

## De drie mooiste modellen deel 2

Medische statistiek  
Lineaire regressie

Hans Burgerhof  
h.burgerhof@home.nl

## Begrijpen we deze tabel?

240

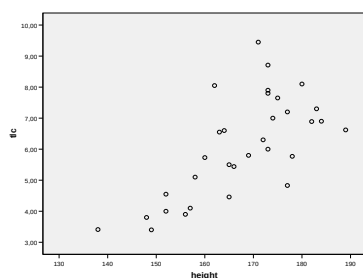
Table 4. Independent determinants of birthweight: results of multiple linear regression analysis

| Predictor variable       | Unit     | Regression coefficient ± SE <sup>a</sup> | p-Value |
|--------------------------|----------|------------------------------------------|---------|
| Gestational age          | +1 week  | 195.7 ± 4.7                              | <0.001  |
| Baby's sex               | Female   | -185.5 ± 17.1                            | <0.001  |
| Smoking during pregnancy | Yes      | -215.2 ± 18.6                            | <0.001  |
| Marital status           | Single   | -46.9 ± 21.4                             | 0.002   |
| Country of birth         | Asia     | -108.5 ± 38.2                            | 0.005   |
| Parity                   | +1 child | 121.9 ± 17.6                             | <0.001  |

Coefficient of determination ( $R^2$ ) of the model was 0.44.

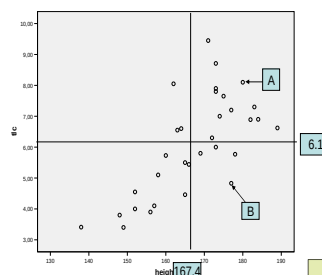
Phung et al.

## Spreidingsdiagram van tlc en lengte



Is er een lineair verband? Hoe sterk?

$$\text{covariantie}(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$



Covariantie is  
afhankelijk van de  
meeteenheid!

## Pearson's correlatiecoëfficiënt

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

- $-1 \leq r \leq 1$
- Maat voor *lineaire* samenhang

## Correlatie in SPSS

$$H_0 : r = 0$$

$$H_1 : r \neq 0$$

Correlations

|                         |                     | height (cm) | total lung capacity (l) |
|-------------------------|---------------------|-------------|-------------------------|
| height (cm)             | Pearson Correlation | 1           | .696**                  |
|                         | Sig. (2-tailed)     |             | .000                    |
|                         | N                   | 32          | 32                      |
| total lung capacity (l) | Pearson Correlation | .696**      | 1                       |
|                         | Sig. (2-tailed)     | .000        |                         |
|                         | N                   | 32          | 32                      |

\*\*. Correlation is significant at the 0.01 level (2-tailed).

## Enkelvoudige lineaire regressie

- Er is een continue responsievariabele Y  
Bijvoorbeeld: gewicht
- Er is een verklarende variabele X  
Bijvoorbeeld: leeftijd
- We zijn geïnteresseerd in de relatie tussen Y en X  
Relatie is niet symmetrisch!
- We beschikken over onafhankelijke waarnemingen
- Een lineair verband lijkt redelijk

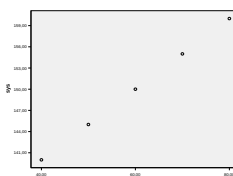
## Lineaire regressie (formule)

- We schatten de populatiere relatie

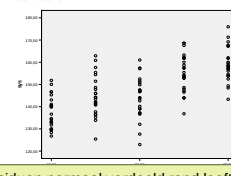
$$Y_i = b_0 + b_1 X_i + e_i$$

- We nemen aan dat  $e \sim N(0, s^2)$

Controleer!

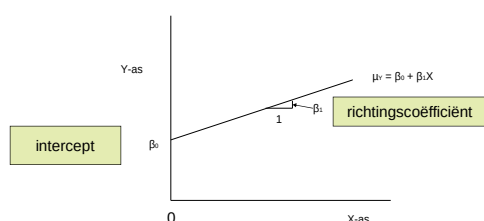


Rechtlijnig verband van de gemiddelden

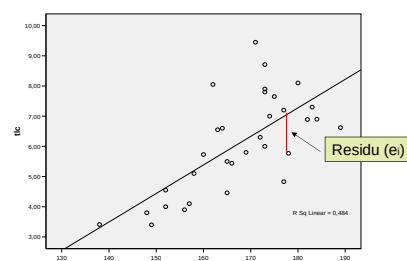


Residuen normaal verdeeld rond leeftijds-gemiddelde, met dezelfde spreiding

## Interpretatie coëfficiënten bij een enkelvoudige lineaire regressie



## SPSS trekt de "best passende lijn"



Kleinste kwadraten criterium: trek die lijn zodanig dat

$$\sum e_i^2 \text{ zo klein mogelijk is}$$

## Output SPSS van enkelvoudige lineaire regressie

| Model Summary |                   |          |                   |                            |
|---------------|-------------------|----------|-------------------|----------------------------|
| Model         | R                 | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1             | .696 <sup>a</sup> | .484     | .467              | 1.18610                    |

