Föreläsning 8

Josef Wilzen

2024-11-20

Outline

Linjär regression

Antaganden? När är det lämpligt att använda linjär regresssion?

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t$$

Vektor och matrisform för regressionsmodellen:

$$y = X\beta + \epsilon = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \vdots \\ \\ 1 & x_{1,N} & x_{2,N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \\ \epsilon_N \end{bmatrix}$$

Kolumnen med ettor har vi för att kunna modellera interceptet β_0.

Antaganden? När är det lämpligt att använda linjär regresssion?

- Feltermen ϵ:
  - Väntevärde 0: $E[\epsilon_i] = 0$ för alla $i = 1, 2, \ldots, N$
  - Variansen är konstant över hela datamaterialet: $V[\epsilon_i] = \sigma^2$ för alla $i = 1, 2, \ldots, N$
  - Statistiskt oberoende
  - Normalfördelad: $\epsilon \sim N(0, \sigma^2)$

Beroende felterm?

- Vi har ett datamaterial och vi vill använda modellen

$$y = X\beta + \epsilon$$

- Anta att väntevärde 0 och variansen är konstant för feltermen ϵ.

- ...men vi har beroenden i feltermen....
  - Hur kan vi undersöka det?

- Vad kan vi göra? Slänga data???

Beroende felterm?

- ...vi antar en ny liknande modell. Gamla modellen:

$$y = X\beta + \epsilon$$

- Nya modellen:

$$y = X\beta + \eta$$

  Låt η modelleras med en AR/ARMA/ARIMA-modell

- FPP3: Chapter 10 Dynamic regression models
- TSAF: kap 3.7-3.7.2, 3.8, 3.10

Regression + ARMA

Modell

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \eta_t$$

ARMA modell för η_t

$$\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \ldots + \phi_p \eta_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_p \epsilon_{t-q} + \epsilon_t$$

Exempel: ARMA(2,1) för η

$$\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

Bakåtskiftsnotation:

$$(1 - \phi_1 B - \ldots - \phi_p B^p)\eta_t = (1 + \theta_1 B + \ldots + \theta_q B^q)\epsilon_t$$

Använder Hyndmans notation för MA-delen nu.

Regression + ARIMA

Modell

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \eta_t$$

ARIMA modell för η_t

$$\left(1 - \phi_1 B - \ldots - \phi_p B^p\right)\left(1 - B\right)^d \eta_t = \left(1 + \theta_1 B + \ldots + \theta_q B^q\right)\epsilon_t$$

d är antal diff

Skattning

Hur skattar vi den vanliga regressionsmodellen?

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t$$

$$y = X\beta + \epsilon$$

Hur får vi fram $\hat{\beta}$ och $\hat{y}$? $\rightarrow$ $\hat{\beta}_{OLS}$!

- Vad är skattningsformeln för $\hat{\beta}_{OLS}$?
- Vad står OLS för här?

Skattning

Hur skattar vi den vanliga regressionsmodellen?

- Minimera kvadratsumman av residualerna
- Residualerna: $r_t = y_t - \hat{y}_t$,
  - Skattning av β: $\hat{\beta}$
  - $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \ldots + \hat{\beta}_k x_{k,t}$
- kvadratsumma:

$$SSE = \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{N} r_t^2$$

- Linjär algebra ger:

$$\hat{\beta}_{OLS} = \left(X^\top X\right)^{-1} X^\top y$$

Där X och y ges av vektor och matrisformen för regressionsmodellen.

Skattning: Regression + ARIMA?

En metod: minimera residualerna för η $\rightarrow$ precis som i fallet med “vanlig” regression

- Detta ger sämre skattningar för β
- Test och konfidensintervall för β stämmer inte längre
  - Ofta leder det till att skattade standardavvikelsen (medelfelet) för β blir för litet vilket leder till för små KI och för låga p-värden $\rightarrow$ inte bra!!!
- Är finns det beroenden mellan observationerna i feltermen? $\rightarrow$ då ska vi inte göra test och intervall på den modellen!

