Multivariat Statistik TAMS39

Jonathan Vikbladh - jonvi252

September 2021

Omgång 1

1

Anta att datamatrisen $\boldsymbol{X}$ är en observation från $\boldsymbol{X} \sim \mathcal{N}_{2,45}(\boldsymbol{\mu}\mathbf{1}'_{45}, \boldsymbol{\Sigma}, \boldsymbol{I})$, då ska vi testa

$$H : \boldsymbol{\mu} = \boldsymbol{\mu}_0 = \begin{bmatrix}190 & 275\end{bmatrix}'$$
$$A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

Vi skattar de okända parametrarna

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{45}\boldsymbol{X}\mathbf{1}_{45}$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \frac{1}{45-1}\boldsymbol{X}(\boldsymbol{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')\boldsymbol{X}'$$

Vi bildar vår teststorhet

$$T^2 = 45(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \sim \frac{(n-1)p}{n-p}F_{p,n-p}$$

Vi förkastar hypotesen då $T^2 > \frac{(n-1)p}{n-p}F_{p,n-p}(0.95)$, vi beräknar allt och får $T^2 = 5.54313 < 6.578471$ den kritiska gränsen, så vi kan inte förkasta nollhypotesen att väntevärdesvektorn är $\boldsymbol{\mu}_0$.

b)

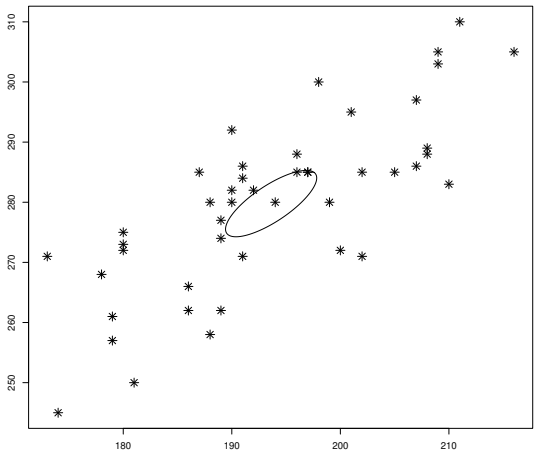

Figure 1: Konfidensellips över honfåglarnas väntevärdesvektor

c)

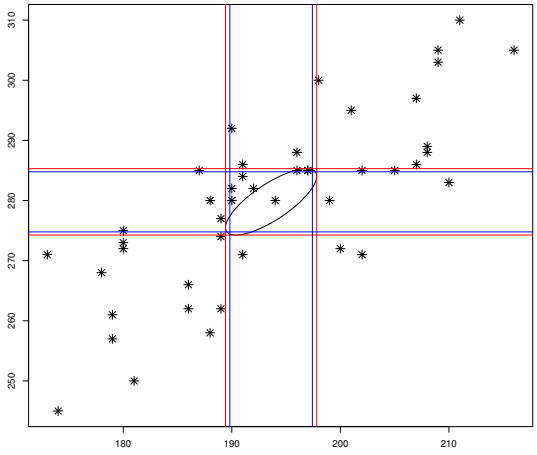

Figure 2: Bonferroni i blått, Roy/T2 i rött

Bonferroni ger smalare intervall, och skulle kunna ge att vi förkastar även om vi befinner oss i ellipsen. Med Roy däremot kommer det inte hända dvs vi kommer inte förkasta fast det är sant dvs typ 1 fel. Vice versa kommer vi med Bonferroni att mer sällan acceptera fast det är falskt dvs typ 2 fel eftersom

hörnen snett upp vänster och snett ner höger har mindre area jämfört Roy. Inget av intervallen tar dock hänsyn till att datan samvarierar dvs ellipsens form.

2

a)

Vi ser att observation 31 avviker från resten av observationerna genom att den har en mycket större svanslängd

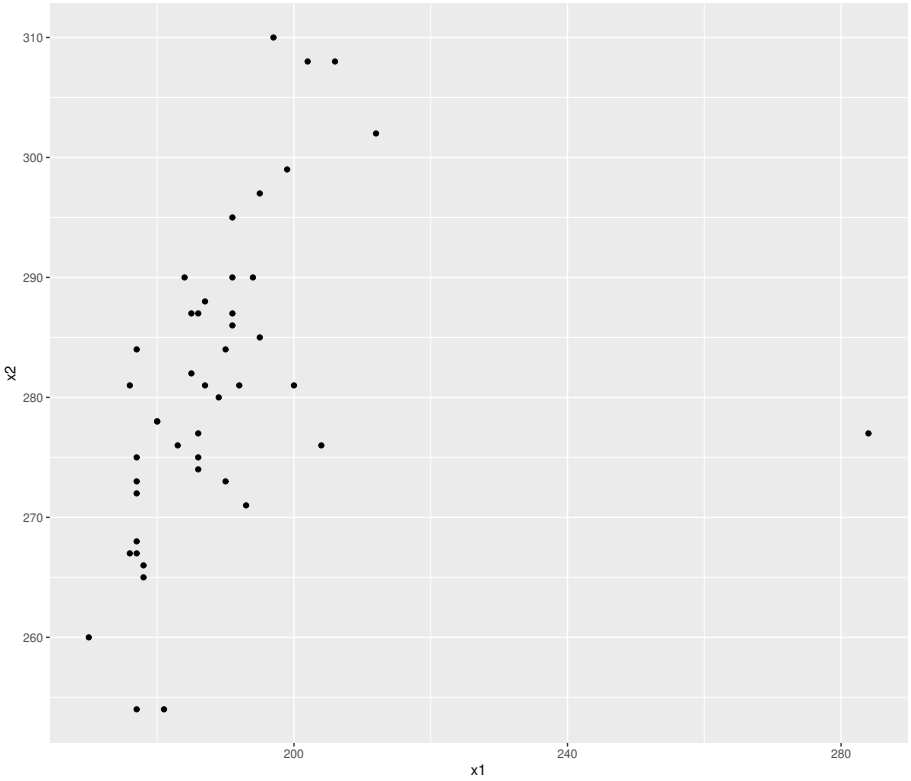

Figure 3: Caption

b)

