8. Multiple regression

Lieven Clement

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

Contents

1 Intro

- Until now: one outcome Y and a single predictor X.
- Often useful to use multiple predictors to model the response. e.g

1. Association between X and Y is affected by confounder: Smoking and age by youngsters are confounded and they both affect the lung capacity
2. Which group of variables is associated with a given outcome. E.g Habitat and human activity on the biodiversity of the rain forest. (Size, age, height of the wood → assess all effects simultaneously.
3. Prediction of outcome for individuals: use as many predictive information simultaneously. E.g prediction of risk on mortality is used on a daily basis in intensive care units to prioritise patient care.

→ Extend simple linear regression to multiple predictors.

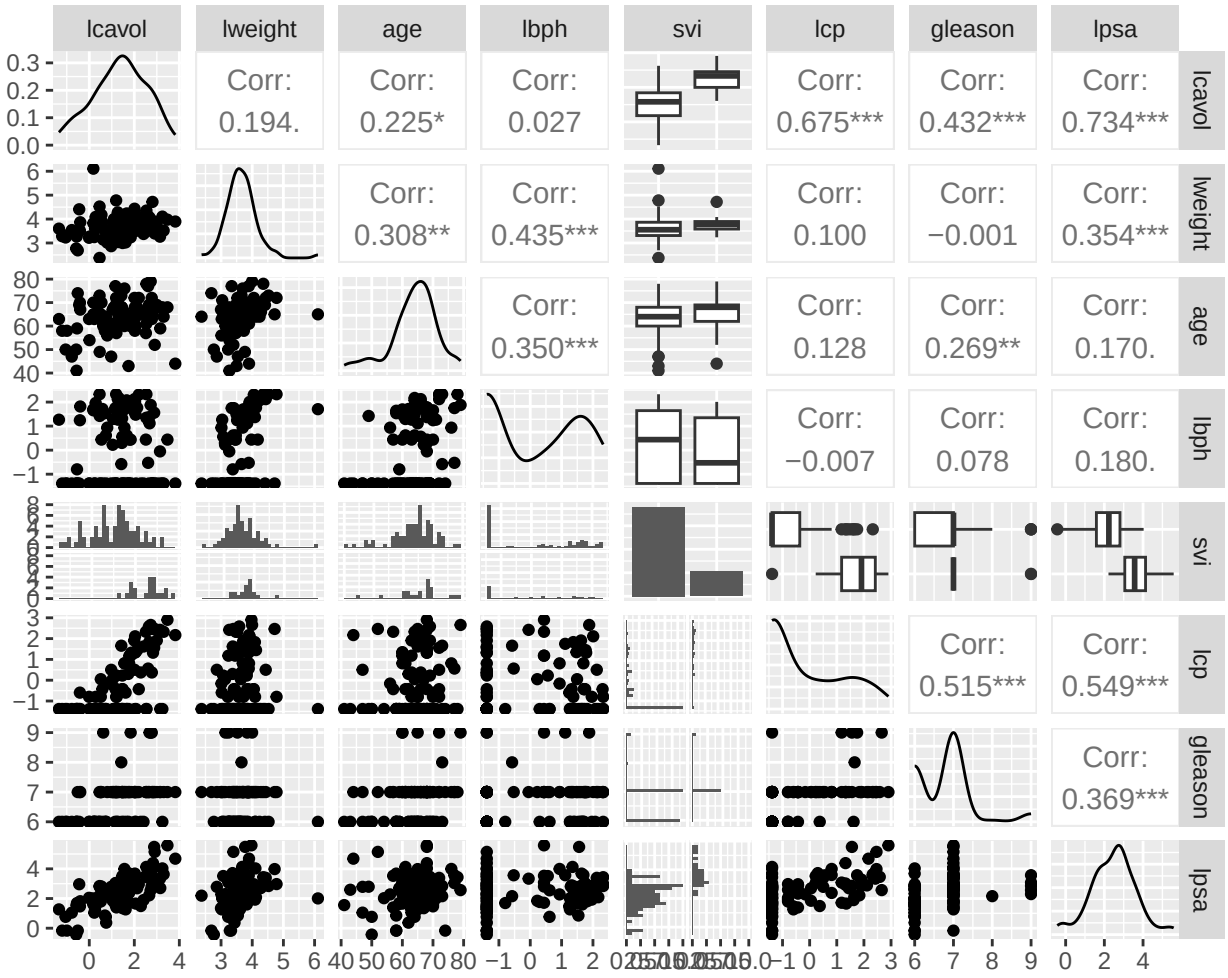
1.1 Prostate cancer example

- Prostate specific antigen (PSA) and a number of clinical variables for 97 males with radical prostatectomy.
- Association of PSA by
  - tumor volume (lcavol)
  - prostate weight (lweight)
  - age
  - benign prostate hypertrophy (lbph)
  - seminal vesicle invasion (svi)
  - capsular penetration (lcp)
  - Gleason score (gleason)
  - precentage gleason score 4/5 (pgg45)

```
prostate <- read_csv("https://raw.githubusercontent.com/GTPB/PSLS20/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
# i 87 more rows
```

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



---

2 Additive multiple linair model

Separate simple linair models, like

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

- Association between lpsa en 1 variabele e.g lcavol.
- More accurate predictions by simultaneously accounting for multiple predictors
- Estimate for parameter β_v does not only capture the effect of tumor volume.
- β_v average difference for log-psa for patients that differ in 1 unit of the log tumor volume.
- Even if lcavol is not associated with lpsa then patients with a higher tumor volume can have a higher lpsa because their semen vesicles are affected (svi status 1). $\rightarrow$ confounding.
- Compare patients with same svi status
- Is posible in multiple linear model

2.1 Statistical model

- $p-1$ predictors $X_1, ..., X_{p-1}$ and outcome Y for n subjecten.

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{ip-1} + \epsilon_i \quad (1)$$

- $\beta_0, \beta_1, ..., \beta_{p-1}$ unknown parameters
- ϵ_i residuals that cannot be explained by predictors
- Estimation by *least squares method*

Model allows to

1. predict the expected outcome for subjects given their values $x_1, ..., x_{p-1}$ for the predictor variables. $E[Y|X_1 = x_1, ...X_{p-1} = x_{p-1}] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_{p-1} x_{p-1}$.
2. Does the average outcome differ between two groups of patients that differ by δ units in predictor X_j but have the same value for the remaining variables $\{X_k, k = 1, ..., p, k \neq j\}$.

$$\begin{aligned} & E(Y|X_1 = x_1, ..., X_j = x_j + \delta, ..., X_{p-1} = x_{p-1}) \\ & \quad -E(Y|X_1 = x_1, ..., X_j = x_j, ..., X_{p-1} = x_{p-1}) \\ & \\ & = \beta_0 + \beta_1 x_1 + ... + \beta_j (x_j + \delta) + ... + \beta_{p-1} x_{p-1} \\ & \quad -\beta_0 - \beta_1 x_1 - ... - \beta_j x_j - ... - \beta_{p-1} x_{p-1} \\ & \\ & = \beta_j \delta \end{aligned}$$

Interpretation β_j:

- difference in mean outcome between subjects that differ in one unit of X_j, but have the same value for the remaining predictors in the model.

or

- Effect of predictor j corrected for the remaining predictors. e.g. effect of cancer volume correct for prostate weight and the svi status.

