Lineární regrese

Matematické metody pro ITS (11MAMY)

Jan Přikryl

s využitím díla G. James et al., *An Introduction to Statistical Learning*

8. přednáška 11MAMY

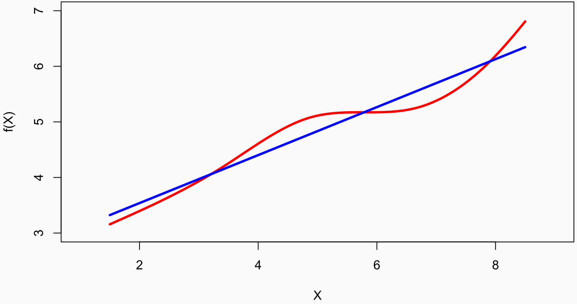
úterý 15. března 2023

verze: 2023-03-13 18:27

Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní

Lineární regrese

- Lineární regrese je jednoduchý přístup k učení s učitelem (supervizovanému učení). Předpokládá, že závislost Y na $X_1, X_2, \ldots, X_p$ (tedy $y = f(\mathbf{x})$ je lineární.
- Skutečné regresní funkce ale nejsou nikdy lineární!



- Ačkoli se může zdát přehnaně zjednodušená, je lineární regrese extrémně užitečná jak svou koncepcí, tak prakticky.

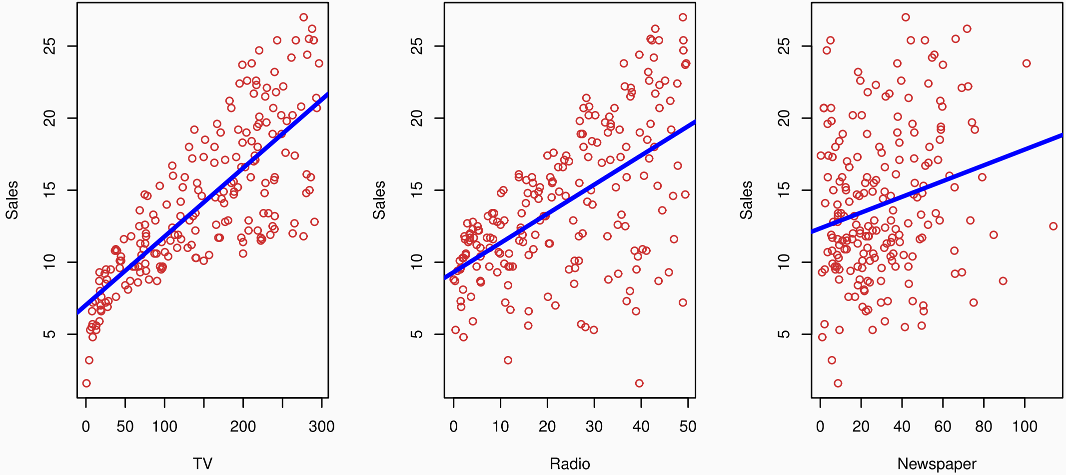
Lineární regrese pro reklamní údaje

Uvažujme reklamní údaje, které ukazuje následující slajd.

Otázky, které si můžeme klást:

- Existuje vztah mezi rozpočtem na reklamu a prodejem?
- Jak silný je vztah mezi rozpočtem na reklamu a prodejem?
- Která média přispívají k prodeji?
- Jak přesně můžeme předpovědět budoucí prodeje?
- Je ten vztah lineární?
- Existuje synergie (efekt společného působení) mezi inzertními médii?

Reklamní údaje



Jednoduchá lineární regrese s jediným prediktorem X

- Budeme uvažovat model
$$Y = \beta_0 + \beta_1 X + \epsilon,$$
kde β_0 a β_1 jsou dvě neznámé konstanty, které představují **regresní konstantu** (absolutní člen) a **sklon** (směrnici), říká se jim také **regresní koeficienty** nebo **parametry**,
- ϵ je chybový člen, často $\epsilon \approx \mathcal{N}(0, \sigma^2)$ (šum modelu).
- Jsou-li dány nějaké odhady $\hat{\beta}_0$ a $\hat{\beta}_1$ koeficientů modelu, předpovídáme budoucí prodeje pomocí vzorce
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$
kde $\hat{y}$ označuje předpověď Y na základě $X = x$, $E[Y|X = x]$. Symbol **stříška** označuje odhadnutou hodnotu.

