Practicum 1: Oefening anova

Alexandre Segers & Lieven Clement

statOmics, Ghent University (https://statomics.github.io)

Contents

1 ANOVA: Kuikentjes **1**

2 Data exploratie **1**

2.1 Datastructuur bekijken 2

2.2 Data filteren 2

2.3 Figuur van de data 2

3 Statische test: **3**

3.1 Welke test kan men uitvoeren om het gemiddelde gewicht simultaan te vergelijken tussen alle soorten voeding? Wat is de link met het lineair model? Geef de nul- en alternatieve hypothese van deze test. 3

3.2 Geef de assumpties voor dit model en ga deze na: 4

3.3 Modelleer de data met het lineair model: 6

3.4 Post-hoc analyse 7

4 Conclusie **10**

1 ANOVA: Kuikentjes

In deze studie (1948) werd de invloed van verschillende soorten voeding op het gewicht van kuikentjes onderzocht. De kuikentjes werden na geboorte willekeurig in één van zes groepen toegekend, waarna deze groepen elk een andere voeding kregen. Het gewicht van deze kuikentjes werd gemeten na zes weken. Wij zullen ons beperken tot drie groepen van voeding: caseïne (casein), lijnzaad (linseed) en sojabonen (soybean).

```
suppressPackageStartupMessages({
library(tidyverse)
library(ggplot2)})
```

2 Data exploratie

```r
data("chickwts")
```

2.1 Datastructuur bekijken

```r
#Bekijk de structuur van dataset chickwts
head(chickwts)
```

```
##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

2.2 Data filteren

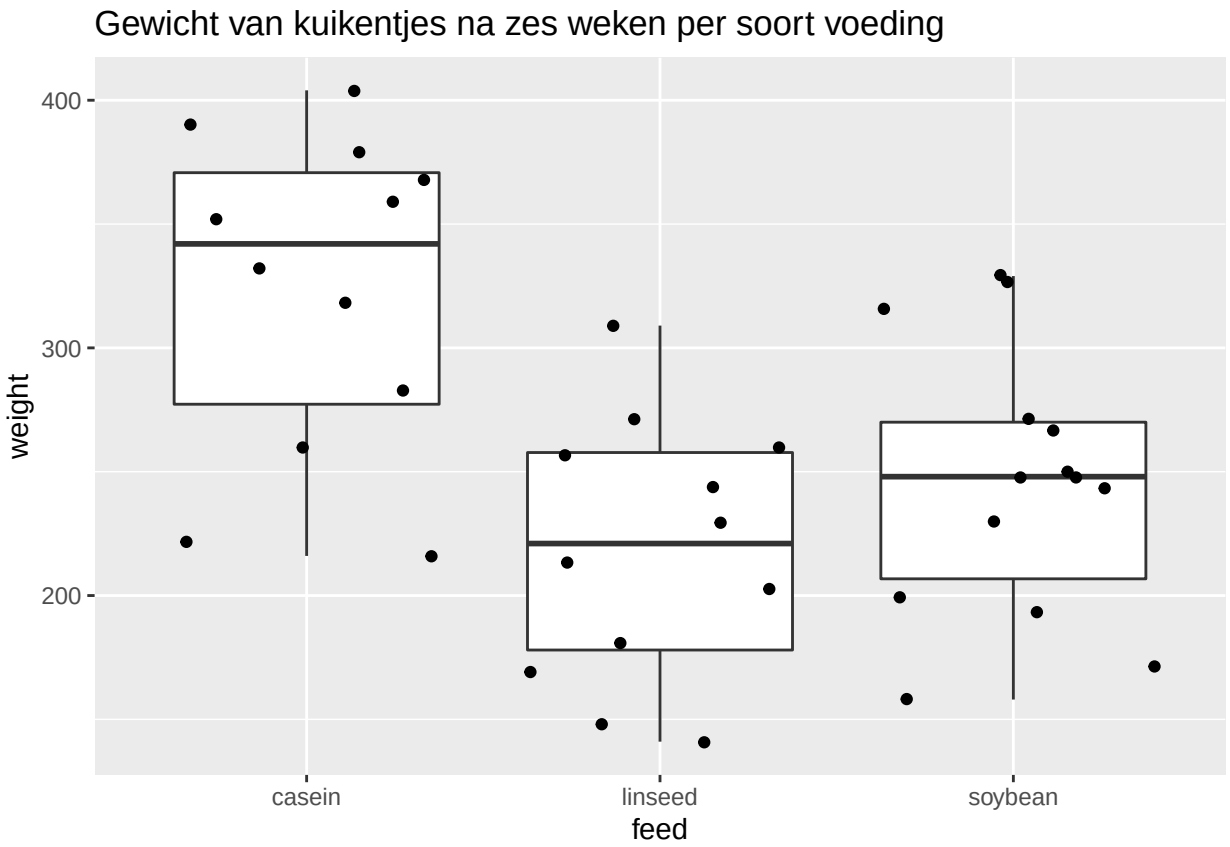
We gaan de analyse beperken tot het vergelijken van voeding caseïne (casein), lijnzaad (linseed) en sojabonen (soybean).

