10. Algemeen lineair model

Lieven Clement

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

Contents

1 Inleiding **2**

1.1 Prostaatkanker voorbeeld 2

2 Additief meervoudig lineaire regressie model **3**

2.1 Statistisch model 4

3 Besluitvorming in algemeen lineair regressiemodellen **6**

3.1 Nagaan van modelveronderstellingen 8

3.2 Het niet-additieve meervoudig lineair regressiemodel 12

4 ANOVA Tabel **17**

4.1 Extra Kwadratensommen 18

5 Diagnostiek **20**

5.1 Multicollineariteit 20

5.2 Invloedrijke Observaties 25

6 Constrasten **29**

6.1 NHANES voorbeeld 29

6.2 Model 29

Home **36**

Link naar webpage/script die wordt gebruik in de kennisclips:

- script Hoofdstuk 10

1 Inleiding

- Tot nu toe: één uitkomst Y en één predictor X.
- Vaak handig om meerdere predictors te gebruiken om de respons te modelleren. bijv

1. Associatie tussen X en Y verstoord door confounder: blootstelling aan asbest (X) op de longfunctie (Y), is leeftijd (C).
2. Welke van een groep variabelen beïnvloedt een gegeven uitkomst. Habitat en menselijke activiteit op biodiversiteit in het regenwoud. (grootte, ouderdom, hoogteligging van het woud → bestudeer het simultane effect van die verschillende variabelen
3. Voorspellen van uitkomst voor individuen: zoveel mogelijk predictieve informatie simultaan gebruiken. Verwante predicties (maar dan voor het risico op sterfte) worden dagdagelijks gebruikt in eenheden intensieve zorgen om de ernst van de gezondheidstoestand van een patiënt uit te drukken.

→ Uitbreiden van enkelvoudige lineaire regressie naar meerdere predictoren.

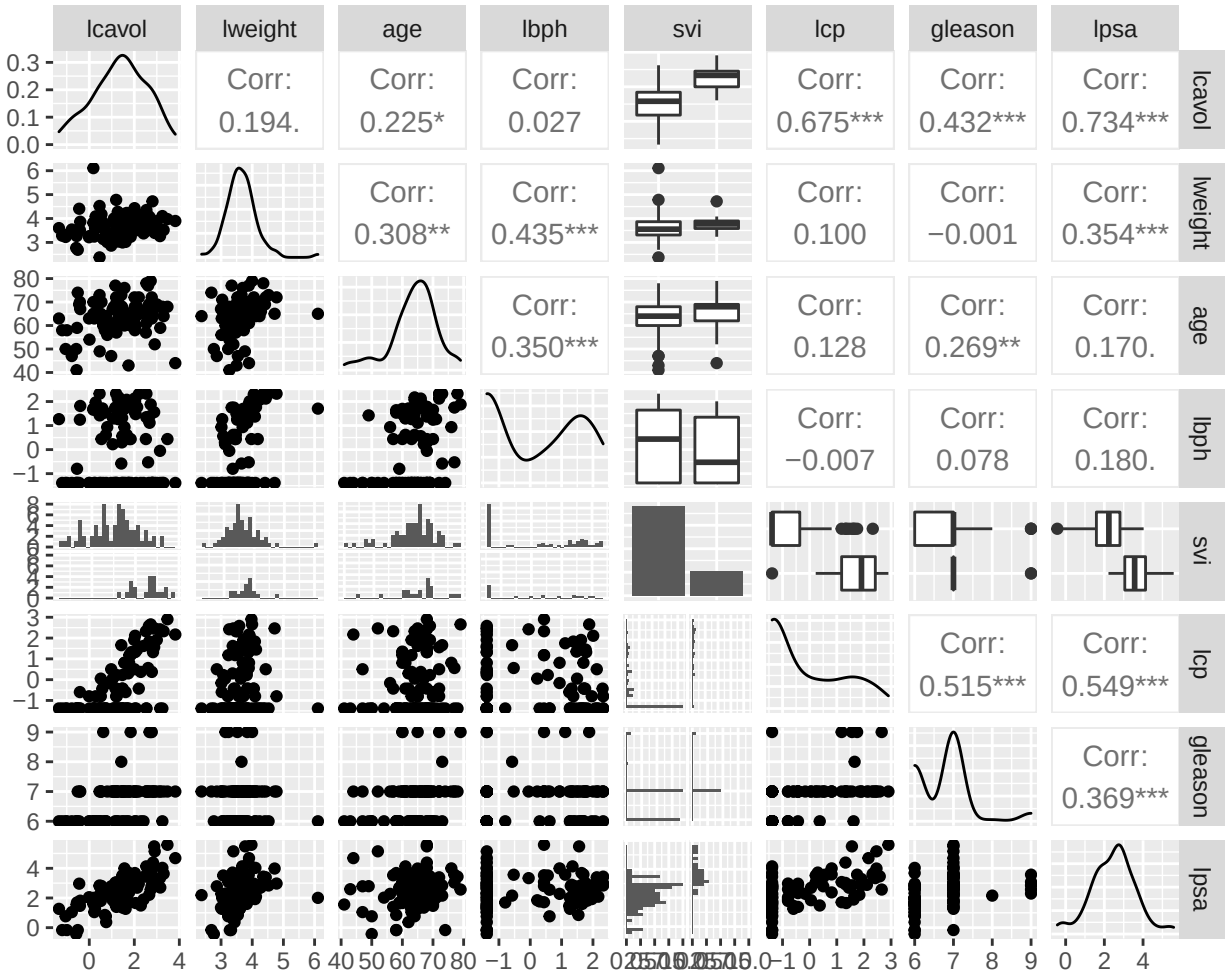
1.1 Prostaatkanker voorbeeld

- Prostaat specific antigen (PSA) en een aantal klinische metingen bij 97 mannen waarvan de prostaat werd verwijderd.
- Associatie van PSA i.f.v.
  - tumor volume (lcavol)
  - het gewicht van de prostaat (lweight)
  - leeftijd (age)
  - de goedaardige prostaathypertrofie hoeveelheid (lbph)
  - een indicator voor de aantasting van de zaadblaasjes (svi)
  - capsulaire penetratie (lcp)
  - Gleason score (gleason)
  - precentage gleason score 4/5 (pgg45)

```
prostate <- read_csv("https://raw.githubusercontent.com/statomics/sbc20/master/data/prostate.csv")
prostate
```

```
# A tibble: 97 x 9
   lcavol lweight   age   lbph svi       lcp gleason pgg45     lpsa
    <dbl>   <dbl> <dbl>  <dbl> <chr>   <dbl>   <dbl> <chr>    <dbl>
 1 -0.580    2.77    50 -1.39  healthy -1.39       6 healthy -0.431
 2 -0.994    3.32    58 -1.39  healthy -1.39       6 healthy -0.163
 3 -0.511    2.69    74 -1.39  healthy -1.39       7 20      -0.163
 4 -1.20     3.28    58 -1.39  healthy -1.39       6 healthy -0.163
 5  0.751    3.43    62 -1.39  healthy -1.39       6 healthy  0.372
 6 -1.05     3.23    50 -1.39  healthy -1.39       6 healthy  0.765
 7  0.737    3.47    64  0.615 healthy -1.39       6 healthy  0.765
 8  0.693    3.54    58  1.54  healthy -1.39       6 healthy  0.854
 9 -0.777    3.54    47 -1.39  healthy -1.39       6 healthy  1.05
10  0.223    3.24    63 -1.39  healthy -1.39       6 healthy  1.05
# ... with 87 more rows
```

```r
prostate$svi <- as.factor(prostate$svi)
```



2 Additief meervoudig lineaire regressie model

Afzonderlijke lineaire regressiemodellen, zoals

$$E(Y|X_v) = \beta_0 + \beta_v X_v$$

- Associatie tussen lpsa en 1 variabele vb (lcavol).
- Meer accurate predicties door meerdere predictoren simultaan in rekening te brengen
- Schatting voor parameter β_v mogelijks geen zuiver effect van tumor volume.
- β_v gemiddeld verschil in log-psa voor patiënten die 1 eenheid in het log tumor volume (lcavol) verschillen.

- Zelfs als lcavol niet is geassocieerd met het lpsa, dan nog kunnen patiënten met een groter tumor volume een hoger lpsa hebben omdat ze bijvoorbeeld een aantasting van de zaadblaasjes hebben (svi status 1). → Confounding.
- Vergelijken van patiënten met zelfde svi status
- Kan eenvoudig via meervoudige lineaire regressiemodellen

2.1 Statistisch model

- $p-1$ predictors $X_1, ..., X_{p-1}$ en uitkomst Y voor n subjecten
- bijvoorbeeld p-1=3: log kanker volume (X_v), log gewicht van de prostaat (X_w) en status van de zaadblaasjes (X_s)

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{ip-1} + \epsilon_i \tag{1}$$
$$Y_i = \beta_0 + \beta_v X_{iv} + \beta_w X_{iw} \beta_s X_{is} + \epsilon_i \tag{2}$$

- $\beta_0, \beta_1, ..., \beta_{p-1}$ ongekende parameters
- ϵ_i residuen die niet verklaard kunnen worden door de predictors
- Schatting met *kleinste kwadraten techniek*

Model staat toe om:

1. de verwachte uitkomst te voorspellen voor subjecten gegeven hun waarden $x_1, ..., x_{p-1}$ voor de predictoren.

$$E[Y|X_1 = x_1, ... X_{p-1} = x_{p-1}] = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_{p-1} x_{p-1}$$

2. Verschilt gemiddelde uitkomst tussen 2 groepen subjecten die δ eenheden verschillen in een verklarende variabele X_j maar dezelfde waarden hebben voor alle andere variabelen $\{X_k, k = 1, ..., p, k \neq j\}$.

$$\begin{aligned} &E(Y|X_v = x_v + \delta, X_w = x_w, X_s = x_s) \\ &\quad -E(Y|X_v = x_v, X_w = x_w, X_s = x_s) \\ &= \beta_0 + \beta_v(x_v + \delta) + \beta_w x_w + \beta_s x_s \\ &\quad -\beta_0 - \beta_v x_v - \beta_w x_w - \beta_s x_s \\ &= \beta_v \delta \end{aligned}$$

Interpretatie β_v:

- verschil in gemiddelde uitkomst tussen subjecten die in één eenheid van log tumor volume (X_v) verschillen, maar dezelfde waarde hebben voor de overige verklarende variabelen (X_w en X_s) in het model.

of

- Effect van predictor log tumor volume waarbij gecorrigeerd wordt voor de overige predictoren, hier dus associatie van tumor volume na correctie voor prostaatgewicht en svi-status.

