

depth(M) =  $\frac{n+1}{2}$   
 depth(H) =  $\frac{[depth(M)]+1}{2}$   
 $H_u \rightarrow spread = H_u - H_L$   
 $H_u - M$ ; measure of skewness;  
 $M - H_L \approx 1$  grade transformatie.

Box-Cox:  $X \rightarrow X^{(p)}$   
 $= \frac{X^p - 1}{p}$ , met  $X^{(0)} = \ln(X)$   
 Als  $\log(H_u - H_L) \approx a + b \log(M)$   
 $\Rightarrow p = -b$   
 $\begin{cases} p > 1 & \text{neg, left} \\ p < 1 & \text{pos, right} \end{cases}$

simple monotone lineariteit:  $y \uparrow$  up,  $x$  down;  $y \downarrow$  down,  $x$  up;  
 $\log(X)$ ,  $\sqrt{x}$ ,  $\log(y)$

Residual  $E_i = y_i - \hat{y}_i$   
 $A = \bar{y} - B\bar{x}$   
 $B = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$   
 Residual standard error  $SE = \sqrt{\frac{\sum E_i^2}{n-2}}$

RegSS: = variatie y-waarden "explained" door model  
 $RSS = \sum (y_i - \bar{y})^2$  niet verhaard door model.  
 $\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$   
 ofwel:  $TSS = RegSS + RSS$

$r^2/R^2 = RegSS / TSS$   
 r correlatie  $y_i$  en  $x_i$ ,  
 R correlatie  $y_i$  en  $\hat{y}_i$   
 $Cov(X, Y) = E(XY) - E(X)E(Y)$   
 $Cov(X, X) = Var(X)$ , geeft aan hoe 2 random variabelen gerelateerd zijn (positief, negatief)

Multiple Regression: ANOVA  
 source SS df MS F  
 Regression RegSS k  $RegSS/k$   $RegSS/RMS$   
 Residuals RSS n-k-1  $RSS/(n-k-1)$   
 Total TSS n-1

n model equations:  $y = X\beta + \epsilon$   
 $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ ,  $X = \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$  of X.  
 $b = (X'X)^{-1} X'y$   
 Null-Hypothesis:  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$   
 $F_0 = \frac{RegSS/k}{RSS/(n-k-1)} = \frac{RegMS}{RMS}$   
 $F_0 \sim F(k, n-k-1)$

Null-Hypothesis on subset of slopes:  
 $H_0: \beta_1 = \dots = \beta_q = 0$   
 $F_0 = \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-k-1)}$   
 $F_0 \sim F(q, n-k-1)$   
 single slope:  $H_0: \beta_j = 0$  tegen  $H_1: \beta_j \neq 0$ : reject  $H_0$  als  $T \ll 0$  of  $T \gg 0$  en  $n-k-1$

adding a slope: ... de exponentiële familie.  
 $F = \frac{RegSS_j - RegSS_{j-1}}{RSS/(n-k-1)}$   
 $F_0 \sim F(1, n-k-1)$   
 factors: Dichotomous (gender); Polytomous (type profession)

Dummy: 3 level profession-tipe:  $y_i = prestige$ ,  $x_1 = income$ ,  $x_2 = education$   
 $H_1$  / Full-model / met interactie:  
 $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \delta_{11} x_{i1} D_{i1} + \delta_{12} x_{i1} D_{i2} + \delta_{21} x_{i2} D_{i1} + \delta_{22} x_{i2} D_{i2} + \epsilon_i$   
 $H_0$ : geen interactie:  $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$   
 $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$   
 $H_0: \gamma_1 = \gamma_2 = 0$   $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$  := variabele type irrelevant  
 $H_0: \beta_1 = 0$   $\delta_{11} = \delta_{12} = 0$  := variabele income zelf is irrelevant.

One-Way-ANOVA:  
 source SS df MS F  
 Groups  $\sum n_j (\bar{y}_j - \bar{y})^2$  m-1  $RegSS/m-1$   $RegMS/RMS$   
 Residuals  $\sum \sum (y_{ij} - \bar{y}_j)^2$  n-m  $RSS/n-m$   
 Total  $\sum \sum (y_{ij} - \bar{y})^2$  n-1  
 $F \sim F(m-1, n-m)$  to test  $H_0: \alpha_1 = \dots = \alpha_m = 0$ .