```r
#Filter de dataset zodat enkel datapunten van de relevante voeding (feed) aanwezig zijn.
chickwts <- chickwts %>% filter(feed %in% c("casein","linseed","soybean"))
```

2.3 Figuur van de data

We gaan eerst de data bekijken zodat we een idee hebben waarmee we te maken hebben.

```r
#Maak een boxplot van het gewicht voor elke groep van voeding. Plot ook de individuele observaties.
chickwts %>% ggplot(aes(x=feed,y=weight)) +
  geom_boxplot() +
  geom_jitter() +
  ggtitle("Gewicht van kuikentjes na zes weken per soort voeding")
```



3 Statische test:

3.1 Welke test kan men uitvoeren om het gemiddelde gewicht simultaan te vergelijken tussen alle soorten voeding? Wat is de link met het lineair model? Geef de nul- en alternatieve hypothese van deze test.

In vorige oefening zagen we enkel de two-sample t-test om twee gemiddelden met elkaar vergelijken. We hebben echter ook reeds gezien dat de two-sample t-test een specifieke versie is van een lineair model, namelijk van een lineair model waarbij de covariaat een categorische variabele X is met 2 levels, i.e.

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

Dit lineair model kan echter ook makkelijk veralgemeend worden naar factoren met meerdere levels.

$$E[Y_i] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Waarbij X_{1i} gelijk is aan 1 wanneer de voeding linseed is en anders 0. Anderzijds is X_{2i} gelijk aan 1 bij soybeans en gelijk aan 0 bij de andere voedingen.