2.1.1 Prostate voorbeeld

```
lmV <- lm(lpsa ~ lcavol, prostate)
summary(lmV)
```

```
Call:
lm(formula = lpsa ~ lcavol, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.67624 -0.41648  0.09859  0.50709  1.89672

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.50730    0.12194   12.36   <2e-16 ***
lcavol       0.71932    0.06819   10.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7875 on 95 degrees of freedom
Multiple R-squared:  0.5394,    Adjusted R-squared:  0.5346
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

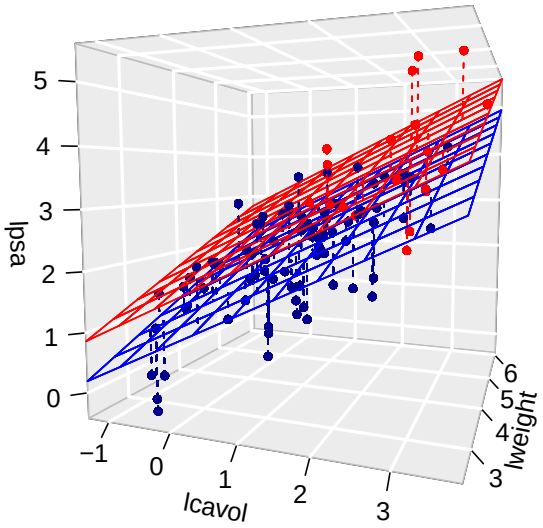
```
lmVWS <- lm(lpsa~lcavol + lweight + svi, prostate)
summary(lmVWS)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72966 -0.45767  0.02814  0.46404  1.57012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26807    0.54350  -0.493  0.62301
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
sviinvasion  0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,    Adjusted R-squared:  0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

Na terug transformatie

```
lmVWS$coef %>% exp
```

```
(Intercept)      lcavol     lweight sviinvasion
  0.7648524   1.7360954   1.6628548   1.9467442
```



---

3 Besluitvorming in algemeen lineair regressiemodellen

Als gegevens representatief zijn dan zijn kleinste kwadraten schatters voor het intercept en de hellingen onvertekend.

$$E[\hat{\beta}_j] = \beta_j, \quad j = 0, \ldots, p-1.$$

- Om resultaten uit de steekproef te kunnen veralgemenen naar de populatie is inzicht nodig in de verdeling van de parameterschatters.
- Om dat op basis van slechts één steekproef te kunnen doen zijn bijkomende veronderstellingen nodig.

1. *Lineariteit*
2. *Onafhankelijkheid*
3. *Homoscedasticiteit* of *gelijke variantie*

4. *Normaliteit*: residuen ϵ_i zijn normaal verdeeld

Onder deze aannames geldt:

$$\epsilon_i \sim N(0, \sigma^2).$$

en

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{ip-1}, \sigma^2)$$

- Hellingen zullen opnieuw nauwkeuriger worden geschat als de observaties meer gespreid zijn.
- De conditionele variantie (σ^2) opnieuw schatten op basis van de *mean squared error* (MSE):

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - ... - \hat{\beta}_{p-1} X_{ip-1}\right)^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}.$$

Opnieuw toetsen en betrouwbaarheidsintervallen via

$$T_k = \frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)} \text{ met } k = 0, ..., p-1.$$

Als aan alle aannames is voldaan dan volgen deze statistieken T_k een t-verdeling met $n-p$ vrijheidsgraden.

Wanneer niet is voldaan aan de veronderstelling van normaliteit maar wel aan lineariteit, onafhankelijkheid en homoscedasticiteit dan kunnen we voor inferentie opnieuw beroep doen op de centrale limietstelling die zegt dat de statistiek T_k bij benadering een standaard Normale verdeling zal volgen wanneer het aantal observaties voldoende groot is.

Voor het prostaatkanker voorbeeld kunnen we de effecten in de steekproef opnieuw veralgemenen naar de populatie toe door betrouwbaarheidsintervallen te bouwen voor de hellingen:

$$[\hat{\beta}_j - t_{n-p,\alpha/2} \text{SE}_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-p,\alpha/2} \text{SE}_{\hat{\beta}_j}]$$

.

```
confint(lmVWS)
```

```
                2.5 %    97.5 %
(Intercept) -1.3473509 0.8112061
lcavol       0.4033628 0.6999144
lweight      0.2103288 0.8067430
sviinvasion  0.2495824 1.0827342
```

Formele hypothese testen:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

met test statistiek

$$T = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

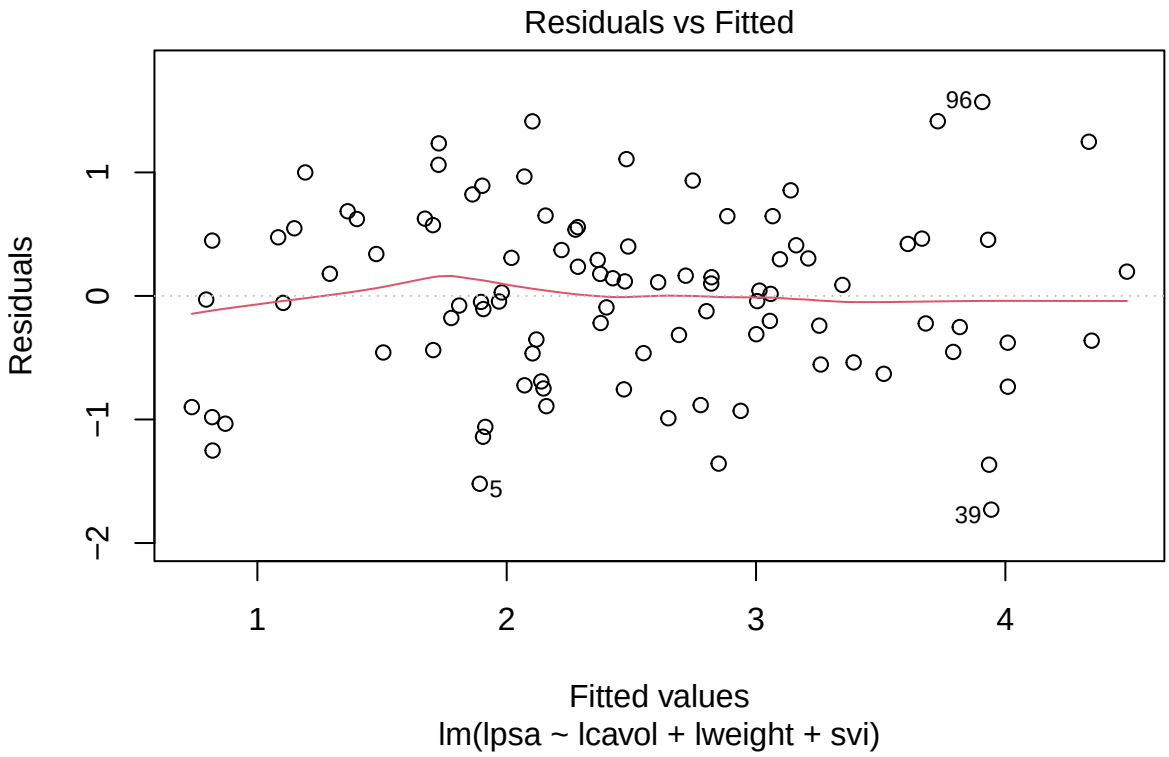
die een t-verdeling volgt met $n - p$ vrijheidsgraden onder H_0