Skattning: Regression + ARIMA?

Bättre alternativ: minimera residualerna för ϵ

- Kom ihåg: η modelleras med en AR/ARMA/ARIMA modell
- Cochrane–Orcutt estimering:
  - Först skatta $\hat{\beta}$ genom att minimera residualerna för η genom att använda vanlig OLS $\rightarrow$ detta ger oss en skattning av residualerna $\hat{r}$
  - Skatta en ARIMA modell på vektorn $\hat{r}$
  - Upprepa dessa två steg tills parametervärdena inte ändras nämnvärt eller max antal iterationer uppnås
- Denna metod funkar, men inte bästa lösningen

Skattning: Regression + ARIMA?

- Bättre: Skatta $\hat{\beta}$ och parameterar i ARIMA **samtidigt** genom att minimera residualerna för ϵ direkt $\rightarrow$ numerisk optimering av likelihoodfunktionen

- Likelihoodfunktionen:
  - Sannolikhetsfördelning för data givet värden på modellens parametrar: $l\left(y|\beta,\phi,\theta,\sigma^2\right)$
  - Vi vill hitta den uppstättning parametervärden $\left(\beta,\phi,\theta,\sigma^2\right)$ som gör att funktionen $log\left(l\left(y|\beta,\phi,\theta,\sigma^2\right)\right)$ har ett så stort värde som möjligt

    $\rightarrow$ kallas för maximum likelihood skattning (MLE)
  - Detta görs numerisk med en *optimeringsalgoritm*
  - Vanligt att anta att likelihoodfunktionen är normalfördelad om $y \in \mathbb{R}$
  - Paketet `fable`, exempel: `ARIMA(y ~ x1 + x2 + pdq(1,1,2))`

Inferens

Modell

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \eta_t$$

ARMA modell för η_t

$$\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \ldots + \phi_p \eta_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_p \epsilon_{t-q} + \epsilon_t$$

Om vi har en lämplig modell för η_t och vi har skattningar för alla parametrar i den modellen, då har vi:

$$\hat{\beta}_{GLS} = \left(X^\top \hat{\Sigma}^{-1} X\right)^{-1} X^\top \hat{\Sigma}^{-1} y$$

$$V\left[\hat{\beta}_{GLS}\right] = \left(X^\top \hat{\Sigma}^{-1} X\right)^{-1} \qquad \hat{\Sigma} = \Sigma\left(\hat{\phi}, \hat{\theta}\right)$$

GLS = Generalized least squares, (se kap 3.7 i TSAF)

Att tänka på

Krav på våra förklarande variabler: $x_1, x_2, \ldots, x_k$

- Om vi anser att x är deterministisk $\rightarrow$ OK!
  - Ex: Vanliga trendfunktioner för tiden tex linjär, kvadratisk, kubisk, logaritmisk trend
  - Ex: Dummyvariabler för månader, veckodagar, helgdagar etc
  - Ex: trigonometriska funktioner som vi använder för att modellera säsonger

- Om vi anser att x är stokastisk (slumpmässig) $\rightarrow$ då måste vi vara försiktiga!
  - Ex: vi låter en aktiekurs vara vår responsvariabel och vi låter en annan aktiekurs vara vår förklarande variabel
  - y och x har tydliga trender och är korrelerade $\rightarrow$ det kan vara en falsk korrelation!
    - Korrelation implicerar inte kausalitet!