- Nechť $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ je předpověď Y založená na i-té hodnotě X. Pak $e_i = y_i - \hat{y}_i$ představuje i-té **reziduum**.
- Definujeme **reziduální součet čtverců** (RSS) jako

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

nebo ekvivalentně jako

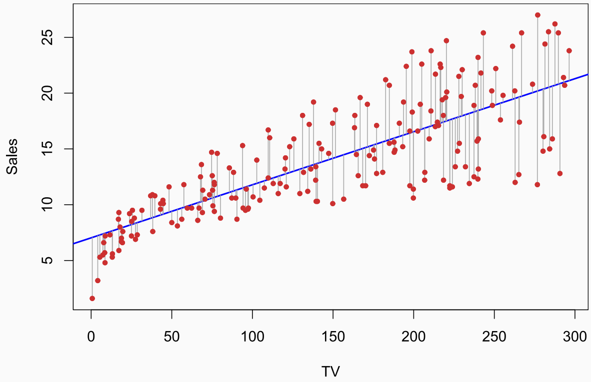
$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- Metoda nejmenších čtverců volí $\hat{\beta}_0$ a $\hat{\beta}_1$ tak, aby hodnota RSS byla minimální. Dá se ukázat, že tyto minimalizující hodnoty jsou

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

kde $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ jsou výběrové průměry.

Příklad: reklamní údaje



Výsledek metody nejmenších čtverců pro regresi položky `sales` vůči `TV`. V tomto případě lineární aproximace zachycuje podstatu vzájemného vztahu, i když na levém konci grafu je poněkud závadná.

- Směrodatná chyba odhadu odráží to, jak se odhad mění při opakovaném vzorkování. Máme

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right],$$

  kde $\sigma^2 = \mathrm{var}(\epsilon)$.
- Tyto směrodatné chyby se mohou použít k výpočtu **intervalů spolehlivosti**. Interval spolehlivosti 95 % se definuje jako takový rozsah hodnot, že s pravděpodobností 95 % bude tento obor obsahovat skutečnou neznámou hodnotu daného parametru. Pro β_1 má tvar

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

Intervaly spolehlivosti – pokračování

Znamená to, že je přibližně 95 % možnost, že interval

$$\left\langle \hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1) \right\rangle$$

bude obsahovat skutečnou hodnotu β_1 (ve scénáři, kdy jsme dostali opakované vzorky jako je současný vzorek).

Pro naše reklamní data je 95 % interval spolehlivosti pro β_1 roven $\langle 0{,}042, 0{,}053\rangle$.

- Směrodatné chyby mohou být také použity k *testování hypotéz* o koeficientech. Nejběžnější test hypotézy je testování **nulové hypotézy** tvaru

  H_0: Mezi X a Y není žádný vzájemný vztah

  vůči **alternativní hypotéze**

  H_A: Existuje nějaký vztah mezi X a Y.
- Matematicky to odpovídá testování

$$H_0 : \beta_1 = 0$$

vůči

$$H_\mathrm{A} : \beta_1 \neq 0,$$

neboť pokud $\beta_1 = 0$, model se redukuje na $Y = \beta_0 + \epsilon$ a X s Y není propojeno.

- K testu nulové hypotézy vypočítáme t-**statistiku** danou vztahem

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}.$$

- Ta bude mít t-rozdělení s $n - 2$ stupni volnosti, za předpokladu, že $\beta_1 = 0$.
- Pomocí statistického softwaru se snadno vypočítá pravděpodobnost, že budeme pozorovat jakoukoli hodnotu rovnou $|t|$ nebo větší. Tato pravděpodobnost se nazývá p-**hodnota**.

