UNIVERSITEIT GENT

BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

Modelselectie

Lieven Clement

2^{de} bach. in de Biologie, Chemie, Biochemie en Biotechnologie en Biomedische Wetenschappen

Automatische selectieprocedures

1. De uitkomst op basis van de predictoren te voorspellen
2. Associaties tussen de uitkomst en predictoren te beschrijven

- Geobserveerde dataset (steekproef) gebruiken om tot een model te komen dat ruimer toepasbaar is dan enkel op de geobserveerde data.
- Geselecteerde model moet toepasbaar zijn op ruimere populatie waaruit de steekproefdata bekomen is en waarop de onderzoeksvraag van toepassing is.

Modelselectie

- Methode voor het selecteren van model dat best geschikt is voor beantwoorden van onderzoeksvraag
- We beperken ons tot meervoudige lineaire regressiemodellen waarin sommige van de $p-1$ effecttermen interactietermen voorstellen.
- Het totaal aantal mogelijke modellen is dan (zonder hiërarchie-restrictie) 2^{p-1}
- We onderscheiden de volgende algemene procedures:
  1. Alle mogelijke modellen worden geëvalueerd. Deze methode zal enkel haalbaar blijken als het aantal kandidaat modellen niet te groot is.
  2. Niet alle modellen worden geëvalueerd (stapsgewijze procedures):
     - Initialisatie met een model
     - Regel en een evaluatiecriterium om pad doorheen modelruimte te bepalen
- Hiërarchisch modelleren als we interactietermen toelaten: hogere orde term (vb. interactie) nooit in model zonder lagere orde termen (vb. hoofdeffecten).

Modelselectie op basis van hypothesetesten

Voorbeeld: lpsa waarbij lcavol, lweight en svi als predicoren worden toegelaten, alsook alle paarsgewijze interactietermen. Drie strategieën:

1. Voorwaartse modelselectie
2. Achterwaartse modelselectie
3. Stapsgewijze modelselectie

Voorwaartse modelselectie

1. Start met het minimale model m_1 met enkel het intercept
2. Test voor alle modellen met slechts 1 predictor (er zijn dus $p-1$ dergelijke modellen) of de toegevoegde regressor significant verschillend is van 0 op het α_{IN} significantieniveau.
3. Indien geen enkele van de p-waarden kleiner is dan α_{IN}, stop de procedure. Het finale model is het minimale model m_1,

$$Y_i = \beta_0 + \epsilon_i$$

   met ϵ_i i.i.d. $N(0, \sigma^2)$.
4. Indien er minstens 1 p-waarde kleiner is dan α_{IN}, selecteer het model dat overeenkomt met de kleinste p-waarde. Dit model wordt m_2 genoemd.
5. Stel $k = 2$.

6. Definieer de verzameling van modellen met 1 predictor toegevoegd aan m_k. Test dat de parameter van de toegevoegde regressor significant verschillend is van 0 op het α_{IN} significantieniveau.
7. Indien geen enkel van de p-waarden kleiner is dan α_{IN}, stop de procedure. Het finale model is

$$Y_i = \beta_0 + \sum_{j \in m_k} \beta_j x_{ij} + \epsilon_i$$

   met ϵ_i i.i.d. $N(0, \sigma^2)$.
8. Indien er minstens één p-waarde kleiner is dan α_{IN}, selecteer het model dat overeenkomt met de kleinste p-waarde. Dit model wordt m_{k+1} genoemd.
9. Verhoog k met 1 (i.e. $k \leftarrow k + 1$) en ga naar stap 6

Voorbeeld voor lpsa.