Spurious regression

Från FPP3: länk

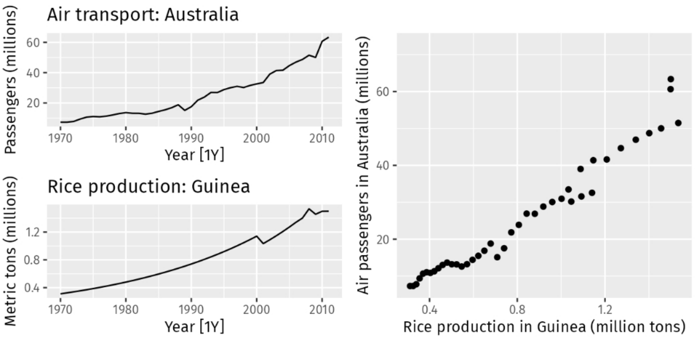


Figure 7.12: Trending time series data can appear to be related, as shown in this example where air passengers in Australia are regressed against rice production in Guinea.

Confounding factor?

Att tänka på

- Använd differentiering om y eller någon stokastisk x är icke-stationär
- Om y eller någon x behöver differentiering: då är det vanligt att differentiera alla $(y, x_1, x_2, \ldots, x_p)$ lika mycket

- Modell:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \ldots + \beta_k x_{k,t} + \eta_t$$

$$\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \ldots + \phi_p \eta_{t-p}$$

- Anpassade värden: vi måste kombinera regressionsdelen och ARMA/ARIMA delen
  - $\hat{y}_t = \hat{y}_{t,reg} + \hat{\eta}_t$
  - $\hat{y}_{t,reg} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \ldots + \hat{\beta}_k x_{k,t}$
  - $\hat{\eta}_t = \hat{\phi}_1 \eta_{t-1} + \hat{\phi}_2 \eta_{t-2} + \ldots + \hat{\phi}_p \eta_{t-p} + \hat{\theta}_1 \epsilon_{t-1} + \hat{\theta}_2 \epsilon_{t-2} + \ldots + \hat{\theta}_p \epsilon_{t-q}$

Prognoser

- Hur gör vi prognoser för $\hat{y}_{reg}$?
    - Om x_i är deterministisk: projicera in i framtiden, exempel: fortsätt trendfunktionen in i framtiden
    - Om x_i är stokastisk: Då måste vi ha någon tidseriemodell för x_i som kan göra en separat prognos in i framtiden för just x_i
- Givet att vi har lämpliga projektioner/prognoser för alla x, då använder vi bara regressionsekvationen för $\hat{y}_{reg}$
- $\hat{\eta}_t$ ges av de vanliga prediktionsformlerna för ARMA/ARIMA modeller

Demo i R

- Finns flera olika funktioner i R som kan skatta modellen
- Vi kollar på `fable`/`fpp3`

Utvärdera tidseriemodeller

Scenario: Ni ska hjälpa ett företag att göra prognos för hur många kunder som kommer att besöka deras butik de närmaste 6 veckorna. Detta för att kunna planera hur mycket personal som behövs. Tre stycken tidseriemodeller för “antal kunder per vecka” finns skattade och redo att användas. Dessa är skattade på två år av veckodata (104 obs).

- Vilken modell ska vi välja?
- Hur vet vi om en modell är bra i framtiden?

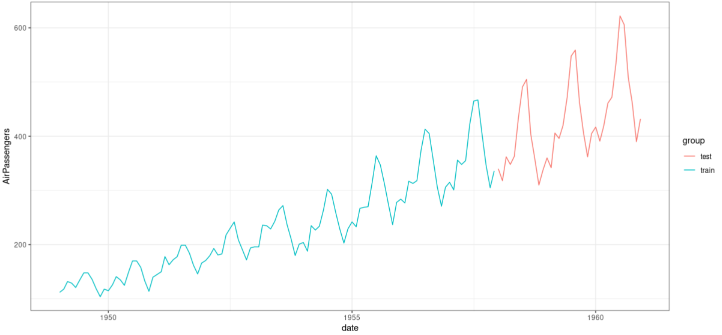
Utvärdera tidseriemodeller

- Utvärdera på historiska data: dvs den tidserie som vi har observerat just nu.
    - Olika utvärderingsmått, tex MSE på residualerna
    - Om modellen är på bra på historiska data, så kan vi hoppas att den är bra i framtiden
- Kan vi “simulera” framtiden på något sätt?