Two-Way-ANOVA:  
 source df SS  $H_0$   
 R r-1  $SS(R|\beta, \gamma)$  all  $\alpha_j = 0$   
 C c-1  $SS(C|\beta, \gamma)$  all  $\gamma_{jk} = 0$   
 RC (r-1)(c-1)  $SS(RC|\beta, \gamma)$  all  $\gamma_{jk} = 0$   
 Residuals n-rc  $TSS - SS(R, C, RC)$   
 $F = \frac{\text{Incremental Sum of Squares/df}}{RSS/(n-rc)}$

q: = verschil in parameters null- en alternatief model.  
 $n-k-1$  RSS degrees of freedom  
 univariate outlier: value of Y (or X) die unconditionally unusual is.  
 Regression outlier: observation whose y value is "... given the value of X.  
 High-Leverage: observation with outlying X-value.  
 Influence on coefficients = Leverage

Regression Outlyingness  
 Generalized Lindeberg  
 models  $\rightarrow$  GLM's verdeeld volgens  
 GLM ("Generalized Linear model"): Een GLM bestaat uit 3 componenten: (1) Een will. component die de voorwaardelijke distributie van de responsvariabele,  $y_i$ , specificiert, wanneer de waarden van de verklarende variabelen bekend zijn in het model. In de aan vankelijke formulering van de GLM's, was de verdeling van  $y_i$  een lid van

de steekproefvariantie van de coëfficiënt  $\beta_j$  is het j<sup>e</sup> diagonale element van  $V(b)$ :  $V(b_j) = \sigma_e^2$ .  
 $\sum_{i=1}^n m_{ji}^2 + \sum_{i=1}^n a_{ji}^2$   
 Beide sommaties zijn kwadraten sommen en dus is, zijn alle  $a_{ji}$  gelijk aan 0. Dit geldt voor elke coëfficiënt in  $b$  en dus geldt dat elke rij van  $A$  gelijk is aan 0, dus  $A=0$ , dus  $b = (0+M)y = My = b$ .  $\square$   
 $b = \log(M_{u, x_1} - H_{u, x_1}) - \log(M_{u, x_2} - H_{u, x_2}) / \log(M_{u, x_1} - \log(M_{u, x_2}))$   
 $U^2 \sim t(20)$ ,  $U^2 \sim F(1, 20)$   
 $\min S(A, B_1, B_2, \dots, B_k)$   
 $= \sum_{i=1}^n [y_i - (A + B_1 x_{i1} + B_2 x_{i2} + \dots + B_k x_{ik})]^2$  Volgende vgl 'en:  
 $A + B_1 \sum x_{i1} + B_2 \sum x_{i2} + \dots + B_k \sum x_{ik} = \sum y_i$  (1)  
 $A \sum x_{i1} + B_1 \sum x_{i1}^2 + B_2 \sum x_{i1} x_{i2} + \dots + B_k \sum x_{i1} x_{ik} = \sum x_{i1} y_i$  (2)  
 $A \sum x_{i2} + B_1 \sum x_{i2} x_{i1} + B_2 \sum x_{i2}^2 + \dots + B_k \sum x_{i2} x_{ik} = \sum x_{i2} y_i$  (3)  
 $\vdots$   
 $A \sum x_{ik} + B_1 \sum x_{ik} x_{i1} + B_2 \sum x_{ik} x_{i2} + \dots + B_k \sum x_{ik}^2 = \sum x_{ik} y_i$

De steekproefvariantie van de coëfficiënt  $\beta_j$  is het j<sup>e</sup> diagonale element van  $V(b)$ :  $V(b_j) = \sigma_e^2$ .  
 $\sum_{i=1}^n m_{ji}^2 + \sum_{i=1}^n a_{ji}^2$   
 Beide sommaties zijn kwadraten sommen en dus is, zijn alle  $a_{ji}$  gelijk aan 0. Dit geldt voor elke coëfficiënt in  $b$  en dus geldt dat elke rij van  $A$  gelijk is aan 0, dus  $A=0$ , dus  $b = (0+M)y = My = b$ .  $\square$   
 $b = \log(M_{u, x_1} - H_{u, x_1}) - \log(M_{u, x_2} - H_{u, x_2}) / \log(M_{u, x_1} - \log(M_{u, x_2}))$   
 $U^2 \sim t(20)$ ,  $U^2 \sim F(1, 20)$   
 $\min S(A, B_1, B_2, \dots, B_k)$   
 $= \sum_{i=1}^n [y_i - (A + B_1 x_{i1} + B_2 x_{i2} + \dots + B_k x_{ik})]^2$  Volgende vgl 'en:  
 $A + B_1 \sum x_{i1} + B_2 \sum x_{i2} + \dots + B_k \sum x_{ik} = \sum y_i$  (1)  
 $A \sum x_{i1} + B_1 \sum x_{i1}^2 + B_2 \sum x_{i1} x_{i2} + \dots + B_k \sum x_{i1} x_{ik} = \sum x_{i1} y_i$  (2)  
 $A \sum x_{i2} + B_1 \sum x_{i2} x_{i1} + B_2 \sum x_{i2}^2 + \dots + B_k \sum x_{i2} x_{ik} = \sum x_{i2} y_i$  (3)  
 $\vdots$   
 $A \sum x_{ik} + B_1 \sum x_{ik} x_{i1} + B_2 \sum x_{ik} x_{i2} + \dots + B_k \sum x_{ik}^2 = \sum x_{ik} y_i$