- We laten interactietermen toe
- Modelbouw moet zich houden aan hiërarchie van hoofd- en interactietermen
- We stellen $\alpha_{\text{IN}} = 5\%$.
- We starten van model zonder predictor

```
m1 <- lm(lpsa ~ 1, data=prostate)
add1(m1,scope=~lweight+lcavol+svi+lweight:lcavol+lweight:svi +lc
```

```
## Single term additions
##
## Model:
## lpsa ~ 1
##         Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>                127.918  28.838
## lweight  1    16.041 111.877  17.841  13.621  0.000373 ***
## lcavol   1    69.003  58.915 -44.366 111.267 < 2.2e-16 ***
## svi      1    41.011  86.907  -6.658  44.830 1.499e-09 ***
## ---
```

```r
m2 <- update(m1,~. + lcavol)
add1(m2,scope=~lweight+lcavol+svi+lweight:lcavol+lweight:svi +lc
```

```
## Single term additions
##
## Model:
## lpsa ~ lcavol
##         Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                58.915 -44.366
## lweight  1    5.9484 52.966 -52.690 10.5567 0.001606 **
## svi      1    5.2375 53.677 -51.397  9.1719 0.003172 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```r
m3 <- update(m2,~ . + lweight)
add1(m3,scope=~lweight+lcavol+svi+lweight:lcavol+lweight:svi +lc
```

```
## Single term additions
##
## Model:
## lpsa ~ lcavol + lweight
##                Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                      52.966 -52.690
## svi             1    5.1814 47.785 -60.676  10.084 0.002029 *
## lweight:lcavol  1    0.1222 52.844 -50.914   0.215 0.643962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
m4 <- update(m3,~ . + svi)
add1(m4,scope=~lweight+lcavol+svi+lweight:lcavol+lweight:svi +lc
```

```
## Single term additions
##
## Model:
## lpsa ~ lcavol + lweight + svi
##                Df Sum of Sq    RSS     AIC F value Pr(>F)
## <none>                      47.785 -60.676
## lweight:lcavol  1   0.36606 47.419 -59.422  0.7102 0.4016
## lweight:svi     1   1.16445 46.621 -61.069  2.2979 0.1330
## lcavol:svi      1   0.02433 47.761 -58.725  0.0469 0.8291
```

```r
summary(m4)
```

```
##
## 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 ' '
##
## Residual standard error: 0.7168 on 93 degrees of freedom
```

Achterwaartse modelselectie

- Achterwaartse modelselectie ook achterwaartse eliminatie genoemd (Engels: *backward elimination*).
- Procedure start van het maximale model
- Gaat na welke predictoren uit het model kunnen worden verwijderd.

1. Start met het maximale model m_1 met alle $p-1$ regressoren
2. Bouw alle modellen waar 1 predictor wordt weggelaten en ga via een F-test na of het verwijderen van de predictoren significant is op het α_{OUT} significantieniveau. Merk op dat deze test in model m_1 moet uitgevoerd worden.
3. Indien geen enkele van de p-waarden groter is dan α_{OUT}, stop de procedure. Het finale model is het maximale model m_1.
4. Indien er minstens één p-waarde groter dan α_{OUT} is, selecteer het model dat overeenkomt met de grootste p-waarde. Dit model wordt m_2 genoemd.
5. Stel $k = 2$

6. Bouw alle modellen waar 1 predictor wordt weggelaten uit m_k en ga via een F-test na of het verwijderen van de predictoren significant is op het α_{OUT} significantieniveau. Merk op dat deze test in model m_k moet uitgevoerd worden.
7. Indien geen enkel van de p-waarden groter is dan α_{OUT}, stop de procedure. Het finale model is m_k
8. Indien er minstens één p-waarde groter is dan α_{OUT}, selecteer het model dat overeenkomt met de grootste p-waarde. Dit model wordt m_{k+1} genoemd.
9. Verhoog k met één (i.e. $k \leftarrow k + 1$) en ga naar stap 6.

```r
alphaOut <- 0.05
m<- lm(lpsa~lweight+lcavol+svi+lweight:lcavol+lweight:svi +lcavo
dropHlp<-drop1(m,test="F")
dropHlp
```

```
## Single term deletions
##
## Model:
## lpsa ~ lweight + lcavol + svi + lweight:lcavol + lweight:svi
##     lcavol:svi
##                Df Sum of Sq    RSS     AIC F value Pr(>F)
## <none>                      46.478 -57.365
## lweight:lcavol  1   0.00012 46.479 -59.365  0.0002 0.9878
## lweight:svi     1   0.87404 47.353 -57.558  1.6925 0.1966
## lcavol:svi      1   0.14093 46.619 -59.071  0.2729 0.6027
```

```
while (sum(dropHlp[,6]>=alphaOut,na.rm=TRUE))
{
dropVarName<-rownames(dropHlp)[which.max(dropHlp[,6])]
m <- update(m,as.formula(paste("~.-",dropVarName)))
dropHlp<-drop1(m,test="F")
print(dropHlp)
}
```

```
## Single term deletions
##
## Model:
## lpsa ~ lweight + lcavol + svi + lweight:svi + lcavol:svi
##              Df Sum of Sq    RSS     AIC F value Pr(>F)
## <none>                     46.479 -59.365
## lweight:svi  1    1.2820 47.761 -58.725  2.5100 0.1166
## lcavol:svi   1    0.1419 46.621 -61.069  0.2778 0.5994

## Single term deletions
##
## Model:
## lpsa ~ lweight + lcavol + svi + lweight:svi
##              Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                     46.621 -61.069
## lcavol       1   28.1995 74.820 -17.184 55.6484 4.711e-11 ***
## lweight:svi  1    1.1644 47.785 -60.676  2.2979     0.133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
## Single term deletions
##
## Model:
## lpsa ~ lweight + lcavol + svi
##         Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                47.785 -60.676
## lweight  1    5.8923 53.677 -51.397  11.468  0.001039 **
## lcavol   1   28.0446 75.830 -17.884  54.581 6.304e-11 ***
## svi      1    5.1814 52.966 -52.690  10.084  0.002029 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Stapsgewijze modelselectie

- Combinatie van voorwaartse en achterwaartse modelselectie
- In elke stap zal eerst worden gekeken of een predictor kan worden toegevoegd aan het model.
- Na toevoeging wordt vervolgens nagegaan of een predictor kan worden verwijderd van het model.
- Kan starten van het maximale of het minimale model.
- We implementeren de methode, startende van het minimale model.