Vi antar att datan på hannarna $\boldsymbol{X}_1 = (\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{45,1})$ där varje observation är en kolumn som är multivariat normalfördelad $\boldsymbol{x}_{j,1} \sim \mathcal{N}_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $j = 1, \ldots, 45$ och observationerna är oberoende sinsemellan. På samma sätt för honorna, $\boldsymbol{X}_2 = (\boldsymbol{x}_{1,2}, \ldots, \boldsymbol{x}_{45,2})$ med $\boldsymbol{x}_{j,1} \sim \mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $j = 1, \ldots, 45$ oberoende obser-

vationer. Vi skattar parametrarna för hannarna respektive honorna

$$\bar{\boldsymbol{x}}_1 = \hat{\boldsymbol{\mu}}_1 = \frac{1}{45}\boldsymbol{X}_1\mathbf{1}_{45}$$

och

$$\bar{\boldsymbol{x}}_2 = \hat{\boldsymbol{\mu}}_2 = \frac{1}{45}\boldsymbol{X}_2\mathbf{1}_{45}$$

samt

$$V_1 = \boldsymbol{X}_1(\boldsymbol{I}_{45} - \frac{1}{45}\mathbf{1}_{45}\mathbf{1}'_{45})\boldsymbol{X}'_1$$

$$V_2 = \boldsymbol{X}_2(\boldsymbol{I}_{45} - \frac{1}{45}\mathbf{1}_{45}\mathbf{1}'_{45})\boldsymbol{X}'_2$$

och under $H : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ kan vi poola så vi bildar även

$$\boldsymbol{V} = \boldsymbol{V}_1 + \boldsymbol{V}_2$$

och låter $n = n_1 + n_2 = 45 + 45 = 90$, och $p = 2$ ty 2 variabler
Vi gör ett likelihood ratio test och eftersom vi har lika många observationer från båda populationerna får vi teststorheten

$$\Lambda = \frac{\det(\boldsymbol{V}_1)^{45/2}\det(\boldsymbol{V}_2)^{45/2}(90)^{90}}{\det(\boldsymbol{V})^{90/2}(45)^{90}}$$

Vi använder en Box-korrigering för snabbare asymptotik mot en χ^2-fördelning och förkastar nollhypotesen då

$$z = 2f^{-1}m\ln\Lambda > \chi^2(0.95; df)$$

där $2f^{-1}m$ är Box-korrigeringen och df är skilladen i antalet skattade oberoende parameterar i alternativhypotesen och nollhypotesen. Vi beräknar allt och får att vår teststorhet

$$z = 28.16144 > 7.814728$$

är större än den kritiska gränsen så vi förkastar nollhypotesen att kovariansmatriserna för populationerna är densamma

c)

Vi ändrar först den avvikande observationen genom att sätta $x_1 = 184$ istället för 284. Sen genomför vi samma test som i a) med nya skattningar, eftersom det fortfarande är samma antal observationer från populationerna. Vi får teststorheten

$$z = 1.250097 < 7.814728$$

så vi kan inte förkasta nollhypotesen att populationerna har samma kovariansmatriser så vi fortsätter med testet för väntevärdesvektorerna. Vi testar

$$H: \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

mot $A \neq H$. Vi antar som tidigare att observationerna från de två olika populationerna är multivariat normalfördelade men nu kommer de från $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ respektive $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Dvs vi antar att de har samma kovariansmatris. Alltså kan vi poola våra stickprovskovariansmatriser $\boldsymbol{S} = (45-1)\boldsymbol{S}_1 + (45-1)\boldsymbol{S}_2$ och bilda

$$\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, (\frac{1}{45} + \frac{1}{45})\boldsymbol{\Sigma})$$

och fortsätta med ett T^2-test, vi får att teststorheten

$$T^2 = 25.66253 > 6.273886$$

dvs den kritiska gränsen, så vi kan förkasta nollhypotesen att populationerna har samma väntevärdesvektor.
Sedan tar vi bort hela observationen, och genomför testet $H : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, testet blir som i a) fast för att få väntevärdesriktiga skattningar för noll- respektive alternativhypotesen byter vi ut n_i mot $f_i = n_i - 1$ i teststorheten. Vi får att vår Box-korrigerade teststorhet är

$$z = 1.043102 < 7.814728$$

så vi kan inte heller här förkasta att kovariansmatriserna är samma, vi fortsätter med ett T^2-test som innan och får

$$T^2 = 24.9649 > 6.27725$$

så vi kan återigen förkasta att väntevärdesvektorerna är samma. Verkar alltså inte spela någon roll om vi tar bort eller ändrar observationen, väntevärdesvektorerna verkar vara annorlunda.

d)