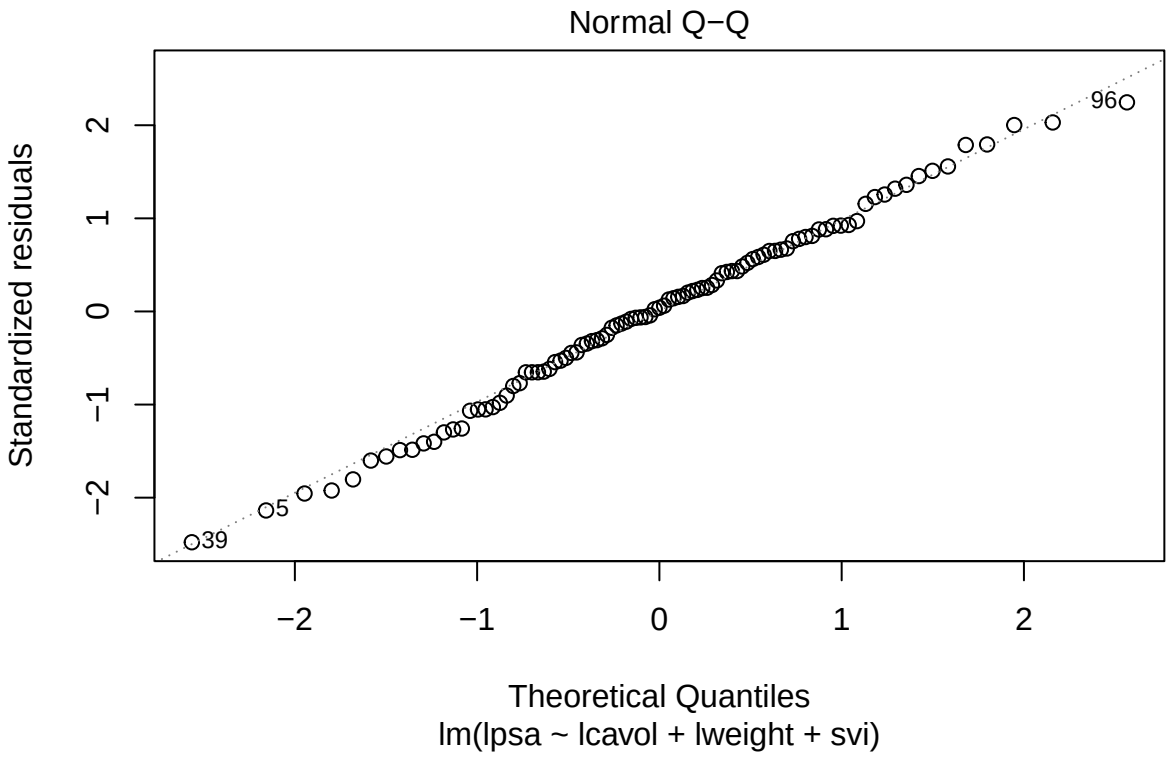
```
summary(lmVWS)
```

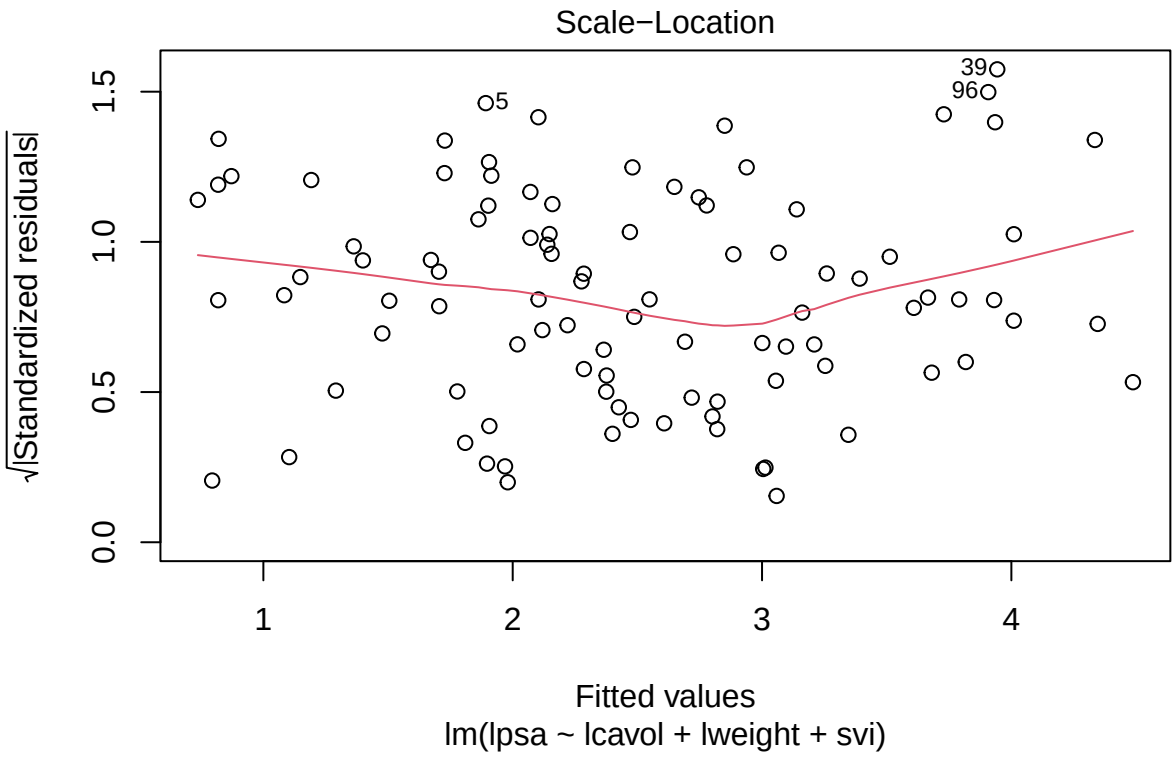
```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72966 -0.45767  0.02814  0.46404  1.57012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26807    0.54350  -0.493  0.62301
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
sviinvasion  0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,    Adjusted R-squared:  0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

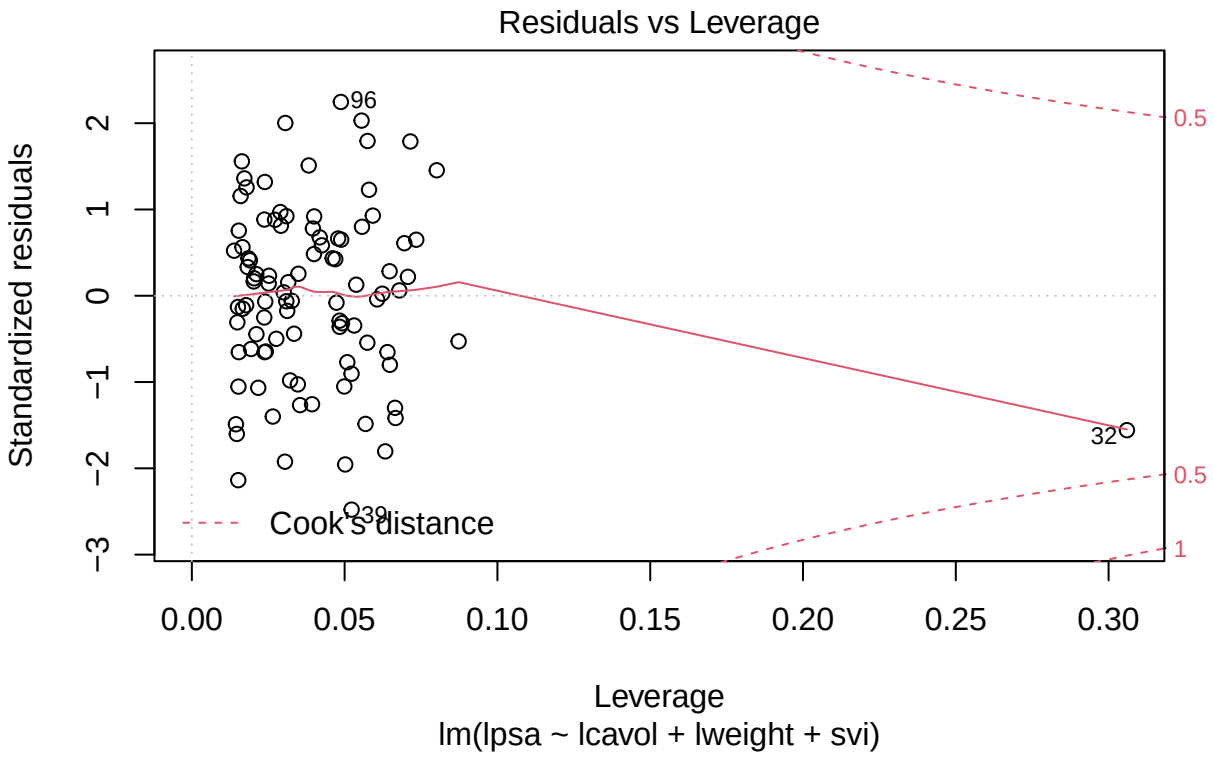
3.1 Nagaan van modelveronderstellingen

```
plot(lmVWS)
```


Residuals vs Fitted
Residuals
Fitted values
lm(lpsa ~ lcavol + lweight + svi)
96
5
39
1
0
−1
−2
1
2
3
4


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(lpsa ~ lcavol + lweight + svi)
96
5
39
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1
2
3
4
5
39
96
Fitted values
lm(lpsa ~ lcavol + lweight + svi)



3.2 Het niet-additieve meervoudig lineair regressiemodel

3.2.1 Interactie tussen een continue variabele en een factor variabele

- Het vorige model wordt het additief model genoemd omdat de bijdrage van het kanker volume in lpsa niet afhangt van de hoogte van het prostaat gewicht en de status van de zaadblaasjes.
- De helling voor lcavol hangt m.a.w. niet af van de hoogte van het log prostaat gewicht en de status van de zaadblaasjes.

$$\begin{aligned} &E[Y|X_v = x_v + \delta_v, X_w = x_w, X_S = x_s] - E[Y|X_v = x_v, X_w = x_w, X_s = x_s] = \\ &\quad [\beta_0 + \beta_v(x_v + \delta_v) + \beta_w x_w + \beta_s x_s] - [\beta_0 + \beta_v x_v + \beta_w x_w + \beta_s x_s] \\ &\quad = \beta_v \delta_v \end{aligned}$$

De svi status en de hoogte van het log-prostaatgewicht (x_w) heeft geen invloed op de bijdrage van het log-tumorvolume (x_v) in de gemiddelde log-prostaat antigeen concentratie en vice versa.

- Het zou nu echter kunnen zijn dat de associatie tussen lpsa en lcavol wel afhangt van de status van de zaadblaasjes.

- De gemiddelde toename in lpsa tussen patiënten die één eenheid van log-tumorvolume verschillen zou bijvoorbeeld lager kunnen zijn voor patiënten met aangetaste zaadblaasjes dan voor patienten met niet-aangetaste zaadblaasjes.
- Het effect van het tumorvolume op de prostaat antigeen concentratie hangt in dit geval af van de status van de zaadblaasjes.

Om deze **interactie** of **effectmodificatie** tussen variabelen X_v en X_s, en X_w en X_s statistisch te modelleren, kan men de producten van beide variabelen in kwestie aan het model toevoegen

$$Y_i = \beta_0 + \beta_v x_{iv} + \beta_w x_{iw} + \beta_s x_{is} + \beta_{vs} x_{iv} x_{is} + \beta_{ws} x_{iw} x_{is} + \epsilon_i$$

Deze termen kwantificeren de *interactie-effecten* van respectievelijk de predictoren x_v en x_s, en, x_v en x_s op de gemiddelde uitkomst.

In dit model worden de termen $\beta_v x_{iv}$, $\beta_w x_{iw}$ en $ $ sx{is} dikwijls de *hoofdeffecten* van de predictoren x_v, x_w en x_s genoemd.

```
lmVWS_IntVS_WS <- lm(
  lpsa ~
    lcavol +
    lweight +
    svi +
    svi:lcavol +
    svi:lweight,
  data = prostate)

summary(lmVWS_IntVS_WS)
```

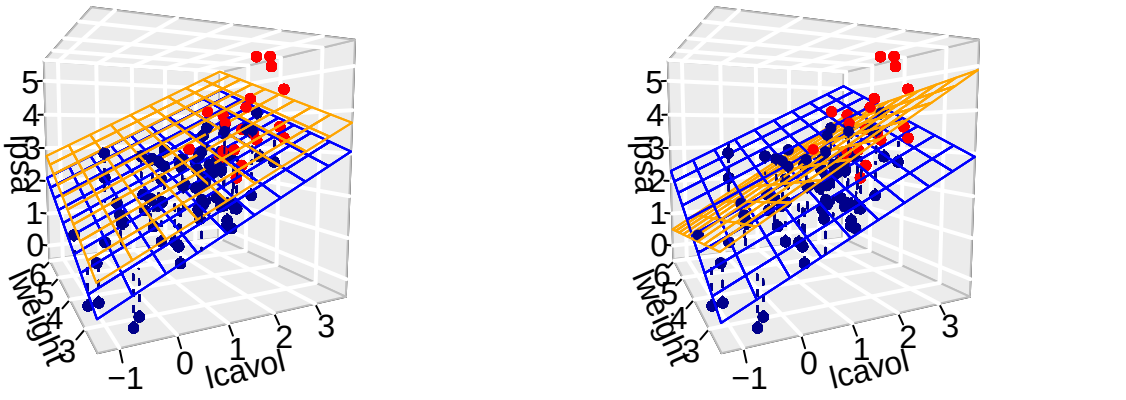
```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi + svi:lcavol + svi:lweight,
    data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50902 -0.44807  0.06455  0.45657  1.54354

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.52642    0.56793  -0.927 0.356422
lcavol               0.54060    0.07821   6.912 6.38e-10 ***
lweight              0.58292    0.15699   3.713 0.000353 ***
sviinvasion          3.43653    1.93954   1.772 0.079771 .
lcavol:sviinvasion   0.13467    0.25550   0.527 0.599410
lweight:sviinvasion -0.82740    0.52224  -1.584 0.116592
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7147 on 91 degrees of freedom
Multiple R-squared:  0.6367,    Adjusted R-squared:  0.6167
F-statistic: 31.89 on 5 and 91 DF,  p-value: < 2.2e-16
```

Omdat X_S een dummy variabele is, verkrijgen we verschillende regressievlakken:

1. Model voor $X_s = 0$:
$$Y = \beta_0 + \beta_v X_v + \beta_w X_w + \epsilon$$
waar de hoofdeffecten de hellingen voor lcavol en lweight zijn

2. en het model voor $X_s = 1$:
$$\begin{aligned} Y &= \beta_0 + \beta_v X_v + \beta_s + \beta_w X_w + \beta_{vs} X_v + \beta_{ws} X_w + \epsilon \\ &= (\beta_0 + \beta_s) + (\beta_v + \beta_{vs}) X_v + (\beta_w + \beta_{ws}) X_w + \epsilon \end{aligned}$$
met intercept $\beta_0 + \beta_s$ en hellingen $\beta_v + \beta_{vs}$ en $\beta_w + \beta_{ws}$

Additive model **Model met lcavol:lweight interactie**



3.2.2 Interactie tussen twee continue variabelen

- Het zou nu echter kunnen zijn dat de associatie tussen lpsa en lcavol afhangt van het prostaatgewicht.
- De gemiddelde toename in lpsa tussen patiënten die één eenheid van log-tumorvolume verschillen zou bijvoorbeeld lager kunnen zijn voor patiënten met een hoog prostaatgewicht dan bij patiënten met een laag prostaatgewicht.

- Het effect van het tumorvolume op de prostaat antigeen concentratie hangt in dit geval af van het prostaatgewicht.

Om deze **interactie** of **effectmodificatie** tussen 2 variabelen X_v en X_w statistisch te modelleren, kan men het product van beide variabelen in kwestie aan het model toevoegen

$$Y_i = \beta_0 + \beta_v x_{iv} + \beta_w x_{iw} + \beta_s x_{is} + \beta_{vw} x_{iv} x_{iw} + \epsilon_i$$

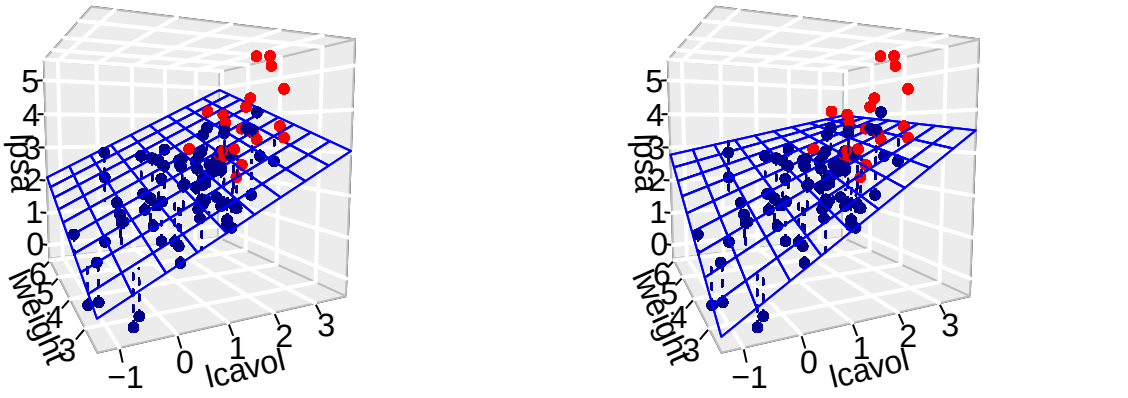
Deze term kwantificeert het *interactie-effect* van de predictoren x_v en x_w op de gemiddelde uitkomst.

Het effect van een verschil in 1 eenheid in X_v op de gemiddelde uitkomst bedraagt nu:

$$\begin{aligned} &E(Y|X_v = x_v + 1, X_w = x_w, X_s = x_s) - E(Y|X_v = x_v, X_w = x_w, X_s = x_s) \\ &\quad = [\beta_0 + \beta_v(x_v + 1) + \beta_w x_w + \beta_s x_s + \beta_{vw}(x_v + 1)x_w] - [\beta_0 + \beta_v x_v + \beta_w x_w + \beta_s x_s + \beta_{vw}(x_v)x_w] \\ &\quad = \beta_v + \beta_{vw} x_w \end{aligned}$$

