Inhoudsopgave

8. The confidence interval: meer onderwerpen

Population means kunnen afhankelijk zijn en onafhankelijk.

Samples zijn afhankelijk als een waarde beïnvloedbaar is door andere waarden. Afhankelijke samples zijn *matched pairs,* bijvoorbeeld gewicht en lengte, of hetzelfde individu of objecten die twee keer getest zijn, bijvoorbeeld medicijnen. Het individu of object moet voor én na de test gemeten worden. De twee objecten/individuën worden aangegeven met x en y en omdat ze afhankelijk zijn van andere waarden, krijgen ze de letter d.

$d_i = x_i - y_i$, hieruit volgt dat de *confidence interval for the difference between means* is:

$$\bar{d} - \frac{t_{n-1,\alpha/2} s_d}{\sqrt{n}} < \mu_d < \bar{d} + \frac{t_{n-1,\alpha/2} s_d}{\sqrt{n}} \text{ en } \mu_{\bar{d}} = \mu_x - \mu_y$$

De *margin of error* ziet er dan als volgt uit: $\frac{t_{n-1,\alpha/2} s_d}{\sqrt{n}}$

Onafhankelijke samples komen in drie situaties voor, met de objecten x en y, n_x en n_y, μ_x en μ_y, $\bar{x}$ en $\bar{y}$, σ^2_x en σ^2_y:

1. Beide populatie *variances* zijn bekend

De *mean* van de variable bereken je als volgt: $E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_x - \mu_y$

De *variance* met de formule: $Var(\bar{X} - \bar{Y}) = Var(\bar{X}) - Var(\bar{Y}) = \frac{\sigma^2_x}{n_x} + \frac{\sigma^2_y}{n_y}$

Omdat er sprake is van een normale verdeling, kan ook de z-waarde berekend worden: $Z = \frac{(\bar{X} - \bar{Y}) - (\mu x - \mu y)}{\sqrt{\frac{\sigma^2_x}{n_x} + \frac{\sigma^2_y}{n_y}}}$

En de *confidence interval*: $(\bar{x} - \bar{y}) - z_{\alpha/2} \sqrt{\frac{\sigma^2_x}{n_x} + \frac{\sigma^2_y}{n_y}} < \mu_x - \mu_y < (\bar{x} - \bar{y}) + z_{\alpha/2} \sqrt{\frac{\sigma^2_x}{n_x} + \frac{\sigma^2_y}{n_y}}$, waar de *margin of error* $z_{\alpha/2} \sqrt{\frac{\sigma^2_x}{n_x} + \frac{\sigma^2_y}{n_y}}$ is.

2. Beide populatie *variances* zijn onbekend, maar wel gelijk aan elkaar

σ^2 is niet bekend, waardoor de *mean* ook niet bekend is. Wel valt te concluderen dan $\sigma^2_x = \sigma^2_y = \sigma^2$. De *mean* zal nog altijd $\mu_x - \mu_y$ blijven, maar de *variance* formule veranderd:

$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) - Var(\bar{Y}) = \frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}$, wat ook een verandering in de z-waarde met zich mee brengt: $Z = \frac{(\bar{X} - \bar{Y}) - (\mu x - \mu y)}{\sqrt{\frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}}}$.

De *confidence interval* wordt dan als volgt:

$(\bar{x} - \bar{y}) - t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} < \mu_x - \mu_y < (\bar{x} - \bar{y}) + t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$

En de *margin of error*:

$t_{n_x+n_y-2\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$

De populaties samen hebben ook een *variance*, de *pooled sample variance*, s_p^2:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

De *Student's t distribution* functie ziet er als volgt uit:

$\frac{(\bar{x}-\bar{y}) - (\mu x - \mu y)}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$ = t, waarbij $n_x + n_y - 2$ de vrijheidsgraad is

3. Beide populatie *variances* zijn onbekend en niet gelijk aan elkaar

In de meeste gevallen zijn de *variances* niet bekend, maar zijn ze wel normaal verdeeld. De *confidence interval* heeft de volgende formule:

$(\bar{x} - \bar{y}) - t_{v,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} < \mu_x - \mu_y < (\bar{x} - \bar{y}) + t_{v,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$, waarbij

$t_{v,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$ de *margin of error* is.

De vrijheidsgraad wordt weergegeven met:

$v = \frac{(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y})^2}{\frac{(\frac{s_x^2}{n_x})^2}{(n_x-1)} + \frac{(\frac{s_y^2}{n_y})^2}{(n_y-1)}}$, wanneer de *samples sizes* gelijk zijn, wordt v: $v = (1 + \frac{2(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y})^2}{\frac{s_x^2}{s_y^2} + \frac{s_y^2}{s_x^2}})$ x (n-1)

Om de *proportions* te vinden bij de twee populaties, worden de volgende formules gebruikt:

Mean: $E(\hat{p}_x - \hat{p}_y) = P_x - P_y$

Variance: $Var(\hat{p}_x - \hat{p}_y) = \frac{P_x(1-P_x)}{n_x} + \frac{P_y(1-P_y)}{n_y}$

Bij grote *sample sizes* is de *random variable*: $Z = \frac{(\hat{p}_x - \hat{p}_y) - (P_x - P_y)}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}$

Confidence interval: $(\hat{p}_x - \hat{p}_y) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$,

waarbij de *margin of error*: $z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$ is.

Het is niet altijd duidelijk hoeveel *samples* er zijn gebruikt. Er is een formule die dat wel weergeeft, met een bekende variance:

$n = \frac{z^2_{\alpha/2}\sigma^2}{ME^2}$, dit moet een geheel getal zijn. Is dit niet het geval, rond dan af naar boven.

Ook bij de *population proportions* zijn de hoeveelheid gebruikte samples niet altijd gegeven. De formule hiervoor is: $n = \frac{z^2_{\frac{\alpha}{2}}(0.25)}{ME^2}$

Om de gehele populatie *mean* te achterhalen, zijn het aantal deelnemers, N, en de *variance* nodig:

$n = \frac{N\sigma^2}{(N-1)\sigma_{\bar{x}}^2 + \sigma^2}$ of $n = \frac{N * n_0}{n_0 + (N-1)}$, waarbij $n_0 = \frac{z^2_{\frac{\alpha}{2}}\sigma^2}{ME^2}$

De *variance* van de populatie wordt berekend aan de hand van de volgende formule:

$Var(\hat{p}) = \sigma^2_{\hat{p}} = \frac{P(1-P)}{n} \times \frac{(N-n)}{(N-1)}$

De hoeveelheid samples bereken aan de hand van de *proportions*, P, heeft de formule:

$n = \frac{NP(1-P)}{(N-1)\sigma_{\bar{p}}^2} + P(1-P)$ en de grootste waarde voor n: $n_{max} = \frac{0.25N}{(N-1)\sigma_{\bar{p}}^2} + 0.25$

9. De hypothese testen van een enkele populatie

Bij *hypothesis testing*
Null hypothesis: de eerste hypothese over een bepaalde parameter, die gehandhaafd wordt totdat er bewijs is gevonden dat deze niet klopt.

Alternative hypothesis: de tweede hypothese die de *null hypothesis* moet vervangen, als dat nodig is.

Het kan voorkomen dat de *null hypothesis* niet altijd afgekeurd kan worden, omdat de testen niet goed worden uitgevoerd. De *null hypothesis* kan dan ook nog altijd fout zijn en de *alternative hypothesis* goed.

De *null hypothesis* heeft als symbool H_0 : parameter (bv. μ) en de *alternative hypothesis* H_1 : parameter. Als de hypothese een specifiek getal is, is er sprake van een *simple hypothesis*. Als de hypothese groter of kleiner dan één getal is, is er sprake van een *one-sided composite alternative* (H1: μ </> 9) en als de hypothese alle andere mogelijke getallen kan zijn, is er sprake van *two-sided composite alternative hypothesis* (H1: $\mu \neq 9$).

De *null hypothesis* moet vaak genoeg aangetoond kunnen worden in de testen, anders zal de hypothese verworpen worden. Wanneer de *null hypothesis* meerdere malen aangetoond kan worden, zal deze behouden blijven. De testen zijn echter gebaseerd op *samples* en de sample parameter verschilt van de populatie parameter. Gedurende de testen zal hier rekening mee moeten worden gehouden, de *types of error*:

- **Type I error**: de *null hypothesis* wordt afgekeurd, terwijl deze wel waar is. De kans op het 'falen' van de hypothese is α, *significance level*, en is klein. De kans op het falen van het verwerpen van de *null hypothesis* is (1-α)
- **Type II error**: de *null hypothesis* wordt goedgekeurd, terwijl deze fout is. De kans hierop is β en de kans om deze foute hypothese alsnog te verwerpen is (1- β), wat de *power of the test* is.

De *types of error* formuleer je als: Reject H_0 if (sample parameter) </> xx.xx (getal).
Er bestaat een verband tussen de *types of error*. De waarden van α en β zijn met elkaar verbonden, bij een toename van α zal β afnemen en andersom. Dit is geen lineair verband.

