Statistiske Modeller
Efterår 2024
Eksamen, skriftligt oplæg

13. december – 18. december 2024

Formalia

Eksamen er en mundtlig eksamen med udgangspunkt i et skriftligt oplæg, som er en besvarelse af nærværende opgavesæt. Opgaverne er baseret på tre hjemmeopgaver som er stillet i løbet af semesteret[1].

Besvarelsen udarbejdes i grupper af 3–5 personer. Sidetallet af selve besvarelsen må ikke overstige 15 normalsider[2]. Det er tilladt derudover at inkludere bilag, men disse tæller som udgangspunkt ikke med i bedømmelsen. Besvarelsen bør være i form af en PDF-fil.

Besvarelsen af opgaven skal afleveres i Digital Eksamen indenfor tidsfristen. Dette er en forudsætning for at deltage i den mundtlige eksamen.

R-koder og output kan indgå i besvarelsen, men regnes *ikke* som fyldestgørende besvarelse.

[1] I forhold til hjemmeopgaverne er der foretaget enkelte ændringer. Følgende delopgaver udgår: Hj2:1.4 (`gls`), Hj2:2.7 (modelsøgning), Hj2:3.6–8 (andet led i loglikelihood) og Hj3:1.3 (grænsefordeling for betafordelingen) udgår. Hj1, 1.7 og Hj2, 1.3 har reducerede krav, og der er derudover rettet enkelte formuleringer og trykfejl.

[2] Dvs 15 *fysiske* sider, med i gennemsnit max 2275 anslag inkl. mellemrum per side, osv. `https://studentcbs.sharepoint.com/sites/CoursesAndExams/SitePages/Formalia.aspx`

Opgave 1

Data til opgaven er datasættet `Grunfeld` som findes i `AER` pakken[3] Disse data kommer oprindelig fra Y. Grünfelds Ph.D. afhandling (1958) og indeholder data om 11 virksomheders investeringer m.m. over 20 år fra 1935–1954 (såkaldte *panel data*). Når pakken er installeret, kan data hentes ind på følgende vis:

```
install.packages("AER") # Kun første gang!
library(AER)
data(Grunfeld, package="AER")
```

Se evt. hjælpesiden `help(Grunfeld)` for flere detaljer.

De to sidste virksomheder er meget mindre end de øvrige så vi udelader dem af datasættet i denne opgave

```
Gr <- head(Grunfeld,-40)   # drop last 2 firms
Gr$firm <- factor(Gr$firm) # drop factor levels
```

1° Benyt R funktionen `interaction.plot()` til at lave en figur, hvor de 20 observationer af `log(invest)` for hver virksomhed tegnes mod tiden.

2° Fit en additiv lineær model `fit`, hvori `log(invest)` beskrives via `year` og `firm`, dvs modelformlen skal være

```
log(invest) ~ year + firm
```

Beskriv output af `summary(fit)` og forklar modellens antagelser (benyt evt. igen funktionen `interaction.plot` på `fitted(fit)`. Angiv hvordan man ud fra `summary`-output kan finde den prædikterede værdi af `log(invest)` for Chrysler i 1944.

3° Plot `residuals(fit)` med `interaction.plot` og diskuter hvad det viser om modellens fit til data.

4° Udvid modellen ved at erstatte den lineære variabel `year` med en kategorisk variabel (`factor(year)`). Find også i denne model den prædikterede værdi for Chrysler i 1944.

5° Hvis den udvidede model kaldes `fitx`, kan vi sammenligne de to modeller med `anova(fit, fitx)`. Forklar hvordan output herfra skal læses. Angiv også læsningen af `anova(fit)` og `anova(fitx)`.

6° Udfør modelkontrol for `fitx`, specielt plot at residualer som før, og QQ-plot mod normalfordeling.

[3] Materiale fra Kleiber & Zeileis (2008): *Applied Econometrics with R*, Springer-Verlag, New York.

7° Undersøg om der er *seriel korrelation* i data ved at tegne residualerne op mod de umiddelbart foregående residualer for samme firma.