```
lmVWS_IntVW <- lm(
  lpsa ~ lcavol +
    lweight +
    svi +
    lcavol:lweight,
  prostate)

summary(lmVWS_IntVW)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi + lcavol:lweight,
    data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65886 -0.44673  0.02082  0.50244  1.57457

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.6430     0.7030  -0.915  0.36278
lcavol           1.0046     0.5427   1.851  0.06734 .
lweight          0.6146     0.1961   3.134  0.00232 **
sviinvasion      0.6859     0.2114   3.244  0.00164 **
lcavol:lweight  -0.1246     0.1478  -0.843  0.40156
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7179 on 92 degrees of freedom
Multiple R-squared:  0.6293,    Adjusted R-squared:  0.6132
F-statistic: 39.05 on 4 and 92 DF,  p-value: < 2.2e-16
```



- Merk op, dat het interactie effect dat geobserveerd wordt in de steekproef echter statistisch niet significant is (p=0.4).
- Gezien de hoofdeffecten die betrokken zijn in een interactie term niet los van elkaar kunnen worden geïnterpreteerd is de conventie om een interactieterm uit het model te verwijderen wanneer die niet significant is.
- Na verwijdering van de niet-significante interactieterm kunnen de hoofdeffecten worden geïnterpreteerd.

3.2.3 Interactie tussen twee factor variabelen

Zie aparte presentatie over factoriële designs

4 ANOVA Tabel

De totale kwadratensom SSTot is opnieuw

$$\text{SSTot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Ook de residuele kwadratensom is zoals voorheen.

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Dan geldt de volgende decompositie van de totale kwadratensom,

$$\text{SSTot} = \text{SSR} + \text{SSE},$$

met

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

Voor de vrijheidsgraden en de gemiddelde kwadratensommen geldt:

- SSTot heeft $n-1$ vrijheidsgraden en $\text{SSTot}/(n-1)$ is een schatter voor de variantie van Y (van de marginale distributie van Y).
- SSE heeft $n-p$ vrijheidsgraden en $\text{MSE} = \text{SSE}/(n-p)$ is een schatter voor de residuele variantie van Y gegeven de regressoren (i.e. een schatter voor de residuele variantie σ^2 van de foutterm ϵ).
- SSR heeft $p-1$ vrijheidsgraden en $\text{MSR} = \text{SSR}/(p-1)$ is de gemiddelde kwadratensom van de regressie.

De determinatiecoëfficiënt blijft zoals voorheen, i.e.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SSTot}} = \frac{\text{SSR}}{\text{SSTot}}$$

is de fractie van de totale variabiliteit in de uitkomsten die verklaard wordt door het regressiemodel.

De teststatistiek $F = \text{MSR}/\text{MSE}$ is onder $H_0 : \beta_1 = ... = \beta_{p-1} = 0$ verdeeld als $F_{p-1;n-p}$.

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72966 -0.45767  0.02814  0.46404  1.57012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26807    0.54350  -0.493  0.62301
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
```

```
sviinvasion  0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,    Adjusted R-squared:  0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

4.1 Extra Kwadratensommen

Beschouw de volgende twee regressiemodellen voor regressoren x_1 en x_2:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i,$$

met ϵ_i iid $N(0, \sigma_1^2)$, en

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

met ϵ_i iid $N(0, \sigma_2^2)$.

Voor het eerste (gereduceerde) model geldt de decompositie

$$\text{SSTot} = \text{SSR}_1 + \text{SSE}_1$$

en voor het tweede (niet-gereduceerde) model

$$\text{SSTot} = \text{SSR}_2 + \text{SSE}_2$$

(SSTot is uiteraard dezelfde in beide modellen omdat dit niet afhangt van het regressiemodel).

Definitie extra kwadratensom De *extra kwadratensom* (Engels: *extra sum of squares*) van predictor x_2 t.o.v. het model met enkel x_1 als predictor wordt gegeven door

$$\text{SSR}_{2|1} = \text{SSE}_1 - \text{SSE}_2 = \text{SSR}_2 - \text{SSR}_1.$$

Einde definitie

Merk eerst op dat $\text{SSE}_1 - \text{SSE}_2 = \text{SSR}_2 - \text{SSR}_1$ triviaal is gezien de decomposities van de totale kwadratensommen.

De extra kwadratensom $\text{SSR}_{2|1}$ kan eenvoudig geïnterpreteerd worden als de extra variantie van de uitkomst die verklaard kan worden door regressor x_2 toe te voegen aan een model waarin regressor x_1 reeds aanwezig is.

Met dit nieuw soort kwadratensom kunnen we voor het model met twee predictoren schrijven

$$\text{SSTot} = \text{SSR}_1 + \text{SSR}_{2|1} + \text{SSE}.$$

Dit volgt rechtstreeks uit de definitie van de extra kwadratensom $\text{SSR}_{2|1}$.

Uitbreiding: Zonder in te boeten in algemeenheid starten we met de regressiemodellen ($s < p-1$)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_s x_{is} + \epsilon_i$$

met ϵ_i iid $N(0, \sigma_1^2)$, en ($s < q \leq p-1$)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_s x_{is} + \beta_{s+1} x_{is+1} + \cdots \beta_q x_{iq} + \epsilon_i$$

met ϵ_i iid $N(0, \sigma_2^2)$.

De **extra kwadratensom** van predictoren $x_{s+1}, \ldots, x_q$ t.o.v. het model met enkel de predictoren $x_1, \ldots, x_s$ wordt gegeven door

$$\text{SSR}_{s+1,\ldots,q|1,\ldots,s} = \text{SSE}_1 - \text{SSE}_2 = \text{SSR}_2 - \text{SSR}_1.$$

4.1.1 Type I Kwadratensommen

Stel dat $p-1$ regressoren beschouwd worden, en beschouw een sequentie van modellen ($s = 2, \ldots, p-1$)

$$Y_i = \beta_0 + \sum_{j=1}^{s} \beta_j x_{ij} + \epsilon_i$$

met ϵ_i iid $N(0, \sigma^2)$.

- De overeenkomstige kwadratensommen worden genoteerd als SSR_s en SSE_s.
- De modelsequentie geeft ook aanleiding tot extra kwadratensommen $\text{SSR}_{s|1,\ldots,s-1}$.
- Deze laatste kwadratensom wordt een type I kwadratensom genoemd. Merk op dat deze afhangt van de volgorde (nummering) van regressoren.

Er kan aangetoond worden dat voor Model met $s = p-1$ geldt

$$\text{SSTot} = \text{SSR}_1 + \text{SSR}_{2|1} + \text{SSR}_{3|1,2} + \cdots + \text{SSR}_{p-1|1,\ldots,p-2} + \text{SSE},$$

met SSE de residuele kwadratensom van het model met alle $p-1$ regressoren en

$$\text{SSR}_1 + \text{SSR}_{2|1} + \text{SSR}_{3|1,2} + \cdots + \text{SSR}_{p-1|1,\ldots,p-2} = \text{SSR}$$

met SSR de kwadratensom van de regressie van het model met alle $p-1$ regressoren.

- Interpretatie van iedere term afhangt van de volgorde van de regressoren in de sequentie van regressiemodellen.

- Iedere type I SSR heeft betrekking op het effect van 1 regressor en heeft dus 1 vrijheidsgraad.
- Voor iedere type I SSR term kan een gemiddelde kwadratensom gedefinieerd worden als $\text{MSR}_{j|1,\ldots,j-1} = \text{SSR}_{j|1,\ldots,j-1}/1$.
- De teststatistiek $F = \text{MSR}_{j|1,\ldots,j-1}/\text{MSE}$ is onder $H_0 : \beta_j = 0$ met $s = j$ verdeeld als $F_{1;n-(j+1)}$.
- Deze kwadratensommen worden standaard weergegeven door de anova functie in R.

4.1.2 Type III Kwadratensommen

De type III kwadratensom van regressor x_j wordt gegeven door de extra kwadratensom

$$\text{SSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1} = \text{SSE}_1 - \text{SSE}_2$$

- SSE_2 de residuele kwadratensom van regressiemodel met alle $p-1$ regressoren.
- SSE_1 de residuele kwadratensom van regressiemodel met alle $p-1$ regressoren, uitgezonderd regressor x_j.

De type III kwadratensom $\text{SSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1}$ kwantificeert dus het aandeel van de totale variantie van de uitkomst dat door regressor x_j verklaard wordt en dat niet door de andere $p-2$ regressoren verklaard wordt.

De type III kwadratensom heeft ook 1 vrijheidsgraad omdat het om 1 β-parameter gaat.

Voor iedere type III SSR term kan een gemiddelde kwadratensom gedefinieerd worden als $\text{MSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1} = \text{SSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1}/1$.

De teststatistiek $F = \text{MSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1}/\text{MSE}$ is onder $H_0 : \beta_j = 0$ verdeeld as $F_{1;n-p}$.

```r
library(car)
Anova(lmVWS,type=3)
```

```
Anova Table (Type III tests)

Response: lpsa
            Sum Sq Df F value    Pr(>F)
(Intercept)  0.125  1  0.2433  0.623009
lcavol      28.045  1 54.5809 6.304e-11 ***
lweight      5.892  1 11.4678  0.001039 **
svi          5.181  1 10.0841  0.002029 **
Residuals   47.785 93
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-waarden identiek aan die van tweezijdige t-testen

Merk echter op dat alle dummy's voor factoren met meerdere niveaus in één keer uit het model worden gehaald. De type III sum of squares heeft dus evenveel vrijheidsgraden als het aantal dummy's en er wordt een omnibustest uitgevoerd voor het effect van de factor.