a. Predictors: (Constant), height (cm)

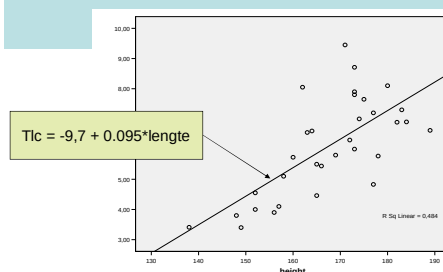
| ANOVA <sup>a</sup> |            |                |    |             |        |                   |
|--------------------|------------|----------------|----|-------------|--------|-------------------|
| Model              |            | Sum of Squares | df | Mean Square | F      | Sig.              |
| 1                  | Regression | 36.656         | 1  | 36.656      | 26.114 | .000 <sup>a</sup> |
|                    | Residual   | 42.205         | 30 | 1.407       |        |                   |
|                    | Total      | 84.264         | 31 |             |        |                   |

a. Predictors: (Constant), height (cm)

b. Dependent Variable: total lung capacity (l)

| Coefficients <sup>a</sup> |             |                             |                           |        |      |
|---------------------------|-------------|-----------------------------|---------------------------|--------|------|
| Model                     |             | Unstandardized Coefficients | Standardized Coefficients | t      | Sig. |
| 1                         | (Constant)  | -9.742                      |                           | -3.255 | .003 |
|                           | height (cm) | .095                        | .696                      | 5.303  | .000 |

a. Dependent Variable: total lung capacity (l)



| Coefficients <sup>a</sup> |             |                             |                           |        |      |
|---------------------------|-------------|-----------------------------|---------------------------|--------|------|
| Model                     |             | Unstandardized Coefficients | Standardized Coefficients | t      | Sig. |
| 1                         | (Constant)  | -9.742                      |                           | -3.255 | .003 |
|                           | height (cm) | .095                        | .696                      | 5.303  | .000 |

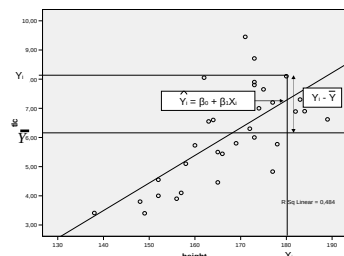
a. Dependent Variable: total lung capacity (l)

ANOVA

| Model           | Sum of Squares       | df | Mean Square | F      | Sig. |
|-----------------|----------------------|----|-------------|--------|------|
| 1               | 111.111 <sup>a</sup> | 1  | 111.111     | 15.160 | .000 |
| Total           | 111.111              | 19 |             |        |      |
| Total Corrected | 111.111              | 18 |             |        |      |

a. Predictors: (Constant), height (cm)

b. Dependent Variable: total lung capacity (l)



$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \bar{Y}_i + \bar{Y}_i - \bar{Y})^2 = \sum (Y_i - \bar{Y}_i)^2 + \sum (\bar{Y}_i - \bar{Y})^2$$

## Model summary

R<sup>2</sup> = proportie  
verklaarde variantie

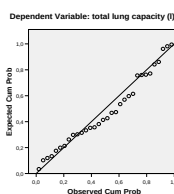
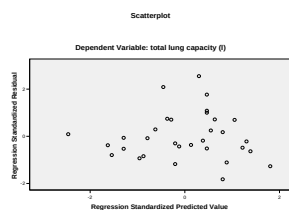
Model Summary

| Model | R                 | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1     | .696 <sup>a</sup> | .484     | .467              | 1.18610                    |

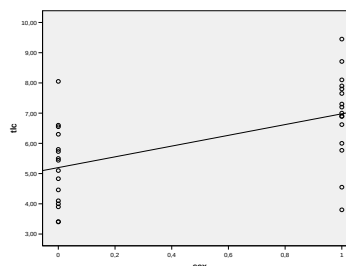
a. Predictors: (Constant), height (cm)

R = de absolute waarde van de  
correlatiecoëfficiëntSchatter voor het  
gemiddelde absolute  
residu

## Controle van de voorwaarden

Residuen normaal  
verdeeld?Homogene spreiding van de  
residuen?

