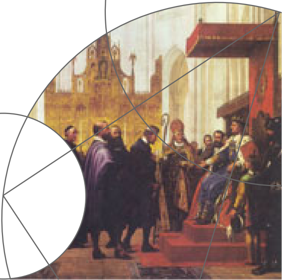
Modelkontrol og prædiktion

Anders Tolver
Institut for Matematiske Fag



I dag

Formiddag:

- Statistiske modeller (repetition/oversigt)
- Modelkontrol
- Prædiktion af nye observationer

Der er ingen Quiz til kursusuge 4. Lav i stedet Quiz 1-3 eller brug tid på den frivillige afleveringsopgave.

Det er muligt, at jeg uploader video som samler op på

- dele af sidste slides (31-41) fra mandag d. 25/9-2023.

Ikke noget nyt stof på mandag, men derimod opsamling, repetition og eksempler.

Overblik

Vi skal have "'udfyldt"' følgende skema over modeller (rækker) og statistiske begreber (søjler):

	Intro	Model	Est.+SE	KI	Test	Kontrol	Præd.
En stikprøve	✓	✓	✓	✓	✓	✓	nu
Ensidet ANOVA	✓	✓	✓	✓	✓	nu	nu
Lineær regr.	✓	✓	✓	✓	✓	nu	nu
To stikprøver	✓	✓	✓	✓	✓	nu	nu
Multipel regr.							
Tosidet ANOVA							

Statistiske begreber

- Population og stikprøve
- Gennemsnit, stikprøvespredning, median, kvartiler
- Statistisk model og parametre
- Estimater og standard error (SE) for estimater
- Konfidensinterval
- Hypotesetest
- Modelkontrol (residualplot, QQ-plot)
- Prædiktion, prædiktionsintervaller

Liste er faktisk færdig nu!

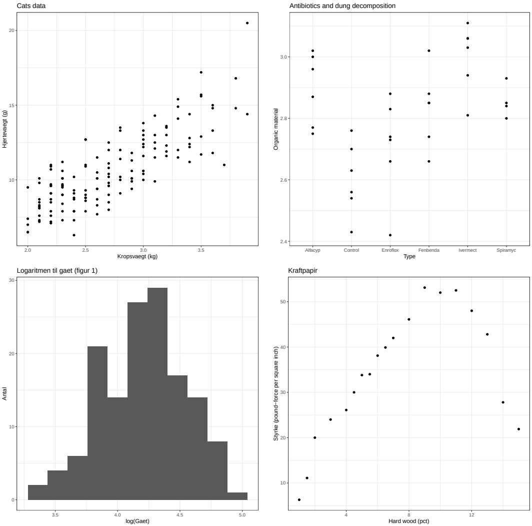
Intro/motivation: forskellige data og statistiske modeller

Dataeksempler

Et par dataeksempler som vi har diskuteret tidligere (+ et nyt):

- Fra `MASS` R-pakken:
  Sammenhæng mellem hjertevægt og kropsvægt for katte
- Eksempel 3.2: Gødning og antibiotika
- Gæt på antal punkter i figur 1:
  Data fra 143 studerende på StatData1 i 2017
- **Nyt Eksempel 8.3:**
  Styrke af kraftpapir ved forskelligt indhold af hårdt træ

Dataeksempler



Samme struktur af modellerne

De modeller vi har snakket om indtil videre, har **samme struktur:**

$$y_i = \text{middelværdi} + e_i$$

hvor restleddene $e_1, \ldots, e_n$ er iid. $N(0, \sigma^2)$, dvs. uafhængige og allesammen normalfordelt med middelværdi 0 og spredning σ.

Samme struktur af modellerne

De modeller vi har snakket om indtil videre, har **samme struktur:**

$$y_i = \text{middelværdi} + e_i$$

hvor restleddene $e_1, \ldots, e_n$ er iid. $N(0, \sigma^2)$, dvs. uafhængige og allesammen normalfordelt med middelværdi 0 og spredning σ.

