

Notat 3.

Usikkerhedsvurderingen i projekt 1

— Anden reviderede version, der var desværre en fejl i den første.

Der er mange, som ikke har været i stand til at foretage en rimelig vurdering af vejetællingsmetodens usikkerhed, som metode til at fastsætte det årlige antal rejser. Det er heller ikke så nemt, må jeg indrømme, det kræver en del beskidte tricks og approksimationer. Jeg må indrømme, at jeg nok har overvurderet Jeres statistiske erfaring i den forbindelse, og jeg har været nødt til i nogen grad at se bort fra denne del af opgaven ved vurderingen af Jeres besvarelser (bortset fra de tilfælde hvor I har skrevet noget som helt åbenlyst er forkert, eller slet ikke har forsøgt). Dette notat handler om hvad I måske kunne have gjort.

Lad $i = 1, \dots, n$ betegne samtlige afgang fra en bestemt station i et bestemt år. Med x_i betegnes det faktiske antal passagerer i toget, med p_i resultatet af en (i de fleste tilfælde rent fiktiv) perrontælling, og med $v_i = v_{2i} + v_{4i}$ togets samlede vægt, opdelt på vægten af 2-vogns moduler og 4-vogns moduler.

Vi vil holde os til det problem der består i at estimere det samlede antal passagerer i toget ved disse afgang. Det er ganske vist ikke helt det tal vi er ude efter. Men hvis vi kan vurdere den relative fejl på dette tal, vil den jo nok være ret tæt på den relative fejl på det samlede antal rejser. Vi går jo ud fra, at omregningen fra disse stationstællinger til det samlede antal rejser foregår på en måde, som er helt uden fejl

Vi har to regressionsmodeller til rådighed for usikkerhedsvurderingen. Dels har vi fået at vide, at perrontællingerne som funktion af kontroltællinger inde i toget (som vi her vil gå ud fra er det samme som de sande værdier, altså x_i 'erne) kan beskrives ved modellen

$$(1) \quad p_i = \gamma x_i + 0.639\sqrt{x_i}U_i, \quad U_i \sim N(0, 1)$$

hvor koefficienten γ er estimeret til 0.988 plus/minus en normalfordelt estimationsfejl med middelværdi 0 og standardafvigelse 0.002. Vi har selv, til beskrivelse af perrontællingerne som funktion af vejninger, fundet (som en af mange, der er andre der kan bruges) modellen

$$(2) \quad p_i = \beta_2 v_{2i} + \beta_4 v_{4i} + 12.99Z_i, \quad Z_i \sim N(0, 1)$$

hvor koefficienterne er estimeret til

$\hat{\beta}_2 = 0.01135$ med standardafvigelse 0.0000546
 $\hat{\beta}_4 = 0.01279$ med standardafvigelse 0.0000154.

Hvis vi sætter fejleddene til 0 i begge modeller og simpelthen sætter de to udtryk for p_i lig med hinanden uden hensyn til estimationsfejl får vi relationen

$$0.988x_i = 0.01135v2_i + 0.01279v4_i$$

som ved summation m.h.t. $i = 1, \dots, n$ giver

$$(*) \quad 0.988 \sum x_i = \sum (0.01135v2_i + 0.01279v4_i).$$

Denne formel kan vi så bruge til at “beregne” det samlede passagertal $\sum x_i$ ved at dividere over med 0.988 (eller simpelthen erstatte 0.988 med et éttal; det sidste er nu en dårlig idé, som vi skal se).

Det er klart, at denne metode på en eller anden måde er det bedste vi kan præstere ud fra de oplysninger vi har. Spørgsmålet er, hvor stor en fejl vi begår ved at gøre det på den måde. Der er flere fejlkilder at tage hensyn til her. For det først har vi fjernet fejleddene på begge sider af lighedstegnet. De skal naturligvis indgå i overvejelserne. For det andet har vi indsat estimerne for γ , β_2 og β_4 uden hensyn til at de er behæftet med estimationsfejl. Det skal vi naturligvis også tage hensyn til.

Antag nu at modellen (1) bruges til, f.eks., at danne et skøn over hvad perrontællingerne ville have givet hvis vi havde talt samtlige afgange fra denne bestemte station i et år:

$$\sum p_i = \gamma \sum x_i + 0.639 \sum \sqrt{x_i} U_i$$

hvor de indgående summer er taget over sådan ca. (lavt sat) $365 \times 30 \approx 10000$ afgange. Hvis vi sætter det samlede passagertal til $80 \times 10000 = 800000$ får vi standardafvigelsen

$$\sqrt{\text{var} \left(0.639 \sum \sqrt{x_i} U_i \right)} = 0.639 \times \sqrt{800000} \approx 570.$$

Det betyder at relationen

$$\sum p_i = \gamma \sum x_i$$

(hvor begge sider er af størrelsesorden 800000) holder med en relativ fejl på $570/800000 = 0.07\%$.