```
alphaIn <- alphaOut <- 0.05
m <- lm(lpsa ~ 1 ,prostate)
notConverged <- TRUE
while (notConverged)
{
addHlp <- add1(m,scope= ~lweight+lcavol+svi+
        lweight:lcavol+lweight:svi +lcavol:svi,test="F")
print(addHlp)
addVarName<-rownames(addHlp)[which.min(addHlp[,6])]
if ((sum(addHlp[,6] < alphaIn,na.rm=TRUE))>0)
        m <- update(m,as.formula(paste("~.+",addVarName)))
dropHlp<-drop1(m,test="F")
print(dropHlp)
if ((sum(dropHlp[,6]>=alphaOut,na.rm=TRUE))>0)
        m <- update(m,as.formula(paste("~.-",dropVarName)))
if ((sum(addHlp[,6] < alphaIn,na.rm=TRUE) + sum(dropHlp[,6]>=alp
        notConverged <- TRUE else notConverged <- FALSE
}
```

```
## Single term additions
##
## Model:
## lpsa ~ 1
##          Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>                 127.918  28.838
## lweight   1    16.041 111.877  17.841  13.621  0.000373 ***
## lcavol    1    69.003  58.915 -44.366 111.267 < 2.2e-16 ***
## svi       1    41.011  86.907  -6.658  44.830 1.499e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
## Single term deletions
##
## Model:
## lpsa ~ lcavol
##         Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>                58.915 -44.366
## lcavol   1    69.003 127.918  28.838  111.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
## Single term additions
##
## Model:
## lpsa ~ lcavol
##         Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                58.915 -44.366
## lweight  1    5.9484 52.966 -52.690 10.5567 0.001606 **
## svi      1    5.2375 53.677 -51.397  9.1719 0.003172 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
## Single term deletions
##
## Model:
## lpsa ~ lcavol + lweight
##          Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>                 52.966 -52.690
## lcavol    1    58.910 111.877  17.841 104.549 < 2.2e-16 ***
## lweight   1     5.948  58.915 -44.366  10.557  0.001606 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
## Single term additions
##
## Model:
## lpsa ~ lcavol + lweight
##                Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                      52.966 -52.690
## svi             1    5.1814 47.785 -60.676  10.084 0.002029 *
## lweight:lcavol  1    0.1222 52.844 -50.914   0.215 0.643962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
## Single term deletions
##
## Model:
## lpsa ~ lcavol + lweight + svi
##         Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                47.785 -60.676
## lcavol   1   28.0446 75.830 -17.884  54.581 6.304e-11 ***
## lweight  1    5.8923 53.677 -51.397  11.468  0.001039 **
## svi      1    5.1814 52.966 -52.690  10.084  0.002029 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
## Single term additions
##
## Model:
## lpsa ~ lcavol + lweight + svi
##                Df Sum of Sq    RSS     AIC F value Pr(>F)
## <none>                      47.785 -60.676
## lweight:lcavol  1   0.36606 47.419 -59.422  0.7102 0.4016
## lweight:svi     1   1.16445 46.621 -61.069  2.2979 0.1330
## lcavol:svi      1   0.02433 47.761 -58.725  0.0469 0.8291
```

```
## Single term deletions
##
## Model:
## lpsa ~ lcavol + lweight + svi
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                47.785 -60.676
## lcavol    1   28.0446 75.830 -17.884  54.581 6.304e-11 ***
## lweight   1    5.8923 53.677 -51.397  11.468  0.001039 **
## svi       1    5.1814 52.966 -52.690  10.084  0.002029 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Opmerkingen