Er bestaat een manier waarbij we **alle levels simultaan kunnen testen**, men zal namelijk testen of de gehele categorische variabele een invloed heeft op de respons variabele. In de context van ons voorbeeld, zal men kunnen testen of de voeding-soort überhaupt een effect heeft op het gemiddelde gewicht van kuikentjes. Zo'n een test heet een one-way ANOVA.

Stel dat μ_1 het gemiddelde gewicht van kuikens voor caseïne voorstelt, en idem voor μ_2 en μ_3. De nul- en alternatieve hypothese voor een ANOVA kan men dan voorstellen als

H_0: $\mu_1 = \mu_2 = \mu_3$

H_A: Voor minstens één $i \neq j$ is $\mu_i \neq \mu_j$: met andere woorden $\mu_1 \neq \mu_2$ of $\mu_1 \neq \mu_3$ of $\mu_2 \neq \mu_3$

In woorden, zegt de nulhypothese dat het gemiddelde gewicht van kuikens onafhankelijk is van de voeding: er is geen systematisch verschil in gemiddeld gewicht van het kuiken. De alternatieve hypothese zegt dat het gemiddelde gewicht verschilt tussen **minstens twee voeding-soorten**. Merk op dat men bij het verwerpen van de nulhypothese **niet weet tussen welke soorten** er een verschil is!

Indien we de ANOVA testen op basis van het lineair model kunnen we ook de volgende nul- en alternatieve hypothese opstellen:

H_0: $\beta_1 = \beta_2 = 0$

H_A: $\beta_1 \neq 0$ of $\beta_2 \neq 0$

3.2 Geef de assumpties voor dit model en ga deze na:

Zoals beschreven in de cursus, veronderstelt ANOVA een locatie-shift. Dit wil zeggen dat elke groep een gelijke variantie heeft en er enkel shifts in gemiddelde kunnen optreden. Een andere assumptie is dat de data van elke groep normaal verdeeld is, en dat de observaties onafhankelijk van elkaar zijn.

