01RAD - přednáška 11, 19.11.2024

4.5 Korelované chyby

- zejména v časových nebo ekonomických datech se často objevuje korelace jednotlivých hodnot
- potom není splněn předpoklad nezávislosti chyb
- tento stav je třeba detekovat (někdy pomohou grafy reziduí)
- modely pro korelovaná data: Analýza časových řad

Pokud je přítomna autokorelace a chyby mají konstantní rozptyl, platí:

1) OLS odhad $\widehat{\beta}$ je nestranný, ale neplatí Gauss-Markovova věta, tzn. $\widehat{\beta}$ nemá nejmenší rozptyl
2) MSE $= \frac{1}{n-m-1} SSE$ (odhad σ^2) může být podstatně menší než skutečná hodnota σ^2
3) v důsledku bodu 2) mohou být zvětšeny hodnoty t-statistik, takže t-testy a IS nefungují
4) protože jsou chyby závislé, F-testy a t-testy nejsou přesně platné ani když jsou chyby normální

Durbin-Watson statistika

- omezíme se na pozorování získaná v čase $t = 1, 2, \ldots, n$
- a případ, že chyby e_t splňují podmínky autoregresního procesu 1. řádu (AR1), tj.

$$e_t = \varrho e_{t-1} + u_t, \quad |\varrho| < 1,$$

  kde ϱ je autokorelační koeficient, $u_t \sim N(0, \sigma_u^2)$ jsou nezávislé, $t = 1, \ldots, n$,
  a u_t je nezávislé na $e_t, t \geq 1$

- pro test $H_0 : \varrho = 0 \quad \times \quad H_1 : \varrho > 0$ se užívá Durbinova-Watsonova statistika

$$d = \frac{\sum_{t=2}^{n} (\widehat{e}_t - \widehat{e}_{t-1})^2}{\sum_{t=1}^{n} \widehat{e}_t^2}, \qquad \text{kde } \widehat{e}_t \text{ jsou rezidua modelu LR}$$

- pokud zamítneme H_0, odhadne se ϱ pomocí

$$\widehat{\varrho} = \frac{\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum_{t=2}^{n} \widehat{e}_t^2}$$

Poznámka: Dá se ukázat, že $d \approx 2(1 - \widehat{\varrho})$:

- kritické hodnoty určené Durbinem a Watsonem jsou tabelované

Test: $(H_0 : \varrho = 0 \quad \times \quad H_1 : \varrho > 0)$

1) spočítat hodnotu d
2) nalézt kritické hodnoty (d_L, d_U) pro dané n a $m + 1$
3)
   a) zamítnout H_0, pokud $d < d_L$
   b) nezamítnout H_0, pokud $d > d_U$
   c) pro $d_L < d < d_U$ test nerozhodne

Poznámka 4.17

- pro test $H_0 : \varrho = 0 \quad \times \quad H_1 : \varrho < 0$ lze použít popsaný test pro $d' = 4 - d$
- metody pro korekci autokorelace: **Cochrane-Orcutt**

5. Výběr regresního modelu

- budeme se zabývat výběrem nejvhodnější množiny regresorů
- špatná specifikace modelu (použití jiného než skutečného modelu) má dva hlavní důsledky:
  1) při vynechání některých proměnných modelu, jsou odhady parametrů ostatních proměnných vychýlené
  2) pokud jsou v modelu nějaké proměnné navíc, jsou obecně rozptyly odhadů parametrů pro ostatní proměnné velké
- volba „nejlepšího“ modelu je hledání kompromisu mezi

  a) přesností modelu b) jednoduchostí modelu (parsimony)

- „ideální model“ by měl mít nejmenší možný počet regresorů, který umožňuje adekvátní interpretaci (nebo predikci)
- obvykle neexistuje jednoznačný nejlepší model
- ani jednoznačná statistická procedura, jak ho najít

Poznámka 5.1

Důsledky vynechání proměnných ze skutečného, i když neznámého, modelu:

5.1 Kritéria pro porovnávání modelů

- předpokládejme, že máme k dispozici T proměnných (regresorů) včetně interceptu
- uvažujme podmnožinu p proměnných (včetně interceptu)

A) koeficient vícenásobné determinace R^2

$$R_p^2 = \frac{SSR_p}{SST} = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{SSE_p}{SST}$$

