Statistiske Modeller
Efterår 2024
Eksamen

Navn og Studienumre:

S161770 (Michael Jiaxin Lin Wang)

S162378 (Jonas Johan Kjæreby Kühn)

S162465 (Sharma Harshdeep Kumar)

S160844 (Tommy Nguyen)

S161309 (Patrick Thorsøe Christiansen)

Uddannelsens navn: HA(mat.) Copenhagen Business School

Dato for aflevering af opgaven: 18-12-2024

Gruppenummer: Fri-158077-23

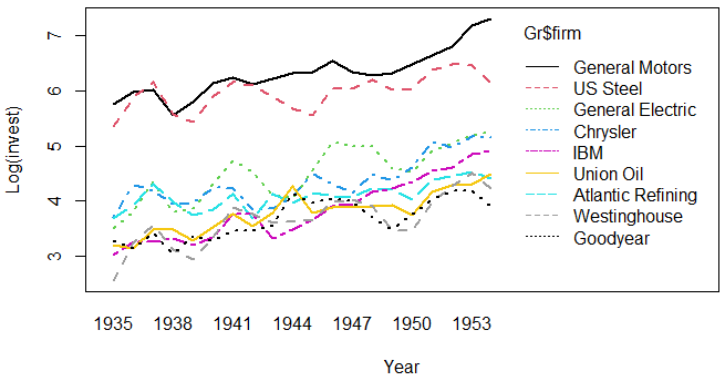
Antal anslag: 31.283 tegn (med mellemrum)

Antal sider: 15

Opgave 1

1.1

Vi benytter interaction.plot() til at tegne de 20 observationer af *log(invest)* for hver virksomhed mod årene. Vi observerer, at alle virksomhederne følger en positiv tilnærmelsesvis lineær tilvækst på *log(invest).* General Motors har generelt haft en mere stabil vækst sammenlignet med de øvrige virksomheder, hvor den har reageret mindre på økonomiske forhold.



1.2

En additiv lineær model er estimeret, hvor "log(invest)" forklares via "year" og "firm" med følgende resultater. Summary-output kan findes på Figur 1.1 under bilag. Alle variable er stærkt signifikante med høje t-værdier. Variablen "*year"* er positiv, hvilket tyder på, at virksomhedernes investeringer stiger årligt (Se også Figur 1.2). Fordi "General Motors" er valgt som corner point, og da den har bedst vækst, er koefficienterne for de øvrige virksomheder negative relativt til "General Motors". Standardafvigelserne er lave, hvilket styrker modellens troværdighed. Modellen er et godt fit, som kan ses på at *justeret-*$R^2 = 94{,}12\ \%$.

Forudsætningerne for modellen er, at residualerne er normalfordelte med samme middelværdi og varians.

Den prædikterede værdi af $\log(invest)$ for Chrysler i 1944 beregnes som:

$$\hat{y} = Intercept + \beta_{year} * year + \beta_{firmChrysler}$$

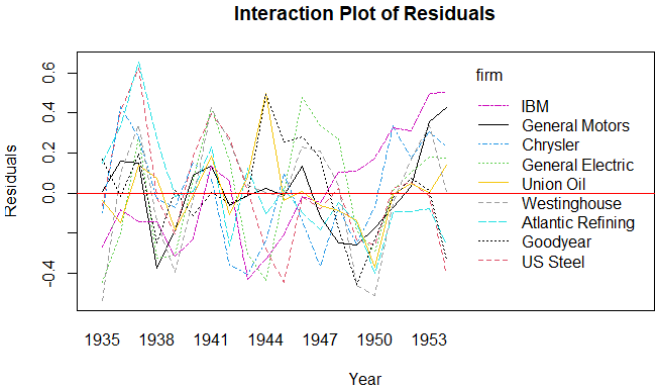
$$\hat{y} = -108{,}8223196 + 0{,}0592108 * 1944 - 1{,}9575647 = 4{,}32591$$

Alternativt kan den prædikterede værdi findes via "*fitted(fit)[70]"*. Se Figur 1.3.

Dette betyder at Chryslers log-investeringer, givet at $year = 1944$, er estimeret til $4{,}32591$. Dette er vores bedste gæt på Chryslers log-investeringer, hvis vi kun har $year = 1944$ og selve linjen til vores rådighed.

1.3

Vi ser at residualerne er ligeligt fordelt omkring førsteaksen, $y = 0$, og der er ingen tydelige systematiske mønstre. Variansen ser ud til at være stabil, dvs. ingen tegn på heteroskedasticitet. Enkelte residualer er dog relativt store, men generelt ligger de tæt omkring nul. Modellen synes derfor at passe godt til dataene ifølge modelantagelse for en lineær model.



1.4

Variablen "*year* " ændres nu fra en lineær variabel til en kategorisk variabel. Vi opstiller en ny lineær model "fitx" og bruger summary til at estimere koefficienterne, så vi kan bestemme den prædikterede værdi for Chrysler i 1944. Se tabellen i Figur 1.4.

Vi har her en model med 152 frihedsgrader og en "Adjusted R-squared" på 0,9553, hvilket er meget højt. Koefficienterne for factor(year) udtrykker, hvor meget mere (eller mindre) log(invest) er for de pågældende år sammenlignet med 1935. Koefficienterne for de forskellige firmaer udtrykker hvor meget log(invest) afviger fra reference-variable (General Motors log(invest)).

Den prædikterede værdi for Chrysler i 1944 udregnes, og vi ved at "year" nu er en kategorisk variabel, kan vi direkte bruge koefficienterne for 1944 uden at skulle gange med 1944.

$$\hat{y} = 5{,}62651 + 0{,}61440 - 1{,}95756 = 4{,}28335$$

Alternativ kan den prædikterede værdi findes gennem fitted(fitx)[70] (se Figur 1.5).

1.5

Vi har nu både opstillet en model, hvor variablen "year" antager talværdier, og en udvidet model, hvor "year" er en kategorisk variabel. Vi vil undersøge om den udvidede model gør en forskel ift. modellens fit ved brug af ANOVA-test (Se Figur 1.6). RSS er 10,6370 for fit-modellen, mens den er 7,2302 for fitx-modellen. En lavere RSS-værdi er bedst, og dermed er fitx-modellen bedre til at forklare dataene end fit-modellen.

Det ses også, at F-teststørrelsen $F = 3{,}9789$ med en p-værdi på $p = 1{,}109e - 06 < 0{,}05$.

Med en lille p-værdi, kan man forkaste nulhypotesen, hvilket vil sige, at det har en påvirkning på modellen, når variablen "year" ændres til en kategorisk variabel.

Alt i alt er fitx-model med "year" som kategorisk variabel altså signifikant bedre end fit-modellen.

Vi tager også ANOVA af hver af de to fittede modeller individuelt, hvor vi undersøger om koefficienterne tilhørende de forklarende variable er lig med 0, altså:

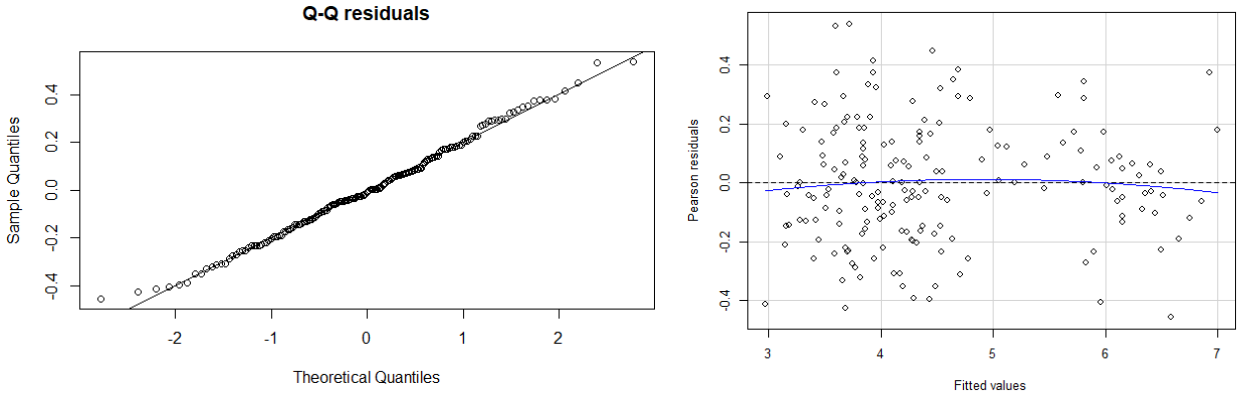
$$fit: H_0: \beta_1 = \beta_2 = 0 \quad og \quad fitx: H_0: \beta_1 = \beta_2 = 0$$

Resultaterne ses i Figur 1.7. P-værdierne for *year* og *firm* er $2{,}2e - 16 < 0{,}05$ for både *fit* og *fitx*.

Dermed kan vi forkaste vores nulhypoteser, hvilket vil sige, koefficienterne β_1 og β_2 i begge modeller må med en stor sikkerhed antage værdier, der er forskellige fra 0. De er individuelt signifikante.

1.6

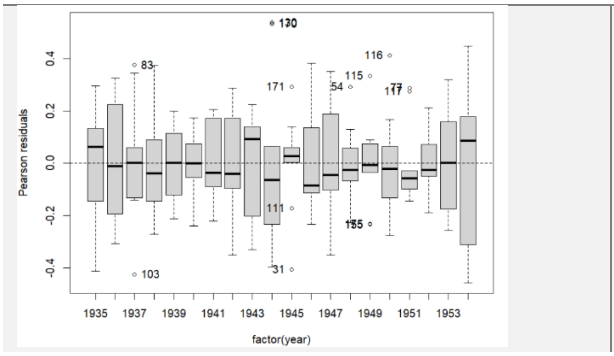
Først tjekkes der, om residualerne er normalfordelte, hvilket kan gøres gennem et QQ-plot af residual fordeling. Derefter undersøges der residualernes varians og middelværdi ved at foretage et residualplot, hvor vi plotter Pearson residualerne mod de fittede værdier, vha. *"> residualPlots(fitx, terms = ~ . - Gr$year)"*:



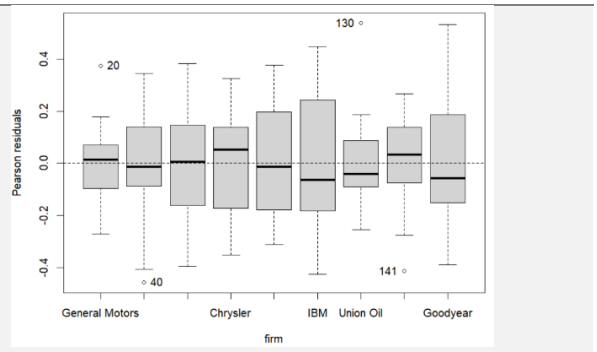
I QQ-plottet ses det, at punkterne ligger tæt langs den rette linje, dvs. residualerne er normalfordelte. I residualplottet ønsker vi, at residualernes middelværdi er "ens", hvilket ses ved, at den blå linje smyger sig langs nul. Derudover skal variansen se nogenlunde ens ud, hvilket vi også vurdere den for at være.

Vi tester nu residualernes varianser for de enkelte uafhængige variable. Vi ser at varianserne ser ens ud og de er uafhængige for begge variable.

RESIDUALER MOD ÅR



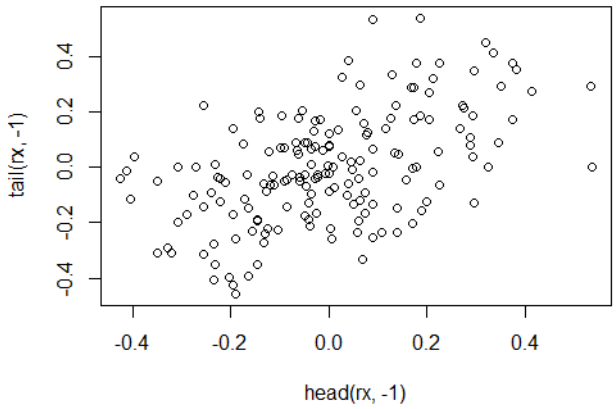
RESIDUALER MOD FIRM



Vi udregner også middelværdien og variansen for residualerne for hver kategori i variablen "firm" og "year", som ses i Figur 1.8 ("year" til venstre og "firm" til højre). Det ses hermed at middelværdien for residualerne for alle kategoriske variable og variansen mellem residualerne er meget ens for alle kategorierne. Dermed er denne modelkrav opfyldt for datasættet. Til sidst så kan vi også se i summary(fitx) at vi har en R-kvadreret værdi på 96,2%, hvilket siger at modellen er meget relevant for datasættet.

1.7

Vi plotter residualerne op mod de foregående residualer for samme firma i *R*. Der ses, at punkterne har en positiv trend, hvilket indikerer at der er en svag seriel korrelation. Hvis punkterne derimod fremstår tilfældige, er der ingen seriel korrelation.



Opgave 2

2.1

Vi skal finde variansmatricen for OLS estimatoren $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$, altså $Var(\hat{\beta}_{OLS})$. Vi ved at variansmatricen for OLS estimatoren er bestemt ved:

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

Kalder vi det Σ, som vi fik givet i opgaven for Σ_{opg}, har vi:

$$Var(\hat{\beta}_{OLS}) = (X'X)^{-1}\,X'\Sigma_{opg}X(X'X)^{-1}$$

Når $\Sigma = \sigma^2 I$, bliver variansmatricen simplere:

$$Var(\beta) = \sigma^2(X'X)^{-1}$$

Da vi har fået designmatricen X og variansmatricen Σ givet, kan vi nu ved brug af *R* opstille variansmatricen $Var(\hat{\beta}_{OLS})$, både for den Σ som vi har fået givet og for $\Sigma = \sigma^2 I$.