- Prostaat voorbeeld: De drie methoden resulteerden in hetzelfde finale geselecteerde model (met enkel hoofdeffecten van lcavol, lweight en svi). In het algemeen leiden de drie methoden niet noodzakelijk tot eenzelfde finale model.
-
- De *p*-waarden in het geselecteerde model mogen niet geïnterpreteerd worden.
  - We hebben immers meerdere hypothese testen uitgevoerd om tot dit model te komen.
  - Bovendien zorgt de modelselectie ervoor dat enkel "significante" termen in het finale model staan; de selectieprocedure is dus een doelbewuste aanrijkingsprocedure.
  - Definitie van de *p*-waarde: de kans onder H_0 dat de teststatistiek "door zuiver toeval" minstens zo extreem is dan deze die waargenomen werd.
  - Door de aanrijking via de modelselectieprocedure, is het niet langer "zuiver toeval" dat de teststatistiek groter is dan de geobserveerde.

- Er is geen theoretische basis voor de selectieprocedure.
  - Intuïtief lijkt de procedure zinvol omdat we gewend zijn om beslissingen te nemen aan de hand van hypothesetesten, maar hypothesetesten zijn eigenlijk enkel geldig wanneer ze doelgericht toegepast worden om een vooraf (i.e. voor het observeren van de data) duidelijk omschreven onderzoeksvraag te beantwoorden.
  - Bij modelselectie wordt het pad doorheen de modellenruimte bepaald door de geobserveerde data, waardoor de p-waarden hun betekenis verliezen
  - Hypotheses worden niet vooraf door de onderzoekers geformuleerd, maar ze worden door de data bepaald.

- De analyse van Ipsa in het vorige hoofdstuk zou ook als modelselectie kunnen worden bekeken.
- We startten met het model met de hoofdeffecten en met interactietermen.
- Vervolgens werd getest of de interactieterm nul was.
- Indien de interactieterm niet significant was, dan werd deze verwijderd en werden de hoofdeffecten getest.
- Dit lijkt op een achterwaartse modelselectie, maar in dat voorbeeld moeten we de procedure interpreteren als een klassieke hypothesetest (voor interactie) waarvan de nulhypothese in de onderzoeksvraag verscholen lag en dus niet door de data gesuggereerd werd.
- In de oefeningensessies zullen we eveneens een alternatieve aanpak zien waarbij men het interactie-effect niet uit het model verwijderd en gebruik maakt van contrasten om marginale effecten te schatten.