Hvis vi for de samme ca. 10000 afgange adderer relationerne (2) får vi

$$\sum p_i = \sum (\beta_2 v2_i + \beta_4 v4_i) + (12.99 \sum Z_i).$$

Her har fejleddet standardafvigelse

$$\sqrt{\text{var}\left(12.99 \sum Z_i\right)} = 12.99\sqrt{10000} = 1299$$

hvilket betyder at relationen

$$\sum p_i = \sum (\beta_2 v_2 + \beta_4 v_4)$$

holder med en relativ fejl på $1299/800000 = 0.17\%$.

Hertil kommer estimationsfejlen:

For γ : $0.002/0.988 = 0.002$,

for β_2 : $0.0000546/0.01135 = 0.005$

og for β_4 : $0.0000154/0.01279 = 0.001$.

Da γ indgår som proportionalitetsfaktor skal den relative fejl på $\hat{\gamma}$ indgå direkte som et bidrag til den relative fejl på estimatet for det samlede passagertal. For β_2 og β_4 kan vi passende tage den største af de to relative fejl, 0.005, som et (sikkert alt for stort) skøn over den relative fejl på en hvilken som helst linearkombination af disse to koefficienter. Alt i alt får vi, ved at kombinere de fire relative fejl som om de var uafhængige, en samlet relativ fejl på

$$\sqrt{0.0007^2 + 0.0017^2 + 0.002^2 + 0.005^2} = 0.0057 \approx 0.6\%$$

der skal fortolkes som en standardafvigelse på en approksimativt normalfordelt afvigelse, så vi kan sige at den relative fejl på estimatet givet ved (*) af det samlede passagertal skal forsynes med en fejlangivelse på omkring $\pm 1.96 \times 0.0057$, dvs. cirka plus/minus en procent. Det er ikke mere end den systematiske fejl man ville begå ved at erstatte koefficienten 0.988 med 1, så det må anbefales at man undlader dette.

Her har vi kombineret de forskellige fejlkilder som om de var stokastisk uafhængige, og det er de jo nok ikke i virkeligheden. Men denne simplificering betyder en overvurdering af fejlen snarere end en undervurdering. I særdeleshed ser man at fejlen stammende fra perrontællemetoden på en måde indgår to gange med modsat fortegn. Hvis vi havde haft adgang til vejebaserede målinger kontrolleret direkte ved tællinger inde i toget ville denne usikkerhed helt forsvinde.

Hvis man absolut vil gå med både livrem og seler må man benytte formelen

$$0.0007 + 0.0017 + 0.002 + 0.005 = 0.0094 = 0.94\%$$

som svarer til “worst case” scenariet hvor alle fejlene er positivt korrelerede med korrelation 1. Så bliver fejlen altså næsten dobbelt så stor.

Nogle gode råd til DSB S-tog, som den perfekte rapport nok burde ende med (med forbehold for hvilken regressionsmodel man har valgt at bruge):

Det anbefales at benytte formelen

$$\hat{x}_i = \frac{1}{0.988} (0.01135v_{2_i} + 0.01279v_{4_i}) = 0.01149v_{2_i} + 0.01295v_{4_i}$$

til estimation af passagertallet ved afgang som funktion af vægtene v_{2_i} og v_{4_i} af togets henholdsvis korte og lange moduler. De passagertal man beregner på årsbasis ved brug af denne formel kan anses for nøjagtige inden for en fejlmargen (dobbelte standardafvigelse) på $\pm 1\%$ (hvor et *meget* forsigtigt skøn siger $\pm 1.8\%$).

Så længe kontrollen med vejemetoden foretages ved hjælp af perrontællinger, bør disse løbende kontrolleres (da de nemt kan ændre sig på grund af udskiftning af personale osv.) ved kontroltællinger inde i toget.

Korrektion for den afvigelse man herved finder bør foretages. I øjeblikket er denne afvigelse på -1.2% (svarende til at perrontællingerne undertæller i forhold til optællingerne inde i toget) og dermed ret stor i forhold til den tilfældige variation, som målt i standardafvigelse er på $0.5-1.0\%$.

Undersøgelsen kompliceres af at der ikke foreligger samhørende kontroltællinger inde i toget og vejebaserede tællinger. I samme øjeblik sådanne findes i tilstrækkeligt omfang, synes det at være en unødigt komplikation at foretage vurderingen af de vejebaserede tællinger ved hjælp af perrontællinger. Man kan lige så godt direkte basere kontrollen på tællinger inde i toget, og eventuelt helt afskaffe perrontællingerne. Det vil jo også være mere effektivt at kontrollere de vejebaserede tællinger ved tællinger der (ligesom disse) kan foretages ved en hvilken som helst station. Ligesom det vil give en sikrere bestemmelse af koefficienterne β_2 og β_4 , som formodentlig vil føre til et sikrere (og formodentlig mindre) skøn for usikkerheden.