Supplement til Kapitel 10 - Pearsons korrelationskoefficient og regression.

I kapitel 10 i *Statistik* om regressionsanalyse er forklaringsgraden r^2 et mål for regressionslinjens tilnærmelse til datapunkterne. I kapitlet præsenteres også et mål for sammenhængen mellem to variable, Pearsons korrelationskoefficient $\rho(X,Y)$. På side 248 i *Statistik* påstås det, at $(corr(X,Y))^2 = r^2$ for en lineær sammenhæng, hvilket fx er efterprøvet i Excel på side 303.

Sætning

For en lineær sammenhæng gælder:

$(corr(X,Y))^2 = r^2$.

Bevis

Vi antager, at vi har et sæt af data (x_i, y_i), og der er en lineær sammenhæng, $y = f(x) = \hat{a} \cdot x + \hat{b}$, mellem *X* og *Y*. Vi bruger samme betegnelser som i kapitel 10, dvs. $\overline{y}$ er middelværdien af *y*'erne og $\hat{y}_i = f(x_i)$ er de af modellen forudsagte *y*-værdier.

1. Forklaringsgraden

Vi husker at totalvariationen er: $T = \sum_{i=1}^{N} (y_i - \overline{y})^2$. Idet vi bruger formlen for kvadratet på den toleddede størrelse kan det skrives ud som:

$$T = \sum_{i=1}^{N} (y_i - \overline{y})^2 = \sum_{i=1}^{N} ((y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}))^2$$
$$= \sum_{i=1}^{N} ((y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2 \cdot (y_i - \hat{y}_i) \cdot (\hat{y}_i - \overline{y}))$$
$$= \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2 + 2 \cdot \sum_{i=1}^{N} (y_i - \hat{y}_i) \cdot (\hat{y}_i - \overline{y}))$$

Det første led er residual sum af kvadrater $R = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ det andet led er den forklarede sum af kvadrater $E = \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2$. Det tredje led er nul, hvilket vi beviser sidst i afsnittet, og vi får derfor i alt $T = R + E$.

Divideres igennem med *T* fås: $1=\frac{R}{T}+\frac{E}{T}$ som er henholdsvis den brøkdel af variationen, der ikke forklares med den lineære model og den brøkdel, der er forklaret. Flytter vi om på ligningen fås $\frac{E}{T}=1-\frac{R}{T}=\frac{(T-R)}{T}$ som er den formel for forklaringsgraden vi giver på side 248 i *Statistik*.

Skrevet ud bliver ligningen øverst på siden:

$$1=\frac{\sum_i (y_i-\hat{y}_i)^2}{\sum_i (y_i-\bar{y})^2}+\frac{\sum_i (\hat{y}_i-\bar{y})^2}{\sum_i (y_i-\bar{y})^2}.$$

Hvor det altså er det andet led, der er forklaringsgraden.

2. Korrelationen

Fra kapitel 10 finder vi formlen for kovariansen og korrelationen af y_i og $\hat{y}_i$ som vi skal bruge:

$\text{cov}(y,\hat{y})=\sum_i (y_i-\bar{y})\cdot(\hat{y}_i-\bar{y})$ (s. 235) og $corr(y,\hat{y})=\frac{\text{cov}(y,\hat{y})}{\sqrt{\sum_i (y_i-\bar{y})^2\cdot(\hat{y}_i-\bar{y})^2}}$ (s. 238). De sættes sammen til

$$\begin{aligned}
corr(y,\hat{y}) &= \frac{\sum_i (y_i-\bar{y})\cdot(\hat{y}_i-\bar{y})}{\sqrt{\sum_i (y_i-\bar{y})^2\cdot(\hat{y}_i-\bar{y})^2}} \\
&= \frac{\sum_i (y_i-\hat{y}_i+\hat{y}_i-\bar{y})\cdot(\hat{y}_i-\bar{y})}{\sqrt{\sum_i (y_i-\bar{y})^2\cdot(\hat{y}_i-\bar{y})^2}} \\
&= \frac{\sum_i (y_i-\hat{y}_i)\cdot(\hat{y}_i-\bar{y})+(\hat{y}_i-\bar{y})^2}{\sqrt{\sum_i (y_i-\bar{y})^2\cdot(\hat{y}_i-\bar{y})^2}} \\
&= \frac{\sum_i (\hat{y}_i-\bar{y})^2}{\sqrt{\sum_i (y_i-\bar{y})^2\cdot(\hat{y}_i-\bar{y})^2}} \\
&= \sqrt{\frac{\sum_i (\hat{y}_i-\bar{y})^2}{\sum_i (y_i-\bar{y})^2}}.
\end{aligned}$$

Vi har altså vist at $(corr(y,\hat{y}))^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$ som er den samme formel som vi fandt for forklaringsgraden.