Det kan ses i variansmatricerne i Figur 2.1, at varianserne i $Var(\hat{\beta}_{OLS})$ for Σ_{opg} (0,8567 og 0,0211) er større sammenlignet med varianserne i $Var(\hat{\beta}_{OLS})$ for $\Sigma = \sigma^2 I$ (0,4666 og 0,0121). Vi ved at OLS estimatoren er mest effektiv under Gauss-Markov antagelser, herunder homoskedasticitet, dvs. når $\Sigma = \sigma^2 I$. Under Gauss-Markov antagelserne giver OLS estimatoren den mindste varians. Når der er seriel korrelation, er der ikke

længere homoskedasticitet, hvilket betyder at OLS estimatoren ikke længere er den mest effektive estimator. Derfor ses der større varianser under Σ_{opg}.

2.2

Undersøgelse om $\tilde{\sigma}^2$ er central

For at estimatoren $\tilde{\sigma}^2$ skal være central, skal $E(\tilde{\sigma}^2) = \sigma^2$. Vi finder derfor først middelværdien af $\tilde{\sigma}^2$:

$$E(\tilde{\sigma}^2) = E\left(\frac{RSS}{N-2}\right) = \frac{1}{N-2}E(RSS)$$

Vi har at $RSS = Y'(I-P)Y$, hvor $P = X(X'X)^{-1}X'$. Dimensionerne for RSS er $1 \times N \cdot N \times N \cdot N \times 1 = 1 \times 1$. Dermed er RSS en enkelt talværdi. For en skalar c gælder der, at $tr(c) = c$. Da RSS er 1×1, kan vi anvende dette:

$$E(RSS) = E(Y'(I-P)Y) = tr(E(Y'(I-P)Y))$$

Da $E\big(tr(A)\big) = tr\big(E(A)\big)$ har vi $E(RSS) = E\big(tr(Y'(I-P)Y)\big)$. Der gælder også $tr(ABC) = tr(CAB) = tr(BCA)$ – altså er spor invariant under cykliske ombytninger. Derfor har vi:

$$E(RSS) = E\left(tr\big((I-P)YY'\big)\right) = tr(E((I-P)YY')) = tr((I-P)E[YY']) = tr((I-P)\Sigma)$$

Vi skal nu undersøge om $E(\tilde{\sigma}^2) = E(\frac{RSS}{N-2}) = \frac{tr((I-P)\Sigma)}{N-2}$ er centralt.

Det der afgør om $\frac{tr((I-P)\Sigma)}{N-2}$ er central eller ej, er hvordan Σ ser ud. Hvis $\Sigma = \sigma^2 I$ har vi:

$$E(\tilde{\sigma}^2) = \frac{tr((I-P)\sigma^2 I)}{N-2} = \frac{\sigma^2 tr((I-P)I)}{N-2} = \frac{\sigma^2 tr(I-P)}{N-2} = \frac{\sigma^2(tr(I)-tr(P))}{N-2}$$

Vi ved fra forelæsningen at $dim\ L = tr(P)$. Da vores lineære model er $Y_t = \beta_0 + \beta_1 t$ må $tr(P) = 2 = dim\ L$. Da $tr(I) = N$, giver det os: $E(\tilde{\sigma}^2) = \frac{\sigma^2(N-2)}{N-2} = \sigma^2$

Dermed er estimatoren for variansen central, hvis $\Sigma = \sigma^2 I$. Dermed er $\tilde{\sigma}^2 = \frac{RSS}{N-2}$ ikke en central estimator, hvis $\Sigma \neq \sigma^2 I$. Så med Σ_{opg} er $\tilde{\sigma}^2$ ikke en central estimator.

Det generelle tilfælde

I det generelle tilfælde hvor vi ikke nødvendigvis har $E(Y) = 0$, dvs. $E(YY') \neq \Sigma$ har vi i stedet:

$$YY' = (X\beta + \varepsilon)(X\beta + \varepsilon)' = (X\beta + \varepsilon)(\varepsilon' + (X\beta)') = (X\beta + \varepsilon)(\varepsilon' + \beta'X')$$
$$= X\beta\varepsilon + X\beta\beta'X' + \varepsilon\varepsilon' + \varepsilon\beta'X'$$

Vi tager middelværdien af dette:

$$E(YY') = E(X\beta\varepsilon) + E(X\beta\beta'X') + E(\varepsilon\varepsilon') + (\varepsilon\beta'X')$$

Vi antager at $\varepsilon_i \sim N(0, \sigma^2)$, dvs. ε-vektoren har fordelingen $\varepsilon \sim N(0, \Sigma)$. Vi fik jo oplyst, at $E(YY') = \Sigma$ når $E(Y) = X\beta = 0$, hvor $Y \sim N(X\beta, \Sigma)$ – det vil sige når $Y \sim N(0, \Sigma)$, så er $E(YY') = \Sigma$. Så da $\varepsilon \sim N(0, \Sigma)$ må $E(\varepsilon\varepsilon') = \Sigma$. De andre led, som ε indgår i, bliver bare 0, da $E(\varepsilon) = 0$. Da X og β ikke er stokastiske, har vi:

$$E(YY') = E(X\beta\beta'X') + \Sigma = X\beta\beta'X' + \Sigma$$

I udregningen af E(RSS) fås:

$$E(RSS) = tr((I-P)E[YY']) = tr((I-P)(X\beta\beta'X' + \Sigma)) = tr((I-P)X\beta\beta'X' + (I-P)\Sigma)$$

Da $I - P = P_2 \in L_2 = L^{\perp}$ og $X \in spanX = L$ er $(I-P)X = 0$, idet de er ortogonale. Derfor er $((I-P)X)\beta\beta'X' = 0$, hvilket samlet set giver: $E(RSS) = tr(0 + (I-P)\Sigma) = tr((I-P)\Sigma)$

Altså ses det, at det samme også gælder i det generelle tilfælde, som når $E(Y) = 0$.

2.3

Der antages at $\Sigma \neq \sigma^2 I$. Der anvendes derfor i stedet generalized least squares, der er givet ved:

$$\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

Vi ved at variansen af en generalized least square er bestemt ud fra formlen:

$$Var(\hat{\beta}_{GLS}) = (X'\Sigma^{-1}X)^{-1}$$

Vi beregner $Var(\hat{\beta}_{GLS})$ med $\Sigma = \Sigma_{opg}$ (Se variansmatricen i Figur 2.2). Når vi sammenligner resultaterne med variansmatricerne for OLS estimatoren, $Var(\hat{\beta}_{OLS})$, ser vi at resultaterne er en smule mindre end varianserne i $Var(\hat{\beta}_{OLS})$ for Σ_{opg}. Varianserne er 0,81566 og 0,01989 med GLS, mens de er 0,85667 og 0,02105 med OLS. Denne forskel opstår, da OLS estimatoren ikke længere er den mest efficiente estimator (derfor større varians), når Gauss-Markov antagelserne ikke er opfyldt ($\Sigma \neq \sigma^2 I$). Variansmatricen af GLS estimatoren, $Var(\hat{\beta}_{GLS})$, tager højde for at Gauss-Markov antagelserne ikke er opfyldt. Så givet at der er seriel korrelation, som kommer af $\Sigma = \Sigma_{opg}$, er GLS estimatoren mere efficient end OLS estimatoren. Derfor vil variansmatricen af GLS estimatoren, $Var(\hat{\beta}_{GLS})$, have lavere varianser.

2.4

For at finde den betinget fordeling af $Y_2 = [Y_{10}]$ givet $Y_1 = [Y_1, \dots Y_9]^T$, antages det først at $EY = 0$, så Y kan findes fra normalfordelingen $Y \sim N(0, \Sigma)$. Vi kan nu finde komponenterne "den betinget middelværdi" og "den betinget varians", for at bestemme den betinget fordeling. Der gælder at hvis $[Y_1 \ \ Y_2]^T \sim N([\mu_1 \ \mu_2]^T, [(\Sigma_{11} \ \Sigma_{12}), (\Sigma_{21} \ \Sigma_{22})])$ og Σ_{22} er regulær, er den betingede fordeling af Y_2 givet Y_1 en normalfordeling med:

$$E(Y_2|Y_1) = \mu_2 + \Sigma_{21}{\Sigma_{11}}^{-1}(Y_1 - \mu_1) \quad og \quad Var(Y_2|Y_1) = \Sigma_{22} - \Sigma_{21}{\Sigma^{-1}}_{11}\Sigma_{12}$$

Da $Y_2 = [Y_{10}]$ og $Y_1 = [Y_1, \dots, Y_9]$ med $\mu_1 = \mu_2 = 0$ har vi:

$$E(Y_{10}|Y_1 \dots Y_9) = \Sigma_{21}{\Sigma_{11}}^{-1}Y_1 \quad og \quad Var(Y_{10}|Y_1 \dots Y_9) = \Sigma_{22} - \Sigma_{21}{\Sigma^{-1}}_{11}\Sigma_{12}$$

Vi får hermed følgende resultater (se også koden i Figur 2.3Figur 2.1 i bilag):

$$E(Y_{10}|Y_1 \dots Y_9) = 0{,}1020962 \quad og \quad Var(Y_{10}|Y_1 \dots Y_9) = 0{,}8236$$

Vi får altså at den betinget fordeling af Y_2 givet Y_1, følger en normalfordeling $N(0.2217, 0.8236)$. Kører man koden flere gange, bemærkes man, at den betingede middelværdi ændrer sig hver gang men ikke variansen. Den betingede middelværdi af Y_{10} givet $Y_1, \dots, Y_9$ ændrer sig hver gang, fordi den er afhængig af de observerede værdier i $\boldsymbol{Y_1}$. Vi udregner $E(Y_{10}|Y_1 \dots Y_9) = \Sigma_{21}\Sigma_{11}^{-1}Y_1$ og får $E(Y_2|Y_1) = E(Y_{10}|Y_1, \dots, Y_9) = (0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0{,}42)^T Y_1 = 0{,}42$ (Se Figur 2.4). Den betingede middelværdi af Y_{10} afhænger kun af Y_9.

Opgave 3

3.1

Vi definerer en udvidet funktion *expand_fitx*, hvor vi tilføjer leddene *log(capital)* og *log(value)*. Vi udfører summary for både *fitx* og *expand_fitx*. For at teste for modelreduktion, skal vi teste hypoteserne:

$$H_0: E(Y) = X_R\beta_R \quad mod \quad H_A: E(Y) = X\beta$$

Hvor X og X_R er designmatricen for henholdsvis den fulde model og den reducerede model. I vores tilfælde er *expand_fitx* den fulde model, mens *fitx* er den reducerede model. Vi skal så finde F-teststørrelsen:

$$F = \frac{\frac{(RSS_R - RSS)}{df_R - df}}{\frac{RSS}{df}} \sim F(df_R - df, df)$$

Hvor $RSS = \left|\left|Y - X\hat{\beta}\right|\right|^2 = \sum(y_i - \hat{\mu}_i)^2$ og $RSS_R = \left|\left|Y - X_R\hat{\beta}_R\right|\right|^2 > RSS$. Dermed fås en F-teststørrelse på $F = 12{,}273$ med en p-værdi på $p = 1{,}158 \cdot 10^{-5} < 0{,}05$ (Se Figur 3.1). Det vil sige, at variablene $\log(value)$ og $\log(capital)$ er statistisk signifikante og bør derfor inkluderes i modellen. Vi finder så et 95% konfidensinterval for *log(value)*. Konfidensintervallet for *log(value)* lig med $[0{,}2013;\ 0{,}5236]$. Vi ser, at 0 ikke er i intervallet, dvs. endnu engang kan vi konkludere, at *log(value)* er statistisk signifikant.

3.2

Vi inkluderer nu *firm*(log(value)+log(capital))* i modellen, som vi opstillede i 1), hvor vi herefter bruger *drop1-funktionen* til at undersøge for vekselvirkning. Vi har modellen:

$\log(invest) = \beta_0 + \beta_1 year + \beta_2 firm + \beta_3 \log(value) + \beta_4 \log(capital) + \beta_5(firm \cdot \log(value)) + \beta_6(firm \cdot \log(capital))$

Outputtet viser signifikansen af termerne ved at fjerne dem enkeltvis fra modellen (Se Figur 3.2Figur 3.1).

- *"**Firm:log(value)**"* er interaktionstermen $\beta_5(firm \cdot \log(value))$ og har $F = 1{,}7192$ samt $p = 0{,}0994 > 0{,}05$. Denne term er dermed ikke statistisk signifikant, hvilket betyder at der ikke er evidens for vekselvirkning mellem *firm* og *log(value)*.
- *"**Firm:log(capital)**"* er interaktionstermen $\beta_6(firm \cdot \log(capital))$ og har $F = 2{,}8150$ samt $p = 0{,}0065 < 0{,}05$. Det vil sige, at denne term er statistisk signifikant, og der er altså evidens for vekselvirkning mellem *firm* og *log(capital)*.

Vekselvirkningen for interaktionen *Firm:log(capital)* betyder, at udover den gennemsnitlige effekt, som kapital har på investering, så er der også en ekstra effekt afhængig af firma. Kapitalens indflydelse på investeringer er dermed ikke den samme for alle firmaer.