GLM ("Generalized Linear model"): Een GLM bestaat uit 3 componenten: (1) Een will. component die de voorwaardelijke distributie van de responsvariabele,  $y_i$ , specificiert, wanneer de waarden van de verklarende variabelen bekend zijn in het model. In de aan vankelijke formulering van de GLM's, was de verdeling van  $y_i$  een lid van

de steekproefvariantie van de coëfficiënt  $\beta_j$  is het j<sup>e</sup> diagonale element van  $V(b)$ :  $V(b_j) = \sigma_e^2$ .  
 $\sum_{i=1}^n m_{ji}^2 + \sum_{i=1}^n a_{ji}^2$   
 Beide sommaties zijn kwadraten sommen en dus is, zijn alle  $a_{ji}$  gelijk aan 0. Dit geldt voor elke coëfficiënt in  $b$  en dus geldt dat elke rij van  $A$  gelijk is aan 0, dus  $A=0$ , dus  $b = (0+M)y = My = b$ .  $\square$   
 $b = \log(M_{u, x_1} - H_{u, x_1}) - \log(M_{u, x_2} - H_{u, x_2}) / \log(M_{u, x_1} - \log(M_{u, x_2}))$   
 $U^2 \sim t(20)$ ,  $U^2 \sim F(1, 20)$   
 $\min S(A, B_1, B_2, \dots, B_k)$   
 $= \sum_{i=1}^n [y_i - (A + B_1 x_{i1} + B_2 x_{i2} + \dots + B_k x_{ik})]^2$  Volgende vgl 'en:  
 $A + B_1 \sum x_{i1} + B_2 \sum x_{i2} + \dots + B_k \sum x_{ik} = \sum y_i$  (1)  
 $A \sum x_{i1} + B_1 \sum x_{i1}^2 + B_2 \sum x_{i1} x_{i2} + \dots + B_k \sum x_{i1} x_{ik} = \sum x_{i1} y_i$  (2)  
 $A \sum x_{i2} + B_1 \sum x_{i2} x_{i1} + B_2 \sum x_{i2}^2 + \dots + B_k \sum x_{i2} x_{ik} = \sum x_{i2} y_i$  (3)  
 $\vdots$   
 $A \sum x_{ik} + B_1 \sum x_{ik} x_{i1} + B_2 \sum x_{ik} x_{i2} + \dots + B_k \sum x_{ik}^2 = \sum x_{ik} y_i$

opg: gegeven: Nation Sample variance  $H_L$  Median  $M_u$   
 Beland was us 19331 993 427.0 1123 3586  
 dat de log de other 434 987 058 566.5 1592 5250  
 grote transformatie voor X was. Nu, willen we b weten zodat we een machts transformatie  $p = 1-b$  kunnen toepassen. Bereken:  
 us other  
 $\log(M)$  7.032624 7.372746  
 $\log(H_u - H_L)$  8.061171 8.451801  
 voor beide groepen, dus  $b = 8.061171 - 8.451801 = -1.148498$   
 $7.032624 - 7.372746$   
 De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan

opg: Stel  $(X_1, X_2, X_3) \sim N(\mu, \Sigma)$ , met  $\mu = (-1, 0, 1)$  en  $\Sigma = \begin{pmatrix} 2 & -1 & 4 \\ -1 & 2 & -1 \\ 4 & -1 & 2 \end{pmatrix}$ . (a) Bereken  $P(X_1 - 2X_2 + 4X_3 < 2)$ . antwoord: Stel  $Y_i = X_i - 2X_2 + 4X_3$ , dan  $E(Y) = 3$  en  $Y = \begin{pmatrix} 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ , dus  $Var(Y) = \begin{pmatrix} 1 & -2 & -4 \end{pmatrix} \Sigma \begin{pmatrix} -2 \\ -1 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 & -2 & -4 \end{pmatrix} \begin{pmatrix} 20 \\ -9 \\ 14 \end{pmatrix} = 94$ , dus  $Y \sim N(3, 94)$ . Dan:  $P(Y < 2) = P(\frac{Y-3}{\sqrt{94}} < \frac{2-3}{\sqrt{94}}) = P(Z < -0.1) = 0.4602$ .