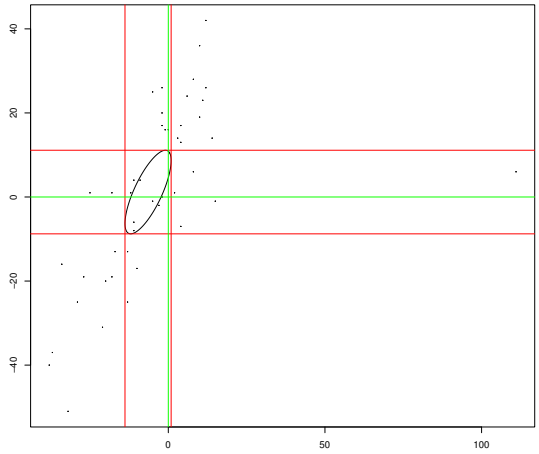

Figure 4: Gröna linjerna korsar där vår teststorhet är

Ellipsen är konfidensregionen för skillnaden i väntevärde och de röda linjerna är de simultana konfidensintervallen över komponenterna.

e)

Vi ser att honor och hanar inte är lika stora, men att hanar har längre vingspann och honor har längre svansar. Dock är honornas svansar mycket längre än hanarnas jämfört med hur mycket vingspannen skiljer, så honorna är generellt sett större åtminstone sett till svansen.

3

a)

Vi testar

$$H: \quad \boldsymbol{\mu} = \begin{bmatrix} 2 \\ 2/3 \end{bmatrix} \text{ mot } A: \quad \boldsymbol{\mu} \neq \begin{bmatrix} 2 \\ 2/3 \end{bmatrix}$$

Vi har att

$$\bar{\boldsymbol{X}} \sim \mathcal{N}_2(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

så

$$\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{I})$$

Vi bildar vår teststorhet som

$$Z = (\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}))'(\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})) \sim \chi^2(2)$$

Så vi förkastar vår teststorhet då $z = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) > \chi^2(0.95; 2)$. Vi beräknar vår teststorhet och får $z = 4.777778 < 5.991465$ så vi kan inte förkasta nollhypotesen

b)

$$H: \quad \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 = \frac{7}{2}$$

Vi har att $\bar{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$, låt $\boldsymbol{a}' = \begin{bmatrix} 1 & 1 \end{bmatrix}$ så fås att $\boldsymbol{a}'\bar{\boldsymbol{X}} \sim \mathcal{N}(\boldsymbol{a}'\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$.

Så under H har vi $\boldsymbol{a}'\bar{\boldsymbol{X}} \sim \mathcal{N}(\frac{7}{2}, \frac{1}{n}\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$, bilda teststorheten

$$Z = \frac{\boldsymbol{a}'\bar{\boldsymbol{x}} - \frac{7}{2}}{\sqrt{\frac{1}{n}\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}} \sim \mathcal{N}(0, 1)$$

Förkasta nollhypotesen då $|Z| > \Phi^{-1}(1 - \frac{\alpha}{2}) = 1.96$, vi får $Z = -2\sqrt{3} \approx -3.46 \implies |Z| > 1.96$ vi kan förkasta nollhypotesen.

c)

$$H: \quad \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \frac{1}{2}$$

Vi använder $\boldsymbol{a}' = \begin{bmatrix} 1 & -1 \end{bmatrix}$ och får samma (univariata normal-)fördelning som ovan. Teststorheten blir

$$Z = \frac{\boldsymbol{a}^T\bar{\boldsymbol{x}} - \frac{1}{2}}{\sqrt{\frac{1}{n}\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}}} \sim \mathcal{N}(0, 1)$$

Vi får $z = 0 \implies |z| \leq 1.96$ vi kan ej förkasta nollhypotesen

d)

Med $\boldsymbol{a}' = \begin{bmatrix} 1 & 0 \end{bmatrix}$ fås på samma sätt som tidigare uppgifter $z = -\sqrt{3} \approx -1.73 \implies |z| < 1.96$, vi inte kan förkasta nollhypotesen.

4

a)

Vi ser att profilerna ser ganska parallella ut, men de ser inte ut att vara samma och inte heller plana.

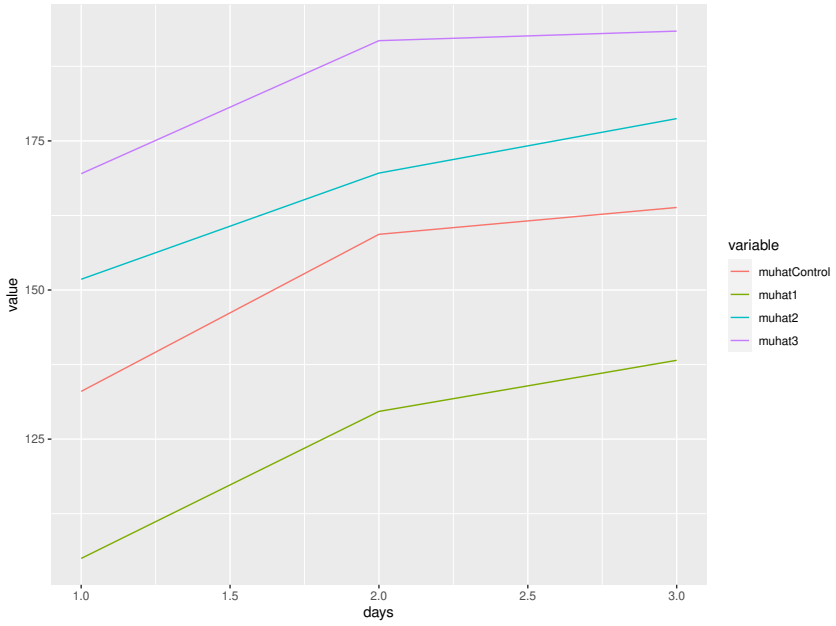

Figure 5: Caption

b)

