STK1000: Løsningsforslag Uke 45

2022

Check In - Oppgave 10.6

Feilmarginen (margin of error) er halvparten av bredden til konfidensintervallet. For $x = 9.0$ har vi derfor $m = (24.4 - 23)/2 = 0.7$.

Konfidensintervallet er smalest for $\bar{x}$ og øker når vi beveger oss vekk fra $\bar{x}$. Nå kan jeg ikke sett at $\bar{x}$ er oppgitt noe sted, men fra figur 10.7 ser vi at konfidensintervallet er smallest rundt $x = 9$ så $\bar{x} \approx 9$. Derfor vil marginen for $x = 11$ være større enn for $x = 9$.

Check In - Oppgave 10.7

Marginer for prediksjonintervallet er $(31 - 16.4)/2 = 7.3$.

Når n vokser vil man forvente at marginen blir mindre. For et konfidensintervall for μ vil marginen gå mot 0 når n øker, men for et prediksjonsintervall for y vil ikke dette være tilfellet. Dette er fordi punktene er spredd rundt linja, og uansett hvor godt vi estimerer linja vår, vil det alltid være usikkerhet knyttet til den individuelle variasjonen til en ny observasjon. Altså, et prediksjonsintervall inneholder både usikkerheten til den estimerte linja og usikkerheten til hvordan punktene er spredd rundt linja.

For $n = 400$ kan vi derfor utelukke at marginen er dobbelt så stor. Det er ikke like lett å si om marginen vil bli halvparten så stor eller uendret. Med tanke på hvor liten marginen til linja var i Oppgave 10.6, kan det virke som mesteparten av usikkerheten i prediksjonsintervallet her kommer fra spredningen av punktene rundt den sanne men ukjente populasjonsregresjonslinja. Det er derfor fristende å påstå at prediksjonsintervallet ikke vil forandre seg spesielt mye for økende n.

Oppgave 10.44 (R)

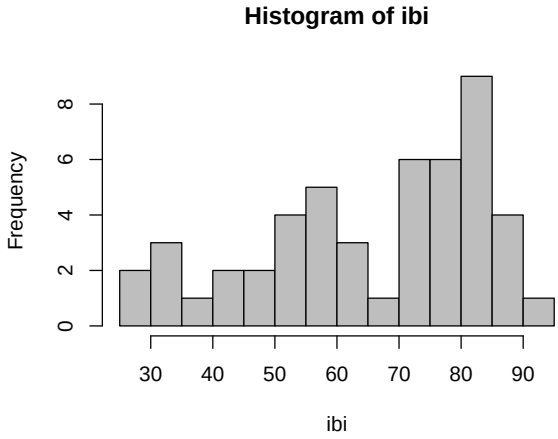
a)

Vi laster inn dataene og bruker `summary` og `histogram` for å studere IBI og areal.

```
data = read.csv('../ips10e_csv_data_sets/ips10e_ch10_csv_data_sets/EX10-044IBI.csv')
ibi = data$IBI
area = data$Area
summary(ibi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   55.00   71.00   65.94   82.00   91.00
```

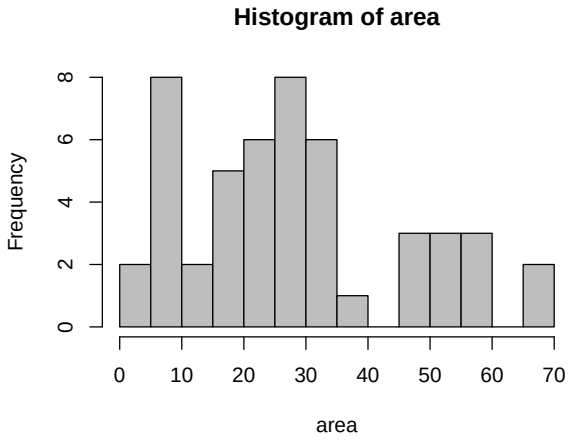
```
hist(ibi, col = 'gray', breaks = 20)
```




```
summary(area)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   16.00   26.00   28.29   34.00   70.00
```

```
hist(area, col = 'gray', breaks = 20)
```

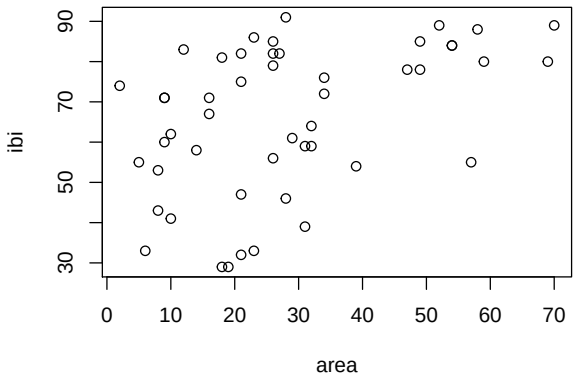


Vi ser fra at IBI er litt venstreforskjøvet mens areal er litt høyreforskjøvet, og ingen av dem er tilsynelatende normalfordelt. Det ser heller ikke ut til å være noen outliers her.

b)