## Lineaire regressie met een dichotome verklarende variabele?



## Lineaire regressie op geslacht

Coefficients<sup>a</sup>

| Model      | Unstandardized Coefficients |            | Standardized Coefficients |  | t      | Sig. |
|------------|-----------------------------|------------|---------------------------|--|--------|------|
|            | B                           | Std. Error | Beta                      |  |        |      |
| 1          |                             |            |                           |  |        |      |
| (Constant) | 5.198                       | .343       |                           |  | 15.160 | .000 |
| sex        | 1.779                       | .485       | .557                      |  | 3.669  | .001 |

a. Dependent Variable: total lung capacity (l)

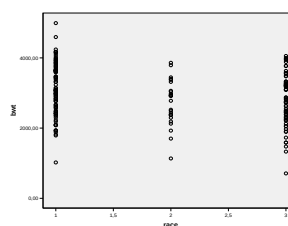
Vergeleken met de  
t-toets

Group Statistics

|                         | sex    | N  | Mean   | Std. Deviation | Std. Error Mean |
|-------------------------|--------|----|--------|----------------|-----------------|
| total lung capacity (l) | female | 16 | 5.1981 | 1.30062        | .32521          |
|                         | male   | 16 | 6.9775 | 1.43882        | .35970          |

| Independent Samples Test |                                |                                            |      |                              |        |      |          |                 |                    |                          |
|--------------------------|--------------------------------|--------------------------------------------|------|------------------------------|--------|------|----------|-----------------|--------------------|--------------------------|
|                          |                                | Levene's Test for<br>Equality of Variances |      | t-test for Equality of Means |        |      |          |                 |                    |                          |
|                          |                                |                                            |      | f                            | Sig.   | t    | df       | Sig. (2-tailed) | Mean<br>Difference | Std. Error<br>Difference |
|                          |                                |                                            |      |                              |        |      |          |                 | Lower              | Upper                    |
| total lung capacity (l)  | Equal variances<br>assumed     | .001                                       | .976 | -3.669                       | 30     | .001 | -1.77938 | .48492          | -2.76972           | -.78904                  |
|                          | Equal variances<br>not assumed |                                            |      | -3.669                       | 29.707 | .001 | -1.77938 | .48492          | -2.77013           | -.78882                  |

## X heeft meer dan 2 categorieën



## Dummy coding

| Ras | d1 | d2 |
|-----|----|----|
| 1   | 0  | 0  |
| 2   | 1  | 0  |
| 3   | 0  | 1  |

$geboortegewicht = b_0 + b_1 * d1 + b_2 * d2$

| ANOVA <sup>a</sup> |                |     |             |       |                   |
|--------------------|----------------|-----|-------------|-------|-------------------|
| Model              | Sum of Squares | df  | Mean Square | F     | Sig.              |
| 1                  | 5070608        | 2   | 2535303,816 | 4,972 | ,008 <sup>a</sup> |
| Regression         | 9E+007         | 186 |             |       |                   |
| Residual           | 1E+008         |     |             |       |                   |
| Total              |                |     |             |       |                   |

a. Predictors: (Constant), d2, d1

b. Dependent Variable: birth weight in grams

| Model |            | Unstandardized Coefficients |            | Standardized Coefficients |  | t      | Sig. |
|-------|------------|-----------------------------|------------|---------------------------|--|--------|------|
|       |            | B                           | Std. Error | Beta                      |  |        |      |
| 1     | (Constant) | 3103,740                    | 77,887     |                           |  | 47,586 | ,000 |
|       | d1         | -384,047                    | 157,874    | -.182                     |  | -2,433 | ,016 |
|       | d2         | -289,725                    | 113,678    | -.197                     |  | -2,637 | ,009 |

a. Dependent Variable: birth weight in grams

## Interpretatie $geboortegewicht = b_0 + b_1 * d1 + b_2 * d2$

- Als we de gebruikte codes invullen :
- Blank:  $geboortegewicht = 3104$
- Zwart:  $geboortegewicht = 3104 + -384 = 2720$
- Ander:  $geboortegewicht = 3104 + -300 = 2804$