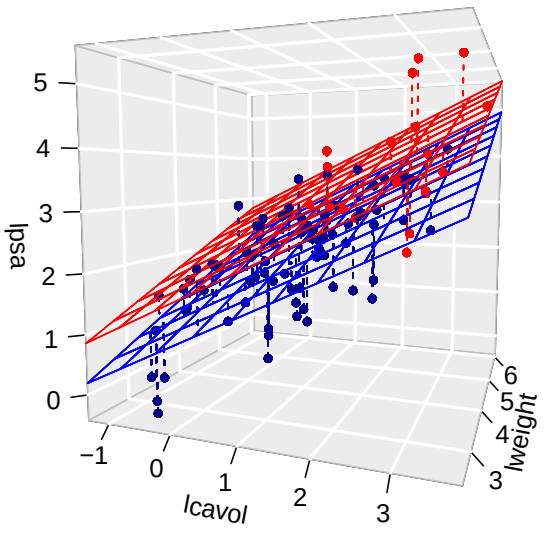
2.1.1 Prostate example

```
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
```

```r
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
```



3 Inference in multiple linear models

If data are representative than the least squares estimators for the intercept and slopes are unbiased.

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

- Gain insight in the distribution of the parameter estimators so as to generalize the effect in the sample to the population.
- Additional assumptions are needed for inference.

1. *Linearity*
2. *Independence*
3. *Homoscedasticity* of *equal variance*
4. *Normality*: residuals ϵ_i are normally distributed.

Under these assumptions:

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

and

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

- Slopes are again more precise if the predictor values have a larger range.
- Conditional variance (σ^2) can again be estimated based on the *mean squared error* (MSE):

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

Again hypothesis tests and confidence intervals by

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

If all assumptions are satisfied than the statistics T_k t-distributed with $n-p$ degrees of freedom.

When normality thus not hold, but lineariteit, independence and homoscedasticity are valid we can again adopt the CLT that states that statistic T_k is approximately normally distributed in large samples.

We can build confidence intervals on the slopes by:

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

```
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
```

Formal hypothesis tests:

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

With test statistic

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

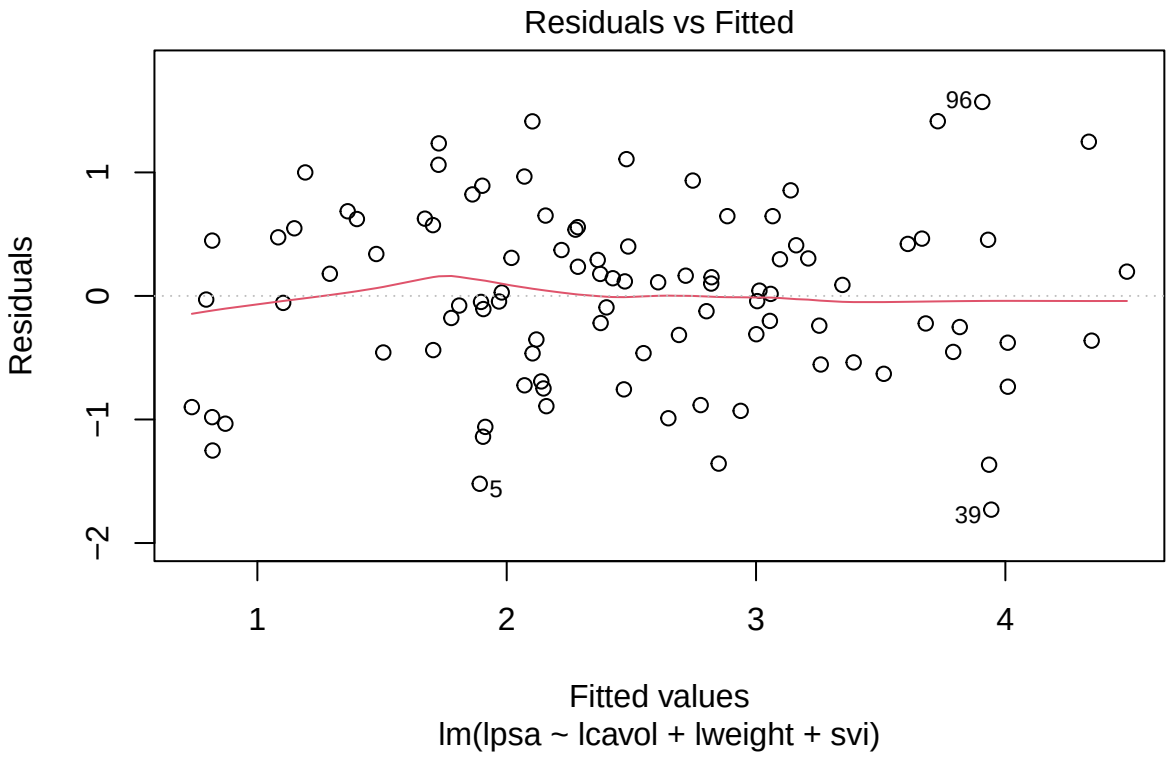
which follows a t-distribution with $n - p$ degrees of freedom under H_0