Träning och test

Kan vi "simulera" framtiden på något sätt? $\rightarrow$ ja!

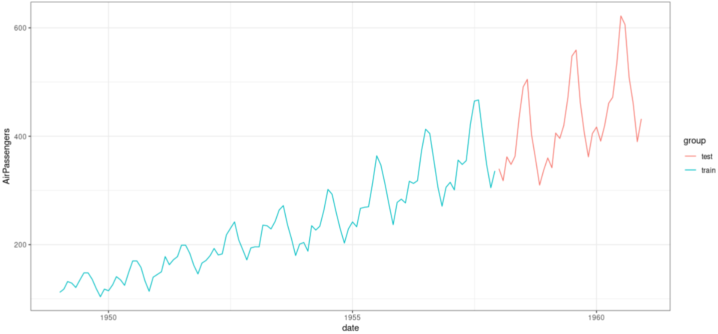
- Vi delar in tidserien i två delar:
  - Träningsdata[1]: skattningsdata, används för att skatta olika modeller
  - Testdata: Används för att utvärdera modeller, vår "simulerade framtid". Ofta så får testdata vara 5-30 % av de senaste obs i tidserien



[1]"Train" kommer ifrån maskininlärning, och används istället för "estimate"

Träning och test

Hur ska vi använda testdata? Några förslag?



Utvärdera tidseriemodeller

- Anpassade värden: $\hat{y}_t$
  - använd $\hat{y}_t$ för att beräkna residualer: $\hat{r}_t$
  - "residualen just nu", kallas innovationsresidualer
- Steg-1 prognoser: $\hat{y}_{t+1}$
  - Definerar anpassade värden för många tidseriemodeller
  - Använd $\hat{y}_{t+1}$ för att beräkna residualer: $\hat{r}_{t+1}$
  - $\hat{r}_{t+1}$ kallas ofta bara för residualer för många tidseriemodeller
- Prognoser:
  - Hur många tidssteg vill vi kunna göra bra prognoser?
  - Olika problem har olika *prognoshorisont* (H)
  - Låt data ha T tidpunkter, prognoser:

$$\hat{y}_{T+1|T}, \hat{y}_{T+2|T}, \dots, \hat{y}_{T+h|T}, \dots, \hat{y}_{T+H-1|T}, \hat{y}_{T+H|T}$$

- Vårt mål är att flerstegsprognoserna

$$\hat{y}_{T+1|T}, \hat{y}_{T+2|T}, \ldots, \hat{y}_{T+h|T}, \ldots, \hat{y}_{T+H-1|T}, \hat{y}_{T+H|T}$$

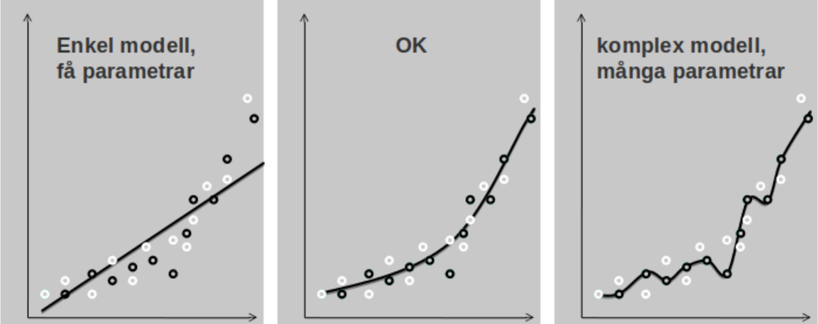
för framtida observationer ska vara bra.

- Vi vill att våra tidseriemodeller ska ha en bra *generaliserbarhet*.
- Vi kan skatta generliserbarheten genom att använda en extern testdatamängd.

