Forecasting: Principles and Practice (2. vyd.)

Kapitola 5 Regresní modely časových řad

V této kapitole se zabýváme regresními modely. Základním konceptem je, že předpovídáme časové řady zájmu y za předpokladu, že má lineární vztah s jinými časovými řadami x.

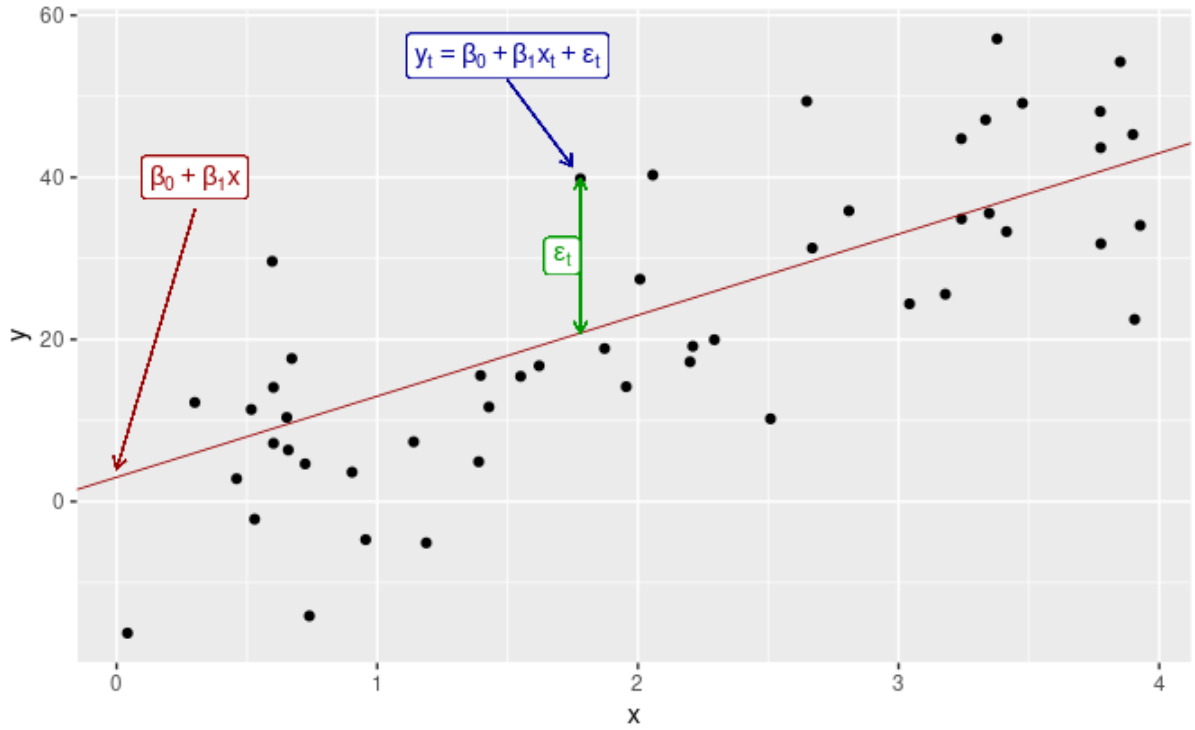
Můžeme například chtít předpovídat měsíční tržby y pomocí celkových výdajů na reklamu x jako prediktoru. Nebo můžeme předpovídat denní poptávku po elektřině pomocí teploty (x_1) a dne v týdnu (x_2) jako prediktorů.

Proměnná prognózy y se někdy také nazývá regresní, závislá nebo vysvětlená proměnná. Proměnné **prediktoru** x se někdy také nazývají regresory, nezávislé nebo vysvětlující proměnné. V této knize je budeme vždy označovat jako proměnnou "prognózy" a proměnné "prediktoru".

5.1 Lineární model

Jednoduchá lineární regrese

V nejjednodušším případě regresní model umožňuje lineární vztah mezi předpovědní proměnnou y a jedinou prognostorovou proměnnou x: $$ y_t = \beta_0 + \beta_1 x_t + \varepsilon_t. $$ Umělý příklad dat z takového modelu je znázorněn na obrázku 5.1.
Koeficienty β_0 a β_1 označují průsečík a sklon přímky. Intercept β_0 představuje předpokládanou hodnotu y když $x=0$. Sklon β_1 představuje průměrnou předpokládanou změnu v y vyplývající z nárůstu o jednu jednotku v x.



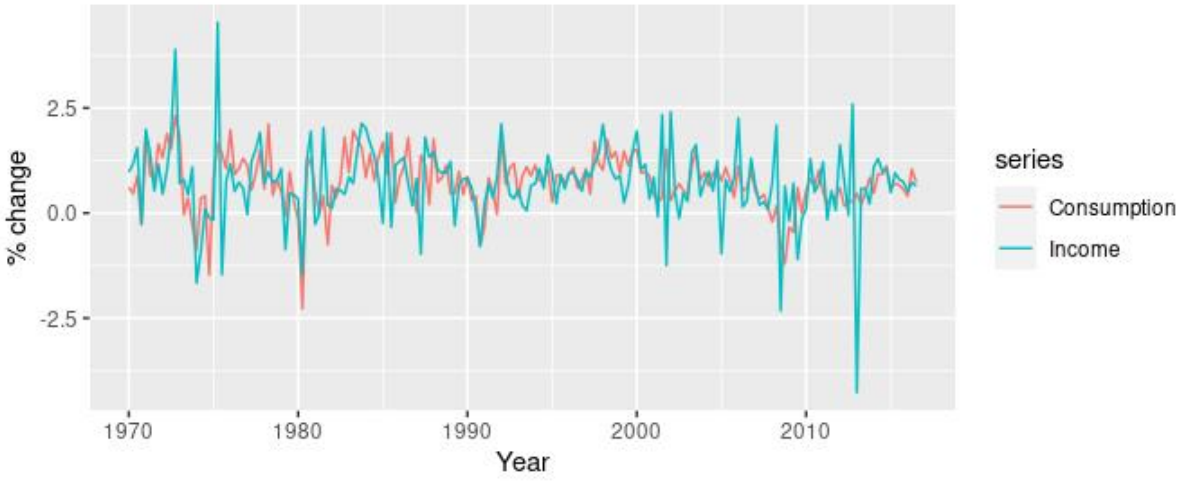
Obrázek 5.1: Příklad dat z jednoduchého lineárního regresního modelu.

Všimněte si, že pozorování neleží na přímce, ale jsou rozptýlena kolem ní. Každé pozorování si můžeme představit \(y_t\) jako skládající se ze systematické nebo vysvětlené části modelu, \(\beta_0+\beta_1x_t\)a náhodné "chyby" \(\varepsilon_t\). Termín "chyba" neznamená chybu, ale odchylku od základního modelu přímky. Zachycuje vše, co může ovlivnit \(y_t\) jiné než \(x_t\).

Příklad: Výdaje na spotřebu v USA

Obrázek 5.2 ukazuje časové řady čtvrtletních procentních změn (tempa růstu) reálných výdajů na osobní spotřebu, \(y\)a reálného osobního disponibilního důchodu, \(x\), pro USA od 1. čtvrtletí 1970 do 3. čtvrtletí 2016.

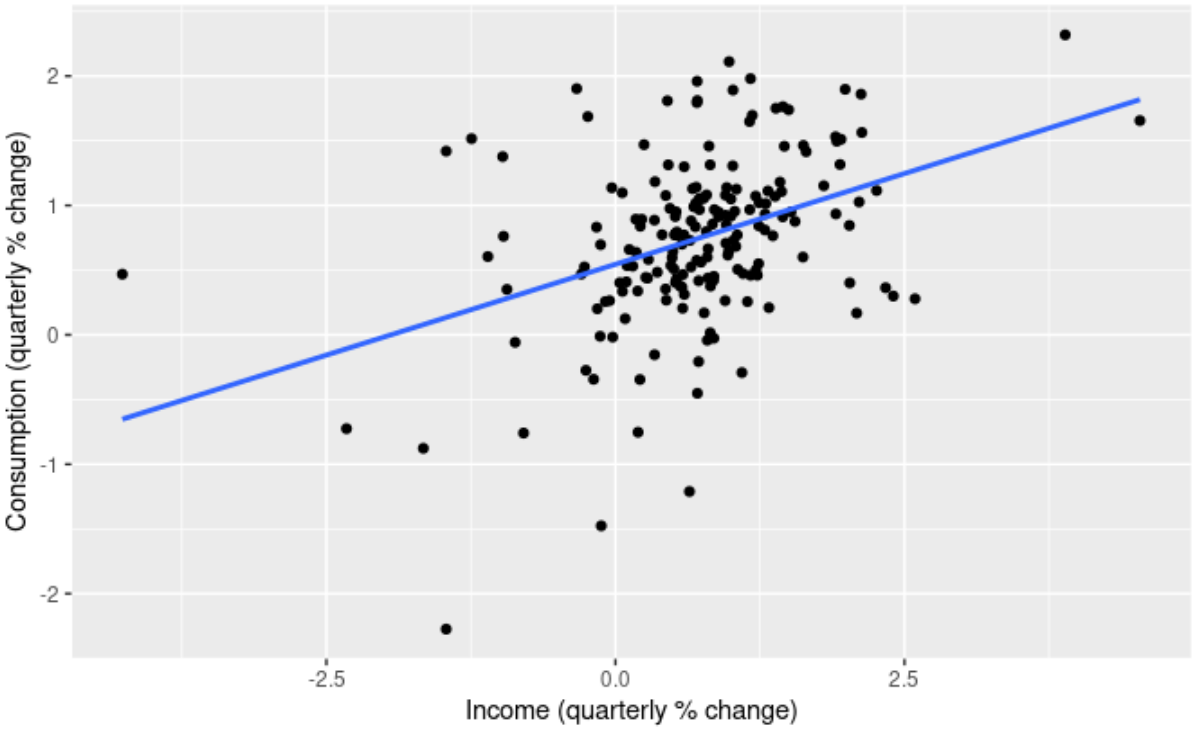
```
autoplot(uschange[,c("Consumption","Income")]) +
  ylab("% change") + xlab("Year")
```



Obrázek 5.2: Procentuální změny výdajů na osobní spotřebu a osobních příjmů v USA.

Rozptylový graf změn spotřeby proti změnám příjmů je znázorněn na obrázku 5.3 spolu s odhadovanou regresní přímkou \[\hat{y}_t=0,55 + 0.28x_t. \] (Nad \(y\) umístíme "klobouk", který označuje, že se jedná o hodnotu \(y\) předpovězenou modelem.)

```
uschange %>%
  as.data.frame() %>%
  ggplot(aes(x=Income, y=Consumption)) +
    ylab("Consumption (quarterly % change)") +
    xlab("Income (quarterly % change)") +
    geom_point() +
    geom_smooth(method="lm", se=FALSE)
#> `geom_smooth()` using formula 'y ~ x'
```



Obr. 5.3: Rozptyl čtvrtletních změn spotřebních výdajů oproti čtvrtletním změnám osobních důchodů a přizpůsobené regresní linii.

Rovnice se odhaduje v R pomocí funkce:`tslm()`

```
tslm(Consumption ~ Income, data=uschange)
#>
#> Call:
#> tslm(formula = Consumption ~ Income, data = uschange)
#>
#> Coefficients:
#> (Intercept)       Income
#>       0.545        0.281
```

O tom, jak výpočet koeficientů pojednáme, budeme diskutovat v části 5.2.`tslm()`

Osazená linka má kladný sklon, který odráží pozitivní vztah mezi příjmem a spotřebou. Koeficient sklonu ukazuje, že zvýšení o jednu jednotku v x (zvýšení osobního disponibilního příjmu o 1 procentní bod) vede v průměru k nárůstu o 0,28 jednotky v y (průměrný nárůst výdajů na osobní spotřebu o 0,28 procentního bodu). Alternativně odhadovaná rovnice ukazuje, že hodnota 1 pro x (procentní nárůst osobního disponibilního příjmu) bude mít za následek předpokládanou hodnotu $0,55 + 0,28 \krát 1 = 0,83$ pro y (procentní nárůst výdajů na osobní spotřebu).

Interpretace průsečíku vyžaduje, aby hodnota $x=0$ dávala smysl. V tomto případě, kdy $x=0$ (tj. když od posledního čtvrtletí nedošlo ke změně osobního disponibilního příjmu), je předpokládaná hodnota y 0,55 (tj. průměrný nárůst výdajů na osobní spotřebu o 0,55%). I když $x=0$ nedává smysl, zachycení je důležitou součástí

modelu. Bez ní může být koeficient sklonu zbytečně zkreslen. Zachycení by mělo být vždy zahrnuto, pokud není požadavkem vynutit regresní čáru "přes počátek". V následujícím textu předpokládáme, že v modelu je vždy zahrnuta odposlech.

Vícenásobná lineární regrese

Pokud existují dvě nebo více proměnných prediktoru, model se nazývá **vícenásobný regresní model**. Obecná forma modelu s vícenásobnou regresí je $$y_t = \beta_{0} + \beta_{1} x_{1,t} + \beta_{2} x_{2,t} + \cdots + \beta_{k} x_{k,t} + \varepsilon_t, \tag{5.1}$$ kde y je proměnná, která má být předpovězena, a $x_{1},\dots,x_{k}$ jsou proměnné prediktoru k. Každá z proměnných prediktoru musí být číselná. Koeficienty $\beta_{1},\dots,\beta_{k}$ měří účinek každého prediktoru po zohlednění účinků všech ostatních prediktorů v modelu. Koeficienty tedy měří *mezní účinky* proměnných prediktoru.

