Context: Assume that we are working for a breeding organisation. We want to develop a new breeding program or improve an existing breeding program. We are interested in including a new trait in our breeding goal. The question is "what are the necessary steps to be able to include a new trait in an existing breeding goal". Examples for such new traits: Mastitis resistence (dairy cattle), fat coverage (beef cattle), more new traits in the future: Ketosis resistence, (dairy cattle), feed efficiency (dairy cattle), ...

Model Selection

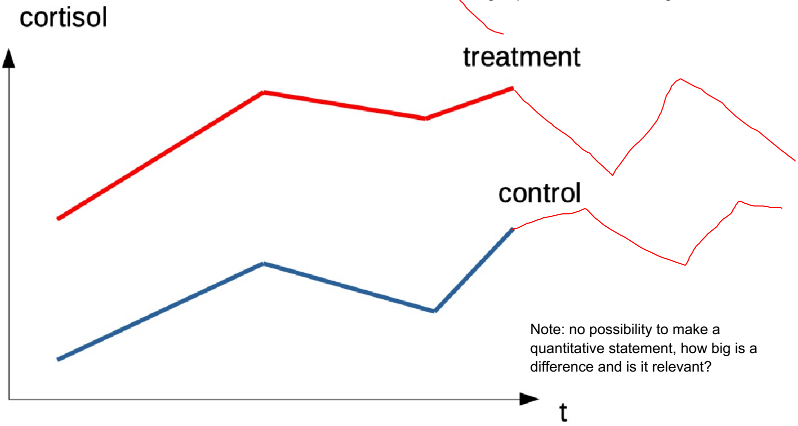
Peter von Rohr

19.04.2021

Why Statistical Modelling?

Some people believe, they do not need statistics. For them it is enough to look at a diagram

group of animals that undergo a certain treatment



Note: no possibility to make a quantitative statement, how big is a difference and is it relevant?

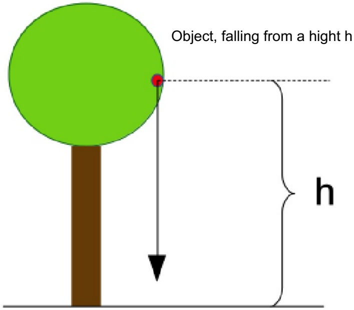
Statistical Modelling Because . . .

Two types of dependencies between physical quantities

1. deterministic
2. stochastic

Deterministic Versus Stochastic

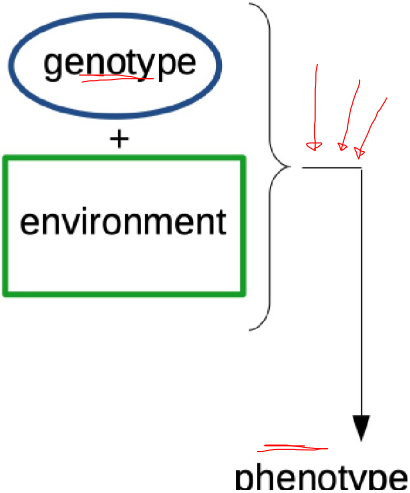
deterministic



Law of gravity

stochastic

account for the different sources of variation that might occur



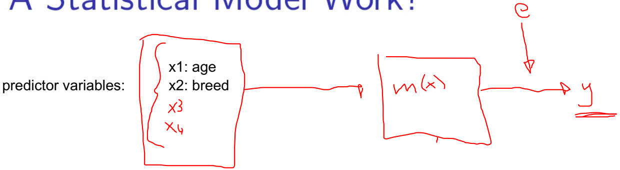
Statistical Model

Stochastic systems can be quantified by statistical models

- stochastic systems contains many sources of uncertainty
- statistical models can handle uncertainty
- components of a statistical model
  - response variable y observations or measurements of a trait from animals
  - predictor variables $x_1, x_2, \ldots, x_k$
    other characteristics or properties from animals (age, sex, herd, season, breed, ...)
  - error term e
  - function $m(x)$

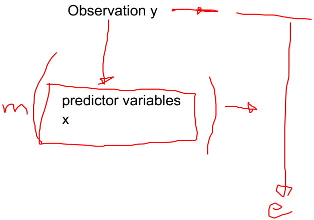
source of uncertainty

How Does A Statistical Model Work?



