# Projekt 2: BMI undersøgelse II

## Formaliteter, struktur og forventninger – 2. obligatoriske opgave

I dette projekt fortsætter vi med at analysere data fra BMI-undersøgelsen. Målet er at opstille en passende multipel lineær regressionsmodel for BMI. Opgaven skal i praksis løses ved hjælp af den statistiske software R. Rundt omkring i opgaven er der givet forslag til R-koden, men udover det er det en god idé at se på R-koden fra projekt 1 samt f.eks. kapitel 5 og 6 i bogen.

Besvarelsen skal dokumentere den gennemførte analyse ved tabeller, grafer, passende matematisk notation og tekst der beskriver analysens resultater. Relevante grafer og tabeller skal indgå i sammenhæng med teksten – ikke som bilag. Præsenter resultaterne fra jeres analyser på samme måde, som I ville videreformidle dem til andre fagfæller. Inddel besvarelsen i et underafsnit for hvert af de stillede spørgsmål.

Besvarelsen skal afleveres som pdf-fil. R-kode bør ikke indgå i besvarelsen, men vedlægges som bilag (i form af en .R-fil). Besvarelsen samt bilag afleveres under Opgaver/Assignments på CampusNet ved:
`Opgaver > Aktive opgaver > Obligatorisk opgave nr. 2:`
`BMI undersøgelse II > Besvar > Besvar opgave`

En samlet besvarelse bør ikke overstige 6 sider (ekskl. plots, tabeller og bilag). En side udgør 2400 anslag.

Grafer og tabeller kan IKKE stå alene - det er altså vigtigt, at I beskriver og fortolker outputtet fra R med ord.

Når I bliver bedt om at angive en formel, indsætte tal og derefter foretage en beregning er det vigtigt, at I viser I har gjort dette ved at inkludere nogle mellemregninger. (Disse

steder er det ikke nok at anføre resultater aflæst i R). Husk også at et hypotesetest består af følgende elementer: Angivelse af hypotese og signifikansniveau ($\alpha$), teststørrelse inkl. dennes fordeling og $p$-værdi, samt en konklusion med ord.

Grafer og tabeller indgår ikke i opgørelsen af besvarelsens længde. Det er dog IKKE i sig selv en fordel at medtage mange plots, hvis de ikke er relevante!

I må gerne arbejde sammen i grupper, men besvarelsen af opgaven skal skrives individuelt. Spørgsmål omkring projektet kan rettes til hjælpelæren, se retningslinjerne på siden *Projects* på kursets hjemmeside.

## Data

Indlæs datasættet `bmi2_data.csv`. Følgende R-kode kan benyttes:

```
# Indlæs 'bmi2_data.csv' filen med data
D <- read.table("bmi2_data.csv", header = TRUE, sep = ";")
```

Datasættet til dette projekt omfatter observationer af følgende variable:

- id: Respondentens nr. (kan bruges til identifikation)
- `bmi`: Respondentens BMI (i $\mathrm{kg/m^2}$)
- `age`: Respondentens alder (i år)
- `fastfood`: Hyppighed af respondentens besøg ved fastfood restauranter (dage/år)

Variablen `fastfood` var oprindeligt en kategoriseret variabel, men den er blevet rekodet (se projekt 1), så den kan anvendes som en kontinuert variabel. Vi vil i det følgende referere til den som *"fastfood-forbrug"*.

Tilføj følgende variabel med logaritme-transformeret BMI til datasættet:

```
# Tilføj log-BMI til datasættet
D$logbmi <- log(D$bmi)
```

# Statistisk analyse

a) Lav en kort deskriptiv analyse og opsummering af data for variablene `logbmi`, `age` og `fastfood`. Inkluder scatterplots af logaritmen til BMI mod de to andre variable, samt histogrammer og boxplots af alle tre variable. Der skal også være en tabel med opsummerende størrelser, som for hver variabel inkluderer antal observationer, gennemsnit, standardafvigelse, median, samt 25%- og 75%-fraktiler.