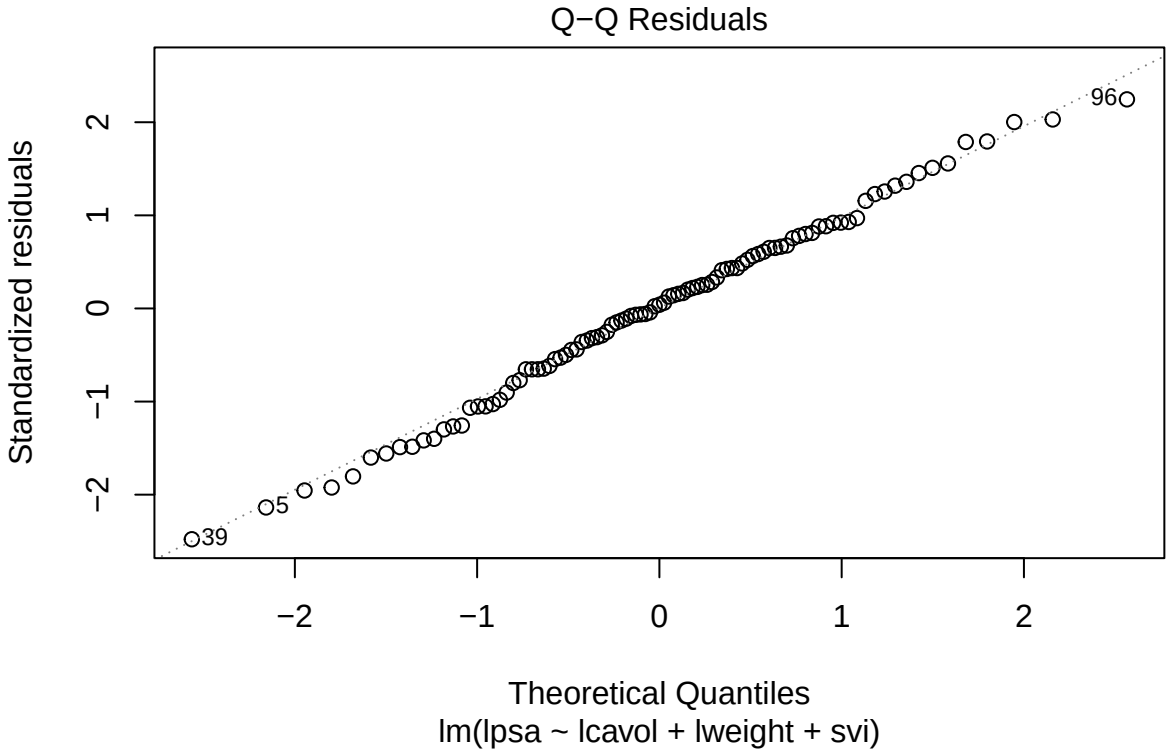
```
summary(lmVWS)
```

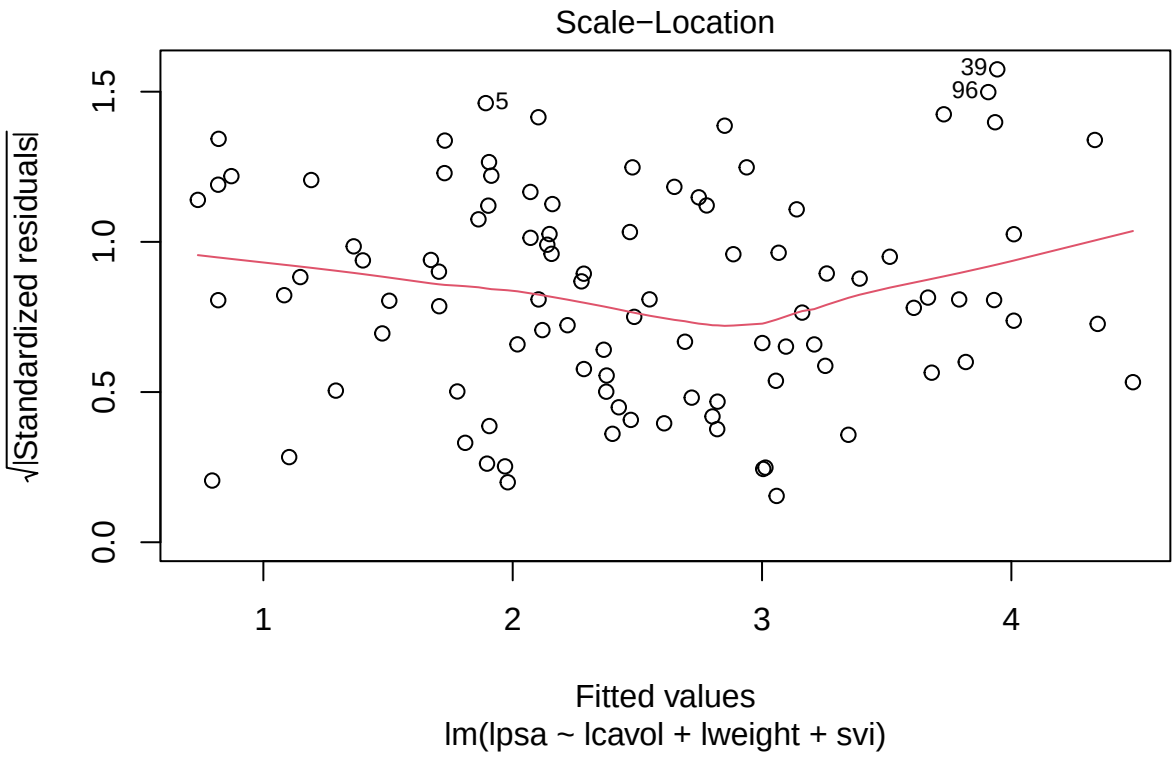
```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

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
```

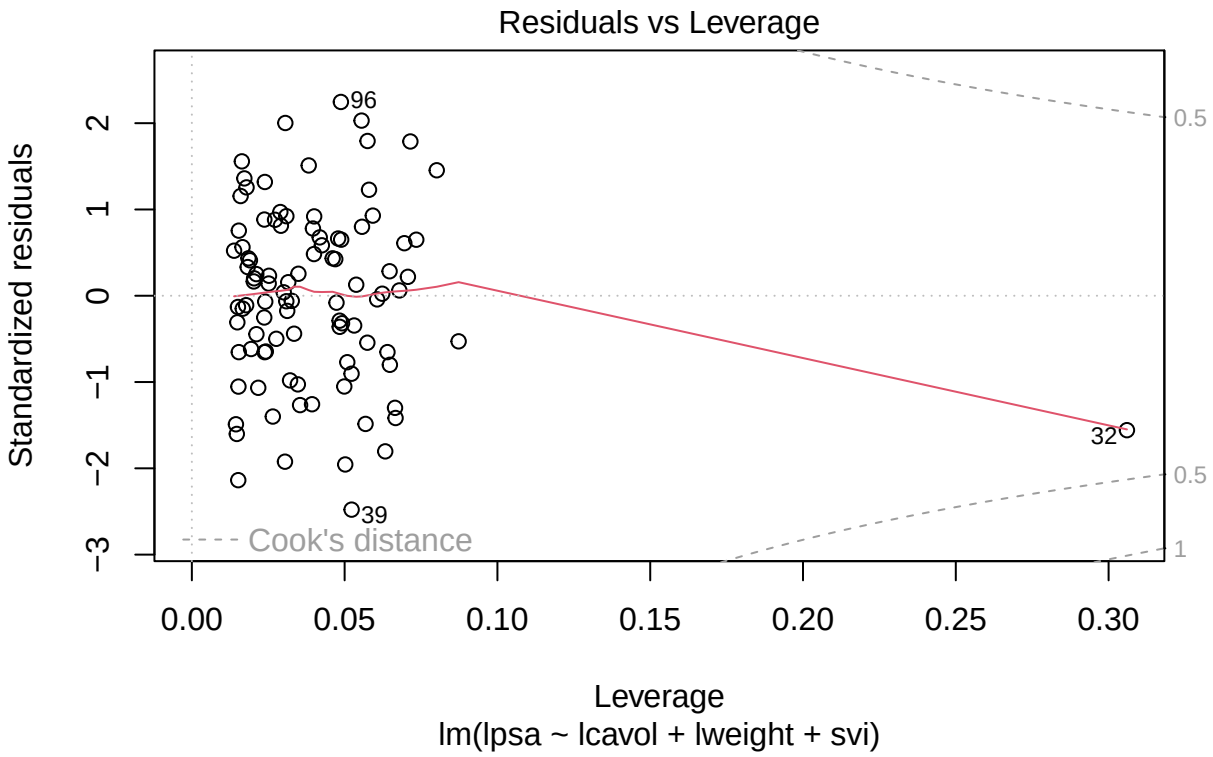
3.1 Assess the model assumptions

```
plot(lmVWS)
```


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


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


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 The non additive multiple linear model

3.2.1 Interaction between two continuous variables

The previous model is additive because the contribution of the cancer volume on lpsa does not depend on the height of the prostate weight and the svi status.

The slope for lcavol does not depend on log prostate weight and svi.

$$\beta_0 + \beta_v(x_v + \delta_v) + \beta_w x_w + \beta_s x_s - \beta_0 - \beta_v x_v - \beta_w x_w - \beta_s x_s = \beta_v \delta_v$$

The svi status and the log-prostategewicht (x_w) do not influence the contribution of the log-tumor volume (x_v) to the average log-PSA and vice versa.

- It is however possible that the association of lpsa and lcavol depends on the prostate weight.
- The average difference in lpsa for patients that differ in one unit of the log-tumor volume can for instance can be higher for patients wiht a high tumor weight then for those with a low tumor weight.
- The effect of the tumor volume on the PSA depends on the prostate weight.

To model this **interactie** or **effect modification** we can add a product term of both variables to the model

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

This term quantifies the *interactie-effect* of predictors x_v en x_w on the mean outcome.

Terms $\beta_v x_{iv}$ and $\beta_w x_{iw}$ are referred to as *main effects* of predictors x_v and x_w.