- predictor variables $x_1, x_2, \ldots, x_k$ are transformed by function $m(x)$ to explain the response variable y
- uncertainty is captured by error term.
- as a formula, for observation i

$$y_i = m(x_i) + e_i$$



Which function $m(x)$?

What should we choose for m(x) to transform the predictors x?

exp(x)

b*x^2

a*x

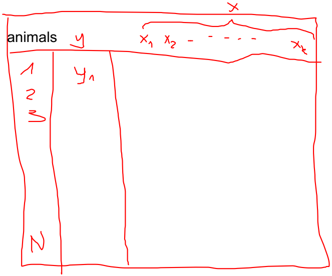
- class of functions that can be used as $m(x)$ is infinitely large
- restrict to linear functions of predictor variables

The answer to the question what is a good choice for m(x) depends on the problem and the nature of the data.

In genetic evaluation: the basic model from quantitative genetics tells that an phenotype is influenced by very many different genes and for the genetic evaluation (prediction breeding values) only the additive effects of a gene are relevant. ==> the linear function suits our problems in an optimal way.

Which predictor variables?

dataset contains all available information



- Question, about which predictor variables to use is answered by model selection

Why Model Selection

- Many predictor variables are available
- Are all of them `relevant`?
- What is the meaning of `relevant` in this context?

Example Dataset

$M_1: y_{BW} = m(x_{BC}) + e$

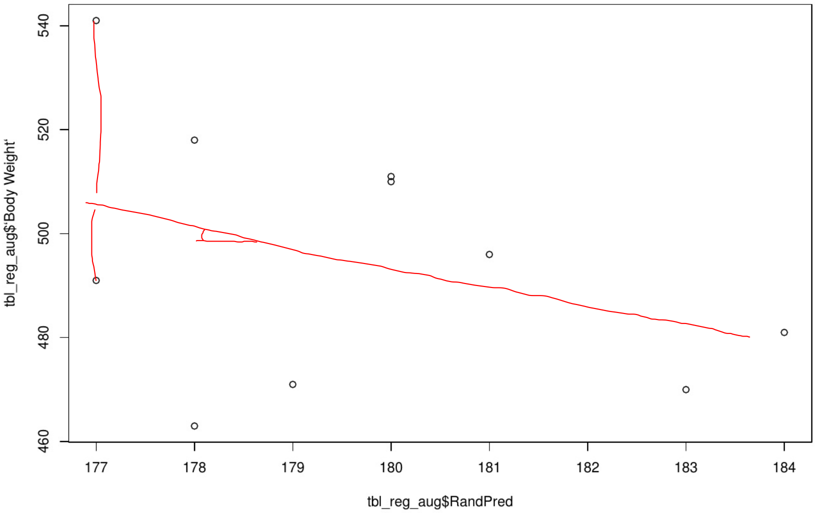
additional predictor

Animal	Breast Circumference	Body Weight	RandPred
1	176	471	179
2	177	463	178
3	178	481	184
4	179	470	183
5	179	496	181
6	180	491	177
7	181	518	178
8	182	511	180
9	183	510	180
10	184	541	177