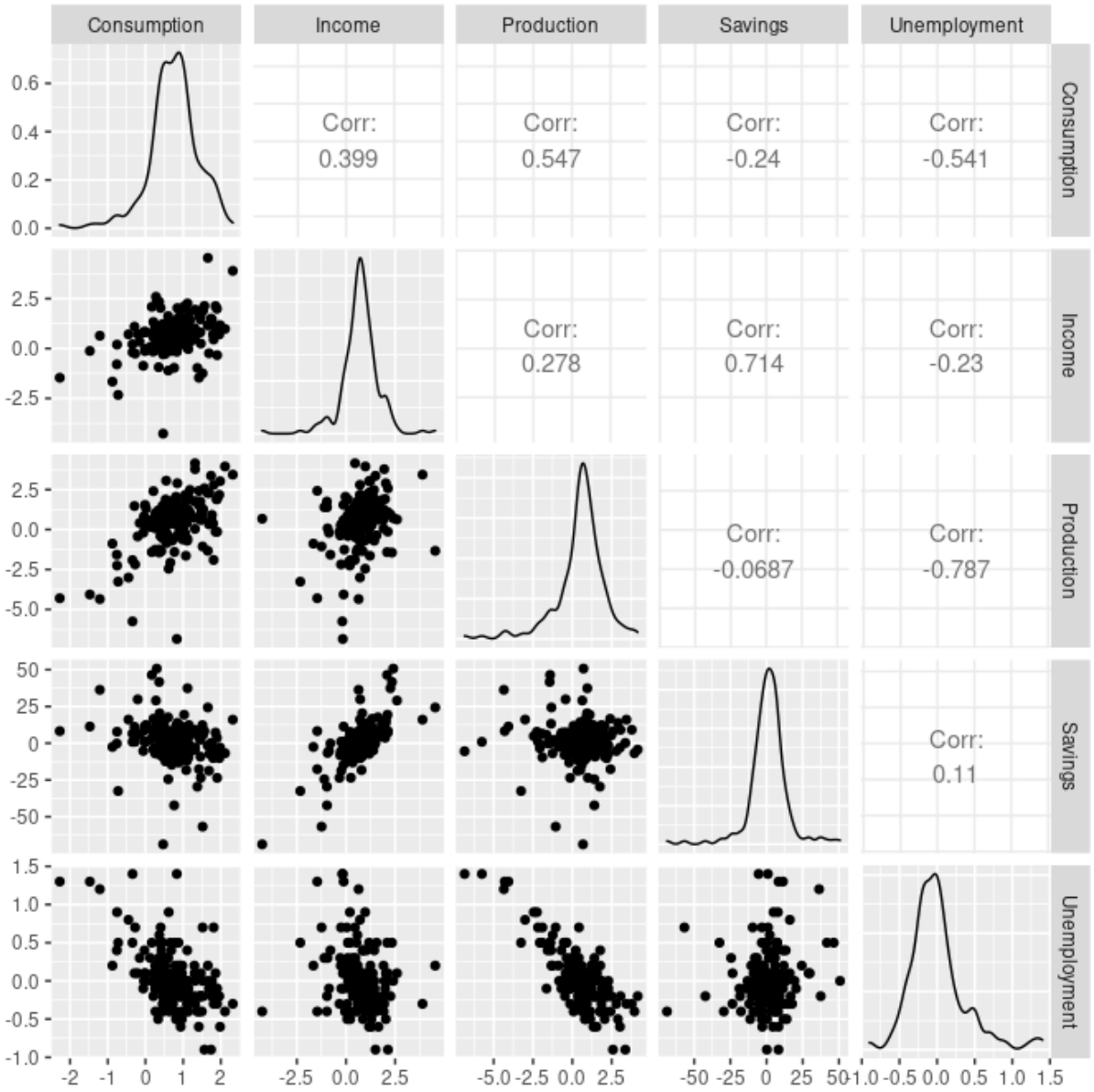
Příklad: Výdaje na spotřebu v USA

Obrázek 5.4 ukazuje další prediktory, které mohou být užitečné pro prognózu výdajů na spotřebu v USA. Jedná se o čtvrtletní procentní změny v průmyslové výrobě a osobních úsporách a čtvrtletní změny míry nezaměstnanosti (protože to je již procento). Vytvoření vícenásobného lineárního regresního modelu může potenciálně generovat přesnější prognózy, protože očekáváme, že výdaje na spotřebu budou záviset nejen na osobním příjmu, ale i na dalších prediktorech.

Obrázek 5.4: Čtvrtletní procentní změny průmyslové výroby a osobních úspor a čtvrtletní změny míry nezaměstnanosti v USA v období 1. až 3. 1. 2016.

Obrázek 5.5 je matice scatterplot pěti proměnných. První sloupec zobrazuje vztahy mezi proměnnou prognózy (spotřeba) a každým z prediktorů. Rozptylové grafy ukazují pozitivní vztah k příjmu a průmyslové výrobě a negativní vztahy s úsporami a nezaměstnaností. Síla těchto vztahů je znázorněna korelačními koeficienty v prvním řádku. Zbývající scatterploty a korelační koeficienty ukazují vztahy mezi prediktory.

```
uschange %>%
  as.data.frame() %>%
  GGally::ggpairs()
```



Obrázek 5.5: Matice rozptylových grafů amerických spotřebních výdajů a čtyř prediktorů.

Předpoklady

Když použijeme lineární regresní model, implicitně vytváříme některé předpoklady o proměnných v rovnici (5.1).

Za prvé, předpokládáme, že model je rozumnou aproximací k realitě; to znamená, že vztah mezi předpovědní proměnnou a proměnnými prediktoru splňuje tuto lineární rovnici.

Za druhé, děláme následující předpoklady o chybách $(\varepsilon_{1},\dots,\varepsilon_{T})$:

- mají střední nulu; jinak budou prognózy systematicky zkreslené.

- nejsou autokorelované; jinak budou prognózy neefektivní, protože v datech je více informací, které lze využít.
- nesouvisejí s proměnnými prediktoru; jinak by bylo více informací, které by měly být zahrnuty do systematické části modelu.

Je také užitečné, aby chyby byly normálně distribuovány s konstantním rozptylem \(\sigma^2\), aby bylo možné snadno vytvářet intervaly predikce.

Dalším důležitým předpokladem v lineárním regresovém modelu je, že každý prediktor \(x\) není náhodná veličina. Pokud bychom prováděli kontrolovaný experiment v laboratoři, mohli bychom kontrolovat hodnoty každého \(x\) (takže by nebyly náhodné) a pozorovat výsledné hodnoty \(y\). S pozorovacími daty (včetně většiny dat v podnikání a ekonomice) není možné kontrolovat hodnotu \(x\), jednoduše ji pozorujeme. Proto z toho děláme předpoklad.

5.2 Odhad nejmenších čtverců

V praxi samozřejmě máme sbírku pozorování, ale neznáme hodnoty koeficientů \(\beta_0,\beta_1, \dots, \beta_k\). Ty je třeba odhadnout z údajů.

Princip nejmenších čtverců poskytuje způsob efektivního výběru koeficientů minimalizací součtu druhých mocnin. To znamená, že vybereme hodnoty \(\beta_0, \beta_1, \dots, \beta_k\), které minimalizují \[\sum_{t=1}^T \varepsilon_t^2 = \sum_{t=1}^T (y_t - \beta_{0} - \beta_{1} x_{1,t} - \beta_{2} x_{2,t} - \cdots - \beta_{k} x_{k,t})^2. \]

To se nazývá odhad **nejmenších čtverců,** protože dává nejmenší hodnotu součtu druhých mocnin chyb. Nalezení nejlepších odhadů koeficientů se často nazývá "přizpůsobení" modelu datům nebo někdy "učení" nebo "trénování" modelu. Čára znázorněná na obrázku 5.3 byla získána tímto způsobem.

Když odkazujeme na *odhadované* koeficienty, použijeme zápis \(\hat\beta_0, \dots, \hat\beta_k\). Rovnice pro ně budou uvedeny v části 5.7.

Funkce přizpůsobuje lineární regresní model datům časových řad. Je podobná funkci, která je široce používána pro lineární modely, ale poskytuje další zařízení pro manipulaci s časovými řadami.`tslm()` `lm()` `tslm()`

Příklad: Výdaje na spotřebu v USA

Vícenásobný lineární regresní model pro spotřebu v USA je \[y_t=\beta_0 + \beta_1 x_{1,t}+ \beta_2 x_{2,t}+ \beta_3 x_{3,t}+ \beta_4 x_{4,t}+\varepsilon_t, \] kde \(y\) je procentní změna reálných výdajů na osobní spotřebu, \(x_1\) je procentní změna reálného osobního disponibilního důchodu, (x_2) je procentní změna průmyslové výroby, (x_3) je procentní změna osobních úspor a (x_4) je změna míry nezaměstnanosti.

Následující výstup poskytuje informace o namontovaného modelu. První sloupec poskytuje odhad každého koeficientu \(\beta\) a druhý sloupec udává jeho standardní chybu (tj. směrodatnou odchylku, která by byla získána opakovaným odhadem koeficientů \(\beta\) na podobných datových souborech). Standardní chyba udává míru nejistoty v odhadovaném koeficientu \(\beta\).`Coefficients`

```
fit.consMR <- tslm(
  Consumption ~ Income + Production + Unemployment + Savings,
  data=uschange)
summary(fit.consMR)
#>
#> Call:
#> tslm(formula = Consumption ~ Income + Production + Unemployment +
#>     Savings, data = uschange)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.8830 -0.1764 -0.0368  0.1525  1.2055
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   0.26729    0.03721    7.18  1.7e-11 ***
#> Income        0.71448    0.04219   16.93  < 2e-16 ***
#> Production    0.04589    0.02588    1.77    0.078 .
#> Unemployment -0.20477    0.10550   -1.94    0.054 .
#> Savings      -0.04527    0.00278  -16.29  < 2e-16 ***
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.329 on 182 degrees of freedom
#> Multiple R-squared:  0.754,  Adjusted R-squared:  0.749
#> F-statistic:  139 on 4 and 182 DF,  p-value: <2e-16
```

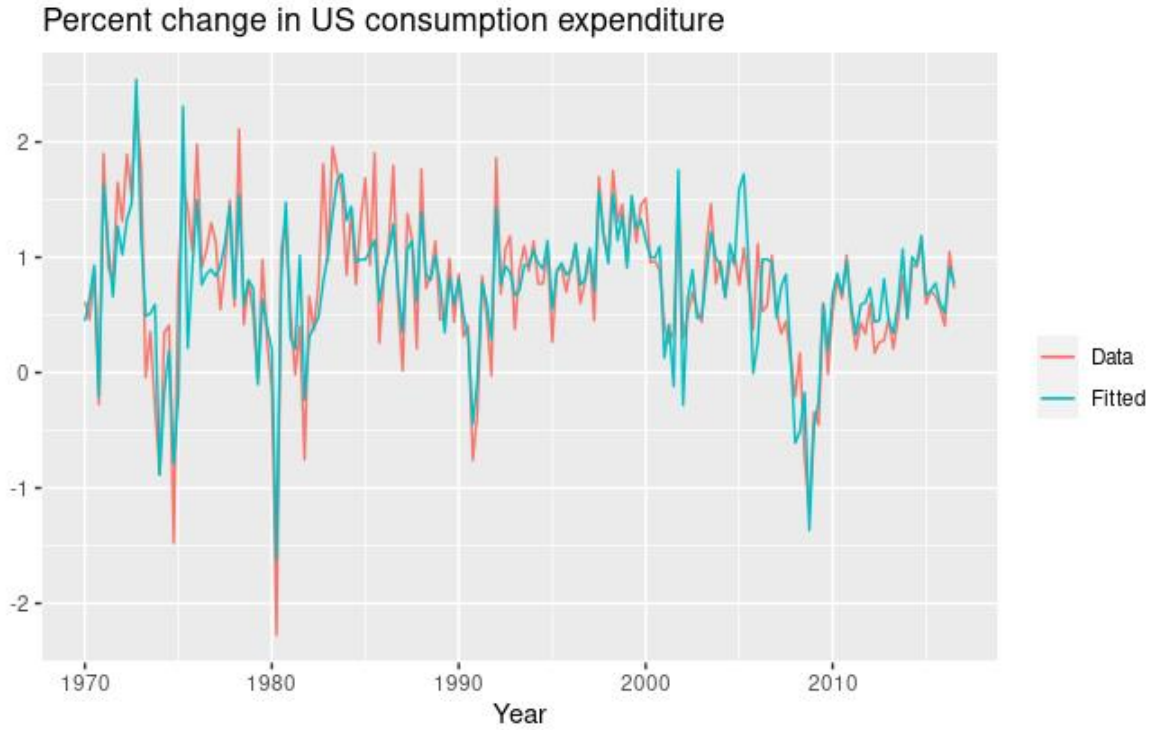
Pro účely prognózování jsou poslední dva sloupce jen omezeně zajímavé. "Hodnota t" je poměr odhadovaného koeficientu \(\beta\) k jeho standardní chybě a poslední sloupec udává hodnotu p: pravděpodobnost, že odhadovaný koeficient \(\beta\) bude tak velký, jako je, pokud neexistuje žádný skutečný vztah mezi spotřebou a odpovídajícím prediktorem. To je užitečné při studiu účinku každého prediktoru, ale není zvláště užitečné pro prognózování.

Montované hodnoty

Předpovědi \(y\) lze získat použitím odhadovaných koeficientů v regresní rovnici a nastavením chybového členu na nulu. Obecně píšeme: \[\begin{equation} \hat{y}_t = \hat\beta_{0} + \hat\beta_{1} x_{1,t} + \hat\beta_{2} x_{2,t} + \cdots + \hat\beta_{k} x_{k,t}. \tag{5.2} \end{equation}\] Zapojení hodnot \(x_{1,t},\dots,x_{k,t}\) pro \(t=1,\dots,T\) vrátí předpovědi \(y_t\) v rámci tréninkového vzorku, označované jako *montované hodnoty*. Všimněte si, že se jedná o předpovědi dat použitých k odhadu modelu, nikoli o skutečné předpovědi budoucích hodnot \ (y \).

Následující grafy ukazují skutečné hodnoty ve srovnání s hodnotami pro procentuální změnu v řadě výdajů na spotřebu v USA. Časový graf na obrázku 5.6 ukazuje, že osazené hodnoty sledují skutečná data poměrně pečlivě. To je ověřeno silným pozitivním vztahem, který ukazuje rozptylový graf na obrázku 5.7.

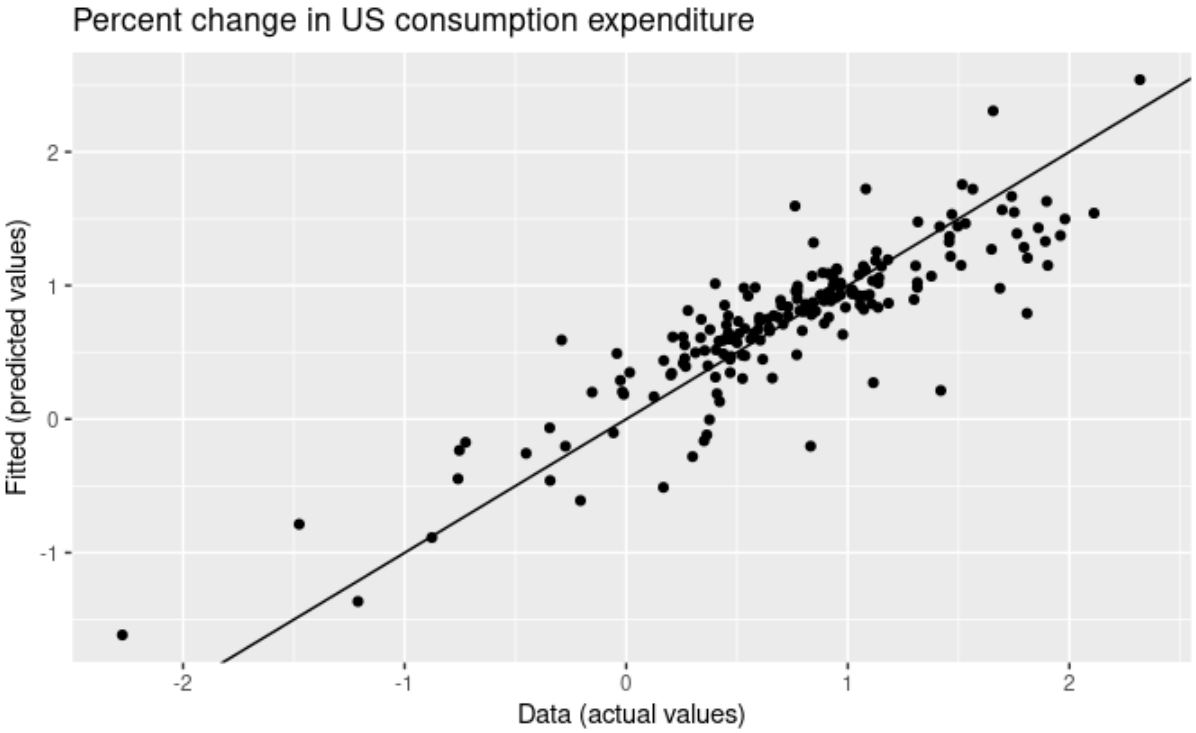
```
autoplot(uschange[,'Consumption'], series="Data") +
  autolayer(fitted(fit.consMR), series="Fitted") +
  xlab("Year") + ylab("") +
  ggtitle("Percent change in US consumption expenditure") +
  guides(colour=guide_legend(title=" "))
```



Obrázek 5.6: Časový graf skutečných spotřebních výdajů v USA a předpokládaných spotřebních výdajů USA.

```
cbind(Data = uschange[,"Consumption"],
      Fitted = fitted(fit.consMR)) %>%
```

```
as.data.frame() %>%
ggplot(aes(x=Data, y=Fitted)) +
  geom_point() +
  ylab("Fitted (predicted values)") +
  xlab("Data (actual values)") +
  ggtitle("Percent change in US consumption expenditure") +
  geom_abline(intercept=0, slope=1)
```



Obrázek 5.7: Skutečné výdaje na spotřebu v USA vynesené proti předpokládaným spotřebním výdajům USA.