3.3

Vi skal finde en transformationsmatrix T_0 til at transformere en korrelationsmatrix, således:

$$T_0 \psi T_0' = I$$

I koden defineres først $\rho = 0{,}5$ og en 20×20 matrix ψ: *rho <- .5; Psi <- diag(20)*

Herefter beregnes elementerne i ψ, hvor $\psi_{ij} = \rho^{|i-j|} = 0{,}5^{|i-1|}$. Det giver korrelationsmatricen:

$$\psi = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots \\ \rho & 1 & \rho & \cdots \\ \rho^2 & \rho & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

F.eks. er $\psi_{12} = \rho^{|1-2|} = \rho$. Herefter bestemmes Cholesky faktoriseringen af ψ, som er en 20×20 øvre trekantsmatrix C. Med baglæns substitution finder vi så den matrix T^*, der løser $CT^* = I$. Vi finder så T_0 ved at transponere T^*: *T0 <- t(backsolve(chol(Psi),diag(20)))*

Med følgende kode, ses det, at $T_0 \psi T_0' = I$: zapsmall(T0 %*% Psi %*% t(T0))

Til sidst opstilles 180×180 blokdiagonalmatricen T, hvor T_0 gentages 9 gange langs diagonalen:

T <- kronecker(diag(9),T0)

Vi opstiller følgende to lineære modeller:

$$\log(invest) = year + firm + \log(value) + \log(capital) + firm(\log(value) + \log(capital))$$

$$\log(invest) = year + firm + \log(value) + \log(capital) + firm \cdot \log(value)$$

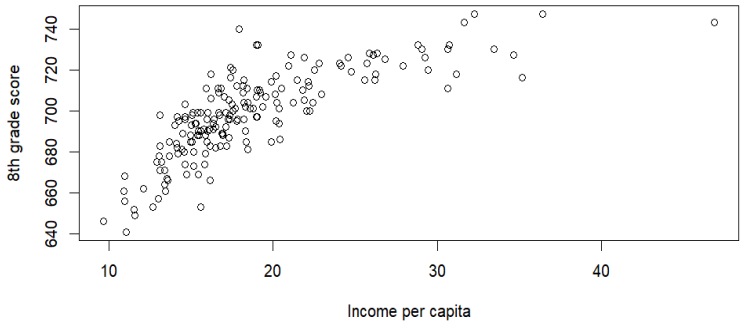
Vi kalder modellerne for *serialfitvv* (med vekselvirkning) og *serialfit* (uden vekselvirkning). Designmatricen X bestemmes for begge modeller, hvor ANOVA bruges til at teste H_0: $ingen\ vekselvirkning$.

Ved transformation af designmatricen, bliver første søjlevektor, der normalt består af 1'ere ændret, så søjlen nu består af 1 og 0,577 (Figur 3.3), altså har vi et ikke-konstant intercept. Dette giver problemer med fortolkning af interceptet og koefficienterne. Derfor skal den fjernes.
Vi tester H_0: $ingen\ forskel\ mellem\ modellerne$ $(\beta_{fitvv} = \beta_{fit})$ (Figur 3.4). Med $F = 1{,}3678$ og $p = 0{,}2162 > 0{,}05$ findes ingen evidens for vekselvirkning mellem *firm* og *log(capital)*, hvilket er modsat resultatet fra før. Den generelle formel for variansen af $\hat{\beta}$ er $Var(\hat{\beta}) = (X'\Sigma^{-1}X)^{-1}$. I 3.2 anvendte vi OLS til modelfit, dvs. vi antog $\Sigma = \sigma^2 I$, hvilket fører til, at man undervurderer variansen og dermed standardfejlene. Dette giver problemer, da estimaterne vil fremstå mere præcise og tilsyneladende signifikante. Her i 3.3 tages der højde for korrelation, så vi undgår at undervurdere variansen.

Opgave 4

4.1

Vi plotter *totsc8* mod percap fra datasættet. Figuren viser en positiv, men marginalt aftagende sammenhæng mellem *tostc8* og *percap*, dvs. ikke en lineær sammenhæng. Så der er altså korreleret variable i datasættet. Intuitionen er at jo højere indkomst per indbygger der er i den givet distrikt jo højere gennemsnitlige karakter i 8. klasse har den distrikt, med marginalt faldende sammenhæng.



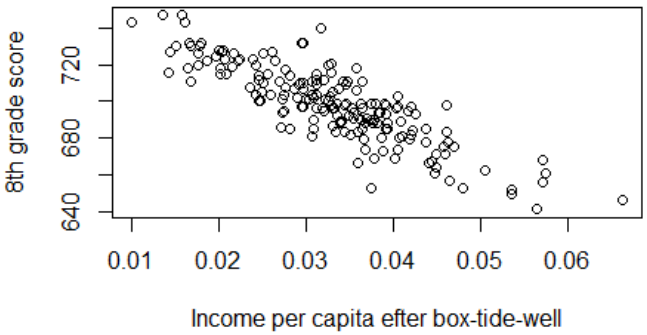
4.2

I Box-Tidwell transformation bestemmes et sæt $\lambda = (\lambda_1, \dots, \lambda_k)$, så $Y_i = \alpha + \beta_1 x_{1k}^{\gamma_1} + \cdots + \beta_k x_{kk}^{\gamma_k}$ er en lineær sammenhæng. Ideen er at vi transformerer vores data, så vi får et lineært sammenhæng, hvilket vi ikke havde før. En af antagelserne ved standard regression er, at X eller transformeret X har et lineært sammenhæng med Y eller transformeret Y. Vi så *totsc8* ikke havde den relation med *percap* og vælger derfor at transformere den. Idet vi kun har *totsc8* mod *percap*, skal der kun bestemmes én γ_1:

$$totsc8 = \alpha + \beta_1 percap^{\gamma_1}$$

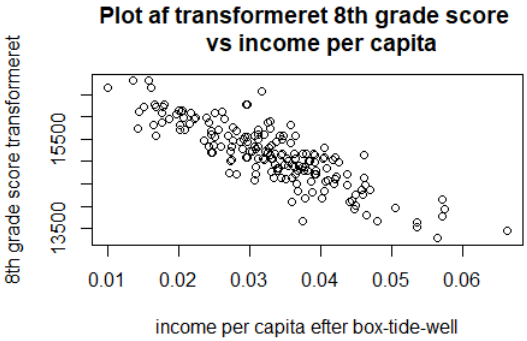
I R fås $\lambda = \lambda_1 = -1{,}1959$ (Figur 4.1), dvs. $totsc8 = \alpha + \beta_1 percap^{-1{,}1959}$. Vi opretter derfor den nye variabel $percap^{-1{,}1959}$ og plotter *totsc8* igen mod den nye variabel.

Med transformationen har vi nu en lineær, men dog negativ, sammenhæng mellem den gennemsnitlige karakter i 8. klasse og percap_boxtidwell.



4.3

Vi laver et Box-Cox transformation og har derfor brug for at transformere Y. Vi bruger transformationen $\frac{Y^{\lambda}-1}{\lambda}$, som for $\lambda \to 0$ går mod $\ln(Y)$. Vi estimerer λ med maximum likelihood. Vi får $\lambda = 1{,}535 > 1$. Vi bruger Box-Cox transformationen på det allerede transformerede data fra 2), så der fås figuren til højre.



Det ses, at plottet er "identisk" med plottet fra spørgsmål 4.2, udover skaleringen af totsc8. Dermed er Box-Cox transformationen ikke nødvendig, idet den ikke forbedrer lineariteten. I virkeligheden er der også allerede skabt en fin lineær sammenhæng mellem de to variable efter første transformation og endnu en transformation er ikke nødvendigt.

4.4

Modellen udvides med de øvrige variable i datasættet på nær de første 3 variable og totsc4:

```
fit <- lm(totsc8 ~ . -code - municipa - district - totsc4, data = MCAS)
```

4.5

Vi kigger på Variansinflation på de indgående variable:

$$Var(\tilde{\beta}_x) = \frac{\sigma^2}{SSD_x} * \underbrace{\frac{1}{1-R^2_{x,Z}}}_{VIF}$$

Hvor $R^2_{x,Z}$ er R^2 for regression af x på øvrige kovariater. Hvis der er en høj R^2, indikerer der en høj korrelation mellem en variabel og de andre, hvilket resulterer en høj varians for den estimatet af x, altså β_x. Dette betyder at, hvis man har en høj *VIF*, så har man en stor usikkerhed på estimat og det er svært at opnå signifikans. Vi kan se på Figur 4.2, at vi har to høje VIF-scores, nemlig for variablerne *regday* (udgift pr. elev, regulær) og *totday* (udgift pr. elev, samlet) med henholdsvis 19,663 og 24,666, mens de øvrige variable har VIF<10, hvilket anses som små. De høje VIF-værdier indikerer multikollinearitet med de øvrige variable og måske hinanden. Det ville give logisk mening, da ændringer i "udgift pr. elev, regulær", vil naturligvis påvirke "udgift pr. elev, samlet".

4.6

Vi kan undersøge indflydelsesrige observationer ved at tjekke følgende målinger; *hat-værdier af leverage, Dffit, DFBETA,* $CovRatio_i$*, Cooks afstand* D_i*:*

$$leverage: h_i = x_i' = (X'X)^{-1}x_i, \quad DFBETAS: \hat{\beta} - \hat{\beta}_{-i} = \frac{(X'X)^{-1}x_i\hat{\varepsilon}_i}{(1-h_i)}, \quad Dffit: \hat{\mu}_i - \hat{\mu}_{-i} = x_i'(\hat{\beta} - \hat{\beta}_{-i})$$

$$CovRatio_i = \frac{1}{(1-h_i)\left(1+\frac{\hat{\varepsilon}_i^2/\tilde{\sigma}_{-i}^2(1-h_i)-1}{n-d}\right)^{k+1}}, \quad Cooks\ afstand: D_i = \frac{\frac{1}{\tilde{\sigma}^2}(\hat{\beta}-\hat{\beta}_{-i})'X'X(\hat{\beta}-\hat{\beta}_{-i})}{d}$$

Observationerne har kritiske værdier (er indflydelsesrige), hvis de opfylder mindst én af følgende kriterier:

Hat-værdier	DFFITS	DFBETAS	CovRatio	Cooks afstand
$> 3\frac{d}{n}$	$> 3\sqrt{d/(n-d)}$	$\lvert\cdot\rvert > 1$	$\lvert CovRatio_i\rvert > \frac{3d}{n-d}$	$>$ medianen i $F(d, n-d)$ fordelingen

Vi opstiller en oversigt over alle observationer, der har mindst én kritisk værdi (Figur 4.3 og Figur 4.4). Det ses at observation 37 er særlig indflydelsesrig ifølge testene: hatværdier af leverage, dfb.blng, dffit, CovRatio og Cooks afstand. At observation 37 er særlig indflydelsesrig, må skyldes antallet af tosprogede elever i Carver-distriktet (295.140), hvor $dfb.\,blng$ = −5,75. Denne værdi er langt mere negativ end de øvrige. Idet DFBETA indgår i dffit og Cooks afstand, forklarer det, hvorfor de også er kritiske. Andre indflydelsesrige værdier er 22, 41, 81, 88, 169 og 208, som ses i Figur 4.3.

Opgave 5

5.1

Der gælder følgende sætning om transformationer af normalfordelte variable:

Antag at $Y \sim N_p(\xi, \Sigma)$ *og lad* $a \in \mathbb{R}^q$ *og* B *være en* $q \times p$*-matrix. Så gælder:*

$$a + BY \sim N_q(a + B\xi, B\Sigma B')$$

Da ses det, at:

$$TY \sim N(TX\beta, T\Sigma T') \quad \Rightarrow \quad TY \sim N(0, T\Sigma T')$$

Vi udtrykker nu likelihoodfunktionen for TY. Likelihoodfunktionen defineres ved $L(\theta; y) = f(y, \theta)$, hvor f er den simultane tæthed for Y. Da er log-likelihoodfunktionen $l(\theta; y) = \log L(\theta; y)$. Den multivariate tæthed for en normalfordeling:

$$f(y) = \frac{1}{\left(\sqrt{2\pi}\right)^n |\Omega|^{\frac{1}{2}}} e^{-\frac{1}{2}(y-\mu)'\Omega^{-1}(y-\mu)} \quad \Rightarrow \quad L = \frac{1}{\left(\sqrt{2\pi}\right)^n |\mathrm{T\Sigma T'}|^{\frac{1}{2}}} e^{-\frac{1}{2}(TY)'(T\Sigma T')^{-1}(TY)}$$

Hvor y er TY, middelværdien $\mu = 0$ og variansmatricen $\Omega = T\Sigma T'$.

$$l = \log\left(\frac{1}{\left(\sqrt{2\pi}\right)^n |\mathrm{T\Sigma T'}|^{\frac{1}{2}}} e^{-\frac{1}{2}(TY)'(T\Sigma T')^{-1}(TY)}\right) = \log\frac{1}{\left(\sqrt{2\pi}\right)^n} + \log\frac{1}{|\mathrm{T\Sigma T'}|^{\frac{1}{2}}} - \frac{1}{2}(TY)'(T\Sigma T')^{-1}(TY)$$

Vi lader $const = \log\frac{1}{\left(\sqrt{2\pi}\right)^n}$. Da $\log\frac{1}{|\mathrm{T\Sigma T'}|^{\frac{1}{2}}} = \log(1) - \frac{1}{2}\log|T\Sigma T'| = -\frac{1}{2}\log|T\Sigma T'|$ fås endelig REML log-likelihoodfunktionen, som er hvad vi skulle vise:

$$l = const - \frac{1}{2}\log|T\Sigma T'| - \frac{1}{2}(TY)'(T\Sigma T')^{-1}TY$$

5.2

Søjlerne i T' udgør en basis for det ortogonale komplement $L^\perp$ til $L = span\ X$ netop fordi $TX = 0$. For ortogonale vektorer gælder der $u'v = 0$. Elementerne 0_{ij} i matricen $\mathbf{0}$ udregnes ved $0_{ij} = \sum_{k=1}^{n} T_{ik}\, X_{jk}$
Vi opskriver hele matrixmultiplikationen $TX = 0$ op på matrixform:

$$TX = 0 \Leftrightarrow \begin{bmatrix} T_{11} & \cdots & T_{1,n} \\ \vdots & \ddots & \vdots \\ T_{n-d,1} & \cdots & T_{n-d,n} \end{bmatrix} \begin{bmatrix} X_{11} & \cdots & X_{1,d} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & T_{n,d} \end{bmatrix} = \begin{bmatrix} 0_{11} & \cdots & 0_{1,d} \\ \vdots & \ddots & \vdots \\ 0_{n-d,1} & \cdots & 0_{n-d,d} \end{bmatrix}$$