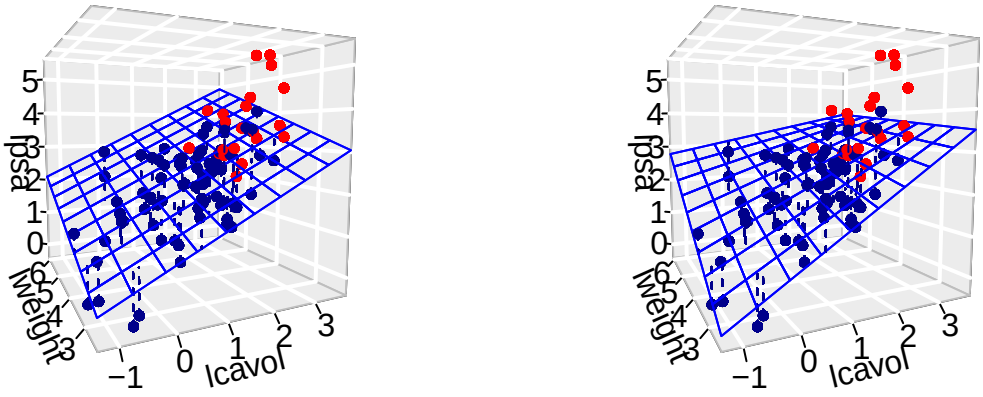
The difference in lpsa for patients that differ 1 unit in X_v and have an equal log prostate weight and the same svi status now becomes:

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

```
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
```

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



- Note that the interaction effect that is observed is not statistically significant (p=0.4).
- The main effects that are involved in the interaction cannot be interpreted separately from one another.
- We will therefore remove non-significant interaction terms from the model.
- Upon removal of non-significant interaction terms the main effects can be interpreted.

3.3 Interaction between a continuous variable and a factor variable

Interaction between lcavol $\leftrightarrow$ svi and lweight $\leftrightarrow$ svi.

The model becomes

$$Y = \beta_0 + \beta_v X_v + \beta_w X_w + \beta_s X_s + \beta_{vs} X_v X_s + \beta_{ws} X_w X_s + \epsilon$$

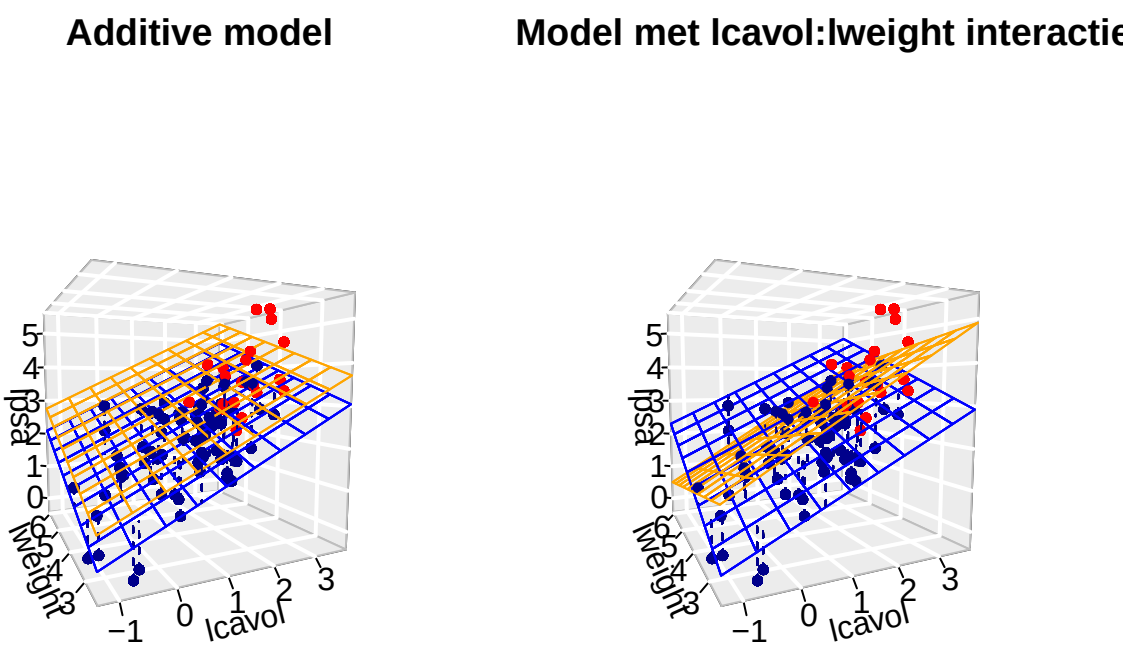
```
lmVWS_IntVS_WS <- lm(lpsa ~ lcavol + lweight + svi + svi:lcavol + svi:lweight, data = prostate)
summary(lmVWS_IntVS_WS)
```

```
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
```

Because X_S is a dummy variabele we obtain to distinct regression planes:

1. Model for $X_s = 0$:
$$Y = \beta_0 + \beta_v X_v + \beta_w X_w + \epsilon$$
where the main effects are the slope for lcavol and lweight
2. and model for $X_s = 1$:
$$\begin{aligned} Y &= \beta_0 + \beta_v X_v + \beta_s + \beta_w X_w + \beta_{vs} X_v + \beta_{ws} X_w + \epsilon \\ &= (\beta_0 + \beta_s) + (\beta_v + \beta_{vs}) X_v + (\beta_w + \beta_{ws}) X_w + \epsilon \end{aligned}$$
with intercept $\beta_0 + \beta_s$ and slopes $\beta_v + \beta_{vs}$ and $\beta_w + \beta_{ws}$



---

4 ANOVA table

The total SSTot is again

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

The residual sum of squares remains similar

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

Again the total sum of squares can be decomposed in ,

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

with

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

We have following degrees of freedom and mean sum of squares:

- SSTot has $n-1$ degrees of freedom and SSTot$/(n-1)$ is an estimator for the total variance in Y (marginal distribution of Y).
- SSE has $n-p$ degrees of freedom and MSE $=$ SSE$/(n-p)$ is an schatter for the residual variance of Y given the predictores (i.e. an estimator for the residual variance σ^2 of the error term ϵ).
- SSR has $p-1$ degrees of freedom and MSR $=$ SSR$/(p-1)$ is the mean sum of squares of the regression.

The determination coefficients remains as before, i.e.

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

and is the fraction of the total variability that can be explained by the regression model.

Teststatistic $F = \text{MSR/MSE}$ is under $H_0 : \beta_1 = ... = \beta_{p-1} = 0$ distributed by an F distribution: $F_{p-1;n-p}$.

```
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 Additional sums of squares

Consider 2 models for the predictors x_1 en x_2:

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

with ϵ_i iid $N(0, \sigma_1^2)$, and

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

with ϵ_i iid $N(0, \sigma_2^2)$.

for the first (gereduceerde) model we have decomposition

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

en for the second non-reduced model we have

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

(SSTot is of course the same because it only depends on the response and not of the models).

Definition of additional sum of squares The *additional sum of squares* of predictor x_2 as compared to the model with only x_1 as predictor is given by

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

Note that, $\text{SSE}_1 - \text{SSE}_2 = \text{SSR}_2 - \text{SSR}_1$ is triviaal is because of the decomposition of the total sum of squares.

The additional sum of squares $\text{SSR}_{2|1}$ can simply be interpreted as the additional variability that can be explained by adding predictor x_2 to the model with predictor x_1.

With this sum of squares we can further decompose the total sum of squares

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

which follows directly from the definition $\text{SSR}_{2|1}$.

Extension: $(s < p-1)$

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

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

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

with ϵ_i iid $N(0, \sigma_2^2)$.

The **additional sum of squares** of predictor $x_{s+1}, \ldots, x_q$ compared to a model with only predictors $x_1, \ldots, x_s$ is given by

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

4.1.1 Type I Sums of Squares

Suppose that $p-1$ predictors are considered, and suppose the following sequence of models $(s = 2, \ldots, p-1)$

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

wuth ϵ_i iid $N(0, \sigma^2)$.

- The corresponding sum of squares are denoted as SSR_s and SSE_s.
- The sequence of models gives rise to the following sums of squares: $\text{SSR}_{s|1,\ldots,s-1}$.
- The latter sum of squares is referred to as type I sums of squares. Note that they depend on the order in which the models were added to the model.

We can show for model Model with $s = p-1$ that

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

with SSE the residual sum of squares of the model with all $p-1$ predictors

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

with SSR the sum of squares of all $p-1$ predictors.

- The interpretation of each term depends on the order of the sequence of the regression models.

- Each type I SSR involves 1 predictor and has 1 degree of freedom (note that multiple dummies for a factor are typically removed together).
- For each type I SSR term the mean sum of squares is defined by $\text{MSR}_{j|1,\ldots,j-1} = \text{SSR}_{j|1,\ldots,j-1}/1$.
- And teststatistic $F = \text{MSR}_{j|1,\ldots,j-1}/\text{MSE}$ follows a $F_{1;n-(j+1)}$ distribution under $H_0 : \beta_j = 0$ with $s = j$.
- These sums of squares are the default sum of squares in the anova function of R.

4.1.2 Type III Sums of squares

Type III sum of squares for predictor x_j are given by the additional sum of squares

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

- SSE_2 the sum of squares of the residuals of the model with all $p-1$ predictors.
- SSE_1 sum of squares of the residuals with all $p-1$ predictors, except for predictor x_j.

The type III sum of squares $\text{SSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1}$ quantify the contribution in the total variance of the outcome explained by x_j that cannot be explained by the remaining $p-2$ predictors.

The type III sum of squares has 1 degree of freedom because it involves 1 β-parameter.

For each type III SSR term the mean sum of squares is defined by $\text{MSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1} = \text{SSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1}/1$.

Teststatistiek $F = \text{MSR}_{j|1,\ldots,j-1,j+1,\ldots,p-1}/\text{MSE}$ is $F_{1;n-p}$ distributed under $H_0 : \beta_j = 0$.

4.2 We can obtain these sums of squares using the `Anova` function from the `car` package.