Vi antar att datan är matrisnormalfördelad $\boldsymbol{X} \sim \mathcal{N}_{p,n}(\boldsymbol{MA}, \boldsymbol{\Sigma}, \boldsymbol{I})$, där

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1, \ldots, \boldsymbol{X}_4 \end{bmatrix}$$

är datan för respektive grupp, är oberoende varandra men med $\boldsymbol{X}_i \sim \mathcal{N}_{p,n_i}(\boldsymbol{\mu}_i \mathbf{1}'_{n_i}, \boldsymbol{\Sigma}, \boldsymbol{I})$, $i = 1, \ldots, 4$ dvs ett beroende mellan variablerna. Där $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\mu}_1 & \ldots & \boldsymbol{\mu}_4 \end{bmatrix}$ och

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{1}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} \\ \mathbf{0}'_{n_1} & \mathbf{1}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{1}'_{n_3} & \mathbf{0}'_{n_4} \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{1}'_{n_4} \end{bmatrix}$$

Vi testar om profilerna är parallella dvs hypotesen

$$H_1: \quad \boldsymbol{\mu}_i - \boldsymbol{\mu}_k = \gamma_i \mathbf{1}, \quad i = 1, \ldots, k-1$$

mot $A \neq H$. Med en kontrastmatris $\boldsymbol{C}$ så att $\boldsymbol{C}\mathbf{1} = \mathbf{0}$ (vi använder $\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$) kan testet ställas upp som

$$H_1: \quad \boldsymbol{C}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k) = \gamma_i \boldsymbol{C}\mathbf{1} = \mathbf{0}, \quad i = 1, \ldots, k-1$$

dvs ett equality of means test efter ett "variabelbyte" $\boldsymbol{X} \mapsto \boldsymbol{CX}$

$$H_1: \quad \boldsymbol{C}\boldsymbol{\mu}_i = \boldsymbol{C}\boldsymbol{\mu}_k, \quad i = 1, \ldots, k-1$$

mot $A \neq H$. Eftersom vi har 4 grupper testar vi med ett LRT, med

$$\Lambda_{H_1} = \frac{\det(\boldsymbol{CVC}')}{\det(\boldsymbol{CVC}' + \boldsymbol{CHC}')}$$

Där within-sum of squares skattas med $\boldsymbol{V} = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{A}'(\boldsymbol{AA}')^{-1}\boldsymbol{A})\boldsymbol{X}$ och between-sum of squares skattas med $\boldsymbol{H} = \boldsymbol{Y\Xi}^{-1}\boldsymbol{Y}'$. Där $\boldsymbol{Y} = \left[\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_4, \ldots, \bar{\boldsymbol{x}}_3 - \bar{\boldsymbol{x}}_4,\right]$ och $\boldsymbol{\Xi}^{-1} = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}\begin{bmatrix} n_1 & n_2 & n_3 \end{bmatrix}$ där $n = n_1 + n_2 + n_3 + n_4$.
Vi använder Box-korrigeringen och får att vår teststorhet som är asymptotiskt χ^2-fördelad är

$$z = (-n - \frac{1}{2}(k+p+1))\ln\Lambda_{H_1} = 2.238067 < 12.59159 = \chi^2(0.95; (p-1)(k-1))$$

Vi kan inte förkasta att de är parallella

c)

Eftersom vi klarade parallell-testet kan vi fortsätta med testet för likhet, vi testar

$$H_2|H_1: \quad \boldsymbol{\mu}_i - \boldsymbol{\mu}_k = \boldsymbol{0}, \quad i = 1, \ldots, k-1$$

Vilket också blir ett equality of means test som vi testar med

$$\Lambda_{H_2|H_1} = \frac{\det(\boldsymbol{CVC}' + \boldsymbol{CHC}')}{\det(\boldsymbol{CVC}')}\frac{\det(\boldsymbol{V})}{\det(\boldsymbol{V} + \boldsymbol{H})}$$

som är stor (nära ett) då "variabelbytet" med kontrastmatrisen inte gör någon skillnad eftersom γ redan var noll. Vi ställer upp vår teststorhet som

$$F = \frac{1 - \Lambda_{H_1|H_1}}{\Lambda_{H_1|H_1}}$$

vi beräknar $F = 0.2181818 < \frac{n-k-p+1}{k-1}F_{k-1,n-k-p+1}(0.95) = 0.2188514$ så vi kan (precis) inte förkasta nollhypotesen att profilerna sammanfaller.

d)

På samma sätt, eftersom vi klarade parallell-testet kan vi fortsätta med testet för planhet. Vi ställer upp nollhypotesen som att de inte är plana

$$H_3|H_1: \quad \frac{1}{n}\sum_{n=1}^{k} n_i\boldsymbol{\mu}_i = \gamma_k\boldsymbol{1}_p$$

mot $A \neq H_3|H_1$. Vi bildar teststorheten som

$$z = n\bar{\boldsymbol{x}}'\boldsymbol{C}'(\boldsymbol{CVC}' + \boldsymbol{CHC}')^{-1}\boldsymbol{C}\bar{\boldsymbol{x}}$$

där $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{\boldsymbol{x}}_i$. Vi beräknar och får $z = 1.579214 > 0.1495107 = \frac{p-1}{n-p+1}F_{p-1,n-p+1}(0.95)$ så vi kan förkasta nollhypotesen att de inte är plana.