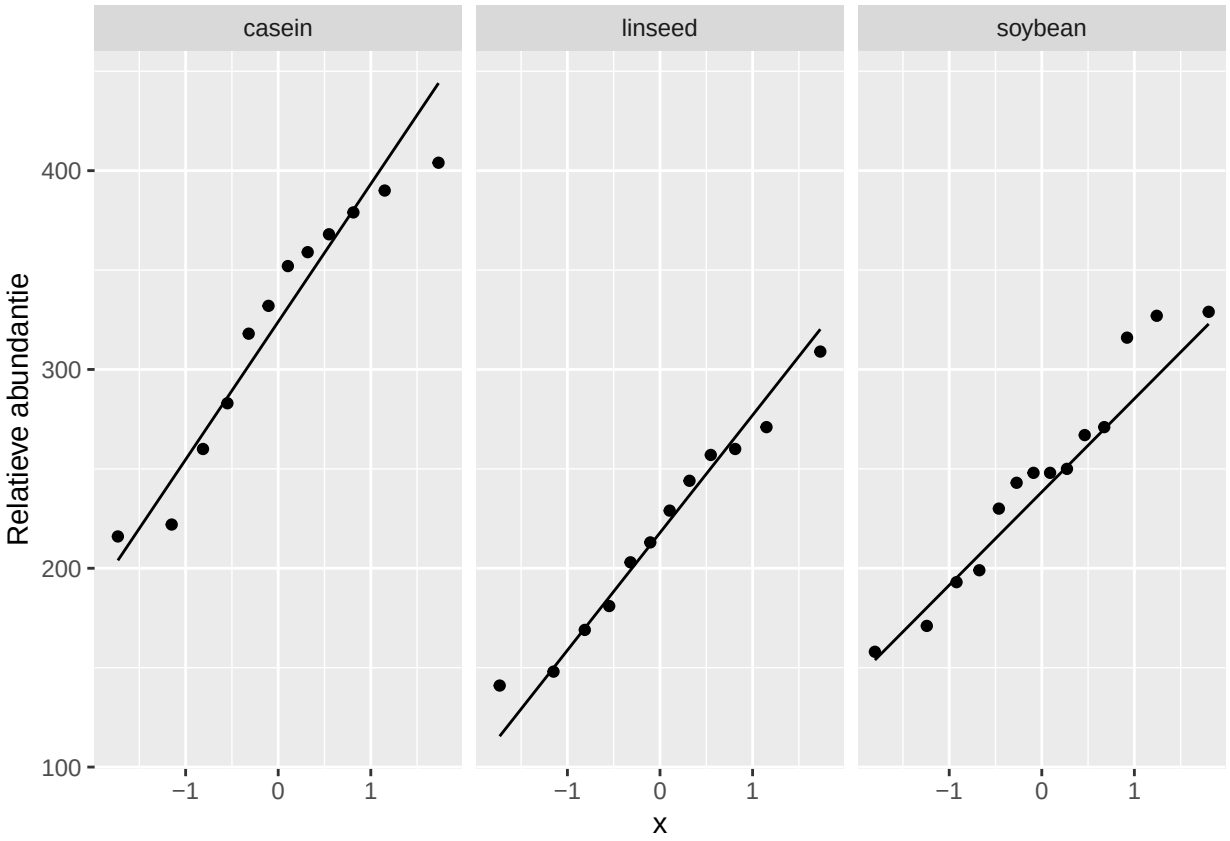
Er zijn dus drie assumpties die moeten voldaan zijn. Aangezien we in de data niet kunnen waarnemen of de observaties onafhankelijk zijn van elkaar, moeten we dit veronderstellen dat de onderzoeker dit correct heeft uitgevoerd. Daarnaast moeten we controleren dat:

- elke groep normaal is.
- de groepen een gelijke variantie hebben. (homoscedasticiteit)

We gaan gelijkheid van varianties na met boxplots. Dit lijkt alvast in orde te zijn.

We zullen nu de assumptie van normale verdeling binnen elke groep nagaan.

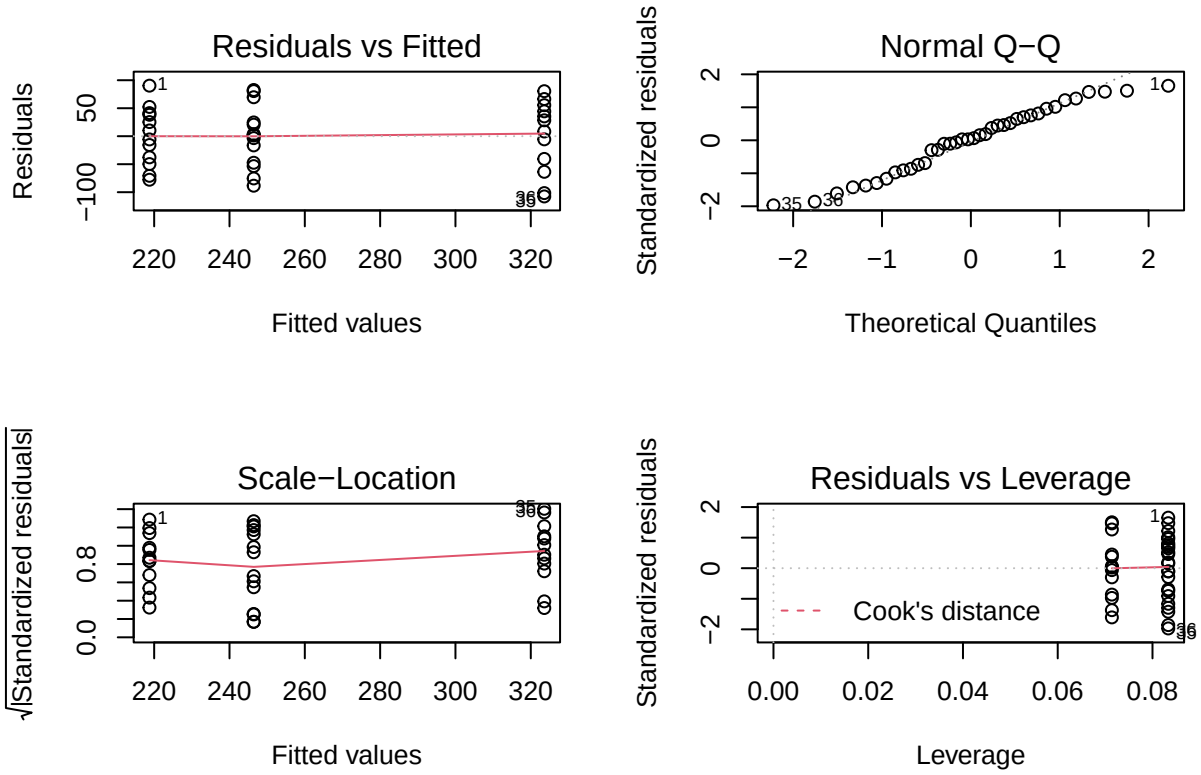
```
#Maak een QQplot voor het gewicht per voedingssoort.
chickwts %>%
  ggplot(aes(sample = weight)) +
  geom_qq() +
  geom_qq_line() +
  facet_grid(.~feed) +
  ylab("Relatieve abundantie")
```



De afwijkingen die we in onze qqplot zien lijken niet zeer uitzonderlijk. Daarom kunnen we stellen dat elke groep een normale verdeling lijkt te volgen.

Indien men veel groepen moet vergelijken, kan het makkelijker zijn om slechts één plot te moeten beoordelen. In dat geval kan men ervoor kiezen om niet voor elke groep apart een QQ-plot te maken, maar kan men de residuen van het lineair model checken. Merk op dat men dan checkt voor een normale distributie van ***alle*** residuen van het gewicht ten opzichte van hun groepsgemiddelde, en dus niet voor een normale distributie binnen elke groep apart.