```
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
```

The p-values are identical to those of two-sided t-tests

Note, however, that all dummies for factors with multiple levels will be taken out of the model at once. So then the type III sum of squares will have as many degrees of freedom as the number of dummies and an omnibus test is performed for the effect of the factor.

---

5 Diagnostics

5.1 Multicollinearity

```
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
```

- Estimates are different from those in the additive model and the standard errors are much higher!
- This is caused by the multicollinearity problem.
- If 2 predictors are strongly correlated than they share a lot of information.

- It is therefore difficult to estimate the individual contribution of each predictor on the outcome.
- Least squares estimators become instable.
- Standard errors become inflated.
- As long as we only do predictions on the basis of the regression model without extrapolating beyond the range of the predictors observed in the sample multicolinearity is not problematic.
- But for inference it is problematic.

```
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
```

- High correlation between log-tumor volume and interaction.
- It is a known problem for higher order terms (interactions and quadratic terms)

- Detect multicollinearity based on the correlation matrix or scatterplot matrix is suboptimal.
- In models with 3 or more predictors, say X1, X2, X3 we can have high multicollinearity while alle pairswise correlations between the predictors are low.
- We also have multicollinearity if there is a high correlation between X1 and a linair combination of X2 and X3.

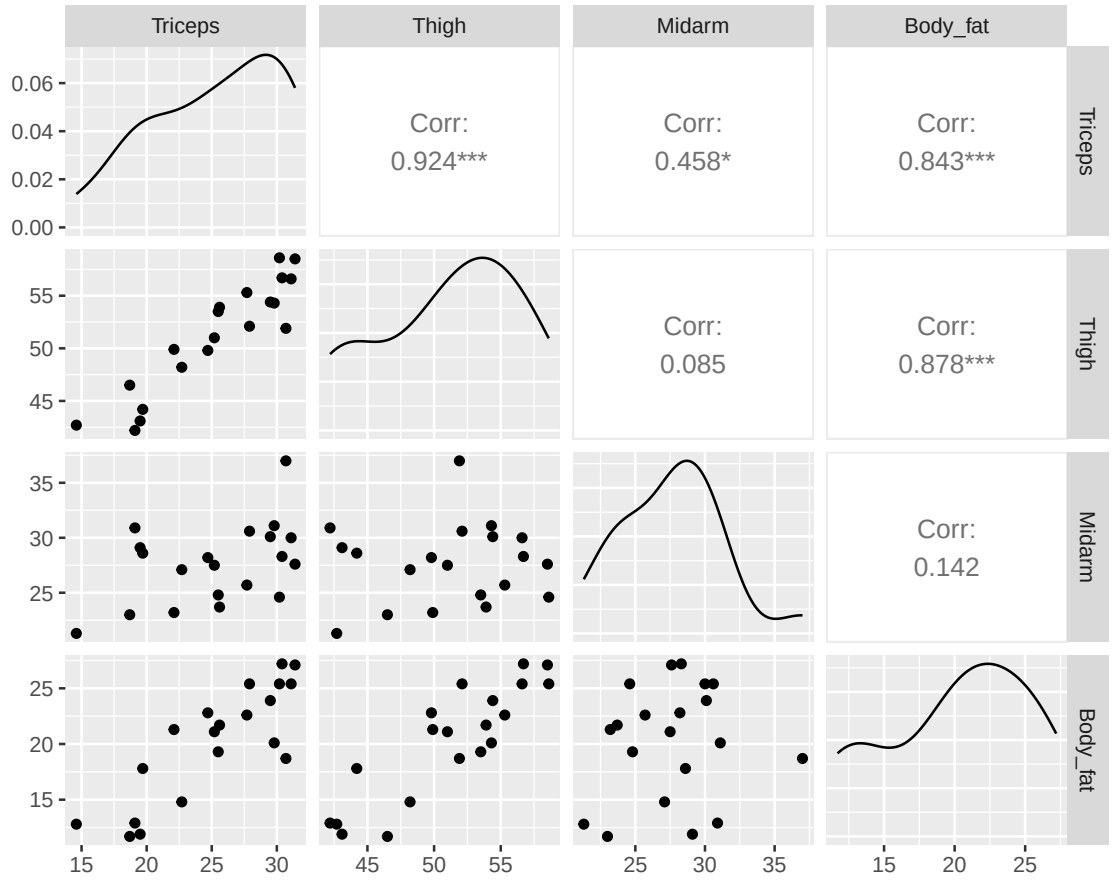
5.1.1 Variance inflation factor (VIF)

For parameter j in de regression model

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

- In this expression R_j^2 is the multiple determination coefficient of the linear regression of predictor j on the remaining predictors in the model.
- VIF is 1 if predictor j is not linear associated with the remaining predictors in the model.
- VIF is larger than 1 in all andere cases.
- VIF is the factor with which the observed variance inflates as compared to a model for which all predictoren would be independend.
- VIF $> 10 \rightarrow$ strong multicollinearity.

5.1.2 Body fat example



```
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 evaluate the VIF in the prostate cancer example for the additive model and the model with interactie.

```
vif(lmVWS)
```