(b) U stochast met student t-verdeling met 20 vrijheidsgraden. Geef q, zdd.  $P(U^2 > q) = 0.05$ . antw:  $U \sim t(20)$ , dus  $U^2 \sim F(1, 20)$ .  $\Rightarrow$  zoek waarde tabel.  
 opg: Nieuw type camera in 4 regio's; elke regio 5 dealerbieden. Verkoop weergegeven:

| Regio | 1        | 2        | 3       |
|-------|----------|----------|---------|
| 1     | 4323     | 8867     | 1210911 |
| 2     | 8766     | 3214     | 6787    |
| 3     | 10141315 | 3324     | 3234    |
| 4     | 3212     | 15161214 | 8776    |

met de onderstaande ANOVA tabel  
 (a) vul verder in:  
 antw: gebruik income en region van significante invloed is.  
 antw:  $H_1: y_i = \alpha + \beta x_i + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \gamma_3 D_{i3} + \delta_1 x_i D_{i1} + \delta_2 x_i D_{i2} + \delta_3 x_i D_{i3} + \epsilon_i$   
 $H_0: \delta_1 = \delta_2 = \delta_3 = 0$   
 RegSS volledige model is  $RegSS_1 = 47.003 + \dots + 0.122 = 60.612$ .  
 $RegSS_0 = 47.003 + \dots + 1.541 = 57.789$  (Dit kan doordat alles een voor een is toegevoegd.)  
 $q = 3$ , dus:  
 $F_0 = \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-rc)} = \frac{2.82713}{0.356} = 2.647$   
 $F_c = F(3, 93) \approx F(3, 90) = 2.71$ . We verwerpen  $H_0$  dus niet.  
 (c) ondertoeck of 'region' sign. invloed heeft.  
 antw: we testen  $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0$  tegen  $H_1: y_i = \alpha + \beta x_i + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \gamma_3 D_{i3}$   
 $RegSS_1 = 47.003 + \dots + 1.541 = 57.789$   
 $RegSS_0 = 47.003$   
 $RSS_1 = 33.152 + 2.705 + 0.000 + 0.122 = 35.979$  en heeft 93+1+1+1=96 df.. Verder  $q = 3$   
 $F_0 = \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-k-1)} = \frac{10.70613}{35.979/96} = 9.522$   
 $F(3, 96) \approx F(3, 100) = 2.70$ . We kunnen de  $H_0$  hypothese verwerpen.

| Source     | SS      | df | MS     | F     |
|------------|---------|----|--------|-------|
| regio      | 34.729  | 3  | 11.576 | 8.292 |
| conc. niv. | 4.292   | 2  | 2.146  | 1.537 |
| interactie | 730.302 | 6  | 121.7  | 87.19 |
| RSS        | 50.250  | 36 | 1.396  |       |
| total      | 817.67  | 47 |        |       |

interactie, gebruik  $\alpha = 0.05$ . Formuleer hypothesen en modellen. antw: D: = regio, C: = conc niv.  
 we testen  $H_0: y_i = \alpha + \beta_1 D_{i1} + \beta_2 D_{i2} + \beta_3 D_{i3} + \gamma_1 C_{i1} + \gamma_2 C_{i2} + \gamma_3 C_{i3} + \delta_{11} D_{i1} C_{i1} + \delta_{12} D_{i1} C_{i2} + \delta_{13} D_{i1} C_{i3} + \delta_{21} D_{i2} C_{i1} + \delta_{22} D_{i2} C_{i2} + \delta_{23} D_{i2} C_{i3} + \delta_{31} D_{i3} C_{i1} + \delta_{32} D_{i3} C_{i2} + \delta_{33} D_{i3} C_{i3} + \epsilon_i$   
 We hebben al een F-waarde in de tabel,  $F = 87.19$ . Deze is  $F(6, 36)$  verdeeld.  $F_c \approx F(3, 40) = 2.34$ , dus we kunnen  $H_0$  verwerpen.

