Eksempel vejledende besvarelse

Bemærk at jeg i denne løsning ikke altid har output med. Tanken er, at I skal se løsningen og selv prøve at køre kommandoerne (og dermed undgår jeg også at dette dokument bliver alt for langt).

Opgave 1

1.1

Vi indlæser data og undersøger, om der er nogle missing values:

```
d <- read.csv('http://publicifsv.sund.ku.dk/~sr/regression/datasets/vitamind.csv')
```

Da der ikke er angivet NA's for nogle af variablene i ovenstående summary, har vi altså ikke nogen missing values.

Antallet af deltagere fra hvert land fås med

```
table(d$country)
```

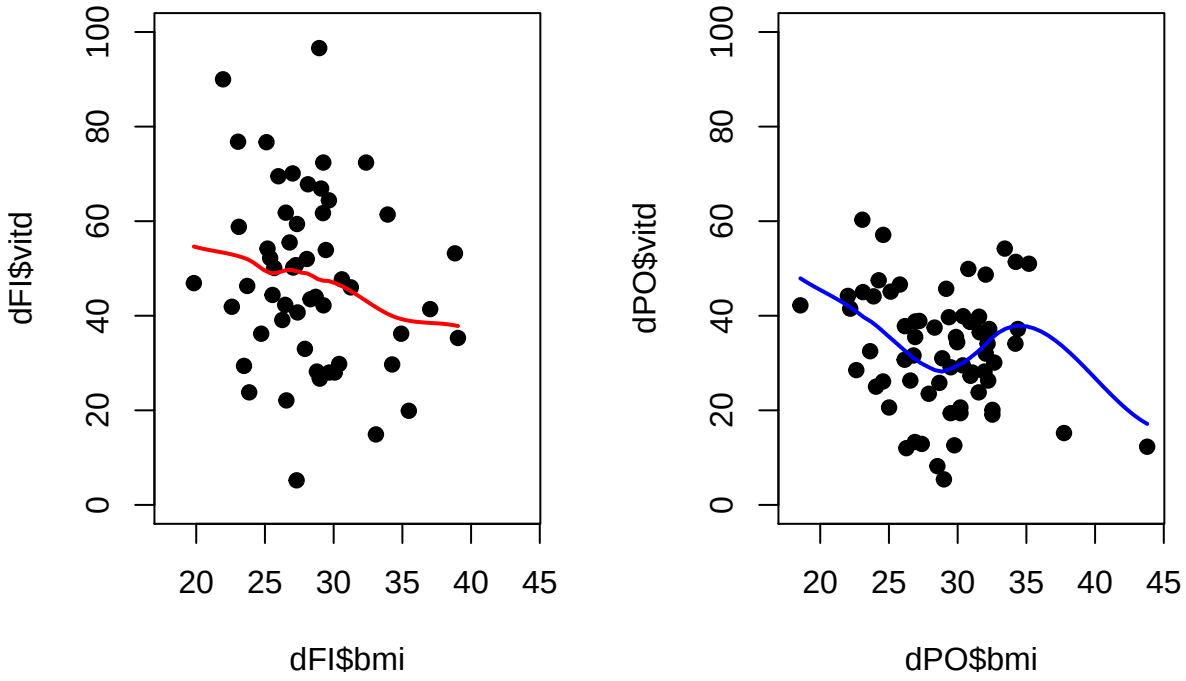
dvs 54 finske og 65 polske kvinder.

1.2

For hvert land laver vi en illustration af associationen mellem BMI og vitamin D. Vi sørger for at gøre plottene sammenlignelige ved at vælge samme x- og y-akse og sætte plottene ved siden af hinanden med `par(mfrow=c(1,2))`

```
dFI <- subset(d, country==2)
dPO <- subset(d, country==6)

par(mfrow=c(1,2))
scatter.smooth( dFI$vitd ~ dFI$bmi, pch=19,
                xlim=c(18,44),ylim=c(0,100), lpars=c(lwd=2,col=2))
scatter.smooth( dPO$vitd ~ dPO$bmi, pch=19,
                xlim=c(18,44),ylim=c(0,100), lpars=c(lwd=2,col=4) )
```



Associationen ser meget lineær ud for Finland, hvorimod den for Polen ikke er helt så entydig. Vi noterer også, at spredningen ser ud til at være større for de finske kvinder.