Dobrá shoda

Obvyklý způsob, jak shrnout, jak dobře lineární regresní model odpovídá datům, je pomocí koeficientu stanovení nebo \ (R ^ 2 \). To lze vypočítat jako druhou mocninu korelace mezi pozorovanými hodnotami \(y\) a předpokládanými hodnotami \(\hat{y}\). Alternativně může být také vypočtena jako, \[R^2 = \frac{\sum(\hat{y}_{t} - \bar{y})^2}{\sum(y_{t}-\bar{y})^2}, \], kde součty jsou nad všemi pozorováními. Odráží tedy podíl změn ve proměnné prognózy, který je zohledněn (nebo vysvětlen) regresním modelem.

V jednoduché lineární regresi je hodnota \(R^2\) také rovna čtverci korelace mezi \(y\) a \(x\) (za předpokladu, že byl zahrnut průsečík).

Pokud se předpovědi blíží skutečným hodnotám, očekávali bychom, že \(R^2\) se bude blížit 1. Na druhou stranu, pokud předpovědi nesouvisejí se skutečnými hodnotami, pak \(R^2=0\) (opět za předpokladu, že existuje zachycení). Ve všech případech leží \(R^2\) mezi 0 a 1.

Hodnota \(R^2\) se při prognózování používá často, i když často nesprávně. Hodnota \(R^2\) se při přidání dalšího prediktoru do modelu nikdy nesníží, což může vést k nadměrnému přizpůsobení. Neexistují žádná nastavená pravidla pro to, co je dobrá hodnota \(R^2\) a typické hodnoty \(R^2\) závisí na typu použitých dat. Ověření výkonu prognóz modelu na testovacích datech je mnohem lepší než měření hodnoty \(R^2\) na trénovacích datech.

Příklad: Výdaje na spotřebu v USA

Obrázek 5.7 vykresluje hodnoty skutečných výdajů na spotřebu oproti hodnotám. Korelace mezi těmito proměnnými je \(r=0,868\) tedy \(R^2= 0,754\) (zobrazeno ve výstupu výše). V tomto případě model odvádí vynikající práci, protože vysvětluje 75,4% odchylek v údajích o spotřebě. Porovnejte to s hodnotou \ (R^ 2 \) 0,16 získanou z jednoduché regrese se stejnou datovou sadou v oddílu 5.1. Přidání tří dalších prediktorů umožnilo vysvětlit mnohem více odchylek v údajích o spotřebě.

Standardní chyba regrese

Dalším měřítkem toho, jak dobře model odpovídal datům, je směrodatná odchylka reziduí, která je často známá jako "zbytková standardní chyba". To je ukázáno ve výše uvedeném výstupu s hodnotou 0,329.

Vypočítá se pomocí \[\begin{equation} \hat{\sigma}_e=\sqrt{\frac{1}{T-k-1}\sum_{t=1}^{T}{e_t^2}}, \tag{5.3} \end{equation}\], kde \(k\) je počet prediktorů v modelu. Všimněte si, že dělíme \(T-k-1\), protože jsme odhadli \(k+1\) parametry (průsečík a koeficient pro každou proměnnou prediktoru) při výpočtu reziduí.

Standardní chyba souvisí s velikostí průměrné chyby, kterou model produkuje. Tuto chybu můžeme porovnat se střední hodnotou vzorku \(y\) nebo se směrodatnou odchylkou \(y\), abychom získali určitý pohled na přesnost modelu.

Standardní chyba bude použita při generování predikčních intervalů, které jsou popsány v kapitole 5.6.

5.3 Vyhodnocení regresního modelu

Rozdíly mezi pozorovanými hodnotami \(y\) a odpovídajícími hodnotami \(\hat{y}\) jsou chyby trénovací sady nebo "rezidua" definované jako \[\begin{align*} e_t &= y_t - \hat{y}_t \\ &= y_t - \hat\beta_{0} - \hat\beta_{1} x_{1,t} - \hat\beta_{2} x_{2,t} - \cdots - \hat\beta_{k} x_{k, t} \end{align*}\] pro \(t=1,\dots,T\). Každý zbytek je nepředvídatelnou složkou souvisejícího pozorování.

Rezidua mají některé užitečné vlastnosti, včetně následujících dvou: \[\sum_{t=1}^{T}{e_t}=0 \quad\text{and}\quad \sum_{t=1}^{T}{x_{k,t}e_t}=0\qquad\text{for all k}. \] V důsledku těchto vlastností je jasné, že průměr reziduí je nulový a že korelace mezi rezidui a pozorováními pro proměnnou prediktoru je také nulová. (To nemusí být nutně pravda, pokud je zachycení z modelu vynecháno.)

Po výběru regresních proměnných a přizpůsobení regresního modelu je nutné vykreslit rezidua, aby se zkontrolovalo, zda byly splněny předpoklady modelu. Existuje řada grafů, které by měly být vytvořeny, aby bylo možné zkontrolovat různé aspekty osazeného modelu a základní předpoklady. Nyní budeme diskutovat o každém z nich.

ACF graf reziduí

U dat časových řad je vysoce pravděpodobné, že hodnota proměnné pozorovaná v aktuálním časovém období bude podobná její hodnotě v předchozím období, nebo dokonce v období před ním a tak dále. Proto při montáži regresního modelu na data časových řad je běžné najít autokorelaci ve zbytekech. V tomto případě odhadovaný model porušuje předpoklad, že v chybách není žádná autokorelace, a naše prognózy mohou být neefektivní - zbývají některé informace, které by měly být v modelu zohledněny, aby bylo možné získat lepší prognózy. Předpovědi z modelu s automaticky korelovanými chybami jsou stále nezaujaté, a proto nejsou "špatné", ale obvykle budou mít větší intervaly předpovědí, než je nutné. Proto bychom se měli vždy podívat na graf zbytků ACF.

Dalším užitečným testem autokorelace v reziduích určených k zohlednění regresního modelu je **Breusch-Godfreyův** test, označovaný také jako LM (Lagrangeův multiplikátor) test pro sériovou korelaci. Používá se k testování společné hypotézy, že ve zbytek neexistuje autokorelace až do určitého stanoveného pořadí. Malá hodnota p indikuje, že ve zbytek zůstává významná autokorelace.

Breusch-Godfreyův test je podobný testu Ljung-Box, ale je speciálně navržen pro použití s regresními modely.

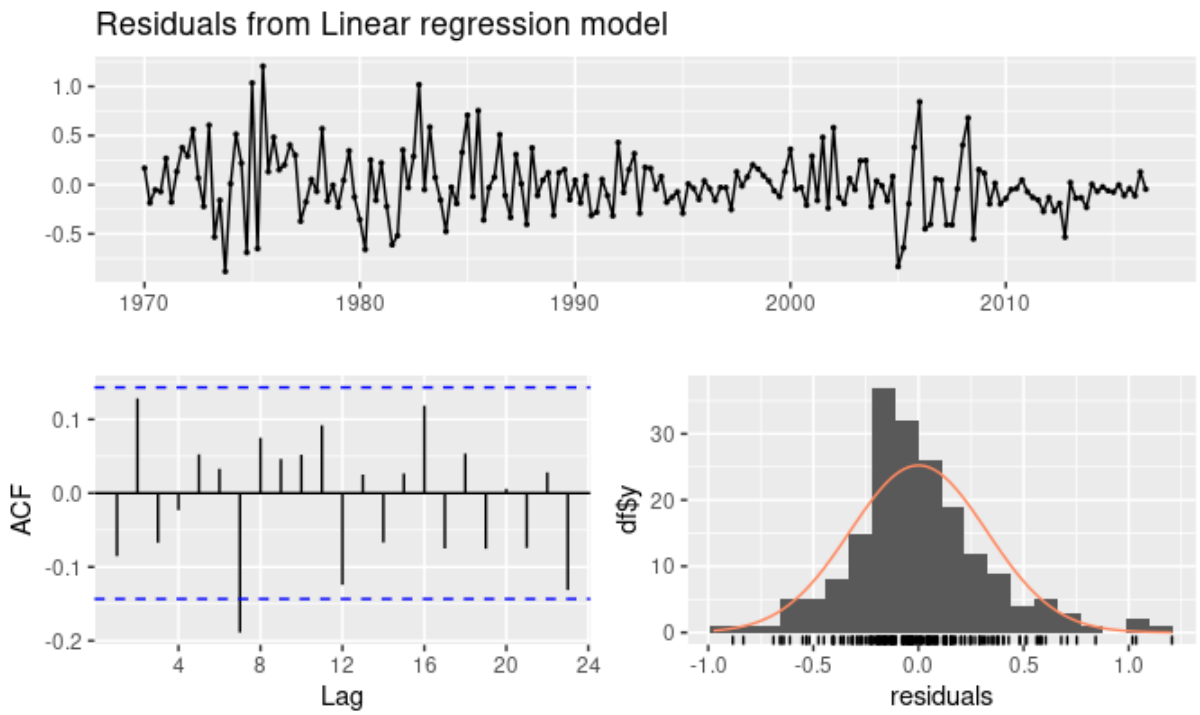
Histogram reziduí

Vždy je dobré zkontrolovat, zda jsou zbytky normálně distribuovány. Jak jsme vysvětlili dříve, není to nezbytné pro prognózování, ale usnadňuje to výpočet predikčních intervalů.

Příklad

Pomocí funkce zavedené v části 3.3můžeme získat všechny užitečné zbytkové diagnostiky uvedené výše.`checkresiduals()`

```
checkresiduals(fit.consMR)
```



Obrázek 5.8: Analýza reziduí z regresního modelu pro čtvrtletní spotřebu v USA.

```
#>
#>  Breusch-Godfrey test for serial correlation of
#>  order up to 8
#>
#> data:  Residuals from Linear regression model
#> LM test = 15, df = 8, p-value = 0.06
```

Obrázek 5.8 ukazuje časový graf, ACF a histogram reziduí z modelu vícenásobné regrese, který se hodí k americkým čtvrtletním údajům o spotřebě, stejně jako Breusch-Godfreyův test pro společné testování autokorelace až do 8. řádu. (Funkce bude používat Breusch-Godfreyův test pro regresní modely, ale Ljung-Boxův test jinak.)`checkresiduals()`
Časový graf ukazuje některé měnící se variace v průběhu času, ale jinak je relativně nepozoruhodný. Tato heteroscedasticita potenciálně způsobí, že pokrytí predikčního intervalu bude nepřesné.

Histogram ukazuje, že rezidua se zdají být mírně zkosená, což může také ovlivnit pravděpodobnost pokrytí intervalů předpovědi.

Autokorelační graf ukazuje významný nárůst při zpoždění 7, ale není to dost na to, aby Breusch-Godfrey byl významný na úrovni 5%. V každém případě není autokorelace nijak zvlášť velká a při zpoždění 7 je nepravděpodobné, že by měla znatelný dopad na prognózy nebo intervaly předpovědí. V kapitole 9 se zabýváme dynamickými regresními modely používanými pro lepší zachycení informací zanechaných ve zbytcích.

Zbytkové grafy proti prediktorům

Očekávali bychom, že zbytky budou náhodně rozptýleny, aniž by vykazovaly nějaké systematické vzorce. Jednoduchý a rychlý způsob, jak to zkontrolovat, je prozkoumat rozptylové grafy reziduí proti každé z proměnných prediktoru. Pokud tyto

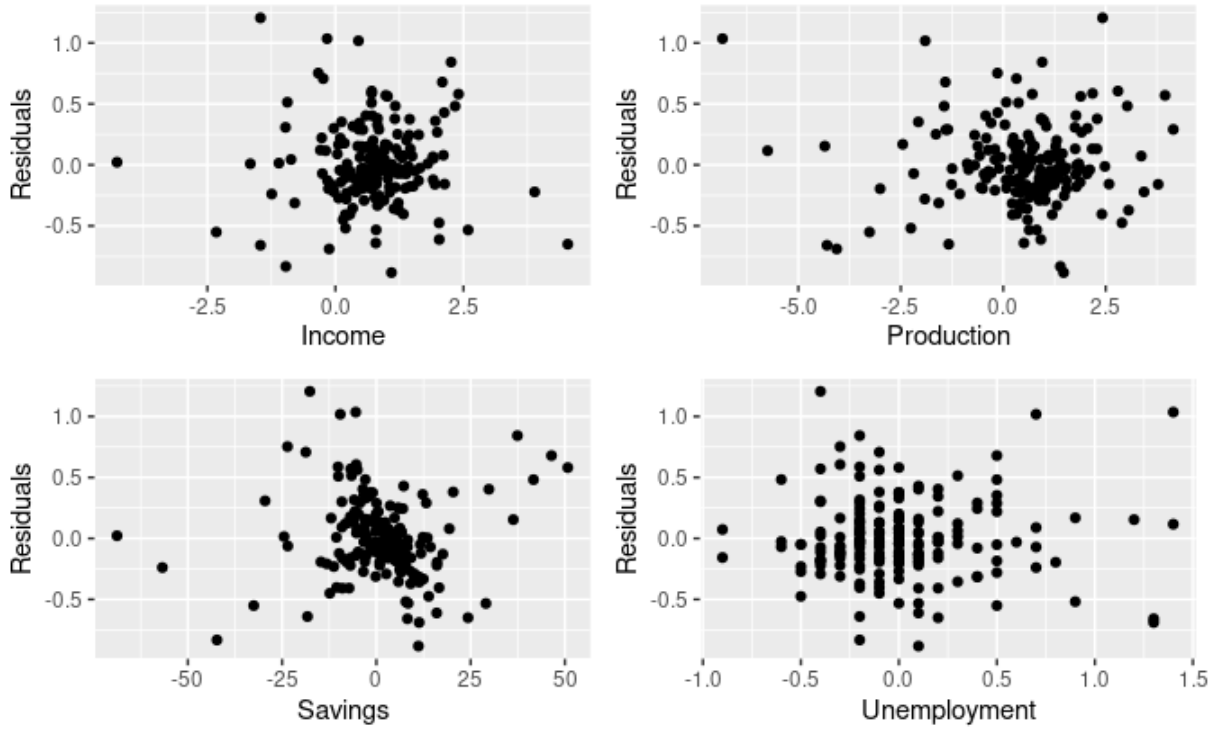
scatterploty vykazují vzor, pak vztah může být nelineární a model bude muset být odpovídajícím způsobem upraven. Viz oddíl 5.8 pro diskusi o nelineární regresi.