I denne opgave skal den statistiske model kun opstilles på baggrund af de første 840 observationer i datasættet. De sidste syv observationer skal vi senere bruge til at vurdere modellens evne til at prædiktere. Benyt f.eks. følgende R-kode til at dele datasættet op i et nyt deldatasæt, der benyttes til at estimere modellen (`D_model`), og et der benyttes til validering af modellens prædiktionsevne (`D_test`):

```
# Deldatasæt med de første 840 observationer (til model)
D_model <- subset(D, id <= 840)

# Deldatasæt med de sidste 7 observationer (til validering)
D_test <- subset(D, id >= 841)
```

b) Opstil en multipel lineær regressionsmodel med logaritmen til BMI som responsvariabel ($Y_i$), og med alder og fastfood-forbrug som forklarende variable (hhv. $x_{1,i}$ og $x_{2,i}$). Husk at angive forudsætningerne/de statistiske antagelser for modellen. (Se bemærkning 5.6, ligning (6-1) og eksempel 6.1).

c) Estimer modellens parametre, som består af regressionskoefficienterne, her kaldet $\beta_0$, $\beta_1$, $\beta_2$, og residualernes varians, $\sigma^2$. Brug evt. følgende R-kode:

```
# Estimer multipel lineær regressionsmodel
fit <- lm(logbmi ~ age + fastfood, data = D_model)

# Vis estimerede parametre mm.
summary(fit)
```

Giv en fortolkning af estimaterne $\hat{\beta}_0$, $\hat{\beta}_1$ og $\hat{\beta}_2$, hvor du forklarer, hvad de siger om relationen mellem logaritmen til BMI og de to forklarende variable i modellen. (Se bemærkning 6.14). Angiv også de estimerede standardafvigelser for $\hat{\beta}_0$, $\hat{\beta}_1$ og $\hat{\beta}_2$, frihedsgraderne anvendt til estimatet af residualernes varians $\hat{\sigma}^2$, samt modellens forklarede varians, $R^2$.

d) Foretag modelkontrol for at undersøge, om forudsætningerne for modellen (modellens antagelser) er opfyldte. Benyt de plots, der kan laves ved hjælp af R-koden nedenfor, som udgangspunkt for din vurdering. (Se afsnit 6.4 om residualanalyse).

```
# Plots til modelkontrol

# Observationer mod fittede værdier
plot(fit$fitted.values, D_model$logbmi, xlab = "Fittede værdier",
       ylab = "log(BMI)")

# Residualer mod hver af de forklarende variable
plot(D_model$FORKLARENDE_VARIABEL, fit$residuals,
        xlab = "INDSÆT TEKST", ylab = "Residualer")

# Residualer mod fittede værdier
plot(fit$fitted.values, fit$residuals, xlab = "Fittede værdier",
     ylab = "Residualer")

# Normal QQ-plot af residualerne
qqnorm(fit$residuals, ylab = "Residualer", xlab = "Z-scores",
       main = "")
qqline(fit$residuals)
```

e) Angiv formlen for et 95% konfidensinterval for koefficienten for alder, her kaldet $\beta_1$. (Se metode 6.5). Indsæt tal i formlen og beregn konfidensintervallet. Benyt derefter nedenstående R-kode til at kontrollere resultatet og til at bestemme konfidensintervaller for de to andre koefficienter i modellen.

```
# Konfidensintervaller for modellens koefficienter
confint(fit, level = 0.95)
```

f) Man er interesseret i, om $\beta_1$ kunne have værdien 0.001. Opstil den tilsvarende hypotese. Anvend signifikansniveauet $\alpha = 0.05$. Angiv formlen for den relevante teststørrelse (se metode 6.4), indsæt tal og beregn teststørrelsen. Angiv fordelingen af teststørrelsen (inkl. frihedsgrader), beregn $p$-værdien og konkluder.

g) Undersøg ved *backward selection* om modellen kan reduceres. (Se eksempel 6.13). Husk at reestimere modellen undervejs, hvis der kan foretages reduktion af modellen. Angiv slutmodellen og estimater for dens parametre.

h) Tag udgangspunkt i din slutmodel fra forrige spørgsmål. Bestem prædiktioner og 95% prædiktionsintervaller for logaritmen til BMI for hver af de syv observationer i valideringsdatasættet (`D_test`). Se eksempel 6.8, metode 6.9 og R-koden nedenfor. Sammenlign prædiktionerne med de observerede log-BMI værdier for disse syv observationer og lav en vurdering af modellens evne til at prædiktere.

```
# Prædiktioner og 95% prædiktionsintervaller
pred <- predict(SLUTMODEL, newdata = D_test, interval = "prediction",
                level = 0.95)

# Observerede værdier sammen med prædiktioner
cbind(id = D_test$id, logbmi = D_test$logbmi, pred)
```

Dvs. skriv ikke formlerne ind i rapporten, men istedet, at I har brugt R funktionen `predict` til beregningerne. Formlerne kræver en matrix formulering, som rækker ud over pensum (for at udlede formlerne kan ligningerne (6-48) og (6-49) bruges sammen med udledningerne der fører til ligningerne (5-57) og (5-58)).