Descriptives

| birth weight in grams |     |          |                |            |                                  |           |      |
|-----------------------|-----|----------|----------------|------------|----------------------------------|-----------|------|
|                       | N   | Mean     | Std. Deviation | Std. Error | 95% Confidence Interval for Mean |           |      |
| white                 | 46  | 3103,740 | 77,77474       | 74,77364   | 2964,5688                        | 3241,9102 | 1021 |
| black                 | 26  | 2719,462 | 136,60086      | 175,79637  | 2461,7773                        | 2977,6627 | 1196 |
| other                 | 67  | 2804,015 | 211,30115      | 88,12096   | 2628,0758                        | 2979,9541 | 709  |
| Total                 | 139 | 2944,656 | 279,02242      | 83,02858   | 2840,0486                        | 3049,2636 | 4990 |

ANOVA

| birth weight in grams |                |     |             |       |      |
|-----------------------|----------------|-----|-------------|-------|------|
|                       | Sum of Squares | df  | Mean Square | F     | Sig. |
| Between Groups        | 5070608        | 2   | 2535303,816 | 4,972 | ,008 |
| Within Groups         | 9E+007         | 186 |             |       |      |
| Total                 | 1E+008         | 188 |             |       |      |

Vergelijking met ANOVA

## Samenvattend voor de enkelvoudige lineaire regressie

- Y continue variabele
- X continu – relatie met correlatie
- X dichotoom – relatie t-toets
- X nominaal – relatie ANOVA
- Controleer voorwaarden!

## Multivariate statistiek

|              | Therapie A | Therapie B | totaal |
|--------------|------------|------------|--------|
| genezen      | 31         | 69         | 100    |
| niet genezen | 49         | 61         | 110    |
| totaal       | 80         | 130        | 210    |

## Multivariate statistiek

- Genezen bij A:  
 $31/80 * 100 \% = 39 \%$
- Genezen bij B:  
 $69/130 * 100 \% = 53 \%$

Wat schrijven we voortaan voor?

## Paradox van Simpson

| Licht zieken   | Therapie A | Therapie B | totaal |
|----------------|------------|------------|--------|
| genezen        | 16 (75 %)  | 63 (70 %)  | 79     |
| niet genezen   | 4          | 27         | 31     |
| totaal         | 20         | 90         | 110    |
|                |            |            |        |
| Ernstig zieken | Therapie A | Therapie B | totaal |
| genezen        | 15 (25 %)  | 6 (15 %)   | 21     |
| niet genezen   | 45         | 34         | 79     |
| totaal         | 60         | 40         | 100    |

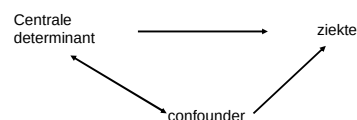
## Meervoudige lineaire regressie

- Vaak willen we weten hoe Y samenhangt met meerdere variabelen of
- Willen we het effect van  $X_1$  weten, gecorrigeerd voor  $X_2$

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + e$$

## In epidemiologische termen

Een confounder is een versturende variabele, die verantwoordelijk is voor een vertekende weergave tussen de centrale determinant en de ziekte en niet op het causale pad ligt. (Bouter e.a.)



## Regressie van tlc op lengte en geslacht

| ANOVA <sup>a</sup> |            |                |    |             |        |
|--------------------|------------|----------------|----|-------------|--------|
| Model              |            | Sum of Squares | df | Mean Square | F      |
| 1                  | Regression | 42.810         | 2  | 21.405      | 15.935 |
|                    | Residual   | 38.354         | 29 | 1.323       |        |
|                    | Total      | 81.164         | 31 |             |        |

a. Predictors: (Constant), sex, height (cm)

b. Dependent Variable: total lung capacity (l)

| Coefficients <sup>a</sup> |             |                             |            |                           |        |
|---------------------------|-------------|-----------------------------|------------|---------------------------|--------|
| Model                     |             | Unstandardized Coefficients |            | Standardized Coefficients | t      |
|                           |             | B                           | Std. Error | Beta                      |        |
| 1                         | (Constant)  | -7.034                      | 2.409      |                           | -2.962 |
|                           | height (cm) | .076                        | .021       | .560                      | 3.607  |
|                           | sex         | .772                        | .495       | .241                      | 1.556  |