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

Opgave 2

Tidsrækker udviser ofte seriel korrelation, hvilket kan være et problem for Least Squares (OLS) metoder, der forudsætter ukorrelerede observationer.

En simpel model for tidsrækker er en *autoregression af orden 1*, AR(1). I denne model har variansmatricen formen

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots \\ \rho & 1 & \rho & \rho^2 & \ddots \\ \rho^2 & \rho & 1 & \rho & \ddots \\ \rho^3 & \rho^2 & \rho & 1 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

sådan at $\mathrm{Cor}(Y_i, Y_j) = \rho^{|i-j|}$.

Vi kan generere en variansmatrix af denne form i R, således:

```
N <- 10
rho <- .42
sigma2 <-1
Sigma <- diag(N)
Sigma <- sigma2 * rho^abs(row(Sigma)-col(Sigma))
```

Vi antager endvidere en lineær model, hvor $Y_t = \beta_0 + \beta_1 t$, altså med designmatrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & N \end{bmatrix}.$$

I R kan vi bruge

```
X <- cbind(1, 1:N)
```

1° Antag $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$ med $\boldsymbol{\Sigma}$ som beskrevet ovenfor (med de benyttede værdier af N, ρ og σ^2), og at vi anvender OLS til at estimere $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$. Find variansmatricen for $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. Sammenlign med resultatet af standardformlen for $\mathrm{Var}\,\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ når $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$.

2° Undersøg, om estimatoren $\tilde{\sigma}^2 = \text{RSS}/(N-2)$ er central hvis kovariansmatricen er $\boldsymbol{\Sigma}$. Vis først at $\text{E}(\text{RSS}) = \text{tr}((\boldsymbol{I}-\boldsymbol{P})\boldsymbol{\Sigma})$, hvor $\boldsymbol{P}$ er projektionsmatricen ved OLS. (Vink: Skriv $\text{RSS} = \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ og udnyt at (a) der gælder $\text{tr}(\boldsymbol{BA}) = \text{tr}(\boldsymbol{AB})$, og (b) $\text{E}(\boldsymbol{YY'}) = \boldsymbol{\Sigma}$, når $\text{E}\,\boldsymbol{Y} = \boldsymbol{0}$. Antag først, at $\text{E}\,\boldsymbol{Y} = \boldsymbol{0}$ og overvej derefter det generelle tilfælde.)

3° Gauss-Markov sætningen gælder ikke når $\boldsymbol{\Sigma} \neq \sigma^2\boldsymbol{I}$. Man kan vise, at det i stedet er optimalt[4] at bruge GLS (generalized least squares):

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

Beregn $\text{Var}\,\hat{\boldsymbol{\beta}}_{\text{GLS}}$ og sammenlign med resultaterne i spørgsmål 1°

4° Antag i dette spørgsmål, at $\boldsymbol{\beta} = \boldsymbol{0}$, sådan at $\text{E}\,\boldsymbol{Y} = \boldsymbol{0}$. Udregn (numerisk) den betingede fordeling af $\boldsymbol{Y}_2 = \begin{bmatrix} Y_{10} \end{bmatrix}$ givet

$$\boldsymbol{Y}_1 = \begin{bmatrix} Y_1 \\ \vdots \\ Y_9 \end{bmatrix}.$$

Ved beregning i R kan man beregne de relevante delmatricer af $\boldsymbol{\Sigma}$ som

```
Sigma11 <- Sigma[-10, -10]
Sigma22 <- Sigma[10,10]
Sigma12 <- Sigma[-10, 10]
Sigma21 <- Sigma[10, -10]
```

Hvad lægger man mærke til?

Opgave 3

Data til denne opgave er de samme som i opgave 1.

Vi tager udgangspunkt i modellen med "fixed effects" af `firm` og `year`

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

Denne gang vil vi inddrage den potientielle effekt af de to kovariater `value` og `capital`.