F.eks. udregnes det første element i $\mathbf{0}$ udregnes ved $\left(T_{11} \cdots T_{1,n}\right)\left(X_{11} \cdots X_{n,1}\right)' = 0_{11}$. Altså, enhver rækkevektor i T og enhver søjlevektor i Xopfylder $u'v = 0$. Det betyder, at rækkerne i T er basis for $L^\perp = span\ X$. Tilsvarende er søjlevektorerne i T' basis for $L^\perp$, da enhver søjlevektor i X og enhver søjlevektor i T' jo opfylder $u'v = 0$.
Vi lader $L_1 = L = span\ X$ og $L_2 = L^\perp = span\ T'$. Vi ved fra undervisningen, at $P_1 = X(X'X)^{-1}X'$ og $P_2 = I - P_1$, hvor P_1 og P_2 er projektionsmatricen på hhv. L_1 og L_2. Da T' udgør en basis for L_2 må der gælde $P_2 = T'(TT')^{-1}T$. Vi får dermed, hvad vi skulle vise:

$$T'(TT')^{-1}T = I - X(X'X)^{-1}X' \quad \Leftrightarrow \quad X(X'X)^{-1}X' = I - T'(TT')^{-1}T$$

5.3

Vi lader $T^* = \begin{bmatrix} T \\ X'\Sigma^{-1} \end{bmatrix}$ og skal vise, at $T^*\Sigma T^{*\prime} = \begin{bmatrix} T\Sigma T' & 0 \\ 0 & X'\Sigma^{-1}X \end{bmatrix}$. Med $T^{*\prime} = [T' \quad (X'\Sigma^{-1})']$ har vi:

$$T^*\Sigma T^{*\prime} = \begin{bmatrix} T \\ X'\Sigma^{-1} \end{bmatrix} \Sigma [T' \quad (X'\Sigma^{-1})'] = \begin{bmatrix} T\Sigma \\ X' \end{bmatrix} [T' \quad (X'\Sigma^{-1})'] = \begin{bmatrix} T\Sigma \\ X' \end{bmatrix} [T' \quad (\Sigma^{-1})'X] = \begin{bmatrix} T\Sigma \\ X' \end{bmatrix} [T' \quad \Sigma^{-1}X]$$

idet Σ^{-1} er en symmetrisk matrix. Da fås så:

$$T^*\Sigma T^{*\prime} = \begin{bmatrix} T\Sigma \\ X' \end{bmatrix} [T' \quad \Sigma^{-1}X] = \begin{bmatrix} T\Sigma T' & T\Sigma\Sigma^{-1}X \\ X'T' & X'\Sigma^{-1}X \end{bmatrix} = \begin{bmatrix} T\Sigma T' & 0 \\ 0 & X'\Sigma^{-1}X \end{bmatrix}$$

idet $T\Sigma\Sigma^{-1}X = TIX = TX = 0$ og $X'T' = (TX)' = 0$. Det var det første vi skulle vise.

Nu bruges regneregel (4 / 5) til at vise $\left|T^*\Sigma T^{*\prime}\right| = |T\Sigma T'||X'\Sigma^{-1}X|$:

$$\left|T^*\Sigma T^{*\prime}\right| = \begin{vmatrix} T\Sigma T' & 0 \\ 0 & X'\Sigma^{-1}X \end{vmatrix} = |T\Sigma T'||X'\Sigma^{-1}X|$$

Vi bruger regneregel (1) og (2) til at vise $\left|T^*\Sigma T^{*\prime}\right| = |T^*|^2|\Sigma|$:

$$\left|T^*\Sigma T^{*\prime}\right| = |T^*||\Sigma||T^{*\prime}| = |T^*||\Sigma||T^*| = |T^*|^2|\Sigma|$$

Dermed har vi også vist, at determinanten $\left|T^*\Sigma T^{*\prime}\right|$ kan skrives på de to måder.

5.4

Vi ganger $T^* = \begin{bmatrix} T \\ X'\Sigma^{-1} \end{bmatrix}$ med $T^{*\prime} = [T' \quad \Sigma^{-1}X]$:

$$T^*T^{*\prime} = \begin{bmatrix} T \\ X'\Sigma^{-1} \end{bmatrix} [T' \quad \Sigma^{-1}X] = \begin{bmatrix} TT' & T\Sigma^{-1}X \\ X'\Sigma^{-1}T' & X'\Sigma^{-1}\Sigma^{-1}X \end{bmatrix} = \begin{bmatrix} TT' & T\Sigma^{-1}X \\ X'\Sigma^{-1}T' & X'\Sigma^{-2}X \end{bmatrix}$$

Det var det første vi skulle vise. Regnereglen (4) giver:

$$\left|T^*T^{*\prime}\right| = |TT'||X'\Sigma^{-2}X - X'\Sigma^{-1}T'(TT')^{-1}T\Sigma^{-1}X| = |TT'||X'\Sigma^{-1}(\Sigma^{-1}X - T'(TT')^{-1}T\Sigma^{-1}X)|$$

Da der gælder $(A - BA) = (I - B)A$ fås endelig determinanten:

$$\left|T^*T^{*\prime}\right| = |TT'||X'\Sigma^{-1}(I - T'(TT')^{-1}T)\Sigma^{-1}X|$$

Bruges regnereglen (1) og (2) viser vi også, at $\left|T^*T^{*\prime}\right| = |T^*|^2$:

$$\left|T^*T^{*\prime}\right| = |T^*||T^{*\prime}| = |T^*||T^*| = |T^*|^2$$

Dermed har vi vist, hvad vi skulle vise.

5.5

I spørgsmål 3 og 4 viste vi, at $|T^*|^2|\Sigma| = |T\Sigma T'||X'\Sigma^{-1}X|$ og $|T^*|^2 = |TT'||X'\Sigma^{-1}(I - T'(TT')^{-1}T)\Sigma^{-1}X|$, dvs. vi kan opstille følgende:

$$|TT'||X'\Sigma^{-1}(I - T'(TT')^{-1}T)\Sigma^{-1}X||\Sigma| = |T\Sigma T'||X'\Sigma^{-1}X|$$

Vi bruger (3) og får:

$$|T\Sigma T'| = |TT'||\Sigma||X'\Sigma^{-1}(I - T'(TT')^{-1}T)\Sigma^{-1}X||(X'\Sigma^{-1}X)^{-1}|$$

Vi bruger (1) og får:

$$|T\Sigma T'| = |TT'||\Sigma||X'\Sigma^{-1}(I - T'(TT')^{-1}T)\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}|$$

$$= |TT'||\Sigma||X'\Sigma^{-1}(X(X'X)^{-1}X')\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}| = |TT'||\Sigma||X'\Sigma^{-1}X(X'X)^{-1}(X'\Sigma^{-1}X)(X'\Sigma^{-1}X)^{-1}|$$

$$= |TT'||\Sigma||X'\Sigma^{-1}X(X'X)^{-1}| = |TT'||\Sigma||X'\Sigma^{-1}X||(X'X)^{-1}|$$

Endelig fås, hvad vi skulle vise:

$$|T\Sigma T'| = |TT'||\Sigma||X'\Sigma^{-1}X|/|X'X|$$

Opgave 6

6.1

Vi tager logaritmen af $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$: $\log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}\right)$

$$= \log\big(\Gamma(\alpha+\beta)\big) - \log\big(\Gamma(\alpha)\Gamma(\beta)\big) + \log\big(y^{\alpha-1}(1-y)^{\beta-1}\big)$$
$$= \log\Gamma(\alpha+\beta) - (\log\Gamma(\alpha) + \log\Gamma(\beta)) + (\alpha-1)\log y + (\beta-1)\log(1-y)$$
$$= \log\Gamma(\alpha+\beta) - (\log\Gamma(\alpha) + \log\Gamma(\beta)) + \alpha\log y - \log y + \beta\log(1-y) - \log(1-y)$$

Vi lader $\alpha(y) = \begin{pmatrix}\log y \\ \log(1-y)\end{pmatrix} \quad \Rightarrow \quad \alpha\log y + \beta\log(1-y) = \alpha(y)'\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$

Samlet set har vi så:

$$\alpha(y)'\theta = \alpha\log y + \beta\log(1-y)$$
$$c(\theta) = \log\Gamma(\alpha+\beta) - (\log\Gamma(\alpha) + \log\Gamma(\beta))$$
$$d(y) = -(\log y + \log(1-y))$$

$\alpha(y)'\theta$ består af de led, der afhænger af både y og parametrene α og β. $c(\theta)$ består af led, der udelukkende har α og β, og leddene i $d(y)$ afhænger kun af y.

Vi har dermed vist, at $\log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}\right)$ kan skrives på formen $\boldsymbol{\alpha}(y)'\boldsymbol{\theta} + c(\boldsymbol{\theta}) + d(y)$.

Derfor kan $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$ skrives på formen: $e^{\boldsymbol{\alpha}(y)'\boldsymbol{\theta}+c(\boldsymbol{\theta})+d(y)}$

6.2

Vi lader $\phi = \alpha+\beta$, og vi ved, at $Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Vi udtrykker først $\mu(1-\mu)$:

$$\mu = \frac{\alpha}{\alpha+\beta} \;\Rightarrow\; \mu(1-\mu) = \frac{\alpha}{\alpha+\beta} - \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha(\alpha+\beta)-\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2}$$

Dermed kan variansen udtrykkes ved μ og ϕ:

$$Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{\alpha+\beta+1}\frac{\alpha\beta}{(\alpha+\beta)^2} = \frac{1}{1+\phi}\mu(1-\mu)$$

Til den anden del af spørgsmålet, opstiller vi (α,β) som funktion af (μ,ϕ):

$$\mu = \frac{\alpha}{\alpha+\beta} = \frac{\alpha}{\phi} \;\Leftrightarrow\; \alpha = \mu\phi \;\Rightarrow\; \phi = \alpha+\beta = \mu\phi+\beta \;\Leftrightarrow\; \beta = \phi-\mu\phi = \phi(1-\mu)$$

Altså samlet set har vi α og β udtrykt ved μ og ϕ:

$$\alpha = \mu\phi \quad og \quad \beta = \phi(1-\mu)$$

6.3

Vi finder integralet $D(y,\mu)$ manuelt, hvor $v(u) = u(1-u)$. Vi har først og fremmest:

$$D(y,\mu) = 2\int_\mu^y \frac{y-u}{v(u)}du = 2\int_\mu^y \frac{y-u}{u(1-u)}du = 2\int_\mu^y \frac{y-u}{u} + \frac{y-u}{1-u}d = 2\left(\int_\mu^y \frac{y-u}{u}du + \int_\mu^y \frac{y-u}{1-u}du\right)$$

Første integral giver:

$$\int_\mu^y \frac{y-u}{u}du = y\int_\mu^y \frac{1}{u}du - \int_\mu^y 1\,du = y[\log u]_\mu^y - [u]_\mu^y = y(\log y - \log u) - y + \mu$$

Andet integral giver:

$$\int_\mu^y \frac{y-u}{1-u}du \;=\; \int_\mu^y \frac{y}{1-u}du - \int_\mu^y \frac{u}{1-u}du \;=\; y\int_\mu^y \frac{1}{1-u}du - \left(\int_\mu^y \frac{1}{1-u} - \frac{1-u}{1-u}du\right)$$

$$= \; y\int_{\mu}^{y}\frac{1}{1-u}du - \left(\int_{\mu}^{y}\frac{1}{1-u}du - \int_{\mu}^{y}1\,du\right) \; = \; y[-\log(1-u)]_{\mu}^{y} - \left([-\log(1-u)]_{\mu}^{y} - [u]_{\mu}^{y}\right)$$
$$= y(\log(1-\mu) - \log(1-y)) - (\log(1-\mu) - \log(1-y) - y + \mu)$$

Vi kan så samle de to integraler og finde det samlede udtryk for $D(y,\mu)$:

$$\int_{\mu}^{y}\frac{y-u}{u}du + \int_{\mu}^{y}\frac{y-u}{1-u}du$$
$$= (y(\log y - \log\mu) - y + \mu) + \left(y(\log(1-\mu) - \log(1-y)) - (\log(1-\mu) - \log(1-y) - y + \mu)\right)$$
$$= y(\log y - \log\mu) - y + \mu - y(\log(1-y) - \log(1-\mu)) + \log(1-y) - \log(1-\mu) + y - \mu$$
$$= y(\log y - \log\mu) - y(\log(1-y) - \log(1-\mu)) + \log(1-y) - \log(1-\mu)$$
$$= y(\log y - \log\mu) + (1-y)(\log(1-y) - \log(1-\mu))$$

Dermed er:

$$D(y,\mu) = 2\int_{\mu}^{y}\frac{y-u}{u}du + 2\int_{\mu}^{y}\frac{y-u}{1-u}du = 2y(\log y - \log\mu) + 2(1-y)(\log(1-y) - \log(1-\mu))$$
$$= \begin{cases} -2\log(1-\mu)\,, & hvis\ y = 0 \\ -2\log\mu\,, & hvis\ y = 1 \end{cases}$$

Så hvis vi observerer $y = 0$, så er unit deviance $-2\log(1-\mu)$. Observeres $y = 1$, så er unit deviance $-2\log\mu$. Siden vi har, at y kun kan tage værdierne 0 og 1, så må $0 \le \mu \le 1$. Middelværdien μ kan forstås som sandsynligheden for succes ($y = 1$), så fx hvis $\mu = 1$, så er $D(1,1) = 0$. Altså er afstanden mellem den observerede værdi y og middelværdien μ lig med 0 i det tilfælde.