Daarnaast is de *power of the test* te bereken bij de *type II error*: Power = P (Reject H_0 | μ, ($\mu \subset H_1$), waarbij P de power is en ⊂ betekent dat μ een element van H_1 is, maar geen H_1 zelf mag zijn.
Het vergroten van de *sample size* vergroot tevens de *power*.

Counterfactual argumemt: wanneer H_0 wordt verworpen en het bewijs dit ook laat zien, heeft het veel minder consequenties dan wanneer H_0 niet wordt verworpen en deze wel fout blijkt te zijn.

De H_0 wordt verworpen. wanneer deze boven de $\overline{x}_c$ waarde uitkomt. De $\overline{x}_c$ waarde is de *critical value*. Bij een normale verdeling en een bekende populatie *variance* kan de *critical* value als volgt worden berekend:

Formuleer de *null hypothesis*: $H_0 : \mu = \mu_0$

Formuleer de *alternative hypthesis*: $H_1 : \mu > \mu_0$
Formuleer de *decision rule*: Reject H_0 if $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha$ of Reject H_0 if $\bar{x} > \overline{x_c} = \mu_0 + z_\alpha(\sigma/\sqrt{n})$, waarbij z de *standard normal random variable* en is gegeven in de tabel in het boek

Er is echter nog een manier om de zekerheid van de *null hypothesis* te testen. Deze test wordt gedaan met behulp van de *p-value*.
P-value: de mogelijkheid voor het verkrijgen van een waarde van de test statistiek net iets groter dan de eigenlijke verkregen waarde wanneer de *null hypothesis* goed is. Met andere woorden: de kleinste *significance level* waarbij de *null hypothesis* verworpen kan worden.

De formule voor de *p-value* is: $P(\bar{x} > xx \mid H_c: \mu = xx) = P(Z > \frac{\bar{x}-\mu_c}{\sigma/\sqrt{n}})$, hier komt een z- waarde uit en in de *normal probability table* in je boek, staat het oppervlak erbij. De p wordt dan 1- oppervlak. Om de *null hypothesis* te testen wordt de *decision rule* gebruikt:
Reject H_0 if p-value $< \alpha = \mu_0 + z_p(\sigma/\sqrt{n})$.

Vaak zul je tegenkomen dat de *alternative hypothese* groter moet zijn dan de *null hypothesis*, maar andersom. Als de *sample mean* al kleiner is dan μ_0 zal de *null hypothesis* per direct vervallen. Om de juiste *decision value* te achterhalen, maak je gebruik van de formule:
$Z = \frac{\bar{x}-\mu_c}{\sigma/\sqrt{n}}$. Als Z een groot negatief getal is, dan volgt $P(Z < -z_\alpha) = \alpha$, wat de *significance level* is van de *null hypothesis*. De *decision rule* die dan wordt toegepast is:
Reject H_0 if $\frac{\bar{x}-\mu_c}{\sigma/\sqrt{n}} < -z_\alpha$ of Reject H_0 if H_1 if $\bar{x} < \overline{x_c} = \mu_0 - z_\alpha(\sigma/\sqrt{n})$, waarbij $\overline{x_c}$ de *critical value* is. Als σ bekend is, trek je deze van de F(z) af. zie tabel voor standaard z- waarden in boek, en vindt je de z: $z = F(z) - \alpha$. Deze manier van testen is volgens de *Type I error*.

Naast de enkele *alternative hypothesis* bestaat er ook de *two-sided alternative hypothesis*. In de standaard situtatie geldt:
$H_0 : \mu = \mu_0$ en $H_1 : \mu \neq \mu_0$

Als $\bar{x}$ veel groter of kleiner dan μ_1 is, zal de *null hypothesis* per direct verworpen worden. Bij een kleiner verschil tussen $\bar{x}$ en μ_0 zal H_c getest moeten worden, dit is wel bij een bekende variance. De *significance level* α wordt weer gebruikt, maar wordt wel door 2 gedeeld:
$P(Z > z_{\alpha/2}) = \alpha/2$ en $P(Z < z_{\alpha/2}) = \alpha/2$

De *decision rule* voor het testen wordt dan:
Reject H_0 if $\frac{\bar{x}-\mu_c}{\sigma/\sqrt{n}} < -z_{\alpha/2}$ en Reject H_0 if $\frac{\bar{x}-\mu_c}{\sigma/\sqrt{n}} > -z_{\alpha/2}$
of Reject H_0 if $x < \mu_0 - z_{\alpha/2}(\sigma/\sqrt{n})$ en Reject H_0 if $x > \mu_0 + z_{\alpha/2}(\sigma/\sqrt{n})$

Ook via *Type II error* kan de *p-value* berekend worden:
p-value = $2P(|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}| > z_{p/2} \mid H_0 : \mu = \mu_0)$, waarbij $z_{p/2}$ de kleinste waarde is waar de H_0 door verworpen wordt.

Hypotheses testen wanneer de variance onbekend is,kunnen ook voorkomen. Hiervoor is de *Student's t distribution* nodig. Bij *samples sizes* groter dan 100, zal de normale verdeling gebruikt worden om de *Student's t distribution* te berekenen: $t_{n-1} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$.

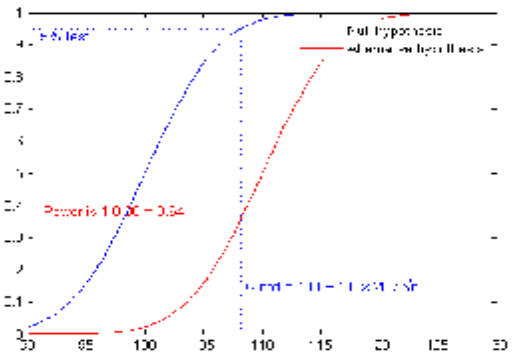
Enkele manieren om de *null hypothesis* te testen zijn:

a. $H_0 : \mu = \mu_0$ of $H_0 : \mu \leq \mu_0$ en $H_1 : \mu > \mu_0$
   Decision rule: Reject H_0 if $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} > t_{n-1,\alpha}$ of Reject H_0 if H_0 if $\bar{x} > \overline{x_c} = \mu_0 + t_{n-1,\alpha}(s/\sqrt{n})$
b. $H_0 : \mu = \mu_0$ of $H_0 : \mu \geq \mu_0$ en $H_1 : \mu < \mu_0$
   Decision rule: Reject H_0 if $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} < t_{n-1,\alpha}$ of Reject H_0 if H_0 if $\bar{x} < \overline{x_c} = \mu_0 + t_{n-1,\alpha}(s/\sqrt{n})$
c. $H_0 : \mu = \mu_0$ en $H_1 : \mu \neq \mu_0$
   Decision rule: Reject H_0 if $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} > t_{n-1,\alpha/2}$ en Reject H_0 if $s < -t_{n-1,\alpha/2}$
   of Reject H_0 if H_0 if $\bar{x} > \mu_0 + t_{n-1,\alpha/2}(s/\sqrt{n})$ en Reject H_0 if H_0 if $\bar{x} < \mu_0 - t_{n-1,\alpha/2}(s/\sqrt{n})$

De *p-value* bereken je zoals deze ook bij *Type II error* is berekend.

Ook de *population proportions* hebben hypothesen. Eerder is gemeld dat P de *mean* is bij de *proportions* en $\hat{p}$ de *sample proportion*. Bij de standaard normale statistiek is z: $Z = \frac{\hat{p}-P}{\sqrt{P(1-P)/n}}$. De hypothesen die hierbij opgesteld kunnen worden zijn:

a. $H_0 : P = P_0$ of $H_0 : P \leq P_0$ en $H_1 : P > P_0$
   en de *decision rule* die hierbij hoort: Reject H_0 if $\frac{\hat{p}-P}{\sqrt{P(1-P)/n}} > z_\alpha$
b. $H_0 : P = P_0$ of $H_0 : P \geq P_0$ en $H_1 : P < P_0$
   en de *decision rule* die hierbij hoort:
   Reject H_0 if $\frac{\hat{p}-P}{\sqrt{P(1-P)/n}} < -z_\alpha$
c. $H_0 : P = P_0$ en $H_1 : P \neq P_0$
   en de *decision rule* die hierbij hoort:
   Reject H_0 if $\frac{\hat{p}-P}{\sqrt{P(1-P)/n}} > z_{\alpha/2}$ en
   Reject H_0 if $\frac{\hat{p}-P}{\sqrt{P(1-P)/n}} < -z_{\alpha/2}$



Hier boven zijn verschillende manieren om de geldigheid van de *null hypothesis* aan te kunnen tonen. Maar het kan ook voorkomen dat er wordt gefaald om deze *null hypothesis* te verwerpen en er dus sprake is van een *Type II* error. De *Type II error* formule is:

$\beta = P(\bar{x} < \bar{x}_c \mid \mu = \mu^*) = P(z < \frac{x_c-\mu^*}{\sigma/\sqrt{n}})$ als $H_0 : \mu = \mu$ en $H_1 : \mu > \mu$, waarbij $\mu^* = \mu$ en $x_c = > \mu_0 + z_\alpha(\sigma/\sqrt{n})$.