Je také nutné vykreslit rezidua proti všem prediktorům, které *nejsou* v modelu. Pokud některý z nich vykazuje vzor, může být nutné přidat odpovídající prediktor do modelu (případně v nelineární formě).

Příklad

Zbytky z modelu vícenásobné regrese pro předpovídání spotřeby v USA vynesené proti každému prediktoru na obrázku 5.9 se zdají být náhodně rozptýlené. Proto jsme s nimi v tomto případě spokojeni.

```
df <- as.data.frame(uschange)
df[,"Residuals"]  <- as.numeric(residuals(fit.consMR))
p1 <- ggplot(df, aes(x=Income, y=Residuals)) +
  geom_point()
p2 <- ggplot(df, aes(x=Production, y=Residuals)) +
  geom_point()
p3 <- ggplot(df, aes(x=Savings, y=Residuals)) +
  geom_point()
p4 <- ggplot(df, aes(x=Unemployment, y=Residuals)) +
  geom_point()
gridExtra::grid.arrange(p1, p2, p3, p4, nrow=2)
```



Obrázek 5.9: Scatterplots reziduí versus každý prediktor.

Zbytkové grafy proti namontovaným hodnotám

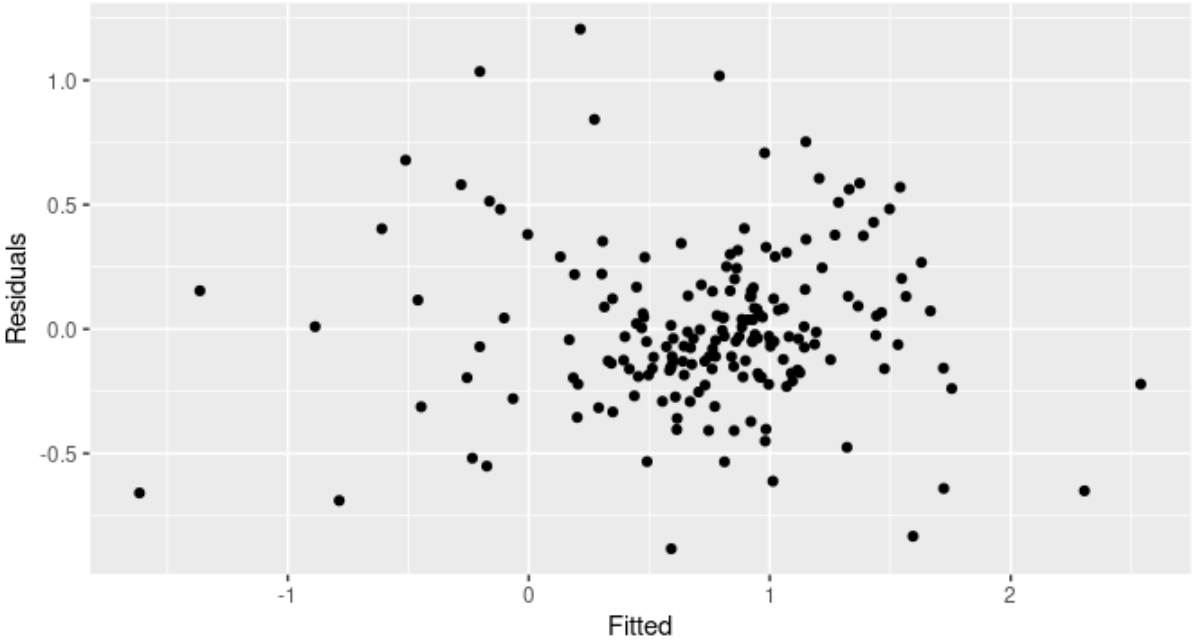
Graf zbytků proti namontovaným hodnotám by také neměl vykazovat žádný vzor. Pokud je pozorován vzorec, může být v chybách "heteroscedasticita", což znamená, že rozptyl reziduí nemusí být konstantní. Pokud k tomuto problému dojde, může být

vyžadována transformace proměnné prognózy, například logaritmus nebo druhá odmocnina (viz Část 3.2.)

Příklad

V návaznosti na předchozí příklad ukazuje obrázek 5.10 zbytky vykreslené proti namontovaným hodnotám. Náhodný rozptyl naznačuje, že chyby jsou homoscedastické.

```
cbind(Fitted = fitted(fit.consMR),
      Residuals=residuals(fit.consMR)) %>%
  as.data.frame() %>%
  ggplot(aes(x=Fitted, y=Residuals)) + geom_point()
```



Obrázek 5.10: Rozptylové grafy reziduí versus namontované hodnoty.

Odlehlé hodnoty a vlivná pozorování

Pozorování, která mají extrémní hodnoty ve srovnání s většinou dat, se nazývají **odlehlé hodnoty**. Pozorování, která mají velký vliv na odhadované koeficienty regresního modelu, se nazývají **vlivná pozorování**. Vlivná pozorování jsou obvykle také odlehlé hodnoty, které jsou extrémní ve směru \(x\).

Existují formální metody pro detekci odlehlých hodnot a vlivných pozorování, které jsou mimo rozsah této učebnice. Jak jsme navrhli na začátku kapitoly 2, seznámení se s vašimi daty před provedením jakékoli analýzy má zásadní význam. Bodový graf \(y\) proti každému \(x\) je vždy užitečným výchozím bodem v regresní analýze a často pomáhá identifikovat neobvyklá pozorování.

Jedním ze zdrojů odlehlých hodnot je nesprávné zadání dat. Jednoduchá popisná statistika vašich dat může identifikovat minima a maxima, která nejsou rozumná. Pokud je takové pozorování identifikováno a bylo zaznamenáno nesprávně, mělo by být okamžitě opraveno nebo odstraněno ze vzorku.

Odlehlé hodnoty se také vyskytují, když jsou některá pozorování jednoduše odlišná. V tomto případě nemusí být moudré, aby tato pozorování byla odstraněna. Pokud bylo pozorování identifikováno jako pravděpodobná odlehlá hodnota, je důležité jej prostudovat a analyzovat jeho možné důvody. Rozhodnutí odstranit nebo zachovat pozorování může být náročné (zejména pokud jsou odlehlé hodnoty vlivnými pozorováními). Je moudré hlásit výsledky jak s odstraněním takových pozorování, tak bez něj.

Příklad

Obrázek 5.11 zdůrazňuje vliv jediné odlehlé hodnoty při regresi americké spotřeby na příjmy (příklad uvedený v oddíle 5.1). V levém panelu je odlehlá hodnota pouze extrémní ve směru \(y\), protože procentuální změna spotřeby byla nesprávně zaznamenána jako -4%. Červená čára je regresní čára přizpůsobená datům, která zahrnuje odlehlé hodnoty, ve srovnání s černou čárou, což je čára přizpůsobená datům bez odlehlé hodnoty. V pravém panelu je nyní odlehlá hodnota také extrémní ve směru \(x\) s 4% poklesem spotřeby odpovídajícím 6% nárůstu příjmů. V tomto případě je odlehlá hodnota extrémně vlivná, protože červená čára se nyní podstatně odchyluje od černé čáry.

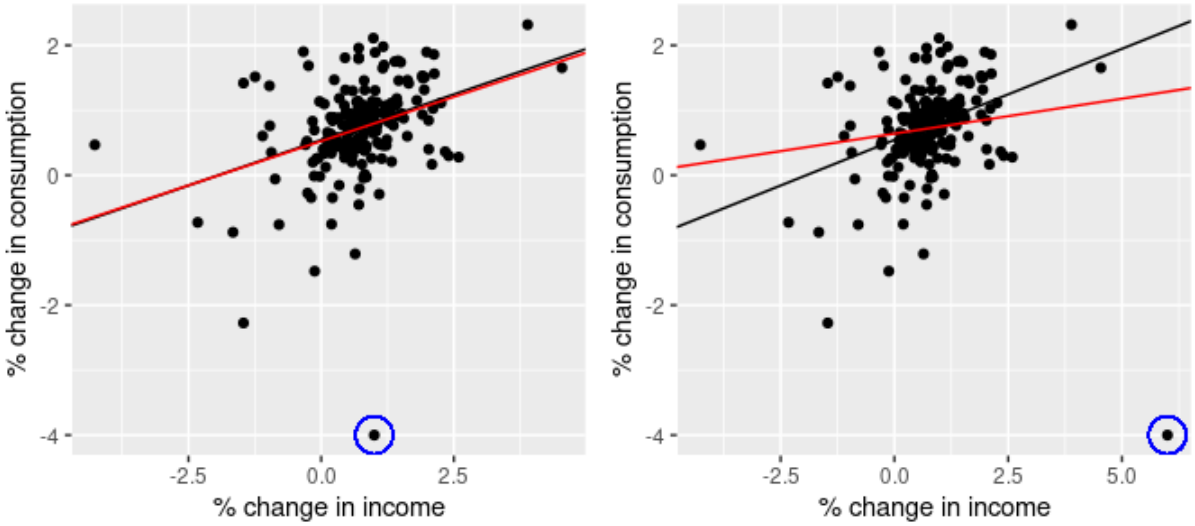
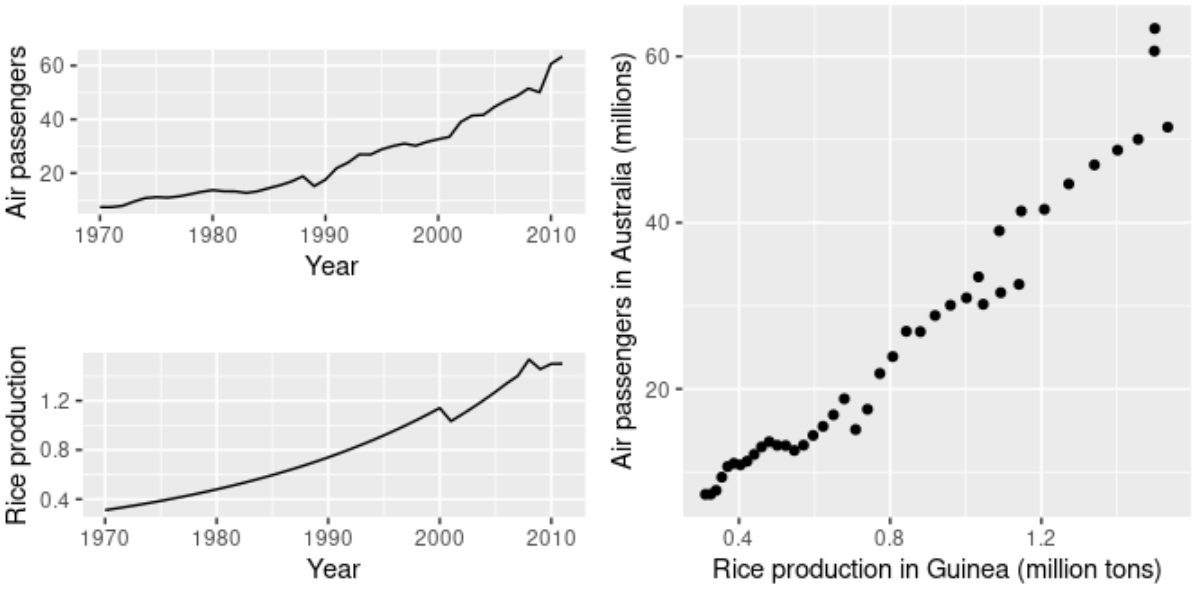


Figure 5.11: The effect of outliers and influential observations on regression

Falešná regrese

Více často než ne, data časových řad jsou "nestacionární"; to znamená, že hodnoty časové řady nekolísají kolem konstantní střední hodnoty nebo s konstantním rozptylem. Stacionaritou časových řad se budeme podrobněji zabývat v kapitole 8, ale zde se musíme zabývat účinkem, který mohou mít nestacionární data na regresní modely.

Zvažte například dvě proměnné vykreslené na obrázku 5.12. Zdá se, že spolu souvisí jednoduše proto, že oba směřují nahoru stejným způsobem. Letecká osobní doprava v Austrálii však nemá nic společného s produkcí rýže v Guineji.

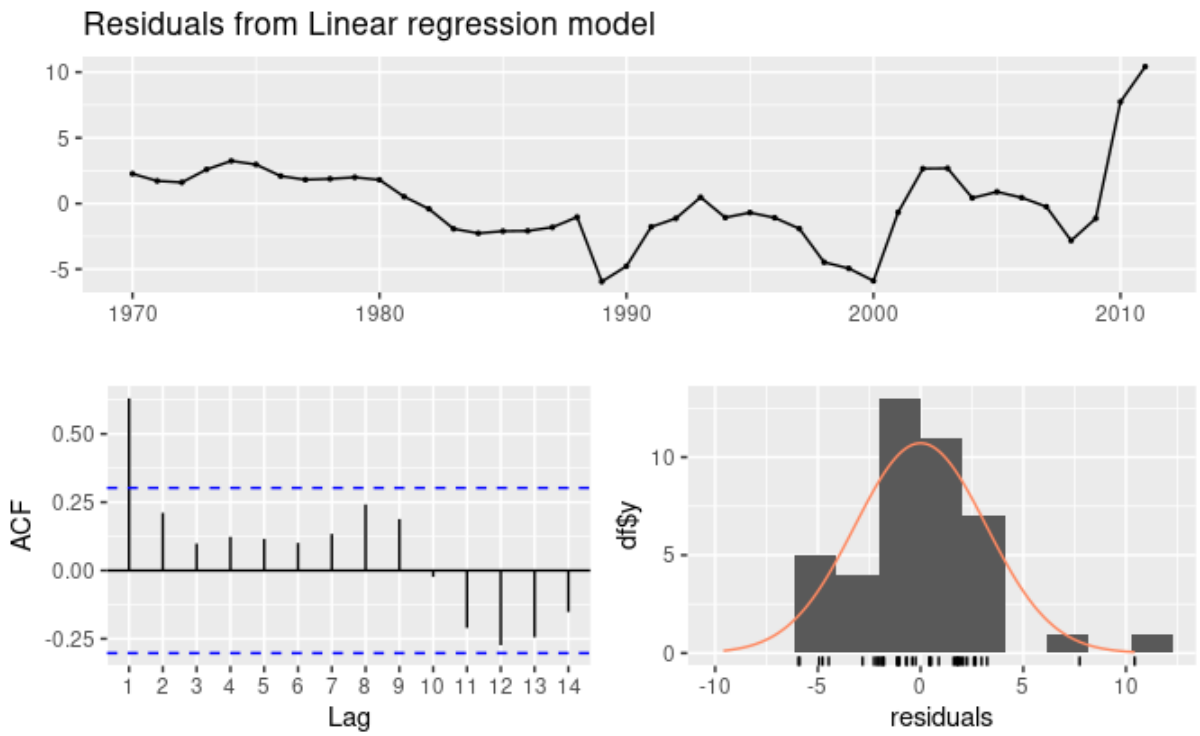


Obrázek 5.12: Může se zdát, že údaje o trendových časových řadách spolu souvisí, jak ukazuje tento příklad, kdy jsou cestující v letecké dopravě v Austrálii regresováni proti produkci rýže v Guineji.