Eksempler:

- Katte (**lineær regression**): $Hwt_i = \alpha + \beta \cdot Bwt_i + e_i$
- Gødning (**ensidet ANOVA**): $\text{org}_i = \alpha_{\text{type}_i} + e_i$
- Gæt på punkter (**en stikprøve**): $\log(\text{gæt}_i) = \gamma + e_i$.
- Kraftpapir (**kvadratisk regression**): $\text{str}_i = \alpha + \beta_1 \cdot \text{hw}_i + \beta_2 \cdot \text{hw}_i^2 + e_i$

Statistisk inferens

Metoder:

- **Middelværdiparametre estimeres** med mindste kvadraters metode (least squares)
- **Spredningen estimeres** via **residualkvadratsummen**
- **95% konfidensintervaller:**
  estimat $\pm\ t_{0.975,n-p} \cdot \mathrm{SE}(\text{estimat})$ hvor p er antalle af parametre i middelværdien
- **Hypotesetest** udføres som t-test eller F-test

Statistisk inferens

Metoder:

- **Middelværdiparametre estimeres** med mindste kvadraters metode (least squares)
- **Spredningen estimeres** via **residualkvadratsummen**
- **95% konfidensintervaller:** estimat $\pm\, t_{0.975,n-p} \cdot \mathrm{SE}$(estimat) hvor p er antalle af parametre i middelværdien
- **Hypotesetest** udføres som t-test eller F-test

Standard errors, konfidensintervaller og hypotestest er **kun valide hvis modelantagelserne er OK.**

Derfor er det vigtigt at lave **modelkontrol!**

Modelkontrol

Hvad er antagelserne egentlig?

Model:

$$y_i = \text{middelværdi} + e_i$$

hvor restleddene $e_1, \ldots, e_n$ er iid. $N(0, \sigma^2)$, dvs. uafhængige og allesammen normalfordelt med middelværdi 0 og spredning σ.

Hvad er antagelserne egentlig?

Model:

$$y_i = \text{middelværdi} + e_i$$

hvor restleddene $e_1, \ldots, e_n$ er iid. $N(0, \sigma^2)$, dvs. uafhængige og allesammen normalfordelt med middelværdi 0 og spredning σ.

Antagelserne, pindet ud:

1. Alle e_i (eller y_i) er **normalfordelte**
2. **Middelværdien** har den rette form
3. Alle e_i (eller y_i) har **samme spredning**
4. $e_1, \ldots, e_n$ (eller $y_1, \ldots, y_n$) **uafhængige**

Rimeligheden af de tre første antagelser — men ikke den sidste — kan undersøges vha. data.

Antagelse 4: Uafhængighed

Antagelsen on uafhængighed er mest et spørgsmål om **designet af eksperimentet** eller **dataindsamlingen.**

- Er der flere målinger på samme eksperimentelle enhed (person, mark, dyr, plante, ...)?
- Er individerne i familie med hinanden?
- Hører observationerne naturligt sammen i grupper (som ikke behandlingsgrupperne)?

Antagelse 4: Uafhængighed

Antagelsen on uafhængighed er mest et spørgsmål om **designet af eksperimentet** eller **dataindsamlingen.**

- Er der flere målinger på samme eksperimentelle enhed (person, mark, dyr, plante, ...)?
- Er individerne i familie med hinanden?
- Hører observationerne naturligt sammen i grupper (som ikke behandlingsgrupperne)?

Hvis ja, så er der måske problemer med uafhængighedsantagelsen, og der skal andre modeller til → Mixed models / tilfældige effekter

En enkelt stikprøve

For en enkelt stikprøve:

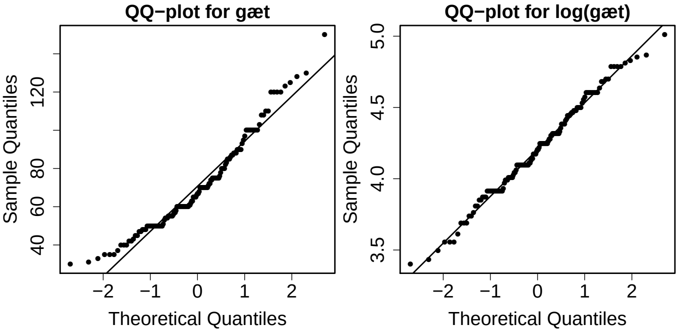
- Uafhængighed: Overvej dataindsamling
- Alle y_i kommer fra samme normalfordeling: QQ-plot.