Výsledky pro reklamní údaje

	Koeficient	Směr. chyba	t-statistika	p-hodnota
Regr.konst.	7,0325	0,4578	15,36	$< 0{,}0001$
`TV`	0,0475	0,0027	17,67	$< 0{,}0001$

- Vypočítáme **reziduální směrodatnou chybu**

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  kde **reziduální součet čtverců** je $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- **R kvadrát** neboli koeficient determinace je

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}},$$

  kde $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ je **celkový součet čtverců**.
- Dá se ukázat, že v této jednoduché lineární regresní situaci je $R^2 = r^2$, kde r je korelace mezi X a Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2)}}.$$

Výsledky pro reklamní údaje

Veličina	Hodnota
Reziduální směrodatná chyba	3,26
R^2	0,612
F-statistika	312,1

- Náš model zde je

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- Interpretujeme β_j jako *průměrný* vliv jednoho jednotkového růstu X_j na Y, za předpokladu, že **všechny ostatní prediktory se nemění**. V příkladu s reklamou nabývá model tvaru

$$\texttt{prodeje} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{rozhlas} + \beta_3 \times \texttt{noviny} + \epsilon.$$

Interpretace regresních koeficientů

- Ideální scénář je tehdy, když prediktory nejsou korelovány — **vyvážený plán**:
  - Každý koeficient může být odhadnut a testován odděleně.
  - Jsou možné interpretace typu **„jednotková změna v X_j je spojena se změnou Y o β_j, přičemž všechny ostatní proměnné zůstávají beze změny“**.
- Korelace mezi prediktory působí problémy:
  - Rozptyl všech koeficientů má tendenci růst, někdy dramaticky.
  - Interpretace se stávají hazardními — když se změní X_j, změní se všechno ostatní.
- Měli bychom se vyhnout **kauzalitě** v pozorovaných datech.

„Data Analysis and Regression“ Mosteller a Tukey 1977

- Regresní koeficient β_j odhaduje očekávanou změnu Y při jednotkové změně X_j, **přičemž všechny ostatní prediktory jsou zafixovány**. Ale prediktory se obvykle mění společně!

Příklad

Y je celková částka v drobných ve vaší kapse; X_1 je počet mincí; X_2 je počet korun, dvoukorun a pětikorun. Sám o sobě bude regresní koeficient Y vzhledem k X_2 kladný. Ale co se dá říci o X_1 v modelu?

Trápení s interpretací regresních koeficientů

„Data Analysis and Regression" Mosteller a Tukey 1977

- Regresní koeficient β_j odhaduje očekávanou změnu Y při jednotkové změně X_j, **přičemž všechny ostatní prediktory jsou zafixovány**. Ale prediktory se obvykle mění společně!

Příklad

$Y =$ počet zákroků fotbalového hráče během sezóny; W a H jsou jeho váha a výška. Proložený regresní model je $Y = \hat{\beta}_0 + 0{,}50W - 0{,}10H$. Jak interpretujeme $\hat{\beta}_2 < 0$?

Dva citáty známých statistiků

„V zásadě jsou všechny modely špatné, ale některé jsou užitečné.“

George Box

„Jediný způsob, jak zjistit, co se děje při porušení složitého systému, je porušit ten systém a ne jej pouze pasivně pozorovat.“

Fred Mosteller a John Tukey, jako parafráze George Boxe

Odhad a předpověď pro vícenásobnou regresi

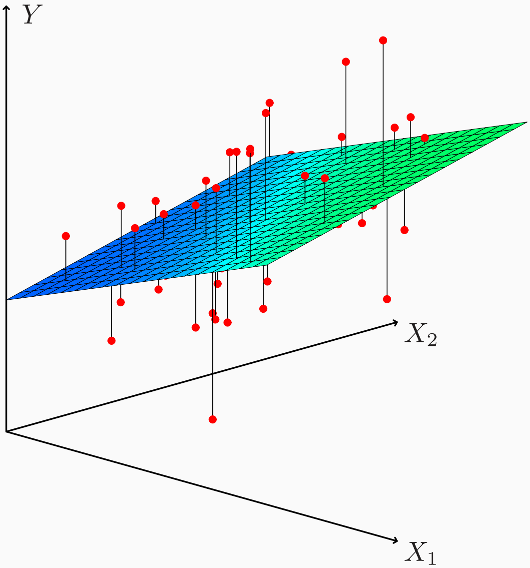
- Při daných odhadech $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ můžeme počítat predikce pomocí vzorce