- při použití je třeba si uvědomit, že R_p^2 je rostoucí funkcí p
- maxima tedy nabývá pro $p = T$
- hledáme model, ve kterém přidání dalšího regresoru už nezpůsobí podstatný nárůst R_p^2
- často se používá upravený koeficient determinace

$$\overline{R}_p^2 = 1 - \frac{\frac{SSE_p}{n-p}}{\frac{SST}{n-1}}$$

B) $(R)MSE$

$$MSE_p = \frac{SSE_p}{n-p} = s_n^2, \qquad RMSE_p = s_n$$

- hledáme model s minimální hodnotou s_n^2

C) F-test pro vnořené modely

pro $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ a $\beta = (\beta_1^T, \beta_2^T)^T$ umíme otestovat $H_0: \beta_2 = 0$ pomocí F-testu

- R:
  `anova(mod1, mod2)` - pozor na vnořenost modelů
  `anova(mod)` - záleží na pořadí v jakém přidáváme proměnné do modelu

D) Mallows C_p, AIC, BIC

- kritéria beroucí více v poraz počet použitých regresorů
- lze je použít i pro nevnořené modely

- Mallows C_p

$$C_p = \frac{SSE_p}{\widehat{\sigma}^2} - n + 2p, \qquad \widehat{\sigma}^2 = \frac{SSE_T}{n-T}$$

Vlastnosti C_p:

1) snadno se počítá, SSE_p a $\widehat{\sigma}^2$ jsou implementované
2) pokud je $\widehat{\sigma}^2$ konzistentní odhad σ^2 (nezávisející na p), má C_p následující interpretaci:
   - porovnává, co zbývá vysvětlit pomocí modelů s p a T parametry
   - zvýhodňuje počet dostupných dat
   - penalizuje počet parametrů, které je třeba odhadnout
3) při zvyšování počtu regresorů: $\widehat{\sigma}^2$ je konst., SSE_p klesá, p roste
4) $C_T = T$
5) pokud je správný model s p parametry, dá se ukázat, že $C_p \approx p$ pro $n \gg T$
6) v praxi se volí model s nejmenším C_p ve skupině modelů splňujících $C_p \approx p$

Poznámka 5.2

Nevýhoda: pro dobrou interpretaci je třeba spočítat C_p pro všechny nebo většinu podmnožin regresorů.

- Akaikeho informační kritérium AIC

- obecná definice je
$$\mathrm{AIC} = -2\ell(\widehat{\theta}) + 2p^*,$$

- $\widehat{\theta}$ je MLE odhad parametru θ v modelu

- ℓ je logaritmus věrohodnostní funkce

- p^* je počet parametrů, které je třeba odhadnout $(p^* = p + 1$, počítáme i $\sigma^2)$

Pro náš model LR:

- odvodili jsme

$$\mathrm{AIC} = n \ln 2\pi + n + n \ln \frac{SSE}{n} + 2p^* \qquad (\text{alt. } \mathrm{AIC} = n \ln \frac{SSE}{n} + 2p^*)$$

- hledáme model s minimální hodnotou AIC
- AIC není mírou kvality modelu, je užitečná pro porovnávání modelů

AIC v R:

`AIC(.)` počítá $\mathrm{AIC} = n \ln 2\pi + n + n \ln \frac{SSE}{n} + 2p^*$, kde p^* je počet parametrů β, σ^2

`extractAIC(.)` počítá $\mathrm{AIC} = n \ln \frac{SSE}{n} + 2p$, kde p je jen počet parametrů β

- (Schwarzovo) bayesovské informační kritérium BIC

$$\mathrm{BIC} = -2\ell(\widehat{\theta}) + p^* \ln n$$

- více penalizuje počet parametrů $\Longrightarrow$ vybírá jednodušší modely s jednodušší interpretací než AIC
- BIC vyžaduje významnější příspěvek proměnné, aby byla zařazena do modelu
- R: `BIC(.)` nebo `AIC(.)`, `extractAIC(.)` s volbou `k = log(nobs(fit))`

E) PRESS statistika

- pokud je pro nás důležitá kvalita predikce, lze použít pro srovnání modelů statistiku

$$\mathrm{PRESS} = \sum_{i=1}^{n} \widehat{e}^2_{(-i)} = \sum_{i=1}^{n} \left(\frac{\widehat{e}_i}{1 - h_{ii}} \right)^2$$

- vybírá se model s minimální hodnotou této statistiky

5.2 Metody výběru modelu

1) Vyhodnocení všech možných modelů

- pro T dostupných regresorů tzn. naladit 2^T modelů, pak je porovnat pomocí nějakého kritéria
- náročné pro velká T (například $T = 10$ znamená 1024 modelů)

2) Zpětná eliminace (backward elimination)

- začneme s plným modelem a v každém kroku odstraníme jednu proměnnou
- tu, která nejméně přispívá modelu (měřeno F stat)
- nebo jejíž odstranění znamená největší zlepšení modelu (měřeno AIC)

Algoritmus:

1) naladíme model se všemi proměnnými
2) pro každou proměnnou spočteme částečnou F statistiku (nebo t-statistiku) jako by právě byla přidána do modelu, tzn. za předpokladu, že ostatní proměnné tam už jsou
3) pokud je nějaká F-statistika menší, než kritická hodnota F_{out}, vynecháme z modelu proměnnou s nejnižší hodnotou F

   ($F_{out} = F_{1-\alpha_{out}}(1, n-p)$, kde p je aktuální počet regresorů v modelu, $\alpha_{out} = 0.05, 0.1, \dots$)
4) opakujeme kroky 2) a 3), dokud všechny částečné F statistiky nejsou větší, než příslušná kritická hodnota F_{out}, tzn. nelze už vyřadit žádnou proměnnou

Poznámka: Místo F lze používat AIC.

3) Dopředná regrese (forward regression)

- začneme pouze s interceptem (nebo nutným minimálním modelem)
- v každém kroku přidáme jednu proměnnou, která má za následek největší zlepšení modelu (největší nárůst F nebo největší pokles AIC)
- tato metoda neumožňuje odstranit proměnnou, která už do modelu byla jednou přidána

Algoritmus:

1) naladíme minimální model
2) pro každou dostupnou proměnnou spočteme F statistiku pro test významnosti jejího přidání do modelu
3) pokud některá z těchto F statistik překračuje kritickou hodnotu F_{in}, přidáme do modelu proměnnou s nejvyšší hodnotou F statistiky
4) opakujeme kroky 2) a 3), dokud všechny F-statistiky pro zbývající proměnné nebudou menší než F_{in} nebo dokud nezbude žádná proměnná na přidání do modelu

Poznámka: I když tento postup zjednodušuje výběr modelu, často bohužel vede na zařazení proměnných, které nemají významný příspěvek, jakmile jsou zařazeny další proměnné.

4) Postupná regrese (stepwise regression)

- kombinace dvou předchozích metod
- v každém kroku algoritmu přidáme jednu proměnnou a poté zkontrolujeme, zda není možné nějakou odebrat
- budeme potřebovat dvě kritické hodnoty F_{in}, F_{out} (pro použití F statistiky)

Algoritmus:

1) naladíme minimální model
2) zjistíme, zda přidání nějaké další proměnné může zlepšit model (F nebo AIC)
   pokud ano, přidáme do modelu proměnnou, která má za následek největší zlepšení modelu
   (největší pokles AIC)
3) v novém modelu zjistíme, zda nelze některou proměnnou vynechat (opět pomocí AIC nebo F)
   pokud ano, vynecháme proměnnou, jejíž vyřazení má za následek největší zlepšení modelu
   (největší pokles AIC)
4) opakujeme kroky 2) a 3) do té doby, až nebude možné přidat ani ubrat žádnou proměnnou

Poznámka 5.3 (Princip marginality)

- pokud jsou v modelu vyšší mocniny nějakého regressoru, měly by tam být obsaženy i všechny jeho nižší mocniny (i když jsou případně nevýznamné)
- pokud je v modelu obsažena interakce dvou regressorů, měly by tam být i oba individuální regresory
- s každou interakcí vyššího řádu by měl model obsahovat i všechny interakce řádu nižšího $(a : b : c \;\rightarrow\; a : b, a : c, b : c)$

Poznámka 5.4

Jakmile nalezneme optimální model, je třeba řádně ověřit adekvátnost.

Příklad: data cement - backward elimination

```
drop1(lm(y ~ x1 + x2 + x3 + x4), test="F")

## Model: y ~ x1 + x2 + x3 + x4
##        Df Sum of Sq    RSS    AIC F value  Pr(>F)
## <none>              47.864 26.944
## x1      1   25.9509 73.815 30.576  4.3375 0.07082 .
## x2      1    2.9725 50.836 25.728  0.4968 0.50090
## x3      1    0.1091 47.973 24.974  0.0182 0.89592
## x4      1    0.2470 48.111 25.011  0.0413 0.84407

drop1(lm(y ~ x1 + x2 + x4), test="F")

## Model: y ~ x1 + x2 + x4
##        Df Sum of Sq    RSS    AIC  F value    Pr(>F)
## <none>               47.97 24.974
## x1      1    820.91 868.88 60.629 154.0076 5.781e-07 ***
## x2      1     26.79  74.76 28.742   5.0259   0.05169 .
## x4      1      9.93  57.90 25.420   1.8633   0.20540

drop1(lm(y ~ x1 + x2), test="F")

## Model: y ~ x1 + x2
##        Df Sum of Sq     RSS    AIC F value    Pr(>F)
## <none>                57.90 25.420
## x1      1    848.43  906.34 59.178  146.52 2.692e-07 ***
## x2      1   1207.78 1265.69 63.519  208.58 5.029e-08 ***
```

```
min.model <- lm(y ~ 1)
max.model <- lm(y ~ x1 + x2 + x3 + x4)

auto.backward <- step(max.model, direction = "backward",
        scope = list(lower=min.model, upper=max.model))

## Start:  AIC=26.94
## y ~ x1 + x2 + x3 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x3    1    0.1091 47.973 24.974
## - x4    1    0.2470 48.111 25.011
## - x2    1    2.9725 50.836 25.728
## <none>              47.864 26.944
## - x1    1   25.9509 73.815 30.576
##
## Step:  AIC=24.97
## y ~ x1 + x2 + x4
##
##        Df Sum of Sq    RSS    AIC
## <none>               47.97 24.974
## - x4    1      9.93  57.90 25.420
## - x2    1     26.79  74.76 28.742
## - x1    1    820.91 868.88 60.629
```