Modelselectie voor predictie

Voorbeeld: *early drug development*

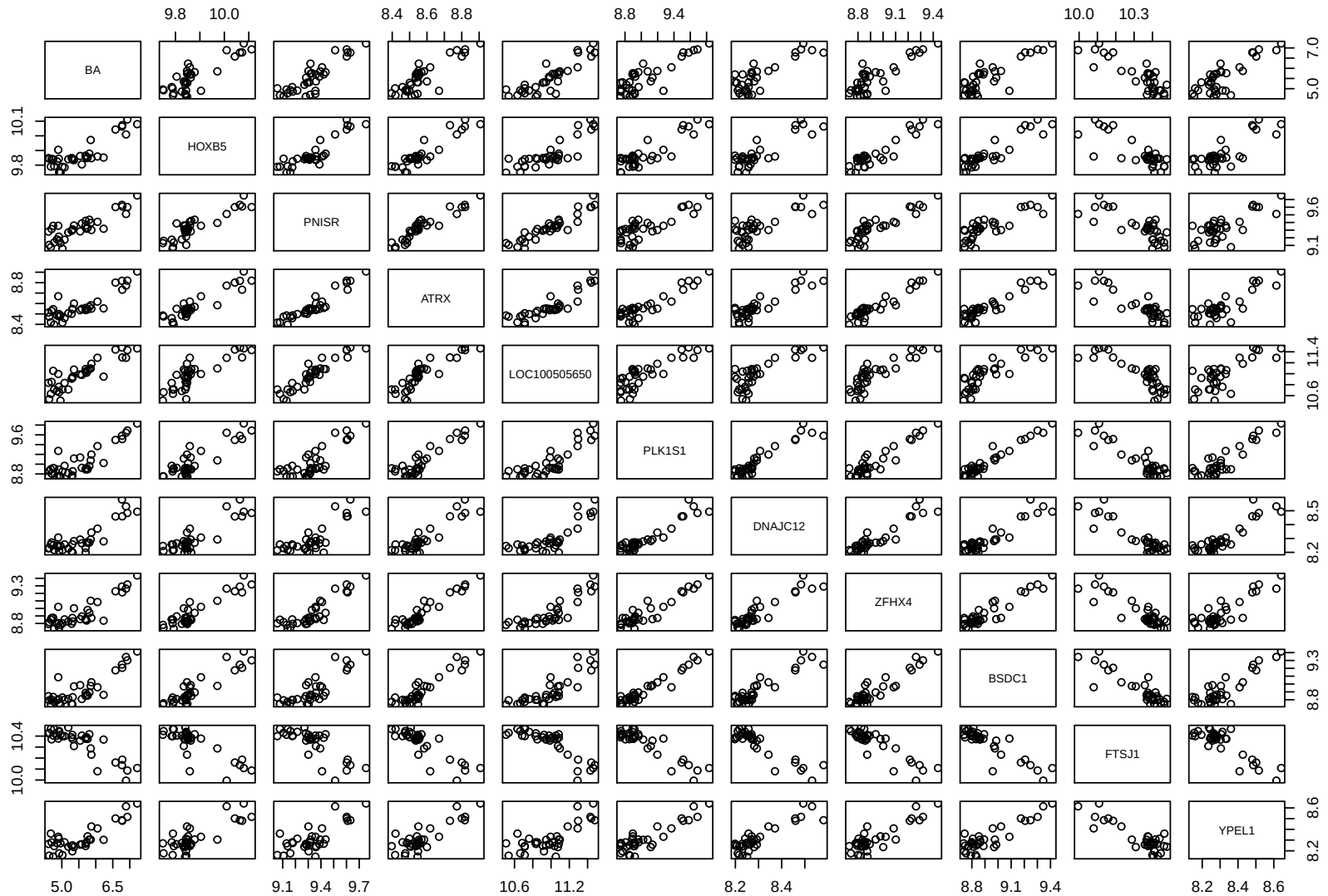
- Effect van de werkzame stof (*compound*) op genexpressie.
- Dieper inzicht geven in werking en/of vroeg toxiciteit detecteren (toxicogenomics).
- Gewenste activiteit bepalen via bio-assay bv. bindingscapaciteit van de compound aan celwandreceptor (target, IC_{50}).
- Vroege fase: 20 tot 50 compounds
- O.b.v. in vitro resultaten beslissen hoe tot een betere compound te komen (hogere on-target activiteit en minder toxiciteit).
- Kleine variaties in moleculaire structuur geven aanleiding tot variaties in BA en genexpressies.
- Doelstelling: opstellen van een model dat toelaat de bioactiviteit (BA) te voorspellen aan de hand van genexpressies in een levercellijn.

- 35 compounds getest op de genexpressie van levercellen.
- Levercellen werden gedurende 24 uur blootgesteld aan de compound, waarna genexpressie gemeten werd m.b.v. microarrays.
- On-target bioactiviteit van iedere compounds via een in-vitro bioassay (IC_{50}).
- We beperken ons hier tot 10 genen (geselecteerd door het onderzoeksteam op basis van a-priori kennis van de pathways betrokken bij de ontwikkeling van de ziekte.

```
mcx<-read.table("dataset/mcx.txt",header=TRUE)
head(mcx)
```

```
##      BA    HOXB5    PNISR     ATRX LOC100505650   PLK1S1  DNAJ
## 1 5.70 9.864744 9.421823 8.557665     11.08555 8.892243 8.198
## 2 5.87 9.847176 9.301424 8.550591     11.18236 9.196701 8.343
## 3 5.70 9.859438 9.360302 8.536606     11.03822 8.903690 8.231
## 4 5.54 9.862448 9.283498 8.535854     10.98804 9.136609 8.292
## 5 5.32 9.836761 9.309442 8.543522     10.94489 8.855404 8.197
## 6 5.73 9.848315 9.318502 8.545635     11.01789 8.919469 8.272
##       BSDC1    FTSJ1    YPEL1
## 1 8.848922 10.38470 8.265589
## 2 9.022556 10.23077 8.426856
## 3 8.888275 10.37585 8.244617
## 4 8.985628 10.35854 8.257426
## 5 8.832436 10.37179 8.250075
## 6 8.864538 10.38892 8.280436
```

plot(mcx)