1.3

Vi udfører en simpel lineær regression for hvert land for sig og husker også at bestemme konfidensinterval:

```
lm0.FI <- lm( vitd ~ bmi, data=dFI )
summary(lm0.FI)
confint(lm0.FI)

lm0.PO <- lm( vitd ~ bmi, data=dPO )
summary(lm0.PO)
confint(lm0.PO)
```

I begge tilfælde ser vi en negativ association, som dog ikke er signifikant.

Vi undersøger om antagelsen om normalfordeling, varianshomogenitet og linearitet er opfyldt for hvert land for sig:

Finland:

```
par(mfrow=c(1,2))
plot(lm0.FI,which=2)
plot(lm0.FI,which=1)
```

Polen:

```
par(mfrow=c(1,2))
plot(lm0.PO,which=2)
plot(lm0.PO,which=1)
```

Begge qq-plots er nydelige, mens der er lidt slinger i residualplottet for Polen. Der er dog ikke tegn på, at variansen falder eller stiger med værdien af de fittede (prædikterede) værdier, hvorfor varianshomogeniteten ikke ser ud til at kunne anfægtes. At kurven krummer kan skyldes, at associationen mellem BMI og vitamin

D muligvis ikke er lineær som funktion af BMI for de polske kvinder (jvf scatter plottet ovenfor). Denne afvigelse er dog næppe signifikant.

Vi supplerer med et formelt test for linearitet for hver gruppe (bemærk her at jeg ikke definerer kvadratet på BMI i datasættet, men sætter R til selv at regne den ud i `lm()` ved at sætte et `I()` omkring):

```
summary( lm( vitd ~ bmi+I(bmi*bmi), data=dFI ) )
summary( lm( vitd ~ bmi+I(bmi*bmi), data=dPO ) )
```

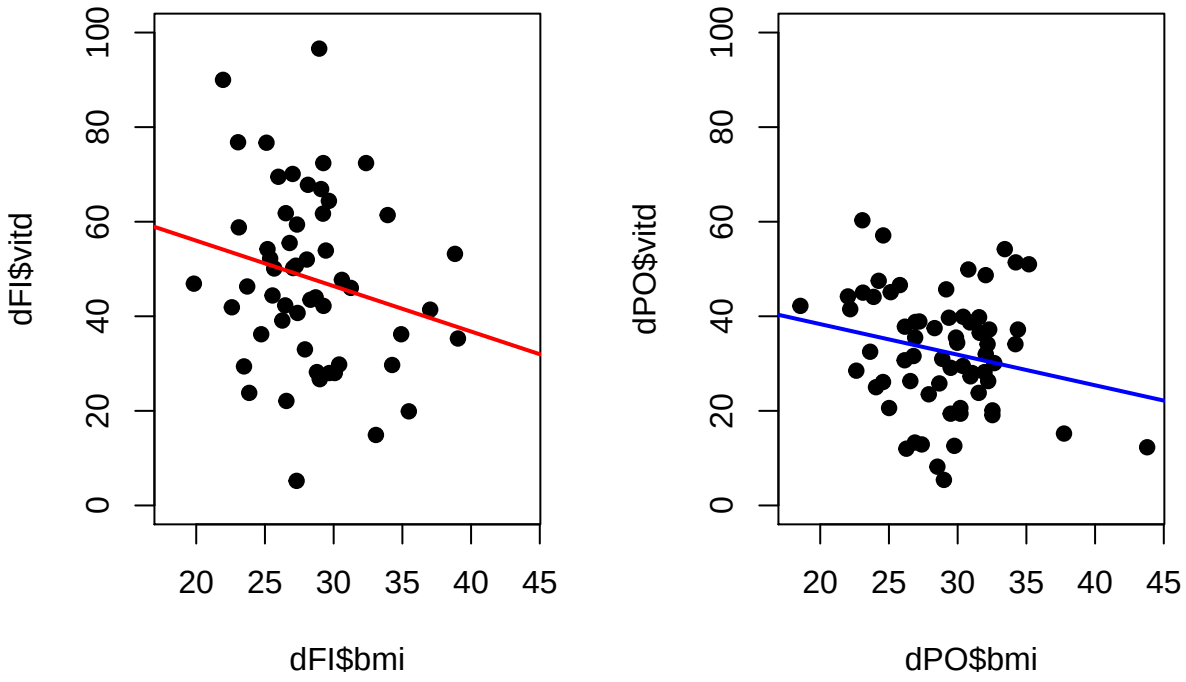
Bemærk at et andengradspolynomie ikke er den bedste funktion til at finde en afvigelse fra linearitet for de polske kvinder grundet de to kvinder med meget høj BMI og meget lav vitamin D.