1° Udvid modellen til en model, hvori `log(value)` og `log(capital)` indgår additivt. Test for modelreduktion ved hjælp af den relevante del af `summary()`-output og angiv et konfidensinterval for regressionskoefficienten for `log(value)`.

[4] Forudsat $\boldsymbol{\Sigma}$ er kendt. I praksis vil man skulle estimere ρ, hvilket gør forholdene mere komplicerede.

2° Undersøg om der er vekselvirkning mellem de to kovariater og `firm` variablen (dvs, inkluder `firm*(log(value)+log(capital))` i modelformlen og brug `drop1()` funktionen). Angiv fortolkningen af sådanne vekselvirkninger.

3° Analyserne i de foregående spørgsmål adresserer ikke den mulige serielle korrelation i data. For at få en fornemmelse af hvor meget det betyder, kan vi prøve at fitte en model med en AR(1) korrelationsstruktur inden for firma. Dvs. i stedet for at antage $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ kunne vi antage

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Psi} \end{bmatrix},$$

hvor $\boldsymbol{\Psi}$ er en 20×20 korrelationsmatrix for en AR1 model (som diskuteret i opgave 2) og der er 9 blokke svarende til de 9 firmaer. Ideelt set burde man estimere korrelationsparameteren ρ, men det fører for vidt, så vi vælger at sætte $\rho = 0.5$.

Modeller hvor $\boldsymbol{\Sigma}$ er kendt pånær en proportionalitetsfaktor er principielt ret lette at fitte fordi man kan bruge en "transform both sides" tilgang: Hvis $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{\Sigma}_0)$ er $\boldsymbol{TY} \sim N(\boldsymbol{TX\beta}, \sigma^2\boldsymbol{T\Sigma}_0\boldsymbol{T}')$, og hvis vi vælger $\boldsymbol{T}$ så $\boldsymbol{T\Sigma}_0\boldsymbol{T}' = \boldsymbol{I}$ så er vi på kendt grund.

I vores tilfælde kan man vælge $\boldsymbol{T}$ til at bestå af identiske blokke, som hver beregnes ud fra en Choleski faktorisering af $\boldsymbol{\Psi}$. Følgende kode kan benyttes

```
N <- 20
Psi <- diag(20)
Psi <- .5^abs(row(Psi)-col(Psi))
T0 <- t(backsolve(chol(Psi),diag(20)))
T <- kronecker(diag(9),T0)
```

Man kan herefter udtrække designmatricen $\boldsymbol{X}$ via `model.matrix()`, og beregne[5]

```
Y <- log(Gr$invest)
TY <- T %*% Y
TX <- T %*% X
serialfit <- lm(TY ~ TX - 1)
```

(Bemærk at det er nødvendigt at fjerne interceptet. Forklar hvorfor.)

Prøv med denne teknik at fitte modellen fra spørgsmål 2°, og fit samme model uden vekselvirkningen mellem `firm` og `log(capital)`. Test hypotesen om ingen vekselvirkning med brug af `anova(model1, model2)` og sammenlign resultatet med de tidligere analyser, hvor man antog at der ikke var korrelation.

[5] Dette er inefficient, men da datamaterialet er ret lille, kan vi slippe afsted med det.

Opgave 4

I denne opgave bruges `MCAS` datasættet fra `Ecdat` pakken. Det gøres tilgængeligt via

```
install.packages("Ecdat") ## Kun første gang!
library(Ecdat)
```

1° Lav et plot af `totsc8` mod `percap`. Diskuter hvad figuren viser og om relationen ser ud til at være lineær.

2° Benyt Box-Tidwell metoden til at finde en transformation af `percap` der gør sammenhængen mere lineær, og tegn figuren igen med den transformerede variabel.

3° Ser det ud som om y-aksen også skal transformeres? Prøv med Box-Cox.

4° Udvid modellen til en multipel lineær regressionsanalyse ved at tilføje en række af de øvrige variable i datasættet (bemærk at de tre første variable ikke kan bruges, og at variablen `totsc4` kan give fortolkningsmæssige problemer.)