Vi plotter IBI mot areal og ser at det er en svak positiv trend. Det er litt større spredning for lave verdier av areal, og det er tilsynelatende ikke noe outliere.

```
plot(area, ibi)
```



c)

Modellen er $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, der $\epsilon_i \sim N(0, \sigma^2)$ altså normalfordelt med forventning 0 og varians σ^2.

d)

Vi har $H_0 : \beta_1 = 0$ mot $H_a : \beta_1 \neq 0$.

e)

```
model = lm(ibi~area)
summary(model)
```

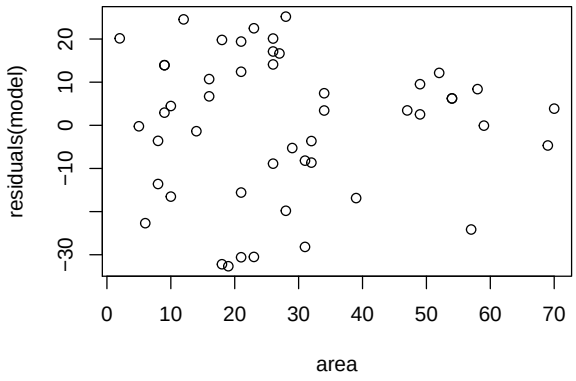
```
##
## Call:
## lm(formula = ibi ~ area)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.666  -8.887   3.432  12.414  25.193
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  52.9230     4.4835  11.804 1.17e-15 ***
## area          0.4602     0.1347   3.415  0.00132 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.53 on 47 degrees of freedom
## Multiple R-squared:  0.1988, Adjusted R-squared:  0.1818
## F-statistic: 11.67 on 1 and 47 DF,  p-value: 0.001322
```

Fra sammendraget ser har vi estimert linjen $\hat{y} = 52.9230 + 0.4602\,x$, med $s = 16.53$ og en p-verdi for hypotesetesten i d) på 0.00132 (som er rimelig lavt). Men hvis vi ser på r^2 ser vi at bare 19.88% av variasjonen i IBI er forklart av arealet. Dette stemmer bra med at det er en svak positiv sammenheng.

f)

Vi plotter residualene mot arealet og ser at det er større variasjon for mindre verdier av areal. Dette strider mot antagelsen om at residualene har same σ.

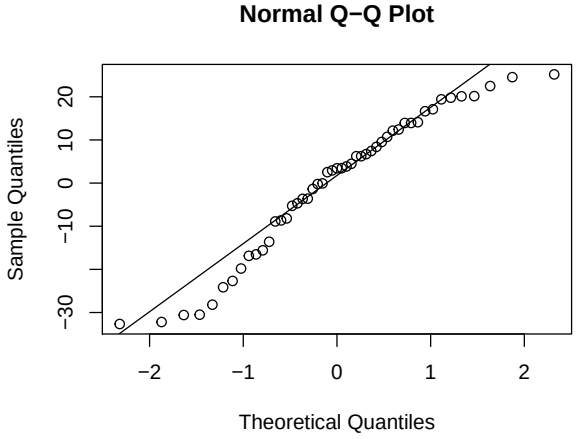
```
plot(area, residuals(model))
```



g)

Som vi sa i f) er størrelsen på residualene avhengig av arealet, og antagelsen om normalfordelingen passer derfor ikke. Vi kan også vise dette med å se på et QQ Plot av residualene, hvor vi ser at punktene ikke følger linjen spesielt bra:

```
qqnorm(residuals(model))
qqline(residuals(model))
```



h)

Vi antok i c) at $\epsilon_i \sim N(0, \sigma^2)$, altså at residualene var normalfordelte. Vi har konkludert med at dette ikke stemmer og derfor stemmer ikke modellantagelsene våre.

Oppgave 10.48 (R)

a)

Fra forelesningsslidene om Kapittel 10.1 får vi oppgitt R-kode for å regne ut et 95% konfidensintervall for gjennomsnittlig respons (intervallet er gitt av (`lwr`, `upr`)):

```
predict(model, newdata = data.frame(area=40), interval='confidence')
```

```
##        fit      lwr      upr
## 1 71.32916 65.61416 77.04417
```

b)

Et tilsvarende prediksjonsintervall finnes også i samme forelesningsslides

```
predict(model, newdata = data.frame(area=40), interval='prediction')
```

```
##        fit      lwr    upr
## 1 71.32916 37.57836 105.08
```

c)

For mange bekker med areal på 40 km^2 vil gjennomsnittlig IBI trolig være mellom 65.61 og 77.04. De fleste IBIene for de individuelle bekkene vil trolig mellom 37.78 og 105.08.

d)

Vi kan ikke anta (med sikkerhet) at disse resultatene generaliserer til bekker utenfor Ozark Highland, ettersom regionen trolig spiller en rolle på IBI-verdiene. Med andre ord, det er urimelig å anta at vannkvaliteten er uavhengig av lokasjonen.

Eksamen 2015 oppgave 3 a,b,c, Eksamen 2012 oppgave 3

Løsningsforslag til eksamensoppgavene er på emnets semesterside https://www.uio.no/studier/emner/matnat/math/STK1000/oppgaver/losningsforslag/