---

5 Diagnostiek

5.1 Multicollineariteit

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72966 -0.45767  0.02814  0.46404  1.57012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26807    0.54350  -0.493  0.62301
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
sviinvasion  0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,    Adjusted R-squared:  0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi + lcavol:lweight,
    data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65886 -0.44673  0.02082  0.50244  1.57457

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.6430     0.7030  -0.915  0.36278
lcavol           1.0046     0.5427   1.851  0.06734 .
lweight          0.6146     0.1961   3.134  0.00232 **
sviinvasion      0.6859     0.2114   3.244  0.00164 **
lcavol:lweight  -0.1246     0.1478  -0.843  0.40156
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7179 on 92 degrees of freedom
Multiple R-squared:  0.6293,    Adjusted R-squared:  0.6132
F-statistic: 39.05 on 4 and 92 DF,  p-value: < 2.2e-16
```

- Schattingen verschillend van additief model en standaardfouten zijn veel groter!
- De oorzaak is probleem van multicollineariteit.
- Als 2 predictoren sterk gecorreleerd zijn, dan delen ze voor een groot stuk dezelfde informatie
- Moeilijk om de afzonderlijke effecten van beiden op de uitkomst te schatten.
- Kleinste kwadratenschatters onstabiel wordt
- Standaard errors kunnen worden opgeblazen

- Zolang men enkel predicties tracht te bekomen op basis van het regressiemodel zonder daarbij te extrapoleren buiten het bereik van de predictoren is multicollineariteit geen probleem.
- Wel probleem voor inferentie

```
cor(cbind(prostate$lcavol,prostate$lweight,prostate$lcavol*prostate$lweight))
```

```
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.1941283 0.9893127
[2,] 0.1941283 1.0000000 0.2835608
[3,] 0.9893127 0.2835608 1.0000000
```

- hoge correlatie tussen log-tumorvolume en interactieterm.
- Is een gekend probleem voor hogere orde termen (interacties en kwadratische termen)

- Multicollineariteit opsporen a.d.h.v. correlatie matrix of scatterplot matrix is niet ideaal.
- Geen idee in welke mate de geobserveerde multicollineariteit de resultaten onstabiel maakt.
- In modellen met 3 of meerdere predictoren, zeg X1, X2, X3 kan er zware multicollineariteit optreden ondanks dat alle paarsgewijze correlaties tussen de predictoren laag zijn.
- Ook multicollineariteit als er een hoge correlatie is tussen X1 en een lineaire combinatie van X2 en X3.

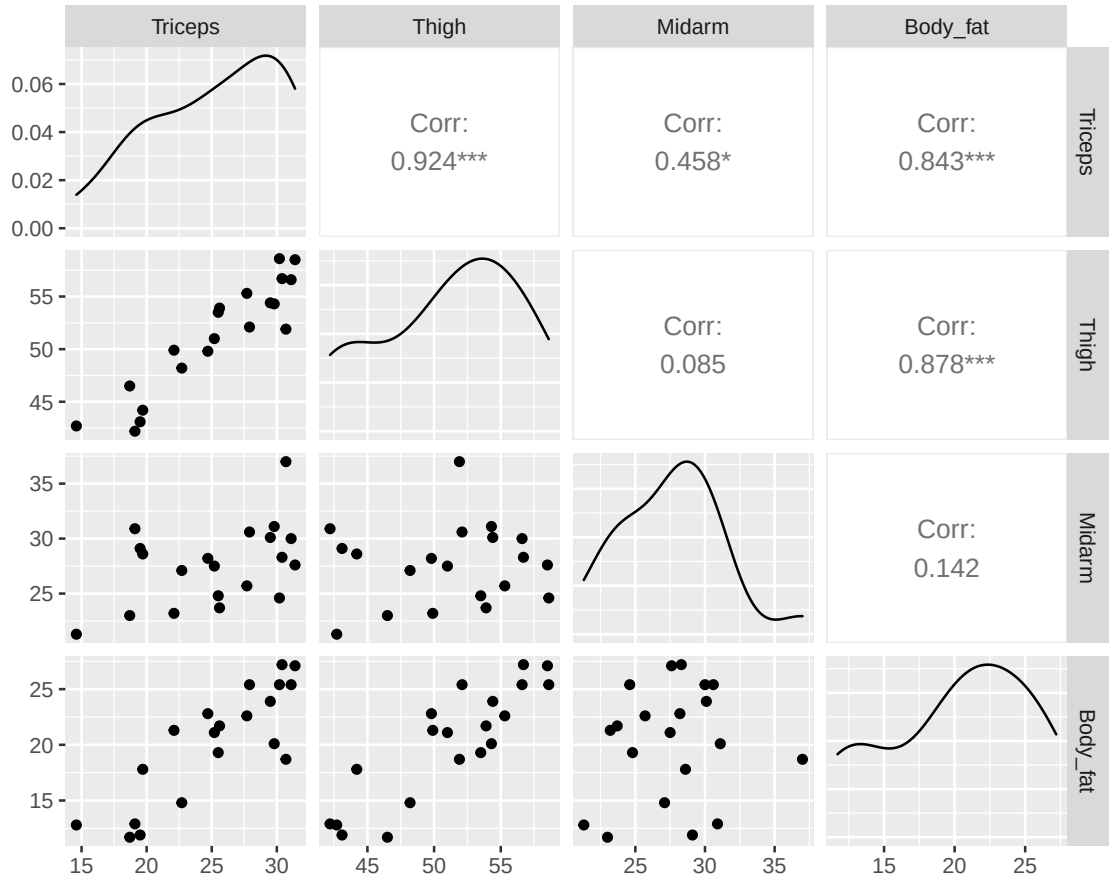
5.1.1 Variance inflation factor (VIF)

Voor de j-de parameter in het regressiemodel gedefinieerd wordt als

$$\text{VIF}_j = \left(1 - R_j^2\right)^{-1}$$

- In deze uitdrukking stelt R_j^2 de meervoudige determinatiecoëfficiënt voor van een lineaire regressie van de j-de predictor op alle andere predictoren in het model.
- VIF is 1 als j-de predictor niet lineair geassocieerd is met de andere predictoren in het model.
- VIF is groter dan 1 in alle andere gevallen.
- VIF is factor waarmee geobserveerde variantie groter is dan wanneer alle predictoren onafhankelijk zouden zijn.
- VIF $> 10 \rightarrow$ ernstige multicollineariteit.

5.1.2 Body fat voorbeeld



```
Call:
lm(formula = Body_fat ~ Triceps + Thigh + Midarm, data = bodyfat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437    0.170
Thigh         -2.857      2.582  -1.106    0.285
Midarm        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared:  0.8014,    Adjusted R-squared:  0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

```
vif(lmFat)
```

```
 Triceps    Thigh   Midarm
708.8429 564.3434 104.6060
```

```
Call:
lm(formula = Midarm ~ Triceps + Thigh, data = bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58200 -0.30625  0.02592  0.29526  0.56102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.33083    1.23934   50.29   <2e-16 ***
Triceps      1.88089    0.04498   41.82   <2e-16 ***
Thigh       -1.60850    0.04316  -37.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.377 on 17 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9893
F-statistic: 880.7 on 2 and 17 DF,  p-value: < 2.2e-16
```

We evalueren nu de VIF in het prostaatkanker voorbeeld voor het additieve model en het model met interactie.

```
vif(lmVWS)
```

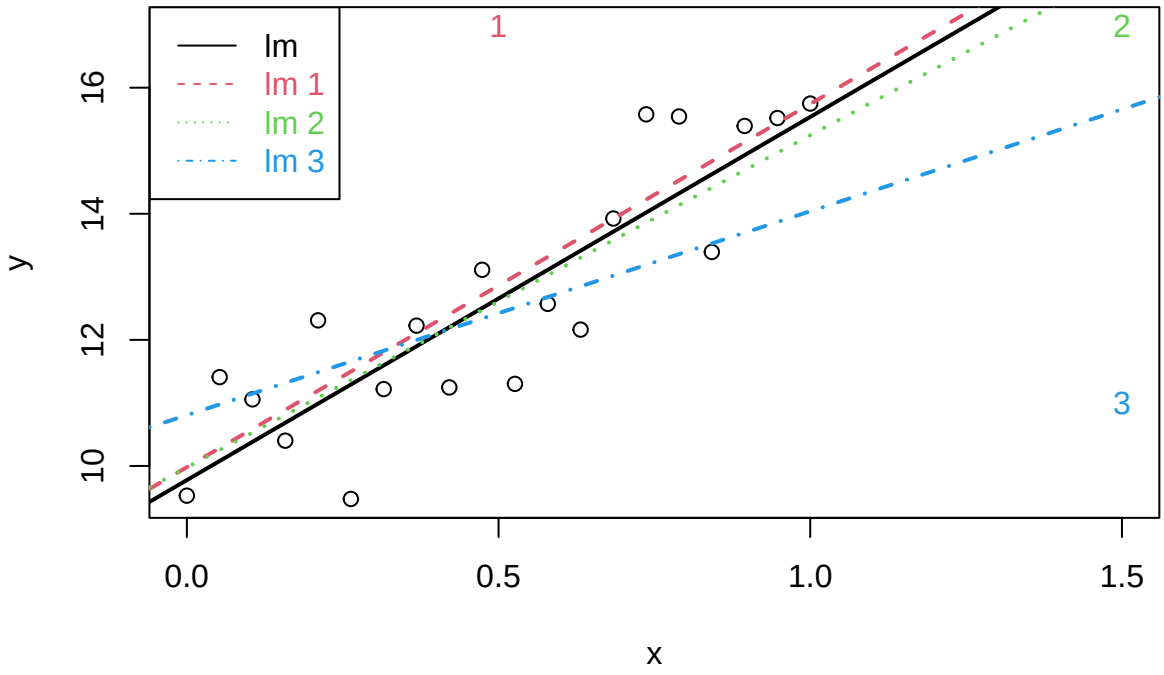
```
  lcavol  lweight      svi
1.447048 1.039188 1.409189
```

```
vif(lmVWS_IntVW)
```

```
       lcavol        lweight            svi lcavol:lweight
    76.193815       1.767121       1.426646      80.611657
```

- Inflatie voor interactietermen wordt vaak veroorzaakt door het feit dat het hoofdeffect een andere interpretatie krijgt.