1.4

Vi laver nu en illustration af de to estimerede regressionslinier. Igen stiller vi plottene ved siden af hinanden, så linierne er lettere at sammenligne:

```
par(mfrow=c(1,2))
plot( dFI$vitd ~ dFI$bmi, pch=19, xlim=c(18,44),ylim=c(0,100) )
abline( lm0.FI, lwd=2, col=2 )

plot( dPO$vitd ~ dPO$bmi, pch=19, xlim=c(18,44),ylim=c(0,100) )
abline( lm0.PO, lwd=2, col=4 )
```



Vi ser at linierne ser nogenlunde parallelle ud, men at linien for de finske kvinder ser ud til at ligge højere.

1.5

Vi formulerer nu en multivariabel model på det samlede materiale. Land er en kvalitativ variabel og skal derfor indgå som en faktor.

```
lm1 <- lm( vitd ~ factor(country)+bmi, data=d )
summary(lm1)
```

Vi bemærker at der er en forskel på 15 i vitamin D niveauet mellem de to lande (Polen ligger lavest). Forskellen er højsignifikant. Vi bemærker også, at hældningen nu er blevet signifikant - vitamin D falder med 0.78 for en stigning af BMI på 1. Det skyldes, at vi nu kan estimere hældningen pba et større materiale, end når vi ser på hvert materiale for sig.

Husk konfidensintervaller her når tallene rapporteres!

En anden måde at fitte modellen på er ved at lave en dummy-variabel for land. En dummy-variabel er en variabel, som kun har kode 0 og 1. For en sådan variabel er det ligegyldigt om vi betragter den som kvantitativ og laver regression på den eller sætter den ind i analysen som en faktor (hvorfor??).

Hvis vi vil have Finland som reference kan vi definere

```
d$polen <-0
d$polen[ d$country==6 ] <- 1
```

og derefter benytte denne variabel som en regressionsvariabel:

```
lm1a <- lm( vitd ~ polen+bmi, data=d )
summary(lm1a)
```

Vi får samme resultat, hvis vi sætter `factor()` omkring Polen (fordi laveste niveau 0 bliver referencen).

1. Sammenligning af en finsk og en polsk kvinde hver med et BMI på 25: `a+c*25 - (a+b+c*25) = -b = 14.95` således at den finske har et forventet vitamin D niveau som er 15 enheder højere end den finske. Vi kan lave den samme sammenligning for to kvinder hver især med et BMI på 30 - og får igen de 14.95.
2. Sammenligning af to finske kvinder med BMI på hhv 25 og 30: `a+c*30 - (a+c*25) = 5*c = 5*(-.78) = -3.92`. Dvs den 'tykke kvinde' har et forventet vitamin D niveau som er 3.9 lavere end den 'normalvægtige kvinde'.
3. Samme sammenligning for Polske kvinder: `a+b+c*30 - (a+b+c*25) = 5*c = 5*(-.78) = -3.92`.

1.6

Vi inkluderer nu også alder i modellen:

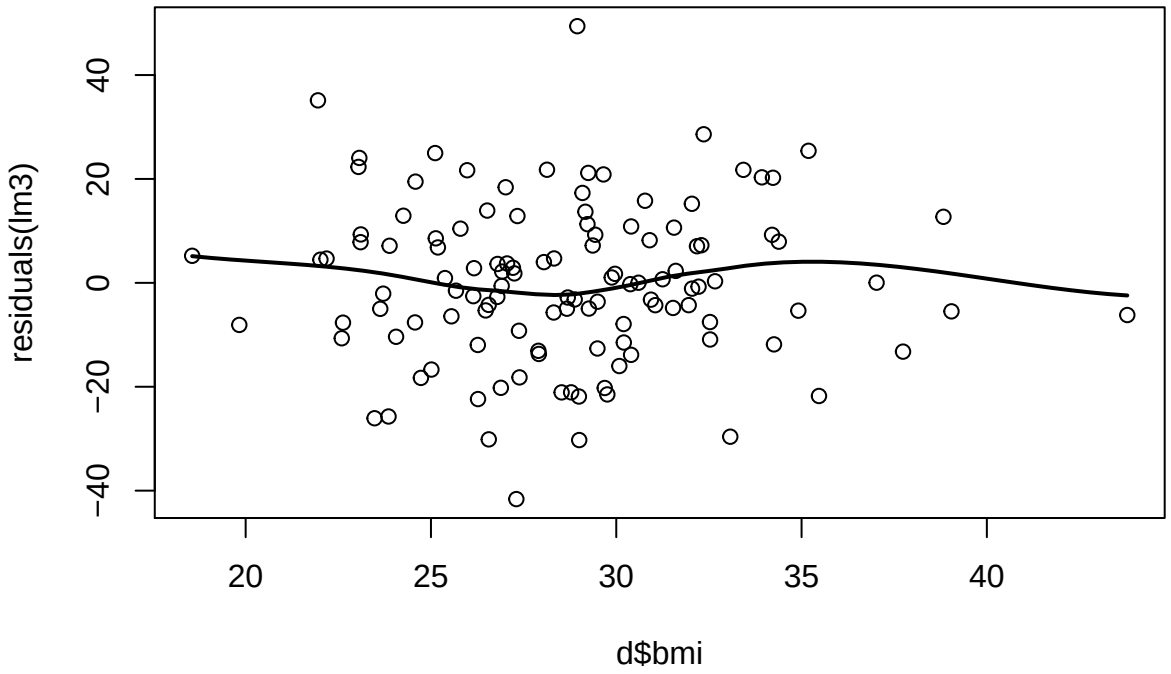
```
lm3 <- lm( vitd ~ factor(country) + bmi +age , data=d )
summary(lm3)
```

og ser at der - justeret for land og bmi - ikke er nogen association mellem alder og vitamin D niveau.

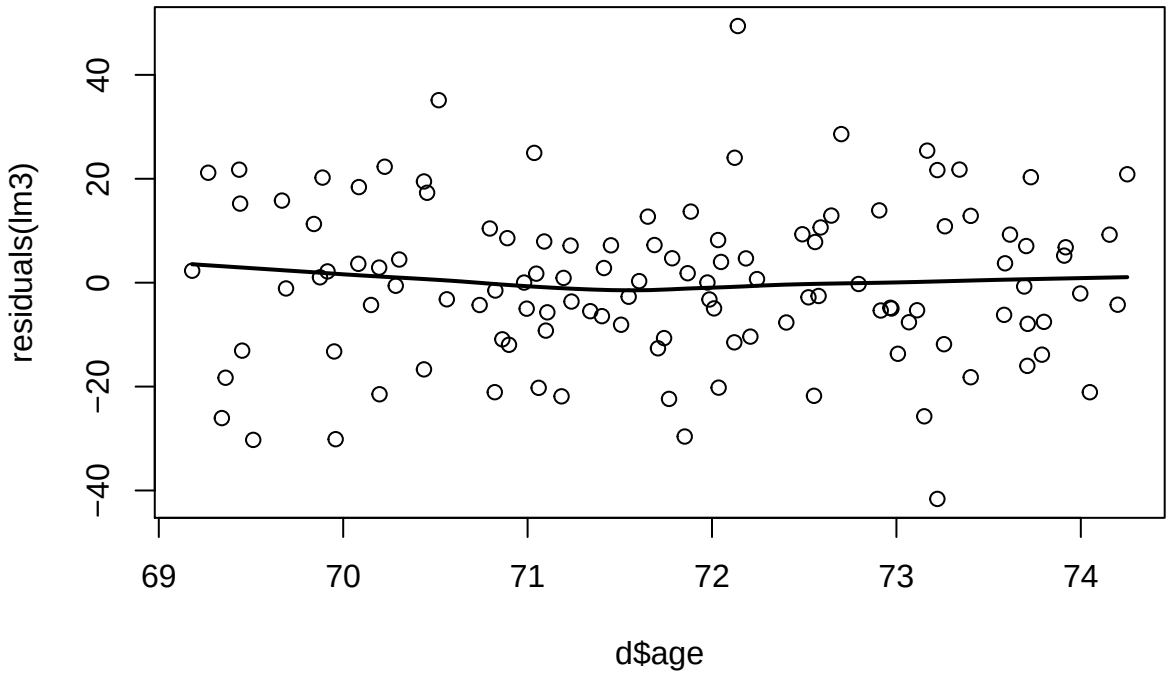
1.7

Vi har antaget, at både alder og BMI kan indgå lineært i modellen. Vi har ovenfor undersøgt BMi for hvert land for sig i en model hvor vi ikke to højde for alder. Vi udfører nu et check baseret på hele materialet. Vi ser derfor på residualplottene:

```
scatter.smooth( residuals(lm3) ~ d$bmi, lpars = c(lwd=2) )
```




```
scatter.smooth( residuals(lm3) ~ d$age, lpars = c(lwd=2) )
```