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_p x_p.$$

- Odhadujeme $\beta_0, \beta_1, \ldots, \beta_p$ jako hodnoty, které minimalizují součet kvadrátů reziduí

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

Dělá se to pomocí standardního statistického softwaru. Hodnoty $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, které minimalizují RSS, jsou odhady regresních koeficientů vícenásobnou metodou nejmenších čtverců.



Y
X_2
X_1

Výsledky pro reklamní data

	Koeficient	Směrod.chyba	t-statistika	p-hodnota
Regr. konst.	2,939	0,3119	9,42	$< 0{,}0001$
`TV`	0,046	0,0014	32,81	$< 0{,}0001$
`rozhlas`	0,189	0,0086	21,90	$< 0{,}0001$
`noviny`	-0,001	0,0059	-0,18	0,8599

Korelace:

	`TV`	`rozhlas`	`noviny`	`prodeje`
`TV`	1,0000	0,0548	0,0567	0,7822
`rozhlas`		1,0000	0,3541	0,5762
`noviny`			1,0000	0,2283
`prodeje`				1,0000

Některé důležité otázky

a) *Je alespoň jeden z prediktorů* $X_1, X_2, \ldots, X_p$ *užitečný při předpovídání odpovědi?*

b) *Pomáhají všechny prediktory vysvětlit* Y, *nebo je užitečná pouze nějaká podmnožina prediktorů?*

c) *Jak dobře model aproximuje data?*

d) *Je-li dán soubor hodnot prediktorů, jakou hodnotu odpovědi bychom měli předpovědět a jak přesná je naše předpověď?*

Je alespoň jeden prediktor užitečný?

Pro první otázku můžeme použít F-statistiku

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}\,/(n - p - 1)} \sim F_{p,n-p-1}$$

Veličina	Hodnota
Reziduální směrodat. chyba	1,69
R^2	0,897
F-statistika	570

Rozhodování o důležitých proměnných

- Nejpřímější přístup se nazývá regrese se **všemi podmnožinami** nebo s **nejlepší podmnožinou**: počítáme aproximace metodou nejlepších čtverců pro všechny možné podmnožiny a pak z nich vybereme pomocí nějakého kritéria, které vyvažuje trénovací chybu s velikostí modelu.
- Často však nemůžeme vyšetřit všechny možné modely, protože je jich 2^p; například pro $p = 40$ existuje přes miliardu modelů!
  Místo toho potřebujeme automatizovaný přístup, který prohledává nějakou jejich podmnožinu. V dalším probereme dva běžně používané přístupy.

Dopředná selekce

- Začni s **nulovým modelem** — modelem, který obsahuje regresní konstantu, ale žádné prediktory.
- Prolož p jednoduchých lineárních regresí a přidej k nulovému modelu tu proměnnou, která vede k nejnižšímu RSS.
- Přidej k tomu modelu proměnnou, která vede k nejnižšímu RSS mezi všemi modely s dvěma proměnnými.
- Pokračuj, dokud není splněno nějaké zastavovací kritérium, například že všechny zbývající proměnné mají p-hodnotu nad nějakou hranicí.

Zpětná eliminace

- Začni se všemi proměnnými v modelu.
- Odeber proměnnou s největší p-hodnotou – to jest proměnnou, která je nejméně statisticky významná.
- Prolož nový model s $p-1$ proměnnými a odeber proměnnou s největší p-hodnotou.
- Pokračuj do splnění nějakého zastavovacího kritéria. Můžeme například zastavit, když všechny zbývající proměnné mají významnou p-hodnotu definovanou jako nějaká hranice významnosti.