BA
HOXB5
PNISR
ATRX
LOC100505650
PLK1S1
DNAJC12
ZFHX4
BSDC1
FTSJ1
YPEL1

- $E[Y \mid (X_1, \ldots, X_{p-1})] = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_j$.
- Hypothesetesten voor $H_0 : \beta_j = 0$ gaan dus na of predictor X_j (conditioneel op de andere regressoren in het model) gerelateerd is tot de gemiddelde uitkomst.
- Al dan niet verwerpen van H_0 hangt niet enkel van werkelijke waarde van β_j, maar ook van de kracht van de hypothesetest en steekproefgrootte n.
- Hypothesetesten maken een afweging van de evidentie in de data tegen/voor de gestelde hypotheses die betrekking hebben op de gemiddelde uitkomst.
- Voor een goed predictiemodel is het minder belangrijk dat model de werkelijkheid weerspiegelt
- Het is enkel belangrijk dat het model goede predicties geeft voor individuele uitkomsten.
- Daarom gebruik maken van ander criterium dan statistische hypothese testen

Selectie-criterium

- Aanlokkelijk om R^2 te gebruiken
- Maar hoe meer predictoren er worden opgenomen hoe hoger R^2 wordt:

$$R^2 = 1 - SSE/SSTot = SSR/SSTot$$

- R^2 zou dus steeds leiden tot meest complexe model
- Overfitting: model sluit te goed aan bij de data en veralgemeent niet goed naar nieuwe data

Akaike Information Criterium (AIC)

$$AIC = -2\ln\, L(\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}, \hat{\sigma}^2) + 2(p+1),$$

- Eerste term het natuurlijke logaritme is van likelihood
- Voor een steekproef van n i.i.d. normaal verdeelde observaties met dichtheidsfunctie $f(Y_i|x_{i1} \ldots x_{ip-1}, \beta_0, \ldots, \beta_{p-1}, \sigma^2)$ en conditioneel gemiddelde $E[Y_i|x_{i1} \ldots x_{ip-1}] = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_j$ gedefinieerd is als:

$$L(\beta_0, \ldots, \beta_{p-1}, \sigma^2) = \prod_{i=1}^{n} f(y_i \ldots x_{ip-1}, \beta_0, \ldots, \beta_{p-1}, \sigma^2)$$

- De likelihood geeft dus weer hoe waarschijnlijk het is om de data te observeren onder het normale regressiemodel.
- Voor Normale lineaire regressie model kunnen we aantonen dat $-2\ln\, L(\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}, \hat{\sigma}^2)$ evenredig is met SSE.
- Tweede term straft af voor model complexiteit: p parameters voor gemiddelde + 1 parameter voor variantie (MSE)

$$AIC = -2\ln\ L(\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}, \hat{\sigma}^2) + 2(p+1),$$

- Data analysist zal verschillende kandidaat modellen met elkaar vergelijken.
- Model geselecteren waaronder het zo waarschijnlijk mogelijk is om de geobserveerde data te trekken in een steekproef (eerste term minimaal, SSE zo laag mogelijk), maar zonder dat de modelcomplexiteit daarbij te hoog wordt (tweede term).
- AIC zo laag mogelijk!
- Opnieuw kan men gebruik maken van voorwaarts, achterwaartse of stapsgewijze modelselectie.

Voorwaarts

```
m1 <- lm(BA~1,mcx)
mfull <- lm(BA~.,mcx)
mForward <- step(m1,scope=formula(mfull),direction="forward")
```

```
Start:  AIC=-19.33
BA ~ 1

                Df Sum of Sq     RSS     AIC
+ LOC100505650   1    14.668  4.3588 -68.910
+ PNISR          1    14.376  4.6505 -66.643
+ ATRX           1    14.104  4.9231 -64.650
+ ZFHX4          1    13.869  5.1574 -63.022
+ PLK1S1         1    13.860  5.1671 -62.956
+ FTSJ1          1    13.781  5.2459 -62.426
+ BSDC1          1    13.703  5.3234 -61.913
+ DNAJC12        1    13.412  5.6141 -60.052
+ HOXB5          1    13.400  5.6263 -59.976
+ YPEL1          1    13.049  5.9774 -57.858
<none>                       19.0266 -19.333
```

```
Step:  AIC=-68.91
BA ~ LOC100505650

           Df Sum of Sq    RSS     AIC
+ YPEL1     1   1.20478 3.1540 -78.233
+ DNAJC12   1   0.91062 3.4482 -75.112
+ PLK1S1    1   0.81030 3.5485 -74.109
+ FTSJ1     1   0.78960 3.5692 -73.905
+ BSDC1     1   0.75759 3.6012 -73.592
+ HOXB5     1   0.69933 3.6595 -73.031
+ ZFHX4     1   0.69895 3.6599 -73.027
+ ATRX      1   0.63276 3.7261 -72.400
+ PNISR     1   0.36333 3.9955 -69.956
<none>                 4.3588 -68.910
```

```
Step:  AIC=-78.23
BA ~ LOC100505650 + YPEL1

           Df Sum of Sq    RSS     AIC
<none>                  3.1540 -78.233
+ PNISR     1  0.133032 3.0210 -77.741
+ HOXB5     1  0.116332 3.0377 -77.548
+ FTSJ1     1  0.063065 3.0910 -76.940
+ ATRX      1  0.048977 3.1051 -76.781
+ DNAJC12   1  0.044347 3.1097 -76.729
+ ZFHX4     1  0.029234 3.1248 -76.559
+ BSDC1     1  0.029032 3.1250 -76.557
+ PLK1S1    1  0.025778 3.1283 -76.520
>
```