5.2 Invloedrijke Observaties



- Het is niet wenselijk dat een enkele waarneming het resultaat van een lineaire regressieanalyse grotendeels bepaald
- Diagnostieken die ons toelaten om extreme observaties op te sporen
- *Studentized residu's* om outliers op te sporen
- *leverage (invloed, hefboom)* om observaties met extreem covariaatpatroon op te sporen

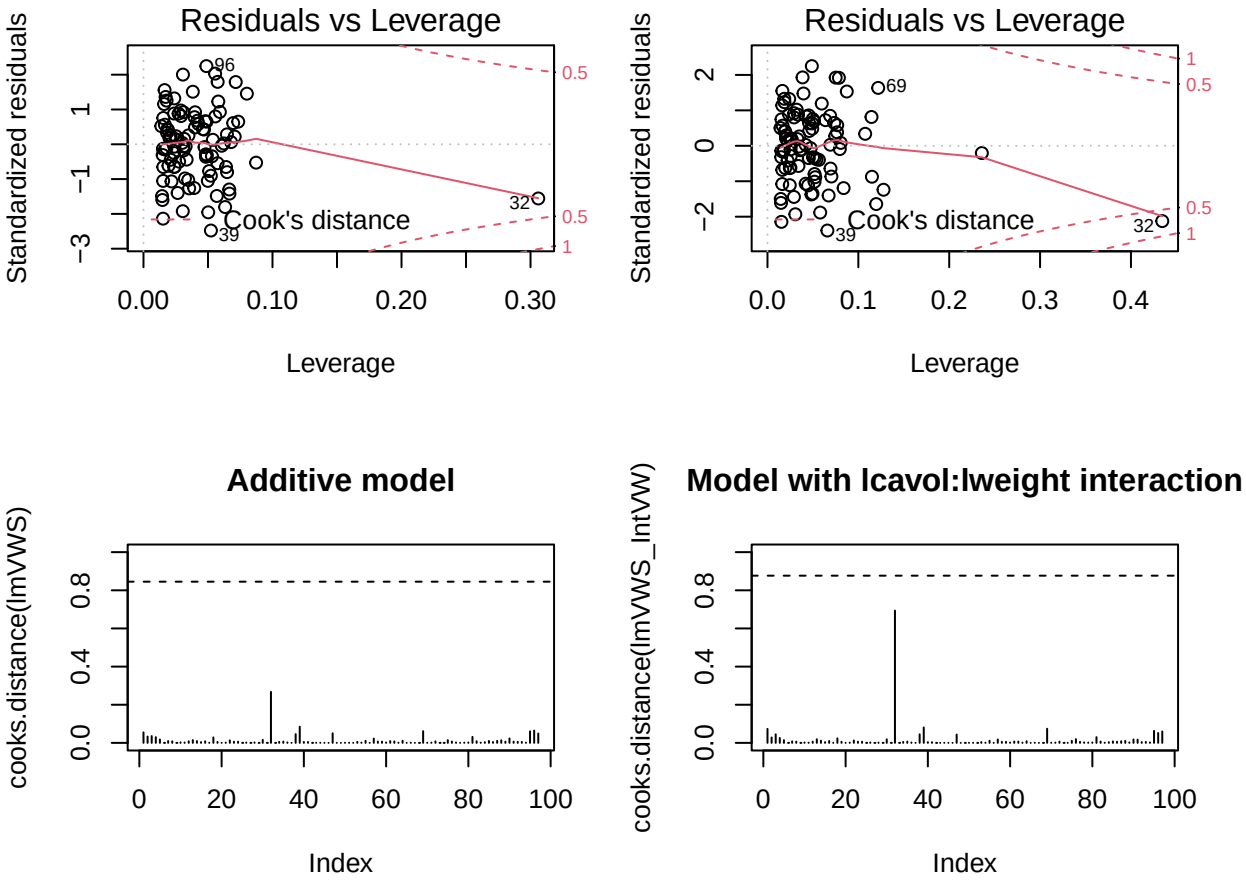
---

5.2.1 Cook's distance

- Een meer rechtstreekse maat om de invloed van elke observatie op de regressie-analyse uit te drukken
- Cook's distance voor i-de observatie is een diagnostische maat voor de invloed van die observatie op alle predicties of voor haar invloed op *alle* geschatte parameters.

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\text{MSE}}$$

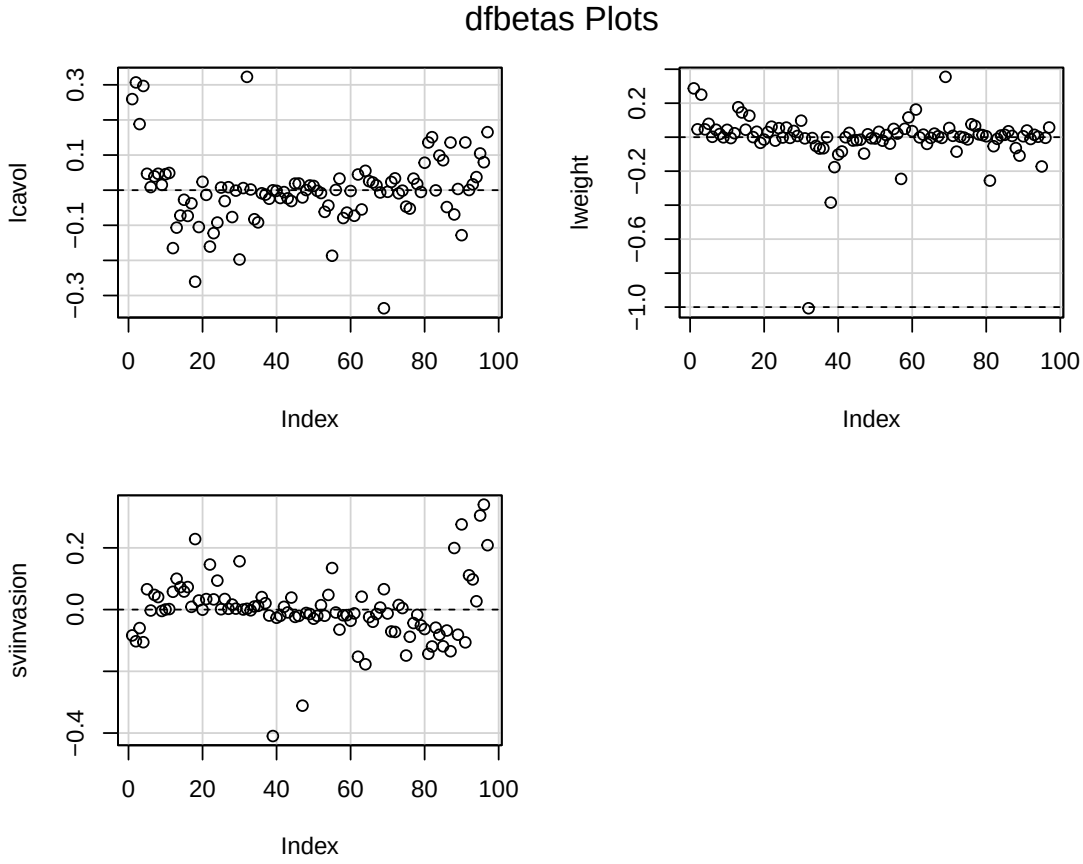
- Als Cook's distance D_i groot is, dan heeft de i-de observatie een grote invloed op de predicties en geschatte parameters.
- Extreme Cook's distance als het het 50% percentiel van de $F_{p,n-p}$-verdeling overschrijdt.

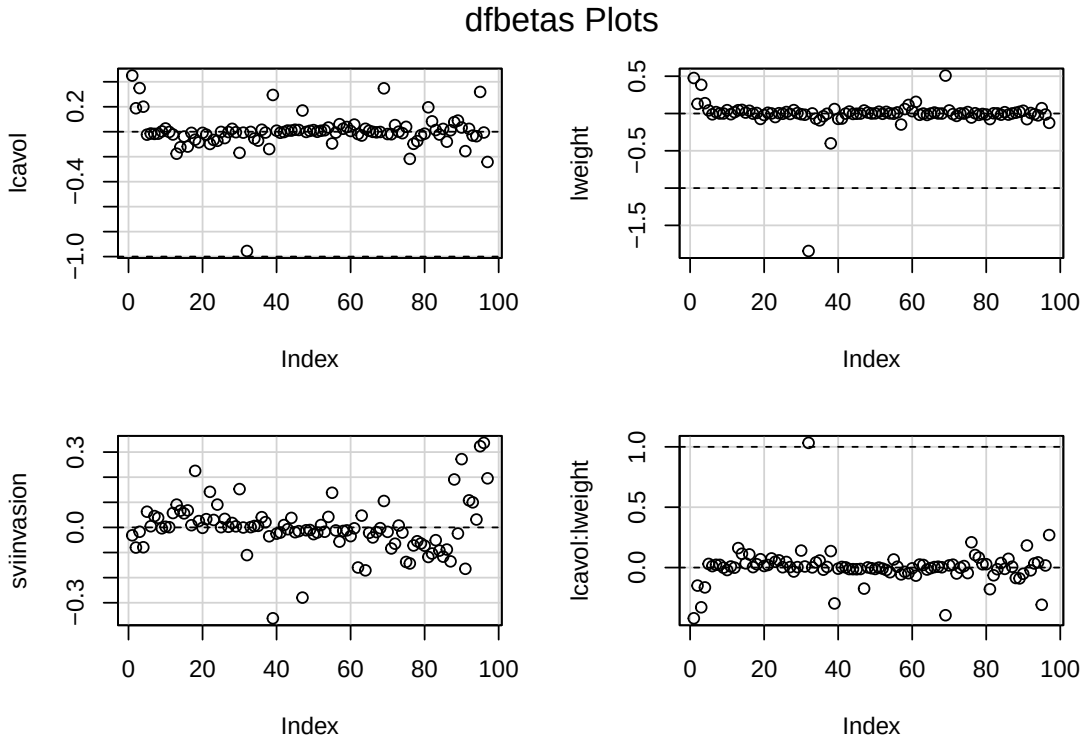


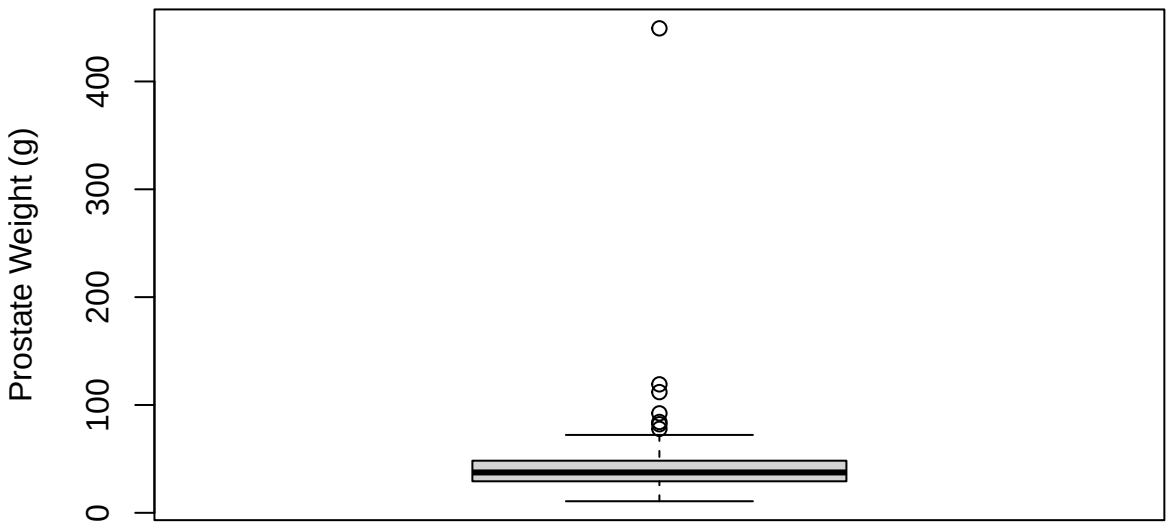
- Eenmaal men vastgesteld heeft dat een observatie invloedrijk is, kan men zogenaamde *DFBETAS* gebruiken om te bepalen op welke parameter(s) ze een grote invloed uitoefent.
- DFBETAS van de i-de observatie vormen een diagnostische maat voor de invloed van die observatie *op elke regressieparameter afzonderlijk*

$$\text{DFBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\text{SD}(\hat{\beta}_j)}$$

- DFBETAS extreem is wanneer ze 1 overschrijdt in kleine tot middelgrote datasets en $2/\sqrt{n}$ in grote datasets


dfbetas Plots
lcavol
lweight
svi1nvasion
Index
Index
Index


dfbetas Plots
lcavol
lweight
svi
lcavol:lweight
Index



---

6 Constrasten

- Bij meer complexe algemene lineaire modellen wenst men dikwijls meerdere hypothesen te toetsen.
- Bovendien vertalen de onderzoekshypotheses zich hierbij niet steeds in één parameter, maar in een lineaire combinatie van modelparameters.
- Een lineaire combinatie van modelparameters wordt ook een contrast genoemd.