Opgave 7

7.1

Vi fitter intercept-modellen og får følgende summary (Figur 7.1). Vi får interceptet $\beta_0 = -0{,}605877$. Vi ved at glm-funktionen i R bruger logit-link funktion som standard. Her er den lineære prædiktor $\eta = X\beta = \beta_0$. For logit-funktionen gælder $\eta = g(\pi) = \ln\frac{\pi}{1-\pi} = \beta_0$.

Vi vil gerne isolere π, så vi finder $\pi = g^{-1}(\beta_0)$:

$$\ln\frac{\pi}{1-\pi} = \beta_0 \Leftrightarrow \frac{\pi}{1-\pi} = e^{\beta_0} \Leftrightarrow \pi + \pi e^{\beta_0} = e^{\beta_0} \Leftrightarrow \pi = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

I GLM gælder der $\mu = g^{-1}(\eta) = g^{-1}(\beta_0)$, altså er $\mu = \pi$. Derfor er:

$$\mu = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{e^{-0{,}605877}}{1+e^{-0{,}605877}} = 0{,}3530002$$

Altså har vi her sammenhængen mellem middelværdien og interceptet.

Det ses også i outputtet, at dispersionsparameteren for quasibinomial familien er $\psi = 0{,}09116$. Der gælder for eksponentielle dispersionsfamilier:

$$Var(Y) = \psi v(\mu)$$

Vi fik før i 6.3 givet, at $v(u) = u(1-u)$. Den genbruger vi. Dvs. vi har for quasibinomialmodellen:

$$Var(Y) = \psi\mu(1-\mu)$$

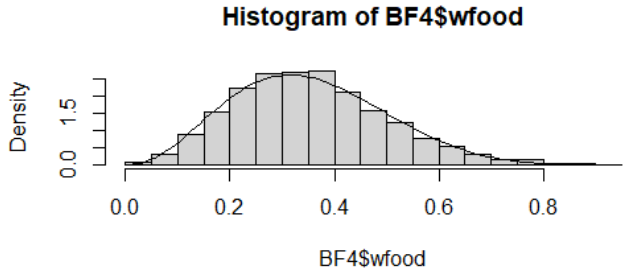
Til sammenligning med betafordelingen i 6.2, havde vi:

$$Var(Y) = \frac{1}{1+\phi}\mu(1-\mu)$$

Da må sammenhængen mellem ϕ og ψ være:

$$\psi = \frac{1}{1+\phi} \quad \Leftrightarrow \quad \phi = \frac{1}{\psi} - 1 = \frac{1}{0{,}09116} - 1 = 9{,}96994$$

Vi udregner et 95% konfidensinterval for interceptet ved brug af først (quasi-)likelihood profiling, som gøres ved direkte at tage confint af GLM-modellen. Konfidensintervallet gennem Wald-test fås gennem confint.default. Vi får: $CI_{GLM} = \{-0{,}6226240;\ -0{,}5891578\}$ og $CI_{Wald} = \{-0{,}6226102;\ -0{,}5891442\}$. Det ses hermed, at de to konfidensintervaller er næsten helt ens, og forskellen ligger nede i de små decimaler. Vi tegner histogrammet for wfood. Vi finder så $\phi = 9{,}96994$ og $\mu = 0{,}3530002$, hvorefter vi så får $\alpha = \mu\phi = 3{,}519391$ og $\beta = \phi(1-\mu) = 6{,}450549$ til betafordelingen:



Det ses tydeligt, at histogrammet følger en betafordeling med $\alpha = 3{,}519391$ og $\beta = 6{,}450549$.

7.2

Da man kan finde parametrene α og β ved at sætte den teoretiske middelværdi og varians med den empiriske middelværdi og varians, har vi:

$$E(Y) = \frac{\alpha}{\alpha+\beta} = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{\mu} \quad og \quad Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = s^2$$

Vi fandt jo før ud af, at når $\phi = \alpha + \beta$, så er $Var(Y) = \frac{1}{1+\phi}\mu(1-\mu)$. Dermed har vi så:

$$\frac{1}{1+\phi}\mu(1-\mu) = s^2 \quad \Leftrightarrow \quad \phi = \frac{\mu(1-\mu)}{s^2} - 1$$

Da kan vi finde α og β:

$$\frac{\alpha}{\alpha+\beta} = \frac{\alpha}{\phi} = \overline{\mu} \quad \Leftrightarrow \quad \alpha = \overline{\mu}\phi = \overline{\mu}\left(\frac{\mu(1-\mu)}{s^2} - 1\right)$$

$$\phi = \alpha + \beta = \overline{\mu}\phi + \beta \quad \Leftrightarrow \quad \beta = \phi(1-\overline{\mu}) = \left(\frac{\mu(1-\mu)}{s^2} - 1\right)(1-\overline{\mu})$$

Vi prøver det af, på det konkrete eksempel. Det ses hermed (Figur 7.2), at ved at bruge momentmetoden, fås de samme α og β som i 7.1. Middelværdien μ, som vi fandt ud fra interceptet i 7.1, er også lig med den empiriske middelværdi $\overline{\mu}$ her. Vi skal bruge deltametoden til at finde standard error for det estimerede intercept, som vi fandt til at være $\beta_0 = -0{,}605877$. Deltametoden i én dimension siger:

$\tau(\hat{\theta}) \sim_{as} N\left(\tau(\theta), \left(\frac{\partial\tau}{\partial\theta}\right)^2 \frac{\sigma_\theta^2}{n}\right)$, hvor σ_θ^2 er variansen for θ. I vores tilfælde er $\theta = \mu$ og $\beta_0 = \tau(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$:

$$\frac{\partial\tau}{\partial\mu} = \frac{\partial}{\partial\mu}\log\left(\frac{\mu}{1-\mu}\right) = \frac{1}{\mu-\mu^2}$$

I én dimension er variansen bare $Var\left(\tau(\hat{\theta})\right) = \left(\frac{\partial\tau}{\partial\theta}\right)^2 \frac{Var(\theta)}{n} \quad \Rightarrow \quad SE\left(\tau(\hat{\theta})\right) = \frac{\partial\tau}{\partial\theta}\frac{SE(\theta)}{\sqrt{n}}$, altså:

$$SE(\beta_0) = \frac{1}{\mu-\mu^2}\frac{SE(\mu)}{\sqrt{n}} = \frac{1}{\mu-\mu^2}\frac{SE(\overline{y})}{\sqrt{n}}$$

Vi udregner $SE(\beta_0)$ manuelt og sammenligner det med *deltaMethod*-funktionen. Som ses i Figur 7.3, får vi $SE(\beta_0) = 0{,}0085374$ med begge metoder. Dette passer også med det output, vi fik i spørgsmål 1.

7.3

Vi bruger *betareg*-funktionen (Resultatet ses også på Figur 7.4). *Betareg* estimerer interceptet til $\beta_0 = -0{,}603420$, hvor $\phi = 9{,}8213$. Til sammenligning med resultatet med quasibinomialmodel, fik vi $\beta_0 = -0{,}605877$ og $\phi = 9{,}96994$. Vi ser, at der er en lille forskel mellem de to resultater, hvilket tyder på, at quasibinomialmodellen var en god "approksimation" til betafordelingen.

7.4

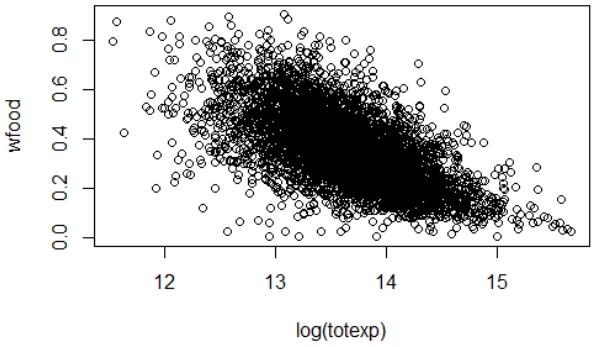
Vi opstiller først en funktion, der beregner den negative log-likelihood. Vi sætter $\mu = 0{,}5$ og $\phi = 1$ – initialværdien for μ og ϕ er tilfældigt sat. Den negative log-likelihood er:

$$-\sum_{i=1}^{n} \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1}\right)$$

Ideen er så, at vi med maksimum likelihood estimation finder den optimale μ og ϕ. Med MLE, får vi altså $\mu = 0{,}3535618$. Vi bruger dette μ og får $\beta_0 = \log\left(\frac{\mu}{1-\mu}\right) = -0{,}6034194$. Vi får også $\phi = 9{,}82134$. Med *Betareg* estimerede vi interceptet til $\beta_0 = -0{,}603420$ og $\phi = 9{,}8213$. Det ses hermed, at de to metoder er ækvivalente med hinanden.

7.5

Vi tegner scatterplottet. Det ses, at der er en negativ sammenhæng mellem "andelen af budgettet, der bruges på mad" og "samlede udgifter". Når samlede udgifter stiger, falder andelen af budgettet, der bruges på nødvendighedsvarer som mad. Dette skyldes nok, at folk med højere samlede udgifter (højst sandsynligt rigere) har mere at bruge på luksusvarer og ikke-essentielle varer, der ikke er mad.



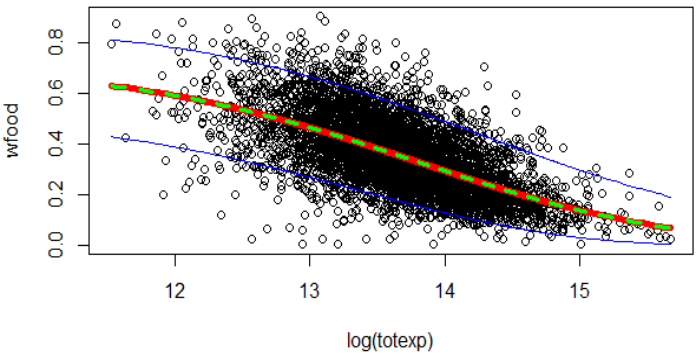
7.6

Vi fitter *wfood* mod 2. grads polynomiet af *log(totexp)* og får følgende resultat (Se også Figur 7.5):
Regressionen bruger en logit-linkfunktion til at modellere den forventede værdi af *wfood*:

$$logit(\mu) = \ln\frac{\mu}{1-\mu} = -0{,}6259 - 30{,}2312 poly_1(\log totexp) - 3{,}7058 poly_2(\log totexp)$$

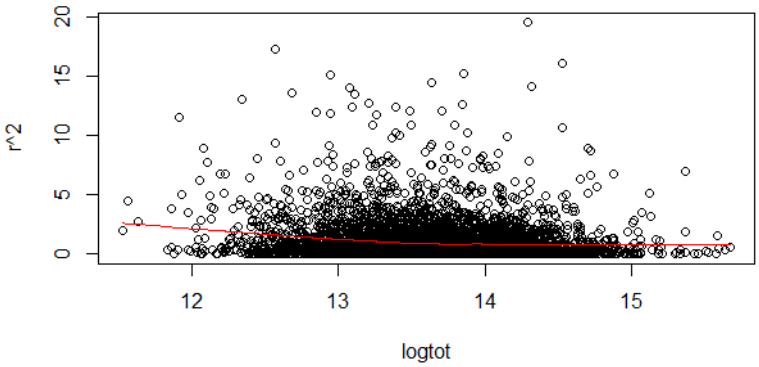
hvor $poly_1(\log totexp)$ er et førstegradspolynomium, og $poly_2(\log totexp)$ er et andengradspolynomium.

Vi tegner herefter middelværdikurven, hvor vi plotter de prædikterede værdier af μ fra modellen mod *log(totexp)*. Vi får så følgende middelværdikurve, som ses i figuren til højre.
"*quantile*"-varianten bruges til at udregne øvre 95% kvantil og nedre 5% kvantil af betafordelingen. Prædiktionsbåndet angiver det interval, hvor vi med 90% sikkerhed kan forvente, at den sande af *wfood* ligger i dette interval.



7.7

Vi benytter koden og tegner følgende figur til højre. Vi har taget Pearson-residualerne af modellen ($\frac{y-\mu}{\sqrt{var(\mu)}}$), hvorefter vi plottede dem imod *log(totexp)*.



Vi bruger LOWESS (Locally Weighted Scatterplot Smoothing) til at tegne, hvordan de kvadrerede Pearson residualer ændrer sig som funktion af *log(totexp)*. LOWESS er en polynomiel regression, som består af en masse lokale polynomielle regressioner. Så den beskriver den generelle trend af r^2 mod *log(totexp)*. Det ses at R^2 har en svag negativ sammenhæng med *log(totexp)*, dvs. man kan ikke betragte præcisionsparameteren ϕ som konstant og uafhængig af *log(totexp)*. Præcisionsparameteren ϕ angiver koncentrationen af observationerne omkring modellen. Da residualerne i "gennemsnit" er større i starten, må observationerne være mere spredt omkring modellen i starten, dvs. ϕ er ikke konstant.