De *power* van de test is 1 – β en zal voor elke μ* anders zijn en grafisch gezien ziet deze er als volgt uit:

Enkele kenmerken van de power functie zijn:

- Hoe verder μ van μ_0 af ligt, hoe groter de power
- Hoe kleiner α, hoe kleiner de power
- Hoe groter σ, hoe kleiner de power

- Hoe groter de sample size, hoe groter d e power
- Als $x_c = 0.5$, dan is de power ook 0.5

Als laatst kan ook de variance, σ^2, van de populatie getest worden. De *null hypothesis* wordt:
$H_0 : \sigma^2 = \sigma^2_0$.

Als de *null hypothesis* waar is, dan is de *random variable*:
$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$, waarbij n-1 = v.
Enkele manieren om de *variance* te berekenen:

- $H_0 : \sigma^2 = \sigma^2_0$ of $H_0 : \sigma^2 \leq \sigma^2_0$ en $H1 : \sigma^2 > \sigma^2_0$

de decision rule is dan Reject H_0 if $\frac{(n-1)s^2}{\sigma^2_0} > \chi^2_{n-1,\alpha}$

- $H_0 : \sigma^2 = \sigma^2_0$ of $H_0 : \sigma^2 \geq \sigma^2_0$ en $H1 : \sigma^2 < \sigma^2_0$

de decision rule is dan Reject H_0 if $\frac{(n-1)s^2}{\sigma^2_0} < \chi^2_{n-1,1-\alpha}$

- $H_0 : \sigma^2 = \sigma^2_0$ en $H1 : \sigma^2 \neq \sigma^2_0$

de decision rule is dan Reject H_0 if $\frac{(n-1)s^2}{\sigma^2_0} > \chi^2_{n-1,\alpha/2}$ en Reject H_0 if $\frac{(n-1)s^2}{\sigma^2_0} < \chi^2_{n-1,\alpha/2}$

Waar χ^2_{n-1} een *chi-square random variable* is en $P(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha$

10. Testen van hypotheses met twee populaties

In het voorgaande hoofdstuk is het testen van hypotheses besproken. De *null hypothesis* en de *alternative hypothesis* worden opgesteld en de null hypothesis wordt getest. Tot nu toe gebeurde dit altijd met één parameter, maar dit kan ook met meerdere parameters. Zo zal in een situatie waar twee gemiddeldes zijn de null hypothesis en de alternative hypothesis er als volgt uit zien:
$H_0 = \mu_1 - \mu_2 = 0$ en $H_1 = \mu_1 - \mu_2 > 0$

Ook is besproken dat variabelen afhankelijk en onafhankelijk van elkaar kunnen zijn. Als eerst worden de afhankelijke samples besproken.

In een situatie met n als de hoeveelheid matched pairs van x en y en de population means μ_x en μ_y, kan de *observed sample mean* berekend worden. De observed sample mean, $\bar{d}$, is het verschil tussen de sample mean van x en de sample mean van y. Ook de variance is bij het testen van twee parameters is belangrijk. Als het om twee populaties gaat, is het verschil in de variance, s_d, groter tussen dan bij twee matched pairs.

Om de null hypothesis te testen met verschillende parameters, is het significantie niveau α nodig, s_d is bekend en de T-tabel. De volgende formules zijn bij het testen van toepassing:

- $H_0 : \mu_x - \mu_y = D_0$ of $H_0 : \mu_x - \mu_y \leq D_0$
  $H_1: \mu_x - \mu_y > D_0$
  Hieruit volgt de decision rule: Reject H_0 **if** $\frac{\bar{d} - D_0}{s_d/\sqrt{n}} > t_{n-1,\alpha}$
- $H_0 : \mu_x - \mu_y = D_0$ of $H_0 : \mu_x - \mu_y \geq D_0$
  $H_1 : \mu_x - \mu_y < D_0$
  Hieruit volgt de decision rule: Reject H_0 **if** $\frac{\bar{d} - D_0}{s_d/\sqrt{n}} < -t_{n-1,\alpha}$
- En het twee-zijdige alternatief
  $H_0 : \mu_x - \mu_y = D_0$
  $H_1 : \mu_x - \mu_y \neq D_0$
  Hieruit volgt de decision rule: Reject H_0 **if** $\frac{\bar{d} - D_0}{s_d/\sqrt{n}} > t_{n-1,\alpha/2}$ of $\frac{\bar{d} - D_0}{s_d/\sqrt{n}} < -t_{n-1,\alpha/2}$

In veel gevallen zal D_0 0 zijn, omdat de population means gelijk aan elkaar zijn. De sample means zijn dit vaak niet!

Met onafhankelijke variabelen met een normale verdeling, is de Z-tabel nodig. De population variance van x en y zijn gegeven, net zoals de population means. De formule $Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$ kan bij elke hypothese gebruikt worden, zo lang de sample size maar groot is. Als n > 100 mogen de population variances vervangen worden door de sample variances. Om de null hypothese te testen, zijn er weer drie formules van toepassing waarbij α en σ^2 van x en y gegeven zijn:

- $H_0 : \mu_x - \mu_y = D_0$ of $H_0 : \mu_x - \mu_y \leq D_0$
  $H_1 : \mu_x - \mu_y > D_0$
  Hieruit volgt de decision rule: Reject H_0 **if** $\frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} > z_\alpha$

- $H_0 : \mu_x - \mu_y = D_0$ of $H_0 : \mu_x - \mu_y \geq D_0$
  $H_1 : \mu_x - \mu_y < D_0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} < -z_\alpha$

- En het twee-zijdige alternatief
  $H_0 : \mu_x - \mu_y = D_0$
  $H_1 : \mu_x - \mu_y \neq D_0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} < -z_{\alpha/2}$ of $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} > z_{\alpha/2}$

Population variances zijn niet altijd bekend. Bij onbekende population variances en n < 100 is de T-tabel nodig. Er vanuit gaande dat de population variances gelijk zijn aan elkaar, kan alsnog een t-waarde berekend worden en de null hypothese getest worden. De variance estimator wordt weergegeven met s_p^2 en is als volgt te bereken:

$s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{(n_x+n_y-2)}$

De null hypothese, met de variance estimator en significance level α, is te testen aan de hand van de volgende formules:

- $H_0 : \mu_x - \mu_y = D_0$ of $H_0 : \mu_x - \mu_y \leq D_0$
  $H_1 : \mu_x - \mu_y > D_0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} > t_{n_x+n_y-2,\alpha}$

- $H_0 : \mu_x - \mu_y = D_0$ of $H_0 : \mu_x - \mu_y \geq D_0$
  $H_1 : \mu_x - \mu_y < D_0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} < -t_{n_x+n_y-2,\alpha}$

- En het twee-zijdig alternatief
  $H_0 : \mu_x - \mu_y = D_0$
  $H_1 : \mu_x - \mu_y \neq D_0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} > t_{n_x+n_y-2,\alpha}$

  of $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} < -t_{n_x+n_y-2,\alpha}$

Let hierbij op de $n_x + n_y - 2$! Het gaat om een kleinere populatiegrootte en twee verschillende variances, σ_x en σ_y, én niet om de difference variance, s_d.

Bij een aanname dat de population variances niet gelijk aan elkaar zijn, zal de degrees of freedom aangepast worden om de juiste kritieke t-waarde te krijgen. Bij gegeven sample variances en significance level α wordt de degrees of freedom als volgt berekend:

$$v = \frac{\left[\left(\frac{s_x^2}{n_x}\right)+\left(\frac{s_y^2}{n_y}\right)\right]^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x-1}+\frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y-1}}$$

De null hypoteses kunnen getest worden met de volgende formules:

- $H_0 : \mu_x - \mu_y = D_0$ of $H_0 : \mu_x - \mu_y \leq D_0$
  $H_1 : \mu_x - \mu_y > D_0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} > t_{v,\alpha}$
- $H_0 : \mu_x - \mu_y = D_0$ of $H_0 : \mu_x - \mu_y \geq D_0$
  $H_1 : \mu_x - \mu_y < D_0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} < -t_{v,\alpha}$
- En het twee-zijdige alternatief
  $H_0 : \mu_x - \mu_y = D_0$
  $H_1 : \mu_x - \mu_y \neq D_0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} > t_{v,\alpha/2}$ of $\frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} < -t_{v,\alpha/2}$

Eerder is besproken dat ook de population proportions getest kunnen worden. Hierbij zijn echter wel grote sample sizes nodig. Ook is eerder de formule voor de proportions besproken: $Z = \frac{(\hat{p}_x-\hat{p}_y)-(P_x-P_y)}{\sqrt{\frac{P_x(1-P_x)}{n_x}+\frac{P_y(1-P_y)}{n_y}}}$

Om aan te tonen dat population proportions P_x en P_y aan elkaar gelijk zijn, worden zowel P_x als P_Y weergegeven met P_0. De estimator van P_3 is $\hat{p}_0$ en kun je gebruiken als P_0 onbekend is. Om $\hat{p}_0$ te berekenen, zijn de sample proportions van x en y nodig:

$$\hat{p}_3 = \frac{n_x\hat{p}_x+n_y\hat{p}_y}{n_x+n_y}$$

Het testen van de null hypothese en bij de sample size van '$nP_0(1-P_0) > 5$' zijn de volgende formules van belang, bij een gegeven significance level van α:

- $H_0 : P_x - P_y = 0$ of $H_0 : P_x - P_y \leq 0$
  $H_1 : P_x - P_y > 0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\hat{p}_x-\hat{p}_y)}{\sqrt{\frac{p_0(1-p_0)}{n_x}+\frac{p_0(1-p_0)}{n_y}}} > z_\alpha$
- $H_0 : P_x - P_y = 0$ of $H_0 : P_x - P_y \geq 0$
  $H_1 : P_x - P_y < 0$

  Hieruit volgt de decision rule: Reject H_0 if $\frac{(\hat{p}_x-\hat{p}_y)}{\sqrt{\frac{p_0(1-p_0)}{n_x}+\frac{p_0(1-p_0)}{n_y}}} < -z_\alpha$
- En het twee-zijdig alternatief
  $H_0 : P_x - P_y = 0$
  $H_1 : P_x - P_y \neq 0$

Hieruit volgt de decision rule: Reject H_0 if $\frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\bar{p}_0(1-\bar{p}_0)}{n_x} + \frac{\bar{p}_0(1-\bar{p}_0)}{n_y}}} > z_{\alpha/2}$

of $\frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\bar{p}_0(1-\bar{p}_0)}{n_x} + \frac{\bar{p}_0(1-\bar{p}_0)}{n_y}}} < -z_{\alpha/2}$

Het vergelijken van de variances zijn erg belangrijk voor de regressie analyse. Regressie analyse komt in de volgende hoofdstukken aan bod. Om aan te kunnen tonen dat twee variances gelijk aan elkaar zijn, wordt er een test uitgevoerd, de F probability distribution test. $F = \frac{\frac{s_x^2}{\sigma_x^2}}{\frac{s_y^2}{\sigma_y^2}}$

De degrees of freedom voor de teller is $n_x - 1 = v_1$ en voor de noemer $n_y - 1 = v_2$. In formule vorm zal F er als volgt uit zien: $F_{v1,v2,\alpha}$, waarbij α het significantie niveau is. In tabel 9 van het boek staan de cuttoff points van de F distributie. Elke α heeft zijn eigen tabel. Bij gelijke population variances verandert F, $F = \frac{s_x^2}{s_y^2}$. De nieuwe F-formule gebruik je bij het uitvoeren van de test. Een test uitvoeren is echter niet nodig wanneer de sample variances van x die van y zeer overschrijdt of andersom. Om te testen of de variances van twee populaties gelijk aan elkaar zijn, kan gebruik gemaakt worden van twee formules:

- $H_0 : \sigma_x^2 = \sigma_y^2$ of $H_0 : \sigma_x^2 \leq \sigma_y^2$
  $H_1 : \sigma_x^2 > \sigma_y^2$

  Hieruit volgt de decision rule: Reject H_0 if $F = \frac{s_x^2}{s_y^2} > F_{v1,v2,\alpha}$
- $H_0 : \sigma_x^2 = \sigma_y^2$
  $H_1 : \sigma_x^2 \neq \sigma_y^2$

  Hieruit volgt de decision rule: Reject H_0 if $F = \frac{s_x^2}{s_y^2} > F_{v1,v2,\alpha/2}$

11. Regressie analyse met twee variabelen

Eerder is besproken hoe de correlatie coëfficiënt de relatie tussen twee variabelen weergeeft. Deze lineaire relatie wordt weergegeven door de functie: $Y = \beta_0 + \beta_1 X$. Hierbij is Y de endogene variabele, X de exogene variabele, β_0 de Y-intercept en β_1 de helling van de lijn. De helling van de grafiek is voor bedrijven belangrijk om te weten, ze kunnen door een verandering in X bepalen wat het met Y gaat doen. β_0 is een schatting van het gemiddelde niveau van de output van alle niveaus van de input variabele. De lineaire relatie wordt ook wel de least squares regression genoemd. Met regressie probeer je de beste waarden voor β_0 en β_1 te vinden. De least squares regression line, bij een sample, wordt weergegeven door:
$\hat{y} = b_0 + b_1 x$. waarbij b_1 de helling en b_0 de y-intercept
$b_1 = \frac{Cov(x,y)}{s_x^2}$
$b_1 = \bar{y} - b_1 \bar{x}$

In de economie heb je geleerd dat Y de hoeveelheid verkochte goederen zijn en X het inkomen. Echter zijn er in de realiteit nog meer factoren die van invloed zijn op Y. Deze externe factoren worden aangeduid met ε (èta). De least squares regression ziet er dan als volgt uit: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. Wanneer je de regressie lijn tekent, zullen veel punten niet op de lineaire lijn van $Y = \beta_0 + \beta_1 X$ liggen. Deze rechte lijn is de geschatte waarde van de Y waarde en is in formule vorm als volgt: $E(Y|X=x) = \beta_0 + \beta_1 X$. Door ε als error term te nemen, zal de verwachte lineaire regressie lijn ontstaan. De geschatte y waarde is dan: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. De error term geldt alleen als het gemiddelde 0, $E[\varepsilon_i] = 0$, is en de variance hetzelfde,
$E[\varepsilon_i^2] = \sigma^2$. Ook mogen verschillende error terms niet in verband staan met elkaar, dus $E[\varepsilon_i \varepsilon_j] = 0$ en $i \neq j$. Het verschil tussen het daadwerkelijke punt (x_i, y_i) met de regressie lijn in een sample heet ook wel de residual en wordt aangeduid met e_i. Let wel op de e_i niet hetzelfde is als ε. In de samples zitten ook fouten en deze omvat e_i wel en ε niet. De formule van de geschatte regressie lijn in een sample is: $\hat{y}_i = b_0 + b_1 x_i$ en $e_i = y_i - \hat{y}_i$. Het is mogelijk dat bepaalde punten van de grafiek zo'n grote invloed op de regressielijn hebben dat het de hele lijn doet verschuiven. Het is daarom van belang om zo veel mogelijk punten te gebruiken om de regressielijn te plotten.

Om de juiste regressie lijn te vinden, is het van belang om alle deviations, e_i, bij elkaar op te tellen. Echter kunnen deze deviations ook negatief zijn. Door ze te kwadrateren en te sommeren, ontstaat de error sum of squares (SSE).
$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$.

Vervolgens kun je b_1 anders definiëren:
$b_1 = \frac{Cov(x,y)}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} y_i = r_{xy}\frac{s_y}{s_x}$, waarbij r_{xy} de sample correlatie is.
De constante estimator b_0 kan omgeschreven worden naar:
$b_1 = \bar{y} - b_1 \bar{x}$. wat uiteindelijk leidt naar de functie van de geschatte waarde van y:
$\hat{y}_i = \bar{y} + b_1 (x_i - \bar{x})$

Nu de ontwikkeling van de meting die aangeeft hoe effectief the variabele x het gedrag van y bepaald, kunnen we de capaciteit van x meten die bepaald y bepaald. Door middel

van de ANOVA, de analyse van de variance, kan de totale variatie van Y verdeeld worden in een component dat uitgelegd kan worden en een error component:
$y_i - \bar{y} = (\hat{y}_i - \bar{y}_i) + (y_i - \hat{y}_i)$

Vervolgens wordt de functie gekwadrateerd om negatieve getallen positief te maken en gesommeerd:
$\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y}_i)^2 + \Sigma(y_i - \hat{y}_i)^2$

SST is de totale som van de kwadraten en is onder te verdelen in de regressie som van alle kwadraten, SSR, en de error som van alle kwadraten, SSE:
SST = SSR + SSE
SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 = b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
SSE = $\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

Bij gegeven waardes van de Y staat SST vast. Als SSR groter wordt en SSE kleiner, komt de regressie vergelijking dichterbij de gevonden data. Door SSR te delen door SST ontstaat de coefficient of determination, R^2.
$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
R^2 kan een waarde aannemen tussen de 0 en 1. Hoe hoger de waarde, hoe beter de regressie. Let wel op dat als SSE klein is en/of SST groot dit niet opgaat. Door de gevonden R^2 waarde te vermenigvuldigen met 100%, weet je het percentage uitgelegde variatie.

Voor simpele regressie is de coefficient of determination gelijk aan de correlatie, dus $R^2 = r_{xy.}$

De variance van de populatie kan nu herschreven worden:
$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$, vergeet niet n-2 te doen in plaats van n-1, omdat het om twee parameters gaat, b_0 en b_1.

