dříve. Náhodné faktory se však v mnoha věcech odlišují. Náhodný faktor je vždy nominální. **Pokud o nějakém faktoru řekneme, že je náhodný, předpokládáme, že existuje rozsáhlá populace úrovní tohoto faktoru a že velikosti regresních vah těchto úrovní mají v dané populaci normální rozdělení.** My však máme k dispozici jen několik náhodě vylosovaných úrovní. V našem případě bychom za náhodný regresor mohli považovat proměnnou proband. Existuje velká populace lidí, my jsme ale „vylosovali“ jen n z nich. Navíc můžeme očekávat, že rozmanitost srdeční frekvence je mezi různými lidmi normálně rozdělená.

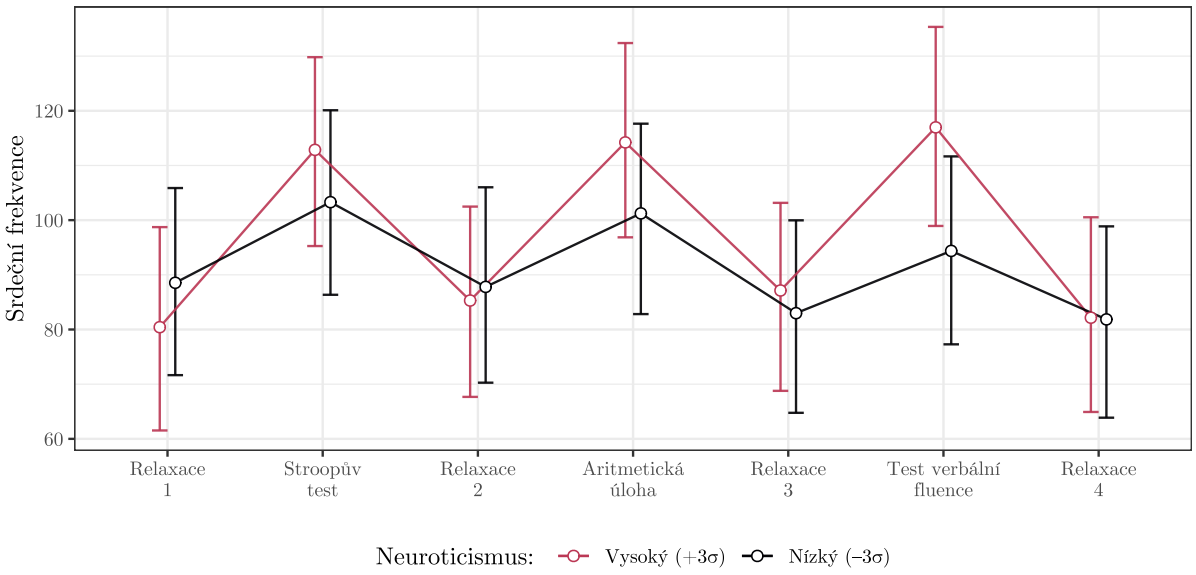
Zásadní výhodou náhodných faktorů je to, že se na ně nevztahuje podmínka nepřítomnosti kolinearity. Váhy náhodných faktorů můžeme odhadovat i v případě, kdy jsou plně lineárně závislé na libovolné skupině pevných nebo náhodných faktorů.

Při výpočtu modelu se smíšenými efekty pracujeme se dvěma druhy chybových rozptylů: odhadujeme nám již známý reziduální rozptyl, ale také odhadujeme rozptyl vah úrovní náhodného efektu. Tedy v našem případě se nesnažíme jen minimalizovat nepřesnost predikce (σ^2_ϵ), ale taky, aby odhad rozptylu srdečních frekvencí ($\sigma^2_{\text{proband}}$) jednotlivých probandů byl co nejmenší.

Při interpretaci výsledků obvykle nehledíme na váhy úrovní náhodného efektu, ale zabýváme se jen váhami pevných efektů. Vyplatí se nicméně věnovat pozornost odhadu parametru σ_{proband}, jelikož ten nám pomůže si udělat představu o tom, jak si stojí pevné efekty ve srovnání s náhodnými efekty. V našem případě jsme interindividuální rozmanitost tepové frekvence odhadli na přibližně 7.5 tepů za minutu. Z výsledků vyplývá, že

rozdíly v tepové frekvenci mezi jedinci skórujícími na opačných koncích spektra neuroticismu se pohybují mezi 10 a 20 body. Můžeme tedy konstatovat, že pozorované efekty dosahují vedle statistické významnosti i významnosti praktické.

Obrázek 24: Srdeční frekvence při jednotlivých podmínkách u probandů s nízkou a s vysokou mírou neuroticismu



Modely se smíšenými efekty mají širší uplatnění, než naznačuje náš příklad. Kromě designů s opakovanými měřeními nachází popsaný postup uplatnění v designech, kdy pracujeme s hierarchicky uspořádanými vzájemně *zahnízděnými* kategoriemi. Typicky by šlo třeba o situaci, kdy pozorujeme žáky z několika různých škol a v rámci každé z těchto škol pracujeme s několika třídami. Regresory škola a třída jsou v takovém případě dokonale závislé, a tedy pomocí metody nejmenších čtverců nezkoumatelné. V těchto designech se často setkáme s označením *hierarchical models* (nezaměňovat s hierarchickou regresí) a *nested models*.