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

```
vif(lmVWS_IntVW)
```

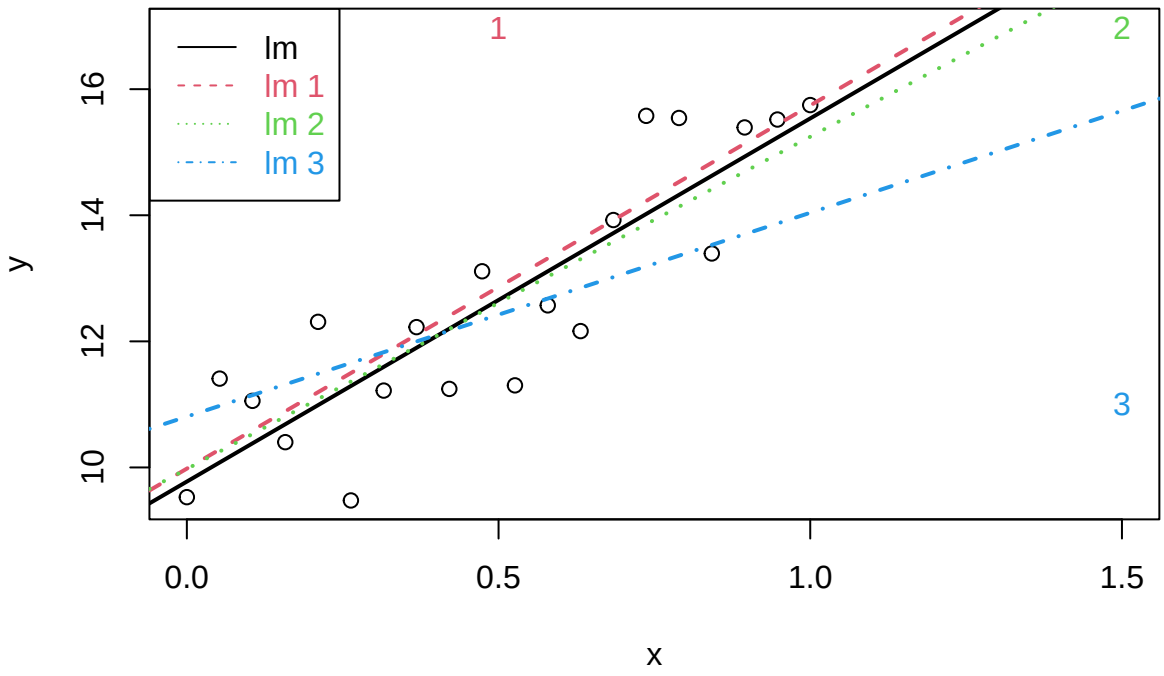
```
       lcavol        lweight            svi lcavol:lweight
    76.193815       1.767121       1.426646      80.611657
```

- Inflation in interaction terms often caused because main effect get another interpretation.

5.2 Influential observations

```
set.seed(112358)
nobs <- 20
sdy <- 1
x <- seq(0, 1, length = nobs)
y <- 10 + 5 * x + rnorm(nobs, sd = sdy)
x1 <- c(x, 0.5)
y1 <- c(y, 10 + 5 * 1.5 + rnorm(1, sd = sdy))
x2 <- c(x, 1.5)
y2 <- c(y, y1[21])
x3 <- c(x, 1.5)
y3 <- c(y, 11)
plot(x, y, xlim = range(c(x1, x2, x3)), ylim = range(c(y1, y2, y3)))
points(c(x1[21], x2[21], x3[21]), c(y1[21], y2[21], y3[21]), pch = as.character(1:3), col = 2:4)
```

```r
abline(lm(y ~ x), lwd = 2)
abline(lm(y1 ~ x1), col = 2, lty = 2, lwd = 2)
abline(lm(y2 ~ x2), col = 3, lty = 3, lwd = 2)
abline(lm(y3 ~ x3), col = 4, lty = 4, lwd = 2)
legend("topleft", col = 1:4, lty = 1:4, legend = paste("lm", c("", as.character(1:3))), text.col = 1:4)
```



- It is not desirable that a single observation largely influences the result of a linear regression analysis
- Diagnostics allow us to detect extreme observations.
- *Studentized residuals* to spot outliers
- *Leverage* to spot observations with extreem covariate pattern

---

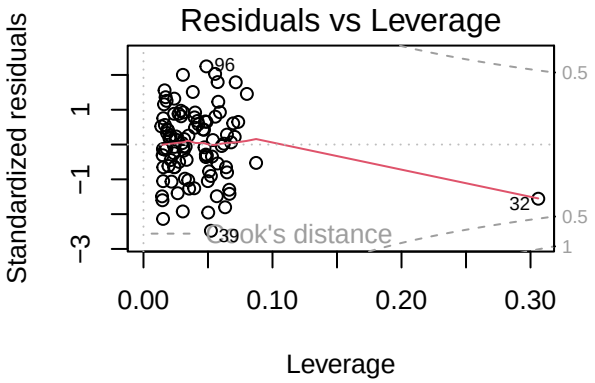
5.2.1 Cook's distance

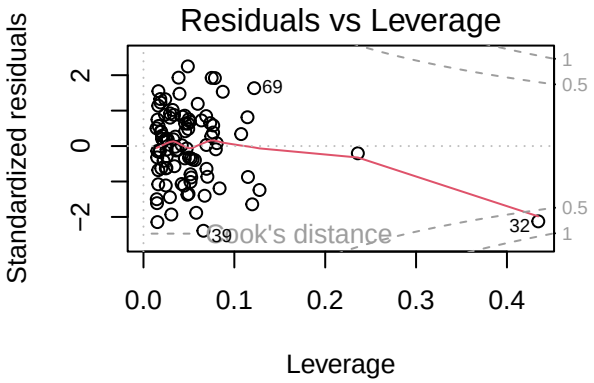
- A statistics to assess the influence the effect of a single observation on the regression analysis
- Cook's distance for observation i is diagnostic measure for this particular observation on all all predictions or on *all* estimated parameters.

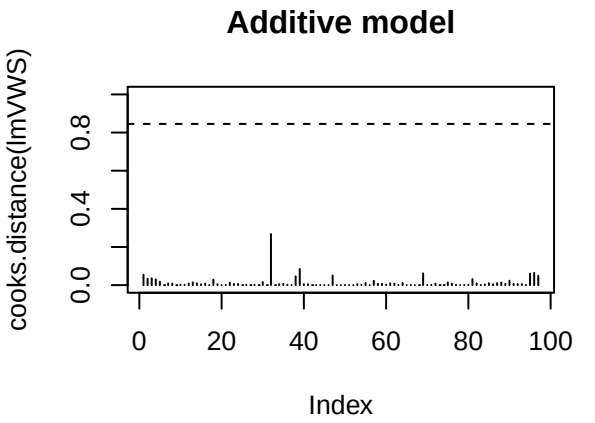
$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\text{MSE}}$$

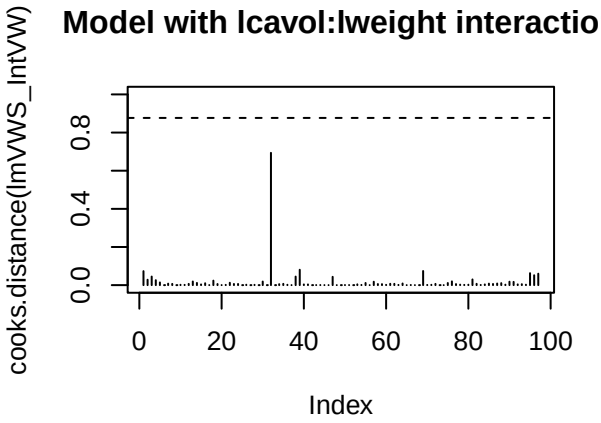
- Observation i has a large influence on the regression parameters and predictions if the Cook's distance D_i is large.

- Extreme Cook's distance if it is larger than the 50% quantile of an $F_{p+1,n-(p+1)}$-distribution.