For andre modeltyper er det mere kompliceret:

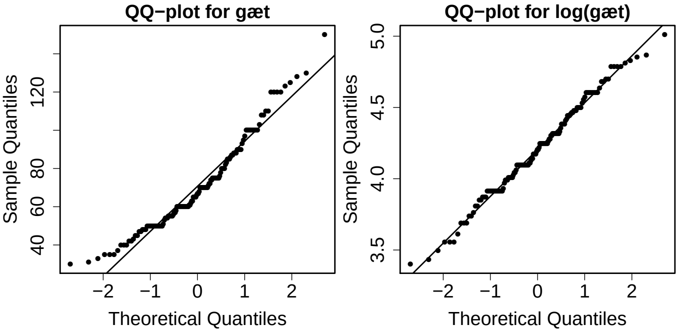
- For ensidet ANOVA: data fra hver gruppe skal være normalfordelte
- Histogram/QQ-plot for **alle** data vil være en blanding af normalfordelinger med forskellig gruppemiddelværdi

Løsning: vi ser i stedet på residualer (om lidt)

Eksempel: Gæt på antal punkter



Eksempel: Gæt på antal punkter



- QQ-plottet er meget pænere for log-transformerede gæt, så vi bruger de log-transformerede gæt i analysen
- Husk at føre resultaterne tilbage til den oprindelige skala. Vi gjorde dette onsdag eftermiddag i uge 2.

De andre modeller

For de andre modeller giver det ikke mening blot at lave QQ-plots for responsvariablen.

Vi antager jo netop at observationerne kan have forskellige middelværdier.

Men vi antager at alle restleddene har samme fordeling, så vi bruger **residualerne** til at lave modelkontrol.

Fittede værdier, residualer, standardiserede resid.

Fittede/forventede værdier er estimaterne for middelværdier, $\hat{y}_i$

Eksempler:

- Katte: $\widehat{Hwt}_i = \hat{\alpha} + \hat{\beta} \cdot \mathrm{Bwt}_i$
- Gødning: $\widehat{\mathrm{org}}_i = \hat{\alpha}_{\mathrm{type}_i}$
- Punktplot: $\widehat{\log(\mathrm{gæt})}_i = \hat{\gamma}$
- Kraftpapir: $\widehat{\mathrm{str}}_i = \hat{\alpha} + \hat{\beta}_1 \cdot \mathrm{hw}_i + \hat{\beta}_2 \cdot \mathrm{hw}_i^2$

Fittede værdier, residualer, standardiserede resid.

Fittede/forventede værdier er estimaterne for middelværdier, $\hat{y}_i$

Eksempler:

- Katte: $\widehat{Hwt}_i = \hat{\alpha} + \hat{\beta} \cdot \mathrm{Bwt}_i$
- Gødning: $\widehat{\mathrm{org}}_i = \hat{\alpha}_{\mathrm{type}_i}$
- Punktplot: $\widehat{\log(\text{gæt})}_i = \hat{\gamma}$
- Kraftpapir: $\widehat{\mathrm{str}}_i = \hat{\alpha} + \hat{\beta}_1 \cdot \mathrm{hw}_i + \hat{\beta}_2 \cdot \mathrm{hw}_i^2$

Rå residualer:

$$r_i = y_i - \hat{y}_i = \text{observeret} - \text{fitted}$$

Residualerne kan **standardiseres** så de har spredning 1 (via R):

$$\tilde{r}_i = \frac{r_i}{\mathrm{SE}(r_i)}$$

Residualer som gæt på restled

Model:

$$y_i = \text{middelværdi} + e_i$$

Residualer, standardiserede residualer

$$r_i = y_i - \hat{y}_i, \quad \tilde{r}_i = \frac{r_i}{\mathrm{SE}(r_i)}$$

Residualerne er vores bedste "gæt" på e_i'erne. Vi undersøger antagelserne på e_i via r_i og $\tilde{r}_i$

- **Hvis** modelantagelserne er OK, **så** vil alle $\tilde{r}_i$ være (cirka) normalfordelte med middelværdi 0 og spredninger 1.
- Ser efter **afvigelser** fra at std. residualer har middelværdier lig 0, spredning er lig 1, og normalfordelingen

