IMM, 2005 -02-27
Poul Thyregod

Vedr. opgave om bilpriser

Udfordringen synes at være spørgsmålet om bestemmelse af den fittede værdi af prisen på en 7 år gammel Ford, og specielt usikkerheden herpå. Dette notat angiver forskellige metoder til besvarelse af dette spørgsmål.

Fittet værdi og prædikteret pris.

Første lige en bemærkning om begreberne "fittet værdi" og "prædikteret værdi" og deres usikkerheder. Ved *den fittede værdi* forstås sædvanligvis det estimerede kurvepunkt (liniepunkt), $\widehat{\eta}$, som jo netop er udtryk for middelværdien (her af de logaritmerede værdier). Dette kurvepunkt er behæftet med en estimationsusikkerhed, $\mathsf{V}[\widehat{\eta}]$ (hidrørende fra estimationsusikkerheden, $\mathsf{V}[\widehat{\boldsymbol{\beta}}]$ på koefficienterne).

Vi vil sædvanligvis bruge den fittede værdi til at *prædiktere* (forudsige) en værdi af logaritmen til prisen på en tilfældig 7 år gammel Ford. Men nu kommer der et yderligere variansbidrag fra den tilfældige variation omkring den forudsagte middelpris, hvorfor usikkerheden på prædiktionen af log-prisen på en tilfældig 7 år gammel Ford bliver $\mathsf{V}[\widehat{\eta}] + \sigma^2$.

I dette notat vil vi kun betragte usikkerheden, $\mathsf{V}[\widehat{\eta}]$ på den fittede værdi.

Bestemmelse af usikkerheden på den fittede værdi

Udgangspunkter er den lineære normalfordelingsmodel for logaritmen til prisen:
$\mathsf{Lpris} = \mathsf{Fabr}\ \mathsf{Alder}$
med estimaterne

$$\begin{array}{lcl} \mathsf{Intercept} & = & 12.5449 \\ \mathsf{Ford} & = & -0.4379 \\ \mathsf{Alder} & = & -0.1887 \end{array}$$

og med den estimerede kovariansmatrix for estimaterne

$$\widehat{\mathsf{D}[\widehat{\boldsymbol{\beta}}]} = \begin{bmatrix} 0.0116 & -0.0052 & -0.0010 \\ -0.0052 & 0.0085 & 7.613e-5 \\ -0.0010 & 7.613e-5 & 0.0001 \end{bmatrix}$$

Fittet værdi af log pris og varians herpå

Direkte fremgangsmåde

Den fittede værdi fås som

$$\widehat{\eta} = [1 \;\; 1 \;\; 7] \begin{bmatrix} 12.5449 \\ -0.4379 \\ -0.1887 \end{bmatrix} = 12.5449 - 0.379 - 7 \cdot 0.1887 = 10.7861$$

Variansen på den fittede værdi fås som

$$\widehat{\mathsf{V}[\widehat{\eta}]} = [1 \;\; 1 \;\; 7] \; \widehat{\mathsf{D}[\widehat{\boldsymbol{\beta}}]} \; [1 \;\; 1 \;\; 7]' = 0.0038 = 0.0615^2$$

idet vi benytter, at $\mathsf{V}[\mathbf{x}\,\widehat{\boldsymbol{\beta}}] = \mathbf{x}\,\mathsf{D}[\widehat{\boldsymbol{\beta}}]\,\mathbf{x}'$ (jvf formel (1.7) side 9 i noten).

Den elegante fremgangsmåde

Man kan lade SAS-systemet klare nogle af beregningerne ved at udvide datasættet med en "observation", hvor fabrikat er Ford, Alder er 7, men prisen er uoplyst (et punktum). Hvis man i menuen for output variable under Fit beder om "predicted" får man netop værdien 10.7862 for denne observation.

Det følger af udtrykket efter formel (3.23) i noten, at variansen for en fittet værdi kan udtrykkes ved diagonalelementet i hat-matricen som:

$$\mathsf{V}[\widehat{\eta}_i] = \sigma^2 \; h_{ii} \; ;$$

Diagonalelementet i hat-matricen svarende til "observationen" for den 7 år gamle Ford fås ved - i menuen for outputvariable - at bede om hatdiagonalelementerne. Man får $h_{28,28} = 0.0670$ og da $\widehat{\sigma}^2 = 0.0563$ har man altså

$$\widehat{\mathsf{V}[\widehat{\eta}_{28}]} = \widehat{\sigma}^2 \cdot h_{28,28} = 0.0038 = 0.0615^2$$

Fittet værdi af pris og usikkerhed herpå

Ved konfidensinterval

Man får et 95% konfidensinterval for log pris som

$$\widehat{\eta} \pm \mathrm{t}_{24,.975} \cdot \widehat{\mathsf{V}[\widehat{\eta}_{28}]} = 10.7861 \pm 2.048 \cdot 0.0615 = \begin{cases} 10.66 \\ 10.91 \end{cases}$$

Dette interval gælder eksakt under forudsætning af normalfordeling af logaritmerne.

