Het gebruik van cijferanalyse ten tijde van Corona: zorg voor een plan B!

Gegevensgerichte cijferanalyse is te zien als een vorm van machine learning. Er wordt een relatie verondersteld tussen te controleren gegevens en andere gegevens, die geschat wordt op een trainingsset van betrouwbare data, en vervolgens als norm gebruikt voor nieuwe gegevens.

Een voorbeeld: een grootwinkelbedrijf beschikt voor alle filialen per periode de omzet en het aantal gewerkte uren. Die uren per periode zijn op juistheid en volledigheid gecontroleerd. De relatie tussen deze variabelen kan worden gebruikt in een cross-sectie model of in een tijdreeksmodel.

In een *cross-sectie* model wordt in een gekozen periode voor een gekozen aantal filialen de omzet gecontroleerd en wordt met (lineaire) regressie de omzet uit de uren voorspeld. Voor de overige filialen wordt de omzet op basis van de geschatte regressielijn voorspeld uit de uren. De voorspelde omzet per filiaal wordt gebruikt als norm voor de geadministreerde omzet.
In een *tijdreeks*model wordt voor een gekozen filiaal voor een gekozen aantal voorbije perioden de omzet gecontroleerd en wordt met (lineaire) regressie de omzet uit de uren voorspeld. Voor het huidige controlejaar wordt de omzet op basis van de geschatte regressielijn voorspeld uit de uren. De voorspelde omzet per periode wordt gebruikt als norm voor de geadministreerde omzet.

In beide gevallen zal de regressie een formule opleveren in de vorm y = a + b x waarbij y de voorspelde omzet en x het (gecontroleerde) aantal uren in die periode voor dat filiaal.

De term b in de formule geeft aan hoeveel omzet extra wordt gegenereerd door 1 extra gewerkt uur. Het is de marginale bijdrage van een extra uur, niet te verwarren met de gemiddelde omzet per uur (y/x = a/x + b). Als een accountant assurance wil ontlenen aan dit model, zal hij of zij dit bedrag per uur moeten begrijpen en kunnen uitleggen aan de gecontroleerde. Ik denk dat een zinnige marge een essentiële voorwaarde is voor het gebruik van cijferanalyse.

De term a in de formule vind ik minder spannend. Letterlijk geeft het de omzet aan bij 0 gewerkte uren. Daarom zou het fijn zijn als de uitkomst dicht bij 0 ligt, en/of dat de schatting statistisch niet significant is. Dat betekent dat de steekproef een getal ongelijk 0 opleverde maar dat niet aangetoond is dat dat getal in de populatie ook ongelijk 0 is.

Maar, als de uitkomst a niet getalsmatig te negeren is, is er nog steeds niet veel aan de hand. Het (lineaire) regressiemodel is gebaseerd op data in een range van uren per periode, en alleen binnen die range is de geschatte formule de beste schatting voor de relatie tussen y en x. Hoe verder het fictieve datapunt x=0 uren van (het gemiddelde van) de echte data voor x verwijderd ligt, des te minder zeker is de schatting dat bij x=0 uren ook y = a omzet hoort.

Die zelfde relevant range is op het ogenblik een belangrijke factor bij het beoordelen van de bruikbaarheid van cijferanalyse in de controle. Als in de huidige maanden zowel de omzet als de uren gehalveerd zijn ten opzichte van volgend jaar kan dat nieuwe datapunt misschien mooi op de doorgetrokken regressielijn liggen, er is niet noodzakelijk voldoende reden om aan te nemen dat de lineaire relatie tussen omzet en uren in de data die werden gebruikt om het model te schatten ook bestaat in datapunten buiten die range.

Het is dus niet ondenkbaar dat bij de controle van cijfers over 2020 het gebruik van cijferanalyse in een tijdreeksmodel met data uit 2018 en 2019 niet leidt tot voldoende controlezekerheid. Evengoed zal het lastig worden om cijfers uit 2020 in de toekomst te gebruiken bij het voorspellen van volgende jaren. Ik denk dat accountants op tijd op zoek moeten gaan naar een plan B. Een cross-sectie model kan daar een voorbeeld van zijn.