Regresní nestacionární časové řady mohou vést k falešným regresím. Produkce regresivních australských cestujících v letecké dopravě na produkci rýže v Guineji je znázorněna na obrázku 5.13. Vysoká \(R^2\) a vysoká zbytková autokorelace mohou být známkami falešné regrese. Všimněte si těchto funkcí ve výstupu níže. O otázkách týkajících se nestacionárních dat a falešných regresí se podrobněji zabýváme v kapitole 9.

Případy falešné regrese se mohou zdát jako rozumné krátkodobé prognózy, ale obecně nebudou pokračovat v práci do budoucna.

```
aussies <- window(ausair, end=2011)
fit <- tslm(aussies ~ guinearice)
summary(fit)
#>
#> Call:
#> tslm(formula = aussies ~ guinearice)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -5.945 -1.892 -0.327  1.862 10.421
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    -7.49       1.20   -6.23  2.3e-07 ***
#> guinearice     40.29       1.34   30.13  < 2e-16 ***
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.24 on 40 degrees of freedom
#> Multiple R-squared:  0.958,  Adjusted R-squared:  0.957
#> F-statistic:  908 on 1 and 40 DF,  p-value: <2e-16
checkresiduals(fit)
```



Obrázek 5.13: Zbytky z falešné regrese.

```
#>
#>  Breusch-Godfrey test for serial correlation of
#>  order up to 8
#>
#> data:  Residuals from Linear regression model
#> LM test = 29, df = 8, p-value = 3e-04
```

5.4 Některé užitečné prediktory

Existuje několik užitečných prediktorů, které se vyskytují často při použití regrese pro data časových řad.

Trend

Je běžné, že data časových řad jsou trendy. Lineární trend lze modelovat jednoduše pomocí \(x_{1,t}=t\) jako prediktoru, \[y_{t}= \beta_0+\beta_1t+\varepsilon_t, \] kde \(t=1,\dots,T\). Proměnnou trendu lze ve funkci zadat pomocí prediktoru. V části 5.8 diskutujeme o tom, jak můžeme také modelovat nelineární trendy.`tslm()` `trend`

Fiktivní proměnné

Zatím jsme předpokládali, že každý prediktor má číselné hodnoty. Ale co když je prediktor kategorická proměnná, která má pouze dvě hodnoty (např. "ano" a "ne")? Taková proměnná může vzniknout například při prognózování denních prodejů a chcete vzít v úvahu, zda je den **státním svátkem** nebo ne. Takže prediktor má hodnotu "ano" o státním svátku a "ne" jinak.

Tuto situaci lze stále řešit v rámci více regresních modelů vytvořením "fiktivní proměnné", která má hodnotu 1 odpovídající "ano" a 0 odpovídající "ne". Fiktivní proměnná je také známá jako "proměnná indikátoru".

Fiktivní proměnnou lze také použít k **zohlednění odlehlé hodnoty** v datech. Místo vynechání odlehlé hodnoty fiktivní proměnná odstraní její účinek. V tomto případě má fiktivní proměnná hodnotu 1 pro toto pozorování a 0 všude jinde. Příkladem je případ, kdy došlo ke zvláštní události. Například při předpovídání příjezdů turistů do Brazílie budeme muset zohlednit vliv letních olympijských her v Rio de Janeiru v roce 2016.

Pokud existuje více než dvě kategorie, pak může být proměnná kódována pomocí několika fiktivních proměnných (o jednu méně než celkový počet kategorií). automaticky zpracuje tento případ, pokud zadáte proměnnou faktoru jako prediktor. Obvykle není třeba ručně vytvářet odpovídající fiktivní proměnné.`tslm()`

Sezónní fiktivní proměnné

Předpokládejme, že předpovídáme denní data a chceme jako prediktor zohlednit den v týdnu. Poté lze vytvořit následující fiktivní proměnné.

	d_{1,t}	d_{2,t}	d_{3,t}	d_{4,t}	d_{5,t}	d_{6,
Pondělí	1	0	0	0	0	0
Úterý	0	1	0	0	0	0
Středa	0	0	1	0	0	0
Čtvrtek	0	0	0	1	0	0
Pátek	0	0	0	0	1	0
Sobota	0	0	0	0	0	1
Neděle	0	0	0	0	0	0
Pondělí	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Všimněte si, že pro kódování sedmi kategorií je zapotřebí pouze šest fiktivních proměnných. Je to proto, že sedmá kategorie (v tomto případě neděle) je zachycena zachycením a je určena, když jsou všechny fiktivní proměnné nastaveny na nulu.

Mnoho začátečníků se pokusí přidat sedmou proměnnou figurína pro sedmou kategorii. Toto je známé jako "fiktivní proměnná past", protože způsobí selhání regrese. Bude existovat příliš mnoho parametrů, které lze odhadnout, pokud je zahrnut také zachycení. Obecným pravidlem je použít o jednu fiktivní proměnnou méně než kategorie. Pro čtvrtletní údaje tedy použijte tři fiktivní proměnné; pro měsíční údaje použijte 11 fiktivních proměnných; a pro denní data použijte šest fiktivních proměnných a tak dále.

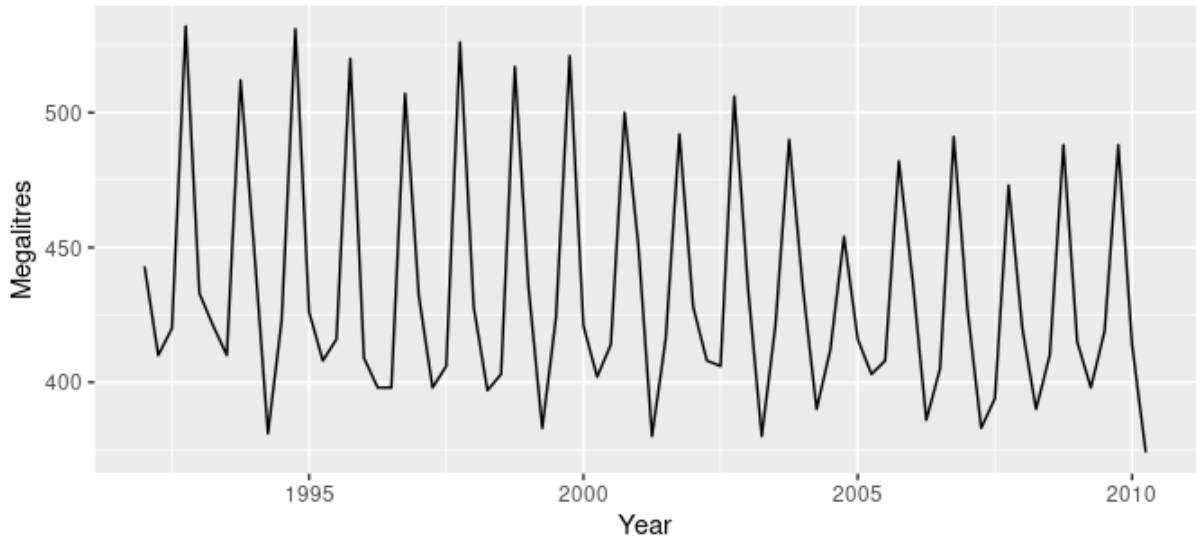
Interpretace každého z koeficientů spojených s fiktivními proměnnými spočívá v tom, že se jedná *o míru účinku této kategorie vzhledem k vynechané kategorii*. Ve výše uvedeném příkladu bude koeficient \(d_{1,t}\) spojený s pondělím měřit vliv pondělí na proměnnou prognózy ve srovnání s vlivem neděle. Následuje příklad interpretace odhadovaných variabilních koeficientů zachycujících čtvrtletní sezónnost australské výroby piva.

Funkce tuto situaci automaticky zvládá, pokud zadáte prediktor `.tslm()season`

Příklad: Australská čtvrtletní výroba piva

Připomeňme si australské čtvrtletní údaje o výrobě piva, které jsou opět uvedeny na obrázku 5.14.

```
beer2 <- window(ausbeer, start=1992)
autoplot(beer2) + xlab("Year") + ylab("Megalitres")
```



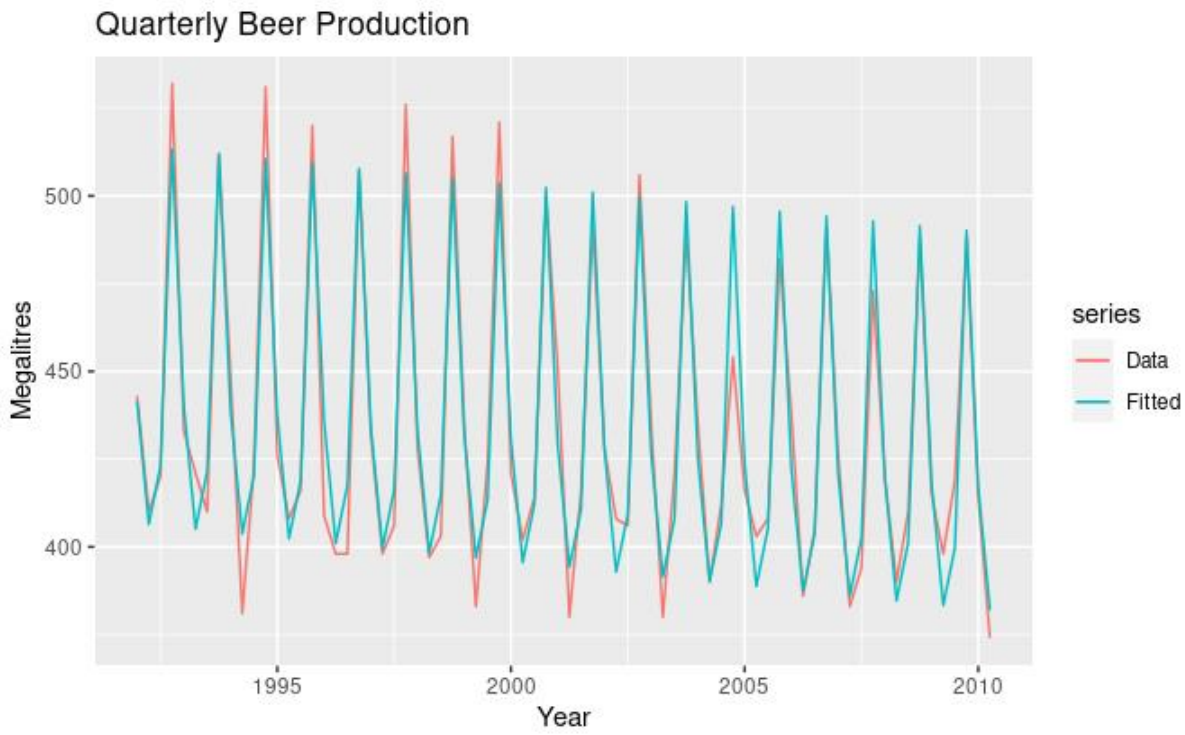
Obrázek 5.14: Australská čtvrtletní výroba piva.

Chceme předpovídat hodnotu budoucí výroby piva. Tato data můžeme modelovat pomocí regresního modelu s lineárním trendem a čtvrtletními fiktivními proměnnými, \[y_{t} = \beta_{0} + \beta_{1} t + \beta_{2}d_{2,t} + \beta_3 d_{3,t} + \beta_4 d_{4,t} + \varepsilon_{t}, \] kde \(d_{i,t} = 1\), pokud \(t\) je ve čtvrtletí \(i\) a 0 jinak. Proměnná za první čtvrtletí byla vynechána, takže koeficienty spojené s ostatními čtvrtletími jsou měřítkem rozdílu mezi těmito čtvrtletími a prvním čtvrtletím.

```
fit.beer <- tslm(beer2 ~ trend + season)
summary(fit.beer)
#>
#> Call:
#> tslm(formula = beer2 ~ trend + season)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -42.90  -7.60  -0.46   7.99  21.79
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 441.8004     3.7335  118.33  < 2e-16 ***
#> trend        -0.3403     0.0666   -5.11  2.7e-06 ***
#> season2     -34.6597     3.9683   -8.73  9.1e-13 ***
#> season3     -17.8216     4.0225   -4.43  3.4e-05 ***
#> season4      72.7964     4.0230   18.09  < 2e-16 ***
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 12.2 on 69 degrees of freedom
#> Multiple R-squared:  0.924,  Adjusted R-squared:  0.92
#> F-statistic:  211 on 4 and 69 DF,  p-value: <2e-16
```

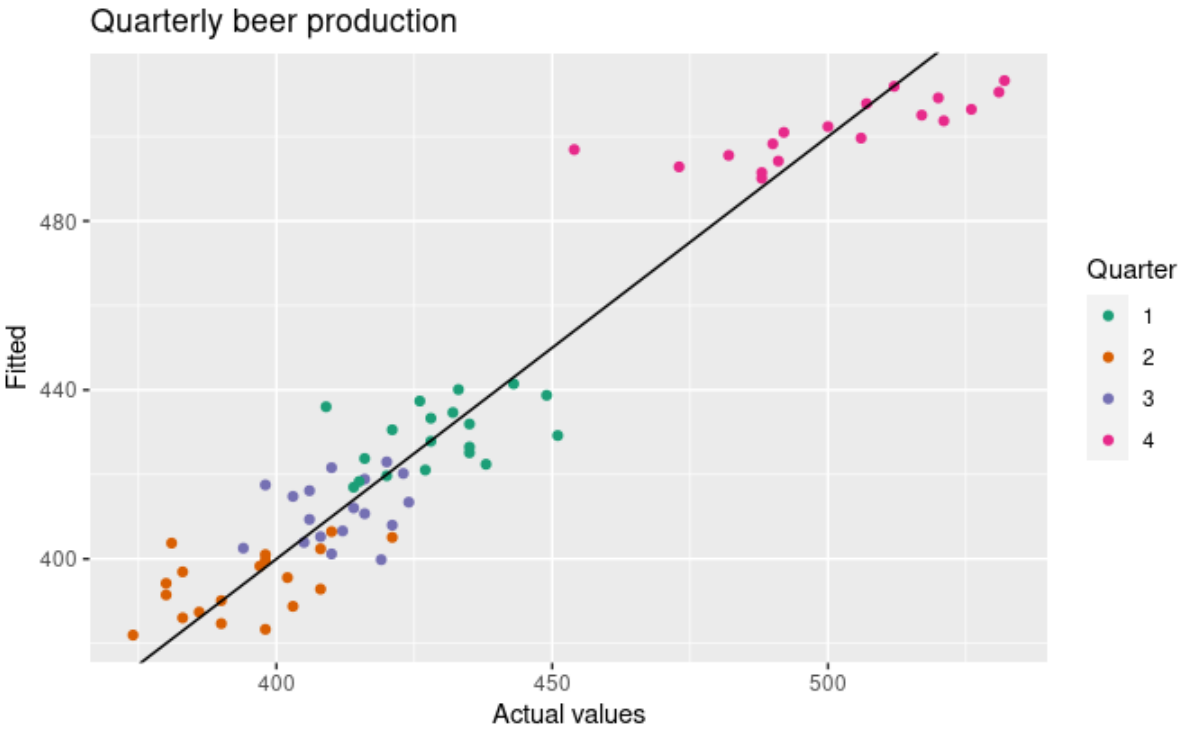
Všimněte si, že a nejsou objekty v pracovním prostoru R; jsou vytvořeny automaticky, pokud jsou zadány tímto způsobem.`trendseasontslm()`
Průměrný klesající trend je -0,34 megalitrů za čtvrtletí. Ve druhém čtvrtletí je produkce o 34,7 megalitrů nižší než v prvním čtvrtletí, ve třetím čtvrtletí je produkce o 17,8 megalitrů nižší než v prvním čtvrtletí a ve čtvrtém čtvrtletí je produkce o 72,8 megalitrů vyšší než v prvním čtvrtletí.

```
autoplot(beer2, series="Data") +
  autolayer(fitted(fit.beer), series="Fitted") +
  xlab("Year") + ylab("Megalitres") +
  ggtitle("Quarterly Beer Production")
```



Obrázek 5.15: Časový graf výroby piva a předpokládané výroby piva.

```
cbind(Data=beer2, Fitted=fitted(fit.beer)) %>%
  as.data.frame() %>%
  ggplot(aes(x = Data, y = Fitted,
             colour = as.factor(cycle(beer2)))) +
    geom_point() +
    ylab("Fitted") + xlab("Actual values") +
    ggtitle("Quarterly beer production") +
    scale_colour_brewer(palette="Dark2", name="Quarter") +
    geom_abline(intercept=0, slope=1)
```



Obrázek 5.16: Skutečná výroba piva vykreslená proti předpokládané produkci piva.