5

a)

Vi ser att vinden verkar vara mer eller mindre okorrelered med de andra variablerna, och att solstrålning är förhållandevis (jämfört med de andra) starkt positivt korrelerad med ozon (O3) vilket är rimligt eftersom solstrålning skapar ozon från syremolekyler i atmosfären. CO verkar vara positivt korrelerad med NO, NO2 och ozon och NO och NO2 positivt korrelerade med varandra. NO2 verkar också vara positivt korrelerad med alla andra gaser, kanske mest med ozon. Vilket är rimligt eftersom alla gaser uppstår vid motortrafik

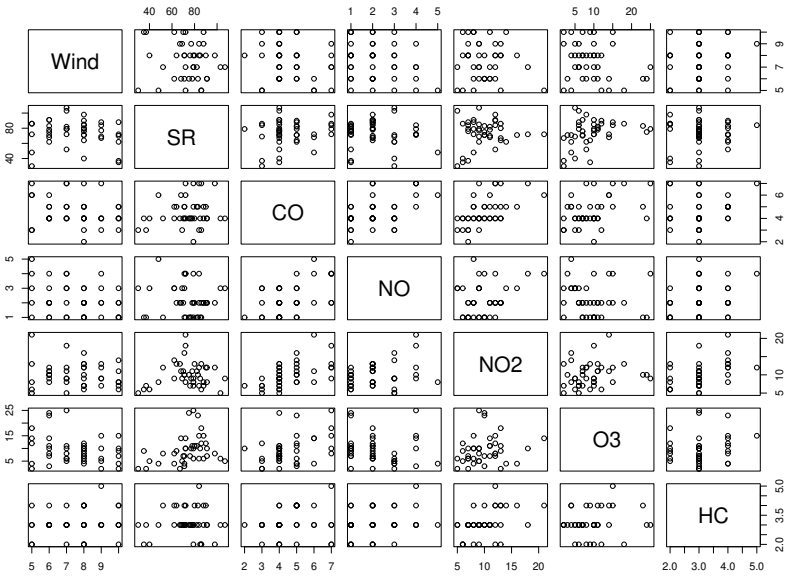

Figure 6: Caption

b)

Vi får korrelationsmatrisskattningen

$$\boldsymbol{R} = \begin{bmatrix} 1. & -0.10 & -0.19 & -0.27 & -0.11 & -0.25 & 0.16 \\ -0.10 & 1. & 0.18 & -0.07 & 0.12 & 0.32 & 0.05 \\ -0.19 & 0.18 & 1. & 0.50 & 0.56 & 0.41 & 0.17 \\ -0.27 & -0.07 & 0.50 & 1. & 0.30 & -0.13 & 0.23 \\ -0.11 & 0.12 & 0.56 & 0.30 & 1. & 0.17 & 0.45 \\ -0.25 & 0.32 & 0.41 & -0.13 & 0.17 & 1. & 0.15 \\ 0.16 & 0.05 & 0.17 & 0.23 & 0.45 & 0.15 & 1. \end{bmatrix}$$

vilket verkar överensstämma med scatterplottarna, att vinden är lite negativt korrelerad med gaserna, solstrålning mest (positivt) korrelerad med ozon, och gaserna är positivt korrelerade med varandra (kanske pga gemensam orsak, dvs biltrafik).

c)

Med responsvariabeln $y_1 = \text{NO2}$ och förklaringsvariablerna $x_1 = \text{vind}$, $x_2 = \text{sol-}$strålning ser vi först att förklaringsvariablerna inte verkar vara linjärt beroende (korrelation -0.1) och att båda förklaringsvariablerna har en viss korrelation med responsvariablen (-0.11 respektive 0.11) så det är rimligt att ansätta en regressionsmodell.

Vi ansätter den linjära modellen

$$\boldsymbol{y}_1 = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n,1} \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{bmatrix} \begin{bmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{21} \end{bmatrix} + \boldsymbol{\epsilon}$$

och antar att vår data är normalfördelad $\boldsymbol{y}_1 \sim \mathcal{N}_{n_1}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$och skattar $\boldsymbol{\beta}$ med MLE-skattningen $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_1$. Vi skattar residualerna med $\boldsymbol{y}_1 - \hat{\boldsymbol{y}}_1 = \hat{\boldsymbol{\epsilon}}$ och får att residualerna inte ser ut att vara normalfördelade genom att studera histogram och QQ-plot