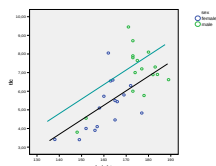
a. Dependent Variable: total lung capacity (l)

$$Tlc = -7,03 + 0,08 * \text{lengte} + 0,77 * \text{geslacht}$$

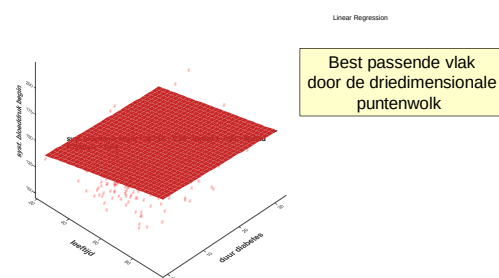
Vrouw = 0  
Man = 1

Vrouwen:  $tlc = -7,03 + 0,08 * \text{lengte}$

Mannen:  $tlc = -7,03 + 0,08 * \text{lengte} + 0,77 = -6,26 + 0,08 * \text{lengte}$



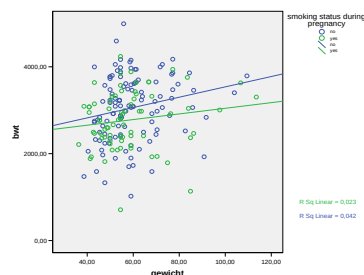
## Twee continue verklarende variabelen



## Opbouw van een regressiemodel

- Voer univariate analyses uit met al je potentiële verklarende variabelen
- Selecteer de belangrijkste variabelen op grond van theoretische overwegingen en de gevonden P-waarden (neem een ruime  $\alpha$ , bijvoorbeeld 0,25)
- Bouw je model stap voor stap op, begin bij je belangrijkste (meest significante) variabele
- Voeg variabelen toe zolang het model significant beter wordt  
Je kunt ook "backward" te werk gaan; begin met alle variabelen in je model en verwijder één voor één de minst significante variabele tot je een model overhoudt met alleen maar significante of theoretisch relevante variabelen.
- Als alle relevante variabelen in het model zitten, kijk dan eventueel naar interacties (zie hieronder)
- Controleer de residuen van het uiteindelijke model op normaliteit en homogeniteit van de spreiding

## Interactie (effectmodificatie)



## Interpretatie van interactie

Intgewsmoke = gewicht \* smoke

$$\text{geboortegewicht} = b_0 + b_1 * \text{gewicht} + b_2 * \text{rookgedrag} + b_3 * \text{Intgewsmoke}$$

| Coefficients <sup>a</sup> |                                 |                             |            |                           |      |
|---------------------------|---------------------------------|-----------------------------|------------|---------------------------|------|
| Model                     |                                 | Unstandardized Coefficients |            | Standardized Coefficients | Sig. |
|                           |                                 | B                           | Std. Error | Beta                      |      |
| 1                         | (Constant)                      | 2347.507                    | 312.717    |                           | .000 |
|                           | gewicht                         | 11.905                      | 5.143      | .227                      | .022 |
|                           | smoking status during pregnancy | 47.867                      | 451.163    | .032                      | .916 |
|                           | intgewsmoke                     | -2.456                      | 3.388      | -.223                     | .470 |

a. Dependent Variable: birth weight in grams

Niet Significant!

voor niet-rooksters :

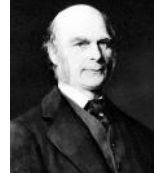
$$\begin{aligned} \text{geboortegewicht} &= 2348 + 11,9 * \text{gewicht} + 48 * 0 - 2,5 * \text{gewicht} * 0 = \\ &= 2348 + 11,9 * \text{gewicht} \end{aligned}$$

voor rooksters

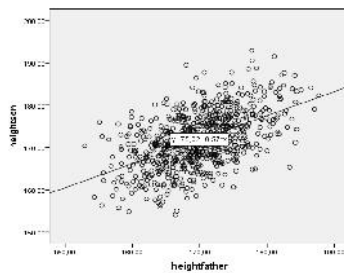
$$\begin{aligned} \text{geboortegewicht} &= 2348 + 11,9 * \text{gewicht} + 48 * 1 - 2,5 * \text{gewicht} * 1 = \\ &= 2396 + 9,4 * \text{gewicht} \end{aligned}$$

## Waar komt de term regressie vandaan?

- Regressie = terugkeer, terugval
- Francis Galton (1822 – 1911)
- Onderzoek naar lengtes van vaders en hun zonen



## Regression towards the mean (regressie naar het gemiddelde)



De term is een historisch misverstand!