Intervenční proměnné

Často je nutné modelovat intervence, které mohly ovlivnit proměnnou, která má být prognóza. Například činnost konkurence, výdaje na reklamu, protestní akce a tak dále, to vše může mít vliv.

Když efekt trvá pouze jedno období, použijeme proměnnou "spike". Jedná se o fiktivní proměnnou, která má hodnotu jedna v období intervence a nulu jinde. Proměnná špičky je ekvivalentní fiktivní proměnné pro zpracování odlehlé hodnoty.

Další intervence mají okamžitý a trvalý účinek. Pokud zásah způsobí posun úrovně (tj. hodnota řady se od okamžiku zásahu náhle a trvale změní), pak použijeme proměnnou "krok". Kroková proměnná má nulovou hodnotu před zásahem a jednu od okamžiku zásahu dále.

Další formou trvalého účinku je změna sklonu. Zde je intervence řešena pomocí po částech lineárního trendu; trend, který se v době zásahu ohýbá, a proto je nelineární. Budeme o tom diskutovat v oddíle 5.8.

Obchodní dny

Počet obchodních dnů v měsíci se může značně lišit a může mít podstatný vliv na údaje o prodeji. Aby to bylo možné, může být jako prediktor zahrnut počet obchodních dnů v každém měsíci.

U měsíčních nebo čtvrtletních dat funkce vypočítá počet obchodních dnů v každém období.`bizdays()`
Alternativa, která umožňuje účinky různých dnů v týdnu, má následující prediktory: \[\begin{align*} x_{1} &= \text{počet pondělků v měsíci;} \\ x_{2} &= \text{počet úterků v měsíci;} \\ & \vdots \\ x_{7} &= \text{počet nedělí v měsíci.} \end{align*}\]

Distribuovaná zpoždění

Často je užitečné zahrnout výdaje na reklamu jako prediktor. Vzhledem k tomu, že účinek reklamy může trvat déle než samotná kampaň, musíme zahrnout zpožděné hodnoty výdajů na reklamu. Lze tedy použít následující prediktory. \[\begin{align*} x_{1} &= \text{reklama za předchozí měsíc;} \\ x_{2} &= \text{reklama za dva měsíce předtím;} \\ & \vdots \\ x_{m} &= \text{reklama na m měsíce předtím.} \end{align*}\]

Je běžné požadovat, aby se koeficienty snižoval s rostoucí prodlevou, i když to je nad rámec této knihy.

Velikonoce

Velikonoce se liší od většiny svátků, protože se každoročně nekonalo ve stejný den a jejich účinek může trvat několik dní. V tomto případě lze použít fiktivní proměnnou s hodnotou jedna, kde svátek spadá do konkrétního časového období a jinak nula.

S měsíčními údaji, pokud Velikonoce připadají na březen, pak fiktivní proměnná má hodnotu 1 v březnu, a pokud spadá v dubnu, fiktivní proměnná má hodnotu 1 v dubnu. Když Velikonoce začínají v březnu a končí v dubnu, fiktivní proměnná je rozdělena proporcionálně mezi měsíce.

Funkce vypočítá fiktivní proměnnou za vás.`easter()`

Fourierova řada

Alternativou k použití sezónních fiktivních proměnných, zejména pro dlouhá sezónní období, je použití Fourierových termínů. Jean-Baptiste Fourier byl francouzský matematik, narozený v roce 1700, který ukázal, že

řada sinusových a kosinusových termínů správných frekvencí se může přiblížit jakékoli periodické funkci. Můžeme je použít pro sezónní vzory.

Pokud je \(m\) sezónní období, pak prvních několik Fourierových výrazů je dáno \[x_{1,t} = \sin\left(\textstyle\frac{2\pi t}{m}\right), x_{2,t} = \cos\left(\textstyle\frac{2\pi t}{m}\right), x_{3,t} = \sin\left(\textstyle\frac{4\pi t}{m}\right), \] \[x_{4,t} = \cos\left(\textstyle\frac{4\pi t}{m}\right), x_{5,t} = \sin\left(\textstyle\frac{6\pi t}{m}\right), x_{6,t} = \cos\left(\textstyle\frac{6\pi t}{m}\right), \] a tak dále. Pokud máme měsíční sezónnost a použijeme prvních 11 z těchto prediktorových proměnných, pak dostaneme přesně stejné předpovědi jako při použití 11 fiktivních proměnných.

U Fourierových termínů často potřebujeme méně prediktorů než u fiktivních proměnných, zvláště když je \(m\) velký. Díky tomu jsou užitečné pro týdenní data, například kde \(m\cca 52\). Pro krátká sezónní období (např. čtvrtletní údaje) je použití Fourierových termínů jen malou výhodou oproti sezónním fiktivním proměnným.

Tyto Fourierovy termíny jsou vytvářeny pomocí funkce. Například údaje o australském pivu lze modelovat takto.`fourier()`

```
fourier.beer <- tslm(beer2 ~ trend + fourier(beer2, K=2))
summary(fourier.beer)
#>
#> Call:
#> tslm(formula = beer2 ~ trend + fourier(beer2, K = 2))
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -42.90  -7.60  -0.46   7.99  21.79
#>
#> Coefficients:
#>                            Estimate Std. Error t value
#> (Intercept)                446.8792     2.8732  155.53
#> trend                       -0.3403     0.0666   -5.11
#> fourier(beer2, K = 2)S1-4    8.9108     2.0112    4.43
#> fourier(beer2, K = 2)C1-4   53.7281     2.0112   26.71
#> fourier(beer2, K = 2)C2-4   13.9896     1.4226    9.83
#>                            Pr(>|t|)
#> (Intercept)                 < 2e-16 ***
#> trend                       2.7e-06 ***
#> fourier(beer2, K = 2)S1-4   3.4e-05 ***
#> fourier(beer2, K = 2)C1-4   < 2e-16 ***
#> fourier(beer2, K = 2)C2-4   9.3e-15 ***
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 12.2 on 69 degrees of freedom
#> Multiple R-squared:  0.924,  Adjusted R-squared:  0.92
#> F-statistic:  211 on 4 and 69 DF,  p-value: <2e-16
```

První argument, který umožňuje identifikovat sezónní období \(m\) a délku návratu prediktorů. Druhý argument určuje, kolik párů termínů sin a cos zahrnout. Maximální povolený počet je \(K=m/2\), kde \(m\) je sezónní období. Protože jsme zde použili maximum, výsledky jsou totožné s těmi, které jsme získali při použití sezónních fiktivních proměnných.`fourier()K`
Pokud se použijí pouze první dva Fourierovy výrazy (\(x_{1,t}\) a \(x_{2,t}\)), sezónní vzor bude následovat jednoduchou sinusovou vlnu. Regresní model obsahující Fourierovy termíny se často nazývá **harmonická regrese,** protože po sobě jdoucí Fourierovy členy představují harmonické prvních dvou Fourierových termínů.

5.5 Výběr prediktorů

Pokud existuje mnoho možných prediktorů, potřebujeme nějakou strategii pro výběr nejlepších prediktorů pro použití v regresním modelu.

Běžným přístupem, který *se nedoporučuje,* je vykreslit proměnnou prognózy proti určitému prediktoru a pokud neexistuje žádný znatelný vztah, vypusťte tento prediktor z modelu. To je neplatné, protože není vždy možné vidět vztah z rozptylového plástu, zejména pokud nebyly zohledněny účinky jiných prediktorů.

Dalším běžným přístupem, který je také neplatný, je provést vícenásobnou lineární regresi na všech prediktorech a ignorovat všechny proměnné, jejichž hodnoty \(p\)-jsou větší než 0,05. Za prvé, statistická významnost neznamená vždy prediktivní hodnotu. I když prognózování není cílem, není to dobrá strategie, protože hodnoty \(p\)-mohou být zavádějící, když jsou dva nebo více prediktorů vzájemně korelovány (viz oddíl 5.9).

Místo toho použijeme míru prediktivní přesnosti. V tomto oddíle je zavedeno pět takových opatření. Lze je vypočítat pomocí funkce, která je zde použita pro model pro spotřebu v USA:`CV()`

```
CV(fit.consMR)
#>         CV        AIC       AICc        BIC      AdjR2
#>     0.1163 -409.2980 -408.8314 -389.9114     0.7486
```

Tyto hodnoty porovnáváme s odpovídajícími hodnotami z jiných modelů. Pro míry CV, AIC, AICc a BIC chceme najít model s nejnižší hodnotou; pro Adjusted \(R^2\)hledáme model s nejvyšší hodnotou.

Upravené R\(^2\)

Počítačový výstup pro regresi bude vždy dávat hodnotu \(R^2\), která je popsána v Kapitole 5.2. Není to však dobré měřítko prediktivní schopnosti

modelu. Měří, jak dobře model odpovídá historickým datům, ale ne to, jak dobře bude model předpovídat budoucí data.

Kromě toho R^2 neumožňuje "stupně volnosti".
Přidání *libovolné* proměnné má tendenci zvyšovat hodnotu R^2, i když je tato proměnná irelevantní. Z těchto důvodů by prognostici neměli používat R^2 k určení, zda model poskytne dobré předpovědi, protože to povede k nadměrnému vybavení.

Ekvivalentní myšlenkou je vybrat model, který dává minimální součet čtvercových chyb (SSE), daný $$ \text{SSE} = \sum_{t=1}^T e_{t}^2. $$

Minimalizace SSE je ekvivalentní maximalizaci R^2 a vždy zvolí model s nejvíce proměnnými, a proto není platným způsobem výběru prediktorů.

Alternativou, která je navržena k překonání těchto problémů, je upravený R^2 (také nazývaný "R-bar-squared"): $$ \bar{R}^2 = 1-(1-R^2)\frac{T-1}{T-k-1}, $$ kde T je počet pozorování a k je počet prediktorů. Jedná se o zlepšení na $ (R ^ 2 $, protože se již nebude zvyšovat s každým přidaným prediktorem. Při použití této míry bude nejlepším modelem model s největší hodnotou $\bar{R}^2$.
Maximalizace $\bar{R}^2$ je ekvivalentní minimalizaci standardní chyby $\hat{\sigma}_e$ uvedené v rovnici (5.3).

Maximalizace $\bar{R}^2$ funguje docela dobře jako metoda výběru prediktorů, i když má tendenci se mýlí na straně výběru příliš mnoha prediktorů.

Křížové ověření

Křížová validace časových řad byla představena v kapitole 3.4 jako obecný nástroj pro určení prediktivní schopnosti modelu. Pro regresní modely je také možné použít klasickou křížovou validaci leave-one-out na prediktory výběru (Bergmeir, Hyndman, & Koo, 2018). To je rychlejší a efektivněji využívá data. Postup používá následující kroky:

1. Odeberte pozorování t ze sady dat a použijte zbývající data do modelu. Pak vypočítejte chybu ($e_{t}^*=y_{t}-\hat{y}_{t}$) pro vynechané pozorování. (To není totéž jako zbytkové, protože tth pozorování nebylo použito při odhadu hodnoty $\hat{y}_{t}$.)
2. Opakujte krok 1 pro $t=1,\dots,T$.
3. Vypočítejte MSE z $e_{1}^*,\dots,e_{T}^*$. Budeme tomu říkat **životopis**.

Ačkoli to vypadá jako časově náročný postup, existují rychlé metody výpočtu životopisu, takže to netrvá déle než přizpůsobení jednoho modelu úplnému souboru dat. Rovnice pro efektivní výpočet CV je uvedena v

kapitole 5.7. Podle tohoto kritéria je nejlepším modelem ten, který má nejmenší hodnotu životopisu.

Akaikeho informační kritérium

Úzce související metodou je Akaikeho informační kritérium, které definujeme jako $$ \text{AIC} = T\log\left(\frac{\text{SSE}}{T}\right) + 2(k+2), $$ kde T je počet pozorování použitých pro odhad a k je počet prediktorů v modelu. Různé počítačové balíčky používají mírně odlišné definice pro AIC, i když by všechny měly vést k výběru stejného modelu. K části rovnice $k+2$ dochází, protože v modelu jsou parametry $k+2$: koeficienty k pro prediktory, průsečík a rozptyl reziduí. Myšlenkou je penalizovat vhodnost modelu (SSE) počtem parametrů, které je třeba odhadnout.

Model s minimální hodnotou AIC je často nejlepším modelem pro prognózování. U velkých hodnot Tje minimalizace AIC ekvivalentní minimalizaci hodnoty CV.

Opraveno informační kritérium Akaike

Pro malé hodnoty Tmá AIC tendenci vybírat příliš mnoho prediktorů, a proto byla vyvinuta zkreslená verze AIC, $$ \text{AIC}_{\text{c}} = \text{AIC} + \frac{2(k+2)(k+3)}{T-k-3}. $$ Stejně jako u AIC by měl být AICc minimalizován.

Schwarzovo bayesovské informační kritérium

Souvisejícím měřítkem je Schwarzovo Bayesovské informační kritérium (obvykle zkráceně BIC, SBIC nebo SC): $$ \text{BIC} = T\log\left(\frac{\text{SSE}}{T}\right) + (k+2)\log(T). $$ Stejně jako u AIC je minimalizace BIC určena k poskytnutí nejlepšího modelu. Model zvolený BIC je buď stejný jako model zvolený AIC, nebo model s menším počtem termínů. Je to proto, že BIC penalizuje počet parametrů silněji než AIC. Pro velké hodnoty Tje minimalizace BIC podobná křížovému ověřování leave-v-out, když $v = T[1-1/(\log(T)-1)]$.

Jaké opatření bychom měli použít?

Zatímco $\bar{R}^2$ je široce používán a existuje déle než ostatní míry, jeho tendence vybrat příliš mnoho proměnných prediktoru jej činí méně vhodným pro prognózování.