Nu σ^2, b_0 en b_1 behandeld zijn, kunnen er testen mee uitgevoerd worden met behulp van de de Student's t distributie:

- $H_0 : \beta_1 = \beta_1^*$ of $H_0 : \beta_1 \leq \beta_1^*$
  $H_1 : \beta_1 > \beta_1^*$
  Hieruit volgt de decision rule: Reject H_0 if $\frac{b_1 - \beta_1^*}{s_{b_1}} > t_{n-2,\alpha}$
- $H_0 : \beta_1 = \beta_1^*$ of $H_0 : \beta_1 \geq \beta_1^*$
  $H_1 : \beta_1 < \beta_1^*$
  Hieruit volgt de decision rule: Reject H_0 if $\frac{b_1 - \beta_1^*}{s_{b_1}} < -t_{n-2,\alpha}$
- $H_0 : \beta_1 = \beta_1^*$
  $H_1 : \beta_1 \neq \beta_1^*$
  Hieruit volgt de decision rule: Reject H_0 if $\frac{b_1 - \beta_1^*}{s_{b_1}} > t_{n-2,\alpha/2}$ of $\frac{b_1 - \beta_1^*}{s_{b_1}} < -t_{n-2,\alpha/2}$

Het confidence interval wordt bepaald door de error:
$b_1 - (t_{n-2,\alpha/2})s_{b1} < \beta_1 < b_1 + (t_{n-2,\alpha/2})s_{b1}$, waarbij α de significance level is en n-2 de degrees of freedom.

De hypothese kan ook getest worden aan de hand van de F distributie. Door de null hypothese gelijk te stellen aan 0 en aan te nemen dat deze waar is, kan de mean square for regression opgesteld worden:

MSR = SSR/1 = SSR

SSR deel je door 1, omdat de helling van de lijn enkel is. Vervolgens is op te merken dat MSR een schatting van de variance is. Ook de mean square of error kan opgesteld worden:

MSE = SSE/n-2 = s_{ε}^2

Door MSR te delen door de MSE kan de F ratio gemaakt worden, met significance level α, en de degrees of freedom van 1 en n-2:

F = MSR/MSE = SSR/s_{ε}^2
$F_{\alpha,1,n-2} = t^2_{\alpha/2,n-2}$

Het testen van de F ziet er als volgt uit:

- $H_0 : \beta_1 = 0$
  $H_1 : \beta_1 \neq 0$
  Hieruit volgt de decision rule: Reject H_0 if $F \geq F_{\alpha,1,n-2}$ of $F \geq t^2_{b1}$

Regressie modellen gebruik je om waardes mee te voorspellen. De normale regressie lijn is $y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$, maar bij voorspellingen verandert n in n+1, dus:
$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$, wat als verwachting de volgede formule geeft:
$E[y_{n+1} | x_{n+1}] = = \beta_0 + \beta_1 x_{n+1}$

Er zijn twee opties mogelijk bij de verwachte waarde:

- Het voorspellen van de waarde voor een enkele observatie, y_{n+1}
- Het schatten van de conditionele verwachte waarde, $E[y_{n+1} | x_{n+1}]$, wanneer x_{n+1} vast staat

Bij het voorspellen van de waarde voor een enkele observatie is het bereik groter dan die van de conditionele verwachte waarde.

Het confidence interval voor de enkele observatie wordt weergegeven door:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} s_e$$

Het confidence interval voor de conditionele verwachte waarde wordt weergegeven door:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} s_e$$

In beide gevallen geldt: $\bar{x} = \sum_{i-1}^{n} x_i / n$ en $\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$

Hoe groter het confidence interval is, hoe groter de onzekerheid over de punten is. Tevens zijn uit de formules een aantal punten belangrijke punten te halen:

- Hoe groter n, hoe nauwkeuriger het voorspelde interval en de confidence interval
- Hoe groter s_e^2, hoe groter het voorspelde interval en de confidence interval
- $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is meervoud van alle sample variances van de observatie van de onafhankelijke variabele. Hoe groter de variance, hoe groter het bereik, hoe nauwkeuriger het voorspelde interval en de confidence interval
- Hoe groter $(x_{n+1} - \bar{x})^2$, hoe groter het voorspelde interval en confidence interval

De correlatie coëfficiënt geeft de relatie weer tussen twee variabelen. De correlatie tussen de twee variabelen is ook te testen. Allereerst moet je zeker weten dat er een lineair verband is tussen x en y. Daarna kan de null hypothese opgesteld worden met de population correlation, ρ_{xy}:

- $H_0 : \rho = 0$
  $H_1 : \rho > 0$
  Hieruit volgt de decision rule: Reject H_0 if $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{\alpha,n-2}$
- $H_0 : \rho = 0$
  $H_1 : \rho < 0$
  Hieruit volgt de decision rule: Reject H_0 if $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{\alpha,n-2}$
- $H_0 : \rho = 0$
  $H_1 : \rho \neq 0$
  Hieruit volgt de decision rule: Reject H_0 if $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \geq t_{\alpha,n-2}$ of $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{\alpha,n-2}$

Bij het testen van de null hypothesis kan de rule of thumb gebruikt worden: $|r| > \frac{2}{\sqrt{n}}$

In de financiële wereld zijn er metingen ontwikkeld en analyses gemaakt om investeerder te helpen en het financiële risico van een investering weer te geven. Deze risico's worden in een portfolio genoteerd en hier geldt ook dat hoe meer portfolio's er zijn gebruikt, hoe nauwkeuriger de data. Ook de regressielijn kan getekend worden, vaak is de helling hiervan positief. Als het bedrijf en de markt precies gelijk lopen is het beta coëfficiënt gelijk aan 1, gaat het bedrijf sneller dan de markt, dan is het beta coëfficiënt groter dan 1.

Het model van Capital Asset Pricing geeft de teruggave van een investering aan:
(Required return on Investment) = (risk free rate) + [(beta for investment) x ((market return) – (risk free rate))]. Hoe groter beta is, hoe groter return on investment is.

X is niet altijd een kleine waarde, het komt vaak voor dat deze een extreme waarde is. Om achter de verwachte waarde van y te komen bij deze extreme x waarde, wordt de leverage, h_i, gebruikt:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

De leverage vergroot de standard deviation van de verwachte waarde van het punt. Als het punt van x met $h_i > 3\ p/n$ is, spreken we van een extreme waarde. P is hierin het aantal voorspelling inclusief de constante zelf.

De y waarde van deze extreme punten, wordt bepaald door het de standaard residual:

$$e_{is} = \frac{e_i}{s_e\sqrt{1-h_i}}$$

12. Veelvoudige variabele regressie analyse

In het lineaire model $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$ zijn alle b_j constante lineaire coëfficiënten van de onafhankelijke variabele X_j die het conditionele effect van elke onafhankelijke variabele op de bepaling van de afhankelijke variabele, Y, in de populatie indiceert. Anders gezegd kunnen we stellen dat de coëfficiënten b_j onafhankelijke parameters zijn in het *lineair regression model* (lineaire regressie model). De strategie voor een modelspecificatie zal worden beïnvloed door de doelstellingen van het model. Een objectief is een voorspelling van een afhankelijke of resulterende variabele. Applicaties omvatten voorspellende sales, output, totale consumptie, totale investeringen en veel andere economische voorkomende criteria. Een tweede objectief schat het marginale effect van elke onafhankelijke variabele. Dit model heeft twee belangrijke resultaten. Ten eerste is er een geschatte lineaire vergelijking dat een afhankelijke variabele, Y, voorspelt als een functie van K geobserveerde onafhankelijke variabelen, X_j, waarbij j = 1, 2, 3, …, K: $\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + ... + b_Kx_{Ki}$ waarbij i = 1, 2, 3, …, n observaties. Ten tweede is de marginale verandering in de afhankelijke variabele, Y, die gerelateerd is aan de verandering in de onafhankelijke variabelen, geschat door de coëfficiënten b_j. In het multiple regression zijn deze coëfficiënten afhankelijk van andere variabelen in het model, De coëfficiënt b_j indiceert de verandering in Y, gegeven een unit verandering in X_j, terwijl er gelijk een correctie is voor het gelijktijdige effect van de andere onafhankelijke variabelen. In sommige problemen zijn beide resultaten even belangrijk, maar over het algemeen zal een van de twee overheersen. Een marginale verandering is belangrijker om te schatten omdat de onafhankelijke variabelen niet alleen gerelateerd zijn aan de afhankelijke variabelen maar ook aan elkaar. Als twee of meer onafhankelijke variabelen veranderen in een direct lineair verband met elkaar, dan is het individuele effect van elke onafhankelijke variabele op de afhankelijke variabele moeilijk te bepalen.

Wanneer we gebruik maken van *multiple regression*, dan construeren we een model om de veranderlijkheid in een afhankelijke variabele te kunnen uitleggen. Wanneer we dit doen willen de gelijktijdige en individuele invloeden van verschillende onafhankelijke variabelen hier ook in verwerken. Het multiple regression model definieert de relatie tussen een afhankelijke / endogene variabele, Y, en een set van onafhankelijke /exogene variabelen, X_j, waarbij j = 1, 2, 3, …, K getallen. De termen x_{ji} zijn vaste getallen; Y is een random variabele die gedefinieerd is voor elke observatie, i, waarbij i = 1, 2, 3, …, n getallen; en n is het aantal observaties. Het model definiëren we als volgt: $y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + ... + b_Kx_{Ki} + e_i$ waar de coëfficiënten b_j constanten zijn en de gevallen van e_i zijn random variabelen met een gemiddelde 0 en met eenzelfde variantie s^2.