7.8

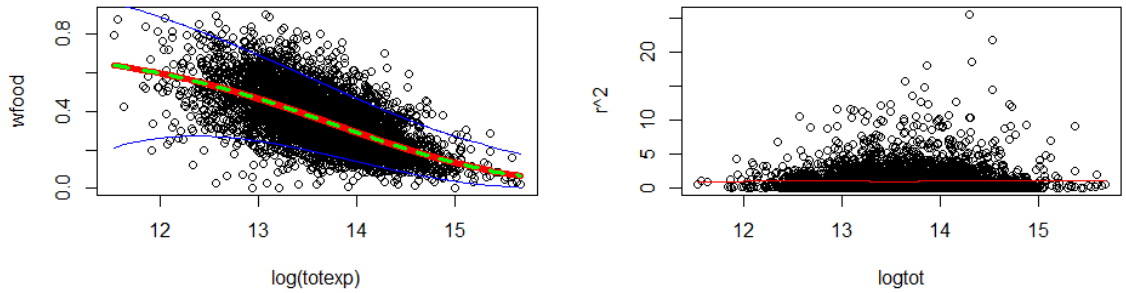
Vi udvider modellen fra 7.6, hvor vi tilføjer "|poly(log(totexp),2)". Vi får her koefficienterne for $logit(\mu)$ og $logit(\phi)$ – se Figur 7.6. Koefficienterne for $logit(\mu)$ er stort det samme som dem i spørgsmål 7.6:

$$logit(\mu) = \ln\frac{\mu}{1-\mu} = -0{,}6327 - 30{,}9664poly_1(\log totexp) - 3{,}6899poly_2(\log totexp)$$

Vi opstiller også modellen for $logit(\phi)$:

$$logit(\phi) = \ln\frac{\phi}{1-\phi} = 2{,}8150 + 15{,}5949poly_1(\log totexp) - 6{,}3262poly_2(\log totexp)$$

Vi tegner herefter den fittede kurve på scatterplottet samt 90% prædiktionsinterval, samt et plot over de kvadrerede residualer mod *log(totexp)*:



Som ses, har middelværdikurven for *wfood* stort set ikke ændret sig, men prædiktionsintervallet har ændret sig markant. Dette skyldes, at vi i modellen nu har tilføjet en variansmodel, der tillader præcisionsparameteren ϕ at variere for forskellige *log(totexp)*, så den altid tilpasses til hver *log(totexp)*. Prædiktionsintervallerne spreder sig ud i starten, fordi observationerne for *wfood* ligger mere spredt i starten. Modellen fanger nu denne variation bedre. Betragter vi de kvadrerede Pearson residualer, ses det, at LOWESS kurven nu er helt flad – det er også tydeligt at r^2 nu ligger mindre spredt ud sammenlignet med før. Efter at have korrigeret for forskellen i spredningen af observationerne for forskellige *log(totexp)*, er den generelle trend for r^2 nu konstant overfor *log(totexp)*, dvs. ϕ er konstant overfor *log(totexp)*.

Bilag

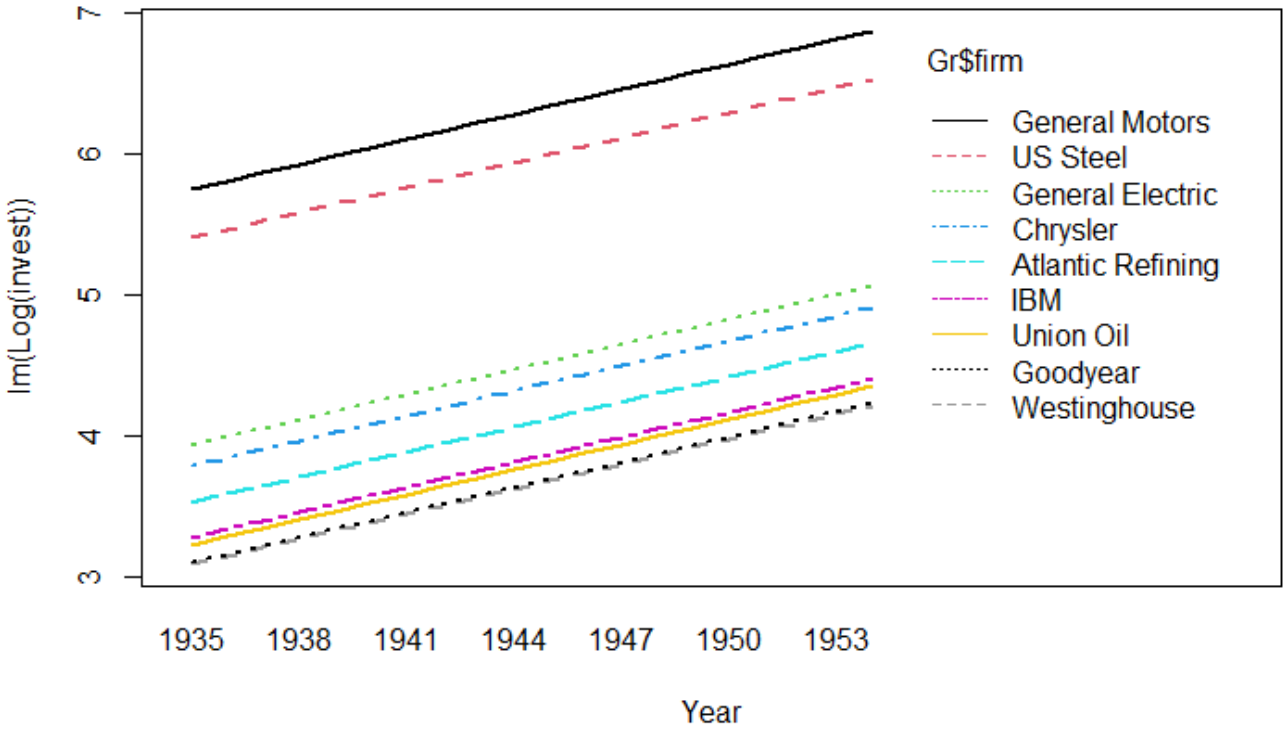
Opgave 1

Figur 1.1

```
> fit <- lm(log(invest) ~ year + firm, data = Gr)
> summary(fit)

Call:
lm(formula = log(invest) ~ year + firm, data = Gr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53761 -0.15616 -0.00953  0.16100  0.65626

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.088e+02  6.287e+00 -17.308  < 2e-16 ***
year                   5.921e-02  3.233e-03  18.313  < 2e-16 ***
firmUS Steel          -3.427e-01  7.910e-02  -4.332 2.52e-05 ***
firmGeneral Electric  -1.805e+00  7.910e-02 -22.824  < 2e-16 ***
firmChrysler          -1.958e+00  7.910e-02 -24.747  < 2e-16 ***
firmAtlantic Refining -2.218e+00  7.910e-02 -28.039  < 2e-16 ***
firmIBM               -2.468e+00  7.910e-02 -31.196  < 2e-16 ***
firmUnion Oil         -2.519e+00  7.910e-02 -31.849  < 2e-16 ***
firmWestinghouse      -2.653e+00  7.910e-02 -33.544  < 2e-16 ***
firmGoodyear          -2.640e+00  7.910e-02 -33.373  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2501 on 170 degrees of freedom
Multiple R-squared:  0.9442,    Adjusted R-squared:  0.9412
F-statistic: 319.3 on 9 and 170 DF,  p-value: < 2.2e-16
```

Figur 1.2

```
interaction.plot(Gr$year,Gr$firm,fitted(fit),ylab = "lm(Log(invest))",xlab = "Year",
                 col = 1:9, lty = 1:9,lwd = 2)
```



Figur 1.3

Først findes den række-nummer, der har year=1944 og firm="Chrysler" vha. 'which'. Vi får output til 70. Så vi kan bruge rækken 70 til at finde værdien hos fitted(fit)

```
> ##Finder data, der er Gr$year == 1944 & Gr$firm == "Chrysler"
> (data_nummer<- which(Gr$year == 1944 & Gr$firm == "Chrysler"))
[1] 70
> ## Vi finder  prædikterede værdi ved data=70
> fitted(fit)[data_nummer]
70
4.325885
```

Figur 1.4

```
> fitx <- lm(log(invest)~factor(year)+firm,Gr)
> summary(fitx)

Call:
lm(formula = log(invest) ~ factor(year) + firm, data = Gr)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4570 -0.1348 -0.0098  0.1346  0.5382

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.62651    0.08602  65.410  < 2e-16 ***
factor(year)1936     0.29307    0.10281   2.851  0.00497 **
factor(year)1937     0.52306    0.10281   5.088 1.06e-06 ***
factor(year)1938     0.19725    0.10281   1.919  0.05692 .
factor(year)1939     0.17478    0.10281   1.700  0.09117 .
factor(year)1940     0.43252    0.10281   4.207 4.41e-05 ***
factor(year)1941     0.70278    0.10281   6.836 1.86e-10 ***
factor(year)1942     0.52827    0.10281   5.138 8.40e-07 ***
factor(year)1943     0.49758    0.10281   4.840 3.17e-06 ***
factor(year)1944     0.61440    0.10281   5.976 1.56e-08 ***
factor(year)1945     0.67796    0.10281   6.594 6.66e-10 ***
factor(year)1946     0.87022    0.10281   8.464 2.02e-14 ***
factor(year)1947     0.82037    0.10281   7.979 3.32e-13 ***
factor(year)1948     0.87203    0.10281   8.482 1.82e-14 ***
factor(year)1949     0.72796    0.10281   7.080 4.97e-11 ***
factor(year)1950     0.77587    0.10281   7.546 3.82e-12 ***
factor(year)1951     1.12218    0.10281  10.915  < 2e-16 ***
factor(year)1952     1.22878    0.10281  11.952  < 2e-16 ***
factor(year)1953     1.36882    0.10281  13.314  < 2e-16 ***
factor(year)1954     1.30290    0.10281  12.673  < 2e-16 ***
firmUS Steel        -0.34269    0.06897  -4.969 1.79e-06 ***
firmGeneral Electric -1.80538   0.06897 -26.177  < 2e-16 ***
firmChrysler        -1.95756    0.06897 -28.383  < 2e-16 ***
firmAtlantic Refining -2.21793  0.06897 -32.158  < 2e-16 ***
firmIBM             -2.46767    0.06897 -35.779  < 2e-16 ***
firmUnion Oil       -2.51927    0.06897 -36.528  < 2e-16 ***
firmWestinghouse    -2.65339    0.06897 -38.472  < 2e-16 ***
firmGoodyear        -2.63986    0.06897 -38.276  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2181 on 152 degrees of freedom
Multiple R-squared:  0.962,     Adjusted R-squared:  0.9553
F-statistic: 142.7 on 27 and 152 DF,  p-value: < 2.2e-16
```

Figur 1.5

```
> fitx <- lm(log(invest) ~ factor(year) + firm, data = Gr)
> fitted(fitx)[70]
       70
4.283354
```

Figur 1.6

```
> anova(fit,fitx)
Analysis of Variance Table

Model 1: log(invest) ~ year + firm
Model 2: log(invest) ~ factor(year) + firm
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    170 10.6370
2    152  7.2302 18    3.4068 3.9789 1.109e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figur 1.7

```
> anova(fit)
Analysis of Variance Table

Response: log(invest)
           Df  Sum Sq Mean Sq F value    Pr(>F)
year        1  20.983 20.9829  335.35 < 2.2e-16 ***
firm        8 158.838 19.8548  317.32 < 2.2e-16 ***
Residuals 170  10.637  0.0626
```

```
> anova(fitx)
Analysis of Variance Table

Response: log(invest)
              Df Sum Sq Mean Sq F value    Pr(>F)
factor(year)  19  24.39  1.2837  26.986 < 2.2e-16 ***
firm           8 158.84 19.8548 417.405 < 2.2e-16 ***
Residuals    152   7.23  0.0476
---
```

Figur 1.8

```
     Mean Residual Variance Residual
1935  6.514790e-16         0.2468573
1936  4.258806e-17         0.2399957
1937  4.972648e-17         0.2476322
1938 -2.359224e-16         0.2037721
1939 -6.168056e-17         0.1409022
1940 -9.618078e-17         0.1299365
1941  1.445452e-17         0.1588740
1942 -4.163336e-17         0.2210752
1943 -5.243247e-17         0.2173360
1944 -6.323308e-17         0.3388827
1945 -3.989864e-17         0.1955739
1946 -1.538965e-18         0.2020503
1947 -3.077930e-18         0.2172185
1948 -5.551115e-17         0.1546973
1949  1.060410e-17         0.1724530
1950 -4.644969e-17         0.2023323
1951 -6.398901e-17         0.1642191
1952 -2.332108e-17         0.1126992
1953 -3.720169e-17         0.2045735
1954 -2.775558e-17         0.3272316
```

	Mean Residual	Variance Residual
General Motors	1.864828e-16	0.02110553
US Steel	-8.785832e-17	0.04529627
General Electric	-3.461553e-17	0.04796307
Chrysler	-2.471981e-17	0.04150622
Atlantic Refining	-3.259925e-17	0.04622460
IBM	1.390489e-18	0.06796666
Union Oil	-2.367830e-17	0.02856018
Westinghouse	-2.003673e-17	0.03187847
Goodyear	-7.833361e-19	0.05003658

Opgave 2

Figur 2.1

```
> VarbetaAR1 <- solve(t(X)%*%X)%*%t(X)%*%Sigma%*%X%*%solve(t(X)%*%X);
> VarbetaAR1
           [,1]        [,2]
[1,]  0.8566722 -0.11578374
[2,] -0.1157837  0.02105159
> VarbetaStd <- sigma2*solve(t(X)%*%X); VarbetaStd
           [,1]        [,2]
[1,]  0.46666667 -0.06666667
[2,] -0.06666667  0.01212121
```

Figur 2.2

```
> VariansbetaAR1 <- solve(t(X) %*% solve(Sigma) %*% X); VariansbetaAR1
           [,1]       [,2]
[1,]  0.8156579 -0.1094186
[2,] -0.1094186  0.0198943
```

Figur 2.3

```
> library(mvtnorm)
> Sigma11 <- Sigma[-10, -10]
> Sigma22 <- Sigma[10, 10]
> Sigma12 <- Sigma[-10, 10]
> Sigma21 <- Sigma[10, -10]
> mu <- rep(0, N)
> # Vi genererer Y med middelværdi 0 og kovariansmatrix lig Sigma
> Y <- rmvnorm(1, mean = mu, sigma = Sigma)
> Betinget_middel <- Sigma21 %*% solve(Sigma11) %*% Y[-10]; Betinget_middel
          [,1]
[1,] 0.1020962
> Betinget_varians <- Sigma22 - Sigma12 %*% solve(Sigma11) %*% Sigma21; Betinget_varians
       [,1]
[1,] 0.8236
```

Figur 2.4