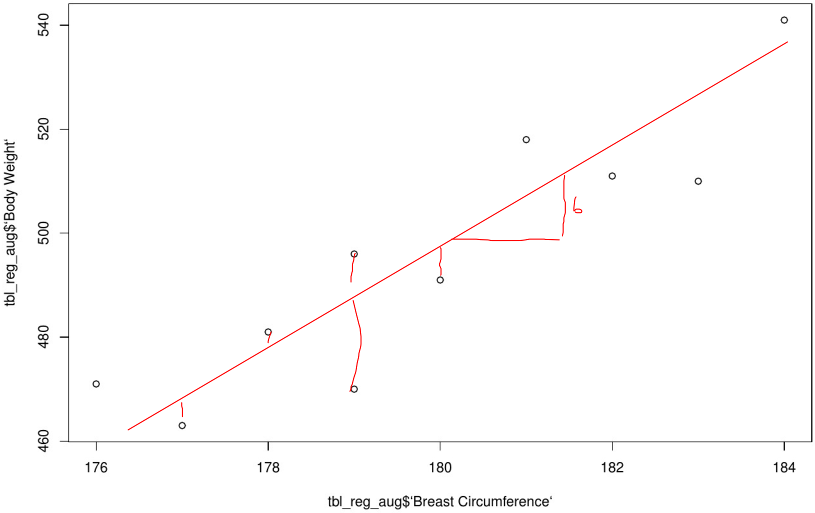
$M_2: y_{BW} = m(x_{BC}, x_{RP}) + e$

If M1 is better than M2 ==> RandPred is not relevant

No Relevance of Predictors



Relevance of Predictors



Fitting a Regression Model

```
##
## Call:
## lm(formula = `Body Weight` ~ RandPred, data = tbl_reg_aug)
##
## Residuals:
##      Min       1Q  Median      3Q     Max
## -39.163 -14.365   4.769  15.981  34.741
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1231.246    602.814   2.042   0.0754 .
## RandPred      -4.096      3.354  -1.221   0.2568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

$\sqrt{\sigma_e^2}$

```
## Residual standard error: 24.21 on 8 degrees of freedom
## Multiple R-squared:  0.1571, Adjusted R-squared:  0.05175
## F-statistic: 1.491 on 1 and 8 DF,  p-value: 0.2568
```

Fitting a Regression Model II

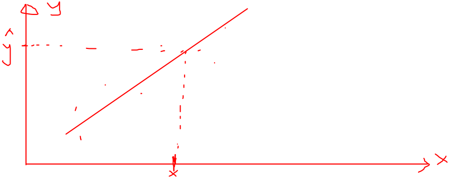
```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference`, data = tbl_reg_aug)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3941  -6.5525  -0.0673   9.3707  13.2594
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1065.115    255.483  -4.169 0.003126 **
## `Breast Circumference`     8.673      1.420   6.108 0.000287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 8 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8014
## F-statistic: 37.31 on 1 and 8 DF,  p-value: 0.000287
```

Multiple Regression

relevant not