Het population mutiple regression model $y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + ... + b_Kx_{Ki} + e_i$ heeft n aantal sets van observaties beschikbaar. Vervolgens kunnen we deze *assumpties* maken. (1) De termen x_{ji} zijn vaste getallen, *of* ze zijn verwezenlijkingen van random variabelen, X_j, die onafhankelijk zijn van de error termen e_i. In het laatste geval wordt de

gevolgtrekking conditioneel uitgevoerd op de geobserveerde waarden van de x_{ji}'s. (2) De verwachte waarde van de random variabele Y is een lineaire functie van de onafhankelijke X_j variabelen. (3) De error termen zijn normaal verdeelde random variabelen met een gemiddelde 0 en eenzelfde variantie s^2.

Deze variantie noemen we de uniform / homoscedasticity variance. De formule voor het gemiddelde is $E[e_i]=0$ en de formule voor de variantie is $E[e^2]=s_i^2$ voor i = 1, 2, 3, …, n. (4) De random error termen e_i zijn niet gecorreleerd met elkaar. Hierom geldt dat $E[e_i e_j]=0$ voor i = j. (5) Het is niet mogelijk om een set van niet-nul getallen, $c_0, c_1, c_2, \ldots, c_K$, te verkrijgen zodanig dat $c_0+c_1x_{1i}+c_2x_{2i}+c_3x_{3i}+\ldots+c_Kx_{Ki}=0$. Dit geeft aan dat er geen lineaire relatie tussen de X_j variabelen is.

De eerste vier assumpties zijn in principe hetzelfde als de assumpties voor *simple regression*. De error terms in assumptie drie worden aangenomen normaal gedistribueerd te zijn voor statistische gevolgtrekkingen. Later in het hoofdstuk zal duidelijk worden dat net als met simple regression door middel van de *central limit theorem* (CLT) we deze assumptie kunnen versimpelen mits de sample grootte maar groot genoeg is. Assumptie vijf sluit bepaalde zaken uit in welke er lineaire relaties tussen de voorspellende variabelen zijn.

Voor *least squares estimation* (kleinste kwadraten schatting) en de *sample multiple regression* (sample meervoudige regressie) beginnen we met een sample van *n* observaties, $x_{1i}, x_{2i}, x_{3i}, \ldots, x_{Ki}$, en met Y_i waarbij i = 1, 2, 3, …, n, gemeten voor een proces waarvan de population multiple regression model $y_i=b_0+b_1x_{1i}+b_2x_{2i}+b_3x_{3i}+\ldots+b_Kx_{Ki}+e_i$ is. De least squares estimates van de coëfficiënten $b_1, b_2, b_3, \ldots, b_K$ zijn de waarden $b_0, b_1, b_2, b_3, \ldots, b_k$ voor welk de sum of the squared deviations (som van de gekwadrateerde deviatie)

$SSE=\sum_{i=1}^{n}\left(y_i-b_0-b_1x_{1i}-b_2x_{2i}-b_3x_{3i}-\ldots b_Kx_{Ki}\right)^2$ een minimum is. De resulterende vergelijking $\hat{y}_i=b_0+b_1x_{1i}+b_2x_{2i}+b_3x_{3i}+\ldots+b_Kx_{Ki}$ is de sample multiple regression van Y op $X_1, X_2, X_3, \ldots, X_K$.

We kijken hoe deze resulterende vergelijking ontstaat. We gaan even uit van een regression model met slechts twee voorspellende variabelen $\hat{y}_i=b_0+b_1x_{1i}+b_2x_{2i}$. De *coëfficiënt schatters* kunnen worden gevonden door middel van de volgende formule:

$b_1=\frac{s_y(r_{x_1y}-r_{x_1x_2}r_{x_2y})}{s_{x_1}(1-r_{x_1x_2}^2)}$, $b_2=\frac{s_y(r_{x_2y}-r_{x_1x_2}r_{x_1y})}{s_{x_2}(1-r_{x_1x_2}^2)}$ en $b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2$, waarbij r_{x_1y} de sample correlatie tussen X_1 en Y is, waarbij r_{x_2y} de sample correlatie tussen X_2 en Y is, waarbij $r_{x_1x_2}$ de sample correlatie tussen X_1 en X_2 is, waarbij s_{x_1} de sample standaard deviatie voor X_1 is en waarbij s_{x_2} de sample standaard deviatie voor X_2 is. In deze formules zien we dat de richtingscoëfficiënt schatter, b_1, niet alleen afhangt van de correlatie tussen Y en X_1, maar ook wordt beïnvloed door de correlatie tussen X_1 en X_2 en ook nog de correlatie tussen X_2 en Y. Als de correlatie tussen X_1 en X_2 gelijk is aan nul, dan zijn de b_1 en b_2 gelijk

aan de coëfficiënt voor simple regression. Als de correlatie tussen X_1 en X_2 gelijk is aan één, dan zijn b_1 en b_2 onbepaald, dit komt slechts voort uit een slechte modelspecificatie en is in strijd met assumptie vijf in het meervoudige regressie model.

De *estimated regression model from the sample* is $y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + ... + b_Kx_{Ki} + e_i$, een verkortte versie hiervoor is $y_i = \hat{y}_i + e_i$ waarbij $\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + ... + b_Kx_{Ki}$ de voorspelde waarde van de afhankelijke variabele is en resterende e_i het verschil tussen de geobserveerde en de voorspelde waarden is. De *model variability* kan worden opgedeeld in twee componenten: SST = SSR + SSE. De *total sum of squares* (totale som van kwadraten) $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ bestaat uit twee delen. (1) De *error sum of squares* (de error som van kwadraten) $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}e_i^2$. (2) De *regression sum of squares* (regressie som van kwadraten) $SSR = \sum_{i=1}^{n}(y_i - \bar{y})^2$. De ontbinding kan als volgt geïnterpreteerd worden: Total sample variability = explained variability + unexplained variability. De *coefficient of determination*, R^2, van de *fitted regression* is gedefinieerd als de proportie van de totale sample variability uitgelegd door de regressie $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$, hierbij ligt de R^2 tussen de 0 en 1 oftewel $0 \le R^2 \le 1$.

De *standard error of the estimate*, s_e – met gegeven het populatie multiple regression model $y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + ... + b_Kx_{Ki} + e_i$ met de standaard regression assumpties, waarbij s^2 de algemene variantie van de error term e_i – is de wortel van de volgende formule $s_e^2 = \frac{\sum_{i=1}^{n}e_i^2}{n-K-1} = \frac{SSE}{n-K-1}$ waarbij K het aantal van onafhankelijke variabelen in het regression model is. Oftewel de standard error of the estimate is gelijk aan $s_e = \sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n-K-1}} = \sqrt{\frac{SSE}{n-K-1}}$.

De *adjusted coefficient of determination*, $\bar{R}^2$, heeft de volgende formule $\bar{R}^2 = 1 - \frac{\frac{SSE}{(n-K-1)}}{\frac{SST}{(n-1)}}$.

We gebruiken deze coëfficiënt om te corrigeren voor het feit dat de irrelevante onafhankelijke variabelen zullen resulteren in een kleine reductie in de error sum of squares. De aangepaste $\bar{R}^2$ geeft daarom een betere vergelijking tussen meervoudige regressie modellen met verschillende aantallen onafhankelijke variabelen.

De *coefficient of multiple correlation* is de correlatie tussen de voorspelde waarde en de

geobserveerde waarde van de afhankelijke variabele: $R = r(\hat{y}, y) = \sqrt{R^2}$; het is gelijk aan de wortel van de multiple coefficient of determination. We gebruiken R als een andere maatregel om de sterkte van de relatie tussen de afhankelijke variabele en de onafhankelijke variabele te meten. Het is vergelijkbaar met de correlatie tussen X en Y in de simple regression.
De *coefficient variance estimator* (coëfficiënt variantie schatter) is voor b_1 uit de formule $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$

als volgt te berekenen: $s_{b_1}^2 = \frac{s_e^2}{(n-1)s_{x_1}^2(1-r_{x_1x_2}^2)}$ en ook is deze coefficient variance estimator te berekenen voor b_2: $s_{b_2}^2 = \frac{s_e^2}{(n-1)s_{x_2}^2(1-r_{x_1x_2}^2)}$, voor b_2 is dit hetzelfde principe als voor b_1.