Mnoho statistiků rádo používá BIC, protože má tu vlastnost, že pokud existuje skutečný základní model, BIC vybere tento model s dostatkem dat. Ve skutečnosti však existuje jen zřídka, pokud vůbec, skutečný základní

model, a i kdyby existoval skutečný základní model, výběr tohoto modelu nemusí nutně poskytnout nejlepší prognózy (protože odhady parametrů nemusí být přesné).

Proto doporučujeme použít jednu ze statistik AICc, AIC nebo CV, z nichž každá má jako cíl prognózu. Pokud je hodnota \(T\) dostatečně velká, všechny povedou ke stejnému modelu. Ve většině příkladů v této knize používáme hodnotu AICc k výběru modelu prognózy.

Příklad: Spotřeba v USA

V příkladu vícenásobné regrese pro předpovídání spotřeby v USA jsme uvažovali o čtyřech prediktorech. Se čtyřmi prediktory existují \(2^4=16\) možné modely. Nyní můžeme zkontrolovat, zda jsou všechny čtyři prediktory skutečně užitečné, nebo zda můžeme jeden nebo více z nich vypustit. Bylo namontováno všech 16 modelů a výsledky jsou shrnuty v tabulce 5.1. "1" označuje, že prediktor byl zahrnut do modelu, a "0" znamená, že prediktor nebyl zahrnut do modelu. První řádek tedy ukazuje míry prediktivní přesnosti pro model zahrnující všechny čtyři prediktory.

Výsledky byly seřazeny podle AICc. Proto jsou nejlepší modely uvedeny v horní části tabulky a nejhorší ve spodní části tabulky.

Tabulka 5.1: Všech 16 možných modelů pro předpovídání spotřeby v USA se 4 prediktory.

Příjem	Výroba	Úspory	Nezaměstnanost	CV	AIC	Aicc	MODERÁTOR
1	1	1	1	0.116	-409.3	-408.8	-389.9
1	0	1	1	0.116	-408.1	-407.8	-391.9
1	1	1	0	0.118	-407.5	-407.1	-391.3
1	0	1	0	0.129	-388.7	-388.5	-375.8
1	1	0	1	0.278	-243.2	-242.8	-227.0
1	0	0	1	0.283	-237.9	-237.7	-225.0
1	1	0	0	0.289	-236.1	-235.9	-223.2
0	1	1	1	0.293	-234.4	-234.0	-218.2
0	1	1	0	0.300	-228.9	-228.7	-216.0
0	1	0	1	0.303	-226.3	-226.1	-213.4
0	0	1	1	0.306	-224.6	-224.4	-211.7
0	1	0	0	0.314	-219.6	-219.5	-209.9

Tabulka 5.1: Všech 16 možných modelů pro předpovídání spotřeby v USA se 4 prediktory.

Příjem	Výroba	Úspory	Nezaměstnanost	CV	AIC	Aicc	MODERÁTOR
0	0	0	1	0.314	-217.7	-217.5	-208.0
1	0	0	0	0.372	-185.4	-185.3	-175.7
0	0	1	0	0.414	-164.1	-164.0	-154.4
0	0	0	0	0.432	-155.1	-155.0	-148.6

Nejlepší model obsahuje všechny čtyři prediktory. Bližší pohled na výsledky však odhaluje některé zajímavé rysy. Mezi modely v prvních čtyřech řadách a těmi níže je jasné oddělení. To naznačuje, že příjmy a úspory jsou důležitější proměnné než výroba a nezaměstnanost. Také první dva řádky mají téměř identické hodnoty CV, AIC a AICc. Takže bychom mohli vypustit proměnnou Produkce a získat podobné prognózy. Všimněte si, že výroba a nezaměstnanost jsou vysoce (negativně) korelovány, jak ukazuje obrázek 5.5, takže většina prediktivních informací ve výrobě je také obsažena ve proměnné Nezaměstnanost.

Nejlepší regrese podmnožiny

Tam, kde je to možné, by měly být namontovány všechny potenciální regresní modely (jak bylo provedeno ve výše uvedeném příkladu) a nejlepší model by měl být vybrán na základě jednoho z diskutovaných opatření. Toto je známé jako regrese "nejlepších podmnožin" nebo regrese "všech možných podmnožin".

Postupná regrese

Pokud existuje velký počet prediktorů, není možné přizpůsobit všechny možné modely. Například 40 prediktorů vede k 1 bilionu možných modelů 2^{40} >\). V důsledku toho je nutná strategie, která omezí počet modelů, které mají být prozkoumány.

Přístup, který funguje docela dobře, je *zpětná postupná regrese:*

- Začněte s modelem obsahujícím všechny potenciální prediktory.
- Odeberte jeden prediktor najednou. Udržujte model, pokud zlepšuje míru prediktivní přesnosti.
- Iterujte, dokud nedojde k dalšímu zlepšení.

Pokud je počet potenciálních prediktorů příliš velký, pak zpětná postupná regrese nebude fungovat a místo toho lze použít *postupnou regresi vpřed.* Tento postup začíná modelem, který obsahuje pouze zachycení. Prediktory se přidávají jeden po druhém a ten, který nejvíce zlepšuje míru

prediktivní přesnosti, je v modelu zachován. Postup se opakuje, dokud nebude dosaženo dalšího zlepšení.

Alternativně pro směr dozadu nebo dopředu může být výchozím modelem model, který obsahuje podmnožinu potenciálních prediktorů. V tomto případě je třeba zahrnout další krok. Pro zpětnou proceduru bychom měli také zvážit přidání prediktoru s každým krokem a pro postup vpřed bychom měli také zvážit vypuštění prediktoru s každým krokem. Tyto postupy jsou označovány jako *hybridní* postupy.

Je důležité si uvědomit, že žádný postupný přístup není zaručen, že povede k nejlepšímu možnému modelu, ale téměř vždy vede k dobrému modelu. Další podrobnosti viz James, Witten, Hastie, & Tibshirani (2014).

Beware of inference after selecting predictors

V této knize se nezabýváme statistickým odvozeným odvozem prediktorů (např. při pohledu na p-hodnoty spojené s každým prediktorem). Pokud se chcete podívat na statistickou významnost prediktorů, dejte si pozor na to, že *jakýkoli* postup zahrnující výběr prediktorů nejprve zneplatní předpoklady za hodnotami p-. Postupy, které doporučujeme pro výběr prediktorů, jsou užitečné, pokud se model používá pro prognózování; nejsou užitečné, pokud chcete studovat vliv jakéhokoli prediktoru na proměnnou prognózy.

Bibliografie

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). Poznámka k platnosti křížové validace pro vyhodnocení autoregresivní predikce časových řad. *Výpočetní statistika a analýza dat*, *120*, 70–83. [DOI]

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *Úvod do statistického učení: S aplikacemi v R*. New York: Springer. [Amazon]

5.6 Prognózování s regresí

Připomeňme, že předpovědi y lze získat pomocí $$ \hat{y_t} = \hat\beta_{0} + \hat\beta_{1} x_{1,t} + \hat\beta_{2} x_{2,t} + \cdots + \hat\beta_{k} x_{k,t}, $$, který obsahuje odhadované koeficienty a ignoruje chybu v regresní rovnici. Zapojení hodnot proměnných prediktoru $x_{1,t},\dots,x_{k,t}$ pro $t=1,\dots,T$ vrátilo namontované (training-sample) hodnoty y. Co nás zde však zajímá, je předpovídání *budoucích* hodnot y.

Prognózy ex ante versus ex post

Při použití regresních modelů pro data časových řad musíme rozlišovat mezi různými typy prognóz, které lze vytvořit, v závislosti na tom, co se předpokládá, že je známo při výpočtu prognóz.

Předběžné prognózy jsou ty, které jsou vytvořeny pouze s využitím informací, které jsou k dispozici předem. Například předběžné prognózy procentní změny spotřeby v USA za čtvrtletí následující po ukončení vzorku by měly používat pouze informace, které byly k dispozici až do 3. čtvrtletí 2016 *včetně.* Jedná se o skutečné předpovědi, vytvořené předem s využitím jakýchkoli informací, které jsou v té době k dispozici. Proto, aby bylo možné generovat předběžné prognózy, model vyžaduje prognózy prediktorů. K jejich získání můžeme použít jednu z jednoduchých metod zavedených v kapitole 3.1 nebo sofistikovanější přístupy k čistým časovým řadám, které následují v kapitolách 7 a 8. Alternativně mohou být k dispozici a mohou být použity prognózy z jiného zdroje, jako je vládní agentura.

Ex post prognózy jsou ty, které jsou vytvořeny pomocí pozdějších informací o prediktorech. Například ex-post prognózy spotřeby mohou používat skutečná pozorování prediktorů, jakmile jsou pozorována. Nejedná se o skutečné prognózy, ale jsou užitečné pro studium chování prognostických modelů.

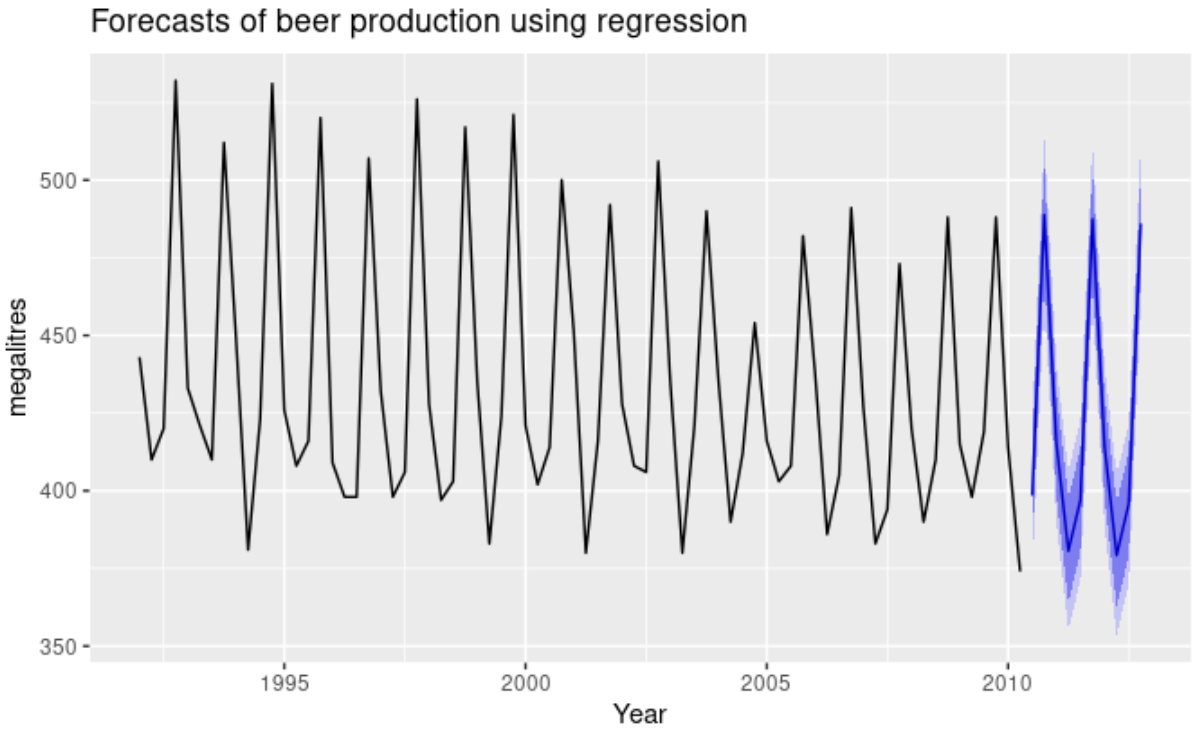
Model, z jehož plošných prognóz se vypěřují předběžné prognózy, by neměl být odhadován na bázi údajů z období prognózy. To znamená, že prognózy ex post mohou předpokládat znalost proměnných prediktoru (proměnné \(x\), ale neměly by předpokládat znalost dat, která mají být předpovězena (proměnná \(y\).

Srovnávací hodnocení prognóz ex ante a ex post prognóz může pomoci oddělit zdroje nejistoty prognóz. To ukáže, zda chyby prognóz vznikly v důsledku špatných předpovědí prediktoru nebo v důsledku špatného předpovědního modelu.

Příklad: Australská čtvrtletní výroba piva

Za normálních okolností nemůžeme při vytváření předběžných předpovědí použít skutečné budoucí hodnoty prediktorových proměnných, protože jejich hodnoty nebudou předem známy. Všechny speciální prediktory zavedené v oddíle 5.4 jsou však známy předem, protože jsou založeny na proměnných kalendáře (např. sezónní fiktivní proměnné nebo ukazatele státních svátků) nebo deterministických funkcích času (např. časový trend). V takových případech není rozdíl mezi prognózami ex ante a ex post.

```
beer2 <- window(ausbeer, start=1992)
fit.beer <- tslm(beer2 ~ trend + season)
fcast <- forecast(fit.beer)
autoplot(fcast) +
  ggtitle("Forecasts of beer production using regression") +
  xlab("Year") + ylab("megalitres")
```



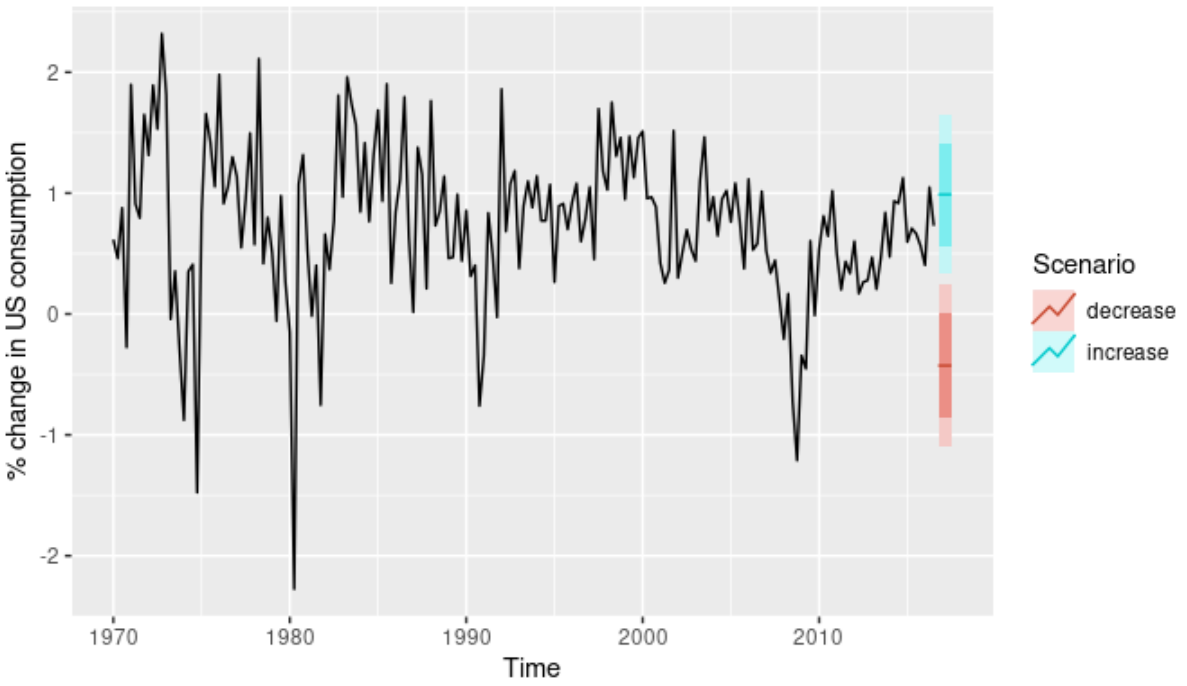
Obrázek 5.17: Prognózy z regresního modelu výroby piva. Tmavě stínovaná oblast vykazuje 80 % intervalů předpovědi a světle stínovaná oblast 95 % intervalů předpovědí.