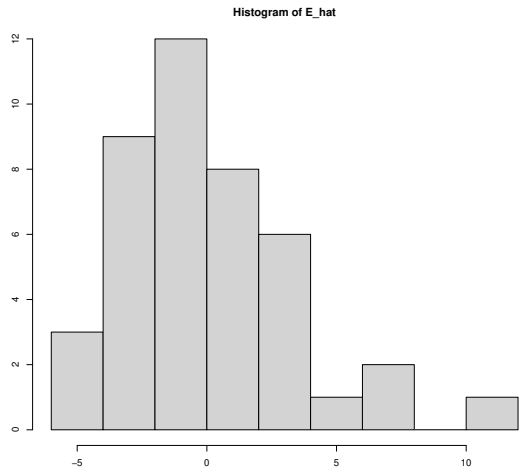

Figure 7: Histogram över skattade residualer med respons y_1

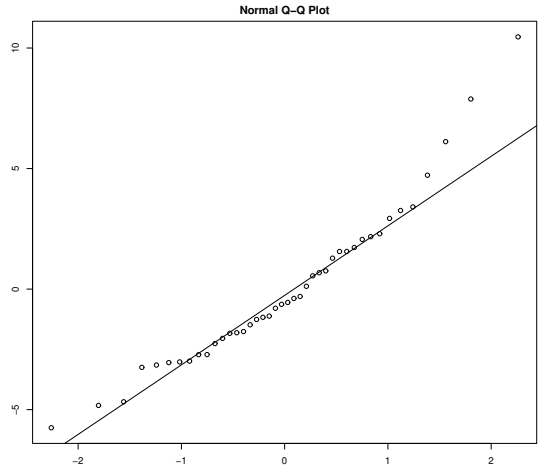

Figure 8: QQ-plot över skattade residualer med respons y_1

Om vi istället transformerar responsvariablen med $\tilde{\boldsymbol{y}}_1 = \ln(\boldsymbol{y}_1)$ får vi plottar som visar på mer normalfördelade residualer. Vi har sen att för den linjära

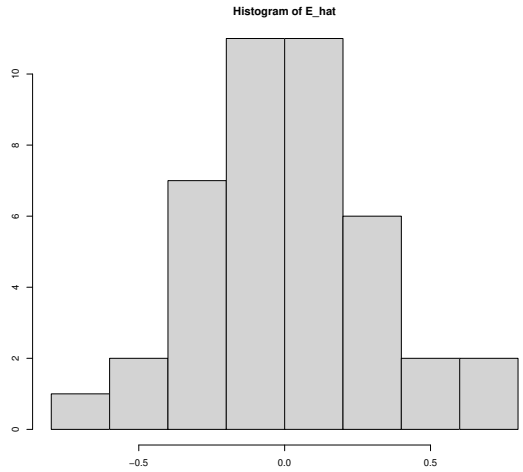

Figure 9: Histogram över skattade residualer med respons $\ln(y_1)$

responsvariablen

$$\boldsymbol{y}_0 - \boldsymbol{x}_0'\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0))$$

vi skattar $s^2 = \frac{1}{n-r}\boldsymbol{y}_0'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y}_0$ där r är antalet oberoende kolumnvektorer i $\boldsymbol{X}$ och får prediktionsintervallet

$$PI_{y_0|x_0} = (\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm t(0.975, n - r - r)s\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0})$$

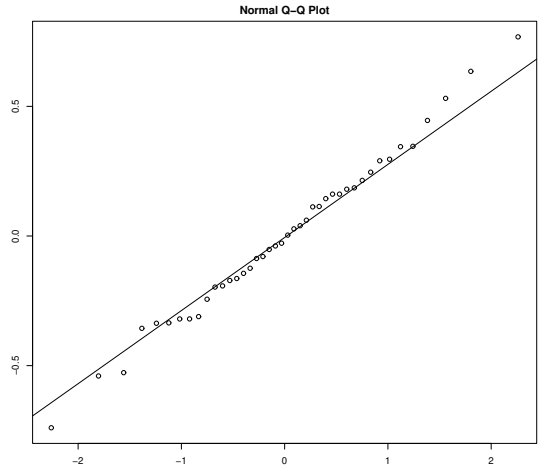

Figure 10: QQ-plt över skattade residualer med respons $\ln(y_1)$

vi beräknar och får $PI_{y_0|x_0} = (2.42112; 16.87013)$

c)

Vi ansätter den linjära modellen

$$\boldsymbol{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} \\ y_{2,1} & y_{2,2} \\ \vdots & \vdots \\ y_{n,1} & y_{n,1} \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} + \boldsymbol{E}$$

och antar att

$$\boldsymbol{Y} \sim \mathcal{N}_{n,2}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$$

och skattar $\boldsymbol{\beta}$ med MLE-skattningen $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$

Vi får att residualerna inte ser ut att vara univariat normalfördelade genom att studera histogram och QQ-plot och scatterplotten ser inte heller ut att vara ellipsoidisk, däremot om vi logaritmerar responsvariablerna först får vi likt i c) mer normalfördelade residualer.

Vi har att $\boldsymbol{Y}_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \sim \mathcal{N}_2(\boldsymbol{0}, (1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)\boldsymbol{\Sigma})$ och ställer upp prediktionsellipsen som

$$(\boldsymbol{Y}_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}})'\boldsymbol{S}^{-1}(\boldsymbol{Y}_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) \leq (1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)\frac{p(n-r-1)}{n-r-p}F_{p,n-r-p}(0.95)$$

Vi ser att prediktionsintervallet från c) överensstämmer med värdena för y_1 i prediktionsellipsen.