```
> Sigma21%*%solve(Sigma11)
              [,1]         [,2]         [,3]          [,4]         [,5]          [,6]         [,7]          [,8]
[1,] 2.487996e-35 2.494713e-19 8.756653e-19 -1.133408e-18 5.367048e-18 -3.669797e-18 4.273873e-17 -1.110223e-16
     [,9]
[1,] 0.42
```

Opgave 3

Figur 3.1

```
> #F-teststørrelse
> Fvalue <- ((RSSr-RSS)/(dfr-df))/(RSS/df); Fvalue
[1] 12.27275
> #p-værdi
> pf(Fvalue, df1 = dfr-df, df2 = df, lower=FALSE)
[1] 1.157965e-05
> #til sammenligning med facit output:
> anova(fitx, expand_fitx)
Analysis of Variance Table

Model 1: log(invest) ~ factor(year) + firm
Model 2: log(invest) ~ factor(year) + firm + log(value) + log(capital)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    152 7.2302
2    150 6.2135  2    1.0168 12.273 1.158e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #konfidensinterval for log(value)
> confint(expand_fitx,level=0.95)["log(value)",]
    2.5 %    97.5 %
0.2013022 0.5236323
```

Figur 3.2

```
Single term deletions

Model:
log(invest) ~ factor(year) + firm * (log(value) + log(capital))
                  Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                          4.2517 -582.22
factor(year)      19    3.5146 7.7662 -511.77  5.8299 2.032e-10 ***
firm:log(value)    8    0.4364 4.6881 -580.63  1.7192  0.099386 .
firm:log(capital)  8    0.7145 4.9662 -570.25  2.8150  0.006456 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figur 3.3

```
> print(TX1[,1])
  [1] 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [14] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [27] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [40] 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [53] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503
 [66] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [79] 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [92] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503
[105] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
[118] 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
[131] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503
[144] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
[157] 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
[170] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
> print(TX2[,1])
  [1] 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [14] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [27] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [40] 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [53] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503
 [66] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [79] 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
 [92] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503
[105] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
[118] 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
[131] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503
[144] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
[157] 0.5773503 0.5773503 0.5773503 0.5773503 1.0000000 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
[170] 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503 0.5773503
```

Figur 3.4

```
Model 1: TY ~ TX2 - 1
Model 2: TY ~ TX1 - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    142 5.3210
2    134 4.9193  8    0.4017 1.3678 0.2162
```

Opgave 4

Figur 4.1

```
> boxTidwell(MCAS$totsc8~MCAS$percap, data=MCAS)
 MLE of lambda Score Statistic (t)  Pr(>|t|)
       -1.1959             -7.9751 1.821e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  3
> percap_boxtidwell <- MCAS$percap^(-1.1959)
> plot(percap_boxtidwell, MCAS$totsc8)
```

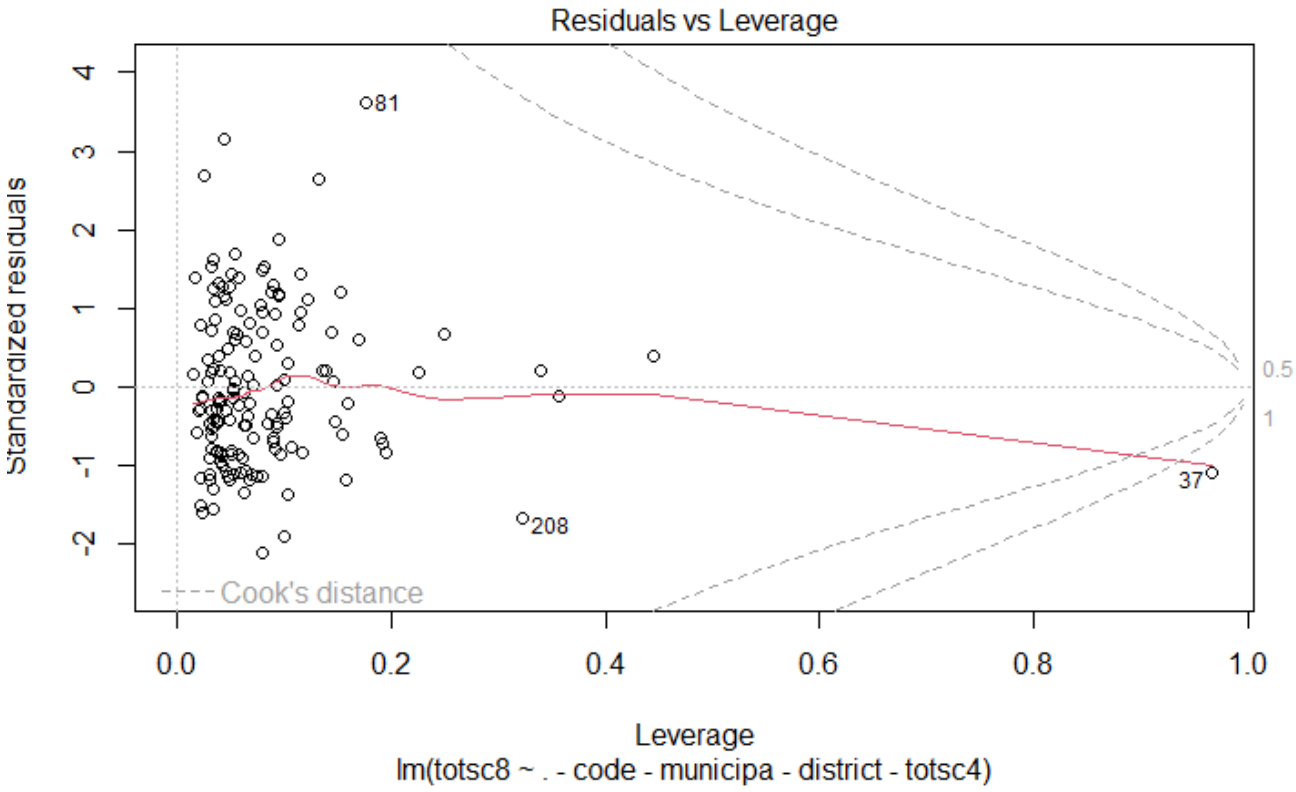
Figur 4.2

```
> vif(fit)
   regday  specneed  bilingua  occupday    totday       spc    speced   lnchpct  tchratio
19.663586  2.750799  1.041924  1.563100 24.666184  1.092560  1.419987  3.773392  1.698637
   percap avgsalary     pctel
 3.093565  1.896306  2.312229
```

Figur 4.3

```
> summary(fit.infl)
Potentially influential observations of
         lm(formula = totsc8 ~ . - code - municipa - district - totsc4,      data = MCAS) :

    dfb.1_ dfb.rgdy dfb.spcn dfb.blng dfb.occp dfb.ttdy dfb.spc dfb.spcd dfb.lnch dfb.tchr
22   -0.06   -0.05    -0.02     0.00     0.00     0.07    -0.01     0.01     0.00     0.07
37   -0.14   -0.19    -0.04    -5.75_*   0.17     0.18    -0.06     0.11     0.05     0.00
38   -0.06    0.07    -0.05    -0.01     0.05    -0.12     0.04     0.12     0.16     0.08
40    0.01   -0.01    -0.03     0.00     0.05     0.02     0.02     0.02    -0.04    -0.01
41    0.02   -0.04    -0.08     0.00    -0.01     0.06    -0.01    -0.04    -0.01     0.00
81    0.54    0.11     0.37     0.14     0.04     0.14    -0.01    -0.38    -0.61     0.42
88    0.06    0.05    -0.03    -0.01     0.02    -0.05     0.03     0.02    -0.07    -0.06
97    0.02    0.02     0.03     0.01     0.02    -0.04    -0.07    -0.07     0.04     0.01
107  -0.07    0.09     0.01    -0.01    -0.01    -0.06     0.08     0.06     0.08     0.00
122  -0.02    0.12     0.00    -0.01    -0.21    -0.11    -0.06     0.03     0.11     0.09
123   0.01    0.10    -0.28    -0.10    -0.01    -0.08    -0.30     0.15    -0.08     0.07
136   0.00   -0.01    -0.01     0.00    -0.02     0.01     0.00    -0.01     0.02     0.01
144   0.00    0.05     0.19    -0.01    -0.01    -0.09     0.09    -0.06    -0.21     0.07
155   0.01    0.00    -0.02     0.00     0.08     0.00    -0.01    -0.03     0.02    -0.01
169  -0.06   -0.19    -0.25    -0.01     0.03     0.26     0.24    -0.04     0.68     0.03
203  -0.14    0.29    -0.02    -0.03    -0.01    -0.23     0.01     0.10     0.07     0.21
208   0.17   -0.25     0.15    -0.04    -0.01     0.11     0.08    -0.01    -0.44    -0.21
    dfb.prcp dfb.avgs dfb.pctl dffit   cov.r   cook.d  hat
22   -0.01     0.01     0.00     0.14    1.65_*  0.00    0.34_*
37    0.00     0.16     0.09    -5.87_* 29.89_*  2.65_*  0.97_*
38    0.20     0.02    -0.11    -0.41    1.28_*  0.01    0.20
40   -0.01    -0.01    -0.01    -0.09    1.30_*  0.00    0.16
41    0.00    -0.01     0.01    -0.08    1.70_*  0.00    0.36_*
81   -0.14    -1.35_*   0.04     1.76_*  0.38_*  0.22    0.18
88    0.00    -0.03     0.29     0.35    1.94_*  0.01    0.44_*
97    0.00     0.02     0.11     0.27    1.28_*  0.01    0.17
107  -0.02     0.01    -0.27    -0.35    1.29_*  0.01    0.19
122   0.06    -0.06    -0.03    -0.31    1.30_*  0.01    0.19
123  -0.04     0.11     0.04     0.69    0.45_*  0.03    0.04
136   0.00     0.00    -0.01     0.03    1.28_*  0.00    0.15
144  -0.08     0.02     0.15     0.45    0.57_*  0.01    0.03
155   0.01     0.02    -0.03     0.10    1.41_*  0.00    0.23
169   0.39    -0.25    -0.35     1.06_*  0.66_*  0.08    0.13
203   0.00    -0.03    -0.05     0.39    1.40_*  0.01    0.25
208  -0.79     0.38     0.27    -1.16_*  1.25    0.10    0.32_*
```

Figur 4.4



Opgave 7

Figur 7.1

```
Call:
glm(formula = wfood ~ 1, family = quasibinomial(), data = BF4)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.605877   0.008537  -70.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.0911582)

    Null deviance: 517.8  on 5475  degrees of freedom
Residual deviance: 517.8  on 5475  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 3
```

Figur 7.2

```
> mu <- mean(BF4$wfood); mu
[1] 0.3530002
> s2 <- var(BF4$wfood); phi <- ((mu*(1-mu)/s2)-1)
> alpha <- mu*phi; alpha
[1] 3.5194
> beta <- phi*(1-mu); beta
[1] 6.450565
```

Figur 7.3

Manuelt	deltaMethod

`> mu <- mean(BF4$wfood)`
`> psi <- summary(BF4model)$dispersion`
`> var_y <- psi*mu*(1-mu)`
`> n <- length(BF4$wfood)`
`> se_y <- sqrt(var_y/n); se_y`
`[1] 0.00194987`
`> se_mu <- (1/(mu-mu^2))*se_y; se_mu`
`[1] 0.008537419`
	`> b0 <- coef(BF4model)`
	`> deltaMethod(BF4model, "b0", parameterNames = c("b0"))`
	`Estimate SE 2.5 % 97.5 %`
	`b0 -0.6058772 0.0085374 -0.6226102 -0.5891`

Figur 7.4

```
Call:
betareg(formula = wfood ~ 1, data = BF4)

Quantile residuals:
    Min      1Q  Median      3Q     Max
-5.0325 -0.6451 -0.0116  0.6239  4.1624

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.603420   0.008566  -70.44   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   9.8213     0.1797   54.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  2916 on 2 Df
Number of iterations: 9 (BFGS) + 2 (Fisher scoring)
```

Figur 7.5

```
Call:
betareg(formula = wfood ~ poly(log(totexp), 2), data = BF4)

Quantile residuals:
    Min      1Q  Median      3Q     Max
-7.5828 -0.6163 -0.0508  0.5955  4.4820

Coefficients (mean model with logit link):
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.625936   0.006977 -89.719  < 2e-16 ***
poly(log(totexp), 2)1 -30.231211   0.537924 -56.200  < 2e-16 ***
poly(log(totexp), 2)2  -3.705778   0.551146  -6.724 1.77e-11 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  16.1497     0.3008   53.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  4265 on 4 Df
Pseudo R-squared: 0.3915
Number of iterations: 13 (BFGS) + 3 (Fisher scoring)
> |
```

Figur 7.6

```
> betaregmodel <- betareg(BF4$wfood~poly(log(totexp), 2)|poly(log(totexp), 2), data = BF4)
> summary(betaregmodel)

Call:
betareg(formula = BF4$wfood ~ poly(log(totexp), 2) | poly(log(totexp), 2), data = BF4)

Quantile residuals:
    Min      1Q  Median      3Q     Max
-6.6279 -0.6297 -0.0452  0.6111  4.3900

Coefficients (mean model with logit link):
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.632674   0.006918 -91.459  < 2e-16 ***
poly(log(totexp), 2)1 -30.966404   0.563686 -54.936  < 2e-16 ***
poly(log(totexp), 2)2  -3.689996   0.626599  -5.889 3.89e-09 ***

Phi coefficients (precision model with log link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.81504    0.01862 151.220  < 2e-16 ***
poly(log(totexp), 2)1 15.59492    1.37347  11.354  < 2e-16 ***
poly(log(totexp), 2)2 -6.32619    1.34710  -4.696 2.65e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  4351 on 6 Df
Pseudo R-squared: 0.3915
Number of iterations: 13 (BFGS) + 2 (Fisher scoring)
```