Antagelse 2 og 3: Middelværdi og spredning

Checkes med **residualplot:**

- Residualplottet er punktplottet med fittede værdier på x-aksen og standardiserede residualer på y-aksen
- Der må ikke være noget mønster i den lodrette variation. **Mønstre er tegn på at antagelserne ikke er OK**

Antagelse 2 og 3: Middelværdi og spredning

Checkes med **residualplot:**

- Residualplottet er punktplottet med fittede værdier på x-aksen og standardiserede residualer på y-aksen
- Der må ikke være noget mønster i den lodrette variation. **Mønstre er tegn på at antagelserne ikke er OK**
- Typisk mønster: Punkterne udgør en slags kurve (fx. parabel).

  $\rightarrow$ Tegn på at middelværdien er forkert.
- Typisk mønster: Punkterne udgør en slags tragt, så lodret variation er mindre "til venstre" end "til højre".

  $\rightarrow$ Tegn på at spredningen vokser med middelværdien. Variansinhomogenitet.

Antagelse 2 og 3: Middelværdi og spredning

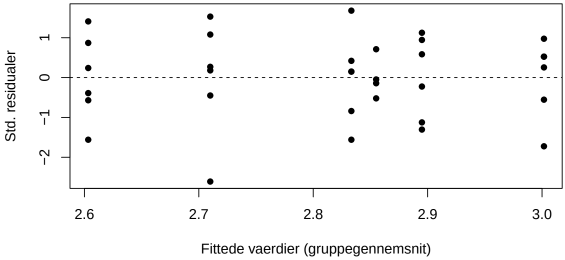
Checkes med **residualplot:**

- Residualplottet er punktplottet med fittede værdier på x-aksen og standardiserede residualer på y-aksen
- Der må ikke være noget mønster i den lodrette variation. **Mønstre er tegn på at antagelserne ikke er OK**
- Typisk mønster: Punkterne udgør en slags kurve (fx. parabel).

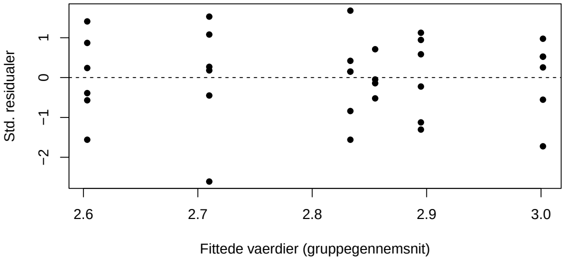
  $\rightarrow$ Tegn på at middelværdien er forkert.
- Typisk mønster: Punkterne udgør en slags tragt, så lodret variation er mindre "til venstre" end "til højre".

  $\rightarrow$ Tegn på at spredningen vokser med middelværdien. Variansinhomogenitet.
- Kig også efter **meget store/små std. residualer.** Outliers.

Eksempel: Gødning

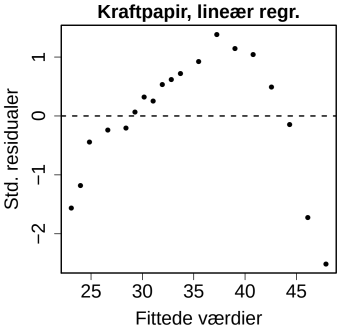


Eksempel: Gødning

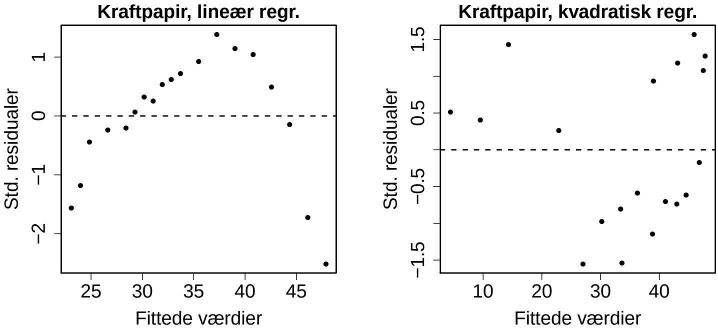


- Punkter ligger nogenlunde symmetrisk om nul overalt i plottet
- Cirka samme lodrette variation overalt i plottet
- Ingen ekstreme residualer