5° Angiv VIF for de indgående variable og diskuter om de høje VIF scores har en logisk forklaring.

6° Undersøg om der er observationer, der er særligt indflydelsesrige og diskuter hvad grunden er hertil.

Opgave 5

Betragt en lineær model med en generel kovariansmatrix

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

hvor $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ afhænger af en eller flere parametre.

Vi ved at MLE for sådanne modeller bliver biased for kovariansparametrene $\boldsymbol{\theta}$. Dette kan forklares ved, at den del af data der bruges til at estimere $\boldsymbol{\beta}$ ikke også kan også bruges til at estimere $\boldsymbol{\Sigma}$.

Den såkaldte REML metode (restricted ML) består i at fjerne middelværdikomponenten i data *før* man maksimerer likelihood. Hertil skal man bruge en *middelværdifjernende transformation*, som er en matrix $\boldsymbol{T}$ der opfylder $\boldsymbol{T}\boldsymbol{X} = \boldsymbol{0}$ og har maksimal rang, dvs $(n - d)$ når $\boldsymbol{X}$ er $n \times d$.

REML likelihoodfunktionen kan omskrives til en form, hvori matricen $\boldsymbol{T}$ ikke indgår eksplicit, hvilket gør beregningerne håndterbare.

Vi får brug for en række regneregler for determinanter[6]. Disse kan benyttes uden bevis:

$$|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}| \tag{1}$$
$$|\boldsymbol{A'}| = |\boldsymbol{A}| \tag{2}$$
$$|\boldsymbol{A}^{-1}| = 1/|\boldsymbol{A}| \tag{3}$$
$$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{B} & \boldsymbol{D} \end{vmatrix} = |\boldsymbol{A}||\boldsymbol{D} - \boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{C}| \tag{4}$$
$$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{vmatrix} = |\boldsymbol{A}||\boldsymbol{D}|. \tag{5}$$

1° Gør rede for at

$$\boldsymbol{TY} \sim N(\boldsymbol{0}, \boldsymbol{T\Sigma T'})$$

og at log-likelihood baseret på $\boldsymbol{TY}$ er

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

2° Gør rede for, at søjlerne i $\boldsymbol{T'}$ er en basis for det ortogonale komplement $L^{\perp}$ til $L = \text{span}\,\boldsymbol{X}$, og det derfor gælder, at $\boldsymbol{T'}(\boldsymbol{TT'})^{-1}\boldsymbol{T} = \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$ og omvendt, at $\boldsymbol{I} - \boldsymbol{T'}(\boldsymbol{TT'})^{-1}\boldsymbol{T} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$.

3° Lad $\boldsymbol{T}^* = \begin{bmatrix} \boldsymbol{T} \\ \boldsymbol{X'\Sigma}^{-1} \end{bmatrix}$ og vis, at

$$\boldsymbol{T}^*\boldsymbol{\Sigma}\boldsymbol{T}^{*\prime} = \begin{bmatrix} \boldsymbol{T\Sigma T'} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X'\Sigma}^{-1}\boldsymbol{X} \end{bmatrix},$$

hvis determinant kan skrives på to måder:

$$|\boldsymbol{T}^*\boldsymbol{\Sigma}\boldsymbol{T}^{*\prime}| = |\boldsymbol{T}^*|^2|\boldsymbol{\Sigma}| = |\boldsymbol{T\Sigma T'}||\boldsymbol{X'\Sigma}^{-1}\boldsymbol{X}|.$$

4° Vis, at

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

og at dens determinant er

$$\begin{aligned} |\boldsymbol{T}^*\boldsymbol{T}^{*\prime}| &= |\boldsymbol{T}^*|^2 \\ &= |\boldsymbol{TT'}||\boldsymbol{X'\Sigma}^{-2}\boldsymbol{X} - \boldsymbol{X'\Sigma}^{-1}\boldsymbol{T'}(\boldsymbol{TT'})^{-1}\boldsymbol{T\Sigma}^{-1}\boldsymbol{X}| \\ &= |\boldsymbol{TT'}||\boldsymbol{X'\Sigma}^{-1}(\boldsymbol{I} - \boldsymbol{T'}(\boldsymbol{TT'})^{-1}\boldsymbol{T})\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|. \end{aligned}$$

[6] Vi bruger her den korte notation $|\boldsymbol{A}| = \det \boldsymbol{A}$.