```
model_lm <- lm(weight ~ feed, data = chickwts)
par(mfrow=c(2,2))
plot(model_lm) # Enkel figuur rechts boven is relevant
```




```
par(mfrow=c(1,1))
```

De QQ-plot vertoont geen systematische afwijkingen van een normale distributie. Dit is geen garantie dat de data normaal verdeeld is binnen elke groep, maar het is een benadering die we kunnen gebruiken in het geval dat:

- er te veel groepen zijn om de assumpties te checken binnen elke groep;
- er te weinig observaties zijn per groep om binnen elke groep de assumpties na te gaan.

Merk op dat je in principe de assumptie van gelijke varianties ook op basis van de plot linksboven zou kunnen checken: elke 'kolom' van punten stelt een soort voor (1 soort heeft 1 geschat gemiddelde) en de punten stellen de residuen voor ten opzichte van hun groepsgemiddelde. Men kan deze plot dus ook gebruiken om te kijken of er groepen (soorten) zijn die een verschillende variantie hebben ten opzichte van andere groepen.

3.3 Modelleer de data met het lineair model:

```
summary(model_lm)
```

```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.583  -45.717    2.571   40.500   90.250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   323.58      16.46  19.660  < 2e-16 ***
## feedlinseed  -104.83      23.28  -4.504 7.11e-05 ***
## feedsoybean   -77.15      22.43  -3.440  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.01 on 35 degrees of freedom
## Multiple R-squared:  0.3854, Adjusted R-squared:  0.3503
## F-statistic: 10.97 on 2 and 35 DF,  p-value: 0.0001996
```

De p-waarden die we hier krijgen voor de parameters komen niet overeen met de ANOVA-test die we willen uitvoeren. De p-waarde bij de F-statistic is wel de juiste p-waarde. We kunnen dit ook verkrijgen door de volgende code:

```
#Voer de ANOVA uit op het lineare model.
anova(model_lm)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## feed       2  71351   35676  10.975 0.0001996 ***
## Residuals 35 113773    3251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We voerden de ANOVA test uit aan de hand van het lineair regressiemodel. In principe testen we dan volgende nulhypothese:

$$H_0 : \beta_1 = \beta_2 = 0$$

Merk op dat deze nulhypothese evenwaardig is aan de nulhypothese die we eerder formuleerden. Als alle gemiddelden $\mu_1, \mu_2, \mu_3 = 0$, betekent dit dat beide regressieparameters β_1 en β_1 gelijk zijn aan 0.

De p-waarde van deze ANOVA test is bijzonder klein. We besluiten dat we de nulhypothese kunnen verwerpen ($p << 0.001$) en dat het gemiddelde gewicht van kuikens verschilt tussen minstens twee van de voedingspatronen op het 5% significantieniveau.