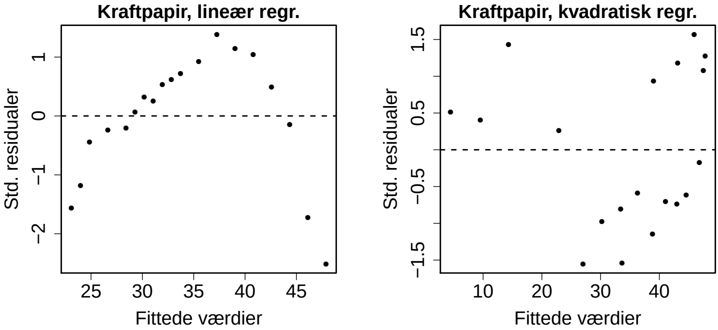
Eksempel: Kraftpapir



Eksempel: Kraftpapir

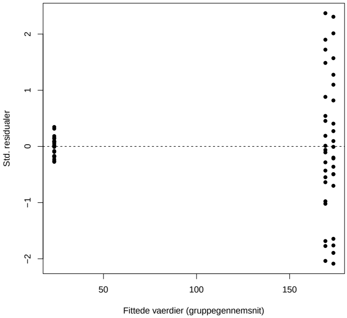


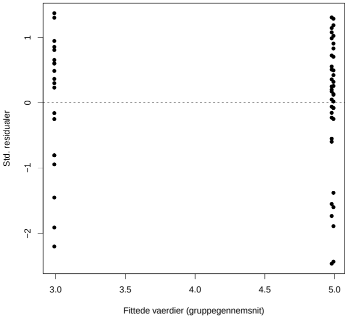
Eksempel: Kraftpapir



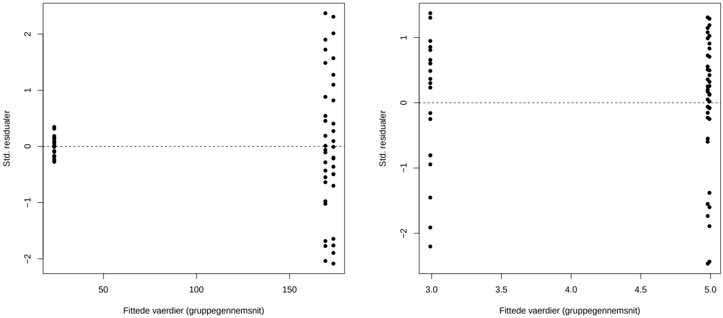
- Lineær regression: Tydeligt parabel-agtigt mønster
- Kvadratisk regression: Meget bedre, men ikke så pænt symmetrisk om nul. "Mangler" punkter nederst til venstre.

Opgave 7.4: Bedre eksempler på variansproblemer





Opgave 7.4: Bedre eksempler på variansproblemer



- Venstre: Tydelig varians-inhomogenitet (ANOVA)
- Højre: Det hjælper det at log-transformere responsen

Antagelse 1: Normalfordeling

Husk:

- Antagelsen er restleddende e_i er normalfordelte
- Residualerne r_i er vores gæt på restleddene, de standardiserede residualer $\tilde{r}_i$ har spredning 1.

Hvis modellen er OK, så vil de standardiserede residualer $\tilde{r}_i$ være normalfordelte med middelværdi 0 og spredning 1.

Antagelse 1: Normalfordeling

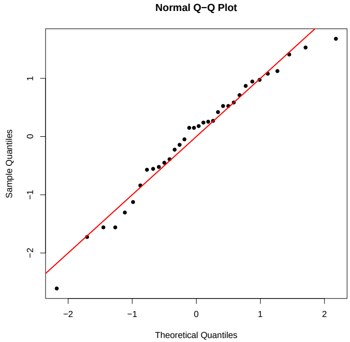
Husk:

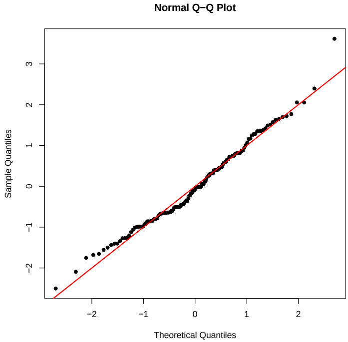
- Antagelsen er restleddende e_i er normalfordelte
- Residualerne r_i er vores gæt på restleddene, de standardiserede residualer $\tilde{r}_i$ har spredning 1.