hvor vi ikke ser nogen tydelig tendens til krumninger (måske lidt slinger for BMI (som nok stammer fra de polske kvinder)). Vi supplerer med kvadratledstest (her tilføjes begge kvadratled på en gang, og vi forsøger at fjerne dem en af gangen):

```
lm4 <- lm( vitd ~ factor(country) + bmi + age + I(age*age) + I(bmi*bmi) , data=d )
summary( lm4 )
```

og ser at vi må fjerne kvadratleddet for alder med p=0.85.

```
lm5 <- lm( vitd ~ factor(country) + bmi + age + I(bmi*bmi) , data=d )
summary( lm5 )
```

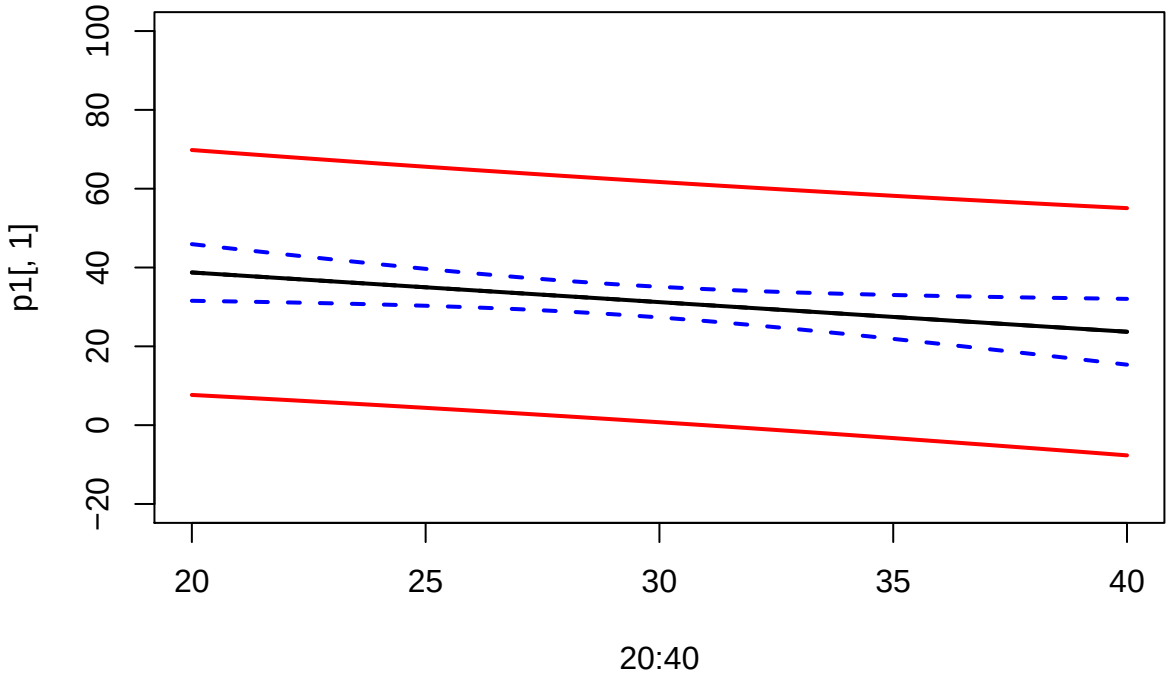
Heller ingen problemer med lineariten for BMI. Vores endelige model er derfor den, hvor både alder og BMI indgår lineært (`lm3`).