Da transformationen $\mu = \exp(\eta)$ er monoton, overføres 95% konfidensintervallet for η i et (ligeleds eksakt) 95% konfidensinterval for μ ved denne transformation, hvorfor intervallet for kronebeløbet bliver (42 623 kr; 54 833 kr)

Ved brug af fejlophobningsloven

Fejlophobningsloven angiver approximationen

$$\mathsf{E}[g(X)] \approx g(\mathsf{E}[X]) \quad \mathsf{V}[g(X)] \approx g'(\mathsf{E}[X])^2 \mathsf{V}[X]$$

For $Y = g(X) = \exp(X)$ med $g'(X) = \exp(X)$ får man

$$\mathsf{E}[Y] \approx \exp(\mu_x) \quad \mathsf{V}[Y] \approx \exp(2\mu_x)\mathsf{V}[X]$$

hvorfor

$$\widehat{\mu}_{pris} = \exp(10.7861) = 48\ 344 \text{ kr}$$

og

$$\widehat{\mathsf{V}[\widehat{\mu}_{pris}]} \approx \exp(2 \cdot 10.7861) \cdot 0.0038 = (2\ 980 \text{ kr})^2$$

Ved angivelse af eksakt middelværdi og varians

I denne situation med normaltfordelte logaritmiske værdier kan man godt beskrive den eksakte fordeling af kronebeløbene, idet den såkaldte logaritmiske normalfordeling er velbeskrevet. Den logaritmisk normale fordeling, $\mathrm{LN}(\alpha, \beta^2)$ betegner den fordeling, hvis logaritmerede værdier følger en $\mathrm{N}(\alpha, \beta^2)$ jvf. afsnit 1.4.1 i noten.

Lad $X \in \mathrm{N}(\mu_x, \sigma_x^2)$, da gælder for $Y = \exp(X)$ at $Y \in \mathrm{LN}(\alpha_x, \beta_x^2)$ og at

$$\begin{aligned} \mathsf{E}[Y] &= \exp(\alpha_x + \frac{1}{2}\beta_x^2) \\ \mathsf{V}[Y] &= \big(\mathsf{E}[X]\big)^2 \left\{ \exp(\beta_x^2) - 1 \right\} \end{aligned}$$

Vi får derfor estimaterne

$$\widehat{\mu}_{pris} = \exp(10.7861 + 0.5 \cdot 0.0038) = 48\ 436 \text{ kr}$$

og

$$\widehat{\mathsf{V}[\widehat{\mu}_{pris}]} = (48\ 436)^2 \cdot [\exp(0.0038) - 1] = (2\ 989 \text{ kr})^2$$

Belysning af approximationsfejlen

I denne situation med normaltfordelte logaritmiske værdier kan vi altså eksplicit vurdere approximationsfejlen.

Lad $X \in \mathrm{N}(\mu_x, \sigma_x^2)$, da gælder for $Y = \exp(X)$ at $Y \in \mathrm{LN}(\alpha_x, \beta_x^2)$ og at

$$\begin{aligned} \mathsf{E}[Y] &= \exp(\alpha_x + \frac{1}{2}\beta_x^2) \\ \mathsf{V}[Y] &= \big(\mathsf{E}[X]\big)^2 \left\{ \exp(\beta_x^2) - 1 \right\} = \exp(2\alpha_x + \beta_x^2)\left\{ \exp(\beta_x^2) - 1 \right\} \end{aligned}$$

dvs

$$\mathsf{E}[\exp(X)] \geq \exp(\mathsf{E}[X])$$

hvor lighedstegnet kun gælder for $\mathsf{V}[X] = 0$

Den *relative fejl* ved at benytte $\exp(\mathsf{E}[X])$ som approximation til $\mathsf{E}[\exp(X)]$ er $1 - \exp(-\alpha_x^2/2)$. Jo større varians på X, desto større er den relative fejl.

Approximationen til variansen er

$$\mathsf{V}[Y] \approx \beta_x^2 \exp(2\alpha_x)$$

Jo større varians, β_x^2, desto større er afvigelsen fra den eksakte værdi

$$\mathsf{V}[Y] = \exp(2\alpha_x + \beta_x^2)\left\{ \exp(\beta_x^2) - 1 \right\}$$