AIC: = Akaike Information Criterion def  $AIC = -2 \ln(L) - 2s$ , met L de likelihood v.h. model en s het aantal parameters.  
 opgave: Appendix (ANOVA tabellen). Coëfficiënten:  
 value std. Error t-value Pr(>|t|)  
 Intercept 7.027  
 income -0.532  
 d1 -2.008  
 d2 -0.522  
 d3 -0.825  
 income:d1 0.520  
 income:d2 0.123  
 income:d3 0.141

De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan  
 $b = 8.061171 - 8.451801 = -1.148498$   
 $7.032624 - 7.372746$   
 De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan

De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan  
 $b = 8.061171 - 8.451801 = -1.148498$   
 $7.032624 - 7.372746$   
 De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan

opg: gegeven: Nation Sample variance  $H_L$  Median  $M_u$   
 Beland was us 19331 993 427.0 1123 3586  
 dat de log de other 434 987 058 566.5 1592 5250  
 grote transformatie voor X was. Nu, willen we b weten zodat we een machts transformatie  $p = 1-b$  kunnen toepassen. Bereken:  
 us other  
 $\log(M)$  7.032624 7.372746  
 $\log(H_u - H_L)$  8.061171 8.451801  
 voor beide groepen, dus  $b = 8.061171 - 8.451801 = -1.148498$   
 $7.032624 - 7.372746$   
 De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan

opg: Stel  $(X_1, X_2, X_3) \sim N(\mu, \Sigma)$ , met  $\mu = (-1, 0, 1)$  en  $\Sigma = \begin{pmatrix} 2 & -1 & 4 \\ -1 & 2 & -1 \\ 4 & -1 & 2 \end{pmatrix}$ . (a) Bereken  $P(X_1 - 2X_2 + 4X_3 < 2)$ . antwoord: Stel  $Y_i = X_i - 2X_2 + 4X_3$ , dan  $E(Y) = 3$  en  $Y = \begin{pmatrix} 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ , dus  $Var(Y) = \begin{pmatrix} 1 & -2 & -4 \end{pmatrix} \Sigma \begin{pmatrix} -2 \\ -1 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 & -2 & -4 \end{pmatrix} \begin{pmatrix} 20 \\ -9 \\ 14 \end{pmatrix} = 94$ , dus  $Y \sim N(3, 94)$ . Dan:  $P(Y < 2) = P(\frac{Y-3}{\sqrt{94}} < \frac{2-3}{\sqrt{94}}) = P(Z < -0.1) = 0.4602$ .

(b) U stochast met student t-verdeling met 20 vrijheidsgraden. Geef q, zdd.  $P(U^2 > q) = 0.05$ . antw:  $U \sim t(20)$ , dus  $U^2 \sim F(1, 20)$ .  $\Rightarrow$  zoek waarde tabel.  
 opg: Nieuw type camera in 4 regio's; elke regio 5 dealerbieden. Verkoop weergegeven:

| Regio | 1        | 2        | 3       |
|-------|----------|----------|---------|
| 1     | 4323     | 8867     | 1210911 |
| 2     | 8766     | 3214     | 6787    |
| 3     | 10141315 | 3324     | 3234    |
| 4     | 3212     | 15161214 | 8776    |

met de onderstaande ANOVA tabel  
 (a) vul verder in:  
 antw: gebruik income en region van significante invloed is.  
 antw:  $H_1: y_i = \alpha + \beta x_i + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \gamma_3 D_{i3} + \delta_1 x_i D_{i1} + \delta_2 x_i D_{i2} + \delta_3 x_i D_{i3} + \epsilon_i$   
 $H_0: \delta_1 = \delta_2 = \delta_3 = 0$   
 RegSS volledige model is  $RegSS_1 = 47.003 + \dots + 0.122 = 60.612$ .  
 $RegSS_0 = 47.003 + \dots + 1.541 = 57.789$  (Dit kan doordat alles een voor een is toegevoegd.)  
 $q = 3$ , dus:  
 $F_0 = \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-rc)} = \frac{2.82713}{0.356} = 2.647$   
 $F_c = F(3, 93) \approx F(3, 90) = 2.71$ . We verwerpen  $H_0$  dus niet.  
 (c) ondertoeck of 'region' sign. invloed heeft.  
 antw: we testen  $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0$  tegen  $H_1: y_i = \alpha + \beta x_i + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \gamma_3 D_{i3}$   
 $RegSS_1 = 47.003 + \dots + 1.541 = 57.789$   
 $RegSS_0 = 47.003$   
 $RSS_1 = 33.152 + 2.705 + 0.000 + 0.122 = 35.979$  en heeft 93+1+1+1=96 df.. Verder  $q = 3$   
 $F_0 = \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-k-1)} = \frac{10.70613}{35.979/96} = 9.522$   
 $F(3, 96) \approx F(3, 100) = 2.70$ . We kunnen de  $H_0$  hypothese verwerpen.