1.8

Vi vil nu prædiktere en regressionsline for polske kvinder på 72 år. Vi illustrerer den for passende valgte BMI-værdier, her mellem 20 og 40 (skal helst passe med data range, så vi ikke ekstrapolerer vores model for meget):

```
pdata <- data.frame( age=72, country=6, bmi=20:40)
p1 <- predict( lm3, newdata=pdata, interval='prediction')
p2 <- predict( lm3, newdata=pdata, interval='confidence')

plot( 20:40, p1[,1], ylim=c(-20,100), type='l')
matlines( 20:40, p1, lwd=2, col=c(1,2,2), lty=1)
matlines( 20:40, p2, lwd=2, col=c(1,4,4), lty=2)
```



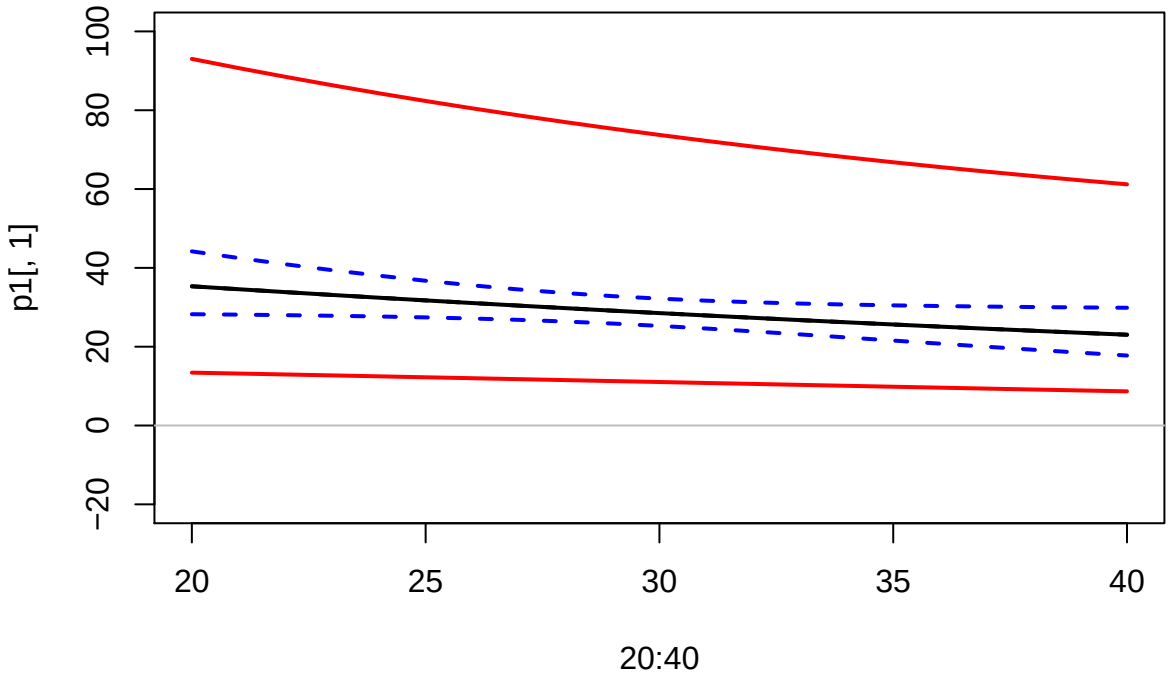
Vi ser, at referenceområdet i den tunge ende af BMI-skala'en ryger under 0. Det giver selvfølgelig ikke mening, da man ikke kan have negative vitamin D niveauer.

En måde at undgå negativt prædiktionsinterval er ved at logaritme-transformere outcome. Når man regner tilbage til den oprindelige skala med antilog får man altid noget positivt. Vi kan prøve:

```
d$logvitd <- log(d$vitd)
lm3a <- lm( logvitd ~ factor(country) + bmi +age , data=d )
p1 <- exp( predict( lm3a, newdata=pdata, interval='prediction') )
p2 <- exp( predict( lm3a, newdata=pdata, interval='confidence') )

plot( 20:40, p1[,1], ylim=c(-20,100), type='l')
matlines( 20:40, p1, lwd=2, col=c(1,2,2), lty=1)
```

```r
matlines( 20:40, p2, lwd=2, col=c(1,4,4), lty=2)
abline( h=0, col='gray')
```



og ser at referenceområdet nu holder sig pænt over 0.

NB: Selve den prædikterede kurve (den sorte) kan ikke længere fortolkes som en middelværdi - men svarer faktisk til medianen! Undersøger du hvordan modellen passer med logaritmetransformationen, vil du se at normalfordelingsantagelsen bliver problematisk. Og derfor giver referenceområdet ikke mening, fordi dette beregnes under forudsætningen af en normalfordeling.

Vi kan derfor ikke lave et fornuftigt referenceområde på baggrund af nogle af disse modeller!