De wortel van $s_{b_1}^2$ en $s_{b_2}^2$ wordt de *coefficient standard errors* genoemd, s_{b_1} en s_{b_2}.
De variance of the coefficient estimators stijgt direct gelijk aan de afstand tussen de punten en de lijn, gemeten door de s_e^2 de *estimated error variance*. Daarbij, een bredere spreiding van de onafhankelijke variabele waarden – gemeten door $s_{x_1}^2$ of door $s_{x_2}^2$ - verlaagt de coëfficiënt variantie. De variantie van de coëfficiënt schatter stijgt als de correlatie tussen de twee onafhankelijke variabelen in het model stijgt. Als de correlatie tussen twee onafhankelijke variabelen stijgt, wordt het moeilijker om het effect van de individuele variabelen te scheiden om de afhankelijke variabelen te voorspellen. Wanneer het aantal onafhankelijke variabelen in een model stijgt, blijft de invloed van deze variabelen op de coefficient variance estimator belangrijk, maar de algebraïsche structuur wordt zeer complex en wordt daarom in het boek niet uitgelegd. Het correlatie effect leidt tot het resultaat dat de coefficient estimators conditioneel zijn op de andere onafhankelijke variabelen in het model. Bedenk hierbij dat de eigenlijke coefficient estimators ook conditioneel zijn op de andere onafhankelijke variabelen in het model, dit ook door het effect van de correlaties tussen de onafhankelijke variabelen.

Als we uitgaan van het populatie regressie model
$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + ... + b_K x_{Ki} + e_i$ met b_0, b_1, b_2, b_3, …, b_k de kleinste kwadratische schatters van de populatie parameters en $s_{b_0}, s_{b_1}, s_{b_2}, s_{b_3}, ..., s_{b_K}$ de geschatte standaard deviaties van de kleinste kwadratische schatters. Dan is, uitgaand van de standaard regressie assumpties en een normaal gedistribueerde error termen, e_j, $t_{bj} = \frac{b_j - b_j}{s_{b_j}}$ waarbij j = 1, 2, 3, …, K gedistribueerd als een Student's t distributie met (n – K - 1) mate van vrijheid (degrees of freedom).

De 100(1 - a)% tweezijdige *confidence intervallen* voor de regressie coëfficiënten, b_j, kunnen afgeleid worden van de vergelijking $b_j - t_{n-K-1,\frac{a}{2}} \cdot s_{b_j} < b_j < b_j + t_{n-K-1,\frac{a}{2}} \cdot s_{b_j}$ waarbij $t_{n-K-1,\frac{a}{2}}$ het aantal is waarbij $P(t_{n-K-1} > t_{n-K-1,\frac{a}{2}}) = \frac{a}{2}$ en waarbij de random variabele t_{n-K-1} een Student's t distributie volgt met (n – K – 1) mate van vrijheid. Hierbij geldt dat de

populatie regressie errors, e_j, normaal gedistribueerd zijn en de standaard regressie assumpties van toepassing zijn.

Voor regressie coëfficiënten kunnen ook *hypotheses* getest worden. Hiervoor gaan we ervan uit dat de regressie errors, e_j, normaal gedistribueerd zijn en de standaard regressie assumpties van toepassing zijn. Vervolgens kunnen we hypotheses opstellen met een significantie niveau a. De nul hypothese, H_0, testen we tegen het alternatief, H_1. Hiervoor hebben we drie situaties die hieronder worden beschreven.

(1) $H_0: b_j = b^*$ Eenzijdige test

$H_1: b_j \leq b^*$ Decision Rule: Verwerp H_0 als $\frac{b_j - b^*}{s_{b_j}} > t_{n-K-1,a}$

(2) $H_0: b_j = b^*$ Eenzijdige test

$H_1: b_j \geq b^*$ Decision Rule: Verwerp H_0 als $\frac{b_j - b^*}{s_{b_j}} < -t_{n-K-1,a}$

(3) $H0: b_j = b^*$ Tweezijdige test

$H_1: b_j \neq b^*$ Decision Rule: Verwerp H_0 als $\frac{b_j - b^*}{s_{b_j}} > t_{n-K-1,\frac{a}{2}}$ of verwerp H_0

als $\frac{b_j - b^*}{s_{b_j}} < -t_{n-K-1,\frac{a}{2}}$

Deze testen kunnen we interpreteren. Analisten beweren dat we de variabele niet zouden moeten opnemen in het model wanneer we niet de conditionele hypothese dat de coëfficiënt 0 is kunnen verwerpen, De Student's t statistiek voor deze test is typisch samengesteld in meeste regressieprogramma's en is naast de coëfficiënt variantie schatter afgedrukt, daarbij is de p-waarde voor de hypothese test typisch opgenomen. Bij gebruik van de Student's t statistiek of de p-waarde, kunnen we meteen concluderen of een bepaalde voorspellende variabele conditioneel significant is of niet, waarbij de andere variabelen in het regressie model gegeven zijn. Er zijn duidelijk andere procedures om te bepalen of een onafhankelijke variabele wel of niet moet worden opgenomen in het regressie model. We hebben gezien dat de voorgaande selectieprocedure het type II error negeert – de populatie coëfficiënt is niet gelijk aan 0, maar we falen in het verwerpen als de nul hypothese gelijk is aan 0. Dit is een specifiek probleem in een model gebaseerd op een economische of een andere theorie dat met zorg is gespecificeerd om bepaalde onafhankelijke variabelen in het model te insluiten. Dan, omdat de error, e, groot is of omdat de correlatie tussen de onafhankelijke variabelen groot is, of beide gevallen, kunnen we niet de hypothese verwerpen dat de coëfficiënt gelijk is aan 0. In dit geval zullen veel analisten de onafhankelijke variabele in het model opnemen, omdat het originele model specificatie gebaseerd op de economische theorie of ervaring domineert. Dit is een moeilijk probleem en dit vereist goede besluitvorming gebaseerd op statistische resultaten en theorie die de onderliggende relatie die gemodelleerd is in ogenschouw neemt.

Het is belangrijk om te benadrukken dat de hypothese testen gebaseerd zijn op de specifieke set variabelen die in het regressie model zijn opgenomen. Met extra variabelen in het model zal de geschatte coëfficiënt en hun geschatte standaard deviaties anders zijn

en daarmee zal ook de Student's t statistiek anders zijn.

Het is mogelijk om tests te doen op alle coëfficiënten. Hiervoor gaan we uit van het model $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + ... + b_K x_{Ki} + e_i$. We beginnen bij het opstellen van de nul hypothese dat alle coëfficiënten gelijk zijn aan 0. H_0: $b_1 = b_2 = b_3 = ... = b_K = 0$. Deze hypothese leidt tot de conclusie dat de voorspellende variabelen in het regressie model statistisch significant is en daarmee geen nuttige informatie verstrekt. Als dit wel zou voorkomen, dan zouden we opnieuw moeten kijken naar de model specificatie en deze hierop aanpassen door een nieuwe set van voorspellende variabelen te creëren. Om de hypothese H_0: $b_1 = b_2 = b_3 = ... = b_K = 0$ te kunnen testen kunnen we gebruik maken van SST = SSR + SSE, een formule die eerder in dit hoofdstuk is behandeld. Hierbij was SSR de hoeveelheid variabiliteit verklaard door de regressie en SSE is de hoeveelheid onverklaarde variabiliteit. Bedenk hierbij dat de variantie van het regressie model kan worden berekend door het volgende $s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-K-1} = \frac{SSE}{n-K-1}$.

Als de nul hypothese klopt, dit betekent dat alle coëfficiënten gelijk zijn aan nul, dan is de *mean square regression* (gemiddelde kwadratische regressie) gelijk aan $MSR = \frac{SSR}{K}$ wat ook een maatstaaf is voor de error met K degrees of freedom. Een resulterende vergelijking is de ratio $F = \frac{SSR/K}{SSE/(n-K-1)} = \frac{MSR}{s_e^2}$; een F distributie met K degrees of freedom voor de teller en (n – K – 1) degrees of freedom voor de noemer. Als de nul hypothese waar is, dan verstrekken zowel de teller en de noemer schattingen van de populatie variantie. Als de ratio van de onafhankelijke sample varianties van populaties met een gelijke populatie variantie is, dan volgt een F distributie onder de voorwaarde dat de populaties een normale verdeling hebben. De samengestelde waarde van F wordt vergeleken met de kritieke waarde van F uit tabel 9 in de appendix in het boek op significatie niveau a. Als de samengestelde waarde de kritieke waarde die in de tabel gevonden is, overschrijdt dan kunnen we de nul hypothese verwerpen en concluderen dat ten minste één coëfficiënt *niet* gelijk is aan nul. Oftewel de nul hypothese is H_0: $b_1 = b_2 = b_3 = ... = b_K = 0$ en de alternatieve hypothese is H_1: ten minste één $b_j \neq 0$. Op significantie niveau a gebruiken we de decision rule: verwerp H_0 als $F_{K,n-K-1} = \frac{MSR}{s_e^2} > F_{K,n-K-1,a}$ waarbij $F_{K,n-K-1,a}$ de kritieke waarde van F is gevonden in tabel 9 in de appendix van het einde van het boek voor welke $P(F_{K,n-K-1} > F_{K,n-K-1,a}) = a$. De samengestelde random variabele $F_{K,n-K-1}$ volgt een F verdeling waarbij de teller K degrees of freedom heeft en de noemer (n – K – 1) degrees of freedom heeft.