5° Benyt de foregående resultater til at vise, at [7]

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

Opgave 6

Betafordelingen er en fordeling på intervallet $(0,1)$ som bruges til at modellere data som er proportioner.

Vi betragter en stokastisk variabel Y som er betafordelt, altså med tæthed

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

Det kan benyttes uden nærmere redegørelse[8], at

$$\begin{aligned} \mathrm{E}\, Y &= \frac{\alpha}{\alpha+\beta} \\ \mathrm{Var}\, Y &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned}$$

1° Gør rede for, at betafordelingen kan opskrives som en to-parameter eksponentiel familie, hvis man sætter $\boldsymbol{a}(y) = (\log y, \log(1-y))$, altså på vektorformen

$$\exp(\boldsymbol{a}(y)'\boldsymbol{\theta} + c(\boldsymbol{\theta}) + d(y)).$$

Man kunne i princippet bygge en slags generaliseret lineær model på denne repræsentation, men man foretrækker at modellere $\mu = \mathrm{E}\, Y$ direkte, hvilket er ideen i resten af opgaven.

2° Gør rede for at hvis vi sætter $\phi = \alpha + \beta$, så kan vi repræsentere variansen som

$$\frac{1}{1+\phi}\mu(1-\mu).$$

Opskriv desuden de oprindelige parametre (α, β) som funktion af (μ, ϕ).

[7] I hjemmeopgaven vistes yderligere, at med $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$ gælder omskrivningen $(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = (\boldsymbol{T}\boldsymbol{Y})'(\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}')^{-1}\boldsymbol{T}\boldsymbol{Y}$. Dette overspringes her af pladshensyn.

[8] https://en.wikipedia.org/wiki/Beta_distribution

Ligheden med binomialfordelingen gør det nærliggende at opskrive modeller som er lineære i $\log \mu/(1-\mu)$ og hvor ϕ bruges som en underspredningsparameter. Sådanne modeller kan fittes med maximum likelihood, men også som en quasi-binomial model, hvor man bruger standard `glm()` teknikker.

Fra forelæsningerne ved vi, at unit deviance i en naturlig eksponentiel familie kan beregnes som

$$D(y,\mu) = 2\int_{\mu}^{y} \frac{y-u}{v(u)}\,du$$

og at loglikelihood-bidrag for den tilsvarende dispersionsfamilie[9] kan skrives

$$\ell(\mu;y,\psi) = -\frac{1}{2\psi}D(y,\mu) + k(y,\psi).$$

For modeller der ikke er naturlige eksponentielle familier kan man bruge de samme formler til at definere en *quasilikelihoodfunktion*. (Bemærk at man kun behøver at kende variansfunktionen for at definere unit deviance.)

3° Udfør beregningen af $D(\mu,y)$ konkret i tilfældet hvor variansfunktionen er

$$v(u) = u(1-u)$$

(slutresultatet er selvfølgelig velkendt fra binomialfordelingen.)

Vink: Det er lettere at finde det ubestemte integral hvis man benytter at $\frac{1}{u(1-u)} = \frac{1}{u} + \frac{1}{1-u}$.

Opgave 7

I `Ecdat` pakken findes datasættet `BudgetFood`. Variablen `wfood` indeholder den andel af husholdningens budget som anvendes på mad. Nogle få observationer har `wfood` sat til 0, hvilket kunne give problemer med estimationen, så vi vælger at udelade dem. Til denne opgave laver vi yderligere den restriktion at vi kun ser på husholdninger af størrelse 4.

```
library(Ecdat)
BF4 <- subset(BudgetFood, wfood > 0 & size==4)
```

Vi vil prøve at modellere `wfood` ved hjælp af betafordelingen, med udnyttelse af resultaterne fra opgave 6.