Achterwaarts

```
mBackward <- step(mfull,direction="backward")
```

```
Start:  AIC=-67.34
BA ~ HOXB5 + PNISR + ATRX + LOC100505650 + PLK1S1 + DNAJC12 +
    ZFHX4 + BSDC1 + FTSJ1 + YPEL1

               Df Sum of Sq    RSS     AIC
- BSDC1         1  0.003369 2.7288 -69.302
- DNAJC12       1  0.005688 2.7312 -69.272
- LOC100505650  1  0.014735 2.7402 -69.156
- PLK1S1        1  0.025044 2.7505 -69.025
- ZFHX4         1  0.030840 2.7563 -68.951
- HOXB5         1  0.039762 2.7652 -68.838
- ATRX          1  0.063271 2.7887 -68.542
- YPEL1         1  0.139909 2.8654 -67.593
- FTSJ1         1  0.157697 2.8832 -67.376
<none>                      2.7255 -67.345
- PNISR         1  0.211967 2.9374 -66.723
```

```
Step:  AIC=-69.3
BA ~ HOXB5 + PNISR + ATRX + LOC100505650 + PLK1S1 + DNAJC12 +
    ZFHX4 + FTSJ1 + YPEL1

                Df Sum of Sq    RSS     AIC
- DNAJC12        1  0.006916 2.7357 -71.213
- LOC100505650   1  0.014255 2.7431 -71.119
- PLK1S1         1  0.025052 2.7539 -70.982
- ZFHX4          1  0.028026 2.7569 -70.944
- HOXB5          1  0.036934 2.7658 -70.831
- ATRX           1  0.087988 2.8168 -70.191
- YPEL1          1  0.136972 2.8658 -69.587
- FTSJ1          1  0.155963 2.8848 -69.356
<none>                      2.7288 -69.302
- PNISR          1  0.228042 2.9569 -68.492
```

```
Step:  AIC=-71.21
BA ~ HOXB5 + PNISR + ATRX + LOC100505650 + PLK1S1 + ZFHX4 + FTSJ
    YPEL1

                Df Sum of Sq    RSS     AIC
- LOC100505650   1  0.014893 2.7506 -73.023
- ZFHX4          1  0.025561 2.7613 -72.887
- PLK1S1         1  0.031313 2.7671 -72.815
- HOXB5          1  0.044757 2.7805 -72.645
- ATRX           1  0.081205 2.8170 -72.189
<none>                      2.7357 -71.213
- YPEL1          1  0.171775 2.9075 -71.082
- FTSJ1          1  0.183246 2.9190 -70.944
- PNISR          1  0.221272 2.9570 -70.491
```

```
Step:  AIC=-73.02
BA ~ HOXB5 + PNISR + ATRX + PLK1S1 + ZFHX4 + FTSJ1 + YPEL1

          Df Sum of Sq    RSS     AIC
- ZFHX4    1   0.02825 2.7789 -74.665
- HOXB5    1   0.04052 2.7912 -74.511
- PLK1S1   1   0.05611 2.8068 -74.316
- ATRX     1   0.13350 2.8841 -73.364
- YPEL1    1   0.15739 2.9080 -73.075
<none>                 2.7506 -73.023
- FTSJ1    1   0.31935 3.0700 -71.178
- PNISR    1   0.90015 3.6508 -65.114
```

```
Step:  AIC=-74.67
BA ~ HOXB5 + PNISR + ATRX + PLK1S1 + FTSJ1 + YPEL1

          Df Sum of Sq    RSS     AIC
- HOXB5    1   0.02391 2.8028 -76.365
- PLK1S1   1   0.03749 2.8164 -76.196
- YPEL1    1   0.14939 2.9283 -74.832
<none>                 2.7789 -74.665
- ATRX     1   0.19488 2.9738 -74.293
- FTSJ1    1   0.29613 3.0750 -73.121
- PNISR    1   0.89851 3.6774 -66.860
```

```
Step:  AIC=-76.37
BA ~ PNISR + ATRX + PLK1S1 + FTSJ1 + YPEL1

          Df Sum of Sq    RSS     AIC
- PLK1S1   1   0.05400 2.8568 -77.698
- YPEL1    1   0.15995 2.9628 -76.423
<none>                 2.8028 -76.365
- ATRX     1   0.17201 2.9748 -76.281
- FTSJ1    1   0.28291 3.0857 -75.000
- PNISR    1   1.01508 3.8179 -67.548
```

```
Step:  AIC=-77.7
BA ~ PNISR + ATRX + FTSJ1 + YPEL1

         Df Sum of Sq    RSS     AIC
- ATRX    1   0.11974 2.9766 -78.260
<none>                2.8568 -77.698
- YPEL1   1   0.26844 3.1252 -76.554
- FTSJ1   1   0.50227 3.3591 -74.029
- PNISR   1   0.96156 3.8184 -69.543

Step:  AIC=-78.26
BA ~ PNISR + FTSJ1 + YPEL1

         Df Sum of Sq    RSS     AIC
<none>                2.9766 -78.260
- YPEL1   1   0.18969 3.1662 -78.098
- FTSJ1   1   0.39421 3.3708 -75.907
- PNISR   1   1.59191 4.5685 -65.266
```