Stel we hebben een gegeven regressie model waarin de onafhankelijke variabele opgedeeld zijn in *X en Z subsets*, oftewel als formule

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + ... + b_K x_{Ki} + a_1 z_{1i} + a_2 z_{2i} + a_3 z_{3i} + ... + a_r z_{ri} + e_i.$$

We willen de nul hypothese testen $H_0: a_1 = a_2 = a_3 = ... = a_r = 0$; wat betekent dat de regressie parameters in een bepaalde subset (in dit geval de subset Z) tegelijkertijd gelijk zijn aan 0, tegen het alternatief H_1 : tenminste één $a_r \neq 0$ waarbij j = 1, 2, 3, …, r. Dan stellen we de *error sum of squares* voor het gehele model samen met de error sum of squares voor het beperkte model samen. Dit als volgt (1) voer een regressie uit voor het complete model, die alle onafhankelijke variabelen bevat, en bereken de error sum of squares, SSE en (2) voer een beperkt model uit, welke de Z variabelen bevat, waarvan de coëfficiënten a's zijn en het aantal variabelen die uitgesloten zijn is r, bereken hiervan de *restricted error sum of squares* , SSE(r). Stel daarna de F statistiek samen en pas de decision rule voor het significantie niveau a samen: Verwerp H_0 als F = $\frac{(SSE(r) - SSE)/r}{s_e^2} > F_{r,n-K-1,a}$.

Wanneer r =1 in de vergelijking $\frac{(SSE(r) - SSE)/r}{s_e^2} > F_{r,n-K-1,a}$ dan kunnen we de hypothese testen dat een variabele, X_j, niet de voorspelling van een afhankelijke variabele beïnvloed, waarbij de andere onafhankelijke variabelen in het model bekend zijn. In formules hebben we de volgende hypotheses:

$H_0: b_j = 0 \mid b_l \neq 0$	$j \neq l$	l = 1, 2, 3, …, K
$H_1: b_j \neq 0 \mid b_l \neq 0$	$j \neq l$	l = 1, 2, 3, …, K

Deze test kan worden gedaan door gebruik te maken van een Student's t test. Uiteindelijk kan worden aangetoond dat de corresponderende F en t testen zullen dezelfde conclusies geven met betrekking tot de hypothese test voor een enkele variabele. Daarbij is de samengestelde t statistiek voor de coëfficiënt b_j gelijk aan de vierkantswortel van de corresponderende samengestelde F statistiek: $t_{b_j}^2 = F_{x_j}$ waarbij F_{x_j} de F statistiek is die samengesteld is uit de vergelijking $\frac{(SSE(r) - SSE)/r}{s_e^2} > F_{r,n-K-1,a}$ wanneer de variabele x_j uit het model is geëlimineerd en daardoor r = 1. De statistische distributie theorie laat ook zien dat een F random variabele met 1 degree of freedom in de teller het kwadraat van een t random variabele is met de zelfde degree of freedom als de noemer van de F random variabele. Dit betekent dat de F en de t testen altijd dezelfde conclusies zullen opleveren met betrekking tot de hypothese test voor één enkele onafhankelijke variabele in een meervoudig regressie model.

Stel we hebben het populatie regressie model $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + ... + b_K x_{Ki} + e_i$, de standaard regressie assumpties zijn van toepassing, met b_0, b_1, b_2, b_3, …, b_k als de kleinste kwadratische schatters van de model coëfficiënten, b_j, waarbij j = 1, 2, 3, …, K, gebaseerd op $x_{1i}, x_{2i}, x_{3i}, ..., x_{Ki}$ data punten waarbij i = 1, 2, 3, …, n. Dan is - gegeven dat de nieuwe observatie van een data punt gelijk is aan $x_{1,n+1}, x_{2,n+1}, x_{3,n+1}, ..., x_{K,n+1}$ - de beste *linear unbiased forecast of* $\hat{y}_{n+1}$ te bereken door middel van de formule $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + ... + b_K x_{Ki}$ waarbij *i = n + 1*. Het is zeer riskant om forecast te

verkrijgen die gebaseerd zijn op X waarden buiten het bereik van de data die wordt gebruikt om de model coëfficiënten te schatten omdat we geen databewijs hebben om het lineaire model op deze punten te ondersteunen.

Veel processen kunnen het best weergegeven worden door niet-lineaire vergelijkingen.

Het *kwadratische model* voor een aantal economische en bedrijfseconomische relaties is $Y = b_0 + b_1 X_1 + b_2 X_2^2 + e_i$. Dit kwadratisch model kan worden getransformeerd in een lineaire meervoudige regressie model door nieuwe variabelen te definiëren: $z_1 = x_1$ en $z_2 = x_1^2$. Vervolgens moet je dan ook het model specificeren: $y_i = b_0 + b_1 z_{1i} + b_2 z_{2i} + e_i$; dit is een lineaire functie. De getransformeerde kwadratische variabelen kunnen worden gecombineerd met andere variabelen in het meervoudige regressie model. We kunnen op die manier een passend meervoudig kwadratische regressie vinden door middel van getransformeerde variabelen. Het doel is om modellen te vinden die lineair zijn in andere wiskundige vormen van een variabele.

Bij het transformeren van variabelen kunnen we een lineair meervoudige regressie model schatten en de resultaten gebruiken als een niet-lineair model. De interferentie procedures voor getransformeerde kwadratische modellen zijn hetzelfde als die we eerder hebben genoemd voor lineaire modellen. Op deze manier is verwarring vermeden. De coëfficiënten moeten gecombineerd worden voor een interpretatie. Dus wanneer we een kwadratisch model hebben dan is het effect van een variabele, X, aangegeven door de coëfficiënten van zowel de lineaire als de kwadratische termen. We kunnen ook een simpele hypothese test doen om te bepalen of een kwadratisch model een betere versie is dan een lineair model. De Z_2 of de X_1^2 variabele is slechts een extra variabele waarvan de coëfficiënt getest kan worden. Hiervoor is de nul hypothese $H_0 : b_2 = 0$. Door gebruik te maken van de conditionele Student's t of F statistiek kan dit worden getest. Als een kwadratisch model beter is voor de data dan een lineair model, dan zal de coëfficiënt van de kwadratische variabele $Z_2 = X_1^2$ significant verschillen van 0. Dezelfde methode geldt voor variabelen als $Z_3 = X_1^3$ of $Z_4 = X_1^2 X_2$.

Sommige modellen worden met *exponentiele functies* uitgedrukt. Deze hebben de vorm $Y = b_0 \ X_1^{b_1} \ X_2^{b_2} + e$ en kan worden geschat door eerst het logaritme van beide kanten ten nemen zodat een vergelijking ontstaat die lineair is in de logaritmen van de variabelen: $\log(Y) = \log(b_0) + b_1 \log(X_1) + b_2 \log(X_2) + \log(e)$. Door de formule op deze manier te schrijven kunnen we de regressie van het logaritme van Y op de logaritmen van de twee X variabelen berekenen en daarmee een schatter voor de coëfficiënten b_1, b_2 direct uit de regressie analyse verkrijgen. De coëfficiënten zijn elasticiteiten waardoor economisten deze vorm van het model gebruiken wanneer ze kunnen aannemen dat de elasticiteiten constant zijn in het gehele bereik van de data. Hierbij is op te merken dat de schattingsprocedure vereist dat de random errors vermenigvuldigbaar / multiplicatief in het originele exponentiele model. Oftwel de error term, e, is uitgedrukt als een percentage stijging of daling in plaats van de opgetelde of afgetrokken hoeveelheid van een random error, zoals in voorgaande lineaire regressie modellen wel het geval was.

Een belangrijke vorm van het exponentiele model is de *Cobb-Douglas productie functie*, met de formule $Q = b_0 \ L^{b_1} \ K^{b_2}$ waarbij Q de hoeveelheid is die geproduceerd wordt, L de hoeveelheid arbeid die gebruikt wordt voor de productie en K de hoeveelheid kapitaal die gebruikt wordt voor de productie. Een speciaal geval is de som van de coëfficiënten gelijk aan 1, dan is er sprake van constant returns to scale (constante schaalvoordelen). In dat geval zijn b_1 en b_2 de procentuele bijdrage arbeid en kapitaal om productiviteit te laten stijgen. De schatting van de coëfficiënten wanneer hun som gelijk is aan 1 is één voorbeeld van beperkte schatting in regressie modellen. De vergelijking $\log(Y) = \log(b_0) + b_1 \log(X_1) + b_2 \log(X_2) + \log(e)$ is gewijzigd door de beperking $b_1 + b_2 = 1$ en daarom is de substitutie van de vorm $b_2 = 1 - b_1$ opgenomen in de vergelijking waardoor de nieuwe vergelijking als volgt wordt: $\log(Y) = \log(b_0) + b_1 \log(X_1) + (1 - b_1) \log(X_2) + \log(e)$ die uiteindelijk hervormd kan worden tot $\log\left(\frac{Y}{X_2}\right) = \log(b_0) + b_1 \log\left(\frac{X_1}{X_2}\right) + \log(e)$. Het is namelijk zo dat de b_1 coëfficiënt wordt verkregen door middel van regressie $\log\left(\frac{Y}{X_2}\right)$ op $\log\left(\frac{X_1}{X_2}\right)$. Vervolgens wordt b_2 berekend door $b_2 = 1 - b_1$.