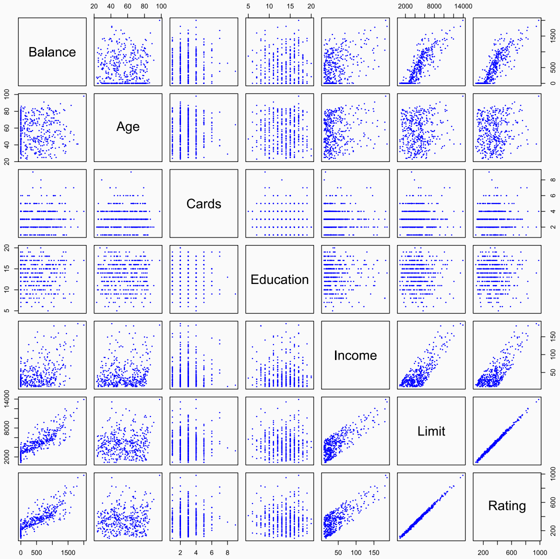
- Později probereme systematičtější kritéria pro volbu „optimálního" modelu na dráze modelů vytvořených postupně dopřednou selekcí nebo zpětnou eliminací.
- Patří k nim **Mallowovo** C_p, **Akaike informační kritérium (AIC)**, **Bayesovské informační kritérium (BIC)**, **upravené** R^2 a **křížová validace (CV)**.

Kvalitativní prediktory

- Některé prediktory nejsou **kvantitativní**, ale jsou **kvalitativní**, nabývají hodnot v diskrétní množině.
- Nazývají se také **kategoriální** prediktory nebo **proměnné faktory**.
- Viz například matici bodových grafů s údaji o kreditních kartách na následujícím slajdu.

  Kromě sedmi kvantitativních proměnných, jež jsou v matici uvedeny, jsou v datech čtyři kvalitativní proměnné: `gender` (pohlaví), `student` (studentský status), `status` (rodinný stav) a `ethnicity` (původ – kavkazský, afroamerický (AA) nebo asijský).

Údaje o kreditních kartách



Příklad (Model s kvalitativními prediktory)

Vyšetřete rozdíl v zůstatku na kreditní kartě mezi muži a ženami, přičemž budete ignorovat ostatní proměnné.

Utvoříme novou proměnnou x_i, rovnou 1, je-li i-tá osoba žena, rovnou 0, je-li i-tá osoba muž. Výsledný model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{je-li } i\text{-tá osoba žena,} \\ \beta_0 + \epsilon_i & \text{je-li } i\text{-tá osoba muž.} \end{cases}$$

Q: Interpretace?

Údaje o kreditních kartách — pokračování

Výsledky pro model podle pohlaví:

	Koef.	SE	t-statistika	p-hodnota
Regresní konst.	509,80	33,13	15,389	$< 0{,}0001$
`gender` [žena]	19,73	46,05	0,429	0,6690

ČVUT ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

- Při více než dvou úrovních utvoříme dodatečné fiktivní proměnné. Tak například pro proměnnou `ethnicity` utvoříme dvě fiktivní proměnné. První by mohla být

$$x_{i1} = \begin{cases} 1 & \text{je-li } i\text{-tá osoba asijského původu,} \\ 0 & \text{není-li } i\text{-tá osoba asijského původu,} \end{cases}$$

a druhá by mohla být

$$x_{i2} = \begin{cases} 1 & \text{je-li } i\text{-tá osoba kavkazského původu,} \\ 0 & \text{není-li } i\text{-tá osoba kavkazského původu.} \end{cases}$$

- Pak mohou být v regresní rovnici použity obě tyto proměnné, takže dostaneme model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{je-li } i\text{-tá osoba Asiat} \\ \beta_0 + \beta_2 + \epsilon_i & \text{je-li } i\text{-tá osoba běloch} \\ \beta_0 + \epsilon_i & \text{je-li } i\text{-tá osoba Afroameričan.} \end{cases}$$

- Fiktivních proměnných bude vždy o jednu méně než je počet úrovní. Úroveň bez fiktivní proměnné — v tomto příkladu Afroameričani — je známa jako **výchozí úroveň**.