For det andet lighedstegn lagde vi det samme led til og trak det fra i summen i tælleren. Ved det tredje lighedstegn gangede vi den sidste parentes ind i den første. Ved det fjerde lighedstegn benyttede vi igen sætningen som bevises i del 3 af beviset. For det sidste lighedstegn brugte vi $\frac{x}{\sqrt{x}} = \sqrt{x}$.

Vi har altså nu vist at $(corr(y,\hat{y}))^2 = \frac{E}{T} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$.

Nu mangler vi kun at argumentere for, at korrelationen mellem *y* og $\hat{y}$ er det samme som korrelationen mellem *x* og *y*.

I formlen for varians kommer der et a^2 når man erstatter x med y-hat = ax+b, det bliver så numerisk værdi a ekstra i nævneren. Kovariansen vokser også med en faktor a (s. 235). For at vise det skal man vise at middelværdien af y-hat = a * middelværdien af x og E(Yhat*Y) = a*E(X*Y)

Der gælder:

$Cov(a \cdot X + b, c \cdot Y + d) = a \cdot c \cdot Cov(X, Y)$ og da $\hat{y}_i = \hat{a} \cdot x_i + \hat{b}$ får vi endelig:

$$corr(y,\hat{y}) = \frac{\text{cov}(y,\hat{y})}{\text{var}(y)\cdot\text{var}(\hat{y})} = \frac{\text{cov}(y,\hat{a}\cdot x+\hat{b})}{|\hat{a}|\cdot\text{var}(y)\cdot\text{var}(x)} = \frac{\hat{a}\cdot\text{cov}(y,x)}{|\hat{a}|\cdot\text{var}(y)\cdot\text{var}(x)} = \frac{\pm\text{cov}(y,x)}{\text{var}(y)\cdot\text{var}(x)} = \pm corr(x,y).$$

Da vi kvadrerer korrelationen forsvinder minusset.

3. Bevis for sidste led i *T* er lig 0

For en lineær sammenhæng har vi $\hat{y}_i = \hat{a} \cdot x_i + \hat{b}$ og $\bar{y} = \hat{a} \cdot \bar{x} + \hat{b} \Leftrightarrow \hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$. Det benytter vi til nedenstående omskrivninger.

$$\sum_{i=1}^{N} (y_i - \hat{y}_i) \cdot (\hat{y}_i - \bar{y}))$$

Først skriver vi første parentes om:

$$
\begin{aligned}
(y_i - \hat{y}_i) &= y_i - (\hat{a} \cdot x_i + \hat{b}) \\
&= y_i - (\hat{a} \cdot x_i + \overline{y} - \hat{a} \cdot \overline{x}) \\
&= y_i - \overline{y} - (\hat{a} \cdot x_i - \hat{a} \cdot \overline{x}) \\
&= y_i - \overline{y} - \hat{a} \cdot (x_i - \overline{x}).
\end{aligned}
$$

Dernæst den anden parentes:

$$
\begin{aligned}
(\hat{y}_i - \overline{y}) &= (\hat{a} \cdot x_i + \hat{b}) - \overline{y} \\
&= (\hat{a} \cdot x_i + \overline{y} - \hat{a} \cdot \overline{x}) - \overline{y} \\
&= (\hat{a} \cdot x_i - \hat{a} \cdot \overline{x}) \\
&= \hat{a} \cdot (x_i - \overline{x}).
\end{aligned}
$$

Og samler faktorerne:

$$
\begin{aligned}
\sum_{i=1}^{N} (y_i - \hat{y}_i) \cdot (\hat{y}_i - \overline{y})) &= \sum_{i=1}^{N} \hat{a} \cdot (x_i - \overline{x}) \cdot (y_i - \overline{y} - \hat{a} \cdot (x_i - \overline{x})) \\
&= \sum_{i=1}^{N} \hat{a} \cdot (x_i - \overline{x}) \cdot (y_i - \overline{y}) - \hat{a} \cdot (x_i - \overline{x})^2)
\end{aligned}
$$

Nu får vi brug for endnu en egenskab ved lineær regression. For mindste sum af kvadrater metoden er estimatet for hældningen givet ved:

$$
\hat{a} = \frac{\sum_{i=1}^{N} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \Leftrightarrow \hat{a} \cdot \sum_{i=1}^{N} (x_i - \overline{x})^2 = \sum_{i=1}^{N} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \ .
$$

Indsættes dette i sidste led i udledningen får vi

$$
\begin{aligned}
&= \sum_{i=1}^{N} \hat{a} \cdot ((x_i - \overline{x}) \cdot (y_i - \overline{y}) - (x_i - \overline{x}) \cdot (y_i - \overline{y})) \\
&= \hat{a} \cdot 0 \\
&= 0.
\end{aligned}
$$

Hvilket skulle vises.