Generaliserbarhet

- Avancerade modeller: överanpassar (overfitting)
  - Modellerar bruset i data $\rightarrow$ dåliga prediktioner
- Enkla modeller: underanpassar (underfitting)
  - Hittar inte mönstret/signalen i data $\rightarrow$ dåliga prediktioner
- Vi vill ha lagom komplicerade modeller!

Exempel: regression



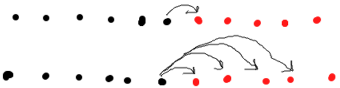
Träning och test

- Vi bestämmer en prognoshorisont (forecast horizon) för vårt problem
  - Hur långt in i framtiden vill vi göra bra prognoser?
- Testa vald prognoshorisont på testdata

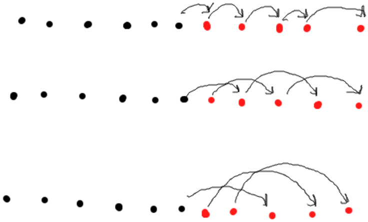
Träning och test

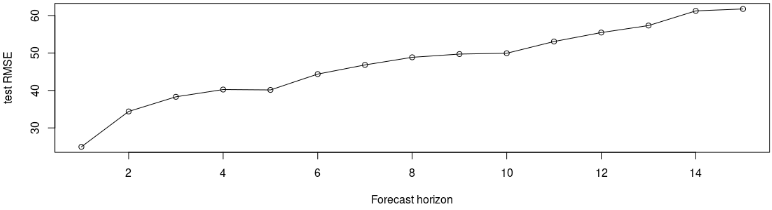
Modellens parameterar skattas bara på svarta punkter

Röd är testdata

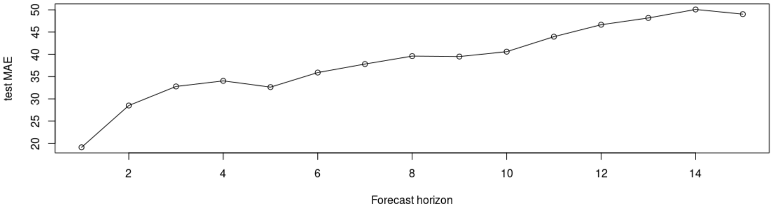


Bättre att använda alla testpunkter. Gör 1,2, ..., h-stegs prognoser för alla testpunkter. Beräkna sen utvärderingsmått för varje prognoslängd.



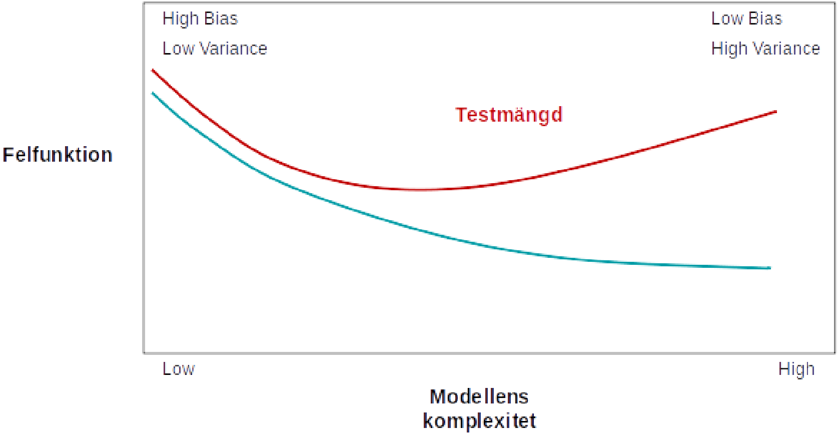


test RMSE
30
40
50
60
2
4
6
8
10
12
14
Forecast horizon



test MAE
20
25
30
35
40
45
50
2
4
6
8
10
12
14
Forecast horizon

Träning och test



Utvärdera tidseriemodeller

Utvärderingsmått som baseras på residualer:

- RMSE, MAE, MAPE
- Kan beräknas för olika typer av residualer

Interna utvärderngsmått för att skatta framtida prognosfel:

- Formler nedan för regressionsmodeller, k är antal förklarande variabler
- $AIC = T \cdot log\left(\frac{SSE}{T}\right) + 2\left(k+2\right)$
- Corrected AIC: $AIC_c = AIC + \frac{2(k+2)(k+3)}{T-k+3}$
- $BIC = T \cdot log\left(\frac{SSE}{T}\right) + \left(k+2\right) log\left(T\right)$

Modellval

- Skatta flera tidseriemodeller på träningsdata
- Undersök om de har en bra anpassning på träningsdata
  - Ser residualerna bra ut? Är de oberende? Är den skattade felvariansen hyfsat liten?
  - Fångar modellerna upp relevanta strukturer i data? tex trender och säsonger
  - Vi vill undvika att modellena underanpassar
- För att undvika överanpassnig och få bra framtida prognoser:
  - Interna utvärderngsmått
  - Använd valideringsdata $\rightarrow$ extern validering

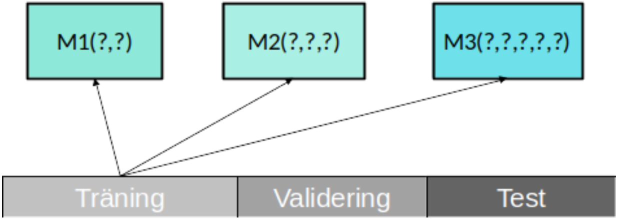
Modellval: Extern validering

Uppdelning av data:

- Träningsdata: används för att skatta modellens parametrar
- Valideringsdata: används för att välja modell
  - Använd något utvärderingsmått för lämpliga residualer
    $\rightarrow$ jämför valideringsfel för alla kandidatmodeller
- Testdata: används för att skatta generliserbarhet på den bästa modellen som vi valt

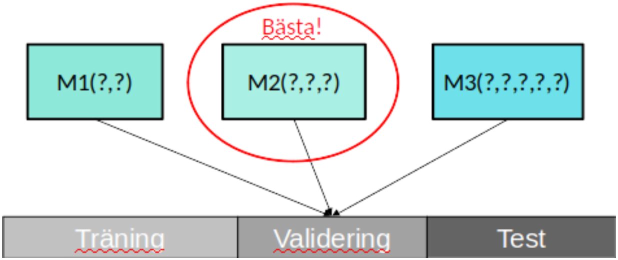
Modellval: Extern validering

Skatta tre modeller på träningsdata:



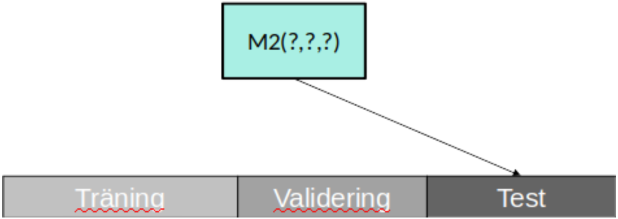
Modellval: Extern validering

Utvärdera på valideringdata:



Modellval: Extern validering

Skatta generliseringsfel på testdata:



I detta fal så använder vi modellen M2 för framtida prognoser.

Hyperparametrar

- En “högre ordningens parametrar”, anger övergripande egenskaper hos modellen
  - modelhyperparametrar
  - algoritmhyperparametrar
- Kan oftast inte skattas på “vanligt sätt” på träningsdata utan att riskera överanpassning
- Olika modellklasser har olika hyperparametrar, exempel:
  - Regression: antal förklarande variabler, val av designmatris
  - $ARIMA(p, d, q)$: Val av p, q och d.
- Exempel:
  - Regression, vilken trendfunktion?
    $\beta_1 \cdot t$ eller $\beta_1 \cdot t + \beta_2 \cdot t^2$ eller $\beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3$?
  - Ska vi välja AR(3), ARMA(2,3), eller ARMA(6,4) ?