```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference` + RandPred,
##     data = tbl_reg_aug)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.778 -10.062   2.941   7.955  11.139
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -721.333    449.542  -1.605 0.152618
## `Breast Circumference`    8.269      1.496   5.529 0.000879 ***
## RandPred                 -1.509      1.617  -0.933 0.381831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.17 on 7 degrees of freedom
## Multiple R-squared:  0.843,  Adjusted R-squared:  0.7981
## F-statistic: 18.79 on 2 and 7 DF,  p-value: 0.001535
```

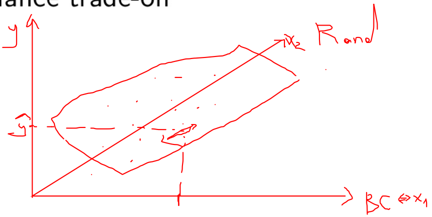
Which model is better?



for a regression model: slope b

Why not taking all predictors?

- Additional parameters must be estimated from data
- Predictive power decreased with too many predictors (cannot be shown for this data set, because too few data points)
- Bias-variance trade-off

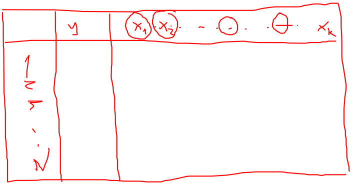


Bias-variance trade-off

From the k available predictors, we select q (q < k)

- Assume, we are looking for optimum prediction

dataset



$$s_i = \sum_{r=1}^{q} \hat{\beta}_{j_r} x_{ij_r}$$

with q relevant predictor variables

- Average mean squared error of prediction s_i

Optimality criterion

$$MSE = n^{-1} \sum_{i=1}^{n} E\left[(m(x_i) - s_i)^2 \right]$$

prediction from q predictor variables

where $m(.)$ denotes the linear function of the unknown true model.

Bias-variance trade-off II

- MSE can be split into two parts

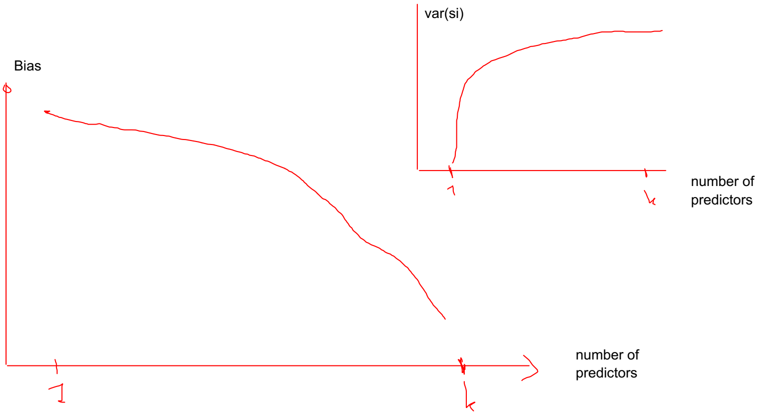
including all k predictors, converge to 0

$$MSE = \overset{1}{\boxed{n^{-1}\sum_{i=1}^{n}\left(E\left[s_i\right] - m(x_i)\right)^2}} + \overset{2}{\boxed{n^{-1}\sum_{i=1}^{n} var(s_i)}}$$

where $n^{-1}\sum_{i=1}^{n}\left(E\left[s_i\right] - m(x_i)\right)^2$ is called the squared **bias** 1

- Increasing q leads to reduced bias but increased variance $(var(s_i))$
- Hence, find s_i such that MSE is minimal
- Problem: cannot compute MSE because $m(.)$ is not known

$\rightarrow$ estimate MSE


Bias
number of predictors
var(si)
number of predictors

Mallows C_p statistic

We do not know the true model (m(x)) ==> MSE cannot be computed exactly. But we want to estimate it from the data.

- For a given model $\mathcal{M}$, $SSE(\mathcal{M})$ stands for the residual sum of squares.
- MSE can be estimated as

the number of predictors included in the model (q)

$$\widehat{MSE} = n^{-1} SSE(\mathcal{M}) - \hat{\sigma}^2 + 2\hat{\sigma}^2 |\mathcal{M}|/n$$

where $\hat{\sigma}^2$ is the estimate of the error variance of the full model, $SSE(\mathcal{M})$ is the residual sum of squares of the model $\mathcal{M}$, n is the number of observations and $|\mathcal{M}|$ stands for the number of predictors in $\mathcal{M}$

number of observations in the dataset

optimal model means that Mallow Cp statistic is as small as possible

$$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\hat{\sigma}^2} - n + \underline{2|\mathcal{M}|}$$

Searching The Best Model

- Exhaustive search over all sub-models might be too expensive
- For p predictors there are $2^p - 1$ sub-models
- With $p = 16$, we get 6.5535×10^4 sub-models

$\rightarrow$ step-wise approaches

Forward Selection

y	$x_1, x_2, \ldots, x_k$

1. $y = x_0 + e \Rightarrow SSE(M_0)$

empty model, model with just an intercept

2. M_1: $y = x_0 + x_1 + e \Rightarrow SSE$
$y = x_0 + x_2 + e \Rightarrow SSE$
$\vdots$
$y = x_0 + x_k + e \Rightarrow SSE$

min $\Rightarrow M_1$

1. Start with smallest sub-model $\mathcal{M}_0$ as current model
2. Include predictor that reduces SSE the most to current model
3. Repeat step 2 until all predictors are chosen

$\rightarrow$ results in sequence $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \ldots$ of sub-models

4. Out of sequence of sub-models choose the one with minimal C_p

For k predictor variables: $M_0, M_1, M_2, \ldots, M_k$

$\downarrow$ C_p C_p C_p C_p

min $\rightarrow$ optimal model

Backward Selection

Full model: containing all k predictors

1: $\mathcal{M}_0$: $y = x_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

2. $\mathcal{M}_1$: $y = x_0 + \beta_2 x_2 + \ldots \beta_k x_k \rightarrow SSE$

$y = x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k$ SSE $\rightarrow \mathcal{M}_1$

1. Start with full model $\mathcal{M}_0$ as the current model
2. Exclude predictor variable that increases SSE the least from current model
3. Repeat step 2 until all predictors are excluded (except for intercept)

$\rightarrow$ results in sequence $\mathcal{M}_0 \supseteq \mathcal{M}_1 \supseteq \mathcal{M}_2 \supseteq \ldots$ of sub-models

4. Out of sequence choose the one with minimal C_p

Considerations

- Whenever possible, choose **backward** selection, because it leads to better results
- If $p \geq n$, only forward is possible, but then consider LASSO

Alternative Selection Criteria

When comparing two models, so far, we have used Mallow Cp

AIC: Akaike Information Criterion
BIC: Bayes Information Criterion

- AIC or BIC, requires distributional assumptions.
- AIC is implemented in `MASS::stepAIC()`
- Adjusted R^2 is a measure of goodness of fit, but sometimes is not conclusive when comparing two models
- Try in exercise