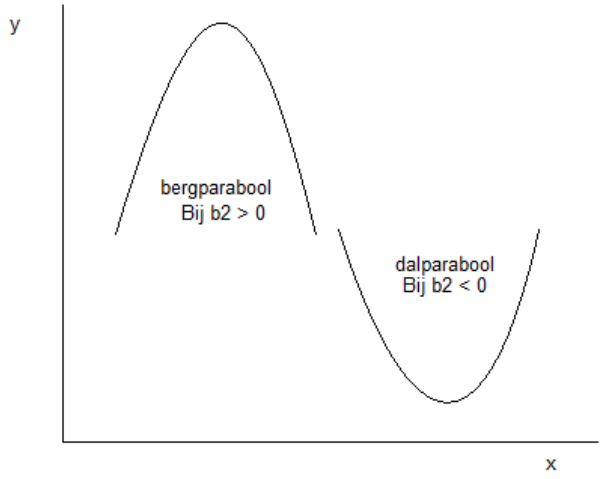
14. Model Building with Multiple Regression

Regressie gaat er van uit dat er sprake is van een lineair verband, of dat de relatie in ieder geval te benaderen is met een lineair verband. Maar soms is dat niet zo, en wijkt het verband te ver af van een rechte lijn. Het is belangrijk dat je dan niet uitgaat van lineariteit, omdat je dan hele verkeerde conclusies gaat trekken. De correlatie zal niet kloppen, en je schattingen zullen te ver afwijken.

Kwadratisch regressie model
We kijken naar kwadratische regressie modellen, ook wel polynomiale regressie modellen genoemd. De regressiefunctie ziet er dan zo uit : $E(y) = a + b1(x) + b2(x^2)$. Dat is als we even uitgaan van 1 verklarende variabele. Kwadratische verbanden zien er in een grafiek zo uit:



Wanneer de coëfficiënt van x^2 positief is, zal de data lopen in de vorm van een bergparabool. Als de coëfficiënt negatief is, zal de data lopen in de vorm van een dalparabool.

Het interpreteren van de coëfficiënten is nu iets verandert, in die zin dat we niet meer kunnen zeggen dat een toename van 1 op x, leidt tot een toename op y van b1. Want de toename op y, hangt nu meer af van de waarde van x, omdat we ook met een kwadraat van x zitten die verandert als x met 1 toeneemt.

De top van de parabool, het maakt niet uit of dat een berg- of dalparabool is, kan worden berekent met de volgende formule: $x = -b12(b2)$. Je neemt dus de negatieve coëfficiënt van b1, en deelt die door tweemaal de coëfficiënt b2. Op die manier bereken je de top. Deze x-waarde geeft dan aan waar de grafiek de maximale of minimale waarde van y bereikt.

Inferentie
Bij kwadratische modellen kijkt R^2 niet naar de sterkte van het verband, maar beschrijft het de vermindering van schattingsfouten door het gebruiken van een kwadratisch verband in plaats van een lineair verband.

De nulhypothese stelt dat er geen sprake van zal zijn dat het kwadratische verband iets toevoegt aan het model. De coëfficiënt van de gekwadrateerde waarde zal dan 0 moeten zijn: H0: b2 = 0. De significantie wordt berekent aan de hand van de t-statistiek, die wordt berekent door de coëfficiënt van b2 te delen door z'n bijbehorende standaardfout. De t-waarde en p-waarde worden ook gewoon gegeven in SPSS.

Wanneer je vermoedt dat je er sprake is van een kwadratisch verband, moet je eerst kijken of het significant is (je moet het verband dus vaststellen). Vervolgens kijk je naar de vorm van het verband (dalparabool, bergparabool). Daarna kijk je naar waar jouw data eigenlijk ligt. Zo kun je bijvoorbeeld wel een bergparabool hebben, met een piek bij x = 50. Maar als jouw data begint bij x = 45, dan zul je niet de hele bergparabool in je data weerzien. Daarna kun je het verband gaan beschrijven. Kijk daarbij naar het minimum/maximum, en beschrijf de invloed van de variabelen op y.