[9] Bemærk at ϕ (phi) her er omdøbt til ψ (psi) for ikke at komme i konflikt med notationen for betafordelingen.

1° Fit en intercept-only model, `wfood ~ 1`, som en generaliseret lineær model med "quasibinomial" familie. Beskriv `summary()` output fra modellen, specielt sammenhængen mellem dispersionsparameteren ψ i quasibinomialmodellen og præcisionsparameteren ϕ i betafordelingen og mellem interceptet i modellen og middelværdien μ.

Find et 95% konfidensinterval for interceptet, både via (quasi-)likelihood profiling og Wald test. Indtegn den fittede betafordeling på et histogram over `wfood`:

```
hist(BF4$wfood, freq=FALSE)
curve(dbeta(x, alpha, beta), add=TRUE)
```

(Omregning fra (μ,ϕ) til de sædvanlige formparametre (α,β) blev diskuteret i opgave 6.)

2° Når modellen er så simpel at den kun består af interceptet, bliver quasibinomialmetoden ækvivalent med momentmetoden, dvs at man kan finde parametrene ved at sætte den teoretiske middelværdi og varians lig med den empiriske `mean(y)`, henholdsvis `var(y)`. Eftervis dette i det konkrete eksempel. Benyt deltametoden til at finde standard error for det estimerede intercept ud fra s.e.$(\bar{y})$.

3° R pakken `betareg` kan bruges til at fitte tilsvarende modeller ved "rigtig" maximum likelihood. Fit igen en intercept-only model og sammenhold resultaterne med dem fra quasi-likelihood metoden.

4° Maximum likelihood estimation kan også udføres med `mle()` funktionen i `stats4` pakken. Den kræver at man laver en funktion der kan beregne den negative loglikelihood, efter devisen

```
mll <- function(mu=0.5, phi=1){ -sum(....)}
summary(mle(mll))
```

Gennemfør dette og vis at resultatet er ækvivalent med det fra `betareg()`.

5° Man kan let udvide modellen til en regressionsmodel både i `betareg()` og i quasibinomial-metodologien. Vi nøjes her med `betareg()` og med at se på effekten af den totale budgetstørrelse. Start med at tegne et scatterplot af `wfood` mod `log(totexp)` og diskuter hvad det viser.

6° Fit herefter en `betareg` model hvori `wfood` beskrives ved et polynomium i `log(totexp)`. Her kan man med fordel bruge `poly(...., n)` hvor et andengradspolynomium formentlig er tilstrækkeligt. Indtegn den fittede middelkurve på scatterplottet[10], og tilføj et approksimativt 90% prædiktionsbånd. Hertil kan man benytte

[10] Dette kan eventuelt, lidt primitivt, gøres med points(), i en anden farve og med et lille plottesymbol.

```
predict(...., type="response")})
predict(...., type="quantile", at=.....).
```

Forklar hvad det er "quantile"-varianten gør.

7° I modellen antages præcisionsparameteren ϕ at være konstant, specielt ikke afhængig af regressionsvariablen. For at checke antagelsen kan man plotte de kvadrerede Pearson residualer mod `log(totexp)` og indlægge en udglatningskurve som dem man ser i modelkontrolplots for `lm()`. Følgende skitse-kode kan give et udgangspunkt:

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

Hvad ser man på figuren?

8° `betareg()` tillader foruden modellering af middelværdien μ også modellering af ϕ (som default via lineære modeller med logaritmisk link funktion). Dette kan gøres ved at modificere modelformlen så den består af to dele adskilt af en lodret streg, altså af formen

```
y ~ mean model | variance model
```

Prøv dette med et andengradspolynomium i `log(totexp)` for variansleddet.

Indtegn igen den fittede kurve og prædiktionsbånd på scatterplot (evt. samme scatterplot som før for at spare plads), og tegn plottet med kvadrerede Pearson residualer og udglatningskurve. Diskuter resultatet.