Hvis modellen er OK, så vil de standardiserede residualer $\tilde{r}_i$ være normalfordelte med middelværdi 0 og spredning 1.

Derfor laver vi **QQ-plot for std. residualer og sammenligner med linien med skæring 0 og hældning 1.**

QQ-plots — ser her ganske fornuftige ud





- Venstre: Gødning (ensidet ANOVA)
- Højre: Katte (lineær regression)

Hvilke antagelser er vigtigst...

... for at vi kan stole på KI og test?

- Uafhængighed, middelværdi, spredning er vigtigt
- Normalfordelingen er lidt mindre vigtig
- Med andre ord: **Residualplottet er vigtigere end QQ-plottet**

Hvad gør man...

...hvis modellen viser sig at passe dårligt?

- Problemer med **middelværdi:** Tilføjelse af kvadratisk led eller transformation af x og/eller y
- Problemer med **varians:** Log- eller potenstransf. kan hjælpe
- Problemer med **uafhængighed:** Tilføjelse af tilfældige effekter (ikke på dette kursus)

Hvad gør man...

...hvis modellen viser sig at passe dårligt?

- Problemer med **middelværdi:** Tilføjelse af kvadratisk led eller transformation af x og/eller y
- Problemer med **varians:** Log- eller potenstransf. kan hjælpe
- Problemer med **uafhængighed:** Tilføjelse af tilfældige effekter (ikke på dette kursus)

Husk at **transformation ændrer fortolkningen** af parametrene, jf. analyser af punktgæt (uge 2, onsdag) og opgaver.

Kan ikke altid finde en model der passer rigtigt godt! Må sommetider være pragmatisk og gøre det så godt som muligt.

R

```
model <- lm(...)

### Fittede værdier, residualer, std. residualer
fitted(model)
residuals(model)
rstandard(model)

### Residualplot med vandret linie
plot(fitted(model), rstandard(model))
abline(h=0, lty=2)

### QQ-plot med 0/1-line
qqnorm(rstandard(model))
abline(0,1, lty=2)
```

Opsummering, modelkontrol

- Det er vigtigt at kontrollere modelantagelserne fordi vi kun kan stole på resultaterne hvis modellen er fornuftig
- Residualplottet $(\hat{y}_i, \tilde{r}_i)$ er særligt vigtigt: Punkter skal ligge omkring 0, med cirka samme lodrette variation overalt
- QQ-plot over std. residualer: Punkter skal ligge linien med skæring 0 og hældning 1
- Kig også efter ekstreme std. residualer
- Transformation kan være nyttigt, især ved problemer med middelværdi og varians
- Der er grænser for hvor meget man kan sige når man kun har få observationer

Prædiktion

Hvad er prædiktion?

Prædiktion = **forudsigelse af nye observationer**, med angivelse af usikkerhed.

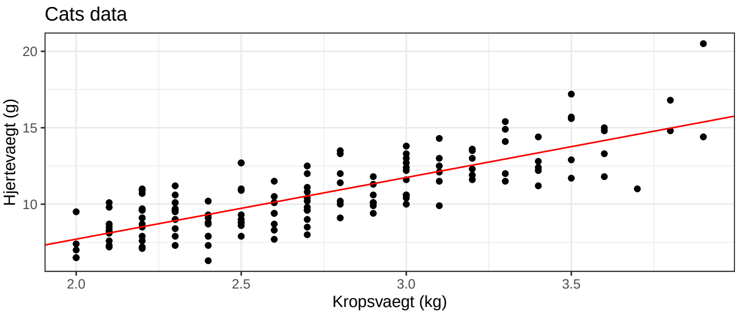
Man snakker mest om prædition ifm. regressionsmodeller. Givet en værdi af x, hvad kan vi sige om det tilhørende y?

- Hvad er vores **bedste gæt** på y?
- I hvilket **interval** vil vi forvente at y havner?