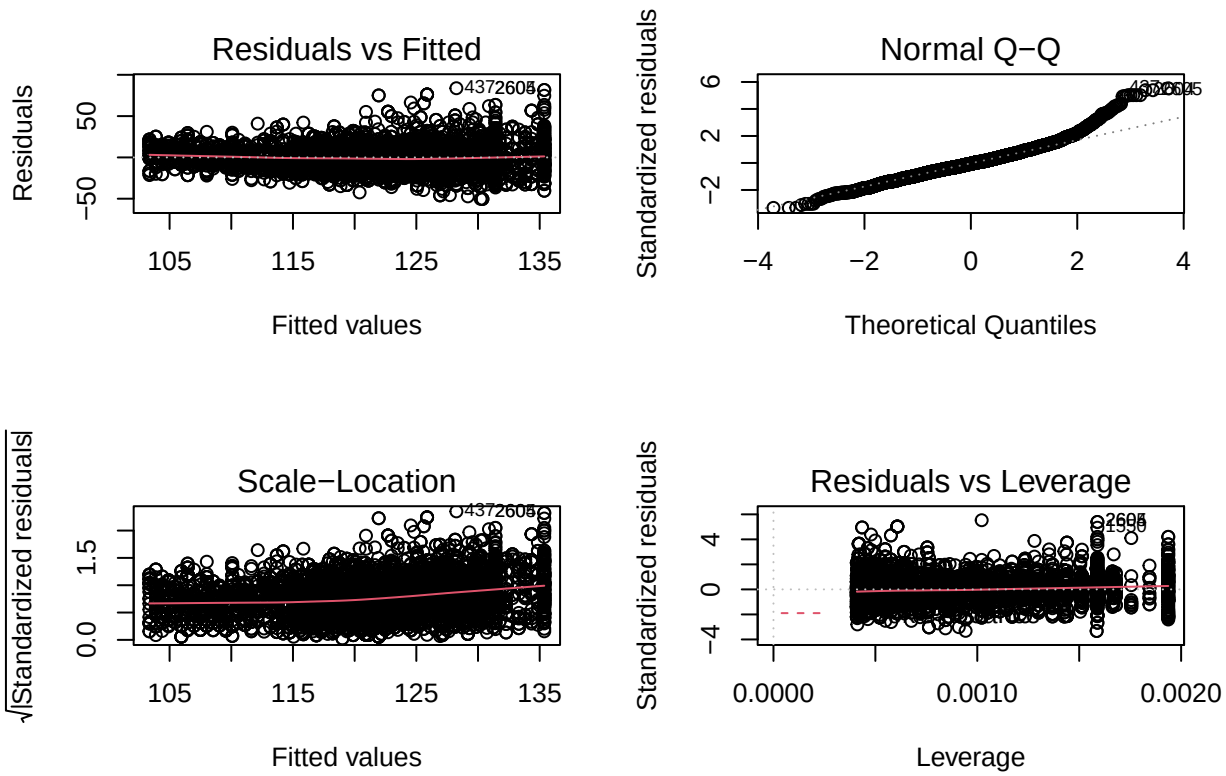
6.1 NHANES voorbeeld

- Stel dat de onderzoekers de associatie tussen de leeftijd en de bloeddruk wensen te bestuderen. Mogelijks is die associatie anders is bij mannen dan vrouwen.
- De onderzoekers wensen de volgende onderzoeksvragen te beantwoorden:
  - Is er een associatie tussen leeftijd en de bloeddruk bij vrouwen?
  - Is er een associatie tussen leeftijd en de bloeddruk bij mannen?
  - Is de associatie tussen leeftijd en de bloeddruk verschillend bij mannen dan bij vrouwen?

6.2 Model

We fitten een model op basis van de gemiddelde systolische bloeddruk (`BPSysAve`) in functie van de leeftijd, geslacht en een interactie tussen leeftijd en geslacht voor volwassen blanke subjecten uit de NHANES studie.

```r
library(NHANES)
bpData <- NHANES %>%
filter(
  Race1 =="White" &
    Age >= 18 &
    !is.na(BPSysAve)
    )

mBp1 <- lm(BPSysAve ~ Age*Gender, bpData)
par(mfrow = c(2,2))
plot(mBp1)
```

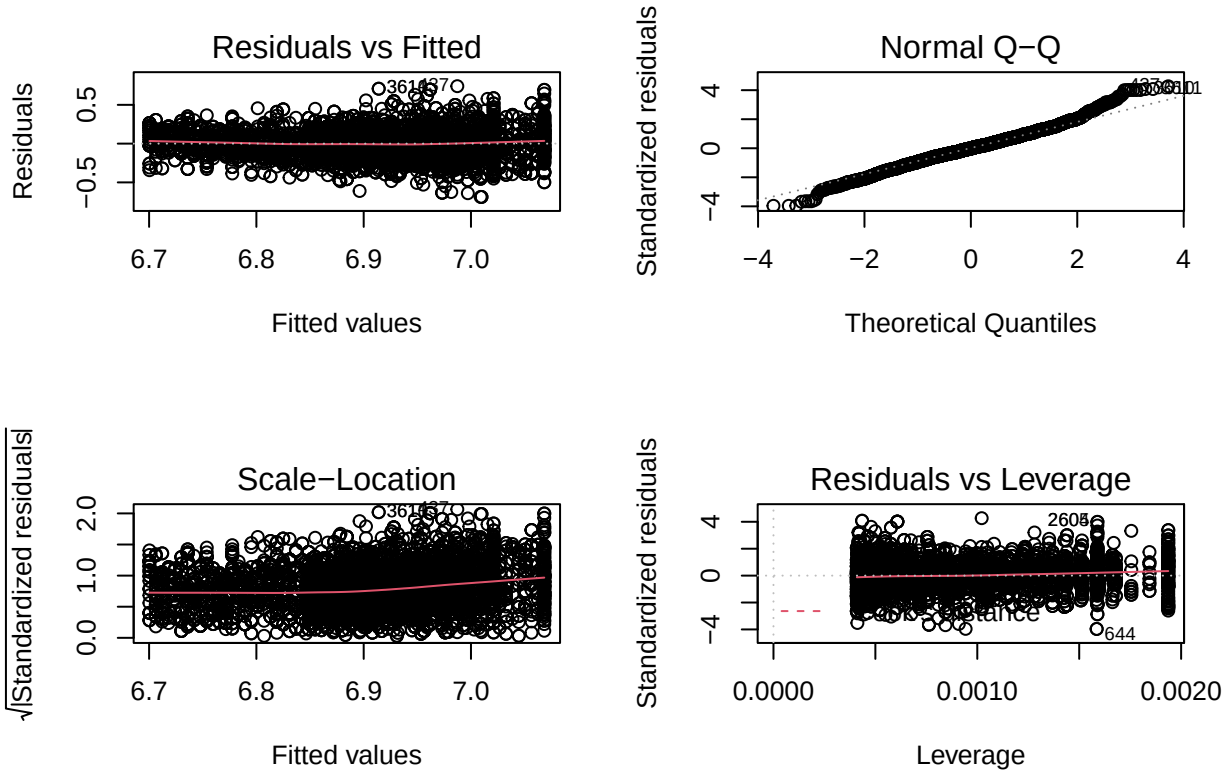


- Aannames van het model zijn niet voldaan!
  - lineariteit lijkt ok
  - heteroscedasticiteit
  - Geen normaliteit: scheve verdeling met staart naar rechts.
  - Grote dataset we kunnen steunen op de CLT

6.2.1 Transformatie

We fitten een model op basis van de log2 getransformeerde gemiddelde systolische bloeddruk (`BPSysAve`) in functie van de leeftijd, geslacht en een interactie tussen leeftijd en geslacht.

```
mBp2 <- lm(BPSysAve %>% log2 ~ Age*Gender, bpData)
par(mfrow = c(2,2))
plot(mBp2)
```



- De residuen tonen nog steeds heteroscedasticiteit.

6.2.2 Remediëren voor heteroscedasticiteit

- Als de plot van de residuen i.f.v. de geschatte waarden een toetervorm vertoont kan men toch correcte inferentie bekomen voor grote steekproeven als men de variantie van de response kan schatten.
- De inverse variantie voor elke observatie kan dan als gewicht worden gebruikt in de lm functie.

1. We zullen daarom de standard deviatie modelleren in functie van het gemiddelde.
2. Dat kan door de absolute waarde van de residuen te modelleren in functie van de gefitte waarden.
3. We kunnen de variantie van Y schatten voor elke observatie d.m.v de kwadraten van de predicties voor alle data punten a.d.h.v model voor de standard deviatie.
4. De inferentie blijft asymptotisch geldig.

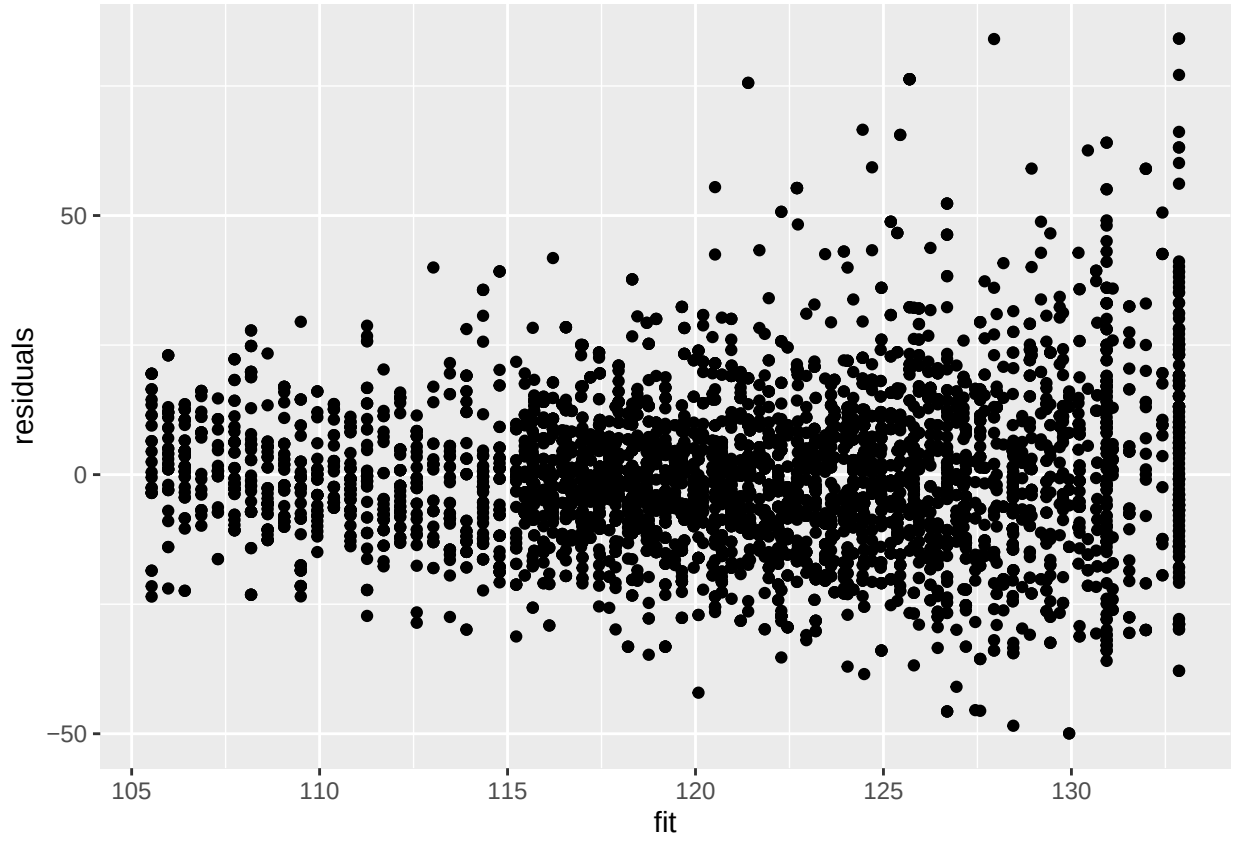
```
mSd <- lm(mBp1$res %>% abs ~ mBp2$fitted)
```