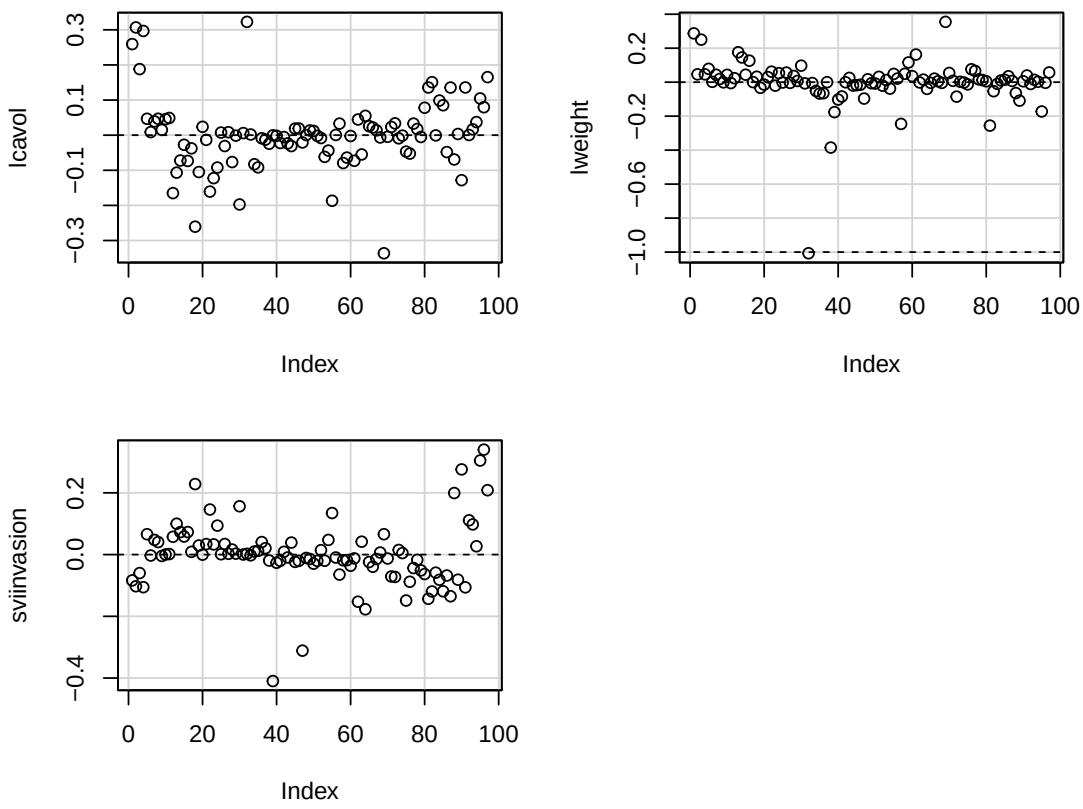


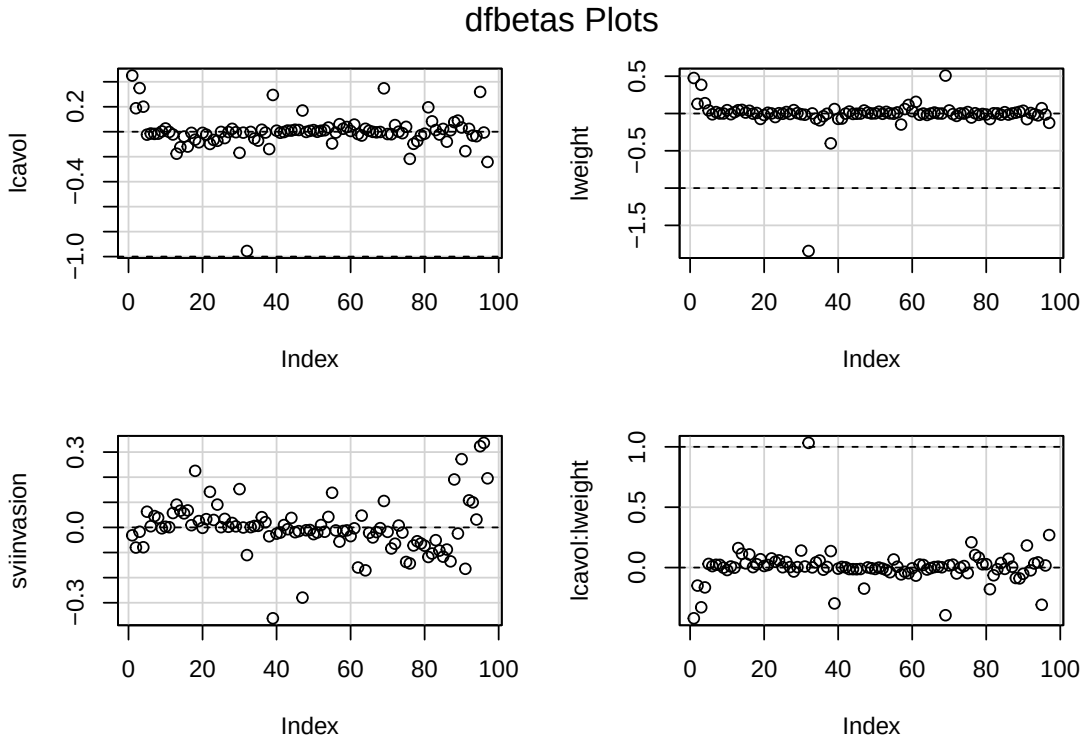


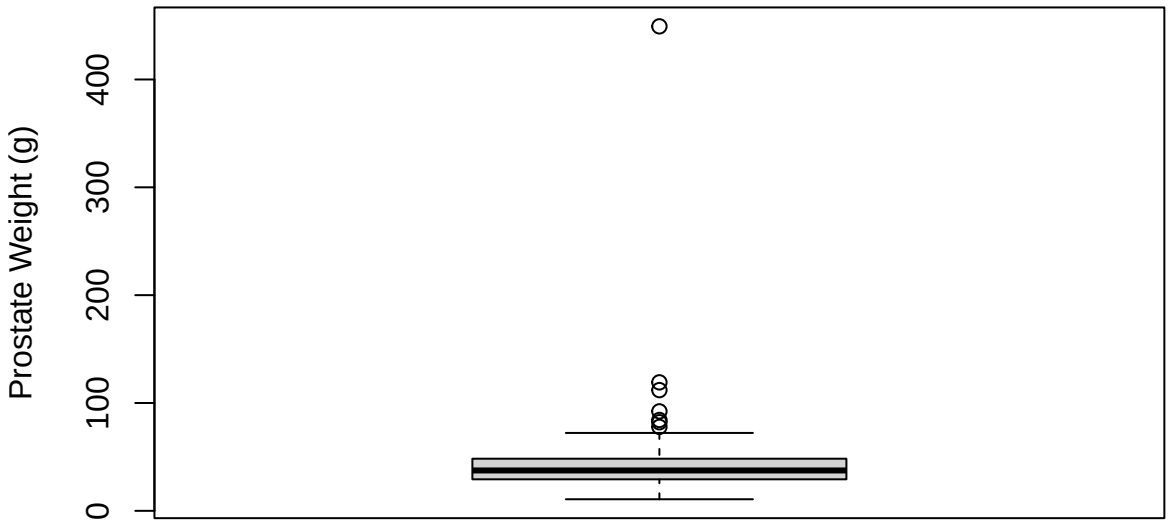
- Once we established that an observation is influential we can use *DFBETAS* to find the parameters for which the estimates are largely affected by the observation
- DFBETAS of observatie i is a diagnostic measure for *each model parameter separately.*

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

- DFBETAS is extreme when it is larger than 1 in small to moderate datasets or exceeds $2/\sqrt{n}$ in large datasets.


dfbetas Plots
lcavol
lweight
svi
Index
Index
Index


dfbetas Plots
lcavol
lweight
svi:invasion
lcavol:lweight
Index
Index
Index
Index


Prostate Weight (g)
0
100
200
300
400

6 Constrasts

- In more complex designs that are modelled using general linear models one often has to assess multiple hypotheses.
- Moreover these hypotheses can typically not always be translated into a test on one parameter, but in a linear combination of model parameters.
- A linear combination of model parameters is also referred to as a contrast.