Man man kan også lave prædiktion for en stikprøve og ANOVA (afsnit 7.2.3, side 206).

Dagens R program diskuterer også prædiktion for ensidet ANOVA.

Eksempel: hjertevægt og kropsvægt af katte



- Model: $\text{Hwt}_i = \alpha + \beta \cdot \text{Bwt}_i + e_i$
- Kat med $\text{Bwt} = 2.5$ kg: hvad kan vi sige om Hwt?

Selve prædiktionen

- Data: $n = 144$ par af $(x, y) = (\text{Bwt}, \text{Hwt})$
- Model: $\text{Hwt}_i = \alpha + \beta \cdot \text{Bwt}_i + e_i$ eller $y_i = \alpha + \beta x_i + e_i$
- Estimater: $\hat{\alpha} = -0.3567, \hat{\beta} = 4.0341, \hat{\sigma} = s = 1.452$
- Ny værdi $x = x_0 = 2.5$
- Bedste gæt på Hwt:

$$\hat{y} = \hat{\alpha} + \hat{\beta} \cdot x_0 = \hat{\alpha} + \hat{\beta} \cdot 2.5 = 9.728$$

Men hvor meget anderledes kunne observationen blive? Vi vil lave et 95% prædiktionsinterval.

Konfidensinterval for middelværdi

I sidste uge fandt vi **konfidensinterval** for middelværdien, dvs. 95% for $\alpha + \beta \cdot x_0$:

$$\hat{y} \pm t_{0.975,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_x}}$$

Vi får:

$$9.728 \pm \ 1.976 \cdot 0.134 = (9.465, 9.992)$$

Konfidensinterval for middelværdi

I sidste uge fandt vi **konfidensinterval** for middelværdien, dvs. 95% for $\alpha + \beta \cdot x_0$:

$$\hat{y} \pm t_{0.975,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_x}}$$

Vi får:

$$9.728 \pm \ 1.976 \cdot 0.134 = (9.465, 9.992)$$

KI udtaler sig om **middelværdien** — altså om gennemsnit for populationen af katte med kropsvægten Bwt $= 2.5$ kg.

Et prædiktionsinterval handler derimod om en **ny observation!**

Prædiktionsinterval

En ny observation falder ikke på den rette linie, men enten over/under. Husk den biologiske variation:

$$y_i = \alpha + \beta x_i + e_i$$

Prædiktionsinterval

En ny observation falder ikke på den rette linie, men enten over/under. Husk den biologiske variation:

$$y_i = \alpha + \beta x_i + e_i$$

95%-prædiktionsinterval for ny observation:

$$\hat{y} \pm t_{0.975,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_x}}$$

Vi får:

$$9.728 \pm \; 1.976 \cdot 1.458 = (6.845, 12.612)$$

Prædiktionsinterval

En ny observation falder ikke på den rette linie, men enten over/under. Husk den biologiske variation:

$$y_i = \alpha + \beta x_i + e_i$$

95%-prædiktionsinterval for ny observation:

$$\hat{y} \pm t_{0.975,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_x}}$$

Vi får:

$$9.728 \pm \ 1.976 \cdot 1.458 = (6.845, 12.612)$$

Fortolkning:
For en (ny) kat med kropsvægt Bwt $= 2.5$ kg vil en **ny observation** af hjertevægt med 95% ssh. havne mellem 6.845 og 12.612.

Konfidensinterval vs. prædiktionsinterval

PI altid bredere end KI:

- PI udtaler sig om ny observation, KI udtaler sig om middelværdi. Det sidste er "nemmere"
- KI inddrager kun estimationsusikkerhed, PI også den direkte biologiske variation
- Se på formlerne

R

```r
library(isdals)
data(cats)
linreg1 <- lm(Hwt ~ Bwt, data = cats)
newData <- data.frame(Bwt = 2.5)
### praediktion og konfidensinterval
predict(linreg1, newData, interval = "c")

##        fit      lwr      upr
## 1 9.728494 9.464902 9.992087

### praediktion og praediktionsinterval
predict(linreg1, newData, interval = "p")

##        fit      lwr      upr
## 1 9.728494 6.845352 12.61164
```