Výsledky pro etnickou příslušnost

	Koef.	SE	t-statistika	p-hodnota
Regresní konst.	531,00	46,32	11,464	$< 0{,}0001$
`ethnicity` [asijská]	-18,69	65,02	-0,287	0,7740
`ethnicity` [kavkazská]	-12,50	56,68	-0,221	0,8260

Odstraníme předpoklad aditivity: **interakce** a **nelinearita**

Interakce:

- V naší předchozí analýze reklamních dat jsme předpokládali, že vliv zvýšení prostředků jednoho reklamního média na `sales` (prodeje) nezávisí na objemu prostředků vynaložených na zbylá média.
- Tak například, lineární model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

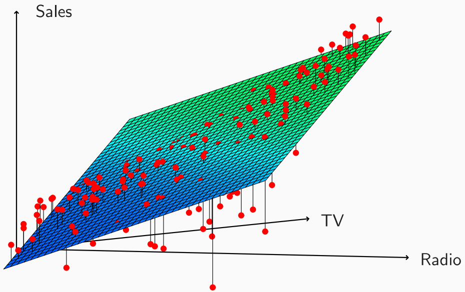
říká, že průměrný efekt jednotkového vzrůstu reklamních nákladů v `TV` na `sales` je vždy β_1, nezávisle na množství prostředků vynaložených na `radio`.

Interakce

ČVUT ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

- Ale předpokládejme, že peníze vynaložené na rozhlasovou reklamu ve skutečnosti zvyšují efektivitu TV reklamy, takže koeficient sklonu pro `TV` by měl s růstem hodnoty `radio` růst.
- V této situaci, máme-li dán pevný rozpočet $\$100\,000$, investice poloviny do `radio` a poloviny do `TV` může zvýšit `sales` více, než použití celé částky na `TV` nebo na `radio`.
- V marketingu se tomuhle říká efekt **synergie**, ve statistice se o tom mluví jako o efektu **interakce**.

Interakce v reklamních datech?



Když je úroveň `TV` nebo `radio` nízká, pak jsou skutečné hodnoty `sales` nižší, než jak předpovídá lineární model.

Ale když se reklama rozdělí mezi tato dvě média, pak má model tendenci `sales` podhodnocovat.

Model nabývá tvaru

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.\end{aligned}$$

Výsledky:

	Koef.	SE	t-statistika	p-hodnota
Regres. konst.	6,7502	0,248	27,23	$< 0{,}0001$
TV	0,0191	0,002	12,70	$< 0{,}0001$
radio	0,0289	0,009	3,24	0,0014
TV × radio	0,0011	0,000	20,73	$< 0{,}0001$

- Výsledky v této tabulce naznačují, že interakce jsou důležité.
- p-hodnota pro interakční člen `TV` $\times$ `radio` je extrémně nízká, což ukazuje, že jsou tu silné náznaky platnosti $H_A : \beta_3 \neq 0$.
- Hodnota R^2 pro interakční model je 96.8 % ve srovnání s pouhými 89.7 % u modelu, který předpovídá `sales` pomocí `TV` a `radio` bez interakčního členu.
- To znamená, že $(96{,}8 - 89{,}7)/(100 - 89{,}7) = 69\,\%$ variability v proměnné `sales`, která zůstává po proložení aditivního modelu, se vysvětlilo interakčním členem.

Odhady koeficientů v tabulce naznačují, že

- zvýšení prostředků na reklamu v televizi o $1 000 je spojeno se zvýšením prodejů o $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1{,}1 \times \texttt{radio}$ jednotek,
- zvýšení prostředků na reklamu v rozhlase o $1 000 bude spojeno s růstem prodejů o $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1{,}1 \times \texttt{TV}$ jednotek.

Hierarchie

- Někdy se stane, že interakční člen má velmi malou p-hodnotu, ale přidružené hlavní efekty (v tomto případě `TV` a `radio`) nikoliv.
- **Princip hierarchie:**
  Pokud zahrneme do modelu nějakou interakci, měli bychom také zahrnout hlavní efekty, a to i tehdy, nejsou-li p-hodnoty spojené s jejich koeficienty významné.