| Source     | SS      | df | MS     | F     |
|------------|---------|----|--------|-------|
| regio      | 34.729  | 3  | 11.576 | 8.292 |
| conc. niv. | 4.292   | 2  | 2.146  | 1.537 |
| interactie | 730.302 | 6  | 121.7  | 87.19 |
| RSS        | 50.250  | 36 | 1.396  |       |
| total      | 817.67  | 47 |        |       |

interactie, gebruik  $\alpha = 0.05$ . Formuleer hypothesen en modellen. antw: D: = regio, C: = conc niv.  
 we testen  $H_0: y_i = \alpha + \beta_1 D_{i1} + \beta_2 D_{i2} + \beta_3 D_{i3} + \gamma_1 C_{i1} + \gamma_2 C_{i2} + \gamma_3 C_{i3} + \delta_{11} D_{i1} C_{i1} + \delta_{12} D_{i1} C_{i2} + \delta_{13} D_{i1} C_{i3} + \delta_{21} D_{i2} C_{i1} + \delta_{22} D_{i2} C_{i2} + \delta_{23} D_{i2} C_{i3} + \delta_{31} D_{i3} C_{i1} + \delta_{32} D_{i3} C_{i2} + \delta_{33} D_{i3} C_{i3} + \epsilon_i$   
 We hebben al een F-waarde in de tabel,  $F = 87.19$ . Deze is  $F(6, 36)$  verdeeld.  $F_c \approx F(3, 40) = 2.34$ , dus we kunnen  $H_0$  verwerpen.

AIC: = Akaike Information Criterion def  $AIC = -2 \ln(L) - 2s$ , met L de likelihood v.h. model en s het aantal parameters.  
 opgave: Appendix (ANOVA tabellen). Coëfficiënten:  
 value std. Error t-value Pr(>|t|)  
 Intercept 7.027  
 income -0.532  
 d1 -2.008  
 d2 -0.522  
 d3 -0.825  
 income:d1 0.520  
 income:d2 0.123  
 income:d3 0.141

De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan  
 $b = 8.061171 - 8.451801 = -1.148498$   
 $7.032624 - 7.372746$   
 De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan

De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan  
 $b = 8.061171 - 8.451801 = -1.148498$   
 $7.032624 - 7.372746$   
 De kleinste kwadraten regressie lijn moet door  $\log(M)$  en de  $\log(H_u - H_L)$  gaan

Terms added sequentially:  
 Region D1 D2 D3  
 Africa 1 0 0  
 America 0 1 0  
 Asia 0 0 1  
 Europe 0 0 0

↑ kinderterfte per 1000 met var income en factor region. (output regressie)  
 (a) geef een schatting voor het kindersterftecijfer per 1000 voor een Europees gezin met een jaarinkomen van 12.  
 antw: model voor Europa is  $y_i = \alpha + \beta_1 \cdot inc$ .  
 Nu  $\alpha = 7.027$  en  $\beta_1 = -0.532$ , dus  $X = 12$  geeft  $V = 0.643$  dus cijfer is gelijk aan  $2 \cdot 0.643 = 1.562$

|           | Df | SS     | MS    | F-value | Pr(>F)  |
|-----------|----|--------|-------|---------|---------|
| income    | 1  | 47.003 | zelfd | 132.08  | 0.00000 |
| d1        | 1  | 7.833  | "     | 21.97   | 0.00001 |
| d2        | 1  | 1.332  | "     | 3.74    | 0.05631 |
| d3        | 1  | 1.541  | "     | 4.32    | 0.04038 |
| income:d1 | 1  | 2.705  | "     | 7.59    | 0.00706 |
| income:d2 | 1  | 0.000  | "     | 0.00    | 0.97515 |
| income:d3 | 1  | 0.122  | "     | 0.34    | 0.55971 |
| Residuals | 93 | 33.152 | 0.356 |         |         |

(b) ondertoeck met  $\alpha = 0.$