Aan de hand van dit resultaat weten we echter niet tussen welke voedingen er een verschil optreedt, en hiervoor zullen we een **post-hoc analyse** moeten uitvoeren. Een post-hoc analyse voert men enkel uit indien de ANOVA test significant was, en bestaat erin om paarsgewijze vergelijkingen uit te voeren tussen de groepen. In dit geval komt dit overeen met het vergelijken of er een verschil is tussen groep 1 en 2, tussen groep 1 en 3 en tussen groep 2 en 3.

3.4 Post-hoc analyse

De post-hoc analyse bestaat eruit om paarsgewijze testen uit te voeren. Indien men over k groepen beschikt is het totaal aantal paarsgewijze vergelijkingen gelijk aan $k(k-1)/2$. Bij ons is $k = 3$ waardoor we 3

paarsgewijze vergelijkingen zullen uitvoeren. We kunnen echter niet elke test op het 5% significantieniveau uitvoeren vanwege het meervoudig toetsen probleem. Indien we 3 vergelijkingen zouden testen elk op het 5% significantieniveau, dan is de kans dat we minstens één nulhypothese onterecht zouden verwerpen niet langer gelijk aan ons significantieniveau (5%). In ons geval, zou deze kans gelijk zijn aan:

```
alpha <- 0.05
nComparisons <- 3
1-(1-alpha)^nComparisons
```

```
## [1] 0.142625
```

Dus indien we elke test op het 5% significantieniveau zouden uitvoeren hebben we, als alle nulhypotheses waar zouden zijn, een kans van 14.3% dat we minstens één nulhypothese verkeerd zouden verwerpen. Om deze kans globaal gezien (dit is, over alle paarsgewijze vergelijkingen) op 5% te houden, kunnen we bijvoorbeeld de Bonferroni correctie uitvoeren.