Hierarchie – pokračování

- Odůvodněním tohoto principu je skutečnost, že interakce se v modelu bez hlavních efektů obtížně interpretují – jejich smysl se změní.
- Speciálně, interakční členy také obsahují hlavní efekty i tehdy, když model nemá členy s hlavními efekty.

Interakce mezi kvalitativními a kvantitativními proměnnými

Uvažujme soubor dat `Credit` a předpokládejme, že chceme předpovědět `balance` na základě `income` (kvantitativní) a `student` (kvalitativní).

Bez interakčního členu bude model mít tvar

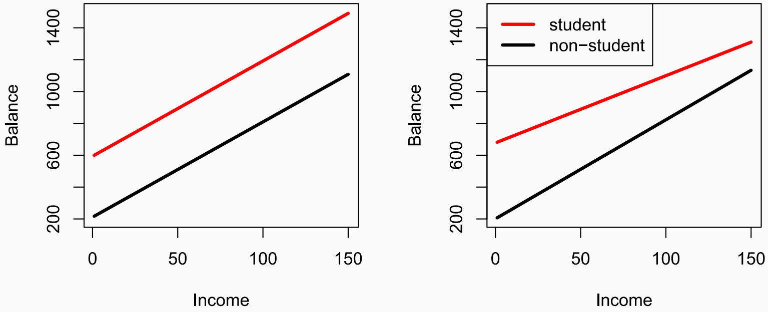
$$
\begin{aligned}
\texttt{balance}_i &= \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{je-li } i\text{-tá osoba student,} \\ 0 & \text{není-li } i\text{-tá osoba student,} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{je-li } i\text{-tá osoba student,} \\ \beta_0 & \text{není-li } i\text{-tá osoba student.} \end{cases}
\end{aligned}
$$

Uvažujme soubor dat `Credit` a předpokládejme, že chceme předpovědět `balance` na základě `income` (kvantitativní) a `student` (kvalitativní).

S interakcemi bude model mít tvar

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{je-li student} \\ 0 & \text{není student} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{je-li } i\text{-tá osoba student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{není-li student} \end{cases}
\end{aligned}$$

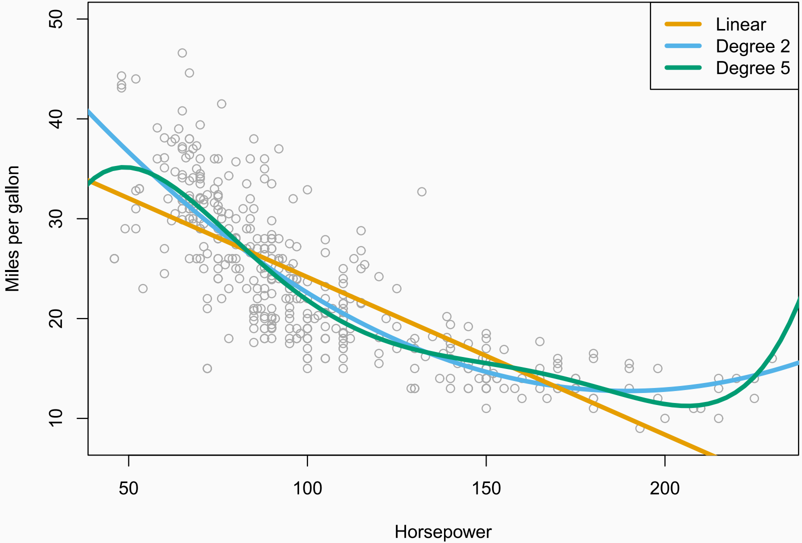
Kreditní data podle příjmu a studia



Vlevo: žádná interakce mezi `income` a `student`.

Vpravo: s interakčním členem mezi `income` a `student`.

Nelineární efekty prediktorů



Obrázek naznačuje, že

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

může dávat lepší aproximaci.

	Koeficient	SE	t-statistika	p-hodnota
Regres. konst.	56,9001	1,8004	31,6	$< 0,0001$
horsepower	-0,4662	0,0311	-15,0	$< 0,0001$
$\texttt{horsepower}^2$	0,0012	0,0001	10,1	$< 0,0001$