Stepwise

```
mForStep <- step(m1,scope=list(lower=formula(m1),upper=formula(m
```

```
Start:  AIC=-19.33
BA ~ 1

                Df Sum of Sq     RSS     AIC
+ LOC100505650   1    14.668  4.3588 -68.910
+ PNISR          1    14.376  4.6505 -66.643
+ ATRX           1    14.104  4.9231 -64.650
+ ZFHX4          1    13.869  5.1574 -63.022
+ PLK1S1         1    13.860  5.1671 -62.956
+ FTSJ1          1    13.781  5.2459 -62.426
+ BSDC1          1    13.703  5.3234 -61.913
+ DNAJC12        1    13.412  5.6141 -60.052
+ HOXB5          1    13.400  5.6263 -59.976
+ YPEL1          1    13.049  5.9774 -57.858
<none>                       19.0266 -19.333
```

```
Step:  AIC=-68.91
BA ~ LOC100505650

                Df Sum of Sq     RSS     AIC
+ YPEL1          1    1.2048  3.1540 -78.233
+ DNAJC12        1    0.9106  3.4482 -75.112
+ PLK1S1         1    0.8103  3.5485 -74.109
+ FTSJ1          1    0.7896  3.5692 -73.905
+ BSDC1          1    0.7576  3.6012 -73.592
+ HOXB5          1    0.6993  3.6595 -73.031
+ ZFHX4          1    0.6990  3.6599 -73.027
+ ATRX           1    0.6328  3.7261 -72.400
+ PNISR          1    0.3633  3.9955 -69.956
<none>                        4.3588 -68.910
- LOC100505650   1   14.6678 19.0266 -19.333
```

```
Step:  AIC=-78.23
BA ~ LOC100505650 + YPEL1

                Df Sum of Sq    RSS     AIC
<none>                       3.1540 -78.233
+ PNISR          1   0.13303 3.0210 -77.741
+ HOXB5          1   0.11633 3.0377 -77.548
+ FTSJ1          1   0.06306 3.0910 -76.940
+ ATRX           1   0.04898 3.1051 -76.781
+ DNAJC12        1   0.04435 3.1097 -76.729
+ ZFHX4          1   0.02923 3.1248 -76.559
+ BSDC1          1   0.02903 3.1250 -76.557
+ PLK1S1         1   0.02578 3.1283 -76.520
- YPEL1          1   1.20478 4.3588 -68.910
- LOC100505650   1   2.82334 5.9774 -57.858
```

Alternatieve criteria

- Recent zijn binnen machine learning diverse alternatieve technieken voor modelbouw ontwikkeld
- Crossvalidatie, d.i. een techniek waarbij men telkens het regressiemodel schat op basis van een subset van de observaties en haar performantie evalueert door na te gaan hoe goed ze de resterende observaties voorspelt.
- Deze methoden zijn nog beter geschikt voor het bouwen van predictiemodellen.
- Ze evalueren het model immers op data die het model nog niet heeft gezien tijdens de training (fitting) fase.