Prognózy založené na scénářích

V tomto nastavení prognostik předpokládá možné scénáře pro proměnné prediktoru, které jsou zajímavé. Například tvůrce politik v USA by mohl mít zájem porovnat předpokládanou změnu spotřeby, pokud dochází k neustálému růstu příjmů o 1 % a 0,5 % beze změny míry zaměstnanosti oproti příslušnému poklesu o 1 % a 0,5 % pro každé ze čtyř čtvrtletí následujících po skončení vzorku. Výsledné prognózy jsou vypočteny níže a jsou uvedeny na obrázku 5.18. Měli bychom poznamenat, že predikční intervaly pro prognózy založené na scénářích nezahrnují nejistotu spojenou s budoucími hodnotami proměnných prediktoru. Předpokládají, že hodnoty prediktorů jsou známy předem.

```
fit.consBest <- tslm(
  Consumption ~ Income + Savings + Unemployment,
  data = uschange)
h <- 4
newdata <- data.frame(
    Income = c(1, 1, 1, 1),
    Savings = c(0.5, 0.5, 0.5, 0.5),
    Unemployment = c(0, 0, 0, 0))
fcast.up <- forecast(fit.consBest, newdata = newdata)
newdata <- data.frame(
    Income = rep(-1, h),
    Savings = rep(-0.5, h),
    Unemployment = rep(0, h))
fcast.down <- forecast(fit.consBest, newdata = newdata)
autoplot(uschange[, 1]) +
  ylab("% change in US consumption") +
  autolayer(fcast.up, PI = TRUE, series = "increase") +
```

```
autolayer(fcast.down, PI = TRUE, series = "decrease") +
guides(colour = guide_legend(title = "Scenario"))
```



Obrázek 5.18: Prognóza procentuálních změn výdajů na osobní spotřebu pro USA v rámci prognóz založených na scénářích.

Vytvoření prediktivního regresního modelu

Velkou výhodou regresních modelů je, že je lze použít k zachycení důležitých vztahů mezi předpokládanou proměnnou zájmu a proměnnými prediktoru. Hlavní výzvou však je, že pro generování předběžných předpovědí vyžaduje model budoucí hodnoty každého prediktoru. Pokud je prognóza založená na scénářích zajímavá, pak jsou tyto modely velmi užitečné. Pokud je však hlavním cílem prognózování ex ante, může být získání prognóz prediktorů náročné (v mnoha případech může být generování prognóz pro proměnné prediktoru náročnější než přímé předpovídání proměnné prognózy bez použití prediktorů).

Alternativní formulace je použít jako prediktory jejich zpožděné hodnoty. Za předpokladu, že máme zájem o generování \(h\)-krok dopředu prognózy, napíšeme \[
y_{t+h}=\beta_0+\beta_1x_{1,t}+\dots+\beta_kx_{k,t}+\varepsilon_{t+h}
\] pro \(h=1,2\dots\). Sada prediktorů je tvořena hodnotami \(x\)s, které jsou pozorovány \(h\) časovými obdobími před pozorováním \(y\). Proto, když je odhadovaný model projektován do budoucnosti, tj. za koncem vzorku \(T\), jsou k dispozici všechny hodnoty prediktoru.

Zahrnutí zpožděných hodnot prediktorů nejenže činí model funkčním pro snadné generování prognóz, ale také jej činí intuitivně přitažlivým. Například účinek změny politiky s cílem zvýšit výrobu nemusí mít okamžitý účinek na výdaje na spotřebu. Je velmi pravděpodobné, že se to stane se zpožděním. Dotkli jsme se toho v části 5.4, když jsme stručně představili distribuované zpoždění jako prediktory.

Několik směrů pro zobecnění regresních modelů pro lepší začlenění bohaté dynamiky pozorované v časových řadách je popsáno v části 9.

Intervaly předpovědi

S každou prognózou změny spotřeby na obrázku 5.18jsou také zahrnuty intervaly předpovědi 95% a 80%. Obecná formulace výpočtu predikčních intervalů pro více regresních modelů je uvedena v kapitole 5.7. Vzhledem k tomu, že se jedná o pokročilou maticovou algebru, uvádíme zde případ pro výpočet predikčních intervalů pro jednoduchou regresi, kde lze prognózu vygenerovat pomocí rovnice, $$ \hat{y}=\hat{\beta}_0+\hat{\beta}_1x. $$ Za předpokladu, že regresní chyby jsou normálně rozloženy, přibližný 95% interval předpovědi spojený s touto předpovědí je dán $$\begin{equation} \hat{y} \pm 1.96 \hat{\sigma}_e\sqrt{1+\frac{1}{T}+\frac{(x-\bar{x})^2}{(T-1)s_x^2}}, \tag{5.4} \end{equation}$$ kde T je celkový počet pozorování, $\bar{x}$ je průměr pozorovaných hodnot x, s_x je směrodatná odchylka pozorovaných hodnot x a $\hat{\sigma}_e$ je standardní chyba regrese dané rovnicí (5.3) . Podobně lze získat 80% predikční interval nahrazením 1,96 za 1,28. Jiné intervaly předpovědi lze získat nahrazením hodnoty 1,96 příslušnou hodnotou uvedenou v tabulce 3.1. Pokud se k získání predikčních intervalů použije R, získají se přesnější výpočty (zejména pro malé hodnoty T), než je dáno rovnicí (5.4).

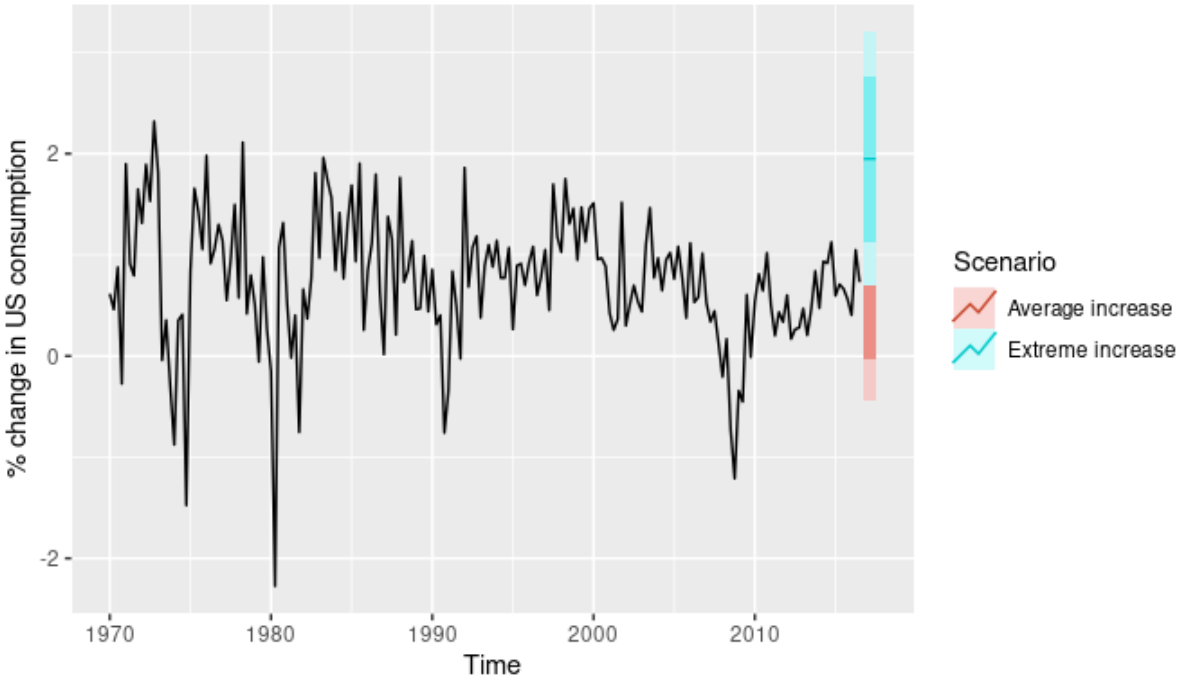
Rovnice (5.4) ukazuje, že interval předpovědi je širší, když x je daleko od $\bar{x}$. To znamená, že jsme si jistější našimi prognózami, když zvažujeme hodnoty proměnné prediktoru blízké jejímu průměru vzorku.

Příklad

Odhadovaná jednoduchá regresní čára v příkladu spotřeby v USA je $$ \hat{y}_t=0,55 + 0.28x_t. $$

Za předpokladu, že v příštích čtyřech čtvrtletích se osobní důchod zvýší o svou historickou střední hodnotu ($\bar{x}=0,72\%$), spotřeba se podle prognóz zvýší o $0,75\%$ a odpovídající intervaly předpovědi 95% a 80% jsou $[-0,45,1,94]$ a $[-0,03,1,52]$ (vypočteno pomocí R). Pokud předpokládáme extrémní nárůst příjmů o 5%, pak jsou predikční intervaly podstatně širší, jak ukazuje obrázek 5.19.

```
fit.cons <- tslm(Consumption ~ Income, data = uschange)
h <- 4
fcast.ave <- forecast(fit.cons,
  newdata = data.frame(
    Income = rep(mean(uschange[,"Income"]), h)))
fcast.up <- forecast(fit.cons,
  newdata = data.frame(Income = rep(5, h)))
autoplot(uschange[, "Consumption"]) +
  ylab("% change in US consumption") +
  autolayer(fcast.ave, series = "Average increase",
    PI = TRUE) +
  autolayer(fcast.up, series = "Extreme increase",
    PI = TRUE) +
  guides(colour = guide_legend(title = "Scenario"))
```



Obrázek 5.19: Predikční intervaly, pokud se příjem zvýší o svůj historický průměr \(0,72\%\) oproti extrémnímu nárůstu o 5%.

5.7 Formulace matrice

Upozornění: jedná se o pokročilejší, volitelnou část a předpokládá znalost maticové algebry.

Připomeňme, že vícenásobný regresní model lze zapsat jako $$ y_{t} = \beta_{0} + \beta_{1} x_{1,t} + \beta_{2} x_{2,t} + \cdots + \beta_{k} x_{k,t} + \varepsilon_{t} $$ kde ε_{t} má střední nulu a rozptyl σ^2. To vyjadřuje vztah mezi jednou hodnotou proměnné prognózy a prediktory.

Může být vhodné to napsat ve formě matice, kde jsou všechny hodnoty proměnné prognózy uvedeny v jedné rovnici. Let $\bm{y} = (y_{1},\dots,y_{T})'$, $\bm{\varepsilon} = (\varepsilon_{1},\dots,\varepsilon_{T})'$, $\bm{\beta} = (\beta_{0},\dots,\beta_{k})'$ a $$ \bm{X} = \left[\begin{matrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1}\\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2}\\ \vdots& \vdots& \vdots&& \vdots\\ 1 & x_{1, T}& x_{2,T}& \dots& x_{k,T} \end{matrix}\right]. $$ Pak $$ \bm{y} = \bm{X}\bm{\beta} + \bm{\varepsilon}. $$ kde $\bm{\varepsilon}$ má střední hodnotu $\bm{0}$ a rozptyl $\sigma^2\bm{I}$. Všimněte si, že matice $\bm{X}$ má řádky T odrážející počet pozorování a $k+1$ sloupce odrážející průsečík, který je reprezentován sloupcem jedniček plus počtem prediktorů.

Odhad nejmenších čtverců

Odhad nejmenších čtverců se provádí minimalizací výrazu $\bm{\varepsilon}'\bm{\varepsilon} = (\bm{y} - \bm{X}\bm{\beta})'(\bm{y} - \bm{X}\bm{\beta})$. Lze ukázat, že je to minimalizováno, když $\bm{\beta}$ má hodnotu $$ \hat{\bm{\beta}} = (\bm{X}'\bm{X})^{-1}\bm{X}'\bm{y} $$ Toto je někdy známé jako "normální rovnice". Odhadované koeficienty vyžadují inverzi matice $\bm{X}'\bm{X}$. Pokud $\bm{X}$ nemá plné pořadí sloupců, pak matice $\bm{X}'\bm{X}$ je singulární a model nelze odhadnout. K tomu dojde například tehdy, pokud se dostanete do "pasti fiktivní proměnné", tj. máte stejný počet fiktivních proměnných, jako existují kategorie kategorického prediktoru, jak je popsáno v části 5.4.

Zbytková odchylka se odhaduje pomocí $$ \hat{\sigma}_e^2 = \frac{1}{T-k-1}(\bm{y} - \bm{X}\hat{\bm{\beta}})' (\bm{y} - \bm{X}\hat{\bm{\beta}}). $$

Namontované hodnoty a křížová validace

Normální rovnice ukazuje, že namontované hodnoty lze vypočítat pomocí $$ \bm{\hat{y}} = \bm{X}\hat{\bm{\beta}} = \bm{X}(\bm{X}'\bm{X})^{-1}\bm{X}'\bm{y} = \bm{H}\bm{y}, $$ kde $\bm{H} = \bm{X}(\bm{X}'\bm{X})^{-1}\bm{X}'$ je známá jako "hat-matrix", protože se používá k výpočtu $\bm{\hat{y}}$ ("y-hat").

Pokud jsou diagonální hodnoty $\bm{H}$ označeny $h_{1},\dots,h_{T}$, pak lze statistiku křížového ověření vypočítat pomocí $$ \text{CV} = \frac{1}{T}\sum_{t=1}^T [e_{t}/(1-h_{t})]^2, $$ kde e_{t} je zbytek získaný přizpůsobením modelu všem pozorováním T. Při výpočtu statistiky CV tedy není nutné skutečně zapadat do samostatných modelů T.

Předpovědi a intervaly předpovědí

Nechť $\bm{x}^*$ je řádkový vektor obsahující hodnoty prediktorů (ve stejném formátu jako $\bm{X}$), pro které chceme generovat prognózu . Pak je předpověď dána $$ \hat{y} = \bm{x}^*\hat{\bm{\beta}}=\bm{x}^*(\bm{X}'\bm{X})^{-1}\bm{X}'\bm{y} $$ a její odhadovaný rozptyl je dán $$ \hat\sigma_e^2 \left[1 + \bm{x}^* (\bm{X}'\bm{X})^{-1} (\bm{x}^*)'\right]. $$ 95% interval předpovědi lze vypočítat (za předpokladu normálně distribuovaných chyb) jako $$ \hat{y} \pm 1.96 \hat{\sigma}_e \sqrt{1 + \bm{x}^* (\bm{X}'\bm{X})^{-1} (\ bm{x}^*)'}. $$ To bere v úvahu nejistotu způsobenou chybovým termínem ε a nejistotou v odhadech koeficientů. Ignoruje však všechny chyby v $\bm{x}^*$. Pokud jsou tedy budoucí hodnoty prediktorů nejisté, pak bude predikční interval vypočtený pomocí tohoto výrazu příliš úzký.