Oceňovanou vlastností modelů se smíšenými efekty je také to, že se elegantně vyrovnávají s chybějícími hodnotami. Například kdyby se v našem příkladu se srdeční frekvencí vyskytovaly poškozené záznamy, které bylo potřeba vyřadit, metodu můžeme beze změny použít. To, zda byl každý proband měřen sedmkrát nebo toto číslo různě kolísá, výsledky nezkresluje.

Seznam použitých symbolů

X	Nezávisle proměná (regresor).
Y	Závisle proměnná.
$\hat{Y}$	Predikovaná (též vyrovnaná) hodnota závisle proměnné.
ϵ	Reziduum. Rozdíl mezi predikcí $\hat{Y}$ a pozorovanou hodnotou Y.
β	Nestandardizovaný regresní koeficient.
β^*	Standardizovaný regresní koeficient.
β_0	Počátek (průměrná hodnota Y, pokud jsou všechny regresory rovny 0).
β_1	V jednoduché regresi parametr sklonu regresní přímky.
n	Počet pozorování (obvykle rozsah souboru).
q	Počet podmínek v modelu s podmínkami.
h	Počet vyřazených regresorů, o které se model liší od svého podmodelu.
k	Počet regresorů.
p	Počet odhadovaných parametrů (obvykle $k+1$). V případě desetinného čísla p-hodnota.
RSČ	Reziduální součet čtverců.
R^2	Koeficient determinace (procento vysvětleného rozptylu závisle proměnné).
ΔR^2	Změna R^2 po přidání či vyloučení regresorů z modelu.
$R^2_{adj.}$	Upravený koeficient determinace.
S^2_ϵ	Odhad reziduálního rozptylu. Též lze zapsat jako $\hat{\sigma}^2_\epsilon$.
σ^2_ϵ	Skutečná hodnota reziduálního rozptylu.
$\mathrm{SČ}_Y$	Součet čtverců odchylek Y od průměru.
F	Statistika F, nebo její odhad, pro test podmodelu.
t	Waldova statistika pro test významnosti regresoru.
H_0	Nulová hypotéza.
α	Hladina významnosti.
sv	Stupně volnosti.
D_j	Cookova vzdálenost.
$\mathbf{H}$	Projekční matice.
h_j	Leverage. Vliv pozorování (j-tý diagonální prvek matice $\mathbf{H}$).
VIF	Variance inflation factor. Ukazatel míry multikolinearity. Její převrácená hodnota se nazývá tolerance.
$\mathbf{x}$	Vektor zadaných hodnot proměnných X pro výpočet predikce.
$\mathbf{X}$	Designová matice. Odpovídá matici všech proměnných X doplněné prvním sloupcem se samými jedničkami.
$\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}})$	Odhad rozptylu varianční matice regresních vah.
$I_{1-\alpha}$	Obecně konfidenční interval (může být pro jeden parametr, pro regresní přímku i kolem regresní přímky).
$P_{1-\alpha}$	Predikční interval.

Doporučená literatura

Řada témat popisovaných v těchto skriptech zůstala nedovysvětlena nebo byla čtenáři předána s notnou dávkou zjednodušení. Pokud toužíte tato bílá místa zaplnit, existuje řada titulů, které tuto úlohu nějakým způsobem dokáží splnit. Jmenujme alespoň několik česky psaných titulů.

Fišerová, E. (2015). *Lineární statistické modely* (2. vydání). Univerzita Palackého v Olomouci. ISBN 978-80-244-4797-1.

- Text poskytuje matematicky přesné definice popisovaných konceptů včetně důkazů.
- Pro porozumění je nezbytná znalost algebry na úrovni nižších ročníků technických a matematických oborů.

Pekár, S., &, Brabec, M. (2020). *Moderní analýza biologických dat 1.* Masarykova univerzita v Brně. ISBN 978-80-210-9622-6.

- Text pokrývá i vybrané nelineární regresní modely.
- Čtenář má možnost pozorovat specifika práce s daty napříč obory (psychologie a biologie).
- Kniha obsahuje návody pro práci se statistickými modely v prostředí jazyka R.
- Od prvního autora též k dispozici druhý a třetí díl, přinášející nespočet dalších postupů.

PhDr. Daniel Dostál, Ph.D.

Lineární statistické modely v psychologii

Odpovědný redaktor a jazyková korektura Otakar Loutocký
Sazbu provedl Daniel Dostál

Vydala Univerzita Palackého v Olomouci, Křížkovského 8, 771 47 Olomouc
vydavatelstvi.upol.cz

1. vydání

Olomouc 2021

DOI: 10.5507/ff.21.24458236

ISBN 978-80-244-5823-6 (online: iPDF)

VUP 2021/0396 (online: iPDF)