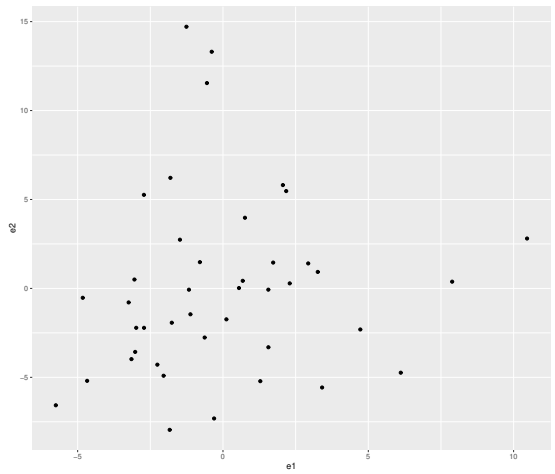

Figure 11: Scatterplots över residualer med responser y_1 och y_2

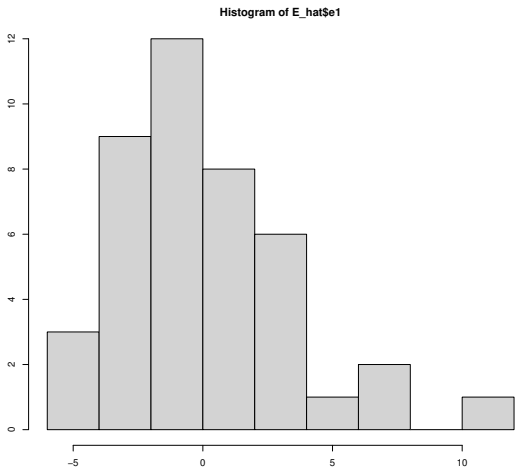

Figure 12: Residualer för y_1

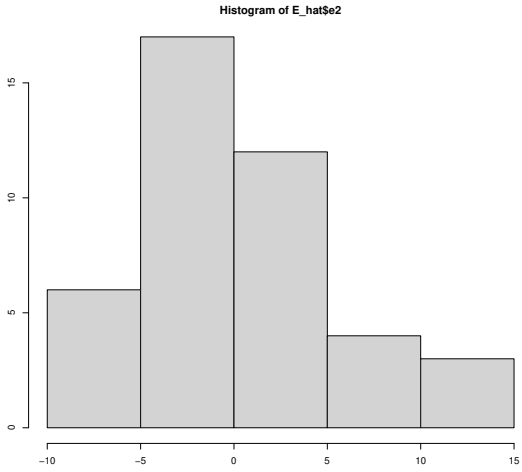

Figure 13: Residualer för y_2

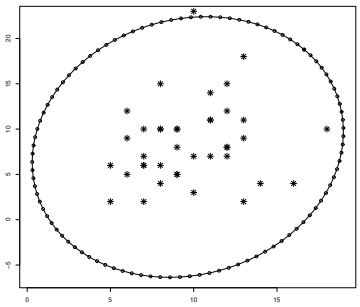

Figure 14: Prediktionsellips

6

a)

Vi ansätter den multivariata linjära modellen

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}$$

Vi har designmatrisen

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{25,1} & x_{25,2} \end{bmatrix}$$

och

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{05} \\ \beta_{11} & \beta_{12} & \dots & \beta_{15} \\ \beta_{21} & \beta_{22} & \dots & \beta_{25} \end{bmatrix}$$

Vi antar att $\boldsymbol{Y} \sim \mathcal{N}_{n,p}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}, \boldsymbol{\Sigma})$ så att vi kan skatta $\boldsymbol{\beta}$ med MLE-skattningen

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

b)

Vi skattar korrelationsmatrisen och ser att dosen koppar är starkt positivt korrelerad med andelen döda vid samtliga tidpunkter, och att $y_1, \dots, y_5$ alla är starkt positivt korrelerade med varandra vilket var väntat. Medelvikten verkar inte ha en särskilt hög korrelation med varken andelen döda eller dosen koppar.

c)

Vi ska undersöka om medelvikten dvs x_2 bidrar till förklaringsgraden dvs testa om koefficienterna framför är noll mot att de är nollskilda, vi får nollhypotesen

$$H: \quad \beta_{21} = \dots = \beta_{25} = 0$$

mot $A \neq H$ vi skriver testet på formen

$$H: \quad \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$$

med

$$\boldsymbol{C} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

Vi gör ett LRT med

$$\Lambda = \frac{\det(\boldsymbol{V})}{\det(\boldsymbol{V} + \boldsymbol{W})}$$

där

$$\boldsymbol{V} = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1}\boldsymbol{C})\boldsymbol{Y}$$

är within-sum of squares och

$$\boldsymbol{W} = \hat{\boldsymbol{\beta}}'\boldsymbol{C}'(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}')^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}$$

är between-sum of squares. Vi förkastar då p-värdet från

$$P(-(n-q-\frac{1}{2}(p-m+1))\ln\Lambda \geq z) \approx P(\chi^2_f \geq z) + \frac{\gamma_2}{\nu^2}(P(\chi^2_{f+4} \geq z) - P(\chi^2_f \geq z))$$

är mindre än 0.05, vi beräknar och får p-värdet $0.1414228 > 0.05$ så vi kan inte förkasta nollhypotesen att medelvikten inte bidrar med till regressionen.

