Lineární regrese — pokrač.

7.12.2012

Úvodní nastavení.

– Otevřete si R Studio. Z internetu si stáhněte data blood.csv a Srazky.csv. Dále budeme potřebovat data Brain.txt, se kterými jsme již dříve pracovali.

Vícenásobná regrese

1. Načtěte si data blood.csv. Datový soubor obsahuje informace o 11 pacientech: výši jejich systolického krevního tlaku, věk a hmotnost.

systolic	systolický krevní tlak
age	věk v letech
weight	hmotnost (v librách)

2. Převeďte hmotnost pacientů na kilogramy.

3. Zajímá nás, jak závisí výše krevního tlaku na věku a hmotnosti.

   (a) Prohlédněte si závislost (graf) výše systolického tlaku na věku a hmotnosti zvlášť.

   (b) Odhadneme model, ve kterém bude závislost systolického tlaku na obou veličinách:

   ```
   m=lm(systolic~age+weight)
   summary(m)
   ```

   Jak budeme interpretovat odhadnuté parametry? Závisí systolický tlak na věku i hmotnosti?

   (c) Kdybychom chtěli dostat interpretaci i pro absolutní člen, můžeme si posunout obě proměnné např. do jejich průměrné hodnoty

   ```
   m=lm(systolic~I(age-mean(age))+I(weight-mean(weight)))
   summary(m)
   ```

   (d) Předpovězte výši systolického tlaku pro 50-letého pacienta, který váží 90 kg.

4. Uvažujte data Brain.txt z předminulé hodiny. Zjistěte, jak závisí velikost mozku na hmotnosti, výšce a celkovém IQ.
   – Které proměnné mají významný vliv na velikost mozku? Jaký je jejich efekt?
   – Předpovězte velikost mozku člověka o výšce 174 cm, hmotnosti 65 kg a IQ 130.

Kategoriální proměnné v regresi

1. Uvažujte data Srazky.csv, kde jsou uvedeny průměrné měsíční srážky a teploty v jižních Čechách v letech 2000–2011.

   Načtěte si data a podívejte se na časový průběh řady množství srážek a na závislost srážek na teplotě. Na základě vhodných obrázků tedy odpovězte na otázky:

- Pozorujeme v časové řadě srážek nějaký dlouhodobý trend?
- Lze v řadě srážek pozorovat nějakou „pravidelnost“?
- Jakým způsobem závisí množství denních srážek na teplotě?

2. Odhadneme si model, ve kterém bude množství srážek záviset na teplotě, daném měsíci a bude obsahovat lineární trend v čase.

```
mod=lm(srazky~factor(mesic)+teplota+I(rok-2000))
summary(mod)
```

- Je časový trend statisticky významný?
- Je významná závislost na teplotě? Jak bychom interpretovali odhadnutý koeficient?
- Vysvětlíme si interpretaci koeficientů u jednotlivých měsíců.

Statistickou významnost kategoriální veličiny mesic zjistíme následovně:

```
anova(mod)
```

Zde opět R používá stejný princip sekvenčního testování, jak jsme již měli u vícenásobné analýzy rozptylu. pro jiný způsob je potom možné použít funkci Anova z knihovny car.

3. Budeme se zabývat odhadnutou sezónností. Odhady absolutních členů pro závislost srážek na teplotě jsou pro jednotlivé měsíce následující:

```
c=coef(mod)[1:12]
a=c(c[1],c[1]+c[2:12])
a
barplot(a,names=1:12)
```

Sezónnost se ale většinou znázorňuje tak, že se udávají rozdíly pro jednotlivá období oproti celkovému průměru. Takové rozdíly bychom dostali následovně:

```
fmesic=factor(mesic)
mod2=lm(srazky~fmesic+teplota+I(rok-2000),contrasts=list(fmesic=contr.sum))
c2=coef(mod2)[2:12]

b=c(c2,-sum(c2))
b
barplot(b,names=1:12)
```

4. Kdybychom chtěli uvažovat pro každý měsíc jinou lineární závislost srážek na teplotě (jiné směrnice přímek), pak bychom použili následující model:

```
mod3=lm(srazky~teplota*fmesic)
```

V tomto případě to však není potřeba, o čemž se přesvědčíme následujícím testem:

```
anova(mod3)
```

5. Nakonec se opět podíváme na splnění předpokladů:

```
r=resid(mod)

plot(r)

library(car)
qqPlot(r,dist="norm")

plot(r~fitted(mod))
```

Z posledního obrázku vidíme, že rozptyl reziduí není konstantní. Tento předpoklad lineární regrese tedy zřejmě není splněn.

Nesplnění předpokladů v regresi

– Porušení normality: Máme-li dostatečný počet pozorování, p-hodnoty zůstávají platné.
– Při porušení shody rozptylů nebo nezávislosti nejsou p-hodnoty v pořádku a je potřeba je „opravit“, např. následovně:

```
library(sandwich)
library(lmtest)

# puvodni testy:
coeftest(mod)

#opraveni heteroskedasticity
coeftest(mod, vcov = vcovHC)

# opraveni mozne zavislosti, tzv. autokorelovanosti

coeftest(mod, vcov = vcovHAC)
coeftest(mod, vcov = NeweyWest)
```