Kode

```
# opgave 1
rm(list=ls())
install.packages("AER")
library(AER)
data(Grunfeld)
View(Grunfeld)
summary(Grunfeld)

Gr <- head(Grunfeld,-40) # drop last 2 firms (fjerne sidste 40 observationer)
Gr$firm <- factor(Gr$firm) # drop factor levels - vi konverterer dataene til
#kategorisk data for sikkerhedsskyld
View(Gr)

# 1.1
interaction.plot(Gr$year,Gr$firm,log(Gr$invest),ylab = "Log(invest)",xlab = "Year",
            col = 1:9, lty = 1:9,lwd = 1.5)

# 1.2
fit <- lm(log(invest)~year+firm,Gr)
summary(fit)

interaction.plot(Gr$year,Gr$firm,fitted(fit),ylab = "lm(Log(invest))",xlab = "Year",
            col = 1:9, lty = 1:9,lwd = 2)

#Vi udskriver summary(fit) med 10 decimaler, så vi kan udregne log(invest) for
# Chrysler i 1944 med mere nøjagtige tal:
summaryfit <- summary(fit)
printCoefmat(summaryfit$coefficients, digits = 10)
fitted(fit)

##Finder data, der er Gr$year == 1944 & Gr$firm == "Chrysler"
(data_nummer<- which(Gr$year == 1944 & Gr$firm == "Chrysler"))
## Vi finder  prædikterede værdi ved data=70
fitted(fit)[data_nummer]

# 1.3
rf <- residuals(fit)
with(Gr, interaction.plot(year, firm, rf,
                    xlab = "Year", #Overskriften på x-aksen
                    ylab = "Residuals", #overskriften på y-aksen
                    main = "Interaction Plot of Residuals", #titel
```

```
                col = 1:5, #Tilføjer farver
                lty = 1:9,
                lwd = 1.5))

with(Gr,abline(lm(rf~year), col="red"),xaxp = c(1935,1953,1))

# 1.4
fitx <- lm(log(invest)~factor(year)+firm,Gr)
summary(fitx)
#Igen finder vi log(invest) for Chrysler i 1944:
fitted(fitx)[70]

# 1.5
fit <- lm(log(invest)~year+firm,Gr)
fitx <- lm(log(invest)~factor(year)+firm,Gr)
anova(fit,fitx)
anova(fit)
anova(fitx)

# 1.6
qqnorm(residuals(fitx), main = "Q-Q residuals")
qqline(residuals(fitx))
residualPlots(fitx, terms = ~ . - Gr$year)

#middelværdi og varians af residualerne for firm
Gr$firm <- factor(Gr$firm)
firms <- levels(Gr$firm); firms
mean_residuals <- sapply(firms, function(firm) mean(fitx$residuals[Gr$firm == firm]))
var_residuals <- sapply(firms, function(firm) var(fitx$residuals[Gr$firm == firm]))
matrix <- cbind(mean_residuals, var_residuals)
rownames(matrix) <- firms
colnames(matrix) <- c("Mean Residual", "Variance Residual")
print(matrix)

#middelværdi og varians af residualerne for year
Grunfeld$year <- factor(Grunfeld$year)
years <- levels(Grunfeld$year); years
mean_residuals_years <- sapply(years, function(year) mean(fitx$residuals[Gr$year == year]))
sd_residuals_years <- sapply(years, function(year) sd(fitx$residuals[Gr$year == year]))
matrix_years <- cbind(mean_residuals_years, sd_residuals_years)
rownames(matrix_years) <- years
colnames(matrix_years) <- c("Mean Residual", "Variance Residual")
print(matrix_years)
```

```
#  1.7
rx <- residuals(fitx)
plot(tail(rx,-1) ~ head(rx,-1), subset=tail(Gr$year,-1) > 1935)

# opgave 2

# 2.1
#Variansmatricen SIGMA
N <- 10
rho <- .42
sigma2 <-1
Sigma <- diag(N)
Sigma <- sigma2 * rho^abs(row(Sigma)-col(Sigma))

#Designmatricen X
X <- cbind(1, 1:N)
VarbetaAR1 <- solve(t(X)%*%X)%*%t(X)%*%Sigma%*%X%*%solve(t(X)%*%X); VarbetaAR1
VarbetaStd <- sigma2*solve(t(X)%*%X); VarbetaStd


# 2.3
VariansbetaAR1 <- solve(t(X) %*% solve(Sigma) %*% X); VariansbetaAR1

# 2.4
library(mvtnorm)
Sigma11 <- Sigma[-10, -10]
Sigma22 <- Sigma[10, 10]
Sigma12 <- Sigma[-10, 10]
Sigma21 <- Sigma[10, -10]
mu <- rep(0, N)
# Vi genererer Y med middelværdi 0 og kovariansmatrix lig Sigma
Y <- rmvnorm(1, mean = mu, sigma = Sigma)
Betinget_middel <- Sigma21 %*% solve(Sigma11) %*% Y[-10]; Betinget_middel
Betinget_varians <- Sigma22 - Sigma12 %*% solve(Sigma11) %*% Sigma21; Betinget_varians

Sigma21 %*% solve(Sigma11)

#opgave 3

rm(list=ls())
library(AER)
data(Grunfeld)
```

```
Gr <- head(Grunfeld,-40) # drop last 2 firms
Gr$firm <- factor(Gr$firm) # drop factor levels
View(Gr)
#lineær model
fitx <- lm(log(invest)~factor(year) + firm, Gr)

# 3.1
expand_fitx <- lm(log(invest)~factor(year) + firm + log(value) + log(capital), Gr)
summary(expand_fitx)
summary(fitx)
expand_fitxfitted <- expand_fitx$fitted.values
fitxfitted <- fitx$fitted.values
RSS <- sum((log(Gr$invest)-expand_fitxfitted)^2); RSS
RSSr <- sum((log(Gr$invest)-fitxfitted)^2); RSSr
df <- expand_fitx$df.residual
dfr <- fitx$df.residual
n <- length(Gr$invest)

#F-teststørrelse
Fvalue <- ((RSSr-RSS)/(dfr-df))/(RSS/df); Fvalue
#p-værdi
pf(Fvalue, df1 = dfr-df, df2 = df, lower=FALSE)
#til sammenligning med facit output:
anova(fitx, expand_fitx)
#konfidensinterval for log(value)
confint(expand_fitx,level=0.95)["log(value)",]

# 3.2
vvfitx <- lm(log(invest)~factor(year) + firm*(log(value) + log(capital)), Gr)
drop1(vvfitx, test="F")
#alternativ
#vvfitx <- lm(log(invest)~factor(year) + firm + log(value) + log(capital) + firm:(log(value) + log(capital)), Gr)

# 3.3
rho <- .5
Psi <- diag(20)
Psi <- rho^abs(row(Psi)-col(Psi))
T0 <- t(backsolve(chol(Psi),diag(20)))

#check at:
zapsmall(T0 %*% Psi %*% t(T0))

#konstruer T
```

```
T <- kronecker(diag(9),T0)
fitvv <- lm(log(invest) ~ factor(year) + firm*(log(value) + log(capital)), Gr)
X1 <- model.matrix(fitvv)
Y <- log(Gr$invest)
TY <- T %*% Y
TX1 <- T %*% X1
serialfitvv <- lm(TY ~ TX1 - 1)
fit <- lm(log(invest) ~ factor(year) + firm*log(value) + log(capital), Gr)
X2 <- model.matrix(fit); X1
Y <- log(Gr$invest)
TY <- T %*% Y
TX2 <- T %*% X2
serialfit <- lm(TY ~ TX2 - 1)
anova(serialfit, serialfitvv)

#vi tjekker første søjle i designmatricerne
print(TX1[,1])
print(TX2[,1])

# 3.4
library(nlme)
glsmodel <- gls(log(invest)~factor(year) + firm*log(value) + log(capital), data = Gr,
            correlation = corAR1(0.5, form = ~ year | firm, fixed = TRUE))
glsmodelvv <- gls(log(invest)~factor(year) + firm*(log(value) + log(capital)), data = Gr,
              correlation = corAR1(0.5, form = ~ year | firm, fixed = TRUE))
anova(glsmodel, glsmodelvv)



# Opgave 4

# 4.1
install.packages("Ecdat")
install.packages("Ecfun")
library(Ecdat)

MCAS <- MCAS
plot(MCAS$percap, MCAS$totsc8, xlab="Income per capita", ylab="8th grade score")

# 4.2
boxTidwell(MCAS$totsc8~MCAS$percap, data=MCAS)
percap_boxtidwell <- MCAS$percap^(-1.1959)
plot(percap_boxtidwell, MCAS$totsc8,
   xlab="Income per capita efter box-tide-well", ylab="8th grade score")
```

```
# 4.3
library(MASS)
bc1_resultat <- boxcox(MCAS$totsc8 ~ percap_boxtidwell, lambda=seq(-2,5))

opt1_lambda <- bc1_resultat$x[which.max(bc1_resultat$y)]
totsc8_transformation1 <- (MCAS$totsc8 ^ opt1_lambda - 1) / opt1_lambda
hist(MCAS$totsc8)
hist(totsc8_transformation1)
plot(totsc8_transformation1 ~ percap_boxtidwell,
    xlab="income per capita efter box-tide-well", ylab="8th grade score transformeret",
    main = "Plot af transformeret 8th grade score vs income per capita")

# 4.4
fit <- lm(totsc8 ~ . -code - municipa - district - totsc4, data = MCAS)


# 4.5
vif(fit)

# 4.6
install.packages("ISwR")
library(Ecdat)
library(ISwR)

#Vi omdømmer fit i stedet for regr, så vi kan bruge Peter's kode:
fit <- lm(totsc8 ~ . -code - municipa - district - totsc4, data = MCAS)
fit.infl <- influence.measures(fit)
summary(fit.infl)

plot(fit, ask=FALSE) # use plot history

# opgave 7

# 7.1
library(Ecdat)
BF4 <- subset(BudgetFood, wfood > 0 & size==4)
BF4model <- glm(wfood~1, family=quasibinomial(), data = BF4)
summary(BF4model)
confint(BF4model)
confint.default(BF4model)
mu <- exp(coef(BF4model)) / (1 + exp(coef(BF4model))); mu
phi <- (1 /(summary(BF4model)$dispersion))-1; phi
```

```
# Beregn alpha og beta
alpha <- mu * phi
beta <- (1 - mu) * phi
alpha
beta
hist(BF4$wfood, freq=FALSE)
x <- seq(0, 1, length.out = 100)
curve(dbeta(x, alpha, beta), add=TRUE)

# 7.2
mu <- mean(BF4$wfood); mu
s2 <- var(BF4$wfood); phi <- ((mu*(1-mu)/s2)-1)
alpha <- mu*phi; alpha
beta <- phi*(1-mu); beta
mu <- mean(BF4$wfood)
psi <- summary(BF4model)$dispersion
var_y <- psi*mu*(1-mu)
n <- length(BF4$wfood)
se_y <- sqrt(var_y/n); se_y
se_mu <- (1/(mu-mu^2))*se_y; se_mu
library(car)
b0 <- coef(BF4model)
deltaMethod(BF4model, "b0", parameterNames = c("b0"))

# 7.3
install.packages("betareg")
library(betareg)
BF4model2 <- betareg(wfood~1, data = BF4)
summary(BF4model2)

# 7.4
library(stats4)
y <- BF4$wfood
mll <- function(mu=0.5, phi=1){
  alpha <- mu * phi
  beta <- (1 - mu) * phi
  lh <- (-sum(dbeta(BF4$wfood, alpha, beta, log = TRUE)))
}
suppressWarnings(summary(mle(mll)))
mle_mu <- suppressWarnings(coef(summary(mle(mll)))[1])
mle_phi <- suppressWarnings(coef(summary(mle(mll)))[2])
intercept <- round(log(mle_mu/(1-mle_mu))); intercept
```

```
mle_phi

# 7.5
plot(log(BF4$totexp), BF4$wfood, xlab = "log(totexp)", ylab = "wfood")

# 7.6
library(betareg)
model <- betareg(wfood ~ poly(log(totexp), 2), data = BF4)
summary(model)
sorted <- order(log(BF4$totexp))
sorted_totexp <- log(BF4$totexp)[sorted]
sorted_mu <- predict(model)[sorted]
lines(sorted_totexp, sorted_mu, col = 'red', lwd = 5)
lines(sorted_totexp, predict(model, type="response")[sorted], col="green", lty="dashed", lwd = 3)
lines(sorted_totexp, predict(model, type="quantile", at=0.95)[sorted], col="blue",)
lines(sorted_totexp, predict(model, type="quantile", at=0.05)[sorted], col="blue")

# 7.7
r <- residuals(model, type="pearson")
logtot <- log(BF4$totexp)
plot(logtot, r^2)
lines(lowess(logtot, r^2, iter=0), col="red")

# 7.8
betaregmodel <- betareg(BF4$wfood~poly(log(totexp), 2)|poly(log(totexp), 2), data = BF4)
summary(betaregmodel)
par(mfrow = c(1,2))
plot(log(BF4$totexp), BF4$wfood, xlab = "log(totexp)", ylab = "wfood")
sorted <- order(log(BF4$totexp))
sorted_totexp <- log(BF4$totexp)[sorted]
sorted_predict <- predict(betaregmodel)[sorted]
lines(sorted_totexp, sorted_predict, col = 'red', lwd = 5)
lines(sorted_totexp, predict(betaregmodel, type="response")[sorted], col="green", lty="dashed", lwd = 3)
lines(sorted_totexp, predict(betaregmodel, type="quantile", at=0.95)[sorted], col="blue",)
lines(sorted_totexp, predict(betaregmodel, type="quantile", at=0.05)[sorted], col="blue")
r <- residuals(betaregmodel, type="pearson")
logtot <- log(BF4$totexp)
plot(logtot, r^2)
lines(lowess(logtot, r^2, iter=0), col="red")
par(mfrow = c(1,1))
```