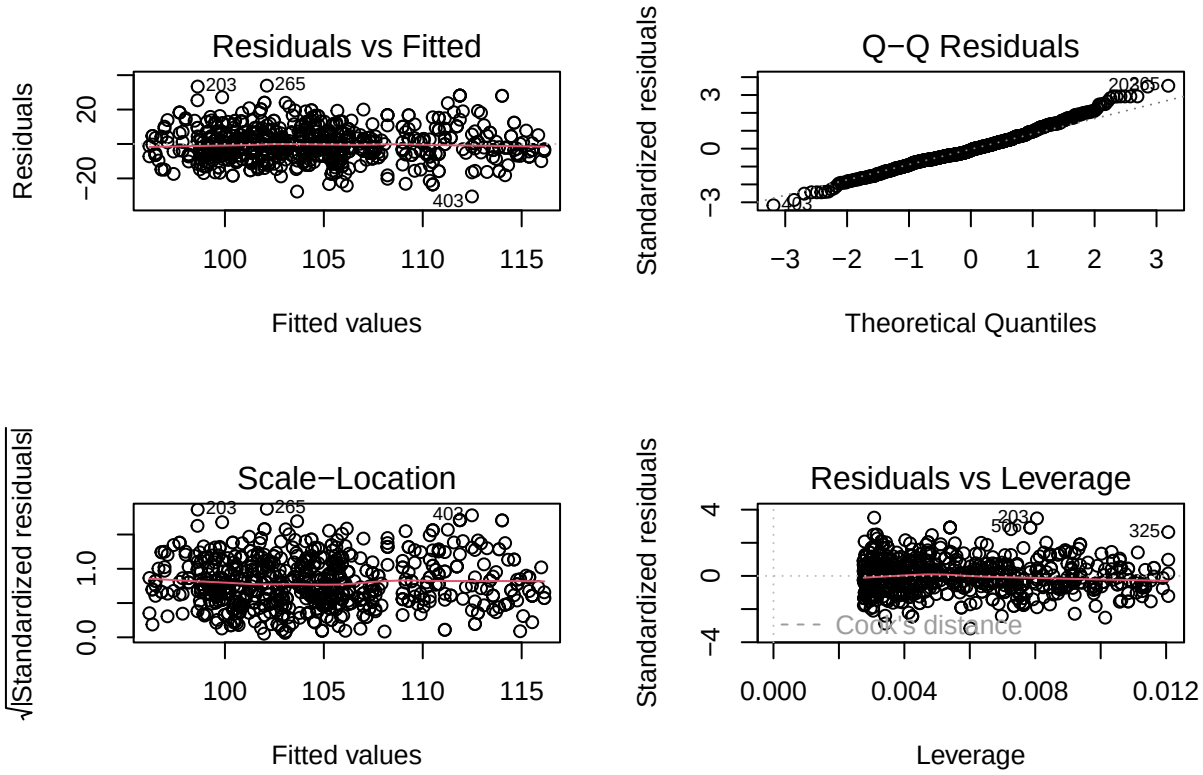
6.1 NHANES example

- Suppose that researchers want to assess the association between age and bloodpressure for American children.
- Possibly this association will differ between boys and girls.
- They want to assess following hypotheses:
  - Is there an association between age and blood pressure for girls?
  - Is there an association between age and blood pressure for boys?
  - Is the association between age and blood pressure different for boys and girls?

6.2 Model

We fit a model for the average systolic blood pressure (`BPSysAve`) using age (in months), gender and the interaction between age and gender for children between 6 and 18 years from the NHANES study.

```
library(NHANES)
bpData <- NHANES %>%
  filter(
      Age >=6 &
      Age <= 18 &
      !is.na(BPSysAve) &
      !is.na(AgeMonths)
  )

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



Assumptions?

- No deviations from Lineariteit
- Assumption of homoscedasticity seems to be valid
- Slight deviations from normality, indication for some tail to the right
- Large dataset (n = 703) so we can adopt the CLT

6.3 Inference

```
summary(mBp1)
```

```
Call:
lm(formula = BPSysAve ~ AgeMonths * Gender, data = bpData)

Residuals:
    Min      1Q  Median      3Q     Max
-30.487  -5.871  -0.890   5.265  33.882

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          92.90682    2.29792  40.431  < 2e-16 ***
AgeMonths             0.05943    0.01371   4.336 1.66e-05 ***
Gendermale          -11.35031    3.25237  -3.490 0.000514 ***
AgeMonths:Gendermale  0.09294    0.01943   4.783 2.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 9.65 on 699 degrees of freedom
Multiple R-squared:  0.1939,    Adjusted R-squared:  0.1904
F-statistic: 56.04 on 3 and 699 DF,  p-value: < 2.2e-16
```

The research questions translate to following nullhypotheses:

1. Association between blood pressure and age for girls?

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

2. Association between blood pressure and age for boys?

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

3. Is the association between blood pressure and age different for girls and boys?

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

- We can assess hypotheses 1 and 3 immediately using the output of the model.
- Hypotheses 2 is a linear combination of two parameters.
- We also need multiple tests for assessing the association between the systolic blood pressure and Age.

We can again use an Anova approach.

1. We first assess the omnibus hypothesis that there is no association between age and blood pressure.

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

- which simplifies to assessing

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

- We can do this by comparing two models: the full model with an effect for Gender, AgeMonths and Gender x AgeMonths interaction against a reduced model with only Gender.

2. If we can reject this hypothesis we can again do a posthoc analysis for each of the contrasts.

6.3.1 Omnibus test

```
mBp0 <- lm(BPSysAve ~ Gender, bpData)
anova(mBp0, mBp1)
```

```
Analysis of Variance Table

Model 1: BPSysAve ~ Gender
Model 2: BPSysAve ~ AgeMonths * Gender
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    701 78239
2    699 65095  2     13145 70.576 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is an extremely significant association between the systolic blood pressure and Age ($p << 0.001$).

6.3.2 Posthoc tests

For the posthoc tests we will again build upon the `multcomp` package.

```r
library(multcomp)
bpPosthoc <- glht(mBp1, linfct = c(
  "AgeMonths = 0",
  "AgeMonths + AgeMonths:Gendermale = 0",
  "AgeMonths:Gendermale = 0"
))
bpPosthoc %>% summary()
```

```
     Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = BPSysAve ~ AgeMonths * Gender, data = bpData)

Linear Hypotheses:
                                     Estimate Std. Error t value Pr(>|t|)
AgeMonths == 0                        0.05943    0.01371   4.336 3.69e-05 ***
AgeMonths + AgeMonths:Gendermale == 0  0.15237    0.01377  11.061  < 1e-05 ***
AgeMonths:Gendermale == 0              0.09294    0.01943   4.783  < 1e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```r
bpPosthocCI <- bpPosthoc %>% confint()
bpPosthocCI
```

```
     Simultaneous Confidence Intervals

Fit: lm(formula = BPSysAve ~ AgeMonths * Gender, data = bpData)

Quantile = 2.3215
95% family-wise confidence level


Linear Hypotheses:
                                     Estimate lwr     upr
AgeMonths == 0                       0.05943  0.02761 0.09124
AgeMonths + AgeMonths:Gendermale == 0 0.15237  0.12039 0.18434
AgeMonths:Gendermale == 0             0.09294  0.04783 0.13805
```

Note that the glht function allows us to define the contrasts by explicitely defining the nullhypotheses using the names of the model parameters.

6.4 Conclusion

We can conclude that the association between age and blood pressure is extremely significant (p « 0.001).

The blood pressure for girls that differ in age is on average 0.059 mm Hg higher per month of age difference for the eldest girl (p « 0.001, 95% CI [0.028, 0.091].

The blood pressure for boys that differ in age is on average 0.152 mm Hg higher per month of age difference for the eldest boy (p « 0.001, 95% CI [0.12, 0.184].

The average blood pressure difference between subjects that differ in age is on average 0.093 mm Hg/month higher for boys than for girls (p « 0.001, 95% CI [0.048, 0.138]).