7

a)

Vi har $n_i, i = 1, \ldots, 4$ univariata observationer från 4 grupper där vi antar att observationerna från varje grupp och mellan grupperna är oberoende och att observationerna från grupp i är univariat normalfördelade $x_i \sim \mathcal{N}(\mu_i, \sigma^2)$. Vi testar

$$H: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4$$

mot $A \neq H$, Vi tar fram alla skattningar som vanligt för och ställer upp between-sum of squares

$$SST = \sum_{i=1}^{4} n_i(\bar{x}_i - \bar{x})^2$$

och within-sum of squares

$$SSE = \sum_{i=1}^{4} (n_i - 1)s_i^2$$

och förkastar vår teststorhet $F = \frac{SST/(k-1)}{SSE/(n-k)}$ då den är större än den kritiska gränsen $F_{k-1,n-k}(0.95)$. Vi beräknar och får $F = 7.664023 > 2.832747$ alltså kan vi förkasta nollhypotesen att grupperna har samma väntevärden

b)

Vi ansätter tillväxtmodellen

$$\boldsymbol{Y} = \boldsymbol{ABC} + \boldsymbol{E}$$

Där, med kvadratiska tillväxtkurvor vi får designmatriserna

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \\ 1 & 8 & 64 \\ 1 & 9 & 81 \\ 1 & 10 & 100 \end{bmatrix}$$

och

$$\boldsymbol{C} = \begin{bmatrix} \mathbf{1}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} \\ \mathbf{0}'_{n_1} & \mathbf{1}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{1}'_{n_3} & \mathbf{0}'_{n_4} \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{1}'_{n_4} \end{bmatrix}$$

och den okända

$$\boldsymbol{B} = \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} & \beta_{04} \\ \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \end{bmatrix}$$

Vi antar att matrisnormalfördelad $\boldsymbol{Y} \sim \mathcal{N}_{p,n}(\boldsymbol{ABC}, \boldsymbol{\Sigma}, \boldsymbol{I}_n)$ och skattar $\boldsymbol{B}$ med MLE-skattningen

$$\hat{\boldsymbol{B}} = (\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{Y}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1}$$

där

$$\boldsymbol{V} = \boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1}\boldsymbol{C})\boldsymbol{Y}'$$

Vi ser i plotten att tillväxtkurvorna för vissa grupper inte följer väntevärdena

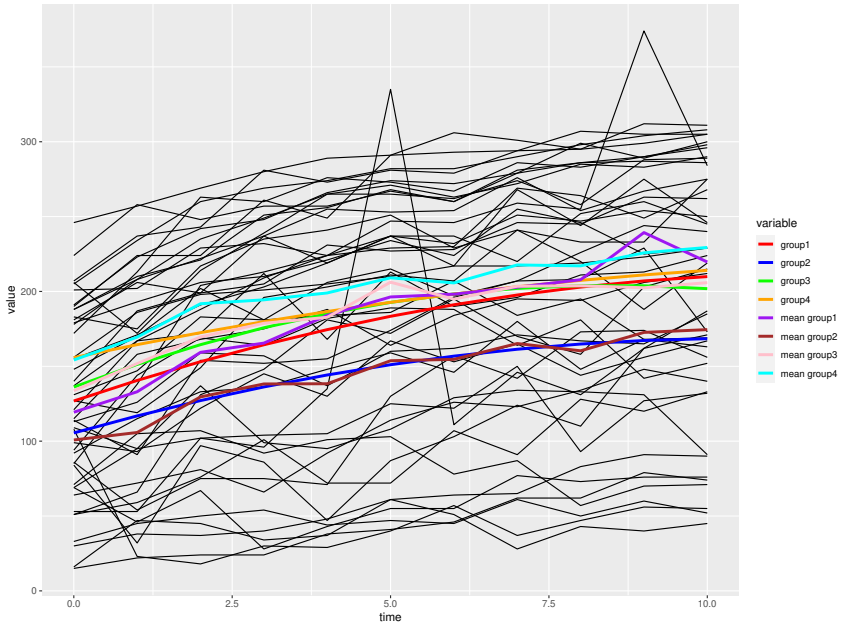

Figure 15: tillväxtkurvor, observationer och väntevärden för de olika grupperna

så bra, antagligen pga outliers bland observationerna för de grupperna.

c)

Vi testar den bilinjära hypotesen

$$H: \quad \boldsymbol{GBH} = \boldsymbol{0}$$

mot $A \neq H$ och specificerar $\boldsymbol{H} = \boldsymbol{I}_k$ där $k = 4$ antalet grupper. Vi sätter

$$\boldsymbol{G} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

som blir $H : \begin{bmatrix} \beta_{21} & \dots & \beta_{24} \end{bmatrix} = \boldsymbol{0}$ dvs $\beta_{21} = \cdots = \beta_{24} = 0$
Vi gör ett LRT med

$$\Lambda = \frac{\det(\boldsymbol{G}(\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A})^{-1}\boldsymbol{G}')}{\det(\boldsymbol{G}(\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A})^{-1}\boldsymbol{G}' + \boldsymbol{G}\hat{\boldsymbol{B}}\boldsymbol{R}^{-1}\hat{\boldsymbol{B}}'\boldsymbol{G}')}$$

där $\boldsymbol{R} = \boldsymbol{C}\boldsymbol{C}^{-1} + \boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{Y}'(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{V}^{-1})\boldsymbol{Y}\boldsymbol{C}'\boldsymbol{C}\boldsymbol{C}^{-1}$
Vi förkastar då

$$z = -(n - k + q - p + \frac{1}{2}(r - t + 1)) \ln \Lambda > \chi^2(0.99; rt)$$

vi beräknar och får $z = 18.12288 > 13.2767$ dvs vi kan förkasta att koefficienterna framför de kvadratiska termerna inte bidrar till modellen, alltså behöver den linjära modellen inte vara bättre.