In `R` kunnen we de post-hoc analyse uitvoeren met behulp van het `multcomp` package aan de hand van de `glht` functie. We specifiëren hier in het `linfct` argument dat we *multiple comparisons* (`mcp`) willen uitvoeren waarbij we alle paarsgewijze vergelijkingen voor de `feed` variabele willen testen aan de hand van de `"Tukey"` methode. De `multcomp` package zorgt ervoor dat deze p-waarden automatisch gecorrigeerd worden voor meervoudig toetsen.

```
suppressPackageStartupMessages(library(multcomp))
mcp <- glht(model_lm,linfct=mcp(feed="Tukey"))
summary <- summary(mcp)
summary
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = weight ~ feed, data = chickwts)
##
## Linear Hypotheses:
##                         Estimate Std. Error t value Pr(>|t|)
## linseed - casein == 0   -104.83      23.28  -4.504 0.000167 ***
## soybean - casein == 0    -77.15      22.43  -3.440 0.004179 **
## soybean - linseed == 0    27.68      22.43   1.234 0.441393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

In de output hiervan zien we de verschillende paarsgewijze vergelijkingen die werden uitvoerd. Bijvoorbeeld `linseed - casein == 0` duidt erop dat voor dit contrast wordt getest of het gemiddelde gewicht voor voeding `linseed` min het gemiddelde gewicht voor soort `casein` verschillend is van nul. In de tweede kolom wordt het verschil in gemiddelden weergegeven, met hun standaard error en teststatistiek in de respectievelijk derde en vierde kolom. De laatste kolom geeft aangepaste p-waarden weer op een globaal significantieniveau van 5%. Aan de hand van de aangepaste p-waarden zien we dat het gemiddelde gewicht bij `casein` verschilt van zowel `linseed` en `soybean` ($p < 0.001$ en $p = 0.004$) op het 5% significantieniveau. We zien ook dat er geen significant verschil is in gemiddeld gewicht bij voeding `soybean` en `linseed` op het 5% significantieniveau.

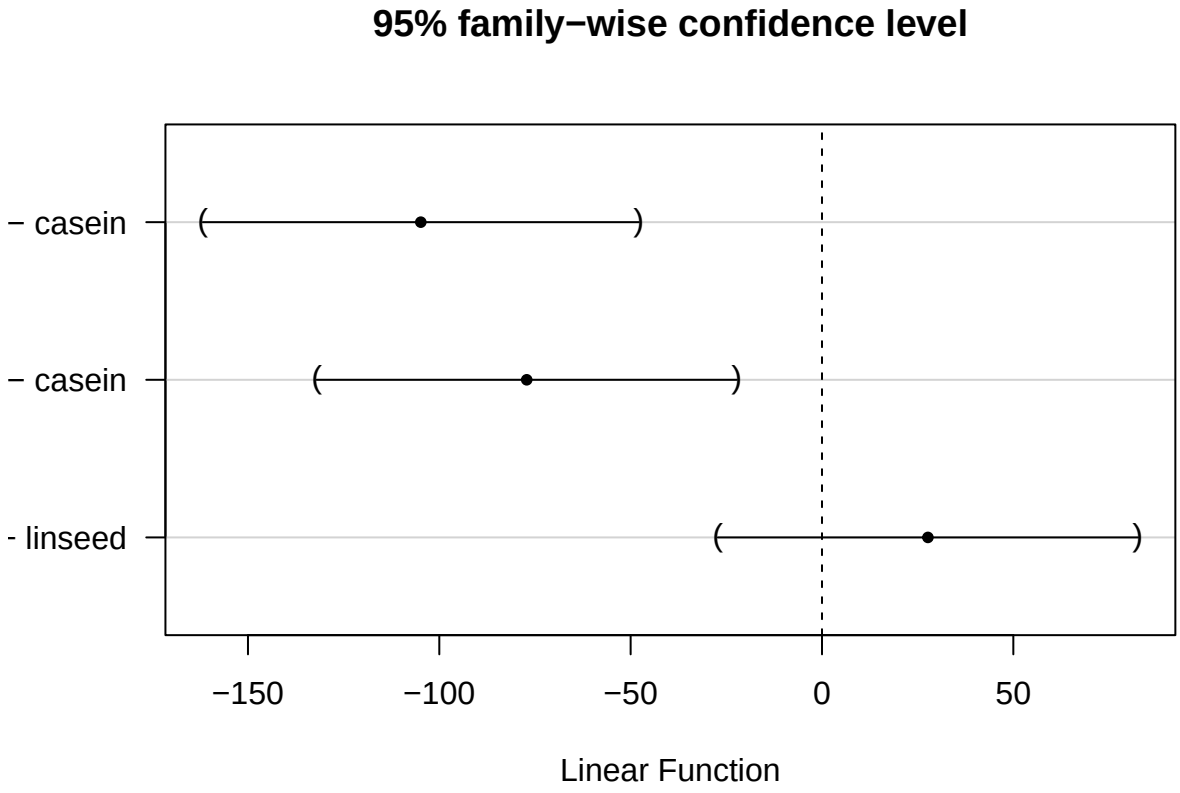
De effectgrootte is voor bij zowel `linseed` als `soybean` negatief, wat erop duidt dat het gemiddelde gewicht van kuikens hoger is bij voedingspatroon `casein`.

De betrouwbaarheidsintervallen van elke paarsgewijze test kunnen we ook makkelijk grafisch voorstellen aan de hand van de `plot` functie die zo op een `glht` object kan toegepast worden.

```
confint(mcp)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = weight ~ feed, data = chickwts)
##
## Quantile = 2.4463
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                         Estimate  lwr       upr
## linseed - casein == 0  -104.8333 -161.7729  -47.8938
## soybean - casein == 0   -77.1548 -132.0231  -22.2864
## soybean - linseed == 0   27.6786  -27.1897   82.5469
```

```
plot(mcp)
```



4 Conclusie

We kunnen concluderen dat er een significant verschil in gewicht is tussen het gewicht van kuikentjes met voeding caseine tegenover de andere voedingen op het 5% significantieniveau. Tussen linseed en soybean is er geen significant verschil in gewicht van de kuikentjes.