We schatten het model nu opnieuw:

```
mBp3 <- lm(BPSysAve ~ Age*Gender, bpData, w = 1/mSd$fitted^2)
```

De residuen vertonen nog steeds heteroscedasticiteit.

```
data.frame(residuals = mBp3$residuals, fit = mBp3$fitted) %>%
  ggplot(aes(fit,residuals)) +
  geom_point()
```



Na het herschalen van de residuen a.d.h.v. de standard deviatie (vermenigvuldigen met vierkantswortel van het gewicht) zijn de geschaalde residuen homoscedastisch.

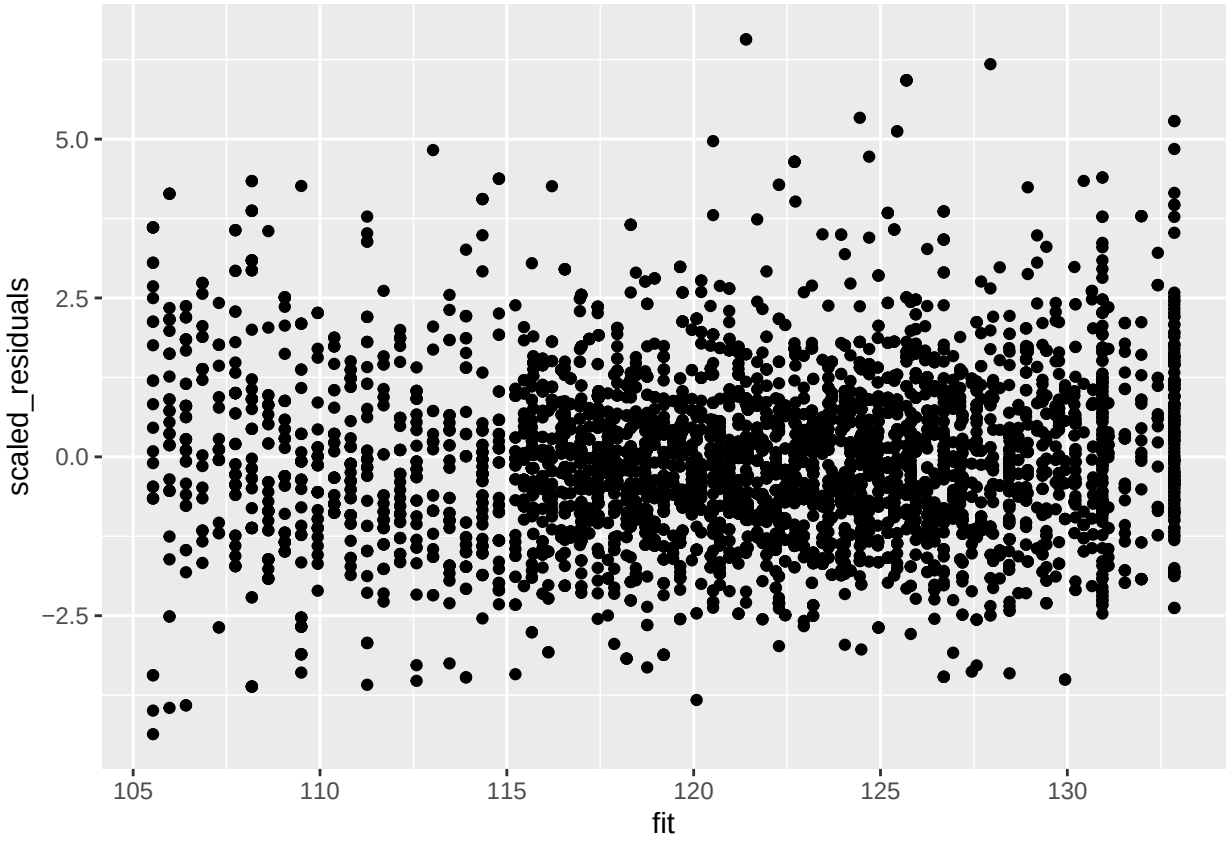
De parameters worden geschat door de gewogen kleinste kwadraten techniek.

$$SSE = \sum_{i=1}^{n} w_i e_i^2$$

met $w_i = 1/\hat{\sigma}_i^2$.

De gewogen regressie zal dus correct rekening houden met heteroscedasticiteit.

```
data.frame(scaled_residuals = mBp3$residuals/mSd$fitted, fit = mBp3$fitted) %>%
  ggplot(aes(fit,scaled_residuals)) +
  geom_point()
```



6.2.3 Besluitvorming

```
summary(mBp3)
```

```
Call:
lm(formula = BPSysAve ~ Age * Gender, data = bpData, weights = 1/mSd$fitted^2)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-4.3642 -0.8494 -0.0940  0.7605  6.5701

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    97.59709    0.63501 153.693  < 2e-16 ***
Age             0.44082    0.01505  29.294  < 2e-16 ***
Gendermale     13.36724    1.09017  12.262  < 2e-16 ***
Age:Gendermale -0.19115    0.02420  -7.899 3.45e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.319 on 4828 degrees of freedom
Multiple R-squared:  0.2182,    Adjusted R-squared:  0.2178
F-statistic: 449.3 on 3 and 4828 DF,  p-value: < 2.2e-16
```

De onderzoeksvragen vertalen zich in de volgende nullhypotheses:

1. Associatie tussen bloeddruk en leeftijd bij de vrouwen?

$$H_0 : \beta_{\text{Age}} = 0 \text{ vs } H_1 : \beta_{\text{Age}} \neq 0$$

2. Associatie tussen bloeddruk en leeftijd bij de mannen?

$$H_0 : \beta_{\text{Age}} + \beta_{\text{Age:Gendermale}} = 0 \text{ vs } H_1 : \beta_{\text{Age}} + \beta_{\text{Age:Gendermale}} \neq 0$$

3. Is de Associatie tussen bloeddruk en leeftijd verschillend bij mannen en vrouwen?

$$H_0 : \beta_{\text{Age:Gendermale}} = 0 \text{ vs } H_1 : \beta_{\text{Age:Gendermale}} \neq 0$$

- We kunnen onderzoeksvraag 1 en 3 onmiddelijk toetsen o.b.v. de model output.
- Onderzoeksvraag 2 is echter een lineaire combinatie van twee parameters.
- Bovendien is er ook het probleem dat we meerdere toetsen nodig hebben om de associatie te bestuderen.

We kunnen opnieuw gebruik maken van een Anova approach.

1. We toetsen eerste de omnibus hypothese dat er geen associatie is tussen leeftijd en de bloeddruk.

$$H_0 : \beta_{\text{Age}} = \beta_{\text{Age}} + \beta_{\text{Age:Gendermale}} = \beta_{\text{Age:Gendermale}} = 0$$

- Dat vereenvoudigt zich tot het toetsen dat

$$H_0 : \beta_{\text{Age}} = \beta_{\text{Age:Gendermale}} = 0$$

- Wat we kunnen evalueren door twee modellen te vergelijken. Een model met enkel het gender effect en volledige model met Gender, Age en Gender x Age interactie.

2. Als we deze hypothese kunnen verwerpen voeren we posthoc analyses uit voor elk van de 3 contrasten.

```
mBp0 <- lm(BPSysAve ~ Gender, bpData, w = 1/mSd$fitted^2)
anova(mBp0, mBp3)
```

6.2.3.1 Omnibus test

```
Analysis of Variance Table

Model 1: BPSysAve ~ Gender
Model 2: BPSysAve ~ Age * Gender
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   4830 10200.5
2   4828  8404.5  2      1796 515.86 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.2.3.2 Posthoc testen De posthoc testen kunnen we opnieuw uitvoeren a.d.h.v. het multcomp pakket.

```
library(multcomp)
bpPosthoc <- glht(mBp3, linfct=c(
  "Age = 0",
  "Age + Age:Gendermale = 0",
  "Age:Gendermale = 0")
  )
bpPosthoc %>% summary
```

```
     Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = BPSysAve ~ Age * Gender, data = bpData, weights = 1/mSd$fitted^2)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
Age == 0                    0.44082    0.01505  29.294   <1e-10 ***
Age + Age:Gendermale == 0   0.24967    0.01895  13.175   <1e-10 ***
Age:Gendermale == 0        -0.19115    0.02420  -7.899   <1e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
bpPosthocBI <- bpPosthoc %>% confint
bpPosthocBI
```

```
     Simultaneous Confidence Intervals

Fit: lm(formula = BPSysAve ~ Age * Gender, data = bpData, weights = 1/mSd$fitted^2)

Quantile = 2.3154
95% family-wise confidence level


Linear Hypotheses:
                           Estimate lwr     upr
Age == 0                    0.4408   0.4060  0.4757
Age + Age:Gendermale == 0   0.2497   0.2058  0.2936
Age:Gendermale == 0        -0.1911  -0.2472 -0.1351
```

Merk op dat de glht functie ons toelaat om de contrasten te definiëren door de nulhypotheses expliciet te formuleren in een karaktervector waarbij gebruik wordt gemaakt van de naam van de pararameters in het model.

6.2.3.3 Conclusie We kunnen besluiten dat er een extreem significante associatie is tussen leeftijd en de bloeddruk (p « 0.001). De bloeddruk bij twee vrouwen die in leeftijd verschillen is gemiddeld 0.44 mm Hg hoger per jaar leeftijdsverschil bij de oudste vrouw en dat verschil is extreem significant (p « 0.001, 95% BI [0.41, 0.48]. De bloeddruk bij mannen die in leeftijd verschillen is gemiddeld 0.25 mm Hg hoger per jaar leeftijdsverschil bij de oudere man. (p « 0.001, 95% BI [0.21, 0.29]. Het gemiddelde bloeddrukverschil tussen personen in leeftijd verschillen is gemiddeld -0.19 mm Hg/jaar hoger bij vrouwen dan mannen (p « 0.001, 95% BI [-0.25, -0.14]).

Home