Příklad: data cement - forward selection

```
add1(min.model, max.model, test = "F")

## Model: y ~ 1
##         Df Sum of Sq     RSS    AIC F value    Pr(>F)
## <none>                2715.76 71.444
## x1       1   1450.08 1265.69 63.519 12.6025 0.0045520 **
## x2       1   1809.43  906.34 59.178 21.9606 0.0006648 ***
## x3       1    776.36 1939.40 69.067  4.4034 0.0597623 .
## x4       1   1831.90  883.87 58.852 22.7985 0.0005762 ***
```

```
add1(lm(y ~ x4), max.model, test = "F")

## Model: y ~ x4
##         Df Sum of Sq    RSS    AIC  F value    Pr(>F)
## <none>               883.87 58.852
## x1       1    809.10  74.76 28.742 108.2239 1.105e-06 ***
## x2       1     14.99 868.88 60.629   0.1725    0.6867
## x3       1    708.13 175.74 39.853  40.2946 8.375e-05 ***
```

```
add1(lm(y ~ x1 + x4), max.model, test = "F")

## Model: y ~ x1 + x4
##         Df Sum of Sq    RSS    AIC F value  Pr(>F)
## <none>               74.762 28.742
## x2       1    26.789 47.973 24.974  5.0259 0.05169 .
## x3       1    23.926 50.836 25.728  4.2358 0.06969 .
```

```
add1(lm(y ~ x1 + x2 + x4), max.model, test = "F")

## Model: y ~ x1 + x2 + x4
##         Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>               47.973 24.974
## x3       1   0.10909 47.864 26.944  0.0182 0.8959
```

```
auto.forward <- step(min.model, direction = "forward",
                     scope = list(lower=min.model, upper=max.model))
```

Příklad: data cement - stepwise regression

```
auto.both <- step(min.model, direction = "both",
    scope = list(lower=min.model, upper=max.model),
    k=log(nobs(max.model)))

signif(coef(auto.both), 3)

## (Intercept)          x1          x2
##      52.600       1.470       0.662

## Start:  AIC=72.01
## y ~ 1
##        Df Sum of Sq     RSS    AIC
## + x4    1   1831.90  883.87 59.982
## + x2    1   1809.43  906.34 60.308
## + x1    1   1450.08 1265.69 64.649
## + x3    1    776.36 1939.40 70.197
## <none>              2715.76 72.009

## Step:  AIC=59.98
## y ~ x4
##        Df Sum of Sq     RSS    AIC
## + x1    1    809.10   74.76 30.437
## + x3    1    708.13  175.74 41.547
## <none>               883.87 59.982
## + x2    1     14.99  868.88 62.324
## - x4    1   1831.90 2715.76 72.009

## Step:  AIC=30.44
## y ~ x4 + x1
##        Df Sum of Sq     RSS    AIC
## + x2    1     26.79   47.97 27.234
## + x3    1     23.93   50.84 27.987
## <none>                74.76 30.437
## - x1    1    809.10  883.87 59.982
## - x4    1   1190.92 1265.69 64.649

## Step:  AIC=27.23
## y ~ x4 + x1 + x2
##        Df Sum of Sq    RSS    AIC
## - x4    1      9.93  57.90 27.115
## <none>               47.97 27.234
## + x3    1      0.11  47.86 29.769
## - x2    1     26.79  74.76 30.437
## - x1    1    820.91 868.88 62.324

## Step:  AIC=27.11
## y ~ x1 + x2
##        Df Sum of Sq     RSS    AIC
## <none>                57.90 27.115
## + x4    1      9.93   47.97 27.234
## + x3    1      9.79   48.11 27.271
## - x1    1    848.43  906.34 60.308
## - x2    1   1207.78 1265.69 64.649
```